TTK4250

Lecture 1

Probability for sensor fusion

Edmund Førland Brekke

19. August 2019

Outline

Formalia

Instructors.

- Lecturer: Edmund Førland Brekke, edmund.brekke@ntnu.no, Rom D335.
- Scientific assistant: Lars-Christian Ness Tokle, lars-christian.n.tokle@ntnu.no.

Blackboard.

All information about the course will be made available on Blackboardso it is recommended to visit this regularly. (Ensure that you receive notifications)

Reference group.

- Reference groups are an important part of NTNUs quality assurance of education
- Referansegruppen consists of three students, with the responsibility to
  - have three meetings during the semester with the lecturer.
  - have a dialogue with the class during the semester.
  - write a reference group report that sums up the students' opinions and recommendations for improvement of the course. This report is included in the final course report.

Expectations and prerequisites

About this course:

This will introduce you to the **probabilistic framework** that dominates in sensor fusion. The course will not cover very much of sensor fusion itself, but there will be a strong focus on **algorithms and methods** that play a central role in typical sensor fusion applications, such as **target tracking**, **inertial navigation** and **SLAM**.

Background and prerequisites that I expect you to have:

- You must have had linear systems theory or any other course where you have either became familiar with the **Kalman filter** or **multivariate Gaussians**.
- The course builds directly on TMA4240 Statistics, but key concepts that often are forgotten will be repeated in the first lectures.
- You are recommended to have additional background in statistics and probability. TTT4275 - Estimation, Detection and Classification is an ideal background, but any statistics course from IMF will also be very useful.
- The course will include 3 fairly challenging programming assignments to be solved in Matlab. If you perceive Matlab more as a challenge then as a useful tool, then you probably need to brush up your Matlab skills immediately.

Overview

Teaching material.

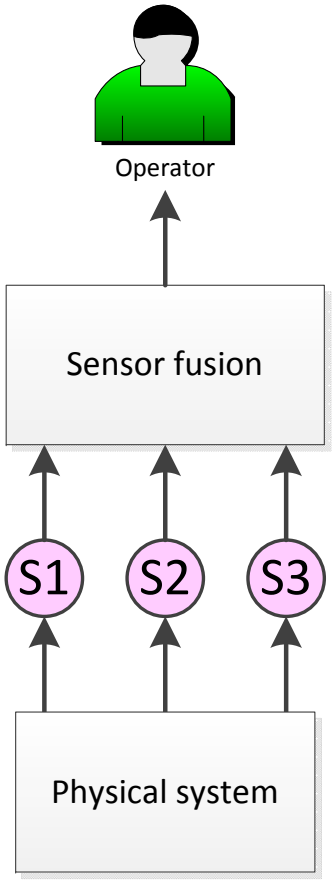
- Edmund Brekke: *Fundamentals of sensor fusion*, 2019 (in writing).
- All lecture notes, assignments and solutions to the assignments.

Topics that we will study.

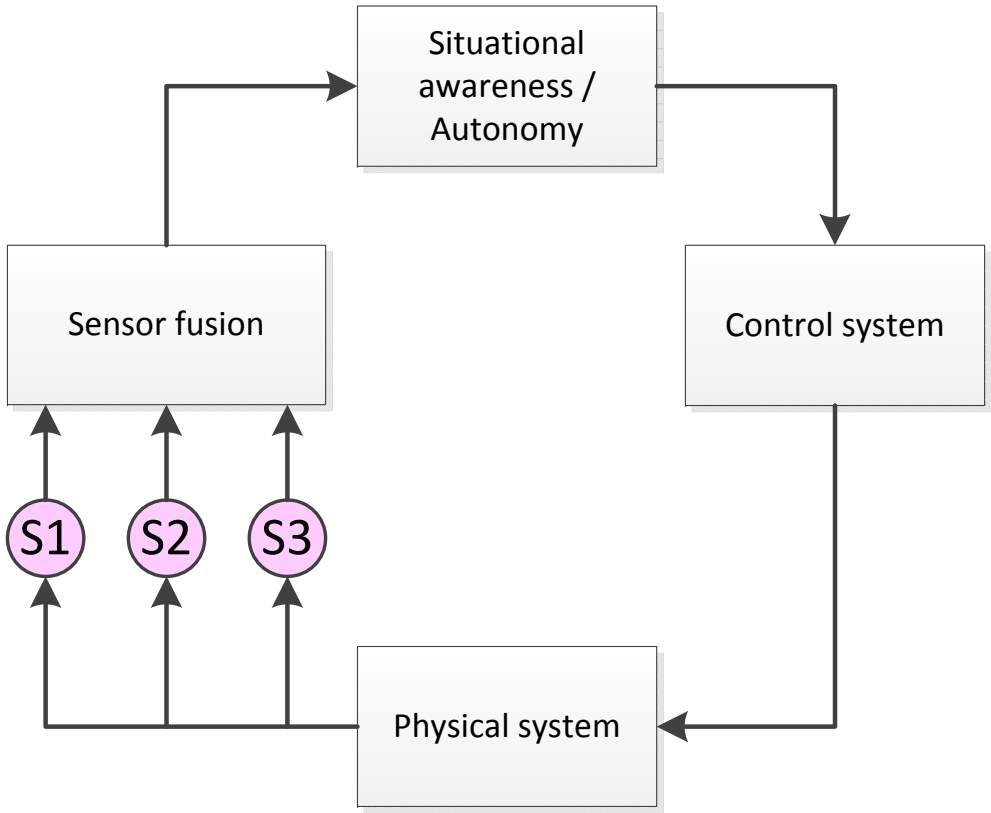
Kapitler i Balchen et al. som er pensum:

1. Probability and estimation.
2. The multivariate Gaussian.
3. The Kalman filter.
4. Nonlinear filters: EKF and particle filters.
5. The Interacting Multiple Models (IMM) method.
6. Single-target tracking.
7. Multi-target tracking.
8. Inertial navigation.
9. Simultaneous localization and mapping (SLAM).

Two roles of sensor fusion



Sensor fusion for surveillance and decision support

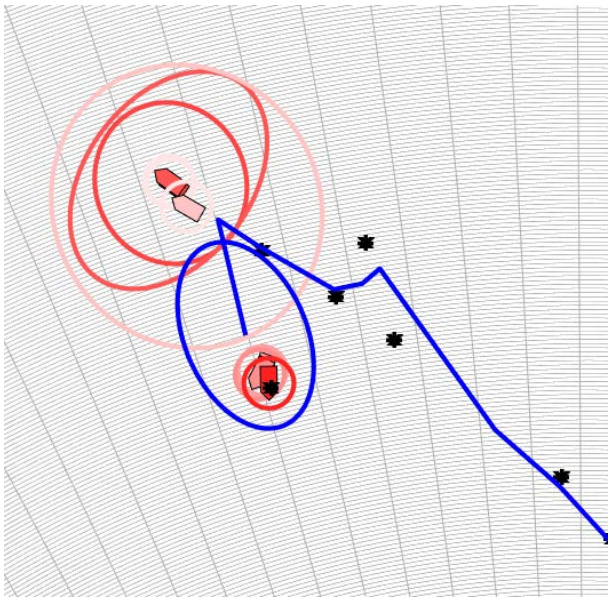


Sensor fusion in a closed-loop system

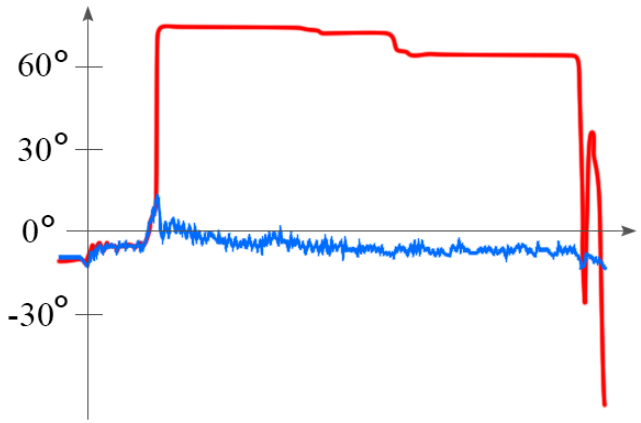
Why is sensor fusion important?

- Because challenges are non-trivial.
- Because poorly designed methods can have disastrous consequences.

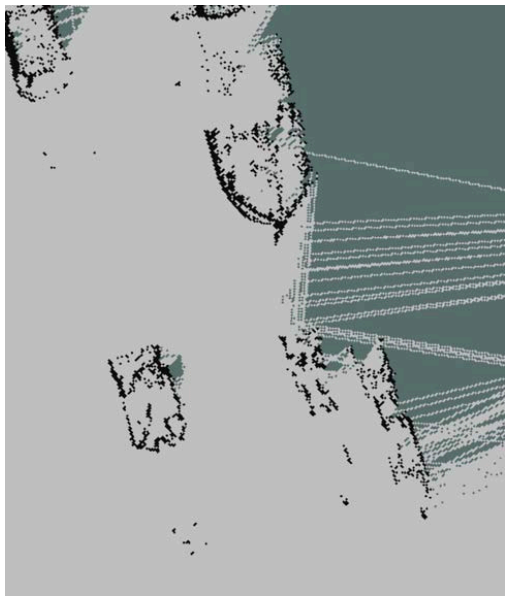
Competing data interpretations



Are estimates physically reasonable?



Quantify lack of knowledge



Why probability?

Most industry-standard methods in target tracking and navigation, and to a large extent also in SLAM, are based on probability theory.

- Probability theory is a language for quantifying uncertainty.

$\longrightarrow$ Enable the algorithms to **hedge** on different possibilities in the same way as a hedge fund manager.

$\longrightarrow$ The uncertainty of the algorithms can have implications for what decisions are rational.

- Correlations, Bayes' rule and other probabilistic tools can help us make inference when limited/noisy data is available.

The axioms of probability

- A probability $\Pr\{\cdot\}$ is **something that we assign to an event** E.
- The union of all possible events is called the outcome space Ω.
- The probability obeys the following three axioms:

1. $\Pr\{E\} \in \mathbb{R}$, $\Pr\{E\} \geq 0$.
2. $\Pr\{\Omega\} = 1$.
3. For any sequence of disjoint events $E_1, \ldots, E_n$ we have

$$\Pr\left\{\bigcup_{i=1}^{n} E_i\right\} = \sum_{i=1}^{n} \Pr\{E_i\}.$$

Bread-and-butter definitions and rules

Definition: Conditional probability

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

Definition: Independence

The events A and B are independent if $\Pr\{A \cap B\} = \Pr\{A\}\Pr\{B\}$.

The total probability theorem

$$\Pr\{A\} = \sum_n \Pr\{A|B_n\}\Pr\{B_n\}$$

Bayes' rule

$$\Pr\{A|B\} = \frac{\Pr\{B|A\}\Pr\{A\}}{\Pr\{B\}}$$

Real analysis: σ-algebras and measures

Why study tools from real analysis?

- Working directly in terms of probabilities is cumbersome.
- It is much more convenient to work with random variables and their probability distributions.
- Mathematicians define the concept of a random variable in terms of these concepts.

σ-algebra

A σ-algebra $\mathcal{F}$ on a set Ω is a collection of subsets of Ω such that

- Ω itself is a member of $\mathcal{F}$.
- If the subset $A \subseteq \Omega$ is a member of $\mathcal{F}$, then its complement $\Omega \backslash A$ is also a member of $\mathcal{F}$.
- If A_1, A_2, A_3, etc are members of $\mathcal{F}$, then the union $\bigcup_n A_n$ is also a member of $\mathcal{F}$.

In probability theory, the elements of the σ-algebra play the role of meaningful events.

Real analysis: σ-algebras and measures

Example: Throws of a dice

The outcome space when we throw a dice is $\Omega = \{1, 2, 3, 4, 5, 6\}$. All the possible subsets of these 6 elements constitute a corresponding σ-algebra:

$$\mathcal{F} = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \{1, 2, 3, 4, 5, 6\}\}. \tag{1}$$

For a single throw, the number of the dice could be part of $\{1\}$, $\{2\}$, $\{1, 2\}$, etc.

Definition: Probability measure

A probability measure P on the σ-algebra $\mathcal{F}$ is a function from $\mathcal{F}$ to the unit interval $[0, 1]$ that obeys the three axioms of probability.

Example: Throws of a dice

For the events listed in (1) the probability measure that describes a fair dice should return

$$P(\{1\}) = 1/6, \quad P(\{2\}) = 1/6, \quad \ldots,$$
$$P(\{1, 2\}) = 1/3, \quad \ldots \quad , \quad P(\{1, 2, 3, 4, 5, 6\}) = 1.$$

Random variables

The mathematical approach to probability theory distinguishes between an **abstract** outcome space Ω and the outcome space that we really are interested.

Definition: Random variable

A random variable X is a function from Ω into another space $\mathbb{O}$, which we henceforth shall know as the outcome space.

This promotes us to be careful in exploring how probability is mapped from Ω onto $\mathbb{O}$.

Definition: Probability measure of a random variable

The probability measure of X is a function $\beta_X(\cdot)$ from subsets of $\mathbb{O}$ to $[0, 1]$ so that

$$\beta_X(S) = \Pr\{X \in S\} = P(X^{-1}(S))$$

Definition: Realization

The output of the random variable X for a particular $\omega \in \Omega$ is called a realization of X. We can write this as $x = X(\omega)$.

Probability distributions

Definition: Cumulative distribution function (cdf)

The cdf, denoted $P(x)$, of $X \in \mathbb{R}$ is the probability $\Pr\{X < x\}$.

Definition: Probability density function (pdf)

The pdf of the scalar random variable X is the derivative

$$p(x) = \frac{\partial P(x)}{\partial x}.$$

Pdf's and probabilities are two different things. We get probabilities when we integrate a pdf over a subset of the outcome space.

The definitions are easily extended to $n = 2, 3, \ldots$ for $\mathbb{R}^n$:

$$P(x, y) = \Pr\{X \leq x, Y \leq y\}$$

$$p(x, y) = \frac{\partial^2}{\partial x \partial y} P(x, y) = \frac{\partial^2}{\partial y \partial x} P(x, y).$$

Examples of discrete probability distributions

- For a discrete random variable we can again define the cdf as $P(x) = \Pr\{X < x\}$.
- The pdf will then contain δ-spikes at all possible values in the outcome space.
- Since the outcome space is discrete, we can just as well think of the pdf as consisting of probabilities for a discrete random variable.

The Bernoulli distribution

A Bernoulli random variable with parameter r has a binary outcome space: $E = \{0, 1\}$, and its probability distribution is given by

$$p(x) = \begin{cases} 1 - r & \text{if } x = 0 \\ r & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

We write $p(x) = \text{Bernoulli}(x; r)$ to signify this distribution.

Examples of discrete probability distributions

The Binomial distribution

A binomial random variable with parameters $r \in [0, 1]$ and $n \in \mathbb{N}$ has outcome space $\{0, \ldots, n\}$ and its probability distribution is given by

$$p(x) = \binom{n}{x} r^x (1 - r)^{n-x}. \tag{3}$$

The Poisson distribution

A Poisson random variable with parameter λ has the countable (that means infinite, but discrete) outcome space $\{0, 1, 2, 3, \ldots\}$ and its probability distribution is given by

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{4}$$

We write $p(x) = \text{Poisson}(x; \lambda)$ to signify this distribution.

Examples of continuous probability distributions

The uniform distribution

A uniformly distributed random variable X on the interval $[a, b]$ has the pdf

$$p(x) = \text{Uniform}(x\,;\,[a,b]) = \frac{1}{b-a}\chi_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The Gaussian distribution

A Gaussian random variable with expectation μ and variance σ^2 has the pdf

$$p(x) = \mathcal{N}(x\,;\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{6}$$

The cdf of the Gaussian does not exist in closed form. It is typically expressed in terms of the so-called error function according to

$$P(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \quad \text{where} \quad \text{erf}(x) = \frac{1}{\sqrt{\pi}}\int_0^x \exp(-t^2)\mathrm{d}t. \tag{7}$$

Examples of continuous probability distributions

The Gamma distribution

A Gamma random variable with shape parameter k and scale parameter θ has the pdf

$$p(x) = \text{Gamma}(x\,;\,k,\theta) = \frac{x^{k-1}\exp(-x/\theta)}{\theta^k \Gamma(k)}. \tag{8}$$

Several special cases are of importance.

- The Rayleigh distribution results if $k = 2$ and $\theta = 2\sigma^2$:

$$p(x) = \frac{x}{\sigma^2}\exp\left(\frac{-x^2}{2\sigma^2}\right)$$

- The exponential distribution results if $k = 1$ and $\theta = 1/\lambda$:

$$p(x) = \lambda e^{-\lambda x}$$

- The χ^2 distribution results if $k = \frac{n}{2}$ and $\theta = 2$:

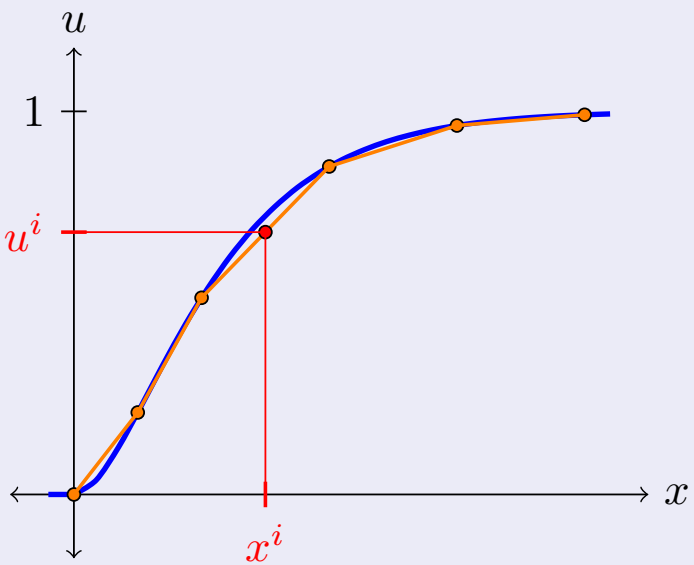
$$p(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1}\exp\left(-\frac{x}{2}\right). \tag{9}$$

Sampling from continuous probability distributions

Option 1

Use inbuilt functions such as `randn(4,1)` and `rand(4,1)`, etc.

Option 2



Use cdf inversion, also known as the Smirnov transform or inverse transform sampling.

Inverse transform sampling is easy if the cdf is invertible in closed form, but can also be used by means of interpolation if this is not the case.

More complicated random variables can often be simulated by exploiting their relationships to other random variables.

Moments

The most important moments are the expectation and the covariance, which for random vectors are

$$E[X] = \int \mathbf{x} p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$\mathrm{Var}[X] = E[(X - E[X])(X - E[X])^{\mathsf{T}}] = \int (\mathbf{x} - E[X])(\mathbf{x} - E[X])^{\mathsf{T}} p(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

These are related to, but not the same as, the sample mean and the sample covariance

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

$$\mathbf{P} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathsf{T}}$$

Moments are useful because the summarize information about the pdf in a single number (vector, matrix, etc.).

Useful rules for expectation and variance

- Expectation of linear combinations

$$E[\mathbf{A}X + \mathbf{B}Y] = \mathbf{A}E[X] + \mathbf{B}E[Y].$$

- Variance of linear combinations

$$\mathrm{Var}[\mathbf{A}X + \mathbf{B}Y] = \mathbf{A}\mathrm{Var}(X)\mathbf{A}^\mathsf{T} + \mathbf{B}\mathrm{Var}(Y)\mathbf{B}^\mathsf{T} + \mathbf{A}\mathrm{Cov}(X, Y)\mathbf{B}^\mathsf{T} + \mathbf{B}\mathrm{Cov}(Y, X)\mathbf{A}^\mathsf{T}.$$

- The law of total expectation

$$E_X[X] = E_Y[E_X[X|Y]].$$

- Independence implies that ...

$$E_{X,Y}[XY] = E_X[E_Y[XY]] = E_Y[E_X[X|Y]Y] = E_X[X]E_Y[Y].$$

- Jensen's inequality: For any convex function $f(\cdot)$ the following must hold:

$$f(E[X]) \leq E[f(X)].$$

Higher-order moments and heavytailedness

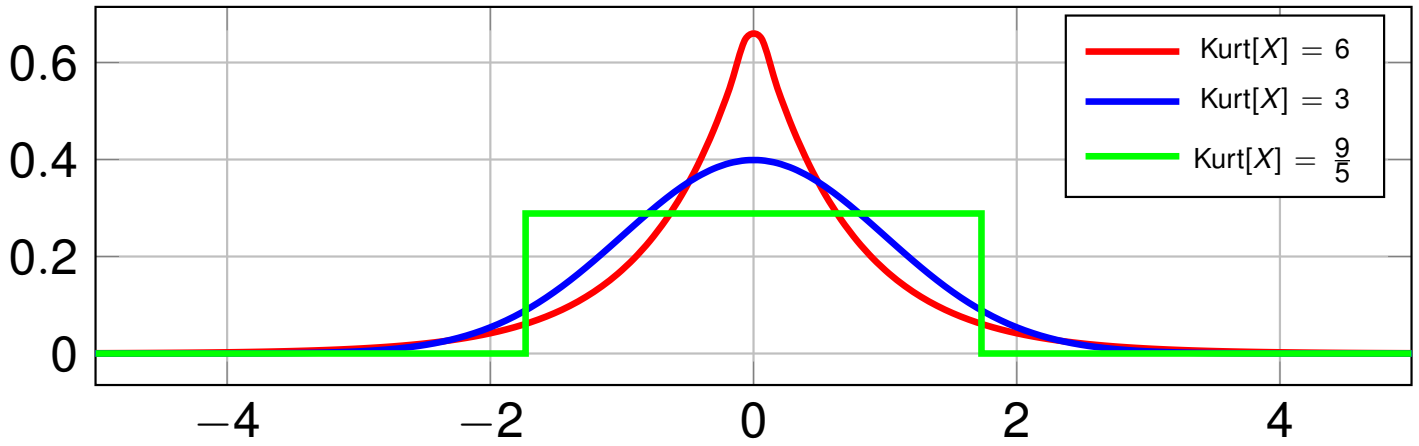
Third and fourth order moments

- The skewness tells us whether the distribution is more spread out on one side of the expectation:

$$\gamma = \frac{E[(X-\mu)^3]}{\sigma^3}.$$

- The kurtosis tells us whether the distribution has heavier or lighter tails than a Gaussian (reference value 3):

$$\text{Kurt}[X] = \frac{E[(X-\mu)^4]}{\sigma^4}$$



Generating functions

Moment-generating function

For a continuous random variable we use $M_X(s) = E_X[e^{sx}] = \int_{-\infty}^{\infty} p(x)e^{sx}\,\mathrm{d}x.$

Probability-generating function

For a discrete random variable we use $G(t) = E_X[t^X] = \sum_{n=-\infty}^{\infty} p(x_n)t^{x_n}.$

Generating functions are useful because

1. The generating function determines the distribution and vice versa, in a manner similar to Laplace- and Z-transforms.
2. The generating function of a sum of independent random variables is the product of the generating functions.
3. The moments can be found by differentiating the generating function.

Examples of use of generating functions

Sum of Gaussians

Let X and Y be two independent Gaussian RVs with expectations a and b, and covariances q and r. What is the distribution of $Z = X + Y$?

Sum of Exponentials

Let X_i, $i = 1, \ldots, N$ be N i.i.d. exponential RVs with parameter λ, and let

$$Y = \sum_{i=1}^{N} X_i.$$

What is the distribution of Y?

Generalizations

- For a vector-valued random variable the moment-generating function is

$$M_X(\mathbf{s}) = E_X[e^{\mathbf{s}^\mathsf{T}\mathbf{x}}] = \int_\infty p(\mathbf{x}) e^{\mathbf{s}^\mathsf{T}\mathbf{x}} \mathrm{d}\mathbf{x}.$$

- In the advanced course TK8102 we shall encounter **probability-generating functionals**.

Transformations of random variables

The other important tool that we have to derive new distributions is the following theorem.

Nonlinear transformations of random variables

Suppose that $\mathbf{y} = \mathbf{f}(\mathbf{x})$ where $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Denote the pdf's of $\mathbf{x}$ and $\mathbf{y}$ by $g(\mathbf{x})$ and $h(\mathbf{y})$, respectively. Then we have that

$$h(\mathbf{y}) = \sum_i g(\mathbf{f}_i^{-1}(\mathbf{y})) \left|\det(\mathbf{F}_i^{-1}(\mathbf{y}))\right|$$

where $\mathbf{f}_i^{-1}(\mathbf{y})$ range over all solutions of $\mathbf{y} = \mathbf{f}(\mathbf{x})$ with respect to $\mathbf{x}$, and $\mathbf{F}_i^{-1}(\mathbf{y})$ is the corresponding Jacobian matrix of the inverse mapping $\mathbf{f}_i^{-1}(\mathbf{y})$.

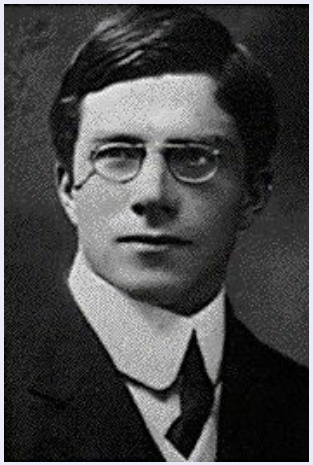
Square of zero-mean univariate Gaussian

If $X \sim \mathcal{N}(0, 1)$, what is then the pdf of $Y = X^2$?

The Bayesian and frequentist paradigms

The frequentist approach

- All probability should be interpretable as a frequency.
- In the frequentist mindset, a quantity to be estimated is not random. Only the data are random.



The Bayesian approach

- Bayesians are generally inclined to represent uncertainty probabilistically, even if this necessitates subjective assignments of probabilities.
- Sensor fusion techniques such as the Kalman filter are fundamentally Bayesian because they rely on *a priori* uncertainties (e.g., in the process model).

Estimators

Definition: Estimator.

We have some data z that were generated at random according to $p(z|x)$. We would like to infer knowledge about x from z. Let $x \in \mathcal{X}$ and let $z \in \mathcal{Z}$. An estimator is a function $\theta : \mathcal{Z} \to \mathcal{X}$ so that $f(z)$ gives an estimate of x.

Estimators that maximize PDFs/probabilities.

- The maximum-likelihood (ML) estimator is given by

$$\theta = \arg\max_x p(z|x).$$

- The maximum *a posteriori* (MAP) estimator is given by

$$\theta = \arg\max_x p(x|z) = \arg\max_x p(z|x)p(x).$$

To implement such estimators, various optimization techniques may be required:

- Steepest descent
- Newton
- Gauss-Newton
- Conjugate gradients
- MILP
- Viterbi algorithm
- Genetic algorithms
- Lagrangian Relaxation
- RANSAC

Estimators are random variables

Example 1: ML estimator of Rayleigh distribution parameter.

Let $\mathbf{z} = [z_1, \ldots, z_M]^\mathsf{T}$ consist of IID samples from a Rayleigh distribution:

$$p(\mathbf{z} \,|\, \eta) = \prod_{i=1}^{M} \frac{z_i}{\eta} \exp\left(-\frac{z_i^2}{\eta^2}\right).$$

By differentiating the logarithm of $p(\mathbf{z} \,|\, \eta)$ and equating the derivative to zero, we get the ML estimator

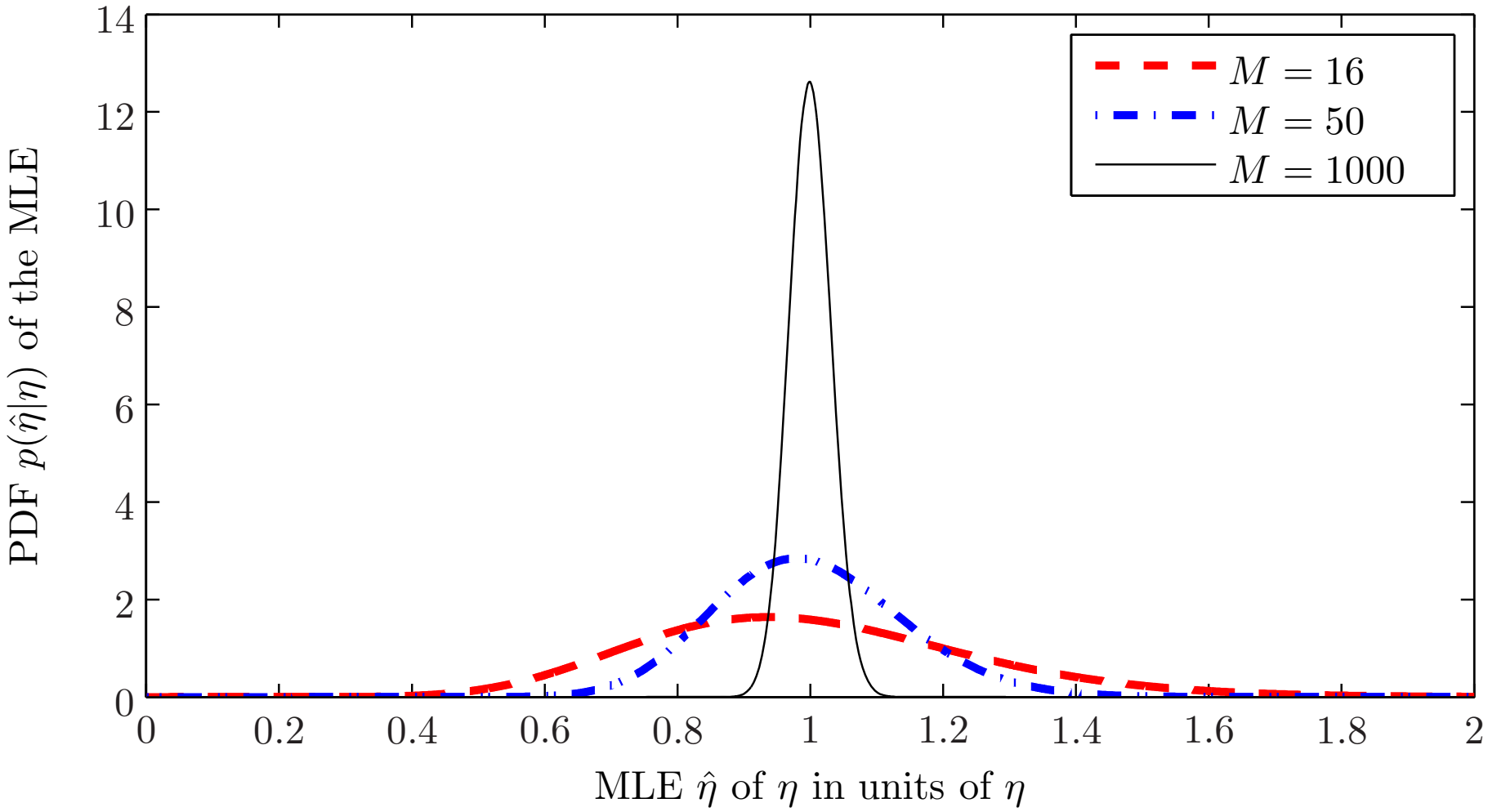
$$\hat{\eta} = \frac{1}{2M} \sum_{i=1}^{M} z_i^2.$$

This entity (a sum of IID exponential random variables) has a Gamma distribution:[a]

$$p(\hat{\eta} \,|\, \eta) = \text{Gamma}\left(\hat{\eta}\,;\, M, \frac{2\eta}{M}\right) = \frac{M^M}{\Gamma(M)} \left(\frac{\hat{\eta}}{\eta}\right)^M \exp\left(-M\frac{\hat{\eta}}{\eta}\right). \tag{10}$$

[a] See Papoulis & Pillai (2002): "Probability, Random Variables and Stochastic Processes" for more on Gamma, Rayleigh, Exponential and all kinds of distributions.

Estimators are random variables



Estimators are random variables

Example 2: MAP classifier beats all other classifiers.

Assume a Bayesian estimation problem with prior $p_X(x)$ and likelihood $p_{Z|X}(z|x)$ where X has a discrete outcome space[a]. Can we find an expression for how often, on average, an estimator θ will succeed in finding the correct value of X?

The success rate is given by

$$\begin{aligned} s &= \Pr\{\theta(Z) \text{ succeeds}\} \\ &= \int \sum_t \Pr\{\theta(z) \text{ succeeds} \mid x_t, z\} P_{X|Z}(x_t \mid z) p_Z(z)\, \mathrm{d}z \\ &= \int \sum_t \delta_{\theta(z),x_t} P_{X|Z}(x_t \mid z) p_Z(z)\, \mathrm{d}z \\ &= \int P_{X|Z}(\theta(z) \mid z) p_Z(z)\, \mathrm{d}z \\ &= E_Z[P_{X|Z}(\theta(Z) \mid Z)] \end{aligned}$$

If $\theta(Z)$ is the MAP estimator, then $P_{X|Z}(\theta(Z) \mid Z) = \max_x P_{X|Z}(x \mid Z)$. Thus, the MAP estimator has a higher success rate than any other estimator.

[a] This is also known as a classification problem. See Figueiredo (2004): "Lecture Notes on Bayesian Estimation and Classification", available online.

Sufficient statistics

The concept of a statistics.

A statistic is a single measure of some attribute of a sample.

- Estimators are statistics. A statistic may or may not corresponding to a meaningful estimator.

A sufficient statistic $g(\mathbf{z})$ for the parameter $\mathbf{x}$ summarizes the information about $\mathbf{x}$ contained in the data $\mathbf{z}$.[1]

The factorization theorem.

If the likelihood of the parameter $\mathbf{x}$ can be factorized as

$$f(\mathbf{z} \mid \mathbf{x}) = f_1(g(\mathbf{z}), \mathbf{x}) f_2(\mathbf{z})$$

then $g(\mathbf{z})$ is sufficient for $\mathbf{x}$.

- The ML-estimate depends only on $g(\mathbf{z})$ and not on the whole data set $\mathbf{z}$.

Example: The Rayleigh distribution, continued.

For the Rayleigh distribution, we found that $\sum_i^M z_i^2$ was a sufficient statistic.

[1] See Papoulis & Pillai (2002) pp. 322-327, or Duda, Hart & Stork (2001): "Pattern Classification", pp. 102-107, for more on sufficient statisics and the factorization theorem.

LS and MMSE estimators

The least squares (LS) estimator.

For any estimation problem on the form $\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{w}$ the least squares estimator is

$$\hat{\mathbf{x}}_{\text{LS}} = \arg\min_{\mathbf{x}} \|\mathbf{z} - \mathbf{h}(\mathbf{x})\|_2^2$$

- The LS estimator does not make any assumptions about the measurement "noise" $\mathbf{w}$. It is therefore of a non-probabilistic nature.
- If $\mathbf{w}$ is IID multivariate Gaussian, then the LS estimator is identical to the MLE.

The minimum mean square error (MMSE) estimator.

This is the probabilistic counterpart of the LS estimator. It is given by

$$\hat{\mathbf{x}}_{\text{MMSE}} = \arg\min_{\hat{\mathbf{x}}} E\left[(\hat{\mathbf{x}} - \mathbf{x})^{\mathsf{T}}(\hat{\mathbf{x}} - \mathbf{x}) \,|\, \mathbf{z}\right] = E[\mathbf{x}|\mathbf{z}] = \int \mathbf{x} p(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{x}$$

MMSE estimator versus MAP estimator.

- MMSE and MAP are equal for all symmetric posterior PDFs.
- The MMSE estimator is a Bayes estimator which minimizes expected Bayes risk.
- Care should be exercised in choosing estimator if $p(\mathbf{x}|\mathbf{z})$ is multimodal or skewed.

Unbiasedness of estimators

Definition of unbiased estimator.

Let $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$ be the estimation error. We then require that $E[\tilde{\mathbf{x}}] = 0$.

- The MMSE estimator is always unbiased.
- The ML and MAP estimators are in general biased.

Example: MLE of variance versus joint MLE of expectation and variance.

Let the likelihood be $p(\mathbf{z} \mid x, \eta) = \frac{1}{(2\pi)^{k/2}\eta^{k/2}} \exp\left(-\frac{1}{2\eta}\sum_{j=1}^{k}(z_j - x)^2\right)$.

First assume that x is known. Then the MLE of η is $\hat{\eta} = \frac{1}{k}\sum_{j=1}^{k}(z_j - x)^2$ and its expectation is η.

Then, let both x and η be unknown. Their MLEs are $\hat{x} = \frac{1}{k}\sum_{j=1}^{k} z_j$ and $\hat{\eta} = \frac{1}{k}\sum_{j=1}^{k}(z_j - \hat{x})^2$. Their expectations are now $E[\hat{x}] = x$ and $E[\hat{\eta}] = \frac{k-1}{k}\eta$.

Unbiasedness is desirable. However, there may exist biased estimators with lower MSE, and there exist estimation problems where the requirement of unbiasedness will lead to unacceptable degradation in MSE.

Variance and MSE of estimators

Scalar case.

Variance and MSE are given by

$$\mathrm{Var}(\hat{x}) = E[(\hat{x} - E[\hat{x}])^2] \quad , \quad \mathrm{MSE}(\hat{x}) = E[(\hat{x} - x)^2].$$

From this it follows that

$$\mathrm{MSE}(\hat{x}) = \mathrm{Var}(\hat{x}) + \mathrm{Bias}(\hat{x}, x)^2.$$

Vector case.

The variance is given by

$$\mathrm{Cov}(\hat{\mathbf{x}}) = E\left[(\hat{\mathbf{x}} - E[\hat{\mathbf{x}}])(\hat{\mathbf{x}} - E[\hat{\mathbf{x}}])^{\mathsf{T}}\right]$$

The MSE is given by

$$\mathrm{MSE}(\hat{\mathbf{x}}) = E\left[(\hat{\mathbf{x}} - \mathbf{x})^{\mathsf{T}}(\hat{\mathbf{x}} - \mathbf{x})\right] = \mathrm{tr}\left(E[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^{\mathsf{T}}]\right)$$

Bayesian MSE.

- If the expectation is conditional on the data $\mathbf{z}$, we call it a conditional MSE.
- Otherwise, we call it an unconditional MSE.

LMMSE estimators

When the MMSE is too complicated we may settle for the best linear estimator. This entails finding the best estimator $\hat{\mathbf{x}}$ on the form

$$\hat{\mathbf{x}} = \mathbf{A}\mathbf{z} + \mathbf{b}$$

that mimimizes

$$\text{MSE}(\hat{\mathbf{x}}) = E\left(\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right).$$

This minimization problem has the solution

$$\hat{\mathbf{x}} = E[\mathbf{x}] + \text{Cov}(\mathbf{x}, \mathbf{z})\text{Cov}(\mathbf{z})^{-1}(\mathbf{z} - E[\mathbf{z}]). \tag{11}$$

To use the LMMSE estimator we need to know the first two moments (expectation and covariance) of the joint PDF $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}, \mathbf{x})p(\mathbf{x})$.

The MSE of the LMMSE estimator.

Let $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$ be the estimation error. The MSE is then given by the matrix

$$E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}] = \text{Cov}(\mathbf{x}) - \text{Cov}(\mathbf{x}, \mathbf{z})\text{Cov}(\mathbf{z})^{-1}\text{Cov}(\mathbf{x}, \mathbf{z})^{\mathsf{T}} \tag{12}$$

- Thus, the LMMSE estimator provides a simple measure of its own performance.
- You recognize (11) and (12) as the update step of the Kalman filter.
- (12) may be misleading if we have wrong values of $\text{Cov}(\mathbf{x})$, $\text{Cov}(\mathbf{x}, \mathbf{z})$ or $\text{Cov}(\mathbf{z})$.

Example of LMMSE estimation[2]

Sensitivity to incorrect prior variance

The random variable x with prior mean $\bar{x}$ and variance σ_0^2 is measured via $z = x + w$ where w is zero mean, with variance σ^2 and independent of x.

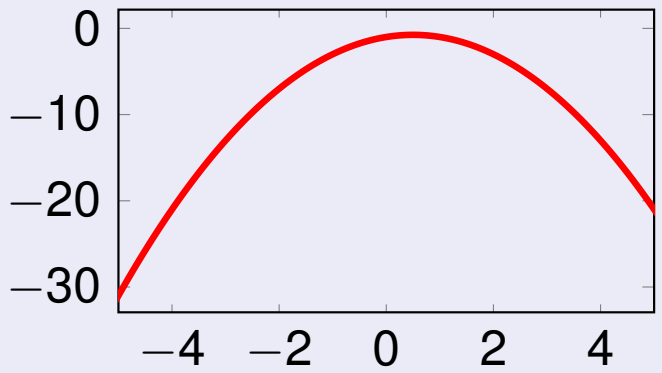
1) Write the LMMSE estimator $\hat{x}$ in terms of z and the MSE σ_1^2 associated with this estimator.
2) Write the estimate x^* of x as above but under the **incorrect** assumption that the prior variance is σ_p^2.
3) Find the actual MSE, σ_a^2 associated with 2), and the MSE σ_c^2 computed by the estimator in 2).
4) Verify the expression for σ_a^2 by inserting $\sigma_p^2 = \sigma_0^2$ and compare with σ_1^2.
5) Investigate how x^* behaves in the limits $s \to 0$ and $s \to \infty$.

[2] Adapted from Exercise 3.2 in Bar-Shalom, Kirubarajan & Li (2001)

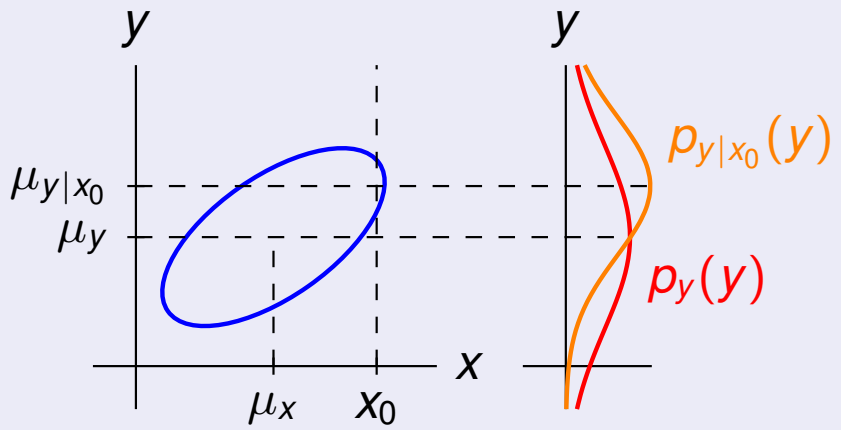
Next week

Next week's topic is the multivariate Gaussian distribution.

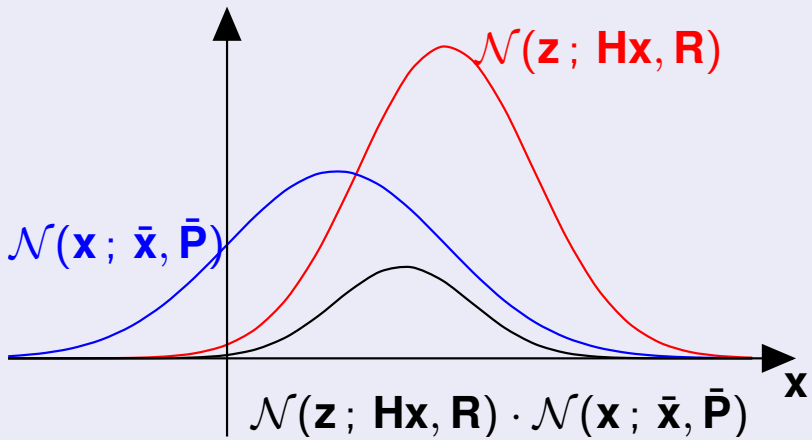
The importance of quadratic forms



Convenient manipulation rules



The product identity



Product identity $\longrightarrow$ Kalman filter