Probability Space

$(\Omega, \mathcal{F}, \Pr)$

probability measure on Ω

Sample space

collection of events I want to assign probabilities to

level of resolution we observe our random experiment at

Axioms

$\boxed{\mathcal{F}}$

(i) $\Omega \in \mathcal{F}$

(ii) If $A \in \mathcal{F}$, $A^c \in \mathcal{F}$

(iii) If $A_i \in \mathcal{F}$, $\bigcup_i A_i \in \mathcal{F}$ ← countable intersections also there! $A \cap B = (A^c \cup B^c)^c$!

[illegible] countable union to belo to inf

$\boxed{\Pr}$ $\mathcal{F} \longrightarrow [0,1]$

(i) $0 \le \Pr(A) \le 1$

(ii) $\Pr(\Omega) = 1$

(iii) If $A_1, A_2, \dots,$ is a sequence of mutually exclusive events then $\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$

if one event occurs the other events do not

Properties that follow

(i) $\Pr(A^c) = 1 - \Pr(A)$

Proof: $\Pr(A \cup A^c) = \Pr(A) + \Pr(A^c)$

$\Pr(\Omega) = 1 = \Pr(A) + \Pr(A^c)$

$\Pr(A^c) = 1 - \Pr(A)$

(ii) $\Pr(\emptyset) = 0$, $\emptyset = \Omega^c$

(iii) If $A \subseteq B$ then $\Pr(A) \le \Pr(B)$

Proof: $B = A \cup (A^c \cap B)$

$\Pr(B) = \Pr(A) + \Pr(A^c \cap B) \ge \Pr(A)$

(iv) $\Pr(A \cup B) = \Pr(A) - \Pr(B) - \Pr(A \cap B)$

Union Bound

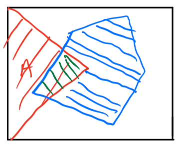
$\Pr(\bigcup_i A_i) \le \sum_{i=1}^{n} \Pr(A_i)$

Conditional Probability

If A and B are two events, $\Pr(B) \neq 0$, then

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

"B becomes the new universe, $B = \Omega'$"



Total Probability Theorem

If $\{E_1, E_2, \dots, E_k\}$ partition Ω then

$$\Pr(A) = \sum_{i=1}^{k} \Pr(A \cap E_i)$$

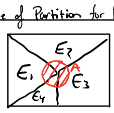
$$= \sum_{i=1}^{k} \Pr(A|E_i)\Pr(E_i)$$

This makes more sense since you know base event which occurred

generative models

Example of Partition for $k=4$

Ω



No intersections $\bigcup_{i=1}^{k} E_i = \Omega$

Bayes Rule

Want to find "ground truth" that gave this observation.

$$\Pr[E_j|A] = \frac{\Pr[A \cap E_j]}{\Pr[A]} = \frac{\Pr[E_j]\Pr(A|E_j)}{\sum_{i=1} \Pr[E_i]\Pr[A|E_i]}$$

prior knowledge, generative model

← $\Pr(A)$ by TPT

Independence

Two events, A_1, A_2, are independent if $\Pr(A_1 \cap A_2) = \Pr(A_1)\Pr(A_2)$

$$\Longrightarrow \Pr[A_1|A_2] = \frac{\Pr[A_1 \cap A_2]}{\Pr[A_2]} = \frac{\Pr(A_1)\Pr(A_2)}{\Pr(A_2)} = \Pr(A_1)$$

Can be extended:

Events $\{A_1, A_2, \dots, A_n\}$ are independent if $\Pr(A_1 \cap A_2 \cap \dots \cap A_k) = \Pr(A_1)\Pr(A_2)\cdots\Pr(A_k)$ $\forall \, k \in [2,n], k \in \mathbb{Z}$

$$F_{X_1,\dots X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

$$\Pr[X_1 \le x_1, \dots, X_n \le x_n] = \Pr[X_1 \le x_1] \cdots \Pr[X_n \le x_n]$$

Random Variables

Given $(\Omega, \mathcal{F}, \Pr)$ a random variable is a function $X: \Omega \to \mathbb{R}$ such that $\forall x \in \mathbb{R}$, $\{\omega \mid X(\omega) \le x\} \in \mathcal{F} \subseteq \Omega$

N^{th} Moment

$$\mathbb{E}[X^n] = \int_{-\infty}^{+\infty} x^n f_X(x)\,dx$$

N^{th} Central Moment

$$\mathbb{E}[(X - \mathbb{E}[X])^n]$$

Cumulative Distribution Function

The CDF is defined as

$$F_X(x) = \Pr(X \le x) \quad \forall x \in \mathbb{R}$$
$$= \Pr(\{\omega \mid X(\omega) \le x\})$$

Properties of CDF

1. $\lim_{x \to \infty} F_X(x) = 0$

   $\lim_{x \to -\infty} F_X(x) = 1$

2. $\forall x < y \quad F_X(x) \le F_X(y)$

3. F is right continuous

   $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$

4. $\Pr[x \le X \le y] = F_X(y) - F_X(x)$

PMF for Discrete Random Variables

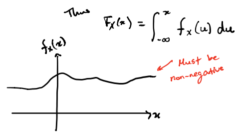
$p_X(x) = \Pr[X = x]$

$F_X(x) = \sum_{u \le x} p_X(u)$

PDF for Continuous Random Variables

$f_X(x) = \frac{dF_X(x)}{dx}$ } measures how fast we accumulate probability

Thus $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$



Properties of PDF

1. $f_X(x) \ge 0 \quad \forall x$
2. $\int_{-\infty}^{+\infty} f_X(x) = 1 = F_X(\infty)$
3. $\Pr[x \le X \le y] = \int_x^y f_X(u)\,du$

Variance

How much a r.v varies from its expectation

$$\mathrm{Var}(X) = \mathbb{E}[\overbrace{(X - \mathbb{E}[X])^2}^{g(x)}]$$

$$= \mathbb{E}[X^2 - 2\mathbb{E}[X] + (\mathbb{E}[X])^2]$$

$$= \mathbb{E}[X^2] + (\mathbb{E}[X])^2$$

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} g(x) f_X(x)\,dx = \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 f_X(x)\,dx$$

Correlation

Correlation between X and Y

$$\mathbb{E}[XY] = \int_{y=-\infty}^{y=+\infty}\int_{x=-\infty}^{x=+\infty} xy\, f_{XY}(x,y)\,dx\,dy$$

$$\mathbb{E}[XY] = \sum_{x_i}\sum_{y_j} x_i y_j P_{X,Y}(x_i, y_j)$$

Covariance

$$\mathrm{Cov}(X,Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

If $\mathbb{E}[XY] = 0$, we say X is orthogonal to Y.

If $\mathrm{Cov}(X,Y) = 0$, we say X is uncorrelated to Y.

Properties

Independence of r.v's X, Y

$f_{XY}(x,y) = f_X(x) f_Y(y) \Rightarrow$ uncorrelatedness

X, Y, Z r.v.'s

$$\mathrm{Cov}(X, aY + bZ) = a\mathrm{Cov}(X,Y) + b\mathrm{Cov}(X,Z)$$

$$\mathrm{Cov}(X - \mathbb{E}[X], Y - \mathbb{E}[Y]) = \mathrm{Cov}(X,Y)$$

Expectation

$$\mathbb{E}[X] = \begin{cases} \sum_k k \Pr[X = k], & X \text{ discrete} \\ \int_{-\infty}^{+\infty} x f_X(x)\,dx, & X \text{ continuous} \end{cases}$$

Properties

LOTUS Rule

If

$$Y = g(X)$$

$f_X(x)$ g

then

$$\mathbb{E}[Y] = \int_{-\infty}^{+\infty} x f_Y(y)\,dy$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\,dx$$

Linearity of Expectation

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y], \quad \alpha, \beta \in \mathbb{R}$$

Preservation of Order

If

$$\Pr[X \ge Y] = 1$$

then

$$\mathbb{E}[X] \ge \mathbb{E}[Y]$$

Integration by Parts

A discrete, non-negative r.v. $X = 0, 1, 2, 3, \dots$

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i \Pr(X = i)$$

$$= \sum_{i=0}^{\infty} \underbrace{\Pr(X > i)}_{1 - F_X(i)} \quad \leftarrow \text{Tail Probabilit.}$$

To see this, observe

$\sum_{i=0}^{\infty} i \Pr(X = i) = 0 \cdot \Pr(X=0) + 1 \cdot \Pr(X=1) + 2 \cdot \Pr(X=2) + \cdots$

$\sum_{i=0}^{\infty} \Pr(X > i)$ $i=0$ ✓ ✓, $i=1$ ✓, ⋮ etc

Thus

$$\mathbb{E}[X] = \int_0^{\infty} (1 - F_X(x))\,dx - \int_{-\infty}^{0} F_X(x)\,dx$$

Conditioning on Random Variables

Suppose X and Y have joint pmf $p_{XY}(x,y)$

The conditional pmf of X given $\{Y = y\}$ is

$$p_{X|Y}(x|y) \triangleq \Pr[\overset{A}{X = x} \mid \overset{B}{Y = y}]$$

$$= \frac{\Pr(\{X = x\} \cap \{Y = y\})}{\Pr(\{Y = y\})} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Similarly for continuous r.v.'s

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \quad \text{if } f_Y(y) \neq 0 \; \forall y$$

Conditional Expectation

$$\mathbb{E}[X|Y=y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)\,dx$$

$\mathbb{E}[X|Y]$ is a random variable which takes on value $\mathbb{E}[X|Y=y]$ w/ density $f_Y(y)$.

Since $\mathbb{E}[X|Y]$ is r.v. can take its expectation.

$$\mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]] = \mathbb{E}[X]$$

Cauchy-Schwartz Inequality

$$\mathbb{E}[XY] \le \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

Correlation Coefficient

normalization comes from Cauchy-Schwarz inequality

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \qquad (|\rho_{X,Y}| \le 1)$$

If $X_1, \dots, X_m$ are pairwise uncorrelated, then

$$Var(X_1+X_2+\dots+X_m) = Var(X_1)+Var(X_2)+\dots+Var(X_m)$$

$$Var\left(\sum_{i=1}^m X_i\right) = \sum_{i=1}^m Var(X_i)$$

Linear Minimum Mean Squared Error

Now our estimator $g(Y)$ takes the form

$$g(Y) = aY+b$$

The MSE becomes

$$MSE = \mathbb{E}[(X-g(Y))^2] = \mathbb{E}[(X-aY-b)^2]$$

Want

$$\{a^*, b^*\} = \underset{a,b}{\operatorname{argmin}}\ \mathbb{E}[(X-aY-b)^2]$$

Use orthogonality principle

$$W \perp g(Y) \Rightarrow X-a^*Y-b^* \perp aY+b \quad \forall a,b$$

Consider the following cases

① $a=0, b=1$

$$X-a^*Y-b^* \perp 1 \Rightarrow \mathbb{E}[X-a^*Y-b^*]=0$$

$\langle X,Y \rangle \triangleq \mathbb{E}[XY]$

$$\mathbb{E}[X]-\mathbb{E}[a^*Y]-\mathbb{E}[b^*]=0$$

$W \Rightarrow \mathbb{E}[W]=0 \Rightarrow$ UNBIASED Estimator

$$\rightarrow b^* = \mathbb{E}[X] - a^*\mathbb{E}[Y]$$

② $a=1, b=0$

$$X-a^*Y-b^* \perp Y \Rightarrow \mathbb{E}[(X-a^*Y-b^*)Y]=0$$

$$\mathbb{E}[(X-a^*Y-b^*)Y]$$

$$= \mathbb{E}[(X-a^*Y-\mathbb{E}[X]-a^*\mathbb{E}[Y])Y]$$

first term W

$$= \mathbb{E}[(X-\mathbb{E}[X]-a^*(Y-\mathbb{E}[Y]))Y]$$

$$\Rightarrow Cov((X-\mathbb{E}[X]-a^*(Y-\mathbb{E}[Y])), Y)=0$$

Note) $Cov(W,Y) = \mathbb{E}[WY]$ here

$$Cov(X,Y) - a^*Cov(Y,Y) = 0$$

$$a^* = Cov(X,Y)/Var(Y) \leftarrow Cov(Y,Y)$$

Aside

$$Cov(X, aY+bZ) = aCov(X,Y) + bCov(X,Z)$$

$$Cov(X-\mathbb{E}[X], Y-\mathbb{E}[Y]) = Cov(X,Y)$$

So

$$\hat{X}^*_{LMMSE} = a^*Y+b^*$$

$$= \frac{Cov(X,Y)}{Var(Y)}Y + \mathbb{E}[X] - \frac{Cov(X,Y)}{Var(Y)}\mathbb{E}[Y]$$

$$= \frac{Cov(X,Y)}{Var(Y)}(Y-\mathbb{E}[Y]) + \mathbb{E}[X]$$

and the MSE is

$$MSE = \mathbb{E}[W^2] = Cov(W,W)$$

$$= Cov((X-\mathbb{E}[X]-a^*(Y-\mathbb{E}[Y])), (X-\mathbb{E}[X]-a^*(Y-\mathbb{E}[Y])))$$

$$= Cov(X,X) + a^{*2}Cov(Y,Y) - 2a^*Cov(X,Y)$$

$$= Var(X) + \left(\frac{Cov(X,Y)}{Var(Y)}\right)^2 Var(Y) - 2\left(\frac{Cov(X,Y)}{Var(Y)}\right)Cov(X,Y)$$

$$= Var(X) - \frac{(Cov(X,Y))^2}{Var(Y)}$$

$$W = X - \hat{X}_{MMSE}$$

$$= X - \left(\mathbb{E}[X] + \frac{Cov(X,Y)}{Var(Y)}(Y-\mathbb{E}[Y])\right)$$

$$= X - \frac{Cov(X,Y)}{Var(Y)}Y - \left(\mathbb{E}[X] + \frac{Cov(X,Y)}{Var(Y)}\mathbb{E}[Y]\right)$$

Know

W & Y are Jointly Gaussian

$\mathbb{E}[W]=0$

$\mathbb{E}[WY]=0$ (orthogonality)

So $Cov(W,Y)=0 \Rightarrow$ independence by property 2

Need W is error given by best MMSE

$\mathbb{E}[Wg(Y)]=0 \quad \forall g(Y) \Rightarrow$

Trivially true since W independent of Y.

$$\mathbb{E}[Wg(Y)] = \mathbb{E}[W]\mathbb{E}[g(Y)] = 0$$

Proof of ④

Hilbert Space

A Hilbert space is a vector space that

① Has an inner product $\langle \cdot, \cdot \rangle$ defined

② Is complete w/ respect to the norm

$\|x\| = \sqrt{\langle x, x \rangle}$ induced by $\langle \cdot, \cdot \rangle$

→ complete in the sense every Cauchy sequence converges

Hilbert Space of Random Variables

All random variables w/ finite 2nd moments form a Hilbert space w/ inner product (r.v.'s NOT vectors)

$$\langle X, Y \rangle \triangleq \mathbb{E}[XY]$$

Inner Product ⇒ norm/length ⇒ distance metric

$$\text{length/norm of } X = \sqrt{\langle X, X \rangle} = \sqrt{\mathbb{E}[X^2]}$$

Angle b/t X, Y; θ_{XY}:

$$\cos(\theta_{XY}) = \frac{\langle X, Y \rangle}{\|X\|\|Y\|} = \frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}}$$

Cauchy-Schwarz ⇒ this ≤ 1

Projection of X onto Y

$$\Pi_Y(X) = \frac{\langle X, Y \rangle}{\|Y\|^2}Y = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y$$

a random variable

Orthogonality

$$\langle X, Y \rangle \triangleq \mathbb{E}[XY] = 0$$

$$\Pi_Y(X) = 0$$

Pythagorean

all X_i pairwise orthogonal

$$\mathbb{E}\left[\left(\sum_{i=1}^m X_i\right)^2\right] = \sum_{i=1}^m \mathbb{E}[X_i^2]$$

Jointly Gaussian Random Variables

If

$$X \sim N(\mu, \sigma^2)$$

then

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

Note) K is our covariance matrix.

Here,

$$K = \begin{pmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{pmatrix} = \mathbb{E}\left[\begin{pmatrix} X-\mu_X \\ Y-\mu_Y \end{pmatrix}\begin{pmatrix} X-\mu_X \\ Y-\mu_Y \end{pmatrix}^T\right] = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Definition: X, Y are jointly Gaussian if $aX+bY$ is Gaussian $\forall a, b \in \mathbb{R}$.

Definition: Joint pdf of such a distribution is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{|K|}} \exp\left(-\frac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}^T K^{-1}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}\right)$$

Properties of Jointly Gaussian Random Variables

① J.G. $\Rightarrow$ M.G. ($\not\Leftarrow$)

② Uncorrelated JG rv's are independent ← special!

independence ⇒ uncorrelated

$\not\Leftarrow$ in general

③ If X,Y jointly Gaussian, then

$a_1X+b_1Y+c_1$, $a_2X+b_2Y+c_2$ are jointly Gaussian

④ For JG X,Y

$$\hat{X}_{MMSE} = \mathbb{E}[X|Y] = \hat{X}_{LMMSE} = \mathbb{E}[X] + \frac{Cov(X,Y)}{Var(Y)}(Y-\mathbb{E}[Y])$$

$$MMSE = Var(X)(1-\rho_{XY}^2) = Var(X) - \frac{Cov^2(X,Y)}{Var(Y)}$$

⑤ For J.G X,Y

$$\mathbb{E}[X|Y=y] = \mathbb{E}[X] + \frac{Cov(X,y)}{Var(y)}(y-\mathbb{E}[y])$$

$$\mathbb{E}[X|Y] = \mathbb{E}[X] + \frac{Cov(X,Y)}{Var(Y)}(Y-\mathbb{E}[Y])$$

$$Var(X|Y=y) = Var(X) - \frac{Cov^2(X,y)}{Var(y)}$$

$$Var(X|Y) = Var(X) - \frac{Cov^2(X,Y)}{Var(Y)}$$

⑥ The conditional distribution of X given Y=y is Gaussian

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi Var(X|Y)}} e^{-(\dots)}$$

Proof) $X = \hat{X}_{MMSE} + W$

where W is independent of Y. Given Y=y (an observation),

- $\hat{X}_{MMSE}$ is a constant
- W maintains its distribution

Thus $X \sim N(\mu_W + \hat{X}_{MMSE}, \sigma_W^2)$

Mean Squared Error (MSE)

Have r.v. X which can't be directly observed. Want to estimate $\hat{X}$.

Note: we are in a Hilbert space of random variables

The estimation error

$$W = X - \hat{X}$$

The MSE of $\hat{X}$ is

$$\mathbb{E}[(X-\hat{X})^2] = \mathbb{E}[W^2] = \langle W, W \rangle = \|W\|^2 = (d(X,\hat{X}))^2$$

Minimum Mean Squared Error

$$\hat{X}_{MMSE} = \underset{\hat{X}}{\operatorname{argmin}}\ \mathbb{E}[(X-\hat{X})^2]$$

MMSE of X using a constant

$$\hat{X}_{MMSE} = a^*$$

where

$$a^* = \underset{a \in \mathbb{R}}{\operatorname{argmin}}\ \mathbb{E}[(X-a)^2] \quad \Big\}\ \text{it's } \mathbb{E}[X]!$$

$$MSE = \mathbb{E}[(X-a)^2]$$

$$= \mathbb{E}[(X-\mathbb{E}[X]+\mathbb{E}[X]-a)^2]$$

$$= \mathbb{E}[(X-\mathbb{E}[X])^2] + \underbrace{\cancel{2\mathbb{E}[(X-\mathbb{E}[X])(\mathbb{E}[X]-a)]}}_{0} + \mathbb{E}[(\mathbb{E}[X]-a)^2]$$

Want to minimize

$$\mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(\mathbb{E}[X]-a)^2]$$

$$\rightarrow Var(X) + (\mathbb{E}[X]-a)^2$$

Take $a = \mathbb{E}[X]$

MMSE of X given Y=y — similar proof

$$\hat{X}_{MMSE} = \mathbb{E}[X|Y]$$

$$MSE(\hat{X}_{MMSE}) = \mathbb{E}[Var(X|Y)]$$

The error

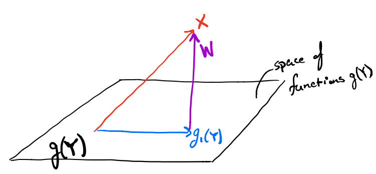
$$W = X - \mathbb{E}[X|Y]$$

has

$$\mathbb{E}[W] = \mathbb{E}[X-\mathbb{E}[X|Y]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$$

we call this type of estimator UNBIASED

VISUALIZATION



$X, Y, \hat{X}$

$W = X - \hat{X}^*$

Want $\min_{\hat{X}} \mathbb{E}[W^2]$

- Orthogonality principle:

$$W \perp g(Y) \quad \forall g \quad \Big\}\ \text{necessary AND sufficient condition}$$

W induced by $\hat{X}^*$ (best estimator)

Claim $X - \mathbb{E}[X|Y] \perp g(Y) \quad \forall g$

Proof

$$\mathbb{E}[(X-\mathbb{E}[X|Y])g(Y)] = 0$$

inner product $\langle X-\mathbb{E}[X|Y], g(Y) \rangle$

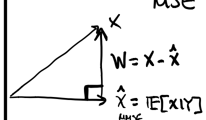
LHS

$$\mathbb{E}_{XY}[Xg(Y)] - \mathbb{E}_Y[(\mathbb{E}_{X|Y}[X|Y])g(Y)]$$

$$\mathbb{E}_Y[\mathbb{E}_{X|Y}[Xg(Y)|Y]] - \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]g(Y)]$$

$$\mathbb{E}_Y[g(Y)\mathbb{E}_{X|Y}[X|Y]] - \mathbb{E}_Y[g(Y)\mathbb{E}_{X|Y}[X|Y]] = 0 \checkmark$$

MSE Using Pythagorean Theory



$\|W\|^2 = \langle W, W \rangle$ in Hilbert Space

$$MSE = \mathbb{E}[W^2] = \mathbb{E}[X^2] - \mathbb{E}[\hat{X}^2]$$

$$= Var(X) + (\mathbb{E}(X))^2 - (Var(\hat{X}_{MMSE}) + (\mathbb{E}[\hat{X}_{MMSE}])^2)$$

$\mathbb{E}[X] = \mathbb{E}[\hat{X}]$

$$= Var(X) - Var(\hat{X}_{MMSE})$$