

1. Statistical Learning

1.1. Definition Statistical Model

Statistical Model:	$\{\mathbb{X}, \mathbb{F}, P_\theta; \theta \in \Theta\}$
Sample Space:	Ω
Observation Space:	\mathbb{X}
Sigma Algebra:	\mathbb{F}
Probability:	P_θ
Test:	$T: \mathbb{X} \mapsto \{\theta_0, \theta_1\}, x \mapsto T(x)$
Null Hypothesis:	$H_0: \theta \in \Theta_0$
Alternative Hypothesis:	$H_1: \theta \in \Theta_1$

Cost Criterion G_T :
 $G_T: \{\theta_0, \theta_1\} \mapsto [0, 1], \theta \mapsto P(\{T(X) = 1\}|\theta)$
 $= E[T(X); \theta] = \int T(x)f_X(x|\theta) dx$

Error Level α : $G_T(\theta_0) \leq \alpha$
Two Error Types:
 False Alarm: $\theta = \theta_0, T(x) = 1$
 $G_T(\theta_0) = P(\{T(X) = 1\}|\theta_0)$
 Detection Error: $\theta = \theta_1, T(x) = 0$
 $1 - G_T(\theta_1) = P(\{T(X) = 0\}|\theta_1)$

1.2. Maximum Likelihood Test

ML Ratio Test Statistic:

$$R(x) = \begin{cases} \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} & ; f_X(x|\theta_0) > 0 \\ \infty & ; f_X(x|\theta_0) = 0 \text{ and } f_X(x|\theta_1) > 0 \end{cases}$$

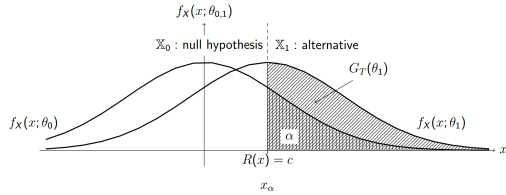
ML Test:
 $T_{ML}: \mathbb{X} \mapsto \{0, 1\}, x \mapsto \begin{cases} 1 & ; R(x) > c \\ 0 & ; \text{otherwise} \end{cases}$
 if $c \neq 1$ False Alarm Error Probability can be adjusted \rightarrow Neyman Pearson Test

1.3. Neyman-Pearson-Test

The best test of P_0 against P_1 is

$$T_{NP}(x) = \begin{cases} 1 & R(x) > c \\ \gamma & R(x) = c \\ 0 & R(x) < c \end{cases} \quad \text{Likelihood-Ratio: } R(x) = \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)}$$

$\gamma = \frac{\alpha - P_0(\{R > c\})}{P_0(\{R = c\})}$ Errorlevel α
 Steps: For α calculate x_α , then $c = R(x_\alpha)$



Maximum Likelihood Detector: $T_{ML}(x) = \begin{cases} 1 & R(x) > 1 \\ 0 & \text{otherwise} \end{cases}$

ROC Graphs: plot $G_T(\theta_1)$ as a function of $G_T(\theta_0)$

1.4. Bayes Test (MAP Test)

Prior knowledge about possible hypotheses:
 $P(\{\theta \in \Theta_0\}) + P(\{\theta \in \Theta_1\}) = 1$

$$T_{\text{Bayes}} = \underset{T}{\operatorname{argmin}} \{P_\epsilon\} = \begin{cases} 1 & ; \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > c \\ 0 & ; \text{otherwise} \end{cases}$$

$= \begin{cases} 1 & ; P(\theta_1|x) > P(\theta_0|x) \\ 0 & ; \text{otherwise} \end{cases}$
 with:
 $P_\epsilon = P(\theta_0)G_T(\theta_0) + P(\theta_1)(1 - G_T(\theta_1)), \quad c = \frac{P(\theta_0)}{P(\theta_1)}$

if $P(\theta_0) = P(\theta_1) \rightarrow T_{\text{Bayes}} = T_{ML}$

Multiple Hypothesis $\{\theta_0, \dots, \theta_k\}; \mathbb{X}_0, \dots, \mathbb{X}_k \in \mathbb{X}$:

$$T_{\text{Bayes}} = \underset{k \in 1, \dots, K}{\operatorname{argmin}} \{P(\theta_k|x)\}$$

Loss Function:

$$L(T(x), \theta) = \begin{cases} L_0 & ; T(x) = 1, \text{ but } \theta = \theta_0 \quad (\text{FALSE ALARM}) \\ L_1 & ; T(x) = 0, \text{ but } \theta = \theta_1 \quad (\text{DETEC. ERROR}) \\ 0 & ; \text{otherwise} \end{cases}$$

L_i denotes the Loss Value in cases where the correct decision parameter θ_i is missed.
 $\text{Risk}(T) = E[L(T(X), \theta)] = E[E[L(T(x), \theta)|x = X]]$

1.5. Linear Alternative Tests

Estimate normal vector \underline{w}^\top and w_0 , which separate \mathbb{X} into \mathbb{X}_0 and \mathbb{X}_1
 $\log R(\underline{x}) = -\frac{1}{2} \ln \left(\frac{\det(\underline{C}_1)}{\det(\underline{C}_0)} \right) - \frac{1}{2} (\underline{x} - \underline{\mu}_1)^\top \underline{C}_1^{-1} (\underline{x} - \underline{\mu}_1) + \frac{1}{2} (\underline{x} - \underline{\mu}_0)^\top \underline{C}_0^{-1} (\underline{x} - \underline{\mu}_0) = \ln \left(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)} \right)$ (separating surface)

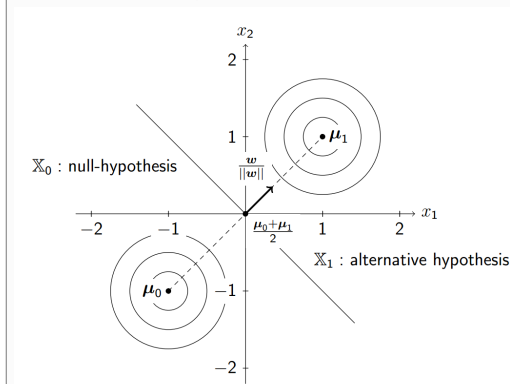
For Gaussian $f_X(x; \mu_k, C_k)$ with θ_0 and θ_1 corresponding to $\{\mu_0, C_0\}$ and $\{\mu_1, C_1\}$, it follows that

- if $C_0 \neq C_1$, $\log R(x) = 0$ is non-linear and the separating surfaces are surfaces of second order: parabolic, hyperbolic, or elliptic surfaces.
- if $C_0 = C_1$, $\log R(x) = 0$ is affine and thus defines a hyperplane in \mathbb{X} which decomposes \mathbb{X} into \mathbb{X}_0 and \mathbb{X}_1 , i.e.,

$$T: \mathbb{X} \rightarrow \mathbb{R}, \underline{x} \mapsto \begin{cases} 1 & \underline{w}^\top \underline{x} > w_0 \\ 0 & \text{otherwise} \end{cases}$$

- case 1: $\underline{C}_0 = \underline{C}_1 = \sigma^2 \underline{I}_N$
 $\underline{w}^\top = (\underline{\mu}_1 - \underline{\mu}_0)^\top$,
 $w_0 = \frac{1}{2} (\underline{\mu}_1^\top \underline{\mu}_1 - \underline{\mu}_0^\top \underline{\mu}_0) - \sigma^2 \ln \left(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)} \right)$
 \underline{w} colinear with $(\underline{\mu}_1 - \underline{\mu}_0)$
 \rightarrow hyperplane orthogonal to $(\underline{\mu}_1 - \underline{\mu}_0)$
- case 2: $\underline{C}_0 = \underline{C}_1 = \underline{C}$
 $\underline{w}^\top = (\underline{\mu}_1 - \underline{\mu}_0)^\top \underline{C}^{-1}$,
 $w_0 = \frac{1}{2} (\underline{\mu}_1 - \underline{\mu}_0)^\top \underline{C}^{-1} (\underline{\mu}_1 + \underline{\mu}_0) - \ln \left(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)} \right)$
 in general \underline{w} not colinear with $(\underline{\mu}_1 - \underline{\mu}_0)$
 \rightarrow hyperplane not orthogonal to $(\underline{\mu}_1 - \underline{\mu}_0)$
- if $C_0 = C_1$ and $\mu_0 = -\mu_1$, $\log R(x) = 0$ is linear and defines a separating hyperplane in \mathbb{X} which contains the origin, i.e.,

$$T: \mathbb{X} \rightarrow \mathbb{R}, \underline{x} \mapsto \begin{cases} 1 & \underline{w}^\top \underline{x} > 0 \\ 0 & \text{otherwise} \end{cases}$$



2. Learning and Generalization

2.1. Empirical Risk Function and Generalization Error

ML scenarios (unknown Stochastic Model) base learning on:
 $\text{Risk}_{\text{emp}}(T; \mathbb{S}) = \frac{1}{M} \sum_{i=1}^M L(T(\underline{x}_i), y_i), \quad (\underline{x}_i, y_i) \in \mathbb{S}$
 $\underline{x} \mapsto T(\underline{x}; \mathbb{S}) \quad T = \underset{T' \in \mathbb{T}}{\operatorname{argmin}} \{\text{Risk}_{\text{emp}}(T'; \mathbb{S})\}$

good Generalization: $\text{Risk}_{\text{emp}}(T; \mathbb{S}_{\text{test}})$ similar to $\text{Risk}_{\text{emp}}(T; \mathbb{S})$
bad Generalization:

- small \mathbb{T} that does not cover $T_{\text{opt}} \rightarrow$ cannot be selected by ML
- \Rightarrow strong mismatch between the desired and derived Test and refers to a sort of **Bias Error Term**
- too rich $\mathbb{T} \rightarrow$ fluctuating of the available data (measurement noise) is interpreted as meaningful information
- \Rightarrow **Overfitting**; leads to an increased **Variance Error Term**

2.2. Bias-Variance Decomposition

$\text{Risk} = E_{\mathbb{S}, X, Y} [L(T(X; \mathbb{S}), Y)] = E_X [1 - P_{Y|X}(Y = T_B(X)) + (1 - P_{\mathbb{S}|X}(T(X; \mathbb{S}) = T_B(X)))] (2P_{Y|X}(Y = T_B(X)) - 1)]$, $T_B(X)$ is the unknown **Bayes Test**
 If the potential set \mathbb{S} would be selected from a distribution such that the derived Test $T(\underline{x}; \mathbb{S})$ and the corresponding Bayes Test $T_B(\underline{x})$ are identical almost surely, then the Risk Function achieves its minimum value which is equal to the **Irreducible Error** $E_X [1 - P_{Y|X}(Y = T_B(X))]$ (denotes the probability that for a given input \underline{x} the Bayes Test $T_B(X)$ decides for the false label y).

3. Classification Trees and Random Forests

3.1. CART Algorithms

Generate Binary Trees by splitting \mathbb{X} at each (internal/root) node:
 $\mathbb{X}_{i, \text{left}} = \{\underline{x} \in \mathbb{X}_i | x_{j_i} < \tau_i\}$ $\mathbb{X}_{i, \text{right}} = \mathbb{X}_i \setminus \mathbb{X}_{i, \text{left}}$

Root/Internal node: Binary decision based on chosen threshold $\tau_i \in \mathbb{R}$, feature $x_{j_i} = [\underline{x}]_{j_i}$ with $j_i \in \mathbb{J} = \{1, \dots, \dim[\mathbb{X}]\}$ aims at minimizing $\text{Risk}_{\text{emp}}(T_{\text{CART}})$

Terminal node: n_i corresponds to subset $\mathbb{X}_i \in \mathbb{X} \rightarrow$ has no more children; outputs a decision
 $\Rightarrow \underline{x} \mapsto n_i(\underline{x})$

Empirical Impurity Measure: choose j_i and τ_i at n_i by:

$$I_{\text{CART}}(\mathbb{S}_i) = \sum_{k=1}^K (1 - \hat{P}_{Y|X}(Y = \theta_k | \{\underline{x} \in \mathbb{X}_i; \mathbb{S}_i\})) \hat{P}_{Y|X}(Y = \theta_k | \{\underline{x} \in \mathbb{X}_i; \mathbb{S}_i\})$$

 with

$$\hat{P}_{Y|X}(Y = \theta_k | \{\underline{x} \in \mathbb{X}_i; \mathbb{S}_i\}) = \frac{M_k(\mathbb{S}_i)}{M(\mathbb{S}_i)} = \frac{|\{(\underline{x}, y) \in \mathbb{S}_i | y = \theta_k\}|}{|\mathbb{S}_i|}$$

$$\Rightarrow \{j_i, \tau_i\} = \underset{j \in \mathbb{J}, \tau \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{k=1}^K \left(1 - \frac{M_k(\mathbb{S}_{i, \text{left}})}{M(\mathbb{S}_i)} \right) \frac{M_k(\mathbb{S}_{i, \text{left}})}{M(\mathbb{S}_i)} + \left(1 - \frac{M_k(\mathbb{S}_{i, \text{right}})}{M(\mathbb{S}_i)} \right) \frac{M_k(\mathbb{S}_{i, \text{right}})}{M(\mathbb{S}_i)} \right\}$$

Overfitting (comes with high purity) can be controlled by a **Test Set \mathbb{S}_{Test}** .

Decision Rule: At terminal node n_i , input \underline{x} is assigned to $T_{\text{CART}}(\underline{x}; \mathbb{S}): \mathbb{X} \mapsto \{1, \dots, K\}, \underline{x} \mapsto \underset{k}{\operatorname{argmax}} \{M_k(\mathbb{S}_i)\}$

Gini Impurity Index: $I_{\text{CART}} =$

$$\sum_{k=1}^K (1 - P_{Y|X}(\underline{y} = \theta_k | \{\underline{x} \in \mathbb{X}\})) P_{Y|X}(\underline{y} = \theta_k | \{\underline{x} \in \mathbb{X}\}) = \sum_{k=1}^K \sum_{j=1, j \neq k}^K P_{Y|X}(\underline{y} = \theta_j | \{\underline{x} \in \mathbb{X}\}) P_{Y|X}(\underline{y} = \theta_k | \{\underline{x} \in \mathbb{X}\})$$

3.2. Random Forests

Avoid **Overfitting** (here: CART) \Rightarrow combine independent **Hypothesis Tests**: e.g. by **Majority Vote**

$T_{\text{maj}}(\underline{x}) = \text{majority} \{T_{\text{CART}}(\underline{x}; \mathbb{S}^{(t)}, \nu^{(t)})\}_{t=1}^{t_{\text{max}}}$
Randomization Parameter ν_t controls an additionally introduced Randomness between the individual Tests.

\Rightarrow **Variance of $T_{\text{avg}}(\underline{x})$** is reduced by $1/t_{\text{max}}$ with respect to the **Variance** of the individual test.

Random Forest Method:

- $T_{\text{RF}}(\underline{x}) = \text{majority} \{T_{\text{CART}}(\underline{x}; \mathbb{S}^{(t)}, \mathbb{J}^{(t)})\}_{t=1}^{t_{\text{max}}}$
- Stochastic Independence by Bootstrapping of training samples (random sampling from \mathbb{S} with replacement) \Rightarrow large t_{max} guarantees excellent performance (yet Tests are still correlated)
- Overfitting not considered (maximum purity) \Rightarrow small bias of RF Method

4. Hypothesis Testing

making a decision based on the observations

4.1. Definition
Null hypothesis $H_0 : \theta \in \Theta_0$ (Assumed first to be true)
Alternate hypothesis $H_1 : \theta \in \Theta_1$ (The one to proof)
Decision rule $\varphi : \mathbb{X} \rightarrow [0, 1]$ with
 $\varphi(x) = 1$: decide for H_1 , $\varphi(x) = 0$: decide for H_0 Error level α with $\mathbb{E}[d(X)|\theta] \leq \alpha, \forall \theta \in \Theta_0$

Error Type	Decision \ Reality	H_1 false (H_0 true)	H_1 true (H_0 false)
1 (FA) False Alarm	H_1 rejected (H_0 accepted)	True Negative $P = 1 - \alpha$	False Negative (Type 2) $P = \beta$
2 (DE) Detection Error	H_1 accepted (H_0 rejected)	False Positive (Type 1) $P = \alpha$	True Positive $P = 1 - \beta$

Power: Sensitivity/Recall/Hit Rate: $\frac{TP}{TP+FN} = 1 - \beta$
Specificity/True negative rate: $\frac{TN}{FP+TN} = 1 - \alpha$
Precision/Positive Prediction rate: $\frac{TP}{TP+FP}$
Accuracy: $\frac{TP+TN}{P+N} = \frac{2-\alpha-\beta}{2}$

4.1.1. Design of a test
Cost criterion $G_\varphi : \Theta \rightarrow [0, 1], \theta \mapsto \mathbb{E}[d(X)|\theta]$
False Positive lower than α : $G_d(\theta)|_{\theta \in \Theta_0} \leq \alpha, \forall \theta \in \Theta_0$
False Negative small as possible: $\max\{G_d(\theta)|_{\theta \in \Theta_1}\}, \forall \theta \in \Theta_1$

4.2. Sufficient Statistics
Sufficiency for a test $T(X)$ means that no other test statistic, i.e., function of the observations \underline{x} , contains additional information about the parameter θ to be estimated:
 $f_{X|T}(x|T(x) = t, \theta) = f_{X|T}(x|T(x) = t)$

5. Math

$\pi \approx 3.141\,59$ $e \approx 2.718\,28$ $\sqrt{2} \approx 1.414$ $\sqrt{3} \approx 1.732$
Binome, Trinome
 $(a \pm b)^2 = a^2 \pm 2ab + b^2$ $a^2 - b^2 = (a - b)(a + b)$
 $(a \pm b)^3 = a^3 \pm 3a^2b + 3ab^2 \pm b^3$
 $(a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc$
Folgen und Reihen
 $\sum_{k=1}^n k = \frac{n(n+1)}{2}$ $\sum_{k=0}^n q^k = \frac{1-q^{n+1}}{1-q}$ $\sum_{n=0}^{\infty} \frac{z^n}{n!} = e^z$
Arithmetische Summenformel Geometrische Summenformel Exponentialreihe

Mittelwerte (\sum von i bis N) (Median: Mitte einer geordneten Liste)
 $\bar{x}_{ar} = \frac{1}{N} \sum x_i \geq \bar{x}_{geo} = \sqrt[N]{\prod x_i} \geq \bar{x}_{hm} = \frac{N}{\sum \frac{1}{x_i}}$
Arithmetisches Geometrisches Mittel Harmonisches
Ungleichungen: Bernoulli-Ungleichung: $(1 + x)^n \geq 1 + nx$
 $||x| - |y|| \leq |x \pm y| \leq |x| + |y|$ $|\underline{x}^\top \cdot \underline{y}| \leq \|\underline{x}\| \cdot \|\underline{y}\|$
Dreiecksungleichung Cauchy-Schwarz-Ungleichung
Mengen: De Morgan: $\overline{A \cap B} = \overline{A} \cup \overline{B}$ $\overline{A \cup B} = \overline{A} \cap \overline{B}$

5.1. Exp. und Log. $e^x := \lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n$ $e \approx 2,71828$
 $a^x = e^{x \ln a}$ $\log_a x = \frac{\ln x}{\ln a}$ $\ln x \leq x - 1$
 $\ln(x^a) = a \ln(x)$ $\ln(\frac{x}{a}) = \ln x - \ln a$ $\log(1) = 0$

5.2. Matrizen $\underline{A} \in \mathbb{K}^{m \times n}$
 $\underline{A} = (a_{ij}) \in \mathbb{K}^{m \times n}$ hat m Zeilen (Index i) und n Spalten (Index j)
 $(\underline{A} + \underline{B})^\top = \underline{A}^\top + \underline{B}^\top$ $(\underline{A} \cdot \underline{B})^\top = \underline{B}^\top \cdot \underline{A}^\top$
 $(\underline{A}^\top)^{-1} = (\underline{A}^{-1})^\top$ $(\underline{A} \cdot \underline{B})^{-1} = \underline{B}^{-1} \underline{A}^{-1}$
 $\dim \mathbb{K} = n = \text{rang } \underline{A} + \dim \ker \underline{A}$ $\text{rang } \underline{A} = \text{rang } \underline{A}^\top$
5.2.1. Quadratische Matrizen $\underline{A} \in \mathbb{K}^{n \times n}$
regulär/invertierbar/nicht-singulär $\Leftrightarrow \det(\underline{A}) \neq 0 \Leftrightarrow \text{rang } \underline{A} = n$
singulär/nicht-invertierbar $\Leftrightarrow \det(\underline{A}) = 0 \Leftrightarrow \text{rang } \underline{A} \neq n$
orthogonal $\Leftrightarrow \underline{A}^\top = \underline{A}^{-1} \Rightarrow \det(\underline{A}) = \pm 1$
symmetrisch: $\underline{A} = \underline{A}^\top$ schief-symmetrisch: $\underline{A} = -\underline{A}^\top$
5.2.2. Determinante von $\underline{A} \in \mathbb{K}^{n \times n}$: $\det(\underline{A}) = |\underline{A}|$
 $\det \begin{bmatrix} \underline{A} & \underline{0} \\ \underline{C} & \underline{D} \end{bmatrix} = \det \begin{bmatrix} \underline{A} & \underline{B} \\ \underline{0} & \underline{D} \end{bmatrix} = \det(\underline{A}) \det(\underline{D})$
 $\det(\underline{A}) = \det(\underline{A}^\top)$ $\det(\underline{A}^{-1}) = \det(\underline{A})^{-1}$
 $\det(\underline{A}\underline{B}) = \det(\underline{A}) \det(\underline{B}) = \det(\underline{B}) \det(\underline{A}) = \det(\underline{B}\underline{A})$
Hat \underline{A} 2 linear abhäng. Zeilen/Spalten $\Rightarrow |\underline{A}| = 0$

5.2.3. Eigenwerte (EW) λ und Eigenvektoren (EV) \underline{v}
 $\underline{A}\underline{v} = \lambda \underline{v}$ $\det \underline{A} = \prod \lambda_i$ $\text{Sp } \underline{A} = \sum a_{ii} = \sum \lambda_i$

Eigenwerte: $\det(\underline{A} - \lambda \underline{1}) = 0$ Eigenvektoren: $\ker(\underline{A} - \lambda_i \underline{1}) = \underline{v}_i$
EW von Dreieck/Diagonal Matrizen sind die Elem. der Hauptdiagonale.
5.2.4. Spezialfall 2×2 Matrix \underline{A}
 $\det(\underline{A}) = ad - bc$ $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \underline{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$
 $\text{Sp}(\underline{A}) = a + d$
 $\lambda_{1/2} = \frac{\text{Sp } \underline{A}}{2} \pm \sqrt{\left(\frac{\text{sp } \underline{A}}{2}\right)^2 - \det \underline{A}}$

5.2.5. Differentiation
 $\frac{\partial \underline{x}^\top \underline{y}}{\partial \underline{x}} = \frac{\partial \underline{y}^\top \underline{x}}{\partial \underline{x}} = \underline{y}$ $\frac{\partial \underline{x}^\top \underline{A} \underline{x}}{\partial \underline{x}} = (\underline{A} + \underline{A}^\top) \underline{x}$
 $\frac{\partial \underline{x}^\top \underline{A} \underline{y}}{\partial \underline{A}} = \underline{x} \underline{y}^\top$ $\frac{\partial \det(\underline{B} \underline{A} \underline{C})}{\partial \underline{A}} = \det(\underline{B} \underline{A} \underline{C}) (\underline{A}^{-1})^\top$

5.2.6. Ableitungsregeln ($\forall \lambda, \mu \in \mathbb{R}$)
Linearität: $(\lambda f + \mu g)'(x) = \lambda f'(x) + \mu g'(x_0)$
Produkt: $(f \cdot g)'(x) = f'(x)g(x) + f(x)g'(x)$
Quotient: $\left(\frac{f}{g}\right)'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}$ $\left(\frac{\text{NAZ}-\text{ZAN}}{\text{N}^2}\right)$
Kettenregel $(f(g(x)))' = f'(g(x))g'(x)$

5.3. Integrale $\int e^x dx = e^x = (e^x)'$
Partielle Integration: $\int u w' = u w - \int u' w$
Substitution: $\int f(g(x))g'(x) dx = \int f(t) dt$

$F(x) - C$	$f(x)$	$f'(x)$
$\frac{1}{q+1} x^{q+1}$	x^q	$q x^{q-1}$
$\frac{2\sqrt{ax^3}}{3}$	\sqrt{ax}	$\frac{a}{2\sqrt{ax}}$
$x \ln(ax) - x$	$\ln(ax)$	$\frac{1}{x}$
$\frac{1}{a^2} e^{ax} (ax - 1)$	$x \cdot e^{ax}$	$e^{ax} (ax + 1)$
$\frac{a^x}{\ln(a)}$	a^x	$a^x \ln(a)$
$-\cos(x)$	$\sin(x)$	$\cos(x)$
$\cosh(x)$	$\sinh(x)$	$\cosh(x)$
$-\ln \cos(x) $	$\tan(x)$	$\frac{1}{\cos^2(x)}$

$\int e^{at} \sin(bt) dt = e^{at} \frac{\sin(bt)+b \cos(bt)}{a^2+b^2}$
 $\int \frac{dt}{\sqrt{at+b}} = \frac{2\sqrt{at+b}}{a}$ $\int t^2 e^{at} dt = \frac{(ax-1)^2+1}{a^3} e^{at}$
 $\int t e^{at} dt = \frac{at-1}{a^2} e^{at}$ $\int x e^{ax^2} dx = \frac{1}{2a} e^{ax^2}$

5.3.1. Volumen und Oberfläche von Rotationskörpern um x-Achse
 $V = \pi \int_a^b f(x)^2 dx$ $O = 2\pi \int_a^b f(x) \sqrt{1 + f'(x)^2} dx$

6. Probability Theory Basics

6.1. Kombinatorik
Mögliche Variationen/Kombinationen um k Elemente von maximal n Elementen zu wählen bzw. k Elemente auf n Felder zu verteilen:

	Mit Reihenfolge	Reihenfolge egal
Mit Wiederholung	n^k	$\binom{n+k-1}{k}$
Ohne Wiederholung	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Permutation von n mit jeweils k gleichen Elementen: $\frac{n!}{k_1! \cdot k_2! \cdot \dots}$
Binomialkoeffizient $\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k! \cdot (n-k)!}$
 $\binom{n}{0} = 1$ $\binom{n}{1} = n$ $\binom{n}{2} = 6$ $\binom{5}{2} = 10$ $\binom{6}{2} = 15$

6.2. Der Wahrscheinlichkeitsraum (Ω, \mathbb{F}, P)		
Ergebnismenge	$\Omega = \{\omega_1, \omega_2, \dots\}$	Ergebnis $\omega_j \in \Omega$
Ereignisalgebra	$\mathbb{F} = \{A_1, A_2, \dots\}$	Ereignis $A_i \subseteq \Omega$
Wahrscheinlichkeitsmaß	$P : \mathbb{F} \rightarrow [0, 1]$	$P(A) = \frac{ A }{ \Omega }$

6.3. Wahrscheinlichkeitsmaß P
 $P(A) = \frac{|A|}{|\Omega|}$ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

6.3.1. Axiome von Kolmogorow
Nichtnegativität: $P(A) \geq 0 \Rightarrow P : \mathbb{F} \mapsto [0, 1]$
Normiertheit: $P(\Omega) = 1$
Additivität: $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$,
wenn $A_i \cap A_j = \emptyset, \forall i \neq j$

6.4. Bedingte Wahrscheinlichkeit
Bedingte Wahrscheinlichkeit für A falls B bereits eingetreten ist:
 $P_B(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$
6.4.1. Totale Wahrscheinlichkeit und Satz von Bayes
Es muss gelten: $\bigcup_{i \in I} B_i = \Omega$ für $B_i \cap B_j = \emptyset, \forall i \neq j$
Totale Wahrscheinlichkeit: $P(A) = \sum_{i \in I} P(A|B_i) P(B_i)$
Satz von Bayes: $P(B_k|A) = \frac{P(A|B_k) P(B_k)}{\sum_{i \in I} P(A|B_i) P(B_i)}$
Multiplikationssatz: $P(A \cap B) = P(A|B) P(B) = P(B|A) P(A)$

6.5. Zufallsvariable
 $X : \Omega \mapsto \Omega'$ ist Zufallsvariable, wenn für jedes Ereignis $A' \in \mathbb{F}'$ im Bildraum ein Ereignis A im Urbildraum \mathbb{F} existiert, sodass $\{\omega \in \Omega | X(\omega) \in A'\} \in \mathbb{F}$

6.6. Distribution		
Bezeichnung	Abk.	Zusammenhang
Wahrscheinlichkeitsdichte	pdf	$f_X(x) = \frac{dF_X(x)}{dx}$
Kumulative Verteilungsfkt.	cdf	$F_X(x) = \int_{-\infty}^x f_X(\xi) d\xi$
Joint CDF: $F_{X,Y}(x, y) = P(\{X \leq x, Y \leq y\})$		

6.7. Relations between $f_X(x), f_{X,Y}(x, y), f_{X|Y}(x|y)$
 $f_{X,Y}(x, y) = f_X|Y(x, y) f_Y(y) = f_Y|X(y, x) f_X(x)$
Joint PDF
 $\int_{-\infty}^{\infty} f_{X,Y}(x, \xi) d\xi = \int_{-\infty}^{\infty} f_{X|Y}(x, \xi) f_Y(\xi) d\xi = f_X(x)$
Marginalization Total Probability

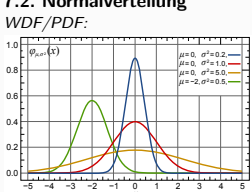
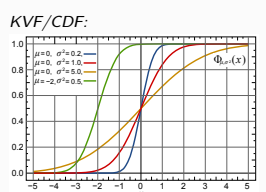
6.8. Bedingte Zufallsvariablen
Ereignis A gegeben: $F_{X|A}(x|A) = P(\{X \leq x\} | A)$
ZV Y gegeben: $F_{X|Y}(x|y) = P(\{X \leq x\} | \{Y = y\})$
 $p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$
 $f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{dF_{X|Y}(x|y)}{dx}$

6.9. Unabhängigkeit von Zufallsvariablen
 X_1, \dots, X_n sind stochastisch unabhängig, wenn für jedes $\underline{x} \in \mathbb{R}^n$ gilt:
 $F_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i)$
 $p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n p_{X_i}(x_i)$
 $f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$

7. Common Distributions

7.1. Binomialverteilung $\mathcal{B}(n, p)$ mit $p \in [0, 1], n \in \mathbb{N}$
Folge von n Bernoulli-Experimenten
 p : Wahrscheinlichkeit für Erfolg k : Anzahl der Erfolge
 $p_X(k) = B_{n,p}(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k \in \{0, \dots, n\} \\ 0 & \text{sonst} \end{cases}$

$E[X] = np$ Erwartungswert	$\text{Var}[X] = np(1-p)$ Varianz	$G_X(z) = (pz + 1 - p)^n$ Wahrscheinlichkeitserz. Funktion
-------------------------------	--------------------------------------	---

7.2. Normalverteilung
WDF/PDF:  KVF/CDF: 
WDF: $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $x \in \mathbb{R}$ $\mu \in \mathbb{R}$ $\sigma > 0$

$E(X) = \mu$ Erwartungswert	$\text{Var}(X) = \sigma^2$ Varianz	$\varphi_X(\omega) = e^{j\omega\mu - \frac{\omega^2\sigma^2}{2}}$ Charakt. Funktion
--------------------------------	---------------------------------------	--

7.3. Sonstiges
Gammadistribution $\Gamma(\alpha, \beta)$: $E[X] = \frac{\alpha}{\beta}$
Exponential: $f(x, \lambda) = \lambda e^{-\lambda x}$ $E[X] = \lambda^{-1}$ $\text{Var}[X] = \lambda^{-2}$

8. Wichtige Parameter

8.1. Erwartungswert (1. zentrales Moment)

gibt den mittleren Wert einer Zufallsvariablen an

μ_X = E[X] = ∑_{x ∈ Ω'} x · P_X(x) ≐ ∫_{ℝ} x · f_X(x) dx

E[αX + βY] = αE[X] + βE[Y] X ≤ Y ⇒ E[X] ≤ E[Y] E[X^2] = Var[X] + E[X]^2 E[X Y] = E[X]E[Y], falls X und Y stochastisch unabhängig

8.1.1. Für Funktionen von Zufallsvariablen g(x)

E[g(X)] = ∑_{x ∈ Ω'} g(x) P_X(x) ≐ ∫_{ℝ} g(x) f_X(x) dx

8.2. Varianz (2. zentrales Moment)

ist ein Maß für die Stärke der Abweichung vom Erwartungswert

σ_X^2 = Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2

Var[αX + βY] = α^2 Var[X] + β^2 Var[Y] + 2αβ Cov[X, Y]

Var[∑_{i=1}^n X_i] = ∑_{i=1}^n Var[X_i] + ∑_{j≠i} Cov[X_i, X_j]

Standard Abweichung: σ = √Var[X]

8.3. Kovarianz

Maß für den linearen Zusammenhang zweier Variablen

Cov[X, Y] = E[(X - E[X])(Y - E[Y])^T] = E[X Y^T] - E[X]E[Y]^T = Cov[Y, X]

Cov[αX + βY, γZ + δW] = αγCov[X, Z] + αδCov[X, W] + βγCov[Y, Z] + βδCov[Y, W]

8.3.1. Korrelation = standardisierte Kovarianz

ρ(X, Y) = Cov[X, Y] / (σ_X σ_Y) = C_{XY} / (σ_X σ_Y) ρ(X, Y) ∈ [-1; 1]

8.3.2. Kovarianzmatrix für z = (x, y)^T

Cov[z] = C_z = [C_{XX} C_{XY}; C_{YX} C_{YY}] = [Cov[X, X] Cov[X, Y]; Cov[Y, X] Cov[Y, Y]]

Immer symmetrisch: C_{yx} = C_{xy}! Für Matrizen: C_{xy} = C_{yx}^T

9. Estimation

9.1. Estimation

Statistic Estimation treats the problem of inferring underlying characteristics of unknown random variables on the basis of observations of outputs of those random variables.

Sample Space Ω	nonempty set of outputs of experiment
Sigma Algebra F ⊆ 2^Ω	set of subsets of outputs (events)
Probability P : F → [0, 1]	
Random Variable X : Ω → X	mapped subsets of Ω
Observations: x_1, ..., x_N	single values of X
Observation Space X	possible observations of X
Unknown parameter θ ∈ Θ	parameter of probability function
Estimator ○→● : X → Θ	○→●(X) = θ̂, finds θ̂ from X

unknown parm. θ	estimation of param. θ̂
R.V. of param. Θ	estim. of R.V. of parm T(X) = Θ̂

9.2. Quality Properties of Estimators

Consistent: If lim_{N→∞} ○→●(x_1, ..., x_N) = θ

Bias Bias(○→●) := E[○→●(X_1, ..., X_N)] - θ

unbiased if Bias(○→●) = 0 (biased estimators can provide better estimates than unbiased estimators.)

Variance Var[○→●] := E[(○→● - E[○→●])^2]

9.3. Mean Square Error (MSE)

The MSE is an extension of the Variance Var[○→●] := E[(○→● - E[○→●])^2]:

ε[○→●] = E[(○→● - θ)^2]^{MSE} = Var(○→●) + (Bias[○→●])^2 = E[(θ̂ - θ)^2]

If θ is also r.v. ⇒ mean over both (e.g. Bayes est.):

Mean MSE: E[(○→●(X) - θ)^2] = E[E[(○→●(X) - θ)^2 | θ]]

9.3.1. Minimum Mean Square Error (MMSE)

Minimizes mean square error: arg min_{θ̂} E[(θ̂ - θ)^2]

E[(θ̂ - θ)^2] = E[θ^2] - 2θE[θ̂] + θ̂^2

Solution: d/dθ E[(θ̂ - θ)^2] ≐ 0 = -2E[θ] + 2θ̂ ⇒ θ̂_{MMSE} = E[θ]

9.4. Maximum Likelihood

Given model {X, F, P_θ; θ ∈ Θ}, assume P_θ(x̄) or f_X(x̄, θ) for observed data x̄. Estimate parameter θ so that the likelihood L(x̄, θ) or L(θ | X = x̄) to obtain x̄ is maximized.

Likelihood Function: (Prob. for θ given x̄)

Discrete: L(x_1, ..., x_N; θ) = P_θ(x_1, ..., x_N)

Continuous: L(x_1, ..., x_N; θ) = f_{X_1, ..., X_N}(x_1, ..., x_N, θ)

If N observations are Identically Independently Distributed (i.i.d.):

L(x̄, θ) = ∏_{i=1}^N P_θ(x_i) = ∏_{i=1}^N f_{X_i}(x_i)

ML Estimator (Picks θ): ○→●_{ML} : X → arg max_{θ ∈ Θ} L(x̄, θ)

= arg max_{θ ∈ Θ} {log L(x̄, θ)} ≐ arg max_{θ ∈ Θ} {∑ log L(x_i, θ)}

Find Maximum: ∂L(x̄, θ)/∂θ = d/dθ log L(x; θ)|_{θ=θ̂} ≐ 0

Solve for θ to obtain ML estimator function θ̂_{ML}

Check quality of estimator with MSE

Maximum-Likelihood Estimator is Asymptotically Efficient. However, there might be not enough samples and the likelihood function is often not known.

9.5. Uniformly Minimum Variance Unbiased (UMVU) Estimators (Best unbiased estimators)

Best unbiased estimator: Lowest Variance of all estimators.

Fisher's Information Inequality: Estimate lower bound of variance if

- L(x, θ) > 0, ∀ x, θ
- L(x, θ) is diffable for θ
- ∫_X ∂/∂θ L(x, θ) dx = ∂/∂θ ∫_X L(x, θ) dx

Score Function:

g(x, θ) = ∂/∂θ log L(x, θ) = (∂/∂θ L(x, θ)) / L(x, θ) E[g(x, θ)] = 0

Fischer Information:

I_F(θ) := Var[g(X, θ)] = E[g(x, θ)^2] = -E[∂^2/∂θ^2 log L(X, θ)]

Cramér-Rao Lower Bound (CRB): (if ○→● is unbiased)

Var[○→●(X)] ≥ (∂E[○→●(X)]/∂θ)^2 / I_F(θ) Var[○→●(X)] ≥ 1/I_F(θ)

For N i.i.d. observations: I_F^{(N)}(x, θ) = N · I_F^{(1)}(x, θ)

9.5.1. Exponential Models

If f_X(x) = h(x) exp(a(θ)t(x)) / exp(b(θ)) then I_F(θ) = ∂a(θ)/∂θ ∂E[t(X)]/∂θ

Some Derivations: (check in exam)

Uniformly: Not diffable ⇒ no I_F(θ)

Normal N(θ, σ^2): g(x, θ) = (x-θ)/σ^2 I_F(θ) = 1/σ^2

Binomial B(θ, K): g(x, θ) = x/θ - K/(1-θ) I_F(θ) = K/(θ(1-θ))

9.6. Bayes Estimation (Conditional Mean)

A Priori information about θ is known as probability f_Θ(θ; σ) with random variable Θ and parameter σ. Now the conditional pdf f_{X|Θ}(x, θ) is used to find θ by minimizing the mean MSE instead of uniformly MSE. Mean MSE for Θ: E[E[(T(X) - Θ)^2 | Θ = θ]]

Conditional Mean Estimator:

T_{CM} : x → E[Θ | X = x] = ∫_Θ θ · f_{Θ|X}(θ | x) dθ

Posterior f_{Θ|X}(θ | x̄) = f_{X|Θ}(x̄) f_Θ(θ) / ∫_Θ f_{X|Ξ}(x̄, ξ) dξ = f_{X|θ}(x̄) f_Θ(θ) / f_X(x̄)

Hint: to calculate f_{Θ|X}(θ | x̄): Replace every factor not containing θ, such as 1/f_X(x̄) with a factor γ and determine γ at the end such that ∫_Θ f_{Θ|X}(θ | x̄) dθ = 1

MMSE: E[|| ○→●_{CM} - Θ ||_2^2] = tr(C_{Θ|X}) = tr(C_{Θ} - C_{Θ,X} C_X^{-1} C_{X,Θ})

Multivariate Gaussian: X, Θ ~ N ⇒ σ_X^2 = σ_X^2|_{Θ=θ} + σ_Θ

○→●_{CM} : x → E[Θ | X = x] = μ_Θ + C_{Θ,X} C_X^{-1} (x - μ_X)

MMSE:

E[|| ○→●_{CM} - Θ ||_2^2] = tr(C_{Θ|X}) = tr(C_{Θ} - C_{Θ,X} C_X^{-1} C_{X,Θ})

Orthogonality Principle:

○→●_{CM}(X) - Θ ⊥ h(X) ⇒ E[(T_{CM}(X) - Θ)h(X)] = 0

MMSE Estimator: θ̂_{MMSE} = arg min MSE

minimizes the MSE for all estimators

9.7. Example:

Estimate mean θ of X with prior knowledge θ ∈ Θ ~ N: X ~ N(θ, σ_X^2|_{Θ=θ}) and Θ ~ N(m, σ_Θ^2)

θ̂_{CM} = E[Θ | X = x̄] = (Nσ_Θ^2 / (σ_X^2|_{Θ=θ} + Nσ_Θ^2)) θ̂_{ML} + (σ_X^2|_{Θ=θ} / (σ_X^2|_{Θ=θ} + Nσ_Θ^2)) m

For N independent observations x_i: θ̂_{ML} = 1/N ∑ x_i Large N ⇒ ML better, small N ⇒ CM better

10. Linear Estimation

t is now the unknown parameter θ, we want to estimate y and x̄ is the input vector... review regression problem ŷ = A x̄ (we solve for x̄), here we solve for t, because x̄ is known (measured)! Confusing... 1. Training → 2. Estimation Training: We observe y and x̄ (knowing both) and then based on that we try to estimate y given x̄ (only observe x̄) with a linear model ŷ = x̄^T t

Estimation: ŷ = x̄^T t + m or ŷ = x̄^T t

Given: N observations (y_i, x̄_i), unknown parameters t, noise m

y = [y_1; ...; y_n]^T X = [x_1^T; ...; x_n^T]^T Note: ŷ ≠ y!!

Problem: Estimate y based on given (known) observations x̄ and unknown parameter t with assumed linear Model: ŷ = x̄^T t

Note y = x̄^T t + m → y = x̄'^T t' with x̄' = (x̄; 1), t' = (t; m)

Sometimes in Exams: ŷ = x̄^T t ⇔ x̄ = T^T y estimate x̄ given y and unknown T

10.1. Least Square Estimation (LSE)

Tries to minimize the square error for linear Model: ŷ_{LS} = x̄^T t_{LS}

Least Square Error: min_t ∑_{i=1}^N (y_i - x̄_i^T t)^2 = min_t ||y - X t||

t_{LS} = (X^T X)^{-1} X^T y

ŷ_{LS} = X t_{LS} ∈ span(X)

Orthogonality Principle: N observations x̄_i ∈ ℝ^d

Y - X t_{LS} ⊥ span(X) ⇔ Y - X t_{LS} ∈ null(X^T), thus

X^T (Y - X t_{LS}) = 0 and if N > d ∧ rang(X) = d:

t_{LS} = (X^T X)^{-1} X^T Y

10.2. Linear Minimum Mean Square Estimator (LMMSE)

Estimate y with linear estimator t, such that ŷ = t^T x̄ + m

Note: the Model does not need to be linear! The estimator is linear!

ŷ_{LMMSE} = arg min_{t,m} E[||y - (t^T x̄ + m)||_2^2]

If Random joint variable z = (x̄; y) with

μ_z = (μ_x; μ_y) and C_z = [C_{xx} C_{xy}; C_{yx} C_{yy}] then

LMMSE Estimation of y given x is

ŷ = μ_y + c_{y,x} C_x^{-1} (x̄ - μ_x) = c_{y,x} C_x^{-1} x̄ - μ_y + c_{y,x} C_x^{-1} μ_x = t^T x̄ + m

Minimum MSE: E[||y - (x̄^T t + m)||_2^2] = c_{yy} - c_{y,x} C_x^{-1} c_{x,y}

Hint: First calculate ŷ in general and then set variables according to system equation.

Multivariate: ŷ = T_{LMMSE} x̄ T_{LMMSE}^T = C_{yx} C_x^{-1}

If μ_z = 0 then

Estimator ŷ = c_{y,x} C_x^{-1} x̄

Minimum MSE: E[c_{y,x}] = c_{yy} - t^T c_{x,y}

10.3. Matched Filter Estimator (MF)

For channel $\mathbf{y} = \mathbf{h}x + \mathbf{v}$, Filtered: $\hat{\mathbf{t}}^\top \mathbf{y} = \hat{\mathbf{t}}^\top \mathbf{h}x + \hat{\mathbf{t}}^\top \mathbf{v}$

Find Filter $\hat{\mathbf{t}}^\top$ that maximizes SNR = $\frac{\|\mathbf{h}x\|}{\|\mathbf{v}\|}$

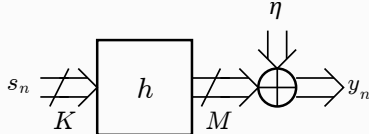
$$\hat{\mathbf{t}}_{\text{MF}} = \max_{\mathbf{t}} \left\{ \frac{\mathbb{E}[(\hat{\mathbf{t}}^\top \mathbf{h}x)^2]}{\mathbb{E}[(\hat{\mathbf{t}}^\top \mathbf{v})^2]} \right\}$$

In the lecture (estimate \mathbf{h}):

$$\hat{\mathbf{T}}_{\text{MF}} = \max_{\mathbf{T}} \left\{ \frac{|\mathbb{E}[\hat{\mathbf{h}}^H \mathbf{h}]|^2}{\text{tr}[\text{Var}[\mathbf{T}\mathbf{n}]]} \right\}$$

$$\hat{\mathbf{t}}_{\text{MF}} = \hat{\mathbf{T}}_{\text{MF}} \mathbf{y} \quad \mathbf{T}_{\text{MF}} \propto \mathbf{C}_{\mathbf{h}} \mathbf{S}^H \mathbf{C}_{\mathbf{n}}^{-1}$$

10.4. Example



System Model: $\mathbf{y}_n = \mathbf{H}\mathbf{s}_n + \eta_n$

with $\mathbf{H} = (h_{m,k}) \in \mathbb{C}^{M \times K}$ ($m \in [1, M], k \in [1, K]$)

Linear Channel Model $\mathbf{y} = \mathbf{S}\mathbf{h} + \mathbf{n}$ with $\mathbf{h} \sim \mathcal{N}(0, \mathbf{C}_{\mathbf{h}})$ and $\mathbf{n} \sim \mathcal{N}(0, \mathbf{C}_{\mathbf{n}})$

Linear Estimator \mathbf{T} estimates $\hat{\mathbf{h}} = \mathbf{T}\mathbf{y} \in \mathbb{C}^{MK}$

$$\mathbf{T}_{\text{MMSE}} = \mathbf{C}_{\mathbf{h}} \mathbf{y} \mathbf{y}^H = \mathbf{C}_{\mathbf{h}} \mathbf{S}^H (\mathbf{S} \mathbf{C}_{\mathbf{h}} \mathbf{S}^H + \mathbf{C}_{\mathbf{n}})^{-1}$$

$$\mathbf{T}_{\text{ML}} = \mathbf{T}_{\text{Cor}} = (\mathbf{S}^H \mathbf{C}_{\mathbf{n}}^{-1} \mathbf{S})^{-1} \mathbf{S}^H \mathbf{C}_{\mathbf{n}}^{-1}$$

$$\mathbf{T}_{\text{MF}} \propto \mathbf{C}_{\mathbf{h}} \mathbf{S}^H \mathbf{C}_{\mathbf{n}}^{-1}$$

For Assumption $\mathbf{S}^H \mathbf{S} = N\sigma_s^2 \mathbf{I}_{K \times M}$ and $\mathbf{C}_{\mathbf{n}} = \sigma_n^2 \mathbf{I}_{N \times M}$

Estimator	Averaged Squared Bias	Variance
ML/Correlator	0	$KM \frac{\sigma_n^2}{N\sigma_s^2}$
Matched Filter	$\sum_{i=1}^{KM} \lambda_i \left(\frac{\lambda_i}{\lambda_1} - 1 \right)^2$	$\sum_{i=1}^{KM} \left(\frac{\lambda_i}{\lambda_1} \right)^2 \frac{\sigma_n^2}{N\sigma_s^2}$
MMSE	$\sum_{i=1}^{KM} \lambda_i \left(\frac{1}{1 + \frac{\sigma_n^2}{\lambda_i N \sigma_s^2}} - 1 \right)^2$	$\sum_{i=1}^{KM} \frac{1}{\left(1 + \frac{\sigma_n^2}{\lambda_i N \sigma_s^2} \right)^2} \frac{\sigma_n^2}{N\sigma_s^2}$

10.5. Estimators

Upper Bound: Uniform in $[0; \theta]$: $\hat{\theta}_{\text{ML}} = \frac{2}{N} \sum x_i$

Probability p for $\mathcal{B}(p, N)$: $\hat{p}_{\text{ML}} = \frac{x}{N}$ $\hat{p}_{\text{CM}} = \frac{x+1}{N+2}$

Mean μ for $\mathcal{N}(\mu, \sigma^2)$: $\hat{\mu}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x_i$

Variance σ^2 for $\mathcal{N}(\mu, \sigma^2)$: $\hat{\sigma}_{\text{ML}}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$

11. Gaussian Stuff

11.1. Gaussian Channel

Channel: $Y = h s_i + N$ with $h \sim \mathcal{N}$, $N \sim \mathcal{N}$

$$L(y_1, \dots, y_N) = \prod_{i=1}^N f_{Y_i}(y_i, h)$$

$$f_{Y_i}(y_i, h) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - h s_i)^2\right)$$

$$\hat{h}_{ML} = \arg\min_h \left\{ \|\mathbf{y} - h\mathbf{s}\|^2 \right\} = \frac{\mathbf{s}^\top \mathbf{y}}{\mathbf{s}^\top \mathbf{s}}$$

If multidimensional channel: $\mathbf{y} = \mathbf{S}\mathbf{h} + \mathbf{n}$:

$$L(\mathbf{y}, \mathbf{h}) = \frac{1}{\sqrt{\det(2\pi\mathbf{C})}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{S}\mathbf{h})^\top \mathbf{C}^{-1}(\mathbf{y} - \mathbf{S}\mathbf{h})\right)$$

$$l(\mathbf{y}, \mathbf{h}) = \frac{1}{2} \left(\log(\det(2\pi\mathbf{C})) - (\mathbf{y} - \mathbf{S}\mathbf{h})^\top \mathbf{C}^{-1}(\mathbf{y} - \mathbf{S}\mathbf{h}) \right)$$

$$\frac{d}{d\mathbf{h}} (\mathbf{y} - \mathbf{S}\mathbf{h})^\top \mathbf{C}^{-1}(\mathbf{y} - \mathbf{S}\mathbf{h}) = -2\mathbf{S}^\top \mathbf{C}^{-1}(\mathbf{y} - \mathbf{S}\mathbf{h})$$

Gaussian Covariance: if $Y \sim \mathcal{N}(0, \sigma^2)$, $N \sim \mathcal{N}(0, \sigma^2)$:

$$\mathbf{C}_Y = \text{Cov}[Y, Y] = \mathbb{E}[(Y - \mu)(Y - \mu)^\top] = \mathbb{E}[Y Y^\top]$$

For Channel $Y = Sh + N$: $\mathbb{E}[Y Y^\top] = S \mathbb{E}[h h^\top] S^\top + \mathbb{E}[N N^\top]$

11.2. Multivariate Gaussian Distributions

A vector \mathbf{x} of n independent Gaussian random variables x_i is jointly Gaussian. If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{C}_{\mathbf{x}})$:

$$f_{\mathbf{x}}(\mathbf{x}) = f_{x_1, \dots, x_n}(x_1, \dots, x_n) = \frac{1}{\sqrt{\det(2\pi\mathbf{C}_{\mathbf{x}})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top \mathbf{C}_{\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\right)$$

Affine transformations $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ are jointly Gaussian with

$$\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_{\mathbf{x}} + \mathbf{b}, \mathbf{A}\mathbf{C}_{\mathbf{x}}\mathbf{A}^\top)$$

All marginal PDFs are Gaussian as well

Contour Lines

Ellipsoid with central point $\mathbb{E}[\mathbf{y}]$ and main axis are the eigenvectors of $\mathbf{C}_{\mathbf{y}}^{-1}$

11.3. Conditional Gaussian

$$\mathbf{A} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{A}}, \mathbf{C}_{\mathbf{A}}), \mathbf{B} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{B}}, \mathbf{C}_{\mathbf{B}})$$

$$\Rightarrow (\mathbf{A}|\mathbf{B}=\mathbf{b}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{A}|\mathbf{B}}, \mathbf{C}_{\mathbf{A}|\mathbf{B}})$$

Conditional Mean:

$$\mathbb{E}[\mathbf{A}|\mathbf{B}=\mathbf{b}] = \boldsymbol{\mu}_{\mathbf{A}|\mathbf{B}=\mathbf{b}} = \boldsymbol{\mu}_{\mathbf{A}} + \mathbf{C}_{\mathbf{AB}} \mathbf{C}_{\mathbf{BB}}^{-1} (\mathbf{b} - \boldsymbol{\mu}_{\mathbf{B}})$$

Conditional Variance:

$$\mathbf{C}_{\mathbf{A}|\mathbf{B}} = \mathbf{C}_{\mathbf{AA}} - \mathbf{C}_{\mathbf{AB}} \mathbf{C}_{\mathbf{BB}}^{-1} \mathbf{C}_{\mathbf{BA}}$$

11.4. Misc

If CDF of gaussian distribution given $\Phi(z) \sim \mathcal{N}(0, 1)$ then for $X \sim \mathcal{N}(1, 1)$ the CDF is given as $\Phi(x - \mu_x)$

12. Sequences

12.1. Random Sequences

Sequence of a random variable. Example: result of a dice is RV, roll a dice several times is a random sequence.

12.2. Markov Sequence $X_n : \Omega \rightarrow X_n$

Sequence of memoryless state transitions with certain probabilities.

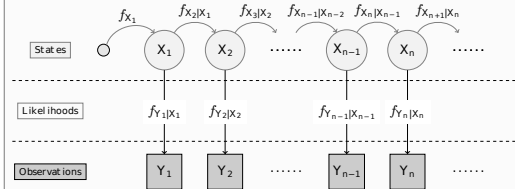
1. state: $f_{X_1}(x_1)$

2. state: $f_{X_2|X_1}(x_2|x_1)$

n. state: $f_{X_n|X_{n-1}}(x_n|x_{n-1})$

12.3. Hidden Markov Chains

Problem: states X_i are not visible and can only be guessed indirectly as a random variable Y_i .



Conditional pdf $f_{\mathbf{x}_n|\mathbf{y}_n}$ Likelihood pdf $f_{Y_n|X_n}$

State-transition pdf $f_{X_n|X_{n-1}}$

Estimation:

$$f_{\mathbf{x}_n|\mathbf{y}_n} \propto f_{\mathbf{y}_n|\mathbf{x}_n} \cdot \int_{\mathbf{x}} f_{\mathbf{x}_n|\mathbf{x}_{n-1}} \cdot f_{\mathbf{x}_{n-1}|\mathbf{y}_{n-1}} d\mathbf{x}_{n-1}$$

13. Recursive Estimation

13.1. Kalman-Filter

recursively calculates the most likely state from previous state estimates and current observation. Shows optimum performance for Gauss-Markov Sequences.

State space:

$$\mathbf{x}_n = \mathbf{G}_n \mathbf{x}_{n-1} + \mathbf{B}_n \mathbf{u}_n + \mathbf{v}_n$$

$$\mathbf{y}_n = \mathbf{H}_n \mathbf{x}_n + \mathbf{w}_n$$

With gaussian process/measurement noise $\mathbf{v}_n/\mathbf{w}_n$

Short notation: $\mathbb{E}[\mathbf{x}_n|\mathbf{y}_{n-1}] = \hat{\mathbf{x}}_{n|n-1}$ $\mathbb{E}[\mathbf{x}_n|\mathbf{y}_n] = \hat{\mathbf{x}}_{n|n}$

$$\mathbb{E}[\mathbf{y}_n|\mathbf{y}_{n-1}] = \hat{\mathbf{y}}_{n|n-1} \quad \mathbb{E}[\mathbf{y}_n|\mathbf{y}_n] = \hat{\mathbf{y}}_{n|n}$$

1. step: Prediction

$$\text{Mean: } \hat{\mathbf{x}}_{n|n-1} = \mathbf{G}_n \hat{\mathbf{x}}_{n-1|n-1}$$

$$\text{Covariance: } \mathbf{C}_{\mathbf{x}_{n|n-1}} = \mathbf{G}_n \mathbf{C}_{\mathbf{x}_{n-1|n-1}} \mathbf{G}_n^\top + \mathbf{C}_{\mathbf{v}}$$

2. step: Update

$$\text{Mean: } \hat{\mathbf{x}}_{n|n} = \hat{\mathbf{x}}_{n|n-1} + \mathbf{K}_n (\mathbf{y}_n - \mathbf{H}_n \hat{\mathbf{x}}_{n|n-1})$$

$$\text{Covariance: } \mathbf{C}_{\mathbf{x}_{n|n}} = \mathbf{C}_{\mathbf{x}_{n|n-1}} + \mathbf{K}_n \mathbf{H}_n \mathbf{C}_{\mathbf{x}_{n|n-1}}$$

$$\hat{\mathbf{x}}_{n|n} = \underbrace{\hat{\mathbf{x}}_{n|n-1}}_{\text{estimation } \mathbb{E}[X_n | Y_{n-1}=y_{n-1}]} + \underbrace{\mathbf{K}_n (\mathbf{y}_n - \mathbf{H}_n \hat{\mathbf{x}}_{n|n-1})}_{\text{correction: } \mathbb{E}[X_n | \Delta Y_n=y_n] \text{ innovation: } \Delta y_n}$$

With optimal Kalman-gain (prediction for \mathbf{x}_n based on Δy_n):

$$\mathbf{K}_n = \mathbf{C}_{\mathbf{x}_{n|n-1}} \mathbf{H}_n^\top (\mathbf{H}_n \mathbf{C}_{\mathbf{x}_{n|n-1}} \mathbf{H}_n^\top + \mathbf{C}_{\mathbf{w}_n})^{-1}$$

Innovation: closeness of the estimated mean value to the real value

$$\Delta \mathbf{y}_n = \mathbf{y}_n - \hat{\mathbf{y}}_{n|n-1} = \mathbf{y}_n - \mathbf{H}_n \hat{\mathbf{x}}_{n|n-1}$$

$$\text{Init: } \hat{\mathbf{x}}_{0|0} = \mathbb{E}[X_0] \quad \sigma_{0|0}^2 = \text{Var}[X_0]$$

$$\text{MMSE Estimator: } \hat{\mathbf{x}} = \int \mathbf{x}_n f_{X_n|Y_{(n)}}(\mathbf{x}_n|\mathbf{y}_{(n)}) d\mathbf{x}_n$$

For non linear problems: Suboptimum nonlinear Filters: Extended KF, Unscented KF, ParticleFilter

13.2. Extended Kalman (EKF)

Linear approximation of non-linear g, h

$$\mathbf{x}_n = g_n(\mathbf{x}_{n-1}, \mathbf{u}_n) \quad \mathbf{u}_n \sim \mathcal{N}$$

$$\mathbf{y}_n = h_n(\mathbf{x}_{n-1}, \mathbf{w}_n) \quad \mathbf{w}_n \sim \mathcal{N}$$

13.3. Unscented Kalman (UKF)

Approximation of desired PDF $f_{X_n|Y_n}(x_n|y_n)$ by Gaussian PDF.

13.4. Particle-Filter

For non linear state space and non-gaussian noise

Non-linear State space:

$$\mathbf{x}_n = g_n(\mathbf{x}_{n-1}, \mathbf{u}_n)$$

$$\mathbf{y}_n = h_n(\mathbf{x}_{n-1}, \mathbf{w}_n)$$

$$\text{Posterior Conditional PDF: } f_{X_n|Y_n}(x_n|y_n) \propto \underbrace{f_{Y_n|X_n}(y_n|x_n)}_{\text{likelihood}} \cdot \underbrace{f_{X_n|X_{n-1}}(x_n|x_{n-1})}_{\text{state transition}} \underbrace{f_{X_{n-1}|Y_{n-1}}(x_{n-1}|y_{n-1})}_{\text{last conditional PDF}} d\mathbf{x}_{n-1}$$

N random Particles with particle weight w_n^i at time n

$$\text{Monte-Carlo-Integration: } I = \mathbb{E}[g(X)] \approx I_N = \frac{1}{N} \sum_{i=1}^N \tilde{g}(x^i)$$

Importance Sampling: Instead of $f_X(x)$ use Importance Density $q_X(x)$

$$I_N = \frac{1}{N} \sum_{i=1}^N \tilde{w}^i g(x^i) \text{ with weights } \tilde{w}^i = \frac{f_X(x^i)}{q_X(x^i)}$$

$$\text{If } \int f_{X_n}(x) dx \neq 1 \text{ then } I_N = \sum_{i=1}^N \tilde{w}^i g(x^i)$$

13.5. Conditional Stochastic Independence

$$P(A \cap B|E) = P(A|E) \cdot P(B|E)$$

Given Y, X and Z are independent if

$$f_{Z|Y,X}(z|y,x) = f_{Z|Y}(z|y) \text{ or } f_{X,Z|Y}(x,z|y) = f_{Z|Y}(z|y) \cdot f_{X|Y}(x|y)$$

$$f_{Z|X,Y}(z|x,y) = f_{Z|Y}(z|y) \text{ or } f_{X|Z,Y}(x|z,y) = f_{X|Y}(x|y)$$

$$f_{Z|X,Y}(z|x,y) = f_{Z|Y}(z|y) \text{ or } f_{X|Z,Y}(x|z,y) = f_{X|Y}(x|y)$$

14. Hypothesis Testing

making a decision based on the observations

14.1. Definition

Null hypothesis $H_0 : \theta \in \Theta_0$ (Assumed first to be true)

Alternate hypothesis $H_1 : \theta \in \Theta_1$ (The one to proof)

Decision rule $\varphi : \mathbb{X} \rightarrow [0, 1]$ with

$\varphi(x) = 1$: decide for H_1 , $\varphi(x) = 0$: decide for H_0 Error level α with $E[d(X)|\theta] \leq \alpha, \forall \theta \in \Theta_0$

Error Type	Decision \ Reality	H_1 false (H_0 true)	H_1 true (H_0 false)
1 (FA) False Alarm	H_1 rejected (H_0 accepted)	True Negative $P = 1 - \alpha$	False Negative (Type 2) $P = \beta$
2 (DE) Detection Error	H_1 accepted (H_0 rejected)	False Positive (Type 1) $P = \alpha$	True Positive $P = 1 - \beta$

Power: Sensitivity/Recall/Hit Rate: $\frac{TP}{TP+FN} = 1 - \beta$

Specificity/True negative rate: $\frac{TN}{FP+TN} = 1 - \alpha$

Precision/Positive Prediction rate: $\frac{TP}{TP+FP}$

Accuracy: $\frac{TP+TN}{P+N} = \frac{2-\alpha-\beta}{2}$

14.1.1. Design of a test

Cost criterion $G_\varphi : \Theta \rightarrow [0, 1], \theta \mapsto E[d(X)|\theta]$

False Positive lower than α : $G_d(\theta)|_{\theta \in \Theta_0} \leq \alpha, \forall \theta \in \Theta_0$

False Negative small as possible: $\max\{G_d(\theta)|_{\theta \in \Theta_1}\}, \forall \theta \in \Theta_1$

14.2. Sufficient Statistics

Sufficiency for a test $T(X)$ means that no other test statistic, i.e., function of the observations \underline{x} , contains additional information about the parameter θ to be estimated:

$$f_{X|T}(x|T(x) = t, \theta) = f_{X|T}(x|T(x) = t)$$

15. Tests

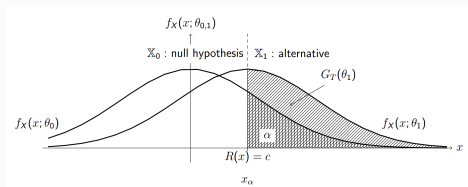
15.1. Neyman-Pearson-Test

The best test of P_0 against P_1 is

$$d_{NP}(x) = \begin{cases} 1 & R(x) > c \\ \gamma & R(x) = c \\ 0 & R(x) < c \end{cases} \quad \text{Likelihood-Ratio: } R(x) = \frac{f_X(x; \theta_1)}{f_X(x; \theta_0)}$$

$$\gamma = \frac{\alpha - P_0(\{R > c\})}{P_0(\{R = c\})} \quad \text{Errorlevel } \alpha$$

Steps: For α calculate x_α , then $c = R(x_\alpha)$



$$\text{Maximum Likelihood Detector: } d_{ML}(x) = \begin{cases} 1 & R(x) > 1 \\ 0 & \text{otherwise} \end{cases}$$

ROC Graphs: plot $G_d(\theta_1)$ as a function of $G_d(\theta_0)$

15.2. Bayes Test (MAP Detector)

Prior knowledge on possible hypotheses: $P(\{\theta \in \Theta_0\}) + P(\{\theta \in \Theta_1\}) = 1$, minimizes the probability of a wrong decision.

$$d_{\text{Bayes}} = \begin{cases} 1 & \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > \frac{c_0 P(\theta_0|x)}{c_1 P(\theta_1|x)} \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & P(\theta_1|x) > P(\theta_0|x) \\ 0 & \text{otherwise} \end{cases}$$

Risk weights c_0, c_1 are 1 by default.

If $P(\theta_0) = P(\theta_1)$, the Bayes test is equivalent to the ML test

$$\text{Loss Function } L(d(x), \theta) = \begin{cases} c_0 & \text{type 1 } d(x) = 1, \text{ but } \theta = \theta_0 \\ c_1 & \text{type 2 } d(x) = 0, \text{ but } \theta = \theta_1 \end{cases}$$

$$\text{risk}(d) = E[L(d(X), \theta)] = E[E[L(d(X), \theta)|x = X]]$$

$$\text{Multiple Hypothesis } d_{\text{Bayes}} = \begin{cases} 0 & x \in \mathbb{X}_0 \\ 1 & x \in \mathbb{X}_1 \\ 2 & x \in \mathbb{X}_2 \end{cases}$$

15.3. Linear Alternative Tests

$$d : \mathbb{X} \rightarrow \mathbb{R}, \underline{x} \mapsto \begin{cases} 1 & \underline{w}^\top \underline{x} - w_0 > 0 \\ 0 & \text{otherwise} \end{cases}$$

Estimate normal vector \underline{w}^\top , which separates \mathbb{X} into \mathbb{X}_0 and \mathbb{X}_1

$$\log R(\underline{x}) = \frac{\ln(\det(\underline{C}_0))}{\ln(\det(\underline{C}_1))} + \frac{1}{2}(\underline{x} - \underline{\mu}_0)^\top \underline{C}_0^{-1}(\underline{x} - \underline{\mu}_0) -$$

$$- \frac{1}{2}(\underline{x} - \underline{\mu}_1)^\top \underline{C}_1^{-1}(\underline{x} - \underline{\mu}_1) = 0$$

$$\text{For 2 Gaussians, with } \underline{C}_0 = \underline{C}_1 = \underline{C}: \underline{w}^\top = (\underline{\mu}_1 - \underline{\mu}_0)^\top \underline{C}$$

$$\text{and constant translation } w_0 = \frac{(\underline{\mu}_1 - \underline{\mu}_0)^\top \underline{C}(\underline{\mu}_1 - \underline{\mu}_0)}{2}$$

