Chapter 2

Probabilistic Methods

2.1 Introduction to Probabilities

2.1.1 Definitions

Preliminaries

In probability theory, for any given experiment, an outcome is the result of that experiment, and the sample space, denoted by S, is the set of all possible outcomes. An event, denoted by A, is a set of outcomes that is also a subset of the sample space, i.e., $A \subset S$. The partition P of S is a set of nonempty disjoint subsets of S whose union is S, such that:

$$P = \left\{ A_i \subset S \Big|_{\bigcup_{i=1}^{n_p} A_i = S}^{\forall i,j \in \{1,\ldots,n_p\}, A_i \cap A_j = \varnothing,} \right\} \tag{2.1}$$

The quantity $Pr(A)$ is the probability of event A to occur, and it satisfies the following three axioms of probability:

$$Pr(A) \geq 0, \quad \forall A \subset S, \tag{2.2}$$

$$Pr(S) = 1, \tag{2.3}$$

$$\bigcup_{i=1}^{n_p} Pr(A_i) = \sum_{i=1}^{n_p} Pr(A_i), \quad \forall A_i \in P. \tag{2.4}$$

From these axioms, the following consequences hold true:

$$Pr(\emptyset) = 0 \tag{2.5}$$

$$Pr(A) = 1 - P(\neg A), \tag{2.6}$$

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B), \tag{2.7}$$

$$Pr(A) \leq Pr(B), \quad \forall A \subset B \tag{2.8}$$

Conditional Probability

Conditional probability is the likelihood of one event occurring given the occurrence of another event, and it is defined for any two events A, B as follows:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} \tag{2.9}$$

where $Pr(B) > 0$.

Mutual Exclusive Events

Two events A, B are said to be mutually exclusive if A and B are disjoint sets, i.e., $A \cap B = \emptyset$. In terms of probability, mutually exclusive events correspond to the following:

$$Pr(A \cap B) = 0, \tag{2.10}$$

$$Pr(A \cup B) = Pr(A) + Pr(B), \tag{2.11}$$

$$Pr(A|B) = 0, \tag{2.12}$$

$$Pr(A|\neg B) = \frac{Pr(A)}{1 - Pr(B)} \tag{2.13}$$

Independent Events

Events A, B are *independent* if the occurrence of event A does not affect the occurrence of event B. This definition translates to probabilities as follows:

$$Pr(A \cap B) = Pr(A)Pr(B), \tag{2.14}$$

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A)Pr(B), \tag{2.15}$$

$$Pr(A|B) = Pr(A). \tag{2.16}$$

Keep in mind that independence does not mean mutual exclusion and vice versa. In fact, for any nonempty mutual exclusive events are not independent, and vice versa.

Chain Rule for Probabilities

Conditional probabilities can be used to derive the chain rule for any two events $A, B \subset S$ as follows:

$$Pr(A \cap B) = Pr(A|B)Pr(B). \tag{2.17}$$

In general, the chain rule for k events in S is:

$$Pr\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} Pr\left(A_i | \bigcap_{j=1}^{i-1} A_j\right). \tag{2.18}$$

Note that A_i in equation(2.18) is not in the partition P, otherwise the result is 0.

Law of Total Probability

The *law of total probability* computes the probability of event B given the set of k joint probabilities $Pr(B \cap A_i)$, where $A_i \in P$, and it is defined as:

$$Pr(B) = \sum_{i=1}^{n_p} Pr(B \cap A_i), \tag{2.19}$$

$$Pr(B) = \sum_{i=1}^{n_p} Pr(B|A_i)Pr(A_i). \tag{2.20}$$

Bayes' Theorem

An important consequence of law of total probability and conditional probability is Bayes' theorem, and it states the following:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)} \tag{2.21}$$

where $Pr(B) > 0$. In general, if $A_i \in P$, then, Bayes' theorem corresponds to the following:

$$Pr(A_j|B) = \frac{Pr(B|A_j)Pr(A_j)}{\sum_{i=1}^{n_p} Pr(B|A_i)Pr(A_i)} \tag{2.22}$$

2.1.2 Discrete Random Variables and Probability Distribution

A *random variable* is a function that associates each outcome in the sample space with a real number. Let X denote a random variable, then, by definition, $X : S \to \mathbb{R}$. A *discrete* random variable can take a countable number of distinct values, and it is generally associated with a *probability mass function* (PMF), denoted by $p_X(x)$, that return the probability of the random variable being exactly equal to some value $x \in \mathbb{R}$, such that:

$$p_X(x) = Pr(X = x), \tag{2.23}$$

$$\sum_{x=-\infty}^{\infty} p_X(x) = 1. \tag{2.24}$$

The *expected value*, or the *mean*, of a random variable is denoted by $E[X]$, and it is defined for a discrete random variable as follows:

$$E[X] = \mu_X = \sum_{i=1}^{\infty} p_X(x_i)x_i. \tag{2.25}$$

The *variance* of a random variable is defined as:

$$Var(X) = \sigma_X^2 = E[(X - \mu_X)^2] = E[X^2] - \mu_X^2 \tag{2.26}$$

For discrete random variable, the variance is computed as follows:

$$Var(X) = \sum_{i=1}^{\infty} p_X(x_i)(x_i - \mu_X)^2 = \sum_{i=1}^{\infty} p_X(x_i)x_i^2 - \mu_X^2 \tag{2.27}$$

2.1.3 Continuous Random Variables and Probability Distribution

A *continuous* random variable takes an infinite number of possible values, and it is usually associated with a *probability density function* (PDF) that is denoted by $p_X(x)$, where $p_X(x) \leq 0; \forall x \in \mathbb{R}$ .The probability of an event with continuous random variables is computed using the cumulative distribution function (CDF) which is defined as follows:

$$Pr(X \leq x) = F_X(x) = \int_{-\infty}^{x} p_X(u)du \tag{2.28}$$

Note that $\lim_{x\to\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.
The expected value and the variance of a continuous random variable is defined in (2.29) and (2.30), respectively.

$$E[X] = \mu_X = \int_{-\infty}^{\infty} p_X(x)dx, \tag{2.29}$$

$$Var(X) = \sigma_X^2 = \int_{-\infty}^{\infty} p_X(x)(x - \mu_X)^2 = \int_{-\infty}^{\infty} p_X(x)x^2dx - \mu_X^2. \tag{2.30}$$

Moreover, the *support* of a continuous random variable is the smallest closed set at which the probability density function is not zero.

Conditional Distribution

Consider two continuous random variables X, Y with PDFs $p_X(x)$, $p_Y(y)$, respectively. Then, the *conditional* probability density function of X given $Y = y$ is:

$$p_X(x|Y = y) = p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \tag{2.31}$$

where $p_{X,Y}(x, y)$ is the density of the *joint probability distribution* of X and Y.

Chain Rule and Bayes' Theorem for Distributions

Let $p_{X|Y}(x|y)$ be the conditional distribution of X given $Y = y$, and $p_Y(y)$ is the PDF of Y, then, the *chain rule* describes the joint distribution of X and Y as:

$$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y). \tag{2.32}$$

Note that this rule is applied to PDFs, and it is similar to the chain rule for probabilities in section 2.1.1. Moreover, Bayes' theorem for PDFs is defined as:

$$p_{X|Y}(x|y) = p_{Y|X}(y|x)\frac{p_X(x)}{(p_Y(y)} \tag{2.33}$$

Independent Random Variables

Two random variables X,Y are said to be independent if and only if:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \tag{2.34}$$

where $p_X(x)$, $p_Y(y)$, are the PDFs of X, Y, respectively; and $p_{X,Y}(x,y)$ is the joint density of X and Y.

Marginal Distribution

The *marginal distribution* of the random variable X is defined as:

$$p_X(x) = \int_y p_{X,Y}(x,y)dy = \int_y p_{X|Y}(x|y)p_Y(y)dy. \tag{2.35}$$

This is equivalent to law of total probability defined in section 2.1.1, however, here it is applied to PDFs.

2.2 Gaussian Distributions

2.2.1 Introduction

A random variable X is normally distributed, or Gaussian, if its probability density function is defined as:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu x)^2)}{2\sigma^2}), \tag{2.36}$$

where, μ_X, σ^2 are the *mean* and *variance*, respectively, they are the distribution parameters. The notation $X \sim \mathcal{N}(\mu_X, \sigma^2)$ means that the random variable X is Gaussian.

Multivariate Gaussian is a generalisation of uni-variate Gaussian by considering the random vector in $\mathbb{R}^n$, $X = [X_1 X_2 \dots X_n]^T$,where every linear combination of its components is uni-variate Gaussian. In that case, the probability density function of X is defined as follows:

$$p_X(x) = \frac{1}{\sqrt{2\pi^2|\boldsymbol{\Sigma_X}|}} exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_x})^T \Sigma_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu_X})), \tag{2.37}$$

where, $\boldsymbol{\mu_X} \in \mathbb{R}^{n \times n}$ is the *mean* vector, and $\boldsymbol{\Sigma_X} \in \mathbb{R}^{n \times n}$ is a symmetric positive definite *covariance* matrix. If $\boldsymbol{\Sigma_X}$ is singular, then, the distribution is *degenerate*. In terms of notation, $X \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ is said to be a Gaussian random vector, or normally distributed random vector.

2.2.2 Properties

Gaussian distributions have a lot of useful properties and they are presented in this section, however, the derivations are omitted although they are easily obtained by applying the definitions in section 2.1.3 with the Gaussian PDF.

2.2.3 Affine Transformation

Consider $X \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ in $\mathbb{R}^n$, and let $Y = \mathbf{A}X + \mathbf{b}$ be the affine transformation, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Then, the random vector $Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ such that:

$$\begin{aligned} \boldsymbol{\mu_Y} &= \boldsymbol{A\mu_X + b}, \\ \boldsymbol{\Sigma_Y} &= \boldsymbol{A\Sigma_X A^T}. \end{aligned} \tag{2.38}$$

Sum of Independent Gaussian Distributions

Let $X \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ be independent random vectors, then, the sum $Z = X + Y$ is Gaussian such that $Z \sim \mathcal{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$ where

$$\begin{aligned} \boldsymbol{\mu_Z} &= \boldsymbol{\mu_X + \mu_Y}, \\ \boldsymbol{\Sigma_Z} &= \boldsymbol{\Sigma_X + \Sigma_Y}. \end{aligned} \tag{2.39}$$

Marginal Distribution

Let $X \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ be a random vector in $\mathbb{R}^n$, that can be partitioned as:

$$X = \begin{bmatrix} U \\ V \end{bmatrix}, \quad \boldsymbol{\mu_X} = \begin{bmatrix} \boldsymbol{\mu_U} \\ \boldsymbol{\mu_V} \end{bmatrix}, \quad \boldsymbol{\Sigma_X} = \begin{bmatrix} \boldsymbol{\Sigma_U} & \boldsymbol{\Sigma_{UV}} \\ \boldsymbol{\Sigma_{UV}^T} & \boldsymbol{\Sigma_V} \end{bmatrix}. \tag{2.40}$$

Then, the *marginal distribution* of the random vectors U and V are Gaussians such that:

$$\begin{aligned} U &\sim \mathcal{N}(\boldsymbol{\mu}_U, \boldsymbol{\Sigma}_U) \\ V &\sim \mathcal{N}(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V) \end{aligned} \tag{2.41}$$

Conditional Distribution

Following the partitioning defined in equation (2.40), the *conditional distribution* of $U|V \sim \mathcal{N}(\boldsymbol{\mu}_U|V,\ \boldsymbol{\Sigma}_U|V)$, where:

$$\begin{aligned} \boldsymbol{\mu}_{U|V} &= \boldsymbol{\mu}_U + \boldsymbol{\Sigma}_{UV}\boldsymbol{\Sigma}_V^{-1} \\ \boldsymbol{\Sigma}_{U|V} &= \boldsymbol{\Sigma}_U - \boldsymbol{\Sigma}_{UV}\boldsymbol{\Sigma}_V^{-1}\boldsymbol{\Sigma}_{UV}^T \end{aligned} \tag{2.42}$$

Note that $\boldsymbol{\Sigma_{U|V}}$ is *Schur complement* of $\boldsymbol{\Sigma_V}$ in $\boldsymbol{\Sigma_X}$.

2.2.4 Chain Rule for Gaussian Distributions

Consider the conditional distribution $U|V \sim \mathcal{N}(\boldsymbol{A}V + \boldsymbol{b}, \boldsymbol{\Sigma}_{U|V})$, and the random vector $V \sim \mathcal{N}(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$; then, the joint distribution of $X = [U^T\ V^T]^T$ is also Gaussian such that $X \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$, where:

$$\boldsymbol{\mu}_X = \begin{bmatrix} \boldsymbol{A\mu}_V + \boldsymbol{b} \\ \boldsymbol{\mu}_V \end{bmatrix}, \quad \boldsymbol{\Sigma}_X = \begin{bmatrix} \boldsymbol{A\Sigma}_V\boldsymbol{A}^T + \boldsymbol{\Sigma_{U|V}} & \boldsymbol{A\Sigma}_V \\ \boldsymbol{A\Sigma}_V{}^T & \boldsymbol{\Sigma}_V \end{bmatrix} \tag{2.43}$$

2.3 Bayes Filter

Bayes filter is a probabilistic approach to recursively estimate some unknown PDFs as new information, such as measurements, becomes available. Consider, for example, a discrete dynamical system that is characterised by two probability distributions: (i) *state transition*, or *motion model* probability $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{u}_k)$, and (ii) *measurement*, or *observation model* probability $p(\mathbf{z}_k|\mathbf{x}_k)$, where $\mathbf{x}$ is the *state* of the system, $\mathbf{u}$ is control input, $\mathbf{z}$ is the measurement, and k is the time step.

The quantities $\mathbf{x}_k$, $\mathbf{z}_k$, and $\mathbf{u}_k$ can be treated as either discrete or continuous random vectors. This filter, also known as *Recursive Bayesian estimator*, is used to estimate the *belief* of the system, denoted by $bel(\mathbf{x}_k)$, which it is defined as the probability distribution of the system state $\mathbf{x}_k$ conditioned on all available data: control inputs $U_k = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ and measurements $Z_k = \{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$,

such that:

$$bel\left(\mathbf{x}_k\right) = p(\mathbf{x}_k|U_k, Z_k) \tag{2.44}$$

This belief can be simplified to include the state transition distribution and measurement distribution. To do so, use Bayes' Theorem defined in equation (2.33) to express the belief as:

$$\begin{aligned} bel(\mathbf{x}_k) &= \frac{p(\mathbf{z}_k|\mathbf{x}_k, U_k, Z_{k-1})p(\mathbf{x}_k|U_k, Z_{k-1})}{p(\mathbf{z}_k|Z_{k-1}, U_k)} \\ &= \eta\ p(\mathbf{z}_k|\mathbf{x}_k, U_k, Z_{k-1})p(\mathbf{x}_k|U_k, Z_{k-1}) \end{aligned} \tag{2.45}$$

where η is a normalisation constant that is independent of $\mathbf{x}_k$. From the measurement probability, the current measurement $\mathbf{z}_k$ is influenced only by the current state $\mathbf{x}_k$, thus:

$$p(\mathbf{z}_k|\mathbf{x}_k) = p(\mathbf{z}_k|\mathbf{x}_k, U_k, Z_{k-1}), \tag{2.46}$$

$$bel\left(\mathbf{x}_k\right) = \eta\ p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|U_k, Z_{k-1}) \tag{2.47}$$

Now, define $\overline{bel}\left(\mathbf{x}_k\right) = p(\mathbf{x}_k|U_k, Z_k)$, and use the marginal distribution in equation (2.35) to expand $\overline{bel}\left(x_k\right)$ as:

$$\overline{bel}\left(\mathbf{x}_k\right) = p(\mathbf{x}_k|U_k, Z_k) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|U_k, Z_{k-1})d\mathbf{x}_{k-1} \tag{2.48}$$

Note that the state transition probability follows *Markov assumption* where the current state $\mathbf{x}_k$ depends only on the immediate previous state and the current control input. Moreover, the dynamical system is *causal*, and current state does not depend on future inputs, thus:

$$\begin{aligned} bel\left(\mathbf{x}_k\right) &= \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|U_k, Z_{k-1})d\mathbf{x}_{k-1} \\ &= \int p(\mathbf{x}_k|\mathbf{x}_{k-1})bel(\mathbf{x}_{k-1})d\mathbf{x}_{k-1}. \end{aligned} \tag{2.49}$$

Finally, by putting everything together, equation (2.44) becomes:

$$bel\left(\mathbf{x}_k\right) = \eta\ p(\mathbf{z}_k|\mathbf{x}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{u}_k)\ bel(\mathbf{x}_{k-1})d\mathbf{x}_{k-1} \tag{2.50}$$

This is the recursive nature of the Bayes filter, that given a prior probability distribution of the system state, the posterior distribution is estimated using the state transition model and observation model. In general, this process assumes

the system executes a control action $\mathbf{u}_k$ first, and then, takes a measurement $\mathbf{z}_k$. Therefore, the Bayes filter has two steps: (i) *prediction* or *control update* to estimate $\overline{bel}\,(\mathbf{x}_k)$, and (ii) *correction* or *measurement update* to estimated $bel\,(\mathbf{x}_k)$. Generally, during the control update step, the belief become less certain, i.e., increase the variance, due to the integration action. On the other hand, in the measurement update step, the belief certainty increases, i.e., the variance decreases, because of the conditional distribution.

2.3.1 Kalman Filter

Kalman filter is a special case of Bayes filter that works with linear Gaussian dynamical systems. It uses the properties of Gaussian distributions to produce a closed-form estimate of the system belief. The Kalman filter is a *parametric* approach where the belief is a PDF described by its parameters, such as the mean and covariance. The state is a continuous random vector $\mathbf{x}_k \in \mathbb{R}^{n_s}$ such that $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The state transition model and the observation model are linear in $\mathbf{x}$ as defined in equation (2.51) and (2.52), respectively, where $\mathbf{u}_k \in \mathbb{R}^{n_i}$ is the control input vector, and $\mathbf{z}_k \in \mathbb{R}^{n_o}$ is the measurement vector.

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u} + \mathbf{q}_k, \tag{2.51}$$

$$\mathbf{z}_k = \mathbf{C}\mathbf{x}_k + \mathbf{r}_k. \tag{2.52}$$

Note that $\mathbf{A} \in \mathbb{R}^{n_s \times n_s}$, $\mathbf{B} \in \mathbb{R}^{n_s \times n_i}$, and $\mathbf{C} \in \mathbb{R}^{n_o \times n_s}$. Moreover, $\mathbf{q}_k \in \mathbb{R}^{n_s}$ and $\mathbf{r}_k \in \mathbb{R}^{n_o}$ represent the noise in the state transition model and observation models. They are assumed to be zero mean, Gaussian white noise, i.e., $\mathbf{q}_k \sim \mathcal{N}(0, Q_k)$, and $\mathbf{r}_k \sim \mathcal{N}(0, R_k)$, where $\mathbf{Q}_k$ and $\mathbf{R}_k$ are covariance matrices.
In terms of probability distributions, the prior state transition model belief, i.e., $bel\,(\mathbf{x}_{k-1})$, is defined in equation (2.53), the state transition model is in equation (2.54), and the observation model is in equation (2.55).

$$\mathbf{x}_{k-1} | U_{k-1}, Z_{k-1} \sim \mathcal{N}(\boldsymbol{\mu}_{k-1}, \boldsymbol{\Sigma}_{k-1}), \tag{2.53}$$

$$\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{u}_k \sim \mathcal{N}(\mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k, \mathbf{Q}_k), \tag{2.54}$$

$$\mathbf{z}_k | \mathbf{x}_k \sim \mathcal{N}(\mathbf{C}\mathbf{x}_k, \mathbf{R}_k) \tag{2.55}$$

Similar to Bayes filter, Kalman filter follows the prediction and correction steps to estimate the belief by incorporating the control signal $\mathbf{u}_k$ and the measurement $\mathbf{z}_k$. However, since the state transition model is linear, and all distributions are assumed to be Gaussian distributions, the properties in section 2.2 are used to derive estimation equations as shown next.

Prediction Step

In this step, only the control input $\mathbf{u}_k$ is incorporated to estimate $\overline{bel}\,(\mathbf{x}_k)$. First, the chain rule in equation (2.43) is used to find the joint distribution of $\mathbf{x}_k|U_k, Z_{k-1}$ and $\mathbf{x}_{k-1}|U_{k-1}, Z_{k-1}$, from the prior belief in equation (2.53) and the state transition probability in equation (2.54) as follows:

$$\begin{bmatrix} \mathbf{x}_k|U_k, Z_{k-1} \\ \mathbf{x}_{k-1}|U_{k-1}, Z_{k-1} \end{bmatrix} \sim \mathcal{N}(\mu_{t1}, \Sigma_{t1}), \tag{2.56}$$

$$\boldsymbol{\mu}_{t1} = \begin{bmatrix} \mathbf{A}\boldsymbol{\mu}_{k-1} + \mathbf{B}\mathbf{u}_k \\ \boldsymbol{\mu}_{k-1} \end{bmatrix} \tag{2.57}$$

$$\boldsymbol{\Sigma}_{t1} = \begin{bmatrix} \mathbf{A}\boldsymbol{\Sigma}_{k-1}\mathbf{A}^T + \mathbf{Q}_k & \mathbf{A}\boldsymbol{\Sigma}_{k-1} \\ (\mathbf{A}\boldsymbol{\Sigma}_{k-1})^T & \boldsymbol{\Sigma}_{k-1} \end{bmatrix} \tag{2.58}$$

Then, define $\widehat{\mathbf{x}}_k = (\mathbf{x}_k|U_k, Z_{k-1})$, and marginalise it using equation(2.40) such that:

$$\widehat{\mathbf{x}}_k \sim \mathcal{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k), \tag{2.59}$$

$$\widehat{\boldsymbol{\mu}}_k = \mathbf{A}\boldsymbol{\mu}_{k-1} + \mathbf{B}\mathbf{u}_k, \tag{2.60}$$

$$\widehat{\boldsymbol{\Sigma}}_k = \mathbf{A}\boldsymbol{\Sigma}_{k-1}\mathbf{A}^T + \mathbf{Q}_k. \tag{2.61}$$

Note that $\overline{bel}\,(\mathbf{x}_k) = p(\widehat{\mathbf{x}}_k)$.

Correction Step

The measurement $\mathbf{z}_k$ is incorporated by computing the joint distribution of $\mathbf{z}_k$ and $\widehat{\mathbf{x}}_k$ using the chain rule, i.e., $p(\mathbf{z}_k, \widehat{\mathbf{x}}_k)$=$p(\mathbf{z}_k|\widehat{\mathbf{x}}_k)\,p(\widehat{\mathbf{x}}_k)$, and then, rearranging

the variables as follows:

$$\begin{bmatrix} \widehat{\mathbf{x}}_k \\ \\ \mathbf{z}_k \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}_{t2}, \boldsymbol{\Sigma}_{t2}), \tag{2.62}$$

$$\boldsymbol{\mu}_{t2} = \begin{bmatrix} \widehat{\boldsymbol{\mu}}_k \\ \\ \mathbf{C}\widehat{\boldsymbol{\mu}}_k \end{bmatrix}, \tag{2.63}$$

$$\boldsymbol{\Sigma}_{t2} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_k & \widehat{\boldsymbol{\Sigma}}_k\mathbf{C}^T \\ \\ \mathbf{C}\widehat{\boldsymbol{\Sigma}}_k & \mathbf{C}\widehat{\boldsymbol{\Sigma}}_k\mathbf{C}^T + \mathbf{R}_k \end{bmatrix}. \tag{2.64}$$

Finally, $p(\widehat{\mathbf{x}}_k|\mathbf{z}_k)$ is computed using the conditional distribution in equation(2.42) as follows:

$$\widehat{\mathbf{x}}_k|\mathbf{z}_k \sim \mathcal{N}((\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.65}$$

$$\boldsymbol{\mu}_k = \widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\Sigma}}_k\mathbf{C}^T(\mathbf{C}\widehat{\boldsymbol{\Sigma}}_k\mathbf{C}^T + \mathbf{R}_k)^{-1}(\mathbf{z}_k - \mathbf{C}\widehat{\boldsymbol{\mu}}_k), \tag{2.66}$$

$$\boldsymbol{\Sigma}_k = \widehat{\boldsymbol{\Sigma}}_k - \widehat{\boldsymbol{\Sigma}}_k\mathbf{C}^T(\mathbf{C}\widehat{\boldsymbol{\Sigma}}_k\mathbf{C}^T + \mathbf{R}_k)^{-1}\mathbf{C}\widehat{\boldsymbol{\Sigma}}_k. \tag{2.67}$$

Note that $(\widehat{\mathbf{x}}_k|\mathbf{z}_k) = (\mathbf{x}|U_k, Z_k)$, therefore, $bel\,(\mathbf{x}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. In Kalman filter convention, the quantity $\mathbf{K}_k = \widehat{\boldsymbol{\Sigma}}_k\mathbf{C}^T(\mathbf{C}\widehat{\boldsymbol{\Sigma}}_k\mathbf{C}^T + \mathbf{R}_k)^{-1}$ is known as the *Kalman gain*.

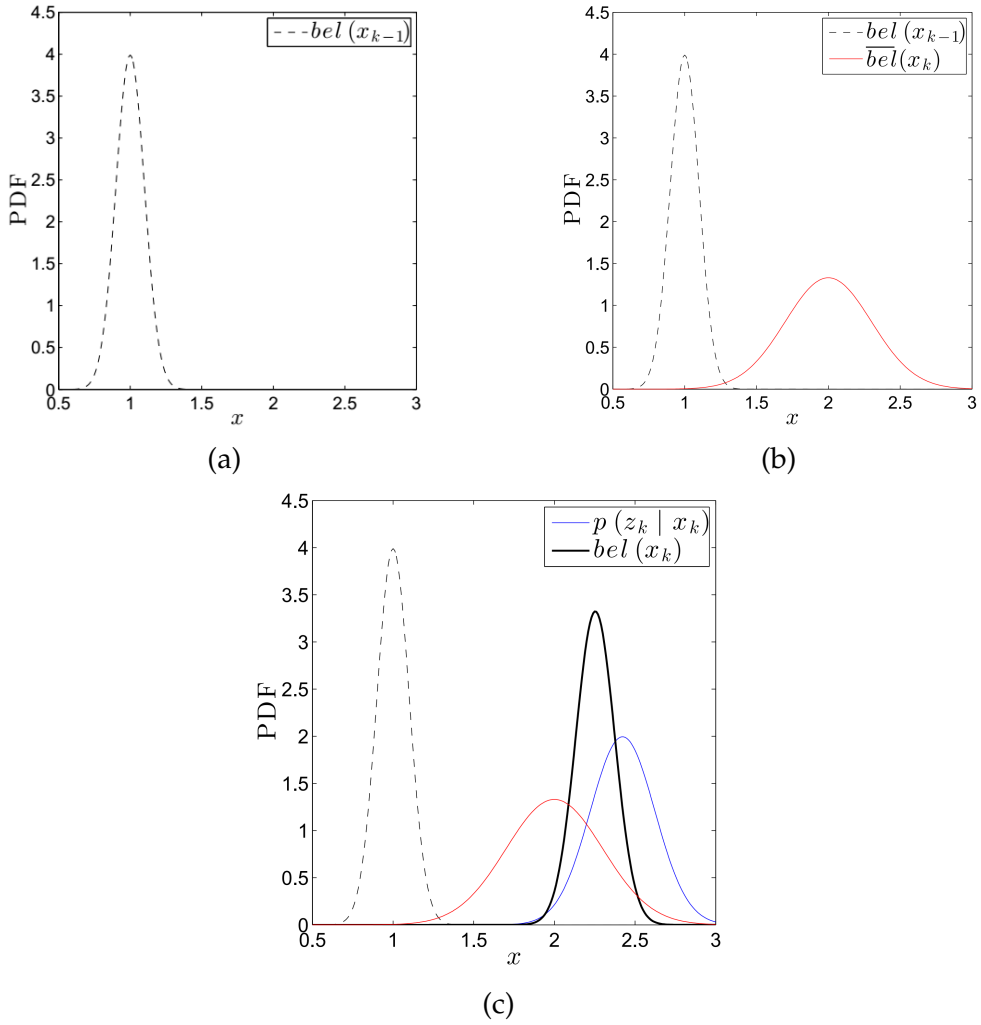

Figure 2.1: Kalman filter example in $\mathbb{R}$. (a) The initial system belief, (b) after integrating state transition model, the belief variance increases, and (c) by incorporating the observation model, the resulting belief $bel\,(x_k)$ has variance less than $\overline{bel}\,(x_k)$ and $p(z_k|x_k)$

Figure 2.1 illustrates the two steps of Kalman filter in $\mathbb{R}$. Note that the belief variance $\mathbf{\Sigma}_k$, also known as the *uncertainty*, is less than the variance after state transition $\widehat{\mathbf{\Sigma}}_k$ and measurement variance $\mathbf{R}_k$. The state transition and observation models can be thought of as having two sources of information about the system state, and Kalman filter fuses these two estimates and use weighted average to compute the best estimate with variance smaller than both.

Keep in mind that not all dynamical state transition models are linear, nor the associated noises are Gaussians. In the case of the former, *Extended Kalman filter* (EKF) is a popular approach to be applied in case of nonlinear systems. If the

noise is non Gaussian as well as the system is non linear, a nonparametric approach known as *particle filter*, or *Sequential Monte Carlo* (SMC), is used to estimate the belief by sampling the state space then assigning different weights to each sample.