Lecture Notes

Statistical Signal Processing

Univ.-Prof. Dr.-Ing. Wolfgang Utschick

Technische Universität München

Summer 2019

Email: utschick@tum.de

Layout by LaTeX2$_\varepsilon$

Part I

Parameter Estimation

1. Statistic Modeling

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

The basic problem of statistical estimation is to infer the Probability Measure P, which the realizations of the respective random variables $X : \Omega \to \mathbb{X}$ are subject to.

The most critical part of any parameter estimation problem is the choice of a proper Statistic Model $(\Omega, \mathbb{F}, \mathsf{P}_\theta)$, with the metric space $(\mathbb{X}, \mathbb{B})$ and

$$\text{Observation Space}: \quad \mathbb{X}, \tag{1.1}$$
$$\text{Sigma Algebra}: \quad \mathbb{F}, \tag{1.2}$$
$$\text{Probability Measure}: \quad \mathsf{P}_\theta,\ \theta \in \Theta. \tag{1.3}$$

The stochastic model is a set of Probability Spaces and the task of statistic estimation is to select the most appropriate candidate.

1.1 Standard Model

Definition. We call the introduced statistical model STANDARD MODEL and the inference problem PARAMETER ESTIMATION, if

$$\Theta \subset \mathbb{R}^D, \tag{1.4}$$

and the random variable X is either DISCRETE or CONTINUOUS.

Commonly used terminology:

- Random variables $X_i : \Omega \to \mathbb{X}$ are STATISTICS, and
- its realizations x_i of X_i are OBSERVATIONS, SAMPLES, MEASUREMENTS, etc.

Definition. A special statistic is the random variable $T : \mathbb{X} \to \Theta$, which maps one or multiple samples to $\theta \in \Theta$ or another parameter depending thereof. The random variable or statistic $T : \mathbb{X} \to \Theta$ is called ESTIMATOR.

1.2 Introductory Example

Given observations $X_1, \ldots, X_N$ of a uniquely distributed random variable

$$X : \Omega \to [0, \theta], \tag{1.5}$$

with $[0, \theta] \subset \mathbb{R}$ such that $F_X(\xi) = \frac{\xi}{\theta}$ and $f_X(\xi) = \frac{1}{\theta}$, if $0 \leq \xi \leq \theta$.

The unknown parameter θ, which describes the random variable X is DETERMINISTIC and UNKNOWN.

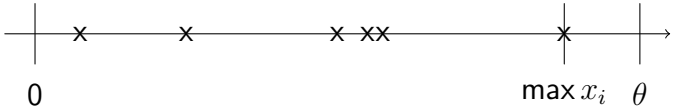


Fig. 1.1: Estimating the upper bound of an intervall.

How to estimate the upper bound? Any guesses?

How to estimate the upper bound?

1. Attempt: Given $\mathrm{E}[X] = \theta/2$, we conclude for the statistics X_i

$$T_1 = 2 \cdot \underbrace{\frac{1}{N} \sum_{i=1}^{N} X_i}_{\text{Average}} : \quad x_1, \ldots, x_N \mapsto \hat{\theta}_1. \tag{1.6}$$

2. Attempt: Since for large N the maximum observed value will be close to the upper bound, we conclude

$$T_2 = \max_{i=1,\ldots,N} \{X_i\} : \quad x_1, \ldots, x_N \mapsto \hat{\theta}_2. \tag{1.7}$$

How realiable are these attempts?

Estimator T_1: According to the Law of Large Numbers (Chebyshev Inequality):

$$P_\theta\left(\left|\frac{T_1}{2}-\frac{\theta}{2}\right|\geq\epsilon\right)\leq\frac{\mathsf{Var}\left[X\right]}{N\epsilon^2}\underset{N\to\infty}{\longrightarrow}0. \tag{1.8}$$

Estimator T_2: Again with an asymptotic approach:

$$\begin{aligned}P_\theta\left(|T_2-\theta|\geq\epsilon\right)&=P_\theta\left(X_1\leq\theta-\epsilon,\ldots,X_N\leq\theta-\epsilon\right)\\&=\prod_{i=1}^{N}P_\theta\left(X_i\leq\theta-\epsilon\right)\\&=\prod_{i=1}^{N}\frac{\theta-\epsilon}{\theta}\\&=\left(\frac{\theta-\epsilon}{\theta}\right)^N\\&=\underbrace{\left(1-\frac{\epsilon}{\theta}\right)}_{<1}^{N}\underset{N\to\infty}{\longrightarrow}0.\end{aligned} \tag{1.9}$$

2. Quality Criteria

2.1 Consistency and Unbiasedness

Definition. CONSISTENT ESTIMATORS are characterized by

$$\lim_{N\to\infty} T(x_1, \ldots, x_N) = \theta. \tag{2.1}$$

Obviously, T_1 and T_2 are consistent.

Definition. UNBIASED ESTIMATORS are characterized by

$$\mathsf{E}\,[T(X_1, \ldots, X_N)] = \theta. \tag{2.2}$$

T_1 and T_2 are random variables which depend on the randomly drawn observations $X_1, \ldots, X_N$. UNBIASEDNESS means that averaging over the entire sample set gets the true parameter:

- T_1 is UNBIASED, since

$$\mathsf{E}\left[T_1(X_1, \ldots, X_N)\right] = \mathsf{E}\left[\frac{2}{N}\sum_{i=1}^{N} X_i\right] = \frac{2}{N}\sum_{i=1}^{N}\mathsf{E}\left[X_i\right] = \frac{2}{N}\cdot N \cdot \frac{\theta}{2} = \theta. \tag{2.3}$$

- T_2 is ASYMPTOTICALLY UNBIASED, since

$$\mathsf{E}\left[T_2(X_1, \ldots, X_N)\right] = \int_0^{\theta} \xi \cdot f_{T_2}(\xi)\ \mathrm{d}\xi = \frac{N}{N+1}\theta, \tag{2.4}$$

  where

$$f_{T_2}(\xi) = \frac{\mathrm{d}}{\mathrm{d}\xi}F_{T_2}(\xi) = \frac{\mathrm{d}}{\mathrm{d}\xi}\left(\frac{\xi}{\theta}\right)^N = N \cdot \frac{\xi^{N-1}}{\theta^N}, \quad \xi \in [0, \theta]. \tag{2.5}$$

- Consequently, we introduce an unbiased version of estimator T_2' by

$$T_2' = \frac{N+1}{N}T_2. \tag{2.6}$$

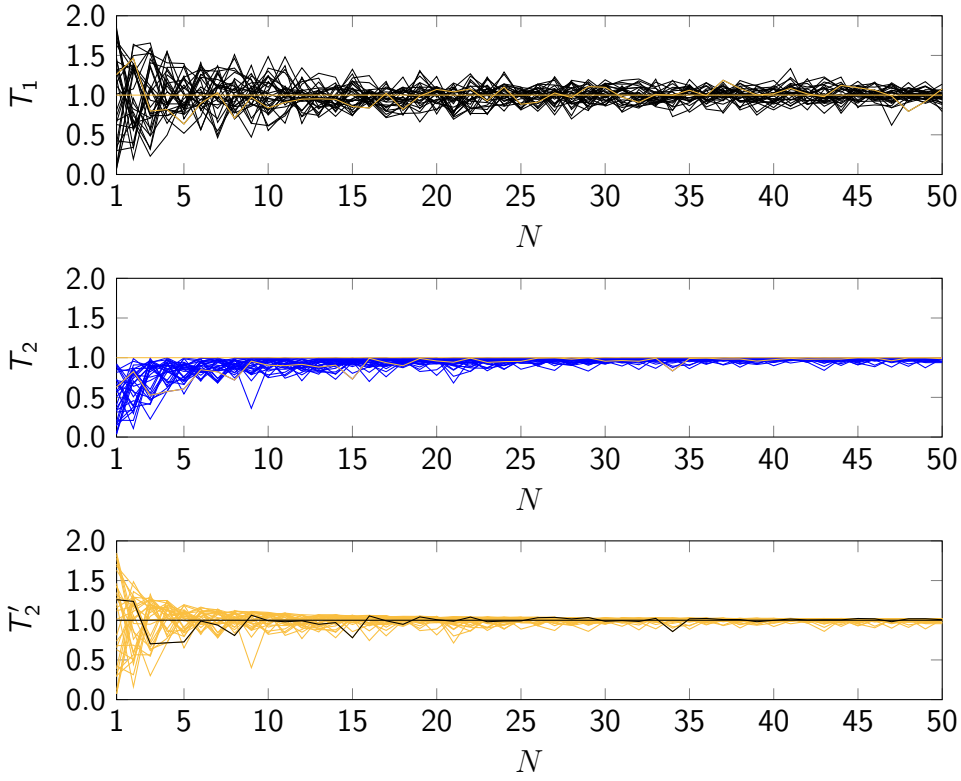


Fig. 2.1: T_1, T_2, and T_2' for 30 uniformly chosen (independent and identically distributed) samples of size $N = 1 \ldots 50$ with $\theta = 1$.

2.2 Variance

Definition. A further quality measure for an estimator is its VARIANCE

$$\mathrm{Var}\left[T(X_1,\ldots,X_N)\right] = \mathrm{E}\left[\left(T-\mathrm{E}\left[T\right]\right)^2\right]. \tag{2.7}$$

For the estimators T_1, T_2 and T_2' of the introductory example we obtain

$$\mathrm{Var}\left[T_1\right] = \frac{\theta^2}{3N}, \tag{2.8}$$

$$\mathrm{Var}\left[T_2\right] = \frac{N\theta^2}{(N+1)^2(N+2)}, \tag{2.9}$$

$$\mathrm{Var}\left[T_2'\right] = \left(\frac{N+1}{N}\right)^2 \mathrm{Var}\left[T_2\right] \tag{2.10}$$

$$= \frac{\theta^2}{N(N+2)}. \tag{2.11}$$

Table 2.1: Variances of the estimators T_1, T_2 and T_2' for $N = 10$ observations.

Estimator T	Variance σ_T^2
T_1	$\frac{5}{150}\theta^2$
T_2	$\frac{5}{726}\theta^2$
T_2'	$\frac{5}{600}\theta^2$

2.3 Mean Square Error

Definition. An extension of the VARIANCE is the MEAN SQUARE ERROR (MSE)

$$\varepsilon[T] = \mathsf{E}\left[(T-\theta)^2\right]. \tag{2.12}$$

For the estimators T_1 and T_2' of the introductory example the MSE is already known, since these estimators are unbiased and thus $\mathsf{E}[T_1] = \mathsf{E}[T_2'] = \theta$ and $\varepsilon[T] = \mathsf{Var}[T]$.

The MSE of the 2nd estimator can be obtained by

$$\begin{aligned}\varepsilon[T_2] &= \mathsf{E}\left[(T_2-\theta)^2\right] \\ &= \mathsf{E}\left[(T_2)^2\right] - 2\,\mathsf{E}[T_2]\,\theta + \theta^2 \\ &= \dots \\ &= \frac{2\theta^2}{(N+2)(N+1)}.\end{aligned} \tag{2.13}$$

Table 2.2: MSEs of the estimators T_1, T_2 and T_2' for $N = 10$ observations.

Estimator T	MSE $\varepsilon[T]$
T_1	$\frac{1}{30}\theta^2$
T_2	$\frac{1}{66}\theta^2$
T_2'	$\frac{1}{120}\theta^2$

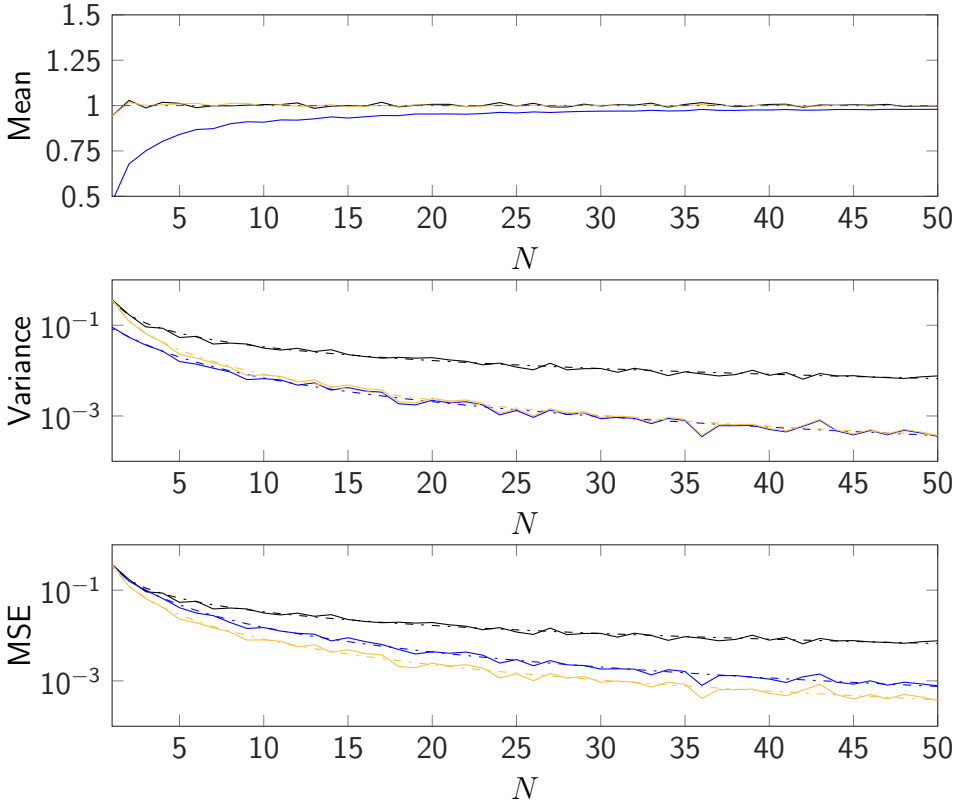


Fig. 2.2: Estimated (solid) and analytical (dashed-dotted) MEAN, VARIANCE, and MSE for T_1 (black), T_2 (blue), and T_2' (yellow) for 30 uniformly chosen (i.i.d.) samples of size $N = 1 \ldots 50$ with $\theta = 1$.

2.4 Bias/Variance Trade–Off

The MSE of an estimator T,

$$\varepsilon[T] = \mathsf{E}\left[(T(X_1, \ldots, X_N) - \theta)^2\right], \tag{2.14}$$

can be decomposed in its BIAS and its VARIANCE, i.e.,[1]

$$\varepsilon[T] = \underbrace{\mathsf{E}\left[(T - \mathsf{E}[T])^2\right]}_{\mathsf{Var}[T]} + \left(\underbrace{\mathsf{E}[T] - \theta}_{\mathsf{Bias}[T]}\right)^2 \tag{2.15}$$

$$= \mathsf{Var}[T] + (\mathsf{Bias}[T])^2. \tag{2.16}$$

Bias and Variance of an estimator cannot be minimized independently.

[1] The decomposition can easily be shown by taking the expectation of the expansion of $(T - \mathsf{E}[T] + \mathsf{E}[T] - \theta)^2$.

2.5 Introductory Example (cont'd)

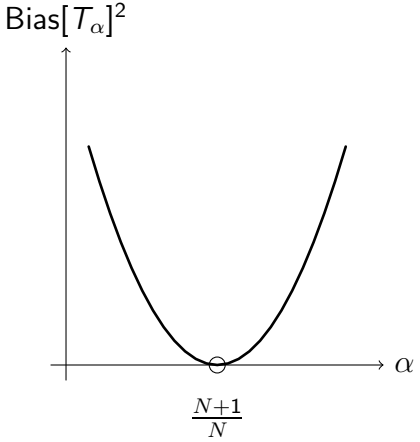
Another estimator T_α can be constructed by

$$T_\alpha = \alpha \cdot T_2 : \quad x_1, \ldots, x_N \to \hat{\theta}_\alpha, \tag{2.17}$$

with

$$\text{Bias}[T_\alpha] = \text{E}[T_\alpha] - \theta = \alpha\,\text{E}[T_2] - \theta = \left(\frac{\alpha N}{N+1} - 1\right)\theta. \tag{2.18}$$

The estimator T_α is obviously unbiased if $\alpha = \frac{N+1}{N}$.



3. Maximum Likelihood Estimation

3.1 Maximum Likelihood Principle

Given the statistical model $\{\mathbb{X}, \mathbb{F}, \mathsf{P}_\theta; \theta \in \Theta\}$, the MAXIMUM-LIKELIHOOD (ML) principle suggest to select a candidate probability measure P_θ, such the observed outcomes of the experiment become most probable.

A maximum likelihood estimator T_{ML} picks the $\hat{\theta} \in \Theta$ which maximizes the LIKELIHOOD FUNCTION, i.e.

$$T_{\mathsf{ML}} : x_1, \ldots, x_N \mapsto \underset{\theta \in \Theta}{\mathsf{argmax}} \{L(x_1, \ldots, x_N; \theta)\}. \tag{3.1}$$

Definition. The LIKELIHOOD FUNCTION depends on the statistical model $\{\mathbb{X}, \mathbb{F}, \mathsf{P}_\theta; \theta \in \Theta\}$, i.e.,

$$\text{DISCRETE R.V.}: \quad L(x_1, \ldots, x_N; \theta) = \mathsf{P}_\theta(x_1, \ldots, x_N), \tag{3.2}$$

$$\text{CONTINUOUS R.V.}: \quad L(x_1, \ldots, x_N; \theta) = f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \theta). \tag{3.3}$$

Assuming that all observations are drawn from IDENTICALLY INDEPENDENTLY DISTRIBUTED (i.i.d.) random variables $X_1, \ldots, X_N$,

$$\mathsf{P}_\theta(x_1, \ldots, x_N) = \prod_{i=1}^{N} \mathsf{P}_\theta(x_i), \tag{3.4}$$

$$f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \theta) = \prod_{i=1}^{N} f_{X_i}(x_i; \theta), \tag{3.5}$$

with $f_{X_1} = \cdots = f_{X_N}$.

Log-Likelihood Function.

Due to the monotonicity of the log-function and assuming i.i.d. random variables, we obtain the following expressions:

$$T_{\mathsf{ML}} : x_1, \ldots, x_N \mapsto \quad \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \log(L(x_1, \ldots, x_N; \theta)) \right\} \tag{3.6}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \log(\prod_{i=1}^{N} L(x_i; \theta)) \right\} \tag{3.7}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \log(L(x_i; \theta)) \right\}. \tag{3.8}$$

Note. The computation of the maximum likelihood estimate obviously involves the solution of an optimization problem.

3.2 Channel Estimation Example

We consider the estimation of the attenuation coefficient h of a Single-Input Single-Output transmission channel with ADDITIVE WHITE GAUSSIAN NOISE (AWGN) $N_i \sim \mathcal{N}(0, \sigma^2)$ at the receiver.

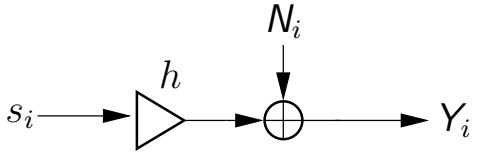


Fig. 3.1: Channel model.

The model of the receiver signal is

$$Y_i = hs_i + N_i, \qquad Y_i \sim \mathcal{N}(hs_i, \sigma^2), \tag{3.9}$$

where s_i denotes the i-th training signal.

Assuming N observations $y_1, \ldots, y_N$ according to the training signals $s_1, \ldots, s_N$, the likelihood function is given by $L(y_1, \ldots, y_N; \theta) = \prod_{i=1}^{N} f_{Y_i}(y_i; \theta)$, with $\theta = h$ and

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \theta s_i\right)^2\right). \tag{3.10}$$

The maximum-likelihood estimation is derived by

$$\begin{aligned}
\hat{h}_{\mathsf{ML}} &= \underset{\theta\in\mathbb{R}}{\mathsf{argmax}} \left\{ \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} \cdot (y_i - \theta s_i)^2 \right) \right) \right\} \\
&= \underset{\theta\in\mathbb{R}}{\mathsf{argmax}} \left\{ \sum_{i=1}^{N} -\frac{1}{2\sigma^2} \cdot (y_i - \theta s_i)^2 \right\} \\
&= \underset{\theta\in\mathbb{R}}{\mathsf{argmin}} \left\{ \sum_{i=1}^{N} \frac{1}{2\sigma^2} \cdot (y_i - \theta s_i)^2 \right\} \\
&= \underset{\theta\in\mathbb{R}}{\mathsf{argmin}} \left\{ \frac{1}{2\sigma^2} \cdot ||\boldsymbol{y} - \theta \boldsymbol{s}||^2 \right\} \\
&= \underset{\theta\in\mathbb{R}}{\mathsf{argmin}} \left\{ ||\boldsymbol{y} - \theta \boldsymbol{s}||^2 \right\}. \qquad (3.11)
\end{aligned}$$

where $\boldsymbol{s} = [s_1, \ldots, s_N]^{\mathsf{T}}$. The respective estimator $\hat{h}_{\mathsf{ML}}$ is given by the PSEUDO INVERSE $(\boldsymbol{s}^{\mathsf{T}}\boldsymbol{s})^{-1}\boldsymbol{s}^{\mathsf{T}}$ of the vector $\boldsymbol{s}$ of training signals,[1] i.e.,

$$\hat{h}_{\mathsf{ML}} = (\boldsymbol{s}^{\mathsf{T}}\boldsymbol{s})^{-1}\boldsymbol{s}^{\mathsf{T}}\boldsymbol{y}. \qquad (3.12)$$

The ML estimator is obviously identical with the LEAST SQUARES estimator. This changes considerably when the statistics Y_i are correlated or when N is non-Gaussian distributed.

[1] Using the optimality condition $\frac{\mathrm{d}}{\mathrm{d}\theta}||\boldsymbol{y} - \theta\boldsymbol{s}||^2 = -2\boldsymbol{s}^{\mathsf{T}}\boldsymbol{y} + 2\theta\boldsymbol{s}^{\mathsf{T}}\boldsymbol{s} = 0$, we obtain $\theta_{\mathsf{ML}} = \frac{\boldsymbol{s}^{\mathsf{T}}\boldsymbol{y}}{\boldsymbol{s}^{\mathsf{T}}\boldsymbol{s}}$.

3.3 Introductory Example (cont'd)

The likelihood function for a single observation ($N = 1$) reads

$$\mathsf{L}(x;\theta) = \begin{cases} \frac{1}{\theta}, & x \leq \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

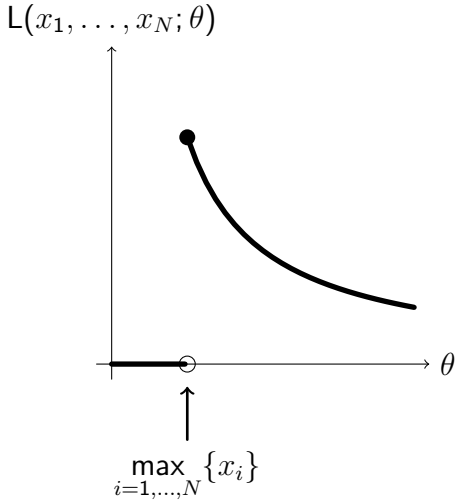
If $N > 1$:

$$\mathsf{L}(x_1,\ldots,x_N;\theta) = \begin{cases} \frac{1}{\theta^N}, & x_i \leq \theta \quad \forall i = 1,\ldots,N \\ \\ 0 & \text{otherwise.} \end{cases}$$

The ML Estimator is given by

$$\begin{aligned} \hat{\theta} &= \underset{\theta\in\mathbb{R}}{\mathsf{argmax}} \left\{\mathsf{L}(x_1,\ldots,x_N;\theta)\right\} \\ &= \max_{i=1,\ldots,N} \{x_1,\ldots,x_N\}, \end{aligned}$$

which corresponds to the 2nd intuitive attempt.



3.4 Bernoulli Experiment

We now consider a BERNOULLI EXPERIMENT with success probability θ, e.g. the transmission of a data packet over a link with probabilities

$$\mathsf{P}\,(\text{no erasure}) = \theta, \tag{3.13}$$

$$\mathsf{P}\,(\text{erasure}) = 1 - \theta. \tag{3.14}$$

We now study the maximum-likelihood framework for estimating the unknown parameter θ based on N independent observations:

The statistical model $\{\mathbb{X},\, \mathbb{F},\, \mathsf{P}_\theta;\, \theta \in \Theta\}$ is

$$\{\underbrace{\{0, 1, \ldots, N\}}_{\substack{\text{number of successes} \\ \text{over } N \text{ observations}}}, \{0, 1, \ldots, N\}^N, \mathsf{B}_{N,\theta};\, \theta \in [0, 1]\},$$

and the respective random variable X, which counts the number of successful Bernoulli trials within N attempts. X is BINOMIALLY DISTRIBUTED according to

$$\mathsf{B}_{N,\theta}(x) = \binom{N}{x} \theta^x (1-\theta)^{N-x}, \quad x \in \{0, 1, \ldots, N\}.$$

Given the likelihood function

$$\mathsf{L}(x;\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x},$$

the log-likelihood function reads

$$\log \mathsf{L}(x;\theta) = \log\binom{N}{x} + x\log\theta + (N-x)\log(1-\theta).$$

The ML-Estimator can be obtained by

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\,\{\log \mathsf{L}(x;\theta)\},$$

i.e.,

$$\begin{aligned} & \frac{\mathsf{d}}{\mathsf{d}\theta}\log \mathsf{L}(x;\theta)\Big|_{\theta=\hat{\theta}} = 0 \\ \Leftrightarrow \quad & \frac{x}{\hat{\theta}} + (N-x)\frac{1}{1-\hat{\theta}}(-1) = 0 \\ \Leftrightarrow \quad & T_{\mathsf{ML}} : x \mapsto \hat{\theta} = \frac{x}{N}. \end{aligned}$$

Quality of $T_{\mathsf{ML}} : x \mapsto \frac{x}{N}$.

Since X is binomially distributed

$$\mathsf{E}[X] = N\theta, \tag{3.15}$$
$$\mathsf{Var}[X] = N\theta(1-\theta), \tag{3.16}$$

we obtain

$$\mathsf{E}[T_{\mathsf{ML}}] = \mathsf{E}\left[\frac{X}{N}\right] = \frac{1}{N}\mathsf{E}[X] = \frac{1}{N}\theta N = \theta, \tag{3.17}$$
$$\mathsf{Var}[T_{\mathsf{ML}}] = \mathsf{Var}\left[\frac{X}{N}\right] = \frac{1}{N^2}\mathsf{Var}[X] = \frac{1}{N^2}N\theta(1-\theta) = \frac{\theta(1-\theta)}{N}. \tag{3.18}$$

Obviously, the ML Estimator of the success probability θ is unbiased, and the MSE is equal to the variance of the estimator,

$$\mathsf{Bias}[T_{\mathsf{ML}}] = 0, \tag{3.19}$$
$$\varepsilon[T_{\mathsf{ML}}] = \mathsf{Var}[T_{\mathsf{ML}}] = \frac{\theta(1-\theta)}{N}. \tag{3.20}$$

Can we improve the estimator?

Alternative Solution: $T' : x \mapsto \frac{x+1}{N+2}$.

$$\mathsf{E}\left[T'\right] = \mathsf{E}\left[\frac{X+1}{N+2}\right] = \frac{N\theta+1}{N+2} \quad \Rightarrow \mathsf{Bias}\left[T'\right] = \frac{1-2\theta}{N+2}, \tag{3.21}$$

$$\begin{aligned}
\varepsilon\left[T'\right] &= \mathsf{Var}\left[T'\right] + \mathsf{Bias}\left[T'\right]^2 \\
&= \mathsf{Var}\left[\frac{X+1}{N+2}\right] + \left(\frac{1-2\theta}{N+2}\right)^2 \\
&= \frac{\mathsf{Var}\left[X+1\right]}{(N+2)^2} + \left(\frac{1-2\theta}{N+2}\right)^2 \\
&= \frac{\mathsf{Var}\left[X\right]}{(N+2)^2} + \left(\frac{1-2\theta}{N+2}\right)^2 \\
&= \frac{N\theta(1-\theta) + (1-2\theta)^2}{(N+2)^2}.
\end{aligned} \tag{3.22}$$

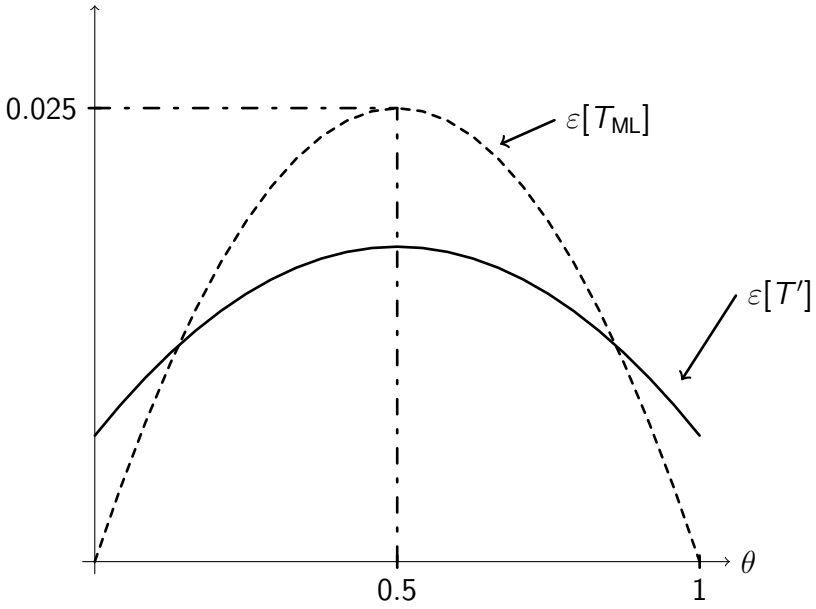


Fig. 3.2: Comparison of the achievable MSEs for $N = 10$ observations.

Note. Biased estimators can provide better estimates than unbiased estimators.

3.5 Best Unbiased Estimator

ML Estimators are not necessarily the best estimators. A wide class of estimators is defined by minimizing the MSE under an unbiasedness constraint.

Definition. Given the statistical model $\{\mathbb{X}, \mathbb{F}, \mathsf{P}_\theta; \theta \in \Theta\}$, we call an estimator BEST UNBIASED ESTIMATOR if $\mathsf{E}[T(X_1, \ldots, X_N)] = \theta$ and

$$\mathsf{Var}[T(X_1, \ldots, X_N)] \leq \mathsf{Var}[S(X_1, \ldots, X_N)], \quad \forall \theta \in \Theta, \tag{3.23}$$

for any alternative unbiased estimator S.

Best unbiased estimators are also referred to as UNIFORMLY MINIMUM VARIANCE UNBIASED (UMVU) Estimators.

The estimator $T_2' = \frac{N+1}{N} T_2$ of the introductory example is a UMVU Estimator.

4. Fisher's Information Inequality

An universal lower bound for the variance of an estimator can be introduced, if the following conditions are fulfilled:

$$\mathsf{L}(x;\theta) > 0, \quad \forall x \in \mathbb{X}, \theta \in \Theta, \tag{4.1}$$

$$\mathsf{L}(x;\theta) \quad \text{differentiable with respect to } \theta, \tag{4.2}$$

$$\int_{\mathbb{X}} \frac{\partial}{\partial\theta}\mathsf{L}(x;\theta)\ \mathrm{d}x = \frac{\partial}{\partial\theta}\int_{\mathbb{X}} \mathsf{L}(x;\theta)\ \mathrm{d}x. \tag{4.3}$$

Definition. Consequently, we define the SCORE FUNCTION as

$$g(x;\theta) = \frac{\partial}{\partial\theta}\log \mathsf{L}(x;\theta) = \frac{\frac{\partial}{\partial\theta}\mathsf{L}(x;\theta)}{\mathsf{L}(x;\theta)}, \tag{4.4}$$

with $\mathsf{E}\left[g(X;\theta)\right] = 0$.

4.1 Cramér-Rao Lower Bound

Given the score function $g(x;\theta)$ of an estimation problem, and the FISHER INFORMATION term $I_{\mathsf{F}}(\theta) = \mathrm{Var}\left[g(X,\theta)\right]$, the variance of any possible estimator can be lower bounded by

$$\mathrm{Var}\left[T(X)\right] \geq \left(\frac{\partial\, \mathrm{E}\left[T(X)\right]}{\partial\theta}\right)^2 \frac{1}{I_{\mathsf{F}}(\theta)}, \tag{4.5}$$

and

$$\mathrm{Var}\left[T(X)\right] \geq \frac{1}{I_{\mathsf{F}}(\theta)}, \tag{4.6}$$

if T is unbiased.

Proof.

$$\begin{aligned}
\frac{\partial \mathsf{E}\left[T(X)\right]}{\partial\theta} &= \frac{\partial}{\partial\theta}\int_{\mathbb{X}} T(x) f_X(x;\theta)\mathrm{d}x = \int_{\mathbb{X}} T(x)\frac{\partial}{\partial\theta} f_X(x;\theta)\mathrm{d}x \\
&= \int_{\mathbb{X}} (T(x) - \mathsf{E}\left[T(X)\right])\frac{\partial}{\partial\theta} f_X(x;\theta)\mathrm{d}x + \underbrace{\int_{\mathbb{X}} \mathsf{E}\left[T(X)\right]\frac{\partial}{\partial\theta} f_X(x;\theta)\mathrm{d}x}_{=0} \\
&= \int_{\mathbb{X}} (T(x) - \mathsf{E}\left[T(X)\right])\sqrt{f_X(x;\theta)}\frac{1}{\sqrt{f_X(x;\theta)}}\frac{\partial}{\partial\theta} f_X(x;\theta)\mathrm{d}x \\
&= \left\langle (T(x) - \mathsf{E}\left[T(X)\right])\sqrt{f_X(x;\theta)}, \frac{1}{\sqrt{f_X(x;\theta)}}\frac{\partial}{\partial\theta} f_X(x;\theta)\right\rangle. \qquad (4.7)
\end{aligned}$$

Appling the Cauchy-Schwarz Inequality, $|\langle a,b\rangle|^2 \leq \langle a,a\rangle \cdot \langle b,b\rangle$, yields:

$$\begin{aligned}
\left(\frac{\partial \mathsf{E}\left[T(X)\right]}{\partial\theta}\right)^2 &\leq \int_{\mathbb{X}} (T(x) - \mathsf{E}\left[T(X)\right])^2 f_X(x;\theta)\mathrm{d}x \;\cdot\; \int_{\mathbb{X}} \left(\frac{\partial}{\partial\theta} f_X(x;\theta)\right)^2 \frac{f_X(x;\theta)}{(f_X(x;\theta))^2}\mathrm{d}x \\
&= \mathsf{Var}\left[T(X)\right] \cdot \int_{\mathbb{X}} \left(g(x,\theta)\right)^2 f_X(x;\theta)\mathrm{d}x. \qquad (4.8)
\end{aligned}$$

$$\Rightarrow \mathsf{Var}\left[T(X)\right] \geq \left(\frac{\partial \mathsf{E}\left[T(X)\right]}{\partial\theta}\right)^2 \frac{1}{\mathsf{Var}\left[g(X,\theta)\right]}. \qquad (4.9)$$

4.2 2nd Version of Fisher Information

An alternative expression for the Fisher Information term $I_{\mathsf{F}}(\theta)$ can be derived,

$$
\begin{aligned}
I_{\mathsf{F}}(\theta) &= \mathsf{Var}\left[g(X,\theta)\right] \\
&= \mathsf{E}\left[g(X,\theta)^2\right] - \underbrace{\mathsf{E}\left[g(X,\theta)\right]^2}_{=0} \\
&= \mathsf{E}\left[g(X,\theta)^2\right] && (4.10) \\
&= -\,\mathsf{E}\left[\frac{\partial^2}{\partial\theta^2}\log \mathsf{L}(X;\theta)\right], && (4.11)
\end{aligned}
$$

which can be interpreted as the negative Mean Curvature of the Log-Likelihood Function.

The last step from (4.10) to (4.11) is not obvious.

Proof.

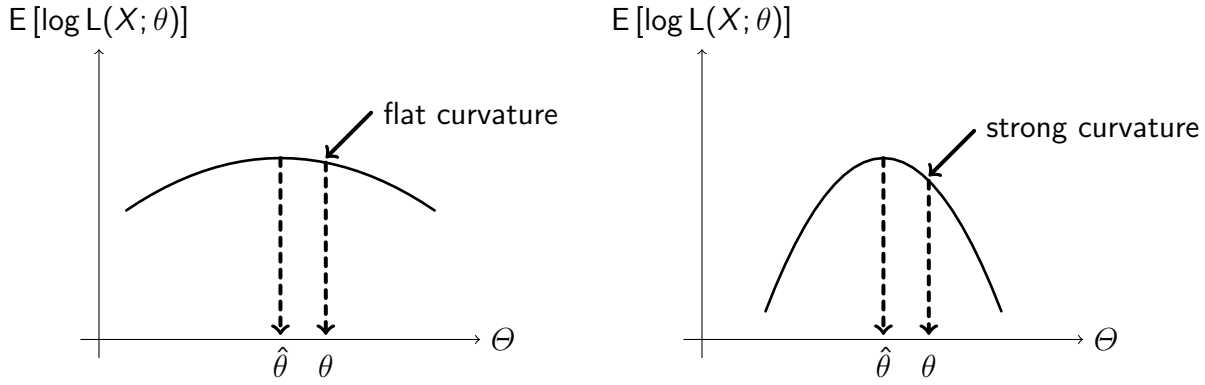
$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\log \mathsf{L}(x;\theta) &= \frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\log \mathsf{L}(x;\theta)\right)\\
&= \frac{\partial}{\partial\theta}\left(\frac{1}{\mathsf{L}(x;\theta)}\frac{\partial}{\partial\theta}\mathsf{L}(x;\theta)\right)\\
&= \frac{\frac{\partial^2}{\partial\theta^2}\mathsf{L}(x;\theta)}{\mathsf{L}(x;\theta)} - \frac{\partial}{\partial\theta}\mathsf{L}(x;\theta)\frac{\frac{\partial}{\partial\theta}\mathsf{L}(x;\theta)}{\mathsf{L}(x;\theta)^2}\\
&= \frac{\frac{\partial^2}{\partial\theta^2}\mathsf{L}(x;\theta)}{\mathsf{L}(x;\theta)} - \frac{\left(\frac{\partial}{\partial\theta}\mathsf{L}(x;\theta)\right)^2}{\mathsf{L}(x;\theta)^2}\\
&= \frac{\frac{\partial^2}{\partial\theta^2}\mathsf{L}(x;\theta)}{\mathsf{L}(x;\theta)} - g(x,\theta)^2. \qquad (4.12)
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \mathsf{E}\left[g(X,\theta)^2\right] &= \mathsf{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}\mathsf{L}(X;\theta)}{\mathsf{L}(X;\theta)} - \frac{\partial^2}{\partial\theta^2}\log \mathsf{L}(X;\theta)\right]\\
&= \underbrace{\int_{\mathbb{X}} \frac{\partial^2}{\partial\theta^2}\mathsf{L}(x;\theta)\mathrm{d}x}_{\frac{\partial^2}{\partial\theta^2}\int_{\mathbb{X}}\mathsf{L}(x;\theta)\mathrm{d}x=\frac{\partial^2}{\partial\theta^2}1=0} - \mathsf{E}\left[\frac{\partial^2}{\partial\theta^2}\log \mathsf{L}(X;\theta)\right]. \qquad (4.13)
\end{aligned}$$

Interpretation.

The 2nd version of the Fisher Information $I_{\mathsf{F}}(\theta)$ as the negative Mean Curvature of the Log-Likelihood Function allows an interpretation of the Cramér-Rao Lower Bound.

$$I_{\mathsf{F}}(\theta) = -\,\mathsf{E}\left[\underbrace{\frac{\partial^2}{\partial\theta^2}\log \mathsf{L}(X;\theta)}_{\text{curvature}}\right].$$



Properties of the Fisher Information:

- The log-likelihood function $\log(L(x_1, \ldots, x_N))$ depends on given observations $x_1, \ldots, x_N$.
- A weak curvature of $\log(L(x_1, \ldots, x_N))$ corresponds to little information in $x_1, \ldots, x_N$.
- A strong curvature of $\log(L(x_1, \ldots, x_N))$ indicates more information in $x_1, \ldots, x_N$.
- The Fisher information $I_F(\theta)$ refers to the negative mean curvature at the true parameter θ.
- $I_F(\theta)$ is based on the expectation of the negative mean curvature of $\log(L(X_1, \ldots, X_N))$ with respect to the observation statistics $X_1, \ldots, X_N$.
- $I_F(\theta)$ is a function of θ.
- $I_F(\theta)$ is monotonically increasing with the number of independent observation statistics, i.e., given the Fisher Information $I_F^{(1)}(\theta)$ for a single observation statistic ($N = 1$), then

$$I_F^{(N)}(\theta) = N \cdot I_F^{(1)}(\theta). \tag{4.14}$$

4.3 Exponential Models

Definition. A EXPONENTIAL MODEL is a statistical model $\{\mathbb{X}, \mathbb{F}, \mathsf{P}_\theta; \theta \in \Theta\}$ with random variable X and

$$f_X(x;\theta) = \frac{h(x)\exp(a(\theta)t(x))}{\exp(b(\theta))}, \tag{4.15}$$

with $b(\theta) = \log \int_{\mathbb{X}} h(x)\exp(a(\theta)t(x))\mathrm{d}x$ for normalization.

Then the respective Fisher Information can be directly obtained by

$$I_{\mathsf{F}}(\theta) = \frac{\partial a(\theta)}{\partial \theta}\frac{\partial \,\mathsf{E}\,[t(X)]}{\partial \theta}. \tag{4.16}$$

For the special case that $f_X(x;\theta)$ can be arranged such that $\mathsf{E}[t(X)] = \theta$, the unbiased estimator

$$T : x \mapsto t(x). \tag{4.17}$$

can directly be obtained. Otherwise, $t(x)$ at least provides a SUFFICIENT STATISTIC for estimating θ. Given an EXPONENTIAL MODEL it can be shown that the

- UNBIASED ESTIMATOR $T : x \mapsto t(x)$
- is a UMVU ESTIMATOR,
- and achieves the CRAMÉR-RAO LOWER BOUND.

4.4 Mean Estimation Example

We consider the estimation of the unknown mean value $\theta = \mu$ of a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ based on a single observation.

The respective PDF can be arranged as follows:

$$\begin{aligned} f_X(x;\theta) = \mathsf{L}(x;\theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right) \\ &= \exp\Bigg[\underbrace{\frac{\theta x}{\sigma^2}}_{a(\theta)t(x)} - \underbrace{\left(\frac{\theta^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}_{b(\theta)}\Bigg] \underbrace{\exp\left[-\frac{1}{2\sigma^2}x^2\right]}_{h(x)}, \end{aligned} \tag{4.18}$$

with $t(x) = x$ and $a(\theta) = \theta/\sigma^2$.

Thus the single observation UMVU Estimator $T^{(1)}$ of $\theta = \mu$ can be obtained as

$$T^{(1)} : \hat{\theta} = x, \qquad \text{with} \qquad \mathsf{Bias}[T^{(1)}(X)] = 0, \tag{4.19}$$

which minimizes both the variance and the MSE criterion to

$$\mathsf{Var}\left[T^{(1)}\right] = \frac{1}{I_{\mathsf{F}}} = \left(\frac{\partial a(\theta)}{\partial \theta}\right)^{-1} = \sigma^2. \tag{4.20}$$

4.5 Asymptotically Efficient Estimators

Definition. An estimator $T(X_1, \ldots, X_N)$ is Asymptotically Efficient if[1]

$$\sqrt{N}\left(T(X_1, \ldots, X_N) - \theta\right) \overset{N\to\infty}{\sim} \mathcal{N}(0, I_{\mathsf{F}}^{(1)}(\theta)^{-1}). \tag{4.21}$$

Given the following requirements are fulfilled,

- Θ is an open set,
- $f_X(x;\theta)$ and $\int_{\mathbb{X}} f_X(x;\theta)\mathrm{d}x$ is twice differentiable with respect to θ,
- $I_{\mathsf{F}}(\theta) < \infty$,
- ...

then a Maximum-Likelihood Estimator is Asymptotically Efficient.

[1] Convergence in distribution.

Part II

Examples

5. ML Principle for Direction of Arrival Estimation

We consider the estimation of the DIRECTION OF ARRIVAL (DoA) θ of an impinging planar wavefront by means of an antenna array with M antennas.

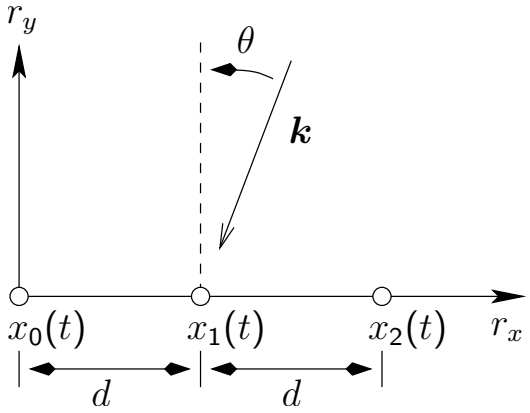


Fig. 5.1: Spatial sampling of a planar wavefront.

The received signal vector at the antenna array at time instant $t \in \mathbb{R}$ is equal to

$$\boldsymbol{x}(t) = \xi \boldsymbol{a} s(t) + \boldsymbol{\eta}(t) \quad \in \mathbb{C}^M, \tag{5.1}$$

with the signal at the mth antenna element as

$$x_m(t) = \xi \, \mathrm{e}^{\mathrm{j}\, 2\pi \sin(\theta) m d/\lambda} \, s(t) + \eta_m(t), \quad m = 0, \ldots, M-1. \tag{5.2}$$

The d and λ denote the distance between two adjacent antenna elements and the wavelength of the assumed NARROWBAND SIGNAL. The received signal is assumed to be corrupted by the Gaussian noise vector $\boldsymbol{\eta}(t) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_\eta)$.

The $\xi \in \mathbb{R}$ represents the attenuation which the transmitted signal $s(t) \in \mathbb{R}$ experiences over the transmission path.

Without loss of generality we thus assume a UNIFORM LINEAR ARRAY (ULA) with M antenna elements, i.e.,

$$\boldsymbol{a} = \begin{bmatrix} \alpha^0 \\ \alpha^1 \\ \vdots \\ \alpha^{M-1} \end{bmatrix}, \quad \text{with} \quad \alpha = \mathrm{e}^{\mathrm{j}\, 2\pi \sin(\theta) d/\lambda}. \tag{5.3}$$

In order to find the ML estimate of the DoA parameter θ, we consider the respective LIKELIHOOD FUNCTION for observations of the received vector $\boldsymbol{x}(t)$ at $t_1, \ldots, t_N$,[1] implicitly assuming a stationary scenario over the respective time intervall,

$$L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \theta) = \frac{1}{\pi^{MN} \det \boldsymbol{C}_\eta} \exp\left(-\sum_{i=1}^{N} (\boldsymbol{x}_i - \xi \boldsymbol{a} s_i)^{\mathsf{H}} \boldsymbol{C}_\eta^{-1} (\boldsymbol{x}_i - \xi \boldsymbol{a} s_i) \right). \tag{5.4}$$

For simplicity reasons we restrict the further analysis to the case of a single observation $N=1$ and ADDITIVE WHITE GAUSSIAN NOISE $\boldsymbol{\eta}(t) \sim \mathcal{N}(\boldsymbol{0}, \sigma_\eta^2 \boldsymbol{I})$, i.e.,

$$L(\boldsymbol{x}; \theta) = \frac{1}{(\pi \sigma_\eta^2)^M} \exp\left(-\frac{(\boldsymbol{x} - \xi \boldsymbol{a} s)^{\mathsf{H}} (\boldsymbol{x} - \xi \boldsymbol{a} s)}{\sigma_\eta^2} \right). \tag{5.5}$$

[1] For simplicity of notation we use $\boldsymbol{x}_i$ instead of $\boldsymbol{x}(t_i)$.

After some simple reformulation steps the ML optimization problem is equal to

$$\min_{\theta} \; (\boldsymbol{x} - \xi \boldsymbol{a} s)^{\mathsf{H}} (\boldsymbol{x} - \xi \boldsymbol{a} s) \tag{5.6}$$

Since the cost function can be expanded in

$$(\boldsymbol{x} - \xi \boldsymbol{a} s)^{\mathsf{H}} (\boldsymbol{x} - \xi \boldsymbol{a} s) = \boldsymbol{x}^{\mathsf{H}} \boldsymbol{x} - 2 \, \mathrm{Re} \left\{ \xi \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a} s \right\} + \xi^2 \boldsymbol{a}^{\mathsf{H}} \boldsymbol{a} s^2, \tag{5.7}$$

and $\boldsymbol{a}^{\mathsf{H}} \boldsymbol{a} = M$, it can be further reduced to

$$\max_{\theta} \; \mathrm{Re} \left\{ \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}(\theta) \right\}. \tag{5.8}$$

The ML Estimator $\hat{\theta} = \underset{\theta}{\mathrm{argmax}} \left\{ \mathrm{Re} \left\{ \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}(\theta) \right\} \right\}$ is obviously a NONLINEAR ESTIMATOR.

Note. In order to estimate the attenuation ξ, we need further information about the TRAINING SIGNAL (PILOT SIGNAL) $s(t)$.

The ML estimator for scenarios with a single impinging wavefront is equal to the so-called CONVENTIONAL BEAMFORMER.

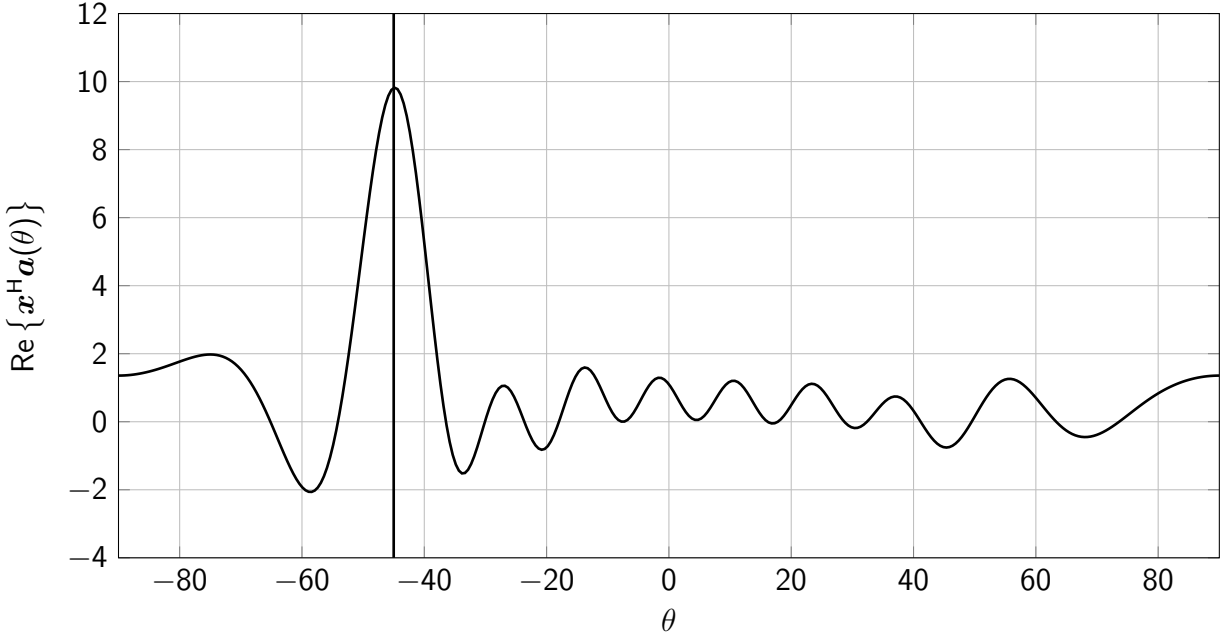


Fig. 5.2: Re $\left\{\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}(\theta)\right\}$ for $M = 10$, $\xi = s = 1$, and $\sigma_\eta^2 = 0.1$ with DoA $\theta_1 = -45$ deg.

In scenarios with more than one impinging wavefront, the beamformer, i.e., the ML estimator for the case of a single wavefront, is only a heuristic method for DOA estimation.

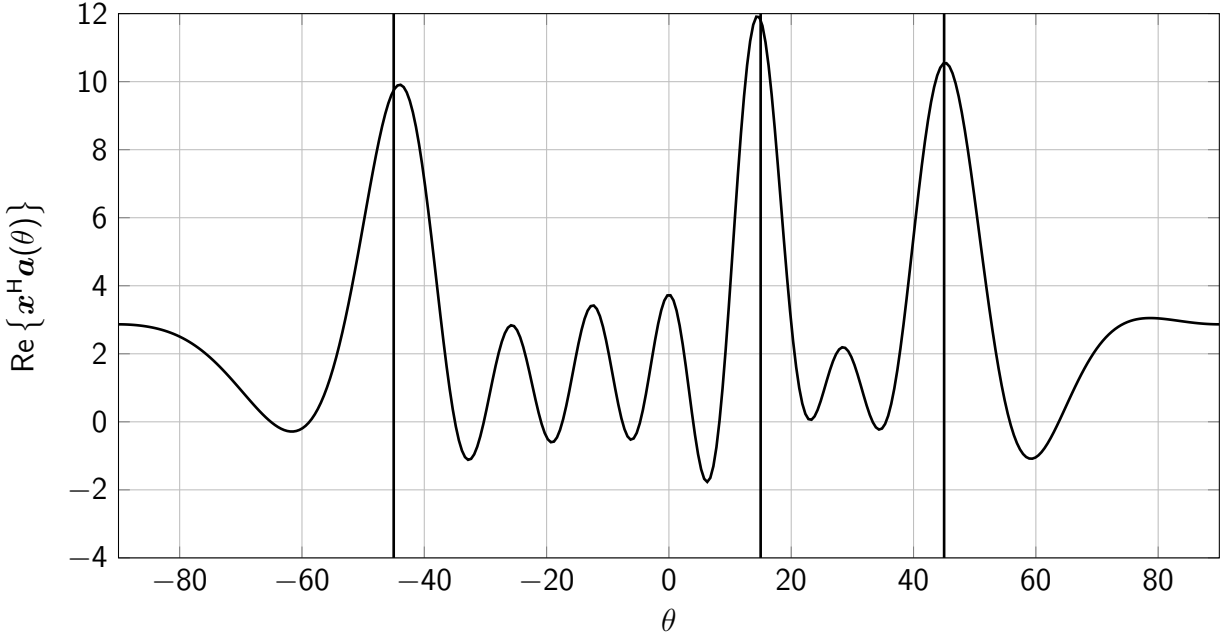


Fig. 5.3: $\mathrm{Re}\left\{\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}(\theta)\right\}$ for $M = 10$, $\xi = s = 1$, and $\sigma_\eta^2 = 0.1$ with DoAs $\theta_1 = -45\,\mathrm{deg}$, $\theta_2 = 15\,\mathrm{deg}$, and $\theta_3 = 45\,\mathrm{deg}$.

5.1 Cramér-Rao Bound for DOA Estimation

The Likelihood function of the given estimation problem obviously belongs to the family of EXPONENTIAL DISTRIBUTIONS. We consider θ as parameter to be estimated. After some reformulation steps $L(\boldsymbol{x};\theta)$ can be expressed as

$$L(\boldsymbol{x};\theta) = \frac{1}{(\pi\sigma_\eta^2)^M}\exp\left(-\frac{(\boldsymbol{x}-\xi\boldsymbol{a}s)^{\mathsf{H}}(\boldsymbol{x}-\xi\boldsymbol{a}s)}{\sigma_\eta^2}\right) = \frac{h(\boldsymbol{x})\exp\left(\boldsymbol{c}^{\mathsf{H}}(\theta)\boldsymbol{t}(\boldsymbol{x})\right)}{\exp(b(\theta))}, \tag{5.9}$$

with

$$h(\boldsymbol{x}) = \frac{1}{(\pi\sigma_\eta^2)^M}\exp\left(-\frac{\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}}{\sigma_\eta^2}\right) \tag{5.10}$$

$$\boldsymbol{c}(\theta) = \frac{1}{\sigma_\eta^2}\begin{bmatrix} \xi\boldsymbol{a}s \\ (\xi\boldsymbol{a}s)^* \end{bmatrix} \tag{5.11}$$

$$\boldsymbol{t}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}^* \end{bmatrix} \tag{5.12}$$

$$b(\theta) = \frac{\xi^2 s^2 \boldsymbol{a}^{\mathsf{H}}\boldsymbol{a}}{\sigma_\eta^2} = \frac{M\xi^2 s^2}{\sigma_\eta^2}. \tag{5.13}$$

Note, that $\mathsf{E}[\boldsymbol{t}(\boldsymbol{x})] = \sigma_\eta^2\boldsymbol{c}(\theta)$.

Consequently, the FISHER INFORMATION of the ML estimation problem can be obtained as[2]

$$I_{\mathsf{F}}(\theta) = \left(\frac{\partial \boldsymbol{c}(\boldsymbol{x})}{\partial \theta}\right)^{\mathsf{H}} \frac{\partial\, \mathsf{E}\left[\boldsymbol{t}(\boldsymbol{x})\right]}{\partial \theta} \tag{5.14}$$

$$= \frac{\xi^2 s^2}{\sigma_\eta^2} \frac{\partial}{\partial \theta} \left(\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{a}^* \end{bmatrix} \right)^{\mathsf{H}} \frac{\partial}{\partial \theta} \left(\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{a}^* \end{bmatrix} \right). \tag{5.15}$$

Since θ is real,

$$I_{\mathsf{F}}(\theta) = \frac{\xi^2 s^2}{\sigma_\eta^2} \left(\frac{\partial}{\partial \theta} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{a}^* \end{bmatrix} \right)^{\mathsf{H}} \left(\frac{\partial}{\partial \theta} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{a}^* \end{bmatrix} \right) \tag{5.16}$$

$$= 2\frac{\xi^2 s^2}{\sigma_\eta^2} \left\| \frac{\partial \boldsymbol{a}}{\partial \theta} \right\|^2. \tag{5.17}$$

Finally, with $\partial \alpha^m / \partial \theta = \mathsf{j}\, m (2\pi d/\lambda) \cos\theta \alpha^m$ and $|\alpha^m| = 1$,

$$I_{\mathsf{F}}(\theta) = 2\frac{\xi^2 s^2}{\sigma_\eta^2} \left(\frac{2\pi d}{\lambda} \cos\theta \right)^2 \sum_{m=1}^{M-1} m^2$$

$$= 2\frac{\xi^2 s^2}{\sigma_\eta^2} \left(\frac{2\pi d}{\lambda} \cos\theta \right)^2 \frac{M(M-1)(2M-1)}{6} \;\propto\; \left(\frac{d}{\lambda}\right)^2 \frac{\xi^2 s^2}{\sigma_\eta^2} (\cos\theta)^2 M^3. \tag{5.18}$$

[2] $\mathsf{E}\left[\boldsymbol{x}\right] = \xi \boldsymbol{a} s$.

References

- S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume I, Prentice Hall Signal Processing Series, 1993.

Part III

Estimation of Random Variables

6. Bayes Estimation

In contrast to the Maximum-Likelihood principle, an A PRIORI information about the unknown parameter $\theta \in \Theta$ is now taken into account. We further assume that the a priori information of the unknown parameter is given by means of the

$$\text{PDF} : f_{\Theta}(\theta; \sigma) \tag{6.1}$$

$$\text{CONDITIONAL PDF} : f_{X|\Theta}(x|\theta), \tag{6.2}$$

where σ parameterizes the statistical model of the (now!) random variable Θ. We now consider the MEAN MSE with respect to the mean of the parameter Θ, i.e.,

$$\mathsf{E}\left[\mathsf{E}\left[(T(X)-\Theta)^2 \mid \Theta\right]\right] = \int_{\Theta} \underbrace{\int_{\mathbb{X}} (T(x)-\theta)^2 f_{X|\Theta}(x|\theta)\mathrm{d}x}_{\mathsf{E}\left[(T(X)-\Theta)^2|\Theta=\theta\right]} f_{\Theta}(\theta;\sigma)\mathrm{d}\theta = \mathsf{E}\left[(T(X)-\Theta)^2\right]. \tag{6.3}$$

6.1 Conditional Mean Estimator – Minimizing the MSE

Theorem. The CONDITIONAL MEAN ESTIMATOR (BAYES ESTIMATOR) T_{CM} is MSE optimal,

$$T_{\mathsf{CM}}: \quad x \mapsto \mathsf{E}[\Theta|X=x] \quad = \int_{\Theta} \theta f_{\Theta|X}(\theta|x)\mathrm{d}\theta, \tag{6.4}$$

i.e., T_{CM} minimizes the MEAN MSE cost criterion

$$\mathsf{E}\left[\mathsf{E}\left[(T(X)-\Theta)^2 \mid \Theta\right]\right]. \tag{6.5}$$

Note. We distinguish

- $\mathsf{E}[(T(X)-\Theta)^2|\Theta=\theta]$, which corresponds to the MSE subject to the condition of the (unknown) deterministic outcome θ of the random variable Θ,
- and $\mathsf{E}[(T(X)-\Theta)^2] = \mathsf{E}[\mathsf{E}[(T(X)-\Theta)^2 \mid \Theta]]$, which assumes Θ to be a random variable, such that the expectation has to be taken over both random variables.

Proof.

$$\mathsf{E}\left[\varepsilon(T(X))\right] = \int_{\Theta}\int_{\mathbb{X}} (T(x)-\theta)^2 \underbrace{f_{X|\Theta}(x|\theta) f_{\Theta}(\theta;\sigma)}_{=f_{X,\Theta}(x,\theta)\quad \text{(Bayes)}} \mathrm{d}x\mathrm{d}\theta$$

$$= \int_{\mathbb{X}}\int_{\Theta} (T(x)-\theta)^2 f_{X,\Theta}(x,\theta)\mathrm{d}\theta\mathrm{d}x = \int_{\mathbb{X}} \underbrace{\int_{\Theta} (T(x)-\theta)^2 f_{\Theta|X}(\theta|x)\mathrm{d}\theta}_{\text{min!}} f_X(x)\mathrm{d}x$$

$$\Rightarrow \quad \boxed{\min_{T(x)} \int_{\Theta} (T(x)-\theta)^2 f_{\Theta|X}(\theta|x)\mathrm{d}\theta,} \quad \text{i.e.,}$$

$$\Rightarrow \quad \boxed{\forall x:\ T_x = T(x)}$$

$$\frac{\partial}{\partial T_x}\int_{\Theta} (T_x-\theta)^2 f_{\Theta|X}(\theta|x)\mathrm{d}\theta \stackrel{!}{=} 0$$

$$2\int_{\Theta} (T_x-\theta) f_{\Theta|X}(\theta|x)\mathrm{d}\theta = 0$$

$$T_x \int_{\Theta} f_{\Theta|X}(\theta|x)\mathrm{d}\theta = \int_{\Theta} \theta f_{\Theta|X}(\theta|x)\mathrm{d}\theta, \text{i.e.,}$$

$$\Rightarrow \quad \boxed{T(x) = \mathsf{E}\left[\Theta|X=x\right].}$$

6.2 Bernoulli Experiment (Cont'd)

In Section 3.4, we found the Maximum Likelihood Estimate of the success probability of a Bernoulli Experiment as

$$\hat{\theta}_{\mathsf{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \log \mathsf{L}(x; \theta) \right\} = \frac{x}{N}, \tag{6.6}$$

given the statistical model by means of the Probability Mass Function

$$\mathsf{B}_{N,\theta}(x) = \binom{N}{x} \theta^x (1-\theta)^{N-x}, \quad x \in \{0, 1, \ldots, N\}.$$

We can expect a better estimator, if we design the estimator with respect to the Mean Squared Error by exploiting the trivial side information

$$f_{\Theta}(\theta) = \begin{cases} 1 & ; \quad \theta \in [0,1] \\ 0 & ; \quad \text{otherwise} \end{cases}.$$

The optimal estimator in terms of the Mean Squared Error can be obtained by means of the Conditional Mean Estimator

$$\hat{\theta}_{\mathsf{CM}} = \mathsf{E}\left[\Theta \mid X = x\right] = \int_0^1 \theta f_{\Theta|X}(\theta \mid x) \, \mathrm{d}\theta. \tag{6.7}$$

The required Conditional PDF is

$$f_{\Theta|X}(\theta) = \frac{p_{X|\Theta}(x \mid \theta)}{p_X(x)} f_{\Theta}(\theta), \quad \theta \in [0,1] \tag{6.8}$$

with[1]

$$p_X(x) = \int_\Theta p_{X|\Theta}(x \mid \theta) f_\Theta(\theta)\,\mathrm{d}\theta, \quad x \in \{0, 1, \ldots, N\} \tag{6.9}$$

$$= \int_0^1 \mathsf{B}_{N,\theta}(x) \cdot 1\,\mathrm{d}\theta = \int_0^1 \binom{N}{x} \theta^x (1-\theta)^{N-x}\,\mathrm{d}\theta \tag{6.10}$$

$$= \binom{N}{x} \frac{x!}{(N-x+1)(N-x+2)\cdots(N+1)} = \frac{1}{N+1}. \tag{6.11}$$

Then the CONDITIONAL MEAN ESTIMATOR equals (cf. alternative solution in Section 3.4)

$$\hat{\theta}_{\mathsf{CM}} = p_X(x)^{-1} \int_0^1 \theta p_{X|\Theta}(x \mid \theta)\,\mathrm{d}\theta \tag{6.12}$$

$$= p_X(x)^{-1} \int_0^1 \binom{N}{x} \theta^{x+1} (1-\theta)^{N-x}\,\mathrm{d}\theta \tag{6.13}$$

$$= p_X(x)^{-1} \binom{N}{x} \frac{(x+1)!}{(N-x+1)(N-x+2)\cdots(N+2)} = \frac{p_X(x)^{-1}}{N+1} \frac{x+1}{N+2} \tag{6.14}$$

$$= \frac{x+1}{N+2} = \frac{N}{N+2}\hat{\theta}_{\mathsf{ML}} + \frac{1}{N+2}. \tag{6.15}$$

[1] m-times applying the partial integration technique ($m, n \in \mathbb{N}$) we obtain

$$\int_0^1 x^m (1-x)^n\,\mathrm{d}x = \frac{m}{n+1} \int_0^1 x^{m-1} (1-x)^{n+1}\,\mathrm{d}x = \cdots = \frac{m!}{(n+1)\cdots(n+m)} \int_0^1 (1-x)^{n+m}\,\mathrm{d}x = \frac{m!}{(n+1)\cdots(n+m+1)}.$$

6.3 Mean Estimation Example

We consider the estimation of the unknown mean value θ of a random variable

$$X \sim \mathcal{N}(\theta, \sigma^2_{X|\Theta=\theta}), \tag{6.16}$$

now based on N i.i.d. observations

$$X_1, \ldots, X_N, \tag{6.17}$$

drawn with respect to the CONDITIONAL PDF[2]

$$f_{X_1,\ldots,X_N|\Theta}(x_1, \ldots, x_N|\theta) = \frac{1}{(\sqrt{2\pi}\sigma_{X|\Theta=\theta})^N} \exp\left(-\frac{1}{2\sigma^2_{X|\Theta=\theta}} \sum_{i=1}^{N} (x_i - \theta)^2\right). \tag{6.18}$$

A priori knowledge about the unknown parameter θ is given by the PDF $\Theta \sim \mathcal{N}(m, \sigma^2_\Theta)$, i.e.,

$$\Theta \sim \frac{1}{\sqrt{2\pi}\sigma_\Theta} \exp\left(-\frac{1}{2\sigma^2_\theta}(\theta - m)^2\right) = f_\Theta(\theta). \tag{6.19}$$

[2]The difference between statistics drawn from a CONDITIONAL DISTRIBUTION and statistics drawn from a UNCONDITIONAL DISTRIBUTION is essential for the understanding of the following results.

Note. Consider the difference between

$$\text{Variance Of } \Theta: \quad \sigma^2_{\Theta}, \tag{6.20}$$

$$\text{Conditional Variance Of } X \text{ Given } \theta: \quad \sigma^2_{X|\Theta=\theta}, \tag{6.21}$$

$$\text{Variance Of } X: \quad \sigma^2_{X}, \tag{6.22}$$

with

$$\sigma^2_X = \sigma^2_{X|\Theta=\theta} + \sigma^2_{\Theta}. \tag{6.23}$$

Interpretation. Since X is the sum of independent random variables

$$X = \Theta + \eta, \tag{6.24}$$

with $\Theta \sim \mathcal{N}(m, \sigma^2_{\Theta})$ and $\eta \sim \mathcal{N}(0, \sigma^2_{X|\Theta=\theta})$, the variance of the sum is equal to the sum of variances and the expectation of the sum is equal to sum of expectations, i.e., $\mu_X = \mathsf{E}[X] = \mathsf{E}[\Theta + \eta] = \mathsf{E}[\Theta] = m$. The covariance between Θ and X equals to the variance of Θ, i.e., $c_{\Theta,X} = \mathsf{Cov}[\Theta, X] = \mathsf{Cov}[\Theta, \Theta + \eta] = \mathsf{Cov}[\Theta, \Theta] + \mathsf{Cov}[\Theta, \eta] = \mathsf{Cov}[\Theta, \Theta] = \mathsf{Var}[\Theta]$. In essence,

$$\mu_X = \mu_{\Theta} \tag{6.25}$$

$$c_{\Theta,X} = \sigma^2_{\Theta}. \tag{6.26}$$

Note. Make sure to identify the difference between the random variable X given $\Theta = \theta$ which is drawn from the PDF $f_{X|\Theta}(x|\theta)$ and the random variable X which is drawn from the PDF $f_X(x)$.

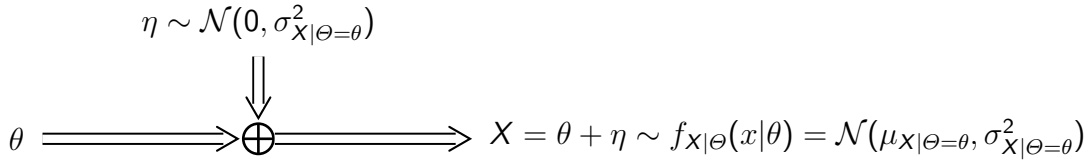


Fig. 6.1: Illustration of the random variable X given θ which is the sum of a specific **realization** θ and **random** additive noise η. Therefore, we have $\mu_{X|\Theta=\theta} = \theta$ and the **conditional variance** $\sigma^2_{X|\Theta=\theta}$.

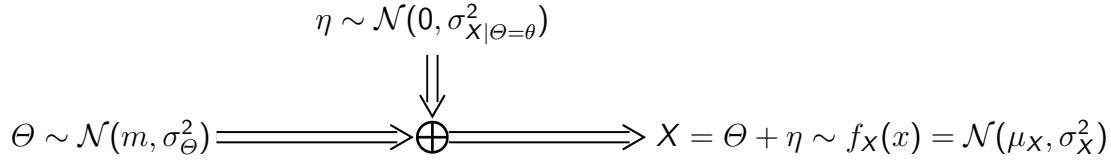


Fig. 6.2: Illustration of the random variable X which is the sum of the **random variable** Θ and **random** additive noise η. Therefore, we have $\mu_X = m$ and the **sum of variances** $\sigma^2_X = \sigma^2_{X|\Theta=\theta} + \sigma^2_\Theta$.

For the computation of the CONDITIONAL MEAN ESTIMATOR we need two steps:

a) Conditional PDF $f_{\Theta|X_1,\ldots,X_N}(\theta|x_1,\ldots,x_N)$:

$$f_{\Theta|X_1,\ldots,X_N}(\theta|x_1,\ldots,x_N) = \frac{f_{X_1,\ldots,X_N|\Theta}(x_1,\ldots,x_N|\theta)f_\Theta(\theta)}{f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)} \tag{6.27}$$

$$= \frac{f_{X_1,\ldots,X_N|\Theta}(x_1,\ldots,x_N|\theta)f_\Theta(\theta)}{\underbrace{\int_\Theta f_{X_1,\ldots,X_N|\Theta}(x_1,\ldots,x_N|\theta)f_\Theta(\theta)\mathrm{d}\theta}_{=f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)\ \text{(Marginalization)}}} \tag{6.28}$$

$$= \gamma \exp\left(-\frac{1}{2\sigma^2_{X|\Theta=\theta}}\sum_{i=1}^{N}(x_i-\theta)^2\right)\exp\left(-\frac{1}{2\sigma^2_\Theta}(\theta-m)^2\right), \tag{6.29}$$

with γ such that $\int_\Theta f_{\Theta|X_1,\ldots,X_N}(\theta|x_1,\ldots,x_N)\mathrm{d}\theta = 1$.

b) Conditional Mean $\mathsf{E}[\Theta|x_1,\ldots,x_N]$:[3]

$$\hat{\theta}_{\mathsf{CM}} = \mathsf{E}[\Theta|x_1,\ldots,x_N] \tag{6.30}$$

$$= \int_{\Theta} \theta f_{\Theta|X_1,\ldots,X_N}(\theta|x_1,\ldots,x_N)\mathrm{d}\theta \quad \text{with CONDITIONAL PDF} \tag{6.31}$$

$$= \frac{\sum_{i=1}^{N} x_i\sigma_{\Theta}^2 + m\sigma_{X|\Theta=\theta}^2}{N\sigma_{\Theta}^2 + \sigma_{X|\Theta=\theta}^2},$$

which is LINEAR IN OBSERVATIONS $x_1,\ldots,x_N$ (!) and after some reformulation steps:

$$\hat{\theta}_{\mathsf{CM}} = \frac{\sigma_{\Theta}^2}{\sigma_{\Theta}^2 + \frac{\sigma_{X|\Theta=\theta}^2}{N}}\hat{\theta}_{\mathsf{ML}} + \frac{\frac{\sigma_{X|\Theta=\theta}^2}{N}}{\sigma_{\Theta}^2 + \frac{\sigma_{X|\Theta=\theta}^2}{N}}m, \quad \text{with} \quad \hat{\theta}_{\mathsf{ML}} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{6.32}$$

Note. For $N = 1$, we obtain $\hat{\theta}_{\mathsf{CM}} = m + \frac{\sigma_{\Theta}^2}{\sigma_{\Theta}^2 + \sigma_{X|\Theta=\theta}^2}(x - m) = \mu_{\Theta} + \frac{c_{\Theta,X}}{\sigma_X^2}(x - \mu_X)$.

[3] Eq. (6.32) can be read out from Eq. (6.29) after some reformulation steps.

Derivation.

$$
\begin{aligned}
f_{\Theta|X_1,\dots,X_N}(\theta|x_1,\dots,x_N) &\propto \exp\left(-\frac{1}{2\sigma^2_{X|\Theta=\theta}}\sum_{i=1}^{N}(x_i-\theta)^2\right)\exp\left(-\frac{1}{2\sigma^2_{\Theta}}(\theta-m)^2\right)\\
&\propto \exp\left(-\frac{1}{2\sigma^2_{X|\Theta=\theta}\sigma^2_{\Theta}}\left(\sum_{i=1}^{N}\left(x_i^2-2x_i\theta+\theta^2\right)\sigma^2_{\Theta}+\left(\theta^2-2m\theta+m^2\right)\sigma^2_{X|\Theta=\theta}\right)\right)\\
&\propto \exp\left(-\frac{1}{2\sigma^2_{X|\Theta=\theta}\sigma^2_{\Theta}}\left(-2\theta\sigma^2_{\Theta}\sum_{i=1}^{N}x_i+N\theta^2\sigma^2_{\Theta}+\theta^2\sigma^2_{X|\Theta=\theta}-2m\theta\sigma^2_{X|\Theta=\theta}+\dots\right)\right)\\
&\propto \exp\left(-\frac{1}{2\sigma^2_{X|\Theta=\theta}\sigma^2_{\Theta}}\left(\theta^2\left(N\sigma^2_{\Theta}+\sigma^2_{X|\Theta=\theta}\right)-2\theta\left(\sigma^2_{\Theta}\sum_{i=1}^{N}x_i+m\sigma^2_{X|\Theta=\theta}\right)+\dots\right)\right)\\
&\propto \exp\left(-\frac{\left(N\sigma^2_{\Theta}+\sigma^2_{X|\Theta=\theta}\right)}{2\sigma^2_{X|\Theta=\theta}\sigma^2_{\Theta}}\left(\theta-\frac{\sigma^2_{\Theta}\sum_{i=1}^{N}x_i+m\sigma^2_{X|\Theta=\theta}}{\left(N\sigma^2_{\Theta}+\sigma^2_{X|\Theta=\theta}\right)}\right)^2\right).
\end{aligned}
$$

A comparison with the Gaussian standard model results into

$$
\mu_{\Theta|X_1,\dots,X_N}=\frac{\sigma^2_{\Theta}\sum_{i=1}^{N}x_i+m\sigma^2_{X|\Theta=\theta}}{N\sigma^2_{\Theta}+\sigma^2_{X|\Theta=\theta}}\quad\text{and}\quad\sigma^2_{\Theta|X_1,\dots,X_N}=\frac{\sigma^2_{X|\Theta=\theta}\sigma^2_{\Theta}}{N\sigma^2_{\Theta}+\sigma^2_{X|\Theta=\theta}}. \tag{6.33}
$$

Discussion.

- Given (a) a large number N of observations or (b) a small conditional variance $\sigma^2_{X|\Theta=\theta}$ of observations or (c) a large variance σ^2_Θ of the a priori distribution of the unknown parameter θ, the derived solution obviously recommends to rely on the MAXIMUM-LIKELIHOOD ESTIMATOR, since

$$\lim_{N\to\infty} \hat{\theta}_{\mathsf{CM}} = \lim_{\sigma^2_{X|\Theta=\theta}\to 0} \hat{\theta}_{\mathsf{CM}} = \lim_{\sigma^2_\Theta\to\infty} \hat{\theta}_{\mathsf{CM}} = \hat{\theta}_{\mathsf{ML}}, \tag{6.34}$$

  with

$$\hat{\theta}_{\mathsf{ML}} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{6.35}$$

- However, given (a) a small variance σ^2_Θ of the a priori distribution of the unknown parameter θ or (b) a large conditional variance $\sigma^2_{X|\Theta=\theta}$ of observations, the derived solution suggests to ignore the given observations and instead to rely solely on the MEAN VALUE of Θ, since

$$\lim_{\sigma^2_\Theta\to 0} \hat{\theta}_{\mathsf{CM}} = \lim_{\sigma^2_{X|\Theta=\theta}\to\infty} \hat{\theta}_{\mathsf{CM}} = m. \tag{6.36}$$

- The respective MINIMUM MEAN SQUARE ERROR (MMSE) is given by[4]

$$\begin{aligned}
\mathsf{E}\left[\mathsf{E}\left[(T_{\mathsf{CM}}-\Theta)^2 \mid \Theta\right]\right] &= \mathsf{E}\left[\mathsf{E}\left[(\mathsf{E}\left[\Theta|X\right]-\Theta)^2|X\right]\right] && (6.37)\\
&= \mathsf{E}\left[\mathsf{E}\left[\Theta|X\right]^2 - 2\,\mathsf{E}\left[\Theta|X\right]\mathsf{E}\left[\Theta|X\right] + \mathsf{E}\left[\Theta^2|X\right]\right] && (6.38)\\
&= \mathsf{E}\left[\mathsf{E}\left[\Theta^2|X\right] - \mathsf{E}\left[\Theta|X\right]^2\right] && (6.39)\\
&= \mathsf{E}\left[\mathsf{Var}\left[\Theta|X\right]\right] && (6.40)
\end{aligned}$$

$$= \underbrace{\frac{\sigma^2_{X|\Theta=\theta}\sigma^2_\Theta}{N\sigma^2_\Theta + \sigma^2_{X|\Theta=\theta}}}_{\text{refer to (6.33)}} = \sigma^2_\Theta - \frac{(\sigma^2_\Theta)^2}{\sigma^2_\Theta + \dfrac{\sigma^2_{X|\Theta=\theta}}{N}}. \qquad (6.41)$$

- Any requisites of the CM Estimator can be found from the mean vector and covariance matrix of the joint distribution of $(X,\Theta)^{\mathsf{T}}$, i.e.,

$$\mathsf{E}\left[\begin{bmatrix} X \\ \Theta \end{bmatrix}\right] = \begin{bmatrix} \mu_X \\ \mu_\Theta \end{bmatrix} = \begin{bmatrix} m \\ m \end{bmatrix}, \qquad (6.42)$$

$$\mathsf{Var}\left[\begin{bmatrix} X \\ \Theta \end{bmatrix}\right] = \begin{bmatrix} c_{X,X} & c_{X,\Theta} \\ c_{\Theta,X} & c_{\Theta,\Theta} \end{bmatrix} = \begin{bmatrix} \sigma^2_\Theta + \sigma^2_{X|\Theta=\theta} & \sigma^2_\Theta \\ \sigma^2_\Theta & \sigma^2_\Theta \end{bmatrix}. \qquad (6.43)$$

[4] Eq. (6.41) can be read out from Eq. (6.29) after some reformulation steps.

6.4 Jointly Gaussian Random Variables

Given two random vectors $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu_X}, \boldsymbol{C_X})$, $\boldsymbol{\Theta} \sim \mathcal{N}(\boldsymbol{\mu_\Theta}, \boldsymbol{C_\Theta})$, and the covariance matrix $\boldsymbol{C_Z}$ from the joint distribution of $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{\Theta})^\mathsf{T} \sim \mathcal{N}(\boldsymbol{\mu_Z}, \boldsymbol{C_Z})$, i.e.,

$$\boldsymbol{C_Z} = \begin{bmatrix} \boldsymbol{C_X} & \boldsymbol{C_{X,\Theta}} \\ \boldsymbol{C_{\Theta,X}} & \boldsymbol{C_\Theta} \end{bmatrix}, \tag{6.44}$$

the Multivariate Conditional Mean Estimator is obtained as

$$\boldsymbol{T}_{\mathsf{CM}}: \quad \boldsymbol{x} \mapsto \mathsf{E}\left[\boldsymbol{\Theta} | \boldsymbol{X} = \boldsymbol{x}\right] = \boldsymbol{\mu_\Theta} + \boldsymbol{C_{\Theta,X}} \boldsymbol{C_X^{-1}} (\boldsymbol{x} - \boldsymbol{\mu_X}). \tag{6.45}$$

The respective MMSE is equal to the Trace of the conditional covariance matrix $\boldsymbol{C_{\Theta|X}}$, this is

$$\mathsf{E}\left[\| \boldsymbol{T}_{\mathsf{CM}} - \boldsymbol{\Theta} \|_2^2\right] = \mathsf{tr}\left[\boldsymbol{C_{\Theta|X=x}}\right] = \mathsf{tr}\left[\boldsymbol{C_\Theta} - \boldsymbol{C_{\Theta,X}} \boldsymbol{C_X^{-1}} \boldsymbol{C_{X,\Theta}}\right]. \tag{6.46}$$

Note. Given Jointly Gaussian Random Variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, the Conditional Mean Estimator $\mathsf{E}\left[\boldsymbol{Y}|\boldsymbol{X}\right]$ is a Linear Function in $\boldsymbol{X}$. This does not hold for arbitrarily jointly distributed random variables!

6.5 Mean Estimation Example (Cont'd)

From a multivariate r.v. perspective, we obtain

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} \sim \mathcal{N}(\mu_{\boldsymbol{X}}, \boldsymbol{C}_{\boldsymbol{X}}) \quad \text{and} \quad \Theta \sim \mathcal{N}(\mu_\Theta, \sigma_\Theta^2), \tag{6.47}$$

and thus

$$\mu_{\boldsymbol{X}} = m\mathbf{1}, \quad \mu_\Theta = m, \quad \boldsymbol{C}_{\Theta,\boldsymbol{X}} = \sigma_\Theta^2 \mathbf{1}^\mathsf{T}, \quad \boldsymbol{C}_{\boldsymbol{X}} = \sigma_\Theta^2 \mathbf{1}\mathbf{1}^\mathsf{T} + \sigma_{X|\Theta=\theta}^2 \boldsymbol{I}, \tag{6.48}$$

where $\mathbf{1} = [1, \ldots, 1]^\mathsf{T}$ and $\boldsymbol{I}$ is the unity matrix. With a closed form solution for the inverse $\boldsymbol{C}_{\boldsymbol{X}}^{-1}$, we obtain the CM Estimator

$$T_{\mathsf{CM}} = m + \sigma_\Theta^2 \mathbf{1}^\mathsf{T} (\sigma_\Theta^2 \mathbf{1}\mathbf{1}^\mathsf{T} + \sigma_{X|\Theta=\theta}^2 \boldsymbol{I})^{-1} (\boldsymbol{x} - m\mathbf{1}) = \frac{\frac{\sum_{i=1}^N x_i}{\sigma_{X|\Theta=\theta}^2} + \frac{m}{\sigma_\Theta^2}}{\frac{N}{\sigma_{X|\Theta=\theta}^2} + \frac{1}{\sigma_\Theta^2}}. \tag{6.49}$$

Note. The computation of $\boldsymbol{C}_{\boldsymbol{X}}^{-1}$ can be obtained in closed form by applying the SHERMAN-MORRISON FORMULA, a special version of the MATRIX-INVERSION LEMMA,[5] i.e.,

$$\boldsymbol{C}_{\boldsymbol{X}}^{-1} = (\sigma_{\Theta}^2 \mathbf{1}\mathbf{1}^{\mathsf{T}} + \sigma_{X|\Theta=\theta}^2 \boldsymbol{I})^{-1} \tag{6.50}$$

$$= \frac{1}{\sigma_{X|\Theta=\theta}^2} \left[\boldsymbol{I} - \mathbf{1}\mathbf{1}^{\mathsf{T}} \frac{\frac{1}{\sigma_{X|\Theta=\theta}^2}}{\frac{1}{\sigma_{\Theta}^2} + \frac{N}{\sigma_{X|\Theta=\theta}^2}} \right], \tag{6.51}$$

where

$$\mathbf{1}\mathbf{1}^{\mathsf{T}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

[5] Given the compatibility of all matrices and the existence of $\boldsymbol{A}^{-1}$, the following equality holds: $(\boldsymbol{A} + \boldsymbol{b}\boldsymbol{c}^{\mathsf{T}})^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{c}^{\mathsf{T}}\boldsymbol{A}^{-1}}{(1 + \boldsymbol{c}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{b})}$.

Note. The random variables $X_1, \ldots, X_N$ given θ are drawn from the conditional joint PDF with respect to the **same realization** θ of Θ, i.e., we have that $X_1, \ldots, X_N$ are drawn from the conditional joint PDF $f_{X_1,\ldots,X_N|\Theta}(x_1, \ldots, x_N|\theta)$.

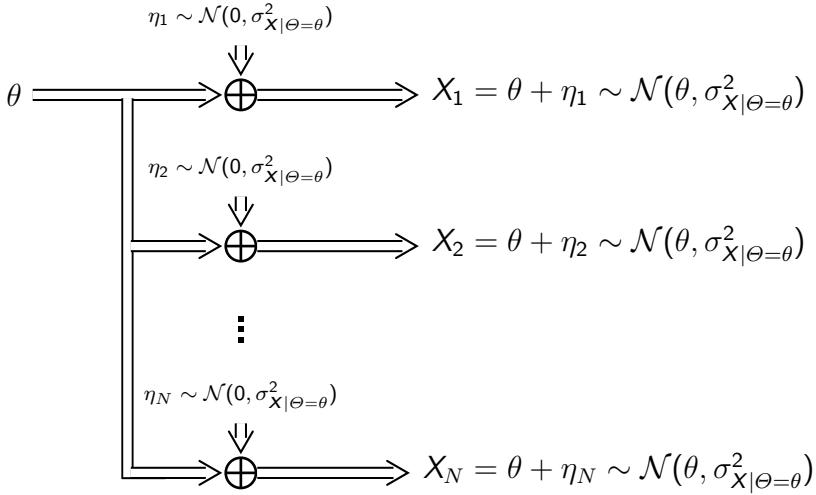


Fig. 6.3: Illustration of the random variables $X_1, \ldots, X_N$ given θ which each are the sum of the **same realization** θ and **independent random** additive noise η_i, i.e., with $\mathrm{E}[\eta_i \eta_j] = 0$ for all $i \neq j$.

Note. Consequently, the random variables $X_1, \ldots, X_N$ and the random variable Θ are drawn from the joint PDF $f_{X_1,\ldots,X_N,\Theta}(x_1, \ldots, x_N, \theta)$.

6.6 Orthogonality Principle

The ORTHOGONALITY PRINCIPLE is an inherent property of the CONDITIONAL MEAN ESTIMATOR. It describes the inherent STOCHASTIC ORTHOGONALITY between the CM estimation error and any observations statistics or functions thereof, i.e.,

$$\mathsf{E}\left[(T_{\mathsf{CM}}(X_1,\ldots,X_N)-\Theta)h(X_1,\ldots,X_N)\right]=0, \tag{6.52}$$

$$\Updownarrow$$

$$T_{\mathsf{CM}}(X_1,\ldots,X_N)-\Theta \quad \perp \quad h(X_1,\ldots,X_N) \tag{6.53}$$

where $h:\mathbb{R}^N\to\mathbb{R}, x_1,\ldots,x_N\mapsto h(x_1,\ldots,x_N)$.

Proof.

$$\begin{aligned}\mathsf{E}\left[(T_{\mathsf{CM}}(X)-\Theta)h(X)\right] &= \mathsf{E}\left[(T_{\mathsf{CM}}(X)h(X)\right]-\mathsf{E}\left[\Theta h(X)\right]=\mathsf{E}\left[\mathsf{E}\left[\Theta|X\right]h(X)\right]-\mathsf{E}\left[\Theta h(X)\right] &(6.54)\\ &= \mathsf{E}\left[\mathsf{E}\left[\Theta h(X)|X\right]\right]-\mathsf{E}\left[\Theta h(X)\right]=\mathsf{E}\left[\Theta h(X)\right]-\mathsf{E}\left[\Theta h(X)\right]=0. &(6.55)\end{aligned}$$

6.7 Mean Estimation Example (Cont'd)

The parameter θ to be estimated and the corresponding observation statistic X are JOINTLY GAUSSIAN DISTRIBUTED according to

$$\begin{bmatrix} X \\ \Theta \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_\Theta \end{bmatrix}, \begin{bmatrix} c_{X,X} & c_{X,\Theta} \\ c_{\Theta,X} & c_{\Theta,\Theta} \end{bmatrix} \right) = \mathcal{N}\left(\begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} \sigma_\Theta^2 + \sigma_{X|\Theta=\theta}^2 & \sigma_\Theta^2 \\ \sigma_\Theta^2 & \sigma_\Theta^2 \end{bmatrix} \right). \tag{6.56}$$

As an SUFFICIENT STATISTIC[6] for estimating the unknown realization of the random variable Θ we use[7]

$$S = \frac{1}{N} \sum_{i=1}^{N} X_i. \tag{6.57}$$

Applying the ORTHOGONALITY PRINCIPLE we obtain

$$\mathrm{E}\left[(T_{\mathrm{CM}} - \Theta)S\right] = \mathrm{E}\left[T_{\mathrm{CM}}S\right] - \mathrm{E}\left[\Theta S\right] = 0. \tag{6.58}$$

Knowing that the MSE optimal estimator is linear in S, we substitute T_{CM} by a linear model

$$T_{\mathrm{CM}}(X) = aS + b. \tag{6.59}$$

For $h(X_1, \ldots, X_N) = S$, the ORTHOGONALITY PRINCIPLE is now equal to

$$\mathrm{E}\left[(T_{\mathrm{CM}} - \Theta)S\right] = \mathrm{E}\left[(aS + b - \Theta)S\right] = 0. \tag{6.60}$$

[6] The concept of SUFFICIENT STATISTIC is later introduced in this course.

[7] Note, that the realizations of the statistics $X_1, \ldots, X_N$ are conditioned on the unknown realization of Θ.

The Orthogonality Principle leaves one Degree of Freedom which is statisfied by a second orthogonality propostion with $h(X_1, \ldots, X_N) = 1$,

$$\mathsf{E}\left[(aS + b - \Theta)1\right] = \mathsf{E}\left[aS + b\right] - \mathsf{E}\left[\Theta\right] = 0. \tag{6.61}$$

Then we obtain

$$a = \frac{c_{\Theta,S}}{\sigma_S^2} \quad \text{and} \quad b = \mu_\Theta - a\mu_S, \tag{6.62}$$

and the Conditional Mean Estimator is given by[8]

$$T_{\mathsf{CM}}: \quad s \mapsto \mu_\Theta + \frac{c_{\Theta,S}}{\sigma_S^2}(s - \mu_S) = m + \frac{\sigma_\Theta^2}{\frac{\sigma_{X|\Theta=\theta}^2}{N} + \sigma_\Theta^2}(s - m) = \frac{\frac{\sum_{i=1}^N x_i}{\sigma_{X|\Theta=\theta}^2} + \frac{m}{\sigma_\Theta^2}}{\frac{N}{\sigma_{X|\Theta=\theta}^2} + \frac{1}{\sigma_\Theta^2}}. \tag{6.63}$$

[8]Since the realizations of the statistics $X_1, \ldots, X_N$ are conditioned on the unknown realization of Θ, we obtain $c_{\Theta,S} = \mathsf{E}\left[\Theta S\right] - \mathsf{E}\left[\Theta\right]\mathsf{E}\left[S\right] = \mathsf{E}\left[\mathsf{E}\left[\Theta S|\Theta\right]\right] - \mathsf{E}\left[\Theta\right]\mathsf{E}\left[\mathsf{E}\left[S|\Theta\right]\right] = \mathsf{E}\left[\Theta^2\right] - (\mathsf{E}\left[\Theta\right])^2 = \sigma_\Theta^2$ and $\sigma_S^2 = \mathsf{E}\left[S^2\right] - (\mathsf{E}\left[S\right])^2 = \mathsf{E}\left[\mathsf{E}\left[S^2|\Theta\right]\right] - (\mathsf{E}\left[\mathsf{E}\left[S|\Theta\right]\right])^2 = \mathsf{E}\left[\mathsf{Var}\left[S|\Theta\right] + \mathsf{E}\left[S|\Theta\right]^2\right] - (\mathsf{E}\left[\Theta\right])^2 = \mathsf{E}\left[\frac{\sigma_{X|\Theta=\theta}^2}{N} + \Theta^2\right] - m^2 = \frac{\sigma_{X|\Theta=\theta}^2}{N} + \sigma_\Theta^2$.

Part IV

Linear Estimation

7. Linear Estimation

In this chapter, we directly focus on linear models, i.e., given an observation $\boldsymbol{x} \in \mathbb{R}^d$ corresponding to the respective random variable, we consider the estimation of the realization of the random variable y by

$$T_{\mathsf{Lin}} : \mathbb{R}^d \to \mathbb{R},\ \boldsymbol{x} \mapsto \hat{y}, \quad \hat{y} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{t} + m. \tag{7.1}$$

In contrast to the Maximum-Likelihood Estimators and the Conditional Mean Estimators, where the inference of the estimator is based on statistical parameters of a given Probability Distribution Function, here the estimation of y, by inference of $\boldsymbol{t}$ and m, is based on N pairs of jointly drawn observations $(\boldsymbol{x}_i, y_i)$, where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N \quad \text{and} \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{N \times d}. \tag{7.2}$$

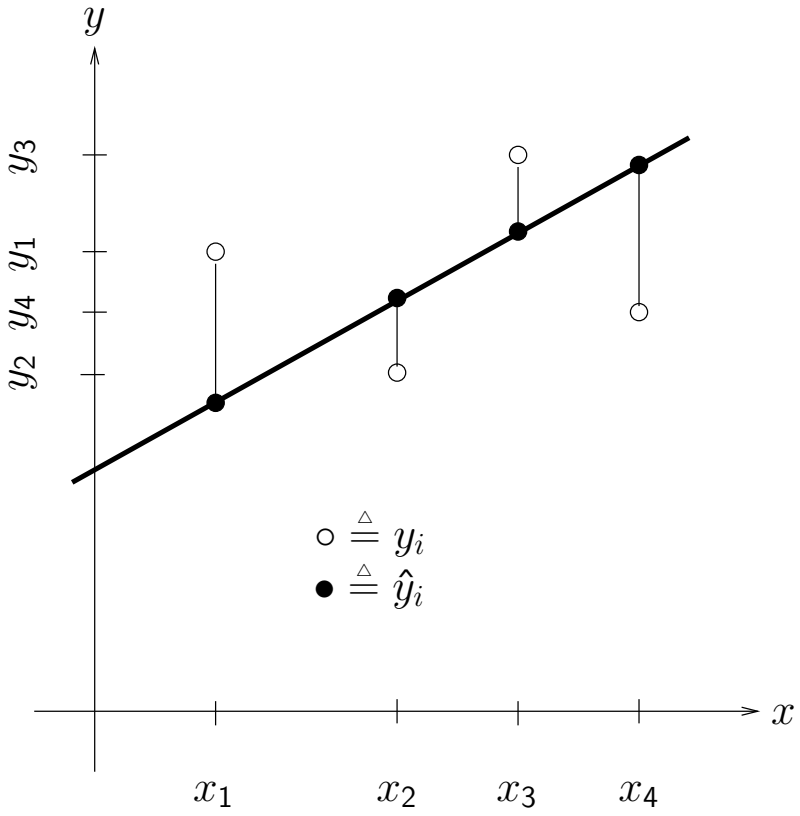


Fig. 7.1: Principle of linear estimation: $\hat{y}_{\mathsf{Lin}} = xt_1 + t_2$

7.1 Least Squares Estimation

In Least Squares Estimation (LS), the inference of $\boldsymbol{t}$ is based on the minimization of the sum of squared errors between N pairs of observations y_i and outcomes $\hat{y}_i$ of the Linear Model[1] $\hat{y} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{t}$ with $\boldsymbol{x}$ and $\boldsymbol{t} \in \mathbb{R}^d$.

The related Optimization Problem is equal to

$$\min_{\boldsymbol{t}\in\mathbb{R}^d} \left\{ \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{t})^2 \right\} \quad \Leftrightarrow \quad \min_{\boldsymbol{t}\in\mathbb{R}^d} \left\| \begin{bmatrix} y_1 - \boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{t} \\ \vdots \\ y_N - \boldsymbol{x}_N^{\mathsf{T}}\boldsymbol{t} \end{bmatrix} \right\|_2^2 \quad \Leftrightarrow \quad \min_{\boldsymbol{t}\in\mathbb{R}^d} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}\|_2^2 . \tag{7.3}$$

From a Subspace Perspective, we search for a vector $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{t}_{\mathsf{LS}} \in \mathsf{range}\,[\boldsymbol{X}] \in \mathbb{R}^N$, which is the Best Approximation of $\boldsymbol{y} \in \mathbb{R}^N$, i.e., $\min_{\hat{\boldsymbol{y}}\in\mathsf{range}[\boldsymbol{X}]} \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$ with

$$\boldsymbol{y} - \hat{\boldsymbol{y}} \perp \mathsf{range}\,[\boldsymbol{X}] \quad \Leftrightarrow \quad \boldsymbol{y} - \hat{\boldsymbol{y}} \in \mathsf{null}\left[\boldsymbol{X}^{\mathsf{T}}\right] \quad \Leftrightarrow \quad \boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{0}. \tag{7.4}$$

The resulting vector solves $\boldsymbol{X}^{\mathsf{H}}\boldsymbol{y} - \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{t}_{\mathsf{LS}} = \boldsymbol{0}$ and thus (under favorable conditions)

$$\boldsymbol{t}_{\mathsf{LS}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}. \tag{7.5}$$

[1] The affine case $y = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{t} + t_0$ can similarly be treated by introducing $y = \boldsymbol{x}'^{,\mathsf{T}}\boldsymbol{t}'$, with $\boldsymbol{t}' = \begin{bmatrix} \boldsymbol{t} \\ t_0 \end{bmatrix}$ and $\boldsymbol{x}' = \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}$ and therefore $d' = d + 1$.

In other words, the BEST APPROXIMATION $\hat{\boldsymbol{y}}$ is found by the ORTHOGONAL PROJECTION onto range $[\boldsymbol{X}]$.

The appropriate LEFT INVERSE of $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ in order to determine the ORTHOGONAL PROJECTOR on range $[\boldsymbol{X}]$ is given by the PSEUDO-INVERSE

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathsf{T}} \quad (\text{LEFT INVERSE of } \boldsymbol{X}), \tag{7.6}$$

i.e.,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{X}^{+}\boldsymbol{y} = \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}}_{\text{ORTHOGONAL PROJECTOR}}\boldsymbol{y}. \tag{7.7}$$

Taking into account the estimation model $\hat{y} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{t}$ and $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{t}$, we consequently again obtain

$$\boldsymbol{t}_{\mathsf{LS}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}. \tag{7.8}$$

Note. In order to obtain a feasible solution, FAVORABLE CONDITIONS are required. To this end, the existance of the INVERSE MATRIX of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ is required, i.e., $N \geq d$ linear independent measurement vectors $\boldsymbol{x}_i$ must be available.

7.2 SVD Perspective

By means of the Singular Value Decomposition[2] of the Observation Matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ with rank $[\boldsymbol{X}] = d$, we obtain

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}\|_2^2, \quad \text{with} \quad \boldsymbol{X} = \left[\boldsymbol{U}, \boldsymbol{U}^{\perp}\right] \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^{\mathsf{T}} \tag{7.9}$$

where $\boldsymbol{U} \in \mathbb{R}^{N \times d}$, $\boldsymbol{U}^{\perp} \in \mathbb{R}^{N \times (N-d)}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, and $\boldsymbol{V} \in \mathbb{R}^{d \times d}$. After some reformulations,[3] the cost function is equal to

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}\|_2^2 &= \left\| \boldsymbol{y} - \left[\boldsymbol{U}, \boldsymbol{U}^{\perp}\right] \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^{\mathsf{T}} \boldsymbol{t} \right\|_2^2 = \left\| \left[\boldsymbol{U}, \boldsymbol{U}^{\perp}\right] \left[\begin{bmatrix} \boldsymbol{U}^{\mathsf{T}} \\ \boldsymbol{U}^{\perp,\mathsf{T}} \end{bmatrix} \boldsymbol{y} - \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^{\mathsf{T}} \boldsymbol{t} \right] \right\|_2^2 \\ &= \left\| \begin{bmatrix} \boldsymbol{U}^{\mathsf{T}} \\ \boldsymbol{U}^{\perp,\mathsf{T}} \end{bmatrix} \boldsymbol{y} - \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^{\mathsf{T}} \boldsymbol{t} \right\|_2^2 = \left\| \begin{bmatrix} \boldsymbol{U}^{\mathsf{T}} \boldsymbol{y} - \boldsymbol{\Sigma} \boldsymbol{V}^{\mathsf{T}} \boldsymbol{t} \\ \boldsymbol{U}^{\perp,\mathsf{T}} \boldsymbol{y} \end{bmatrix} \right\|_2^2. \end{aligned}$$

Since the lower part of the vector does not depend on $\boldsymbol{t}$, the inequality $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}\|_2^2 \geq \left\|\boldsymbol{U}^{\perp,\mathsf{T}}\boldsymbol{y}\right\|_2^2$ holds and the minimum value is achieved if

$$\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{t}_{\mathsf{LS}} = \boldsymbol{0}_d \quad \Leftrightarrow \quad \boldsymbol{t}_{\mathsf{LS}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} \in \mathbb{R}^d. \tag{7.10}$$

Note: The residual error $\left\|\boldsymbol{U}^{\perp,\mathsf{T}}\boldsymbol{y}\right\|_2^2$ corresponds to the orthogonal projection of $\boldsymbol{y}$ onto span $\left[\boldsymbol{U}^{\perp}\right]$.

[2] For real-valued matrices the singular vectors can be chosen real-valued, too.

[3] If $\boldsymbol{Q}$ is unitary, $\|\boldsymbol{Q}\boldsymbol{x}\|_2^2 = \|\boldsymbol{x}\|_2^2 = (\boldsymbol{Q}\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{Q}\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$.

7.3 Examples

Mean Estimation

In order to estimate the mean μ_X from an unknown distribution $f_X(x)$ based on N observations x_i by means of LEAST SQUARES ESTIMATION, we introduce the linear model

$$T_{\text{Lin}} : \mathbb{R} \to \mathbb{R},\ 1 \mapsto \hat{x}, \quad \hat{x}_i = t1, \quad \text{with} \quad t = \hat{\mu}_X, \tag{7.11}$$

and thus the LS optimization problem

$$\min_t \left\{ \sum_{i=1}^{N} (x_i - 1t)^2 \right\} \quad \Leftrightarrow \quad \min_t \left\{ \|\boldsymbol{x} - \mathbf{1}t\|^2 \right\}, \tag{7.12}$$

with

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \quad \text{and} \quad \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{7.13}$$

The resulting LS Estimator is determined by (7.5) and reads

$$\hat{\mu}_X = t_{\text{LS}} = (\mathbf{1}^{\mathsf{T}}\mathbf{1})^{-1}\mathbf{1}^{\mathsf{T}}\boldsymbol{x} = \frac{\mathbf{1}^{\mathsf{T}}\boldsymbol{x}}{\mathbf{1}^{\mathsf{T}}\mathbf{1}} = \frac{\sum_{i=1}^{N} x_i}{\mathbf{1}^{\mathsf{T}}\mathbf{1}} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{7.14}$$

LS Estimation (scalar and linear)

Given a sample set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, i.e., the vector of observations $x_i \in \mathbb{R}$ and the vector of elements $y_i \in \mathbb{R}$,

$$\boldsymbol{x} \triangleq \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix},$$

a linear estimator is denoted by

$$T_{\mathsf{Lin}} : \mathbb{R} \to \mathbb{R},\ x \mapsto \hat{y}, \quad \hat{y} = xt. \tag{7.15}$$

Minimizing the Euklidean Norm of $\boldsymbol{y} - \hat{\boldsymbol{y}}$ with $\hat{\boldsymbol{y}} = \boldsymbol{x}t$ by applying the Best Approximation Principle yields

$$\boldsymbol{y} - \boldsymbol{x}t_{\mathsf{LS}} \perp \mathsf{range}\,[\boldsymbol{x}] \qquad \Leftrightarrow \qquad \boldsymbol{y} - \boldsymbol{x}t_{\mathsf{LS}} \in \mathsf{null}\left[\boldsymbol{x}^{\mathsf{T}}\right],$$

and thus

$$\boldsymbol{x}^{\mathsf{T}}\left(\boldsymbol{y} - \boldsymbol{x}t_{\mathsf{LS}}\right) = 0. \tag{7.16}$$

If follows for the Least Squares optimal weight:

$$t_{\mathsf{LS}} = \frac{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}}{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}}. \tag{7.17}$$

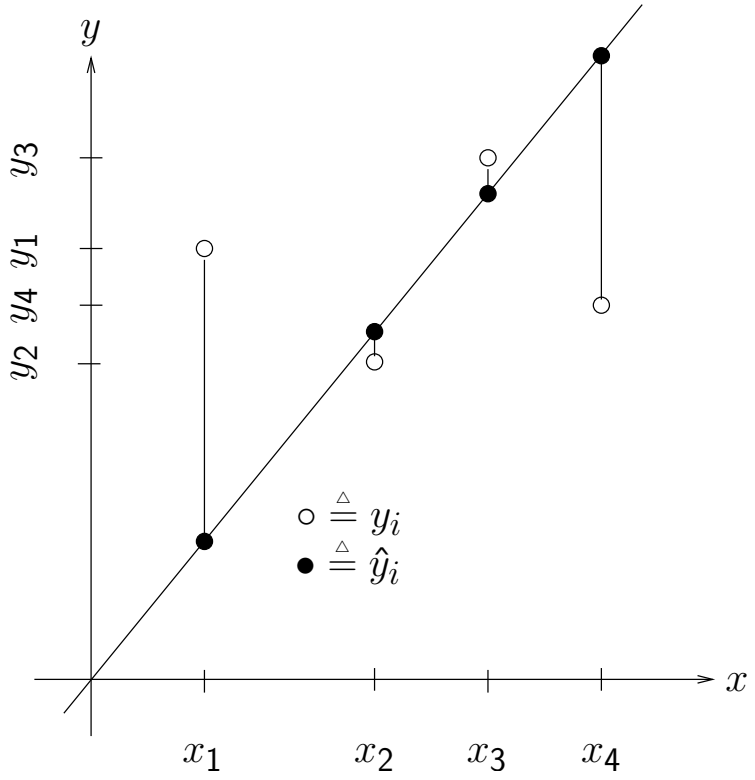


Fig. 7.2: Principle of scalar linear estimation: $\hat{y}_{\mathsf{LS}} = xt_{\mathsf{LS}}$

LS Estimation (scalar and affine)

Assuming the linear estimator as

$$T_{\mathsf{Lin}} : \mathbb{R} \to \mathbb{R},\ x \mapsto \hat{y}, \quad \hat{y} = xt_1 + t_2 = [x\ 1] \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{t}, \tag{7.18}$$

the vector of observations x_i becomes a matrix of observation vectors $\boldsymbol{x}_i^{\mathsf{T}} = [x_i\ 1]$, i.e.,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} [x_1\ 1] \\ [x_2\ 1] \\ \vdots \\ [x_N\ 1] \end{bmatrix}.$$

Since $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{t}$, or in other words $\hat{\boldsymbol{y}} \in \mathsf{range}\,[\boldsymbol{X}]$, the Best Approximation Principle yields:

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}_{\mathsf{LS}} \perp \mathsf{range}\,[\boldsymbol{X}] \qquad \Leftrightarrow \qquad \boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}_{\mathsf{LS}} \in \mathsf{null}\left[\boldsymbol{X}^{\mathsf{T}}\right],$$

and

$$\boldsymbol{X}^{\mathsf{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}_{\mathsf{LS}}) = \boldsymbol{0}. \tag{7.19}$$

Given $N \geq 2$ and at least 2 observation vectors form a linear independent basis of $\mathbb{R}^2$,

$$\boldsymbol{t}_{\mathsf{LS}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}. \tag{7.20}$$

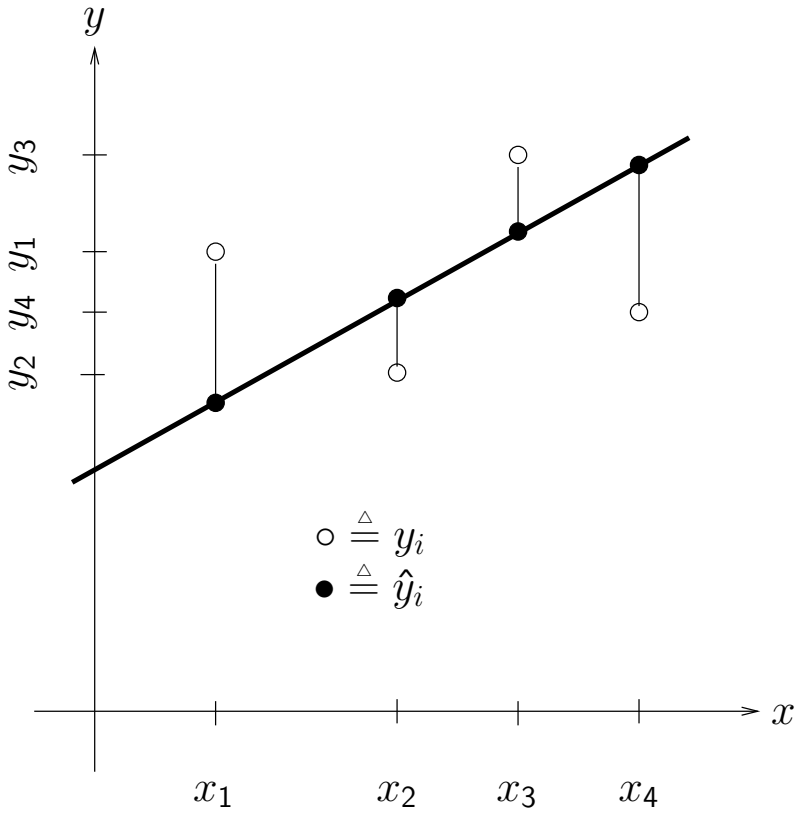


Fig. 7.3: Principle of scalar affine estimation: $\hat{y}_{\mathsf{LS}} = xt_{\mathsf{LS},1} + t_{\mathsf{LS},2}$

LS Estimation (multi-dimensional and linear)

We now consider the most general case with $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$, i.e., the matrix of observation vectors $\boldsymbol{x}_i^{\mathsf{T}}$ and $\boldsymbol{y}_i^{\mathsf{T}}$,

$$\boldsymbol{X} \triangleq \begin{bmatrix} \boldsymbol{x}_1^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathsf{T}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{Y} \triangleq \begin{bmatrix} \boldsymbol{y}_1^{\mathsf{T}} \\ \vdots \\ \boldsymbol{y}_N^{\mathsf{T}} \end{bmatrix}. \tag{7.21}$$

The optimal linear estimator with respect to the LS criterion, i.e.,

$$\hat{\boldsymbol{y}}_{\mathsf{LS}}^{\mathsf{T}} = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{T}_{\mathsf{LS}}, \tag{7.22}$$

is found by applying the BEST APPROXIMIATION PRINCIPLE columnwise:

$$\text{columnwise}: \quad \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{T}_{\mathsf{LS}} \perp \text{range}\left[\boldsymbol{X}\right] \quad \Leftrightarrow \quad \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{T}_{\mathsf{LS}} \in \text{null}\left[\boldsymbol{X}^{\mathsf{T}}\right],$$

and

$$\boldsymbol{X}^{\mathsf{T}} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{T}_{\mathsf{LS}}\right) = \boldsymbol{0}. \tag{7.23}$$

If N is larger than the dimension of observation vectors $\boldsymbol{x}_i$, i.e., if $N > d$ and the columns of $\boldsymbol{X}$ are linear independent such that $\text{rank}\left[\boldsymbol{X}\right] = d$, it follows that

$$\boldsymbol{T}_{\mathsf{LS}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{Y}. \tag{7.24}$$

7.4 Linear Minimum Mean Square Error Estimation

The Linear Minimum Mean Square Error (LMMSE) estimator is the estimator which minimizes the MSE based on a Linear Model for the estimator,

$$T_{\text{Lin}} : \mathbb{R}^d \to \mathbb{R},\ \boldsymbol{x} \mapsto \hat{y}, \quad \hat{y} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{t} + m \quad \text{or} \quad \hat{y} = \boldsymbol{t}^{\mathsf{T}}\boldsymbol{x} + m, \tag{7.25}$$

i.e., the LMMSE Estimator is minimizer of the optimization problem[4] $\min_{\boldsymbol{t},m} \mathsf{E}\left[\left\|y - \boldsymbol{t}^{\mathsf{T}}\boldsymbol{x} - m\right\|^2\right]$.

Given the joint PDF of the random variables $\boldsymbol{z} = (\boldsymbol{x}, y)^{\mathsf{T}}$, with mean values and covariances

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{x}} \\ \mu_y \end{bmatrix} \quad \text{and} \quad \boldsymbol{C}_{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{C}_{\boldsymbol{x}} & \boldsymbol{c}_{\boldsymbol{x},y} \\ \boldsymbol{c}_{y,\boldsymbol{x}} & \boldsymbol{c}_y \end{bmatrix}, \tag{7.26}$$

the LMMSE Estimator is given by

$$\hat{y} = \mu_y + \boldsymbol{c}_{y,\boldsymbol{x}}\boldsymbol{C}_{\boldsymbol{x}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}) = \underbrace{\boldsymbol{c}_{y,\boldsymbol{x}}\boldsymbol{C}_{\boldsymbol{x}}^{-1}}_{=\boldsymbol{t}^{\mathsf{T}}}\boldsymbol{x}\ \underbrace{+\mu_y - \boldsymbol{c}_{y,\boldsymbol{x}}\boldsymbol{C}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}}}_{=m}. \tag{7.27}$$

[4] In contrast to the standard notation, we denote random vectors by $\boldsymbol{x}$ in order to avoid any confusions with matrix notation $\boldsymbol{X}$.

Proof.

Applying the Orthogonality Principle

$$\begin{aligned}\mathsf{E}\left[(y-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{x}-m)\boldsymbol{x}^{\mathsf{T}}\right] &= \boldsymbol{c}_{y,\boldsymbol{x}}+\mu_y\boldsymbol{\mu}_{\boldsymbol{x}}^{\mathsf{T}}-\boldsymbol{t}^{\mathsf{T}}(\boldsymbol{C}_x+\boldsymbol{\mu}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}}^{\mathsf{T}})-m\boldsymbol{\mu}_{\boldsymbol{x}}^{\mathsf{T}}=0\\ \mathsf{E}\left[(y-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{x}-m)1\right] &= \mu_y-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\mu}_{\boldsymbol{x}}-m=0,\end{aligned}$$

we obtain $m=\mu_y-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\mu}_{\boldsymbol{x}}$ and

$$\begin{aligned}\boldsymbol{c}_{y,\boldsymbol{x}}+\mu_y\boldsymbol{\mu}_{\boldsymbol{x}}^{\mathsf{T}}-\boldsymbol{t}^{\mathsf{T}}(\boldsymbol{C}_x+\boldsymbol{\mu}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}}^{\mathsf{T}})-(\mu_y-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\mu}_{\boldsymbol{x}})\boldsymbol{\mu}_{\boldsymbol{x}}^{\mathsf{T}}&=0\\ \boldsymbol{c}_{y,\boldsymbol{x}}-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{C}_x&=0,\end{aligned}$$

and thus $\boldsymbol{t}^{\mathsf{T}}=\boldsymbol{c}_{y,\boldsymbol{x}}\boldsymbol{C}_x^{-1}$. The resulting estimator reads

$$\begin{aligned}\hat{y}&=\boldsymbol{t}^{\mathsf{T}}\boldsymbol{x}+\mu_y-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\mu}_{\boldsymbol{x}}\\ &=\mu_y+\boldsymbol{t}^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}})\\ &=\mu_y+\boldsymbol{c}_{y,\boldsymbol{x}}\boldsymbol{C}_{\boldsymbol{x}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{x}}).\end{aligned}$$

Again applying the Orthogonality Principle, the achievable MMSE is equal to

$$\begin{aligned}\mathsf{E}\left[\left\|y-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{x}-m\right\|^2\right] &= \mathsf{E}\left[(y-\boldsymbol{t}^{\mathsf{T}}\boldsymbol{x}-m)y\right]\\ &= c_y+\mu_y^2-\boldsymbol{t}^{\mathsf{T}}(\boldsymbol{c}_{\boldsymbol{x},y}+\boldsymbol{\mu}_{\boldsymbol{x}}\mu_y)-m\mu_y\\ &= c_y+\mu_y^2-\boldsymbol{c}_{y,\boldsymbol{x}}\boldsymbol{C}_x^{-1}\boldsymbol{c}_{\boldsymbol{x},y}+(m-\mu_y)\mu_y-m\mu_y\\ &= c_y-\boldsymbol{c}_{y,\boldsymbol{x}}\boldsymbol{C}_x^{-1}\boldsymbol{c}_{\boldsymbol{x},y}. \qquad (7.28)\end{aligned}$$

Special Case.

In the case of ZERO-MEAN random variables,

$$\mathsf{E}\left[\begin{bmatrix} \boldsymbol{x} \\ y \end{bmatrix}\right] = \begin{bmatrix} \boldsymbol{\mu_x} \\ \mu_y \end{bmatrix} = \mathbf{0}, \tag{7.29}$$

the LMMSE Estimator and its minimum MSE is

$$\text{LMMSE ESTIMATOR:} \quad \hat{y} = \boldsymbol{c}_{y,\boldsymbol{x}} \boldsymbol{C}_{\boldsymbol{x}}^{-1} \boldsymbol{x}, \tag{7.30}$$

$$\text{MINIMUM MSE:} \quad \mathsf{E}\left[c_{y,\boldsymbol{x}}\right] = c_y - \boldsymbol{t}^{\mathsf{T}} \boldsymbol{c}_{\boldsymbol{x},y} = c_y - \boldsymbol{c}_{y,\boldsymbol{x}} \boldsymbol{C}_x^{-1} \boldsymbol{c}_{\boldsymbol{x},y}. \tag{7.31}$$

Part V

Examples

8. Estimation of a Matrix Channel

The estimation of a MULTIPLE-INPUT MULTIPLE OUTPUT (MIMO) channel is considered. In particular we assume a MIMO channel with K antenna elements at the transmitter and M antenna elements at the receiver, which means KM transmission channels to be estimated.

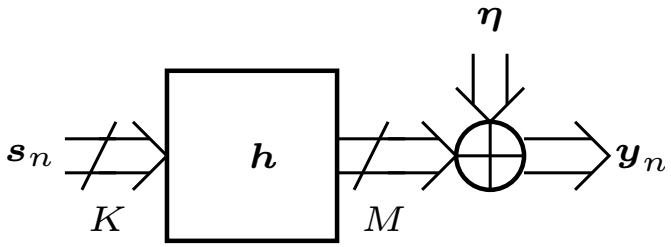


Fig. 8.1: Estimation of a MIMO channel.

Three Linear Estimators,

$$\hat{\boldsymbol{h}} = \boldsymbol{T}\boldsymbol{y} \in \mathbb{C}^{KM}, \tag{8.1}$$

are introduced and compared:

- The Minimum Mean Square Error Estimator (MMSE),
- the Maximum Likelihood Estimator (ML),
- and the "Matched Filter" Estimator (MF).

8.1 Model for Training Channel

The addressed model allows to cover a variety of scenarios, including TIME-VARIANT and DISPERSIVE channels. For the sake of simplicity, we assume the simplest case of an TIME-INVARIANT NON-DISPERSIVE MIMO channel. The task is to find good estimates of the channel coefficients

$$h_{m,k}, \qquad \begin{matrix} m = 1, \ldots, M \\ k = 1, \ldots, K, \end{matrix} \tag{8.2}$$

where $h_{m,k}$ denotes the channel coefficient from the kth transmitter to the mth receiver.

To this end, we assume N vectors of TRAINING SIGNALS $\boldsymbol{s}_n \in \mathbb{C}^K$, $n = 1, \ldots, N$. The estimation of the channel coefficients $h_{m,k}$ is based on the received signal vectors

$$\boldsymbol{y}_n = \boldsymbol{H}\boldsymbol{s}_n + \boldsymbol{\eta}_n, \quad n = 1, \ldots, N, \tag{8.3}$$

where $\boldsymbol{H}$, $\boldsymbol{y}_n$ and $\boldsymbol{\eta}_n$ denotes the matrix of channel coefficients, the received signal vector and the noise corruption at the receiver for the nth training vector, respectively.

The model for the training channel is thus given by

$$\underbrace{[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]}_{\boldsymbol{Y}} = \boldsymbol{H} \underbrace{[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]}_{\bar{\boldsymbol{S}}} + \underbrace{[\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N]}_{\boldsymbol{N}}, \tag{8.4}$$

i.e., $\boldsymbol{Y} = \boldsymbol{H}\bar{\boldsymbol{S}} + \boldsymbol{N}$.

By stacking the column vectors of the matrices, $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$, $\boldsymbol{h} = \mathrm{vec}(\boldsymbol{H})$ and $\boldsymbol{n} = \mathrm{vec}(\boldsymbol{N})$, we obtain[1]

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{h} + \boldsymbol{n} \tag{8.5}$$

$$= (\bar{\boldsymbol{S}}^{\mathsf{T}} \otimes \boldsymbol{I}_M)\boldsymbol{h} + \boldsymbol{n}, \qquad \boldsymbol{I}_M \in \mathbb{C}^{M \times M}. \tag{8.6}$$

The so called KRONECKER PRODUKT $\bar{\boldsymbol{S}}^{\mathsf{T}} \otimes \boldsymbol{I}_M$ means

$$\bar{\boldsymbol{S}}^{\mathsf{T}} \otimes \boldsymbol{I}_M = \begin{bmatrix} s_{1,1}\boldsymbol{I}_M & s_{1,2}\boldsymbol{I}_M & \cdots & s_{1,K}\boldsymbol{I}_M \\ s_{2,1}\boldsymbol{I}_M & s_{2,2}\boldsymbol{I}_M & \cdots & s_{2,K}\boldsymbol{I}_M \\ \vdots & \vdots & & \vdots \\ s_{N,1}\boldsymbol{I}_M & s_{K,2}\boldsymbol{I}_M & \cdots & s_{N,K}\boldsymbol{I}_M \end{bmatrix} \in \mathbb{C}^{NM \times KM}. \tag{8.7}$$

[1] Here we use the following identity for matrix equations: $\boldsymbol{A}\boldsymbol{X}\boldsymbol{B} = \boldsymbol{C} \Leftrightarrow (\boldsymbol{B}^{\mathsf{T}} \otimes \boldsymbol{A})\,\mathrm{vec}(\boldsymbol{X}) = \mathrm{vec}(\boldsymbol{C})$.

Example. K=M=N=2.

$$\begin{aligned}\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} &= \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} + \begin{bmatrix} \eta_{11} & \eta_{12} \\ \eta_{21} & \eta_{22} \end{bmatrix} \\ &= \begin{bmatrix} h_{11}s_{11} + h_{12}s_{21} & h_{11}s_{12} + h_{12}s_{22} \\ h_{21}s_{11} + h_{22}s_{21} & h_{21}s_{12} + h_{22}s_{22} \end{bmatrix} + \begin{bmatrix} \eta_{11} & \eta_{12} \\ \eta_{21} & \eta_{22} \end{bmatrix}\end{aligned}$$

$$\begin{aligned}\begin{bmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \end{bmatrix} &= \begin{bmatrix} h_{11}s_{11} + h_{12}s_{21} \\ h_{21}s_{11} + h_{22}s_{21} \\ h_{11}s_{12} + h_{12}s_{22} \\ h_{21}s_{12} + h_{22}s_{22} \end{bmatrix} + \begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix} \\ &= \begin{bmatrix} s_{11} & 0 & s_{21} & 0 \\ 0 & s_{11} & 0 & s_{21} \\ s_{12} & 0 & s_{22} & \\ 0 & s_{12} & 0 & s_{22} \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{21} \\ h_{12} \\ h_{22} \end{bmatrix} + \begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix} \\ &= \begin{bmatrix} s_{11} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & s_{21} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ s_{12} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & s_{22} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{21} \\ h_{12} \\ h_{22} \end{bmatrix} + \begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix} \\ &= \begin{bmatrix} s_{11} & s_{21} \\ s_{12} & s_{22} \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{21} \\ h_{12} \\ h_{22} \end{bmatrix} + \begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix}.\end{aligned}$$

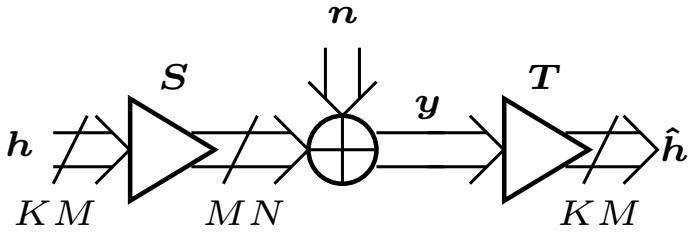


Fig. 8.2: Model for the estimation of a MIMO channel.

Further Assumptions.

We further assume that the stacked vector of channel coefficients $\boldsymbol{h} \in \mathbb{C}^{KM}$ and the stacked vector of distortions $\boldsymbol{n} \in \mathbb{C}^{NM}$ are GAUSSIAN DISTRIBUTED as[2]

$$\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{h}}) \quad \text{and} \quad \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{n}}). \tag{8.8}$$

In the following, we assume no correlations between noise vectors at different time instances, thus leading to a covariance $\boldsymbol{C}_{\boldsymbol{n}} = \boldsymbol{I}_N \otimes \boldsymbol{C}_{\boldsymbol{\eta}} \in \mathbb{C}^{NM \times NM}$.

Additionally, assuming no correlations between distortions of adjacent antenna elements leads to $\boldsymbol{C}_{\boldsymbol{n}} = \sigma_\eta^2 \boldsymbol{I}_{NM}$.

The covariance matrix of the channel vector in general also shows additional structural properties, which are not taken into account in the following.

The channel vectors $\boldsymbol{h}$ and the noise distortions $\boldsymbol{n}$ are assumed to be STOCHASTICALLY INDEPENDENT, and thus UNCORRELATED ($\text{Cov}\left[\boldsymbol{h}, \boldsymbol{n}^{\mathsf{H}}\right] = \boldsymbol{0}$).

Note. Not taking into account these structural properties does not change the channel estimates, however, might be very useful to design more efficient algorithms.

[2] In contrast to the standard notation, we again denote random vectors by $\boldsymbol{x}$ in order to avoid any confusions with matrix notation $\boldsymbol{X}$.

Due to the LINEAR CHANNEL MODEL $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{h} + \boldsymbol{n}$, we conclude that the stacked vector of signal vector at the receiver and the unknown channel vector $\boldsymbol{z} = [\boldsymbol{y}^{\mathsf{T}}, \boldsymbol{h}^{\mathsf{T}}]^{\mathsf{T}}$ are JOINTLY GAUSSIAN DISTRIBUTED. The covariances of the JOINT GAUSSIAN PDF is equal to

$$\boldsymbol{C}_{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{C}_{\boldsymbol{y}} & \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{h}} \\ \boldsymbol{C}_{\boldsymbol{h},\boldsymbol{y}} & \boldsymbol{C}_{\boldsymbol{h}} \end{bmatrix}, \tag{8.9}$$

with

$$\boldsymbol{C}_{\boldsymbol{y}} = \mathsf{Var}[\boldsymbol{y}] = \mathsf{E}\left[\boldsymbol{y}\boldsymbol{y}^{\mathsf{H}}\right] = \mathsf{E}\left[(\boldsymbol{S}\boldsymbol{h} + \boldsymbol{n})(\boldsymbol{S}\boldsymbol{h} + \boldsymbol{n})^{\mathsf{H}}\right] \cdots = \boldsymbol{S}\boldsymbol{C}_{\boldsymbol{h}}\boldsymbol{S}^{\mathsf{H}} + \boldsymbol{C}_{\boldsymbol{n}}, \tag{8.10}$$

$$\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{h}} = \mathsf{Cov}[\boldsymbol{y}, \boldsymbol{h}] = \mathsf{E}\left[\boldsymbol{y}\boldsymbol{h}^{\mathsf{H}}\right] = \mathsf{E}\left[(\boldsymbol{S}\boldsymbol{h} + \boldsymbol{n})\boldsymbol{h}^{\mathsf{H}}\right] \cdots = \boldsymbol{S}\boldsymbol{C}_{\boldsymbol{h}}, \tag{8.11}$$

$$\boldsymbol{C}_{\boldsymbol{h},\boldsymbol{y}} = \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{h}}^{\mathsf{H}} = \boldsymbol{C}_{\boldsymbol{h}}\boldsymbol{S}^{\mathsf{H}}, \tag{8.12}$$

where we intensively applied $\mathsf{Cov}\left[\boldsymbol{h}, \boldsymbol{n}^{\mathsf{H}}\right] = \mathsf{E}\left[\boldsymbol{h}\boldsymbol{n}^{\mathsf{H}}\right] - \boldsymbol{\mu}_{\boldsymbol{h}}\boldsymbol{\mu}_{\boldsymbol{n}}^{\mathsf{T}} = \boldsymbol{0}$.

Note. For the following estimators, we assume full knowledge of the covariance matrices

$$\boldsymbol{C}_{\boldsymbol{y}} \in \mathbb{C}^{NM \times NM} \quad \text{and} \quad \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{h}} \in \mathbb{C}^{NM \times KM}. \tag{8.13}$$

In practice the required covariance matrices must be estimated as well!

8.2 Minimum Mean Square Error Estimator

Since we assume JOINTLY GAUSSIAN DISTRIBUTED RANDOM VARIABLES, the CONDITIONAL MEAN ESTIMATOR is identical with the linear MMSE Estimator $\boldsymbol{T}_{\mathsf{MMSE}}$, the minimizer of

$$\min_{\boldsymbol{T}} \left\{ \mathsf{E}\left[\|\boldsymbol{h} - \boldsymbol{T}(\boldsymbol{S}\boldsymbol{h} + \boldsymbol{n})\|^2 \right] \right\}, \tag{8.14}$$

which is given by

$$\hat{\boldsymbol{h}}_{\mathsf{MMSE}} = \boldsymbol{T}_{\mathsf{MMSE}}\, \boldsymbol{y}, \tag{8.15}$$

$$\begin{aligned} \boldsymbol{T}_{\mathsf{MMSE}} &= \boldsymbol{C}_{\boldsymbol{h},\boldsymbol{y}} \boldsymbol{C}_{\boldsymbol{y}}^{-1} && (8.16)\\ &= \boldsymbol{C}_{\boldsymbol{h}} \boldsymbol{S}^{\mathsf{H}} \left(\boldsymbol{S}\boldsymbol{C}_{\boldsymbol{h}} \boldsymbol{S}^{\mathsf{H}} + \boldsymbol{C}_{\boldsymbol{n}} \right)^{-1} && (8.17)\\ &= \left(\boldsymbol{C}_{\boldsymbol{h}} \boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}}^{-1} \boldsymbol{S} + \boldsymbol{I}_{KM} \right)^{-1} \boldsymbol{C}_{\boldsymbol{h}} \boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}}^{-1}. && (8.18) \end{aligned}$$

Reformulation.

The reformulation of the MMSE ESTIMATOR in (8.18) is based on the MATRIX-INVERSION LEMMA,[3]

$$(\boldsymbol{A}+\boldsymbol{BCD})^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{DA}^{-1}. \tag{8.19}$$

It follows

$$\begin{aligned}
\boldsymbol{T}_{\text{MMSE}} &= \boldsymbol{C_h}\boldsymbol{S}^{\text{H}}\left(\boldsymbol{C_n}+\boldsymbol{S}\boldsymbol{C_h}\boldsymbol{S}^{\text{H}}\right)^{-1} \\
&= \boldsymbol{C_h}\boldsymbol{S}^{\text{H}}\left(\boldsymbol{C_n}^{-1}-\boldsymbol{C_n}^{-1}\boldsymbol{S}\left(\boldsymbol{C_h}^{-1}+\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\right) \\
&= \boldsymbol{C_h}\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}-\boldsymbol{C_h}\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\left(\boldsymbol{C_h}^{-1}+\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1} \\
&= \boldsymbol{C_h}\left(\boldsymbol{I}-\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\left(\boldsymbol{C_h}^{-1}+\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\right)^{-1}\right)\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1} \\
&= \boldsymbol{C_h}\left(\left(\boldsymbol{C_h}^{-1}+\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\right)-\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\right)\left(\boldsymbol{C_h}^{-1}+\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1} \\
&= \left(\boldsymbol{C_h}^{-1}+\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1} \\
&= \left(\boldsymbol{I}_{KM}+\boldsymbol{C_h}\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{C_h}\boldsymbol{S}^{\text{H}}\boldsymbol{C_n}^{-1}.
\end{aligned}$$

[3] Assuming the compatibility of the matrices and the existence of the corresponding inverse matrices, it can be shown that

$$\begin{aligned}
(\boldsymbol{A}+\boldsymbol{BCD})(\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{DA}^{-1}) &= \boldsymbol{I}+\boldsymbol{BCDA}^{-1}-\boldsymbol{B}\left(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{DA}^{-1}-\boldsymbol{BCDA}^{-1}\boldsymbol{B}\left(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{DA}^{-1} \\
&= \boldsymbol{I}+\boldsymbol{BCDA}^{-1}-\boldsymbol{B}\left(\boldsymbol{I}+\boldsymbol{CDA}^{-1}\boldsymbol{B}\right)\left(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{DA}^{-1} \\
&= \boldsymbol{I}+\boldsymbol{BCDA}^{-1}-\boldsymbol{BC}\underbrace{\left(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B}\right)\left(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B}\right)^{-1}}_{\boldsymbol{I}}\boldsymbol{DA}^{-1}=\boldsymbol{I}.
\end{aligned}$$

8.3 Maximum-Likelihood Estimator

If we consider the unknown channel vector as deterministic, the Jointly Distributed Random Variables $\boldsymbol{y}$ and $\boldsymbol{h}$ are replaced by $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{S}\boldsymbol{h}, \boldsymbol{C}_{\boldsymbol{y}|\boldsymbol{h}=\boldsymbol{h}}\right)$, with $\boldsymbol{C}_{\boldsymbol{y}|\boldsymbol{h}=\boldsymbol{h}} = \boldsymbol{C}_{\boldsymbol{n}}$, and the CONDITIONAL PDF

$$f_{\boldsymbol{y}|\boldsymbol{h}}(\boldsymbol{y}|\boldsymbol{h}) = \left(\pi^{NM} \det \boldsymbol{C}_{\boldsymbol{n}}\right)^{-1} \exp\left(-(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{h})^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}}^{-1} (\boldsymbol{y} - \boldsymbol{S}\boldsymbol{h})\right). \tag{8.20}$$

Consequently, the ML Estimator $\boldsymbol{T}_{\mathsf{ML}}$ is the minimizer of the optimization problem

$$\min_{\boldsymbol{T}} \left\{(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{h})^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}}^{-1} (\boldsymbol{y} - \boldsymbol{S}\boldsymbol{h})\right\}, \tag{8.21}$$

which results into

$$\hat{\boldsymbol{h}}_{\mathsf{ML}} = \boldsymbol{T}_{\mathsf{ML}}\, \boldsymbol{y}, \tag{8.22}$$

$$\boldsymbol{T}_{\mathsf{ML}} = \left(\boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}}^{-1} \boldsymbol{S}\right)^{-1} \boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}}^{-1}. \tag{8.23}$$

Note. Given the ML Estimator $\boldsymbol{T}_{\mathsf{ML}}$ the MMSE Estimator, can be denoted as

$$\boldsymbol{T}_{\mathsf{MMSE}} = \left(\boldsymbol{I} + \left(\boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}}^{-1} \boldsymbol{S}\right)^{-1} \boldsymbol{C}_{\boldsymbol{h}}^{-1}\right)^{-1} \boldsymbol{T}_{\mathsf{ML}} = \ldots \tag{8.24}$$

$$= \left(\boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}} \boldsymbol{S}\right)^{-\frac{1}{2}} \left(\boldsymbol{I}_{KM} + \left(\boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}} \boldsymbol{S}\right)^{-\frac{1}{2}} \boldsymbol{C}_{\boldsymbol{h}}^{-1} \left(\boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}} \boldsymbol{S}\right)^{-\frac{1}{2}}\right)^{-1} \left(\boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}} \boldsymbol{S}\right)^{\frac{1}{2}} \boldsymbol{T}_{\mathsf{ML}}. \tag{8.25}$$

8.4 Correlation Estimator

Assuming training signals such that $\boldsymbol{S}^{\mathsf{H}}\boldsymbol{S} \propto N\boldsymbol{I}_{KM\times KM}$ and further assuming AGWN, the ML Estimator is equal to the so called Correlation Estimator (C)

$$\hat{\boldsymbol{h}}_{\mathsf{C}} = \boldsymbol{T}_{\mathsf{C}}\, \boldsymbol{y}, \tag{8.26}$$

$$\boldsymbol{T}_{\mathsf{C}} = \left(\boldsymbol{S}^{\mathsf{H}}\boldsymbol{C}_{\boldsymbol{n}}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{\mathsf{H}}\boldsymbol{C}_{\boldsymbol{n}}^{-1} \propto \boldsymbol{S}^{\mathsf{H}}. \tag{8.27}$$

Note. The Correlation Estimator is the simplest estimator which is often applied beyond cases where it is identical with the ML Estimator.

8.5 Matched Filter Estimator

The Matched Filter Estimator (MF) takes the principle of Strong Correlations between the channel vector $\boldsymbol{h}$ and its $\hat{\boldsymbol{h}}$ as a cost function. In other words, the MF Estimator $\boldsymbol{T}_{\mathsf{MF}}$ is the maximizer of the optimization problem

$$\max_{\boldsymbol{T}} \left\{ \frac{\left| \mathsf{E}\left[\hat{\boldsymbol{h}}^{\mathsf{H}} \boldsymbol{h} \right] \right|^2}{\mathsf{tr}\left[\mathsf{Var}\left[\boldsymbol{T}\boldsymbol{n} \right] \right]} \right\}. \tag{8.28}$$

The corresponding MF estimator is equal to

$$\hat{\boldsymbol{h}}_{\mathsf{MF}} = \boldsymbol{T}_{\mathsf{MF}}\, \boldsymbol{y}, \tag{8.29}$$

$$\boldsymbol{T}_{\mathsf{MF}} \propto \boldsymbol{C}_{\boldsymbol{h}} \boldsymbol{S}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{n}}^{-1}. \tag{8.30}$$

8.6 Special Case

In order to simplify the further analysis, we assume training signals such that

$$\boldsymbol{S}^{\mathsf{H}}\boldsymbol{S} = N\sigma_{\boldsymbol{s}}^2\boldsymbol{I}_{KM} \tag{8.31}$$

$$\boldsymbol{C}_{\boldsymbol{n}} = \sigma_{\eta}^2\boldsymbol{I}_{NM}, \tag{8.32}$$

where $\sigma_{\boldsymbol{s}}^2$ and σ_{η}^2 are the variances of single training and noise distortions signals.

Using these assumptions the proposed estimators can be denoted as follows:

$$\boldsymbol{T}_{\mathsf{MF}} \propto \boldsymbol{C}_{\boldsymbol{h}}\boldsymbol{S}^{\mathsf{H}}, \tag{8.33}$$

$$\boldsymbol{T}_{\mathsf{ML}} = \left(\boldsymbol{S}^{\mathsf{H}}\boldsymbol{C}_{\boldsymbol{n}}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{\mathsf{H}}\boldsymbol{C}_{\boldsymbol{n}}^{-1} = \frac{1}{N\sigma_{\boldsymbol{s}}^2}\boldsymbol{S}^{\mathsf{H}}, \tag{8.34}$$

$$\boldsymbol{T}_{\mathsf{MMSE}} = \left(\boldsymbol{C}_{\boldsymbol{h}}\boldsymbol{S}^{\mathsf{H}}\boldsymbol{C}_{\boldsymbol{n}}^{-1}\boldsymbol{S} + \boldsymbol{I}_{KM}\right)^{-1}\boldsymbol{C}_{\boldsymbol{h}}\boldsymbol{S}^{\mathsf{H}}\boldsymbol{C}_{\boldsymbol{n}}^{-1} = \left(\boldsymbol{C}_{\boldsymbol{h}} + \frac{\sigma_{\eta}^2}{N\sigma_{\boldsymbol{s}}^2}\boldsymbol{I}_{KM}\right)^{-1}\boldsymbol{C}_{\boldsymbol{h}}\underbrace{\frac{1}{N\sigma_{\boldsymbol{s}}^2}\boldsymbol{S}^{\mathsf{H}}}_{\boldsymbol{T}_{\mathsf{ML}}}. \tag{8.35}$$

The simplified case allows a convenient asymptotic analysis for the three estimators.

$$\text{Low Noise Regime:} \qquad \lim_{\sigma_\eta^2 \to 0} \boldsymbol{T}_{\mathsf{MMSE}} = \boldsymbol{T}_{\mathsf{ML}} \tag{8.36}$$

$$\text{High Noise Regime :} \qquad \lim_{\sigma_\eta^2 \to \infty} \sigma_\eta^2 \boldsymbol{T}_{\mathsf{MMSE}} \propto \boldsymbol{T}_{\mathsf{MF}}. \tag{8.37}$$

Using the EIGEN VALUE DECOMPOSITION of the channel covariance matrix

$$\boldsymbol{C}_{\boldsymbol{h}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathsf{H}} = \sum_{i=1}^{KM} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{H}}, \tag{8.38}$$

the estimators are equal to

$$\boldsymbol{T}_{\mathsf{MF}} \propto \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{S}^{\mathsf{H}} \tag{8.39}$$

$$\boldsymbol{T}_{\mathsf{ML}} = \frac{1}{N\sigma_s^2}\boldsymbol{S}^{\mathsf{H}}, \tag{8.40}$$

$$\boldsymbol{T}_{\mathsf{MMSE}} = \boldsymbol{U}\left(\boldsymbol{\Lambda} + \frac{\sigma_\eta^2}{N\sigma_s^2}\boldsymbol{I}_{KM}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{T}_{\mathsf{ML}}. \tag{8.41}$$

Interpretation. Denoting the MMSE estimate in a series of the eigenvectors $\boldsymbol{u}_i$, $i = 1, \ldots, KM$,

$$\hat{\boldsymbol{h}}_{\mathsf{MMSE}} = \sum_{i=1}^{KM} \frac{\lambda_i}{\lambda_i + \frac{\sigma_\eta^2}{N\sigma_s^2}} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{H}} \hat{\boldsymbol{h}}_{\mathsf{ML}}, \tag{8.42}$$

the improvement by the MMSE Estimator can be interpreted as a WEIGHTING OF THE ML ESTIMATE which is optimally adapted to the eigenspaces of the random channel vector.

8.7 Bias/Variance Trade–Off

For the analysis of all introduced channel estimators, we consider the decomposition of the MEAN SQUARE ERROR into the estimator's SQUARED BIAS and VARIANCE. Since we compare estimators of different paradigms with respect to A PRIORI INFORMATION about the unkown parameter, we study the mean average of the respective Bias/Variance Decomposition. To this end, the MSE is decomposed by the following steps[4]

$$\mathsf{E}\left[\|\boldsymbol{h}-\boldsymbol{T}(\boldsymbol{S}\boldsymbol{h}+\boldsymbol{n})\|^2\right] = \mathsf{E}\left[\mathsf{E}\left[\|\boldsymbol{h}-\boldsymbol{T}(\boldsymbol{S}\boldsymbol{h}+\boldsymbol{n})\|^2\right] \mid \boldsymbol{h}\right] \tag{8.43}$$

$$= \mathsf{E}\left[\mathsf{E}\left[\left\|\underbrace{(\boldsymbol{I}_{KM}-\boldsymbol{T}\boldsymbol{S})\boldsymbol{h}}_{\text{Bias}}+\boldsymbol{T}\boldsymbol{n}\right\|^2\right] \mid \boldsymbol{h}\right] \tag{8.44}$$

$$= \underbrace{\mathsf{E}\left[\underbrace{\|(\boldsymbol{I}_{KM}-\boldsymbol{T}\boldsymbol{S})\boldsymbol{h}\|^2 \mid \boldsymbol{h}}_{\text{Squared Bias}}\right]}_{\text{Average Squared Bias}} + \underbrace{\mathsf{E}\left[\underbrace{\mathsf{E}\left[\|\boldsymbol{T}\boldsymbol{n}\|^2 \mid \boldsymbol{h}\right]}_{\text{Variance}}\right]}_{\text{Average Variance = Variance}} \tag{8.45}$$

$$= \mathsf{tr}\left[(\boldsymbol{I}_{KM}-\boldsymbol{T}\boldsymbol{S})\boldsymbol{C}_{\boldsymbol{h}}(\boldsymbol{I}_{KM}-\boldsymbol{T}\boldsymbol{S})^{\mathsf{H}}\right] + \mathsf{tr}\left[\boldsymbol{T}\boldsymbol{C}_{\boldsymbol{n}}\boldsymbol{T}^{\mathsf{H}}\right]. \tag{8.46}$$

[4]The last step is found by using the identity $\mathsf{E}\left[\|\boldsymbol{x}\|^2\right] = \mathsf{E}\left[\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}\right] = \mathsf{E}\left[\mathsf{tr}\left[\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\right]\right] = \mathsf{tr}\left[\mathsf{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\right]\right] = \mathsf{tr}\left[\boldsymbol{R}_{\boldsymbol{x}}\right]$ and $\boldsymbol{R}_{\boldsymbol{x}} = \boldsymbol{C}_{\boldsymbol{x}}$ if $\boldsymbol{x}$ is zero mean.

For the simplified assumptions $\boldsymbol{S}^{\mathsf{H}}\boldsymbol{S} = N\sigma_s^2\boldsymbol{I}_{KM}$ and $\boldsymbol{C_n} = \sigma_\eta^2\boldsymbol{I}_{NM}$:[5]

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

The minimum MSE achieved by the MMSE Estimator is given by

$$\mathsf{E}\left[\|\boldsymbol{h}-\boldsymbol{T}(\boldsymbol{S}\boldsymbol{h}+\boldsymbol{n})\|^2\right] = \sum_{i=1}^{KM}\frac{\sigma_\eta^2}{N\sigma_s^2}\lambda_i\left(\lambda_i+\frac{\sigma_\eta^2}{N\sigma_s^2}\right)^{-1}. \tag{8.47}$$

[5]For this analysis the MF estimator is assumed to be scaled as $\boldsymbol{T}_{\mathsf{MF}} = \frac{1}{\lambda_1}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{T}_{\mathsf{ML}}$.

8.8 Numerical Experiments

We consider

- a NON-DISPERSIVE MIMO channel $\boldsymbol{h} = \text{vec}[\boldsymbol{H}]$,
- with $M = K = 8$ antenna elements,
- a UNIFORM LINEAR ARRAYS at the receiver,
- and $N = 16$ training signal vectors $\boldsymbol{s}_n \in \{-1, +1\}^K$ (BINARY).
- The covariance matrix $\boldsymbol{C}_{\boldsymbol{h}}$ corresponds to 8 impinging planar wavefronts at

$$-45^\circ, -32.1^\circ, -19.3^\circ, -6.4^\circ, 6.4^\circ, 19.3^\circ, 32.1^\circ, 45^\circ, \tag{8.48}$$

  where each AZIMUTH ANGLE represents a cluster of wavefronts spread according to a LAPLACE ANGULAR POWER SPECTRUM with spread $\sigma = 5^\circ$

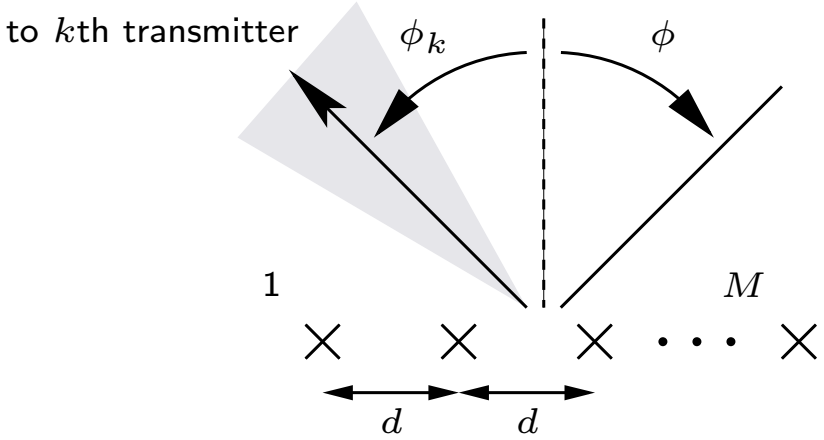


Fig. 8.3: Array geometry.

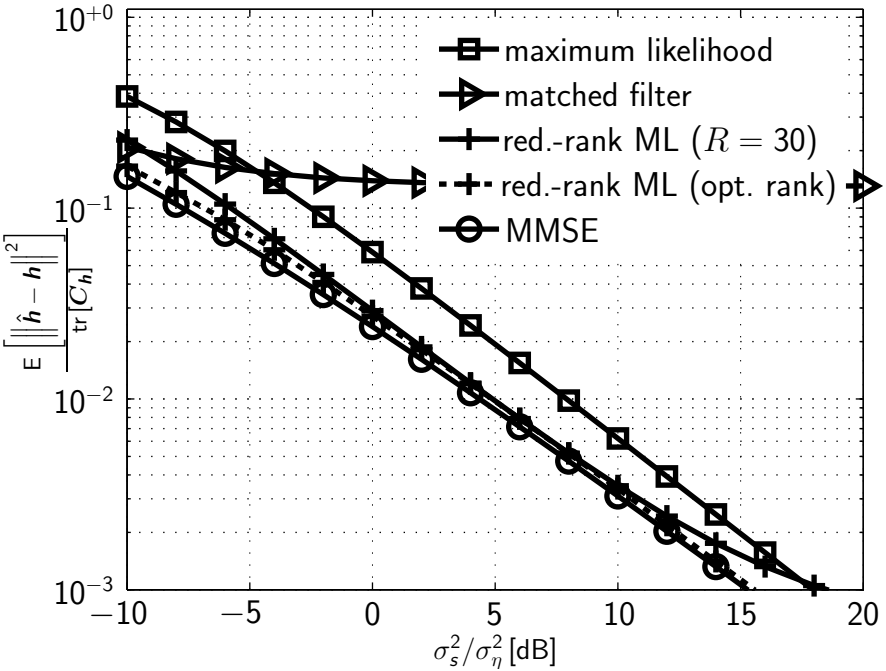


Fig. 8.4: Normalized MSE of channel estimators for $M = 8$, $K = 8$, and $N = 16$.

8.9 Numerical Experiments (cont'd)

We now consider

- a Dispersive Single-Input Multiple-Output (SIMO) channel $\boldsymbol{h} = \text{vec}[\boldsymbol{H}]$,
- with $M = 8$ receiver antenna elements, $K = 1$ transmitter antenna elements,
- a Uniform Linear Arrays at the receiver,
- with dispersion order $L = 4$,
- and $N = 16$ training signal vectors $\boldsymbol{s}_n \in \{-1, +1\}^K$ (Binary).
- The covariance matrix $\boldsymbol{C_h}$ corresponds to 5 impinging planar wavefronts at

$$-45^\circ, -22.5^\circ, 0^\circ, 22.5^\circ, 45^\circ \tag{8.49}$$

per delay, where each Azimuth Angle represents a cluster of wavefronts spread according to a Laplace Angular Power Spectrum with spread $\sigma = 5^\circ$

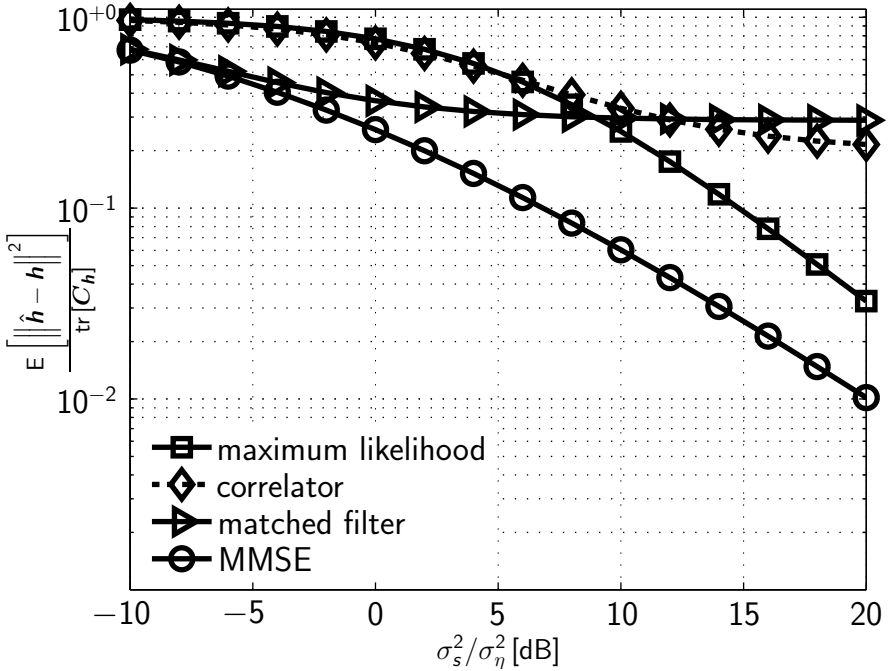


Fig. 8.5: Normalized MSE of channel estimators for $M = 8$, $K = 1$, $L = 4$ (DISPERSIVE), and $N = 16$.

References

- S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume I, Prentice Hall Signal Processing Series, 1993.
- L. L. Scharf. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, 1st edition, Prentice Hall, 1991.
- F. A. Dietrich and W. Utschick. *Pilot assisted channel estimation based on second order statistics*. IEEE Transactions on Signal Processing, 53(3):1178–1193, 2005.

Part VI

Estimation of Random Sequences

9. Random Sequences

In the following, we study SEQUENCES OF RANDOM VARIABLES $X_n : \Omega \to \mathbb{X}_n$, $n \in \mathbb{N}$, with a probability space $(\Omega, \mathbb{F}, \mathsf{P})$ and a measurement space (STATE SPACE) $\mathbb{X} \supseteq \mathbb{X}_n$, which are uniquely defined by the distribution functions

$$F_{X_{i_1},\ldots,X_{i_n}}(x_{i_1},\ldots,x_{i_n}) = \mathsf{P}\left(X_{i_1} \leq x_{i_1},\ldots,X_{i_n} \leq x_{i_n}\right), \tag{9.1}$$

for all $i_1,\ldots,i_n \in \mathbb{N}$ and for all $n \in \mathbb{N}$.

Given the existence of probability density functions of the corresponding distribution functions, the random sequences can be defined alternatively by means of

$$f_{X_{i_1},\ldots,X_{i_n}}(x_{i_1},\ldots,x_{i_n}), \tag{9.2}$$

for all $i_1,\ldots,i_n \in \mathbb{N}$ and for all $n \in \mathbb{N}$.

Note. The definitions of the mean, variance, auto-correlation, and cross-correlation function of random sequences directly follow from the definitions of single random variables.

9.1 Conditional Stochastical Independence

For the subsequent definition of MARKOV SEQUENCES, we introduce the CONDITIONAL STOCHASTICAL INDEPENDENCE of random variables.

Definition. Given Y, the random variables X and Z are conditionally stochastically independent, if

$$f_{Z|Y,X}(z \mid y, x) = f_{Z|Y}(z \mid y), \tag{9.3}$$

and $f_{Y,X}(y, x) > 0$.

Alternatively, following the definition of stochastical independence, the CONDITIONAL STOCHASTICAL INDEPENDENCE of Z and X given Y, we obtain

$$f_{Z,X|Y}(z, x \mid y) = f_{Z|Y}(z \mid y) f_{X|Y}(x \mid y), \tag{9.4}$$

respectively. As a short notation for conditional stochastical independence, we use

$$X \to Y \to Z. \tag{9.5}$$

9.2 Markov Sequences

Definition. A MARKOV SEQUENCE (PROCESS) is a random sequence $X_n : \Omega \to \mathbb{X}_n$, $n \in \mathbb{N}$, with a probability space $(\Omega, \mathbb{F}, \mathsf{P})$ and a state space $\mathbb{X} \supseteq \mathbb{X}_n$, where for any ASCENDING SEQUENCE of indices $n_i \in \mathbb{N}$ and $k > 2$ the following CONDITIONAL STOCHASTICAL INDEPENDENCE holds, namely

$$(X_{n_1}, \ldots, X_{n_{k-2}}) \to X_{n_{k-1}} \to X_{n_k}, \tag{9.6}$$

which means

$$f_{X_{n_k}|X_{n_{k-1}},\ldots,X_{n_1}}(x_{n_k}|x_{n_{k-1}}, \ldots, x_{n_1}) = f_{X_{n_k}|X_{n_{k-1}}}(x_{n_k}|x_{n_{k-1}}), \tag{9.7}$$

and $f_{X_{n_{k-1}},\ldots,X_{n_1}}(x_{n_{k-1}}, \ldots, x_{n_1}) > 0$.[1]

Definition. The conditional $f_{X_{n_k}|X_{n_{k-1}}}(x_{n_k}|x_{n_{k-1}})$ defines the STATE-TRANSITION DENSITY.

[1] In this lecture, we consider so-called FIRST-ORDER MARKOV SEQUENCES with $f_{X_{n_k}|X_{n_{k-1}},\ldots,X_{n_1}} = f_{X_{n_k}|X_{n_{k-1}}}$. If the underlying conditional independence holds only if we condition on more than one previous state, we have HIGHER-ORDER MARKOV SEQUENCES. If, for example, $f_{X_{n_k}|X_{n_{k-1}},\ldots,X_{n_1}} = f_{X_{n_k}|X_{n_{k-1}},X_{n_{k-2}}}$, we have an order of two. The order of a MARKOV SEQUENCE can be interpreted as the memory of the stochastic sequence.

As a consequence of this MARKOV PROPERTY we obtain for the joint PDF of the first n elements

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= f_{X_n|X_{n-1},\ldots,X_1}(x_n|x_{n-1},\ldots,x_1) f_{X_{n-1}|X_{n-2},\ldots,X_1}(x_{n-1}|x_{n-2},\ldots,x_1)\ldots f_{X_1}(x_1) \\ &= f_{X_n|X_{n-1}}(x_n|x_{n-1}) f_{X_{n-1}|X_{n-2}}(x_{n-1}|x_{n-2})\ldots f_{X_1}(x_1) \\ &= f_{X_1}(x_1)\prod_{i=2}^{n} f_{X_i|X_{i-1}}(x_i|x_{i-1}), \end{aligned} \tag{9.8}$$

where

- the first line follows from the MULTIPLICATION RULE[2]
- and the second line follows from the STOCHASTICAL INDEPENDENCE between X_n and X_{n-2} given an event related to X_{n-1}.

[2] $\mathsf{P}(A \cap B \cap C) = \mathsf{P}(A|B \cap C)\,\mathsf{P}(B|C)\,\mathsf{P}(C)$.

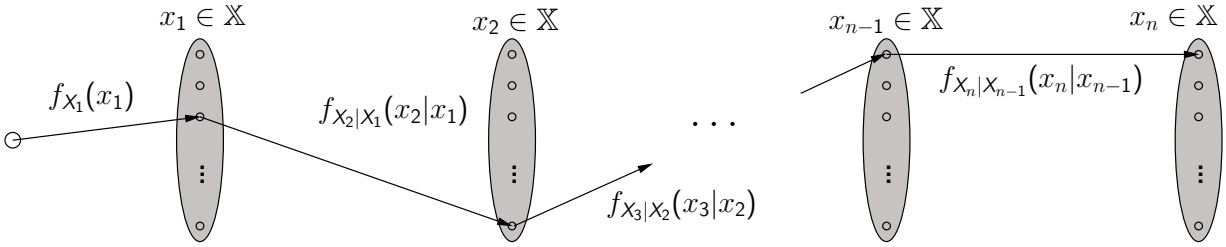


Fig. 9.1: The joint PDF $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$ obviously results from a chain STATE-TRANSITION DENSITIES starting with the PDF of the random variable X_1.

9.3 Chapman-Kolmogorov Equation

Given a Markov sequence $X_n : \Omega \to \mathbb{X}_n$ and $f_{X_n}(x_n) > 0$ for all $n \in \mathbb{N}$, the $(m+\ell)$-step state-transition density starting from X_n depends on the m-step (from X_n) and ℓ-step (from X_{n+m}) state-transition density according to the Chapman-Kolmogorov Equation

$$f_{X_{n+m+\ell}|X_n}(x_{n+m+\ell}|x_n) = \int_{x_{n+m} \in \mathbb{X}_n} f_{X_{n+m+\ell}|X_{n+m}}(x_{n+m+\ell}|x_{n+m})\ f_{X_{n+m}|X_n}(x_{n+m}|x_n)\, \mathrm{d}x_{n+m}. \tag{9.9}$$

The Chapman-Kolmogorov Equations (9.9) can be derived

- by marginalization of $f_{X_{n+m+\ell},X_{n+m}|X_n}(x_{n+m+\ell}, x_{n+m}|x_n)$
- and further taking into account the Markov condition of X_n, i.e.,

$$f_{X_{n+m+\ell}|X_{n+m},X_n}(x_{n+m+\ell}|x_{n+m}, x_n) = f_{X_{n+m+\ell}|X_{n+m}}(x_{n+m+\ell}|x_{n+m}). \tag{9.10}$$

Example.

We consider the 2-step state-transition density based on two 1-step state-transition densities, i.e.,

$$
\begin{aligned}
f_{X_{n+2}|X_n}(x_{n+2}|x_n) &= \underbrace{\int_{-\infty}^{\infty} f_{X_{n+2},X_{n+1}|X_n}(x_{n+2},\xi|x_n)\,\mathrm{d}\xi}_{\text{Marginalization}} \\
&= \int_{-\infty}^{\infty} \underbrace{f_{X_{n+2}|X_{n+1},X_n}(x_{n+2}|\xi,x_n) f_{X_{n+1}|X_n}(\xi|x_n)}_{\text{Conditional PDF} \times \text{Prior PDF}}\,\mathrm{d}\xi \\
&= \int_{-\infty}^{\infty} \underbrace{f_{X_{n+2}|X_{n+1}}(x_{n+2}|\xi)}_{\text{Markov Property}} f_{X_{n+1}|X_n}(\xi|x_n)\,\mathrm{d}\xi.
\end{aligned}
$$

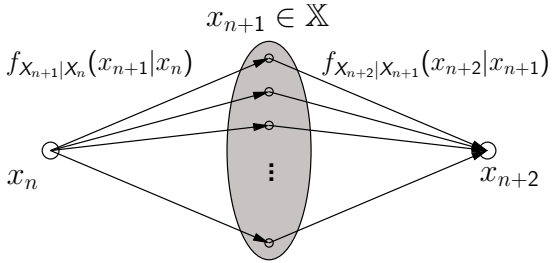


Fig. 9.2: State-Transitions from $\{X_n = x_n\}$ to $\{X_{n+2} = x_{n+2}\}$.

The General Case.

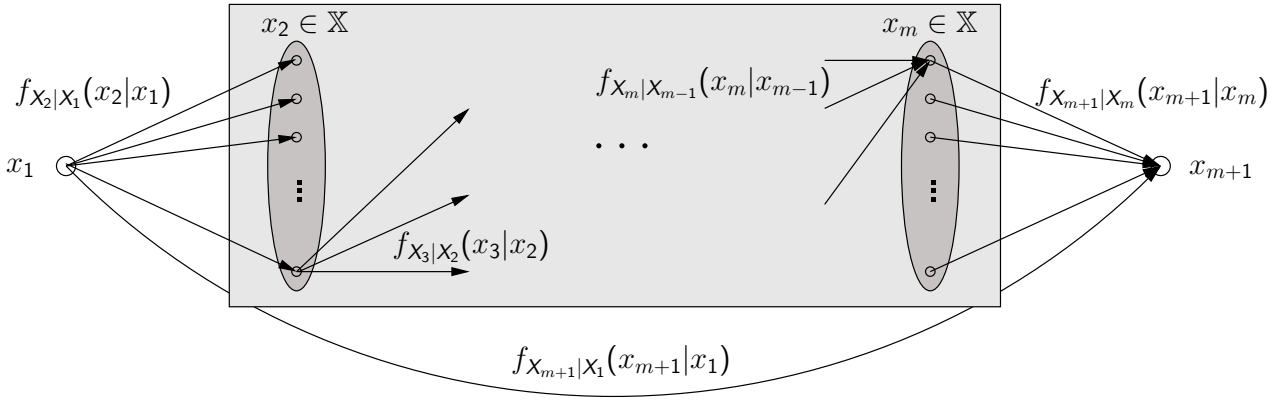


Fig. 9.3: Illustration of (9.9) (m-step state-transition density) for $n = 1$.

9.4 Estimation of Markov Sequences

A Hidden Markov Process, is a Markov Process which states cannot be observed directly, i.e., the states are *hidden*. The state variables $\boldsymbol{X}_n$ are observed indirectly via the random observations $\boldsymbol{Y}_n$ drawn from the conditional PDF $f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n|\boldsymbol{x}_n)$, which is the likelihood of $\boldsymbol{Y}_n$ given that $\boldsymbol{X}_n = \boldsymbol{x}_n$.

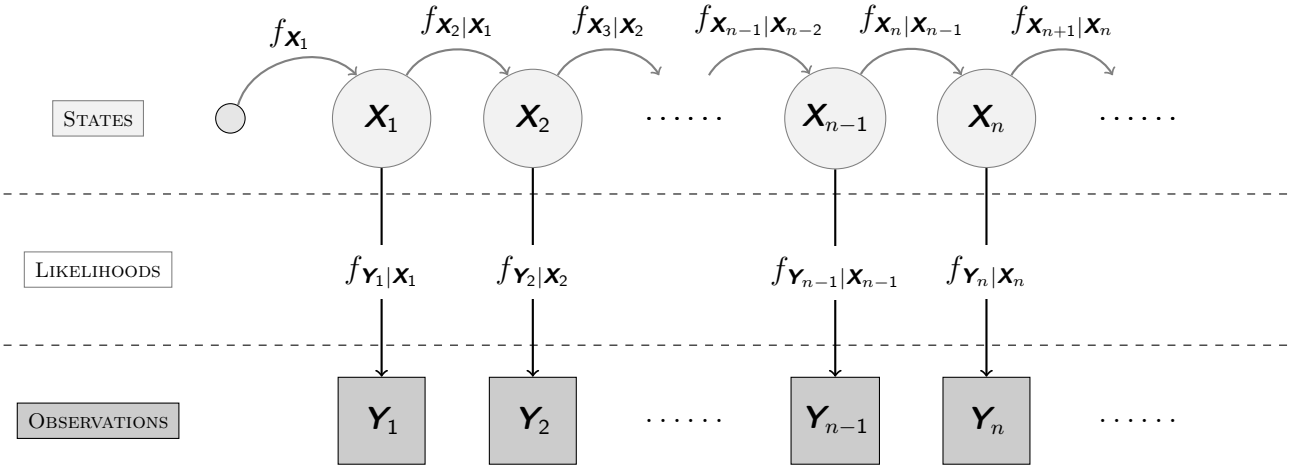


Fig. 9.4: Illustration of a Hidden Markov Process.

The CHAPMAN-KOLMOGOROV EQUATION is the key element of

> *A recursive computation of the conditional PDFs of a Markov sequence $\boldsymbol{X}_n$ conditioned on observations of $\boldsymbol{Y}_{(n)}$ thereof, this is $\boldsymbol{Y}_n$ and $\boldsymbol{Y}_{n-1}, \ldots, \boldsymbol{Y}_1$.*[a]

[a] In the following, the short notation $\boldsymbol{Y}_{(n)}$ will be used instead of $\boldsymbol{Y}_n, \ldots, \boldsymbol{Y}_1$. The same holds accordingly for $\boldsymbol{Y}_{n-1}, \ldots, \boldsymbol{Y}_1$, etc.

The recursive updates can be obtained in two steps:

a) CHAPMANN-KOLMOGOROV EQUATION and the MARKOV CONDITION,

$$f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n-1)}) =$$

$$= \int_{\mathbb{X}} f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{x}_{n-1},\boldsymbol{y}_{(n-1)})\ f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)})\,\mathrm{d}\boldsymbol{x}_{n-1}$$

$$= \int_{\mathbb{X}} f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})\ f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)})\,\mathrm{d}\boldsymbol{x}_{n-1}. \tag{9.11}$$

b) Bayes Rule

$$f_{\boldsymbol{X}_n|\boldsymbol{Y}_n,\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{y}_n,\boldsymbol{y}_{(n-1)}) =$$

$$= \frac{f_{\boldsymbol{Y}_n|\boldsymbol{X}_n,\boldsymbol{Y}_{(n-1)}}(\boldsymbol{y}_n|\boldsymbol{x}_n,\boldsymbol{y}_{(n-1)})\ f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n-1)})}{f_{\boldsymbol{Y}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{y}_n|\boldsymbol{y}_{(n-1)})}$$

$$= \frac{f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n|\boldsymbol{x}_n)\ f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n-1)})}{f_{\boldsymbol{Y}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{y}_n|\boldsymbol{y}_{(n-1)})}. \tag{9.12}$$

Both steps together are equivalent (except some scaling constant) to[3]

$$f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)}) \propto$$
$$f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n|\boldsymbol{x}_n) \int_{\mathbb{X}} f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})\ f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)})\,\mathrm{d}\boldsymbol{x}_{n-1}. \tag{9.13}$$

[3]The denominator of Eq. (9.12) does not depend on the random variable $\boldsymbol{X}_n$ to be estimated and is implicitly given as the normalization constant $f_{\boldsymbol{Y}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{y}_n|\boldsymbol{y}_{(n-1)}) = \int_{\mathbb{X}} f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n|\boldsymbol{x}_n)\ f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n-1)})\,\mathrm{d}\boldsymbol{x}_n$.

Note.

In essence, there are two conditional PDFs and one state-transition PDF which form the basis for the recursive computation of the conditional PDFs of the state variables.

$$\overbrace{f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)})}^{\text{desired conditional PDF in the } n\text{th step}} \propto \underbrace{f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n|\boldsymbol{x}_n)}_{\text{likelihood of the new observation}} \times \int_{\mathbb{X}} \underbrace{f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})}_{\text{state-transition PDF}} \underbrace{f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)})\,\mathrm{d}\boldsymbol{x}_{n-1}}_{\text{conditional PDF from the } (n-1)\text{th-step before}} . \tag{9.14}$$

The principal idea of a variety of estimation methods for Tracking Of State Variables *is to exploit the recursive computation of the conditional PDFs for the recursive estimation of the characteristical parameters of the distributions, cf. Chapter 10 and 12.*

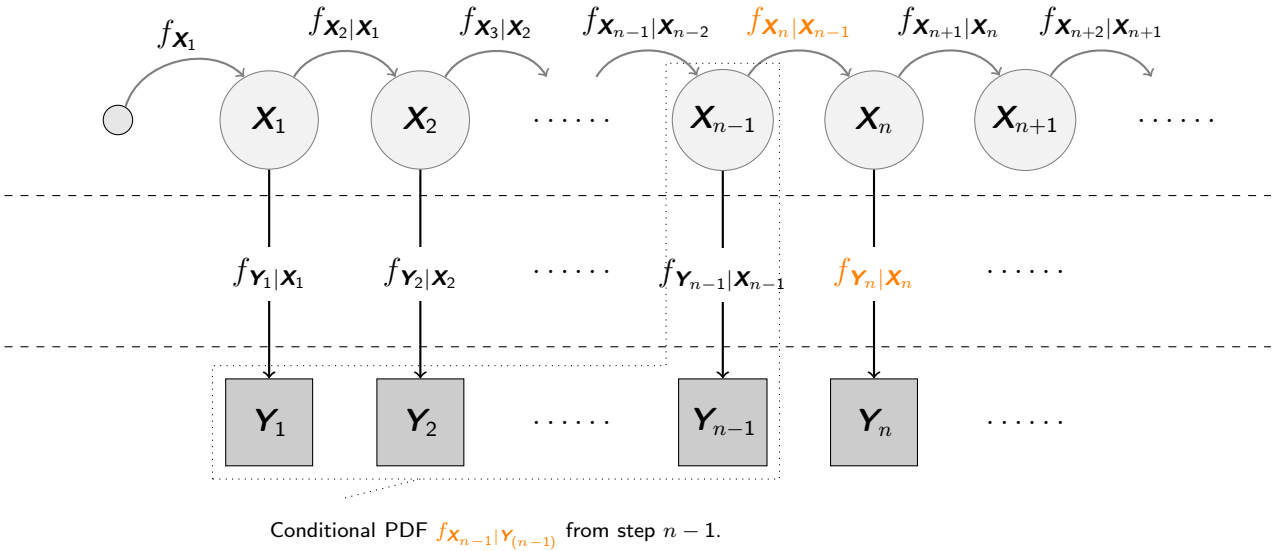


Fig. 9.5: In order to determine the cond. PDF $f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}$, we need the cond. PDF $f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}$ of the previous step, the likelihood $f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}$, and the state-transition PDF $f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}$, cf. Eq. (9.13).

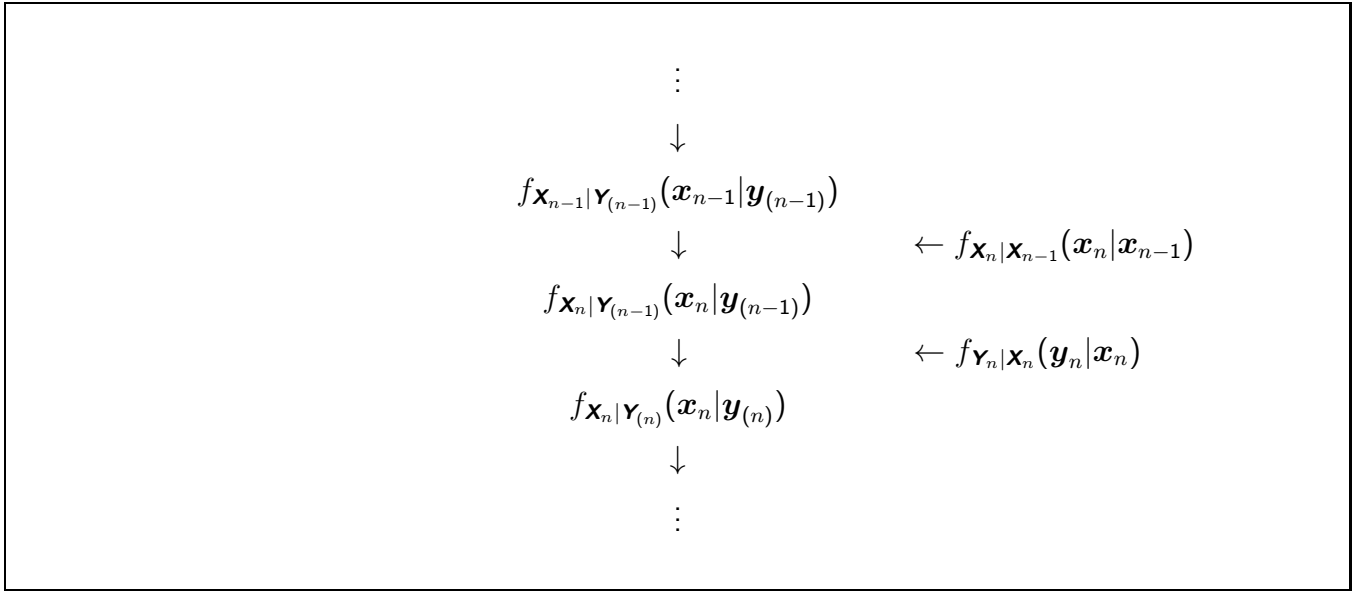


Fig. 9.6: The recursive structure of the computation of the conditional PDFs.

9.5 Gauß-Markov Sequences

Definition. GAUSS-MARKOV SEQUENCES are Gaussian sequences which fulfill the Markov condition, i.., any ensemble of random variables of the sequence is jointly Gaussian distributed and the state-transition probability density functions[4] are fully described by conditional means and covariances and the observation of the last recent random variable of the sequence. Since for jointly Gaussian variables the conditional mean of a random variable is a linear function of the others, $\mu_{n|n-1} = \mathsf{E}[X_n|X_{n-1} = x_{n-1}] = \rho_n x_{n-1}$, we obtain

$$X_0 \sim \mathcal{N}(0, \sigma_0^2) \tag{9.15}$$
$$X_1 \sim \mathcal{N}(\mu_{X_1|X_0=x_0}, \sigma^2_{X_1|X_0=x_0}), \tag{9.16}$$
$$\vdots \tag{9.17}$$
$$X_n \sim \mathcal{N}(\mu_{X_n|X_{n-1}=x_{n-1}}, \sigma^2_{X_n|X_{n-1}=x_{n-1}}). \tag{9.18}$$

Alternatively,[5] the GAUSS-MARKOV SEQUENCE $X_{n-2} \to X_{n-1} \to X_n$ is fully described by

$$X_n = \rho_n X_{n-1} + W_n, \tag{9.19}$$

with $\rho_n = c_{X_n,X_{n-1}}\sigma^{-2}_{X_{n-1}}$ and $W_n \sim \mathcal{N}(0, \sigma^2_{X_n|X_{n-1}=x_{n-1}})$.

[4] W.l.o.g. the further considerations are presented for the case of continuous random variables.

[5] W.l.o.g. we assume zero mean random variables for the unconditioned state variables X_n, i.e., $\mathsf{E}[X_n] = 0$, $\forall n \in \mathbb{N} \cup \{0\}$.

10. Kalman Filter

In the following, we study the problem of estimating the stochastic path of a Gauß-Markov sequence $\boldsymbol{X}_n$ based on linear observations $\boldsymbol{Y}_n$, $\boldsymbol{Y}_{n-1}$, ... of the random state variables,[1]

$$\boldsymbol{X}_n = \boldsymbol{G}_n \boldsymbol{X}_{n-1} + \boldsymbol{V}_n, \tag{10.1}$$

$$\boldsymbol{Y}_n = \boldsymbol{H}_n \boldsymbol{X}_n + \boldsymbol{W}_n, \tag{10.2}$$

with

$$\boldsymbol{V}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{V}_n}),$$
$$\boldsymbol{W}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{W}_n}).$$

In the context of KALMAN FILTERING, the random variables $\boldsymbol{V}_n$ and $\boldsymbol{W}_n$ are called PROCESS NOISE and MEASUREMENT NOISE. The random variables $\boldsymbol{W}_m$, $\boldsymbol{W}_n$, and $\boldsymbol{V}_k$, $\boldsymbol{V}_\ell$ are stochastically independent for different indices m, n and k, ℓ, and $\boldsymbol{V}_m$, $\boldsymbol{W}_n$ are stochastically independent for any m and n.

Note. Eqs. (10.1)-(10.2) are a direct consequence of Eq. (9.19). In the following, we assume that the states $\boldsymbol{X}_n$ are OBSERVABLE.

[1] For the sake of generality, we consider the case of a sequence of random vectors.

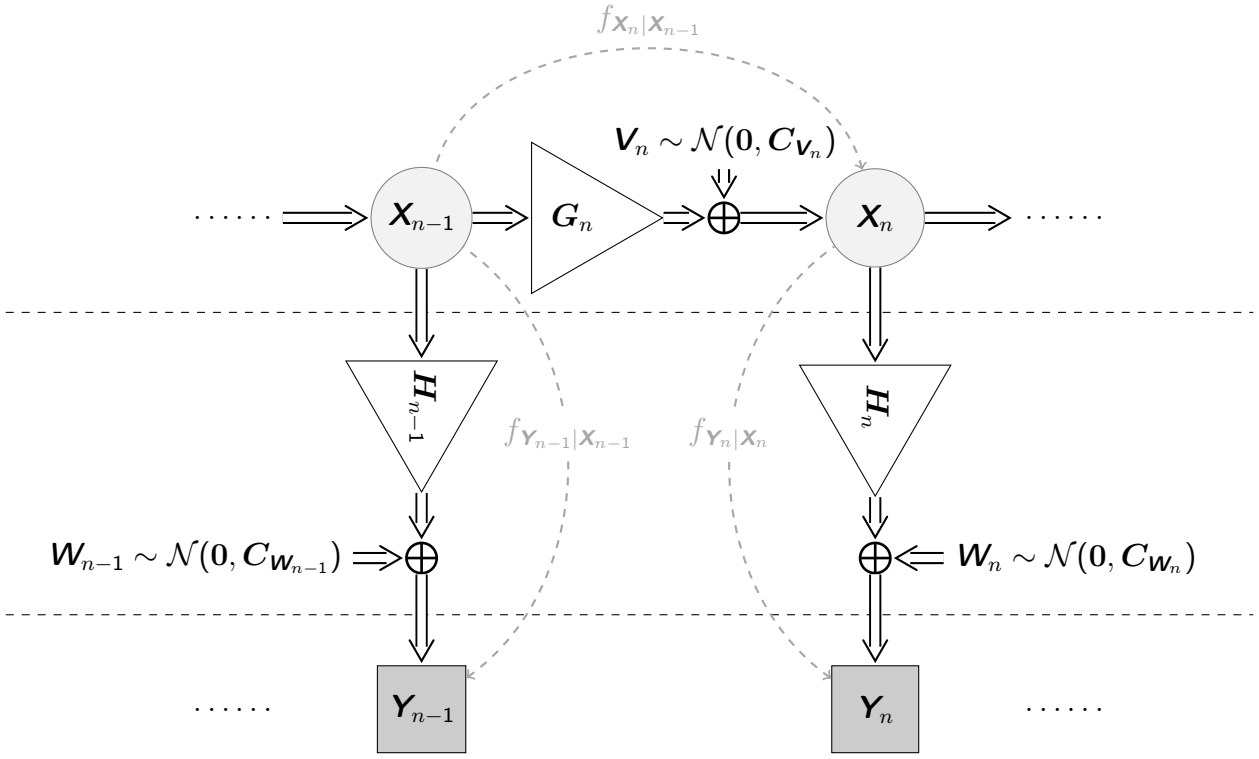


Fig. 10.1: In a HIDDEN GAUSS-MARKOV PROCESS, the states $\boldsymbol{X}_n$ are given as $\boldsymbol{X}_n = \boldsymbol{G}_n \boldsymbol{X}_{n-1} + \boldsymbol{V}_n$ with $\boldsymbol{V}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{V}_n})$ and the observations $\boldsymbol{Y}_n$ are given as $\boldsymbol{Y}_n = \boldsymbol{H}_n \boldsymbol{X}_n + \boldsymbol{W}_n$ with $\boldsymbol{W}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{W}_n})$. Thus, the conditional PDFs $f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}$, the likelihoods $f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}$, and the state-transition PDFs $f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}$ are Gaussian.

In general, the MMSE optimal estimate is obtained by the CONDITIONAL MEAN

$$\mathsf{E}\left[\boldsymbol{X}_n|\,\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}\right]=\int_{\mathbb{X}}\boldsymbol{x}_n f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)})\,\mathrm{d}\boldsymbol{x}_n, \tag{10.3}$$

which can by derived from the CONDITIONAL PDF

$$f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)}). \tag{10.4}$$

Recall that the CONDITIONAL PDF of any the state of the Gauß-Markov sequence is uniquely characterized by the first and second order conditional moments

$$\boldsymbol{\mu}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}}=\mathsf{E}\left[\boldsymbol{X}_n|\,\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}\right], \tag{10.5}$$

$$\boldsymbol{C}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}}=\mathsf{E}\left[\left(\boldsymbol{X}_n-\boldsymbol{\mu}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}}\right)\left(\boldsymbol{X}_n-\boldsymbol{\mu}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}}\right)^{\mathsf{T}}\middle|\,\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}\right]. \tag{10.6}$$

Note.
$\boldsymbol{C}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}}=\mathsf{E}\left[\left(\boldsymbol{X}_n-\boldsymbol{\mu}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}}\right)\left(\boldsymbol{X}_n-\boldsymbol{\mu}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}}\right)^{\mathsf{T}}\mid\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}\right]$ does not depend on the actual realization $\boldsymbol{y}_{(n)}$, since all random variables are Gaussian distributed.

The same holds for $\boldsymbol{C}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}=\boldsymbol{y}_{(n-1)}}$.

Short Notations.

- In the following, $\hat{\boldsymbol{x}}_{n|n}$ and $\hat{\boldsymbol{x}}_{n|n-1}$ are used as short notations for $\boldsymbol{\mu}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}}$ and $\boldsymbol{\mu}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}=\boldsymbol{y}_{(n-1)}}$, i.e.,

$$\begin{aligned}\hat{\boldsymbol{x}}_{n|n} &= \mathsf{E}\left[\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}\right],\\ \hat{\boldsymbol{x}}_{n|n-1} &= \mathsf{E}\left[\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}=\boldsymbol{y}_{(n-1)}\right].\end{aligned}$$

- The same holds for $\boldsymbol{C}_{\boldsymbol{X}_{n|n}}$ and $\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}$, which are short notations for

$$\begin{aligned}\boldsymbol{C}_{\boldsymbol{X}_{n|n}} &= \mathsf{E}\left[\left(\boldsymbol{X}_n-\hat{\boldsymbol{x}}_{n|n}\right)\left(\boldsymbol{X}_n-\hat{\boldsymbol{x}}_{n|n}\right)^{\mathsf{T}} \mid \boldsymbol{Y}_{(n)}=\boldsymbol{y}_{(n)}\right],\\ \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} &= \mathsf{E}\left[\left(\boldsymbol{X}_n-\hat{\boldsymbol{x}}_{n|n-1}\right)\left(\boldsymbol{X}_n-\hat{\boldsymbol{x}}_{n|n-1}\right)^{\mathsf{T}} \mid \boldsymbol{Y}_{(n-1)}=\boldsymbol{y}_{(n-1)}\right].\end{aligned}$$

10.1 Recursive Computation of Conditional Means and Covariances

a) Exploiting the Chapman-Kolmogorov Equation (9.11) and the Linear State Space Model in Eq. (10.1), we obtain

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}=\boldsymbol{y}_{(n-1)}} &= \int_{\mathbb{X}} \underbrace{\int_{\mathbb{X}} \boldsymbol{x}_n \; f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})\,\mathrm{d}\boldsymbol{x}_n}_{\text{cond. expectation of } \boldsymbol{x}_n \text{ (refer to 10.1)}} \; f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)})\,\mathrm{d}\boldsymbol{x}_{n-1} \\ &= \boldsymbol{G}_n \; \boldsymbol{\mu}_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}=\boldsymbol{y}_{(n-1)}},\end{aligned}$$

and thus the State Prediction is given by

$$\hat{\boldsymbol{x}}_{n|n-1} = \boldsymbol{G}_n \hat{\boldsymbol{x}}_{n-1|n-1}. \tag{10.7}$$

The respective Conditional State Covariance Matrix of $\boldsymbol{X}_n$ is obtained as

$$\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} = \boldsymbol{G}_n \boldsymbol{C}_{\boldsymbol{X}_{n-1|n-1}} \boldsymbol{G}_n^{\mathsf{T}} + \boldsymbol{C}_{\boldsymbol{V}_n}. \tag{10.8}$$

b) Since both sides of the BAYES RULE in (9.12) are Gaussian PDFs, a comparison of the exponents,[2]

$$\log\left(f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)})\right) = \log\left(f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n|\boldsymbol{x}_n)\right) + \log\left(f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n-1)})\right) + \ldots,$$

and taking into account the LINEAR OBSERVATION MODEL in Eq. (10.2) results in

$$\hat{\boldsymbol{x}}_{n|n} = \hat{\boldsymbol{x}}_{n|n-1} + \boldsymbol{K}_n\left(\boldsymbol{y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1}\right), \tag{10.9}$$

$$\boldsymbol{C}_{\boldsymbol{X}_{n|n}} = \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} - \boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}, \tag{10.10}$$

with the KALMAN GAIN MATRIX

$$\boldsymbol{K}_n = \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}}\left(\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}} + \boldsymbol{C}_{\boldsymbol{W}_n}\right)^{-1}. \tag{10.11}$$

[2]Note, that we are using (9.12) instead of (9.13).

$$\vdots$$

$$\hat{\boldsymbol{x}}_{n-1|n-1}$$
$$\boldsymbol{C}_{\boldsymbol{X}_{n-1|n-1}}$$

$$\downarrow \qquad \leftarrow \ \boldsymbol{G}_n, \boldsymbol{H}_n, \boldsymbol{C}_{\boldsymbol{V}_n}, \boldsymbol{C}_{\boldsymbol{W}_n}$$

$$\begin{aligned}
\hat{\boldsymbol{x}}_{n|n-1} &= \boldsymbol{G}_n \hat{\boldsymbol{x}}_{n-1|n-1} \\
\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} &= \boldsymbol{G}_n \boldsymbol{C}_{\boldsymbol{X}_{n-1|n-1}} \boldsymbol{G}_n^{\mathsf{T}} + \boldsymbol{C}_{\boldsymbol{V}_n} \\
\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}} &= \boldsymbol{H}_n \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} \boldsymbol{H}_n^{\mathsf{T}} + \boldsymbol{C}_{\boldsymbol{W}_n} \\
\boldsymbol{K}_n &= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} \boldsymbol{H}_n^{\mathsf{T}} \boldsymbol{C}^{-1}_{\boldsymbol{Y}_{n|n-1}}
\end{aligned}$$

$$\downarrow \qquad \leftarrow \ \text{observation } \boldsymbol{y}_n \text{ drawn from } \boldsymbol{Y}_n$$

$$\begin{aligned}
\hat{\boldsymbol{x}}_{n|n} &= \hat{\boldsymbol{x}}_{n|n-1} + \boldsymbol{K}_n \left(\boldsymbol{y}_n - \boldsymbol{H}_n \hat{\boldsymbol{x}}_{n|n-1} \right) \\
\boldsymbol{C}_{\boldsymbol{X}_{n|n}} &= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} - \boldsymbol{K}_n \boldsymbol{H}_n \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}
\end{aligned}$$

$$\vdots$$

Fig. 10.2: The recursive STATE TRACKING procedure. In many applications, the system parameter matrices $\boldsymbol{G}_n, \boldsymbol{H}_n, \boldsymbol{C}_{\boldsymbol{V}_n}, \boldsymbol{C}_{\boldsymbol{W}_n}$ are assumed to be constant and we instead consider $\boldsymbol{G}, \boldsymbol{H}$ and $\boldsymbol{C}_{\boldsymbol{V}}, \boldsymbol{C}_{\boldsymbol{W}}$.

10.2 Proof

A proof for a scalar version is presented. See the appendix for the multivariate version.

a) Prediction step:

$$
\begin{aligned}
\boxed{\hat{x}_{n|n-1}} &= \mathsf{E}\left[X_n \mid Y_{(n-1)} = y_{(n-1)}\right] \\
&= \int_{\mathbb{X}} x_n f_{X_n|Y_{(n-1)}}(x_n \mid y_{(n-1)})\,\mathrm{d}x_n \\
&= \int_{\mathbb{X}}\int_{\mathbb{X}} x_n\ f_{X_n|X_{n-1}}(x_n|x_{n-1})\,\mathrm{d}x_n\ f_{X_{n-1}|Y_{(n-1)}}(x_{n-1}|y_{(n-1)})\,\mathrm{d}x_{n-1} \\
&= \int_{\mathbb{X}} \mathsf{E}\left[X_n \mid X_{n-1} = x_{n-1}\right]\ f_{X_{n-1}|Y_{(n-1)}}(x_{n-1}|y_{(n-1)})\,\mathrm{d}x_{n-1} \\
&= \int_{\mathbb{X}} \mathsf{E}\left[g_n X_{n-1} + V_n \mid X_{n-1} = x_{n-1}\right]\ f_{X_{n-1}|Y_{(n-1)}}(x_{n-1}|y_{(n-1)})\,\mathrm{d}x_{n-1} \\
&= g_n \int_{\mathbb{X}} x_{n-1}\ f_{X_{n-1}|Y_{(n-1)}}(x_{n-1}|y_{(n-1)})\,\mathrm{d}x_{n-1} = g_n\,\mathsf{E}\left[X_{n-1} \mid Y_{(n-1)} = y_{(n-1)}\right] = g_n\,\boxed{\hat{x}_{n-1|n-1}},
\end{aligned}
$$

$$
\begin{aligned}
\boxed{\sigma^2_{X_{n|n-1}}} &= \mathsf{E}\left[(X_n - \hat{x}_{n|n-1})^2 \mid Y_{(n-1)} = y_{(n-1)}\right] \\
&= \mathsf{E}\left[(g_n X_{n-1} + V_n - g_n\hat{x}_{n-1|n-1})^2 \mid Y_{(n-1)} = y_{(n-1)}\right] \\
&= g_n^2\,\mathsf{E}\left[(X_{n-1} - \hat{x}_{n-1|n-1})^2 \mid Y_{(n-1)} = y_{(n-1)}\right] + \mathsf{E}\left[V_n^2 \mid Y_{(n-1)} = y_{(n-1)}\right] = g_n^2\boxed{\sigma^2_{X_{n-1|n-1}}} + \sigma^2_{V_n}.
\end{aligned}
$$

b) Correction step:

$$\log f_{X_n|Y_{(n)}}(x_n|y_{(n)}) = \log f_{Y_n|X_n}(y_n \mid x_n) + \log f_{X_n|Y_{(n-1)}}(x_n|y_{(n-1)}) + \dots$$

$$\begin{aligned}
&= -\frac{(y_n - \mathsf{E}\left[Y_n \mid X_n = x_n\right])^2}{2\,\mathsf{Var}\left[Y_n \mid X_n = x_n\right]} - \frac{(x_n - \hat{x}_{n|n-1})^2}{2\sigma^2_{X_{n|n-1}}} + \dots \\
&= -\frac{(y_n - h_n x_n)^2}{2\sigma^2_{W_n}} - \frac{(x_n - \hat{x}_{n|n-1})^2}{2\sigma^2_{X_{n|n-1}}} + \dots \\
&= -\frac{\sigma^2_{X_{n|n-1}}(y_n^2 - 2y_n h_n x_n + h_n^2 x_n^2) + \sigma^2_{W_n}(x_n^2 - 2x_n\hat{x}_{n|n-1} + \hat{x}^2_{n|n-1})}{2\sigma^2_{W_n}\sigma^2_{X_{n|n-1}}} + \dots \\
&= -\frac{(\sigma^2_{X_{n|n-1}}h_n^2 + \sigma^2_{W_n})x_n^2 - 2(\sigma^2_{X_{n|n-1}}y_n h_n + \sigma^2_{W_n}\hat{x}_{n|n-1})x_n}{2\sigma^2_{W_n}\sigma^2_{X_{n|n-1}}} + \dots \\
&= -\frac{\sigma^2_{X_{n|n-1}}h_n^2 + \sigma^2_{W_n}}{2\sigma^2_{W_n}\sigma^2_{X_{n|n-1}}}\left(x_n - \frac{\sigma^2_{X_{n|n-1}}y_n h_n + \sigma^2_{W_n}\hat{x}_{n|n-1}}{\sigma^2_{X_{n|n-1}}h_n^2 + \sigma^2_{W_n}}\right)^2 + \dots \\
&= -\frac{(x_n - \boxed{\hat{x}_{n|n}})^2}{2\boxed{\sigma^2_{X_{n|n}}}} + \dots \ .
\end{aligned}$$

Comparison:

$$\begin{aligned}
\boxed{\hat{x}_{n|n}} &= \frac{\sigma^2_{X_{n|n-1}} y_n h_n + \sigma^2_{W_n} \hat{x}_{n|n-1}}{\sigma^2_{X_{n|n-1}} h_n^2 + \sigma^2_{W_n}} \\
&= \hat{x}_{n|n-1} + \frac{\sigma^2_{X_{n|n-1}} y_n h_n + \sigma^2_{W_n} \hat{x}_{n|n-1} - (\sigma^2_{X_{n|n-1}} h_n^2 + \sigma^2_{W_n}) \hat{x}_{n|n-1}}{\sigma^2_{X_{n|n-1}} h_n^2 + \sigma^2_{W_n}} \\
&= \hat{x}_{n|n-1} + \frac{\sigma^2_{X_{n|n-1}} y_n h_n - \sigma^2_{X_{n|n-1}} h_n^2 \hat{x}_{n|n-1}}{\sigma^2_{X_{n|n-1}} h_n^2 + \sigma^2_{W_n}} \\
&= \hat{x}_{n|n-1} + \frac{\sigma^2_{X_{n|n-1}} h_n}{\sigma^2_{X_{n|n-1}} h_n^2 + \sigma^2_{W_n}} (y_n - h_n \hat{x}_{n|n-1}) \\
&= \boxed{\hat{x}_{n|n-1}} + \boxed{k_n} (y_n - h_n \boxed{\hat{x}_{n|n-1}}),
\end{aligned}$$

$$\boxed{\sigma^2_{X_{n|n}}} = \frac{\sigma^2_{W_n} \boxed{\sigma^2_{X_{n|n-1}}}}{\boxed{\sigma^2_{X_{n|n-1}}} h_n^2 + \sigma^2_{W_n}} = \frac{\sigma^2_{W_n} \boxed{\sigma^2_{X_{n|n-1}}}}{\boxed{\sigma^2_{Y_{n|n-1}}}},$$

with $\boxed{k_n} = \dfrac{\boxed{\sigma^2_{X_{n|n-1}}} h_n}{\boxed{\sigma^2_{X_{n|n-1}}} h_n^2 + \sigma^2_{W_n}} = \dfrac{\boxed{\sigma^2_{X_{n|n-1}}} h_n}{\boxed{\sigma^2_{Y_{n|n-1}}}}$ and $\boxed{\sigma^2_{Y_{n|n-1}}} = \boxed{\sigma^2_{X_{n|n-1}}} h_n^2 + \sigma^2_{W_n}$.

10.3 Discussion

Innovation Sequence.

The sequence of random variables $\boldsymbol{\Delta Y}_n = \boldsymbol{Y}_n - \hat{\boldsymbol{Y}}_{n|n-1}$ in Eq. (10.9) forms an i.i.d. random sequence which can be constructed by a bijective linear transform of the random observation sequence $\boldsymbol{Y}_n$,

$$\begin{aligned}
\boldsymbol{\Delta Y}_n &= \boldsymbol{Y}_n - \hat{\boldsymbol{Y}}_{n|n-1} \\
&= \boldsymbol{Y}_n - \mathsf{E}\left[\boldsymbol{Y}_n | \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \boldsymbol{Y}_n - \mathsf{E}\left[\boldsymbol{H}_n \boldsymbol{X}_n + \boldsymbol{W}_n | \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \boldsymbol{Y}_n - \boldsymbol{H}_n \underbrace{\mathsf{E}\left[\boldsymbol{X}_n | \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right]}_{\text{linear in } \boldsymbol{Y}_{(n-1)}} \\
&= \boldsymbol{Y}_n - \boldsymbol{H}_n \hat{\boldsymbol{x}}_{n|n-1}, \qquad (10.12)
\end{aligned}$$

and thus fulfills the requirements of an Innovation Sequence according to the observations $\boldsymbol{Y}_n$.

An Innovation Sequence consists of i.i.d. elements and is a bijective linear function of the observations.

Predictor-Corrector-Step.

Since the innovation variable is stochastically independent of previous observations of $\boldsymbol{Y}_{(n-1)}$, Eq. (10.9) can be interpreted as the sum of two stochastically independent contributions for the state estimate $\hat{\boldsymbol{x}}_{n|n}$.

Another interpretation is to consider the state estimate $\hat{\boldsymbol{x}}_{n|n}$ as the accumulative contribution of a PREDICTOR STEP and a CORRECTOR STEP.

$$\hat{\boldsymbol{x}}_{n|n} = \underbrace{\hat{\boldsymbol{x}}_{n|n-1}}_{\text{estimation step: } \mathsf{E}\left[\boldsymbol{X}_n \,|\, \boldsymbol{Y}_{(n-1)}=\boldsymbol{y}_{(n-1)}\right]} + \overbrace{\boldsymbol{K}_n \underbrace{\left(\boldsymbol{y}_n - \boldsymbol{H}_n \hat{\boldsymbol{x}}_{n|n-1}\right)}_{\text{innovation: } \boldsymbol{\Delta y}_n}}^{\text{correction step: } \mathsf{E}\left[\boldsymbol{X}_n \,|\, \boldsymbol{\Delta Y}_n=\boldsymbol{\Delta} y_{(n)}\right]} . \tag{10.13}$$

Kalman Gain Matrix.

The KALMAN GAIN MATRIX $\boldsymbol{K}_n$ in Eq. (10.9) represents the LINEAR PREDICTOR for the state variable $\boldsymbol{X}_n$ based on the single innovation variable $\boldsymbol{Y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1}$.

$$\begin{aligned}
\boldsymbol{K}_n &= \mathsf{E}\left[\boldsymbol{X}_n\boldsymbol{\Delta}\boldsymbol{Y}_n^{\mathsf{T}}\right]\mathsf{E}\left[\boldsymbol{\Delta}\boldsymbol{Y}_n\boldsymbol{\Delta}\boldsymbol{Y}_n^{\mathsf{T}}\right]^{-1} \\
&= \mathsf{E}\left[\boldsymbol{X}_n(\boldsymbol{Y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1})^{\mathsf{T}}\right]\mathsf{E}\left[(\boldsymbol{Y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1})(\boldsymbol{Y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1})^{\mathsf{T}}\right]^{-1} \\
&= \mathsf{E}\left[\boldsymbol{X}_n(\boldsymbol{H}_n\boldsymbol{X}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1})^{\mathsf{T}}\right]\mathsf{E}\left[(\boldsymbol{Y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1})(\boldsymbol{Y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1})^{\mathsf{T}}\right]^{-1} \\
&= \mathsf{E}\left[\boldsymbol{X}_n(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1})^{\mathsf{T}}\right]\boldsymbol{H}_n^{\mathsf{T}}\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}^{-1} \\
&= \mathsf{E}\left[(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1})(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1})^{\mathsf{T}}\right]\boldsymbol{H}_n^{\mathsf{T}}\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}^{-1} = \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}}\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}^{-1}.
\end{aligned} \tag{10.14}$$

The single steps of the derivation:

- from the 1st to 2nd line by the definition of the innovation sequence,
- from the 2nd to 3rd line by the stochastic independence between $\boldsymbol{X}_n$ and $\boldsymbol{W}_n$,
- from the 3rd to 4th line by the definition of the conditional covariance of the observation variable,
- and from the 4th to 5th line by the orthogonality principle and the definition of the conditional covariance of the state variable.

Innovation Covariance Matrix.

The Innovation Covariance Matrix is derived according to the definition of covariance matrices as

$$\begin{aligned}
\boldsymbol{C}_{\boldsymbol{\Delta Y}_{n|n-1}} &= \mathsf{E}\left[\left(\boldsymbol{Y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1}\right)\left(\boldsymbol{Y}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1}\right)^{\mathsf{T}} \mid \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \mathsf{E}\left[\left(\boldsymbol{H}_n\boldsymbol{X}_n + \boldsymbol{W}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1}\right)\left(\boldsymbol{H}_n\boldsymbol{X}_n + \boldsymbol{W}_n - \boldsymbol{H}_n\hat{\boldsymbol{x}}_{n|n-1}\right)^{\mathsf{T}} \mid \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \mathsf{E}\left[\left(\boldsymbol{H}_n(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1}) + \boldsymbol{W}_n\right)\left(\boldsymbol{H}_n(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1}) + \boldsymbol{W}_n\right)^{\mathsf{T}} \mid \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}} + \boldsymbol{C}_{\boldsymbol{W}_n}. \qquad (10.15)
\end{aligned}$$

Note. $\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}} = \boldsymbol{C}_{\boldsymbol{\Delta Y}_{n|n-1}}$, since the random variable $\boldsymbol{\Delta Y}_{n|n-1}$ is zero mean:

$$\begin{aligned}
\mathsf{E}\left[\boldsymbol{\Delta Y}_n \mid \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] &= \mathsf{E}\left[\boldsymbol{Y}_n - \hat{\boldsymbol{Y}}_{n|n-1} \mid \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \mathsf{E}\left[\boldsymbol{Y}_n \mid \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] - \hat{\boldsymbol{Y}}_{n|n-1} \\
&= \boldsymbol{0}.
\end{aligned}$$

Predicted State Covariance Matrix.

The covariance matrix of the predicted state $\boldsymbol{X}_{n|n-1}$ or the Prediction Error Covariance Matrix can be derived by

$$\begin{aligned}
\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} &= \mathsf{E}\left[\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1}\right)\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1}\right)^{\mathsf{T}} \middle| \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \mathsf{E}\left[\left(\boldsymbol{G}_n\boldsymbol{X}_{n-1} + \boldsymbol{V}_n - \boldsymbol{G}_n\hat{\boldsymbol{x}}_{n-1|n-1}\right)\left(\boldsymbol{G}_n\boldsymbol{X}_{n-1} + \boldsymbol{V}_n - \boldsymbol{G}_n\hat{\boldsymbol{x}}_{n-1|n-1}\right)^{\mathsf{T}} \middle| \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \mathsf{E}\left[\left(\boldsymbol{G}_n(\boldsymbol{X}_{n-1} - \hat{\boldsymbol{x}}_{n-1|n-1}) + \boldsymbol{V}_n\right)\left(\boldsymbol{G}_n(\boldsymbol{X}_{n-1} - \hat{\boldsymbol{x}}_{n-1|n-1}) + \boldsymbol{V}_n\right)^{\mathsf{T}} \middle| \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \boldsymbol{G}_n\boldsymbol{C}_{\boldsymbol{X}_{n-1|n-1}}\boldsymbol{G}_n^{\mathsf{T}} + \boldsymbol{C}_{\boldsymbol{V}_n}. \qquad (10.16)
\end{aligned}$$

Filtered State Covariance Matrix.

The resulting Filtered State Covariance Matrix is analogously given by

$$
\begin{aligned}
\boldsymbol{C}_{\boldsymbol{X}_{n|n}} &= \mathsf{E}\left[\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n}\right)\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n}\right)^{\mathsf{T}} \middle| \boldsymbol{Y}_{(n)} = \boldsymbol{y}_{(n)}\right] \\
&= \mathsf{E}\left[\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n}\right)\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n}\right)^{\mathsf{T}}\right] \\
&= \mathsf{E}\left[\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1} - \boldsymbol{K}_n\boldsymbol{\Delta}\boldsymbol{Y}_n\right)\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1} - \boldsymbol{K}_n\boldsymbol{\Delta}\boldsymbol{Y}_n\right)^{\mathsf{T}}\right] \\
&= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} + \boldsymbol{K}_n\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}\boldsymbol{K}_n^{\mathsf{T}} - 2\boldsymbol{K}_n\,\mathsf{E}\left[\boldsymbol{\Delta}\boldsymbol{Y}_n(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1})^{\mathsf{T}}\right] \\
&= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} + \boldsymbol{K}_n\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}\boldsymbol{K}_n^{\mathsf{T}} - 2\boldsymbol{K}_n\boldsymbol{H}_n\,\mathsf{E}\left[(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1})(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1})^{\mathsf{T}}\right] \\
&= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} + \boldsymbol{K}_n\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}\boldsymbol{K}_n^{\mathsf{T}} - 2\boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} \\
&= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} + \boldsymbol{K}_n\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}^{-1}\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} - 2\boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} \\
&= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} - \boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}. \qquad (10.17)
\end{aligned}
$$

11. Two Examples

11.1 Unknown Parameter Estimation

The Kalman Filter approach can be applied for the recursive estimation of a deterministic unknown parameter $\boldsymbol{x}$ by viewing the parameter as a non-changing state variable $\boldsymbol{X}_n = \boldsymbol{X}_{n-1}$, where only the state observations $\boldsymbol{Y}_n = \boldsymbol{H}_n \boldsymbol{X}_n + \boldsymbol{W}_n$ appear as a random sequence.

For the sake of simplicity, we further assume univariate random variables, which leads to a state space formulation as

$$X_n = X_{n-1} \tag{11.1}$$

$$Y_n = h_n X_n + W_n, \quad \text{i.i.d. } W_n \sim \mathcal{N}(0, \sigma^2_{W_n}). \tag{11.2}$$

The respective update rules of the Kalman Filter are

$$\hat{x}_{n|n-1} = \hat{x}_{n-1|n-1} \tag{11.3}$$

$$\sigma^2_{X_{n|n-1}} = \sigma^2_{X_{n-1|n-1}} \tag{11.4}$$

$$\sigma^2_{Y_{n|n-1}} = h_n^2 \sigma^2_{X_{n|n-1}} + \sigma^2_{W_n} \tag{11.5}$$

$$k_n = h_n \frac{\sigma^2_{X_{n|n-1}}}{\sigma^2_{Y_{n|n-1}}} \tag{11.6}$$

$$\hat{x}_{n|n} = \hat{x}_{n|n-1} + k_n(y_n - h_n \hat{x}_{n|n-1}) \tag{11.7}$$

$$\sigma^2_{X_{n|n}} = \sigma^2_{X_{n|n-1}} - k_n h_n \sigma^2_{X_{n|n-1}}. \tag{11.8}$$

which can be derived from the general recursive state tracking procedure of the Kalman Filter.

For the numerical analysis, we assume

- $x = 1$ for the parameter to be estimated,
- the initial conditions $x_1 = 0$ and $\sigma^2_{X_{1|1}} = 1$,
- and the parameters $\sigma^2_{W_n} \equiv 0.01$, $h_n = \sin(\pi n/100)$ for all $n = 1, \ldots, 100$.

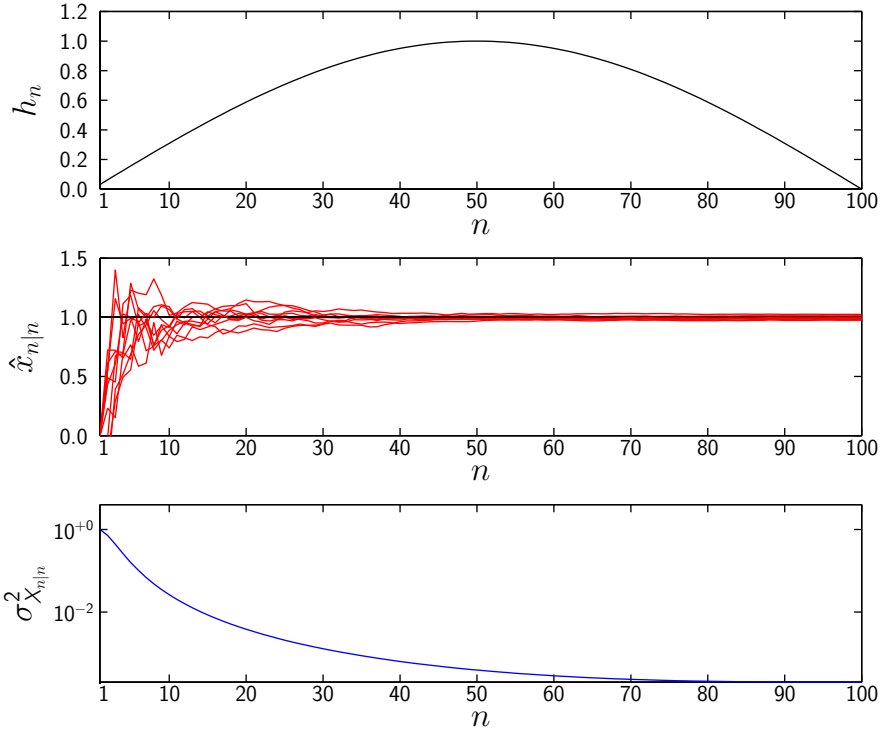


Fig. 11.1: Stochastic paths $\hat{x}_{n|n}$ of the estimation of a constant parameter $x = 1$ for 10 uniformly chosen noise sequences W_n and $Y_n = \sin(\pi n/100) + W_n$, $\forall n = 1, \ldots, 100$, based on the Kalman Filter.

11.2 Tracking Example

The discretization of a stochastically excited dynamical system[1]

$$\ddot{X}_t = V_t, \quad \text{i.i.d. } V_n \sim \mathcal{N}(0, \sigma^2_{V_n}), \tag{11.9}$$

with an equidistant sampling intervall of T, yields the discrete linear system[2]

$$X_n = X_{n-1} + T\dot{X}_{n-1} \tag{11.10}$$

$$\dot{X}_n = \dot{X}_{n-1} + TV_n. \tag{11.11}$$

When applying a Kalman Filter for tracking the state space variable X_n, based on observations

$$Y_n = X_n + W_n, \quad \text{i.i.d. } W_n \sim \mathcal{N}(0, \sigma^2_{W_n}), \tag{11.12}$$

and taking into account Eqs. (10.1)-(10.2), we obtain

$$\boldsymbol{X}_n = \boldsymbol{G}_n \boldsymbol{X}_{n-1} + \begin{bmatrix} 0 \\ T \end{bmatrix} V_n, \tag{11.13}$$

$$Y_n = \boldsymbol{h}_n^{\mathsf{T}} \boldsymbol{X}_n + W_n, \tag{11.14}$$

with $\boldsymbol{X}_n = [X_n \ \dot{X}_n]^{\mathsf{T}}$ and

$$\boldsymbol{G}_n = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{h}_n = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{11.15}$$

[1] Eq. (11.9) denotes a Stochastical Differential Equation (SDE).

[2] The discretization is based on $\dot{X}_n - \dot{X}_{n-1} = T\ddot{X}_{n-1}$ and $X_n - X_{n-1} = T\dot{X}_{n-1}$.

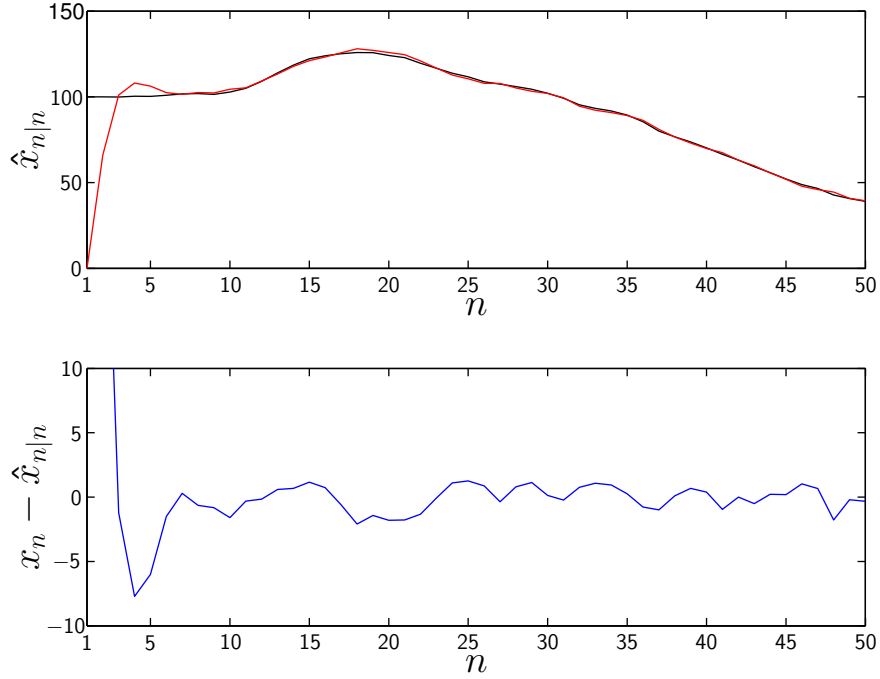


Fig. 11.2: Estimation of the trajetory $\hat{x}_{n|n}$, $n = 1, \ldots, 50$, of one single stochastic path from the SDE $X_n = X_{n-1} + \dot{X}_{n-1}$ and $\dot{X}_n = \dot{X}_{n-1} + V_{n-1}$ for the initial conditions $x_0 = 100$ and $\dot{x}_0 = 0$. The Kalman Filter is based on observations $Y_n = X_n + W_n$ and initial conditions $\hat{\boldsymbol{x}}_{0|0} = \boldsymbol{0}$ and $\boldsymbol{C}_{0|0} = \boldsymbol{I}$. The parameter $T = 1$.

12. Particle Filter

The KALMAN FILTER is a powerful linear approach for estimating the states of a random sequence, otherwise it only shows optimum performance for GAUSS-MARKOV SEQUENCES, this is if there is a linear relation between states of the sequence and if there is an additional Gaussian distortion at most.

For the general state-space model,

$$\boldsymbol{X}_n = g_n(\boldsymbol{X}_{n-1}, \boldsymbol{V}_n), \tag{12.1}$$
$$\boldsymbol{Y}_n = h_n(\boldsymbol{X}_n, \boldsymbol{W}_n), \tag{12.2}$$

where g_n and h_n are nonlinear functions and $\boldsymbol{V}_n$ and $\boldsymbol{W}_n$ are non-Gaussian distributed random variables, alternative techniques beyond the Kalman Filter are required.

The general approach is to substitute the linear Kalman Filter by SUBOPTIMUM NONLINEAR FILTERS.

The three most well known nonlinear alternatives to the KALMAN FILTER are

- the EXTENDED KALMAN FILTER,
- the UNSCENTED KALMAN FILTER,
- and the PARTICLE FILTER.

a) The EXTENDED KALMAN FILTER is typically applied to state-space models like

$$\boldsymbol{X}_n = g_n(\boldsymbol{X}_{n-1}) + \boldsymbol{V}_n, \tag{12.3}$$
$$\boldsymbol{Y}_n = h_n(\boldsymbol{X}_n) + \boldsymbol{W}_n, \tag{12.4}$$

where $\boldsymbol{V}_n$ and $\boldsymbol{W}_n$ are or are assumed to be Gaussian distributed random variables.

The recursive estimation of the state-space variables in a standard KALMAN FILTER type fashion is based on a linear approximation of the nonlinear functions g_{n-1} and h_n in (12.3) and (12.4). The approximation must be performed in each step of the recursive estimation process. Obviously, the EXTENDED KALMAN FILTER in general cannot achieve optimal performance.

b) The so-called UNSCENTED KALMAN FILTER approximates the desired conditional PDF $f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)})$ by a Gaussian PDF of the original recursive computation process.

The recursive estimation process again resembles the standard KALMAN FILTER process, but replacing the estimation of the conditional state covariance matrix in each step by a weighted sample covariance matrix of the non-linear transformed samples of the state-space variables.

c) Whereas the EXTENDED KALMAN FILTER and the UNSCENTED KALMAN FILTER strongly depend on the validity of the Gaussian assumption, the PARTICLE FILTER approach completely breaks with these requirements. It is purely based on the fundamental update rule of all tracking approaches, namely the recursive computation of the posterior conditional PDF.

$$\overbrace{f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)})}^{\text{desired conditional PDF in the } n\text{th step}} \propto \underbrace{f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n|\boldsymbol{x}_n)}_{\text{likelihood of the new observation}} \times \int_{\mathbb{X}} \underbrace{f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})}_{\text{state-transition PDF}} \underbrace{f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)})\,\mathrm{d}\boldsymbol{x}_{n-1}}_{\text{conditional PDF from the } (n-1)\text{th-step before}}. \tag{12.5}$$

It remains the task to compute the integral of the recursive update of the posterior conditional PDF!

12.1 Monte Carlo Integration

Since the integration of (9.13) is in general impossible, the respective integral is solved via Monte Carlo Integration. Monte Carlo Integration means to replace an integral by its sample mean

$$I = \int_{\mathbb{X}} g(\xi) f_X(\xi) \, \mathrm{d}\xi \quad \approx \quad I_N = \frac{1}{N} \sum_{i=1}^{N} g(x^i), \tag{12.6}$$

where the samples x^i, $i = 1, \ldots, N$, are drawn according to the PDF $f_X(x)$, i.e., $x^i \sim f_X(x)$.

Theorem.

If the samples x^i are stochastically independent, the approximation I_N is an unbiased estimate of I with

$$\lim_{N \to \infty} I_N \to I, \tag{12.7}$$

and

$$\lim_{N \to \infty} \sqrt{N}(I_N - I) \sim \mathcal{N}(0, \sigma^2), \tag{12.8}$$

with $\sigma^2 = \mathrm{Var}\,[g(X)]$.

Proof.

The proof is based on the law of large numbers and the fact that $I = \mathrm{E}\,[g(X)]$.

Importance Sampling.

Assume – as usual in the PARTICLE FILTER framework – instead of using the PDF $f_X(x)$ for drawing samples, we use an IMPORTANCE DENSITY $q_X(x)$ with $q_X(x) > 0$ if $f_X(x) > 0$ for all x.[1] In this case, the correct MONTE CARLO INTEGRATION can be still guaranteed by appropriate weights for the samples,

$$I = \int_{\mathbb{X}} g(\xi) f_X(\xi)\, \mathrm{d}\xi = \int_{\mathbb{X}} g(\xi) \frac{f_X(\xi)}{q_X(\xi)} q_X(\xi)\, \mathrm{d}\xi \quad \approx \quad I_N = \frac{1}{N} \sum_{i=1}^{N} g(x^i) \frac{f_X(x^i)}{q_X(x^i)}, \tag{12.9}$$

where the samples x^i, $i = 1, \ldots, N$, are drawn according to the PDF $q_X(x)$, i.e.,

$$I \approx I_N = \frac{1}{N} \sum_{i=1}^{N} \tilde{w}^i g(x^i), \tag{12.10}$$

with $\tilde{w}^i \triangleq \frac{f_X(x^i)}{q_X(x^i)}$ and $x^i \sim q_X(x)$.

[1] This might be motivated by serveral arguments, e.g., drawing samples from $q_X(x)$ might be simpler than drawing samples from $f_X(x)$.

Importance Sampling if $\int_{\mathbb{X}} f_X(\xi)\,\mathrm{d}\xi \neq 1$.

If $f_X(x)$ fulfills the properties of a PDF except a normalization constant, the IMPORTANCE DENSITY results in

$$I \approx I_N = \sum_{i=1}^{N} w^i g(x^i), \tag{12.11}$$

with normalized weights

$$w^i \triangleq \frac{\tilde{w}^i}{\sum_{i=1}^{N} \tilde{w}^i}, \quad \tilde{w}^i \triangleq \frac{f_X(x^i)}{q_X(x^i)} \quad \text{and} \quad x^i \sim q_X(x).$$

Note.

The MONTE CARLO INTEGRAL I_N can be interpreted as the true expected value of $g(X)$ with respect to the approximate PDF[2]

$$\sum_{i=1}^{N} w^i \delta(x - x^i) \approx f_X(x). \tag{12.12}$$

[2] $\mathsf{E}[g(X)] = \int_{\mathbb{X}} g(\xi) \sum_{i=1}^{N} w^i \delta(\xi - x^i)\,\mathrm{d}\xi = \sum_{i=1}^{N} w^i \int_{\mathbb{X}} g(\xi)\delta(\xi - x^i)\,\mathrm{d}\xi = \sum_{i=1}^{N} w^i g(x^i)$.

12.2 Sequential Importance Sampling

We now apply the IMPORTANCE SAMPLING to the POSTERIOR CONDITIONAL PDF $f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)})$. Due to the recursive computation, we obtain a SEQUENTIAL IMPORTANCE SAMPLING rule.

First, we again consider the IMPORTANCE DENSITY and the POSTERIOR CONDITIONAL PDF for the construction of the IMPORTANCE WEIGHTS in a slightly different version, viz.[3]

$$\tilde{w}_n^i = \frac{f_{\boldsymbol{X}_{(n)}|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_{(n)}^i|\boldsymbol{y}_{(n)})}{q_{\boldsymbol{X}_{(n)}|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_{(n)}^i|\boldsymbol{y}_{(n)})} = \frac{f_{\boldsymbol{X}_n,\boldsymbol{X}_{(n-1)}|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n^i,\boldsymbol{x}_{(n-1)}^i|\boldsymbol{y}_{(n)})}{q_{\boldsymbol{X}_n,\boldsymbol{X}_{(n-1)}|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n^i,\boldsymbol{x}_{(n-1)}^i|\boldsymbol{y}_{(n)})}. \tag{12.13}$$

We then expand the nominator by

$$\begin{aligned}
& f_{\boldsymbol{X}_n,\boldsymbol{X}_{(n-1)}|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n,\boldsymbol{x}_{(n-1)} \mid \boldsymbol{y}_{(n)}) \\
\propto\ & f_{\boldsymbol{Y}_n|\boldsymbol{X}_n,\boldsymbol{X}_{(n-1)},\boldsymbol{Y}_{(n-1)}}(\dots) f_{\boldsymbol{X}_n,\boldsymbol{X}_{(n-1)}|\boldsymbol{Y}_{(n-1)}}(\dots) \\
\propto\ & \underbrace{f_{\boldsymbol{Y}_n|\boldsymbol{X}_n,\boldsymbol{X}_{(n-1)},\boldsymbol{Y}_{(n-1)}}(\dots)}_{\text{Markov property!}} \underbrace{f_{\boldsymbol{X}_n|\boldsymbol{X}_{(n-1)},\boldsymbol{Y}_{(n-1)}}(\dots)}_{\text{Markov property!}} f_{\boldsymbol{X}_{(n-1)}|\boldsymbol{Y}_{(n-1)}}(\dots) \\
\propto\ & f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n) f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}) f_{\boldsymbol{X}_{(n-1)}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{(n-1)} \mid \boldsymbol{y}_{(n-1)}),
\end{aligned} \tag{12.14}$$

[3] Note, that this version is equivalent with the original definition of IMPORTANCE WEIGHTS up to scaling factors, since at the n-th time instant the realizations of the $(n-1)$-th state vectors $\boldsymbol{X}_{(n-1)}$ by means of its particles $\boldsymbol{x}_{(n-1)}^i$ are no longer random. The scaling factors are without any impact, due to the subsequent normalization steps.

and approximate the denominator as[4]

$$\begin{aligned} & q_{\boldsymbol{X}_n, \boldsymbol{X}_{(n-1)} \mid \boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n, \boldsymbol{x}_{(n-1)} \mid \boldsymbol{y}_{(n)}) \\ \approx \quad & \underbrace{q_{\boldsymbol{X}_n \mid \boldsymbol{X}_{(n-1)}, \boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{(n-1)}, \boldsymbol{y}_{(n)})}_{\text{Markov property!}} q_{\boldsymbol{X}_{(n-1)} \mid \boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{(n-1)} \mid \boldsymbol{y}_{(n-1)}) \\ = \quad & q_{\boldsymbol{X}_n \mid \boldsymbol{X}_{n-1}, \boldsymbol{Y}_n}(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}, \boldsymbol{y}_n) q_{\boldsymbol{X}_{(n-1)} \mid \boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{(n-1)} \mid \boldsymbol{y}_{(n-1)}). \end{aligned} \tag{12.15}$$

Consequently, the weights are obtained by

$$\tilde{w}_n^i = \frac{f_{\boldsymbol{Y}_n \mid \boldsymbol{X}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n^i) f_{\boldsymbol{X}_n \mid \boldsymbol{X}_{n-1}}(\boldsymbol{x}_n^i \mid \boldsymbol{x}_{n-1}^i) f_{\boldsymbol{X}_{(n-1)} \mid \boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{(n-1)}^i \mid \boldsymbol{y}_{(n-1)})}{q_{\boldsymbol{X}_n \mid \boldsymbol{X}_{n-1}, \boldsymbol{Y}_n}(\boldsymbol{x}_n^i \mid \boldsymbol{x}_{n-1}^i, \boldsymbol{y}_n) q_{\boldsymbol{X}_{(n-1)} \mid \boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{(n-1)}^i \mid \boldsymbol{y}_{(n-1)})}. \tag{12.16}$$

[4] It is an approximation, since it does not equal with the correct expansion of the IMPORTANCE DENSITY. The correct expansion following the BAYES RULE would read $q_{\boldsymbol{X}_n, \boldsymbol{X}_{(n-1)} \mid \boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n, \boldsymbol{x}_{(n-1)} \mid \boldsymbol{y}_{(n)}) = q_{\boldsymbol{X}_n \mid \boldsymbol{X}_{(n-1)}, \boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{(n-1)}, \boldsymbol{y}_{(n)}) q_{\boldsymbol{X}_{(n-1)} \mid \boldsymbol{Y}_{(n)}}(\boldsymbol{x}_{(n-1)} \mid \boldsymbol{y}_{(n)})$. In other words, the defined version neglects the information from the latest observation $\boldsymbol{y}_n$.

Finally, taking into account the definition in (12.13) the weights of the PARTICLES (samples) $\boldsymbol{x}_n^i$ in the nth step are given by

$$\tilde{w}_n^i = \tilde{w}_{n-1}^i \times \frac{f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n^i) f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n^i \mid \boldsymbol{x}_{n-1}^i)}{q_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_n}(\boldsymbol{x}_n^i \mid \boldsymbol{x}_{n-1}^i, \boldsymbol{y}_n)}, \quad \forall i = 1, \ldots, N, \tag{12.17}$$

where the particles $\boldsymbol{x}_n^i$ are drawn from the conditional PDF $q_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_n}(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}^i, \boldsymbol{y}_n)$.

Given the particles $\boldsymbol{x}_{n-1}^i$ with respect to $q_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_n}(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}^i, \boldsymbol{y}_n)$ in the latest step of the recursion process, and given an observation $\boldsymbol{y}_n$ with respect to the state-space model, e.g.,

$$\boldsymbol{X}_n = g_n(\boldsymbol{X}_{n-1}) + \boldsymbol{V}_n, \tag{12.18}$$
$$\boldsymbol{Y}_n = h_n(\boldsymbol{X}_n) + \boldsymbol{W}_n, \tag{12.19}$$

where $\boldsymbol{V}_n$ and $\boldsymbol{W}_n$ are random variables with any PDF $f_{\boldsymbol{V}_n}(\boldsymbol{v})$ and $f_{\boldsymbol{W}_n}(\boldsymbol{w})$, the weights of the IMPORTANCE SAMPLING are updated according to (12.17) with

$$f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n^i) = f_{\boldsymbol{W}_n}(\boldsymbol{y}_n - h_n(\boldsymbol{x}_n^i)) \tag{12.20}$$
$$f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n^i \mid \boldsymbol{x}_{n-1}^i) = f_{\boldsymbol{V}_n}(\boldsymbol{x}_n^i - g_n(\boldsymbol{x}_{n-1}^i)). \tag{12.21}$$

12.3 Degeneracy Problem and Resampling

Degeneracy.

It has been shown that using an IMPORTANCE DENSITY with

$$q_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_n}(\dots)q_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\dots) \neq f_{\boldsymbol{X}_n,\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n)}}(\dots)$$

leads to a monotonic increase of the variance of the weights over the number of iterations n. This in turn leads to the DEGENERACY EFFECT, which means that after a number of recursive steps all but one particle have vanishing normalized IMPORTANCE WEIGHTS.

DEGENERACY is typically detected by a threshold test. If

$$\frac{1}{\sum_{i=1}^{N}(w_n^i)^2} \leq w_{\text{thr}} \tag{12.22}$$

is smaller than a certain threshold w_{thr}, a resampling process must be performed.

It is the purpose of the RESAMPLING STEP to eliminate particles with small weights and to replicate particles with strong weights. REPLICATION means that when drawing new PARTICLES those $\boldsymbol{x}_n^i$ with strong weights might have more than one follower in the next recursive step.

Resampling Step.

Particles are randomly drawn (with replacement) from the finite set of particles $\mathbb{X}_n^N \triangleq \{\boldsymbol{x}_n^1, \ldots, \boldsymbol{x}_n^N\}$, where the probability of each particle is equal to its current IMPORTANCE WEIGHT, i.e., $\mathsf{P}(\{\boldsymbol{x}_n^i\}) = w_n^i$.

Using an alternative formulation of the RESAMPLING STEP by means of the respective PMF[5], we obtain

$$\forall j = 1, \ldots, N: \quad \boldsymbol{x}_n^j \sim p_{\boldsymbol{X}_n}(\boldsymbol{x}_n^j) \qquad \text{with} \qquad p_{\boldsymbol{X}_n}(\boldsymbol{x}_n) = \sum_{i=1}^{N} w_n^i \delta(\boldsymbol{x}_n - \boldsymbol{x}_n^i). \tag{12.23}$$

The PMF $p_{\boldsymbol{X}_n}(\boldsymbol{x}_n)$ obviously corresponds to the approximated posterior conditional PDF

$$f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n \mid \boldsymbol{y}_{(n)}) \approx p_{\boldsymbol{X}_n}(\boldsymbol{x}_n). \tag{12.24}$$

See (12.11) and (12.12).[6]

Note.

After the RESAMPLING STEP the IMPORTANCE WEIGHTS are uniformly set to $w_n^i = N^{-1}$. Thereby the phenomenon of degeneracy is broken.

---

[5] PROBABILITY MASS FUNCTION.

[6] Directly following (12.11) and (12.12) leads to $f_{\boldsymbol{X}_{(n)}|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_{(n)} \mid \boldsymbol{y}_{(n)}) \approx \sum_{i=1}^{N} w_n^i \delta(\boldsymbol{x}_{(n)} - \boldsymbol{x}_{(n)}^i) = \sum_{i=1}^{N} w_n^i \delta(\boldsymbol{x}_n - \boldsymbol{x}_n^i)\delta(\boldsymbol{x}_{(n-1)} - \boldsymbol{x}_{(n-1)}^i)$ as we started with $f_{\boldsymbol{X}_{(n)}|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_{(n)} \mid \boldsymbol{y}_{(n)})$ in (12.13). A subsequent marginalization step with respect to $\boldsymbol{x}_{(n-1)}$ finally yields (12.24).

12.4 Importance Density

Optimal Importance Density.

So far we did not discuss the choice of the IMPORTANCE DENSITY $q_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_n}(\dots)$, which is of course essential for the performance of the PARTICLE FILTER. It has been shown that the optimal choice which minimizes the variance of the IMPORTANCE WEIGHTS is equal to the true posterior conditional PDF, i.e.,

$$\begin{aligned} q^{\mathsf{opt}}_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_n}(\boldsymbol{x}_n \mid \boldsymbol{x}^i_{n-1}, \boldsymbol{y}_n) &= f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_n}(\boldsymbol{x}_n \mid \boldsymbol{x}^i_{n-1}, \boldsymbol{y}_n) \\ &= \frac{f_{\boldsymbol{Y}_n|\boldsymbol{X}_n,\boldsymbol{X}_{n-1}}(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{x}^i_{n-1}) f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n \mid \boldsymbol{x}^i_{n-1})}{f_{\boldsymbol{Y}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{y}_n \mid \boldsymbol{x}^i_{n-1})}, \end{aligned} \tag{12.25}$$

with

$$f_{\boldsymbol{Y}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{y}_n \mid \boldsymbol{x}^i_{n-1}) = \int_{\mathbb{X}} f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n \mid \boldsymbol{\xi}) f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{\xi} \mid \boldsymbol{x}^i_{n-1}) \,\mathrm{d}\boldsymbol{\xi}. \tag{12.26}$$

Applying the OPTIMAL IMPORTANCE DENSITY in (12.17) results in the OPTIMAL SEQUENTIAL IMPORTANCE SAMPLING given by

$$\tilde{w}^{\mathsf{opt},i}_n = \tilde{w}^{\mathsf{opt},i}_{n-1} \times f_{\boldsymbol{Y}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{y}_n \mid \boldsymbol{x}^i_{n-1}), \quad \forall i = 1, \dots, N. \tag{12.27}$$

Suboptimal Importance Density.

Since in general (12.25) and (12.26) are hard to compute in real-world application, the Optimal Importance Density can be only applied in special cases.[7]

Otherwise, the most popular Suboptimal Importance Density ist given by

$$q_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1},\boldsymbol{Y}_n}(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}^i, \boldsymbol{y}_n) \stackrel{\triangle}{=} f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}^i). \tag{12.28}$$

Applying this suboptimal version of the Importance Density in (12.17) results in the Suboptimal Sequential Importance Sampling given by

$$\tilde{w}_n^i = \tilde{w}_{n-1}^i \times f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n^i), \quad \forall i = 1, \ldots, N. \tag{12.29}$$

It has been observed that the Suboptimal Sequential Importance Sampling works quite well, except for the case where the Transitional PDF $f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}^i)$ has a broader spread than the likelihood function $f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n \mid \boldsymbol{x}_n)$. For the latter a rapid degeneracy of the weights has been observed.

There is huge room for further suboptimal techniques...

[7] One of this special cases is again the Additive Gaussian Model. The linearity assumption is not required for the Particle Filtering.

13. Tracking the Trajectory of a Moving Object

In the following, we consider the State-Space Models given as

$$\boldsymbol{X}_{n+1} = g(\boldsymbol{X}_n) + \boldsymbol{V}_n, \tag{13.1}$$

$$\boldsymbol{Y}_n = h(\boldsymbol{X}_n) + \boldsymbol{W}_n, \tag{13.2}$$

where for simplicity additive noise is assumed. $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$ are state space and measurements vectors. The functions $g(\,\cdot\,)$ and $h(\,\cdot\,)$ are in general non-linear. $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ are not necessarily Gaussian distributed.

In the following, we will estimate the results of this state-space model using a Kalman Filter (KF) and using a simple Particle Filter (PF) with Sequential Importance Re-sampling (SIR).

13.1 Linear Model

We consider a two-dimensional version of the example in Section 11.2, i.e.,

$$\ddot{\boldsymbol{X}}_t = \boldsymbol{V}_t, \quad \text{i.i.d. } \boldsymbol{V}_n \sim \mathcal{N}(0, \boldsymbol{C}_V). \tag{13.3}$$

The discretization of the Stochastical Differential Equation (13.3) yields

$$\boldsymbol{X}_{n+1} = \boldsymbol{G}\boldsymbol{X}_n + \boldsymbol{V}_n, \tag{13.4}$$
$$\boldsymbol{Y}_n = \boldsymbol{H}\boldsymbol{X}_n + \boldsymbol{W}_n, \tag{13.5}$$

with

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \tag{13.6}$$

The state space vector and the observation vector are

$$\boldsymbol{X}_n = \begin{bmatrix} X_{1,n} \\ X_{2,n} \\ \dot{X}_{1,n} \\ \dot{X}_{2,n} \end{bmatrix} \in \mathbb{R}^4 \quad \text{and} \quad \boldsymbol{Y}_n \in \mathbb{R}^2. \tag{13.7}$$

The $X_{1,n}$ and $X_{2,n}$ represent the random variables for a position on the (x_1,x_2)–plane and $\dot{X}_{1,n}$ and $\dot{X}_{2,n}$ represent the random variables for the respective velocities along the x_1 and x_2 direction. The $Y_{1,n}$ and $Y_{2,n}$ denote noisy observations of $X_{1,n}$ and $X_{2,n}$.

The $\boldsymbol{V}_n$ and $\boldsymbol{W}_n$ are assumed to be Gaussian distributed with

$$\boldsymbol{V}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_V) \quad \text{and} \quad \boldsymbol{W}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_W), \tag{13.8}$$

and

$$\boldsymbol{C}_V = T \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 50 & 0 \\ 0 & 0 & 0 & 50 \end{bmatrix} \quad \text{and} \quad \boldsymbol{C}_W = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}. \tag{13.9}$$

Note.

For the PARTICLE FILTER the number of particles is set to $N = 100$ and resampling is applied if

$$N = \frac{1}{\sum_{i=1}^{N} (w_n^i)^2} \leq 50. \tag{13.10}$$

The w_n^i are the particle weights at time-step n.

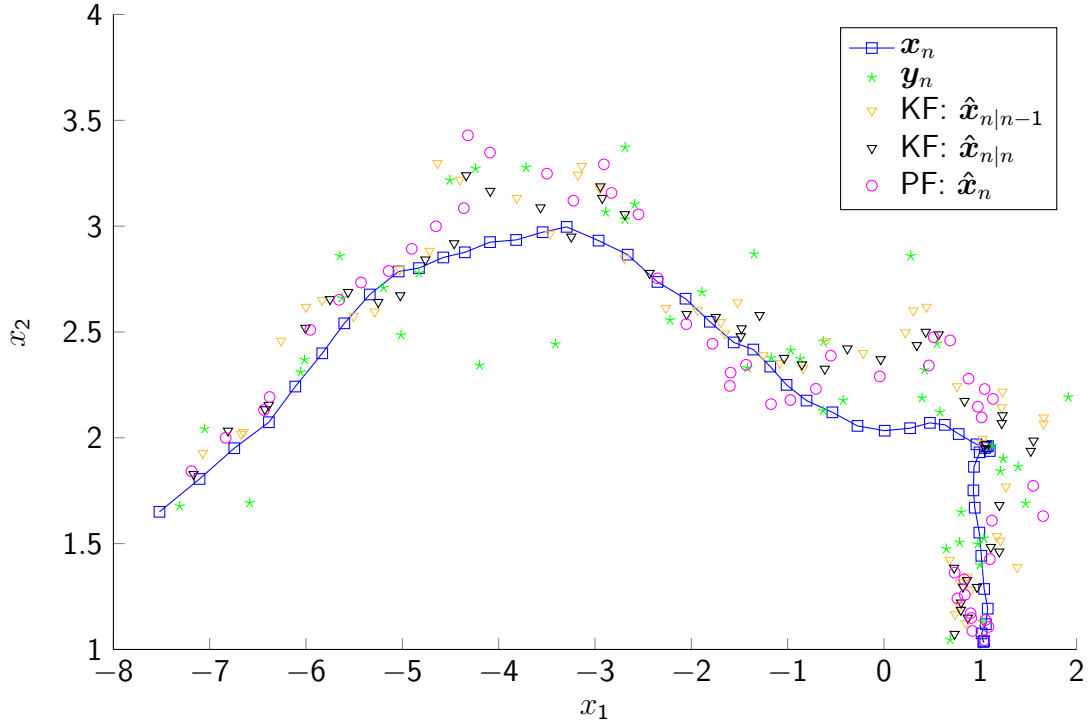


Fig. 13.1: Tracking of an exemplary random trajectory in the (x_1,x_2)–plane based on KALMAN (KF) and PARTICLE FILTERING (PF) applied to observations $\boldsymbol{y}_n$ from a LINEAR SYSTEM MODEL with initial condition $\boldsymbol{x}_0 = [1, 0, 0, 0]^\mathsf{T}$, $T = 0.02$ and $n = 1 \ldots 50$.

13.2 Nonlinear Model

The nonlinear model differs from the model in Section 13.1 by a modification of the first row of Eq. (13.4), i.e., we replace $\dot{X}_{1,n+1} = X_{1,n} + T\dot{X}_{1,n}$ by

$$\dot{X}_{1,n+1} = \sin(X_{1,n}) + T\dot{X}_{1,n}. \tag{13.11}$$

Additionally, the measurement noise covariance matrix is changed to

$$\boldsymbol{C}_W = \begin{bmatrix} 0.025 & 0 \\ 0 & 0.025 \end{bmatrix}. \tag{13.12}$$

Note.

For the PARTICLE FILTER the number of particles is again set to $N = 100$ and resampling is applied if

$$N = \frac{1}{\sum_{i=1}^{N}(w_n^i)^2} \leq 50. \tag{13.13}$$

For comparison we use the Kalman Filter with the linearization (Jacobi matrix) of $g(\boldsymbol{x}_n)$ at the initial condition $\boldsymbol{x}_0$ as constant matrix $\boldsymbol{G}$, i.e.,

$$[\boldsymbol{G}]_{k\ell} = \left.\frac{\partial[g(\boldsymbol{x})]_k}{\partial[\boldsymbol{x}]_\ell}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}. \tag{13.14}$$

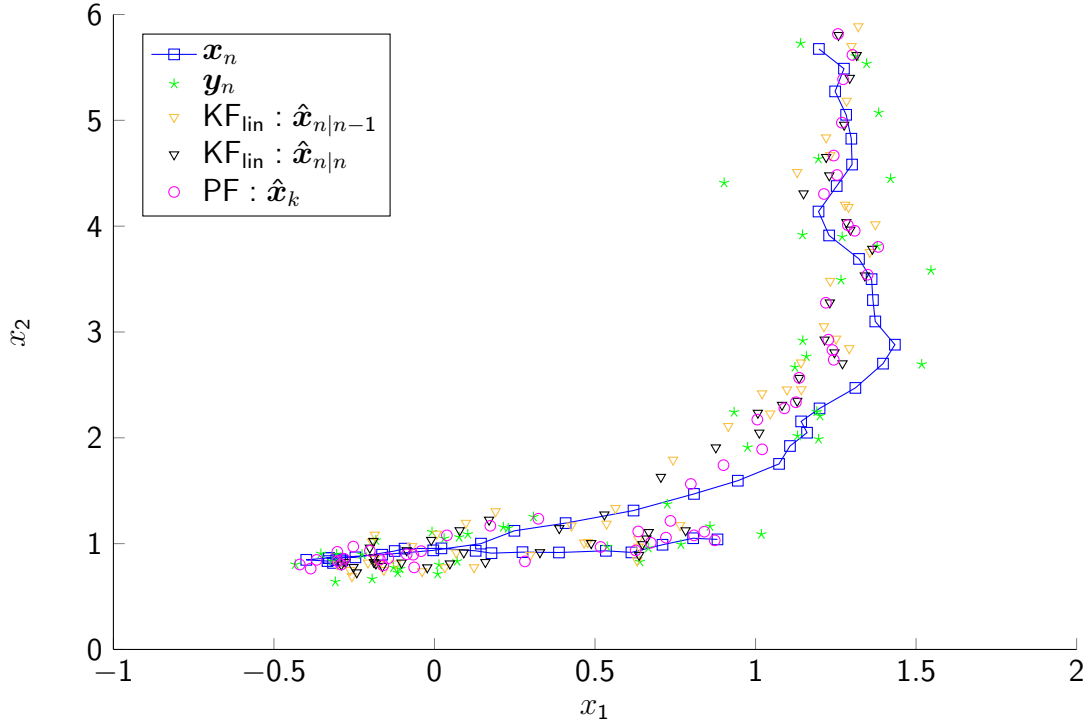


Fig. 13.2: Tracking of an exemplary random trajectory in the (x_1,x_2)–plane based on KALMAN (KF) and PARTICLE FILTERING (PF) applied to observations $\boldsymbol{y}_n$ from a NONLINEAR SYSTEM MODEL with initial condition $\boldsymbol{x}_0 = [1, 0, 0, 0]^{\mathsf{T}}$, $T = 0.02$ and $n = 1 \ldots 50$.

14. Demo

https://syncandshare.lrz.de/dl/fiDKhkL7MKJpUZ9R7YxnUQa9

14.1 Gauss-Markov Model with Standard Measurement Noise

We illustrate the Particle Filter using a very simple Gauss-Markov model

$$\mathsf{X}_\mathsf{n} = \mathsf{X}_{n-1} + \mathsf{V}_n, \tag{14.1}$$
$$\mathsf{Y}_\mathsf{n} = \mathsf{X}_n + \mathsf{W}_n \tag{14.2}$$

with $x_0 = 0$, $\mathsf{V}_\mathsf{n} \sim \mathcal{N}(0,1)$, and $\mathsf{W}_n \sim \mathcal{N}(0,1)$.

- We use the sub-optimal Importance Density $f_{\mathsf{X}_\mathsf{n}|\mathsf{X}_\mathsf{n-1}}(x_n|x_{n-1})$, i.e.,

$$q_{\mathsf{X}_\mathsf{n}|\mathsf{X}_\mathsf{n-1},\mathsf{Y}_\mathsf{n}}(x_n|x_{n-1},y_n) = f_{\mathsf{X}_\mathsf{n}|\mathsf{X}_\mathsf{n-1}}(x_n|x_{n-1}). \tag{14.3}$$

- We use $N_{\mathsf{part}} = 500$ particles with $w_{\mathsf{thr}} = 25$.

- The PARTICLE FILTER approximates the optimal KALMAN FILTER which in turn estimates the true trajectory is $\mathsf{X}_n = x_n$ with $n = 1, \ldots, 51$.
- After each new measurement $\mathsf{Y}_n = y_n$, particles are propagated by using the IMPORTANCE DENSITY $f_{\mathsf{X}_n|\mathsf{X}_{n-1}}(x_n|x_{n-1}^i)$.
- In this example, $f_{\mathsf{X}_n|\mathsf{X}_{n-1}}(x_n|x_{n-1}^i) = \mathcal{N}(x_{n-1}^i, 1)$.
- Then, the weights are updated by using

$$\tilde{w}_n^i = w_{n-1}^i \times f_{\mathsf{Y_n}|\mathsf{X_n}}(y_n|x_n^i) \tag{14.4}$$

and scaled by using

$$w_n^i = \frac{\tilde{w}_n^i}{\sum w_n^i}. \tag{14.5}$$

- In this example, $f_{\mathsf{Y_n}|\mathsf{X_n}}(y_n|x_n^i) = \mathcal{N}(x_n^i, 1)$ and $w_0^i = \frac{1}{N}$.
- Eventually, $\mathsf{E}[\mathsf{X_n}|\mathsf{Y}_{(n)} = y_{(n)}]$ is approximated by using $\hat{x}_{n|n} \approx \sum x_n^i w_n^i$.
- If WEIGHT DEGENERACY occurs, i.e.,

$$\frac{1}{\sum (w_n^i)^2} \leq w_{\mathsf{thr}} = 25, \tag{14.6}$$

particles are resampled by means of $\mathsf{P}(x_n^i) = w_n^i$ and the weights are set to $w_n^i = \frac{1}{N}$.

14.2 Gauss-Markov Model with Non-Standard Measurement Noise

We again illustrate the Particle Filter using a very simple Gauss-Markov model

$$\mathsf{X_n} = \mathsf{X}_{n-1} + \mathsf{V}_n, \tag{14.7}$$

$$\mathsf{Y_n} = \mathsf{X}_n + \mathsf{W}_n \tag{14.8}$$

but now with $x_0 = 0$, $\mathsf{V_n} \sim \mathcal{N}(0, 2)$, and $\mathsf{W}_n \sim \frac{1}{2}\mathcal{N}(0, 0.1) + \frac{1}{2}\mathcal{N}(10, 0.1))$.

In this case of a Bimodal Gaussian Distribution, the Particle Filter will clearly outperform the Kalman Filter, which is not able to interpret the likelihood of the measurements in an appropriate way.

In this example,

- $f_{\mathsf{Y_n}|\mathsf{X_n}}(y_n|x_n^i) = \frac{1}{2}\mathcal{N}(x_n^i, 0.1) + \frac{1}{2}\mathcal{N}(x_n^i + 10, 0.1)$ with $w_0^i = \frac{1}{N}$.

References

- D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.
- B. Hajek. *An Exploration of Random Processes for Engineers*. Lecture Notes at University of Illinois, Urbana-Champaign. URL: http://www.freetechbooks.com.
- G. R. Grimmett, D. R. Stirzaker. *Probability and Random Processes*, Oxford University Press. (↑advanced reader!)
- P. M. Djurić et al. "Particle Filtering: A review of the theory and how it can be used for solving problems in wireless communications", *IEEE Signal Processing Magazine*, p. 19–38, September 2003.
- B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter*, Artech House, 2004.
- D. Simon. *Optimal State Estimation*, Wiley, 2006. (↑advanced reader!)

Part VII

Hypothesis Testing

15. Hypothesis Testing – Decision Making

Contrary to parameter estimation, hypothesis testing is about making a decision based on the observations, so-called decision making problems. Again the statistical model is in the main focus.

15.1 Statistical Model

Given the statistical model $\{\mathbb{X}, \mathbb{F}, \mathsf{P}_\theta; \theta \in \Theta\}$, we can decompose the model parameter $\Theta = \Theta_0 \cup \Theta_1$ into a so-called null hypothesis Θ_0 and alternative Θ_1. In this lecture we restrict ourselves to the singleton sets $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.

Definition. A test is a statistic

$$d : \mathbb{X} \to [0, 1].$$

The statistic d corresponds to a decision rule and the set $\{x \in \mathbb{X} | d(x) = 1\}$ is called rejection region.

Definition. The cost criterion for a test d is

$$G_d : \Theta \to [0,1], \theta \mapsto \mathsf{E}\left[d(X)|\theta\right].$$

Definition. A test has an error level α if

$$\mathsf{E}\left[d(X)|\theta_0\right] \leq \alpha$$

Two Error Types

false alarm: Although $\theta \in \Theta_0$, the null hypothesis is canceled, that is $d(x) = 1$. The false alarm error probability is

$$G_d(\theta_0) = \mathsf{E}\left[d(X)|\theta_0\right].$$

detection error: Although $\theta \in \Theta_1$, the alternative is not detected, that is $d(x) = 0$. The detection error probability is

$$1 - G_d(\theta_1) = 1 - \mathsf{E}\left[d(X)|\theta_1\right].$$

15.2 Design of a Test

A test should be designed under the following requirements

- the false alarm error probability should not exceed a given value α:

$$G_d(\theta_0) \leq \alpha.$$

- the detection error probability should be as small as possible:

$$\max \{G_d(\theta_1)\}.$$

Thus, the design of a test can be formulated as an optimization problem:

$$\begin{array}{ll} \max\limits_{d} & G_d(\theta_1) \\ \text{s. t.} & G_d(\theta_0) \leq \alpha \end{array}$$

Definition. A test d of θ_0 against θ_1 is called uniformly most powerful (UMP) test in case

$$G_d(\theta_1) \geq G_{d'}(\theta_1)$$

for all tests $d' \in \mathbb{D}$, which satisfy $G_d(\theta_0) = G_{d'}(\theta_0)$.

Definition. A test is called a true test with error level α if

$$G_d(\theta_0) \leq \alpha \leq G_d(\theta_1)$$

A true test decides with higher probability for the alternative, if it is correct, than incorrectly for the null hypothesis. This means, a correct detection is more probable than a false alarm.

15.3 Alternative Test

Here we regard the two probability measures $\mathsf{P}(\theta_0), \mathsf{P}(\theta_1)$ respectively $\mathsf{P}_0, \mathsf{P}_1$. The statistical model is $\{\mathbb{X}, \mathbb{F}, \{\mathsf{P}_0, \mathsf{P}_1\}\}$ with probability density functions $f_X(x;\theta_0), f_X(x;\theta_1)$, such that $f_X(x;\theta_0) + f_X(x;\theta_1) > 0,\ \forall x \in \mathbb{X}$.

What is the best test d of P_0 against P_1 for a given error probability α?

Given an observation x, according to the ML-principle one could decide for the alternative in case $f_X(x;\theta_1) > f_X(x;\theta_0)$. The decision threshold is given by the ML-ratio:

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

Neyman-Pearson-Test

Given a statistical model $\{\mathbb{X}, \mathbb{F}, \{\mathsf{P}_0, \mathsf{P}_1\}\}$ with discrete hypothesis and alternative, for any error level $0 < \alpha < 1$:

a) The best test of P_0 against P_1 is

$$d(x) = \begin{cases} 1 & ; \quad R > c \\ \gamma & ; \quad R = c \\ 0 & ; \quad R < c, \end{cases}$$

where $\gamma = \dfrac{\alpha - \mathsf{P}_0(\{R > c\})}{\mathsf{P}_0(\{R = c\})}$ or $\gamma = 0$ if $\mathsf{P}_0(\{R = c\}) = 0$. [1]

b) There exists a test that fully utilizes the error level α.

c) Any Neyman-Pearson (NP) test with error level α is a best test.

---

[1]

$$\begin{aligned}
\mathsf{E}[d(x)|\theta_0] &= 1 \cdot \mathsf{P}_0(\{R > c\}) + \gamma \cdot \mathsf{P}_0(\{R = c\}) \\
&= \begin{cases} \alpha & ; \quad \mathsf{P}_0(\{R = c\}) = 0 \\ \mathsf{P}_0(\{R > c\}) & ; \quad \mathsf{P}_0(\{R = c\}) > 0 \end{cases} + \mathsf{P}_0(\{R = c\}) \cdot \begin{cases} 0 & ; \quad \mathsf{P}_0(\{R = c\}) = 0 \\ \dfrac{\alpha - \mathsf{P}_0(\{R > c\})}{\mathsf{P}_0(\{R = c\})} & ; \quad \mathsf{P}_0(\{R = c\}) > 0 \end{cases} \\
&= \alpha.
\end{aligned}$$

Example 1

$$f_X(x;\theta_0) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}x^2\right)$$
$$f_X(x;\theta_1) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-1)^2\right).$$

The Neyman-Pearson test is

$$d(x) = \begin{cases} 1 & ; \quad R(x) > c \\ 0 & ; \quad R(x) < c. \end{cases}$$

with decision rule

$$R(x) = \frac{f_X(x;\theta_1)}{f_X(x;\theta_0)} = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-1)^2\right)}{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}x^2\right)} = \exp\left(\frac{1}{\sigma^2}\left(x-\frac{1}{2}\right)\right) \gtrless c.$$

Further, we have

$$x \geq x_\alpha \Rightarrow R(x) \geq R(x_\alpha),$$

where $R(x_\alpha) = c$. The α-fractil of $\mathcal{N}(0,\sigma^2)$ is x_α and therefore

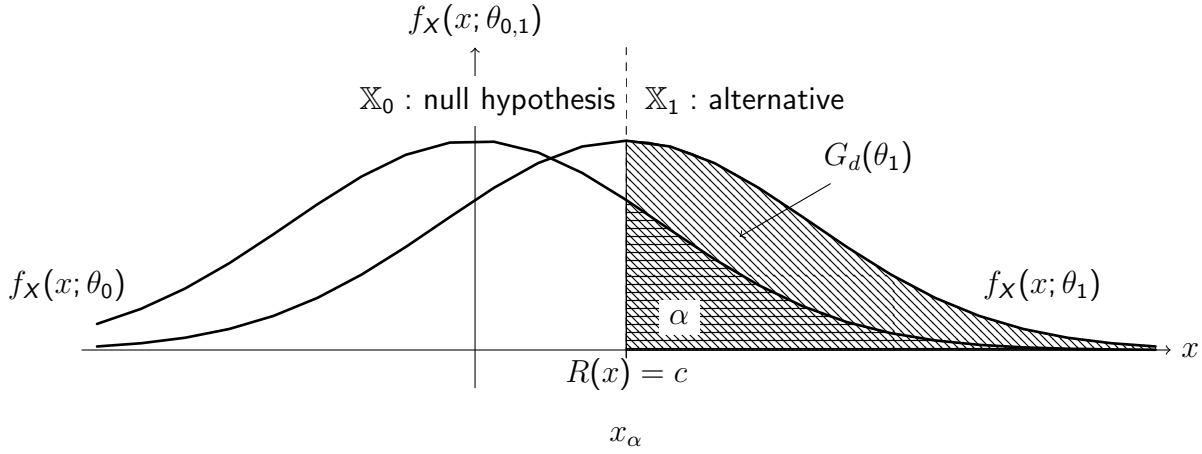
$$\alpha = \int_{x_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}x^2\right)\mathrm{d}x = Q\left(\frac{x_\alpha}{\sigma}\right)$$
$$\Rightarrow x_\alpha = \sigma Q^{-1}(\alpha).$$

As $R(x_\alpha) = c$, we obtain

$$c = \exp\left(\frac{1}{\sigma^2}\left(x_\alpha - \frac{1}{2}\right)\right),$$

and the Neyman-Pearson test can be written as

$$d(x) = \begin{cases} 1 & ; \quad R(x) > \exp\left(\frac{1}{\sigma^2}\left(\sigma Q^{-1}(\alpha) - \frac{1}{2}\right)\right) \\ 0 & ; \quad \text{otherwise.} \end{cases}$$



Example 2

$$f_X(x;\theta_0) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{1}{2\sigma_0^2}x^2\right)$$
$$f_X(x;\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2\sigma_1^2}x^2\right).$$

The decision rule is

$$R(x) = \frac{\frac{1}{\sigma_1}\exp\left(-\frac{1}{2\sigma_1^2}x^2\right)}{\frac{1}{\sigma_0}\exp\left(-\frac{1}{2\sigma_0^2}x^2\right)} \gtrless c$$
$$-\frac{1}{2\sigma_1^2}x^2 + \frac{1}{2\sigma_0^2}x^2 \gtrless \ln c + \ln\frac{\sigma_1}{\sigma_0}$$
$$-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)x^2 \gtrless \ln c + \frac{1}{2}\ln\frac{\sigma_1^2}{\sigma_0^2}.$$

In case $\sigma_1^2 > \sigma_0^2$, we have

$$x^2 \gtrless \frac{2\ln c + \ln\frac{\sigma_1^2}{\sigma_0^2}}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}}.$$

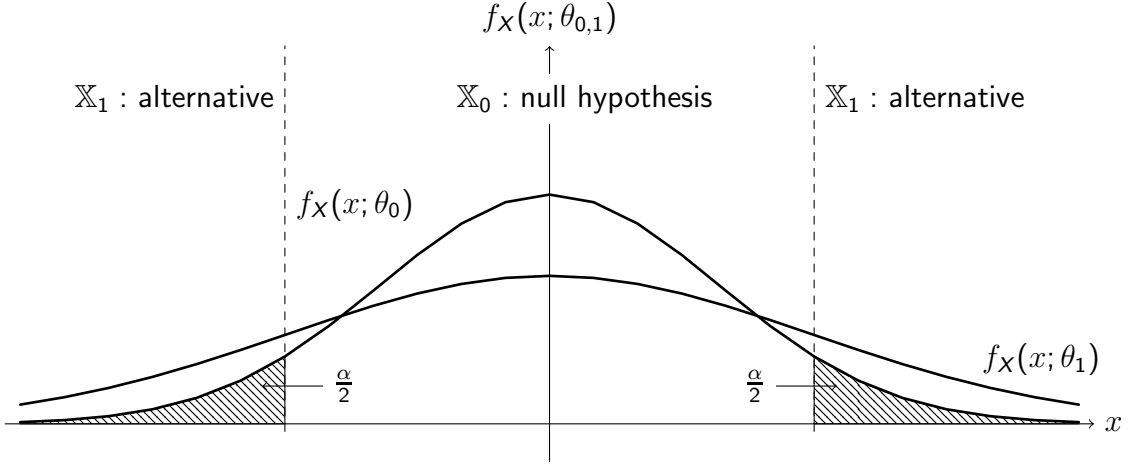
From $R(x_\alpha) = c$, we conclude

$$\ln c = -\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) x_\alpha^2 - \frac{1}{2}\ln\frac{\sigma_1^2}{\sigma_0^2},$$

where x_α is the α-fractil of $\mathcal{N}(0, \sigma_0^2)$. Therefore, the Neyman-Pearson test can be formulated as

$$d(x) = \begin{cases} 1 & ; \quad x^2 > \frac{2\ln c + \ln\frac{\sigma_1^2}{\sigma_0^2}}{\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}} = x_\alpha^2, \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

Note that $R(x)$ is monotone increasing/decreasing in $\mathbb{R}_+/\mathbb{R}_-$.



Example 3

$$\mathsf{B}_{N,\theta_0}(x) = \binom{N}{x}\theta_0^x(1-\theta_0)^{N-x}, \quad x \in \{0,1,\ldots,N\}$$
$$\mathsf{B}_{N,\theta_1}(x) = \binom{N}{x}\theta_1^x(1-\theta_1)^{N-x}, \quad x \in \{0,1,\ldots,N\}.$$

The Neyman-Pearson test subject to an error level α is given as[2]

$$d(x) = \begin{cases} 1 & ; \quad R(x) > c(\alpha) \\ \gamma & ; \quad R(x) = c(\alpha) \\ 0 & ; \quad R(x) < c(\alpha), \end{cases}$$

with the likelihood ratio given as

$$R(x) = \frac{\mathsf{B}_{N,\theta_1}(x)}{\mathsf{B}_{N,\theta_0}(x)} = \frac{\binom{N}{x}\theta_1^x(1-\theta_1)^{N-x}}{\binom{N}{x}\theta_0^x(1-\theta_0)^{N-x}} = \left(\frac{\theta_1}{\theta_0}\right)^x \left(\frac{1-\theta_1}{1-\theta_0}\right)^{N-x} \gtrless c(\alpha), \quad x \in \{0,1,\ldots,N\}.$$

We observe that if $\theta_1 > \theta_0$, $R(x)$ is monotonically increasing in x as $\frac{\theta_1}{\theta_0} > 1$ and $\frac{1-\theta_1}{1-\theta_0} < 1$. Similarly, if $\theta_1 < \theta_0$, $R(x)$ is monotonically decreasing in x. In the following, let $\theta_1 > \theta_0$.

[2]Note that for the Binomial distribution it may occur that $\mathsf{P}(\{R(x) = c(\alpha)\}) > 0$ if $\alpha, \theta_0, \theta_1 \in (0,1)$.

Therefore, we can determine the decision threshold x_{T} of a Neyman-Pearson test subject to an error level of α by looking for an α-fractil of $\mathsf{B}_{N,\theta_0}(x)$, i.e., by considering

$$\mathsf{P}_{\theta_0}(\{X \geq x_{\mathsf{T}}\}) = \sum_{x \geq x_{\mathsf{T}}}^{N} \binom{N}{x} \theta_0^x (1-\theta_0)^{N-x} \gtrsim \alpha.$$

For instance, let $N = 10$ and $\theta_0 = \frac{1}{2}$. It follows that

$$\begin{aligned}
\mathsf{P}_{\theta_0}(\{X = 10\}) &= \binom{10}{10} \left(\frac{1}{2}\right)^{10} = \frac{1}{1024}, \\
\mathsf{P}_{\theta_0}(\{X = 9\}) &= \binom{10}{9} \left(\frac{1}{2}\right)^{10} = \frac{10}{1024}, \\
\mathsf{P}_{\theta_0}(\{X = 8\}) &= \binom{10}{8} \left(\frac{1}{2}\right)^{10} = \frac{45}{1024}.
\end{aligned}$$

Using simple decision thresholds, we can only construct tests that achieve certain error levels. For example, using $x_\mathsf{T} = 8$ and $x_\mathsf{T} = 9$ we achieve error levels of $\mathsf{P}_{\theta_0}(\{X \geq 8\}) = \frac{56}{1024}$ and $\mathsf{P}_{\theta_0}(\{X \geq 9\}) = \frac{11}{1024}$. The Neyman-Pearson test subject to an error level of, for example, $\alpha = \frac{50}{1024}$ is given as

$$d(x) = \begin{cases} 1 & ; \quad x > 8 \\ \gamma & ; \quad x = 8 \\ 0 & ; \quad x < 8, \end{cases}$$

with

$$\gamma = \frac{\alpha - \mathsf{P}_{\theta_0}(\{X > 8\})}{\mathsf{P}_{\theta_0}(\{X = 8\})} = \frac{\frac{50}{1024} - \frac{11}{1024}}{\frac{45}{1024}} = \frac{39}{45}.$$

The decision at the threshold $x_\mathsf{T} = 8$ has to be randomized in order to fully utilize the error level $\alpha = \frac{50}{1024}$ of the Neyman-Pearson test. If we observe $X = 8$, we reject the null hypothesis with probability $\gamma = \frac{39}{45}$ in order to achieve that

$$\begin{aligned} \alpha &= \mathsf{P}_{\theta_0}(\{X > 8\}) + \gamma \mathsf{P}_{\theta_0}(\{X = 8\}) \\ &= \frac{11}{1024} + \frac{39}{45}\frac{45}{1024} = \frac{50}{1024}. \end{aligned}$$

Special Case

The test

$$d(x) = \begin{cases} 1 & ; \quad R(x) > 1 \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

is called ML detector, for instance set $c = 1$ in example 1.

Receiver-Operating-Characteristic (ROC) Graphs

The performance of an NP test can be illustrated by a so-called Receiver-Operating-Characteristics, which plots $G_d(\theta_1)$ as a function of $G_d(\theta_0)$.

ROC for Example 1 ($\sigma^2 = 1$)

$$\mathsf{P}_{\mathsf{FA}} = G_d(\theta_0) = \mathsf{E}\left[d(X)|\theta_0\right] = \int_{x_\alpha}^{\infty} 1 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \mathrm{d}x = Q(x_\alpha) = \alpha$$

$$\Rightarrow x_\alpha = Q^{-1}(\mathsf{P}_{\mathsf{FA}}).$$

$$\begin{aligned} \mathsf{P}_{\mathsf{D}} = G_d(\theta_1) = \mathsf{E}\left[d(X)|\theta_1\right] &= \int_{x_\alpha}^{\infty} 1 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-1)^2\right) \mathrm{d}x \\ &= \int_{x_\alpha - 1}^{\infty} 1 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}x^2\right) \mathrm{d}x = Q(x_\alpha - 1) \end{aligned}$$

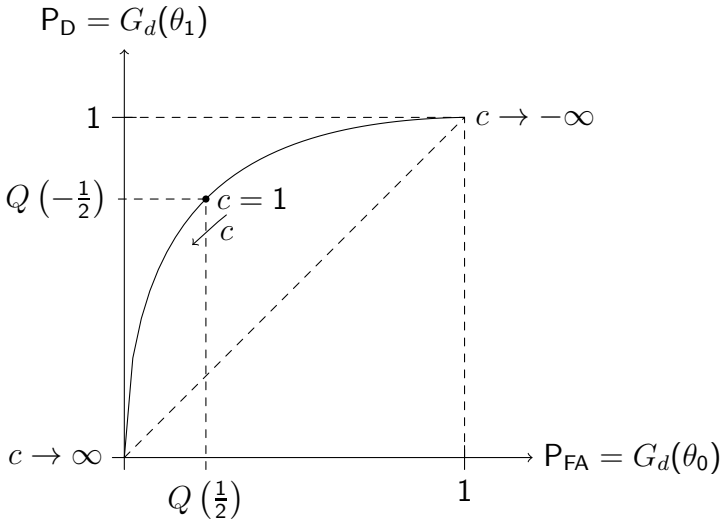
$$\Rightarrow \mathsf{P}_{\mathsf{D}} = Q(Q^{-1}(\mathsf{P}_{\mathsf{FA}}) - 1).$$

For $c = 1$, we have (see page 181):

$$1 = \exp\left(\frac{1}{\sigma^2}\left(x_\alpha - \frac{1}{2}\right)\right) \Rightarrow x_\alpha = \frac{1}{2}$$

$$\Rightarrow Q(x_\alpha) = Q\left(\frac{1}{2}\right)$$

$$\Rightarrow \mathsf{P_D} = Q\left(x_\alpha - 1\right) = Q\left(-\frac{1}{2}\right).$$



15.4 Bayes Test

Similar to Bayes estimation, we regard the case where prior knowledge on possible hypotheses are available, given by so-called *a priori probabilities*. These are $\mathsf{P}(\theta_0)$ and $\mathsf{P}(\theta_1)$, with $\mathsf{P}(\theta_0) + \mathsf{P}(\theta_1) = 1$

Definition. A test $d : \mathbb{X} \to \{0, 1\}$ that, given $\mathsf{P}(\theta_0)$ and $\mathsf{P}(\theta_1)$, minimizes the probability of a wrong decision, is called Bayes test or Bayes detector:

$$d_{\text{Bayes}} = \underset{d}{\operatorname{argmin}} \{\mathsf{P}_\varepsilon\},$$

with $\mathsf{P}_\varepsilon = \mathsf{E}\left[G_d(\theta)(1 - \Pi(\theta)) + (1 - G_d(\theta))\Pi(\theta)\right]$ and

$$\Pi(\theta) = \begin{cases} 0 & ; \quad \theta = \theta_0 \\ 1 & ; \quad \theta = \theta_1. \end{cases}$$

For P_ε we can write:

$$\begin{aligned} \mathsf{P}_\varepsilon &= \mathsf{P}(\theta_0)\left[G_d(\theta_0)(1 - \Pi(\theta_0)) + (1 - G_d(\theta_0))\Pi(\theta_0)\right] + \mathsf{P}(\theta_1)\left[G_d(\theta_1)(1 - \Pi(\theta_1)) + (1 - G_d(\theta_1))\Pi(\theta_1)\right] \\ &= \mathsf{P}(\theta_0)\, G_d(\theta_0) + \mathsf{P}(\theta_1)\,(1 - G_d(\theta_1)). \end{aligned}$$

The decision rule d partitions the observation space $\mathbb{X} = \mathbb{X}_{d,0} \cup \mathbb{X}_{d,1}$, such that $\mathbb{X}_{d,0} \cap \mathbb{X}_{d,1} = \varnothing$.

As $G_d(\theta) = \mathsf{E}[d(X)|\theta] = \int_{d(\mathbb{X})} d\, f_d(d|\theta)\mathrm{d}d = \int_{\mathbb{X}} d(x)\, f_X(x|\theta)\mathrm{d}x$, we have:

$$\begin{aligned}
\mathsf{P}_\varepsilon &= \mathsf{P}(\theta_0)\,\mathsf{E}[d(X)|\theta_0] + \mathsf{P}(\theta_1)\,(1 - \mathsf{E}[d(X)|\theta_1]) \\
&= \mathsf{P}(\theta_0) \int_{\mathbb{X}_{d,1}} 1\, f_X(x|\theta_0)\mathrm{d}x + \mathsf{P}(\theta_1) - \mathsf{P}(\theta_1) \int_{\mathbb{X}_{d,1}} 1\, f_X(x|\theta_1)\mathrm{d}x \\
&= \int_{\mathbb{X}_{d,1}} (\mathsf{P}(\theta_0)\, f_X(x|\theta_0) - \mathsf{P}(\theta_1)\, f_X(x|\theta_1))\,\mathrm{d}x + \mathsf{P}(\theta_1)\,.
\end{aligned}$$

P_ε is minimized, by choosing $\mathbb{X}_{d,1}$ such, that the integrand is negative for every x. We therefore get

$$\mathbb{X}_{d,1} = \{x|\, \mathsf{P}(\theta_0)\, f_X(x|\theta_0) - \mathsf{P}(\theta_1)\, f_X(x|\theta_1) < 0\}.$$

The resulting Bayes detector is

$$d_{\mathsf{Bayes}} = \begin{cases} 1 & ; \quad \dfrac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > \dfrac{\mathsf{P}(\theta_0)}{\mathsf{P}(\theta_1)} \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

For the special case, where $\mathsf{P}(\theta_0) = \mathsf{P}(\theta_1)$ the Bayes test is equivalent to the ML test.

By applying the Bayes rule,

$$\mathsf{P}\left(\theta_i|x\right) = \frac{f_X(x|\theta_i))\,\mathsf{P}\left(\theta_i\right)}{f_X(x)} \text{ (with } f_X(x) \neq 0),$$

on the test condition, we obtain

$$d_{\mathsf{Bayes}} = \begin{cases} 1 & ; \quad \mathsf{P}\left(\theta_1|x\right) > \mathsf{P}\left(\theta_0|x\right), \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

Therefore the Bayes test is also known as *Maximum A Posteriori Detector* or *MAP Detector*.

Bayes Risk Function

In the following we intend to minimize the risk of a wrong decision, instead of the error probability. Therefore the two error types are assigned different weights C_0 and C_1 and the risk function is

$$\begin{aligned}\text{risk}(d) &= \mathsf{E}\left[C_0 G_d(\theta)(1 - \Pi(\theta)) + C_1(1 - G_d(\theta))\Pi(\theta)\right] \\ &= C_0 G_d(\theta_0)\,\mathsf{P}\left(\theta_0\right) + C_1(1 - G_d(\theta_1))\,\mathsf{P}\left(\theta_1\right).\end{aligned}$$

The resulting decision rule is

$$d_{\text{Bayes}} = \begin{cases} 1 & ; \quad \dfrac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > \dfrac{C_0\,\mathsf{P}\left(\theta_0\right)}{C_1\,\mathsf{P}\left(\theta_1\right)}, \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

While for a transmitted bit, the risk of a wrong decision will have the same weight, this is different for scenarios where the alternative corresponds to a hazardous incident, in this case $\Rightarrow C_1 \gg C_0$!

15.5 Bayes Test for Multiple Hypothesis

As a preparation we start with a reformulation of the alternative test:

$$\begin{aligned}
\text{Risk}(d) &= C_0 \,\mathsf{E}\left[d(X)|\theta_0\right] \mathsf{P}\left(\theta_0\right) + C_1 \,\mathsf{E}\left[1 - d(X)|\theta_1\right] \mathsf{P}\left(\theta_1\right) \\
&= \mathsf{E}\left[C_0 d(X)|\theta_0\right] \mathsf{P}\left(\theta_0\right) + \mathsf{E}\left[C_1(1 - d(X))|\theta_1\right] \mathsf{P}\left(\theta_1\right) \\
&= \mathsf{E}\left[L(d(X), \theta_0)\right] \mathsf{P}\left(\theta_0\right) + \mathsf{E}\left[L(d(X), \theta_1)\right] \mathsf{P}\left(\theta_1\right), \\
&= \mathsf{E}\left[\mathsf{E}\left[L(d(X), \theta)|\theta\right]\right] \\
&= \mathsf{E}\left[L(d(X), \theta)\right],
\end{aligned}$$

where $L(d(x), \theta)$ is a loss function defined as

$$L : \{0, 1\} \times \{\theta_0, \theta_1\} \to \mathbb{R}_+ : L(d(x), \theta) \mapsto \begin{cases} C_0 & ; \quad \text{for a false alarm, } d(x) = 1, \text{ but } \theta = \theta_0, \\ C_1 & ; \quad \text{for a detection error, } d(x) = 0, \text{ but } \theta = \theta_1, \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

In the following, we generalize this result to multiple hypothesis, where $\Theta = \{\theta_1, \ldots, \theta_K\}$, and the loss function is

$$L : \{1, \ldots, K\} \times \Theta \to \mathbb{R}_+ : L(d(x), \theta_k) \mapsto \begin{cases} C_k & ; \quad d(x) \neq \theta_k, \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

The expected risk is:

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

The risk is minimized, if by the choice of d the term $\mathsf{E}\left[L(d(x),\theta)|x\right]$ is minimized for all $x \in \mathbb{X}$, that is $\forall x \in \mathbb{X}$ and $d(x) = k' \in \{1,\ldots,K\}$:

$$d : x \mapsto \underset{k' \in \{1,\ldots,K\}}{\mathsf{argmin}} \{\mathsf{E}\left[L(k',\theta)|x\right]\}.$$

As we have,

$$\begin{aligned}
\mathsf{E}\left[L(k',\theta)|x\right] &= \sum_{k \in \{1,\ldots,K\}} P(\theta_k|x) \underbrace{L(k',\theta_k)}_{\in\{C_k,0\}} \\
&= \sum_{k \in \{1,\ldots,K\}\setminus k'} P(\theta_k|x) C_k \\
&= \sum_{k \in \{1,\ldots,K\}\setminus k'} P(\theta_k|x) C_k + P(\theta_{k'}|x) C_{k'} - P(\theta_{k'}|x) C_{k'} \\
&= \mathsf{const} - P(\theta_{k'}|x) C_{k'},
\end{aligned}$$

the decision rule becomes

$$d : x \mapsto \underset{k' \in \{1,\ldots,K\}}{\mathsf{argmin}} \{\mathsf{E}\left[L(k',\theta)|x\right]\} = \underset{k' \in \{1,\ldots,K\}}{\mathsf{argmax}} \{P(\theta_{k'}|x) C_{k'}\}.$$

For the special case $C_k = 1,\ \forall \theta_k \in \Theta$, the loss function corresponds to the error probability and the Bayes test is

$$d : x \mapsto \underset{k' \in \{1,\ldots,K\}}{\mathsf{argmax}} \{P(\theta_{k'}|x)\}.$$

This test is known as the *maximum a posteriori (MAP) detector*.

Example

$$f_{X|\Theta}(x|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\theta_k)^2\right),$$

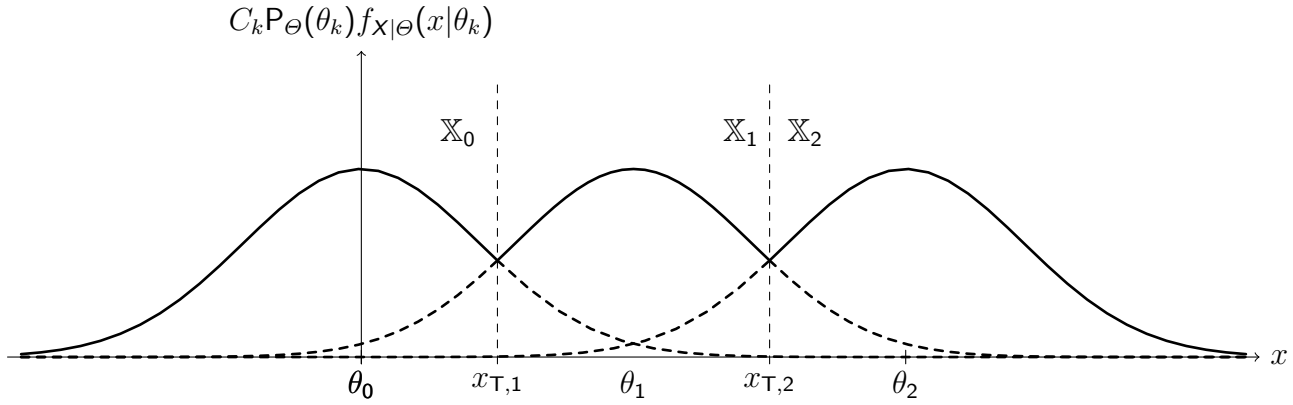
with $\theta_0 < \theta_1 < \theta_2$, a prior $\mathsf{P}_\Theta(\theta_k)$ for Θ, and costs C_k. The Bayes test is given as

$$d(x) = \underset{k\in\{0,1,2\}}{\text{argmax}} \left\{C_k \mathsf{P}_{\Theta|X}(\theta_k|x)\right\} = \underset{k\in\{0,1,2\}}{\text{argmax}} \left\{C_k \mathsf{P}_\Theta(\theta_k) f_{X|\Theta}(x|\theta_k)\right\}.$$

The Bayes test $d(x)$ decomposes the observation space $\mathbb{X}$ into three subsets $\mathbb{X}_k$ with

$$d(x) = \begin{cases} 0 & ; \quad x \in \mathbb{X}_0 \\ 1 & ; \quad x \in \mathbb{X}_1 \\ 2 & ; \quad x \in \mathbb{X}_2. \end{cases}$$

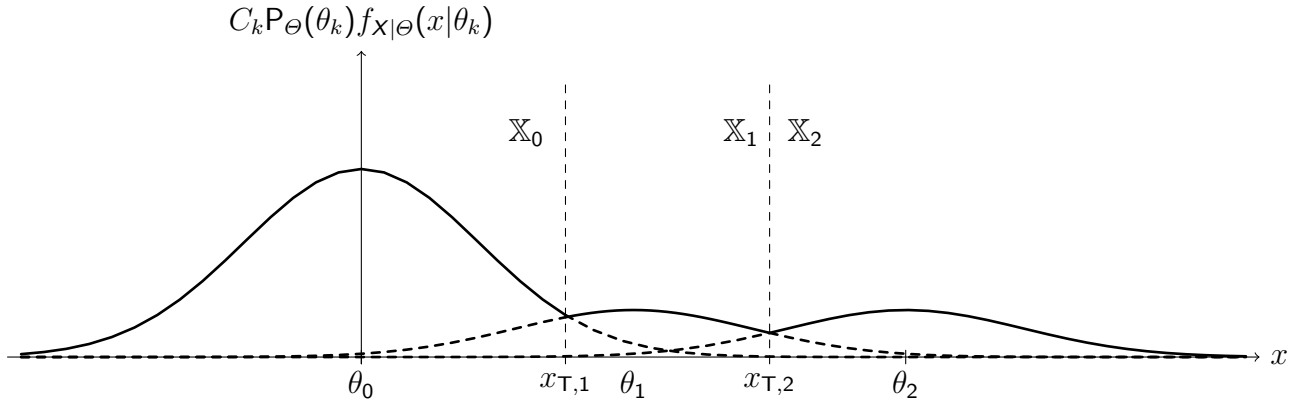
First, consider $\mathsf{P}_\Theta(\theta_k) = \frac{1}{3}$ and $C_k = 1\ \forall k$.



For the case of a uniform prior and equal costs, the Bayes test equals a Maximum-Likelihood test. The test statistic is given as

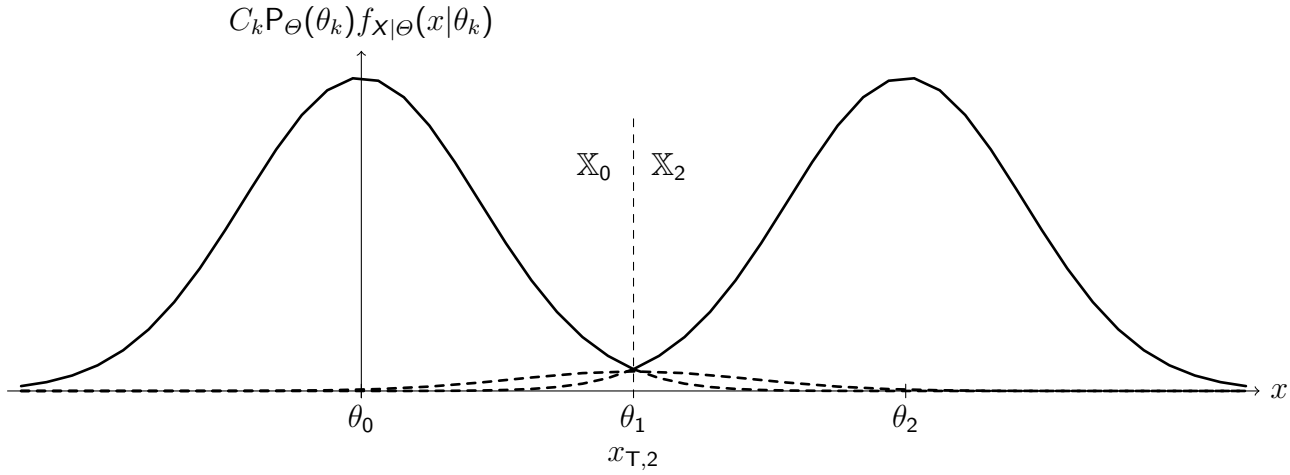
$$d(x) = \begin{cases} 0 & ; \quad x \leq x_{\mathsf{T},1} \\ 1 & ; \quad x_{\mathsf{T},1} < x < x_{\mathsf{T},2} \\ 2 & ; \quad x \geq x_{\mathsf{T},2}. \end{cases}$$

Now, consider $\mathsf{P}_\Theta(\theta_0) = \frac{2}{3}$, $\mathsf{P}_\Theta(\theta_1) = \mathsf{P}_\Theta(\theta_2) = \frac{1}{6}$, and $C_k = 1\ \forall k$.



As $\mathsf{P}_\Theta(\theta_0) > \mathsf{P}_\Theta(\theta_1) = \mathsf{P}_\Theta(\theta_3)$, the threshold $x_{\mathsf{T},1}$ is shifted towards θ_1.

Finally, assume that $\mathsf{P}_{\Theta}(\theta_0) = \mathsf{P}_{\Theta}(\theta_2) = \frac{48}{100}$, $\mathsf{P}_{\Theta}(\theta_1) = \frac{4}{100}$, and $C_k = 1; \forall k$.



Although $\mathsf{P}_{\Theta}(\theta_1) > 0$, the Bayes test is given as

$$d(x) = \begin{cases} 0 & ; \quad x \in \mathbb{X}_0 \\ 2 & ; \quad x \in \mathbb{X}_2 \end{cases}$$

as we have that $\mathbb{X}_1 = \emptyset$.

15.6 Linear Alternative Tests

In the following, we consider the special case of *linear alternative tests*, i.e., tests with

$$d : \mathbb{X} \to \mathbb{R}, \boldsymbol{x} \mapsto \begin{cases} 1 & ; \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} - w_0 > 0 \\ 0 & ; \text{otherwise.} \end{cases} \tag{15.1}$$

Motivation: Gaussian $f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$

Consider the ML test (Neyman-Pearson test with $c = 1$) for Gaussian PDFs, i.e., we have

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \underbrace{\boldsymbol{\mu}_k, \boldsymbol{C}_k}_{\theta_k}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{C}_k)}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{C}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right) \tag{15.2}$$

with $\theta_0 = \{\boldsymbol{\mu}_0, \boldsymbol{C}_0\}$ and $\theta_1 = \{\boldsymbol{\mu}_1, \boldsymbol{C}_1\}$. The likelihood ratio $R(\boldsymbol{x})$ is given as

$$R(\boldsymbol{x}) = \frac{\sqrt{\det(2\pi\boldsymbol{C}_0)}}{\sqrt{\det(2\pi\boldsymbol{C}_1)}} \cdot \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{C}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)}{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{C}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\right)} \gtrless 1. \tag{15.3}$$

Using the log-likelihood ratio $\log R(\boldsymbol{x})$, the separating surface in $\mathbb{X} = \mathbb{R}^N$ which decomposes $\mathbb{X}$ into $\mathbb{X}_0$ and $\mathbb{X}_1$ can be described as the set of all $\boldsymbol{x} \in \mathbb{X}$ for which

$$\log R(\boldsymbol{x}) = \frac{1}{2}\ln\left(\frac{\det(\boldsymbol{C}_0)}{\det(\boldsymbol{C}_1)}\right) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{C}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{C}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) = 0. \tag{15.4}$$

Motivation: $\log R(\boldsymbol{x})$ as an affine function of $\boldsymbol{x}$

In the special case of $\boldsymbol{C_0} = \boldsymbol{C_1} = \boldsymbol{C}$, the quadratic terms in $\log R(\boldsymbol{x})$ cancel out, and we have

$$\log R(\boldsymbol{x}) = \frac{1}{2} \ln \left(\frac{\det(\boldsymbol{C}_0)}{\det(\boldsymbol{C}_1)} \right) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_0)^\mathsf{T} \boldsymbol{C}_0^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_0) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\mathsf{T} \boldsymbol{C}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) \tag{15.5}$$

$$= \boldsymbol{w}^\mathsf{T} \boldsymbol{x} - w_0 \gtrless 0 \tag{15.6}$$

where the normal vector of the separating hyperplane is given as

$$\boldsymbol{w}^\mathsf{T} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\mathsf{T} \boldsymbol{C}^{-1} \tag{15.7}$$

and the constant translation w_0 with respect to the origin is given as

$$w_0 = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\mathsf{T} \boldsymbol{C}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)}{2}. \tag{15.8}$$

The ML estimator reduces to the affine decision rule given as

$$d : \mathbb{X} \to \mathbb{R}, \boldsymbol{x} \mapsto \begin{cases} 1 & ; \boldsymbol{w}^\mathsf{T} \boldsymbol{x} - w_0 > 0 \\ 0 & ; \text{otherwise.} \end{cases} \tag{15.9}$$

Motivation: $\log R(\boldsymbol{x})$ as a linear function of $\boldsymbol{x}$

If in addition to $\boldsymbol{C_0} = \boldsymbol{C_1}$, we have that $\boldsymbol{\mu}_0 = -\boldsymbol{\mu}_1$, the log-likelihood ratio simplifies further to

$$\log R(\boldsymbol{x}) = \boldsymbol{w}^\mathsf{T}\boldsymbol{x} \gtrless 0 \tag{15.10}$$

as from $\boldsymbol{\mu}_0 = -\boldsymbol{\mu}_1$ it follows that $w_0 = 0$. Therefore, the ML estimator reduces to the linear decision rule

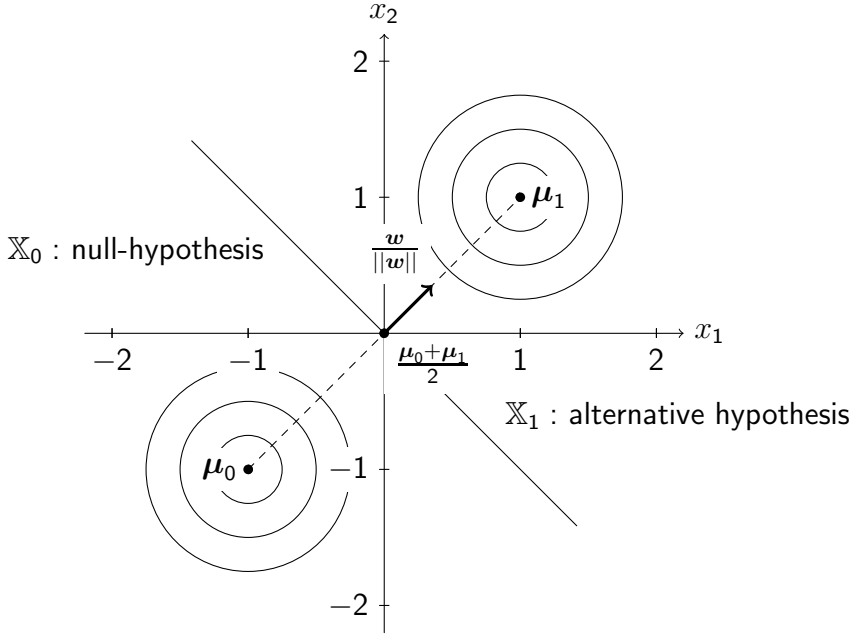
$$d : \mathbb{X} \to \mathbb{R}, \boldsymbol{x} \mapsto \begin{cases} 1 & ; \boldsymbol{w}^\mathsf{T}\boldsymbol{x} > 0 \\ 0 & ; \text{otherwise.} \end{cases} \tag{15.11}$$

For Gaussian $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{C}_k)$ with $\theta_0 = \{\boldsymbol{\mu}_0, \boldsymbol{C}_0\}$ and $\theta_1 = \{\boldsymbol{\mu}_1, \boldsymbol{C}_1\}$, it follows that

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

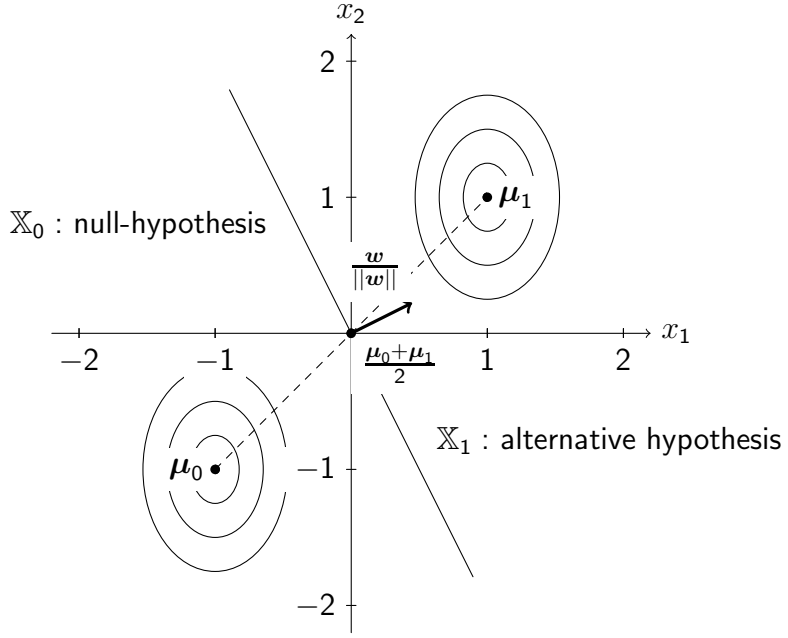
Example 1:

$\boldsymbol{\mu}_0 = (-1,-1)^\mathsf{T}, \boldsymbol{\mu}_1 = (1,1)^\mathsf{T}, \boldsymbol{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{w}^\mathsf{T} = (2,2)$



Example 2:

$\boldsymbol{\mu}_0 = (-1,-1)^\mathsf{T}, \boldsymbol{\mu}_1 = (1,1)^\mathsf{T}, \boldsymbol{C} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \boldsymbol{w}^\mathsf{T} = (2,1)$



Statistical Properties of $T(\boldsymbol{X}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}$

Consider an alternative test with Gaussian $f_{\boldsymbol{X}}(\boldsymbol{x}; \theta_k)$, with $\theta_0 = \{\boldsymbol{\mu}_0, \boldsymbol{C}_0\}$, and $\theta_1 = \{\boldsymbol{\mu}_1, \boldsymbol{C}_1\}$ where $\boldsymbol{C} = \boldsymbol{C}_0 = \boldsymbol{C}_1$. In the following, we will discuss a test statistic given as

$$T(\boldsymbol{X}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{X} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}. \tag{15.12}$$

As $\boldsymbol{X}$ is Gaussian, $T(\boldsymbol{X})$ is Gaussian as well (linear transformation of Gaussian RV is Gaussian). Thus, the distribution of $T(\boldsymbol{X})$ is determined by

$$\mathsf{E}\left[T(\boldsymbol{X})\right] = \mathsf{E}\left[\boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}\right] = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_k = \mu_{T,k}\,,\ k = 0, 1, \tag{15.13}$$

$$\mathsf{Var}\left[T(\boldsymbol{X})\right] = \boldsymbol{w}^{\mathsf{T}}\,\mathsf{Var}\left[\boldsymbol{X}\right]\boldsymbol{w} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = \sigma_T^2. \tag{15.14}$$

The PDF of $T = T(\boldsymbol{X})$ is given as

$$f_{T,k}(t) = \frac{1}{\sqrt{2\pi}\sigma_T}\exp\left(-\frac{1}{2\sigma_T^2}(t - \mu_{T,k})^2\right). \tag{15.15}$$

The ML test can be expressed as

$$
\begin{aligned}
\ln(R(T)) = \frac{f_{T,1}(t)}{f_{T,0}(t)} = \ln\left(\frac{\exp\left(-\frac{1}{2\sigma_T^2}(T-\mu_{T,1})^2\right)}{\exp\left(-\frac{1}{2\sigma_T^2}(T-\mu_{T,0})^2\right)}\right) &\gtrless 0 \\
\Rightarrow \quad (\mu_{T,1}-\mu_{T,0})T - \frac{(\mu_{T,1}^2-\mu_{T,0}^2)}{2} &\gtrless 0 \\
\Rightarrow \quad T - \frac{\mu_{T,1}+\mu_{T,0}}{2} &\gtrless 0 \\
\Rightarrow \quad T - w_0 &\gtrless 0 \\
\Rightarrow \quad T(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} &\gtrless w_0
\end{aligned}
$$

Note that that $w_0 = \frac{\mu_{T,1}+\mu_{T,0}}{2} = \frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_0)}{2}$ as $\mu_{T,k} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_k$ and $\boldsymbol{w}^{\mathsf{T}} = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{C}^{-1}$.

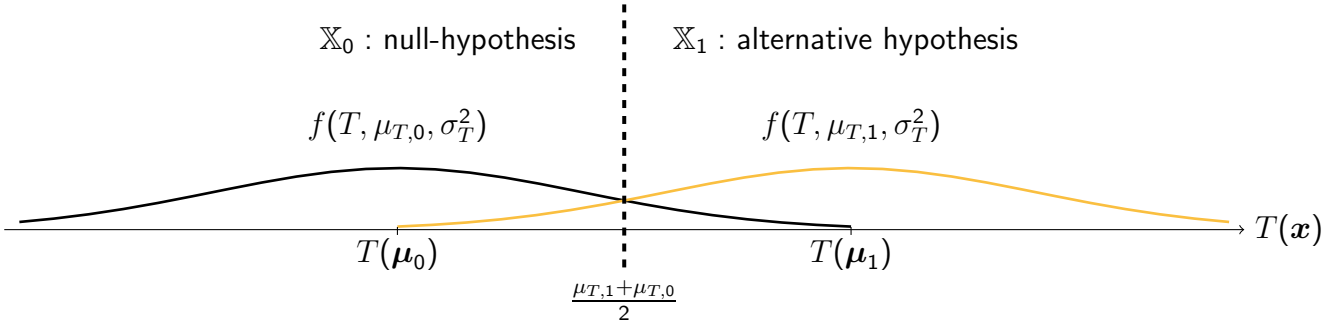


Fig. 15.1: Illustration of the decision rule $T(\boldsymbol{x}) \gtrless \frac{\mu_{T,1}+\mu_{T,0}}{2}$ for Gaussian PDFs with $\boldsymbol{\mu}_0 = -\boldsymbol{\mu}_1$ and $\boldsymbol{C_0} = \boldsymbol{C_1}$.

15.7 Sufficient Statistics

An important property of a test statistic is whether it is *sufficient* in order to decide for a parameter $\theta \in \Theta$ or not. In this context, sufficiency means that no other test statistic, i.e., function of the observations, contains additional information about the parameter to be estimated.

Definition: Sufficient Statistic

Consider a statistical model $\{\mathbb{X}, \mathbb{F}, \mathsf{P}_\theta; \theta \in \Theta\}$. Let $\boldsymbol{X}$ be a random sample drawn from $f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$. If for a test statistic $T(\boldsymbol{X})$, $f_{\boldsymbol{X}|T}(\boldsymbol{x}|T(\boldsymbol{x}) = t)$ is independent of θ for all $\theta \in \Theta$, i.e., if

$$f_{\boldsymbol{X}|T}(\boldsymbol{x}|T(\boldsymbol{x}) = t, \theta) = f_{\boldsymbol{X}|T}(\boldsymbol{x}|T(\boldsymbol{x}) = t), \tag{15.16}$$

the test statistic $T(\boldsymbol{X})$ is a SUFFICIENT STATISTIC for the underlying parameter θ, or more precisely, for the family of probability distributions P_θ which are parametrized by θ.

Example:

Transmission of $d = \pm 1$ from a source to a sink:

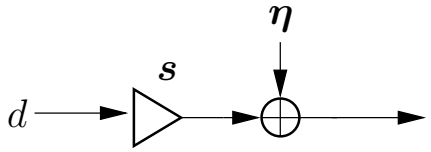


Fig. 15.2: System model.

The receive signal at the sink is given as

$$\boldsymbol{x} = d\boldsymbol{s} + \boldsymbol{\eta}. \tag{15.17}$$

As $d \in \{-1, +1\}$ and $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, the ML test is given as:

$$R(\boldsymbol{x}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{s})^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{s})\right)}{\exp\left(-\frac{1}{2}(\boldsymbol{x}+\boldsymbol{s})^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}+\boldsymbol{s})\right)} \gtrless 1 \tag{15.18}$$

$$\Leftrightarrow \quad \log R(\boldsymbol{x}) = \boldsymbol{s}^{\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{x} \gtrless 0 \tag{15.19}$$

Using a *Whitening Matched Filter* $\boldsymbol{g}^{\mathsf{T}} = \boldsymbol{s}^{\mathsf{T}}\boldsymbol{C}^{-1}$, the ML test is given as[3]:

$$d : \boldsymbol{x} \mapsto \begin{cases} +1 & ; \boldsymbol{g}^{\mathsf{T}}\boldsymbol{x} > 0 \\ -1 & ; \text{otherwise.} \end{cases} \tag{15.20}$$

[3]Here, we define the test as a function which maps to the transmit alphabet $\{-1, +1\}$ directly.

Remark: The test statistic $G(\boldsymbol{x}) = \boldsymbol{g}^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{s}^{\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{x}$ is a sufficient statistic, i.e., for $G(\boldsymbol{x}) = c$, it follows that $f_{\boldsymbol{X}}(\boldsymbol{x}|G(\boldsymbol{x}) = c)$ is independent of d.

Sketch of Proof: Assume that $G(\boldsymbol{x}) = 0$, i.e., that

$$\boldsymbol{x} \in \{\boldsymbol{x}|G(\boldsymbol{x}) = 0\} = \{\boldsymbol{x}|\boldsymbol{g}^{\mathsf{T}}\boldsymbol{x} = 0\} = \{\boldsymbol{x}|\boldsymbol{g}^{\perp}t,\, t \in \mathbb{R}\}, \tag{15.21}$$

where $\boldsymbol{g}^{\perp}$ such that $\boldsymbol{g}^{\mathsf{T}}\boldsymbol{g}^{\perp} = 0$. It remains to show that $f(\boldsymbol{x}|G(\boldsymbol{x}) = 0; d)$ is independent of d:

$$\begin{aligned} f(\boldsymbol{x} \mid \boldsymbol{x} = \boldsymbol{g}^{\perp}t; d) &\propto \exp\left[-\frac{1}{2}(\boldsymbol{g}^{\perp}t - \boldsymbol{s}d)^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{g}^{\perp}t - \boldsymbol{s}d)\right] \\ &\propto \exp\left[-\frac{1}{2}\boldsymbol{g}^{\perp,\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{g}^{\perp}t^2 + d\underbrace{\boldsymbol{s}^{\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{g}^{\perp}}_{=0}t\right] \\ &\propto \exp\left[-\frac{1}{2}\boldsymbol{g}^{\perp,\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{g}^{\perp}t^2\right] = f(\boldsymbol{x} \mid \boldsymbol{x} = \boldsymbol{g}^{\perp}t) = f_t(t). \end{aligned}$$

Appendix

A. Probability Theory and Stochastics

A.1 Probability Space

A Probability Space,

$$(\Omega, \mathbb{F}, \mathsf{P}), \tag{A.1}$$

represents a mathematical model consisting of an

- Observation Space Ω:
  the nonempty set of potential outputs/observations of an experiment,
- Sigma Algebra $\mathbb{F}$:
  a set of subsets of potential outputs/observations of an experiment,
- Probability Measure P:
  a function which maps each element of $A \in \mathbb{F}$ to the interval $[0, 1]$.

Definition. A SIGMA ALGEBRA $\mathbb{F}$ is a set of subsets (EVENTS) of Ω with

$$\Omega \in \mathbb{F}, \tag{A.2}$$

$$A \in \mathbb{F} \quad \Rightarrow \quad A^{\mathsf{c}} \in \mathbb{F}, \tag{A.3}$$

$$A_1, \ldots, A_k \in \mathbb{F} \quad \Rightarrow \quad \bigcup_{i=1}^{k} A_i \in \mathbb{F}. \tag{A.4}$$

Definition. A PROBABILITY MEASURE P maps $A \in \mathbb{F}$ to $\mathsf{P}(A)$, the PROBABILITY of the EVENT A, with

$$\text{Nonnegativity :} \qquad \mathsf{P}(A) \geq 0, \tag{A.5}$$

$$\text{Normation :} \qquad \mathsf{P}(\Omega) = 1, \tag{A.6}$$

$$\text{Additivity :} \qquad \mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i), \quad \text{wenn } A_i \cap A_j = \varnothing, \forall i \neq j. \tag{A.7}$$

Consequently, the following general properties of probability measures are obtained:

$$\mathsf{P}\left(A^{\mathsf{c}}\right) = 1 - \mathsf{P}\left(A\right), \tag{A.8}$$

$$\mathsf{P}\left(\varnothing\right) = 0, \tag{A.9}$$

$$\mathsf{P}\left(A \setminus B\right) = \mathsf{P}\left(A \cap B^{\mathsf{c}}\right) = \mathsf{P}\left(A\right) - \mathsf{P}\left(A \cap B\right), \tag{A.10}$$

$$\mathsf{P}\left(A \cup B\right) = \mathsf{P}\left(A\right) + \mathsf{P}\left(B\right) - \mathsf{P}\left(A \cap B\right), \tag{A.11}$$

$$A \subset B \Rightarrow \mathsf{P}\left(A\right) \leq \mathsf{P}\left(B\right), \tag{A.12}$$

$$\mathsf{P}\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \mathsf{P}\left(A_i\right). \tag{A.13}$$

A.2 Conditional Probability and Stochastical Independence

Conditional Probability

The probability of an event changes when we are informed about the realization of another event within the same probability space.

Definition. Given $\mathsf{P}(A) > 0$ for $A \in \mathbb{F}$, the CONDITIONAL PROBABILITY of B conditioned by A is defined as

$$\mathsf{P}(B \mid A) \stackrel{\triangle}{=} \mathsf{P}_A(B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(A)}. \tag{A.14}$$

Bayes Theorem

Given a partition of Ω in disjoint sets $B_i, i \in I$, such that

$$\bigcup_{i \in I} B_i = \Omega, \tag{A.15}$$

$$B_i \cap B_j = \varnothing, \forall i \neq j, \tag{A.16}$$

then for any $A \in \mathbb{F}$ with $\mathsf{P}(A) > 0$:

$$\text{Theorem of total probability}: \quad \mathsf{P}(A) = \sum_{i \in I} \mathsf{P}(A \mid B_i)\,\mathsf{P}(B_i), \tag{A.17}$$

$$\text{Bayes Theorem}: \quad \mathsf{P}(B_j \mid A) = \frac{\mathsf{P}(A \mid B_j)\,\mathsf{P}(B_j)}{\sum_{i \in I} \mathsf{P}(A \mid B_i)\,\mathsf{P}(B_i)}. \tag{A.18}$$

Special Case. Given $A, B \subset \Omega$ and $\Omega = B \cup B^{\mathsf{c}}$, the Bayes Theorem is given by

$$\mathsf{P}(B \mid A) = \frac{\mathsf{P}(A \mid B)\,\mathsf{P}(B)}{\mathsf{P}(A)} = \frac{\mathsf{P}(A \mid B)\,\mathsf{P}(B)}{\mathsf{P}(A \mid B)\,\mathsf{P}(B) + \mathsf{P}(A \mid B^{\mathsf{c}})\,(1 - \mathsf{P}(B))}. \tag{A.19}$$

Illustration

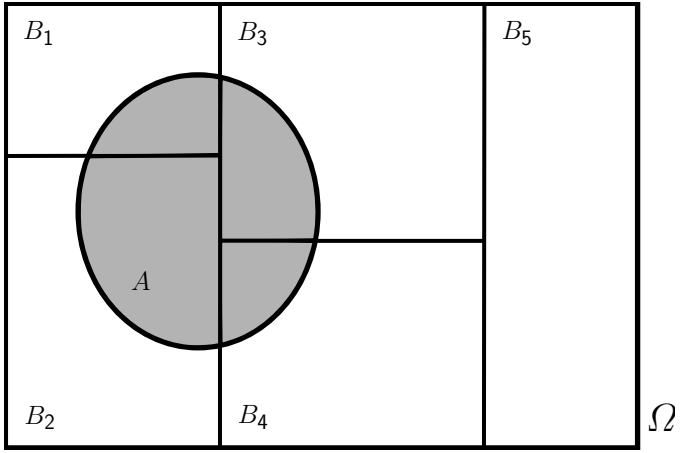


Fig. A.1: Partition of Ω in $B_1, \ldots, B_5$

$$\mathsf{P}(A) = \sum_{i=1}^{4} \mathsf{P}(A \mid B_i)\,\mathsf{P}(B_i) + 0 \cdot \mathsf{P}(B_5)\,, \tag{A.20}$$

$$\mathsf{P}(B_2 \mid A) = \frac{\mathsf{P}(A \mid B_2)\,\mathsf{P}(B_2)}{\sum_{i=1}^{4} \mathsf{P}(A \mid B_i)\,\mathsf{P}(B_i)}. \tag{A.21}$$

Stochastical Independence

A und B are stochastically independent if the probability measure for B is not changed by any knowledge about the realization of A, and vice verca, i.e., given $\mathsf{P}(A) > 0$,

$$\mathsf{P}(B \mid A) = \mathsf{P}(B). \tag{A.22}$$

Otherwise, with

$$\mathsf{P}(B \mid A) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(A)} \tag{A.23}$$

we obtain the following

Definition. Given $(\Omega, \mathbb{F}, \mathsf{P})$ is a probability space, then $A, B \in \mathbb{F}$ are Stochastically Independent with respect to P, if

$$\mathsf{P}(A \cap B) = \mathsf{P}(A)\,\mathsf{P}(B). \tag{A.24}$$

A.3 Random Variables and Distributions

Definition. Given a probability space $(\Omega, \mathbb{F}, \mathsf{P})$ and a measure space $(\Omega', \mathbb{F}')$, we call the mapping $X : \Omega \to \Omega'$ a RANDOM VARIABLE, if and only if for each $A' \in \mathbb{F}'$ there exists an element $A \in \mathbb{F}$, such that $A = \{\omega \in \Omega \mid X(\omega) \in A'\} \in \mathbb{F}$ or — in shorter notation — $A = \{X \in A'\} \in \mathbb{F}$.

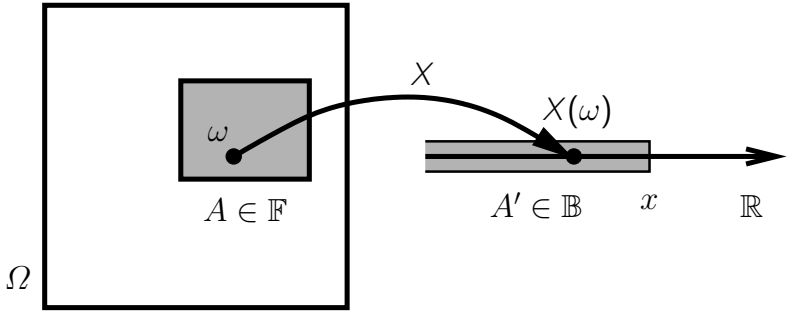


Fig. A.2: Real-valued random variable: $x : \Omega \to \mathbb{R}$

Distribution of Real-Valued Random Variables

Definition. Given a real-valued random variable X on $(\Omega, \mathbb{F}, \mathsf{P})$, the function $F_X : \mathbb{R} \to [0, 1]$, with

$$F_X(x) = \mathsf{P}\left(\{X \leq x\}\right), \tag{A.25}$$

defines the Cumulative Distribution Function (CDF) of X.

Properties of a CDF $F_X : \mathbb{R} \to [0, 1]$:

- $F_X(x)$ is monotonically increasing.
- $\lim_{h \to 0} F_X(x + h) = F_X(x), \quad \forall x \in \mathbb{R}$.
- $\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$.

Definition. We call a random variable X CONTINUOUS if its CDF F_X can be determined by means of

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi)\,\mathrm{d}\xi, \tag{A.26}$$

with $f_X : \mathbb{R} \to [0, \infty[$, which is the PROBABILITY DENSITY FUNCTION (PDF) of X.

If F_X is continous and can be differentiated almost everywhere, then the random variable X is continuous and

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}. \tag{A.27}$$

Examples

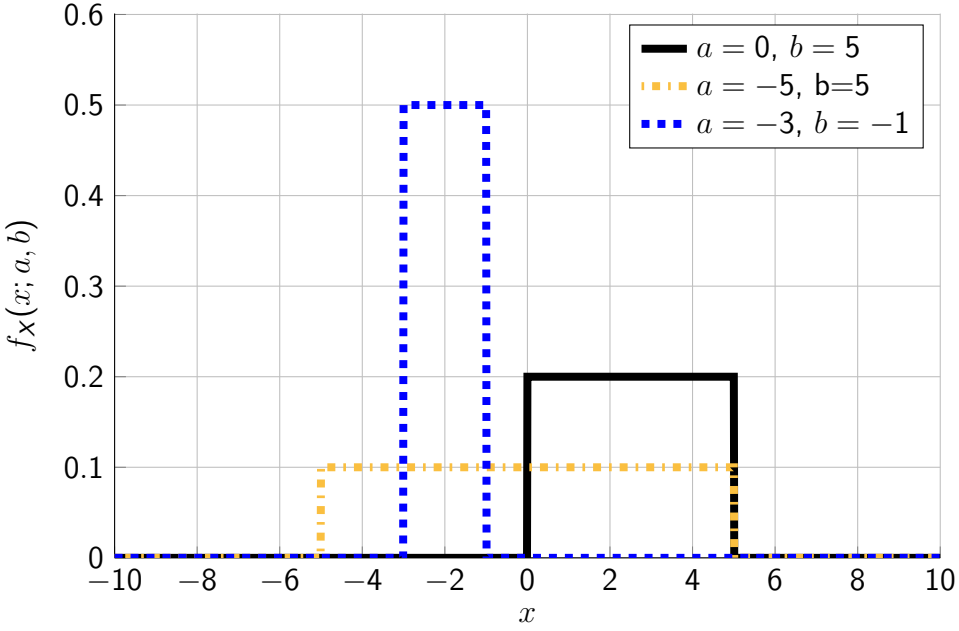


Fig. A.3: Uniform PDF: $f_X(x) = \frac{1}{b-a},\ a \leq x \leq b$

Examples (Cont'd)

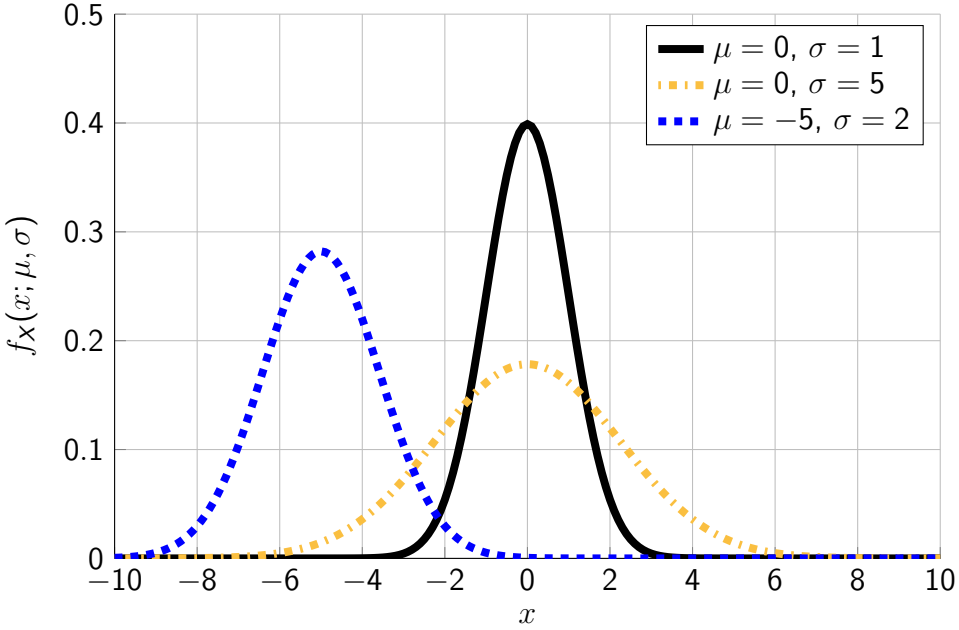


Fig. A.4: Gaussian PDF: $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$

Examples (Cont'd)

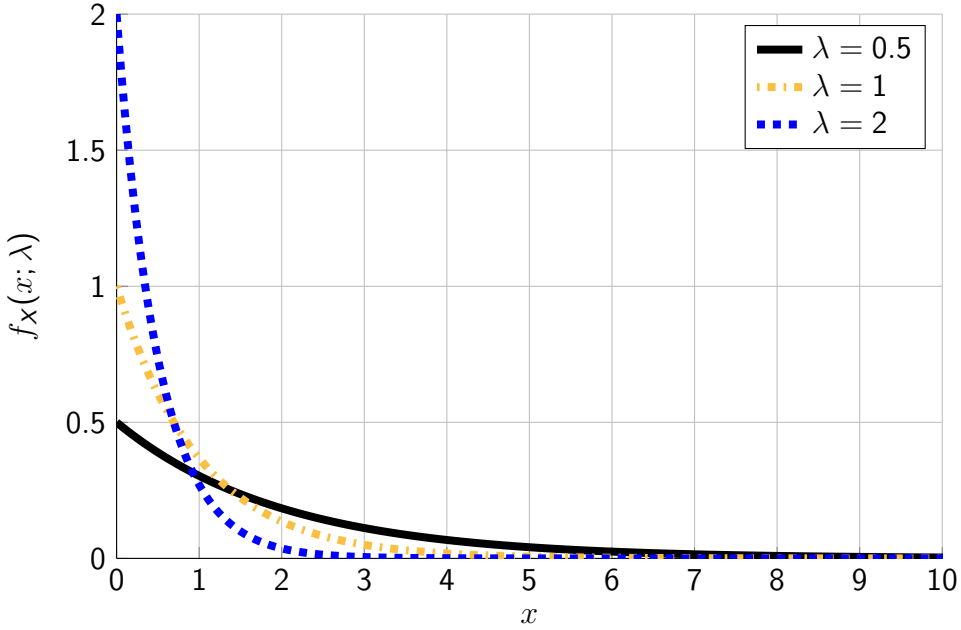


Fig. A.5: Exponential PDF: $f_X(x) = \lambda\, \mathrm{e}^{-\lambda x}, \quad x \geq 0$

Multidimensional Distributions

Definition. Given a n-dimensional real-valued random variable (random vector) $\boldsymbol{X} = [X_1, \ldots, X_n]^{\mathsf{T}}$ on the probability $(\Omega, \mathbb{F}, \mathsf{P})$, then the JOINT CDF $F_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0, 1]$ is denoted as

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathsf{P}\left(\{\boldsymbol{X} \leq \boldsymbol{x}\}\right) \stackrel{\triangle}{=} \mathsf{P}\left(\{X_1 \leq x_1, \ldots, X_n \leq x_n\}\right). \tag{A.28}$$

The respective JOINT PDF is defined as

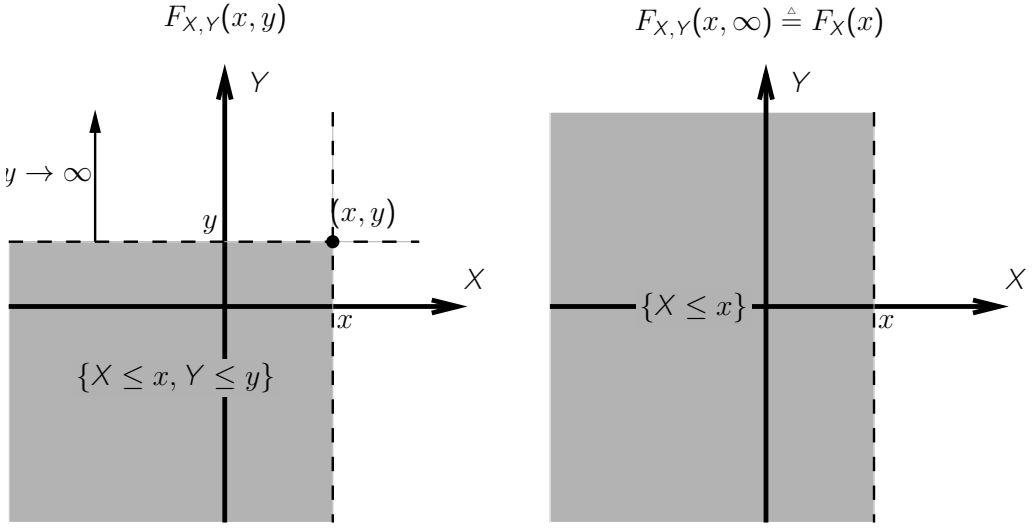
$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}(\xi_1, \ldots, \xi_n)\, \mathrm{d}\xi_n \ldots \mathrm{d}\xi_1 \tag{A.29}$$

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n F_X(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}. \tag{A.30}$$

Marginal CDF and PDF

Definition. Given a pair of r.v. $[X, Y]$, the MARGINAL CDF and PDF of X can be obtained by

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x, y), \quad f_X(x) = \lim_{y\to\infty} \int_{-\infty}^{y} f_{X,Y}(x, \eta)\, \mathrm{d}\eta = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y. \tag{A.31}$$



Conditional CDF and PDF

Definition. Given two jointly distributed random variables X and Y, then the CONDITIONAL CDF of X conditioned on $\{Y = y\}$ is given by $F_{X|Y} : \mathbb{R} \to [0, 1]$, with

$$F_{X|Y}(x|y) = \mathsf{P}\left(\{X \leq x\} \mid \{Y = y\}\right), \quad \text{if } Y \text{ only takes discrete values.} \tag{A.32}$$

If Y is a continuous random variable, the probability of $\{Y = y\}$ is zero, and thus the CONDITIONAL CDF must be defined alternatively as

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} \frac{f_{X,Y}(\xi, y)}{f_Y(y)} \, \mathrm{d}\xi, \tag{A.33}$$

assuming the existence of the PDFs and $f_Y(y) > 0$. Obviously, the CONDITIONAL PDF is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(\xi, y)}{f_Y(y)}. \tag{A.34}$$

A.4 Mean and Covariance

Mean Value of Random Variables

Definition. The MEAN VALUE of a continuous real-valued random variable X is

$$\mu_X = \mathsf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x. \tag{A.35}$$

Properties.

$$\mathsf{E}[\alpha X + \beta Y] = \alpha\,\mathsf{E}[X] + \beta\,\mathsf{E}[Y], \tag{A.36}$$

$$X \leq Y \Rightarrow \mathsf{E}[X] \leq \mathsf{E}[Y], \tag{A.37}$$

for all $\alpha, \beta \in \mathbb{R}$.

Definition. The VARIANCE of random variable X is

$$\sigma_X^2 = \mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2, \tag{A.38}$$

Definition. The COVARIANCE and CORRELATION of two random variables X and Y is

$$c_{X,Y} = \mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y], \tag{A.39}$$
$$r_{X,Y} = \mathrm{E}[XY]. \tag{A.40}$$

Properties. Given the random variables X, Y, U and V and $\alpha, \beta, \gamma, \delta \in \mathbb{R}$, it follows

$$\mathrm{Cov}[\alpha X + \beta, \gamma Y + \delta] = \alpha\gamma\,\mathrm{Cov}[X, Y], \tag{A.41}$$
$$\mathrm{Cov}[X + U, Y + V] = \mathrm{Cov}[X, Y] + \mathrm{Cov}[X, V] + \mathrm{Cov}[U, Y] + \mathrm{Cov}[U, V]. \tag{A.42}$$

Special Case. $\mathrm{Var}[\alpha X + \beta] = \alpha^2\,\mathrm{Var}[X]$.

Correlation of Random Variables

Definition. Random variables X and Y are UNCORRELATED if

$$c_{X,Y} = 0 \quad \Leftrightarrow \quad \mathsf{E}[XY] = \mathsf{E}[X]\,\mathsf{E}[Y], \tag{A.43}$$

otherwise we call X and Y CORRELATED.

Definition. By normalization we obtain the CORRELATION COEFFICIENT $\rho_{X,Y} \in [-1, 1]$

$$\rho_{X,Y} = \frac{c_{X,Y}}{\sigma_X \sigma_Y}. \tag{A.44}$$

Stochastic interrelations of random variables:

$$\text{INDEPENDENT}: \quad F_{XY} = F_X F_Y \tag{A.45}$$
$$\text{UNCORRELATED}: \quad \mu_{XY} = \mu_X \mu_Y \tag{A.46}$$
$$\text{ORTHOGONAL}: \quad \mu_{XY} = 0. \tag{A.47}$$

Example

The joint CDF of n real-valued Gaussian random variables $\boldsymbol{X} = [X_1, \ldots, X_n]^\mathsf{T}$ is given by

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det \boldsymbol{C}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{A.48}$$

The respective MEAN VECTOR and COVARIANCE MATRIX is given by

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \mathsf{E}[X_1] \\ \vdots \\ \mathsf{E}[X_n] \end{bmatrix} = \mathsf{E}\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \mathsf{E}[\boldsymbol{X}] \tag{A.49}$$

and

$$\boldsymbol{C} = \mathsf{E}\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\mathsf{T}\right] = \mathsf{E}\left[\begin{bmatrix} X_1 - \mu_1 \\ \vdots \\ X_n - \mu_n \end{bmatrix} [X_1 - \mu_1, \ldots, X_n - \mu_n]\right] = \boldsymbol{C}^\mathsf{T},$$

respectively. A common shorthand notation is

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}).$$

Covariance matrices of uncorrelated random variables $X_1, \ldots, X_n$ are diagonal

$$\boldsymbol{C} = \text{diag}\left[\sigma_1^2, \ldots, \sigma_n^2\right], \tag{A.50}$$

with variances $\sigma_i^2 = \text{Var}[X_i]$, $i \in \{1, \ldots, n\}$ as main diagonal elements.

Given n uncorrelated Gaussian random variables, the joint PDF is

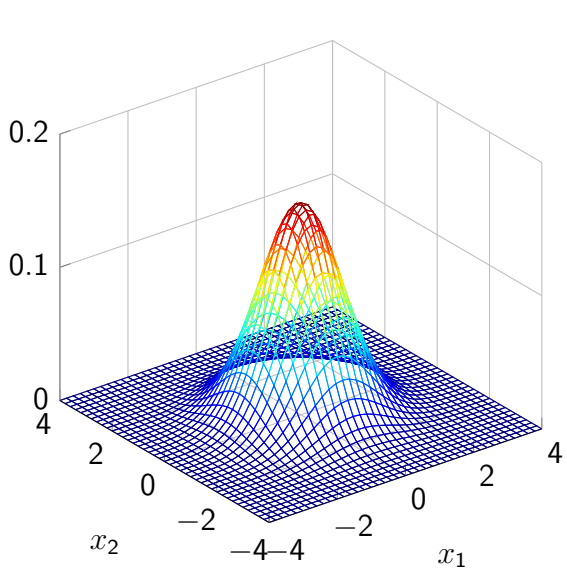
$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \frac{1}{\prod_{i=1}^n \sqrt{2\pi}\sqrt{\sigma_i^2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu_i)\sigma_i^{-2}(x_i - \mu_i)\right) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}}\, \mathrm{e}^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} = \prod_{i=1}^n f_{X_i}(x_i). \end{aligned}$$

Special Case.

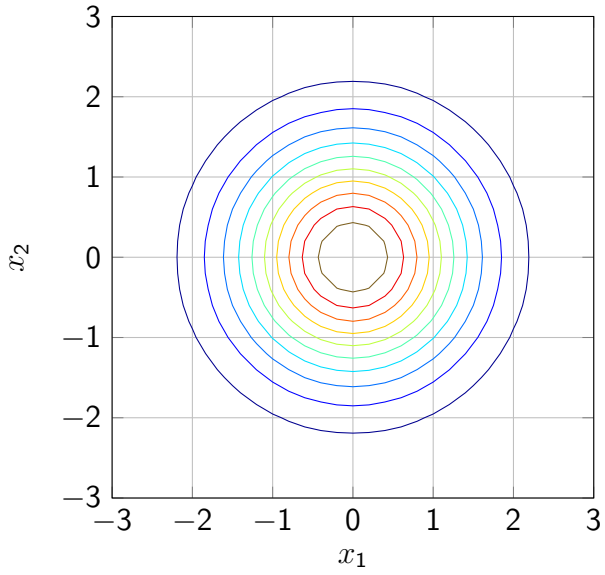
Uncorrelated Gaussian random variables are always Stochastically Independent!

Correlation between the random variables $X_1, \ldots, X_n$ leads to a non-diagonal covariance matrix $\boldsymbol{C}$. Since $\boldsymbol{C}$ is in general Positive Semidefinite (PSD), $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ can be transformed to $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, by the linear transformation

$$\boldsymbol{Y} = \boldsymbol{C}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu}). \tag{A.51}$$

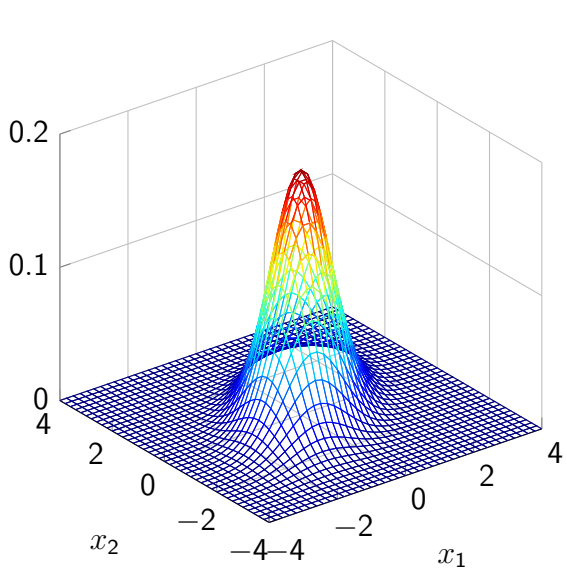


(a) Joint PDF $f_{X_1,X_2}(x_1, x_2)$

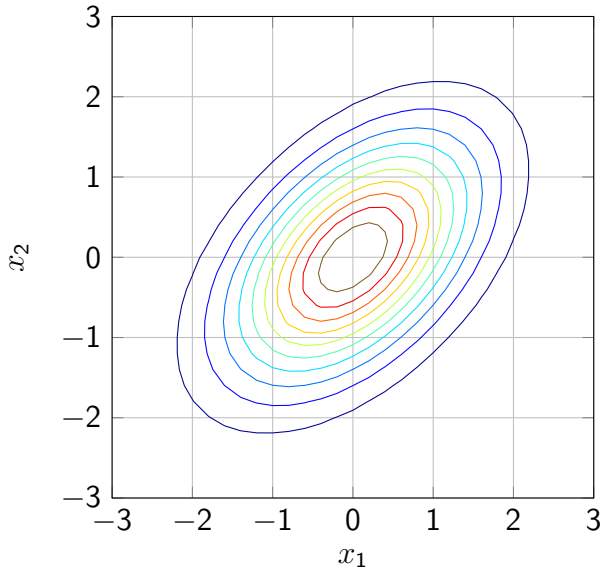


(b) Contour lines of $f_{X_1,X_2}(x_1, x_2)$

Fig. A.6: Joint PDF of the UNCORRELATED bivariate standard normal distribution.



(a) Joint PDF $f_{X_1,X_2}(x_1, x_2)$



(b) Contour lines of $f_{X_1,X_2}(x_1, x_2)$

Fig. A.7: Joint PDF of the CORRELATED bivariate standard normal distribution.

A.5 Conditional Expectation

Definition. The CONDITIONAL EXPECTATION of a random variable X, given an observation y of random variable Y is

$$\mathrm{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x. \tag{A.52}$$

The CONDITIONAL EXPECTATION $\mathrm{E}[X|Y]$ is obviously a function of the observation Y, i.e., $\mathrm{E}[X|Y]$ is still a random variable due to the random nature of the conditional argument.

The EXPECTATION of $\mathrm{E}[X|Y]$ with respect to the random variable Y obtains the expectation of the unconditioned random variable X, i.e.,

$$\mathrm{E}[\mathrm{E}[X|Y]] = \mathrm{E}[X]\,. \tag{A.53}$$

Proof.

$$
\begin{aligned}
\mathsf{E}\left[\int_{-\infty}^{\infty} x f_{X|Y}(x|Y)\,\mathrm{d}x\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x f_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \underbrace{f_{X|Y}(x|y) f_Y(y)}_{f_{X,Y}(x,y)}\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y}_{\text{Marginalization}}\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \\
&= \mathsf{E}[X]\,.
\end{aligned}
$$

The general version of this results for a given function $g(X, Y)$ is

$$
\mathsf{E}[\mathsf{E}[g(X,Y)|Y]] = \mathsf{E}[\mathsf{E}[g(X,Y)|X]] = \mathsf{E}[g(X,Y)]\,, \tag{A.54}
$$

which in the special case above results into $\mathsf{E}[\mathsf{E}[X|Y]] = \mathsf{E}[X]$.

B. Linear Operators – Matrices

B.1 Definition

Definition. A Linear Operator is a function $A : \mathbb{S} \to \mathbb{S}'$ between Vector Spaces $\mathbb{S}$ and $\mathbb{S}'$, such that for all $x, y \in \mathbb{S}$ and $\alpha \in \mathbb{C}$ (or $\mathbb{R}$) the following holds,

$$\text{Additivity} : A(x + y) = Ax + Ay \tag{B.1}$$
$$\text{Scalability} : A(\alpha x) = \alpha Ax. \tag{B.2}$$

In this tutorial, we only consider Linear Operators between finite-dimensional Complex Vector Spaces $\mathbb{C}^N$ and $\mathbb{C}^M$ with $\alpha \in \mathbb{C}$. Consequently, a Linear Operator A is represented by a Matrix $\boldsymbol{A} : \mathbb{C}^N \to \mathbb{C}^M$.

Derivation.

Assume $\{u_n\}_{n=1}^N$ and $\{u'_m\}_{m=1}^M$ is an ORTHONORMAL BASIS of an abstract[1] N-dimensional vector space $\mathbb{S}$ and the abstract M-dimensional vector space $\mathbb{S}'$, respectively, and $\mathsf{A} : \mathbb{S} \to \mathbb{S}'$ is a LINEAR OPERATOR with $y = \mathsf{A}x$ for all $x \in \mathbb{S}$ and $y \in \mathbb{S}'$.

Then, with $\boldsymbol{x} = [x_1, \ldots, x_N]^\mathsf{T}$ and $\boldsymbol{y} = [y_1, \ldots, y_M]^\mathsf{T}$ representing the coordinates of x and y with respect to the basis $\{u_n\}_{n=1}^N$ and $\{u'_m\}_{m=1}^M$, respectively, we obtain the Matrix $\boldsymbol{A}$ with $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ as follows:

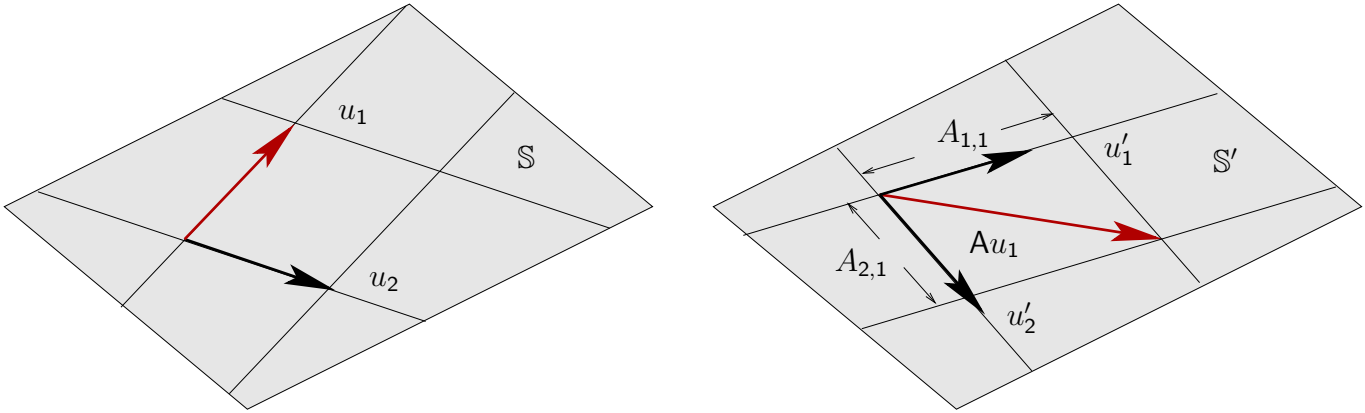
$$y_m = \langle y, u'_m \rangle = \langle \mathsf{A}x, u'_m \rangle = \left\langle \mathsf{A} \sum_{n=1}^{N} x_n u_n, u'_m \right\rangle = \sum_{n=1}^{N} \langle \mathsf{A}u_n, u'_m \rangle x_n \tag{B.3}$$

Consequently, the elements of matrix $\boldsymbol{A}$ are obtained by

$$[\boldsymbol{A}]_{m,n} \triangleq \langle \mathsf{A}u_n, u'_m \rangle \quad \text{and} \quad [\boldsymbol{A}^\mathsf{T}]_{n,m} \triangleq \langle \mathsf{A}u_n, u'_m \rangle. \tag{B.4}$$

[1] For instance the vector space of bandlimited periodic signals.

$$\mathsf{A} : \mathbb{S} \to \mathbb{S}',\ u_1 \mapsto \mathsf{A}u_1 = A_{1,1}u'_1 + A_{2,1}u'_2$$
$$A_{1,1} \triangleq \langle \mathsf{A}u_1, u'_1 \rangle,\ A_{2,1} \triangleq \langle \mathsf{A}u_1, u'_2 \rangle$$



B.2 Elementwise Perspective on Matrices

$$\boldsymbol{A} \in \mathbb{C}^{M \times N} \text{ with } [\boldsymbol{A}]_{m,n} = a_{m,n} = \mathsf{Re}\{a_{m,n}\} + \mathsf{j}\,\mathsf{Im}\{a_{m,n}\} \in \mathbb{C}, \tag{B.5}$$

and

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{x} &= \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & & \vdots \\ a_{M,1} & a_{M,2} & \cdots & a_{M,N} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \\
&= \begin{bmatrix} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,N}x_N \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,N}x_N \\ \vdots \\ a_{M,1}x_1 + a_{M,2}x_2 + \cdots + a_{M,N}x_N \end{bmatrix} \\
&= \begin{bmatrix} \sum_{n=1}^{N} a_{1,n}x_n \\ \sum_{n=1}^{N} a_{2,n}x_n \\ \vdots \\ \sum_{n=1}^{N} a_{M,n}x_n \end{bmatrix}.
\end{aligned}$$

Definition. Matrix, Transpose Matrix, Adjoint Matrix

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & & \vdots \\ a_{M,1} & a_{M,2} & \cdots & a_{M,N} \end{bmatrix},$$

$$\boldsymbol{A}^{\mathsf{T}} = \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{M,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{M,2} \\ \vdots & \vdots & & \vdots \\ a_{1,N} & a_{2,N} & \cdots & a_{M,N} \end{bmatrix},$$

$$\boldsymbol{A}^{\mathsf{H}} = \begin{bmatrix} a^*_{1,1} & a^*_{2,1} & \cdots & a^*_{M,1} \\ a^*_{1,2} & a^*_{2,2} & \cdots & a^*_{M,2} \\ \vdots & \vdots & & \vdots \\ a^*_{1,N} & a^*_{2,N} & \cdots & a^*_{M,N} \end{bmatrix}.$$

The Adjoint Matrix $\boldsymbol{A}^{\mathsf{H}} \in \mathbb{C}^{N \times M}$ of a matrix $\boldsymbol{A} \in \mathbb{C}^{M \times N}$ has the constituting property

$$\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{A}^{\mathsf{H}}\boldsymbol{y} \rangle \quad \Leftrightarrow \quad \boldsymbol{y}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{x} = \left(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{y}\right)^{\mathsf{H}}\boldsymbol{x}, \qquad \forall \boldsymbol{x} \in \mathbb{C}^N, \boldsymbol{y} \in \mathbb{C}^M. \tag{B.6}$$

Matrix Columns, Matrix Rows

$$\boldsymbol{A} = [\boldsymbol{a}_1 \; \boldsymbol{a}_2 \cdots \boldsymbol{a}_N] \quad \text{with} \quad \boldsymbol{a}_n = \begin{bmatrix} a_{1,n} \\ a_{2,n} \\ \vdots \\ a_{M,n} \end{bmatrix}$$

$$\boldsymbol{A}^{\mathsf{T}} = \begin{bmatrix} \boldsymbol{a}_1^{\mathsf{T}} \\ \boldsymbol{a}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{a}_N^{\mathsf{T}} \end{bmatrix} \quad \text{with} \quad \boldsymbol{a}_n^{\mathsf{T}} = [a_{1,n} \; a_{2,n} \cdots a_{M,n}]$$

$$\boldsymbol{A}^{\mathsf{H}} = \begin{bmatrix} \boldsymbol{a}_1^{\mathsf{H}} \\ \boldsymbol{a}_2^{\mathsf{H}} \\ \vdots \\ \boldsymbol{a}_N^{\mathsf{H}} \end{bmatrix} \quad \text{with} \quad \boldsymbol{a}_n^{\mathsf{T}} = \left[a_{1,n}^* \; a_{2,n}^* \cdots a_{M,n}^*\right],$$

with $\left(\boldsymbol{A}^{\mathsf{T}}\right)^{\mathsf{T}} = \boldsymbol{A}$ and $\left(\boldsymbol{A}^{\mathsf{H}}\right)^{\mathsf{H}} = \boldsymbol{A}$.

Definition. Identity Matrix, Zero Matrix, Inverse Matrix, Selection Vector

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathsf{diag}\,[1\ 1 \cdots 1]$$

$$\mathbf{0} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \quad \text{and} \quad \boldsymbol{A} + (-\boldsymbol{A}) = \mathbf{0}$$

$$\boldsymbol{e}_i = [\underbrace{0 \cdots 0}_{i-1 \text{ zeros}}\ 1\ \underbrace{0 \cdots 0}_{N-i \text{ zeros}}]^{\mathsf{T}} \in \mathbb{C}^N \quad \text{with}$$

$$\boldsymbol{A}\boldsymbol{e}_i = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & & \vdots \\ a_{M,1} & a_{M,2} & \cdots & a_{M,N} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{1,i} \\ a_{2,i} \\ \vdots \\ a_{M,i} \end{bmatrix} = \boldsymbol{a}_i.$$

B.3 Fundamental Subspace Perspective on Matrices

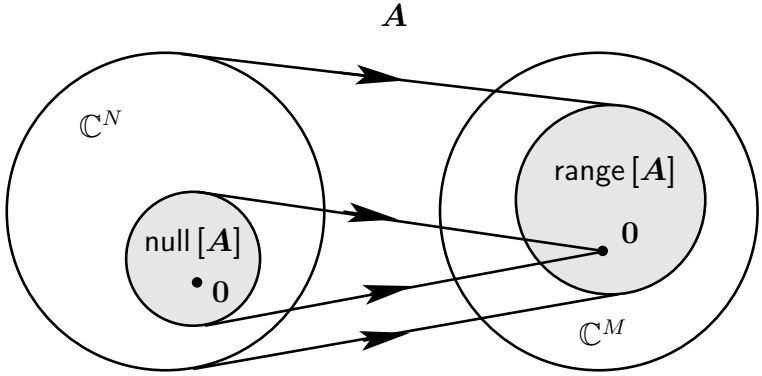
Any Linear Operator $\boldsymbol{A} : \mathbb{C}^N \to \mathbb{C}^M$ is related to Four Fundamental Subspaces, viz.

$$\text{Range Space of } \boldsymbol{A}: \quad \text{range}\,[\boldsymbol{A}] = \left\{ \boldsymbol{A}\boldsymbol{x} \middle| \, \boldsymbol{x} \in \mathbb{C}^N \right\} \tag{B.7}$$

$$\text{Null Space of } \boldsymbol{A}: \quad \text{null}\,[\boldsymbol{A}] = \left\{ \boldsymbol{x} \in \mathbb{C}^N \middle| \, \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \right\} \tag{B.8}$$

$$\text{Orthogonal Complement of } \text{range}\,[\boldsymbol{A}]: \quad \text{range}\,[\boldsymbol{A}]^{\perp} = \text{null}\left[\boldsymbol{A}^{\mathsf{H}}\right] \tag{B.9}$$

$$\text{Orthogonal Complement of null}\,[\boldsymbol{A}]: \quad \text{null}\,[\boldsymbol{A}]^{\perp} = \text{range}\left[\boldsymbol{A}^{\mathsf{H}}\right]. \tag{B.10}$$



B.4 Eigenvectors and Eigenvalues

Definition. An EIGENVECTOR of a matrix $\boldsymbol{A} \in \mathbb{C}^{M \times M}$ is a nonzero vector $\boldsymbol{v} \in \mathbb{C}^N$ such that

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v} \quad \text{for some} \quad \lambda \in \mathbb{C}. \tag{B.11}$$

The constant λ is the EIGENVALUE to the EIGENVECTOR $\boldsymbol{v}$. The $\boldsymbol{v}$ and its corresponding λ form a so-called EIGENPAIR of the matrix $\boldsymbol{A}$.

For finite dimensional matrices the SET OF EIGENVALUES is discrete, the maximum number of different eigenvalues is M.

The EIGENVALUES of SELF-ADJOINT MATRICES $\boldsymbol{A}^{\mathsf{H}} = \boldsymbol{A}$ are real-valued and the EIGENVECTORS form an ONB.

Proof.

(1) $\lambda \boldsymbol{v}^{\mathsf{H}}\boldsymbol{v} = \boldsymbol{v}^{\mathsf{H}}(\lambda \boldsymbol{v}) = \boldsymbol{v}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}^{\mathsf{H}}\boldsymbol{A}^{\mathsf{H}}\boldsymbol{v} = (\boldsymbol{A}\boldsymbol{v})^{\mathsf{H}}\,\boldsymbol{v} = (\lambda \boldsymbol{v})^{\mathsf{H}}\,\boldsymbol{v} = \lambda^{*}\boldsymbol{v}^{\mathsf{H}}\boldsymbol{v} \quad \Rightarrow \quad \lambda = \lambda^{*}$

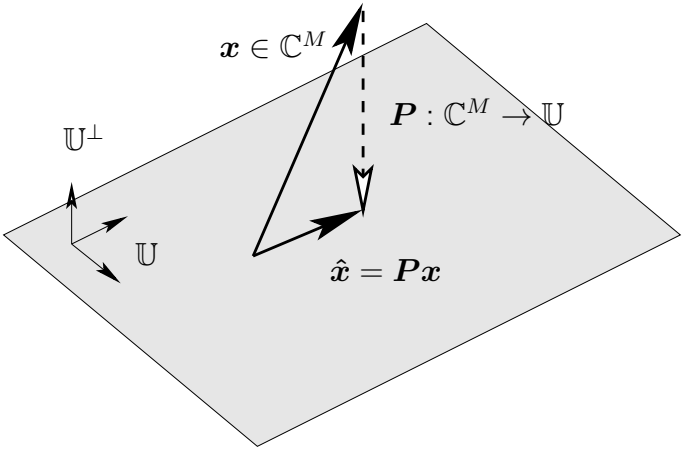
(2) $\lambda_i \boldsymbol{v}_j^{\mathsf{H}}\boldsymbol{v}_i = \boldsymbol{v}_j^{\mathsf{H}}(\lambda_i \boldsymbol{v}_i) = \boldsymbol{v}_j^{\mathsf{H}}\boldsymbol{A}\boldsymbol{v}_i = \boldsymbol{v}_j^{\mathsf{H}}\boldsymbol{A}^{\mathsf{H}}\boldsymbol{v}_i = (\boldsymbol{A}\boldsymbol{v}_j)^{\mathsf{H}}\,\boldsymbol{v}_i = (\lambda_j \boldsymbol{v}_j)^{\mathsf{H}}\,\boldsymbol{v}_i = \lambda_j^{*}\boldsymbol{v}_j^{\mathsf{H}}\boldsymbol{v}_i \quad \Rightarrow \quad \boldsymbol{v}_j^{\mathsf{H}}\boldsymbol{v}_i = 0$

since λ_i and λ_j are real-valued (1st part) and different for $i \neq j$ by assumption.

B.5 Orthogonal Projectors

Best Approximation and Orthogonal Projection

The PROJECTION OPERATOR $\boldsymbol{P} : \mathbb{C}^M \to \mathbb{C}^M$ is related to the BEST APPROXIMATION problem, i.e., find $\hat{\boldsymbol{x}} \in \mathbb{U} \subset \mathbb{C}^M$ that is closest to a given $\boldsymbol{x} \in \mathbb{C}^M$. As a result, we obtain $\hat{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{x} \perp \boldsymbol{x} - \boldsymbol{P}\boldsymbol{x}$.



According to the Closest Point Theorem[2] we obtain

$$\text{Existence}: \ \exists \hat{\boldsymbol{x}} \in \mathbb{U} \text{ such that } \|\boldsymbol{x} - \hat{\boldsymbol{x}}\| \leq \|\boldsymbol{x} - \boldsymbol{s}\| \text{ for every } \boldsymbol{s} \in \mathbb{U} \tag{B.12}$$

$$\text{Orthogonality}: \ \boldsymbol{x} - \hat{\boldsymbol{x}} \perp \mathbb{U} \tag{B.13}$$

$$\text{Linearity}: \ \hat{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{x} \text{ and } \boldsymbol{P} \text{ only depends on } \mathbb{U} \tag{B.14}$$

$$\text{Indempotency}: \ \boldsymbol{P}^2 = \boldsymbol{P} \tag{B.15}$$

$$\text{Self-Adjointness}: \ \boldsymbol{P}^{\mathsf{H}} = \boldsymbol{P}. \tag{B.16}$$

Note.

Projection Operators are always Indempotent.

Orthogonal Projectors must be Indempotent and Self-Adjoint.

Oblique Projectors (non-orthogonal projectors) are only indempotent, i.e., $\boldsymbol{P}^2 = \boldsymbol{P}$, but Self-Adjointness does not hold.

Oblique Projectors are not considered in this tutorial.

[2]The Closest Point Theorem applies for the category of Convex Sets which Linear Subspaces are belonging to.

Orthogonal Projection on Subspaces.

Given a pair of matrices $\boldsymbol{A} : \mathbb{C}^N \to \mathbb{C}^M$ and $\boldsymbol{B} : \mathbb{C}^M \to \mathbb{C}^N$ and $\boldsymbol{A}$ is a LEFT-INVERSE of $\boldsymbol{B}$, i.e., $\boldsymbol{AB} = \boldsymbol{I}$, then

$$\boxed{\boldsymbol{BA} \text{ is a PROJECTOR onto range}\,[\boldsymbol{B}] \subset \mathbb{C}^N,} \tag{B.17}$$

since $(\boldsymbol{BA})^2 = \boldsymbol{BABA} = \boldsymbol{B}(\boldsymbol{AB})\boldsymbol{A} = \boldsymbol{BA}$.

If $\boldsymbol{BA}$ is also SELF-ADJOINT, then

$$\boxed{\boldsymbol{BA} \text{ is an ORTHOGONAL PROJECTOR onto range}\,[\boldsymbol{B}] \subset \mathbb{C}^N,} \tag{B.18}$$

since $(\boldsymbol{BA})^2 = \boldsymbol{BABA} = \boldsymbol{B}(\boldsymbol{AB})\boldsymbol{A} = \boldsymbol{BA}$ and $(\boldsymbol{BA})^{\mathsf{H}} = \boldsymbol{BA}$.

Example. Assume the column vectors of $\boldsymbol{U}$ form an ONB of $\mathbb{U}$, then $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^{\mathsf{H}}$ is the ORTHOGONAL PROJECTOR onto $\mathbb{U}$. For a proof consider the plugins

$$\boldsymbol{A} = \boldsymbol{U}^{\mathsf{H}} \quad \text{and} \quad \boldsymbol{B} = \boldsymbol{U}. \tag{B.19}$$

Orthogonal Projection via Pseudo-Inverse.

If $(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})^{-1}$ exists, then $\boldsymbol{A}^{+}_{\text{left}} = (\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{H}}$ is a LEFT-INVERSE of $\boldsymbol{A}$, i.e.,

$$\text{(1)} \quad \boldsymbol{A}\boldsymbol{A}^{+}_{\text{left}} = \boldsymbol{A}(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{H}} \text{ is the ORTHOGONAL PROJECTOR onto range}\,[\boldsymbol{A}] \tag{B.20}$$

with so-called PSEUDO-INVERSE $\boldsymbol{A}^{+}_{\text{left}}$.

Otherwise, if $(\boldsymbol{A}\boldsymbol{A}^{\mathsf{H}})^{-1}$ exists, then $\boldsymbol{A}^{+}_{\text{right}} = \boldsymbol{A}^{\mathsf{H}}(\boldsymbol{A}\boldsymbol{A}^{\mathsf{H}})^{-1}$ is a RIGHT-INVERSE of $\boldsymbol{A}$, i.e.,

$$\text{(2)} \quad \boldsymbol{A}^{+}_{\text{right}}\boldsymbol{A} = \boldsymbol{A}^{\mathsf{H}}(\boldsymbol{A}\boldsymbol{A}^{\mathsf{H}})^{-1}\boldsymbol{A} \text{ is the ORTHOGONAL PROJECTOR onto range}\,\left[\boldsymbol{A}^{\mathsf{H}}\right] \tag{B.21}$$

with the PSEUDO-INVERSE $\boldsymbol{A}^{+}_{\text{right}}$.

Weighted Sum of Disjoint Orthogonal Projectors

Assume rank-R matrices $\boldsymbol{A} : \mathbb{C}^M \to \mathbb{C}^M$ that can be represented by means of a WEIGHTED SUM of rank-R_i DISJOINT ORTHOGONAL PROJECTORS $\boldsymbol{P}_i : \mathbb{C}^M \to \mathbb{C}^M$,

$$\boldsymbol{A} = \sum_{i=1}^{P} w_i \boldsymbol{P}_i, \tag{B.22}$$

with rank $R = \sum_{i=1}^{P} R_i$ and range$[\boldsymbol{P}_i] \perp$ range$[\boldsymbol{P}_j]$ for all $i \neq j$, $i = 1, \ldots, P$, i.e., all projectors are mutually orthogonal to each other.

A weighted sum of disjoint orthogonal projectors $\boldsymbol{A} : \mathbb{C}^M \to \mathbb{C}^M$ referring to (B.22) is a NORMAL MATRIX with

$$\boldsymbol{A}\boldsymbol{A}^{\mathsf{H}} = \boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}. \tag{B.23}$$

The respective projectors and weights depend on $\boldsymbol{A}$.

Note. The weights take COMPLEX VALUES, i.e., $\boxed{w_i \in \mathbb{C}}$.

B.6 Special Matrices

Definition. $\boldsymbol{A} : \mathbb{C}^M \to \mathbb{C}^M$ referring to (B.22) is a Self-Adjoint Matrix when

$$\boldsymbol{A}^{\mathsf{H}} = \boldsymbol{A}. \tag{B.24}$$

Self-Adjoint Matrices are also Normal Matrices.

Note. The weights strictly take Real Values, i.e., $\boxed{w_i \in \mathbb{R}}$.

Definition. $\boldsymbol{A} : \mathbb{C}^M \to \mathbb{C}^M$ referring to (B.22) is a Positive Definite Matrix when it is Self-Adjoint[3] and

$$\boldsymbol{x}^{\mathsf{H}} \boldsymbol{A} \boldsymbol{x} \geq 0, \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{C}^M. \tag{B.25}$$

Note. The weights strictly take Nonnegative Values, i.e., $\boxed{w_i \geq 0}$.

[3] There exist more general definitions of Positive Definite Matrices.

Eigenpairs of Normal Matrices

NORMAL MATRICES $\boldsymbol{A}$ are LINEAR COMBINATIONS of ORTHOGONAL PROJECTORS, i.e., $\boldsymbol{A} = \sum_{i=1}^{P} w_i \boldsymbol{P}_i$.

Consequently, any $\boldsymbol{u}_j \in \text{range}\,[\boldsymbol{P}_j]$ and its corresponding weight w_j form an EIGENPAIR of $\boldsymbol{A}$, since for every element of $\text{range}\,[\boldsymbol{P}_j]$ we obtain $\boldsymbol{P}_j \boldsymbol{u}_j = \boldsymbol{u}_j$ by definition and $\boldsymbol{P}_i \boldsymbol{u}_j = \boldsymbol{0}$ for $i \neq j$ due to the orthogonality of projectors, and thus

$$\boldsymbol{A}\boldsymbol{u}_j = \left(\sum_{i=1}^{P} w_i \boldsymbol{P}_i \right) \boldsymbol{u}_j \tag{B.26}$$

$$= \sum_{i=1}^{P} w_i (\boldsymbol{P}_i \boldsymbol{u}_j) = w_j \boldsymbol{u}_j. \tag{B.27}$$

Obviously, each w_j and $\boldsymbol{u}_j$ form an EIGENPAIR consisting of an EIGENVECTOR and an EIGENVALUE of the matrix:

$$\boldsymbol{A}\boldsymbol{u}_j = w_j \boldsymbol{u}_j, \quad \text{for all} \quad j = 1, \ldots, P. \tag{B.28}$$

Spectral Theorem of Normal Matrices

Since (1) NORMAL MATRICES $\boldsymbol{A}$ are LINEAR COMBINATIONS of ORTHOGONAL PROJECTORS $\boldsymbol{P}_i$ and any $\boldsymbol{u}_j \in \text{range}\,[\boldsymbol{P}_j]$ and its corresponding weight w_j form an EIGENPAIR of $\boldsymbol{A}$, and

(2) since any $\boldsymbol{P}_i$ can be represented as a LINEAR COMBINATION of rank-1 projectors $\boldsymbol{u}_{i,j}\boldsymbol{u}_{i,j}^{\mathsf{H}}$ with respect to elements of an ONB $\{\boldsymbol{u}_{i,j}\}_{j=1}^{R_i}$, with

$$\text{range}\,[\boldsymbol{P}_i] = \text{range}\,[\boldsymbol{U}_i] = \text{span}\,[\boldsymbol{u}_{i,1} \ \cdots \ \boldsymbol{u}_{i,R_i}]\,, \tag{B.29}$$

with $\boldsymbol{U}_i = [\boldsymbol{u}_{i,1}, \ldots, \boldsymbol{u}_{i,R_i}]$ and R_i is the dimension of $\text{range}\,[\boldsymbol{P}_i]$, we obtain the main result of the SPECTRAL THEOREM of NORMAL MATRICES,

$$\boldsymbol{A} = \sum_{i=1}^{P} w_i \boldsymbol{P}_i \tag{B.30}$$

$$= \sum_{i=1}^{P} w_i \boldsymbol{U}_i \boldsymbol{U}_i^{\mathsf{H}} = \sum_{i=1}^{P} \sum_{j=1}^{R_i} w_i \boldsymbol{u}_{i,j} \boldsymbol{u}_{i,j}^{\mathsf{H}} \tag{B.31}$$

$$= \begin{bmatrix} \boldsymbol{U}_1 & \cdots & \boldsymbol{U}_P \end{bmatrix} \begin{bmatrix} w_1 \boldsymbol{I}_{R_1 \times R_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_P \boldsymbol{I}_{R_P \times R_P} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^{\mathsf{H}} \\ \vdots \\ \boldsymbol{U}_P^{\mathsf{H}} \end{bmatrix}. \tag{B.32}$$

B.7 Singular Value Decomposition

Derivation and Properties of the SVD

Any Compact Linear Operator A can be decomposed by means of the Singular Value Decomposition (SVD) and any Finite Dimensional Matrix $\boldsymbol{A} : \mathbb{C}^N \to \mathbb{C}^M$ is a compact linear operator.

(1) Since the product of matrices $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$—the so-called Gramian Matrix—is Normal, Self-Adjoint and Positive Definite,[4] the Spectral Theorem can be applied such that

$$\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A} = \sum_{i=1}^{P} \lambda_i \boldsymbol{P}_i = \sum_{i=1}^{R} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{H}}, \tag{B.33}$$

with rank $R = R_1 + \cdots + R_P \leq \min(M, N)$ and at least P different weights $\lambda_i > 0$.

The vectors $\{\boldsymbol{v}_i\}_{i=1}^{R}$ form an ONB of range $\left[\boldsymbol{A}^{\mathsf{H}}\right]$.

[4] $\left(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}\right)^{\mathsf{H}} = \boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$ and $\boldsymbol{v}^{\mathsf{H}}\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}^{\mathsf{H}}\lambda\boldsymbol{v} = \lambda\boldsymbol{v}^{\mathsf{H}}\boldsymbol{v} = \lambda \|\boldsymbol{v}\|^2 \geq 0$.

(2) Given the ORTHONORMAL VECTORS $\{\boldsymbol{v}_i\}_{i=1}^R$, the vectors $\{\boldsymbol{u}_i\}_{i=1}^R$, with

$$\boldsymbol{u}_i \triangleq \lambda_i^{-\frac{1}{2}} \boldsymbol{A}\boldsymbol{v}_i, \tag{B.34}$$

again constitute an ONB:

$$\begin{aligned}
\boldsymbol{u}_j^{\mathsf{H}}\boldsymbol{u}_i &= (\lambda_j^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{v}_j)^{\mathsf{H}}\lambda_i^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{v}_i \\
&= \lambda_j^{-\frac{1}{2}}\lambda_i^{-\frac{1}{2}}\boldsymbol{v}_j^{\mathsf{H}}\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{v}_i \\
&= \lambda_j^{-\frac{1}{2}}\lambda_i^{-\frac{1}{2}}\lambda_i\boldsymbol{v}_j^{\mathsf{H}}\boldsymbol{v}_i \\
&= \begin{cases} 1 & ; \quad i=j \\ 0 & ; \quad \text{otherwise} \end{cases}.
\end{aligned}$$

Definition.

The elements of an ONB $\{\boldsymbol{v}_i\}_{i=1}^R$ with EIGENVALUES $\{\lambda_i\}_{i=1}^R$ of a GRAMIAN MATRIX $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$ are called RIGHT SINGULAR VECTORS of $\boldsymbol{A}$.

The elements of the ONB $\{\boldsymbol{u}_i = 1/\sqrt{\lambda_i}\boldsymbol{A}\boldsymbol{v}_i\}_{i=1}^R$ with EIGENVALUES $\{\lambda_i\}_{i=1}^R$ of a GRAMIAN MATRIX $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$ are called corresponding LEFT SINGULAR VECTORS of $\boldsymbol{A}$.

The SINGULAR VALUE DECOMPOSITION (SVD) of an arbitrary matrix $\boldsymbol{A} : \mathbb{C}^N \to \mathbb{C}^M$ is constituted by

$$\boldsymbol{A} = \sum_{i=1}^{R} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{H}}, \quad \text{with matrix rank} \quad R \tag{B.35}$$

$$\boldsymbol{u}_j^{\mathsf{H}} \boldsymbol{u}_i = \delta_{i,j} \quad \text{and} \quad \boldsymbol{v}_j^{\mathsf{H}} \boldsymbol{v}_i = \delta_{i,j} \quad \text{and} \quad \sigma_i > 0 \quad i,j = 1,\ldots,Q. \tag{B.36}$$

with $\delta_{i,j} = 1$ when $i = j$ and zero otherwise.

The set of LEFT SINGULAR VECTORS $\{\boldsymbol{u}_i\}_{i=1}^R$ of nonzero SINGULAR VALUES $\sigma_i > 0$ provides an ONB for the IMAGE SPACE ($\text{range}\,[\boldsymbol{A}] \in \mathbb{C}^M$) of the matrix $\boldsymbol{A}$, i.e.,

$$\text{range}\,[\boldsymbol{A}] = \text{span}\,[\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_R]\,. \tag{B.37}$$

Correspondingly, the RIGHT SINGULAR VECTORS $\{\boldsymbol{v}_i\}_{i=1}^R$ of nonzero SINGULAR VALUES $\sigma_i > 0$ provides an ONB for the IMAGE SPACE ($\text{range}\left[\boldsymbol{A}^{\mathsf{H}}\right] \in \mathbb{C}^N$) of the matrix $\boldsymbol{A}^{\mathsf{H}}$, which is equal to the COMPLEMENT of the NULL SPACE ($\text{null}\,[\boldsymbol{A}]^{\perp} \equiv \text{range}\left[\boldsymbol{A}^{\mathsf{H}}\right] \in \mathbb{C}^N$) of the matrix $\boldsymbol{A}$, i.e.,

$$\text{range}\left[\boldsymbol{A}^{\mathsf{H}}\right] = \text{null}\,[\boldsymbol{A}]^{\perp} = \text{span}\,[\boldsymbol{v}_1\ \boldsymbol{v}_2 \cdots \boldsymbol{v}_R]\,. \tag{B.38}$$

Derivation.

Applying the matrix $\boldsymbol{A}$ to a vector $\boldsymbol{x}$ can be equivalently expressed by

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}(\boldsymbol{x}_- + \boldsymbol{x}_\perp) = \boldsymbol{A}\boldsymbol{x}_-,$$

with $\boldsymbol{x}_- \in \text{range}\left[\boldsymbol{A}^{\mathsf{H}}\right] \equiv \text{null}\left[\boldsymbol{A}\right]^\perp$ and $\boldsymbol{x}_\perp \in \text{null}\left[\boldsymbol{A}\right]$.

Since $\{\boldsymbol{v}_i\}_{i=1}^R$ forms an ONB for $\text{range}\left[\boldsymbol{A}^{\mathsf{H}}\right] \equiv \text{null}\left[\boldsymbol{A}\right]^\perp$, we obtain

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{x} &= \boldsymbol{A}\boldsymbol{x}_- \\
&= \boldsymbol{A}\sum_{i=1}^{R} \boldsymbol{v}_i\boldsymbol{v}_i^{\mathsf{H}}\boldsymbol{x}_- \\
&= \sum_{i=1}^{R} \boldsymbol{A}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathsf{H}}\boldsymbol{x}_- \\
&= \sum_{i=1}^{R} \sqrt{\lambda_i}\boldsymbol{u}_i\boldsymbol{v}_i^{\mathsf{H}}\boldsymbol{x}_- \\
&= \sum_{i=1}^{R} \sqrt{\lambda_i}\boldsymbol{u}_i\boldsymbol{v}_i^{\mathsf{H}}\boldsymbol{x},
\end{aligned}$$

with $\sigma_i = \sqrt{\lambda_i}$.

Sorted Matrix Representation

By sorting the SINGULAR VALUES according to

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R \quad \text{with} \quad R = \mathsf{rank}\,[\boldsymbol{A}]\,,$$

we obtain the SORTED MATRIX REPRESENTATION of the SVD,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{H}}, \quad \text{with} \tag{B.39}$$

$$\boldsymbol{U} = [\boldsymbol{u}_1\ \boldsymbol{u}_2 \cdots \boldsymbol{u}_R] \in \mathbb{C}^{M\times R}, \tag{B.40}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_R \end{bmatrix} = \mathsf{diag}\,[\sigma_1\ \sigma_2 \cdots \sigma_R] \in \mathbb{C}^{R\times R} \tag{B.41}$$

$$\boldsymbol{V}^{\mathsf{H}} = \begin{bmatrix} \boldsymbol{v}_1^{\mathsf{H}} \\ \vdots \\ \boldsymbol{v}_R^{\mathsf{H}} \end{bmatrix} \in \mathbb{C}^{R\times N}. \tag{B.42}$$

Extended Sorted Matrix Representation. By accomplishing the ONB of the LEFT SINGULAR VECTORS (column vectors of $\boldsymbol{U}$) and RIGHT SINGULAR VECTORS (column vectors of $\boldsymbol{V}$) by their ORTHOGONAL COMPLEMENTS, and taking into account that $Q = \mathsf{min}(M, N)$,

$$\boldsymbol{U}_{R<Q}^{\perp} = [\boldsymbol{u}_{R+1}\ \boldsymbol{u}_{R+2} \cdots \boldsymbol{u}_{Q}] \quad \text{and} \quad \boldsymbol{U}_{Q<M}^{\perp} = [\boldsymbol{u}_{Q+1}\ \boldsymbol{u}_{Q+2} \cdots \boldsymbol{u}_{M}] \tag{B.43}$$

$$\boldsymbol{V}_{R<Q}^{\perp} = [\boldsymbol{v}_{R+1}\ \boldsymbol{v}_{R+2} \cdots \boldsymbol{v}_{Q}] \quad \text{and} \quad \boldsymbol{V}_{Q<N}^{\perp} = [\boldsymbol{v}_{Q+1}\ \boldsymbol{v}_{Q+2} \cdots \boldsymbol{v}_{N}] \tag{B.44}$$

$$\mathsf{range}\,[\boldsymbol{U}] \oplus \mathsf{range}\left[\boldsymbol{U}_{R<Q}^{\perp}\right] \oplus \mathsf{range}\left[\boldsymbol{U}_{Q<M}^{\perp}\right] = \mathbb{C}^{M} \quad \text{and} \tag{B.45}$$

$$\mathsf{range}\,[\boldsymbol{V}] \oplus \mathsf{range}\left[\boldsymbol{V}_{R<Q}^{\perp}\right] \oplus \mathsf{range}\left[\boldsymbol{V}_{Q<N}^{\perp}\right] = \mathbb{C}^{N}, \tag{B.46}$$

we obtain the EXTENDED SORTED MATRIX REPRESENTATION of the **SVD of matrices <u>with</u> maximum rank** ($R = Q$)

$$\boldsymbol{A} = \begin{cases} \left[\boldsymbol{U}, \boldsymbol{U}_{Q<M}^{\perp}\right] & \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} & \boldsymbol{V}^{\mathsf{H}} & ; \quad M \geq N \quad (\text{TALL}) \\ \boldsymbol{U} & \boldsymbol{\Sigma} & \boldsymbol{V}^{\mathsf{H}} & ; \quad M = N \quad (\text{SQUARE}) \\ \boldsymbol{U} & \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \end{bmatrix} & \begin{bmatrix} \boldsymbol{V}^{\mathsf{H}} \\ \boldsymbol{V}_{Q<N}^{\perp,\mathsf{H}} \end{bmatrix} & ; \quad M \leq N \quad (\text{WIDE}). \end{cases} \tag{B.47}$$

SVD of Matrices <u>without</u> maximum Rank.

For TALL MATRICES $\boldsymbol{A}_{M\geq N}$ with linear dependent columns or WIDE MATRICES $\boldsymbol{A}_{M\leq N}$ with linear dependent rows, i.e., with

$$\mathsf{rank}\,[\boldsymbol{A}] < Q = \mathsf{min}\{M, N\}, \tag{B.48}$$

we obtain the cases

$$\boldsymbol{A}_{M>N} = \left[\boldsymbol{U}, \boldsymbol{U}^{\perp}_{R<Q}, \boldsymbol{U}^{\perp}_{Q<M}\right] \left[\begin{array}{c} \left[\begin{array}{cc} \boldsymbol{\Sigma}_{R\times R} & \boldsymbol{0}_{R\times N-R} \\ \boldsymbol{0}_{N-R\times R} & \boldsymbol{0}_{N-R\times N-R} \end{array} \right] \\ \left[\begin{array}{cc} \boldsymbol{0}_{M-N\times R} & \boldsymbol{0}_{M-N\times N-R} \end{array} \right] \end{array} \right] \left[\begin{array}{c} \boldsymbol{V}^{\mathsf{H}} \\ \boldsymbol{V}^{\perp,\mathsf{H}}_{R<Q} \end{array} \right] \tag{B.49}$$

$$\boldsymbol{A}_{M=N} = \left[\boldsymbol{U}, \boldsymbol{U}^{\perp}_{R<Q}\right] \left[\begin{array}{cc} \boldsymbol{\Sigma}_{R\times R} & \boldsymbol{0}_{R\times N-R} \\ \boldsymbol{0}_{N-R\times R} & \boldsymbol{0}_{N-R\times N-R} \end{array} \right] \left[\begin{array}{c} \boldsymbol{V}^{\mathsf{H}} \\ \boldsymbol{V}^{\perp,\mathsf{H}}_{R<Q} \end{array} \right] \tag{B.50}$$

$$\boldsymbol{A}_{M<N} = \left[\boldsymbol{U}, \boldsymbol{U}^{\perp}_{R<Q}\right] \left[\begin{array}{cc} \left[\begin{array}{cc} \boldsymbol{\Sigma}_{R\times R} & \boldsymbol{0}_{R\times M-R} \\ \boldsymbol{0}_{M-R\times R} & \boldsymbol{0}_{M-R\times M-R} \end{array} \right] & \left[\begin{array}{c} \boldsymbol{0}_{R\times N-M} \\ \boldsymbol{0}_{M-R\times N-M} \end{array} \right] \end{array} \right] \left[\begin{array}{c} \boldsymbol{V}^{\mathsf{H}} \\ \boldsymbol{V}^{\perp,\mathsf{H}}_{R<Q} \\ \boldsymbol{V}^{\perp,\mathsf{H}}_{Q<N} \end{array} \right]. \tag{B.51}$$

Fundamental Subspaces (continued)

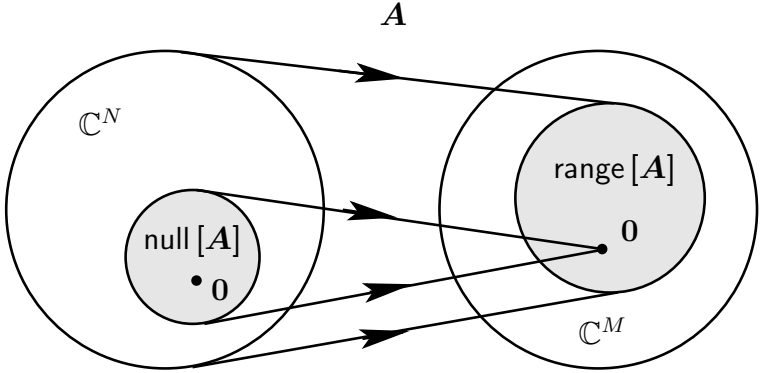
Any Linear Operator $\boldsymbol{A} : \mathbb{C}^N \to \mathbb{C}^M$ is related to Four Fundamental Subspaces, viz.

$$\text{Image Space of } \boldsymbol{A}: \quad \mathsf{range}\,[\boldsymbol{A}] = \left\{ \boldsymbol{A}\boldsymbol{x} \middle| \, \boldsymbol{x} \in \mathbb{C}^N \right\} \tag{B.52}$$

$$\text{Null Space of } \boldsymbol{A}: \quad \mathsf{null}\,[\boldsymbol{A}] = \left\{ \boldsymbol{x} \in \mathbb{C}^N \middle| \, \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \right\} \tag{B.53}$$

$$\text{Orthogonal Complement of } \mathsf{range}\,[\boldsymbol{A}]: \quad \mathsf{range}\,[\boldsymbol{A}]^{\perp} = \mathsf{null}\left[\boldsymbol{A}^{\mathsf{H}}\right] \tag{B.54}$$

$$\text{Orthogonal Complement of } \mathsf{null}\,[\boldsymbol{A}]: \quad \mathsf{null}\,[\boldsymbol{A}]^{\perp} = \mathsf{range}\left[\boldsymbol{A}^{\mathsf{H}}\right]. \tag{B.55}$$



ONB.

The column vectors of $\boldsymbol{U}_{\text{ext}} = \left[\boldsymbol{U}, \boldsymbol{U}^{\perp}\right]$ with $\boldsymbol{U}^{\perp} = \left[\boldsymbol{U}^{\perp}_{R<Q}, \boldsymbol{U}^{\perp}_{Q<M}\right]$ and $\boldsymbol{V}_{\text{ext}} = \left[\boldsymbol{V}, \boldsymbol{V}^{\perp}\right]$ with $\boldsymbol{V}^{\perp} = \left[\boldsymbol{V}^{\perp}_{R<Q}, \boldsymbol{V}^{\perp}_{Q<N}\right]$ from the EXTENDED SVD of a rank-R matrix $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ constitute the ORTHONORMAL BASES for the FOUR FUNDAMENTAL SUBSPACES range$[\boldsymbol{A}]$, range$[\boldsymbol{A}]^{\perp}$, null$[\boldsymbol{A}]$, and null$[\boldsymbol{A}]^{\perp}$, viz.

$$\begin{aligned}
\text{range}\left[\boldsymbol{A}\right] &= \text{span}\left[\boldsymbol{U}_{\text{ext}}\boldsymbol{e}_i \mid i = 1, \ldots, R\right] && \text{(B.56)}\\
\text{range}\left[\boldsymbol{A}\right]^{\perp} &= \text{span}\left[\boldsymbol{U}_{\text{ext}}\boldsymbol{e}_i \mid i = R+1, \ldots, Q\right] \oplus \text{span}\left[\boldsymbol{U}_{\text{ext}}\boldsymbol{e}_i \mid i = Q+1, \ldots, M\right] && \text{(B.57)}\\
\text{null}\left[\boldsymbol{A}\right] &= \text{span}\left[\boldsymbol{V}_{\text{ext}}\boldsymbol{e}_i \mid i = R+1, \ldots, Q\right] \oplus \text{span}\left[\boldsymbol{V}_{\text{ext}}\boldsymbol{e}_i \mid i = Q+1, \ldots, N\right] && \text{(B.58)}\\
\text{null}\left[\boldsymbol{A}\right]^{\perp} &= \text{span}\left[\boldsymbol{V}_{\text{ext}}\boldsymbol{e}_i \mid i = 1, \ldots, R\right]. && \text{(B.59)}
\end{aligned}$$

The FOUR FUNDAMENTAL SUBSPACES of the matrices $\boldsymbol{A} : \mathbb{C}^N \to \mathbb{C}^M$ and $\boldsymbol{A}^{\mathsf{H}} : \mathbb{C}^M \to \mathbb{C}^N$ form ORTHOGONAL COMPLEMENTS of the N- and M-dimensional VECTOR SPACES:

$$\begin{aligned}
\mathbb{C}^N &= \text{null}\left[\boldsymbol{A}\right] \oplus \text{range}\left[\boldsymbol{A}^{\mathsf{H}}\right], && \text{(B.60)}\\
\mathbb{C}^M &= \text{range}\left[\boldsymbol{A}\right] \oplus \text{null}\left[\boldsymbol{A}^{\mathsf{H}}\right]. && \text{(B.61)}
\end{aligned}$$

C. Kalman Filter

C.1 Recursive Computation of Conditional Means and Covariances

a) Exploiting the Chapman-Kolmogorov Equation (9.11) and the Linear State Space Model in Eq. (10.1), we obtain

$$
\begin{aligned}
\hat{\boldsymbol{x}}_{n|n-1} &= \int_{\mathbb{X}} \boldsymbol{x}_n \; f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n-1)}) \, \mathrm{d}\boldsymbol{x}_n \\
&= \int_{\mathbb{X}} \int_{\mathbb{X}} \boldsymbol{x}_n \; f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \; f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)}) \, \mathrm{d}\boldsymbol{x}_n \, \mathrm{d}\boldsymbol{x}_{n-1} \\
&= \int_{\mathbb{X}} \underbrace{\int_{\mathbb{X}} \boldsymbol{x}_n \; f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \, \mathrm{d}\boldsymbol{x}_n}_{\text{cond. expectation of } \boldsymbol{x}_n \text{ (refer to 10.1)}} \; f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)}) \, \mathrm{d}\boldsymbol{x}_{n-1} \\
&= \boldsymbol{G}_n \underbrace{\int_{\mathbb{X}} \boldsymbol{x}_{n-1} \; f_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_{n-1}|\boldsymbol{y}_{(n-1)}) \, \mathrm{d}\boldsymbol{x}_{n-1}}_{\text{cond. expectation of } \boldsymbol{x}_{n-1}} = \boldsymbol{G}_n \; \boldsymbol{\mu}_{\boldsymbol{X}_{n-1}|\boldsymbol{Y}_{(n-1)}=\boldsymbol{y}_{(n-1)}}.
\end{aligned}
\tag{C.1}
$$

The respective CONDITIONAL STATE COVARIANCE MATRIX of $\boldsymbol{X}_n$ is obtained as

$$\begin{aligned}
\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} &= \mathsf{E}\left[\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1}\right)\left(\boldsymbol{X}_n - \hat{\boldsymbol{x}}_{n|n-1}\right)^{\mathsf{T}} \middle| \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \mathsf{E}\left[\left(\boldsymbol{G}_n(\boldsymbol{X}_{n-1} - \hat{\boldsymbol{x}}_{n-1|n-1}) + \boldsymbol{V}_n\right)\left(\boldsymbol{G}_n(\boldsymbol{X}_{n-1} - \hat{\boldsymbol{x}}_{n-1|n-1}) + \boldsymbol{V}_n\right)^{\mathsf{T}} \middle| \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \mathsf{E}\left[\left(\boldsymbol{G}_n(\boldsymbol{X}_{n-1} - \hat{\boldsymbol{x}}_{n-1|n-1})\right)\left(\boldsymbol{G}_n(\boldsymbol{X}_{n-1} - \hat{\boldsymbol{x}}_{n-1|n-1})\right)^{\mathsf{T}} \middle| \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] + \\
&\quad \mathsf{E}\left[\left(\boldsymbol{V}_n\right)\left(\boldsymbol{V}_n\right)^{\mathsf{T}} \middle| \boldsymbol{Y}_{(n-1)} = \boldsymbol{y}_{(n-1)}\right] \\
&= \boldsymbol{G}_n \boldsymbol{C}_{\boldsymbol{X}_{n-1|n-1}} \boldsymbol{G}_n^{\mathsf{T}} + \boldsymbol{C}_{\boldsymbol{V}_n}, \qquad \text{(C.2)}
\end{aligned}$$

since $\boldsymbol{X}_{n-1} - \hat{\boldsymbol{x}}_{n-1|n-1}$ and $\boldsymbol{V}_n$ are obviously stochastically independent.

b) Since both sides of the BAYES RULE in (9.12) are Gaussian PDFs, a comparison of the exponents,[1]

$$\log\left(f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n)})\right) \propto \log\left(f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}(\boldsymbol{y}_n|\boldsymbol{x}_n)\right) + \log\left(f_{\boldsymbol{X}_n|\boldsymbol{Y}_{(n-1)}}(\boldsymbol{x}_n|\boldsymbol{y}_{(n-1)})\right) + \ldots,$$

and taking into account the LINEAR OBSERVATION MODEL in Eq. (10.2) results in

$$\hat{\boldsymbol{x}}_{n|n} = \hat{\boldsymbol{x}}_{n|n-1} + \boldsymbol{K}_n\left(\boldsymbol{y}_n - \boldsymbol{H}_n \hat{\boldsymbol{x}}_{n|n-1}\right), \qquad \text{(C.3)}$$

$$\boldsymbol{C}_{\boldsymbol{X}_{n|n}} = \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} - \boldsymbol{K}_n \boldsymbol{H}_n \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}, \qquad \text{(C.4)}$$

with the KALMAN GAIN MATRIX

$$\boldsymbol{K}_n = \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} \boldsymbol{H}_n^{\mathsf{T}} \left(\boldsymbol{H}_n \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} \boldsymbol{H}_n^{\mathsf{T}} + \boldsymbol{C}_{\boldsymbol{W}_n}\right)^{-1}. \qquad \text{(C.5)}$$

[1] Note, that we are using (9.12) instead of (9.13).

C.2 Proof of the Multivariate Version

From a comparison of the exponents of both sides of (9.12), we obtain

$$
\begin{aligned}
&(\boldsymbol{x}_n - \boldsymbol{\mu}_{\boldsymbol{X}_{n|n}})^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{X}_{n|n}}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_{\boldsymbol{X}_{n|n}}) \propto \\
&(\boldsymbol{y}_n - \underbrace{\boldsymbol{\mu}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}}_{\text{refer to 10.2}})^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}^{-1} (\boldsymbol{y}_n - \boldsymbol{\mu}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}) + (\boldsymbol{x}_n - \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}})^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}}) = \\
&(\boldsymbol{y}_n - \boldsymbol{H}_n \boldsymbol{x}_n)^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}^{-1} (\boldsymbol{y}_n - \boldsymbol{H}_n \boldsymbol{x}_n) + (\boldsymbol{x}_n - \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}})^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}}),
\end{aligned}
$$

and also

$$
\begin{aligned}
&\boldsymbol{x}_n^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{X}_{n|n}}^{-1} \boldsymbol{x}_n - 2\boldsymbol{x}_n^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{X}_{n|n}}^{-1} \boldsymbol{\mu}_{\boldsymbol{X}_{n|n}} + \cdots \propto \\
&\boldsymbol{x}_n^{\mathsf{T}} \left(\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}^{-1} + \boldsymbol{H}_n^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}^{-1} \boldsymbol{H}_n \right) \boldsymbol{x}_n - 2\boldsymbol{x}_n^{\mathsf{T}} \left(\boldsymbol{H}_n^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}^{-1} \boldsymbol{y}_n + \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}^{-1} \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} \right) + \dots \ .
\end{aligned}
$$

By comparing the respective terms, we eventually obtain

$$
\begin{aligned}
\boldsymbol{C}_{\boldsymbol{X}_{n|n}} &= \left(\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}^{-1} + \boldsymbol{H}_n^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}^{-1} \boldsymbol{H}_n \right)^{-1} \\
\boldsymbol{\mu}_{\boldsymbol{X}_{n|n}} &= \boldsymbol{C}_{\boldsymbol{X}_{n|n}} \left(\boldsymbol{H}_n^{\mathsf{T}} \boldsymbol{C}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}^{-1} \boldsymbol{y}_n + \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}^{-1} \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} \right).
\end{aligned}
$$

Finally, the MATRIX INVERSION LEMMA[2] yields

$$\begin{aligned}
\boldsymbol{C}_{\boldsymbol{X}_{n|n}} &= \left(\boldsymbol{C}^{-1}_{\boldsymbol{X}_{n|n-1}} + \boldsymbol{H}_n^{\mathsf{T}}\boldsymbol{C}^{-1}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}\boldsymbol{H}_n\right)^{-1} \\
&= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} - \underbrace{\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}}\overbrace{\left(\underbrace{\boldsymbol{C}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}}_{\boldsymbol{C}_{\boldsymbol{W}_n}} + \boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}}\right)^{-1}}^{\boldsymbol{C}^{-1}_{\boldsymbol{Y}_{n|n-1}} \quad (*)}}_{\text{Kalman Gain: } \boldsymbol{K}_n}\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{X}_{n|n}} &= \left(\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}} - \boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\right)\left(\boldsymbol{H}_n^{\mathsf{T}}\boldsymbol{C}^{-1}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}\boldsymbol{y}_n + \boldsymbol{C}^{-1}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}}\right) \\
&= \boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}}\boldsymbol{C}^{-1}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}\boldsymbol{y}_n + \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} - \boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}}\boldsymbol{C}^{-1}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}\boldsymbol{y}_n - \boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} \\
&= \boldsymbol{K}_n\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}}\boldsymbol{C}^{-1}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}\boldsymbol{y}_n + \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} - \boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}}\boldsymbol{C}^{-1}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}\boldsymbol{y}_n - \boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} \\
&= \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} + \boldsymbol{K}_n\underbrace{\left(\boldsymbol{C}_{\boldsymbol{Y}_{n|n-1}} - \boldsymbol{H}_n\boldsymbol{C}_{\boldsymbol{X}_{n|n-1}}\boldsymbol{H}_n^{\mathsf{T}}\right)}_{\boldsymbol{C}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n} \text{ (refer to } (*))}\boldsymbol{C}^{-1}_{\boldsymbol{Y}_n|\boldsymbol{X}_n=\boldsymbol{x}_n}\boldsymbol{y}_n - \boldsymbol{K}_n\boldsymbol{H}_n\boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} \\
&= \boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}} + \boldsymbol{K}_n\left(\boldsymbol{y}_n - \boldsymbol{H}_n\boldsymbol{\mu}_{\boldsymbol{X}_{n|n-1}}\right).
\end{aligned}$$

[2] $(\boldsymbol{A}+\boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{C}^{-1} + \boldsymbol{DA}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{DA}^{-1}$.

References

- A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes*, 4th edition, Mc Graw Hill, 2002.
- H. Stark and J. W. Woods. *Probability, Random Processes, and Estimation Theory for Engineers*, 2nd edition, Prentice Hall, 1994.
- W. Utschick. *Stochastische Signale*, Aktuelles Manuskript zur gleichnamigen Vorlesung an der Technischen Universität München, 2012.
- G. Strang. *Linear Algebra and Its Applications*, Harcourt Brace Jovanovich, Publishers, 1988.
- L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*, Society for Industrial and Applied Mathematic, 1997.
- W. Utschick. *Mathematische Methoden der Signalverarbeitung*, Aktuelles Manuskript zur gleichnamigen Vorlesung an der Technischen Universität München, 2012.