Short Review of Probability and Stochastic Processes

Puneet Singla, Manoranjan Majji
Department of Aerospace Engineering
Pennsylvania State University, University Park, PA-16802
Texas A&M University, College Station, TX-77843

Workshop: New Advances in Uncertainty Analysis & Estimation

Air-force Research Laboratories, Kirtland, NM

July 18-19, 2017

Basic Probability Concepts

- Probability are numbers assigned to events that indicate "how likely" it is that event will occur when a random experiment is performed.
  - The statement "E has probability P(E)" then mean that if we perform the experiment very often, it is practical certain that the relative frequency is approximately equal to P(E).
- What do we mean by *Relative Frequency?*
  - The relative frequency is at least equal to 0 and at most equal to 1.

$$0 \leq P(E) \leq 1 \tag{1}$$

  - Frequency Function: It shows how the values of the samples are distributed.

$$f(x) = \begin{cases} \bar{f}_j & when\ x = x_j \\ 0 & \text{for any value of } x \text{ not appearing in the sample} \end{cases}$$

  - Sample Distribution Function: $F(x) = \sum_{t \leq x} \bar{f}(t)$

Basic Probability Concepts

- Probability are numbers assigned to events that indicate "how likely" it is that event will occur when a random experiment is performed.
  - The statement "E has probability P(E)" then mean that if we perform the experiment very often, it is practical certain that the relative frequency is approximately equal to P(E).
- What do we mean by *Relative Frequency?*
  - The relative frequency is at least equal to 0 and at most equal to 1.

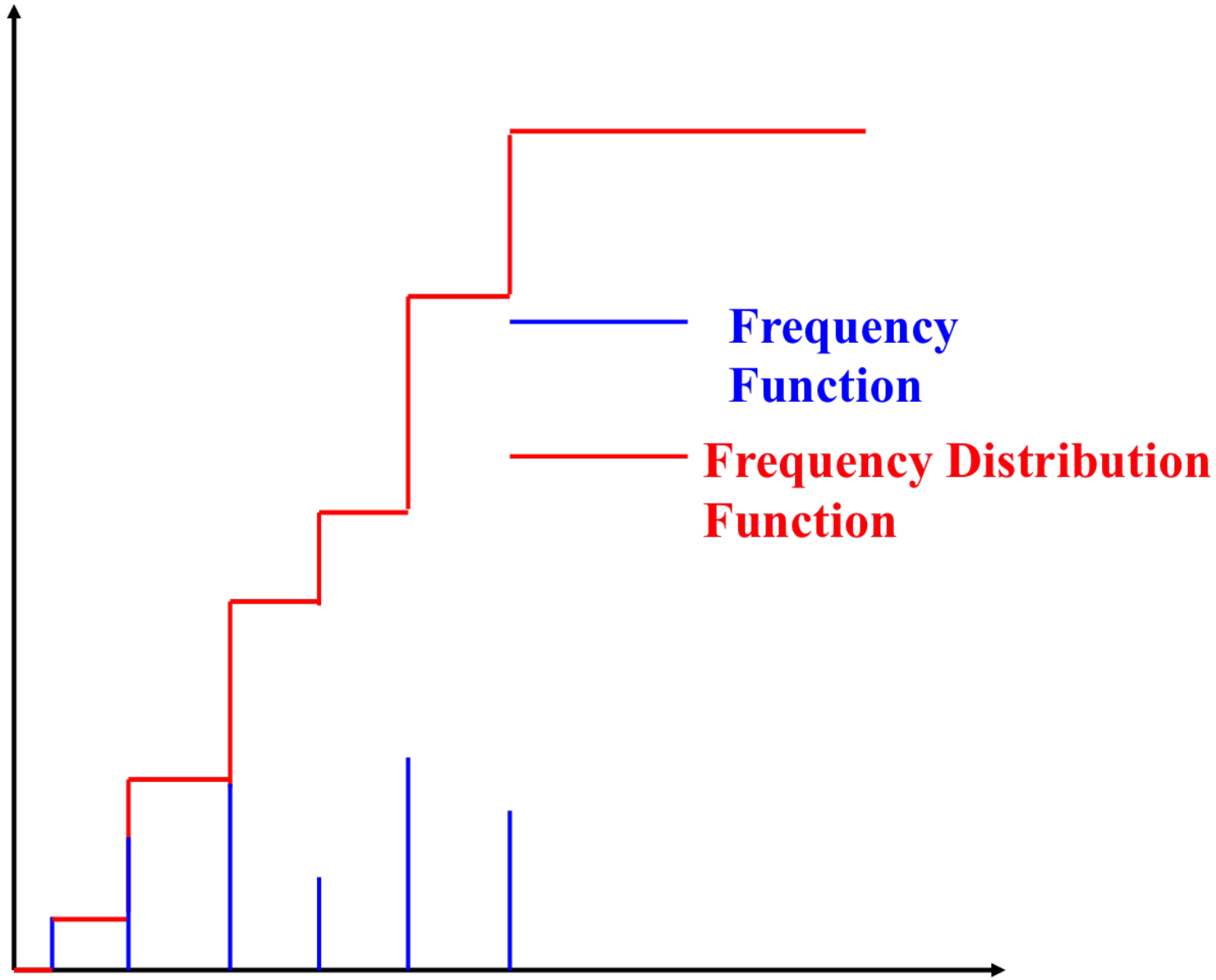
$$0 \leq P(E) \leq 1 \tag{1}$$

  - Frequency Function: It shows how the values of the samples are distributed.

$$f(x) = \begin{cases} \bar{f}_j & \text{when } x = x_j \\ 0 & \text{for any value of } x \text{ not appearing in the sample} \end{cases}$$

  - Sample Distribution Function: $F(x) = \sum_{t \leq x} \bar{f}(t)$

Basic Probability Concepts



Random Experiment

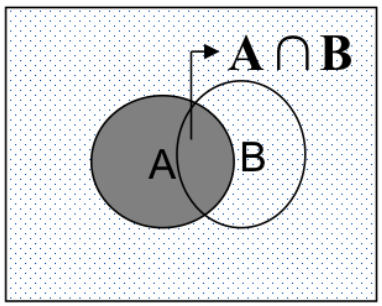
- Random Experiment or Random Observation:
  - It is performed according to a set of rules that determines the performance completely.
  - It can be repeated arbitrarily often.
  - The result of each performance depends on "chance" (that is, on influences which we can not control) and can therefore not be uniquely predicted.
- The result of single performance of the experiment is called the outcome of that experiment.
- The set of all possible outcomes of an experiment is called the sample space of the experiment.
- In most practical problems, we are not interested in the individual outcomes of the experiment but in whether an outcome belongs to a certain set of outcomes. Such a set is called an "Event"

Conditional Probability

- The probability of an event B under the condition that an event A occurs is given by

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

  - P(B/A) is called the conditional probability of B given A.

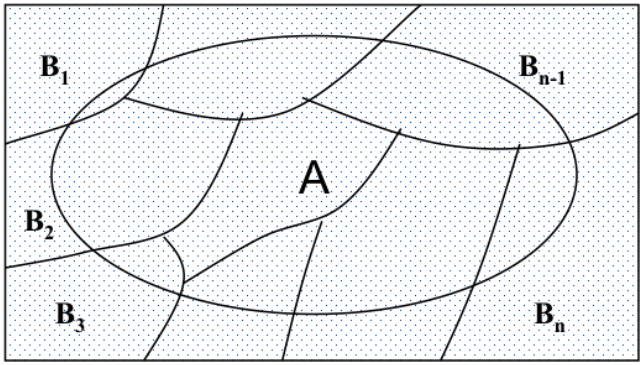


  - In this case, event A serves as a new sample space and event B becomes A∩B.
  - A and B are called independent event if

$$P(B/A) = P(B) \qquad P(A/B) = P(A)$$

$$P(A \cap B) = P(A)P(B)$$

Theorem of Total Probability



- Let $B_1, B_2 \ldots\ldots\ldots, B_n$ be mutually exclusive events s.t.

$$\bigcup_{i=1}^{n} B_i = S$$

- The probability of an event A can be represented as:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots\ldots\ldots\ldots + P(A \cap B_n)$$

- and, therefore

$$P(A) = P(A/B_1)P(B_1) + \ldots\ldots\ldots\ldots + P(A/B_n)P(B_n) = \sum_{i=1}^{n} P(A/B_i)P(B_i)$$

Bayes Theorem

- Let us assume there are *m* mutually exclusive states of nature (classes) labeled ω_j (j=1,2,.....,m).
- Let *P(x)* be the probability that an event assumes the specific value *x*.
- Definitions:
  - Prior Probability: $P(\omega_j)$.
  - Posterior Probability: $P(\omega_j/x)$ (of class ω_j given observation x)
  - Likelihood Probability: $P(x/\omega_j)$ (conditional probability of observation x given class ω_j).

Bayes Theorem

- Bayes's Theorem: gives the relationship between the m prior probabilities $P(\omega_j)$, the m likelihood probabilities $P(x/\omega_j)$ and one posterior probability of interest.

$$P(\omega_j / x) = \frac{P(x \cap \omega_j)}{P(x)} = \frac{P(\omega_j)P(x/\omega_j)}{\sum_{k=1}^{m} P(\omega_k)P(x/\omega_k)}$$

- One would like to choose w_i with highest $P(w_i/x)$.

Random Variable

- A random variables are functions that associate a numerical value to each outcome of an experiment.
  - Function values are real numbers and depend on "chance".
- The function that assigns value to each outcome is fixed and deterministic.
  - The randomness is due to the underlying randomness of the argument of the function X.
  - If we roll a pair of dice then the sum of two face values is a random variable.
- Random numbers can be Discrete or Continuous.
  - Discrete: Countable Range.
  - Continuous: Uncountable Range.

Statistical Characterization of Random Variables

- Expected Value:
  - The expected value of a discrete random variable, x is found by multiplying each value of random variable by its probability and then summing over all values of x.

$$\text{Expected value of x:} \quad E[x] = \sum_{\forall x} xP(x) = \sum_{\forall x} xf(x)$$

  - Expected value is equivalent to center of mass concept.

$$r \times \sum m_i = \sum r_i \times m_i$$

  - That's why name first moment also.
  - Body is perfectly balanced abt. Center of mass
    - The expectation value of x is the "balancing point" for the probability mass function of x
  - Expected value is equal to the point of symmetry in case of symmetric pmf/pdf.

- Joint Probability Functions:
  - Joint Probability Distribution Function:

$$F(\vec{X}) = P[\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\} \cap ..... \cap \{X_n \leq x_n\}]$$

  - Joint Probability Density Function:

$$f(\vec{x}) = \frac{\partial^n F(\vec{X})}{\partial X_1 \partial X_2 ... \partial X_n}$$

- Marginal Probability Functions: A marginal probability functions are obtained by integrating out the variables that are of no interest.

$$F(x) = \sum_{\forall y} P(x,y) \quad \text{or} \quad \int_{y=-\infty}^{y=\infty} f(x,y)\,dy$$

- Mean Vector:

$$E[\mathbf{x}] = [E[x_1] \quad E[x_2] \quad ..... \quad E[x_n]]$$

- Expected value of $g(x_1, x_2, ..., x_n)$ is given by

$$E[g(\mathbf{x})] = \sum_{\forall x_n} \sum_{\forall x_{n-1}} ..... \sum_{\forall x_1} g(\mathbf{x}) f(\mathbf{x}) \quad \text{or} \quad \int_{x_n} \int_{x_{n-1}} ..... \int_{x_1} g(\mathbf{x}) f(\mathbf{x}) dx$$

- Covariance Matrix:

$$\text{cov}[\mathbf{x}] = \mathbf{P} = E[(\mathbf{x} - \vec{\mu})(\mathbf{x} - \vec{\mu})^T] = E[\mathbf{x}\mathbf{x}^T] - \vec{\mu}\vec{\mu}^T$$

where, $S = E[\mathbf{x}\mathbf{x}^T]$ is known as autocorrelation matrix.

$$\text{NOTE: } \mathbf{P} = \Gamma R \Gamma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix}$$

R is the correlation matrix

Covariance Matrix

- Covariance matrix indicates the tendency of each pair of dimensions in random vector to vary together i.e. "co-vary".
- Properties of covariance matrix:
  - Covariance matrix is square.
  - Covariance matrix is always +ive definite i.e. $x^T P x > 0$.
  - Covariance matrix is symmetric i.e. $P = P^T$.
  - If x_i and x_j tends to increase together then $P_{ij} > 0$.
  - If x_i and x_j are uncorrelated then $P_{ij} = 0$.

Independent and Uncorrelated Variables

- Recall, two random variables are said to be independent if knowing values of one tells you nothing about the other variable.
  - Joint probability density function is product of the marginal probability density functions.
  - Cov(X,Y)=0 if X and Y are independent.
  - E(XY)=E(X)E(Y).
- Two variables are said to be uncorrelated if cov(X,Y)=0.
  - **Independent variables are uncorrelated but vice versa is not true.**
- Cov(X,Y)=0➔Integral=0.
  - It tells us that distribution is balanced in some way but says nothing abt. Distribution values.
  - Example: (X,Y) uniformly distributed on unit circle.

$\mathbf{y} = f(\mathbf{x}), \quad \mathbf{x}, \mathbf{y} \in \mathrm{R}^n$

- f^{-1} exist and both f and f^{-1} are continuously differentiable.
- $p_y(\mathbf{y}) = p_x\left(f^{-1}(\mathbf{y})\right) \|\mathbf{J}\|, \quad \mathbf{J} = \dfrac{\partial f^{-1}(\mathbf{y})}{\partial y}.$

EXAMPLE: LET $y = ax^2$ AND $p_x(x) = \dfrac{1}{\sigma_x\sqrt{2\pi}} \exp\left(-\dfrac{x^2}{2\sigma_x^2}\right)$

$$p_y(y) = \frac{1}{2\sigma_x\sqrt{2\pi a y}} \exp\left(-\frac{y}{2a\sigma_x^2}\right), \ \forall y > 0$$
$$= 0, \quad otherwise$$

- **Stochastic Process:** $\{x_t, t \in T\} \longleftrightarrow \{x_t(\omega), t \in T, \omega \in \Omega\}$ is a family of random variables indexed by the parameter set T.
  - For each t, x_t is a random variable (vector).
  - For each ω, $x(\omega)$ is a realization of the process.
  - If x_t can be discrete or continuous random variable.
- Parameter T may also be discrete $(T = \{1, 2, \cdots, n\})$ or continuous $(T = [0, t_f])$
  - If T is discrete $\longrightarrow$ discrete parameter process
  - If T is continuous $\longrightarrow$ continuous parameter process

  We will mainly consider Continuous State Space, which can be divided into two parts:
  - Discrete process
  - Continuous process

DETERMINISTIC DYNAMIC SYSTEMS $\dot{x} = f(x(t))$

- Rate of change of x at time t depends ONLY on current value of x.
- *Finite Memory:* $x(t_2) = g(t_2, x(t_1), t_1)$

MARKOV PROCESS

A discrete or continuous process $\{x_t, t \in T\}$ is called a Markov Process if, for any finite parameter set $\{t_i \mid t_i < t_{i+1}\} \in T$ and for every real λ,

$$pr(x_{t_n}(w) \leq \lambda \mid x_{t_1}, \cdots, x_{t_n}) = pr(x_{t_n}(w) \leq \lambda \mid x_{t_{n-1}})$$

This means future can be predicted from the knowledge of *ONLY PRESENT STATE*.

DETERMINISTIC DYNAMIC SYSTEMS $\dot{x} = f(x(t))$

- Rate of change of x at time t depends ONLY on current value of x.
- *Finite Memory:* $x(t_2) = g(t_2, x(t_1), t_1)$

MARKOV PROCESS

A discrete or continuous process $\{x_t, t \in T\}$ is called a Markov Process if, for any finite parameter set $\{t_i \mid t_i < t_{i+1}\} \in T$ and for every real λ,

$$pr(x_{t_n}(w) \leq \lambda \mid x_{t_1}, \cdots, x_{t_n}) = pr(x_{t_n}(w) \leq \lambda \mid x_{t_{n-1}})$$

This means future can be predicted from the knowledge of *ONLY PRESENT STATE*.

MARKOV PROCESS

In terms of density functions,

$$p(x_{t_n}|x_{t_1},x_{t_2},\cdots,x_{t_{n-1}}) = p(x_{t_n} \mid x_{t_{n-1}})$$

Now let us consider the joint pdf,

$$p(x_{t_n},\cdots,x_{t_1}) = p(x_{t_n} \mid x_{t_{n-1}},\cdots x_{t_1})p(x_{t_n},\cdots,x_{t_1})$$

This implies that

$$p(x_{t_n},\cdots,x_{t_1}) = p(x_{t_n} \mid x_{t_{n-1}})p(x_{t_{n-1}} \mid x_{t_{n-2}})\cdots p(x_{t_2} \mid x_{t_1})p(x_{t_1})$$

Chapman-Kolmogorov Equation (CKE):

$$p(x_n \mid x_m) = \int p(x_n \mid x_{n-1})p(x_{n-1} \mid x_m)dx_{n-1} \qquad m \leq n-2$$

$$\Rightarrow p(x_n) = \int p(x_n \mid x_{n-1})p(x_{n-1})dx_{n-1}$$

- *Discrete System:* $\mathbf{x}_{k+1} = \Phi(\mathbf{x}_k, t_k, t_{k+1}) + \mathbf{w}_k$

$$p(\mathbf{x}_{k+1}) = \int p(\mathbf{x}_{k+1} \mid \mathbf{x}_k) p(\mathbf{x}_k) d\mathbf{x}_k$$

$$p(\mathbf{x}_{k+1} \mid \mathbf{x}_k) = p_{\mathbf{w}_k}(\mathbf{x}_{k+1} - \Phi(\mathbf{x}_k, t_k, t_{k+1}))$$

- *Continuous System:* $d\mathbf{x}_t = f(\mathbf{x}_t, t)dt + G(\mathbf{x}_t, t)d\beta_t$

$$\frac{\partial p(x,t)}{\partial t} = \underbrace{-\sum_{i=1}^{n} \frac{\partial p f_i}{\partial x_i}}_{\text{Drift Term}} + \underbrace{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\left(\partial^2 p[GQG]_{ij}\right)}{\partial x_i \partial x_j}}_{\text{Diffusion Term}}$$

This is the differential form of the CKE which is known as the Fokker-Planck-Kolmogorov Equation (FPKE) or *Kolmogorov's Forward Equation*.

ASSUMPTION

- $\mathbf{y} = H\mathbf{x}_0 + \boldsymbol{\nu},\ \mathbf{y} \in \mathbb{R}^m,\ \mathbf{x} \in \mathbb{R}^n.$
- $E[\boldsymbol{\nu}] = 0,\ E\left[\boldsymbol{\nu}\boldsymbol{\nu}^T\right] = R.$
- Best Linear Unbiased Estimator (BLUE).
  - Best: min $0.5\Sigma = 0.5E\left[(\mathbf{x}-\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)^T\right].$
  - Linear: $\mathbf{x} = M\mathbf{y} + \mathbf{n}.$
  - Unbiased: $E[\mathbf{x}] = \mathbf{x}_0 \ \Rightarrow MH = I,\ \mathbf{n} = 0,\ \Sigma = MRM^T.$

$\min_M J = Tr\left(0.5MRM^T + \Lambda(I - MH)\right)$

- Identities: $\frac{\partial Tr(BAC)}{\partial A} = B^T C^T,\ \frac{\partial Tr(ABA^T)}{\partial A} = A(B + B^T)$
- $\frac{\partial J}{\partial M} = MR - \Lambda^T H^T = 0,\ \frac{\partial J}{\partial \Lambda} = I - MH = 0.$
- $\Lambda^T = (H^T R^{-1} H)^{-1},\ M = (H^T R^{-1} H)^{-1} H^T R^{-1}$

MEASUREMENT MODEL: $\mathbf{y} = h(\mathbf{x}) + \boldsymbol{\nu}$, $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$

- Main idea: Maximize the likelihood function.

$$\max_{\mathbf{x}} \; p(\mathbf{y}|\mathbf{x}) \equiv \max_{\mathbf{x}} \; \ln p(\mathbf{y}|\mathbf{x})$$

- Depending upon $p(\mathbf{y}|\mathbf{x})$, we get different optimization problems.

GAUSSIAN MEASUREMENT MODEL: $\mathbf{y} = h(\mathbf{x}) + \boldsymbol{\nu}$ WITH $\boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{\nu} : \mathbf{0}, R)$

- $\min_{\mathbf{x}} J = \underbrace{\frac{1}{2} (\mathbf{y} - h(\mathbf{x}))^T R^{-1} (\mathbf{y} - h(\mathbf{x}))}_{\text{Weighted Least Square}}$
- Linear Model ($h(\mathbf{x}) = H\mathbf{x}$): $\min_{\mathbf{x}} J = \underbrace{\frac{1}{2} (\mathbf{y} - H\mathbf{x})^T R^{-1} (\mathbf{y} - H\mathbf{x})}_{\text{Weighted Linear Least Square}}$

GAUSSIAN MEASUREMENT MODEL: LINEAR SYSTEM

- Linear Model: $\min_{\mathbf{x}} J = \|\mathbf{y} - H\mathbf{x}\|_R = \underbrace{\frac{1}{2}(\mathbf{y} - H\mathbf{x})^T R^{-1} (\mathbf{y} - H\mathbf{x})}_{\text{Weighted Linear Least Square}}$
- Analytical expression for optimal value of $\mathbf{x}$, $\hat{\mathbf{x}}$ can be obtained.
- *Necessary Condition:* $\frac{dJ}{d\mathbf{x}}|_{\mathbf{x}=\hat{\mathbf{x}}} = 0 \Rightarrow H^T R^{-1} (\mathbf{y} - H\hat{\mathbf{x}}) = 0.$
- $\hat{\mathbf{x}} = \left(H^T R^{-1} H\right)^{-1} H^T R^{-1} \mathbf{y}$
- Show that $E\left[(\mathbf{x} - \hat{\mathbf{x}})^T (\mathbf{x} - \hat{\mathbf{x}})\right] = \left(H^T R^{-1} H\right)^{-1}$
- MLE is also an optimal estimator according to the Cramer-Rao bound for Gaussian Likelihood function.
- BLUE is same as the MLE for Gaussian Likelihood function.

LAPLACIAN MEASUREMENT MODEL: $p(\boldsymbol{\nu}) = \dfrac{1}{2^m \prod_{i=1}^{m} \sigma_i} e^{-\sum_{i=1}^{m} \frac{|\nu_i|}{\sigma_i}}$

- MLE leads to following cost function: $\min_{\mathbf{x}} = J = \sum_{i=1}^{m} \dfrac{|\mathbf{y}_i - \mathbf{h}_i \mathbf{x}|}{\sigma_i}$.
- No Analytical Solution but aforementioned problem is a convex optimization problem.

LINEAR PROGRAMMING: $\min_{\mathbf{x}} = J = \sum_{i=1}^{m} \dfrac{|\mathbf{y}_i - \mathbf{h}_i \mathbf{x}|}{\sigma_i}$

$$\min \sum_{i=1}^{m} t_i, \text{ s.t.}$$

$$\frac{\mathbf{y}_i - \mathbf{h}_i \mathbf{x}}{\sigma_i} \leq t_i, \quad \frac{\mathbf{y}_i - \mathbf{h}_i \mathbf{x}}{\sigma_i} \geq -t_i, \quad \forall i$$

UNIFORM MEASUREMENT MODEL: $p(\boldsymbol{\nu}) = \prod_{i=1}^{m} \mathscr{U}(a_i, b_i)$

- w.l.o.g. assume that $a_i = -1$ and $b_i = 1$.
- MLE leads to following constraints: $-1 \leq \mathbf{y}_i - \mathbf{h}_i\mathbf{x} \leq 1$.
- This is equivalent to minimizing
  $\min_{\mathbf{x}} J = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_\infty = \max_i |\mathbf{y}_i - \mathbf{h}_i\mathbf{x}|$
- No Analytical Solution but aforementioned problem is a convex optimization problem.

LINEAR PROGRAMMING: $\min_{\mathbf{x}} J = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_\infty$

$$\min t, \text{s.t.}$$

$$\mathbf{y}_i - \mathbf{h}_i\mathbf{x} \leq t,\ \mathbf{y}_i - \mathbf{h}_i\mathbf{x} \geq -t,\ \forall i$$

MEASUREMENT MODEL: $\mathbf{y} = h(\mathbf{x}) + \boldsymbol{\nu}$

- Main idea: Maximize the posterior density function.

$$\max_{\mathbf{x}} \; p(\mathbf{x}/\mathbf{y}) \equiv \max_{\mathbf{x}} \; \ln p(\mathbf{x}/\mathbf{y})$$

- Use Bayes' Rule: $\max_{\mathbf{x}} \ln p(\mathbf{x}) + \ln p(\mathbf{y}/\mathbf{x})$
- The first term acts like *preconditioner* or *Regularization*.
- Depending upon $p(\mathbf{x})$ and $p(\mathbf{y}/\mathbf{x})$, we get different optimization problems.

LINEAR MEASUREMENT MODEL

$\mathbf{y} = H\mathbf{x} + \boldsymbol{\nu}$ where

- $\boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{\nu} : \mathbf{0}, R)$, $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} : \hat{\mathbf{x}}, P)$
- $\min_{\mathbf{x}} J = \underbrace{(\mathbf{x} - \hat{\mathbf{x}})^T P^{-1} (\mathbf{x} - \hat{\mathbf{x}}) + (\mathbf{y} - H\mathbf{x})^T R^{-1} (\mathbf{y} - H\mathbf{x})}_{\text{Tikhonov Regularization}}$

SOME USEFUL EXAMPLES

- Prior: Laplacian and Likelihood: Gaussian

$$\min_{\mathbf{x}} J = \underbrace{\|\mathbf{x}-\hat{\mathbf{x}}\|_1 + \|\mathbf{y}-H\mathbf{x}\|_2}_{\text{Sparse Approximation}}$$

- Prior: Uniform and Likelihood: Gaussian

$$\min_{\mathbf{x}} J = \|\mathbf{x}-\hat{\mathbf{x}}\|_\infty + \|\mathbf{y}-H\mathbf{x}\|_2$$

MEASUREMENT UPDATE

Starting with the prior pdf,

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} : \hat{\mathbf{x}}\ , P)$$

the posterior pdf from Bayes' rule is given by

$$p(\mathbf{x}|y) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}}$$

where the measurement likelihood pdf $p(\mathbf{y}|\mathbf{x})$ can be derived from the measurement model equations as:

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y} : H\mathbf{x},\ R)$$

Numerator of Bayes' rule is simplified as:

$$p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \mathcal{N}(\mathbf{y} : H\mathbf{x},\ R)\mathcal{N}(\mathbf{x} : \hat{\mathbf{x}}\ , P)$$

$$= \frac{1}{\sqrt{|2\pi R|}} \cdot \frac{1}{\sqrt{|2\pi P|}} \cdot \exp\Big[-\frac{1}{2}\left(\mathbf{y} - H\mathbf{x}\right)^T R^{-1}\left(\mathbf{y} - H\mathbf{x}\right) - \frac{1}{2}\left(\mathbf{x} - \hat{\mathbf{x}}\right)^T P^{-1}\left(\mathbf{x} - \hat{\mathbf{x}}\right)\Big]$$

The exponent is:

$$\Rightarrow -\frac{1}{2}\left(\mathbf{y} - H\mathbf{x}\right)^T R^{-1}\left(\mathbf{y} - H\mathbf{x}\right) - \frac{1}{2}\left(\mathbf{x} - \hat{\mathbf{x}}\right)^T P^{-1}\left(\mathbf{x} - \hat{\mathbf{x}}\right)$$

$$\Rightarrow -\frac{1}{2}\Big[\mathbf{y}^T R^{-1}\mathbf{y} + \mathbf{x}^T \underbrace{\left(P^{-1} + H^T R^{-1} H\right)}_{A}\mathbf{x} - 2\underbrace{\left(\mathbf{y}^T R^{-1} H + \hat{\mathbf{x}}^T P^{-1}\right)}_{b^T}\mathbf{x} + \hat{\mathbf{x}}^T P^{-1}\hat{\mathbf{x}}\Big]$$

$$\Rightarrow -\frac{1}{2}\Big[\mathbf{y}^T R^{-1}\mathbf{y} + \mathbf{x}^T A\mathbf{x} - 2b^T\mathbf{x} + \hat{\mathbf{x}}^T P^{-1}\hat{\mathbf{x}}\Big]$$

The denominator of Bayes' rule is given as:

$$\int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x} = \frac{1}{\sqrt{|2\pi R|}} \cdot \frac{1}{\sqrt{|2\pi P|}} \cdot$$
$$\int \exp\Big[-\frac{1}{2}\mathbf{y}^T R^{-1}\mathbf{y} - \frac{1}{2}\mathbf{x}^T A\mathbf{x} + b^T\mathbf{x} - \frac{1}{2}\hat{\mathbf{x}}^T P^{-1}\hat{\mathbf{x}}\Big]\,d\mathbf{x}$$
$$= \frac{1}{\sqrt{|2\pi R|}} \cdot \frac{1}{\sqrt{|2\pi P|}} \cdot \exp\Big[-\frac{1}{2}\mathbf{y}^T R^{-1}\mathbf{y} - \frac{1}{2}\hat{\mathbf{x}}^T P^{-1}\hat{\mathbf{x}}\Big]$$
$$\cdot \int \exp\Big[-\frac{1}{2}\mathbf{x}^T A\mathbf{x} + b^T\mathbf{x}\Big]\,d\mathbf{x}$$
$$= \frac{\sqrt{|2\pi A^{-1}|}}{\sqrt{|2\pi R||2\pi P|}} \cdot \exp\Big[-\frac{1}{2}\mathbf{y}^T R^{-1}\mathbf{y} - \frac{1}{2}\hat{\mathbf{x}}^T P^{-1}\hat{\mathbf{x}} + \frac{1}{2}b^T A^{-T} b\Big]$$

The posterior pdf from Bayes' rule is given as:

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{\sqrt{|2\pi A^{-1}|}} \cdot \exp\Big[-\frac{1}{2}\mathbf{x}^T A\mathbf{x} + b^T\mathbf{x} - \frac{1}{2}b^T A^{-T} b\Big]$$
$$= \frac{1}{\sqrt{|2\pi A^{-1}|}} \cdot \exp\Big[-\frac{1}{2}\big(\mathbf{x} - A^{-1}b\big)^T A\big(\mathbf{x} - A^{-1}b\big)\Big]$$

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{\sqrt{|2\pi A^{-1}|}} . \exp\left[-\frac{1}{2}\left(\mathbf{x} - A^{-1}b\right)^T A \left(\mathbf{x} - A^{-1}b\right)\right]$$

$$b^T = \left(\mathbf{y}^T R^{-1} H + \hat{\mathbf{x}}^T P^{-1}\right),\ A = P^{-1} + H^T R^{-1} H$$

Useful Identities:

$$\left(I + PH^T R^{-1} H\right)^{-1} = I - PH^T \left(HPH^T + R\right)^{-1} H$$

$$\left(I + PH^T R^{-1} H\right)^{-1} P = P - I - PH^T \left(HPH^T + R\right)^{-1} H$$

$$\left(I + PH^T R^{-1} H\right)^{-1} PH^T R^{-1} = PH^T \left(HPH^T + R\right)^{-1}$$

$$A^{-1} = P - \underbrace{PH^T \left(R + HPH^T\right)^{-1}}_{K} HP = P - KHP$$

$$A^{-1}b = \left(P - KHP\right)\left(H^T R^{-1}\mathbf{y} + P^{-1}\hat{\mathbf{x}}\right) = \hat{\mathbf{x}} + K\left(\mathbf{y} - H\hat{\mathbf{x}}\right)$$

Hence the posterior pdf still remains Gaussian even after Bayes' Rule update, with mean and covariance as

$$\text{Posterior Mean: } \mu = A^{-1}b = \hat{\mathbf{x}} + K\left(\mathbf{y} - H\hat{\mathbf{x}}\right)$$

$$\text{Posterior Covariance: } \boldsymbol{\Sigma} = A^{-1} = P - KHP$$

- This paves the way to sequentially process the measurement data.
- After processing, one set of measurement data, the posterior density serves as the prior for the next set of measurement data.

In summary, starting with prior $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} : \boldsymbol{\mu_0}\ \boldsymbol{\Sigma}_0,)$ with $\boldsymbol{\mu}_0 = \hat{\mathbf{x}},\ \boldsymbol{\Sigma}_0 = P$

MEASUREMENT UPDATE

After the k^{th} iteration: $p(\mathbf{x}|Y_k) = \mathcal{N}(\mathbf{x} : \mu_{k|k}\ , \boldsymbol{\Sigma}_{k|k})$

$$\boldsymbol{\mu}_{k+1|k+1} = \boldsymbol{\mu}_{k|k} + K_k\left(\mathbf{y}_{k+1} - H\boldsymbol{\mu}_{k|k}\right)$$

$$\boldsymbol{\Sigma}_{k+1|k+1} = \boldsymbol{\Sigma}_{k|k} - K_k H \boldsymbol{\Sigma}_{k|k}$$

$$K_k = \boldsymbol{\Sigma}_{k|k} H^T \left(R + H\boldsymbol{\Sigma}_{k|k} H^T\right)^{-1}$$

Y_k is the measurement data up to the k^{th} iteration.

MINIMUM VARIANCE ESTIMATE

$$\min_{\hat{\mathbf{x}}_{k+1/k+1}} Tr\left\{E[(\mathbf{x}_{k+1}-\boldsymbol{\mu})(\mathbf{x}_{k+1}-\boldsymbol{\mu})^T]\right\}$$

i.e. find an estimate that minimizes the posterior variance

ESTIMATOR: LINEAR UPDATE

$$\boldsymbol{\mu} = \hat{\mathbf{x}}_{k+1/k} + K(\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1})$$

The assumed estimator is unbiased.

$$\min_K Tr\Big\{E[(\mathbf{x}_{k+1}-\boldsymbol{\mu})(\mathbf{x}_{k+1}-\boldsymbol{\mu})^T]\Big\}$$
$$with \quad \boldsymbol{\mu} = \hat{\mathbf{x}}_{k+1/k} + K(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1})$$

$$\min_K Tr\Big\{E[(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k}-K(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1}))(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k}-K(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1}))^T]\Big\}$$
$$\min_K Tr\Big\{E[\big(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k}\big)\big(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k}\big)^T] + KE[(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1})(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1})^T]K^T$$
$$- KE[(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1})(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k})^T] - E[\big(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k}\big)(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1})^T]K^T\Big\}$$
$$\min_K Tr\Big\{P + KP^yK^T - KP^{yx} - P^{xy}K^T\Big\}$$

The optimal gain K is given by

$$K = P^{xy}(P^y)^{-1}$$

- no assumptions on the state pdf.
- All the expectations are with respect to the prior pdf $p(\mathbf{x})$.

$$
\begin{aligned}
\hat{\mathbf{y}}_{k+1} &= E[\mathbf{y}] = E[Hx+\nu] = HE[x] = H\hat{\mathbf{x}} \\
P^y &= E[(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1})(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1})^T] \\
&= E[\left(H\mathbf{x}+\nu-H\hat{\mathbf{x}}\right)\left(H\mathbf{x}+\nu-H\hat{\mathbf{x}}\right)^T] \\
&= HE[\left(\mathbf{x}-\hat{\mathbf{x}}\right)\left(\mathbf{x}-\hat{\mathbf{x}}\right)^T]H^T + E[\nu\nu^T] \\
&= HPH^T + R \\
P^{xy} &= E[\left(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k}\right)(\mathbf{y}_{k+1}-\hat{\mathbf{y}}_{k+1})^T] \\
&= E[\left(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k}\right)\left(H(\mathbf{x}_{k+1}-\hat{\mathbf{x}}_{k+1/k})+\nu\right)^T] \\
&= E[\left(\mathbf{x}-\hat{\mathbf{x}}\right)\left(\mathbf{x}-\hat{\mathbf{x}}\right)^T]H = PH^T
\end{aligned}
$$

The gain matrix is then given as:

$$K = P^{xy}(P^y)^{-1} = PH^T\left(HPH^T+R\right)^{-1}$$

Minimum variance estimator is same as the MAP and hence optimal for linear system with Gaussian pdfs

Minimum variance estimator for Nonlinear Measurement model, $\mathbf{y} = h(\mathbf{x}) + \boldsymbol{\nu}$:

$$\min_K Tr\Big\{E[(\mathbf{x}_{k+1} - \boldsymbol{\mu})(\mathbf{x}_{k+1} - \boldsymbol{\mu})^T]\Big\}$$
$$with \quad \boldsymbol{\mu} = \hat{\mathbf{x}}_{k+1/k} + K(\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1})$$

$$\min_K Tr\Big\{P + KP^y K^T - KP^{yx} - P^{xy}K^T\Big\}$$

The optimal gain K is given by

$$K = P^{xy}(P^y)^{-1}$$

Using Taylor Series Expansion of $h(x)$ about the current estimate at time $k+1$ i.e. $\hat{\mathbf{x}}$

$$h(\mathbf{x}) = h(\hat{\mathbf{x}}) + H(\mathbf{x} - \hat{\mathbf{x}}) \quad where \quad H \equiv \frac{\partial h}{\partial \mathbf{x}}|_{\mathbf{x}=\hat{\mathbf{x}}}$$

$$\hat{\mathbf{y}}_{k+1} = E[\mathbf{y}] = E[h(\mathbf{x})] + E[\boldsymbol{\nu}] \approx h(\hat{\mathbf{x}})$$

$$
\begin{aligned}
P^y &= E[(\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1})(\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1})^T] \\
&\approx E[\left(h(\mathbf{x}) + \nu - h(\hat{\mathbf{x}})\right)\left(h(\mathbf{x}) + \nu - h(\hat{\mathbf{x}})\right)^T] \\
&= HE[\left(\mathbf{x} - \hat{\mathbf{x}}\right)\left(\mathbf{x} - \hat{\mathbf{x}}\right)^T]H^T + E[\nu\nu^T] \\
&= HPH^T + R \\
P^{xy} &= E[\left(\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1/k}\right)(\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1})^T] \\
&\approx E[\left(\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1/k}\right)\left(h(\mathbf{x}) + \nu - h(\hat{\mathbf{x}})\right)^T] \\
&= E[\left(\mathbf{x} - \hat{\mathbf{x}}\right)\left(\mathbf{x} - \hat{\mathbf{x}}\right)^T]H = PH^T
\end{aligned}
$$

The linear gain is then given as:

$$K = P^{xy}(P^y)^{-1} = PH^T\left(HPH^T + R\right)^{-1}$$

In summary,

MEASUREMENT UPDATE

$$\boldsymbol{\mu} = \hat{\mathbf{x}}_{k+1/k} + K(\mathbf{y}_{k+1} - h(\hat{\mathbf{x}}))$$
$$\boldsymbol{\Sigma} = P - KHP$$
$$K = PH^T \left(HPH^T + R\right)^{-1}$$
$$H \equiv \frac{\partial h}{\partial \mathbf{x}}|_{\mathbf{x}=\hat{\mathbf{x}}}$$

- Only mean and covariance are updated.
- All expectation expression $E[.]$ evaluated by linearizations $\Rightarrow$ analytical expressions
- Estimates can quickly diverge due to linearizations involved.

MOTIVATION FOR UNSCENTED AND QUADRATURE RULES

- $\Rightarrow$ Avoid linearization altogether and evaluate these expectation integrals directly using appropriate quadrature scheme.
- *Unscented Transform, Gauss-hermite Quadratures or Conjugate Unscented Transform*