University of Cyprus

Department of Electrical and Computer Engineering

Generalized Robust Estimation for a Class of Systems

Yiannis Socratous

Dissertation submitted as part of the requirements for the degree of Doctor of Philosophy at the University of Cyprus

September, 2010

Πανεπιστήμιο Κύπρου

Τμήμα Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών

Γενικευμένη Εύρωστη Εκτίμηση για μια Τάξη Συστημάτων

Γιάννης Σωκράτους

Διατριβή η οποία υποβλήθηκε προς απόκτηση διδακτορικού
τίτλου σπουδών στο Πανεπιστήμιο Κύπρου

Σεπτέμβρης, 2010

ABSTRACT

One of the most common signal processing tasks arising in various applications is the estimation of a signal from a noisy measurement. Most of the time imprecise $\acute{a}$ priori knowledge of input characteristics results in degradation of performance. This thesis presents a minimax approach to the design of robust estimation. Minimax methods are useful because they lead to constructive procedures for designing robust schemes. The main goal of the thesis is to derive robust least-square estimators for situation when the statistics or internal dynamics describing the signal and observations are not exactly known. Even though the primal focus is on robust minimax estimation, some aspect of other well-known estimation techniques, like the Maximum $\acute{A}$ Posteriori, the Maximum Likelihood estimation techniques, are also investigated. There are three main contributions in this thesis: 1) Modeling of uncertainty of a system using stochastic kernels, and joint distributions, and derivation of robust least-square estimators for various uncertainty sets, through a minimax approach. These uncertainty sets are defined by a Kullback-Leibler distance constraint. The results include existence of the optimal measures, and properties associated with the estimate of the true measure. Various examples, which also include MIMO communication models, are used to illustrate how the results apply to practical problems; 2) Applications of minimax theory developed to finite-dimensional autoregressive channel models in order to derive robust least-square estimators for a class of uncertain models. The methodology presented invokes a change of probability measure technique to derive recursive equations for the conditional distribution of nonlinear filtering problem. The conditional mean equation is solved explicitly to derive envelope and phase estimates for specific models such as the linear gaussian model and the non-coherent multipath model. For the non-coherent multipath model a connection with the classical lest-square estimation is also presented; 3) Derivation of a generalized Maximum $\acute{A}$ Posteriori estimator, and a generalized Maximum Likelihood estimator are derived. The methodology used involves the introduction of an exponential function in the cost definition and the

likelihood function of the the Maximum $\acute{A}$ Posteriori estimation and the Maximum Likelihood estimation technique, respectively. A connection with the minimax approach is also presented and some examples are solved to illustrate the application of the results to theoretical problems.

Περιληψη

Μία από τις πιο κοινές εργασίες στην επεξεργασία σήματος, η οποία προκύπτει σε διάφορες εφαρμογές, είναι η εκτίμηση ενός σήματος από μια μέτρηση που υπόκειται σε θόρυβο. Τις περισσότερες φορές ανακριβής γνώση των εκ των προτέρων χαρακτηριστικών εισαγωγής οδηγεί στην ανακριβή εκτίμηση ενός σήματος και στη μείωση της απόδοσης της σχετικής εφαρμογής. Αυτή η διατριβή παρουσιάζει μια προσέγγιση ελαχιστοποίησης -μεγιστοποίησης (minimax) στο σχεδιασμό εύρωστης εκτίμησης. Οι μέθοδοι minimax είναι πολύ χρήσιμες καθώς οδηγούν σε εποικοδομητικές διαδικασίες για το σχεδιασμό εύρωστων πλάνων. Ο κύριος στόχος της διατριβής είναι η εξαγωγή εύρωστων εκτιμητών ελαχίστων-τετραγώνων που θα μπορούν να χρησιμοποιηθούν σε περιπτώσεις όπου τα στατιστικά ή οι εσωτερικές δυναμικές που χαρακτηρίζουν ένα σήμα καθώς και οι παρατηρήσεις δεν είναι ακριβώς γνωστές. Αν και ο πρωταρχικός στόχος εστιάζεται στην εύρωστη εκτίμηση minimax, η διατριβή ερευνά και κάποιες πτυχές άλλων γνωστών τεχνικών εκτίμησης όπως είναι οι τεχνικές Μέγιστης εκ των Υστέρων (MAP) εκτίμησης και εκτίμησης Μέγιστης Πιθανοφάνειας (ML). Υπάρχουν τρεις κύριες συνεισφορές σε αυτή τη διατριβή: 1) Μοντελοποίηση της αβεβαιότητας ενός συστήματος χρησιμοποιώντας στοχαστικούς πυρήνες (stochastic kernels) και απόκοινού διανομές και εξαγωγή εύρωστων εκτιμητών ελαχίστων-τετραγώνων για διάφορα σύνολα αβεβαιότητας μέσω μιας προσέγγισης minimax. Αυτά τα σύνολα αβεβαιότητας καθορίζονται από τον περιορισμό της απόστασης Kullback-Leibler. Τα αποτελέσματα περιλαμβάνουν την ύπαρξη βέλτιστων μέτρων, καθώς και ιδιότητες που συνδέονται με την εκτίμηση του αληθινού μέτρου. Επιπρόσθετα, παρουσιάζονται διάφορα παραδείγματα, τα οποία περιλαμβάνουν και μοντέλα MIMO, όπου επιδεικνύεται η εφαρμογή των αποτελεσμάτων σε πρακτικά προβλήματα. 2) Εφαρμογή της θεωρίας minimax σε πεπερασμένα-διαστατικά (finite-dimensional) αυτοανάδρομα (autoregressive) μοντέλα καναλιών για την εξαγωγή εύρωστων εκτιμητών ελαχίστων-τετραγώνων για μια τάξη αβέβαιων μοντέλων. Η μεθοδολογία που παρουσιάζεται χρησιμοποιεί μια τεχνική αλλαγής του μέτρου πιθανότητας για την εξαγωγή αναδρομικών εξισώσεων για την υπό όρους διανομή ενός μη γραμμικού προβλήματος φιλτραρίσματος. Η υπό όρους μέση

εξίσωση λύνεται ρητά για να εξαχθούν εκτιμήσεις της περιβάλλουσας και της φάσης για συγκεκριμένα μοντέλα όπως το γραμμικό μοντέλο Gaussian και το μη-συνεκτικό (non-coherent) μοντέλο πολλαπλών διαδρομών. Επίσης, παρουσιάζεται η σχέση του μη-συνεκτικού μοντέλου πολλαπλών διαδρομών με την κλασσική εκτίμηση ελαχίστων-τετραγώνων. 3) Εξαγωγή γενικευμένων εκτιμητών MAP και ML. Η μεθοδολογία που χρησιμοποιείται περιλαμβάνει την εισαγωγή μιας εκθετικής συνάρτησης στον ορισμό του κόστους και της συνάρτησης πιθανότητας των εκτιμητών MAP και ML αντίστοιχα. Τέλος, παρουσιάζεται η σχέση με την προσέγγιση minimax και επιλύονται επίσης μερικά παραδείγματα που επιδεικνύουν την εφαρμογή των αποτελεσμάτων σε θεωρητικά προβλήματα.

ACKNOWLEDGEMENT

First and foremost, I would like to thank my academic advisor Professor Charalambos D. Charalambous, for his patience, continuous guidance and support throughout these years. I have benefited tremendously from his profound way of thinking, comprehensive knowledge, vision and enthusiasm. Also, I would like to thank all the faculty, administrative staff and fellow colleagues at the University of Cyprus.

I would like to express a special gratitude to my friends and colleagues Stelios, Michalis Markou, Ioannis, Kiriakos, Costas, Michalis Michaelides and Stojan for their help and support through past years. Furthermore, I would like to thank the student group of Professor Charalambos D. Charalambous, for their patience (especially during my presentation rehearsals) and their valuable comments.

Finally, I would like to thank my family and friends for their support. Special thanks goes to my wife Eleni, for her unconditional love, for showing patience throughout these years and for keeping me focused on my goals and what is truly important in life. Last but not least, I would like to thank my sons Marcos and Paris (he will be shortly arriving in our life) for keeping me alert and giving me an extra motive in order to complete this difficult journey.

CONTENTS

List of Figures

CHAPTER 1

INTRODUCTION

Signal processing is an area of electrical engineering, systems engineering, and applied mathematics that deals with operations on or analysis of signals, in either discrete or continuous time to perform useful operations on those signals. Signals of interest can include sound, images, time-varying measurement values and sensor data, for example, biological data such as, electrocardiograms, control system signals, telecommunication transmission signals, such as radio signals, and many others. Signals are analog or digital representations of time-varying or spatial-varying physical quantities.

One of the most common signal processing tasks arising in applications is that of estimating (e.g., filtering, predicting, or smoothing) a signal waveform from a noisy measurement. Signal estimation is the area of study that deals with the processing of signals that contain information in order to extract this information from them. Applications of the theory of estimation is found in many areas such as communications, controls, and signal processing. This task arises, for example, in radar and sonar tracking systems, in observers for automatic control systems, in demodulators for analog communication systems, and in medical imaging systems.

In communications applications such as data transmission or radar, estimation provides the theoretical and analytical basis for the design of effective communication transmitters and receivers. Most of the time estimation applications involves taking decisions based on observations that are distorted or corrupted by noise. Moreover, the information that one wishes to extract from such observations is unknown to the observer. In such problems it is useful to formulate estimation problems in a probabilistic framework in which the unknown behavior is modeled by probability distributions.

Basic to the study of signal estimation theory is the concept of a random observation $Y : (\Omega, \mathcal{F}) \rightarrow (\mathcal{Y}, \Sigma_{\mathcal{Y}})$, where $(\Omega, \mathcal{F})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ are measurable spaces. Here $\mathcal{Y}$ maybe a set of vectors, real numbers, or any other set. From the observation of Y one wishes to extract information for some phenomenon related to Y, and in the case of estimation problems one wishes to estimate the value of a quantity that is not observed directly. This relation between the observation and the desired information is probabilistic in the sense that the statistical behavior of Y is influenced by the value of the quantity to be estimated. This is why a model for this situation must involve a family of probability distributions on $\mathcal{Y}$, the members of which correspond to statistical conditions present under the various values of the quantities to be estimated. Under this model the estimation problem is to find an optimum way of processing the observation Y in order to extract the desired information. The basic features that distinguish such problem from each other are the nature of the desired information (discrete or continuous), the amount of $\acute{a}$ priori knowledge that is available about their desired quantities and the performance criteria by which various estimation procedures are graded. If the information we want to extract are some static parameters that do not change with time, then the problem is defined as parameter estimation. If the parameters need to be estimated are dynamic or time-varying then the problem is defined as signal estimation.

Conventional design procedures for optimum signal estimation algorithms often require an exact knowledge of the statistical behavior both of the signal of interest and of the noise corrupting the observations. For example, in the design of optimum linear estimation algorithms someone must know the spectral or autocorrelation properties of the signal and noise in order to specify the optimum procedures, and procedures designed to be optimum for a given model can be undesirably sensitive to inaccuracies or uncertainties in the model.

The two basic sources of uncertainty are noisy data and imprecise knowledge of the model of the underlying physical system. Often, $\acute{a}$ priori knowledge is available to describe, in a probabilistic sense, the two aforementioned sources of uncertainty, e.g., bounds, moments, mixed moments, distribution, stationarity, dependence, and spectral properties. Of course, the application of such information in the estimation procedure should yield more reliable solutions

Because the signals and noise in signal processing applications are usually modeled as random processes, performance measures usually involve probabilistic quanti-

ties (such as mean squared error or probability of error). The theory of statistics has played a fundamental role in the development of optimum signal processing techniques.

Suppose a parameter estimator for a signal with known waveform in additive noise, is designed to give optimum performance for noise possessing a specific statistical description. For example, one widespread model for noise is the Gaussian process. An important question that arises is, how sensitive is the performance of such an optimum scheme to deviations in the signal and noise characteristics from those for which the scheme is designed? This is an important question because in practice one rarely has perfect knowledge of, say, the noise characteristics; the Gaussian or any other specific model is usually a nominal assumption which may at best be approximately valid most of the time. Unfortunately, it turns out that in many cases nominally optimum signal processing schemes can suffer a drastic degradation in performance even for apparently small deviations from nominal assumptions. It is this basic observation that motivates the search for robust signal processing techniques; that is, techniques with good performance under any nominal conditions and acceptable performance for signal and noise conditions other than the nominal, which can range over the whole of allowable classes of possible characteristics. Thus, in seeking robust methods it is recognized at the outset that a single, precise characterization of signal and noise conditions is unrealistic, and so classes of possible signal and noise characterizations are constructed and considered.

Often a class of allowable characteristics, say for a noise power density function, is constructed by starting with a nominal characteristic and then including in the class all other characteristics that are "close," in some well-defined sense, to this nominal one. Then a signal processing scheme that is robust may have performance at the nominal which is not quite as good as the scheme that is optimum for the nominal case, but its overall performance with respect to the defined class of characteristics will be good or acceptable. This loose definition of robustness is perfectly reasonable, but it does not provide a systematic approach to obtaining robust schemes. In order for this to be achieved first a measure of "overall" performance of a scheme with respect to a class of allowable conditions at the input is specified. One such measure that has been widely used and which leads to interesting and useful results in many situations is the worst case performance of a scheme over a class of input conditions. Clearly, if its worst case performance is good, then it maybe concluded that a given scheme is robust. On the other hand, such a robust scheme can be found by looking

for the scheme that optimizes worst case performance. This approach leads to what are known as minimax robust schemes. A scheme that minimizes the maximum possible value of a loss function is called minimax; if performance is measured by a gain function then a maximin scheme would be sought. The term minimax is used as a general description for such schemes in all cases. Implicit in the association of minimax schemes with robust schemes is the expectation that the worst case performance of a minimax scheme will be acceptably good, being the best that can be achieved. Another expectation one has in defining robust schemes in this way is that at any nominal operating point, the performance of the minimax scheme will not be very far below that of the nominally optimum scheme, which on the other hand will have much poorer performance away from the nominal point.

The rest of the chapter is organized as follows. In Section 1.1, the review of related literature is given. In Section 1.2, the thesis goals are introduced and motivated. In Section 1.3, the statement of the problems are presented. In Section 1.4, the main contributions of the thesis are outlined.

1.1 Survey of Related Research

This section summarizes the models and results on classical estimation methods and robust minimax estimation techniques which are relevant to this thesis.

1.1.1 Classical Estimation Methods

In many situations arising in practice one is interested in making a choice among a continuum of possible states of nature. In particular, given a family of distributions on the observation space indexed by a parameter or set of parameters, the true value of the parameter has to be determined as precisely as possible from the observations. Such problems are known as parameter estimation problems. In this thesis two basic approaches to parameter estimation are presented, first the Bayesian, in which the parameter is assumed to be a random quantity related statistically to the observation, and a second, the Maximum Likelihood estimation, in which the parameter is assumed to be unknown but without being characterized with a probabilistic

structure.

Bayesian Estimation

In estimation theory and decision theory, a Bayes estimator or a Bayes rule is an estimator or decision rule that minimizes the posterior expected value of a cost function. Suppose an unknown Random Variable (RV) $X : (\Omega, \mathcal{F}) \rightarrow (\mathcal{X}, \Sigma_{\mathcal{X}})$ is known to have a prior distribution P_X. Let $\Phi(Y)$ be an estimator of X (based on the measurements of a RV, Y), and let $C(X, \Phi(Y))$ be a cost function. The Bayes risk of $\Phi(Y)$ is defined as $E[C(X, \Phi(Y))]$, where the expectation is taken over the probability distribution of X: this defines the risk function as a function of $\Phi(Y)$. An estimator $\Phi(Y)$ is said to be a Bayes estimator if it minimizes the Bayes risk among all estimators.

One of the most popular Bayesian estimation techniques is the Least-Square estimation, also known as Minimum-Mean-Squared-Error estimation. In classical Least-Square estimation of RV's one is interested in finding the best estimate, $\Phi^*(Y)$, of a RV $X : (\Omega, \mathcal{F}) \rightarrow (\mathcal{X}, \Sigma_{\mathcal{X}})$ from the measurements of a RV Y, by minimizing the expected value of the least-square cost function $C(X, \Phi) \stackrel{\triangle}{=} \|X - \Phi(Y)\|^2_{\mathcal{X}}$ over all functions $\Phi : \mathcal{Y} \rightarrow \mathcal{X}$,$Y \mapsto \Phi(Y)$, which is a function of Y ($\|x\|^2_{\mathcal{X}}$ denotes the norm on $\mathcal{X}$). By the orthogonal projection theorem, the solution is given by the conditional expectation:

$$\Phi^*(Y) = E[X|Y] = \int_{\mathcal{X}} x \; dP_{X|Y}(x|y) \tag{1.1}$$

where $P_{X|Y}$ is the conditional distribution of X given Y [1], [2], [3].

The solution to this estimation problem is stated in terms of the *á* posteriori distribution function $P_{X|Y}$. This distribution contains all information available to estimate X or any nonlinear function of it. The objective is thus to find, recursively in time, the evaluation equation of the *á* posteriori distribution and to solve it (for the case of random processes). However, it is only in a few special cases that this distribution can be solved explicitly by parameterizing it using a finite number of statistics. A well known example is the case with linear dynamics and observations in additive Gaussian noise. In this case all densities involved are Gaussian, and hence the conditional distribution can be parameterized using the corresponding mean and covariance. The equations of the finite statistics are given by the Kalman Filter [1], [4], which are going to be explained in more details in the following paragraphs.

For the case of nonlinear systems, there are difficulties in obtaining the solution of the $\acute{a}$ posteriori distribution recursion in closed form. Often, approximations have to be made to find sub-optimal nonlinear estimators. The standard method is to use the Taylor series expansion and apply linear filtering theory, giving rise to the so-called Extended Kalman Filter [5], [6], [7]. Other more sophisticated sub-optimal estimation techniques are available, e.g. reiteration, higher order filters, and statistical methods [5].

Another cost function that is sometimes applied is the absolute error, given by $C(X, \Phi(Y)) = |X - \Phi(Y)|, \ \ (X, \Phi(Y)) \in \Re^2$. The Bayes risk here is $E[|\Phi(Y) - X|]$, a quantity known as the mean-absolute-error, so the corresponding Bayes estimate is known as the Minimum-Mean-Absolute-Error (MMAE) estimate. The Bayes estimate in this case, denoted by $\Phi_{ABS}(y)$, is any point such that

$$\begin{aligned} P(X < t|Y = y) \leq P(X > t|Y = y), \ \ t < \Phi_{ABS}(y) \\ P(X < t|Y = y) \geq P(X > t|Y = y), \ \ t > \Phi_{ABS}(y) \end{aligned} \tag{1.2}$$

Note that a point $\Phi_{ABS}(y)$ satisfying (1.2) is a median of the conditional distribution of X given $Y = y$. Thus the MMAE estimate is a conditional median estimate. This estimate coincides with the Least-Square (LS) estimate only when the distribution of X given $Y = y$ has the same value for the mean and median.

Moreover, another estimation method that, although not properly a Bayes estimate, fits within the Bayesian framework is Maximum $\acute{A}$ Posteriori (MAP) probability estimation. In Bayesian statistics, a MAP estimate is a mode of the posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data. It is closely related to Maximum Likelihood estimation, which is presented in the following paragraphs, but employs an augmented optimization objective which incorporates a prior distribution over the quantity one wants to estimate. MAP estimation can therefore be seen as a regularization of Maximum Likelihood (ML) estimation.

It is assumed that an unobserved RV X has to be estimated on the basis of observations Y. Applying a uniform cost criterion, leads to a procedure for estimating X as that value which maximizes the $\acute{a}$ posteriori (discrete or continuous) density $\eta(y, dx)$ (also referred in this dissertation as stochastic kernel). Therefore,

$$\Phi_{MAP}(y) = \arg\max_{x \in \mathcal{X}} \eta(y, dx) \tag{1.3}$$

In modeling a given statistical situation usually someone starts with the family of conditional distributions (or stochastic kernels) of Y given $X = x$, and for the Bayesian formulation with a prior distribution for X also. The conditional distribution of X given Y (or the stochastic kernel $\eta(y, dx)$) is obtained from the prior and the conditional of Y given X (or the stochastic kernel) $\mu(x, dy)$) by applying Bayes' formula.

$$\eta(y, dx) = \frac{\mu(x, dy)dP_X(x)}{\int_{\mathcal{X}} \mu(x, dy)dP_X(x)} = \frac{\mu(x, dy)dP_X(x)}{dP_Y(y)} \tag{1.4}$$

The MAP estimator can be obtained using (1.4) but without the computation of the denominator of the posterior distribution ($dP_Y(y)$) since this term does not depend on x and therefore plays no role in the optimization. That is,

$$\Phi_{MAP}(y) = \arg\max_{x \in \mathcal{X}} \mu(x, dy)dP_X(x) \tag{1.5}$$

The above MAP estimate of x coincides with the ML estimate when the prior $dP_X(x)$ is uniform (that is, a constant function). The MAP estimate is a limit of Bayes estimators under a sequence of $0 - 1$ cost functions, but not a Bayes estimator per se.

Maximum Likelihood Estimation

For many observation models arising in practice, it is not possible to apply the above results either because of intractability of the required analysis or because of the lack of useful complete sufficient statistic. For such models, an alternative method for seeking good estimators is needed. One very commonly used method of designing estimators is the Maximum Likelihood (ML) method. Maximum Likelihood estimation is a popular statistical method used for fitting a statistical model to data, and providing estimates for the model's parameters.

The ML estimator it is one of the most fundamental estimators in statistics. For a fixed set of data and underlying probability model, ML picks the values of the model parameters that make the data "more likely" than any other value of the parameters would make them. ML estimation gives a unique and easy way to determine the solution in the case of the normal distribution and many other problems, although in very complex problems this may not be the case. If a uniform prior distribution is assumed over the parameters, the ML estimate coincides with the most probable values thereof.

Consider a family of probability distributions parameterized by an unknown parameter x (which could be vector-valued), associated with either a known probability density function (continuous distribution) or a known probability mass function (discrete distribution), denoted as $\mu(x, \cdot)$. A sample $\{y_1, y_2, \ldots, y_n\}$ of n values is drawn from this distribution, and then $\mu(x, \cdot)$ is used for computing the (multivariate) probability density associated with the observed data, $\mu(x, dy)$. As a function of x with $y_1, \ldots, y_n$ fixed, this is the likelihood function

$$\mathcal{L}(x) = \mu(x, dy).$$

The method of maximum likelihood estimates x by finding the value of x that maximizes $\mathcal{L}(x)$. This is the Maximum Likelihood estimator of x:

$$\widehat{x}_{ML}(y) = \arg\max_x \mathcal{L}(x). \tag{1.6}$$

Kalman Filtering

All the estimation methods discussed so far are used for designing estimators for static parameters, that is, for parameters that are not changing with time, although these methods are applied to sequences. In many applications the related problems of estimating dynamic or time-varying parameters is of interest. In the traditional terminology, a dynamic parameter is usually called a signal, so the above problem is known as signal estimation. The dynamic nature of the parameter in signal estimation problems adds a new dimension to the statistical modeling of the problems where the dynamic properties of the signal must be modeled statistically in order to obtain meaningful signal estimation procedures. Unlike the static case, an estimator of a signal is not expected to be perfect as the number of observations becomes infinite because of the time variation in the signal.

Kalman filtering provides a very useful algorithm for estimating signals that are generated by finite-dimensional linear dynamically models. The Kalman filter is an algorithm in control theory introduced by Kalman (1960) and refined by Kalman and Bucy (1961). It is a minimum-mean-square-error estimator and it is an algorithm which makes optimal use of imprecise data on a linear (or nearly linear) system with Gaussian errors, and continuously update the best estimate of the system's current state.

Kalman filter theory is based on a state-space approach in which a state equation models the dynamics of the signal generation process and an observation equation

models the noisy and distorted observation signal. For a signal x_k and noisy observation y_k, equations describing the state process model and the observation model are defined as

$$\begin{aligned} x_k &= Ax_{k-1} + w_{k-1}, \quad x_0 \sim N(Ex_0; cov(x_0)) \\ y_k &= Hx_k + n_k, \end{aligned} \tag{1.7}$$

where, x_k is the p-dimensional signal vector, or the state parameter, at time k, A is a $p \times p$ dimensional state transition matrix that relates the states of the process at times $k-1$ and k, w_k (process noise) is the p-dimensional uncorrelated input excitation vector of the state equation. w_k is assumed to be a normal (Gaussian) process $p(w_k) \sim N(0, Q)$, Q being the $p \times p$ covariance matrix of w_k or process noise covariance. y_k is the M dimensional noisy observation vector, H is a $M \times p$ dimensional matrix n_k is the M-dimensional noise vector, also known as measurement noise, n_k is assumed to have a normal distribution $p(n_k) \sim N(0, R)$ and R is the $M\times$ covariance matrix of n_k (measurement noise covariance). Often, (n_k, w_k, x_0) are assumed independent.

Here, $\hat{x}_{k|k-1}$ is defined as the *á* priori estimate (prediction) at step k from the previous trajectory of x given measurements $y_0, \ldots, y_{k-1}$, and $\hat{x}_{k|k}$ as the *á* posteriori state estimate at step k given measurements $y_0, \ldots, y_k$. Note that $\hat{x}_{k|k-1}$ is a prediction of the value of x_k which is based on the previous values and not on the current observation at time k. $\hat{x}_{k|k}$ on the other hand, uses the information in the current observation and previous $y_0, \ldots, y_{k-1}$.

The Kalman filter is often derived by beginning with the goal of finding an equation that computes an *á* posteriori state estimate as a linear combination of an *á* priori estimate (prediction) and a weighted difference between an actual measurement and a measurement prediction (innovation). Hence, each estimate consists of a fraction which is predictable from the previous values and does not contain new information and a fraction that contains the new information extracted from the observation.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k\Big(y_k - H\hat{x}_{k|k-1}\Big). \tag{1.8}$$

The difference $y_k - H\hat{x}_{k|k-1}$ in (1.8) is called the measurement innovation. The innovation reflects the discrepancy between the predicted value and the actual measurement. The $P \times M$ matrix, K_k, in (1.8) is chosen to be the gain or blending factor that minimizes the *á* posteriori error covariance. One form of the resulting K_k is

given by

$$K_k = \Sigma_{k|k-1} H^T \Big(H \Sigma_{k|k-1} H^T + R \Big)^{-1} \tag{1.9}$$

where $\Sigma_{k|k-1}$ is the covariance of the prediction error, $x_k - \hat{x}_{k|k-1}$, conditioned on y_0^{k-1}.

The Kalman filter estimates a process by using a form of feedback control: the filter estimates the process state at some time and then obtains feedback in the form of (noisy) measurements. As such, the equations for the Kalman filter fall into two groups: time update equations (prediction) and measurement update equations (correction). The time update equations are responsible for projecting forward (in time) the current state and error covariance estimates to obtain the $\acute{a}$ priori estimates for the next time step. The measurement update equations are responsible for the feedback i.e. for incorporating a new measurement into the $\acute{a}$ priori estimate to obtain an improved $\acute{a}$ posteriori estimate.

The time update equations can also be thought of as predictor equations, while the measurement update equations can be thought of as corrector equations. Indeed the final estimation algorithm resembles that of a predictor-corrector algorithm for solving numerical problems as shown below.

Time update (predict)

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1}, \quad \Sigma_{k|k-1} = A\Sigma_{k-1|k-1}A^T + Q, \;\; \Sigma_{0|-1} = cov(x_0), \tag{1.10}$$

where $\Sigma_{k|k} = \Sigma_{k|k-1} - \Sigma_{k|k-1} H^T \Big(H \Sigma_{k|k-1} H^T + R \Big)^{-1} H \Sigma_{k|k-1}$.

Measurement update (correct)

$$\begin{aligned} K_k &= \Sigma_{k|k-1} H^T \Big(H \Sigma_{k|k-1} H^T + R \Big)^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k \Big(y_k - H\hat{x}_{k|k-1} \Big), \;\; x_{0|-1} = Ex_0. \end{aligned} \tag{1.11}$$

1.1.2 Uncertainty Models

In practical applications one of the important issues in estimation problems is the choice of an appropriate system or channel model. When the channel is uncertain, this is not a trivial problem. The uncertainty models can be divided into two

categories, the parametric uncertainty models and the nonparametric uncertainty model. In the following paragraphs some typical examples from the two categories are presented.

Parametric Uncertainty Models

In the estimation theory literature a basic model, which is often used to describe a parametric uncertainty in the channel is the one depicted in Fig. 1.1 [8]. A variable θ, which belongs to a certain set Θ, parameterizes a conditional distribution $Q(y|x,\theta)$. Hence, instead of dealing with a fixed known system, one considers the estimation for the class of systems $\{Q(y|x,\theta) : \theta \in \Theta\}$. Generally speaking, there are two classes of uncertainty models; the class of compound systems, and the class of arbitrarily varying systems. Compound and Arbitrarily Varying Channels are further classified into discrete memoryless and finite-state systems [8].

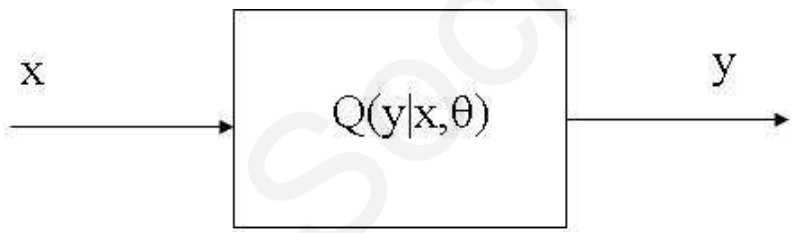


Figure 1.1: Probabilistic representation of a communication channel

A family of discrete memoryless systems

$$\{Q(y|x,\theta) : x \in \mathcal{X}^n,\ y \in \mathcal{Y}^n,\ \theta \in \Theta\}_{n=1}^{\infty} \tag{1.12}$$

where

$$Q(y|x,\theta) = \prod_{t=1}^{n} Q(y_t|x_t,\theta_t), \tag{1.13}$$

and $\{Q(y_t|x_t,\theta_t) : x_t \in \mathcal{X},\ y_t \in \mathcal{Y},\ \theta \in \Theta\}$ is a suitable subset of the set of all stochastic matrices $\mathcal{X} \times \Theta \mapsto \mathcal{Y}$, is called a discrete memoryless compound system. Thus, compound systems assume that the true system $Q_{true}(y|x,\theta)$ is unknown, though, it is assumed that $Q_{true}(y|x,\theta)$ belongs to the family of systems (1.12) and remains unchanged during the course of a transmission.

Arbitrarily Varying Channels (AVCs) are a generalization of (1.12) to include time variations in which the system changes during subsequent transmitted letters x_k.

Assume that Σ is a finite set of system states and $\theta = \Sigma^{\infty}$. Then an AVC is determined by

$$Q(y|x,s) = \prod_{t=1}^{n} Q(y_t|x_t,s_t), \tag{1.14}$$

where $s = (s_1, \ldots, s_n)$, and $\{Q(y_t|x_t,s_t) : x_t \in \mathcal{X},\ y_t \in \mathcal{Y},\ s_t \in \Sigma\}$ is a suitable subset of the set of all stochastic matrices $\mathcal{X} \times \Sigma \mapsto \mathcal{Y}$. Hence, at each moment t, the transmission matrix $Q(y_t|x_t,s_t)$ is unknown, and it is determined by the system state $s_t \in \Sigma$.

Continuous alphabet uncertainty systems received much less attention in the literature than the discrete counterparts. Most of the results are related to Gaussian uncertainty systems, which are briefly described below.

A Gaussian compound channel is defined by

$$Y = HX + W, \tag{1.15}$$

The uncertainty is introduced by assuming that a linear transformation H in unknown, although it is known to belong to some pre-specified class of linear transformations, e.g. $\|H\| \leq \gamma$, or $H = \tilde{H} + \delta H, \|\delta H\| \leq p$ and $\tilde{H}$ is known. X and Y are in general continuous signals, and W is an additive Gaussian noise. These models can also be used when H is known and there is uncertainty in X, which can be described with a constraint on the weighted norm $\|X\| \leq L$ or a constraint on the covariance of X, $C_X \leq S$.

A generalized parametric uncertainty model is being used in [9] which accounts for uncertainties in the data $\{A, b\}$. The uncertainties in these data, expressed as $\{\delta A, \delta b\}$, are assumed to lie within certain balls of radii $\{\eta, \eta_b\}$, i.e., they are known to be bounded and satisfy $\|\delta A\| \leq \eta$, $\|\delta b\| \leq \eta_b$. A special case is also given in [9] where the perturbations $\{\delta A, \delta b\}$ are assumed to satisfy a model of the form

$$[\delta A \ \ \delta b] = HS[E_a \ \ E_b], \tag{1.16}$$

where S is an arbitrary contraction, $\|S\| \leq 1$, and $\{H, E_a, E_b\}$ are known quantities of appropriate dimensions. As a brief motivation, one application of state-space estimation is succinctly described in [9], with full details provided in [10]. For this case the following uncertainty model is considered

$$\begin{aligned} X_{i+1} &= (F_i + \delta F_i)X_i + (G_i + \delta G_i)U_i \ \ i \geq 0, \\ Y_i &= H_i + V_i \end{aligned} \tag{1.17}$$

where the perturbations in $\{F_i, G_i\}$ are modeled as

$$[\delta F_i \;\; \delta G_i] = M_i \Delta_i [E_{f,i} \;\; E_{g,i}] \tag{1.18}$$

for some known matrices $\{M_i, E_{f,i}, E_{g,i}\}$ and for an arbitrary contraction Δ_i, $\|\Delta_i\| \leq 1$.

Of special interest are Multiple-Input Multiple-Output (MIMO) channel models, due to the applications of multiple-antenna systems for wireless communication. Initially, multiple-antenna systems promised considerable gain for fading channels as compared to single-antenna systems. However, at a later stage, it has been shown that this gain depends on the level of knowledge that the transmitter and the receiver have about the channel. The received signal of a flat fading channel is given by

$$y = HX + W \tag{1.19}$$

where x is a transmitted vector in $\mathbb{C}^n$, w and y are random variables in $\mathbb{C}^d$, and H is a channel matrix in $\mathbb{C}^{d\times n}$. The additive noise w and channel matrix H are ergodic and stationary, and their entries are i.i.d., zero mean, circularly symmetric complex Gaussian random variables. If the channel matrix H is not perfectly known to the transmitter and /or receiver, the uncertainty is modeled as additive.

$$H = \hat{H} + E. \tag{1.20}$$

Here, $\hat{H}$ represents the estimation of H, while E is an estimation error.

An example of a flat-fading MIMO uncertainty model is given in [11]. Given a channel matrix H, the noisy observation can be expressed as (1.19) where for this problem x is the η_T-dimensional symbol vector transmitted during a signaling period and demotes a zero-mean complex circular Gaussian noise vector with covariance matrix R. It is assumed that R is positive definite so that the noise affects all observation components, and x has zero-mean and normalized covariance matrix I_{η_T} and is independent of w. This paper considers a robust Mean-Square-Error (MSE) equalizer design for MIMO communication systems with imperfect channel and noise information at the receiver. When the estimated channel $\hat{H}$ and the noise covariance matrix $\hat{R}$ are different from the actual channel H and noise covariance matrix R, respectively, the MSE objective function is expressed as a function of H_Δ, which represents the difference between the actual channel H and the estimated channel $\hat{H}$, i.e. $H_\Delta = H - \hat{H}$. It is also assumed that the estimated covariance matrix $\hat{R}$ is invertible. It also uses the Kullback-Leibler (KL) divergence to measure the distance

between the actual model (H, R) and the estimated model $(\hat{H}, \hat{R})$. Conditioned on the knowledge of x, the KL divergence between the actual model $y \sim N(Hx, R)$ and the estimated model $y \sim N(\hat{H}x, \hat{R})$ takes the form

$$D(f(y|x), \hat{f}(y|X)) = \int \ln \left[\frac{f(y|x)}{\hat{f}(y|x)} \right] f(y|x) dy \tag{1.21}$$

In this respect, it is worth noting that the KL divergence has been used also in [12] to develop a minimax formulation of robust detection. The use of the KL divergence is rather natural as a metric for model mismatch since it is commonly used by statisticians [13] for fitting statistical models, and by using a differential geometric viewpoint it is argued in [14] that the KL divergence is the natural geometric "distance" between systems.

Nonparametric Uncertainty Models

It should be emphasized that the classes of allowable characteristics one deals with in robust signal processing are generally nonparametric function classes, such as the class of all power spectral density functions with specified total power (area under the function), which lie between specified upper and lower bounding functions. The uncertainty class for nonparametric uncertainty models can be specified in a variety of ways [15]. It can be based on an $\epsilon - contamination$ model of the type originally proposed by Huber, a total variational model, a spectral band model wherein the power spectral densities specifying the signal and observations are required to stay within a band centered on the nominal Power Spectral Density (PSD) or even, with a probability density model, where the probability density specifying the channel is also required to stay within a constraint set with reference to the nominal Probability Density Function (PDF).

Consider the model

$$Y_k = S_k + N_k. \tag{1.22}$$

The processes $\{S_k\}$ and $\{N_k\}$ represent signal and noise, respectively and f_S and f_N are the power spectral densities of the signal and the noise. The spectral chosen for designing the estimator may differ from the true signal and noise spectral. This spectral uncertainty is modeled by choosing appropriate classes of PSD's, $\mathcal{G}$ and $\mathcal{N}$, respectively $f_S \in \mathcal{G}$ and $f_N \in \mathcal{N}$. Next, some specific forms for the uncertainty

classes $\mathcal{G}$ and $\mathcal{N}$ are presented. These forms have been widely used to model uncertainty in both the engineering and statistics literature. These forms will be presented for the class $\mathcal{G}$ but, they could easily be used to model noise spectral uncertainty.

The most commonly used uncertainty class is the $\epsilon - contaminated$ model, which is also known as the $\epsilon - mixture$ or gross-error model. This uncertainty class has the following form

$$\mathcal{G}_\epsilon \stackrel{\triangle}{=} \quad \Big\{ f_S : f_S(\theta) = (1-\epsilon) f_S^0(\theta) + \epsilon f_S'(\theta), \;\; \forall\, \theta \in [-\pi, \pi], \\ \int_{-\pi}^{\pi} f_S'(\theta)\lambda(d\theta) = \int_{-\pi}^{\pi} f_S^0(\theta)\lambda(d\theta) \Big\} \tag{1.23}$$

where f_S^0 is a nominal PSD and ϵ $(0 \leq \epsilon \leq 1)$ is the contamination parameter. This class is probably the most popular for representing uncertainty because it models the idea that there is a fraction ϵ of completely general uncertainty about the choice of the PSD f_S^0.

Another common model is the total variational model which has the form

$$\mathcal{G}_{TV} \stackrel{\triangle}{=} \Big\{ f_S : \frac{1}{2} \int_{-\pi}^{\pi} |f_S^0(\theta) - f_S(\theta)| \lambda(d\theta) \leq \epsilon, \int_{-\pi}^{\pi} f_S(\theta)\lambda(d\theta) = \int_{-\pi}^{\pi} f_S^0(\theta)\lambda(d\theta) \Big\} \tag{1.24}$$

where, again, f_S^0 is a nominal PSD and ϵ an uncertainty parameter.

A third model is the band model which has the form

$$\mathcal{G}_B \stackrel{\triangle}{=} \Big\{ f_S : f_S^L(\theta) \leq f_S(\theta) \leq f_s^U(\theta), \;\; \forall\, \theta, \; \int_{-\pi}^{\pi} f_S(\theta)\lambda(d\theta) = 2\pi w \Big\} \tag{1.25}$$

where $\int_{-\pi}^{\pi} f_S^L \lambda(d\theta) \leq 2\pi w \leq \int_{-\pi}^{\pi} f_S^U(\theta)\lambda(d\theta)$ and w is the known power of the signal. The name band model comes from the interpretation that f_S^L and f_S^U are the lower and upper bounds of a confidence band around a spectral estimate.

A fourth model of interest is the p-point model which has the form

$$\mathcal{G}_B \stackrel{\triangle}{=} \Big\{ f_S : \int_{A_i} f_S(\theta)\lambda(d\theta) = 2\pi w_i, \; i = 1, \ldots, n \Big\} \tag{1.26}$$

where the A_i's are a partition of $[-\pi, \pi]$ and $\sum_{i=1}^{n} w_i = w$, the power of the signals. A p-point class is an appropriate model of uncertainty in situations where, for example, the power w_i in each interval $A_i \stackrel{\triangle}{=} [\theta_{i-1}, \theta_i]$ (where $-\pi = \theta_0 < \theta_1 < \ldots < \theta_n = \pi$) can be accurately measured using a nested bank of low-pass filters or a bank of bandpass filters.

Note that for each of these classes it is assumed that the power is known. Often it is a reasonable assumption that the power can be accurately estimated even though the shape of the PSD is uncertain.

Nonparametric uncertainty models can also be formulated through probability density functions [8]. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets denoting the channel input and output alphabets, respectively. The probability law of a (known) channel is specified by a sequence of conditional probability density functions

$$\{Q(y|x) : x \in \mathcal{X}^n,\ y \in \mathcal{Y}^n\}_{n=1}^{\infty} \tag{1.27}$$

where $Q(\cdot|\cdot)$ denotes the conditional pdf governing channel used through n units of time, i.e., "n uses of the channel." If the known channel is a discrete memoryless channel, then the law is characterized in terms of a stochastic matrix $Q : \mathcal{X} \mapsto \mathcal{Y}$ according to

$$Q(y|x) = \prod_{t=1}^{n} Q(y_t|x_t), \tag{1.28}$$

where $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \ldots, y_n) \in \mathcal{Y}^n$. This kind of uncertainty models assume that the true channel $Q_{true}(y|x)$ is unknown, though, it is assumed that $Q_{true}(y|x)$ belongs to a family of channels (1.27), and remains unchanged during the course of a transmission. In [16] a nonparametric uncertainty model is formulated through a joint probability density function. A nominal statistical model is given together with a neighborhood formed by the perturbed models whose KL divergence with respect to the nominal model is bounded by a fixed constant. A static estimation problem is formulated, where the estimate of a random vector $X \in \Re^n$ needs to be found given an observation model $Y \in \Re^p$. It is assumed that the joint nominal density of X and Y is Gaussian, so that

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \sim N(m_Z, K_Z)$$

where m_Z and K_Z denote, respectively, the mean vector and covariance matrix of Z. Accordingly the nominal probability density of Z takes the form

$$f_Z(z) = \frac{1}{(2\pi)^{(n+p)/2}|K_Z|^{1/2}} \exp\Big(\frac{1}{2}(z - m_Z)^T K_Z^{-1}(z - m_z)\Big). \tag{1.29}$$

The mean vector and covariance matrices of Z can be partitioned in terms of the mean vectors and covariances of X and Y as

$$m_Z = \begin{bmatrix} m_X \\ m_Y \end{bmatrix}, \quad K_Z = \begin{bmatrix} K_X & K_{XY} \\ K_{YX} & K_Y \end{bmatrix}.$$

$\tilde{f}_Z(z)$ denotes the true probability density of Z. The KL divergence or relative entropy of $\tilde{f}_Z$ with respect to f_Z is given by

$$H(\tilde{f}_Z|f_Z) = \int_{\Re^n \times \Re^p} \ln\left(\frac{\tilde{f}_Z}{f_Z}\right) \tilde{f}_Z dz. \tag{1.30}$$

The KL divergence is not symmetric and does not obey the triangle inequality, but it satisfies $H(\tilde{f}_Z|f_Z) \geq 0$ with equality if and only if $\tilde{f}_Z = f_Z$.

1.1.3 Robust Minimax Estimation

One of the major techniques for designing systems that are robust with respect to modeling uncertainties is the minimax approach, in which the goal is the optimization of worst-case performance. Early applications of game theoretic concepts to communications problems can be found, for example, in the classical estimation studies of Yovits and Jackson [17], and Carlton and Follin [18]. However, the statistical works of Huber in estimation [19] and hypothesis testing [20] are generally regarded as the starting point of the area of minimax robustness, which has been applied successfully to a long sequel of problems in detection and estimation.

There are useful formulations of robustness other than the minimax one, most notably the stability or qualitative robustness ideas introduced by Root [21] in the context of signal detection and by Hampel [22] in the context of parameter estimation. These formulations utilize the idea of robustness as a continuity property of some performance measure as a function of the underlying model. However, from the viewpoint of design, the minimax approach has had the most impact on robust signal processing schemes.

Adaptive procedures, may also be used as robust schemes, when input conditions are not precisely known and may be time-varying. Adaptive procedures, which attempt to learn about input conditions and adjust their specific signal processing structure accordingly to maintain good performance, are generally more complex than fixed minimax schemes. Adaptive schemes are more desirable when the $\acute{a}$ priori uncertainty is so large that the guaranteed level of performance of a minimax scheme would be too poor to be acceptable, and when adequate time and data for adapting are available. Conversely, minimax procedures would be more desirable under more constrained uncertainty classes, and especially as robust procedures

to guard against excessive performance degradation of nominally optimum schemes for deviations from nominal assumptions.

As was remarked, minimax robustness of estimation and hypothesis testing schemes are considered by Huber in [19] and [20], and since then a large number of results on minimax and alternative formulations of robustness have been generated in the statistics literature. In [23] the linear regression problem of estimating an unknown, deterministic parameter vector based on measurements corrupted by colored Gaussian noise is being investigated. Blind minimax estimators are presented and analysed, which consist of a bounded parameter set minimax estimator, whose parameter set is itself estimated from measurements.

The measurements

$$Y = HX + W \tag{1.31}$$

are linear combinations of the parameter vector X, to which Gaussian noise W is added. The transformation matrix H and the noise covariance are assumed to be known. The paper seeks an estimate $\hat{X}$ which approximates X in the sense of minimal mean-square error. The parameter to be estimated X is assumed to lie within a compact parameter set S. In this case, a linear minimax estimator over the set S maybe constructed [24], [25]. This is the linear estimator $\hat{X}_M = GY$ minimizing the worst case MSE among all possible values of X in S

$$\hat{X}_M = \arg \min_{\hat{X}=GY} \max_{X \in S} E\{\|\hat{X} - X\|^2\}. \tag{1.32}$$

The authors argue that a closed-form solution of (1.32) has been previously derived for many cases of interest and that it has been shown that any linear minimax estimator achieves lower MSE than that of the LS method, for all values of X in S [24], [26].

The Blind Minimax Estimators (BMEs) presented in this paper utilize minimax estimators when no parameter set is known. This is done in a two-stage process: 1) A parameter set S is estimated from the measurements. 2) A minimax estimator designed for S is used to estimate the parameter vector X.

The paper investigates two distinct cases. In the first one the authors investigate a spherical blind minimax estimator based on a parameter set of the form $S = \{X :$

$\|X\|^2 \leq L^2\}$. For a given value of L the linear minimax estimator was derived in [25]

$$\hat{X}_M = \frac{L^2}{L^2 + \epsilon_0}\hat{X}_{LS} \tag{1.33}$$

where $\hat{X}_{LS}$ is the classical LS estimator and ϵ_0 is the MSE. Now when the bound L^2 is estimated as $L^2 = \|\hat{X}_{LS}\|^2$, the Spherical BME (SBME) is given by

$$\hat{x}_{SBME} = \frac{\|\hat{X}_{LS}\|^2}{\|\hat{X}_{LS}\|^2 + \epsilon_0}\hat{X}_{LS}. \tag{1.34}$$

The authors also argue that the BMEs may be constructed around any constant center point X_0. Furthermore, the paper investigate a second case, the ellipsoidal blind minimax estimator based on a parameter set of the form $S = \{X : \|X\|^2_{Q^b} \leq L^2\}$, for some constant $b < 0$. Here Q^{-1} is the covariance of $\hat{X}_{LS}$ and the bound L^2 is estimated as $L^2 = \|\hat{x}_{LS}\|^2_{Q^b}$. The above cases are being examined in the setting of a linear system of measurements with colored Gaussian noise, and it is shown that the proposed BMEs dominate the LS method, i.e., they achieve lower mean-squared error for any value of the parameter vector. Finally, in [23] the relations of blind minimax techniques to Stein type estimators [27] and least-square regularization is also discussed.

The James-Stein estimator is a nonlinear estimator which can be shown to dominate, or outperform, the “ordinary" (least squares) technique. As such, it is the best-known example of Stein’s phenomenon. In 1961, James and Stein discovered a remarkable estimator that dominates the maximum-likelihood estimate of the mean of a p-variate normal distribution, provided the dimension p is greater than two. Various “extended” James-Stein methods were later constructed for the general non-i.i.d. case. However, none of these approaches has become a standard alternative to the LS estimator, and they are rarely used in practice in engineering applications [27].

The problem of estimating an unknown parameter vector X in a linear model that may be subject to uncertainties, where the vector X is known to satisfy a weighted norm constraint is also investigated in [25]. This paper addresses two different estimation problems. In the fist one, which is similar with the general problem in [23], it is assumed that the model matrix H is known exactly and the linear estimator that minimizes the worst-case mean-squared error across all possible values of X is derived. The difference with [23] is that [25] assumes that the parameter set S is

known, i.e., it is assumed that X is known to satisfy the weighted norm constraint $\|X\|_T \leq L$ for some positive definite matrix T and scalar $L > 0$. The robust estimator for this problem is given by (1.33). The second problem, considers the case in which the model matrix H is subject to uncertainties and seeks the robust linear estimator that minimizes the worst-case MSE across all possible values of X and all possible values of the model matrix. In many engineering applications, the model matrix H is subject to uncertainties, for example, it may have been estimated from noisy data, in which case, H is an approximation to some nominal underlying matrix. If the true data matrix is $H + \delta H$ for some unknown perturbation matrix δH, then the actual performance of an estimator designed based on H alone may perform poorly. In this case, robust estimators are considered that explicitly take uncertainties into account. Specifically, in [25] the matrix model H is not known exactly but is rather given by $H + \delta H$, where $\|\delta H\| \leq p$ and the minimax problem is formulated as

$$\min_{\hat{X}=Gy} \max_{\|X\|_T \leq L, \|\delta H\| \leq p} E\{\|\hat{X} - X\|^2\}. \tag{1.35}$$

A similar minimax approach, and model is being used in [28]. Here two problems are being considered. In the first one the uncertainty lies in the covariance of X, $C_X = \tilde{C}_X + \delta C_X$, where $\tilde{C}_X$ is known and $\|\delta C_X\| \leq \epsilon$. Thus, the minimax MSE problem is formulated as

$$\min_{\hat{X}=GY} \max_{\|\delta C_X\| \leq \epsilon} E(\|\hat{X} - X\|^2). \tag{1.36}$$

In the second problem the covariance matrix C_X and the model matrix H are subject to uncertainty. In this case the model matrix H is given by $H = \tilde{H} + \delta H, \ \ \|\delta H\| \leq p$, where $\tilde{H}$ is known. The minimax problem is formulated as

$$\min_{\hat{X}=GY} \max_{\|\delta C_X\| \leq \epsilon, \|\delta H\| \leq p} E(\|\hat{X} - X\|^2). \tag{1.37}$$

In order to improve the performance over the minimax MSE approach, a competitive approach is also considered in [28]. This approach assumes that H is completely known, and seeks a linear estimator $\hat{X}$ that minimizes the worst case regret in order to partially compensate for the conservative character of the minimax approach. The regret $\mathcal{R}(C_X, G)$ is defined as the difference between the MSE using an estimator $\hat{X} = GY$ and the smallest possible MSE attainable with an estimator of the form $\hat{X} = G(C_X)Y$ when the covariance C_X in known, which is denoted as MSE^0. The

minimax problem is formulated as

$$\min_{G} \max_{\|\delta C_X\| \leq \epsilon} \mathcal{R}(C_X, G) \tag{1.38}$$

where $\mathcal{R}(C_X, G) = E(\|\hat{X} - X\|^2) - MSE^0$. The linear minimax regret estimator is shown to be equal to a Minimum-Mean-Squared-Error (MMSE) estimator corresponding to a certain choice of signal covariance, that depends explicitly on the uncertainty region.

A different approach is introduced in [16], where given a nominal statistical model, the minimax estimation problem consisting of finding the best least-squares estimator for the least favorable statistical model within a neighborhood of the nominal model is being considered. The uncertainty model for this problem has already been described in Section 1.1.2. The neighborhood is formed by placing a bound on the KL divergence between the actual and nominal models described by the uncertainty set $\mathcal{B} = \{\tilde{f}_Z \in \mathcal{F} : H(\tilde{f}_Z | f_Z) \leq c\}$.

The authors seek to estimate a random vector $X \in \Re^n$ given an observation vector $Y \in \Re^p$ such that the joint nominal density of $Z = [X^T, Y^T]^T$ is Gaussian with the parameterization (1.29). They need to find an estimator $\hat{X} = g(Y)$ and a least favorable density $\tilde{f}_Z$ which solve the minimax problem

$$\min_{g \in \mathcal{G}} \max_{\tilde{f}_Z \in \mathcal{B}} J(\tilde{f}_Z, g) \tag{1.39}$$

where

$$J(\tilde{f}_Z, g) = \frac{1}{2} E_{\tilde{f}_Z}[\|X - g(Y)\|^2] = \frac{1}{2} \int \|x - g(y)\|^2 \tilde{f}_Z(z) dz \tag{1.40}$$

For the above minimax problem it is shown that $J(\tilde{f}_Z, g)$ admits a saddle point and that the estimator takes the form

$$g_0(Y) = m_X + G_0(Y - m_Y) \tag{1.41}$$

with $G_0 = K_{XY} K_Y^{-1}$. This estimator is linear and yields the usual least-squares estimate of X given Y for both the nominal density f_Z and the perturbed density $\tilde{f}_0$. In [16] is shown that for a Gaussian nominal model and a finite observations interval, or for a stationary Gaussian process over an infinite interval, the usual non-causal Wiener filter remains optimal. However, it also shows that in the causal case, the usual least-squares estimator or the causal Wiener filter are no longer optimal, and

a characterization is given for the structure of an optimal robust estimator and the matching of least favorable statistical model. The optimal causal filter derived is a risk-sensitive filter, where the risk-sensitivity parameter was selected to match the maximum allowable relative entropy for the perturbed model.

The same minimax approach used in [16] is also applied in [11], the difference is in the uncertainty model used. As described on Section 1.1.2, [11] uses a parametric uncertainty model, where the uncertainty is placed in the channel H and noise R information. Similar with [16], the KL divergence is used to measure the "distance" between the actual and the estimated model, and the uncertainty set is defined as $\mathcal{B} = \{(H_\Delta, R) : D(H, R; \hat{H}, \hat{R}) \leq c\}$. Then the MMSE equilizer problem is formulated by finding as

$$\min_{F} \max_{(H_\Delta, R) \in \mathcal{B}} J(F, H_\Delta, R) \tag{1.42}$$

where F is an MMSE equalizer matrix, and (F, H_Δ, R) is the MSE pay-off function. By using Lagrangian duality, the minimax problem is transformed into an equivalent min-min problem over a convex domain, where the standard convex optimization methods apply. Then, it is shown that the robust MSE equalizer can be obtained by solving numerically a scalar convex minimization problem.

Although robustness issues do not appear explicitly in [29], and the minimax approach is not really implemented per say, it presents some interesting results for filtering problems when there is uncertainty in the exact value of the probability model. In the problems described above the minimax approach first maximizes the MSE in order to derive the worst case measure given a specific uncertainty set. In [29] given a measurable space $(\Omega, \mathcal{F})$ and random variables X, Y, a function of X, $\phi = \phi(X)$ needs to be estimated by random variable $\hat{\phi} \in \mathcal{Y}$, where Y represents the observations or measurements. The true distribution $P_{\alpha_0}$is assumed to belong to a family of probability measures $\{P_\alpha\}_{\alpha \in A}$. The minimum cost estimator is defined as the estimator which minimizes the error cost function $E_{\alpha_d}[p_\alpha(\phi - \hat{\phi}]$, where p_α is a strictly convex function and P_{α_d} is assumed as the true model. Instead of maximizing the above error cost function, it chooses an exponential cost function. As stated in [29], it is known from control theory that if a controller is designed to minimize an average-of-exponential criterion then this controller implies certain robustness properties. Therefore, a minimum risk-sensitive estimator, denoted $\hat{\Phi}^*_{rs}$, is defined as the minimum cost estimator obtained by selecting $p_\alpha(e) = \exp(\mu p(e))$, where $\mu > 0$

is a parameter determining the degree of "risk",

$$\hat{\Phi}^*_{rs} = \arg\min_{\hat{\phi}\in R} \int \exp(\mu p(\phi - r))\pi_{\alpha_d}(dx) \tag{1.43}$$

where π_{α_d} is the conditional distribution of X given Y under P_{α_d}.

Furthermore, it is shown that the risk-sensitive estimators enjoy an error bound which is the sum of two terms, the first of which coincides with an upper bound on the error one would obtain if one knew exactly the underlying probability model, while the second term is a measure of the distance between the true and design probability models. The first term quantifies "good performance" under nominal conditions, and the second term quantifies the "acceptable performance" under non-nominal conditions. Also, the second term plays a major role in determining the class of permissible variations from nominal.

Next, [30] deals with the problem of designing robust linear causal estimators of linear functions of discrete-time wide-sense stationary signals, when the knowledge of the signal and/or noise spectral is inexact. The spectral uncertainty is modeled by choosing appropriate classes of PSD's, $\mathcal{G}$ and $\mathcal{N}$, and assuming that the signal PSD $f_S \in \mathcal{G}$ and the noise PSD $f_N \in \mathcal{N}$. The paper seeks to find the transfer function H^*_R with the smallest possible upper bound on the MSE, $e_D(f_S, f_N; H_R)$ over all f_S in $\mathcal{G}$ and f_N in $\mathcal{N}$, that is the solution to the following minimax formulation

$$\inf_{H\in H^2_+} \sup_{(f_S,f_N)\in\mathcal{G}\times\mathcal{N}} e_D(f_S, f_N; H) \tag{1.44}$$

The solution is given under mild regularity conditions in the terms of the least favorable pair of spectra, (f^L_S, f^L_N), thus reducing the minimax problem to a direct maximization problem which in many cases can be solved easily. The solution is based on the fact that a pair of PSD's $(f^L_S, f^L_N) \in \mathcal{G}\times\mathcal{N}$ and its optimal causal transfer function H^*_L, under specific conditions, form a saddle-point solution to (1.44) if and only if (f^L_S, f^L_N) is least favorable for causal estimation. Furthermore, solutions are given explicitly for the problem of robust causal filtering of an uncertain signal in white noise.

1.1.4 Relation of Mutual Information and MMSE

As was mentioned in Section 1.1.1 one of the most popular Bayesian estimation techniques is the Least-Square estimation, also known as Minimum-Mean-Squared-

Error (MMSE) estimation. Recently, extensive work has been done which connects the mutual information between the input and the output of a channel, and the MMSE in estimating the input given the output. In a wider context, the mutual information and mean-square error are at the core of information theory and estimation theory, respectively. The input-output mutual information is an indicator of how much coded information can be pumped through a channel reliably given a certain input signaling, whereas the MMSE measures how accurately each individual input sample can be recovered using the channel output.

The relation between the mutual estimation and MMSE is presented in [31], which deals with arbitrarily distributed finite power input signals observed through an additive Gaussian noise channel. A new formula is presented that connects the input-output mutual information and the minimum mean-square error achievable by optimal estimation of the input given the output. Given that the input-output mutual information and the MMSE are monotone functions of the signal-to-noise ratio (SNR), denoted by $I(snr)$ and $mmse(snr)$, respectively, the mutual information in nats and the MMSE satisfy the following relationship regardless of the input statistics:

$$\frac{d}{dsnr}I(snr) = \frac{1}{2}mmse(snr) \tag{1.45}$$

That is, the derivative of the mutual information (nats) with respect to the signal-to-noise ratio (SNR) is equal to half the MMSE, regardless of the input statistics. This relationship holds for both scalar and vector signals.

For the scalar signal, [31] considers a pair or real-valued random variables related by

$$Y = \sqrt{snr}X + N \tag{1.46}$$

where $snr \geq 0$ and $N \sim N(0,1)$ is a standard Gaussian random variable independent of X. Then X and Y can be regarded as the input and output, respectively, of a single use of a scalar Gaussian channel with an SNR of snr.

On the other hand, the multiple-input multiple-output (MIMO) system is described in [31] by the vector Gaussian channel

$$Y = \sqrt{snr}HX + N \tag{1.47}$$

where H is a deterministic $L \times K$ matrix and the noise N consists of independent standard Gaussian entries. The input X (with distribution P_X) and the output Y are column vectors of appropriate dimensions.

Various applications of the above relationship are identified in [31], e.g. in relating code-division multiple-access (CDMA) channel spectral efficiencies (mutual information per dimension) under joint and separate decoding in the large-system limit. Also the fact that the mutual information and the MMSE determine each other by a simple formula provides new means to calculate or bound one quantity using the other. An upper bound for the mutual information is immediate by bounding the MMSE for all SNRs using a suboptimal estimator. Lower bounds on the MMSE lead to new lower bounds on the mutual information.

The low-SNR asymptotics of $mmse(X; snr)$ is also studied extensively in [31], mainly for the scalar Gaussian channel. The Taylor expansion of $mmse(X; snr)$ at $snr = 0$ is obtained and the coefficients turn out to depend only on the moments of X. Based on the above work [32] deals with the high-SNR asymptotics of MMSE in Gaussian channels and defines a new information measure called MMSE dimension. The MMSE dimension of X for the scalar channel (1.46) is defined as the limit as $snr \to \infty$ of the product of snr and the MMSE. For discrete, absolutely continuous or mixed X [32] shows that the MMSE dimension equals Rényi's information dimension. For singular X, it shows that the product of snr and MMSE oscillates around information dimension periodically in snr (dB).

Finally, based on the relationship (1.45), [33] presents some new results on mutual information and various regularity properties of the MMSE functional are explored together with its connections to Shannon theory.

1.2 Thesis Motivation

Robust estimation is not a new subject of study, as was already seen, and it is a major problem in statistics and signal processing. The minimax approach, in which the goal is to optimize the worst case performance, is one of the major techniques for designing robust systems with respect to modeling uncertainties and has been applied to many problems in detection and estimation, as has already presented in

Section 1.1.3. This thesis will focus, mainly, on least-square estimation problems when the statistics or internal dynamics describing the signal and observations are not known exactly. A robust estimation approach is being proposed in this thesis which employs a minimax formulation.

A measurable space $(\Omega, \mathcal{F})$ is given on which the unobserved Random Variable, X and the observed RV Y are defined, via $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \Sigma_{\mathcal{X}})$, $Y : (\Omega, \mathcal{F}) \to (\mathcal{Y}, \Sigma_{\mathcal{Y}})$. The objective is to estimate X by a function of the random variable Y. As mentioned in Section 1.1.1, the classical least square estimation problem deals with minimization of the average pay-off which can be expressed in the following ways

$$J(\Phi^*) = \inf_{\Phi \in \mathcal{X}_{ad}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) dP_{X,Y}(x, y) \tag{1.48}$$

$$= \inf_{\Phi \in \mathcal{X}_{ad}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) \mu(x, dy) dP_X(x) \tag{1.49}$$

$$= \inf_{\Phi \in \mathcal{X}_{ad}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) \eta(y, dx) dP_Y(y). \tag{1.50}$$

Clearly if $\ell(x, \Phi(y)) = ||x - \Phi(y)||_{\Re^n}$, then the estimate of X denoted as $\hat{X} = \Phi(Y)$ is given by $\Phi^*(Y) = E[X|Y]$. As was already mentioned, and shown from the literature the signal statistics $P_{X,Y}$ or μ, η are not always known. In this thesis the uncertainty description of the system, and the nominal description of the system are modeled by probability distributions, or general measures, defined on measurable spaces. Moreover, two type of uncertainty models are being considered.

1. Uncertainty Models on Conditional Distributions or otherwise known Stochastic Kernels;
   i) When the conditional probability distribution of the measurement Y given the signal to be estimated X, or channel kernel, is unknown;
   ii) When the *á* posteriori distribution of X given Y is unknown;
2. Uncertainty Models on Joint Distributions.

The goal is to derive new robust estimators for the above uncertainty models (this is described explicitly in Chapter 3). Notice that (1.48) will be used when the uncertainty is on the joint distribution, while (1.49), (1.50) will be used when the uncertainty is on the channel kernel, *á* posteriori distribution, respectively.

Stochastic kernel uncertainty models are appropriate for communication system design, in which the input message has a known distribution, while the channel is unknown but belongs to a certain class of channels. These are nonparametric uncertainty models which so far have not been taken into consideration. Joint distribution uncertainty models are usually employed when both the unobserved and observed random variables are uncertain. Joint uncertainty models are used in [16] and [29], but [29] does not implement a minimax approach and it shows mainly that the mean-square error for the true model is bounded by the sum of two terms, with the first term representing the performance of the risk-sensitive filter with respect to the nominal model, and the second term corresponding to the relative entropy between the actual and nominal models. On the other hand [16] employs a minimax viewpoint but uses the additional assumption that the system considered is Gaussian and it uses a parametric approach to derive its results.

The uncertainty description of the above systems is characterized by the class of uncertain measures which satisfy a relative entropy constraint with respect to a nominal measure. The use of the relative entropy, also known as the KL divergence, is rather natural as a metric for model mismatch since it is commonly used by statisticians for fitting statistical models, and by using a differential geometric viewpoint it is argued that the KL divergence is the natural geometric "distance" between systems. The KL distance constraint has been also used in robust estimation problems in [11], [16] and [29].

By using the KL divergence various constraint sets are defined for each uncertainty model, and for each the minimax estimation problem is solved and the worst case measure of the true model and also the robust estimator are derived. This is done in a generalized framework, while the estimators derived can be used accordingly to various estimation applications.

One particular application which is investigated is the robust nonlinear estimation for finite-dimensional autoregressive channel models found in [34], [35] subject to an uncertainty set. Autoregressive channel models have been used with success to predict fading channel dynamics for the purposes of Kalman filter based channel estimation and for long-range channel forecasting. They have also been used by several authors to simulate correlated Rayleigh fading. A specific example is the non-coherent estimation problem, where there is an attenuated sinusoid in a multipath environment which is subject to an additive Gaussian noise. Here an uncertainty

model on the joint distribution is used and as before, the uncertainty set is being described by a KL constraint.

The above problem, which is a nonlinear problem, will be addressed through a minimax approach. The minimax approach has not been used so far for this specific problem and the difficulty lies in the minimization, and the way of obtaining the solution of the $\acute{a}$ posteriori distribution recursion in closed form. In general when dealing with nonlinear estimation problems, approximations have to be made to find sub-optimal nonlinear estimators. The standard method is to use the Taylor series expansion and apply linear filtering theory, giving rise to the so-called extended Kalman filter (EKF) [5, 6, 7]. Other more sophisticated sub-optimal estimation techniques are available, e.g. reiteration, higher order filters, and statistical linearization [5]. The scope of the thesis is not to investigate general nonlinear estimation techniques but to find a way to tackle the minimization in the minimax formulation. Here the maximization is addressed using variational methods, while the minimization is addressed using a change of probability measure technique [36]. The change of probability measure techniques introduced is being used in order to derive recursive equations for the conditional distribution of nonlinear filtering problems, which helps us to compute the worst-case pay-off functional. Minimizing this worst-case pay-off function gives the desired robust minimax estimators. Special emphasis is given also to the connection between the robust minimax estimation problem and the classical non-coherent estimation.

Even though the primal focus of this thesis is robust least square estimation some other well-known estimation techniques, like the MAP and the ML estimation techniques, are also investigated. As was described in Section 1.1.1, in the classical MAP problem, by applying a uniform cost criterion, the estimator of x is derived through the maximization of the $\acute{a}$ posteriori (discrete or continuous) density $\eta(y, dx)$. In this thesis instead of a uniform cost criterion, an exponential one is being used. As its been mentioned in [29], it is known in control theory that if a controller is designed to minimize an average-of-exponential (or risk-sensitive) criterion, then this controller might have certain robustness properties. Hence, it is natural to consider the use of such exponential criteria in filter design. The objective here is to derive a generalized MAP estimator using an exponential cost criterion and also show a connection with the minimax approach used in the least-square estimation problem. A similar approach is also used for the derivation of generalized ML estimator. The ML estimation technique, which is not a Bayesian approach, is an alternative method for

deriving estimators, when the parameter x is assumed to be unknown but without being characterized by a probabilistic structure. The classical method of maximum likelihood estimates x by finding the value of x that maximizes the likelihood function $\mathcal{L}(x) = \mu(x, dy)$. The goal is to derive a generalized ML estimator when this likelihood function includes also an exponential function and show the connection with the generalized MAP estimator.

1.3 Thesis Objective

The main objective of this thesis is to derive robust estimators for least-square estimation problems in the presence of model uncertainties. Additional to this objective is the investigation of a generalized MAP and ML estimation technique. These three estimation techniques are well known and often used in signal estimation and in signal processing.

Firstly, least-square estimation problems are investigated when the real description of the system is unknown and the only knowledge is that it belongs to an uncertainty set, or a class of systems. The real description of the system, the uncertainty description of the system, and the nominal description of the system are modeled by probability distributions, or general measures, defined on measurable spaces. The uncertainty is described by a KL constraint between the unknown distribution and a fixed nominal distribution.

Secondly, the behavior of the Maximum $\acute{A}$ Posteriori estimation and the Maximum Likelihood estimation technique is investigated, when introducing new elements in the derivation of the classical estimators. Generalized estimators for both techniques are derived by altering the cost function for the MAP case and the likelihood function for the ML case.

Chapter 2 presents background material and explains the main techniques used throughout this thesis, which are change of probability theory and the minimax theory.

Chapter 3 considers least-square estimation problems for classes of models and introduces the concept of uncertainty. The uncertainty is described by a KL distance

constraint between the unknown distribution and a fixed nominal distribution. The theory and contribution of this chapter are developed at two levels of generality; the abstract level and the application level. At the abstract level the uncertainly models used are Stochastic Kernels and Joint Distributions.

Starting with the first uncertainty model, the relation between the unobserved RV X and the observed RV Y is defined via a probabilistic mapping, $\mu : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$, which satisfies the following two conditions:

1. For every $x \in \mathcal{X}$, the set function $\mu(x, \cdot)$ is a probability measure on $\Sigma_{\mathcal{Y}}$ (possibly finite additive);
2. For every $F \in \Sigma_{\mathcal{Y}}$, the function $\mu(\cdot, F)$ is $\mathcal{X}$-measurable.

The mapping μ is called a stochastic kernel or transition probability and represents the nominal system model or mapping, which is fixed. The true kernel denoted by $\nu : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$ is assumed unknown. Envisioned scenarios are communication channels whose nominal behavior is known, while its true conditional distribution is unknown. The KL distance is used as a measure of distance between the true model and uncertainty model, hence the true kernel is assumed to belong to the pointwise uncertainty set,

$$\mathcal{A}^x(\mu) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : H(\nu|\mu)(x) \leq R(x)\Big\} \tag{1.51}$$

where $R : \mathcal{X} \to [0, \infty)$ and $H(\cdot|\cdot)(x) : \mathcal{X} \to [0, \infty]$ is the KL distance between two kernels. Additionally, the following uncertainty set is defined

$$\mathcal{A}(\mu) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : \int_{\mathcal{X}} H(\nu|\mu)(x) dP_X(x) \leq \int_{\mathcal{X}} R(x) dP_X(x) \stackrel{\nabla}{=} r_1\Big\}. \tag{1.52}$$

For the second uncertainty model, uncertainty on the *á* posteriori distribution, the mapping $\eta : \mathcal{Y} \times \Sigma_{\mathcal{X}} \to [0, 1]$ is considered, that satisfies the following two conditions:

1. For every $y \in \mathcal{Y}$, the set function $\eta(y, \cdot)$ is a probability measure on $\Sigma_{\mathcal{X}}$;
2. For every $F \in \Sigma_{\mathcal{X}}$, the function $\eta(\cdot, F)$ is $\mathcal{Y}$-measurable.

This probabilistic kernel $\eta(y, dx)$ represents the nominal system model or mapping (*á* posteriori information) and the true kernel $\nu(y, dx)$, denoted by $\nu : \mathcal{Y} \times \Sigma_{\mathcal{X}} \to [0, 1]$

belongs to an uncertainty set. The following two uncertainty sets are considered.

$$\mathcal{B}^y(\eta) \triangleq \left\{\nu \in \mathcal{P} : H(\nu|\eta)(y) \leq R(y)\right\} \tag{1.53}$$

$$\mathcal{B}(\eta) \triangleq \left\{\nu \in \mathcal{P} : \int_{\mathcal{Y}} H(\nu|\eta)(y) dP_Y(y) \leq \int_{\mathcal{Y}} R(y) dP_Y(y) \stackrel{\nabla}{=} r_2\right\} \tag{1.54}$$

where $R : \mathcal{Y} \to [0, \infty)$ and $H(\cdot|\cdot)(y) : \mathcal{Y} \to [0, \infty]$ is the KL distance between two kernels. Additionally, the next case is investigated, where the true kernel $\nu(y, dx)$ belongs to a new uncertainty set described by

$$\mathcal{B}^R(\eta) \triangleq \left\{\nu \in \mathcal{P} : \int_{\mathcal{Y}} H(\nu|\eta)(y) dP_Y(y) \leq \int_{\mathcal{X}\times\mathcal{Y}} R(x, y)\nu(y, dx) dP_Y(y) + \overline{R} \stackrel{\nabla}{=} r_3\right\} \tag{1.55}$$

where $R : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$, $R \in BC(\mathcal{X} \times \mathcal{Y})$ and $H(\cdot|\cdot)$ is the KL distance between two kernels.

Finally, uncertainty is modeled via the joint distribution of X and Y. This model is appropriate when one wishes to model *á* priori uncertainty. It is assumed that the joint distribution $P_{X,Y}(x, y)$ represents the nominal system model and the true joint distribution denoted by $Q \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, is assumed to belong to an uncertainty set described by

$$\mathcal{C}(P_{X,Y}) \triangleq \{Q_{X,Y} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) : H(Q_{X,Y}|P_{X,Y}) \leq R\}$$

where $R \in [0, \infty)$ and $H(\cdot|\cdot)$ is the KL distance between the two joint distributions.

For all the above uncertainty sets the goal is to formulate the minimax problem and derive first the worst-case measure for the true stochastic kernel ν^* or true joint distribution Q^*, and then to derive the robust estimator for each case. The minimax estimation problem can be formulated as

$$\begin{aligned} J_1(\Phi, \nu) &\triangleq \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y))\nu(x, dy) dP_X(x), \\ J_1(\Phi^*, \nu^*) &= \inf_{\Phi\in\mathcal{X}_{ad}} \sup_{\nu\in\mathcal{A}(\mu)} J_1(\Phi, \nu). \end{aligned} \tag{1.56}$$

This represents one of the uncertainty models. Similar formulation will be used for all the other cases.

First, the appropriate space of measures is introduced and then the maximizing kernel and joint measure are computed explicitly using Lagrangian functionals, and

variational methods. Moreover, important monotonicity properties satisfied by the optimal strategies are presented, which can be used to develop numerical algorithms for computation of the optimal solution, and upper and lower bounds on the optimal solution. Furthermore, in Chapter 3, several problems of estimation theory are formulated, and solutions are sought when the models (conditional distributions, joint distributions) are uncertain, and they belong to specific subsets of the set of conditional or joint distributions.These examples include MIMO communication systems.

Chapter 4 applies the theory developed in Chapter 3 to finite-dimensional autoregressive channel models, and derives robust estimators for a class of uncertain models. Similar to Chapter 3, the uncertainty is described by a relative KL distance between the unknown joint distribution and a fixed nominal joint distribution. The abstract channel model is given by

$$\begin{aligned} x_{k+1} &= f(k+1, x_k) + B_{k+1} w_{k+1}, \quad x_0 \in \Re^n \\ y_k &= h(k, x_k) + D_k v_k, \quad y_0 \in \Re^d. \end{aligned} \tag{1.57}$$

Here $x_0 : \Omega \to \Re^n$ is the initial state and $w : \Omega \times N_0 \to \Re^n, \quad v : \Omega \times N_0 \to \Re^d$, are random noises, all mutually independent.

The uncertainty model being used here is the joint distribution, where P_{x^m,y^m} denotes the nominal (in the absence of modeling uncertainties) joint distribution of the sequences (x^m, y^m), which corresponds to the one induced by model (1.57) and Q_{x^m,y^m} denotes the true joint distribution of the sequences (x^m, y^m), which is unknown. The only available information is that Q_{x^m,y^m} belongs to a class of possible distributions. This class is modelled by the information theoretic KL distance set

$$\mathcal{C}(P_{x^m,y^m}) \stackrel{\triangle}{=} \{Q_{x^m,y^m} : H(Q_{x^m,y^m} | P_{x^m,y^m}) \leq R\}.$$

The minimax estimation problem is defined as

$$J(\tilde{x}^*, Q^*_{x^m,y^m}) = \inf_{\tilde{x}_m \in \mathcal{X}_{ad}} \sup_{Q_{x^m,y^m} \in \mathcal{C}(P_{x^m,y^m})} E_{Q_{x^m,y^m}} \Big\{ \sum_{k=0}^{m} \tilde{\ell}(x_k, \tilde{x}_k) \Big\} \tag{1.58}$$

where $\tilde{\ell}(x_k, \tilde{x}_k)$ is a measure of distance between the state x_k and its estimate $\tilde{x}_k$.

In the abstract setting, the maximization is addressed using variational methods, while the minimization is addressed using a change of probability measure technique

[36]. Through this technique conditional expectations are related via

$$E\Big[\Phi(x_m)\exp\Big(\frac{1}{s}\sum_{k=0}^{m}\tilde{\ell}(x_k,\tilde{x}_k)\Big)\Big|\mathcal{Y}_m\Big]=\frac{\overline{E}\Big[\Phi(x_m)\Lambda_m\exp\Big(\frac{1}{s}\sum_{k=0}^{m}\tilde{\ell}(x_k,\tilde{x}_k)\Big)|\mathcal{Y}_m\Big]}{\overline{E}\Big[\Lambda_m|\mathcal{Y}_m\Big]} \tag{1.59}$$

where Λ is the likelihood function of the complete data and E, $\overline{E}$ denotes expectation under the probability distribution P, $\overline{P}$, respectively. The change of probability measure techniques introduced is being used in order to derive recursive equations for the conditional density of nonlinear filtering problems.

Next, the theory developed is applied to two specific applications. Firstly to a linear Gaussian model and secondly to an attenuated sinusoid in a multipath environment which is subject to an additive Gaussian noise given by

$$y(t_k)=\sum_{i=1}^{N}A_i(t_k)r_i\cos(\omega_c(t_k-\tau_i(t_k))+\theta_i)S(t_k-\tau_i(t_k))+D(t_k)v(t_k). \tag{1.60}$$

Various estimators are being derived for both applications. Furthermore, for the attenuated sinusoid model a connection to least-squares estimation found in [1], [4] for single channel is established by reducing the uncertainty to zero, while generalizing existing results.

Chapter 5 considers the classical Maximum $\acute{A}$ Posteriori estimation and Maximum Likelihood estimation problems. The classical MAP estimator is derived by minimizing a specific uniform cost function. The theory behind maximum $\acute{a}$ posteriori estimation is being investigated when this uniform cost function is modified. An exponential cost function is introduced defined by

$$C(X,\Phi(y))=\begin{cases}e^{\frac{\ell(x)}{s}}, & \text{if } |X-\Phi(y)|>\Delta\\ \\ 0, & \text{if } |X-\Phi(y)|\leq\Delta\end{cases}$$

where $\ell:\mathcal{X}\times\mathcal{X}\to[0,\infty)$ is an $\Sigma_{\mathcal{X}}\times\Sigma_{\mathcal{X}}$-measurable function and $\Delta>0$. By minimizing the average cost function a generalized MAP estimator is derived. Furthermore, a connection to the theory of minimax least-square estimation derived in Chapter 3 is presented.

The classical ML estimator is derived by maximizing the maximum likelihood function. The theory behind ML estimation is investigated, when a modified likelihood

function is being used, defined as $e^{\frac{\ell(x)}{s}}\mu(x, dy)$, and a generalized ML estimator is derived. A relation with the MAP estimation is also presented. Finally, the theory developed is applied to several examples.

Chapter 6 contains the main points of the thesis and suggests directions for future research.

1.4 Contributions

The main contributions of Chapter 3 are the following.

1. New uncertainty models based on stochastic kernels and joint distributions are presented for robust minimax least-square estimation problems;
2. For each uncertainty model, various uncertainty sets are defined using the KL distance constraint;
3. The maximization problem is addressed using variational methods and for each uncertainty model the worst case measure is derived;
4. The worst case pay-off is derived, for each uncertainty model, and by minimizing this pay-off the robust estimators are derived;
5. The derived worst case measures and worst case pay-offs are employed to compute robust estimators for linear problems, which also include examples from MIMO communication systems.

The main contributions of Chapter 4 are the following:

1. A minimax least-square estimation problem is formulated for finite-dimensional autoregressive channel models, when the uncertainty inserted is on the joint distribution, and the uncertainty set is described by the KL distance;
2. The minimization is addressed using a change of probability measure technique, and a recursive equation for the conditional distribution is derived. Using this unnormalized $\acute{a}$ posteriori distribution, the worst case pay-off is derived;

3. The derived unnormalized $á$ posteriori distribution is employed to compute the estimator for a linear gaussian autoregressive channel;

4. The derived unnormalized $á$ posteriori distribution is employed to compute robust phase and envelope estimators for an attenuated sinusoid in a multipath environment, which is subject to an additive Gaussian noise;

5. A connection between the derived robust results and the classical non-coherent estimation problem is presented.

The main contributions of Chapter 5 are the following:

1. The cost function used in the MAP estimation method is defined by an exponential function;

2. A generalized MAP estimator is derived using the exponential cost function;

3. A generalized ML estimator is derived by inserting an exponential function in the maximum likelihood function;

4. The derived generalized estimators are used in several examples.

CHAPTER 2

BACKGROUND MATERIAL

In this chapter, background material is presented and the main techniques used throughout this dissertation are explained. In Section 2.1 the basic mathematical concepts of vector spaces, probability theory, are reviewed following a measure-theoretic approach and the theory behind minimax estimation techniques is described. In Section 2.2, the theory of change of probability measure technique is presented.

2.1 Basic Mathematical Background

In the study of systems, functional analysis plays a fundamental role. The concepts and mathematical tools developed in this section are used in subsequent chapters.

2.1.1 Functional Analysis

Here, the basic elements of mathematics that are needed to understand systems that can be modeled by linear or nonlinear differential equations are presented. To fully describe the state of the process of a system at any point of time a number of variables have to be quantified which are called a vector. Moreover, some important results from functional analysis are presented, which will be frequently used in the following chapters.

Definition 2.1.1. *(Vector Space) A vector space over a field F is a set of vectors $\mathcal{Z}$ together with the operations of*

- *addition in* $\mathcal{Z}$*:* $x + y,\ x, y \in \mathcal{Z}$*,*
- *multiplication by Scalar:* $\alpha \cdot x,\ \alpha \in F,\ x \in \mathcal{Z}$*,*

satisfying the following properties:

1. *Associativity:* $(x + y) + z = x + (y + z),\ \forall x, y \in \mathcal{Z}$.
2. *Commutativity:* $x + y = y + x,\ \forall x, y \in \mathcal{Z}$.
3. *Additive identity: There exists* $\emptyset \in \mathcal{Z}$ *such that* $\emptyset + x = x, \forall x \in \mathcal{Z}$.
4. *Additive inverse:* $\forall x \in \mathcal{Z}$ *there exist* $-x \in \mathcal{Z}$*, such that* $x + (-x) = \emptyset,\ \forall x \in \mathcal{Z}$.
5. *Ass. Scalar:* $\alpha(\beta \cdot x) = (\alpha \cdot \beta)x,\ \forall x \in \mathcal{Z},\ \alpha, \beta \in F$.
6. *Multiplicative identity:* $1 \cdot x = x,\ 0 \cdot x = \emptyset,\ \forall x \in \mathcal{Z}$.
7. *Scalar Mult. Distributive w.r.t. vector addition:* $\alpha(x + y) = \alpha x + \beta y,\ \forall x, y \in \mathcal{Z}$.
8. *Scalar Mult. is Distr. w.r.t. scalar addition:* $(\alpha + \beta)x = \alpha x + abetay,\ \forall \alpha, \beta \in F$.

Usually, a vector space over $F = \mathbb{R}$ *is called a real vector space and a vector space over* $F = \mathbb{C}$ *is called a complex vector space.*

Definition 2.1.2. *(Normed Space) A Normed Space is a vector space* $\mathcal{Z}$ *furnished with a norm* $\|\cdot\|_{\mathcal{Z}}$ *and denoted by* $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$*. The norm* $\|\cdot\|_{\mathcal{Z}} : \mathcal{Z} \to \Re$ *is a real valued function defined on* $\mathcal{Z}$ *which must satisfy the following properties [37]:*

i) $\|x\| \geq 0\ \forall\ x \in \mathcal{Z}$*;*

ii) $\|x\| = 0$ *if and only if* $x = 0$*;*

iii) $\|\alpha x\| = |\alpha|\|x\|,\ \forall\ x \in \mathcal{Z},\ \alpha \in F$*;*

iv) $\|x + y\| \leq \|x\| + \|y\|,\ \forall\ x, y \in \mathcal{Z}$.

Definition 2.1.3. *(Cauchy sequence) Let* $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$ *be a normed space. A sequence* $\{x_n\} \in \mathcal{Z}$ *is said to be a Cauchy sequence if [37]*

$$\lim_{n\to\infty} \|x_{n+p} - x_n\| = 0 \ \textit{for every}\ p \geq 1.$$

Definition 2.1.4. *(Banach Space) A normed space $\mathcal{Z}$ is said to be complete if every Cauchy sequence of $\mathcal{Z}$ has a limit on $\mathcal{Z}$. A complete normed space is called a Banach space [37].*

Definition 2.1.5. *(Hilbert Space) A Hilbert space $\mathcal{H}$, is a Banach space with special structure. It is furnished with an inner (or scalar) product $(,)$ as defined below [37].*

For any two elements $f, g \in \mathcal{H}$, the scalar product (f, g) is a real or complex number. Let $\mathbb{C}$ denote the field of complex numbers. The map $\{f, g\} \rightarrow (f, g)$ has the following properties:

(H1) $(\alpha f, g) = \alpha(f, g)\ (f, \alpha g) = \alpha^*(f, g),\ \alpha \in C,\ f, g \in \mathcal{H}$ *where α^* is the complex conjugate of α;*

(H2) $(f, g_1 + g_2) = (f, g_1) + (f, g_2),\ \forall\ f, g_1, g_2 \in \mathcal{H}$;

(H3) $(f, g) = (g, f)^*,\ \forall\ f, g \in \mathcal{H}$;

(H4) $\|f\|^2_{\mathcal{H}} = (f, f),\ \forall\ f \in \mathcal{H}$.

Definition 2.1.6. *Let $\mathcal{Z}$ be any Banach space and ϕ a real valued function on $\mathcal{Z}$. The function ϕ is said to be [37]:*

i) lower semi continuous at $x \in \mathcal{Z}$ if, for every sequence $\{x_n\}$ converging to x,

$$\phi(x) \leq \lim_{n\rightarrow\infty} \inf \phi(x_n),$$

ii) upper semi continuous at x if

$$\phi(x) \geq \lim_{n\rightarrow\infty} \sup \phi(x_n),$$

iii) lower or upper semi continuous on a set $\Gamma \subset \mathcal{Z}$, if the corresponding statements hold for all $x \in \Gamma$.

Let X be a Banach space with the first and second duals denoted by X^* and X^{**} respectively. A sequence $x^*_n \in X^*$ is said to converge weakly to x^*, denoted by

$$x^*_n \longrightarrow x^*,$$

if, for every $x^{**} \in X^{**}$,

$$x^{**}(x^*_n) \longrightarrow x^{**}(x^*)$$

and it is said to converge in the weak star topology to x^*, denoted by

$$x_n^* \xrightarrow{w^*} x^*,$$

if, for every $x \in X$,

$$x_n^* \longrightarrow x^*(x).$$

Since every element of X induces a continuous linear functional on X^* through the relation $\hat{x}(x^*) \equiv x^*(x)$, the canonical embedding $X \hookrightarrow X^{**}$ exists. Hence the weak star topology is weaker than the weak topology [37].

For simplicity let (Σ, d_Σ) denote a complete separable metric space (a Polish space), and $(\Sigma, \mathcal{B}(\Sigma))$ the corresponding measurable space, in which $\mathcal{B}(\Sigma)$ is the σ-algebra generated by open sets in Σ. The material presented below regarding different spaces and their duals can be generalized to locally compact separable metric spaces (Σ, d_Σ).

Let $\mathcal{X}_0 \stackrel{\triangle}{=} C_0(\Sigma)$ denote the Banach space of continuous functions on Σ that vanish at infinity, $\mathcal{X}_1 \stackrel{\triangle}{=} BC(\Sigma)$ the Banach space of bounded continuous functions on Σ, and $\mathcal{X}_2 \stackrel{\triangle}{=} BM(\Sigma)$ the Banach space of bounded measurable functions on Σ, all equipped with the sup-norm. Clearly, $\mathcal{X}_0 \subset \mathcal{X}_1 \subset \mathcal{X}_2$.

It is known that the dual space $\mathcal{X}_0^*$ is isometrically isomorphic to $\mathcal{M}_{rca}(\Sigma)$, the Banach space of finite signed Borel measures on $(\Sigma, \mathcal{B}(\Sigma))$ (also known as Radon measures), the dual space $\mathcal{X}_1^*$ is isometrically isomorphic to $\mathcal{M}_{rba}(\Sigma)$, the Banach space of finitely additive finite regular signed measures on $(\Sigma, \mathcal{B}(\Sigma))$, and the dual space $\mathcal{X}_2^*$ is isometrically isomorphic to $\mathcal{M}_{ba}(\Sigma)$, the Banach space of finitely additive signed measures on $(\Sigma, \mathcal{B}(\Sigma))$. Note that when Σ is compact then $\mathcal{X}_1^*$ is isometrically isomorphic, the Banach space of countably additive signed measures on $(\Sigma, \mathcal{B}(\Sigma))$[38].

Definition 2.1.7. *A Banach space X is said to be reflexive, if $X^{**} = X$ [37].*

For $1 < p < \infty$, the L_p spaces are reflexive Banach spaces. Indeed, for $(1/p)+(1/q) = 1$, $(L_p)^* = L_q$ and $(L_q)^* = L_p$. Hence $(L_p)^{**} = L_p$ and so these spaces are reflexive.

It is well known that a closed bounded subset of a finite dimensional space is compact. Through this is false in infinite dimensional spaces, there is a similar result with respect to weak topologies. This is presented in the next theorem.

Theorem 2.1.8. *A closed bounded subset of a reflexive Banach space is weakly compact [37].*

Theorem 2.1.9. *Suppose $f : \Re^d \to \bar{\Re}$ is lower semi continuous satisfying the following conditions:*

$$-\infty < f(x) \not\equiv +\infty,\ \forall x \in \Re^d,\ \textit{and}\ \lim_{\|x\|\to\infty} f(x) = +\infty. \tag{2.1}$$

Then f attains its minimum [37].

In a finite dimensional space the closed unit ball is compact. An analogous result in infinite dimensional setting is Alaoglu's theorem.

Theorem 2.1.10. *(Alaoglu's Theorem) The unit ball $B_1(X^*)$ of the dual X^* of the Banach space X is weak star compact [37].*

Definition 2.1.11. *(Convex Set) A set S in a vector space X is called a convex set if the line segment joining any pair of points of S lies entirely in S. The former statement is equivalent to saying that for any pair of vectors $u \in S$, $v \in S$, the vector $(1-t)u+tv \in S,\ \forall t \in [0,1]$.*

Definition 2.1.12. *(Convex Function) Let E be any Banach space and f a (possibly extended) real valued function defined on E. The function f is said to be convex if for every $x, y \in E$ and $\alpha \in [0,1]$*

$$f((1-\alpha)x + ay) \leq (1-\alpha)f(x) + \alpha f(y).$$

Similarly, if E is replaced by a closed convex subset Γ on E and f satisfies the above inequality for all $x, y \in \Gamma$, then f is convex on Γ.

Definition 2.1.13. *(Gateaux differential sub differential) A real valued functional f defined on E is said to be Gateaux differentiable at the point $x_0 \in E$ if, for every $y \in E$, the limit*

$$\lim_{s\to 0}(1/s)\{f(x_0+sy) - f(x_0)\} \equiv df(x_0, y)$$

exists. Further, if $y \to df(x_0, y)$ is continuous and linear, then there exists an $e^* \in E^*$ in the dual space of E, dependent on x_0, such that

$$df(x_0, y) = (e^*, y).$$

The element $e^* \in E^*$ satisfying the preceding identity, is called Gateaux gradient of f at $x_0 \in E$. The function f is said to be Gateux differentiable if it is so at every point $x_0 \in E$ [37].

Theorem 2.1.14. *A real valued function f defined on a Banach space E is weakly lower semi continuous if it is convex and continuously (linearly) Gateux differentiable.*

2.1.2 Minimax Theory

Minimax techniques are often used in decision theory, game theory, statistics, based on the philosophy of minimizing the maximum possible loss. It started from two-player zero-sum game theory, covering both the cases where players take alternate moves and those where they make simultaneous moves. It has also been extended to general decision making in the presence of uncertainty.

Below the main minsup theorem is stated, which is used in this thesis; it is based on a generalization of the von Neumann's minimax theorem.

Theorem 2.1.15. *Let X be a compact Hausdorff space and Y an arbitrary set (not topologized). Let f be a real-valued function on $X \times Y$ such that, for every $y \in Y$, $f(x, y)$ is lower semicontinuous on X. If f is convex on X and concave on Y, then there exists an $x^* \in X$ such that*

$$\min_{x \in X} \sup_{y \in Y} f(x, y) = \sup_{y \in Y} f(x^*, y) = \sup_{y \in Y} \min_{x \in X} f(x, y). \tag{2.2}$$

If in addition Y is a compact Hausdorff space for every $x \in X$, $f(x, y)$ is upper semicontinuous on Y, then there exists an $(x^*, y^*) \in X \times Y$ saddle point, and

$$\min_{x \in X} \max_{y \in Y} f(x, y) = f(x^*, y^*) = \max_{y \in Y} \min_{x \in X} f(x, y). \tag{2.3}$$

Proof. See [39]. □

The following theorem will be invoked to prove the equivalence between constrained and unconstrained optimization problems.

Theorem 2.1.16. *(Lagrange Duality) [40] (page 224-225). Let f be a real-valued convex functional defined on a convex subset Ω of a vector space $\mathcal{X}$ and let G be a convex mapping from $\mathcal{X}$ into a normed space $\mathcal{Z}$. Suppose there exists an $x_1 \in \mathcal{X}$ such that $G(x_1) \prec 0$ (here $\prec$ is used for the ordered vector space $(\mathcal{Z}, \prec)$) and $\inf\{f(x) : G(x) \preceq 0, x \in \Omega\}$ is finite, then*

$$\begin{aligned} &\inf\{f(x) : G(x) \preceq 0, x \in \Omega\} \\ &= \max_{z^* \succeq 0} \Big(\inf_{x \in \Omega} \{f(x) + (G(x), z^*)\} \Big) \end{aligned} \tag{2.4}$$

and the maximum on the right is achieved by some $z_0^* \succeq 0$, $z_0^* \in \mathcal{Z}$.

2.1.3 Measurable Space and Probability Space

The mathematical model for a random experiment is the probability space. Before it is introduced, the class of measurable functions which play a fundamental role in integration theory, and hence in measure theory are reviewed. There is an analogy between the concepts of topological space, open set, and continuous function and measurable space, measurable set, and measurable function.

Definition 2.1.17. *(Topological Space) Let $\mathcal{Z}$ be a set and $\mathcal{B}_\mathcal{Z}$ a collection of subsets of $\mathcal{Z}$. Then $\mathcal{B}_\mathcal{Z}$ is called a topology in $\mathcal{Z}$ if the following properties hold [41].*

i) $\oslash \in \mathcal{B}_\mathcal{Z}$ *and* $\mathcal{Z} \in \mathcal{B}_\mathcal{Z}$*;*

ii) If $\mathcal{Z}_i \in \mathcal{B}_\mathcal{Z}, i = 1, 2, ..., n$*, then* $\bigcap_{i=1}^{n} \mathcal{Z}_i \in \mathcal{B}_\mathcal{Z}$

iii) If $\{\mathcal{Z}_i\}$ *is an arbitrary collection of elements of* $\mathcal{Z}$ *(finite, countable, or uncountable), then* $\bigcup_i \mathcal{Z}_i \in \mathcal{B}_\mathcal{Z}$.

The pair $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ is called a topological space and the members of $\mathcal{B}_\mathcal{Z}$ are called open sets in $\mathcal{Z}$. If $f : (\mathcal{Z}, \mathcal{B}_\mathcal{Z}) \rightarrow (Y, \mathcal{B}_Y)$, then f is continuous provided $f^{-1}(Y_i) \subset \mathcal{Z}$ is an open set for every open set $Y_i \subset Y$. Moreover, f is continuous at the point $x_0 \in \mathcal{Z}$ if for every neighborhood (nbh) A of $f(x_0)$ there exists a nbh B of x_0 such that $f(B) \subset A$.

A topological space $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ is said to be T_2 (or is said to satisfy the T_2 axiom) if given distinct $x, y \in \mathcal{Z}$, there exist disjoint open set $U, V \in \mathcal{B}_\mathcal{Z}$ (that is, $U \cap V = \oslash$) such that $x \in U$ and $y \in V$. A T_2 space is also known as a Hausdorff space. A Hausdorff topology for a set $\mathcal{Z}$ is a topology $\mathcal{B}_\mathcal{Z}$ such that $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ is a Hausdorff space.

Definition 2.1.18. *(Algebra) Let Ω be a set of elementary outcomes and $\mathcal{F}$ be a non-empty collection of subsets of Ω. Then $\mathcal{F}$ is called an Algebra on Ω if the following properties hold [41].*

i) $\Omega \in \mathcal{F}$ *(The Sample Space is an element of $\mathcal{F}$);*

ii) If $A \in \mathcal{F}$ *then* $A^c = \Omega - A \in \mathcal{F}$*, where* A^c *is the complementation of A relative to* Ω *(if a subset of* Ω *belongs to* $\mathcal{F}$*, then so is its complement);*

iii) If $A_i \in \mathcal{F}, i = 1, 2, ..., n$*, then* $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$ *(if a finite number of subsets belong to* $\mathcal{F}$*, then so is their union).*

Clearly, an algebra is a collection of subsets of a set Ω, which a) contains Ω and b) is closed under complementation and finite unions. The members of $\mathcal{F}$ are called $\mathcal{F}$-measurable sets or measurable sets.

The collection of finite unions of half open intervals $(a, b], -\infty < a < b \leq \infty$ in $\Re$ is considered. This collection is a field. However, the open interval $(0, 1) = \bigcup_{n=1}^{\infty}(0, 1 - \frac{1}{n}]$ is not in the collection although it contains each interval $(0, 1 - \frac{1}{n}]$. Similarly, it does not contain the singletons $\{x\}$, although $\{x\} = \bigcap_{n=1}^{\infty}(x - \frac{1}{n}, x]$. Therefore, in order to consider sequences of events and convergence of sequence of events, it is necessary to extend the operations on events to countable set operations. This gives rise to a σ-algebra which is closed under countable unions.

Definition 2.1.19. *(σ-Algebra) An algebra $\mathcal{F}$ on Ω is called a σ-Algebra on Ω if it is closed under countable unions, that is if the following properties hold [41].*

i) $\Omega \in \mathcal{F}$;

ii) If $A \in \mathcal{F}$ *then* $A^c \in \mathcal{F}$;

iii) If $A_i \in \mathcal{F}$, $i = 1, 2, ...$ *then* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

If $\mathcal{F}$ is a field the pair $(\Omega, \mathcal{F})$ is called a measurable space and the elements of $\mathcal{F}$ are called Events and are said to be measurable sets in Ω. Fields and σ-fields are convenient mathematical objects which express how much is known about the outcome of the experiment.

Remark 2.1.20. *If $f : X \to Y, X$ is measurable Space, Y is topological Space, then f is said to be measurable if $f^{-1}(V)$ is a measurable set in X for every open set V in Y.*

Since the intersection of arbitrary σ-algebras of subset of Ω is a σ-algebra of subsets of Ω, then for an arbitrary family $\mathcal{A}$ of subsets of Ω there is a smallest σ-algebra $\mathcal{F}$ in Ω such that $\mathcal{A} \subset \mathcal{F}$.

Theorem 2.1.21. *(Smallest σ-algebra) Let Ω be a sample space and $\mathcal{A}$ be a collection of subsets of Ω. There exists a smallest σ-algebra $\mathcal{F}(\mathcal{A})$ on Ω containing $\mathcal{A}$, which is constructed by*

$$\mathcal{F}(\mathcal{A}) = \bigcap_i \{\mathcal{N}_i; \mathcal{N}_i \text{ is a } \sigma\text{-algebra on } \Omega, \mathcal{A} \subset \mathcal{N}_i\}.$$

This is called the σ-algebra generated by $\mathcal{A}$, and it is often denoted by $\mathcal{F}(\mathcal{A}) = \sigma(\mathcal{A})$.

Borel Set. Let X be a topological space (e.g., $X = \Re^n$; the collection of all n-tuples $\{x = (x_1, x_2, ..., x_n) : x_i \in \Re,\ 1 \leq i \leq n\}$. Then there exists a smallest σ-algebra $\mathcal{F}$ on X such that every open set $\mathcal{A} \subset X$ belongs to $\mathcal{F}$. The elements $A \in \mathcal{F}$ called Borel sets and the σ-algebra $\mathcal{F} = \mathcal{F}(X)$ is called a Borel σ-algebra. For example, if $X = \Re^n$, and $\mathcal{A}$ is the collection of all open sets of $\Re^n$, the Borel σ-algebra denoted by $\mathcal{B}(\Re^n)$, contains all open sets, their complements (closed sets), all the countable unions of open sets, and all the countable unions of closed sets. In fact, $\mathcal{B}(\Re^n)$ =the smallest σ-algebra of subsets of $\Re^n$ containing all sets of the form $\{x : x_1 \in A_1, x_2 \in A_2, ..., x_n \in A_n\}$, where A_j are intervals on $\Re$ which are closed, open, semi-open, points, etc. Clearly, $\mathcal{A}$ =Collection of all open intervals of $\Re^n$ is not a σ-algebra, but there exists many σ-algebra containing $\mathcal{A}$ as a subset. The smallest σ-algebra containing $\mathcal{A}$ is the σ-algebra generated by $\mathcal{A}$. The pair $(\Re^n, \mathcal{B}(\Re^n))$ is a measurable space, called, the Borel measurable space.

Probability Space. In order to grade the possibility of occurrences of events associated with a random experiment a function (a map) has to be defined which attaches a numerical value to events $A \in \mathcal{F}$. A function $\mu : \mathcal{F} \to [0, \infty]$ is called a finite-additive set function, if μ satisfies the following two conditions.

$$(FA1)\ \ \mu(\phi) = 0;$$
$$(FA2)\ \ \mu(A \bigcup B) = \mu(A) + \mu(B),\ \text{if } A, B \in \mathcal{F} \text{ and } A \bigcap B = \oslash.$$

A finite-additive set function μ on an algebra $\mathcal{F}$ (or a σ-algebra) is called a measure, if it is countably-additive and a probability measure if it is countably-additive and $\mu(\Omega) = 1$, hence the following definition.

Definition 2.1.22. *(Probability Measure) Let $(\Omega, \mathcal{F})$ be a measurable space. The map*

$$P : \mathcal{F} \to [0, 1],\ P(A) \in [0, 1],\ \forall A \in \mathcal{F}$$

is called a probability measure on (Ω, F) if it satisfies the following properties.

i) $P(\phi) = 0$*;*

ii) $P(\Omega) = 1$*;*

iii) $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$*, if* $A_i \in \mathcal{F}, \forall i$ *and* $\{A_j\}_{j=1}^{\infty}$ *are disjoint, e.g.,* $A_i \bigcap A_j = \oslash, \forall i \neq j$.

The triple $(\Omega, \mathcal{F}, P)$ is called a probability space [41].

Completeness of Probability Space. A probability Space $(\Omega, \mathcal{F}, P)$ is said to be Complete if whenever $B \in \mathcal{F}$ and $P(B) = 0$ then $A \in \mathcal{F}$ for all $A \subset B$. The subsets A of an event B of zero probability is called a null set, therefore $(\Omega, \mathcal{F}, P)$ is complete if $\mathcal{F}$ includes all events of zero probability. Any probability space $(\Omega, \mathcal{F}, P)$ which is not complete can be uniquely extended to the σ-algebra $\overline{\mathcal{F}} = \mathcal{F} \bigvee \{\text{Null Sets}\}$.

Definition 2.1.23. *Let $(\Omega, \mathcal{F})$ denote a measurable space and P a positive measure μ on Ω. Let f be a measurable function on $(\Omega, \mathcal{F})$. Define*

$$\|f\|_p \stackrel{\triangle}{=} \Big\{ \int_\Omega |f|^p(\omega) d\mu(\omega) \Big\}^p, \;\; 1 \leq p < \infty.$$

$L^p(\Omega, \mathcal{F}, P)$ is the set of all measurable functions f on $(\Omega, \mathcal{F})$ for which $\|f\|_p < \infty$. $\|f\|_p$ denotes the L^p-norm of f.

Definition 2.1.24. *(Mutually Independence) The events $\{A_i\}_{i=1}^n$ are said to be mutually independent if*

$$P(A_{i_1} \bigcap A_{i_2} \cdots \bigcap A_{i_k}) = P(A_{i_1}) P(A_{i_2}) \cdots P(A_{i_k})$$

for all non-empty subsets $\{i_1, i_2, \cdots, i_k\}$ of $\{1, 2, \cdots, n\}$ [41].

Remark 2.1.25. *Pairwise independence between events does not imply mutual independence of all events. Disjoint events are not independent since if A, B are disjoint and independent events then $0 = P(A \bigcap B) = P(A)P(B) \Rightarrow P(A)$ and/or $P(B)$ are zero.*

2.1.4 Random Variables

Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces, and let $f : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$. Then the function f is called $\mathcal{F}_1/\mathcal{F}_2$ or $\mathcal{F}_1$ measurable if

$$f^{-1}(A) \stackrel{\triangle}{=} \{\omega : f(\omega) \in A\} \in \mathcal{F}_1, \;\; \forall A \in \mathcal{F}_2.$$

The set $f^{-1}(A)$ is called the inverse image of $A \in \mathcal{F}_2$. If $f : \Omega \to Y$ where $(\Omega, \mathcal{F})$ is a measurable space, Y is a topological space (e.g., $\Re^n$), then f is $\mathcal{F}/\mathcal{B}(\Re^n)$-measurable provided $f^{-1}(V) \in \mathcal{F}$ for every open set $V \subset Y$.

The σ-algebra $\mathcal{F}(f)$ generated by f is the smallest $\sigma - algebra$ on Ω containing all the sets $\{f^{-1}(V) : V \subset Y \text{ is open}\}$ and f will be $\mathcal{F}(f)/\mathcal{B}(\Re^n)$. Moreover, if $Y = \Re^n$ then

$$\mathcal{F}(f) = \{f^{-1}(V) : V \in \mathcal{B}(\Re^n)\}.$$

Clearly, if $(\Omega, \mathcal{B})$ is a Borel measurable space and $f : \Omega \to Y$, where Y is a topological space and f is a continuous function, then from the definition of continuous function

$$f^{-1}(V) \in \mathcal{B}, \ \ \forall \text{ open set } V \subset Y. \tag{2.5}$$

Hence, every continuous function is Borel measurable, called Borel function, e.g., $f : (\Re^n, \mathcal{B}(\Re^n)) \to (\Re^m, \mathcal{B}(\Re^m))$ is a Borel function.

If a probability measure P on $(\Omega, \mathcal{F})$ is defined, where $X : (\Omega, \mathcal{F}) \to (\Re^n, \mathcal{B}(\Re^n))$, is a measurable function then X is called a Random Variable (RV) defined on the probability space $(\Omega, \mathcal{F}, P)$.

Definition 2.1.26. *(Random Variable) Let $X : \Omega \to \Re^n$ be a function defined on a probability space $(\Omega, \mathcal{F}, P)$. Then X is called an n-dimensional Random Variable (measurable function)*

$$X : (\Omega, \mathcal{F}) \to (\Re^n, \mathcal{B}(\Re^n))$$

if for every $A \in \mathcal{B}(\Re^n)$ the set

$$X^{-1}(A) \stackrel{\triangle}{=} \{\omega : X(\omega) \in A\} \in \mathcal{F}.$$

Clearly, the $\sigma - algebra\, \mathcal{F}^X$ (or $\mathcal{F}(X)$) generated by X is the smallest $\sigma - algebra$ on Ω containing all the sets

$$X^{-1}(A) \ : \ A \subset \Re^n \text{ is open}$$

under which X is measurable. Equivalently,

$$\mathcal{F}^X = X^{-1}(\mathcal{B}(\Re^n)) = \{X^{-1}(B) : B \in \mathcal{B}(\Re^n)\}$$

is the smallest σ-algebra on Ω under which X is measurable. However, $\mathcal{B}(\Re^n)$ is generated by products of open sets of the form

$$\{x : -\infty < x_1 \leq \alpha_1, \ldots, -\infty < x_n \leq \alpha_n\}, \ \alpha_j \in \Re, 1 \leq j \leq n \tag{2.6}$$

therefore for X to be a RV it is sufficient for every set of the form

$$\{\omega : X_1(\omega) \leq \alpha_1, \ldots, X_n(\omega) \leq \alpha_n\},\ \alpha_j \in \Re, 1 \leq j \leq n$$

to be an event. This is because $\mathcal{B}(\Re^n)$ is the family of subsets obtained by starting with (2.6) and taking repeatedly all complements, countable unions, intersections. Also, if $\{X_t : 0 \leq t \leq T\}$ is a family of random variables $\mathcal{F}_t^X \stackrel{\triangle}{=} \sigma(X_t : 0 \leq t \leq T) = \bigvee_{t\in T} \mathcal{F}(X_t) = \sigma(\bigcup_{t\in T} \mathcal{F}(X_t))$ is the smallest σ-algebra on Ω under which $\{X_t : 0 \leq t \leq T\}$ are measurable.

Complex Random Variables. A complex square matrix Q is called Hermitian if $Q = Q^\dagger$ (where † denotes complex conjugate transpose), and has the following properties [42]:

i) The eigenvalues of a Hermitian matrix are real.

ii) The diagonal elements of a Hermitian matrix are real.

iii) The complex conjugate of a Hermitian matrix is a Hermitian matrix.

iv) If Q is a Hermitian matrix, and B is a complex matrix of same order as Q , then BQB is a Hermitian matrix.

v) A matrix is symmetric if and only if it is real and Hermitian.

vi) Hermitian matrices are a vector subspace of the vector space of complex matrices. The real symmetric matrices are a subspace of the Hermitian matrices.

vii) Hermitian matrices are also called self-adjoint since if Q is Hermitian, then in the usual inner product of $\mathbb{C}^n$, we have $(uQv) = (Quv)$ for all $u, v \in \mathbb{C}^n$.

A complex Random Variable (RV) $Z \in \mathbb{C}^n$ is simply a pair of real RVs of $\Re^n$ such that

$$Z = X + jY. \tag{2.7}$$

It is therefore always possible to treat all the problems concerning complex RVs by using a real RV of $\Re^{2n}$ dimension.

A complex random vector $Z \in \mathbb{C}^n$ is said to be Gaussian if the real random vector $\bar{Z} \in \Re^{2n}$ consisting of its real and imaginary parts, $\bar{Z} = \begin{bmatrix} \Re(Z) \\ \Im(Z) \end{bmatrix}$ is Gaussian. Thus,

to specify the distribution of a complex Gaussian random vector Z, it is necessary to specify the expectation and covariance of $\bar{Z}$, namely

$$E[\bar{Z}] \in \Re^{2n} \text{ and } E[(\bar{Z} - E[\bar{Z}])(\bar{Z} - E[\bar{Z}])^T] \in \Re^{2n \times 2n}.$$

A complex Gaussian random vector Z is circularly symmetric if the covariance of the corresponding $\bar{Z}$ has the structure

$$E[(\bar{Z} - E[\bar{Z}])(\bar{Z} - E[\bar{Z}])^T] = \begin{bmatrix} \Re(Q) & -\Im(Q) \\ \Im(Q) & \Re(Q) \end{bmatrix} \tag{2.8}$$

for some Hermitian non-negative definite $Q \in \mathbb{C}^{n \times n}$. Note that the real part of an Hermitian matrix is symmetric and the imaginary part of an Hermitian matrix is anti-symmetric and thus the matrix appearing in (2.8) is real and symmetric. In this case $E[(\bar{Z} - E[\bar{Z}])(\bar{Z} - E[\bar{Z}])^T] = Q$, and thus, a circularly symmetric complex Gaussian random vector Z is specified by prescribing $E[Z]$ and $E[(Z - E[Z])(Z - E[Z])^\dagger]$.

2.1.5 Distribution Function

Let $\Big(\Omega, \mathcal{F}, P\Big)$ be a probability space and let $X : (\Omega, \mathcal{F}) \to (\Omega_1, \mathcal{F}_1)$ be an $\mathcal{F}/\mathcal{F}_1$-measurable RV. From the point of view of computations, it is often convenient to work with an induced measure on $\mathcal{F}_1$. This amounts to defining the probability measure induced by the RV on its range rather than treat points with respect to the measure P, and work with a probability measure on $\mathcal{F}_1$ with $\omega \in \Omega_1$ as its sample values.

For the specific case, the RV, $X : (\Omega, \mathcal{F}) \to (\Omega_1, \mathcal{F}_1)$ induces a probability measure P_X on $(\Omega_1, \mathcal{F}_1)$ by

$$\begin{aligned} P_X(A_1) \stackrel{\triangle}{=} P \circ X^{-1}(A_1) &= P(\{\omega : X(\omega) \in A_1\}) \\ &= P(X \in A_1), \quad A_1 \in \mathcal{F}_1. \end{aligned}$$

If $(\Omega_1, \mathcal{F}_1) = (\Re, \mathcal{B}(\Re))$ then someone can work with a probability measure on $\mathcal{B}(\Re)$ with $x \in \Re$ as its sample points.

Definition 2.1.27. *(Probability Distribution) Let $(\Omega, \mathcal{F}, P)$ be a Probability Space and $X : (\Omega, \mathcal{F}) \to (\Re, \mathcal{B}(\Re))$ a RV. The function $F_X(\cdot)$ defined as*

$$F_X(x) \stackrel{\triangle}{=} P(\{\omega : X(\omega) \leq x\}) = P_X(X \leq x)$$

is called the (cumulative) probability distribution of X.

Thus, the relationship

$$P_X(A) = P(\{\omega : X(\omega) \in A\})$$

defines a probability measure P_X on $(\Re, \mathcal{B}(\Re))$. Note that $F_X(x)$ is a probability distribution defined on $\Re$, e.g., it corresponds to the probability measure corresponding to P induced by $X(\cdot)$ on $\Re$.

Suppose $X_1, X_2, ..., X_n$ are n real-valued RV's and $X = (X_1, X_2, ..., X_n)$, then

$$X : (\Omega, \mathcal{F}) \to (\Re^n, \mathcal{B}(\Re^n))$$

is a measurable function. The function

$$F_X(x) = F_X(x_1, x_2, ...x_n) = P(\{\omega : X_i(\omega) \leq x_i, i = 1, ...n\}), x \in \Re^n$$

is called the joint probability distribution function of X. Similarly as above, the relationship

$$P_X(A) = P(\{\omega : X(\omega) \in A\}), \quad A \in \mathcal{B}(\Re^n)$$

defines a Borel probability measure.

A real-valued RV X is said to be discrete if there exists a countable set $S = \{x_i\}$ such that

$$\sum_{x_i \in S} P(\{\omega : X(\omega) = x_i\}) = 1.$$

If X is discrete, then the distribution function F_X is a function which is constant except for jumps at $x_i, i = 1, 2, ...$, the size of the jump at x_i being $P(\{\omega : X(\omega) = x_i\})$. For an arbitrary Borel set A,

$$P_X(A) = \sum_{x_i \in A \bigcap S} P(\{\omega : X(\omega) = x_i\}).$$

Let P be a probability measure on $(\Re^n, \mathcal{B}(\Re^n))$. It is said to be singular (with respect to the Lebesgue measure) if there exist a set $S \in \mathcal{B}(\Re^n)$ such that $P(S) = 1$ and the Lebesgue measure of S is zero. On the other hand, P is said to be absolutely

continuous (w.r.t. the Lebesgue measure) if Lebesgue measure of (A) implies $P(A) = 0$. Clearly, if $X_1, X_2, ..., X_n$ are discrete RV's, then P_X is singular. If $X_1, X_2, ..., X_n$ are such that P_X is absolutely continuous, then there exists a non-negative Borel function $p_X(x), x \in \Re^n$ such that

$$P_X(A) = \int_A p_X(x)dx, A \in \mathcal{B}(\Re^n).$$

The function p_X is called the probability density function for X. In terms of the distribution

$$F_X(x_1, x_2, ...x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p_X(x_1, ..., x_n)dx_1...dx_n$$

which implies

$$p_X(x_1, ...x_n) \stackrel{\triangle}{=} \frac{\partial^n}{\partial x_1...\partial x_n} F_X(x_1, x_2, ..., x_n).$$

Definition 2.1.28. *(Stochastic Kernel) Given a measurable space $(\Omega, \mathcal{F})$ on which the RVs, X and Y are defined, via $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \Sigma_{\mathcal{X}})$, and $Y : (\Omega, \mathcal{F}) \to (\mathcal{Y}, \Sigma_{\mathcal{Y}})$, respectively, then the relation between the RV X and the RV Y is defined via a probabilistic mapping. The mapping $\mu : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$ satisfies the following two conditions:*

(i) For every $x \in \mathcal{X}$, the set function $\mu(x, \cdot)$ is a probability measure on $\Sigma_{\mathcal{Y}}$ (possibly finite additive);

(ii) For every $F \in \Sigma_{\mathcal{Y}}$, the function $\mu(\cdot, F)$ is $\mathcal{X}$-measurable.

The mapping μ is called a stochastic kernel or transition probability.

2.1.6 Duality Relation Between KL Distance and Free Energy

Next the basic definitions and duality relations are introduced between Kullback-Leibler (KL) distance, free energy, and cumulant moment generating function.

Definition 2.1.29. *Let $\nu, \mu \in \mathcal{M}_1(\Sigma)$ (the set of probability measures) and $\ell : \Sigma \to \Re$ a measurable function.*

1) The moment generating function of ℓ with respect to μ is defined by

$$M_\mu(s) \stackrel{\triangle}{=} E_\mu\Big(e^{s\ell}\Big) = \int_\Sigma e^{s\ell} d\mu \in (0, \infty], \ s \in \Re. \tag{2.9}$$

2) The cumulant generating function of ℓ with respect to μ is defined by

$$\Psi_\mu(s) \stackrel{\triangle}{=} \log M_\mu(s) = \log \int_\Sigma e^{s\ell} d\mu, \ s \in \Re. \tag{2.10}$$

3) The free energy of ℓ with respect to μ is defined by $\mathcal{E}(\ell, \mu) \stackrel{\triangle}{=} \Psi_\mu(1) \in (-\infty, \infty]$.

4) The KL distance of $\nu \in \mathcal{M}_1(\Sigma)$ with respect to $\mu \in \mathcal{M}_1(\Sigma)$ is defined by

$$H(\nu|\mu) \stackrel{\triangle}{=} \begin{cases} \int_\Sigma \log(\frac{d\nu}{d\mu}) d\nu, & \text{if } \nu \ll \mu \\ +\infty, & \text{otherwise.} \end{cases}$$

It can be shown that $\mathcal{E}(\ell, \mu)$ as a function of ℓ is convex, $H(\nu|\mu)$ as a function of $\mu, \nu \in \mathcal{M}(\Sigma)$ is convex in both arguments, $M_\mu(s), \Psi_\mu(s)$ are convex functions of $s \in \Re$, and $H(\nu|\mu) \geq 0$, and $H(\nu|\mu) = 0$, if and only if $\mu = \nu$. Moreover, $H(\nu|\mu)$ is often used as a measure of discrepancy between two probability measures.

The moment generating function (2.9) and cumulant general function (2.10) are often employed as pay-off functions in nonlinear stochastic control problems to achieve robustness. Such pay-off's are called risk-sensitive [43],[44].

2.2 Change of Probability Measure

A basic technique used throughout this dissertation is a change of probability measure starting with probability P. A new probability $\overline{P}$ is defined such that under $\overline{P}$ the observations are independent and identically distributed random variables. Calculations take place in the mathematically ideal world of $\overline{P}$ which allows interchange of expectations and summations. They are then related to the real world by an inverse change of measure. The measure change concept is the key to many of the results in the following Chapters.

Change of measure is a fundamental theorem in measure theory known as Radon-Nikodym Theorem. A version of Radon-Nikodym Theorem suitable for probability measures, is stated here.

Theorem 2.2.1. *(Radon-Nikodym Theorem)*

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\overline{P}$ be another measure defined also on $\mathcal{F}$ such that $\overline{P}$ is absolutely continuous with respect to P ($\overline{P} \ll P$), namely,

$$P(B) = 0 \Rightarrow \overline{P}(B) = 0, \ \ \forall B \in \mathcal{F}. \tag{2.11}$$

Then there exists an $\mathcal{F}$-measurable function $\phi : \Omega \to \Re$, such that $\phi \in L^1(\Omega, \mathcal{F}, P)$ and,

$$\overline{P}(B) = \int_B \phi(\omega) dP(\omega), \ \ \forall B \in \mathcal{F}. \tag{2.12}$$

The function ϕ is unique except on a subset of P-measure zero.

This function ϕ is often written as $\phi = \frac{d\overline{P}}{dP}\Big|_{\mathcal{F}}$ and is called the Radon-Nikodym derivative (RND) since it satisfies

$$\overline{P}(B) = \int_B d\overline{P} = \int_B \phi dP, \ \ \forall B \in \mathcal{F}. \tag{2.13}$$

Lemma 2.2.2. *Let $(\Omega, \mathcal{F})$ be a measurable space, and let P, and $\overline{P}$ two measures defined on $(\Omega, \mathcal{F})$ such that P is absolutely continuous with respect to $\overline{P}$ and vice versa (mutually absolutely continues), namely,*

$$\begin{aligned} P(B) = 0 \Rightarrow \overline{P}(B) = 0, \ \ \forall B \in \mathcal{F}, \\ \overline{P}(B) = 0 \Rightarrow P(B) = 0, \ \ \forall B \in \mathcal{F}. \end{aligned} \tag{2.14}$$

Then, given $\phi(\omega) \stackrel{\triangle}{=} \frac{d\overline{P}}{dP}\Big|_{\mathcal{F}}$ the two measures can be expressed as

$$\begin{aligned} \overline{P}(B) = \int_B \phi(\omega) dP, \\ P(B) = \int_B \phi^{-1}(\omega) d\overline{P}. \end{aligned} \tag{2.15}$$

Equivalently, for any random variable $X : \Omega \to \Re$, the following holds

$$\begin{aligned} \overline{E}[X] = E[\phi X] = E\Big[\frac{d\overline{P}}{dP} X\Big], \\ E[X] = \overline{E}[\phi^{-1} X] = \overline{E}\Big[\frac{dP}{d\overline{P}} X\Big] \end{aligned} \tag{2.16}$$

where $\overline{E}$ and E denote expectations under $\overline{P}$ and P, respectively.

$$\begin{aligned}\overline{P}(\omega) &= \int_{\omega} \phi(\omega) dP(\omega) = 1, \\ P(\omega) &= \int_{\omega} \phi^{-1}(\omega) d\overline{P}(\omega) = 1.\end{aligned} \tag{2.17}$$

Theorem 2.2.3. *(Conditional Bayes Theorem)*

Suppose $(\Omega, \mathcal{F})$ be a measurable space and $\mathcal{G} \subset \mathcal{F}$ is a sub-σ-field. Suppose P, and $\overline{P}$ are two measures defined on $(\Omega, \mathcal{F})$ such that P is absolutely continuous with respect to $\overline{P}$ and vice versa, with RND $\frac{d\overline{P}}{dP} = \Lambda$ and $\frac{dP}{d\overline{P}} = \Lambda^{-1}$. Then if X is any integrable $\mathcal{F}$-measurable random variable,

$$\begin{aligned}\overline{E}[X|\mathcal{G}] &= \frac{E[\Lambda X|\mathcal{G}]}{E[\Lambda|\mathcal{G}]}, \ \ \overline{P} - a.s. \\ E[X|\mathcal{G}] &= \frac{\overline{E}[\Lambda^{-1} X|\mathcal{G}]}{\overline{E}[\Lambda^{-1}|\mathcal{G}]}, \ \ P - a.s.\end{aligned} \tag{2.18}$$

Proof. See [41]. □

2.2.1 Change of Probability Measure for Random Processes

Definition 2.2.4. *A Random Process (RP) $\{\Phi_k\}, \ \ k \in N_0 \stackrel{\triangle}{=} \{0, 1, 2, 3, \ldots\}$ is said to be $\{\mathcal{F}_k\}_{k \geq 0}$ adapted if Φ_k is $\mathcal{F}_k$-measurable for every $k \in N_0$.*

Lemma 2.2.5. *Let $(\Omega, \mathcal{F}, P)$ be a complete probability space and let $\{\mathcal{F}_k\}$, $\{\mathcal{F}_k^{\mathcal{Y}}\}, \ \ k \in N_0$ be complete sub-sigma fields of $\mathcal{F}$ such that $\mathcal{F}_k^{\mathcal{Y}} \subset \mathcal{F}_k \subset \mathcal{F}, \ \ \forall \ \ k \in N_0$.*

Let $\Phi : \Omega \times N_0 \rightarrow \Re$, a RP such that $\{\Phi_k\}$ is $\{\mathcal{F}_k\}$ adapted, and $\Phi \in L^1(\Omega, \mathcal{F}, P)$.

Then

$$E[\Phi_k|\mathcal{F}_k^{\mathcal{Y}}] = \frac{\overline{E}[\Lambda_k \Phi_k|\mathcal{F}_k^{\mathcal{Y}}]}{\overline{E}[\Lambda_k|\mathcal{F}_k^{\mathcal{Y}}]}, \tag{2.19}$$

where $\Lambda_k \stackrel{\triangle}{=} \frac{dP}{d\overline{P}}\Big|_{\mathcal{F}_k}$, $\Lambda_k^{-1} \stackrel{\triangle}{=} \frac{d\overline{P}}{dP}\Big|_{\mathcal{F}_k} \ \ P - a.s.$, $E[\Lambda_k^{-1}] = 1$

2.2.2 Change of Probability Measure for Linear Systems

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space on which $\{x_k\}$, $\{y_k\}$, $k \in N_0$ are defined by

$$x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}, \quad x_0 \in \Re^n \tag{2.20}$$
$$y_k = C_k x_k + D_k v_k, \quad y_0 \in \Re^d. \tag{2.21}$$

Assume the following conditions hold:

(i) $w : \Omega \times N_0 \to \Re^m$ is an independent and identically distributed (iid) sequence with density

$$\Phi_{w_k}(w) = \frac{1}{(2\pi)^{m/2}}\, e^{\frac{-w^T w}{2}}, \quad w_k \sim N(0, I_m);$$

(ii) $v : \Omega \times N_0 \to \Re^d$ is an iid sequence with density

$$\Phi_{v_k}(v) = \frac{1}{(2\pi)^{d/2}}\, e^{\frac{-v^T v}{2}}, \quad v_k \sim N(0, I_d);$$

(iii) $x_0 : \Omega \to \Re^n$ is a Random Variable (RV);

(iv) $\{x_0, \{w_k\}, \{v_k\}\}$ are mutually independent;

(v) $D_k > 0, \quad \forall k \in N_0$.

Next, the σ-algebras are defined

$$\mathcal{G}_k^0 \triangleq \sigma\{x_0, x_1, \ldots, x_k, y_0, y_1, \ldots, y_k\}, \quad \forall k \in N_0$$
$$\mathcal{F}_k^{0,\mathcal{Y}} \triangleq \sigma\{y_0, y_1, \ldots, y_k\}, \quad \forall k \in N_0.$$

$\{\mathcal{G}_k\}$, $\{\mathcal{F}_k^{\mathcal{Y}}\}$, $\forall k \in N_0$ are the complete filtrations generated by $\{x_0, x_1, \ldots, x_k, y_0, y_1, \ldots, y_k\}$, $\{y_0, y_1, \ldots, y_k\}$, respectively, $\forall k \in N_0$.

A. Measure Change of the Observation Process

The RV on $(\{\mathcal{G}_k\}, P)$, is considered, defined by

$$\lambda_k^{-1} \triangleq |D_k| \frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)}, \quad k \in N_0 \tag{2.22}$$

setting also

$$\Lambda_k^{-1} \stackrel{\triangle}{=} \prod_{s=0}^{k} \lambda_s^{-1}. \tag{2.23}$$

Then $\Lambda_k^{-1} > 0$, a.s. and in particular $E[\Lambda_k^{-1}] = 1, \ \ \forall k \in N_0$, which can be shown as follows

$$\begin{aligned} E[\Lambda_k^{-1}] &= E\Big[\prod_{s=0}^{k} |D_s| \frac{\Phi_{v_s}(y_s)}{\Phi_{v_s}(v_s)}\Big] = E\Big[E\Big[\prod_{s=0}^{k} |D_s| \frac{\Phi_{v_s}(y_s)}{\Phi_{v_s}(v_s)}\Big|\mathcal{G}_{k-1}\Big]\Big] \\ &= E\Big[\prod_{s=0}^{k-1} |D_s| \frac{\Phi_{v_s}(y_s)}{\Phi_{v_s}(v_s)} E\Big[|D_k| \frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)}\Big|\mathcal{G}_{k-1}\Big]\Big]. \end{aligned}$$

But given

$$\prod_{s=0}^{k-1} |D_s| \frac{\Phi_{v_s}(y_s)}{\Phi_{v_s}(v_s)} = \Lambda_{k-1}^{-1}$$

and

$$E\Big[|D_k| \frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)}\Big|\mathcal{G}_{k-1}\Big] = \int_{\Re^d} |D_k| \frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)} \Phi_{v_k}(v_k) dv_k = \int_{\Re^d} |D_k| \Phi_{v_k}(y_k) |D_k|^{-1} dy_k = 1$$

(here change of variables is being used, $v_k = D_k^{-1}(y_k - C_k x_k)$, $dv_k = |D_k|^{-1} dy_k$). Therefore

$$E[\Lambda_k^{-1}] = E[\Lambda_{k-1}^{-1}] = \ldots = E[\Lambda_0^{-1}] = 1.$$

Hence, Λ_k^{-1} can be used to define a new probability measure $\overline{P}$ on $(\Omega, \{\mathcal{G}_k\})$ through the RND

$$\frac{d\overline{P}}{dP}\Big|_{\mathcal{G}_k} = \Lambda_k^{-1} \ \text{ or } \ d\overline{P}(A) = \int_A \Lambda_k^{-1}(\omega) dP(\omega), \ \ \forall A \in \mathcal{G}_k. \tag{2.24}$$

Next it is shown that under the new measure $\overline{P}$, the observation process $\{y_k\}, \ \ \forall k \in N_0$ is a sequence of iid RVs with density $\{\Phi_{v_k}(\cdot)\}, \ \ \forall k \in N_0$.

Lemma 2.2.6. *Let $(\Omega, \mathcal{F}, P)$ be a probability space on which (2.20), (2.21) are defined.*

Next, define

$$\frac{d\overline{P}}{dP}\Big|_{\mathcal{G}_k} = \Lambda_k^{-1} = \prod_{s=0}^{k} |D_s| \frac{\Phi_{v_s}(y_s)}{\Phi_{v_s}(v_s)},$$

where $v_k = D_k^{-1}(y_k - C_k x_k)$. Then under measure $\overline{P}$ the sequence $\{y_k)\}, k \in N_0$ is iid with density

$$\Phi_{y_k}(y) = \frac{e^{-y^T y/2}}{(2\pi)^{d/2}} = \Phi_{v_k}(y), \ \ \forall k \in N_0.$$

Proof. For $s \in \Re^d$, the event $\{y_k \leq s\} \stackrel{\triangle}{=} \{y_k^1 \leq s^1, \ldots, y_k^d \leq s^d\}$ is considered

$$\begin{aligned}
\overline{P}\Big(\{y_k \leq S\}|\mathcal{G}_{k-1}\Big) &= \overline{E}\Big[I_{\{\omega: y_k(\omega) \leq s\}}\Big|\mathcal{G}_{k-1}\Big] \\
&= \frac{E\Big[\Lambda_k^{-1} I_{\{\omega: y_k(\omega) \leq s\}}\Big|\mathcal{G}_{k-1}\Big]}{E\Big[\Lambda_k^{-1}\Big|\mathcal{G}_{k-1}\Big]} \\
&= \frac{E\Big[\Lambda_{k-1}^{-1}|D_k|\frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)} I_{\{\omega: y_k(\omega) \leq s\}}\Big|\mathcal{G}_{k-1}\Big]}{E\Big[\Lambda_{k-1}^{-1}|D_k|\frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)}\Big|\mathcal{G}_{k-1}\Big]} \\
&= \frac{E\Big[|D_k|\frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)} I_{\{\omega: y_k(\omega) \leq s\}}\Big|\mathcal{G}_{k-1}\Big]}{E\Big[|D_k|\frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)}\Big|\mathcal{G}_{k-1}\Big]}.
\end{aligned}$$

But given

$$E\Big[|D_k|\frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)}\Big|\mathcal{G}_{k-1}\Big] = \int_{\Re^d} |D_k|\frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)}\Phi_{v_k}(v_k)dv_k = \int_{\Re^d} \Phi_{v_k}(y_k)dy_k = 1$$

(using change of variables, $dy_k = |D_k|dv_k$). Therefore

$$\overline{P}\Big(\{y_k \leq S\}|\mathcal{G}_{k-1}\Big) = \int_{\Re^d} I_{\{\omega: y_k(\omega) \leq s\}}\Phi_{v_k}(y_k)dy_k = \int_{-\infty}^{s^1} \ldots \int_{-\infty}^{s^d} \Phi_{v_k}(y_k)dy_k, \quad q.e.d.$$

□

Remark 2.2.7. *The above result shows the following.*

I. *Start with a complete probability space* $\Big(\Omega, \mathcal{F}, P, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\Big)$ *on which (2.20), (2.21) are defined.*

Define $\Lambda_k^{-1} = \prod_{s=0}^{k} |D_s|\frac{\Phi_{v_s}(y_s)}{\Phi_{v_s}(v_s)}$.

Then Λ_k^{-1} *defines a new measure* $\overline{P} \ll P$ *under which:*

1. *The distribution of* $\{x_k\}$ *is the same under* $\overline{P}$ *and* P.
2. *The sequence* $\{y_k\}$ *is an* $(\{\mathcal{G}_k\}, \overline{P})$ *Gaussian sequence with density*

$$\Phi_{y_k}(y) = \frac{e^{-y^T y/2}}{(2\pi)^{d/2}} = \Phi_{v_k}(y), \quad \forall k \in N_0.$$

3. *The inverse of* Λ_k^{-1} *can be used to define a new measure* $P \ll \overline{P}$ *by setting*

$$\frac{dP}{d\overline{P}}\Big|_{\mathcal{G}_k} = \Lambda_k = \prod_{s=0}^{k} |D_s|^{-1}\frac{\Phi_{v_s}(v_s)}{\Phi_{v_s}(y_s)}.$$

II. Start with a complete probability space $\left(\Omega, \mathcal{F}, \overline{P}, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\right)$ *on which* $\{x_k\}$, $\{y_k\}$, $k \in N_0$ *are defined by*

$$x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}, \quad x_0 \in \Re^n$$

$\{y_k\}$ *is an iid sequence with density* $\Phi_{y_k}(y) = \dfrac{e^{-y^T y/2}}{(2\pi)^{d/2}}$.

Let

$$\Lambda_k = (\Lambda_k^{-1})^{-1} = \prod_{s=0}^{k} |D_s|^{-1} \frac{\Phi_{v_s}(D_s^{-1}(y_s - C_s x_s))}{\Phi_{v_s}(y_s)}.$$

Then $\left.\frac{dP}{d\overline{P}}\right|_{\mathcal{G}_k} = \Lambda_k$ *defines a new measure* $P \ll \overline{P}$ *under which:*

1. *The distribution of* $\{x_k\}$ *is the same under* P *and* $\overline{P}$.
2. *The sequence* $\{v_k\}$ *is an* $(\{\mathcal{G}_k\}, P)$ *Gaussian sequence with density*

$$\Phi_{v_k}(v) = \frac{e^{-v^T v}}{(2\pi)^{d/2}} = \Phi_{v_k}(v),$$

where $v_k \stackrel{\triangle}{=} D_k^{-1}(y_k - C_k x_k) \Rightarrow y_k = C_k x_k + D_k v_k$.

B. Measure Change of the State and Observation Process

Start with a complete probability space $(\Omega, \mathcal{F}, P, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\})$ on which $\{x_k\}$, $\{y_k\}$, $k \in N_0$ are defined by

$$x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}, \quad x_0 \in \Re^n, \tag{2.25}$$

$$y_k = C_k x_k + D_k v_k, \quad y_0 \in \Re^d, \tag{2.26}$$

where

(i) $w : \Omega \times N_0 \to \Re^n$ is an iid sequence with density

$$\Phi_{w_k}(w) = \frac{1}{(2\pi)^{n/2}} e^{\frac{-w^T w}{2}}, \quad w_k \sim N(0, I_n);$$

(ii) $v : \Omega \times N_0 \to \Re^d$ is an iid sequence with density

$$\Phi_{v_k}(v) = \frac{1}{(2\pi)^{d/2}} e^{\frac{-v^T v}{2}}, \quad v_k \sim N(0, I_d);$$

(iii) $x_0 : \Omega \to \Re^n$ a RV;

(iv) $\{x_0, \{w_k\}, \{v_k\}\}$ mutually independent;

(v) $D_k > 0, \ \ \forall k \in N_0$;

(vi) $B_k B_k^T > 0, \ \ \forall k \in N_0$.

Next, define

$$\frac{d\overline{P}}{dP}\Big|_{\mathcal{G}_k} \stackrel{\triangle}{=} \Lambda_k^{-1} = \prod_{s=0}^{k} \lambda^{-1},$$

where $\lambda_k^{-1} = |D_k| \frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)}$ and set $\frac{d\overline{P}}{dP}\Big|_{\mathcal{G}_k} = \Lambda_k^{-1}$. Then, a new probability measure is introduced

$$\Big(\Omega, \mathcal{F}, \overline{P}, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\Big) : \begin{cases} x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1} \\ \\ \{y_k\} \text{ is an iid seq. with density } \Phi_{y_k}(y) = \Phi_{v_k}(y). \end{cases}$$

Next, start with $(\{\mathcal{G}_k\}, \overline{P})$ and define a new measure $\overline{\overline{P}}$ under which

$$\Big(\Omega, \mathcal{F}, \overline{\overline{P}}, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\Big) : \begin{cases} \{x_k\} \text{ is an iid seq. with density} \\ \Phi_{x_k}(x) = \Phi_{w_k}(w) = (2\pi)^{-n/2} \ e^{-x^T x/2} \\ \\ \{y_k\} \text{ is an iid seq. with density} \\ \Phi_{y_k}(y) = \Phi_{v_k}(y) = (2\pi)^{-d/2} \ e^{-y^T y/2} \end{cases}$$

To this end, the $(\{\mathcal{G}_k\}, \overline{P})$ RV is considered, the following is defined

$$\mu_k^{-1} \stackrel{\triangle}{=} |B_k| \frac{\Phi_{w_k}(x_k)}{\Phi_{w_k}(w_k)}, \ \ k \in N_0$$

and set

$$M_k^{-1} = \prod_{s=1}^{k} |B_s| \frac{\Phi_{w_s}(x_s)}{\Phi_{w_s}(w_s)},$$

then $\overline{E}[M_k^{-1}] = 1, \ \ \overline{P} - a.s., \ \ \forall k \in N_0$.

Hence, M_k^{-1} can be used to define a new probability measure on $(\Omega, \{\mathcal{G}_k\})$ through the RND

$$\frac{d\overline{\overline{P}}}{d\overline{P}}\Big|_{\mathcal{G}_k} = M_k^{-1} \text{ or } d\overline{\overline{P}}(A) = \int_A M_k^{-1} d\overline{P}(\omega), \ \forall A \in \mathcal{G}_k.$$

Notice that

$$d\overline{\overline{P}} = M_k^{-1} d\overline{P} = M_k^{-1} \frac{d\overline{P}}{dP} dP = M_k^{-1} \Lambda_k^{-1} dP$$

and hence

$$\frac{d\overline{\overline{P}}}{d\overline{P}} \frac{d\overline{P}}{dP}\Big|_{\mathcal{G}_k} = \frac{d\overline{\overline{P}}}{dP}\Big|_{\mathcal{G}_k} = M_k^{-1}\Lambda_k^{-1}$$

defines a change of measure from P to $\overline{\overline{P}}$.

Next it is shown that under the probability space $\left(\Omega, \mathcal{F}, \overline{\overline{P}}, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\right)$, $\{x_k\}$ is an iid seq. with density $\Phi_{x_k}(x) = \Phi_{w_k}(w)$, and $\{y_k\}$ is an iid seq. with density $\Phi_{y_k}(y) = \Phi_{v_k}(y)$.

Lemma 2.2.8. *Let $\left(\Omega, \mathcal{F}, \overline{P}\right)$ be a probability space on which*

$$x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}, \ \ x_0 \in \Re^n$$
$$\{y_k\} \text{ is an iid seq. with density } \Phi_{y_k}(y) = \Phi_{v_k}(y).$$

Define

$$\frac{d\overline{\overline{P}}}{d\overline{P}}\Big|_{\mathcal{G}_k} = M_k^{-1} = \prod_{s=1}^{k} |B_s| \frac{\Phi_{w_s}(x_s)}{\Phi_{w_s}(w_s)}.$$

Then under measure $\overline{\overline{P}}$ the seq. $\{x_k\}$ is iid with density

$$\Phi_{x_k}(x) = \Phi_{w_k}(x) = \frac{e^{-x^T x/2}}{(2\pi)^{n/2}}, \ \ \forall k \in N_0.$$

Proof. For $s \in \Re^n$ the event $\{x_k \le s\}$ is considered

$$\begin{aligned} \overline{\overline{P}}\Big(\{x_k \le s\}|\mathcal{G}_{k-1}\Big) &= \overline{\overline{E}}\Big[I_{\{\omega: x_k(\omega) \le s\}}\Big|\mathcal{G}_{k-1}\Big] \\ &= \frac{\overline{E}\Big[I_{\{\omega: x_k(\omega) \le s\}} M_k^{-1}\Big|\mathcal{G}_{k-1}\Big]}{\overline{E}\Big[M_k^{-1}\Big|\mathcal{G}_{k-1}\Big]} \end{aligned}$$

$$= \frac{\overline{E}\Big[M_{k-1}^{-1}|B_k|\frac{\Phi_{w_k}(x_k)}{\Phi_{w_k}(w_k)}I_{\{\omega:x_k(\omega)\leq s\}}\Big|\mathcal{G}_{k-1}\Big]}{E\Big[M_{k-1}^{-1}|B_k|\frac{\Phi_{w_k}(x_k)}{\Phi_{w_k}(w_k)}\Big|\mathcal{G}_{k-1}\Big]}$$
$$= \frac{\overline{E}\Big[|B_k|\frac{\Phi_{x_k}(x_k)}{\Phi_{w_k}(w_k)}I_{\{\omega:x_k(\omega)\leq s\}}\Big|\mathcal{G}_{k-1}\Big]}{\overline{E}\Big[|B_k|\frac{\Phi_{w_k}(x_k)}{\Phi_{w_k}(w_k)}\Big|\mathcal{G}_{k-1}\Big]}.$$

But given

$$\overline{E}\Big[|B_k|\frac{\Phi_{w_k}(x_k)}{\Phi_{w_k}(w_k)}\Big|\mathcal{G}_{k-1}\Big] = \int_{\Re^n}|B_k|\frac{\Phi_{w_k}(x_k)}{\Phi_{w_k}(w_k)}\,\Phi_{w_k}(w_k)dw_k = \int_{\Re^n}|B_k|\Phi_{w_k}(x_k)dw_k$$
$$= \int_{\Re^n}\Phi_{w_k}(x_k)dx_k = 1$$

(using the change of variables, $w_k = B_k^{-1}(x_k - A_k x_{k-1}) \Rightarrow dw_k = |B_k|^{-1}dx_k$), which implies the desire result. □

Remark 2.2.9. *The above results show the following*

I. Starting with

$$\Big(\Omega,\mathcal{F},\overline{P},\{\mathcal{G}_k\},\{\mathcal{F}_k^{\mathcal{Y}}\}\Big):\begin{cases} x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1} \\ \\ \{y_k\}\ \textit{is an iid seq. with density}\ \Phi_{y_k}(y)\sim N(0;I_d). \end{cases}$$

Let

$$M_k^{-1} = \prod_{s=1}^{k}|B_s|\frac{\Phi_{w_s}(x_s)}{\Phi_{w_s}(w_s)}.$$

Then M_k^{-1} defines a new measure $\overline{\overline{P}} \ll \overline{P}$ $\Big($e.g., $\frac{d\overline{\overline{P}}}{d\overline{P}}\Big|_{\mathcal{G}_k} = M_k^{-1}\Big)$ under which:

1. *The distribution of $\{y_k\}$ is invariant.*
2. *$\{x_k\}$ is an $\Big(\mathcal{G}_k,\overline{\overline{P}}\Big)$ Gaussian seq. with density $N(0;I_n)$.*
3. *The inverse of M_k^{-1} can be used to define a new probability measure $\overline{\overline{P}} \ll \overline{P}$ by setting*

$$\frac{d\overline{P}}{d\overline{\overline{P}}}\Big|_{\mathcal{G}_k} = M_k = \prod_{s=1}^{k}\frac{\Phi_{w_s}(B_s^{-1}(x_s - A_s x_{s-1}))}{\Phi_{w_s}(x_s)}$$

where $w_k \overset{\triangle}{=} B_k^{-1}(x_k - A_k x_{k-1})$ is an iid seq. with density $N(0;I_n)$.

II. Start with

$$\left(\Omega, \mathcal{F}, \overline{\overline{P}}, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\right) : \begin{cases} \{x_k\} \text{ is an } \sim N(0; I_n) \text{ iid seq.} \\ \\ \{y_k\} \text{ is an } \sim N(0; I_d) \text{ iid seq.} \end{cases}$$

Let

$$M_k = \prod_{s=1}^{k} \frac{\Phi_{w_s}(B_s^{-1}(x_s - A_s x_{s-1}))}{\Phi_{w_s}(x_s)}.$$

Then M_k defines an new probability measure by setting $\frac{d\overline{P}}{d\overline{\overline{P}}}\Big|_{\mathcal{G}_k} = M_k$ under which:

1. *The distribution of $\{y_k\}$ is invariant.*
2. *$\{w_k\}$ is an $\left(\mathcal{G}_k, \overline{P}\right)$ iid seq. with density $N(0; I_n)$ where $w_k \stackrel{\triangle}{=} B_k^{-1}(x_k - A_k x_{k-1}) \Rightarrow x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}$.*

C. Change of Measure from iid Sequences

Here it is shown how the dynamics

$$\left(\Omega, \mathcal{F}, P, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\right) : \begin{cases} x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}, \quad x_0 \in \Re^n \\ \\ y_k = C_k x_k + D_k v_k \quad y_0 \in \Re^d \end{cases} \tag{2.27}$$

of (2.20), (2.21) can be introduced starting with an initial probability space

$$\left(\Omega, \mathcal{F}, \overline{\overline{P}}, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\right) : \begin{cases} x_k \sim N(0; I_n) = \Phi_{w_k}(x) = \frac{e^{-x^T x/2}}{(2\pi)^{n/2}} \\ \\ y_k \sim N(0; I_d) = \Phi_{v_k}(y) = \frac{e^{-y^T y/2}}{(2\pi)^{d/2}}. \end{cases}$$

Define the sigma fields

$$\begin{aligned} \mathcal{G}_k &\stackrel{\triangle}{=} \sigma\{x_0, \ldots, x_k, y_0, \ldots, y_k\} \\ \mathcal{F}_k^{\mathcal{Y}} &\stackrel{\triangle}{=} \sigma\{y_0, \ldots, y_k\}, \end{aligned}$$

where $\mathcal{G}_k$, $\mathcal{F}_k^{\mathcal{Y}}$ are complete and the complete filtrations $\{\mathcal{G}_k\}$, $\{\mathcal{F}_k^{\mathcal{Y}}\}$.

Let

$$\lambda_0 = \frac{\Phi_{v_0}(D_0^{-1}(y_0 - C_0 x_0))}{|D_0|\Phi_{v_0}(y_0)}$$

and

$$\lambda_k = \frac{\Phi_{v_k}(D_k^{-1}(y_k - C_k x_k))}{|D_k|\Phi_{v_k}(y_k)} \times \frac{\Psi_{w_k}(B_k^{-1}(x_k - A_k x_{k-1}))}{|B_k|\Psi_{w_k}(x_k)}, \quad k \geq 1.$$

Set

$$\Lambda_k = \prod_{s=0}^{k} \lambda_s. \tag{2.28}$$

Then define a new measure P on $\Big(\Omega, \{\mathcal{G}_k\}\Big)$ by setting

$$\frac{dP}{d\overline{\overline{P}}}\Big|_{\mathcal{G}_k} = \Lambda_k. \tag{2.29}$$

It will be shown that under P,

$$\begin{aligned} v_k &\stackrel{\triangle}{=} D_k^{-1}(y_k - C_k x_k), \quad k \in N_0 \\ w_k &\stackrel{\triangle}{=} B_k^{-1}(x_k - A_k x_{k-1}), \quad k \in N_0 \end{aligned}$$

are iid seq with density $N(0; I_d)$, $(0; I_n)$, respectively.

Lemma 2.2.10. *On $\Big(\Omega, \mathcal{F}\Big)$ and under measure P the seq. defined by*

$$\begin{aligned} v_k &\stackrel{\triangle}{=} D_k^{-1}(y_k - C_k x_k), \quad k \in N_0 \\ w_k &\stackrel{\triangle}{=} B_k^{-1}(x_k - A_k x_{k-1}), \quad k = 1, \ldots, K \end{aligned}$$

are iid and

$$\Phi_{v_k}(v) = \frac{e^{-v^T v/2}}{(2\pi)^{d/2}}, \quad \Psi_{w_k} = \frac{e^{-w^T w/2}}{(2\pi)^{n/2}}.$$

Proof. $f : \Re^d \to \Re$, $g : \Re^n \to \Re$ are supposed to be bounded and Borel measurable. It is sufficient to show

$$E\Big[g(w_k)f(v_k)\Big|\mathcal{G}_{k-1}\Big] = \int_{\Re^d} \Phi_{v_k}(v)f(v)dv \ \times \ \int_{\Re^n} \Psi_{w_k}(w)g(w)dw.$$

Starting with,

$$E\Big[g(w_k)f(v_k)\Big|\mathcal{G}_{k-1}\Big] = \frac{\overline{\overline{E}}\Big[g(w_k)f(v_k)\Lambda_k\Big|\mathcal{G}_{k-1}\Big]}{\overline{\overline{E}}\Big[\Lambda_k\Big|\mathcal{G}_{k-1}\Big]} = \frac{\overline{\overline{E}}\Big[\lambda_k g(w_k)f(v_k)\Big|\mathcal{G}_{k-1}\Big]}{\overline{\overline{E}}\Big[\lambda_k\Big|\mathcal{G}_{k-1}\Big]}$$

(The fact that Λ_{k-1} is $\mathcal{G}_{k-1}$ measurable is used above).

Also,

$$\begin{aligned}
\overline{\overline{E}}\Big[\lambda_k\Big|\mathcal{G}_{k-1}\Big] &= \overline{\overline{E}}\Big[\frac{\Phi_{v_k}(D_k^{-1}(y_k-C_kx_k))}{|D_k|\Phi_{v_k}(y_k)}\times\frac{\Psi_{w_k}(B_k^{-1}(x_k-A_kx_{k-1}))}{|B_k|\Psi_{w_k}(x_k)}\Big|\mathcal{G}_{k-1}\Big]\\
&= \overline{\overline{E}}\Big\{\overline{\overline{E}}\Big[\frac{\Phi_{v_k}(D_k^{-1}(y_k-C_kx_k))}{|D_k|\Phi_{v_k}(y_k)}\\
&\quad\times\frac{\Psi_{w_k}(B_k^{-1}(x_k-A_kx_{k-1}))}{|B_k|\Psi_{w_k}(x_k)}\Big|\mathcal{G}_{k-1},x_k\Big]\Big|\mathcal{G}_{k-1}\Big\}\\
&= \overline{\overline{E}}\Big\{\frac{\Psi_{w_k}(B_k^{-1}(x_k-A_kx_{k-1}))}{|B_k|\Psi_{w_k}(x_k)}\ \overline{\overline{E}}\\
&\quad\times\Big[\frac{\Phi_{v_k}(D_k^{-1}(y_k-C_kx_k))}{|D_k|\Phi_{v_k}(y_k)}\Big|\mathcal{G}_{k-1},x_k\Big]\Big|\mathcal{G}_{k-1}\Big\}.
\end{aligned}$$

Furthermore,

$$\overline{\overline{E}}\Big[\frac{\Phi_{v_k}(D_k^{-1}(y_k-C_kx_k))}{|D_k|\Phi_{v_k}(y_k)}\Big|\mathcal{G}_{k-1},x_k\Big]=\frac{1}{|D_k|}\int_{\Re^d}\frac{\Phi_{v_k}(D_k^{-1}(y_k-C_kx_k))}{\Phi_{v_k}(y_k)}\Phi_{v_k}(y_k)dy_k=1.$$

Therefore,

$$\begin{aligned}
\overline{\overline{E}}\Big[\lambda_k\Big|\mathcal{G}_{k-1}\Big] &= \frac{1}{|B_k|}\int_{\Re^n}\frac{\Psi_{w_k}(B_k^{-1}(x_k-A_kx_{k-1}))}{\Psi_{w_k}(x_k)}\Psi_{w_k}(x_k)dx_k\\
&= \frac{1}{|B_k|}\int_{\Re^n}\Psi_{w_k}(B_k^{-1}(x_k-A_kx_{k-1}))dx_k=1.
\end{aligned}$$

Consequently,

$$\begin{aligned}
E\Big[g(w_k)f(v_k)\Big|\mathcal{G}_{k-1}\Big] &= \overline{\overline{E}}\Big[\lambda_k\ g(w_k)f(v_k)\Big|\mathcal{G}_{k-1}\Big]\\
&= \overline{\overline{E}}\Big[\frac{\Psi_{w_k}(B_k^{-1}(x_k-A_kx_{k-1}))}{|B_k|\Psi_{w_k}(x_k)}\ \frac{\Phi_{v_k}(D_k^{-1}(y_k-C_kx_k))}{|D_k|\Phi_{v_k}(y_k)}\\
&\quad\times g(B_k^{-1}(x_k-A_kx_{k-1}))f(D_k^{-1}(y_k-C_kx_k))\Big|\mathcal{G}_{k-1}\Big]\\
&= \overline{\overline{E}}\Big\{\overline{\overline{E}}\Big[\frac{\Psi_{w_k}(B_k^{-1}(x_k-A_kx_{k-1}))}{|B_k|\Psi_{w_k}(x_k)}\ \frac{\Phi_{v_k}(D_k^{-1}(y_k-C_kx_k))}{|D_k|\Phi_{v_k}(y_k)}\\
&\quad\times g(B_k^{-1}(x_k-A_kx_{k-1}))f(D_k^{-1}(y_k-C_kx_k))\Big|\mathcal{G}_{k-1},x_k\Big]\Big|\mathcal{G}_{k-1}\Big\}\\
&= \overline{\overline{E}}\Big\{\frac{\Psi_{w_k}(B_k^{-1}(x_k-A_kx_{k-1}))}{|B_k|\Psi_{w_k}(x_k)}\ g(B_k^{-1}(x_k-A_kx_{k-1}))\\
&\quad\times\overline{\overline{E}}\Big[\frac{\Phi_{v_k}(D_k^{-1}(y_k-C_kx_k))}{|D_k|\Phi_{v_k}(y_k)}\ f(D_k^{-1}(y_k-C_kx_k))\Big|\mathcal{G}_{k-1},x_k\Big]\Big|\mathcal{G}_{k-1}\Big\}.
\end{aligned}$$

Isolating

$$\begin{aligned}
&\overline{\overline{E}}\left[\frac{\Phi_{v_k}(D_k^{-1}(y_k - C_k x_k))}{|D_k|\Phi_{v_k}(y_k)}\ f(D_k^{-1}(y_k - C_k x_k))\middle|\mathcal{G}_{k-1}, x_k\right] \\
&= \frac{1}{|D_k|}\int_{\Re^d} \frac{\Phi_{v_k}(D_k^{-1}(y_k - C_k x_k))}{\Phi_{v_k}(y_k)}\ f(D_k^{-1}(y_k - C_k x_k))\ \Phi_{v_k}(y_k) dy_k \\
&= \frac{1}{|D_k|}\int_{\Re^d} \Phi_{v_k}(v_k)\ f(v_k)\ dv_k |D_k| \quad \left[dy_k = |D_k| dv_k\right]
\end{aligned}$$

which is independent of $x_0, x_1, \ldots, x_{k-1}, y_0, \ldots, y_{k-1}$ e.g., it is $\mathcal{G}_{k-1}$ independent.

Therefore,

$$E[g(w_k)f(v_k)\Big|\mathcal{G}_{k-1}] = \int_{\Re^d} \Phi_{v_k}(v) f(v) dv\ \times\ \int_{\Re^n} \Psi_{w_k}(w) g(w) dw.$$

□

D. Change of Drift and Signal

The following two systems are considered

$$\left(\Omega, \mathcal{F}, P, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\right) : \begin{cases} x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}, & x_0 \in \Re^n \\ \\ y_k = C_k x_k + D_k v_k, & y_0 \in \Re^d \end{cases} \tag{2.30}$$

$$\left(\Omega, \mathcal{F}, \tilde{P}, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\right) : \begin{cases} x_{k+1} = \tilde{A}_{k+1}x_k + \tilde{B}_{k+1}w_{k+1}, & x_0 \in \Re^n \\ \\ y_k = \tilde{C}_k x_k + \tilde{D}_k v_k, & y_0 \in \Re^d. \end{cases} \tag{2.31}$$

Next, it will be shown how to define system (2.30) from (2.31), which is equivalent to determining $\frac{dP}{d\tilde{P}}\Big|_{\mathcal{G}_k}$.

Clearly,

$$\frac{dP}{d\tilde{P}}\Big|_{\mathcal{G}_k} = \frac{dP}{d\overline{P}}\ \frac{d\overline{P}}{d\tilde{P}}\Big|_{\mathcal{G}_k},$$

where $\overline{P}$ is a measure under which $\{x_k\}$, $\{y_k\}$ are iid and normal.

Thus,

$$\frac{dP}{d\overline{P}}\Big|_{\mathcal{G}_k} = \Lambda_k = \prod_{s=0}^{k} \lambda_s,$$

where

$$\lambda_0 = \frac{\Phi_{v_0}(D_0^{-1}(y_0 - C_0 x_0))}{|D_0|\Phi_{v_0}(y_0)}$$

and

$$\lambda_k = \frac{\Phi_{v_k}(D_k^{-1}(y_k - C_k x_k))}{|D_k|\Phi_{v_k}(y_k)} \times \frac{\Psi_{w_k}(B_k^{-1}(x_k - A_k x_{k-1}))}{|B_k|\Psi_{w_k}(x_k)}, \quad k \geq 1.$$

Also,

$$\frac{d\overline{P}}{d\tilde{P}}\Big|_{\mathcal{G}_k} = \tilde{\Lambda}_k = \prod_{s=0}^{k} \tilde{\lambda}_s,$$

where

$$\tilde{\lambda}_0 = \frac{|\tilde{D}_0|\Phi_{v_0}(y_0)}{\Phi_{v_0}(\tilde{D}_0^{-1}(y_0 - \tilde{C}_0 x_0))}$$

and

$$\tilde{\lambda}_k = \frac{|\tilde{D}_k|\Phi_{v_k}(y_k)}{\Phi_{v_k}(\tilde{D}_k^{-1}(y_k - \tilde{C}_k x_k))} \times \frac{|\tilde{B}_k|\Psi_{w_k}(x_k)}{\Psi_{w_k}(\tilde{B}_k^{-1}(x_k - \tilde{A}_k x_{k-1}))}.$$

Therefore,

$$\frac{dP}{d\tilde{P}}\Big|_{\mathcal{G}_k} = \prod_{s=0}^{k} \frac{|\tilde{D}_s|}{|D_s|} \frac{\Phi_{v_s}(D_s^{-1}(y_s - C_s x_s))}{\Phi_{v_s}(\tilde{D}_s^{-1}(y_s - \tilde{C}_s x_s))} \frac{|\tilde{B}_s|}{|B_s|} \frac{\Psi_{w_s}(B_s^{-1}(x_s - A_s x_{s-1}))}{\Psi_{w_s}(\tilde{B}_s^{-1}(x_s - \tilde{A}_s x_{s-1}))}. \tag{2.32}$$

This is the complete data likelihood function between models (2.30) and (2.31); it is conplete because it is a function of $\mathcal{G}_k = \{x_0, \ldots, x_k, y_0, \ldots, y_k\}$.

2.2.3 Change of Probability Measure for Nonlinear Systems

A complete probability space $(\Omega, \mathcal{F}, P)$, is considered, on which $\{x_k\}$, $\{y_k\}$, $k \in N_0$ are defined by

$$\begin{aligned} x_{k+1} &= f(x_k, w_{k+1}) \quad x_0 \in \Re^n \\ y_k &= h(x_k, v_k), \quad y_0 \in \Re^d \end{aligned} \tag{2.33}$$

where the following conditions hold:

(i) $x : \Omega \times N_0 \to \Re^n$, $y : \Omega \times N_0 \to \Re^d$;

(ii) $w : \Omega \times N_0 \to \Re^n$ is an indep. seq. of random variables with density $\Psi_{w_k}(w)$;

(iii) $v : \Omega \times N_0 \to \Re^d$ is an indep. seq. of random variables with density $\Phi_{v_k}(v) > 0$;

(iv) $x_0 : \Omega \to \Re^n$ has a density $\Pi_0(x)$;

(v) $f : \Re^n \times \Re^n \to \Re^n$, $h : \Re^n \times \Re^d \to \Re^d$ are Borel measurable;

(vi) $\exists$ an inverse $D : \Re^n \times \Re^n \to \Re^n$ such that $w_{k+1} = D(x_{k+1}, x_k)$

(vii) $\exists$ an inverse $G : \Re^d \times \Re^n \to \Re^n$ such that $v_k = G(y_k, x_k)$;

(viii) the derivatives $\frac{\partial}{\partial y} G(y, x_e)\Big|_{y=y_e}$, $\frac{\partial}{\partial v} h(x_e, v)\Big|_{v=v_e}$ are continuous and nonsingular.

Let

$$\begin{aligned} \mathcal{G}_k &\stackrel{\triangle}{=} \sigma\{x_0, x_1, \ldots, x_k, y_0, \ldots, y_k\} \\ \mathcal{F}_k^{\mathcal{Y}} &\stackrel{\triangle}{=} \sigma\{y_0, \ldots, y_k\} \end{aligned}$$

be complete σ-algebras.

Define

$$\Lambda_k^{-1} = \prod_{s=0}^{k} \frac{\Phi_{v_s}(y_s)}{\Phi_{v_s}(v_s)} \Big| \frac{\partial}{\partial y} G(y, x_s) \Big|_{y=y_s}^{-1}.$$

Then

$$E\Big[\Lambda_k^{-1}\Big] = E\Big[E\Big[\Lambda_k^{-1}\Big|\mathcal{G}_{k-1}\Big]\Big] = E\Big[\Lambda_k^{-1} E\Big[\frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)} \Big| \frac{\partial}{\partial y} G(y, x_k)\Big|_{y=y_k}^{-1} \Big| \mathcal{G}_{k-1}\Big]\Big].$$

But

$$\begin{aligned} E\Big[\frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)} \Big| \frac{\partial}{\partial y} G(y, x_k)\Big|_{y=y_k}^{-1} \Big| \mathcal{G}_{k-1}\Big] &= \int_{\Re^d} \frac{\Phi_{v_k}(y_k)}{\Phi_{v_k}(v_k)} \Big| \frac{\partial}{\partial y} G(y, x_k)\Big|_{y=y_k}^{-1} \Phi_{v_k}(v_k) dv_k \\ &= \int_{\Re^d} \Phi_{v_k}(y_k) dy_k = 1 \end{aligned}$$

$$\begin{aligned} \Rightarrow E\Big[\Lambda_k^{-1}\Big] &= E\Big[\Lambda_{k-1}^{-1}\Big], \ \ \forall k \in N_0 \\ \Rightarrow E\Big[\Lambda_k^{-1}\Big] &= E\Big[\Lambda_0^{-1}\Big] = \int_{\Re^d} \frac{\Phi_{v_0}(y_0)}{\Phi_{v_0}(v_0)} \Big| \frac{\partial}{\partial y} G(y, x_0)\Big|_{y=y_0}^{-1} \Phi_{v_0}(v_0) dv_0 = 1. \end{aligned}$$

Therefore, the new measure can be defined by letting

$$\frac{d\overline{P}}{dP}\Big|_{\mathcal{G}_k} = \Lambda_k^{-1}.$$

Such that under measure $\overline{P}$, the seq. $\{y_k\}$ is independent with density $\phi_{y_k}(y_k)$.

Also, starting with $\overline{P}$, P can be constructed by setting

$$\begin{aligned}\Lambda_k &= \prod_{s=0}^{k} \frac{\Phi_{v_s}(v_s)}{\Phi_{v_s}(y_s)} \Big| \frac{\partial}{\partial v} h(x_s, v) \Big|_{v=v_s}^{-1} \\ v_s &\stackrel{\triangle}{=} G(y_s, x_s)\end{aligned}$$

and defining

$$\frac{dP}{d\overline{P}}\Big|_{\mathcal{G}_k} \stackrel{\triangle}{=} \Lambda_k.$$

Lemma 2.2.11. *The following system is considered,*

$$\Big(\Omega, \mathcal{F}, \overline{P}, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\Big) : \begin{cases} x_{k+1} = f(x_k, w_{k+1}), \;\; x_0 \in \Re^n \\ \\ \{y_k\} \textit{ is an indep. seq. with density } \Phi_{y_k}(y_k). \end{cases}$$

Then under measure P, the seq. $\{v_k\}$ is independent having densities $\{\Phi_{v_k}(\cdot)\}$ and

$$\begin{aligned} v_k &\stackrel{\triangle}{=} G(y_k, x_k) \\ \Rightarrow y_k &= h(x_k, v_k). \end{aligned}$$

Proof. Starting with

$$\begin{aligned} P\Big(v_k \le s \Big| \mathcal{G}_{k-1}\Big) &= E\Big[I_{\{\omega: v_k(\omega) \le s\}} \Big| \mathcal{G}_{k-1}\Big] \\ &= \frac{\overline{E}\Big[I_{\{\omega: x_k(\omega) \le s\}} \Lambda_k \Big| \mathcal{G}_{k-1}\Big]}{\overline{E}\Big[\Lambda_k \Big| \mathcal{G}_{k-1}\Big]} \\ &= \overline{E}\Big[\lambda_k I_{\{\omega: x_k(\omega) \le s\}} \Big| \mathcal{G}_{k-1}\Big] \\ &= \int_{\Re^d} I_{\{\omega: x_k(\omega) \le s\}} \frac{\Phi_{v_k}(v_k)}{\Phi_{v_k}(y_k)} \Big| \frac{\partial h}{\partial v} \Big|^{-1} \Phi_{v_k}(y_k) dy \\ &= P\Big(v_k \le s\Big) \end{aligned}$$

$\Rightarrow$ under P, $\{v_k\}$ is an indep. seq. with density Φ_{v_k}, $v_k \stackrel{\triangle}{=} G(y_k, x_k)$. □

A. Relation Between RND's and Sample Densities of RP's

The following system is considered

$$\Big(\Omega,\mathcal{F},P,\{\mathcal{G}_k\},\{\mathcal{F}_k^{\mathcal{Y}}\}\Big):\begin{cases} x_{k+1}=f(x_k)+g(x_k)w_{k+1}, & x_0\in\Re^n \\ \\ y_k=h(x_k)+D_kv_k, & y_0\in\Re^d \end{cases}$$

with the assumptions that $\{w_k\}$, $\{v_k\}$ are iid seq. with densities

$$\begin{aligned}\Phi_{w_k}(w) &= \frac{1}{(2\pi)^{\frac{n}{2}}}\,e^{-\frac{w^Tw}{2}},\\ \Phi_{v_k}(v) &= \frac{1}{(2\pi)^{\frac{d}{2}}}\,e^{-\frac{v^Tv}{2}}.\end{aligned}$$

If someone is interested in the density of the sample path $\{y_0,\ldots,y_k\}$ given the data $\{x_0,\ldots,x_k\}$.

Then

$$\begin{aligned}P\Big(y_0\le y\Big|x_0\Big) &= P\Big(h(x_0)+D_0v_0\le y\Big|x_0\Big)\\ &= P\Big(v_0\le D_0^{-1}(y-h(x_0)\Big|x_0\Big)\\ &= P\Big(v_0\le D_0^{-1}(y-h(x_0))\Big)\\ &= \int_{-\infty}^{D_0^{-1}(y-h(x_0))}\Phi_{v_0}(v_0)dv\\ \Rightarrow \frac{d}{dy}P\Big(y_0\le y\Big|x_0\Big) &= \frac{d}{dy}\int_{-\infty}^{D_0^{-1}(y-h(x_0))}\Phi_{v_0}(v_0)dv\\ &= \Phi_{v_0}\Big(D_0^{-1}(y-h(x_0))\Big)|D_0|^{-1}\end{aligned}$$

e.g.,

$$p(y_0|x_0)=\frac{d}{dy_0}P\Big(y_0\Big|x_0\Big)=|D_0|^{-1}\Phi_{v_0}\Big(D_0^{-1}(y_0-h(x_0))\Big).$$

Similarly,

$$p\Big(y_0,\ldots,y_k\Big|x_0,\ldots,x_k\Big)=\frac{p\Big(y_0,\ldots,y_k,x_0,\ldots,x_k\Big)}{p\Big(x_0,\ldots,x_k\Big)}$$

$$=\frac{\prod_{s=0}^k p\Big(y_s|y_0,\ldots,y_{s-1},x_0,\ldots,x_s\Big)p\Big(x_s|x_0,\ldots,x_{s-1},y_0,\ldots,y_{s-1}\Big)}{\prod_{s=0}^k p\Big(x_s|x_0,\ldots,x_{s-1}\Big)}.$$

Given that x_s is independent of $\{y_0, \ldots, y_{s-1}\}$ and $\{x_0, \ldots, x_{s-2}\}$ given x_{s-1}, and y_s is independent of $\{y_0, \ldots, y_{s-1}\}$ and $\{x_0, \ldots, x_{s-1}\}$ given x_s, then

$$\begin{aligned} p\Big(y_0, \ldots, y_k \Big| x_0, \ldots, x_k\Big) &= \frac{\prod_{s=0}^k p\Big(y_s|x_s\Big) p\Big(x_s|x_{s-1}\Big)}{\prod_{s=0}^k p\Big(x_s|x_{s-1}\Big)} \\ &= \prod_{s=0}^k P\Big(y_s|x_s\Big) \\ \Rightarrow p\Big(y_0, \ldots, y_k \Big| x_0, \ldots, x_k\Big) &= \prod_{s=0}^k |D_s|^{-1} \Phi_{v_s}\Big(D_s^{-1}(y_s - h(x_s))\Big). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{dP}{d\overline{P}}\Big|_{\mathcal{G}_k} = \Lambda_k &= \frac{p\Big(y_0, \ldots, y_k | x_0, \ldots, x_k\Big)}{p_v\Big(y_0, \ldots, y_k\Big)} \\ &= \prod_{s=0}^k \frac{|D_s|^{-1} \Phi_{v_s}\Big(D_s^{-1}(y_s - h(x_s))\Big)}{\Phi_{v_s}(y_s)} \end{aligned}$$

defines a measure P, starting with measure $\overline{P}$ under which $\{y_k\}$ is an iid seq. with density $\Phi_{v_k}(v)$.

Similarly, suppose the sample path data $\{y_0, \ldots, y_k, x_0, \ldots, x_k\}$ are given. Then its sample path density is

$$p\Big(y_0, \ldots, y_k, x_0, \ldots, x_k\Big) = p\Big(y_0, \ldots, y_k \Big| x_0, \ldots, x_k\Big) P\Big(x_0, \ldots, x_k\Big).$$

But

$$\begin{aligned} p\Big(x_0, \ldots, x_k\Big) &= p\Big(x_k \Big| x_0, \ldots, x_{k-1}\Big) p\Big(x_0, \ldots, x_{k-1}\Big) \\ &= p(x_0) \prod_{s=1}^k p\Big(x_s \Big| x_{s-1}\Big), \end{aligned}$$

$$\begin{aligned} P\Big(x_k \le x \Big| x_{k-1}\Big) &= P\Big(f(x_{k-1}) + g(x_{k-1}) w_k \le x \Big| x_{k-1}\Big) \\ &= P\Big(w_k \le g^{-1}(x_{k-1})(x - f(x_{k-1})) \Big| x_{k-1}\Big) \\ &= \int_{-\infty}^{g^{-1}(x_{k-1})(x - f(x_{k-1}))} \Phi_{w_k}(w) dw. \end{aligned}$$

Let $\frac{d}{dx}P\Big(x_k \le x\Big|x_{k-1}\Big) = p(x_k|x_{k-1})$, then

$$\begin{aligned} p\Big(x_k\Big|x_{k-1}\Big) &= |g(x_{k-1})|^{-1}\Phi_{w_k}\Big(g^{-1}(x_{k-1})(x - f(x_{k-1}))\Big) \\ \Rightarrow p\Big(x_0,\ldots,x_k\Big) &= p(x_0)\prod_{s=1}^{k}|g(x_{s-1})|^{-1}\Phi_{w_s}\Big(g^{-1}(x_{s-1})(x_s - f(x_{s-1}))\Big). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{dP}{d\widehat{P}}\Big|_{\mathcal{F}_k^x} &= M_k = \frac{p\Big(x_0,\ldots,x_k\Big)}{p_{w_k}\Big(x_0,\ldots,x_k\Big)} \\ &= \prod_{s=1}^{k}\frac{|g(x_{s-1})|^{-1}\Phi_{w_s}\Big(g^{-1}(x_{s-1})(x_s - f(x_{s-1}))\Big)}{\Psi_{w_s}(x_s)} \end{aligned}$$

defines measure P, starting with measure $\widehat{P}$ under which $\{x_k\}$ is an iid seq with density $\{\Phi_{w_k}(x_k)\}$.

Finally, the sample density of $\mathcal{G}_k$ given by

$$\begin{aligned} \frac{p\Big(y_0,\ldots,y_k,x_0,\ldots,x_k\Big)}{p_{w_k}\Big(x_0,\ldots,x_k\Big)p_{v_k}\Big(y_0,\ldots,y_k\Big)} &= \prod_{s=0}^{k}\frac{|g(x_{s-1})|^{-1}\Phi_{w_s}\Big(g^{-1}(x_{s-1})(x_s - f(x_{s-1}))\Big)}{\Psi_{w_s}(x_s)} \\ &\quad\times\frac{|D_s|^{-1}\Phi_{v_s}\Big(D_s^{-1}(y_s - h(x_s))\Big)}{\Phi_{v_s}(y_s)} \end{aligned}$$

defines a change of measure from $\overline{\overline{P}}$ under which $\{y_k\}$, $\{x_k\}$ are iid with densities $\Phi_{v_k}(\cdot)$, $\Psi_{w_k}(\cdot)$, respectively, to measure P.

2.2.4 Nonlinear Filtering Prediction and Smoothing

Here, the nonlinear filtering, smoothing and prediction are presented. It is assumed that there is a signal $\{x_k\}$, called the state of the system which is not directly observable. Rather only some noisy function $\{y_k\}$ of $\{x_k\}$ are observed, called the observation process. The objective is to obtain an expression for the "best estimate" of x_s (or $\Phi(x_s)$ for Φ in a certain class of functions), given the history of the observations $\{y_\tau : 0 \le \tau \le k\}$, that is, given the observable σ-field.

$$\mathcal{Y}_k = \sigma\{y_\tau : 0 \le \tau \le k\}.$$

Depending of the relation between the times $''s''$ and $''k''$ the following situations exist:

1. **The filtering problem.** Given the observation $\mathcal{Y}_k$ find the best estimate $\hat{x}_k$ of the state x_k from these observations.
2. **The smoothing problem.** Given the observation $\mathcal{Y}_k$ find the best estimate $\hat{x}_s$ of the state x_s, $s < k$ from these observations.
3. **The prediction problem.** Given the observation $\mathcal{Y}_k$ find the best estimate $\hat{x}_s$ of the state x_s, $s > k$ from these observations.

The computations should be done recursively, in terms of a static $\{\pi_k\}$ which can be updated using only new observations

$$\pi_{s+\tau} = \alpha(s, \tau, \pi_s, \{y_{k+u} : 0 \leq u \leq \tau\} \tag{2.34}$$

in which $s = k$ corresponds to the filtering problem, $s > k$ corresponds to the prediction problem, and $s < k$ corresponds to the smoothing problem. The statistic $\{\pi_k\}$ is then used to calculate, pointwise estimates $\hat{\Phi}(x_s)$ of functionals $\Phi(x_s)$ from the observations $\mathcal{Y}_k$:

$$\hat{\Phi}(x_s) = \beta(k, s, \Phi, y_k, \pi_s). \tag{2.35}$$

Suppose the processes $\{x_k\}$, $\{y_k\}$ are defined on a fixed probability space $\Big(\Omega, \mathcal{F}, P\Big)$ with filtration $\{\mathcal{F}_k\}$ and finite time $k \in N_0^m \stackrel{\triangle}{=} \{0, 1, \ldots, m\}$. The specification of the best estimate $\hat{x}_s$ of x_s from $\mathcal{Y}_k$ is usually done by minimizing a functional of the distance of x_s from the closed subspace generated by $\mathcal{Y}_k \stackrel{\triangle}{=} \{y_s : 0 \leq s \leq k\}$. From the general theory of Hilbert spaces it is known that $\hat{x}_s(\mathcal{Y}_k)$ is the orthogonal projection onto the subspace generated by $\mathcal{Y}_k$.

Theorem 2.2.12. *Let $\Big(\Omega, \mathcal{F}, P\Big)$ be a probability space with filtration $\{\mathcal{F}_k\}$, a sub-sigma algebra $\{\mathcal{F}_k^{\mathcal{Y}}\}$ of $\{\mathcal{F}_k\}$ and $X \in L^2\Big(\Omega, \mathcal{F}, P\Big)$.*

Define

$$M_{\mathcal{Y}_k} \stackrel{\triangle}{=} \{y(j, \cdot) : \Omega \to \Re^d : y(j, \cdot) \in L^2\Big(\Omega, \mathcal{F}, P\Big) \text{ and } y_j \text{ is } \mathcal{F}_j^{\mathcal{Y}} - \text{measurable}, j \in N_0^k\},$$

$$\mathcal{P}_{M_{\mathcal{Y}_k}} : L^2\Big(\Omega, \mathcal{F}, P\Big) \to M_{\mathcal{Y}_k}.$$

Then

$$\mathcal{P}_{M_{\mathcal{Y}_k}}(x_s) = E\Big[x_s\Big|M_{\mathcal{Y}_k}\Big] = E\Big[x_s\Big|\mathcal{F}_k^{\mathcal{Y}}\Big]$$

is P-a.s. unique, e.g.,

$$\int_A \mathcal{P}_{M_{\mathcal{Y}_k}}(x_s)(\omega)dP(\omega) = \int_A x_s(\omega)dP(\omega),\ \ \forall A \in \mathcal{F}_k^{\mathcal{Y}}.$$

Remark 2.2.13. *The minimum variance estimate $\mathcal{P}_{M_{y_k}}(x_s)$ is unbiased, that is, $E[\mathcal{P}_{M_{y_k}}(x_s)] = E[x_s],\ \ \forall k \in N_0$.*

Theorem 2.2.14. *(The minimum variance estimation) Given the observations $\mathcal{Y}_k$, the minimum variance estimate of the state x_s defined by*

$$\inf\Big\{E\Big[\|x_s - m_k\|^2 : m_k \in M_{\mathcal{Y}_k}\Big\} \tag{2.36}$$

is given by $\hat{x}_s(y_k) = E[x_s|\mathcal{F}_k^{\mathcal{Y}}]$.

Proof. Since

$$E\Big[\|x_s - m_k\|^2\Big] = E\Big[\|x_s - \hat{x}_s(\mathcal{Y}_k) + \hat{x}_s(\mathcal{Y}_k) - m_k\|^2\Big] \tag{2.37}$$

$$= E\Big[\|m_k - \hat{x}_s(\mathcal{Y}_k)\|^2\Big] + E\Big[||x_s - \hat{x}_s(\mathcal{Y}_k)\|^2\Big] + 2E\Big[\Big(x_s - \hat{x}_s(\mathcal{Y}_k), \hat{x}_s(\mathcal{Y}_k) - m_k\Big)\Big]. \tag{2.38}$$

Reconditioning on $\mathcal{F}_k^{\mathcal{Y}}$, the last term in (2.38) vanishes; since the second right-side term of (2.38) is fixed the left-side is minimized by setting $m_k = \hat{x}_s(\mathcal{Y}_k)$. Therefore, the conditional expectation $E[x_s|\mathcal{F}_s^{\mathcal{Y}}]$ gives the least minimum variance estimate. □

Definition 2.2.15. *(The risk-sensitive estimation) Given the observations $\mathcal{Y}_k$, the minimum risk-sensitive estimate $\hat{x}_s(\mathcal{Y}_k)$ of the state x_s is defined by*

$$\begin{aligned}&\inf\Big\{E\Big[\exp\Big(\theta\sum_{\tau=0}^{s}\|x_\tau - m_\tau\|^2 : m_k \in M_{[0,k]}, \theta \in \Re\Big\}\\ &= \Big\{E\Big[\exp\Big(\theta\mathcal{C}_{0,k}(x_0^s, m_0^s)\Big)\Big] : m_k \in M_{[0,k]}, \theta \in \Re\Big\}\end{aligned} \tag{2.39}$$

where

$$M_{[0,k]} = \Big\{y : N_0^k \times \Omega \to \Re^d : y(k,\omega) \in L^2((0,k)\times\Omega, \mathcal{F}, P),\ \ \textit{for almost all } k$$
$$y(k,.) \in L^2(\Omega, \mathcal{F}, P) \textit{ and } y_k \textit{ is } \mathcal{Y}_k - \textit{measurable}\Big\}. \tag{2.40}$$

The parameter θ is the so-called sensitive parameter which renders the optimizer optimistic if $\theta < 0$ and pessimistic if $\theta > 0$. In addition, for $\theta > 0$ the sample cost $\exp(\theta\mathcal{C})$ is a convex increase function of $\mathcal{C}$. In the former case the optimizer is concerned with designing under worst scenario, hence the name risk-averse, while in the latter case is concerned with designing under favorable conditions, hence the name risk-seeking. The case $\theta = 0$ lies between the risk-averse and the risk-seeking optimalities.

The following expansion is considered

$$E\Big[\exp(\theta\mathcal{C})\Big] = 1 + \theta E\Big[\mathcal{C}\Big] + \frac{1}{2}\theta^2 E\Big[\mathcal{C}^2\Big] + O(\theta^3). \tag{2.41}$$

Then

$$\theta \log E\Big[\exp(\theta\mathcal{C})\Big] = E\Big[\mathcal{C}\Big] + \frac{1}{2}\theta Var\Big(\mathcal{C}\Big) + O(\theta^2). \tag{2.42}$$

Consequently, in the limit, as $\theta \to 0$, the risk-sensitive estimation problem includes as a special case the risk-neutral case.

Definition 2.2.16. *Let $\Big(\Omega, \mathcal{F}, P, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\}\Big)$ is a complete probability space on which the state process $\{x_k\}, k \in N_0$ and the observation process $\{y_k\}, k \in N_0$, are defined by the following recursions.*

$$\begin{aligned} x_{k+1} &= f(k+1, x_k) + g(k+1, x_k)w_{k+1}, \;\; x_0 \in \Re^n, \\ y_k &= h(k, x_k) + D_k v_k \;\; y_0 \in \Re^d, \end{aligned} \tag{2.43}$$

in which the condition of (2.33) hold, and additionally

$$\begin{aligned} \Phi_k(v) &\stackrel{\triangle}{=} \Phi_{v_k}(v) = \frac{1}{(2\pi)^{d/2}}\, e^{v^T v/2}, \\ \Psi_k(w) &\stackrel{\triangle}{=} \Psi_{w_k}(w) = \frac{1}{(2\pi)^{n/2}}\, e^{w^T w/2}, \\ P(x_0 \le x) &= \int_{-\infty}^{x} \Pi_{x_0}(z)dz. \end{aligned}$$

Clearly, the minimum variance estimate of any Borel bounded function $\Phi : \Re^n \to \Re$ is given by

$$\Pi_k(\Phi) = E\Big[\Phi(x_k)\Big|\mathcal{F}_k^{\mathcal{Y}}\Big]. \tag{2.44}$$

Thus, a program will be introduced which will enable the computation of the conditional distribution of x_k given $\mathcal{F}_k^{\mathcal{Y}}$, namely

$$P(x_k \in A|\mathcal{F}_k^{\mathcal{Y}}), \;\; \forall A \in \mathcal{B}(\Re^n). \tag{2.45}$$

Definition 2.2.17. *For any given Borel measurable function which is bounded,* $\Phi : N_0 \times \Re^n \to \Re$*, the measure-value process* $\Pi_k(\Phi)$ *is related to another measure-valued process through*

$$\Pi_k(\Phi) = \frac{\overline{E}\Big[\Phi(k, x_k)\Lambda_k \Big| \mathcal{F}_k^{\mathcal{Y}}\Big]}{\overline{E}\Big[\Lambda_k \Big| \mathcal{F}_k^{\mathcal{Y}}\Big]} \stackrel{\nabla}{=} \frac{\overline{\Pi}_k(\Phi)}{\overline{\Pi}_k(1)}, \quad P - a.s.$$

where

$$\frac{dP}{d\overline{P}}\Big|_{\mathcal{G}_k} = \Lambda_k.$$

Let $BC(\Re^n) \stackrel{\triangle}{=} \{\Phi : \Re^n \to \Re : \Phi \text{ is bounded and continuous}\}$, equipped with the norm topology $\|\Phi\|_\infty \stackrel{\triangle}{=} \sup\{|\Phi| : x \in \Re^n, \Phi \in BC(\Re^n)\}$. Let $\mathcal{M}_+(\Re^n) \stackrel{\triangle}{=} \{$ denotes the set of positive measures on $(\Re^n, \mathcal{B}(\Re^n)\}$, equipped with the norm topology $\|\mu\| = Var_A\mu = \sup_{A_i \subset A \in \mathcal{B}(\Re^n)} \sum_{i=1}^n \mu(A_i)$, when the upper bound is taken over all finite collections of pairwise disjoint sets $\{A_i\}_{i=1}^n \in \mathcal{B}(\Re^n)$ which are contained in A. It is noted that $Var_A\mu = \mu^+(A) + \mu^-(A), \ \ \forall A \in \mathcal{B}(\Re^n)$. Consequently, $Var_A\mu$ is a measure with respect to A, which is denoted by $|\mu|(A) = Var_A\mu$.

If μ is a countably additive function on the σ-field $\mathcal{B}(\Re^n)$, then μ is finite if and only if:

a. μ^+, μ^- are finite,

b. $|\mu(A)| < \infty$,

c. $Var_A\mu < \infty$.

From this, it follows that if $|\mu|(A) < \infty$ for every $\mathcal{B}(\Re^n)$ measurable function Φ which is bounded on A, and is μ-measurable on A, e.g.,

$$\int_A \Phi(x)d\mu(x) = \int_A \Phi(x)d\mu^+(x) - \int_A \Phi(x)d\mu^-(x)$$

and

$$\Big|\int_A \Phi(x)d\mu(x)\Big| \leq \sup_{x \in A} |\Phi(x)||\mu|(A).$$

Thus, for any $\mu \in \mathcal{M}_+(\Re^n), \Phi \in BC(\Re^n)$ the inner product is defined by

$$(\Phi, \mu) = \int_{\Re^n} \Phi(x)d\mu(x).$$

Remark 2.2.18. *The Bayes formula of Definition (2.2.17) is the starting point of non-linear minimum variance filtering. The process*

$$\pi_k(\Phi) \stackrel{\triangle}{=} \overline{E}\Big[\Phi(x_k)\Lambda_k\Big|\mathcal{F}_k^{\mathcal{Y}}\Big]$$

is the unnormalized measure-valued process of $\Pi_k(\Phi)$ *since*

$$\Pi_k(\Phi) = \frac{\pi_k(\Phi)}{\pi_k(1)} = \frac{\int_{\Re^n}\Phi(z)d\pi_k(z)}{\int_{\Re^n}1d\pi_k(z)} = \frac{(\Phi,\pi)}{(1,\pi)}.$$

Further if the measure $\pi_k(.)$ *is absolutely continues with respect to the Lebeggue measure, then* $\frac{d}{dx}\pi_k(x)$ *exist and*

$$\frac{d}{dx}\Pi_k(x) = \frac{\pi_k(x)}{\int_{\Re^n}\pi_k(x)dx}.$$

Next a recursive equation for $\pi_k(x)$ is derived.

The system of Definition 2.2.16, is considered, starting with measure $\overline{P}$ under which $\{x_k\}$, $\{y_k\}$ are white noise sequences.

Let

$$\begin{aligned}\mathcal{G}_k &= \sigma\{x_0,\ldots,x_k,y_0,\ldots,y_k\},\quad k\in N_0\\ \mathcal{F}_k^{\mathcal{Y}} &= \sigma\{y_0,\ldots,y_k\},\quad k\in N_0\end{aligned}$$

where $\{\mathcal{G}_k\}$, $\{\mathcal{F}_k^{\mathcal{Y}}\}$, $k\in N_0$ are their corresponding filtrations.

Let

$$\begin{aligned}\lambda_0 &= \frac{\Phi_{v_0}\Big(D_0^{-1}(y_0-h(0,x_0))\Big)}{|D_0|\Phi_{v_0}(y_0)},\\ \lambda_s &= \frac{\Phi_{v_s}\Big(D_s^{-1}(y_s-h(s,x_s))\Big)}{|D_s|\Phi_{v_s}(y_s)}\frac{\Psi_{w_s}\Big(g^{-1}(s,x_{s-1})(x_s-f(s,x_{s-1}))\Big)}{|g(s,x_{s-1})|\Psi_{w_s}(x_s)},\quad s\in N_0.\end{aligned}$$

Next, a new probability measure P is defined on $\Big(\Omega,\mathcal{F},\{\mathcal{G}_k\}\Big)$ by introducing the RND

$$\frac{dP}{d\overline{P}}\Big|_{\mathcal{G}_k} = \Lambda_k = \prod_{s=0}^{k}\lambda_s.$$

Define

$$\begin{aligned} w_s &\stackrel{\triangle}{=} g^{-1}(s, x_{s-1})(x_s - f(s, x_{s-1})), \quad s = 1, 2, \dots \\ v_s &\stackrel{\triangle}{=} D_s^{-1}(y_s - h(s, x_s)), \quad s = 0, 1, \dots \end{aligned}$$

Then under measure P, the sequences $\{v_k\}$, $\{w_k\}$ are independent, $N(0; I_d)$, $N(0; I_n)$, respectively.

Theorem 2.2.19. *(Recursive Equation) Consider* $\Big(\Omega, \mathcal{F}, P, \{\mathcal{G}_k\}, \{\mathcal{F}_k^{\mathcal{Y}}\Big\})$ *a complete probability space on which* $\{x_k, y_k\}, \quad k \in N_0$ *are defined by Definition 2.2.16.*

Then for any $\Phi \in BC(\Re^n)$ *the measure valued process* $\{\pi_k(\Phi)\}, k \in N_0$ *satisfies the following recursion*

$$\pi_k(\Phi) = \int_{\Re^n} \Phi(x) \frac{\Phi_k\Big(D_k^{-1}(y_k - h(k, x))\Big)}{|D_k|\Phi_k(y_k)} \pi_{k-1}\Big(\frac{\Psi_k\Big(g^{-1}(k, .)(x - f(k, \cdot))\Big)}{|g(k, \cdot)|}\Big) dx \quad (2.46)$$

with initial condition

$$\pi_0(\Phi) = \int_{\Re^n} \Phi(x) \frac{\Phi_0\Big(D_0^{-1}(y_0 - h(0, x))\Big)}{|D_0|\Phi_0(y_0)} \Pi_{x_0}(x) dx. \quad (2.47)$$

Further, if $\pi_k(\Phi)$ *has a density*

$$\pi_k(\Phi) = \int_{\Re^n} \Phi(x) d\pi_k(x) = \int_{\Re^n} \phi(x) \pi_k(x) dx$$

then $\{\pi_k(x)\}, \quad k \in N_0$ *satisfies the recursion*

$$\pi_k(x) = \frac{\Phi_k\Big(D_k^{-1}(y_k - h(k, x))\Big)}{|D_k|\Phi_k(y_k)} \int_{\Re^n} \frac{\Psi_k\Big(g^{-1}(k, z)(x - f(k, z))\Big)}{|g(k, z)|} \pi_{k-1}(z) dz \quad (2.48)$$

with initial condition

$$\pi_0(x) = \frac{1}{|D_0|\Phi_0(y_0)} \Phi_0\Big(D_0^{-1}(y_0 - h(0, x))\Big) \Pi_{x_0}(x). \quad (2.49)$$

Proof. By definition

$$\begin{aligned} \pi_k(\Phi) &= \overline{E}\Big[\Phi(x_k)\Lambda_k \Big| \mathcal{F}_k^{\mathcal{Y}}\Big] \\ &= \int_{\Re^n} \Phi(x) d\pi_k(x). \end{aligned} \quad (2.50)$$

Further,

$$\pi_k(\Phi) = \overline{E}\Big[\Phi(x_k)\Lambda_k\Big|\mathcal{F}_k^{\mathcal{Y}}\Big] = \overline{E}\Big[\prod_{s=0}^{k}\Phi(x_s)\Lambda_s\Big|\mathcal{F}_k^{\mathcal{Y}}\Big] = E^Q\Big[\prod_{s=0}^{k}\Phi(x_s)\Lambda_s\Big]$$

where Q is the measure induced by the independent sequence $\{x_s\}$ (e.g., $\{x_s\}$ and $\{y_s\}$ are independent.

$$\pi_k(\Phi) = E^Q\Big[\Lambda_{k-1}\lambda_k\Phi(x_k)\Big]$$

$$\begin{aligned}
&= E^Q\Bigg[\Lambda_{k-1}\frac{\Phi_k(D_k^{-1}(y_k - h(k,x_k)))}{|D_k|\Phi_k(y_k)}\ \frac{\Psi_k(g^{-1}(k,x_{k-1})(x_k - f(k,x_{k-1})))}{|g(k,x_{k-1})|\Psi_k(x_k)}\Phi(x_k)\Bigg]\\
&= \frac{1}{|D_k|\Phi_k(y_k)}E^Q\Bigg[\Lambda_{k-1}\Phi_k(D_k^{-1}(y_k - h(k,x_k)))\\
&\quad\times\frac{\Psi_k(g^{-1}(k,x_{k-1})(x_k - f(k,x_{k-1})))}{|g(k,x_{k-1})|\Psi_k(x_k)}\Phi(x_k)\Bigg]\\
&= \frac{1}{|D_k|\Phi_k(y_k)}E^Q\Bigg[\frac{\Lambda_{k-1}}{|g(k,x_{k-1})|}\\
&\quad\times E^Q\Bigg[\frac{\Phi_k(D_k^{-1}(y_k - h(k,x_k)))\Psi_k(g^{-1}(k,x_{k-1})(x_k - f(k,x_{k-1})))}{\Psi_k(x_k)}\\
&\quad\Phi(x_k)\Big|\mathcal{F}_{k-1}^{\mathcal{Y}}, x_{k-1}\Bigg]\Bigg]\\
&= \frac{1}{|D_k|\Phi_k(y_k)}E^Q\Bigg[\frac{\Lambda_{k-1}}{|g(k,x_{k-1})|}\\
&\quad\times E^Q\Bigg[\frac{\Phi_k(D_k^{-1}(y_k - h(k,x_k)))\Psi_k(g^{-1}(k,x_{k-1})(x_k - f(k,x_{k-1})))}{\Psi_k(x_k)}\\
&\quad\Phi(x_k)\Bigg]\Bigg]\\
&= \frac{1}{|D_k|\Phi_k(y_k)}E^Q\Bigg[\frac{\Lambda_{k-1}}{|g(k,x_{k-1})|}\\
&\quad\times\int_{\Re^n}\frac{\Phi_k(D_k^{-1}(y_k - h(k,\xi)))\Psi_k(g^{-1}(k,x_{k-1})(\xi - f(k,x_{k-1})))}{\Psi_k(\xi)}\\
&\quad\Phi(\xi)\Psi_k(\xi)d\xi\Bigg]\\
&= \frac{1}{|D_k|\Phi_k(y_k)}E^Q\Bigg[\frac{\Lambda_{k-1}}{|g(t,x_{k-1})|}\tilde{g}(k,x_{k-1},y_k)\Bigg].
\end{aligned}$$

Define

$$\begin{aligned}
&\tilde{g}(k,x_{k-1},y_k) \triangleq\\
&\int_{\Re^n}\Phi_k(D_k^{-1}(y_k - h(k,\xi)))\Psi_{w_k}(g^{-1}(k,x_{k-1})(\xi - f(k,x_{k-1})))\Phi(\xi)d\xi.
\end{aligned}$$

Then

$$\pi_k(\Phi) = \frac{1}{|D_k|\Phi_k(y_k)} \int_{\Re^n} \frac{\tilde{g}(k,z,y_k)}{|g(k,z)|}\ \pi_{k-1}(z)dz.$$

Hence

$$\begin{aligned}\int_{\Re^n} \Phi(x)\pi_k(x)dx &= \frac{1}{|D_k|\Phi_{v_k}(y_k)} \int_{\Re^n} \frac{\tilde{g}(k,z,y_k)}{|g(k,z)|}\ \pi_{k-1}(z)dz \\ &= \frac{1}{|D_k|\Phi_{v_k}(y_k)} \\ &\quad \times \int_{\Re^n}\int_{\Re^n} \frac{\Phi_k(D_k^{-1}(y_k - h(k,x)))\Psi_w(g^{-1}(k,z)(x - f(k,z)))}{|g(k,z)|} \\ &\quad \times \Phi(x)dx\ \pi_{k-1}(z)dz\end{aligned}$$

$$\begin{aligned}&\int_{\Re^n} \Phi(x) \left[\pi_k(x) - \frac{\Phi_k(D_k^{-1}(y_k - h(k,x)))}{|D_k|\Phi_k(y_k)}\right. \\ &\left. \times \int_{\Re^n} \pi_{k-1}(z) \frac{\Psi_k(g^{-1}(k,z)(x - f(k,z)))}{|g(k,z)|} dz\right] dx = 0.\end{aligned}$$

Since $\Phi \in BC(\Re^n)$ is arbitrary, the only solution for this equation is

$$\pi_k(x) = \frac{\Phi_k(D_k^{-1}(y_k - h(k,x)))}{|D_k|\Phi_k(y_k)} \int_{\Re^n} \frac{\Psi_k(g^{-1}(k,z)(x - f(k,z)))}{|g(k,z)|}\ \pi_{k-1}(z)dz.$$

At $k = 0$,

$$\begin{aligned}\pi_0(\Phi) &= \overline{E}\left[\Lambda_0\Phi(x)|\mathcal{F}_0^{\mathcal{Y}}\right] = \overline{E}\left[\frac{\Phi_0(D_0^{-1}(y_0 - h(0,x)))}{|D_0|\Phi_0(y_0)}\Phi(x)\right] \\ &= \frac{1}{|D_0|\Phi_0(y_0)} \int_{\Re^n} \Phi_0(D_0^{-1}(y_0 - h(0,x)))\Phi(x)d\Pi_{x_0}(x).\end{aligned}$$

Hence,

$$\pi_0(x) = \frac{\Phi_0(D_0^{-1}(y_0 - h(0,x)))}{|D_0|\Phi_0(y_0)}\ \Pi_{x_0}(x).$$

□

Example 2.2.20. *(Linear Gaussian filter) It is assumed that the state and observations are given by*

$$\left(\Omega, \mathcal{F}, \{\mathcal{G}_k\}, P\right) : \begin{cases} x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}, \ \ x_k \in \Re^n \\ \\ y_k = C_k x_k + D_k v_k, \ \ y_k \in \Re^d \end{cases} \tag{2.51}$$

in which the noises are Gaussian distributed as follows, $w_k \sim N(0, I_n)$, $v_k \sim N(0, I_d)$ *while sequences* $\{w_k\}$, $\{v_k\}$ *are mutually independent and independent of* $x_0 \sim N(\overline{x}_0, V_0)$, $V_0 > 0$. *The Linearity and Gaussianity implies that*

$$\Pi_k(x) = \frac{\pi_k(x)}{\int_{re^n} \pi_k(x)dx} = \frac{\overline{\alpha}_k(x)}{\int_{re^n} \overline{\alpha}_k(x)dx}$$

is Gaussian with mean $\hat{x}_{k|k} = E\Big[x_k \Big| \mathcal{F}_k^{\mathcal{Y}}\Big]$ *and Variance* $V_{k|k} = E\Big[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T \Big| \mathcal{F}_k^{\mathcal{Y}}\Big]$.

Recursive equations for $\hat{x}_{k|k}$, $V_{k|k}$ *using the unnormalized conditional density, will be obtained.*

According to Theorem 2.2.19, $\{\overline{\alpha}_k(x)\} \stackrel{\triangle}{=} \{\pi_k(x)\}$ *satisfies the following recursion*

$$\overline{\alpha}_k(x) = \frac{\Xi_{v_k}(D_k^{-1}(y_k - C_k x))}{|D_k|\Xi_{v_k}(y_k)} \int_{\Re_n} \frac{\Psi_{w_k}(B_k^{-1}(x - A_k z))}{|B_k|} \overline{\alpha}_{k-1}(z)dz \tag{2.52}$$

where

$$\Psi_{w_k}(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big(\frac{-x^T x}{2}\Big), \tag{2.53}$$

$$\Xi_{v_k}(y) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\Big(-\frac{y^T y}{2}\Big). \tag{2.54}$$

A solution to (2.52) having the following form, is assumed

$$\overline{\alpha}_k(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|V_{k|k}|^{\frac{1}{2}}} \exp\Big(-(x - \hat{x}_{k|k})^T \frac{(V_{k|k})^{-1}}{2}(x - \hat{x}_{k|k}) + \beta_{k|k}\Big) \tag{2.55}$$

when $\{V_{k|k}, \hat{x}_{k|k}, \beta_{k|k}\}$ *will be identified shortly. Substituting (2.55) into the recursion (2.52) deduces the following recursive relations for* $\{V_{k|k}, \hat{x}_{k|k}, \beta_{k|k}\}$*:*

$$\hat{x}_{k|k} = V_{k|k}\Big[C_k^T (D_k D_k^T)^{-1} y_k + (V_{k|k-1})^{-1} \hat{x}_{k|k-1}\Big], \tag{2.56}$$

$$\hat{x}_{k|k-1} = A_k \hat{x}_{k-1|k-1}, \tag{2.57}$$

$$V_{k|k} = \Big(C_k^T (D_k D_k^T)^{-1} C_k + (V_{k|k-1})^{-1}\Big)^{-1}, \tag{2.58}$$

$$V_{k|k-1} = B_k B_k^T + A_k V_{k-1|k-1} A_k^T \tag{2.59}$$

$$\begin{aligned} &= (B_k B_k^T) + A_k V_{k-1|k-2} A_k^T \\ &\quad - A_k V_{k-1|k-2} C_{k-1}^T (D_{k-1} D_{k-1}^T + C_{k-1} V_{k-1|k-2} C_{k-1}^T)^{-1} C_{k-1} V_{k-1|k-2} A_k^T, \end{aligned} \tag{2.60}$$

$$\begin{aligned}\beta_{k|k} &= -\sum_{i=1}^{k}\frac{1}{2}(y_i - C_i\hat{x}_{i|i-1})^T(C_iV_{i|i-1}C_i^T + D_i^TD_i)^{-1}(y_i - C_i\hat{x}_{i|i-1}) + \sum_{i=1}^{k}\frac{y_i^Ty_i}{2} \\ &\quad -\frac{1}{2}\sum_{i=1}^{k}\log|C_iV_{i|i-1}C_i^T + D_i^TD_i|,\end{aligned} \tag{2.61}$$

with initial conditions

$$\hat{x}_{0|0} = V_{0|0}\Big[C_0^T(D_0D_0^T)^{-1}y_0\Big], \tag{2.62}$$

$$V_{0|0} = \Big(C_0^T(D_0D_0^T)^{-1}C_0\Big)^{-1}, \tag{2.63}$$

$$\begin{aligned}\beta_{0|0} &= -\frac{1}{2}(y_0 - C_0\overline{x}_0)^T(C_0V_0C_0^T + D_0^TD_0)^{-1}(y_0 - C_0\overline{x}_0) + \frac{y_0^Ty_0}{2} \\ &\quad -\frac{1}{2}\log|C_0V_0C_0^T + D_0^TD_0|,\end{aligned} \tag{2.64}$$

thus it is concluded that (2.55) is indeed a solution of the recursion (2.52).

Furthermore, $\hat{x}_{k|k}$ can be written as

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + V_{k|k-1}C_k^T\Big[C_kV_{k|k-1}C_k^T + (D_kD_k^T)\Big]^{-1}(y_k - C_k\hat{x}_{k|k-1}). \tag{2.65}$$

Also

$$\begin{aligned}\hat{x}_{k|k-1} &= A_k\hat{x}_{k-1|k-2} + A_kV_{k-1|k-2}C_{k-1}^T\Big[C_{k-1}V_{k-1|k-2}C_{k-1}^T + (D_{k-1}D_{k-1}^T)\Big]^{-1} \\ &\quad \times(y_{k-1} - C_{k-1}\hat{x}_{k-1|k-2}).\end{aligned} \tag{2.66}$$

This is the solution of the Kalman Filter [1]. See Appendix B for the derivation.

CHAPTER 3

ROBUST LEAST-SQUARE ESTIMATION FOR A CLASS OF SYSTEMS

This chapter considers least-square estimation problems for classes of systems. In Section 3.2 the estimation problem for a class of systems is formulated. A general framework is put forward in which the basic ideas and theory are explained and the fundamental results are derived. In Section 3.3 various examples of estimation theory are introduced, when the models (conditional distributions, joint distributions) are uncertain and they belong to specific subsets of the set of conditional or joint distributions, in order to illustrate the results derived. Finally, in Section 3.4 an overview on MIMO communication systems is presented and the theory developed in previous sections is applied to MIMO systems.

3.1 Introduction

In classical Least-Square estimation of Random Variables one is interested in finding the best estimate, $\Phi^*(Y)$, of a RV X from the measurements of a RV Y, by minimizing the expected value of the least-square error $e(X,\Phi) \stackrel{\triangle}{=} ||X-\Phi(Y)||^2_{\Re^n}$ over all functions $\Phi : \Re^d \rightarrow \Re^n$,$Y \rightarrow \Phi(Y)$, which is a function of Y ($||x||^2_{\Re^n}$ denotes Euclidean norm of $x \in \Re^n$). By the orthogonal projection theorem, the solution is given by the conditional expectation:

$$\Phi^*(Y) = E[X|Y] = \int_{\Re^n} x dP_{X|Y}(x|y) \tag{3.1}$$

where $P_{X|Y}$ is the conditional distribution of X given Y [1, 2].

The solution to this estimation problem is stated in terms of the $\acute{a}$ posteriori distribution function $P_{X|Y}$. This distribution contains all information available to estimate X or any nonlinear function of it. The objective is thus to find, recursively in time, the evaluation equation of the $\acute{a}$ posteriori distribution and to solve it. The classical result assumes that $P_{X,Y}$ and hence $P_{X|Y}$ are completely known. In many real applications of estimation theory, knowledge of $P_{X,Y}$ and/or $P_{X|Y}$ is not available; the only knowledge available to the designer is whether $P_{X,Y}$, (resp. $P_{X|Y}$) belong to specific classes which are subsets of the set of all joint (resp. conditional) distributions.

This chapter is concerned with estimation techniques, in which the uncertainty description of the system, and the nominal description of the system are modeled by probability distributions, or general measures, defined on measurable spaces. The uncertainty description of these systems is characterized by the class of uncertain measures which satisfy a Kullback-Leibler (KL) distance constraint with respect to a nominal measure. The problem of robust estimation is formulated by minimizing over the set of estimators, the maximum of a linear functional of the uncertain measure over the constraint set. Two type of uncertainty models are considered.

1. Uncertainty Models on Conditional Distributions or otherwise known Stochastic Kernels;

   i) When the conditional probability distribution of the measurement Y given the signal to be estimated X, or channel kernel, is unknown;

   ii) When the $\acute{a}$ posteriori distribution of X given Y is unknown;

2. Uncertainty Models on Joint Distributions.

Stochastic kernel uncertainty models are appropriate for communication system design, in which the input message has a known distribution, while the channel is unknown but belongs to a certain class of channels. These are nonparametric uncertainty models which so far have not been taken into consideration. Joint distribution uncertainty models are usually employed when both the unobserved and observed random variables are uncertain.

The minimax technique considered here leads to strategies in which the worst case estimate of the uncertain measure subject to the uncertainty description is sought. The theory and contribution of this chapter are developed at two levels of generality; the abstract level and the application level. At the abstract level a general framework

is put forward in which the basic ideas are explained and the fundamental results are derived. Specifically, the estimation problem is described on abstract Polish spaces, while the uncertain model considered is described by stochastic kernels and joint distributions. First, the appropriate space of measures is introduced and then the maximizing kernel and joint measure are computed explicitly using Lagrangian functionals and variational methods. Moreover, important monotonicity properties satisfied by the optimal strategies are presented, which can be used to developed numerical algorithms for computation of the optimal solution, and upper and lower bounds on the optimal solution. At the application level, the results obtained at the abstract level are applied to simple examples.

As was already discussed in Chapter 1 previous related work in which uncertainty is described by relative entropy can be found in [11], [16] and [29]. However, [29] deals only with systems when the uncertainty is defined on joint distributions. On the other hand, [16] similar to [29] deals only with joint distributions, but employs a minimax viewpoint and uses the additional assumption that the system considered is Gaussian. Even though [16] starts with a nonparametric uncertainty models it uses a parametric approach to derive its results. The least-squares error estimator (conditional mean) is considered in [16] and shown to yield a saddle point. A parametric uncertainty model is implemented in [11] which deals with robust least-square error equalization design for multiple-input multiple-output communication channels, when the channel and noise are uncertain. In [11] the Lagrangian duality is used to transform the min-max problem into an equivalent convex min-min problem over a convex domain, to which standard convex optimization methods apply. This way the problem presented in [11] becomes a scalar minimization problem which can be solved numerically. Minimax estimation techniques for uncertain wide sense stationary processes are also considered in [15] and [30], but uncertainty is defined with respect to power spectrum densities. It is worth mentioning that related work in the context of nonlinear stochastic optimal control is also found in [43].

3.2 Nonlinear Optimization

This section, formulates and seeks solution to the estimation problem for a class of systems as follows. First, appropriate models for conditional distributions are

introduced. Then appropriate spaces are identified and existence of maximizing measure is shown. Then the theory of Lagrangian functional is invoked to find the maximizing measure. In addition, specific properties of the Lagrangian functional are identified.

3.2.1 Formulation on Abstract Spaces

Suppose a measurable space $\left(\Omega, \mathcal{F}\right)$ is given on which the unobserved Random Variable (RV), X and the observed RV Y are defined, via $X : (\Omega, \mathcal{F}) \rightarrow (\mathcal{X}, \Sigma_{\mathcal{X}})$, $Y : (\Omega, \mathcal{F}) \rightarrow (\mathcal{Y}, \Sigma_{\mathcal{Y}})$.

Thus $\mathcal{X}$ is the space of the unobserved RV, and $\mathcal{Y}$ is the space of the observed RV. The relation between the unobserved RV X and the observed RV Y is defined via a probabilistic mapping. The mapping $\mu : \mathcal{X} \times \Sigma_{\mathcal{Y}} \rightarrow [0, 1]$ satisfies the following two conditions:

1. For every $x \in \mathcal{X}$, the set function $\mu(x, \cdot)$ is a probability measure on $\Sigma_{\mathcal{Y}}$ (possibly finite additive);
2. For every $F \in \Sigma_{\mathcal{Y}}$, the function $\mu(\cdot, F)$ is $\mathcal{X}$-measurable.

The mapping μ is called a stochastic kernel or transition probability. Let $\mathcal{P}$ denote the class of all stochastic kernels, $\mathcal{M}_1(\mathcal{X})$ denote the space of probability measures (possibly finite additive) on $\mathcal{X}$, and let the measure induced by X, $P_X \in \mathcal{M}_1(\mathcal{X})$ be fixed. For the given pair $\{\mu, P_X\}$, $\mu \in \mathcal{P}, P_X \in \mathcal{M}_1(\mathcal{X})$, two probability measures may be introduced as follows.

1. The joint probability measure $P_{X,Y} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ defined by
$$P_{X,Y}(G) = (P_X \otimes \mu)(G) \stackrel{\triangle}{=} \int_{\mathcal{X}} \mu(x, G_x) P_X(dx), \ \ \forall\, G \in \Sigma_{\mathcal{X}} \times \Sigma_{\mathcal{Y}}$$
where $\otimes$ denotes convolution, G_x is the x-section of G defined by $G_x = \{y \in \mathcal{Y} : (x, y) \in G\}$.
2. The marginal probability measure $P_Y \in \mathcal{M}_1(\mathcal{Y})$ corresponding to $\mu \in \mathcal{P}$ is given by
$$P_Y(F) = P_{X,Y}(\mathcal{X} \times F) \stackrel{\triangle}{=} \int_{\mathcal{X}} \mu(x, F) P_X(dx), \ \ \forall\, F \in \Sigma_{\mathcal{Y}}$$

The objective is to estimate X by a function of the random variable Y. Let $\hat{X} = \Phi(Y)$ denote the estimate of X. The estimation is done by introducing a pay-off, and then minimizing the pay-off over the class of estimators Φ in the admissible set denoted by $\mathcal{X}_{ad}$. It is assumed that all admissible estimators $\Phi : \mathcal{Y} \to \mathcal{X}$ are $\Sigma_{\mathcal{Y}}$ measurable and continuous. Let $\ell : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be an $\Sigma_{\mathcal{X}} \times \Sigma_{\mathcal{Y}}$-measurable function, which corresponds to the sample pay-off. The classical estimation problem deals with minimization of the average pay-off given by

$$J(\Phi^*) = \inf_{\Phi \in \mathcal{X}_{ad}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) dP_{X,Y}(x, y) \tag{3.2}$$

$$= \inf_{\Phi \in \mathcal{X}_{ad}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) \mu(x, dy) dP_X(x) \tag{3.3}$$

$$= \inf_{\Phi \in \mathcal{X}_{ad}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) \eta(y, dx) dP_Y(y). \tag{3.4}$$

Clearly if $\ell(x, \Phi(y)) = ||x - \Phi(y)||_{\Re^n}$, then $\Phi^*(Y) = E[X|Y]$. On the other hand, if $P_{X,Y}$ or μ, η are unknown, then new estimators should be sought (this is done in Sections 3.2.2, 3.2.3 and 3.2.4). It is noted that (3.2) will be used when the uncertainty is on the joint distribution, while (3.3), (3.4) will be used when the uncertainty is on the channel kernel, *á* posteriori distribution, respectively.

Next the appropriate topologies and function spaces used in this paper are introduced. Throughout the rest of this chapter it is assumed that both $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces (complete separable metric spaces) and therefore normal topological spaces [45]. Let $BC(\mathcal{Y})$ denote the vector space of bounded continuous real valued functions defined on the Polish space $\mathcal{Y}$. Furnished with the sup norm topology, this is a Banach space. Let $\Big(BC(\mathcal{Y})\Big)^*$ denote its topological dual. It is known [45, pg. 262] that $\Big(BC(\mathcal{Y})\Big)^*$ is isometrically isomorphic to the Banach space of finitely additive regular bounded signed measures on $\Sigma_{\mathcal{Y}}$. Denote this by $M_{rba}(\mathcal{Y})$ and let $\Pi_{rba}(\mathcal{Y}) \subset M_{rba}(\mathcal{Y})$ denote the set of regular bounded finitely additive probability measures on $\mathcal{Y}$. Clearly if $\mathcal{Y}$ is compact, then $\Big(BC(\mathcal{Y})\Big)^*$ will be the space of countably additive signed measures. Let $L_1(P_X, BC(\mathcal{Y}))$ denote the space of all P_X integrable functions defined on $\mathcal{X}$ with values in $BC(\mathcal{Y})$. In other words, for each $\phi \in L_1(P_X, BC(\mathcal{Y}))$

$$\| \phi \|_{P_X} \equiv \int_{\mathcal{X}} ||\phi(x)(\cdot)||_{BC(\mathcal{Y})} P_X(dx) < \infty.$$

With respect to this norm topology this is a Banach space. Since the Banach spaces $BC(\mathcal{Y})$ and its dual $M_{rba}(\mathcal{Y})$ do not satisfy the Radon Nikodym property, the dual

of $L_1(P_X, BC(\mathcal{Y}))$ is not $L_\infty(P_X, M_{rba}(\mathcal{Y}))$. However, it follows from the theory of "lifting" [46, Theorem 7, pg. 94;Theorem 9, pg. 97] that the dual of $L_1(P_X, BC(\mathcal{Y}))$ is $L^w_\infty(P_X, M_{rba}(\mathcal{Y}))$, i.e., the space of all $M_{rba}(\mathcal{Y})$ valued functions $\{\mu\}$ which are weak star measurable, in the sense that for each $\phi \in BC(\mathcal{Y}), x \longrightarrow \mu_x(\phi) \equiv \int_{\mathcal{Y}} \phi(z)\mu(x, dz)$ is P_X measurable and P_X-essentially bounded. Now define the admissible set as follows

$$\mathcal{P} \equiv L^w_\infty(P_X, \Pi_{rba}(\mathcal{Y})) \subset L^w_\infty(P_X, M_{rba}(\mathcal{Y})).$$

In other words, $\mathcal{P}$ is the unit sphere in the space $L^w_\infty(P_X, M_{rba}(\mathcal{Y}))$. It is assumed throughout this chapter that $\ell : \mathcal{X} \times \mathcal{Y} \longrightarrow [0, \infty]$ is $\Sigma_{\mathcal{X}} \times \Sigma_{\mathcal{Y}}$ measurable (sample pay-off) function from the class $L_1(P_X, BC(\mathcal{Y}))$.

3.2.2 Uncertainty on the Channel Kernel and Minimax Pay-off

It is assumed that the probabilistic kernel $\mu \in \mathcal{P}$ introduced earlier represents the nominal system model or mapping, which is fixed. The true kernel denoted by $\nu : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$ is assumed unknown. Envisioned scenarios are communication channels whose nominal behavior in known, while its true conditional distribution is unknown. The KL distance will be used as a measure of distance between the true model and uncertainty model, hence the true kernel is assumed to belong to the pointwise uncertainty set defined by

$$\mathcal{A}^x(\mu) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : H(\nu|\mu)(x) \leq R(x)\Big\} \tag{3.5}$$

where $R : \mathcal{X} \to [0, \infty)$ and $H(\cdot|\cdot)(x) : \mathcal{X} \to [0, \infty]$ is the KL distance between two kernels defined by

$$H(\nu|\mu)(x) = \begin{cases} \int_{\mathcal{Y}} \log \frac{\nu(x,dy)}{\mu(x,dy)} \nu(x, dy), & \text{if } \nu(x, \cdot) \ll \mu(x, \cdot),\ P_X - a.s. \\ \\ \infty, & \text{otherwise.} \end{cases}$$

It is also assumed that $R \in BC(\mathcal{X})$. Moreover, $\mu : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$ is the nominal fixed kernel.

Remark 3.2.1. *The value of $R(x)$ can be evaluated by statistical methods as follows. Using experimental data and counting techniques the different possible conditional distributions $\nu(x, dy)$ can be found. On the other hand, if the true distribution $\nu(x, dy)$ is parameterized then by using counting techniques in finding relative entropy, the function $R(x)$ is determined.*

Additionally, the following uncertainty set is defined

$$\mathcal{A}(\mu) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : \int_{\mathcal{X}} H(\nu|\mu)(x) dP_X(x) \leq \int_{\mathcal{X}} R(x) dP_X(x) \stackrel{\nabla}{=} r_1\Big\}. \tag{3.6}$$

It is assumed that $\mathcal{A}(\mu)$ is non-empty. Since, the true kernel $\nu \in A(\mu)$, is unknown, then the estimation problem can be formulated as a minimax problem defined by

$$\begin{aligned} J_1(\Phi, \nu) &\stackrel{\triangle}{=} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) dP_X(x), \\ J_1(\Phi^*, \nu^*) &= \inf_{\Phi\in\mathcal{X}_{ad}} \sup_{\nu\in\mathcal{A}(\mu)} J_1(\Phi, \nu). \end{aligned} \tag{3.7}$$

Next the issue of the existence of a solution to the above problem is discussed. Let $\ell : \mathcal{X}\times\mathcal{Y} \longrightarrow [0, \infty]$ be any $\Sigma_{\mathcal{X}}\times\Sigma_{\mathcal{Y}}$ measurable function from the class $L_1(P_X, BC(\mathcal{Y}))$.

The set $\mathcal{P}$ is w^*-compact. This follows from the Alaoglu's theorem [45, Theorem V.4.2, pg. 424]. Also using the lower semi continuity property of relative entropy [47, Lemma 1.4.3, pg. 36][1], it follows that the set $\mathcal{A}(\mu)$ is a w^*-compact (as a w^*-closed subset of the w^*-compact set $\mathcal{P}$). The next lemma establishes the upper semi-continuity of the cost function.

Lemma 3.2.2. *Let $\mathcal{X}, \mathcal{Y}$ be two Polish spaces and $\ell : \mathcal{X}\times\mathcal{Y} \longrightarrow [0, \infty]$, a measurable, nonnegative, extended real valued function, and $y \to \ell(x, y)$ be continuous, for P_X-almost all $x \in \mathcal{X}$. Also assume $\Phi : \mathcal{Y} \to \mathcal{X}$ is $\Sigma_{\mathcal{Y}}$ measurable and continuous. Then the mapping $\nu \to \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) P_X(dx)$ is upper semi-continuous in the w^*-sense.*

Proof. Let $\{\nu_\alpha\} \in \mathcal{A}(\mu)$ be a net, where $\alpha \in (\mathcal{D}, \leq)$ (a directed set). Since $\mathcal{A}(\mu)$ is weak star compact, there exists a subnet of the net $\{\nu_\alpha\}$, relabeled as the original net, and an element $\nu \in \mathcal{A}(\mu)$ such that $\nu_\alpha \xrightarrow{w^*} \nu$. [2]

$$\begin{aligned} \overline{\lim}_\alpha \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \wedge m \;\; \nu_\alpha(x, dy) P_X(dx) \quad &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \wedge m \;\; \nu(x, dy) P_X(dx) \\ &\leq \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) P_X(dx) \end{aligned}$$

[1] Examining the proof in [47] one can easily verify that the same procedure holds true not only for countably additive measures but also for finitely additive ones.

[2] That is,

$$\lim_\alpha \Big| \int_{\mathcal{X}} \int_{\mathcal{Y}} \phi(x, y) \nu_\alpha(x, dy) P_X(dx) - \int_{\mathcal{X}} \int_{\mathcal{Y}} \phi(x, y) \nu(x, dy) P_X(dx) \Big| = 0$$

for any $\phi \in L_1(P_X; BC(\mathcal{Y}))$.

where $m \in N$ (N denotes non-negative integers) is arbitrary. Then there exists $\alpha_0 \in \mathcal{D}$ such that

$$\sup_{\alpha \geq \alpha_0} \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \wedge m \;\; \nu_\alpha(x, dy) P_X(dx) \leq \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) P_X(dx). \quad (3.8)$$

On the other hand, the following holds

$$\begin{aligned}
&\overline{\lim}_\alpha \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \nu_\alpha(x, dy) P_X(dx) = \overline{\lim}_\alpha \sup_{m \in N} \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \wedge m \quad \nu_\alpha(x, dy) P_X(dx) \\
&\leq \sup_{\alpha \geq \alpha_0} \sup_{m \in N} \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \wedge m \quad \nu_\alpha(x, dy) P_X(dx) \\
&\leq \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) P_X(dx)
\end{aligned}$$

where the last inequality follows from (3.8) and the fact that two supremums can be interchanged.

Thus, since $\mathcal{A}(\mu)$ is w^*-compact, and

$$\nu \to \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) P_X(dx)$$

is upper semicontinuous, from the Weierstrass theorem, existence of solution to the supremum problem in (3.7) is established. □

Remark 3.2.3. *Note that by a generalization of von Neumann's minimax theorem, the problem* (3.7) *with $\mathcal{X}_{ad}$ convex, $\ell(x, \Phi) = \|x - \Phi\|_{\Re^n}$, satisfies (see Theorem 2.1.15)*

$$\begin{aligned}
&\inf_{\Phi \in \mathcal{X}_{ad}} \max_{\nu \in \mathcal{A}(\mu)} \int_{\mathcal{X} \times \mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) dP_X(x) \\
&= \max_{\nu \in \mathcal{A}(\mu)} \min_{\Phi \in \mathcal{X}_{ad}} \int_{\mathcal{X} \times \mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) dP_X(x) \qquad (3.9)
\end{aligned}$$

since

1) $\mathcal{A}(\mu)$ is compact,

2) $J_1(\Phi, \nu)$ as a function on $\mathcal{X}_{ad} \times \mathcal{A}(\mu)$ is convex-concave, and

3) $J_1(\Phi, \nu)$ as a function on $\mathcal{X}_{ad} \times \mathcal{A}(\mu)$ is upper-semicontinuous in ν for each $\Phi \in \mathcal{X}_{ad}$.

Clearly, for the case of this chapter $\mathcal{A}(\mu)$ is convex (because of the convexity of relative entropy [47]), $\mathcal{A}(\mu)$ is $weak^*$ compact (see statement above Lemma 3.2.2), $J_1(\Phi, \nu)$ is convex in $\mathcal{X}_{ad}$ and concave in $\mathcal{A}(\mu)$, and $J_1(\Phi, \nu)$ is upper-semicontinuous in ν for each $\Phi \in \mathcal{X}_{ad}$ (by Lemma 3.2.2). Therefore, the only additional assumption needed for existence of saddle point (hence inf *becomes* min*) is that A1) $\mathcal{X}_{ad}$ is compact and convex, A2) $J_1(\cdot, \nu)$ is continuous in Φ for each $\nu \in \mathcal{A}(\mu)$ and hence lower-semicontinuous.*

Define

$$J_1(\phi, \nu^*) \stackrel{\triangle}{=} \sup_{\nu \in \mathcal{A}(\mu)} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \phi(y)) \nu(x, dy) dP_X(x).$$

The solution of the supremum over $\mathcal{A}(\mu)$ is resolved by introducing a pair of Lagrange multipliers $\{\lambda(x), s\}$, $s \in \Re$, $\lambda \in L_1(P_X)$ and defining the Lagrangian functional

$$\begin{aligned} L_1(\nu, \lambda, s) \stackrel{\triangle}{=} \quad & \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) dP_X(x) - s \int_{\mathcal{X}} (H(\nu|\mu)(x) - R(x)) dP_X(x) \\ & - \int_{\mathcal{X}} \lambda(x) \bigg(\int_{\mathcal{Y}} \nu(x, dy) - 1 \bigg) dP_X(x) \end{aligned} \tag{3.10}$$

and the dual functional

$$L_1(\nu^*, \lambda^*, s^*) = \inf_{s \geq 0} \inf_{\{\lambda \in L_1(P_X),\ \lambda \geq 0\}} \sup_{\nu \in \mathcal{P}} L_1(\nu, \lambda, s). \tag{3.11}$$

Alternatively, if the uncertainty set $\mathcal{A}^x(\mu)$ is considered, the supremum problem is defined as

$$\tilde{J}_1(\Phi, \nu^*) = \sup_{\nu \in \mathcal{A}^x(\mu)} \int_{\mathcal{X}} \int_{\mathcal{Y}} \ell(x, \phi(y)) \nu(x, dy) dP_X(x).$$

In this case the solution of the supremum over $\mathcal{A}^x(\mu)$ is resolved using lagrange multiplier $\{\lambda(x), \tilde{s}\}$, $\lambda \in L_1(P_X)$, $\tilde{s} \in L_1(P_X)$, and defining the Lagrangian functional

$$\begin{aligned} \tilde{L}_1(\nu, \lambda, \tilde{s}) \stackrel{\triangle}{=} \quad & \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) \nu(x, dy) dP_X(x) - \int_{\mathcal{X}} \tilde{s}(x) (H(\nu|\mu)(x) - R(x)) dP_X(x) \\ & - \int_{\mathcal{X}} \lambda(x) \bigg(\int_{\mathcal{Y}} \nu(x, dy) - 1 \bigg) dP_X(x) \end{aligned} \tag{3.12}$$

The dual functional is given by

$$\tilde{L}_1(\nu^*, \lambda^*, \tilde{s}^*) = \inf_{\{\tilde{s} \in L_1(P_X), \tilde{s} \geq 0\}} \inf_{\{\lambda \in L_1(P_X),\ \lambda \geq 0\}} \sup_{\nu \in \mathcal{P}} \tilde{L}_1(\nu, \lambda, \tilde{s}) \tag{3.13}$$

Now, if the infimum over the functions $\tilde{s} \in L_1(P_X)$ is taken over the real numbers, then the following inequality holds

$$\tilde{L}_1(\nu^*, \lambda^*, \tilde{s}^*) \leq \inf_{\tilde{s} \geq 0} \inf_{\{\lambda \in L_1(P_X),\ \lambda \geq 0\}} \sup_{\nu \in \mathcal{P}} \tilde{L}_1(\nu, \lambda, \tilde{s}) = L_1(\nu^*, \lambda, \tilde{s}) \tag{3.14}$$

Hence, uncertainty modeling using $\mathcal{A}(\mu)$ yields higher pay-off than uncertainty modeling using $\mathcal{A}^x(\mu)$.

Next, the equivalence between the constrained problem $J_1(\Phi, \nu^*)$ and the unconstrained problem $L_1(\nu^*, \lambda^*, s^*)$ is established.

Theorem 3.2.4. *Suppose $\ell : \mathcal{X} \times \mathcal{X} \longrightarrow \overline{R}_0 \equiv [0, \infty]$ is continuous in the second argument, and $\Phi : \mathcal{Y} \to \mathcal{X}$ is continuous. If for a given $\Phi \in \mathcal{X}_{ad}$, $\sup_{\nu \in \mathcal{A}(\mu)} J_1(\Phi, \nu)$ is finite, then the constrained problem (inner supremum) in (3.7), is equivalent to an unconstrained problem as stated below:*

$$\sup_{\nu \in \mathcal{A}(\mu)} J_1(\Phi, \nu) = \inf_{s \geq 0} \inf_{\lambda(x)} \sup_{\nu \in \mathcal{A}(\mu)} L_1(\nu, \lambda, s).$$

Further the infimum occurs on the boundary of the set $\mathcal{A}(\mu)$, that is

$$\int_{\mathcal{X}} H(\nu^*|\mu)(x) dP_X(x)\Big|_{s=s^*} = r_1.$$

Proof. The proof is based on Lagrange Duality theorem [40, Theorem 1, pg. 224]. $\overline{\mathcal{X}}$ is chosen as $\overline{\mathcal{X}} \equiv L_\infty^w(P_X, M_{rba}(\mathcal{Y}))$ which is clearly a vector space. For the set Ω the natural choice is the set $\Omega = L_\infty^w(P_X, \Pi_{rba}(\mathcal{Y}) \subseteq \overline{\mathcal{X}}$. Clearly, $\int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y))\nu(x, dy)dP_X(x)$ is a linear functional of ν. Theorem 1, in [40, pg. 224] deals with minimization of a convex functional. Multiplying the equation (4) of [40, pg. 224] by a minus sign, converts the problem to maximization of a concave functional over the set Ω. This can be applied to maximization of $J_1(\Phi, \nu)$ over the constrained set, since $\int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y))\nu(x, dy)dP_X(x)$ is a linear functional of ν (hence concave). Also, it follows that Ω is a convex set. Take $G(\nu) = \int_{\mathcal{X}} H(\nu|\mu)(x) - R(x)dP_X(x)$, where $\nu \in \overline{\mathcal{X}}$. Then $G : \overline{\mathcal{X}} \to \Re$, is a convex mapping from $\overline{\mathcal{X}}$ into the ordered vector space $(\Re, \prec)$ with natural ordering. If $\nu = \mu$ is chosen, then $G(\nu) = -\int_{\mathcal{X}} R(x)dP_X(x) < 0$ [3]. Hence there exists a measure $\nu \in \overline{\mathcal{X}}$ such that $G(\nu) < 0$. So when $\sup_{\nu \in \mathcal{A}(\mu)} J_1(\Phi, \nu)$ is finite, conditions of the theorem are satisfied, and the constrained and unconstrained problems are equivalent. Also according to the same duality theorem, if the supremum is achieved by some $\nu^* \in L_\infty^w(P_X, M_{rba}(\mathcal{Y}))$, then

$$s\Big(\int_{\mathcal{X}} (H(\nu^*|\mu)(x) - R(x))dP_X(x)\Big) = 0.$$

In other words, for non-zero $s \in (0, \infty)$, the solution occurs on the boundary. □

The solution of the maximization over $\nu \in \mathcal{A}(\mu)$ of (3.10) is presented in the next Theorem.

[3] The inequality is strict if R is non-zero on a set of non-zero measure, i.e., if there exists $E \in \Sigma_{\mathcal{X}}$, such that $P_X(E) \neq 0$, and $R(x) \neq 0$ for $x \in E$. If such a set does not exist, then R would be zero P_X almost everywhere, and the problem becomes trivial.

Theorem 3.2.5. *Suppose the condition of Theorem 2.1.16 holds. The supremum of (3.11) over $\nu \in \mathcal{A}(\mu)$ is given by*

$$\nu^*(x, dy) = \frac{e^{\frac{\ell(x,\Phi(y))}{s}}\mu(x, dy)}{\int_{\mathcal{Y}} e^{\frac{\ell(x,\Phi(y))}{s}}\mu(x, dy)} \tag{3.15}$$

where $s \geq 0$ is found by the constraint

$$\int_{\mathcal{X}} H(\nu^*|\mu)(x)dP_X(x)\Big|_{s=s^*} = r_1. \tag{3.16}$$

Moreover,

$$L_1(\nu^*, \lambda^*, s^*) = \inf_{s\geq 0} \int_{\mathcal{X}} s \log \int_{\mathcal{Y}} e^{\frac{\ell(x,\Phi(y))}{s}}\mu(x, dy)dP_X(x) + s\int_{\mathcal{X}} R(x)dP_X(x). \tag{3.17}$$

Proof. The Gateaux derivative of $L_1(\cdot, \lambda, s)$ at any ν^* in the direction $\nu - \nu^*$. is derived by computing the following expression

$$\frac{d}{d\varepsilon}L_1(\nu^* + \varepsilon(\nu - \nu^*), \lambda, s)\Big|_{\varepsilon=0} \stackrel{\triangle}{=} \delta L_1(\nu^*; \nu - \nu^*). \tag{3.18}$$

After some calculations it is deduced to

$$\begin{aligned}
\delta L_1(\nu^*; \nu - \nu^*) \quad &= \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y))(\nu(x, dy) - \nu^*(x, dy))dP_X(x) \\
&\quad - \int_{\mathcal{X}} s\bigg(\int_{\mathcal{Y}} \log\frac{\nu^*(x, dy)}{\mu(x, dy)}(\nu(x, dy) - \nu^*(x, dy))\bigg)dP_X(x) \\
&\quad - \int_{\mathcal{X}} \lambda(x)\int_{\mathcal{Y}} (\nu(x, dy) - \nu^*(x, dy))dP_X(x) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \log\left(e^{\ell(x,\Phi(y))-\lambda(x)}\Big(\frac{\nu^*(x, dy)}{\mu(x, dy)}\Big)^{-s}\right)\Big(\nu(x, dy) \\
&\quad -\nu^*(x, dy))dP_X(x)\bigg).
\end{aligned} \tag{3.19}$$

Since $L_1(\nu, \lambda, s)$ is concave in ν, then from basic principles of calculus of variation a necessary and sufficient condition for ν^* to be the maximizer measure is that $\delta L_1(\nu^*; \nu - \nu^*) = 0,\ \forall\ \nu \in \mathcal{P}$. Since the inequality holds for all ν the Gateaux gradient must vanish, hence

$$\left(\frac{\nu^*(x, dy)}{\mu(x, dy)}\right)^{-s} = e^{-\ell(x,\Phi(y))+\lambda(x)}.$$

Moreover,

$$\frac{\nu^*(x, dy)}{\mu(x, dy)} = e^{\frac{\ell(x,\Phi(y))-\lambda(x)}{s}}.$$

Since $\nu^*(x, dy)$ is a probability measure on $\mathcal{Y}$, then using the constraint $\int_{\mathcal{Y}} \nu^*(x, dy) = 1$, the following function is obtained

$$e^{-\frac{\lambda(x)}{s}} = \frac{1}{\int_{\mathcal{Y}} e^{\frac{\ell(x,\Phi(y))}{s}} \mu(x, dy)}.$$

Thus,

$$\nu^*(x, dy) = \frac{e^{\frac{\ell(x,\Phi(y))}{s}} \mu(x, dy)}{\int_{\mathcal{Y}} e^{\frac{\ell(x,\Phi(y))}{s}} \mu(x, dy)}, \quad \forall\, x \in \mathcal{X}. \tag{3.20}$$

Substituting ν^* into (3.10) yields

$$L_1(\nu^*, \lambda, s) = \int_{\mathcal{X}} s \log \int_{\mathcal{Y}} e^{\frac{\ell(x,\Phi(y))}{s}} \mu(x, dy) dP_X(x) + \int_{\mathcal{X}} sR(x) dP_X(x). \tag{3.21}$$

□

3.2.3 Uncertainty on the *Á* Posteriori Distribution and Minimax Pay-off

In this section an uncertainty model on the *á* posteriori distribution is considered. For this case, one may consider the mapping $\eta : \mathcal{Y} \times \Sigma_{\mathcal{X}} \to [0, 1]$ that satisfies the following two conditions:

1. For every $y \in \mathcal{Y}$, the set function $\eta(y, \cdot)$ is a probability measure on $\Sigma_{\mathcal{X}}$ (possibly finite additive);
2. For every $F \in \Sigma_{\mathcal{X}}$, the function $\eta(\cdot, F)$ is $\mathcal{Y}$-measurable.

This probabilistic kernel $\eta(y, dx)$ represents the nominal system model or mapping (*á* posteriori information). The true kernel $\nu(y, dx)$, denoted by $\nu : \mathcal{Y} \times \Sigma_{\mathcal{X}} \to [0, 1]$ belongs to an uncertainty set described by

$$\mathcal{B}^y(\eta) \stackrel{\triangle}{=} \Big\{ \nu \in \mathcal{P} : H(\nu|\eta)(y) \leq R(y) \Big\}$$

where $R : \mathcal{Y} \to [0, \infty)$ and $H(\cdot|\cdot)(y) : \mathcal{Y} \to [0, \infty]$ is the KL distance between two kernels defined, in a similar way as before, by

$$H(\nu|\eta)(x) = \begin{cases} \int_{\mathcal{X}} \log \frac{\nu(y,dx)}{\eta(y,dx)} \nu(y, dx), & \text{if } \nu(y, \cdot) \ll \eta(y, \cdot),\ P_Y - a.s. \\ \\ \infty, & \text{otherwise.} \end{cases}$$

It is assumed that $R \in BC(\mathcal{Y})$. Moreover, $\eta : \mathcal{Y} \times \Sigma_{\mathcal{X}} \to [0,1]$ is the nominal fixed kernel. The value of $R(y)$ can be evaluated by statistical methods as explained in Remark 3.2.1.

It is also assumed that $\mathcal{B}^y(\eta)$ is non-empty. Since, the true kernel $\nu \in \mathcal{B}^y(\eta)$, is unknown, then the estimation problem can be formulated as a minimax problem defined by

$$J_2(\Phi^*, \nu^*) = \inf_{\Phi \in \mathcal{X}_{ad}} \sup_{\nu \in \mathcal{B}^y(\eta)} \int_{\mathcal{X}} \ell(x, \Phi(y)) \nu(y, dx). \tag{3.22}$$

Existence and equivalence of constrained and unconstrained problem is shown similarly using Theorem 2.1.16 following the same procedure as in Section 3.2.2.

The solution of the supremum over $\mathcal{B}^y(\eta)$ is resolved by introducing a pair of Lagrange multipliers $\{\lambda(y), \tilde{s}(y)\}$, $\lambda, \tilde{s} \in L_1(P_Y)$ and defining the Lagrangian

$$\begin{aligned} L_2(\nu, \lambda, \tilde{s}) \stackrel{\triangle}{=} \quad & \int_{\mathcal{X}} \ell(x, \Phi(y)) \nu(y, dx) - \tilde{s}(y) \Big(H(\nu|\eta)(y) - R(y) \Big) \\ & - \lambda(y) \Big(\int_{\mathcal{X}} \nu(y, dx) - 1 \Big) \end{aligned} \tag{3.23}$$

and the dual functional

$$L_2(\nu^*, \lambda^*, \tilde{s}^*) = \inf_{\{\tilde{s} \in L_1(P_Y),\ \tilde{s} \geq 0\}} \inf_{\{\lambda \in L_1(P_Y),\ \lambda \geq 0\}} \sup_{\nu \in \mathcal{P}} L_2(\nu, \lambda, \tilde{s}). \tag{3.24}$$

The solution of the maximization over $\mathcal{B}^y(\eta)$ of (3.23) is presented in the next Theorem.

Theorem 3.2.6. *Suppose the condition of Theorem 2.1.16 holds. The supremum of (3.24) over $\mathcal{B}^y(\eta)$ is given by*

$$\nu^*(y, dx) = \frac{e^{\frac{\ell(x, \Phi(y))}{\tilde{s}(y)}} \eta(y, dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x, \Phi(y))}{\tilde{s}(y)}} \eta(y, dx)} \tag{3.25}$$

where $\tilde{s} \in L_1(P_Y)$ is found by the constraint

$$H(\nu^*|\eta)(y)\Big|_{\tilde{s}(y) = \tilde{s}^*(y)} = R(y). \tag{3.26}$$

Moreover,

$$L_2(\nu^*, \lambda^*, \tilde{s}^*) = \inf_{\{\tilde{s} \in L_1(P_Y),\ \tilde{s} \geq 0\}} \tilde{s}(y) \log \Big(\int_{\mathcal{X}} e^{\frac{\ell(x, \Phi(y))}{\tilde{s}(y)}} \eta(y, dx) \Big) + \tilde{s}(y) R(y). \tag{3.27}$$

Proof. The proof is similar with the one in section 3.2.2. The Gateaux derivative of $L_2(\cdot, \lambda, \tilde{s})$ at any ν^* in the direction $\nu - \nu^*$ is derived by computing the following expression

$$\frac{d}{d\varepsilon} L_1(\nu^* + \varepsilon(\nu - \nu^*), \lambda, \tilde{s})\Big|_{\varepsilon=0} \stackrel{\triangle}{=} \delta L_1(\nu^*; \nu - \nu^*). \tag{3.28}$$

After some calculations the following expression is deduced

$$\begin{aligned}
\delta L_2(\nu^*; \nu - \nu^*) \quad &= \int_{\mathcal{X}} \ell(x, \Phi(y))(\nu(y, dx) - \nu^*(y, dx)) \\
&\quad -\tilde{s}(y)\bigg(\int_{\mathcal{X}} \log \frac{\nu^*(y, dx)}{\eta(y, dx)}(\nu(y, dx) - \nu^*(y, dx))\bigg) \\
&\quad -\lambda(y)\int_{\mathcal{X}} (\nu(y, dx) - \nu^*(y, dx)) \\
&= \int_{\mathcal{X}} \log\bigg(e^{\ell(x,\Phi(y))-\lambda(y)}\Big(\frac{\nu^*(y, dx)}{\eta(y, dx)}\Big)^{-\tilde{s}(y)}\bigg)(\nu(y, dx) - \nu^*(y, dx)).
\end{aligned} \tag{3.29}$$

Since $L_2(\nu, \lambda, \tilde{s})$ is concave in ν, then from basic principles of calculus of variation a necessary and sufficient condition for ν^* to be the maximizer measure is that $\delta L_2(\nu^*; \nu - \nu^*) = 0,\ \forall\ \nu \in \mathcal{P}$. Since the inequality holds for all ν the Gateaux gradient must vanish, hence

$$\Big(\frac{\nu^*(x, dy)}{\mu(x, dy)}\Big)^{-\tilde{s}(y)} = e^{-\ell(x,\Phi(y))+\lambda(y)}.$$

Moreover,

$$\frac{\nu^*(x, dy)}{\eta(x, dy)} = e^{\frac{\ell(x,\Phi(y))-\lambda(y)}{\tilde{s}(y)}}.$$

Since $\nu^*(y, dx)$ is a probability measure on $\mathcal{X}$, then using the constraint $\int_{\mathcal{X}} \nu^*(y, dx) = 1$, the following function is optained

$$e^{-\frac{\lambda(y)}{\tilde{s}(y)}} = \frac{1}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}}\eta(y, dx)}.$$

Thus,

$$\nu^*(y, dx) = \frac{e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}}\eta(y, dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}}\eta(y, dx)}, \quad \forall\ y \in \mathcal{Y}. \tag{3.30}$$

Substituting ν^* into (3.23) yields

$$L_2(\nu^*, \lambda, \tilde{s}) = \inf_{\{\tilde{s}\in L_1(P_Y),\ \tilde{s}\geq 0\}} \tilde{s}(y) \log\bigg(\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}}\eta(y, dx)\bigg) + \tilde{s}(y)R(y). \tag{3.31}$$

□

Alternatively, one may consider the constraint

$$\mathcal{B}(\eta) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : \int_{\mathcal{Y}} H(\nu|\eta)(y)dP_Y(y) \leq \int_{\mathcal{Y}} R(y)dP_Y(y) \stackrel{\nabla}{=} r_2\Big\}.$$

This estimation problem can be formulated as a minimax problem defined by

$$J_3(\Phi^*, \nu^*) = \inf_{\Phi\in\mathcal{X}_{ad}} \sup_{\nu\in\mathcal{B}(\eta)} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y))\nu(y, dx)dP_Y(y). \tag{3.32}$$

Defining the Lagrangian as before (with the Lagrange multiplier $\tilde{s}$ replaced by a real number); then its supremum over $\nu \in \mathcal{P}$ is given by

$$\nu^*(y, dx) = \frac{e^{\frac{\ell(x,\Phi(y))}{s}}\eta(y, dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{s}}\eta(y, dx)} \tag{3.33}$$

where $s \geq 0$ is found by the constraint

$$\int_{\mathcal{Y}} H(\nu^*|\eta)(y)dP_Y(y)\Big|_{s=s^*} = r_2. \tag{3.34}$$

Moreover,

$$L_3(\nu^*, \lambda^*, s^*) = \inf_{s\geq 0} \int_{\mathcal{Y}} s \log \int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{s}}\eta(y, dx)dP_Y(y) + \int_{\mathcal{Y}} sR(y)dP_Y(y). \tag{3.35}$$

Remark 3.2.7. *Next, a new case is investigated, when the true kernel $\nu(y, dx)$, denoted by $\nu : \mathcal{Y} \times \Sigma_{\mathcal{X}} \to [0, 1]$ belongs to a new uncertainty set described by*

$$\mathcal{B}^R(\eta) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : \int_{\mathcal{Y}} H(\nu|\eta)(y)dP_Y(y) \leq \int_{\mathcal{X}\times\mathcal{Y}} R(x, y)\nu(y, dx)dP_Y(y) + \overline{R} \stackrel{\nabla}{=} r_3\Big\},$$

where $R : \mathcal{X}\times\mathcal{Y} \to [0, \infty)$, $R \in BC(\mathcal{X}\times\mathcal{Y})$. The estimation problem can be formulated using the uncertainty set $\mathcal{B}^*(\eta)$ as follows.

$$J_3^R(\Phi^*, \nu^*) = \inf_{\Phi\in\mathcal{X}_{ad}} \sup_{\nu\in\mathcal{B}^R(\eta)} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y))\nu(y, dx)dP_Y(y). \tag{3.36}$$

Defining the Lagrangian as before, then its supremum over $\nu \in \mathcal{P}$ is given by

$$\nu^*(y, dx) = \frac{e^{\frac{\ell(x,\Phi(y))}{s} - R(x,y)}\eta(y, dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{s} - R(x,y)}\eta(y, dx)} \tag{3.37}$$

where $s \geq 0$ is found by the constraint

$$\int_{\mathcal{Y}} H(\nu^*|\eta)(y)dP_Y(y)\Big|_{s=s^*} = r_3. \tag{3.38}$$

Moreover,

$$L_3^R(\nu^*, \lambda^*, s^*) = \inf_{s\geq 0} \int_{\mathcal{Y}} s \log \int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{s} - R(x,y)}\eta(y, dx)dP_Y(y) + s\overline{R}. \tag{3.39}$$

3.2.4 Uncertainty on the Joint Distribution and Minimax Pay-off

In this section, the uncertainty is modeled via the joint distribution of X and Y. This model is appropriate when one wishes to model $\acute{a}$ priori uncertainty. The pay-off function (3.2) is considered

$$J_4(\Phi, Q_{X,Y}) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) dQ_{X,Y}(x, y)$$

in which $Q_{X,Y}$ is the uncertain joint measure. Assume the joint distribution $P_{X,Y}(x, y)$ represents the nominal system model. The true joint distribution denoted by $Q \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, is assumed to belong to an uncertainty set described by

$$\mathcal{C}(P_{X,Y}) \stackrel{\triangle}{=} \{Q_{X,Y} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) : H(Q_{X,Y}|P_{X,Y}) \leq R\}$$

where $R \in [0, \infty)$ and $H(\cdot|\cdot)$ is the KL distance between the two joint distributions defined by

$$H(Q_{X,Y}|P_{X,Y}) = \begin{cases} \int_{\mathcal{X}\times\mathcal{Y}} \log \frac{Q_{X,Y}(x,y)}{P_{X,Y}(x,y)} Q_{X,Y}(x, dy) & \text{if } Q_{X,Y} \ll P_{X,Y} \\ \infty, & \text{otherwise.} \end{cases}$$

The value of R can be evaluated by statistical methods as explained in Remark 3.2.1.

Since the true probability measure $Q_{X,Y} \in \mathcal{C}(P_{X,Y})$, then the estimation problem can be formulated as a minimax problem defined by

$$J_4(\Phi^*, Q^*_{X,Y}) = \inf_{\Phi\in\mathcal{X}_{ad}} \sup_{Q_{X,Y}\in\mathcal{C}(P_{X,Y})} \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) dQ_{X,Y}(x, y). \tag{3.40}$$

This type of problem is similar to the one pursued in [16] and [29], although the method and results presented here are different. The system considered here is an abstract nonparametric system which can be used for various problems, where as in [16] a specific gaussian system is considered and the results presented are base on a parametric uncertainty model. In [29], even though a nonparametric model is used, the method described does not include a minimax approach and it is only shown that risk-sensitive estimators enjoy an error bound.

Existence and equivalence of constrained and unconstrained problem is shown similarly using Theorem 2.1.16. The supremum in (3.40) is resolved by defining the Lagrangian

$$L_4(Q^*_{X,Y}, \lambda^*, s^*) \stackrel{\triangle}{=} \quad \inf_{s\geq 0} \inf_{\lambda\geq 0} \sup_{Q_{X,Y}\in\mathcal{C}(P_{X,Y})} \Big\{ \int_{\mathcal{X}\times\mathcal{Y}} \ell(x, \Phi(y)) dQ_{X,Y}(x, y)$$

$$-s\Big(H(Q_{X,Y}|P_{X,Y})-R\Big)-\lambda\Big(\int_{\mathcal{X}\times\mathcal{Y}}dQ_{X,Y}(x,y)-1\Big)\Big\}. \tag{3.41}$$

The main results which are derived following the abstract formulation in [43] are provided in the next theorem.

Theorem 3.2.8. *Let $\Sigma \stackrel{\triangle}{=} \mathcal{X}\times\mathcal{Y}$. For every $\frac{\ell}{s}:\Sigma\to\Re$ measurable function, bounded below and for $s>0$*

$$\begin{aligned} L_4(Q^*_{X,Y},\lambda^*,s) &= \sup_{Q_{X,Y}\in\mathcal{C}(P_{X,Y})}\Big\{\int_\Sigma \ell(x,\Phi(y))dQ_{X,Y}(x,y)-sH(Q_{X,Y}|P_{X,Y})\Big\}+sR \\ &= s\log\Big(\int_\Sigma e^{\frac{\ell(x,\Phi(y))}{s}}dP_{X,Y}(x,y)\Big)+sR. \end{aligned} \tag{3.42}$$

Moreover, if $\ell(x,\Phi(y))e^{\frac{\ell(x,\Phi(y))}{s}}\in L_1(P_{X,Y})$, then the supremum in (3.42) is attained by the tilted probability measure $Q^*_{X,Y}$ given by

$$dQ^*_{X,Y}(x,y)=\frac{e^{\frac{\ell(x,\Phi(y))}{s}}dP_{X,Y}(x,y)}{\int_\Sigma e^{\frac{\ell(x,\Phi(y))}{s}}dP_{X,Y}(x,y)} \tag{3.43}$$

Proof. The derivation is similar to the one presented under Theorem 3.2.5. □

Next, several properties associated with the maximization over the set $\mathcal{C}(P_{X,Y})$ are presented.

Lemma 3.2.9. *Suppose for a given $\Phi\in\mathcal{X}_{ad}$,*

$$\sup_{Q_{X,Y}\in\mathcal{C}(P_{X,Y})}J_4(\Phi,Q_{X,Y})<\infty,$$

$s\in\Re$, $\frac{\ell}{s}$ a measurable function bounded from below, the following statements hold.

1) *The dual functional $L_4(Q^*_{X,Y},\lambda^*,s)$ is related to the cumulant generating function of $\ell(x,\Phi(y))$ with respect to $P_{X,Y}\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ via*

$$\begin{aligned} L_4(Q^*_{X,Y},\lambda^*,s) &= s\sup_{\{Q_{X,Y}\in\mathcal{M}_1(\Sigma);H(Q_{X,Y}|P_{X,Y})<\infty\}} \\ &\quad \Big\{\frac{1}{s}\int_\Sigma \ell(x,\Phi(y)dQ_{X,Y}(x,y)-H(Q_{X,Y}|P_{X,Y})\Big\}+sR \\ &= s\log\int_\Sigma e^{\frac{\ell(x,\Phi(y))}{s}}dP_{X,Y}(x,y)+sR \\ &= s\Psi_{P_{X,Y}}(\frac{1}{s})+sR. \end{aligned} \tag{3.44}$$

Moreover, if $\ell e^{\frac{\ell}{s}} \in L_1(P_{X,Y})$ *the supremum in (3.44) is attained at* $Q^*_{X,Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ *and it is given by*

$$dQ^*_{X,Y}(x,y) = \frac{e^{\frac{\ell(x,\Phi(y))}{s}} dP_{X,Y}(x,y)}{\int_\Sigma e^{\frac{\ell(x,\Phi(y))}{s}} dP_{X,Y}(x,y)}. \tag{3.45}$$

In addition, "The average energy of the system" = "The Helmoltz Free Energy" $+s\times$ *"The Relative Entropy of the system", that is,*

$$\begin{aligned}&\int_\Sigma \ell(x,\Phi(y)) dQ^*_{X,Y}(x,y)\\ &= s\log \int_\Sigma e^{\frac{\ell(x,\Phi(y))}{s}} dP_{X,Y}(x,y) + sH(Q^*_{X,Y}|P_{X,Y}), \quad s\in(0,\infty)\end{aligned} \tag{3.46}$$

2) *The dual functional* $L_4(Q^*_{X,Y},\lambda^*,s)$ *is convex in* $s>0$.

3) *The function* $\Gamma_{P_{X,Y}}(s) \stackrel{\triangle}{=} s\Psi_{P_{X,Y}}(\frac{1}{s})$ *is a non-increasing function of* $s\in(0,\infty)$*, that is,*

$$\Gamma_{P_{X,Y}}(s) = s_1 \log E_{P_{X,Y}}\Big\{e^{\frac{\ell(x,\Phi(y))}{s_1}}\Big\} \le s_2 \log E_{P_{X,Y}}\Big\{e^{\frac{\ell(x,\Phi(y))}{s_2}}\Big\} = \Gamma_{P_{X,Y}}(s_2), \quad 0<s_2\le s_1 \tag{3.47}$$

4) *The infimum of the dual functional* $L_4(Q^*_{X,Y},\lambda^*,s)$ *over* $s>0$ *defined by*

$$L_4(Q^*_{X,Y},\lambda^*,s^*) = \inf_{s>0}\Big\{s\Psi_{P_{X,Y}}(\frac{1}{s}) + sR\Big\} \tag{3.48}$$

is a concave functional of $R\ge 0$.

5) *Assume* $\exists \eta>0$ *such that* $\ell e^{\eta\ell}\in L_1(P_{X,Y})$. *Then* $L_4(Q^*_{X,Y},\lambda^*,s^*)$ *evaluated at at* $R=0$ *is given by*

$$L_4(Q^*_{X,Y},\lambda^*,s^*)\Big|_{R=0} = \lim_{s\to\infty} s\log\int_\Sigma e^{\frac{\ell(x,\Phi(y))}{s}} dP_{X,Y}(x,y) = E_{P_{X,Y}}\Big\{\ell(x,\Phi(y))\Big\}. \tag{3.49}$$

6) *Under the assumptions of 5), the supremum of the dual functional* $L_4(Q^*_{X,Y},\lambda^*,s)$ *over* $s>0$ *is bounded above and from below as follows.*

$$E_{P_{X,Y}}\Big\{\ell(x,\Phi(y))\Big\} \le L_4(Q^*_{X,Y},\lambda^*,s^*) \le R + \log E_{p_{X,Y}}\Big\{e^{\ell(x,\Phi(y))}\Big\}. \tag{3.50}$$

Moreover if $\ell(x,\Phi(y)$ *is* $Q_{X,Y}$*-essentially bounded for all* $Q_{X,Y}\in\mathcal{C}$*, then the above bounds become*

$$\begin{aligned}E_{P_{X,Y}}\Big\{\ell(x,\Phi(y))\Big\} \quad &\le L_4(Q^*_{X,Y},\lambda^*,s^*)\\ &\le \min\Big\{R+\log E_{P_{X,Y}}\{\ell(x,\Phi(y))\}, ||\ell(x,\Phi(y))||_\infty\Big\}.\end{aligned}$$

7) If for any $\eta > 0$, $\ell e^{\eta\ell} \in L_1(P_{X,Y})$ and $(\ell)^2 e^{\eta\ell} \in L_1(P_{X,Y})$ then the infimum of the functional $L_4(Q^*_{X,Y}, \lambda^*, s)$ over $s > 0$ is uniquely attained at

$$H(Q^*_{X,Y}|P_{X,Y})|_{s=s^*} = R \tag{3.51}$$

where $Q^*_{X,Y}$ is given by (3.45).
That is, a necessary condition for the infimum of the dual functional $L_4(Q^*_{X,Y}, \lambda^*, s)$ over $s > 0$ is that s^* occurs on the boundary of the relative entropy constraint.
Moreover,

$$\begin{aligned} \frac{d}{ds} s \log \int_\Sigma e^{\frac{\ell(x,\Phi(y))}{s}} dP_{X,Y} &= \log \int_\Sigma e^{\frac{\ell(x,\Phi(y))}{s}} dP_{X,Y} - \frac{1}{s} E_{Q^*_{X,Y}}\{\ell(x,\Phi(y)\} \\ &= -H(Q^*_{X,Y}|P_{X,Y}). \end{aligned} \tag{3.52}$$

8) Under the assumptions of 7), the relative entropy $H(Q^*_{X,Y}|P_{X,Y})$ is a non-increasing function of $s > 0$, that is,

$$0 \le H(Q^*_{X,Y}|P_{X,Y})|_{s=s_2} \le H(Q^*_{X,Y}|P_{X,Y})|_{s=s_1} \le H(Q^*_{X,Y}|P_{X,Y})|_{s=s^*} = R, \\ 0 < s^* \le s_1 \le s_2. \tag{3.53}$$

Proof. The derivations are similar to those in [43] hence they are omitted. □

It is pointed out that 4) indicates the concavity of dual function with respect to the uncertainty radius, while 5) shows that as $s \to \infty$ the estimation problem for a class of distribution will converge to an estimation problem for a single distribution. 6) provides lower and upper bounds on the optimal pay-off as a function of the nominal measure. 8) illustrates the non-increasing property of relative entropy as a function of s, hence this together with (3.51) gives an algorithm for finding the Lagrange multiplier s^*.

3.3 Examples from Estimation Theory

Next, some examples from estimation theory are presented to illustrate how the results of this chapter apply to practical problems.

3.3.1 Estimation of Random Variables

Suppose X and Y are RVs defined on $\Big(\Omega, \mathcal{F}, P\Big)$, which are related via the nominal model

$$Y = HX + W. \tag{3.54}$$

Hence $Y : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^d, \mathcal{B}(\Re^d)\Big)$ the observed RV, $X : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^n, \mathcal{B}(\Re^n)\Big)$ the unobserved RV, and $W : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^d, \mathcal{B}(\Re^d)\Big)$ the noise RV. Assume W and X are independent Gaussian Random Variables, $N(0, \Sigma_W)$, $\Sigma_W > 0$, $N(0, \Sigma_X)$, $\Sigma_X > 0$. Moreover, assume (3.54) denotes the nominal model. Then

$$\mu(x, dy) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_W|^{\frac{1}{2}}} \, e^{-(y-Hx)^T \frac{\Sigma_W^{-1}}{2}(y-Hx)} dy, \tag{3.55}$$

$$P_X(dx) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_X|^{\frac{1}{2}}} \, e^{-x^T \frac{\Sigma_X^{-1}}{2} x} dx. \tag{3.56}$$

Let

$$\ell(x, \Phi(y)) = (x - \Phi(y))^T U (x - \Phi(y)), \quad U = U^T > 0. \tag{3.57}$$

Using the above model four different estimation scenarios are investigated. Complete derivations of the results presented below can be found in Appendix C.

Application 1

Suppose the uncertainty is described by $\mathcal{B}^y(\eta) \stackrel{\triangle}{=} \{\nu \in \mathcal{P} : H(\nu|\eta)(y) \leq R(y)\}$. Then the worst case measure $\nu^*(y, dx)$ is given by (3.25). Using (3.55), (3.56) and (3.57) this worst case measure is given as

$$\nu^*(y, dx) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma^y(\tilde{s})|^{\frac{1}{2}}} \, e^{-(x-m^y(\tilde{s}))^T \frac{\Sigma^y(\tilde{s})^{-1}}{2}(x-m^y(\tilde{s}))} dx \tag{3.58}$$

where $\Sigma^y(\tilde{s}) = \Big(H^T \Sigma_W^{-1} H + \Sigma_X^{-1} - \frac{2U}{\tilde{s}(y)}\Big)^{-1}$ and $m^y(\tilde{s}) = \Sigma^y(\tilde{s})\Big(H^T \Sigma_W^{-1} y - \frac{2U}{\tilde{s}(y)} \Phi(y)\Big)$.

Moreover, using (3.27) together with Bayes' formula $\eta(y, dx) = \frac{\mu(x,dy) dP_X(x)}{\int_{\mathcal{X}} \mu(x,dy) dP_X(x)}$ the pay-off function is given as

$$L_2(\nu^*, \lambda^*, \tilde{s}^*) = \inf_{\Phi(\cdot)} \inf_{\tilde{s}(\cdot)} \left\{ -\frac{1}{2} \tilde{s}(y) \log \frac{|\Sigma - 2\frac{U}{\tilde{s}(y)}|}{|\Sigma|} + \tilde{s}(y) g(\tilde{s}, \Phi, y) + \tilde{s}(y) R(y) \right\} \tag{3.59}$$

where $\Sigma = H^T\Sigma_W^{-1}H + \Sigma_X^{-1}$ and

$$g(\tilde{s},\Phi,y) = \left(H^T\Sigma_W^{-1}y - \frac{2U}{\tilde{s}(y)}\Phi(y)\right)^T \frac{\left(\Sigma - \frac{2U}{\tilde{s}(y)}\right)^{-1}}{2} \left(H^T\Sigma_W^{-1}y - \frac{2U}{\tilde{s}(y)}\Phi(y)\right) + \Phi(y)^T\frac{U}{\tilde{s}(y)}\Phi(y) - y^T\Sigma_W^{-1}H\frac{\Sigma^{-1}}{2}H^T\Sigma_W^{-1}y.$$

$L_2(\nu^*,\lambda^*,\tilde{s})$ is a convex function of Φ for all $\tilde{s}(\cdot) > 0$ and a convex function of $\tilde{s}$ for all $\Phi(\cdot)$.

In order to calculate the best estimate $\Phi^*(y)$ of X, the above pay-off function has to be differentiated and the derivative set to zero as shown below.

$$\frac{d}{d\Phi}L_2(\nu^*,\lambda^*,\tilde{s})|_{\Phi=\Phi^*} = 0, \quad \forall s(\cdot). \tag{3.60}$$

By executing the above differentiation and after some manipulations the best estimate is given by

$$\Phi^*(y) = \left(H^T\Sigma_W^{-1}H + \Sigma_X^{-1}\right)^{-1}H^T\Sigma_W^{-1}y = \Sigma_X H^T\left(H\Sigma_X H^T + \Sigma_W\right)^{-1}y. \tag{3.61}$$

If it is assumed that the nominal model is the true model, the Least-Square estimate of X is given by [3]

$$\begin{aligned}\Phi(y) = E[X|Y=y] = \int_{\mathcal{X}} x\ \eta(y,dx) &= \left(H^T\Sigma_W^{-1}H + \Sigma_X^{-1}\right)^{-1}H^T\Sigma_W^{-1}y \\ &= \Sigma_X H^T\left(H\Sigma_X H^T + \Sigma_W\right)^{-1}y. \end{aligned}\tag{3.62}$$

This shows that the best estimate $\Phi^*(y)$ of X is equal with the expected value of X given Y. The reason for this result is the fact that random variables are being used.

Note that when Σ_W is very large compare to $H\Sigma_X H^T$, then $\Phi^*(y) = E[X] = 0$.

For this problem the minimax problem is equal to the maximin problem therefore, one can perform the minimization of (3.22) with respect to Φ to obtain $\Phi^*(y) = \int x d\nu(y,dx)$, while the maximization is given by $\nu^*(y,dx)$ given by (3.25) in which Φ is replaced by Φ^*. Performing the calculation $\Phi^*(y) = \int x d\nu^*(y,dx)$ in which both left and right side terms involve the term Φ^* and solving for this estimator yields (3.61).

Application 2

It is assumed that the uncertainty is described by

$$\mathcal{A}(\mu) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : \int_{\mathcal{X}} H(\nu|\mu)(x)dP_X(x) \leq \int_{\mathcal{X}} R(x)dP_X(x) \stackrel{\nabla}{=} r_1\Big\}.$$

Then the worst case measure $\nu^*(x, dy)$ is given by (3.15), where $\mu(x, dy)$ is given by (3.55). This case can be applied to estimation problems when the received signal is affected by a class of uncertain channels. Next the connection of relative entropy uncertainty to parametric uncertainty is described. Assume W and X are independent Gaussian Random Variables, $N(0, \Sigma_W)$, $\Sigma_W > 0$, $N(0, \Sigma_X)$, $\Sigma_X > 0$. Suppose the uncertain measure $\nu(x, dy)$ is induced by the following system.

$$Y = (H + \Delta H)X + W, \;\; W \sim N(0; \Sigma_W + \Delta\Sigma_W) \tag{3.63}$$

where ΔH denotes the uncertainty matrix to be defined shortly, and $\Delta\Sigma_W \geq 0$ denotes the noise uncertainty. Suppose $\mu(x, dy)$ corresponds to the nominal system (3.54) ($\Delta H = 0$, $\Delta\Sigma_W = 0$) and $\nu(x, dy)$ corresponds to the uncertain system ($\Delta H \neq 0$, $\Delta\Sigma_W \neq 0$). Then the relative entropy between $\mu(x, dy)$ and $\nu(x, dy)$ is given by the following expression:

$$\begin{aligned} H(\nu|\mu)(x) &= \int \log\Big(\frac{\nu(x, dy)}{\mu(x, dy)}\Big)\nu(x, dy) \\ &= \frac{1}{2}\Big\{\log\frac{|\Sigma_W|}{|\Sigma_W + \Delta\Sigma_W|} + tr((\Sigma_W + \Delta\Sigma_W)(\Sigma_W^{-1} - (\Sigma_W + \Delta\Sigma_W)^{-1})) \\ &\quad + x^T(\Delta H)^T\Sigma_W^{-1}(\Delta H)x\Big\}. \end{aligned} \tag{3.64}$$

Taking the average with respect to x,

$$\begin{aligned} \int_{\mathcal{X}} H(\nu|\mu)(x)dP_X(x) &= \frac{1}{2}\Big\{\log\frac{|\Sigma_W|}{|\Sigma_W + \Delta\Sigma_W|} \\ &\quad + tr((\Sigma_W + \Delta\Sigma_W)(\Sigma_W^{-1} - (\Sigma_W + \Delta\Sigma_W)^{-1})) \\ &\quad + tr\Big(\Sigma_X(\Delta H)^T\Sigma_W^{-1}\Delta H\Big)\Big\}. \end{aligned} \tag{3.65}$$

Now, in the special case when $\Delta\Sigma_W = 0$ the previous expression is equivalent to the weighted norm

$$||\Delta H||_N^2 = \frac{1}{2}tr\Big(\Sigma_X^{\frac{1}{2}}(\Delta H)^T\Sigma_W^{-1}\Delta H\Sigma_X^{\frac{1}{2}}\Big). \tag{3.66}$$

The conclusion by using the above expression is that the norm uncertainty defined by (3.66) is a special case of the relative entropy uncertainty. This can be seen if the following uncertainty set of channels is defined:

$$\mathcal{N} = \Big\{ Z \text{ is a channel characterized by } H + \Delta H, W \sim N(0; \Sigma_W) : \sqrt{2}||\Delta H||_N \leq \sqrt{r_1} \Big\}.$$

Then the output distribution of every channel in $\mathcal{N}$ belongs to $\mathcal{A}(\mu)$, and in order to solve the estimation problem over the whole class of channels $\mathcal{N}$ (given the observations Y), one can solve the original problem as posed in Section 3.2.2. Notice that $\mathcal{N} \subseteq \mathcal{A}(\mu)$ because relative entropy uncertainty allows more general models than (3.63) with $\Delta\Sigma_W = 0$. For example, the uncertain measure noise W can be different from the noise of the nominal channel in (3.54).

Application 3

Suppose the uncertainty is described by

$$\mathcal{C}(P_{X,Y}) \stackrel{\triangle}{=} \{Q_{X,Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) : H(Q_{X,Y}|P_{X,Y}) \leq R\}.$$

The worst case measure $dQ^*_{X,Y}(x,y)$ is given by (3.43). Then using (3.55), (3.56) and (3.57) this worst case measure is given by

$$dQ^*_{X,Y}(x,y) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma(s)|^{\frac{1}{2}}}\, e^{-(x-\kappa^y(s))^T \frac{\Sigma(s)^{-1}}{2}(x-\kappa^y(s))} \frac{e^{\theta(s,\Phi,y)}}{\int_{\mathcal{Y}} e^{\theta(s,\Phi,y)}dy} dydx \qquad (3.67)$$

where

$$\Sigma(s) = \Big(H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - 2\frac{U}{s}\Big)^{-1}$$
$$\kappa^y(s) = \Sigma(s)\Big(H^T\Sigma_W^{-1}y - \frac{2U}{s}\Phi(y)\Big)$$

and

$$\theta(s,\Phi,y) = \kappa^y(s)^T\frac{\Sigma(s)^{-1}}{2}\kappa^y(s) + \Phi(y)^T\frac{U}{s}\Phi(y) - y^T\frac{\Sigma_W^{-1}}{2}y.$$

The average pay-off is given by (3.42). Notice that $L_4(Q^*, \lambda^*, s)$ is a convex function of Φ for all $s > 0$ and a convex function of s for all $\Phi(\cdot)$. So in order to calculate the best estimate $\Phi^*(Y)$ of X the average pay-off (3.42) has to be differentiated and the derivative set to zero as follows:

$$\frac{d}{d\Phi}L_4(Q*, \lambda^*, s)\Big|_{\Phi=\Phi^*} = 0, \quad \forall s. \qquad (3.68)$$

Performing the above differentiation, and using Bayes' formula
$dP_{X,Y}(x,y) = \eta(y,dx)dP_Y(y) = \mu(x,dy)dP_X(x)$ the best estimate is given by

$$\Phi^*(y) = \frac{\int_{\mathcal{X}} x e^{\frac{\ell(x,\Phi^*(y))}{s}} \eta(y,dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s}} \eta(y,dx)}. \tag{3.69}$$

After some manipulations the best estimate $\Phi^*(Y)$ of X is given by

$$\Phi^*(y) = \Big(H^T\Sigma_W^{-1}H + \Sigma_X^{-1}\Big)^{-1} H^T\Sigma_W^{-1}y = \Sigma_X H^T\Big(H\Sigma_X H^T + \Sigma_W\Big)^{-1} y. \tag{3.70}$$

Notice that the $\Phi^*(Y)$ is the same as (3.61) which is the expected value of X given Y.

Application 4

Suppose the uncertainty is described by

$$\mathcal{B}^R(\eta) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : \int_{\mathcal{Y}} H(\nu|\eta)(y)dP_Y(y) \leq \int_{\mathcal{X}\times\mathcal{Y}} R(x,y)\nu(y,dx)dP_Y(y) + \overline{R} \stackrel{\nabla}{=} r_3\Big\}$$

and assume that

$$R(x,y) = (y-\bar{H}x)^T\tilde{U}(y-\bar{H}x),\ \tilde{U} = \tilde{U}^T > 0. \tag{3.71}$$

Then the worst case measure $\nu^*(y,dx)$ is given by (3.37). Using (3.55), (3.56), (3.57) and (3.71) this worst case measure is given by

$$\nu^*(y,dx) = \frac{1}{(2\pi)^{\frac{n}{2}}|\tilde{\Sigma}(s)|^{\frac{1}{2}}}\ e^{-(x-\tilde{m}^y(s))^T\frac{\tilde{\Sigma}(s)^{-1}}{2}(x-\tilde{m}^y(s))}dx \tag{3.72}$$

where

$$\begin{aligned}\tilde{\Sigma}(s) &= \Big(H^T\Sigma_W^{-1}H + \Sigma_X^{-1} + 2\bar{H}^T\tilde{U}\bar{H} - \frac{2U}{s}\Big)^{-1},\\ \tilde{m}^y(s) &= \tilde{\Sigma}(s)\Big(H^T\Sigma_W^{-1}y + 2\bar{H}^T\tilde{U}y - \frac{2U}{s}\Phi(y)\Big).\end{aligned}$$

The average pay-off function is given by (3.39). Like the previous application this pay-off function, $L_3^R(\nu^*,\lambda^*,s)$, is a convex function of Φ for all $s > 0$ and a convex function of s for all $\Phi(\cdot)$.

Next, by differentiating the average pay-off (3.39) and setting the derivative to zero the best estimate $\Phi^*(Y)$ of X is given by

$$\frac{d}{d\Phi}L_3^R(\nu^*,\lambda^*,s)\Big|_{\Phi=\Phi^*} = 0 \quad\Rightarrow\quad \Phi^*(y) = \frac{\int_{\mathcal{X}} x e^{\frac{\ell(x,\Phi^*(y))}{s} - R(x,y)}\eta(y,dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s} - R(x,y)}\eta(y,dx)}. \tag{3.73}$$

Finally, after some manipulations the best estimate $\Phi^*(Y)$ of X is given by

$$\Phi^*(y) = \Big(H^T\Sigma_W^{-1}H + \Sigma_X^{-1} + 2\bar{H}^T\tilde{U}\bar{H}\Big)^{-1}\Big(H^T\Sigma_W^{-1} + 2\bar{H}^T\tilde{U}\Big)y. \tag{3.74}$$

Remark 3.3.1. *The best estimate $\Phi^*(Y)$ of X is given by (3.74). Then,*

i. if it is assumed that $\bar{H} = H$,

$$\begin{aligned}\Phi^*(y) &= \Big(H^T\Sigma_W^{-1}H + \Sigma_X^{-1} + 2H^T\tilde{U}H\Big)^{-1}\Big(H^T\Sigma_W^{-1} + 2H^T\tilde{U}\Big)y \\ &= \Big(H^T(\Sigma_W^{-1} + 2\tilde{U})H + \Sigma_X^{-1}\Big)^{-1}H^T(\Sigma_W^{-1} + 2\tilde{U})y;\end{aligned} \tag{3.75}$$

ii. if it is assumed that $\bar{H} = I$,

$$\Phi^*(y) = \Big(H^T\Sigma_W^{-1}H + \Sigma_X^{-1} + 2\tilde{U}\Big)^{-1}\Big(\Sigma_W^{-1} + 2\tilde{U}\Big)y. \tag{3.76}$$

3.3.2 Estimation of a Sequence of Random Variables

Suppose X is a RV and $Y^m = \{Y_0, Y_1, \ldots, Y_m\}$ is a sequence of RVs defined on $\Big(\Omega, \mathcal{F}, P\Big)$, which are related via the nominal model

$$Y_i = HX + W_i, \quad i = 0, \ldots, m \tag{3.77}$$

Hence $Y^m : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^{(m+1)d}, \mathcal{B}(\Re^{(m+1)d})\Big)$ the observed sequence of RVs, $X : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^n, \mathcal{B}(\Re^n)\Big)$ the unobserved RV, and $W^m : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^{(m+1)d}, \mathcal{B}(\Re^{(m+1)d})\Big)$ the noise sequence of RVs. Assume W_i and X are independent Gaussian Random Variables, $N(0, \Sigma_W)$, $\Sigma_W > 0$, $N(0, \Sigma_X)$, $\Sigma_X > 0$. Moreover, assume (3.77) denotes the nominal model. Then

$$\mu(x, dy^m) = \prod_{i=0}^{m} \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_W|^{\frac{1}{2}}}\, e^{-(y_i - Hx)^T\frac{\Sigma_W^{-1}}{2}(y_i - Hx)}dy_i, \tag{3.78}$$

$$P_X(dx) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_X|^{\frac{1}{2}}}\, e^{-x^T\frac{\Sigma_X^{-1}}{2}x}dx. \tag{3.79}$$

Let

$$\begin{aligned}&\ell(x, \Phi(y^m)) = \sum_{i=0}^{m-1}(x - \Phi_i^*(y^i))^T U(x - \Phi_i^*(y^i)) + (x - \Phi_m(y^m))^T U(x - \Phi_m(y^m)) \\ &U = U^T > 0.\end{aligned} \tag{3.80}$$

Suppose the uncertainty is described by $\mathcal{B}^y(\eta) \stackrel{\triangle}{=} \{\nu \in \mathcal{P} : H(\nu|\eta)(y) \leq R(y)\}$. Then the worst case measure $\nu^*(y^m, dx)$ is given by (3.25). Using (3.78), (3.79) and (3.80) this worst case measure can be expressed as

$$\nu^*(y^m, dx) = \frac{1}{(2\pi)^{\frac{n}{2}} |\widehat{\Sigma}^y(\tilde{s})|^{\frac{1}{2}}} \, e^{-(x-\widehat{m}^y(\tilde{s}))^T \frac{\widehat{\Sigma}^y(\tilde{s})^{-1}}{2}(x-m^y(\tilde{s}))} dx \tag{3.81}$$

where

$$\widehat{\Sigma}^y(\tilde{s}) = \left((m+1) H^T \Sigma_W^{-1} H + \Sigma_X^{-1} - \frac{2(m+1)U}{\tilde{s}(y)} \right)^{-1},$$
$$\widehat{m}^y(\tilde{s}) = \widehat{\Sigma}^y(\tilde{s}) \left(H^T \Sigma_W^{-1} \sum_{i=0}^{m} y_i - \frac{2U}{\tilde{s}(y)} \sum_{i=0}^{m-1} \Phi_i^*(y^i) - \frac{2U\Phi_m(y^m)}{\tilde{s}(y)} \right).$$

Note that $\widehat{\Sigma}^y(\tilde{s})$ is a valid covariance provided it is positive semidefinite. Hence,

$$\begin{aligned}
&\widehat{\Sigma}^y(\tilde{s}) = \left((m+1) H^T \Sigma_W^{-1} H + \Sigma_X^{-1} - \frac{2(m+1)U}{\tilde{s}(y)} \right)^{-1} > 0 \\
&\Rightarrow \left(2(m+1)U \Big(\frac{U^{-1}}{2} H^T \Sigma_W^{-1} H + \frac{U^{-1}}{2(m+1)} \Sigma_X^{-1} - \frac{I}{\tilde{s}(y)} \Big) \right)^{-1} > 0 \\
&\Rightarrow \left(2(m+1)U \Big(A - \frac{I}{\tilde{s}(y)} \Big) \right)^{-1} > 0, \;\; A \stackrel{\triangle}{=} \frac{U^{-1}}{2} H^T \Sigma_W^{-1} H + \frac{U^{-1}}{2(m+1)} \Sigma_X^{-1} \\
&\Rightarrow A - \frac{I}{\tilde{s}(y)} > 0.
\end{aligned} \tag{3.82}$$

Given that A is a square matrix (so it can be written in the form $A = V\Lambda V^{-1}$), then (3.82) can be formulated as

$$V\Lambda V^{-1} - \frac{1}{s} V V^{-1} > 0 \;\; \Rightarrow V(\Lambda - \frac{1}{s}) V^{-1} > 0.$$

Therefore, $s > \frac{1}{\lambda_{min}}$, where λ_{min} is the smallest eigenvalue of A.

The average pay-off function $L_2(\nu^*, \lambda^*, s)$ given by (3.27) is a convex function of $\Phi_m(y^m)$ for all $\tilde{s}(\cdot) > 0$ and a convex function of $\tilde{s}$ for all $\Phi(\cdot)$. Similar with previous applications presented, the best estimate $\Phi_m^*(y^m)$ of X is calculated by differentiating the average pay-off (3.27) and setting the derivative to zero

$$\frac{d}{d\Phi_m(y^m)} L_2(\nu^*, \lambda^*, \tilde{s}) \Big|_{\Phi_m(y^m) = \Phi_m^*(y^m)} = 0 \;\; \Rightarrow \;\; \Phi_m^*(y^m) = \frac{\int_{\mathcal{X}} x e^{\frac{\ell(x, \Phi_m^*(y^m))}{\tilde{s}}} \eta(y^m, dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x, \Phi_m^*(y^m))}{\tilde{s}}} \eta(y^m, dx)}. \tag{3.83}$$

By further manipulating the above expression, the best estimate is given by

$$\Phi_m^*(y^m) = \left((m+1)H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - \frac{2mU}{\tilde{s}(y)}\right)^{-1}\left(H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i - \frac{2U}{\tilde{s}(y)}\sum_{i=0}^{m-1}\Phi_i^*(y^i)\right) \tag{3.84}$$

Clearly, (3.84) is a nonlinear estimator because $\left((m+1)H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - \frac{2mU}{\tilde{s}(y)}\right)$ is a function of the observed sequence of RVs Y^m. When $\tilde{s}(y) \to \infty$, then

$$\Phi^*(y^m) = \left((m+1)H^T\Sigma_W^{-1}H + \Sigma_X^{-1}\right)^{-1} H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i = E[X|Y^m = y^m]. \tag{3.85}$$

Note that, complete derivations of the above results can be found in Appendix C.

3.4 Examples from MIMO Communication Systems

In this section the theory developed is applied to Multiple-Input Multiple-Output communication systems. A short overview of the MIMO technology is introduced, and then examples are presented in order to illustrate the applicability of the results to MIMO communication systems.

3.4.1 Overview of MIMO Communication Systems

In radio communication, Multiple-Input and Multiple-Output (MIMO) systems employ multiple antennas at both the transmitter and receiver to improve communication performance. It is one of several forms of smart antenna technology. MIMO technology has attracted attention in wireless communications, since it offers significant increases in data throughput and link range without additional bandwidth or transmit power. They achieve higher spectral efficiency and link reliability or diversity.

Wireless MIMO systems are capable of delivering large increases in capacity through utilization of parallel communication channels [48], [49], [42]. Appearing first in a series of information theory articles published by members of Bell Labs, MIMO

systems now constitute a major research area in telecommunications. It is also considered to be one of the technologies that have a chance to resolve the bottlenecks of traffic capacity in the forthcoming broadband wireless Internet access networks. Multiple antennas, both at the transmitter and the receiver, create a matrix channel. The key advantage is the possibility of transmitting over several spatial modes of the channel matrix within the same time-frequency slot at no additional power expenditure. In addition, if the channel matrix is known both at the transmitter (TX) and the receiver (RX), certain spatial modes (singular modes) of the matrix channel can be used to maximize the SNR for every realization of the channel. The singular modes can be used to transport independent data streams (to increase data rate), or one may choose to exploit the top mode (associated with the largest singular value) in order to maximize the spatial diversity advantage.

As mentioned above, Multiple-Input Multiple-Output technology has emerged recently as one of the most significant technologies in modern communication. By using MIMO technology an increase in the system capacity and/or an improvement in the quality of service can be achieved. The key to fully utilize the MIMO capacity relies heavily on the requirement of accurate channel estimation. MIMO channel estimation methods can be classified into three categories: training-based methods, blind methods and semi-blind methods. For pure training-based schemes, a long training is necessary in order to obtain a reliable MIMO channel estimate which reduces the system bandwidth efficiency considerably. Blind methods which do not require any training symbols achieve high system throughput at the expense of high computational complexity. Semi-blind schemes on the other hand require less computational complexity than blind methods and fewer training symbols than training-based methods, making them attractive for practical implementation.

In this section the theory developed in previous sections of this chapter is implemented in order to solve some simple examples for MIMO communication systems.

3.4.2 Estimation from MIMO communication Systems

A complex representation of a MIMO communication channel with n transmitting and d receiving antennas, is considered

$$Y = HX + W \tag{3.86}$$

where $Y \in \mathbb{C}^d$ is the received signal, $X \in \mathbb{C}^n$ is the transmitted signal, $H \in \mathbb{C}^{d\times n}$ is the channel matrix, and W is the zero-mean circularly-symmetric Gaussian noise independent of the transmitted signal X. The covariance matrices of X and W are denoted by Σ_X and Σ_W, respectively.

Furthermore, it is assumed that the following are known

$$\mu(x, dy) = \pi^{-d}|\Sigma_W|^{-1}\ e^{-(y-Hx)^\dagger \Sigma_W^{-1}(y-Hx)}dy, \tag{3.87}$$

$$P_X(dx) = \pi^{-n}|\Sigma_X|^{-1}\ e^{-x^\dagger \Sigma_X^{-1}x}dx. \tag{3.88}$$

Let

$$\ell(x, \Phi(y)) = (x - \Phi(y))^\dagger U(x - \Phi(y)),\ \ U = U^\dagger > 0. \tag{3.89}$$

$\dagger$ denotes complex conjugate transpose, $\Sigma_W \in \mathbb{C}^{d\times d}, \Sigma_W = \Sigma_W^\dagger \geq 0$, $\Sigma_X \in \mathbb{C}^{n\times n}, \Sigma_X = \Sigma_X^\dagger \geq 0$.

Using the above model two different estimation scenarios are being investigated. Complete derivations of the results presented below can be found in Appendix C.

Application 1

Suppose the uncertainty is described by

$$\mathcal{B}^R(\eta) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : \int_{\mathcal{Y}} H(\nu|\eta)(y)dP_Y(y) \leq \int_{\mathcal{X}\times\mathcal{Y}} R(x, y)\nu(y, dx)dP_Y(y) + \overline{R} \stackrel{\nabla}{=} r_3\Big\}$$

and assume that

$$R(x, y) = (y - \bar{H}x)^\dagger \tilde{U}(y - \bar{H}x),\ \ \tilde{U} = \tilde{U}^\dagger > 0. \tag{3.90}$$

Then the worst case measure $\nu^*(y, dx)$ is given by (3.37). Using (3.87), (3.88), (3.89) and (3.90) this worst case measure can be expressed as

$$\nu^*(y, dx) = \pi^{-n}|\tilde{\Sigma}_M(s)|^{-1}\ e^{-(x-\tilde{m}_M(s))^\dagger \tilde{\Sigma}_M(s)^{-1}(x-\tilde{m}_M(s))}dx \tag{3.91}$$

where

$$\begin{aligned}\tilde{\Sigma}_M(s) &= \Big(H^\dagger\Sigma_W^{-1}H + \Sigma_X^{-1} + \bar{H}^\dagger\tilde{U}\bar{H} - \frac{U}{s}\Big)^{-1}\\ \tilde{m}_M(s) &= \tilde{\Sigma}(s)\Big(H^\dagger\Sigma_W^{-1}y + \bar{H}^\dagger\tilde{U}y - \frac{U}{s}\Phi(y)\Big).\end{aligned}$$

Notice that the pay-off function $L_3^R(\nu^*, \lambda^*, s)$, given by (3.39), is a convex function of Φ for all $s > 0$ and a convex function of s for all $\Phi(\cdot)$. Therefore, by differentiating the average pay-off (3.39) and setting the derivative to zero the best estimate $\Phi^*(Y)$ of X is given by

$$\frac{d}{d\Phi} L_3^R(\nu^*, \lambda^*, s)\Big|_{\Phi=\Phi^*} = 0 \quad \Rightarrow \quad \Phi^*(y) = \frac{\int_{\mathcal{X}} x e^{\frac{\ell(x,\Phi^*(y))}{s} - R(x,y)} \eta(y, dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s} - R(x,y)} \eta(y, dx)}. \tag{3.92}$$

Finally, after some manipulations of the above expression, the best estimate is given by

$$\Phi^*(y) = \Big(H^\dagger \Sigma_W^{-1} H + \Sigma_X^{-1} + \bar{H}^\dagger \tilde{U} \bar{H}\Big)^{-1} \Big(H^\dagger \Sigma_W^{-1} + \bar{H}^\dagger \tilde{U}\Big) y. \tag{3.93}$$

Remark 3.4.1. *The best estimate $\Phi^*(Y)$ of X is given by (3.93). Then,*

i. if it is assumed that $\bar{H} = H$,

$$\begin{aligned} \Phi^*(y) &= \Big(H^\dagger \Sigma_W^{-1} H + \Sigma_X^{-1} + H^\dagger \tilde{U} H\Big)^{-1} \Big(H^\dagger \Sigma_W^{-1} + H^\dagger \tilde{U}\Big) y \\ &= \Big(H^\dagger (\Sigma_W^{-1} + \tilde{U}) H + \Sigma_X^{-1}\Big)^{-1} H^\dagger (\Sigma_W^{-1} + \tilde{U}) y; \end{aligned} \tag{3.94}$$

ii. if it is assumed that $\bar{H} = I$,

$$\Phi^*(y) = \Big(H^\dagger \Sigma_W^{-1} H + \Sigma_X^{-1} + \tilde{U}\Big)^{-1} \Big(H^\dagger \Sigma_W^{-1} + \tilde{U}\Big) y. \tag{3.95}$$

Application 2

Suppose the uncertainty is described by

$$\mathcal{C}(P_{X,Y}) \stackrel{\triangle}{=} \{Q_{X,Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) : H(Q_{X,Y} | P_{X,Y}) \leq R\}.$$

The worst case measure $Q_{X,Y}^*(x, y)$ is given by (3.43). Then using (3.87), (3.88) and (3.89) this worst case measure is given by

$$dQ_{X,Y}^*(x, y) = (2\pi)^{-n} |\Sigma_M(s)|^{-1} \; e^{-(x-\kappa_M^y(s))^\dagger \Sigma_M(s)^{-1} (x-\kappa_M^y(s))} \frac{e^{\tilde{\theta}(s,\Phi,y)}}{\int_{\mathcal{Y}} e^{\tilde{\theta}(s,\Phi,y)} dy} dy dx \tag{3.96}$$

where

$$\begin{aligned} \Sigma_M(s) &= \Big(H^\dagger \Sigma_W^{-1} H + \Sigma_X^{-1} - \frac{U}{s}\Big)^{-1} \\ \kappa_M^y(s) &= \Sigma_M(s) \Big(H^\dagger \Sigma_W^{-1} y - \frac{U}{s} \Phi(y)\Big) \end{aligned}$$

and

$$\tilde{\theta}(s,\Phi,y)=\kappa_M^y(s)^\dagger\Sigma_M(s)^{-1}\kappa_M^y(s)+\Phi(y)^\dagger\frac{U}{s}\Phi(y)-y^\dagger\Sigma_W^{-1}y.$$

Given that the average pay-off $L_4(\nu^*,\lambda^*,s)$, given by (3.42) is a convex function of Φ for all $s>0$ and a convex function of s for all $\Phi(\cdot)$, in order to calculate the best estimate $\Phi^*(Y)$ of X, one just has to differentiate the average pay-off (3.42) and set the derivative to zero as follows

$$\frac{d}{d\Phi}L_4(\nu^*,\lambda^*,s)\Big|_{\Phi=\Phi^*}=0\quad\Rightarrow\quad\Phi^*(y)=\frac{\int_{\mathcal{X}}xe^{\frac{\ell(x,\Phi^*(y))}{s}}\eta(y,dx)}{\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi^*(y))}{s}}\eta(y,dx)}.\tag{3.97}$$

Then, after some calculations and manipulations the best estimate $\Phi^*(Y)$ of X is given by

$$\Phi^*(y)=\Big(H^\dagger\Sigma_W^{-1}H+\Sigma_X^{-1}\Big)^{-1}H^\dagger\Sigma_W^{-1}y=\Sigma_XH^\dagger\Big(H\Sigma_XH^\dagger+\Sigma_W\Big)^{-1}y.\tag{3.98}$$

3.5 Summary

This chapter considers the problem of least-square estimation for a class of systems which are subject to uncertainty, and employs the KL distance to describe the uncertainty classes. Stochastic kernels, and joint distributions are used to described the uncertainty models and a minimax approach is implemented. Three problems are formulated and their solutions are sought, highlighting some properties associated with the estimate of the true distribution. Classical examples are chosen to illustrate the applicability of the results. The theory developed is also applied to simple examples for MIMO communication systems.

CHAPTER 4

APPLICATIONS OF ROBUST ESTIMATION

In this chapter, the theory developed in Chapter 3 is applied to a nonlinear recursive model in order to derive robust estimators. The methodology presented invokes a change of probability measure technique to derive recursive equations for the conditional density of nonlinear filtering problems. In section 4.2, the mathematical theory is presented by extending the theory of Chapter 3. In Section 4.3, the theory developed in Section 4.2 is applied to a linear Gaussian example. While in Section 4.4, the theory is applied to an attenuated sinusoid in a multipath environment, where various estimators are derived by solving the recursive equation satisfied by the unnormalized version of the a posteriori density.

4.1 Introduction

Chapter 4 deals with estimation techniques, for estimating signals which are generated by finite-dimensional autoregressive channel models found in [34], [35], subject to uncertainty. Autoregressive channel models have been used with success to predict fading channel dynamics for the purposes of Kalman filter based channel estimation and for long-range channel forecasting. They have also been used by several authors to simulate correlated Rayleigh fading. The uncertainty description of these systems is characterized by the class of uncertain measures which satisfy a Kullback-Leibler (KL) distance constraint with respect to a nominal measure. The uncertainty description of the system, and the nominal description of the system are modeled by joint probability distributions.

A minimax approach is implemented here in order to address the above problem. The difficulty with this approach lies in the minimization. Because this is a nonlinear problem it is not easy to obtain the solution of the $\acute{a}$ posteriori distribution recursion in closed form. Very often, when dealing with nonlinear estimation problems, approximations have to be made to find sub-optimal nonlinear estimators. A common method is to use the so-called extended Kalman filter [5, 6, 7]. Other more sophisticated sub-optimal estimation techniques are available, e.g., reiteration, higher order filters, and statistical linearization [5]. The chapter deals specifically with the problem of finding a solution to the minimax formulation and does not take into consideration other nonlinear estimation techniques.

In general, minimax estimation techniques lead to strategies, which are robust with respect to variations in the models as long as these belong to the uncertainty class. As was already described in Chapter 1 Minimax Wiener filtering techniques are given in [30], while blind minimax linear regression problems of estimating deterministic parameters are given in [23]. Relations of blind minimax techniques to Stein type estimators [27] and least-square regularization [9] is discussed in [23]. Related work in which uncertainty is described by relative entropy can be found in [16], [29], [50]. The examples provided in [16], [23], [29] and [50] are linear Gaussian.

In this chapter, in the abstract setting, the maximization is addressed using variational methods, while the minimization is addressed using a change of probability measure technique [36]. The change of probability measure techniques introduced is being used in order to derive a recursive equation for the unnormalized $\acute{a}$ posteriori distribution of nonlinear filtering problems. The theory developed is applied first to a linear Gaussian model and then to an attenuated sinusoid in a multipath environment, which is subject to an additive Gaussian noise. The multipath model employed for non-coherent estimation and detection compliments the work found in the literature [4], [51], in the sense that the classical problem assumes no multipath scenario, while the attenuation of the sinusoidal is assumed to be a known deterministic function. Note that, unlike the classical non-coherent estimation problem [1], [4], which is concerned with a fixed model, this chapter deals with a class of models while the estimation is formulated using minimax techniques. The connection to least-squares estimation found in [1], [4] for single channel, is established by reducing the uncertainty to zero, while generalizing existing results.

4.2 The Minimax Filtering

In this section, the nonlinear model is described. The minimax problem is defined and a change of probability measure technique is introduced, which reformulates the problem under a new fictitious probability measure, where the signal to be estimated and the observations are independent. Finally, a linear recursion for an unnormalized conditional density related to the minimax filtering is derived.

4.2.1 State and Observation Models

Let $\Big(\Omega, \mathcal{F}, P\Big)$ be complete probability space on which the nominal state or unobserved process $\{x_k\}, k \in N_0 \stackrel{\triangle}{=} \{0, 1, 2, 3, \ldots\}$ and the observation process $\{y_k\}, k \in N_0$, are defined by the following recursions:

$$\begin{aligned} x_{k+1} &= f(k+1, x_k) + B_{k+1} w_{k+1}, \quad x_0 \in \Re^n \\ y_k &= h(k, x_k) + D_k v_k, \quad y_0 \in \Re^d. \end{aligned} \tag{4.1}$$

Here $x_0 : \Omega \to \Re^n$ is the initial state and $w : \Omega \times N_0 \to \Re^n, \quad v : \Omega \times N_0 \to \Re^d$, are random noises.

Also $\{w_k\}, \{v_k\}, k \in N_0$ are independent noise sequences of Random Variables (RVs) with densities $\Psi_w(w) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{w^T w}{2}}$, $\Xi_v(v) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{v^T v}{2}}$, respectively, x_0 has density $\pi_{x_0}(x) = \frac{d\Pi_{x_0}(x)}{dx}$, which is also independent of $\{w_k\}$, $\{v_k\}$.

It is assumed throughout this chapter that $f : N_0 \times \Re^n \to \Re^n$ and $h : N_0 \times \Re^d \to \Re^d$, B_k, D_k are Borel measurable functions, and $(B_k B_k^T)^{-1}$, $(D_k D_k^T)^{-1}$ exist.

4.2.2 Definition of Minimax Problem

Let $\{\mathcal{G}_m^0\}$ be the σ-field generated by the complete data $\{x_0, x_1, \ldots, x_m, y_0, y_1, \ldots, y_m\}$ and let $\{\mathcal{G}_m\}, m \in N_0$ denote its complete filtration [36]. Let $\{\mathcal{F}_m^{0,\mathcal{Y}}\}$ define the σ-field generated by the incomplete data $\{y_0, y_1, \ldots, y_m\}$ and let $\{\mathcal{F}_m^{\mathcal{Y}}\}, m \in N_0$ denote its complete filtration. Let y^m denote the sequence $\{y_0, \ldots, y_m\}$ and similarly for other sequences. $\tilde{x}_m$ denotes the estimate of the state x_m given $\{\mathcal{F}_m^{\mathcal{Y}}\}, m \in N_0$; it is assumed recursive estimates which update $\tilde{x}_m$ from knowledge of $\tilde{x}_{m-1}$ and

past and present data $\{y_0, \ldots, y_m\}$, $m \in N_0$. Throughout the chapter $E_Q[.]$ denotes expectation with respect to probability distribution Q.

In the next definition the class of admissible estimators, which are functions of the observation sequences are introduced.

Definition 4.2.1. *The set of admissible estimators $\mathcal{X}_{ad}$ are defined as follows.*

$\mathcal{X}_{ad} \triangleq \left\{\tilde{x} : \Omega \times N_0 \to X \subseteq \Re^n;\ \{\tilde{x}_k\}, \text{ is adapted to } \{\mathcal{F}_k^{\mathcal{Y}}\}, k \in N_0\right\}$ *(e.g., adapted means that at each $k \in N_0$, $\tilde{x}_k$ is a causal function of the data $\{\mathcal{F}_k^{\mathcal{Y}}\}$).*

Let P_{x^m,y^m} denote the nominal (in the absence of modelling uncertainties) joint distribution of the sequences (x^m, y^m), which corresponds to the one induced by model (4.1). Let Q_{x^m,y^m} denote the true joint distribution of the sequences (x^m, y^m), which is unknown. The only available information is that Q_{x^m,y^m} belongs to a class of possible distributions. This class is modelled by the information theoretic relative entropy set

$$\mathcal{C}(P_{x^m,y^m}) \triangleq \{Q_{x^m,y^m} : H(Q_{x^m,y^m}|P_{x^m,y^m}) \leq R\} \tag{4.2}$$

where $R > 0$ and $H(\cdot|\cdot)$ is the KL distance between the two joint distributions defined in Section 3.2.4. Note that this is the same uncertainty set used in Section 3.2.4.

The pay-off is the average of a function of the error over the time horizon $[0, m]$, $E_{Q_{x^m,y^m}}\left\{\sum_{k=0}^m \tilde{\ell}(x_k, \tilde{x}_k)\right\}$, in which the average is taken with the unknown distribution $Q_{x^m,y^m} \in \mathcal{C}(P_{x^m,y^m})$. Here $\tilde{\ell}(x_k, \tilde{x}_k)$ is a measure of distance between the state x_k and its estimate $\tilde{x}_k$ (e.g. $||x_k - \tilde{x}_k||^2_{\Re^n}$).

Next, the minimax estimation problem is defined.

Problem 4.2.1. *Given the nominal probability distribution P_{x^m,y^m} induced by system (4.1), find a probability distribution $Q^*_{x^m,y^m}$ and an estimator $\tilde{x}^*_m \in \mathcal{X}_{ad}$ which solve*

$$J(\tilde{x}^*, Q^*_{x^m,y^m}) = \inf_{\tilde{x}_m \in \mathcal{X}_{ad}} \sup_{Q_{x^m,y^m} \in \mathcal{C}(P_{x^m,y^m})} E_{Q_{x^m,y^m}}\left\{\sum_{k=0}^m \tilde{\ell}(x_k, \tilde{x}_k)\right\} \tag{4.3}$$

when $R \in (0, \infty)$, and $\tilde{\ell} : \Re^n \times \Re^n \to [0, \infty)$, is continuous in $(x, \tilde{x}) \in \Re^n \times \Re^n$ and bounded from bellow.

The assumptions on $\tilde{\ell}$ are sufficient for the existence of estimates of the maximizing distribution $Q^*_{x^m,y^m}$ [43].

4.2.3 Minimax Optimization

Define the sample pay-off $\ell(x,\tilde{x}) \stackrel{\triangle}{=} \sum_{k=0}^{m} \tilde{\ell}(x_k,\tilde{x}_k)$. For a given $\tilde{x}_m \in \mathcal{X}_{ad}$, and a fixed nominal probability distribution P_{x^m,y^m} induced by the nominal system (4.1), let $Q^*_{x^m,y^m}$ denote the distribution which achieves the supremum of the average energy pay-off functional subject to the relative entropy uncertainty as defined by

$$J(\tilde{x}_m, Q^*_{x^m,y^m}) = \sup_{Q_{x^m,y^m}\in\mathcal{C}(P_{x^m,y^m})} \int_{\Re^{(m+1)n}\times\Re^{(m+1)d}} \ell(x^m,\tilde{x}^m) dQ_{x^m,y^m}(x,y). \tag{4.4}$$

The solution to this optimization problem returns the worst case probability distribution among those which satisfy the constraint, as a function of the nominal distribution P_{x^m,y^m}.

Thus, the estimation problem for the class of models $\mathcal{C}(P_{x^m,y^m})$, is to find an estimator $\tilde{x}^*_m \in \mathcal{X}_{ad}$ which solves

$$J(\tilde{x}^*_m) = \inf_{\tilde{x}_m\in\mathcal{X}_{ad}} J(\tilde{x}_m, Q^*_{x^m,y^m}). \tag{4.5}$$

Since (4.4) is a constraint optimization, to find the supremum over $Q_{x^m,y^m} \in \mathcal{C}(P_{x^m,y^m})$ the Lagrangian has to be defined [40]

$$\begin{aligned} L(Q^*_{x^m,y^m},\tilde{x}_m,\lambda^*,s^*) \stackrel{\triangle}{=} & \inf_{s\geq 0}\inf_{\lambda\geq 0}\sup_{Q_{x^m,y^m}\in\mathcal{C}(P_{x^m,y^m})} \Bigg\{ \\ & \int_{\Re^{(m+1)n}\times\Re^{(m+1)d}} \ell(x^m,\tilde{x}^m) dQ_{x^m,y^m}(x,y) \\ & -s\Big(H(Q_{x^m,y^m}|P_{x^m,y^m}) - R\Big) \\ & -\lambda\Big(\int_{\Re^{(m+1)n}\times\Re^{(m+1)d}} dQ_{x^m,y^m}(x,y) - 1\Big)\Bigg\} \end{aligned} \tag{4.6}$$

where $s \geq 0$ is the Lagrange multiplier associated with the constant $\mathcal{C}(P_{x^m,y^m})$ and $\lambda \geq 0$ is the Lagrange multiplier associated with the constraint $\int_{\Re^{(m+1)n}\times\Re^{(m+1)d}} dQ_{x^m,y^m}(x,y) = 1$ (e.g., Q_{x^m,y^m} is a joint distribution and hence it should integrate to 1).

From Section 3.2.4 the worst case measure, $Q^*_{x^m,y^m}$, is given by

$$dQ^*_{x^m,y^m}(x,y) = \frac{e^{\frac{\ell(x^m,\tilde{x}^m)}{s}} dP_{x^m,y^m}(x,y)}{\int_{\Re^{(m+1)n}\times\Re^{(m+1)d}} e^{\frac{\ell(x^m,\tilde{x}^m)}{s}} dP_{x^m,y^m}(x,y)}. \tag{4.7}$$

Then (4.7) is substituted into (4.6) to deduce

$$L(Q^*_{x^m,y^m}, \tilde{x}_m, \lambda^*, s) = s \log \int_{\Re^{(m+1)n} \times \Re^{(m+1)d}} e^{\frac{\ell(x^m, \tilde{x}^m)}{s}} dP_{x^m,y^m}(x,y) + sR. \tag{4.8}$$

Hence, Problem 4.2.1 is equivalent to the dual problem of finding the optimal estimator $\tilde{x}^*_m \in \mathcal{X}_{ad}$ via

$$\inf_{s \geq 0} \inf_{\tilde{x}_m \in \mathcal{X}_{ad}} L(Q^*_{x^m,y^m}, \tilde{x}_m, \lambda^*, s). \tag{4.9}$$

Thus, the rest of this chapter deals with the problem of finding the minimum over $\tilde{x}_m \in \mathcal{X}_{ad}$ of (4.8).

This is achieved by expressing the minimization over $\tilde{x}_m \in \mathcal{X}_{ad}$ in terms of the solution of a recursive equation satisfied by a conditional unnormalized distribution. The next theorem presents intermediate steps using the change of probability measure discussed extensively in Section 2.2.

Theorem 4.2.2. *Let Φ be a bounded continuous function on $\Re^n$ taking values in $\Re$. The likelihood function of the complete data $\{x_0, \ldots, x_m, y_0, \ldots, y_m\}$ is defined by*

$$\Lambda_m \stackrel{\triangle}{=} \prod_{k=0}^{m} \left[\frac{\Xi_v(D_k^{-1}(y_k - h(k, x_k)))}{|D_k| \Xi_v(y_k)} \frac{\Psi_w(B_k^{-1}(x_k - f(k, x_{k-1})))}{|B_k| \Psi_w(x_k)} \right] = \frac{dP(x^m, y^m)}{d\overline{P}(x^m, y^m)} \tag{4.10}$$

where under probability distribution $\overline{P}$, $\{x_k\}$ is i.i.d. $N(0, I_n)$ and $\{y_k\}$ is i.i.d. $N(0, I_d)$ with density functions $\Psi_w(x_k) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp(\frac{-x_k^T x_k}{2})$ and $\Xi_v(y_k) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp(\frac{-y_k^T y_k}{2})$, $k \in N_0$, respectively. Then the following relations hold.

1) Conditional expectations are related via

$$E\Big[\Phi(x_m) \exp\Big(\frac{1}{s} \sum_{k=0}^{m} \tilde{\ell}(x_k, \tilde{x}_k)\Big) \Big| \mathcal{F}_m^{\mathcal{Y}}\Big] = \frac{\overline{E}\Big[\Phi(x_m) \Lambda_m \exp\Big(\frac{1}{s} \sum_{k=0}^{m} \tilde{\ell}(x_k, \tilde{x}_k)\Big) | \mathcal{F}_m^{\mathcal{Y}}\Big]}{\overline{E}\Big[\Lambda_m | \mathcal{F}_m^{\mathcal{Y}}\Big]} \tag{4.11}$$

where E, $\overline{E}$ denotes expectation under the probability distribution P, $\overline{P}$, respectively.

Moreover, the numerator of (4.11) can be written in terms of unnormalized probability distribution $\alpha_m^s(.)$ via

$$\alpha_m^s(\Phi) \stackrel{\triangle}{=} \overline{E}\Big[\Phi(x_m) \Lambda_m \exp\Big(\frac{1}{s} \sum_{k=0}^{m} \tilde{\ell}(x_k, \tilde{x}_k)\Big) | \mathcal{F}_m^{\mathcal{Y}}\Big] = \int_{\Re^n} \Phi(z) d\alpha_m^s(z) \tag{4.12}$$

where the probability distribution $\alpha_m^s(.)$ satisfies the following recursion

$$\begin{aligned}\alpha_m^s(\Phi) = \quad & \int_{\Re^n} \Phi(x) \exp\left(\frac{1}{s}\tilde{\ell}(x,\tilde{x}_m)\right) \frac{\Xi_v(D_m^{-1}(y_m - h(m,x)))}{|D_m|\Xi_v(y_m)} \\ & \times \alpha_{m-1}^s\left(\frac{\Psi_w(B_m^{-1}(x - f(m,\cdot)))}{|B_m|}\right) dx \end{aligned} \tag{4.13}$$

with initial condition

$$\alpha_0^s(\Phi) = \int_{\Re_n} \Phi(x_0) \exp\left(\frac{1}{s}\tilde{\ell}(x_0,\tilde{x}_0)\right) \frac{\Xi_v\left(D_0^{-1}\left(y_0 - h(0,x_0)\right)\right)}{|D_0|\Xi_v(y_0)} \, d\Pi_{x_0}(x_0). \tag{4.14}$$

2) If $\alpha_m^s(\Phi)$ has a density, that is $\frac{d}{dx}\alpha_m^s(x) = \overline{\alpha}_m^s(x)$, then

$$\alpha_m^s(\Phi) = \int_{\Re^n} \Phi(x) d\alpha_m^s(x) = \int_{\Re^n} \Phi(x)\overline{\alpha}_m^s(x)dx \tag{4.15}$$

and the density $\{\overline{\alpha}_m^s(x)\}_{m\geq 0}$ satisfies the recursion

$$\begin{aligned}\overline{\alpha}_m^s(x) = \quad & \frac{\Xi_v(D_m^{-1}(y_m - h(m,x)))}{|D_m|\Xi_v(y_m)} \exp\left(\frac{1}{s}\tilde{\ell}(x,\tilde{x}_m)\right) \\ & \times \int_{\Re^n} \frac{\Psi_w(B_m^{-1}(x - f(m,z)))}{|B_m|} \overline{\alpha}_{m-1}^s(z)dz \end{aligned} \tag{4.16}$$

with initial condition

$$\overline{\alpha}_0^s(x) = \frac{\Xi_v(D_0^{-1}(y_0 - h(0,x_0)))}{|D_0|\Xi_v(y_0)} \exp\left(\frac{1}{s}\tilde{\ell}(x_0,\tilde{x}_0)\right) \pi_{x_0}(x_0). \tag{4.17}$$

Proof. The derivation is a variant of the one found in Theorem 2.2.19 in Section 2.2.4.

1) This follows from the relation of conditional expectation under different probability distribution [43]. (4.12) follows from Theorem 2.2.19 in Section 2.2.4 by absorbing the exponential term $\exp\left(\frac{1}{s}\sum_{k=0}^m \tilde{\ell}(x_k,\tilde{x}_k)\right)$ into the likelihood function Λ_m.

2) Follows from Theorem 2.2.19 in Section 2.2.4 and the above discussion.

□

Notice that $\{\overline{\alpha}_m^s(x); m \in N_0\}$ is a sufficient statistic for the robust (e.g., minimax) estimation problem because the initial optimization Problem 4.2.1 is now expressed

in terms of this quantity. Using Theorem 4.2.2, the dual functional (4.8) is now given by

$$
\begin{aligned}
L(Q^*_{x^m,y^m}, \tilde{x}_m, \lambda^*, s) &= s \log E[e^{\frac{\ell(x^m,\tilde{x}^m)}{s}}] + sR = s \log \overline{E}[e^{\frac{\ell(x^m,\tilde{x}^m)}{s}} \Lambda_m] + sR \qquad (4.18)\\
&= s \log \overline{E}\Big[\overline{E}[e^{\frac{\ell(x^m,\tilde{x}^m)}{s}} \Lambda_m | \mathcal{F}^{\mathcal{Y}}_m]\Big] + sR \qquad (4.19)\\
&= s \log \int_{\Re^{(m+1)d}} \overline{E}[e^{\frac{\ell(x^m,\tilde{x}^m)}{s}} \Lambda_m | \mathcal{F}^{\mathcal{Y}}_m] dP_{y^m}(y) + sR \qquad (4.20)\\
&= s \log \int_{\Re^{(m+1)d}} \alpha^s_m(1) dP_{y^m}(y) + sR. \qquad (4.21)
\end{aligned}
$$

The chain of equalities in (4.18)-(4.21) are obtained as follows. Clearly, (4.18) follows by substituting into $\int e^{\frac{\ell(x^m,\tilde{x}^m)}{s}} dP_{x^m,y^m}$ the density $dP = \Lambda_m d\overline{P}$. Also, (4.19) is a property of conditional expectation. In addition, (4.20) follows from (4.19) since $\overline{E}[e^{\frac{\ell(x^m,\tilde{x}^m)}{s}} \Lambda_m | \mathcal{F}^{\mathcal{Y}}_m] = f(y^m)$. Finally, (4.21) follows from (4.12) by letting $\Phi = 1$.

Further, to find the optimal estimator $\tilde{x}^*$ one needs to solve the recursive equation (4.13), then substitute into (4.21) and perform the minimization over $\tilde{x}_m \in \mathcal{X}_{ad}$. The point to be made in this section is that the minimax estimation yields the estimation problem (4.8) with an exponential cost, while (4.8) is further expressed in terms of the unnormalized conditional distribution $\{\alpha^s_m(x) : m \in N_0\}$.

Remark 4.2.3. *In decision applications, there will be one distribution $\alpha^s_m(x)$ for each hypothesis, e.g., in binary hypothesis the decision rule takes the form*

$$
\frac{\inf_{\tilde{x}_m \in \mathcal{X}_{ad}} \int_{\Re^n} d\alpha^{s,1}_m(x)}{\inf_{\tilde{x}_m \in \mathcal{X}_{ad}} \int_{\Re^n} d\alpha^{s,2}_m(x)} \underset{H_2}{\overset{H_1}{\gtrless}} \gamma
$$

where γ is the threshold. Clearly, $\alpha^s_m(x)$ is a sufficient statistic for estimation and decision problems, since both the least-square estimation and decision rule are contructed from this quantity.

In the next sections the results obtained in this section are applied to different problems where model (4.1) has a specific form.

4.3 Minimax Estimation for Linear Gaussian Models

The partially observed Gaussian version of (4.1) given by the following recursions, is being considered in this section.

$$\Big(\Omega, \mathcal{F}, \{\mathcal{G}_k\}, P\Big) : \begin{cases} x_{k+1} = A_{k+1}x_k + B_{k+1}w_{k+1}, \quad x_k \in \Re^n \\ \\ y_k = C_k x_k + D_k v_k, \quad y_k \in \Re^d. \end{cases} \tag{4.22}$$

The noises are Gaussian distributed as follows, $w_k \sim N(0, I_n), \quad v_k \sim N(0, I_d)$, while sequences $\{w_k\}$, $\{v_k\}$ are mutually independent and independent of $x_0 \sim N(\overline{x}_0, V_0)$

According to Theorem 4.2.2 equation (4.16), $\{\overline{\alpha}_m^s(x)\}$ satisfies the following recursion

$$\begin{aligned} \overline{\alpha}_m^s(x) &= \frac{\Xi_v(D_m^{-1}(y_m - C_m x))}{|D_m|\Xi_v(y_m)} \exp\Big(\frac{1}{s}\tilde{\ell}(x, \tilde{x}_m)\Big) \\ &\quad \times \int_{\Re_n} \frac{\Psi_w(B_m^{-1}(x - A_m z))}{|B_m|} \overline{\alpha}_{m-1}(z) dz \end{aligned} \tag{4.23}$$

where

$$\Psi_w(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big(-\frac{x^T x}{2}\Big), \tag{4.24}$$

$$\Xi_v(y) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\Big(-\frac{y^T y}{2}\Big), \tag{4.25}$$

$$\ell(x, \tilde{x}_m) = (x - \tilde{x}_m)^T \frac{W_m}{2} (x - \tilde{x}_m). \tag{4.26}$$

A solution of (4.23) is assumed to have the following form

$$\overline{\alpha}_m^s(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |V_{m|m}^s|^{\frac{1}{2}}} \exp\Big(-(x - \hat{x}_{m|m}^s)^T \frac{(V_{m|m}^s)^{-1}}{2} (x - \hat{x}_{m|m}^s) + \beta_{m|m}^s\Big) \tag{4.27}$$

where $\{V_{m|m}^s, \hat{x}_{m|m}^s, \beta_{m|m}^s\}$ will be identified shortly. By substituting (4.27) into the recursion (4.23) and using (4.24), (4.25), (4.26) the following recursive relations for $\{V_{m|m}^s, \hat{x}_{m|m}^s, \beta_{m|m}^s\}$ are deduced.

$$\hat{x}_{m|m}^s = V_{m|m}^s \Big[C_m^T (D_m D_m^T)^{-1} y_m + (V_{m|m-1}^s)^{-1} \hat{x}_{m|m-1}^s - \frac{W_m}{s} \tilde{x}_m\Big] \tag{4.28}$$

$$V_{m|m}^s = \Big(C_m^T (D_m D_m^T)^{-1} C_m + (V_{m|m-1}^s)^{-1} - \frac{W_m}{s}\Big)^{-1} \tag{4.29}$$

$$\beta_{m|m}^s = (\hat{x}_{m|m}^s)^T \frac{(V_{m|m}^s)^{-1}}{2} \hat{x}_{m|m}^s - (\hat{x}_{m|m-1}^s)^T \frac{(V_{m|m-1}^s)^{-1}}{2} \hat{x}_{m|m-1}^s + \tilde{x}_m^T \frac{W_m}{2s} \tilde{x}_m$$

$$-\frac{1}{2}\log\Big[|V^s_{m-1|m-1}||A^T_m(B_mB^T_m)^{-1}A+(V^s_{m-1|m-1})^{-1}||(V^s_{m|m})^{-1}|\Big]$$
$$-\log[|D_mB_m|]+\frac{1}{2}y^T_m\Big(I_d-(D_mD^T_m)^{-1}\Big)y_m+\beta^s_{m-1|m-1} \tag{4.30}$$

where

$$\hat{x}^s_{m|m-1} = A_m\hat{x}^s_{m-1|m-1} \tag{4.31}$$
$$V^s_{m|m-1} = B_mB^T_m+A_mV^s_{m-1|m-1}A^T_m \tag{4.32}$$

with initial conditions

$$\hat{x}^s_{0|0} = V^s_{0|0}\Big[C^T_0(D^T_0D_0)^{-1}y_0-\frac{W_0}{s}\tilde{x}_0+V^{-1}_0\overline{x}_0\Big] \tag{4.33}$$
$$V^s_{0|0} = \Big(C^T_0(D^T_0D_0)^{-1}C_0-\frac{W_0}{s}+V^{-1}_0\Big)^{-1} \tag{4.34}$$
$$\beta^s_{0|0} = (\hat{x}^s_{0|0})^T\frac{(V^s_{0|0})^{-1}}{2}\hat{x}^s_{0|0}+\tilde{x}^T_0\frac{W_0}{2s}\tilde{x}_0-\overline{x}_0\frac{V^{-1}_0}{2}\overline{x}_0$$
$$+\frac{1}{2}\log\Big[\frac{V_{0|0}}{V_0}\Big]-\log|D_0||B_0|+\frac{1}{2}y^T_0\Big(I_d-(D^T_0D_0)^{-1}\Big)y_0 \tag{4.35}$$

thus it is concluded that (4.27) is indeed a solution of the recursion (4.23).

Theorem 4.3.1. *The optimal estimate $\tilde{x}^{*,s}_m$ of x_m is given by the following expression*

$$\tilde{x}^{*,s}_m = A_m\tilde{x}^{*,s}_{m-1}+\Big[(V^s_{m|m-1})^{-1}+C^T_m(D_mD^T_m)^{-1}C_m\Big]^{-1}C^T_m(D_mD^T_m)^{-1}$$
$$\times\Big(y_m-C_mA_m\tilde{x}^{*,s}_{m-1}\Big). \tag{4.36}$$

Note that the above estimator can also be re-written by applying the PosDef Identity (see Appendix A) as

$$\tilde{x}^{*,s}_m = A_m\tilde{x}^{*,s}_{m-1}+V^s_{m|m-1}C^T_m\Big[C_mV^s_{m|m-1}C^T_m+(D_mD_m)\Big]\Big(y_m-C_mA_m\tilde{x}^{*,s}_{m-1}\Big). \tag{4.37}$$

This form is the same as the Kalman-Filter solution [1], [4]. The difference can be found in the covariance of the predictor error, $V^s_{m|m-1}$, which for this case it is a function of s.

Let

$$\hat{x}_{m|m}\triangleq\lim_{s\to\infty}\hat{x}^s_{m|m},\ V_{m|m}\triangleq\lim_{s\to\infty}V^s_{m|m},\ \hat{x}_{m|m-1}\triangleq\lim_{s\to\infty}\hat{x}^s_{m|m-1},$$
$$V_{m|m-1}\triangleq\lim_{s\to\infty}V^s_{m|m-1},\ \tilde{x}^*_m\triangleq\lim_{s\to\infty}\tilde{x}^{*,s}_m.$$

Then, when $s \to \infty$,

$$\hat{x}_{m|m} = V^s_{m|m}\Big[C_m^T(D_mD_m^T)^{-1}y_m + (V_{m|m-1})^{-1}\hat{x}_{m|m-1}\Big] \tag{4.38}$$

$$V_{m|m} = \Big(C_m^T(D_mD_m^T)^{-1}C_m + (V_{m|m-1})^{-1}\Big)^{-1} \tag{4.39}$$

where

$$\hat{x}_{m|m-1} = A_m\hat{x}_{m-1|m-1} \tag{4.40}$$

$$V_{m|m-1} = B_mB_m^T + A_mV_{m-1|m-1}A_m^T \tag{4.41}$$

the estimator (4.37) converges to the classical Kalman-Filter solution, given by (1.11).

$$\tilde{x}^*_m = A_m\tilde{x}^*_{m-1} + V_{m|m-1}C_m^T\Big[C_mV^s_{m|m-1}C_m^T + (D_mD_m)\Big]\Big(y_m - C_mA_m\tilde{x}^*_{m-1}\Big). \tag{4.42}$$

Proof. As was mentioned in Section 4.2.2 in order to find the optimal estimator $\tilde{x}^{*,s}$ one needs to solve the recursive equation (4.13), then substitute into (4.21) and perform the minimization over $\tilde{x}_m \in \mathcal{X}_{ad}$. By applying the theory to this specific problem, then

$$\begin{aligned}
\tilde{x}^{*,s}_m &= \arg\min_{\tilde{x}_m} s\log\int_{\Re^{(m+1)d}}\alpha^s_m(1)dP_{y^m}(y) + sR \\
&= \arg\min_{\tilde{x}_m}\int_{\Re^n}\overline{\alpha}^s_m(x)dx, \quad P_{y^m} - a.s. \\
&= \arg\min_{\tilde{x}_m}\int_{\Re^n}\frac{1}{(2\pi)^{\frac{n}{2}}|V^s_{m|m}|^{\frac{1}{2}}}\exp\Big(-(x-\hat{x}^s_{m|m})^T\frac{(V^s_{m|m})^{-1}}{2}(x-\hat{x}^s_{m|m}) + \beta^s_{m|m}\Big)dx \\
&= \arg\min_{\tilde{x}_m}\exp\Big(\beta^s_{m|m}\Big) \\
&= \arg\min_{\tilde{x}_m}\exp\Big((\tilde{x}_m - N_m)^T\frac{\Sigma_m^{-1}}{2}(\tilde{x}_m - N_m) + K(y_m, \hat{x}_{m-1|m-1})\Big)
\end{aligned} \tag{4.43}$$

where

$$\begin{aligned}
\Sigma_m^{-1} &= \frac{W_m}{s}V^s_{m|m}\frac{W_m}{s} + \frac{W_m}{s}, \\
N_m &= \Sigma_m\frac{W_m}{s}V^s_{m|m}\Big[C_m^T(D_mD_m^T)^{-1}y_m + (V^s_{m|m-1})^{-1}\hat{x}^s_{m|m-1}\Big].
\end{aligned}$$

Continuing from (4.43) and by performing the minimization

$$\tilde{x}^{*,s}_m = N_m$$

$$
\begin{aligned}
&= \left(\frac{W_m}{s}V^s_{m|m}\frac{W_m}{s}+\frac{W_m}{s}\right)^{-1}\frac{W_m}{s}V^s_{m|m}\Big[C_m^T(D_mD_m^T)^{-1}y_m+(V^s_{m|m-1})^{-1}\hat{x}^s_{m|m-1}\Big]\\
&= \left((V^s_{m|m})^{-1}+\frac{W_m}{s}\right)^{-1}\Big[C_m^T(D_mD_m^T)^{-1}y_m+(V^s_{m|m})^{-1}\hat{x}^s_{m|m-1}\Big]\\
&= \left((V^s_{m|m-1})^{-1}+C_m^T(D_mD_m^T)^{-1}C_m\right)^{-1}\Big[C_m^T(D_mD_m^T)^{-1}y_m+(V^s_{m|m-1})^{-1}\hat{x}^s_{m|m-1}\Big] \qquad (4.44)\\
&= \hat{x}^s_{m|m-1}+\left((V^s_{m|m-1})^{-1}+C_m^T(D_mD_m^T)^{-1}C_m\right)^{-1}C_m^T(D_mD_m^T)^{-1}\Big[y_m-C_m\hat{x}^s_{m|m-1}\Big]. \qquad (4.45)
\end{aligned}
$$

By using (4.28) and solving for $\hat{x}^s_{m|m}$ (4.45) can be transformed into

$$
\begin{aligned}
\hat{x}^s_{m|m} &= V^s_{m|m}\Big[C_m^T(D_mD_m^T)^{-1}y_m+(V^s_{m|m-1})^{-1}\hat{x}_{m|m-1}-\frac{W_m}{s}\tilde{x}^*_m\Big]\\
&= V^s_{m|m}\Big[C_m^T(D_mD_m^T)^{-1}y_m+(V^s_{m|m-1})^{-1}\hat{x}^s_{m|m-1}\Big]-V^s_{m|m}\frac{W_m}{s}\tilde{x}^*_m. \qquad (4.46)
\end{aligned}
$$

Next, (4.44) can be expressed as

$$
\left((V^s_{m|m-1})^{-1}+C_m^T(D_mD_m^T)^{-1}C_m\right)\tilde{x}^*_m=\Big[C_m^T(D_mD_m^T)^{-1}y_m+(V^s_{m|m-1})^{-1}\hat{x}^s_{m|m-1}\Big] \qquad (4.47)
$$

and when it is substituted into (4.46) it transforms it to

$$
\begin{aligned}
\hat{x}^s_{m|m} &= V^s_{m|m}\Big((V^s_{m|m})^{-1}+C_m^T(D_mD_m^T)^{-1}C_m\Big)\tilde{x}^{*,s}_m-V^s_{m|m}\frac{W_m}{s}\tilde{x}^{*,s}_m\\
&= V^s_{m|m}\Big((V^s_{m|m-1})^{-1}+C_m^T(D_mD_m^T)^{-1}C_m-\frac{W_m}{s}\Big)\tilde{x}^*_m\\
&= V^s_{m|m}(V^s_{m|m})^{-1}\tilde{x}^{*,s}_m=\tilde{x}^{*,s}_m. \qquad (4.48)
\end{aligned}
$$

Finally, (4.36) is derived by substituting (4.31) and (4.48) into (4.45).

□

4.4 Non-Coherent Estimation in Multipath

In this section the theory derived in Section 4.2 is applied to an attenuated sinusoid in a multipath environment which is subject to an additive Gaussian noise. Various

estimators are derived by solving the recursive equation for $\{\overline{\alpha}_m^s(x) : m \in N_0\}$. Further, as a special case, the estimation problem when there is no uncertainty is derived, by taking the limit as $s \to \infty$.

4.4.1 Channel Model

A multipath version of the classical non-coherent model, given by the following equation, is considered in this section.

$$\begin{aligned} y(t_k) &= \sum_{i=1}^{N} A_i(t_k) r_i \cos(\omega_c(t_k - \tau_i(t_k)) + \theta_i) S(t_k - \tau_i(t_k)) + D(t_k)v(t_k) \\ &= \sum_{i=1}^{N} h_i(t_k, \theta_i, r_i) + D(t_k)v(t_k) \end{aligned} \tag{4.49}$$

where ω_c is the carrier frequency, $\{\tau_i(t_k)\}$ denotes the propagation delay, $A_i(t_k)$ denotes a deterministic known signal envelope, $\{r_i\}$, $\{\theta_i\}$ are Random Variables denoting the attenuation and phase, respectively, of the signal received associated with ith path, and $v(t_k) \sim N(0,1)$. Furthermore, the following function is defined $h(N, \boldsymbol{\theta}, \mathbf{r}) \stackrel{\triangle}{=} \sum_{i=1}^{N} h_i(t_k, \theta_i, r_i)$, where $\boldsymbol{\theta} \stackrel{\triangle}{=} (\theta_1, \ldots, \theta_N)'$ is the phase vector, and $\mathbf{r} \stackrel{\triangle}{=} (r_1, \ldots, r_N)'$ is the attenuation vector.

The delays $\{\tau_i(.)\}_{i=1}^N$ are assumed to be fixed and known, while the phases $\theta_i : \Omega \to [0, 2\pi]$ are independent and identically distributed RVs with *á* priori density $\pi_{\theta_0}(\theta_i)$, $\theta_i \in [0, 2\pi]$, while the attenuations $r_i : \Omega \to [0, \infty)$ are independent and identically distributed (iid) RVs with *á* priori density $\pi_{r_0}(r_i)$, for $1 \leq i \leq N$. In addition it is assumed $\{r_i\}_{i=1}^N$ and $\{\theta_i\}_{i=1}^N$ are independent, and also independent of the noise process $\{v(t_k); k \in N_0\}$. A special case of (4.49) in which $N = 1,\ \tau_1 = 0$ and $r_1 = 1$ is the model used for non-coherent detection (see [4], [51]).

Further, it is assumed that the estimation error for $\boldsymbol{\theta}$ and $\boldsymbol{r}$ is additive given by

$$\ell(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^m(t_k), \mathbf{r}, \tilde{\mathbf{r}}^m(t_k)) = \ell_1(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^m(t_k)) + \ell_2(\mathbf{r}, \tilde{\mathbf{r}}^m(t_k))$$

where $\ell_1(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^m(t_k)) = \sum_{k=0}^{m} \sum_{i=1}^{N} \tilde{\ell}_{1,i}(\theta_i, \tilde{\theta}_i(t_k))$, $\ell_2(\mathbf{r}, \tilde{\mathbf{r}}^m(t_k)) = \sum_{k=0}^{m} \sum_{i=1}^{N} \tilde{\ell}_{2,i}(r_i, \tilde{r}_i(t_k))$ and $\tilde{\boldsymbol{\theta}}(t_k)$, $\tilde{\mathbf{r}}(t_k)$ are the estimates of $\boldsymbol{\theta}$, $\mathbf{r}$, respectively at time t_k, $k \in N_0$. Note that $\tilde{\ell}_{1,i}$, $\tilde{\ell}_{2,i}$ are functions of estimation error.

The relation of model (4.49) with model (4.1) is the following. Clearly, sampling time k of previous section is now represented by t_k, and the state vector of (4.49) which

needs to be estimated is $x = \begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{r} \end{pmatrix} \in \Re^{2N}$. However, since $\begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{r} \end{pmatrix}$ are random variables then $x_{k+1} = x_k$, $\forall k \in N_0$.

Next, the theory of Section 4.2 is applied to the above problem.

4.4.2 Minimax Estimation of Phase and Envelope

The density (4.16) is specialized to model (4.49). This is done by replacing the integrand $\Psi_w(B_m^{-1}(x_m - f(t,z)))$, $x \in \Re^{2N}$, $z \in \Re^{2N}$ in (4.16) with the delta measure $\delta(\mathbf{r}-z^1)\times\delta(\boldsymbol{\theta}-z^2) = \prod_{i=1}^N \delta(r_i-z_i^1)\times\delta(\theta_i-z_i^2)$, where $x = \begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{r} \end{pmatrix}$, $z = \begin{pmatrix} z^1 \\ z^2 \end{pmatrix} \in \Re^{2N}$, $f(t,z) = \begin{pmatrix} z^1 \\ z^2 \end{pmatrix}$ to get

$$\begin{aligned} \overline{\alpha}_m^s(\boldsymbol{\theta},\mathbf{r}) &= \pi_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})\pi_{\mathbf{r_0}}(\mathbf{r}) \prod_{k=0}^{m} \Bigg[\frac{\Xi_v(D^{-1}(t_k)(y(t_k) - \sum_{i=1}^N h_i(t_k,\theta_i,r_i)))}{|D(t_k)|\Xi_v(y(t_k))} \\ &\quad \times \exp\left(\frac{1}{s}\sum_{i=1}^N \tilde{\ell}_{1,i}(\theta_i,\tilde{\theta}_i(t_k)) + \frac{1}{s}\sum_{i=1}^N \tilde{\ell}_{2,i}(r_i,\tilde{r}_i(t_k))\right)\Bigg] \end{aligned} \tag{4.50}$$

where $\pi_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) = \prod_{i=1}^N \pi_{\theta_0}(\theta_i)$ is the *á* priori joint density of $(\theta_1,\ldots,\theta_N)$, and $\pi_{\mathbf{r_0}}(\mathbf{r}) = \prod_{i=1}^N \pi_{r_0}(r_i)$ is the *á* priori joint density of $(r_1,\ldots,r_N)$.

Furthermore, the error functions are assumed to have the form $\tilde{\ell}_{1,i}(\theta_i,\tilde{\theta}_i(t_k)) = \frac{(\theta_i-\tilde{\theta}_i(t_k))^2 R_i(t_k)}{2}$, and $\tilde{\ell}_{2,i}(r_i,\tilde{r}_i(t_k)) = \frac{(r_i-\tilde{r}_i(t_k))^2 Q_i(t_k)}{2}$, where $R_i(.)$, $Q_i(.)$ are weighting coefficients on the specific errors, which are positive for each t_k.

The following quantities are needed when presenting the robust estimator.

Definition 4.4.1. *For each $1 \le i,j \le N, i \ne j$, define the following quantities.*

$V_c^i(y^m) \triangleq \sum_{k=0}^m A_i(t_k)D^{-2}(t_k)\cos(\omega_c(t_k-\tau_i(t_k)))S(t_k-\tau_i(t_k))y(t_k)$.
$V_s^i(y^m) \triangleq \sum_{k=0}^m A_i(t_k)D^{-2}(t_k)\sin(\omega_c(t_k-\tau_i(t_k)))S(t_k-\tau_i(t_k))y(t_k)$.
$V_i(y^m) = \sqrt{V_c^i(y^m)^2 + V_s^i(y^m)^2}$, $\gamma_i(y^m) = -\tan^{-1}\left(\frac{V_s^i(y^m)}{V_c^i(y^m)}\right)$.
$W_c^{i,m} \triangleq -\frac{1}{4}\sum_{k=0}^m A_i(t_k)D^{-2}(t_k)\cos(2\omega_c(t_k-\tau_i(t_k)))S^2(t_k-\tau_i(t_k))$.
$W_s^{i,m} \triangleq -\frac{1}{4}\sum_{k=0}^m A_i(t_k)D^{-2}(t_k)\sin(2\omega_c(t_k-\tau_i(t_k)))S^2(t_k-\tau_i(t_k))$.
$W_i^m = \sqrt{(W_c^{i,m})^2 + (W_s^{i,m})^2}$, $\beta_i^m = -\tan^{-1}\left(\frac{W_s^{i,m}}{W_c^{i,m}}\right)$.

$U_c^{ij,m} \triangleq -\frac{1}{2}\sum_{k=0}^{m} A_i(t_k)D^{-2}(t_k)\cos(\omega_c(2t_k-\tau_i(t_k)-\tau_j(t_k)))S(t_k-\tau_i(t_k))S(t_k-\tau_j(t_k))$.
$U_s^{ij,m} \triangleq -\frac{1}{2}\sum_{k=0}^{m} A_i(t_k)D^{-2}(t_k)\sin(\omega_c(2t_k-\tau_i(t_k)-\tau_j(t_k)))S(t_k-\tau_i(t_k))S(t_k-\tau_j(t_k))$.
$U_{ij}^m = \sqrt{(U_c^{ij,m})^2+(U_s^{ij,m})^2}$, $\phi_{ij}^m = -\tan^{-1}\left(\frac{U_s^{ij,m}}{U_c^{ij,m}}\right)$.
$T_c^{ij,m} \triangleq -\frac{1}{2}\sum_{k=0}^{m} A_i(t_k)D^{-2}(t_k)\cos(\omega_c(\tau_j(t_k)-\tau_i(t_k)))S(t_k-\tau_i(t_k))S(t_k-\tau_j(t_k))$.
$T_s^{ij,m} \triangleq -\frac{1}{2}\sum_{k=0}^{m} A_i(t_k)D^{-2}(t_k)\sin(\omega_c(\tau_j(t_k)-\tau_i(t_k)))S(t_k-\tau_i(t_k))S(t_k-\tau_j(t_k))$.
$T_{ij}^m = \sqrt{(T_c^{ij,m})^2+(T_s^{ij,m})^2}$, $\psi_{ij}^m = -\tan^{-1}\left(\frac{T_s^{ij,m}}{T_c^{ij,m}}\right)$.

From Definition 4.4.1 and (4.50) the unnormalized conditional density is given by

$$
\begin{aligned}
\overline{\alpha}_m^s(\boldsymbol{\theta},\mathbf{r}) \;=\; & \pi_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})\pi_{\mathbf{r_0}}(\mathbf{r})\exp\Big(-\frac{1}{4}\sum_{i=1}^{N}\sum_{k=0}^{m} r_i^2 A_i^2(t_k)D^{-2}(t_k)S^2(t_k-\tau_i(t_k))\Big) \\
& \times \exp\Big(\sum_{i=1}^{N} r_i V_i(y^m)\cos(\theta_i-\gamma_i(y^m))\Big)\exp\Big(\sum_{i=1}^{N} r_i^2 W_i^m\cos(2\theta_i-\beta_i^m)\Big) \\
& \times \exp\Big(\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_i r_j U_{ij}^m\cos(\theta_i+\theta_j-\phi_{ij}^m)\Big) \\
& \times \exp\Big(\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_i r_j T_{ij}^m\cos(\theta_i-\theta_j-\psi_{ij}^m)\Big) \\
& \times \exp\Big(\sum_{i=1}^{N}\sum_{k=0}^{m}\frac{(\theta_i-\tilde{\theta}_i(t_k))^2 R_i(t_k)}{2s}+\sum_{i=1}^{N}\sum_{k=0}^{m}\frac{(r_i-\tilde{r}_i(t_k))^2 Q_i(t_k)}{2s}\Big) \\
& \times \exp\Big(\frac{1}{2}\sum_{k=0}^{m} y^2(t_k)[1-D^{-2}(t_k)]-\sum_{k=0}^{m}\log|D(t_k)|\Big). \qquad (4.51)
\end{aligned}
$$

Note that (4.51) should be substituted into (4.21) and then minimized over the class of estimators. Since the exponential term is a quadratic function of $(\boldsymbol{\theta},\mathbf{r})$ and $(\tilde{\boldsymbol{\theta}}^m,\tilde{\mathbf{r}}^m)$, the minimization in (4.51) over $\{\tilde{\boldsymbol{\theta}}^m,\tilde{\mathbf{r}}^m\}$ can be found explicitly provided $\pi_{\boldsymbol{\theta}_0}(.)$, $\pi_{\mathbf{r_0}}(.)$ are Gaussian. However, for this problem this is not the case. Nevertheless, closed form expressions for the optimal estimators when $\boldsymbol{\theta}$ is uniformly distributed over $[0,2\pi]$ and $\boldsymbol{r}$ is arbitrary distributed are going to be derived.

Remark 4.4.2. *Below realistic scenarios that lead to a simplified version of (4.51) are discussed.*

1. *Minimax Estimation Subject to Resolvable Paths.*
   Supposed the received paths are resolvable, where path resolvability for wideband signals is defined as having inter-path delays which are larger than the

reciprocal of the bandwidth of the transmitted signals. Under the resolvability assumption, expressions containing cross terms in Definition 4.4.1 are negligible. This implies that the following terms are zero.

$U_c^{ij,m} = U_s^{ij,m} = 0, \;\; T_c^{ij,m} = T_s^{ij,m} = 0, \;\; 1 \leq i, j \leq N, i \neq j.$

Taking this into consideration, expression (4.51) reduces to

$$\begin{aligned}
\overline{\alpha}_m^s(\boldsymbol{\theta}, \mathbf{r}) &= \pi_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})\pi_{\mathbf{r_0}}(\mathbf{r}) \exp\Big(-\frac{1}{4}\sum_{i=1}^{N}\sum_{k=0}^{m} r_i^2 A_i^2(t_k) D^{-2}(t_k) S^2(t_k - \tau_i(t_k))\Big) \\
&\quad \times \exp\Big(\sum_{i=1}^{N} r_i V_i(y^m)\cos(\theta_i - \gamma_i(y^m))\Big) \\
&\quad \times \exp\Big(\sum_{i=1}^{N} r_i^2 W_i^m \cos(2\theta_i - \beta_i^m)\Big) \\
&\quad \times \exp\Big(\sum_{i=1}^{N}\sum_{k=0}^{m} \frac{(\theta_i - \tilde{\theta}_i(t_k))^2 R_i(t_k)}{2s} + \sum_{i=1}^{N}\sum_{k=0}^{m} \frac{(r_i - \tilde{r}_i(t_k))^2 Q_i(t_k)}{2s}\Big) \\
&\quad \exp\Big(\frac{1}{2}\sum_{k=0}^{m} y^2(t_k)[1 - D^{-2}(t_k)] - \sum_{k=0}^{m}\log|D(t_k)|\Big). \qquad (4.52)
\end{aligned}$$

2. *Minimax Estimation Subject to Negligible Double Frequency Terms.*
Here a "Narrow-band" assumption is employed, which implies $2\omega_c$ terms ("double-frequency" terms) are negligible since the receiver will remove them in the process of reconstructing the transmitted signal.

Taking this into consideration, expression (4.51) reduces to

$$\begin{aligned}
\overline{\alpha}_m^s(\boldsymbol{\theta}, \mathbf{r}) &= \pi_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})\pi_{\mathbf{r_0}}(\mathbf{r}) \exp\Big(-\frac{1}{4}\sum_{i=1}^{N}\sum_{k=0}^{m} r_i^2 A_i^2(t_k) D^{-2}(t_k) S^2(t_k - \tau_i(t_k))\Big) \\
&\quad \times \exp\Big(\sum_{i=1}^{N} r_i V_i(y^m)\cos(\theta_i - \gamma_i(y^m))\Big) \\
&\quad \times \exp\Big(\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_i r_j U_{ij}^m \cos(\theta_i + \theta_j - \phi_{ij}^m)\Big) \\
&\quad \times \exp\Big(\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_i r_j T_{ij}^m \cos(\theta_i - \theta_j - \psi_{ij}^m)\Big) \\
&\quad \times \exp\Big(\sum_{i=1}^{N}\sum_{k=0}^{m} \frac{(\theta_i - \tilde{\theta}_i(t_k))^2 R_i(t_k)}{2s} + \sum_{i=1}^{N}\sum_{k=0}^{m} \frac{(r_i - \tilde{r}_i(t_k))^2 Q_i(t_k)}{2s}\Big) \\
&\quad \times \exp\Big(\frac{1}{2}\sum_{k=0}^{m} y^2(t_k)[1 - D^{-2}(t_k)] - \sum_{k=0}^{m}\log|D(t_k)|\Big). \qquad (4.53)
\end{aligned}$$

If cases (1), (2) *above are combined together the density* $\overline{\alpha}_m^s(\boldsymbol{\theta}, \mathbf{r})$ *is given by (4.53) with cross terms* $U_c^{ij,m}, U_s^{ij,m}$, $T_c^{ij,m}, T_s^{ij,m}$ *being zero. This is treated in the following paragraphs.*

Supposed the received paths are resolvable, and double frequency terms $2\omega_c$ are negligible. Then (4.51) reduces to

$$\overline{\alpha}_m^s(\boldsymbol{\theta}, \mathbf{r}) = \Big\{ \prod_{i=1}^{N} \overline{\alpha}_m^{s,i}(\theta_i, r_i) \Big\} \times \exp\left(\frac{1}{2} \sum_{k=0}^{m} y^2(t_k)[1 - D^{-2}(t_k)] - \sum_{k=0}^{m} \log |D(t_k)| \right) \tag{4.54}$$

where $\overline{\alpha}_m^{s,i}(\theta_i, r_i)$, $1 \leq i \leq N$ is given by

$$\begin{aligned} \overline{\alpha}_m^{s,i}(\theta_i, r_i) &= \pi_{\theta_0}(\theta_i)\pi_{r_0}(r_i) \exp\Big(-\frac{1}{4} \sum_{k=0}^{m} r_i A_i^2(t_k) D^{-2}(t_k) S^2(t_k - \tau_i(t_k)) \Big) \\ &\quad \times \exp\Big(r_i V_i(y^m) \cos(\theta_i - \gamma_i(y^m)) \Big) \\ &\quad \times \exp\Big(\frac{(\theta_i - \tilde{\theta}_i(t_m))^2 R_i(t_m)}{2s} + \sum_{k=0}^{m-1} \frac{(\theta_i - \tilde{\theta}_i(t_k))^2 R_i(t_k)}{2s} \Big) \\ &\quad \times \exp\Big(\frac{(r_i - \tilde{r}_i(t_m))^2 Q_i(t_m)}{2s} + \sum_{k=0}^{m-1} \frac{(r_i - \tilde{r}_i(t_k))^2 Q_i(t_k)}{2s} \Big). \end{aligned} \tag{4.55}$$

Note that path resolvability implies that each path component (θ_i, r_i) can be estimated independently of the rest.

Theorem 4.4.3. *Assuming negligible double frequency terms, and resolvable paths, then the optimal robust estimates* $\{\tilde{\theta}_i^{*,s}(t_m)\}_{i=1}^N$ *and* $\{\tilde{r}_i^{*,s}(t_m)\}_{i=1}^N$ *are given by the following expressions*

$$\tilde{\theta}_i^{*,s}(t_m) = \frac{\int_0^\infty \int_0^{2\pi} \theta_i \pi_{r_0}(r_i) \exp(-r_i^2 K_i^m) \exp\Big(r_i V_i(y^m) \cos(\theta_i - \gamma_i(y^m)) \Big) \exp(\zeta_i^m + \xi_i^m) d\theta_i dr_i}{\int_0^\infty \int_0^{2\pi} \pi_{r_0}(r_i) \exp(-r_i^2 K_i^m) \exp\Big(r_i V_i(y^m) \cos(\theta_i - \gamma_i(y^m)) \Big)) \exp(\zeta_i^m + \xi_i^m) d\theta_i dr_i} \tag{4.56}$$

$$\tilde{r}_i^{*,s}(t_m) = \frac{\int_0^\infty \int_0^{2\pi} r_i \pi_{r_0}(r_i) \exp(-r_i^2 K_i^m) \exp\Big(r_i V_i(y^m) \cos(\theta_i - \gamma_i(y^m)) \Big) \exp(\zeta_i^m + \xi_i^m) d\theta_i dr_i}{\int_0^\infty \int_0^{2\pi} \pi_{r_0}(r_i) \exp(-r_i^2 K_i^m) \exp\Big(r_i V_i(y^m) \cos(\theta_i - \gamma_i(y^m)) \Big) \exp(\zeta_i^m + \xi_i^m) d\theta_i dr_i} \tag{4.57}$$

where

$$\zeta_i^m = \frac{(\theta_i - \tilde{\theta}_i^{*,s}(t_m))^2 R_i(t_m)}{2s} + \sum_{k=0}^{m-1} \frac{(\theta_i - \tilde{\theta}_i^{*,s}(t_k))^2 R_i(t_k)}{2s}, \tag{4.58}$$

$$\xi_i^m = \frac{(r_i - \tilde{r}_i^{*,s}(t_m))^2 Q_i(t_m)}{2s} + \sum_{k=0}^{m-1} \frac{(r_i - \tilde{r}_i^{*,s}(t_k))^2 Q_i(t_k)}{2s}, \tag{4.59}$$

$$K_i^m = \frac{1}{4} \sum_{k=0}^{m} A_i^2(t_k) D^{-2}(t_k) S^2(t_k - \tau_i(t_k)). \tag{4.60}$$

Proof. Substituting (4.54) into (4.21) and using the fact that $P_{y^m}(y)$ is a positive density function, which is independent of $\tilde{\theta}_i$ and $\tilde{r}_i$, then $\tilde{\theta}_i^{*,s}$ and $\tilde{r}_i^{*,s}$ are found from the expression

$$\begin{aligned} \{(\tilde{\theta}_i^{*,s}(t_m), \tilde{r}_i^{*,s}(t_m))\}_{i=1}^N &= \arg \min_{\{(\tilde{\theta}_i(t_m), \tilde{r}_i(t_m))\}_{i=1}^N} \log \prod_{i=1}^N \left[\int_0^\infty \int_0^{2\pi} \overline{\alpha}_m^{s,i}(\theta_i, r_i) d\theta_i dr_i \right] \\ &= \arg \min_{\{(\tilde{\theta}_i(t_m), \tilde{r}_i(t_m))\}_{i=1}^N} \sum_{i=1}^N \left[\log \int_0^\infty \int_0^{2\pi} \overline{\alpha}_m^{s,i}(\theta_i, r_i) d\theta_i dr_i \right]. \end{aligned} \tag{4.61}$$

Hence

$$(\tilde{\theta}_i^{*,s}(t_m), \tilde{r}_i^{*,s}(t_m)) = \arg \min_{(\tilde{\theta}_i(t_m), \tilde{r}_i(t_m))} \log \int_0^\infty \int_0^{2\pi} \overline{\alpha}_m^{s,i}(\theta_i, r_i) d\theta_i dr_i, \quad 1 \le i \le N. \tag{4.62}$$

Differentiating the pay-off function (4.62) with respect to $\tilde{\theta}_i(t_m)$ and $\tilde{r}_i(t_m)$, evaluating at $(\tilde{\theta}_i(t_m), \tilde{r}_i(t_m)) = (\tilde{\theta}_i^{*,s}(t_m), \tilde{r}_i^{*,s}(t_m))$, setting it to zero and then solving for $(\tilde{\theta}_i^{*,s}(t_m), \tilde{r}_i^{*,s}(t_m))$ yields expressions (4.56), (4.57).

Notice that the minimax estimators of the phase $\tilde{\theta}_i^{*,s}(t_m)$ and envelope $\tilde{r}_i^{*,s}(t_m)$ are computed from (4.56) and (4.57). Clearly, the estimators derived in (4.56) and (4.57) are given implicitly because $\tilde{\theta}_i^{*,s}(t_m)$ and $\tilde{r}_i^{*,s}(t_m)$ appear on both left and right side of these expressions. In calculating these quantities, one should start with an initial guess $\tilde{\theta}_i^{*,s} = \tilde{\theta}_i^{1,s}$ and $\tilde{r}_i^{*,s} = \tilde{r}_i^{1,s}$, put them into the right side of (4.56) and (4.57) compute $\tilde{\theta}_i^{2,s}$, $\tilde{r}_i^{2,s}$ and repeat this step until the estimators converge. □

4.4.3 Classical Estimation of Phase and Envelope

Here the problem of jointly estimating the channel parameters, namely phase, $\boldsymbol{\theta}$ and envelope, $\boldsymbol{r}$ for an attenuated sinusoid signal in a multipath environment which is subjected to additive Gaussian noise, when there is no uncertainty, is considered. The case of no uncertainty corresponds to having an uncertainty radius $R = 0$. From Theorem 4.2.2, this is equivalent to taking the limit, as $s \to \infty$ [43]. Thus, the limit $\overline{\alpha}_m^{\infty}(\boldsymbol{\theta}, \mathbf{r}) \stackrel{\triangle}{=} \lim_{s\to\infty} \overline{\alpha}_m^{s}(\boldsymbol{\theta}, \mathbf{r})$ corresponds to the unnormalized conditional density of $(\boldsymbol{\theta}, \mathbf{r})$ given $\mathcal{F}_m^{\mathcal{Y}}$. The mean-square error estimation of $(\boldsymbol{\theta}, \boldsymbol{r})$ is then computed via

$$\tilde{\boldsymbol{\theta}}^*(t_m) = E[\boldsymbol{\theta}|\mathcal{F}_m^{\mathcal{Y}}] = \frac{\int \boldsymbol{\theta}\overline{\alpha}_m^{\infty}(\boldsymbol{\theta}, \mathbf{r})d\boldsymbol{\theta}d\boldsymbol{r}}{\int \overline{\alpha}_m^{\infty}(\boldsymbol{\theta}, \mathbf{r})d\boldsymbol{\theta}d\boldsymbol{r}},$$

and similarly for $\tilde{\boldsymbol{r}}^*(t_m)$.

From (4.54), since it is assumed negligible double frequency terms and resolvable paths, and by reducing the uncertainty to zero (by letting $s \to \infty$), the unnormalized conditional density of $(\boldsymbol{\theta}, \mathbf{r})$ given $\mathcal{F}_m^{\mathcal{Y}}$ is given by

$$\begin{aligned}\overline{\alpha}_m^{\infty}(\boldsymbol{\theta}, \mathbf{r}) &= \prod_{i=1}^{N}\left[\pi_{\theta_0}(\theta_i)\pi_{r_0}(r_i)\exp\left(-r_i^2K_i^m\right)\exp\left(r_iV_i(y^m)\cos(\theta_i - \gamma_i(y^m))\right)\right] \\ &\quad \times \exp\left(\frac{1}{2}\sum_{k=0}^{m} y^2(t_k)[1 - D^{-2}(t_k)] - \sum_{k=0}^{m}\log|D(t_k)|\right) \qquad (4.63)\end{aligned}$$

where $K_i^m = \frac{1}{4}\sum_{k=0}^{m} A_i^2(t_k)D^{-2}(t_k)S^2(t_k - \tau_i(t_k))$.

Note that path resolvability implies that each path component (θ_i, r_i) can be estimated independently of the rest. Below various optimal estimator are derived.

Theorem 4.4.4. *Assuming the phases θ_i are iid RVs with á priori density $\pi_{\theta_0}(\theta_i) = \frac{1}{2\pi}$, $\theta_i \in [0, 2\pi]$ and the attenuations r_i are iid RVs with á priori density $\pi_{r_0}(r_i) = \frac{r_i}{\sigma^2}\exp -\frac{r_i^2}{2\sigma^2}$, $r_i \in [0, \infty)$ (Rayleigh distributed), then the following estimators are obtained.*

(a) The incomplete data likelihood ratio defined by $\hat{\Lambda}(t_m) = \overline{E}[\Lambda(t_m)|\mathcal{F}_m^{\mathcal{Y}}]$ is given by

$$\begin{aligned}\hat{\Lambda}(t_m) &= \alpha_m(1) = \int_{[0,\infty)^n}\int_{[0,2\pi]^n}\overline{\alpha}_m(\boldsymbol{\theta}, \mathbf{r})d\boldsymbol{\theta}d\mathbf{r} \qquad (4.64) \\ &= \prod_{i=1}^{N}\left[\frac{1}{1 + 2\sigma^2K_i^m}\exp\left(\frac{V_i^2(y^m)\sigma^2}{2 + 4\sigma^2K_i^m}\right)\right]\end{aligned}$$

$$\times \exp\Big(\frac{1}{2}\sum_{k=0}^{m} y^2(t_k)[1-D^{-2}(t_k)] - \sum_{k=0}^{m}\log|D(t_k)|\Big). \tag{4.65}$$

(b) The normalized conditional density of $(\boldsymbol{\theta},\mathbf{r})$ given $\mathcal{F}_m^{\mathcal{Y}}$, i.e., $p_N(t_m,\boldsymbol{\theta},\mathbf{r}|\mathcal{F}_m^{\mathcal{Y}})$, is given by

$$\begin{aligned} p_N(t_m,\boldsymbol{\theta},\mathbf{r}|\mathcal{F}_m^{\mathcal{Y}}) &= \frac{\overline{\alpha}_m(\boldsymbol{\theta},\mathbf{r})}{\int\int \overline{\alpha}_m(\boldsymbol{\theta},\mathbf{r})d\boldsymbol{\theta}d\mathbf{r}} = \prod_{i=1}^{N} p_N(t_m,\theta_i,r_i|\mathcal{F}_m^{\mathcal{Y}}) && (4.66)\\ &= \prod_{i=1}^{N}\Big[\frac{r_i(1+2\sigma^2K_i^m)}{2\pi\sigma^2}\exp\Big(-\frac{r_i^2(1+2\sigma^2K_i^m)}{2\sigma^2}\Big) \\ &\quad \times \exp\Big(r_iV_i(y^m)\cos(\theta_i-\gamma_i(y^m))\Big) \\ &\quad \times \exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2K_i^m}\Big)\Big]. && (4.67)\end{aligned}$$

(c) The conditional least-square estimate of the noiseless received signal $h_i(t_m,\theta_i,r_i)$ given $\mathcal{F}_m^{\mathcal{Y}}$, defined by $\hat{h}_i(t_m,\theta_i,r_i)\stackrel{\triangle}{=}E[h_i(t_m,\theta_i,r_i)|\mathcal{F}_m^{\mathcal{Y}}]$ is given by

$$\begin{aligned}\hat{h}_i(t_m,\theta_i,r_i) &= \int_0^{\infty}\int_0^{2\pi} h_i(t_m,\theta_i,r_i)p_N(t_m,\theta_i,r_i|\mathcal{F}_m^{\mathcal{Y}})d\theta_i dr_i && (4.68)\\ &= S(t_m-\tau_i(t_m))\cos(\omega_c(t_m-\tau_i(t_m))+\gamma_i(y^m))\frac{V_i(y^m)\sigma^2}{1+2\sigma^2K_i^m}. && (4.69)\end{aligned}$$

(d) The minimum least-square estimator of θ_i given $\mathcal{F}_m^{\mathcal{Y}}$ is given by

$$\begin{aligned}\tilde{\theta}_i^{*,\infty}(t_m) &= E\Big[\theta_i|\mathcal{F}_m^{\mathcal{Y}}\Big] = \int_0^{\infty}\int_0^{2\pi}\theta_i p_N(t_m,\theta_i,r_i|\mathcal{F}_m^{\mathcal{Y}})d\theta_i dr_i && (4.70)\\ &= \frac{(1+2\sigma^2K_i^m)}{2\pi\sigma^2}\exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2K_i^m}\Big) \\ &\quad \times \int_0^{\infty}\int_0^{2\pi}\theta_i\, r_i\exp\Big(-\frac{r_i^2(1+2\sigma^2K_i^m)}{2\sigma^2}\Big) \\ &\quad \times \exp\Big(r_iV_i(y^m)\cos(\theta_i-\gamma_i(y^m))\Big)d\theta_i dr_i. && (4.71)\end{aligned}$$

(e) The minimum least-square estimator of r_i given $\mathcal{F}_m^{\mathcal{Y}}$ is given by

$$\begin{aligned}\tilde{r}_i^{*,\infty}(t_m) &= E\Big[r_i|\mathcal{F}_m^{\mathcal{Y}}\Big] = \int_0^{2\pi}\int_0^{\infty} r_i p_N(t_m,\theta_i,r_i|\mathcal{F}_m^{\mathcal{Y}})dr_i d\theta_i && (4.72)\\ &= (\sqrt{\pi/2})\Big(\sqrt{\frac{\sigma^2}{1+2\sigma^2K_i^m}}\Big)\exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2K_i^m}\Big){}_1F_1\Big(\frac{3}{2},1;\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2K_i^m}\Big) && (4.73)\end{aligned}$$

where ${}_1F_1\Big(\alpha,\beta;x\Big)$ *is the confluent hypergeometric function [52].*

Proof. See Appendix D. □

Remark 4.4.5. *The results of Theorem 4.4.4 can be used in decision problems and in nonlinear estimation problems.*

The detection problem associated with (4.49) is described by the following binary hypothesis problem.

$$\begin{aligned} H_1:\ y(t_k) &= \sum_{i=1}^{N}\Big[A_i(t_k)r_i\cos(\omega_c(t_k-\tau_i(t_k))+\theta_i)S(t_k-\tau_i(t_k))\Big]+D(t_k)v(t_k)\\ H_2:\ y(t_k) &= D(t_k)v(t_k). \end{aligned}$$

The incomplete data likelihood ratio $\hat{\Lambda}(t_m)$ *can be used in minimum risk Bayes' decision applications, where there will be one distribution* $\alpha_m(x)$ *for each hypothesis. That is, in binary hypothesis testing the decision rule takes the form*

$$\frac{\int_{\Re^n} d\alpha_m^1(x)}{\int_{\Re^n} d\alpha_m^2(x)} = \frac{\hat{\Lambda}^1(t_m)}{\hat{\Lambda}^2(t_m)} \underset{H_2}{\overset{H_1}{\gtrless}} \gamma, \tag{4.74}$$

where γ *is the threshold. Clearly,* $\alpha_m(x)$ *is a sufficient statistic for estimation and decision problems, since both the least-square estimation and decision rule are constructed from this quantity.*

The normalized conditional density $p_N(t_m,x|\mathcal{F}_m^{\mathcal{Y}})$ *can be used to compute various least-square estimates as was shown in (4.71) and (4.73). Hence, the general expression for the least-square estimate can by written as*

$$E[\Phi(x_m)|\mathcal{F}_m^{\mathcal{Y}}] = \int_{\Re^n}\Phi(x_m)p_N(t_m,x|\mathcal{F}_m^{\mathcal{Y}})dx. \tag{4.75}$$

Notice that from (4.69) the following least-squares estimate can also be derived

$$E\Big[r_i\cos(\omega_c(t_k-\tau_i(t_k))+\theta_i)|\mathcal{F}_m^{\mathcal{Y}}\Big] = \cos(\omega_c(t_m-\tau_i(t_m))+\gamma_i(y^m))\frac{V_i(y^m)\sigma^2}{1+2\sigma^2K_i^m}. \tag{4.76}$$

Both estimators, (4.69) and (4.76), can be used in RAKE receiver applications. In a RAKE receiver, one RAKE finger is assigned to each multipath, thus maximizing the amount of received signal energy. Each of these different paths are combined to form

a composite signal which has substantially better characteristics for the purpose of demodulation, when compared to a single path. In previous work [52] in order to combine the different paths meaningfully, the RAKE receiver needs the knowledge of the channel parameters. However, using (4.69), (4.76) a RAKE receiver can be constructed without knowledge of the channel parameters.

4.4.4 Numerical Results and Discussion

Here the performance of the estimators derived in Sections 4.4.2 and 4.4.3 is evaluated. Their performance is evaluated thought the Mean Square Error function of each estimator for each arriving path as a function of time for $N = 100$ realizations. The general expression of the MSE is given by

$$MSE_{Z_i} = \frac{1}{N}\sum_{j=1}^{N}|Z_i - \tilde{Z}^*_{i,j}|^2, \tag{4.77}$$

where N is the number of realizations, Z_i is the real value of the parameter that would be estimated for path i and $\tilde{Z}^*_{i,j}$ is the estimated value of that parameter for path i.

The experiments presented here are for evaluation purposes and do not represent real life scenarios. A set of observations y is created for each case based on the non-coherent model given by (4.49). This model represents the baseband signal and all the various parameters, except the phase θ and attenuation r, are assumed to be known. Therefore, a relatively small carrier frequency, $f_c = 1KHz$, will be used thought this section together with a transmitted signal $S(t_k) = 1$ and a Gaussian noise $v(t_k) \sim N(0, \sigma_n^2)$. Finally, the Signal to Noise Ratio is defined as $SNR = \frac{P_s}{\sigma_n^2}$, were P_s is the power of the transmitted signal.

For all the experiments presented below Wolfram Mathematica is being used.

Minimax Non-Coherent Model

Firstly, the robust phase and envelope estimators derived in Section 4.4.2 are evaluated, that is the derived phase estimator (4.56) and attenuation estimator (4.57).

A single path version of the non-coherent model given by (4.49) is used and a set of observations y is created with reference to this model. The true phase θ and

attenuation r parameters are random variables, rayleigh distributed, with $\acute{a}$ priori densities $\pi_\theta(\theta) = \frac{\theta}{\sigma_\theta^2}\exp\left(-\frac{\theta^2}{2\sigma_\theta^2}\right)$, $\theta \in [0, 2\pi]$ and $\pi_r(r) = \frac{r}{\sigma_r^2}\exp\left(-\frac{r^2}{2\sigma_r^2}\right)$, $r \in [0, \infty)$, respectively. The attenuation parameters of the above rayleigh distributions are taken as $\sigma_\theta = 1.5$ and $\sigma_r = \frac{1}{\sqrt{2}}$, respectively. The true distributions of the phase and the attenuation are unknown to the observer and the estimators use the nominal distributions which are assumed to be a uniform distribution for the phase, with $\acute{a}$ priori density $\pi_{\theta_0}(\theta_i) = \frac{1}{2\pi}$, $\theta_i \in [0, 2\pi]$, and a rayleigh distribution for the attenuation, with an attenuation parameter of $\sigma_r = \frac{5}{\sqrt{2}}$. As was mentioned above, the performance of the derived estimators will be evaluated using the Mean Square Error (4.77) of each estimator over a period of time of $10msec$, for $SNR = 20dB$, and $N = 100$ realizations.

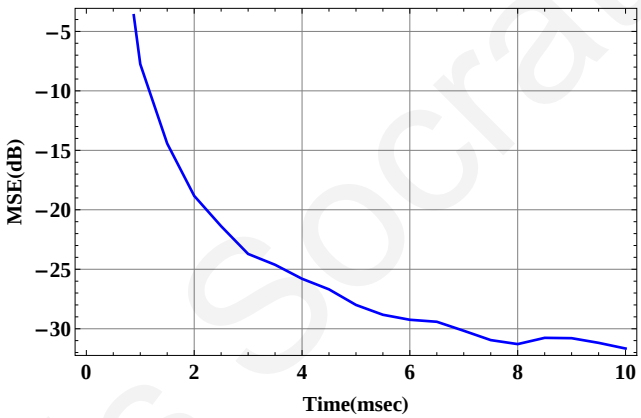


Figure 4.1: MSE of robust minimax phase estimator for a reference SNR=20 dB

Starting with the robust phase estimation, Figure 4.1 displays the performance of the phase estimator (4.56) over time for a reference $SNR = 20dB$. Clearly the MSE decreases with time, as time increases and this shows that the estimated phase converges in mean square sense to the real phase.

Next, the performance of the robust envelope estimator (4.57) is evaluated. Figure 4.2 displays the performance of the attenuation estimator over time for a reference $SNR = 20dB$. The observations are similar with the ones experienced for the robust phase MSE and show that the MSE decreases as time increases and that the estimated attenuation converges in mean square sense to its real value.

Therefore, even though the nominal distributions that were chosen for both the phase and the attenuation were different from the true ones the estimated parameters converges in mean square sense to their real values.

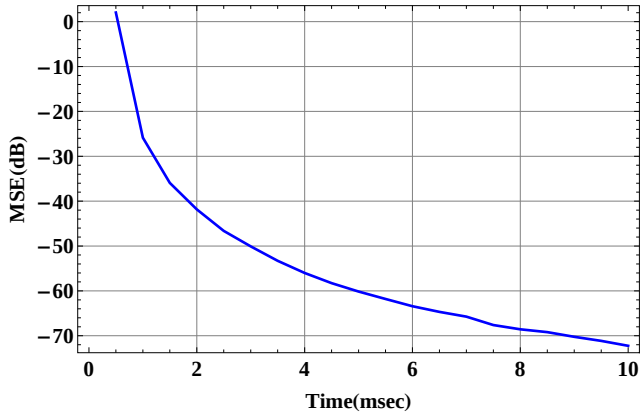


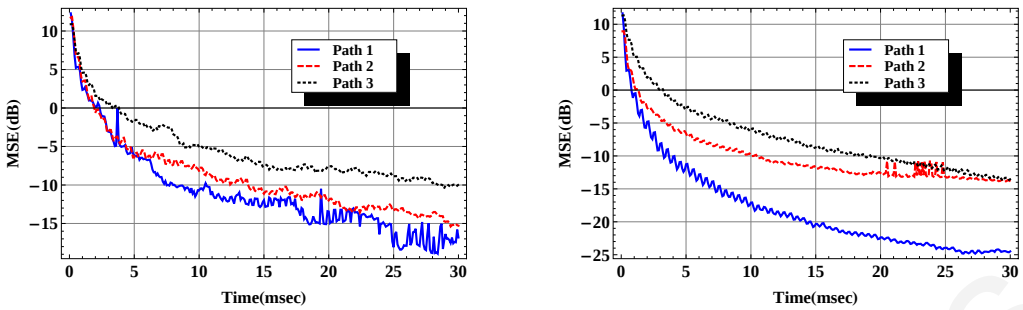
Figure 4.2: MSE of robust minimax attenuation estimator for a reference SNR=20 dB

Classical Non-Coherent Model

Next the phase, envelope and noiseless received signal estimators derived in Section 4.4.3 are evaluated. A multipath version of the classical non-coherent model, given by (4.49) with three resolvable paths, is considered here. It is assumed that the phases θ_i are iid random variables with *á* priori density $\pi_{\theta_0}(\theta_i) = \frac{1}{2\pi}$, $\theta_i \in [0, 2\pi]$ and the attenuations r_i are iid random variables with *á* priori density $\pi_{r_0}(r_i) = \frac{r_i}{\sigma^2_{r_i}} \exp\left(-\frac{r_i^2}{2\sigma^2_{r_i}}\right)$, $r_i \in [0, \infty)$ (Rayleigh distributed). It is also assumed that the attenuation parameter σ_{r_i} and the time delays $\tau_i(t_k)$ for each path, are constant over time and known.

The performance of the derived estimators of Section 4.4.3 is evaluated using the Mean Square Error (4.77) of each estimator for each path over a period of time of $30msec$, for $N = 100$ realizations. The Signal to Noise ratio for the first receiving path is assumed to be $10dB$. A smaller SNR is used compared to the previous experiments since here the real distributions of the parameters are known. This is the reference SNR for the system. The SNR of the second and third receiving paths is $9dB$ and $8dB$ respectively.

Starting with the phase estimation, Figure 4.3(a) displays the performance of the phase estimator (4.71) over time for each arriving path for a reference $SNR = 10dB$. The MSE decreases with time for all three paths, as time increases. This shows that the estimated phase converges in mean square sense to the real phase, as time increases, for all three paths.



(a) Reference SNR=10dB (b) Reference SNR=20dB

Figure 4.3: MSE of the phase estimator

Similarly, the performance of the channel attenuation estimates is evaluated. Figure 4.4 displays the performance of the attenuation estimator (4.73) over time for each arriving path for a reference $SNR = 10dB$. The observations, are similar with the ones experienced for the phase MSE. Clearly, the MSE decreases as time increases.

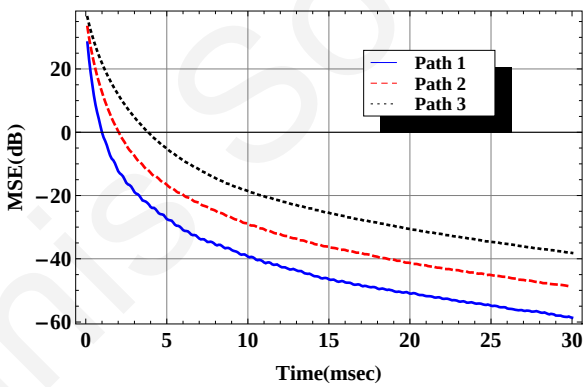


Figure 4.4: MSE of the attenuation estimator, for a reference SNR=10 dB

Next, the conditional least-square estimate of the noiseless received signal for every path, defined by $h_i(t_m, \theta_i, r_i) \stackrel{\triangle}{=} E[h_i(t_m, \theta_i, r_i)|\mathcal{F}_m^{\mathcal{Y}}]$ is considered. Figure 4.5 displays the performance of the estimator (4.69) over time for each arriving path for a reference $SNR = 10dB$. The estimator converges to the real value, in mean square sense, as time increases for all three paths. This is verified by the fact that the MSE decrease to a very small value as time increases.

Finally, when comparing the results of the derived estimators it is obvious that the

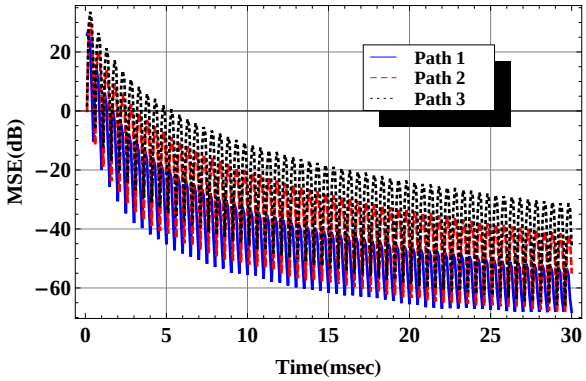


Figure 4.5: MSE of the noiseless received signal estimator, for a reference SNR=10 dB

attenuation estimator and the noiseless received signal estimator perform better than the phase estimator. Since a $10dB$ SNR can be considered as a low to medium system, the reference SNR is increased to $20dB$ and the phase estimator is once again evaluated. Figure 4.3(b) displays the performance of the phase estimator (4.71) over time for each arriving path for a reference $SNR = 20dB$. Notice that the estimator produces better results, in MSE sense, than the previous evaluation. Clearly, SNR influences the performance of the estimator.

4.5 Summary

This chapter considers robust estimation for autoregressive channel models, and employs the KL distance to describe the uncertainty. The methodology presented addresses the maximization problem using variational methods, while for the minimization problem it invokes a change of probability measure technique. The change of probability measure technique introduced is being used in order to derive a recursive equation for the unnormalized $\acute{a}$ posteriori distribution of nonlinear filtering problems. The theory developed is applied to two specific models, the linear Gaussian model and the non-coherent multipath model. The results derived include new robust least-square estimators for both estimation problems. Moreover, the robust minimax estimators for the non-coherent model are linked with classical estimators (e.g., by reducing the model uncertainty to zero).

CHAPTER 5

GENERALIZED MAP AND ML ESTIMATION

In this chapter two well-known estimation techniques, the Maximum $\acute{A}$ Posteriori (MAP) technique and the Maximum Likelihood (ML) technique are being investigated. First, in Section 5.2 the MAP estimation technique is presented and a new generalized estimator is derived. Then, in Section 5.3 a different approach in deriving a ML estimator is presented. These generalized estimators deal with situations when the true distribution is unknown but belongs to a specific set described by relative entropy constraint. They are also obtained by modifying the uniform cost function of the classical MAP technique. The chapter ends with specific examples.

5.1 Introduction

In Bayesian statistics, the maximum $\acute{a}$ posteriori estimate is defined as the mode of the posterior distribution, if posterior is unimodal. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data. It is closely related to Fisher's method of maximum likelihood, but employs an augmented optimization objective which incorporates a prior distribution over the quantity one wants to estimate. MAP estimation can therefore be seen as a regularization of ML estimation.

MAP estimates can be computed in several ways [53].

1. Analytically, when the mode(s) of the posterior distribution can be given in closed form. This is the case when conjugate priors are used.

2. Via numerical optimization such as the conjugate gradient method or Newton's method. This usually requires first or second derivatives, which have to be evaluated analytically or numerically.

3. Via a modification of an expectation-maximization algorithm. This does not require derivatives of the posterior density.

4. Via a Monte Carlo method using simulated annealing.

While MAP estimation is a limit of Bayes estimators (under the $0-1$ cost function), it is not very representative of Bayesian methods in general. This is because MAP estimates are point estimates, whereas Bayesian methods are characterized by the use of distributions to summarize data and draw inferences: thus, Bayesian methods tend to report the posterior mean or median instead, together with credible intervals [1].

The classical maximum $\acute{a}$ posteriori estimation theory is developed by using a uniform cost function of $0-1$. In this chapter a new technique is developed, using the principles of the MAP estimation theory, with the only difference lying in the way that the cost function, is defined. In this chapter instead of a uniform cost function an exponential one is being used. The goal is to derive a generalized MAP estimator which includes as a special case the classical estimator, when the cost function is uniform.

For many observation models arising in practise it is not possible to apply the above technique, either because of intractability of the required analysis or because of the lack of a useful complete statistic. For such models, an alternative non Bayesian method is the maximum likelihood estimation technique. The idea behind maximum likelihood parameter estimation is to determine the parameters that maximize the probability (likelihood) of the sample data. From the statistical point of view, maximum likelihood estimation is regarded to be robust while it yields estimators with good statistical and convergence properties. In other words, ML estimation methods are versatile and apply to most models and to different types of data. In addition, they provide efficient methods for quantifying uncertainty through confidence bounds. Although the methodology for maximum likelihood estimation is simple, the implementation is mathematically intense.

For "large" samples, several results from central limit theorem are applicable yielding ML estimators which have the following properties [53]:

1. ML estimators are asymptotically normally distributed.
2. ML estimators are asymptotically "minimum variance."
3. ML estimators are asymptotically unbiased (ML estimators are often biased, but the bias $\rightarrow$ 0 as n $\rightarrow \infty$).

Maximum likelihood estimation represents the backbone of statistical estimation. It is based on deep theory, originally developed by R. A. Fisher. While beginning classes often focus on least squares estimation ("regression"); likelihood theory is the omnibus approach across the sciences, engineering and medicine. This chapter deal with the methodology used to derive a ML estimator and applies a new approach. The goal is to derive a generalized ML estimator using, like the MAP case, an exponential function.

5.2 Generalized Maximum $\acute{A}$ Posteriori Estimation

In this section the Maximum $\acute{A}$ Posteriori method is being investigated. First the problem is defined by introducing the appropriate spaces and the general concept. Then the generalized estimator is derived by using the theory behind the classical estimator but by applying an exponential cost function.

5.2.1 Abstract Formulation

Suppose a measurable space $\left(\Omega, \mathcal{F}\right)$ is given on which the unobserved Random Variable (RV), X and the observed RV Y are defined, via $X : (\Omega, \mathcal{F}) \rightarrow (\mathcal{X}, \Sigma_{\mathcal{X}})$, $Y : (\Omega, \mathcal{F}) \rightarrow (\mathcal{Y}, \Sigma_{\mathcal{Y}})$.

Thus $\mathcal{X}$ is the space of the unobserved RV, and $\mathcal{Y}$ is the space of the observed RV. The relation between the unobserved RV X and the observed RV Y is defined via a probabilistic mapping. In the Chapter 3 two different probabilistic mappings were used, the mapping $\mu : \mathcal{X} \times \Sigma_{\mathcal{Y}} \rightarrow [0, 1]$ and the mapping $\eta : \mathcal{Y} \times \Sigma_{\mathcal{X}} \rightarrow [0, 1]$. These two mapping, as defined in Sections 3.2.1 and 3.2.3 are also going to be used in this chapter.

The objective is to estimate X by a function of the random variable Y. Let $\hat{X} = \Phi(Y)$ denote the estimate of X. The previous chapters deal with the problem of Least-Square estimation. Here, a different estimation is investigated, which also belongs to the Bayesian group, the Maximum $\acute{A}$ Posteriori estimation method.

The general theory is based on introducing a pay-off, and then minimizing the pay-off over $x \in \mathcal{X}$. As mentioned in previous paragraph, both Least-Square estimation and MAP estimation belong to the same set of estimation methods, the Bayesian Estimation. What distinguishes the different estimation methods is the determination of the appropriate pay-off function.

5.2.2 Derivation of Generalized Estimator

Let $\ell : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be an $\Sigma_{\mathcal{X}} \times \Sigma_{\mathcal{X}}$-measurable function, and let $s \in \Re$. Then a new cost function is introduced defined by

$$C(x, \Phi(y)) = \begin{cases} e^{\frac{\ell(x)}{s}}, & \text{if } |x - \Phi(y)| > \Delta \\ \\ 0, & \text{if } |x - \Phi(y)| \leq \Delta \end{cases} \tag{5.1}$$

where $\Delta > 0$. As shown in Fig. 5.1, the cost function $C(x, \Phi(y))$ for $|x - \Phi(y)| > \Delta$ is a convex increasing function of the error. The classical MAP estimation assumes a uniform cost function $C(x, \Phi(y)) = 1$ for $|x - \Phi(y)| > \Delta$ and 0 for $|x - \Phi(y)| \leq \Delta$. Here it is assumed that the cost function is not uniform but an exponential function of x. Note that the classical case is included in this cost function, since $s \to \infty$, implies $C(x, \Phi(y)) = 1$ for $|X - \Phi(y)| > \Delta$ (red graph in Fig. 5.1).

For the above cost function the average pay-off is given by

$$\begin{aligned} E[C(X, \Phi(Y))] &= \int_{\mathcal{X} \times \mathcal{Y}} e^{\frac{\ell(x)}{s}} I_{\{x:|x-\Phi(y)|>\Delta\}}(x) dP_{X,Y}(x, y) \\ &= \int_{\mathcal{Y}} dP_Y(y) \int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} I_{\{x:|x-\Phi(y)|>\Delta\}}(x) \eta(y, dx) \end{aligned} \tag{5.2}$$

where I is an index function defined by

$$I_{\{x:|x-\Phi(y)|>\Delta\}}(w) = \begin{cases} 1, & \text{if } w \in \{x : |x - \Phi(y)| > \Delta\} \\ \\ 0, & \text{if } w \notin \{x : |x - \Phi(y)| > \Delta\} \end{cases}$$

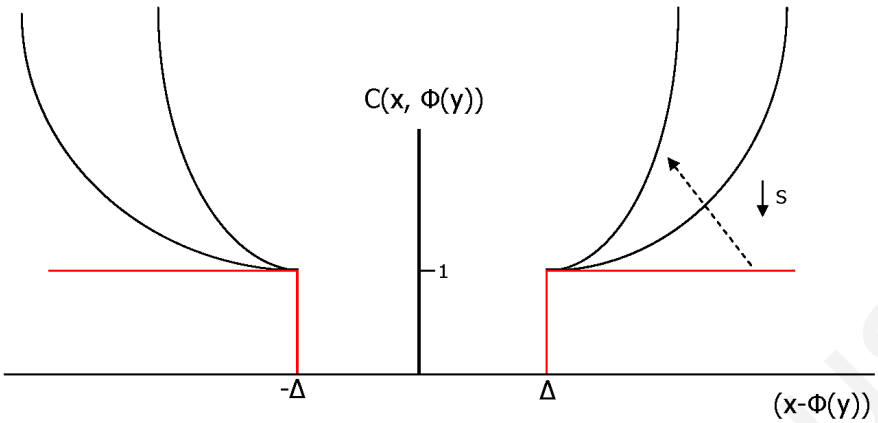


Figure 5.1: MAP Cost Function

Note that for a given $Y = y$, the space $\mathcal{X}$ of the unobserved RV, is partitioned into

$$\mathcal{X} = \{x : |x - \Phi(y)| > \Delta\} \bigcup \{x : |x - \Phi(y)| \leq \Delta\} \tag{5.3}$$

and the index function $I_{\mathcal{X}}$ of the space the unobserved RV is given by

$$I_{\mathcal{X}}(x) = I_{\{x:|x-\Phi(y)|>\Delta\}}(x) + I_{\{x:|x-\Phi(y)|\leq\Delta\}}(x) = 1. \tag{5.4}$$

Therefore, the index function $I_{\{|x-\Phi(y)|>\Delta\}}(x)$ is given by

$$I_{\{x:|x-\Phi(y)|>\Delta\}}(x) = 1 - I_{\{x:|x-\Phi(y)|\leq\Delta\}}(x). \tag{5.5}$$

Using (5.5) with (5.2) the average pay-off is given by

$$\begin{aligned} E[C(X, \Phi(Y))] &= \int_{\mathcal{Y}} dP_Y(y) \int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} \Big[1 - I_{\{x:|x-\Phi(y)|\leq\Delta\}}(x)\Big] \eta(y, dx) \\ &= \int_{\mathcal{Y}} dP_Y(y) \Bigg[\int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} \eta(y, dx) - \int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} I_{\{x:|x-\Phi(y)|\leq\Delta\}}(x) \eta(y, dx) \Bigg]. \end{aligned} \tag{5.6}$$

Then, as mention before, the estimation is done by minimizing the average pay-off (5.6) over $x \in \mathcal{X}$. But, since $dP_Y(y)$ is nonnegative and not a function of $\Phi(y)$, the estimation is done by minimizing the inner integral, which is given by

$$\begin{aligned} E[C(X, \Phi(Y))|Y = y] &= \int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} I_{\{x:|x-\Phi(y)|>\Delta\}}(x) \eta(y, dx) \\ &= \int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} \eta(y, dx) - \int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} I_{\{x:|x-\Phi(y)|\leq\Delta\}}(x) \eta(y, dx). \end{aligned} \tag{5.7}$$

The first part of (5.7) given by $\int_{\mathcal{X}} e^{\frac{\ell(x)}{s}}\eta(y,dx)$ is a nonnegative number that can be estimated and which is also independent of x, so the minimization of the average pay-off (5.2) can be done by maximizing the second part of (5.7) given by

$$\begin{aligned} E\left[e^{\frac{\ell(x)}{s}} I_{\{x:|x-\Phi(y)|>\Delta\}}(x)\right] &= \int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} I_{\{x:|x-\Phi(y)|\leq\Delta\}}(x)\eta(y,dx) \\ &= \int_{\Phi(y)-\Delta}^{\Phi(y)+\Delta} e^{\frac{\ell(x)}{s}}\eta(y,dx). \end{aligned} \tag{5.8}$$

If $\eta(y,dx)$ is a smooth function of x and if $\Delta = dx/2$ is sufficiently small, then

$$\int_{\Phi(y)-dx/2}^{\Phi(y)+dx/2} e^{\frac{\ell(x)}{s}}\eta(y,dz) \simeq e^{\frac{\ell(x)}{s}}\eta(y,dx) \tag{5.9}$$

and the right-hand side is maximized by choosing $\Phi(y)$ to be the value of x maximizing $e^{\frac{\ell(x)}{s}}\eta(y,dx)$, over $x \in \mathcal{X}$.

Therefore,

$$\Phi_{MAP}(y) = \arg\max_{x\in\mathcal{X}} e^{\frac{\ell(x)}{s}}\eta(y,dx). \tag{5.10}$$

However, if $f_\eta(x|y) \stackrel{\triangle}{=} \frac{\eta(y,dx)}{dx}$ exists then

$$\Phi_{MAP}(y) \stackrel{\triangle}{=} \arg\max_{x\in\mathcal{X}} e^{\frac{\ell(x)}{s}} f_\eta(y|x). \tag{5.11}$$

In modeling a given statistical situation the family of conditional distributions (or stochastic kernels) of Y given $X = x$ are needed, and for the Bayesian formulation the prior distribution for X is also needed. The conditional distribution of X given Y (or the stochastic kernel $\eta(y,dx)$) can be obtained from the prior and the conditional of Y given X (or the stochastic kernel $\mu(x,dy)$) by applying Bayes' formula.

$$\eta(y,dx) = \frac{\mu(x,dy)dP_X(x)}{\int_{\mathcal{X}}\mu(x,dy)dP_X(x)} = \frac{\mu(x,dy)dP_X(x)}{dP_Y(y)}. \tag{5.12}$$

The MAP estimator in (5.10) can be obtained using (5.12) but without the computation of $dP_Y(y)$ since this term will not affect the maximization over x. That is,

$$\Phi_{MAP}(y) = \arg\max_{x\in\mathcal{X}} e^{\frac{\ell(x)}{s}}\mu(x,dy)dP_X(x). \tag{5.13}$$

Since the logarithm is an increasing function, $\Phi_{MAP}(y)$ can be obtained by maximizing the following function

$$\log\left[e^{\frac{\ell(x)}{s}}\frac{\mu(x,dy)}{dy}\frac{dP_X(x)}{dx}\right] = \frac{\ell(x)}{s} + \log\frac{\mu(x,dy)}{dy} + \log\frac{dP_X(x)}{dx}. \tag{5.14}$$

5.2.3 Connection with Minimax Approach

In the classical MAP estimation method, the MAP estimator $\Phi_{MAP}(y)$ is obtained by choosing $\Phi(y)$ to be the value of x maximizing the *á* posteriori distribution $\eta(y, dx)$, over $x \in \mathcal{X}$. The connection of the new MAP estimate to robustness is described below.

It is assumed that this probabilistic kernel, $\eta(y, dx)$, represents the nominal system model or mapping (*á* posteriori information) and that true kernel $\nu(y, dx)$, denoted by $\nu : \mathcal{Y} \times \Sigma_{\mathcal{X}} \to [0, 1]$ belongs to an uncertainty set described by

$$\mathcal{B}(\eta) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{P} : \int_{\mathcal{Y}} H(\nu|\eta)(y) dP_Y(y) \leq \int_{\mathcal{Y}} R(y) dP_Y(y) \stackrel{\nabla}{=} r_2\Big\},$$

similar with Section 3.2.3 of Chapter 3. The sample pay-off $\ell(x, \Phi(y))$ used there is replaced by the pay-off $\ell(x)$, then the worst case measure of the true kernel is given by

$$\nu^*(y, dx) = \frac{e^{\frac{\ell(x)}{s}} \eta(y, dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x)}{s}} \eta(y, dx)} \tag{5.15}$$

and the MAP estimator $\Phi_{MAP}(y)$ is obtained by choosing $\Phi(y)$ to be the value of x maximizing $\nu^*(y, dx)$, over $x \in \mathcal{X}$. This is the same as maximizing only the numerator.

That is,

$$\Phi_{MAP}(y) = \arg \max_{x \in \mathcal{X}} e^{\frac{\ell(x)}{s}} \eta(y, dx), \tag{5.16}$$

which is the same result as obtained in Section 5.2.2 above.

5.3 Generalized Maximum Likelihood Estimation

In this section, the Maximum Likelihood estimation method is investigated. First the appropriate spaces are introduced and the problem is set up. Then the generalized estimator is derived by using the theory behind the classical estimator in combination with the theory of the generalized MAP estimator described in the previous section.

5.3.1 Abstract Formulation

For many observation models arising in practice, it is not possible to apply the Bayesian Estimation methods like Least-Square Estimation and Maximum $\acute{A}$ Posteriori Estimation, mainly due to the lack of a useful complete sufficient statistic. For such models a very commonly used method of designing estimators is the Maximum Likelihood method.

Suppose that a measurable space $\Big(\Omega, \mathcal{F}\Big)$ is given on which the unobserved RV, X and the observed RV Y are defined, via $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \Sigma_{\mathcal{X}})$, $Y : (\Omega, \mathcal{F}) \to (\mathcal{Y}, \Sigma_{\mathcal{Y}})$.

Thus $\mathcal{X}$ is the space of the unobserved RV, and $\mathcal{Y}$ is the space of the observed RV. The relation between the unobserved RV X and the observed RV Y is defined via the same probabilistic mapping as the previous section.

The classical Maximum Likelihood estimation technique can be derived from the Maximum $\acute{A}$ Posteriori method. In the absence of any prior information about the RV X which has to be estimated, it is assumed that is uniformly distributed in its range since this represents more or less a worst-case prior. The same assumption is used in the generalized ML estimation technique described in this section.

5.3.2 Derivation of Generalized Estimator

The generalized MAP estimator derived in Section 5.2 is considered here, which is given by (5.13)

$$\Phi_{MAP}(y) = \arg\max_{x \in \mathcal{X}} e^{\frac{\ell(x)}{s}} \mu(x, dy) dP_X(x).$$

In the absence of any prior information about the parameter, it is assumed that it is uniformly distributed in its range since this represents more or less a worst-case prior. Applying the assumption above, the MAP estimate for a given $y \in \mathcal{Y}$ is any value of x that maximizes $e^{\frac{\ell(x)}{s}} \mu(x, dy)$ over $x \in \mathcal{X}$. Since $\mu(x, dy)$ as a function of x is sometimes called the likelihood function, the classical estimate is called maximum likelihood estimate. Here the same name is used for the derived estimate, which is denoted by $\Phi(y)_{ML}$ and is given by

$$\Phi_{ML}(y) = \arg\max_{x \in \mathcal{X}} e^{\frac{\ell(x)}{s}} \frac{\mu(x, dy)}{dy} dy. \tag{5.17}$$

Maximizing $e^{\frac{\ell(x)}{s}}\mu(x,dy)$ is equivalent to maximizing $\log\left[e^{\frac{\ell(x)}{s}}\frac{\mu(x,dy)}{dy}\right] = \frac{\ell(x)}{s} + \log\frac{\mu(x,dy)}{dy}$, and assuming sufficient smoothness of this function, a necessary condition for the maximum likelihood estimate is

$$\frac{\partial}{\partial x}\left[\frac{\ell(x)}{s} + \log\frac{\mu(x,dy)}{dy}\right]\Bigg|_{x=\Phi_{ML}(y)} = 0. \tag{5.18}$$

Note that when $s \to \infty$ then the generalized estimator converges to the classical ML estimator.

5.4 Examples

Next, some examples are presented to illustrate applicability of the theoretical results.

Given X is a RV and $Y^m = \{Y_1, Y_2, \ldots, Y_m\}$ is a sequence of RVs defined on $\Big(\Omega, \mathcal{F}, P\Big)$, which are related via the model

$$Y_i = HX + W_i, \quad i = 0, \ldots, m \tag{5.19}$$

Hence $Y^m : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^{(m+1)d}, \mathcal{B}(\Re^{(m+1)d})\Big)$ the observed sequence of RVs, $X : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^n, \mathcal{B}(\Re^n)\Big)$ the unobserved RV, and $W^m : \Big(\Omega, \mathcal{F}\Big) \to \Big(\Re^{(m+1)d}, \mathcal{B}(\Re^{(m+1)d})\Big)$ the noise sequence of RVs. It is assumed W_i and X are independent Gaussian Random Variables, $N(0, \Sigma_W)$, $\Sigma_W > 0$, $N(0, \Sigma_X)$, $\Sigma_X > 0$. Assuming (5.19) denotes the system model, then

$$\mu(x, dy^m) = \prod_{i=0}^{m} \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_W|^{\frac{1}{2}}}\, e^{-(y_i - Hx)^T \frac{\Sigma_W^{-1}}{2}(y_i - Hx)} dy_i, \tag{5.20}$$

$$P_X(dx) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_X|^{\frac{1}{2}}}\, e^{-x^T \frac{\Sigma_X^{-1}}{2} x} dx. \tag{5.21}$$

Also the following function is defined

$$\ell(x) = x^T U x, \quad U = U^T > 0. \tag{5.22}$$

Complete derivations of the results presented below can be found in Appendix E.

5.4.1 Generalized MAP Estimator

In this first example the generalized MAP estimator for the model (5.19) is computed together with the mean square error for this MAP estimator.

First the MAP estimator $\Phi_{MAP}(y)$ is given by

$$\Phi_{MAP}(y) = \arg\max_{x\in\mathcal{X}} \left\{ \frac{\ell(x)}{s} + \log\frac{\mu(x,dy)}{dy} + \log\frac{dP_X(x)}{dx} \right\}. \tag{5.23}$$

Using (5.20), (5.21) and (5.22) the MAP estimator is given by

$$\Phi_{MAP}(y) = \arg\max_{x\in\mathcal{X}_{ad}} \overline{L}(x, y^m) \tag{5.24}$$

where

$$\begin{aligned} \overline{L}(x,y^m) &= \frac{x^T U x}{s} + \log\left(\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_W|^{\frac{1}{2}}}\right)^{m+1} - \sum_{i=0}^{m}(y_i - Hx)^T\frac{\Sigma_W^{-1}}{2}(y_i - Hx) \\ &\quad + \log\left(\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_X|^{\frac{1}{2}}}\right) - x^T\frac{\Sigma_X^{-1}}{2}x. \end{aligned} \tag{5.25}$$

The MAP estimator $\Phi_{MAP}(y^m)$ of X is computed by differentiating the above pay-off function and setting the derivative to zero as shown below.

$$\frac{\partial}{\partial x}\overline{L}(x,y^m)|_{x=\Phi_{MAP}} = 0. \tag{5.26}$$

So, by differentiating (5.25) the final generalized MAP estimator is derived

$$\Phi_{MAP}(y^m) = \left((m+1)H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - \frac{2U}{s}\right)^{-1} H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i. \tag{5.27}$$

Next the mean square error for $\Phi_{MAP}(y^m)$ is computed. The MSE is given by

$$\begin{aligned} &E\Big[(X - \Phi_{MAP}(Y^m))^T(X - \Phi_{MAP}(Y^m))\Big] = \\ &tr\Big(E\Big[(X - \Phi_{MAP}(Y^m))(X - \Phi_{MAP}(Y^m))^T\Big]\Big). \end{aligned} \tag{5.28}$$

The mean square error can also be written as

$$\begin{aligned} &E\Big[(X - \Phi_{MAP}(Y^m))^T(X - \Phi_{MAP}(Y^m))\Big] = \\ &tr\int_{\Re^n}\int_{\Re^{(m+1)d}}(x - \Phi_{MAP}(y^m))(x - \Phi_{MAP}(y^m))^T\mu(x,dy)dP_X(x). \end{aligned} \tag{5.29}$$

After some manipulation the above function can be expressed as

$$E\Big[(X-\Phi_{MAP}(Y^m))^T(X-\Phi_{MAP}(Y^m))\Big] = \\ tr\Big(\Sigma_X - 2(m+1)H^T\Sigma^T_{\Phi_{MAP}}\Sigma_X + (m+1)^2H^T\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}H\Sigma_X \\ +(m+1)\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}\Sigma_W\Big) \tag{5.30}$$

where $\Sigma_{\Phi_{MAP}} = \Big((m+1)H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - \frac{2U}{s}\Big)^{-1}H^T\Sigma_W^{-1}$.

5.4.2 Generalized ML Estimator

In the second example the generalized ML estimator for the model (5.19) is computed together with the mean square error for this ML estimator. The $Cramér - Rao\ lower\ bound$ is also investigated. For this example it is assumed that $X = x$ is a parameter.

The ML estimator $\Phi_{ML}(y)$ is given by

$$\Phi_{ML}(y) = \arg\max_{x\in\mathcal{X}_{ad}}\left\{\frac{\ell(x)}{s} + \log\frac{\mu(x,dy)}{dy}\right\}. \tag{5.31}$$

Using (5.20) and (5.22) the ML estimator is given by

$$\Phi_{ML}(y) = \arg\max_{x\in\mathcal{X}}\overline{L}_2(x,y^m) \tag{5.32}$$

where

$$\overline{L}_2(x,y^m) = \frac{x^TUx}{s} + \log\left(\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_W|^{\frac{1}{2}}}\right)^{(m+1)} - \sum_{i=0}^{m}(y_i-Hx)^T\frac{\Sigma_W^{-1}}{2}(y_i-Hx). \tag{5.33}$$

The generalized ML estimator $\Phi_{ML}(y)$ of X is computed by differentiating the above pay-off function and setting the derivative to zero as shown below.

$$\frac{\partial}{\partial x}\overline{L}_2(x,y^m)|_{x=\Phi_{ML}} = 0. \tag{5.34}$$

Therefore, by differentiating (5.33) the final generalized ML estimator is derived

$$\Phi_{ML}(y^m) = \left((m+1)H^T\Sigma_W^{-1}H - \frac{2U}{s}\right)^{-1}H^T\Sigma_W^{-1}\sum_{i=0}^{m}y_i. \tag{5.35}$$

Next, the mean square error for $\Phi_{ML}(y^m)$ is computed. The MSE is given by

$$\begin{aligned} & E\Big[(x-\Phi_{ML}(Y^m)^T(x-\Phi_{ML}(Y^m))\Big] \\ & = tr\Big(E\Big[(x-\Phi_{ML}(Y^m))(x-\Phi_{ML}(Y^m))^T\Big]\Big). \end{aligned} \tag{5.36}$$

Given the assumption that in ML estimation X is deterministic, after some manipulation the above function is expressed as

$$\begin{aligned} & E\Big[(x-\Phi_{ML}(Y^m))^T(x-\Phi_{ML}(Y^m))\Big] = \\ & tr\Big(xx^T - 2(m+1)H^T\Sigma_{\Phi_{ML}}^T xx^T + (m+1)^2H^T\Sigma_{\Phi_{ML}}^T\Sigma_{\Phi_{ML}}Hxx^T \\ & +(m+1)\Sigma_{\Phi_{ML}}^T\Sigma_{\Phi_{ML}}\Sigma_W\Big) \end{aligned} \tag{5.37}$$

where $\Sigma_{\Phi_{ML}} = \left((m+1)H^T\Sigma_W^{-1}H - \frac{2U}{s}\right)^{-1} H^T\Sigma_W^{-1}$.

Finally, it is going to be investigated if this new generalized ML estimator satisfies the $Cramér - Rao\ lower\ bound$.

First, it is determined wether the estimator is biased, so the expectation of the estimator $\Phi_{ML}(y^m)$ is computed.

$$E[\Phi_{ML}(Y^m)] = \Sigma_{\Phi_{ML}}\sum_{i=0}^{m}E[Y_i] = \Sigma_{\Phi_{ML}}\sum_{i=0}^{m}Hx = \Sigma_{\Phi_{ML}}(m+1)Hx \neq x. \tag{5.38}$$

This shows that the estimator is biased.

The $Cramér - Rao\ Inequality$ for biased estimator is given by

$$Var[\Phi_{ML}(Y^m)] \geq V_X I_X^{-1} V_X^T \tag{5.39}$$

where

$$[V_X]_{ij} \triangleq \frac{\partial}{\partial x_j}E_X[\Phi_{ML}(y^m)_i], 1 \leq i,j \leq n \tag{5.40}$$

$$[I_X]_{ij} \triangleq -E_X\left[\frac{\partial^2}{\partial x_i \partial x_j}\log\frac{\mu(x,dy^m)}{dy^m}\right]. \tag{5.41}$$

Next the LHS of (5.39) is computed

$$\begin{aligned} Var[\Phi_{ML}(Y^m)] &= E\Big[\Big(\Phi_{ML}(Y^m) - E[\Phi_{ML}(Y^m)]\Big)\Big(\Phi_{ML}(Y^m) - E[\Phi_{ML}(Y^m)]\Big)^T\Big] \\ &= (m+1)\Sigma_{\Phi_{ML}}\Sigma_W\Sigma_{\Phi_{ML}}^T. \end{aligned} \tag{5.42}$$

Then the RHS of (5.39) is computed. Given (5.40),

$$V_X = (m+1)\Sigma_{\Phi_{ML}} H \tag{5.43}$$

and given (5.41),

$$I_X^{-1} = \Big((m+1)H^T\Sigma_w^{-1}H\Big)^{-1} \tag{5.44}$$

provided that H is a square and invertible matrix.

Then

$$V_X I_X^{-1} V_X^T = (m+1)\Sigma_{\Phi_{ML}} H\Big((m+1)H^T\Sigma_w^{-1}H\Big)^{-1}(m+1)H^T\Sigma_{\Phi_{ML}}^T. \tag{5.45}$$

Therefore, assuming that H has size $d \times d$ and is invertible, then

$$\begin{aligned} V_X I_X^{-1} V_X^T &= (m+1)\Sigma_{\Phi_{ML}} H\Big((m+1)H^T\Sigma_w^{-1}H\Big)^{-1}(m+1)H^T\Sigma_{\Phi_{ML}}^T \\ &= (m+1)\Sigma_{\Phi_{ML}} H H^{-1}\Sigma_w (H^T)^{-1}H^T\Sigma_{\Phi_{ML}}^T = (m+1)\Sigma_{\Phi_{ML}}\Sigma_W\Sigma_{\Phi_{ML}}^T. \end{aligned} \tag{5.46}$$

Given the above, the $Cramér - Rao\ lower\ bound$ is satisfied with equality

$$Var[\Phi_{ML}(Y^m)] = V_X I_X^{-1} V_X^T. \tag{5.47}$$

5.5 Summary

This chapter considers a generalized approach to the classical MAP and ML estimation. A new cost function is used for the case of MAP estimation and a generalized estimator is derived. Furthermore, a new approach is used, closely related to the MAP theory developed, in order to derive a generalized estimator for the ML estimation problem. The theory developed is applied to examples in order to illustrate the applicability of the results.

CHAPTER 6

CONCLUSION

In this chapter a summary of the main results of the thesis is presented alongside with some suggestions for future work plans.

6.1 Synopsis

This thesis deals with the subject of signal estimation from noisy measurements under uncertainty. The work presented is focused mainly on the problem of robust least-square estimation for uncertain systems. In addition, it also investigates the methodology behind two well-known and used estimation techniques, the Maximum $\acute{A}$ Posteriori estimation technique and the Maximum Likelihood estimation technique.

Robust estimation is a common problem appearing frequently in statistics and signal processing and its been under study for many years. Usually the performance of a system depends on the $\acute{a}$ priori knowledge of its input characteristics and even small deviations from the assumed conditions can affect it drastically. One of the most widely used ways on dealing with modeling uncertainties is the minimax approach. This approach has been applied to many detection and estimation problems and has been used for designing robust schemes by optimizing the worst case performance. This thesis uses probability distributions, or general measures, defined on measurable spaces to model the the uncertainty description and the nominal description of the system. Specifically, two type of uncertainty models are being considered: 1) Uncertainty Models on Conditional Distributions or otherwise known Stochastic

Kernels; 2) Uncertainty Models on Joint Distributions.

Stochastic kernel uncertainty models are being implemented in the design of communication systems, where the input message has a known distribution, while the channel is unknown but belongs to a certain class of channels. Here two types of uncertain stochastic models are being considered: i) when the conditional probability distribution of the measurement Y given the signal to be estimated X, or channel kernel, is unknown; ii) when the $\acute{a}$ posteriori distribution of X given Y is unknown. Joint distribution uncertainty models, on the other hand, are usually employed when both the unobserved and observed random variables are uncertain. The uncertainty description of these models is characterized by the class of uncertain measures which satisfy a relative entropy constraint with respect to a nominal measure. The problem of robust estimation is formulated by minimizing over the set of estimators, the maximum of a linear functional of the uncertain measure over the Kullback-Leibler (KL) distance constraint set.

In this thesis a general framework is presented where the basic ideas are explained and the fundamental results are derived. Abstract Polish spaces are being used to describe the least-square estimation problem, while,as was mentioned above, the uncertain model considered is described by stochastic kernels and joint distributions. Once the appropriate space of measures are introduced the maximizing kernels and joint measure are computed explicitly using Lagrangian functionals and variational methods. Furthermore, certain properties of the optimal solution are presented, including performance bounds, which are important for the numerical computation of the optimal solution. The theoretical results obtained are applied to various examples. The theory developed is also applied to Multiple-Input Multiple-Output (MIMO) communication systems, and to finite-dimensional autoregressive channel models, in order to derive robust estimators for a class of uncertain models. The methodology is presented, where the maximization is addressed using variational methods, while the minimization is addressed using a change of probability measure technique. The change of probability measure technique introduced is being used in order to derive recursive equations for the conditional distribution of nonlinear filtering problem. The theory is then applied first to a linear Gaussian model and then to an attenuated sinusoid in a multipath environment, which is subject to an additive Gaussian noise, i.e., to a non-coherent multipath model. For the linear Gaussian model a robust estimator is derived, which resembles the Kalman filter. For the non-coherent multipath model robust phase and enveloped estimators are

derived, and a connection with the classical least-square estimation is presented.

Finally, in addition to the robust least-square estimation, this thesis takes a closer look into the derivation of estimators for the Maximum $\acute{A}$ Posteriori estimation method and the Maximum Likelihood estimation method. The MAP estimation technique is based on the minimization of a specific uniform cost function. In this thesis, the classical cost function is replaced by an exponential cost function, and a generalized MAP estimator is derived. A connection with the minimax approach used in the least-square problem is also presented. This generalized estimator includes as a special case the classical MAP estimator. A similar approach is also implemented for the case of the ML estimation technique. The methodology used for ML estimation involves the maximization of a likelihood function. Starting with the MAP estimation method and assuming that the parameter to be estimated is uniformly distributed in its range, the new likelihood function is derived. Through the above approach a generalized likelihood function is introduced, which also includes an exponential function. Examples are presented that illustrate the application of the derived results to various problems.

6.2 Directions for Future Research

The robust least-square estimation method together with the new generalized MAP and ML estimators presented in this thesis raise a few issues related to

1. Robust signal detection;
2. Power Spectral Density uncertainty models;
3. Total variation distance constraint between measures;
4. Application of the models to real-time applications, which involve also MIMO applications;
5. Information theory applications;
6. Relation of Mutual Information and MMSE.

This thesis deals with the subject of signal estimation. Besides signal estimation, one of the most pervasive of functions that signal processing schemes are required to carry out is that of detecting a signal of a generally known type in noisy observations. Most signal detection problems can be cast in the framework of $M-ary$ hypothesis testing, in which there is an observation, usually a vector or function, on the basis of which a decision has to be made among M possible statistical situations describing the observation [1]. Obvious examples of applications in which signal detection is required are provided by radar (detection of echo pulses) and sonar (detection of a random signal present in an array of hydrophones). Numerous other applications may be listed; for example, detection of specified two-level pulse-code sequences in communication systems, and detection of abnormal patterns in medical imaging. The subject of signal detection and estimation deals with the processing of information-bearing signals in order to make inferences about the information that they contain.

The most fundamental problem of signal detection is the determination of the likelihood ratio for detecting signals against a noise background. This is a problem with a rich history in both electrical engineering and mathematics, and its solution has involved, over the years, a wide variety of mathematical tools and flashes of intuition. For one thing, mathematical models are often significant simplifications/ idealizations of complex physical problems. Secondly, even if the model is reasonably good, the knowledge of the parameters in it, e.g., covariance functions, time constants, etc., may not be enough to justify a direct numerical evaluation of formulas derived from the model. The major engineering goal is to obtain structural insights into the mathematical solutions of classes of special problems, with the hope that these insights can then be used to intelligently modify and adapt the mathematical solution to the particular physical problem at hand [4]. The work presented in this thesis can be used in detection problems in order to obtain robust detection scheme. The uncertainty models can be implemented to describe the systems. Also, as was shown in Section 4.2.3 a change of probability measure can be used to obtain the conditional distribution. In decision applications, there will be one distribution $\alpha_m^s(x)$ for each hypothesis, e.g., in binary hypothesis the decision rule takes the form

$$\frac{\inf_{\tilde{x}_m \in \mathcal{X}_{ad}} \int_{\Re^n} d\alpha_m^{s,1}(x)}{\inf_{\tilde{x}_m \in \mathcal{X}_{ad}} \int_{\Re^n} d\alpha_m^{s,2}(x)} \underset{H_2}{\overset{H_1}{\gtrless}} \gamma$$

where γ is the threshold.

Moreover, another area for future research, which is related to signal detection problems, is to device decision rules for a class of probability distribution.

In statistical signal processing and physics, the spectral density, power spectral density is a positive real function of a frequency variable associated with a stationary stochastic process, or a deterministic function of time, which has dimensions of power per Hz. It is often called simply the spectrum of the signal. Intuitively, the spectral density captures the frequency content of a stochastic process and helps identify periodicities.

The concept and use of the power spectrum of a signal is fundamental in electronic engineering, especially in electronic communication systems (radio & microwave communications, radars, and related systems). As was mention in Chapter 1, the classes of allowable characteristics one deals with in robust signal processing are generally nonparametric function classes, such as the class of all power spectral density functions with specified total power (area under the function) and which lie between specified upper and lower bounding functions. Uncertainty models using the concept of PSD for robust estimation problems can be found in [15] and [30]. It will be interesting to combine the minimax approach presented in this thesis with the concept of power spectrum density and also use the concept of KL distance to define the constraint sets. The KL distance has been introduced between spectral density functions of stationary stochastic processes in [54].

In this thesis the uncertainty description of the system, and the nominal description of the system are modeled by probability distributions, or general measures, defined on measurable spaces. The uncertainty description of these models is characterized by the class of uncertain measures which satisfy a KL distance constraint with respect to a nominal measure. Over the last few years, the KL distance uncertainty model has received particular attention due to various properties (convexity, compact level sets), its simplicity and its connection to risk sensitive pay-off, minimax games, and large deviations. Unfortunately, KL distance uncertainty modeling has two disadvantages: 1) it does not define a true metric on the space of measures; 2) relative entropy between two measures is not defined if the measures are not absolutely continuous. The latter rules out the possibility of measures $\nu \in \mathcal{M}_1(\Sigma)$ and $\mu \in \mathcal{M}_1(\Sigma)$ to be defined on different spaces[1]. It is one of the main disadvantages

[1]This corresponds to the case in which the nominal system is a simplified version of the true system and is defined on a lower dimension space.

of employing relative entropy in the context of uncertainty modeling for stochastic controlled diffusions (or stochastic differential equations). Motivated by the above issues, the KL distance constrain can be replaced by an uncertainty model based on the total variation distance defined on the space measures. This uncertainty set is described by a ball with respect to the total variation norm, centered at the nominal measure having positive radius.

Given a known or nominal probability measure $\mu \in \mathcal{M}_1(\Sigma)$ the uncertainty set based on total variation distance is defined by

$$B_R(\mu) \stackrel{\triangle}{=} \Big\{\nu \in \mathcal{M}_1(\Sigma) : ||\nu - \mu|| \leq R\Big\}$$

where $R \in [0, \infty)$. The total variation distance[2] on $\mathcal{M}_1(\Sigma) \times \mathcal{M}_1(\Sigma)$ is defined by

$$||\alpha - \beta|| \stackrel{\triangle}{=} \sup_{P \in \mathcal{P}(\Sigma)} \sum_{F_i \in P} |\alpha(F_i) - \beta(F_i)|, \quad \alpha, \beta \in \mathcal{M}_1(\Sigma)$$

where $\mathcal{P}(\Sigma)$ denotes the collection of all finite partitions of Σ. The above distance satisfies the properties of a metric, and does not require absolute continuity of measures when defining the uncertainty ball, i.e., singular measures are admissible and the measures need not be defined on the same space. It can very well be the case that $\mu \in \mathcal{M}_1(\Sigma)$ and $\nu \in \mathcal{M}_1(\tilde{\Sigma})$ where $\Sigma \subset \tilde{\Sigma}$. Since $\mathcal{M}_1(\Sigma)$ are probability measures then it follows that the radius of uncertainty belongs to the restricted set $R \in [0, 2]$.

Clearly, the total variation distance uncertainty set is larger than the KL distance uncertainty set. This can be concluded from Pinsker's inequality [55] as follows.

$$||\nu - \mu||^2 \leq 2H(\nu|\mu), \quad \nu, \mu \in \mathcal{M}_1(\Sigma), \text{ if } \nu << \mu, \ \log \frac{d\nu}{d\mu} \in L_1(\nu).$$

Hence, even for those measures which satisfy $\nu << \mu$ and $\log \frac{d\nu}{d\mu} \in L_1(\nu)$ the uncertainty set described by relative entropy is a subset of the much larger total variation distance uncertainty set, that is, $A_{\frac{R^2}{2}}(\mu) \subset B_R(\mu)$. In the parlance of stochastic differential equations the total variation distance covers the case when both drift and diffusion coefficients of the stochastic differential equations are uncertain.

This thesis presents mostly theoretical. A natural extension will be to apply the results to real-time applications and evaluate their performance. Special emphasis could be given to MIMO communication systems as they are vastly used in real time.

[2]The definition of total variation distance applies to signed measures as well.

Finally, it is noted that aside from minimax estimation, the tools developed in this thesis can also be used in Information theory. For example, in computing maxmin capacity for a class of channels, and minimax rate distortion for a class of sources.

The mutual information, which is at the core of information theory, is an indicator of how much coded information can be pumped through a channel reliably given a certain input signaling. As was already presented in Section 1.1.4, [31] presents a new formula that connects the input-output mutual information and the minimum mean-square error achievable by optimal estimation of the input given the output. That is, given that the input-output mutual information and the MMSE are monotone functions of the signal-to-noise ratio (SNR), denoted by $I(snr)$ and $mmse(snr)$, respectively, the mutual information in nats and the MMSE satisfy the following relationship regardless of the input statistics:

$$\frac{d}{dsnr}I(snr) = \frac{1}{2}mmse(snr). \tag{6.1}$$

This relationship holds for both scalar and vector signals. Also based on this relationship various other properties of MMSE are investigated in [32] and [33], like the MMSE dimension. All the above results assume that there is no uncertainty in the model being used. The work presented in this thesis, can be used to developed a new relationship between mutual information and MMSE for a class of models which are subject to uncertainty.

APPENDIX A

BASIC MATRIX IDENTITIES

In this thesis the following Matrix identities are being used [56].

1. Basic Identity. Given matrices A and B (both size $n \times n$ and invertible), then

$$(AB)^{-1} = B^{-1}A^{-1}. \tag{A.1}$$

2. Woodbury Identity. Given invertible matrix A (size $n \times n$), matrix C (size $n \times d$), and invertible matrix B (size $d \times d$), then

$$(A + CBC^T)^{-1} = A^{-1} - A^{-1}C(B^{-1} + C^T A^{-1}C)^{-1}C^T A^{-1}. \tag{A.2}$$

3. PosDef Identity. If matrices P and R are assumed to be positive definite and invertible, then

$$(P^{-1} + B^T R^{-1}B)^{-1}B^T R^{-1} = PB^T(BPB^T + R)^{-1}. \tag{A.3}$$

Moreover, the following identities for Determinants are being used in this thesis.

I. Given square matrices A and B, then

$$|A^T| = |A|, \tag{A.4}$$
$$|AB| = |A||B|, \tag{A.5}$$
$$|A^n| = |A|^n. \tag{A.6}$$

II. Given a square and invertible matrix A, then

$$|A|^{-1} = \frac{1}{|A|}. \tag{A.7}$$

III. Given matrix A with size $n \times m$ and matrix B with size $m \times n$, then

$$|I_n + AB| = |I_m + BA| \tag{A.8}$$

where matrix I_n has size $n \times n$ and matrix I_m has size $m \times m$.

APPENDIX B

PROOF OF REMARK 2.2.20

According to Theorem 2.2.19, $\{\overline{\alpha}_k(x)\} \stackrel{\triangle}{=} \{\pi_k(x)\}$ satisfies the following recursion

$$\overline{\alpha}_k(x) = \frac{\Xi_{v_k}(D_k^{-1}(y_k - C_k x))}{|D_k|\Xi_{v_k}(y_k)} \int_{\Re_n} \frac{\Psi_{w_k}(B_k^{-1}(x - A_k z))}{|B_k|} \overline{\alpha}_{k-1}(z) dz \tag{B.1}$$

where

$$\Xi_{v_k}(D_k^{-1}(y_k - C_k x)) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\Big(- \frac{(y_k - C_k x)^T (D_k D_k^T)^{-1} (y_k - C_k x)}{2}\Big), \tag{B.2}$$

$$\Xi_{v_k}(y_k) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\Big(- \frac{y_k^T y_k}{2}\Big), \tag{B.3}$$

$$\Psi_{w_k}(B_k^{-1}(x - A_k z)) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big(\frac{-(x - A_k z)^T (B_k B_k^T)^{-1} (x - A_k z)}{2}\Big). \tag{B.4}$$

The solution of (B.1) is assumed to have the following form.

$$\overline{\alpha}_k(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |V_{k|k}|^{\frac{1}{2}}} \exp\Big(- (x - \hat{x}_{k|k})^T \frac{(V_{k|k})^{-1}}{2} (x - \hat{x}_{k|k}) + \beta_{k|k}\Big). \tag{B.5}$$

Then,

$$\begin{aligned} \overline{\alpha}_{k-1}(z) &= \frac{1}{(2\pi)^{\frac{n}{2}} |V_{k-1|k-1}|^{\frac{1}{2}}} \\ &\quad \times \exp\Big(- (z - \hat{x}_{k-1|k-1})^T \frac{(V_{k-1|k-1})^{-1}}{2} (z - \hat{x}_{k-1|k-1}) + \beta_{k-1|k-1}\Big). \end{aligned} \tag{B.6}$$

The derivation starts with the integral part of (B.1). Then using (B.4) and (B.6) this can be expressed as

$$\int_{\Re_n} \frac{\Psi_{w_k}(B_k^{-1}(x - A_k z))}{|B_k|} \overline{\alpha}_{k-1}(z) dz = \int_{\Re_n} \frac{1}{(2\pi)^{\frac{n}{2}} |B_k| (2\pi)^{\frac{n}{2}} |V_{k-1|k-1}|^{\frac{1}{2}}}$$

$$\times \exp\Big(\frac{-(x-A_kz)^T(B_kB_k^T)^{-1}(x-A_kz)}{2} -(z-\hat{x}_{k-1|k-1})^T\frac{(V_{k-1|k-1})^{-1}}{2}(z-\hat{x}_{k-1|k-1}) +\beta_{k-1|k-1}\Big). \quad \text{(B.7)}$$

Next, the exponential term of B.7 is expanded as shown below.

$$\begin{aligned}
&\exp\Big(\frac{-(x-A_kz)^T(B_kB_k^T)^{-1}(x-A_kz)}{2} - (z-\hat{x}_{k-1|k-1})^T\frac{(V_{k-1|k-1})^{-1}}{2}(z-\hat{x}_{k-1|k-1}) \\
&+\beta_{k-1|k-1}\Big) \\
&= -x^T\frac{(B_kB_k^T)^{-1}}{2}x - z^TA_k^T\frac{(B_kB_k^T)^{-1}}{2}A_kz + z^TA_k^T(B_kB_k^T)^{-1}x - z^T\frac{(V_{k-1|k-1})^{-1}}{2}z \\
&-\hat{x}_{k-1|k-1}^T\frac{(V_{k-1|k-1})^{-1}}{2}\hat{x}_{k-1|k-1} + z^T(V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1} + \beta_{k-1|k-1} \\
&= -\frac{1}{2}z^T\Big((V_{k-1|k-1})^{-1} + A_k^T(B_kB_k^T)^{-1}A_k\Big)z \\
&+z^T\Big(A_k^T(B_kB_k^T)^{-1}x + (V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1}\Big) \\
&-x^T\frac{(B_kB_k^T)^{-1}}{2}x - \hat{x}_{k-1|k-1}^T\frac{(V_{k-1|k-1})^{-1}}{2}\hat{x}_{k-1|k-1} + \beta_{k-1|k-1}.
\end{aligned}$$

Now, the above equation can be written as a quadratic function of z

$$\begin{aligned}
= \quad & -(z-N_1)^T\frac{\Sigma_1^{-1}}{2}(z-N_1) + N_1^T\frac{\Sigma_1^{-1}}{2}N_1 - x^T\frac{(B_kB_k^T)^{-1}}{2}x \\
& -\hat{x}_{k-1|k-1}^T\frac{(V_{k-1|k-1})^{-1}}{2}\hat{x}_{k-1|k-1} + \beta_{k-1|k-1}
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_1^{-1} &= (V_{k-1|k-1})^{-1} + A_k^T(B_kB_k^T)^{-1}A_k, \quad \text{(B.8)} \\
\Sigma_1^{-1}N_1 &= A_k^T(B_kB_k^T)^{-1}x + (V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1} \\
\Rightarrow N_1 &= \Big((V_{k-1|k-1})^{-1} + A_k^T(B_kB_k^T)^{-1}A_k\Big)^{-1} \\
&\quad \times\Big[A_k^T(B_kB_k^T)^{-1}x + (V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1}\Big]. \quad \text{(B.9)}
\end{aligned}$$

Using the fact that $\Sigma_1^T = \Sigma_1$, $(B_kB_k^T)^T = (B_kB_k^T)$ and $(V_{k-1|k-1})^T = (V_{k-1|k-1})$, then

$$\begin{aligned}
N_1^T\frac{\Sigma_1^{-1}}{2}N_1 &= x^T(B_kB_k^T)^{-1}A_k\frac{\Sigma_1}{2}A_k^T(B_kB_k^T)^{-1}x \\
&\quad +\hat{x}_{k-1|k-1}^T(V_{k-1|k-1})^{-1}\frac{\Sigma_1}{2}(V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1} \\
&\quad +x^T(B_kB_k^T)^{-1}A_k\Sigma_1(V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1}. \quad \text{(B.10)}
\end{aligned}$$

Since

$$\int_{\Re_n} \frac{1}{(2\pi)^{n/2}|\Sigma_1|^{1/2}} \exp\Big(-(z-N_1)^T \frac{\Sigma_1^{-1}}{2}(z-N_1)\Big)dz = 1$$

the integral term of (B.1) can be written as

$$\begin{aligned}
&\int_{\Re_n} \frac{\Psi_{w_k}(B_k^{-1}(x-A_kz))}{|B_k|}\overline{\alpha}_{k-1}(z)dz = \frac{|\Sigma_1|^{1/2}}{(2\pi)^{n/2}|B_k||V_{k-1|k-1}|^{1/2}} \\
&\times \exp\Big(N_1^T\frac{\Sigma_1^{-1}}{2}N_1 - x^T\frac{(B_kB_k^T)^{-1}}{2}x - \hat{x}_{k-1|k-1}^T\frac{(V_{k-1|k-1})^{-1}}{2}\hat{x}_{k-1|k-1} + \beta_{k-1|k-1}\Big) \\
&\times \int_{\Re_n} \frac{1}{(2\pi)^{n/2}|\Sigma_1|^{1/2}} \exp\Big(-(z-N_1)^T \frac{\Sigma_1^{-1}}{2}(z-N_1)\Big)dz \\
&= \frac{|\Sigma_1|^{1/2}}{(2\pi)^{n/2}|B_k||V_{k-1|k-1}|^{1/2}} \\
&\times \exp\Big(N_1^T\frac{\Sigma_1^{-1}}{2}N_1 - x^T\frac{(B_kB_k^T)^{-1}}{2}x - \hat{x}_{k-1|k-1}^T\frac{(V_{k-1|k-1})^{-1}}{2}\hat{x}_{k-1|k-1} + \beta_{k-1|k-1}\Big).
\end{aligned} \tag{B.11}$$

Next (B.11) is substituted into (B.1), using also (B.2) and (B.3), resulting in

$$\begin{aligned}
\overline{\alpha}_k(x) \;=\; & \frac{|\Sigma_1|^{1/2}}{(2\pi)^{n/2}|D_k||B_k||V_{k-1|k-1}|^{1/2}} \exp\Big(-(y_k-C_kx)^T\frac{(D_kD_k^T)^{-1}}{2}(y_k-C_kx) \\
& + \frac{y_k^Ty_k}{2} + N_1^T\frac{\Sigma_1^{-1}}{2}N_1 - x^T\frac{(B_kB_k^T)^{-1}}{2}x \\
& - \hat{x}_{k-1|k-1}^T\frac{(V_{k-1|k-1})^{-1}}{2}\hat{x}_{k-1|k-1} + \beta_{k-1|k-1}\Big)
\end{aligned} \tag{B.12}$$

The above expression can be simplified even more. First the exponential of (B.12) is simplified as shown below.

$$\begin{aligned}
&\exp\Big(-(y_k-C_kx)^T\frac{(D_kD_k^T)^{-1}}{2}(y_k-C_kx) + \frac{y_k^Ty_k}{2} + N_1^T\frac{\Sigma_1^{-1}}{2}N_1 - x^T\frac{(B_kB_k^T)^{-1}}{2}x \\
&-\hat{x}_{k-1|k-1}^T\frac{(V_{k-1|k-1})^{-1}}{2}\hat{x}_{k-1|k-1} + \beta_{k-1|k-1}\Big) \\
&= -y_k^T\frac{(D_kD_k^T)^{-1}}{2}y_k - x^TC_k^T\frac{(D_kD_k^T)^{-1}}{2}C_kx + x^TC_k^T(D_kD_k^T)^{-1}y_k + \frac{y_k^Ty_k}{2} \\
&+x^T(B_kB_k^T)^{-1}A_k\frac{\Sigma_1}{2}A_k^T(B_kB_k^T)^{-1}x + \hat{x}_{k-1|k-1}^T(V_{k-1|k-1})^{-1}\frac{\Sigma_1}{2}(V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1} \\
&+x^T(B_kB_k^T)^{-1}A_k\Sigma_1(V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1} - x^T\frac{(B_kB_k^T)^{-1}}{2}x \\
&-\hat{x}_{k-1|k-1}^T\frac{(V_{k-1|k-1})^{-1}}{2}\hat{x}_{k-1|k-1} + \beta_{k-1|k-1}
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}x^T\Big[C_k^T(D_kD_k^T)^{-1}C_k - (B_kB_k^T)^{-1}A_k\frac{\Sigma_1}{2}A_k^T(B_kB_k^T)^{-1} + (B_kB_k^T)^{-1}\Big]x \\
&+x^T\Big[C_k^T(D_kD_k^T)^{-1}y_k + (B_kB_k^T)^{-1}A_k\Sigma_1(V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1}\Big] \\
&+\hat{x}_{k-1|k-1}^T\Big[(V_{k-1|k-1})^{-1}\frac{\Sigma_1}{2}(V_{k-1|k-1})^{-1} - \frac{(V_{k-1|k-1})^{-1}}{2}\Big]\hat{x}_{k-1|k-1} \\
&+y_k^T\Big[\frac{1}{2}I_d - \frac{(D_kD_k^T)^{-1}}{2}\Big]y_k + \beta_{k-1|k-1}.
\end{aligned}
$$

Next, the above equation can be written as a quadratic function of x

$$
\begin{aligned}
&= -(x-\hat{x}_{k|k})^T\frac{(V_{k|k})^{-1}}{2}(x-\hat{x}_{k|k}) \\
&+\hat{x}_{k|k}^T\frac{(V_{k|k})^{-1}}{2}\hat{x}_{k|k} + \hat{x}_{k-1|k-1}^T\Big[(V_{k-1|k-1})^{-1}\frac{\Sigma_1}{2}(V_{k-1|k-1})^{-1} - \frac{(V_{k-1|k-1})^{-1}}{2}\Big]\hat{x}_{k-1|k-1} \\
&+y_k^T\Big[\frac{1}{2}I_d - \frac{(D_kD_k^T)^{-1}}{2}\Big]y_k + \beta_{k-1|k-1} \qquad \text{(B.13)}
\end{aligned}
$$

where

$$
(V_{k|k})^{-1} = C_k^T(D_kD_k^T)^{-1}C_k - (B_kB_k^T)^{-1}A_k\Sigma_1A_k^T(B_kB_k^T)^{-1} + (B_kB_k^T)^{-1} \qquad \text{(B.14)}
$$

$$
\begin{aligned}
&(V_{k|k})^{-1}\hat{x}_{k|k} = C_k^T(D_kD_k^T)^{-1}y_k + (B_kB_k^T)^{-1}A_k\Sigma_1(V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1} \\
&\Rightarrow \hat{x}_{k|k} = (V_{k|k})\Big[C_k^T(D_kD_k^T)^{-1}y_k + (B_kB_k^T)^{-1}A_k\Sigma_1(V_{k-1|k-1})^{-1}\hat{x}_{k-1|k-1}\Big].
\end{aligned} \qquad \text{(B.15)}
$$

Using (B.8) with (B.14) then

$$
\begin{aligned}
(V_{k|k})^{-1} &= C_k^T(D_kD_k^T)^{-1}C_k - (B_kB_k^T)^{-1}A_k\Big((V_{k-1|k-1})^{-1} + A_k^T(B_kB_k^T)^{-1}A_k\Big)^{-1} \\
&\quad\times A_k^T(B_kB_k^T)^{-1} + (B_kB_k^T)^{-1}. \qquad \text{(B.16)}
\end{aligned}
$$

Applying the Woodbury Identity (A.2), the above transforms to

$$
\begin{aligned}
(V_{k|k})^{-1} &= C_k^T(D_kD_k^T)^{-1}C_k + \Big((B_kB_k^T) + A_kV_{k-1|k-1}A_k^T\Big)^{-1} \qquad \text{(B.17)} \\
(V_{k|k}) &= \Big(C_k^T(D_kD_k^T)^{-1}C_k + V_{k|k-1}^{-1}\Big)^{-1} \qquad \text{(B.18)}
\end{aligned}
$$

where

$$
\begin{aligned}
V_{k|k-1} &= (B_kB_k^T) + A_kV_{k-1|k-1}A_k^T \qquad \text{(B.19)} \\
&= (B_kB_k^T) + A_k\Big(C_{k-1}^T(D_{k-1}^TD_{k-1})^{-1}C_{k-1} + V_{k-1|k-2}^{-1}\Big)^{-1}A_k^T
\end{aligned}
$$

$$
\begin{aligned}
&= (B_k B_k^T) + A_k \Big(V_{k-1|k-2} - V_{k-1|k-2} C_{k-1}^T (D_{k-1}^T D_{k-1} + C_{k-1} V_{k-1|k-2} C_{k-1}^T)^{-1} \\
&\quad C_{k-1} V_{k-1|k-2} \Big) A_k^T \\
&= (B_k B_k^T) + A_k V_{k-1|k-2} A_k^T - A_k V_{k-1|k-2} C_{k-1}^T (D_{k-1}^T D_{k-1} \\
&\quad + C_{k-1} V_{k-1|k-2} C_{k-1}^T)^{-1} C_{k-1} V_{k-1|k-2} A_k^T. \qquad \text{(B.20)}
\end{aligned}
$$

Using (B.8) with (B.15), then

$$
\begin{aligned}
\hat{x}_{k|k} &= (V_{k|k}) \Big[C_k^T (D_k D_k^T)^{-1} y_k \\
&\quad + (B_k B_k^T)^{-1} A_k \Big((V_{k-1|k-1})^{-1} + A_k^T (B_k B_k^T)^{-1} A_k \Big)^{-1} (V_{k-1|k-1})^{-1} \hat{x}_{k-1|k-1} \Big]. \qquad \text{(B.21)}
\end{aligned}
$$

Applying the PosDef Identity (A.3)

$$
\begin{aligned}
\hat{x}_{k|k} &= (V_{k|k}) \Big[C_k^T (D_k D_k^T)^{-1} y_k \\
&\quad + \Big((B_k B_k^T) + A_k V_{k-1|k-1} A_k^T \Big)^{-1} A_k V_{k-1|k-1} (V_{k-1|k-1})^{-1} \hat{x}_{k-1|k-1} \Big] \\
&= (V_{k|k}) \Big[C_k^T (D_k D_k^T)^{-1} y_k + \Big((B_k B_k^T) + A_k V_{k-1|k-1} A_k^T \Big)^{-1} A_k \hat{x}_{k-1|k-1} \Big] \\
&= (V_{k|k}) \Big[C_k^T (D_k D_k^T)^{-1} y_k + (V_{k|k-1})^{-1} \hat{x}_{k|k-1} \Big] \qquad \text{(B.22)}
\end{aligned}
$$

where

$$
\hat{x}_{k|k-1} = A_k \hat{x}_{k-1|k-1}. \qquad \text{(B.23)}
$$

The above equation can be manipulated a bit more in order to get the following

$$
\begin{aligned}
\hat{x}_{k|k} &= V_{k|k} \Big[C_k^T (D_k D_k^T)^{-1} y_k + (V_{k|k-1})^{-1} \hat{x}_{k|k-1} \Big] \\
&= \Big[(B_k B_k^T + A_k V_{k-1|k-1} A_k^T)^{-1} + C_k^T (D_k D_k^T)^{-1} C_k \Big]^{-1} \\
&\quad \times \Big[C_k^T (D_k D_k^T)^{-1} y_k + (B_k B_k^T + A_k V_{k-1|k-1})^{-1} A \hat{x}_{k-1|k-1} \Big] \\
&= \Big[(B_k B_k^T + A_k V_{k-1|k-1} A_k^T)^{-1} + C_k^T (D_k D_k^T)^{-1} C_k \Big]^{-1} C_k^T (D_k D_k^T)^{-1} y_k \\
&\quad + \Big[(B_k B_k^T + A_k V_{k-1|k-1} A_k^T)^{-1} + C_k^T (D_k D_k^T)^{-1} C_k \Big]^{-1} (B_k B_k^T + A_k V_{k-1|k-1})^{-1} \\
&\quad \times A \hat{x}_{k-1|k-1}. \qquad \text{(B.24)}
\end{aligned}
$$

By applying the PosDef Identity (A.3) the following holds.

$$\Big[(B_kB_k^T + A_kV_{k-1|k-1}A_k^T)^{-1} + C_k^T(D_kD_k^T)^{-1}C_k\Big]^{-1}C_k^T(D_kD_k^T)^{-1} = (B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T\Big[C_k(B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T + (D_kD_k^T)\Big]^{-1}. \tag{B.25}$$

Then, by applying the Woodbury Identity (A.2) the following holds.

$$\begin{aligned}\Big[(B_kB_k^T + A_kV_{k-1|k-1}A_k^T)^{-1} + C_k^T(D_kD_k^T)^{-1}C_k\Big]^{-1} &= (B_kB_k^T + A_kV_{k-1|k-1}A_k^T)\\ -(B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T&\Big[C_k(B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T + (D_kD_k^T)\Big]^{-1}\\ \times C_k(B_kB_k^T + A_kV_{k-1|k-1}A_k^T).&\end{aligned} \tag{B.26}$$

Finally, (B.25) and (B.26) are substitute into (B.24) to obtain

$$\begin{aligned}\hat{x}_{k|k} &= (B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T\Big[C_k(B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T + (D_kD_k^T)\Big]^{-1}y_k\\ &\quad + A\hat{x}_{k-1|k-1} - (B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T\\ &\quad \times\Big[C_k(B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T + (D_kD_k^T)\Big]^{-1}C_kA\hat{x}_{k-1|k-1}\\ &= A\hat{x}_{k-1|k-1} + (B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T\\ &\quad \times\Big[C_k(B_kB_k^T + A_kV_{k-1|k-1}A_k^T)C_k^T + (D_kD_k^T)\Big]^{-1}(y_k - C_kA\hat{x}_{k-1|k-1}) \qquad \text{(B.27)}\\ &= \hat{x}_{k|k-1} + V_{k|k-1}C_k^T\Big[C_kV_{k|k-1}C_k^T + (D_kD_k^T)\Big]^{-1}(y_k - C_k\hat{x}_{k|k-1}). \qquad \text{(B.28)}\end{aligned}$$

Substituting (B.28) into (B.23) yields

$$\begin{aligned}\hat{x}_{k|k-1} &= A_k\hat{x}_{k-1|k-2} + A_kV_{k-1|k-2}C_{k-1}^T\Big[C_{k-1}V_{k-1|k-2}C_{k-1}^T + (D_{k-1}^TD_{k-1})\Big]^{-1}\\ &\quad \times(y_{k-1} - C_{k-1}\hat{x}_{k-1|k-2}).\end{aligned} \tag{B.29}$$

This is the solution of the Kalman Filter.

Next, the term of equation (B.12) which is outside the integral is expanded.

$$\begin{aligned}\frac{|\Sigma_1|^{1/2}}{(2\pi)^{n/2}|D_k||B_k||V_{k-1|k-1}|^{1/2}} &= \frac{|\Sigma_1|^{1/2}|V_{k|k}|^{1/2}}{(2\pi)^{n/2}|D_k||B_k||V_{k-1|k-1}|^{1/2}|V_{k|k}|^{1/2}}\\ &= \frac{1}{(2\pi)^{n/2}|V_{k|k}|^{1/2}}\exp\Big(\log\frac{|\Sigma_1|^{1/2}|V_{k|k}|^{1/2}}{|D_k||B_k||V_{k-1|k-1}|^{1/2}}\Big).\end{aligned} \tag{B.30}$$

Substituting (B.30) and (B.13) into (B.12) yields

$$\begin{aligned}\overline{\alpha}_k(x) &= \frac{1}{(2\pi)^{n/2}|V_{k|k}|^{1/2}}\exp\Big(-(x-\hat{x}_{k|k})^T\frac{(V_{k|k})^{-1}}{2}(x-\hat{x}_{k|k})\\ &\quad+\hat{x}_{k|k}^T\frac{(V_{k|k})^{-1}}{2}\hat{x}_{k|k}+\hat{x}_{k-1|k-1}^T\Big[(V_{k-1|k-1})^{-1}\frac{\Sigma_1}{2}(V_{k-1|k-1})^{-1}\\ &\quad-\frac{(V_{k-1|k-1})^{-1}}{2}\Big]\hat{x}_{k-1|k-1}+y_k^T\Big[\frac{1}{2}I_d-\frac{(D_kD_k^T)^{-1}}{2}\Big]y_k+\beta_{k-1|k-1}\\ &\quad+\log\frac{|\Sigma_1|^{1/2}|V_{k|k}|^{1/2}}{|D_k||B_k||V_{k-1|k-1}|^{1/2}}\Big)\\ &= \frac{1}{(2\pi)^{n/2}|V_{k|k}|^{1/2}}\exp\Big(-(x-\hat{x}_{k|k})^T\frac{(V_{k|k})^{-1}}{2}(x-\hat{x}_{k|k})+\beta_{k|k}\Big)\end{aligned}\tag{B.31}$$

where

$$\begin{aligned}\beta_{k|k} &= \hat{x}_{k|k}^T\frac{(V_{k|k})^{-1}}{2}\hat{x}_{k|k}+\hat{x}_{k-1|k-1}^T\Big[(V_{k-1|k-1})^{-1}\frac{\Sigma_1}{2}(V_{k-1|k-1})^{-1}\\ &\quad-\frac{(V_{k-1|k-1})^{-1}}{2}\Big]\hat{x}_{k-1|k-1}+y_k^T\Big[\frac{1}{2}I_d-\frac{(D_kD_k^T)^{-1}}{2}\Big]y_k\\ &\quad+\log\frac{|\Sigma_1|^{1/2}|V_{k|k}|^{1/2}}{|D_k||B_k||V_{k-1|k-1}|^{1/2}}+\beta_{k-1|k-1}.\end{aligned}\tag{B.32}$$

Now, $\beta_{k|k}$ will be reshaped in a better form, starting from the expression below,

$$\begin{aligned}&\hat{x}_{k-1|k-1}^T\Big[(V_{k-1|k-1})^{-1}\frac{\Sigma_1}{2}(V_{k-1|k-1})^{-1}-\frac{(V_{k-1|k-1})^{-1}}{2}\Big]\hat{x}_{k-1|k-1}\\ &=\hat{x}_{k-1|k-1}^T\Big[(V_{k-1|k-1})^{-1}\frac{\Big((V_{k-1|k-1})^{-1}+A_k^T(B_kB_k^T)^{-1}A_k\Big)^{-1}}{2}(V_{k-1|k-1})^{-1}\\ &-\frac{(V_{k-1|k-1})^{-1}}{2}\Big]\hat{x}_{k-1|k-1}.\end{aligned}$$

Next, the Woodbury Identity (A.2) is applied in order to get

$$\begin{aligned}&=\hat{x}_{k-1|k-1}^T\frac{1}{2}\Big[(V_{k-1|k-1})^{-1}\Big(V_{k-1|k-1}-V_{k-1|k-1}A_k^T\Big((B_kB_k^T)+A_kV_{k-1|k-1}A_k^T\Big)^{-1}\\ &A_kV_{k-1|k-1}\Big)\times(V_{k-1|k-1})^{-1}-(V_{k-1|k-1})^{-1}\Big]\hat{x}_{k-1|k-1}\\ &=\hat{x}_{k-1|k-1}^T\frac{1}{2}\Big[(V_{k-1|k-1})^{-1}-A_k^T\Big((B_kB_k^T)+A_kV_{k-1|k-1}A_k^T\Big)^{-1}A_k\\ &-(V_{k-1|k-1})^{-1}\Big]\hat{x}_{k-1|k-1}\\ &=-\hat{x}_{k-1|k-1}^T\Big[\frac{A_k^T(V_{k|k-1})^{-1}A_k}{2}\Big]\hat{x}_{k-1|k-1}\end{aligned}$$

$$\begin{aligned}\Rightarrow\quad&\hat{x}_{k-1|k-1}^T\Big[(V_{k-1|k-1})^{-1}\frac{\Sigma_1}{2}(V_{k-1|k-1})^{-1}-\frac{(V_{k-1|k-1})^{-1}}{2}\Big]\hat{x}_{k-1|k-1}=\\ &-\hat{x}_{k|k-1}^T\Big[\frac{(V_{k|k-1})^{-1}}{2}\Big]\hat{x}_{k|k-1}.\end{aligned}\tag{B.33}$$

Next the expression $\hat{x}_{k|k}^T \frac{(V_{k|k})^{-1}}{2} \hat{x}_{k|k}$ is reshaped into a better form, by applying the Woodbury Identity (A.2) and the PosDef Identity (A.3).

$$\begin{aligned}
&\hat{x}_{k|k}^T \frac{(V_{k|k})^{-1}}{2} \hat{x}_{k|k} = \frac{1}{2}\Big[\hat{x}_{k|k-1} + V_{k|k-1}C_k^T\Big(C_k V_{k|k-1} C_k^T + (D_k D_k^T)\Big)^{-1} y_k \\
&-V_{k|k-1}C_k^T\Big(C_k V_{k|k-1} C_k^T + (D_k D_k^T)\Big)^{-1} C_k \hat{x}_{k|k-1}\Big]^T \Big[C_k^T (D_k D_k^T)^{-1} C_k + V_{k|k-1}^{-1}\Big] \\
&\Big[\hat{x}_{k|k-1} + V_{k|k-1}C_k^T\Big(C_k V_{k|k-1} C_k^T + (D_k D_k^T)\Big)^{-1} y_k - V_{k|k-1}C_k^T\Big(C_k V_{k|k-1} C_k^T \\
&+(D_k D_k^T)\Big)^{-1} C_k \hat{x}_{k|k-1}\Big] \\
&= \frac{1}{2}\hat{x}_{k|k-1}^T V_{k|k-1}^{-1} \hat{x}_{k|k-1} - \frac{1}{2} y_k^T \Big(C_k V_{k|k-1} C_k^T + (D_k D_k^T)\Big)^{-1} y_k + \frac{1}{2} y_k^T (D_k D_k^T)^{-1} y_k \\
&-\frac{1}{2}\hat{x}_{k|k-1}^T C_k^T \Big(C_k V_{k|k-1} C_k^T + (D_k D_k^T)\Big)^{-1} C_k \hat{x}_{k|k-1} \\
&+\hat{x}_{k|k-1}^T C_k^T \Big(C_k V_{k|k-1} C_k^T + (D_k D_k^T)\Big)^{-1} y_k.
\end{aligned}$$

Finally,

$$\begin{aligned}
\Rightarrow \hat{x}_{k|k}^T \frac{(V_{k|k})^{-1}}{2} \hat{x}_{k|k} \;&=\; \frac{1}{2}\hat{x}_{k|k-1}^T V_{k|k-1}^{-1} \hat{x}_{k|k-1} + \frac{1}{2} y_k^T (D_k D_k^T)^{-1} y_k \\
&\quad -\frac{1}{2}(y_k - C_k \hat{x}_{k|k-1})^T (C_k V_{k|k-1} C_k^T + D_k D_k^T)^{-1} (y_k - C_k \hat{x}_{k|k-1}).
\end{aligned} \tag{B.34}$$

Next, the logarithmic term of (B.32) will be reshaped into a better form

$$\begin{aligned}
&\log \frac{|\Sigma_1|^{1/2} |V_{k|k}|^{1/2}}{|D_k||B_k||V_{k-1|k-1}|^{1/2}} \\
&= \log \frac{|V_{k-1|k-1}^{-1} + A_k^T (B_k B_k^T)^{-1} A_k|^{-1/2} |V_{k|k-1}^{-1} + C_k^T (D_k D_k^T)^{-1} C_k|^{-1/2}}{|D_k||B_k||V_{k-1|k-1}|^{1/2}} \\
&= \log \frac{|V_{k-1|k-1}^{-1}(I + V_{k-1|k-1} A_k^T (B_k B_k^T)^{-1} A_k)|^{-1/2} |V_{k|k-1}^{-1} + C_k^T (D_k D_k^T)^{-1} C_k|^{-1/2}}{|D_k||B_k||V_{k-1|k-1}|^{1/2}} \\
&= \log \frac{|I + V_{k-1|k-1} A_k^T (B_k B_k^T)^{-1} A_k|^{-1/2} |V_{k|k-1}^{-1} + C_k^T (D_k D_k^T)^{-1} C_k|^{-1/2}}{|D_k||B_k|} \\
&= \log \frac{|I + (B_k B_k^T)^{-1} A_k V_{k-1|k-1} A_k^T|^{-1/2} |V_{k|k-1}^{-1} + C_k^T (D_k D_k^T)^{-1} C_k|^{-1/2}}{|D_k||B_k|} \\
&= \log \frac{|(B_k B_k^T)^{-1}((B_k B_k^T) + A_k V_{k-1|k-1} A_k^T)|^{-1/2} |V_{k|k-1}^{-1} + C_k^T (D_k D_k^T)^{-1} C_k|^{-1/2}}{|D_k||B_k|} \\
&= \log \frac{|(B_k B_k^T)^{-1} V_{k|k-1}(V_{k|k-1}^{-1} + C_k^T (D_k D_k^T)^{-1} C_k)|^{-1/2}}{|D_k||B_k|}
\end{aligned}$$

$$
\begin{aligned}
&= \log \frac{|(B_k B_k^T)^{-1}(I + V_{k|k-1}C_k^T(D_kD_k^T)^{-1}C_k)|^{-1/2}}{|D_k||B_k|} \\
&= \log \frac{|(B_k B_k^T)|^{1/2}|(I + (D_kD_k^T)^{-1}C_kV_{k|k-1}C_k^T)|^{-1/2}}{|D_k||B_k|} \\
&= \log \frac{|(B_k B_k^T)|^{1/2}|(D_kD_k^T)^{-1}(C_kV_{k|k-1}C_k^T + D_kD_k^T)|^{-1/2}}{|D_k||B_k|} \\
&= \log \frac{|(B_k B_k^T)|^{1/2}||(D_kD_k^T)|^{1/2}|(C_kV_{k|k-1}C_k^T + D_kD_k^T)|^{-1/2}}{|D_k||B_k|} \\
&= \log |(C_kV_{k|k-1}C_k^T + D_kD_k^T)|^{-1/2}
\end{aligned}
$$

$$
\Rightarrow \log \frac{|\Sigma_1|^{1/2}|V_{k|k}|^{1/2}}{|D_k||B_k||V_{k-1|k-1}|^{1/2}} = -\frac{1}{2}\log |C_kV_{k|k-1}C_k^T + D_kD_k^T|.
$$

When all the above, (B.33), (B.34) and (B.35), are put together into (B.32) this transforms to

$$
\begin{aligned}
\beta_{k|k} &= -\frac{1}{2}\hat{x}_{k|k-1}^T V_{k|k-1}^{-1}\hat{x}_{k|k-1} + \frac{1}{2}\hat{x}_{k|k-1}^T V_{k|k-1}^{-1}\hat{x}_{k|k-1} \\
&\quad + y_k^T\frac{(D_kD_k^T)^{-1}}{2}y_k - \frac{1}{2}(y_k - C_k\hat{x}_{k|k-1})^T(C_kV_{k|k-1}C_k^T + D_kD_k^T)^{-1} \\
&\quad \times (y_k - C_k\hat{x}_{k|k-1}) \\
&\quad + y_k^T\Big[\frac{I_d}{2} - \frac{(D_kD_k^T)^{-1}}{2}\Big]y_k - \frac{1}{2}\log |C_kV_{k|k-1}C_k^T + D_kD_k^T| + \beta_{k-1|k-1} \\
&= -\frac{1}{2}(y_k - C_k\hat{x}_{k|k-1})^T(C_kV_{k|k-1}C_k^T + D_kD_k^T)^{-1}(y_k - C_k\hat{x}_{k|k-1}) + \frac{y_k^Ty_k}{2} \\
&\quad -\frac{1}{2}\log |C_kV_{k|k-1}C_k^T + D_kD_k^T| + \beta_{k-1|k-1}. \qquad \text{(B.35)}
\end{aligned}
$$

The above expression, (B.35), can also be written as

$$
\begin{aligned}
\beta_{k|k} &= -\sum_{i=1}^{k}\frac{1}{2}(y_i - C_i\hat{x}_{i|i-1})^T(C_iV_{i|i-1}C_i^T + D_i^TD_i)^{-1}(y_i - C_i\hat{x}_{i|i-1}) + \sum_{i=1}^{k}\frac{y_i^Ty_i}{2} \\
&\quad -\frac{1}{2}\sum_{i=1}^{k}\log |C_iV_{i|i-1}C_i^T + D_i^TD_i|. \qquad \text{(B.36)}
\end{aligned}
$$

APPENDIX C

DERIVATIONS OF EXAMPLES OF CHAPTER 3

This Appendix presents the derivations of the solutions of the Examples of Chapter 3.

Derivations of Examples of Section 3.3.1 - Estimation of Random Variables

Here, the derivations of the solutions of the examples of Section 3.3.1, are presented.

Application 1

The worst case measure, $\nu^*(y, dx)$, is given by (3.25). By substituting Bayes' formula $\eta(y, dx) = \frac{\mu(x,dy)dP_X(x)}{\int_{\mathcal{X}} \mu(x,dy)dP_X(x)}$ into (3.25) the worst case measure becomes

$$\nu^*(y, dx) = \frac{e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}} \mu(x, dy)dP_X(x)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}} \mu(x, dy)dP_X(x)}. \tag{C.1}$$

The computation begins from the numerator of (C.1). By substituting (3.55), (3.56) and (3.57) into (C.1) the numerator is given by

$$\begin{aligned} e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}} \mu(x, dy)dP_X(x) &= (2\pi)^{-\frac{d}{2}}(2\pi)^{-\frac{n}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}} \\ &\quad \times \exp\Big\{ -(y - Hx)^T \frac{\Sigma_W^{-1}}{2}(y - Hx) \end{aligned}$$

$$
\begin{aligned}
& -x^T\frac{\Sigma_X^{-1}}{2}x + (x-\Phi(y))^T\frac{U}{\tilde{s}(y)}(x-\Phi(y))\Big\}dxdy \\
= \; & (2\pi)^{-\frac{n}{2}}\exp\Big\{ -x^T\Big(-\frac{U}{\tilde{s}(y)} + H^T\frac{\Sigma_W^{-1}}{2}H + \frac{\Sigma_X^{-1}}{2}\Big)x \\
& +x^T\Big(-2\frac{U}{\tilde{s}(y)}\Phi(y) + H^T\Sigma_W^{-1}y\Big) \\
& +\Phi(y)^T\frac{U}{\tilde{s}(y)}\Phi(y) - y^T\frac{\Sigma_W^{-1}}{2}y \\
& +\log\Big[(2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\Big]\Big\}dxdy. \qquad \text{(C.2)}
\end{aligned}
$$

The function (C.2), is equivalent to

$$
\begin{aligned}
& (2\pi)^{-\frac{n}{2}}\exp\Big\{ -(x-m^y(\tilde{s}))^T\frac{\Sigma^y(\tilde{s})^{-1}}{2}(x-m^y(\tilde{s})) + g_1(\tilde{s},\Phi,y) \\
& +\log|\Sigma^y(\tilde{s})|^{-\frac{1}{2}}\Big\}dxdy \qquad \text{(C.3)}
\end{aligned}
$$

where

$$
\begin{aligned}
\Sigma^y(\tilde{s}) &= (H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - \frac{2U}{\tilde{s}(y)})^{-1}, \\
m^y(\tilde{s}) &= \Sigma^y(\tilde{s})(H^T\Sigma_W^{-1}y - \frac{2U}{\tilde{s}(y)}\Phi(y)), \\
g_1(\tilde{s},\Phi,y) &= m^y(\tilde{s})^T\frac{\Sigma^y(\tilde{s})^{-1}}{2}m^y(\tilde{s}) + \Phi(y)^T\frac{U}{\tilde{s}(y)}\Phi(y) - y^T\frac{\Sigma_W^{-1}}{2}y \\
& \quad +\log\Big[(2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\Sigma^y(\tilde{s})^{\frac{1}{2}}\Big].
\end{aligned}
$$

The denominator of (C.1) is equivalent to

$$
\begin{aligned}
& \int_{\mathcal{X}}(2\pi)^{-\frac{n}{2}}\exp\Big\{ -(x-m^y(\tilde{s}))^T\frac{\Sigma^y(\tilde{s})^{-1}}{2}(x-m^y(\tilde{s})) + g_1(\tilde{s},\Phi,y) \\
& +\log|\Sigma^y(\tilde{s})|^{-\frac{1}{2}}\Big\}dxdy. \qquad \text{(C.4)}
\end{aligned}
$$

Now, given that

$$
\int_{\mathcal{X}}\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma^y(\tilde{s})|^{\frac{1}{2}}}\; e^{-(x-m^y(\tilde{s}))^T\frac{\Sigma^y(\tilde{s})^{-1}}{2}(x-m^y(\tilde{s}))}dx = 1. \qquad \text{(C.5)}
$$

Then, the worst case measure, $\nu^*(y,dx)$ equals to

$$
\begin{aligned}
\nu^*(y,dx) &= \frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma^y(\tilde{s})|^{\frac{1}{2}}}\; e^{-(x-m^y(\tilde{s}))^T\frac{\Sigma^y(\tilde{s})^{-1}}{2}(x-m^y(\tilde{s}))+g_1(\tilde{s},\Phi,y)}dxdy}{e^{g_1(\tilde{s},\Phi,y)}dy} \qquad \text{(C.6)} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma^y(\tilde{s})|^{\frac{1}{2}}}\; e^{-(x-m^y(\tilde{s}))^T\frac{\Sigma^y(\tilde{s})^{-1}}{2}(x-m^y(\tilde{s}))}dx. \qquad \text{(C.7)}
\end{aligned}
$$

Next, using Bayes' formula the pay-off function (3.27) can be written as

$$L_2(\nu^*,\lambda^*,\tilde{s}^*)=\inf_{\Phi(\cdot)}\inf_{\tilde{s}(\cdot)}\Big\{\tilde{s}(y)\log\frac{\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}}\mu(x,dy)P_X(dx)}{\int_{\mathcal{X}}\mu(x,dy)P_X(dx)}+\tilde{s}(y)R(y)\Big\}. \tag{C.8}$$

The function $\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}}\mu(x,dy)P_X(dx)$ has already been derived and it is equal to $e^{g_1(\tilde{s},\Phi,y)}dy$. Next, the following function is derived

$$\begin{aligned}\int_{\mathcal{X}}\mu(x,dy)P_X(dx) &= \int_{\mathcal{X}}(2\pi)^{-\frac{d}{2}}(2\pi)^{-\frac{n}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\\ &\quad\times\exp\Big\{-(y-Hx)^T\frac{\Sigma_W^{-1}}{2}(y-Hx)-x^T\frac{\Sigma_X^{-1}}{2}x\Big\}dxdy\end{aligned} \tag{C.9}$$

which is equivalent to

$$\begin{aligned}&\int_{\mathcal{X}}(2\pi)^{-\frac{n}{2}}\exp\Big\{-(x-m_2^y(\tilde{s}))^T\frac{\Sigma_2^y(\tilde{s})^{-1}}{2}(x-m_2^y(\tilde{s}))+g_2(\tilde{s},y)\\ &+\log|\Sigma_2^y(\tilde{s})|^{-\frac{1}{2}}\Big\}dxdy=e^{g_2(\tilde{s},y)}dy\end{aligned} \tag{C.10}$$

where

$$\begin{aligned}\Sigma_2^y(\tilde{s}) &= (H^T\Sigma_W^{-1}H+\Sigma_X^{-1})^{-1},\\ m_2^y(\tilde{s}) &= \Sigma_2^y(\tilde{s})(H^T\Sigma_W^{-1}y),\\ g_2(\tilde{s},y) &= m_2^y(\tilde{s})^T\frac{\Sigma_2^y(\tilde{s})^{-1}}{2}m_2^y(\tilde{s})-y^T\frac{\Sigma_W^{-1}}{2}y+\log\Big[(2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\Sigma_2^y(\tilde{s})^{\frac{1}{2}}\Big].\end{aligned}$$

Therefore,

$$\frac{\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi(y))}{\tilde{s}(y)}}\mu(x,dy)P_X(dx)}{\int_{\mathcal{X}}\mu(x,dy)P_X(dx)}=\frac{e^{g_1(\tilde{s},\Phi,y)}dy}{e^{g_2(\tilde{s},y)}dy}=e^{g_1(\tilde{s},\Phi,y)-g_2(\tilde{s},y)}. \tag{C.11}$$

Next, the exponent of the above function is isolated and further manipulated,

$$\begin{aligned}g_1(\tilde{s},\Phi,y)-g_2(\tilde{s},y) &= m^y(\tilde{s})^T\frac{\Sigma^y(\tilde{s})^{-1}}{2}m^y(\tilde{s})+\Phi(y)^T\frac{U}{\tilde{s}(y)}\Phi(y)-y^T\frac{\Sigma_W^{-1}}{2}y\\ &\quad+\log\Big[(2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\Sigma^y(\tilde{s})^{\frac{1}{2}}\Big]+y^T\frac{\Sigma_W^{-1}}{2}y\\ &\quad-m_2^y(\tilde{s})^T\frac{\Sigma_2^y(\tilde{s})^{-1}}{2}m_2^y(\tilde{s})-\log\Big[(2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\Sigma_2^y(\tilde{s})^{\frac{1}{2}}\Big]\\ &= (H^T\Sigma_W^{-1}y-\frac{2U}{\tilde{s}(y)}\Phi(y))^T\Big[(H^T\frac{\Sigma_W^{-1}}{2}H+\frac{\Sigma_X^{-1}}{2}-\frac{U}{\tilde{s}(y)})^{-1}\Big]^T\end{aligned}$$

$$
\begin{aligned}
& -y^T[\Sigma_W^{-1}]^T H\Big[(H^T\frac{\Sigma_W^{-1}}{2}H+\frac{\Sigma_X^{-1}}{2})^{-1}\Big]^T(H^T\Sigma_W^{-1}y)\\
& +\Phi(y)^T\frac{U}{\tilde{s}(y)}\Phi(y)-\frac{1}{2}\log\frac{|H^T\Sigma_W^{-1}H+\Sigma_X^{-1}-\frac{2U}{\tilde{s}(y)}|}{|H^T\Sigma_W^{-1}H+\Sigma_X^{-1}|}\\
= \;& g(\tilde{s},\Phi,y)-\frac{1}{2}\log\frac{|\Sigma-\frac{2U}{\tilde{s}(y)}|}{|\Sigma|}
\end{aligned}
$$

where

$$
\begin{aligned}
g(\tilde{s},\Phi,y) &= (H^T\Sigma_W^{-1}y-\frac{2U}{\tilde{s}(y)}\Phi(y))^T\frac{(\Sigma-\frac{2U}{\tilde{s}(y)})^{-1}}{2}(H^T\Sigma_W^{-1}y-\frac{2U}{\tilde{s}(y)}\Phi(y))\\
&\quad +\Phi(y)^T\frac{U}{\tilde{s}(y)}\Phi(y)-y^T\Sigma_W^{-1}H\frac{\Sigma^{-1}}{2}H^T\Sigma_W^{-1}y,\\
\Sigma &= H^T\Sigma_W^{-1}H+\Sigma_X^{-1}.
\end{aligned}
$$

Next, the derived pay-off function has to be differentiated and the derivative set to zero, as shown below

$$
\begin{aligned}
& \frac{d}{d\Phi}L_2(\nu^*,\lambda^*,\tilde{s})|_{\Phi=\Phi^*}=0,\;\;\forall s(\cdot)\\
& \Rightarrow \frac{d}{d\Phi}\tilde{s}(y)g(\tilde{s},\Phi,y)|_{\Phi=\Phi^*}=0.
\end{aligned}
$$

Performing the above differentiation results to

$$
\begin{aligned}
& -2U\Big(\Sigma-\frac{2U}{\tilde{s}(y)}\Big)^{-1}\Big(H^T\Sigma_W^{-1}y-\frac{2U}{\tilde{s}(y)}\Phi^*(y)\Big)+2U\Phi^*(y)=0\\
& \Rightarrow \Phi^*(y)=\Big(\Sigma-\frac{2U}{\tilde{s}(y)}\Big)^{-1}\Big(H^T\Sigma_W^{-1}y-\frac{2U}{\tilde{s}(y)}\Phi^*(y)\Big)\\
& \Rightarrow \Big(\Sigma-\frac{2U}{\tilde{s}(y)}\Big)\Phi^*(y)=H^T\Sigma_W^{-1}y-\frac{2U}{\tilde{s}(y)}\Phi^*(y)\\
& \Rightarrow \Sigma\Phi^*(y)=H^T\Sigma_W^{-1}y\\
& \Rightarrow \Phi^*(y)=\Sigma^{-1}H^T\Sigma_W^{-1}y=\Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}\Big)^{-1}H^T\Sigma_W^{-1}y. \qquad \text{(C.12)}
\end{aligned}
$$

Finally, applying the PosDef Identity

$$
\Phi^*(y)=\Sigma_X H^T\Big(H\Sigma_X H^T+\Sigma_W\Big)^{-1}y. \qquad \text{(C.13)}
$$

Application 2

Assuming that the nominal distribution $\mu(x, dy)$ is given by (3.55), then the true distribution $\nu(x, dy)$ is given by

$$\nu(x, dy) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_W + \Delta\Sigma_W|^{\frac{1}{2}}}\, e^{-(y-(H+\Delta H)x)^T \frac{(\Sigma_W+\Delta\Sigma_W)^{-1}}{2}(y-(H+\Delta H)x)}dy. \quad \text{(C.14)}$$

The derivation of (3.64) is presented here. The relative entropy between $\mu(x, dy)$ and $\nu(x, dy)$ is given by the following expression:

$$\begin{aligned}
H(\nu|\mu)(x) &= \int_{\mathcal{Y}} \log\Big(\frac{\nu(x,dy)}{\mu(x,dy)}\Big)\nu(x,dy) \\
&= \int_{\mathcal{Y}} \nu(x,dy)\log\nu(x,dy) - \int_{\mathcal{Y}} \nu(x,dy)\log\mu(x,dy) \\
&= \int_{\mathcal{Y}} \nu(x,dy)\Big\{ -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_W + \Delta\Sigma_W| \\
&\quad -(y-(H+\Delta H)x)^T\frac{(\Sigma_W+\Delta\Sigma_W)^{-1}}{2}(y-(H+\Delta H)x)\Big\} \\
&\quad -\int_{\mathcal{Y}} \nu(x,dy)\Big\{ -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_W| - (y-Hx)^T\frac{\Sigma_W^{-1}}{2}(y-Hx)\Big\} \\
&= \frac{1}{2}\log|\Sigma_W| - \frac{1}{2}\log|\Sigma_W + \Delta\Sigma_W| \\
&\quad -\int_{\mathcal{Y}} \nu(x,dy)(y-(H+\Delta H)x)^T\frac{(\Sigma_W+\Delta\Sigma_W)^{-1}}{2}(y-(H+\Delta H)x) \\
&\quad +\int_{\mathcal{Y}} \nu(x,dy)(y-Hx)^T\frac{\Sigma_W^{-1}}{2}(y-Hx). \qquad \text{(C.15)}
\end{aligned}$$

Next, each term of the above function is treated separately, starting from the term $\int_{\mathcal{Y}} \nu(x, dy)(y - Hx)^T\frac{\Sigma_W^{-1}}{2}(y - Hx)$ which can be expanded as followes

$$\begin{aligned}
&\int_{\mathcal{Y}} \nu(x,dy)(y-Hx)^T\frac{\Sigma_W^{-1}}{2}(y-Hx) = \\
&\int_{\mathcal{Y}} \nu(x,dy)\Big((y-(H+\Delta H)x) - (Hx-(H+\Delta H)x)\Big)^T\frac{\Sigma_W^{-1}}{2}\Big((y-(H+\Delta H)x) \\
&-(Hx-(H+\Delta H)x)\Big) \\
&= \int_{\mathcal{Y}} \nu(x,dy)\Big\{(y-(H+\Delta H)x)^T\frac{\Sigma_W^{-1}}{2}(y-(H+\Delta H)x) \\
&+(Hx-(H+\Delta H)x)^T\frac{\Sigma_W^{-1}}{2}(Hx-(H+\Delta H)x) \\
&-2(Hx-(H+\Delta H)x)^T\frac{\Sigma_W^{-1}}{2}(y-(H+\Delta H)x)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathcal{Y}} \nu(x,dy)(y-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(y-(H+\Delta H)x) \\
&+ \int_{\mathcal{Y}} \nu(x,dy)(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(Hx-(H+\Delta H)x) \\
&-2 \int_{\mathcal{Y}} \nu(x,dy)(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2} y \\
&+2 \int_{\mathcal{Y}} \nu(x,dy)(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(H+\Delta H)x. \qquad \text{(C.16)}
\end{aligned}
$$

Note that

1.

$$
\begin{aligned}
&\int_{\mathcal{Y}} \nu(x,dy)(y-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(y-(H+\Delta H)x) \\
&= tr\Big\{(\Sigma_W+\Delta\Sigma_w)\frac{\Sigma_W^{-1}}{2}\Big\}. \qquad \text{(C.17)}
\end{aligned}
$$

2.

$$
\begin{aligned}
&\int_{\mathcal{Y}} \nu(x,dy)(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(Hx-(H+\Delta H)x) \\
&= (Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(Hx-(H+\Delta H)x) \\
&= (\Delta Hx)^T \frac{\Sigma_W^{-1}}{2}(\Delta Hx). \qquad \text{(C.18)}
\end{aligned}
$$

3.

$$
\begin{aligned}
&-2 \int_{\mathcal{Y}} \nu(x,dy)(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2} y \\
&= -2(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2} \int_{\mathcal{Y}} y\nu(x,dy) \\
&= -2(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(Hx-(H+\Delta H)x). \qquad \text{(C.19)}
\end{aligned}
$$

4.

$$
\begin{aligned}
&2 \int_{\mathcal{Y}} \nu(x,dy)(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(H+\Delta H)x \\
&= 2(Hx-(H+\Delta H)x)^T \frac{\Sigma_W^{-1}}{2}(H+\Delta H)x. \qquad \text{(C.20)}
\end{aligned}
$$

Therefore, by putting all the above together, (C.16) equals to

$$tr\Big\{(\Sigma_W+\Delta\Sigma_w)\frac{\Sigma_W^{-1}}{2}\Big\}+(\Delta Hx)^T\frac{\Sigma_W^{-1}}{2}(\Delta Hx). \tag{C.21}$$

Similarly,

$$\begin{aligned}&-\int_{\mathcal{Y}}\nu(x,dy)(y-(H+\Delta H)x)^T\frac{(\Sigma_W+\Delta\Sigma_W)^{-1}}{2}(y-(H+\Delta H)x)\\&=-tr\Big\{(\Sigma_W+\Delta\Sigma_w)\frac{(\Sigma_W+\Delta\Sigma_W)^{-1}}{2}\Big\}.\end{aligned} \tag{C.22}$$

Finally, putting all the above together results to (3.64),

$$\begin{aligned}H(\nu|\mu)(x) &= \int\log\Big(\frac{\nu(x,dy)}{\mu(x,dy)}\Big)\nu(x,dy)\\&=\frac{1}{2}\Big\{\log\frac{|\Sigma_W|}{|\Sigma_W+\Delta\Sigma_W|}+tr((\Sigma_W+\Delta\Sigma_W)(\Sigma_W^{-1}-(\Sigma_W+\Delta\Sigma_W)^{-1}))\\&\quad+x^T(\Delta H)^T\Sigma_W^{-1}(\Delta H)x\Big\}.\end{aligned}$$

Application 3

The worst case measure, $dQ^*_{X,Y}(x,y)$, is given by (3.43). By substituting Bayes' formula $dP_{X,Y}(y,x)=\mu(x,dy)dP_X(x)$ into (3.43) the worst case measure becomes

$$dQ^*_{X,Y}(x,y)=\frac{e^{\frac{\ell(x,\Phi(y))}{s}}\mu(x,dy)P_X(dx)}{\int_{\mathcal{X}\times\mathcal{Y}}e^{\frac{\ell(x,\Phi(y))}{s}}\mu(x,dy)P_X(dx)}. \tag{C.23}$$

Then the procedure is the same as the one used in Application 1 above, for deriving (3.58), the only difference is that the integration is performed with respect to $\mathcal{X}\times\mathcal{Y}$ and that s is not a function of y. So, the denominator of (C.23) is given by C.2 ($\tilde{s}(y)$ is replaced by s) and is equivalent to

$$\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma(s)|^{\frac{1}{2}}}\;e^{-(x-\kappa^y(s))^T\frac{\Sigma(s)^{-1}}{2}(x-\kappa^y(s))+\theta_1(s,\Phi,y)}dydx \tag{C.24}$$

where

$$\begin{aligned}\Sigma(s) &= (H^T\Sigma_W^{-1}H+\Sigma_X^{-1}-2\frac{U}{s})^{-1},\\\kappa^y(s) &= \Sigma(s)(H^T\Sigma_W^{-1}y-\frac{2U}{s}\Phi(y)),\\\theta_1(s,\Phi,y) &= \kappa^y(s)^T\frac{\Sigma(s)^{-1}}{2}\kappa^y(s)+\Phi(y)^T\frac{U}{s}\Phi(y)-y^T\frac{\Sigma_W^{-1}}{2}y\\&\quad+\log\Big[(2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\Big].\end{aligned}$$

The denominator of (C.23) is given by

$$\int_{\mathcal{X}\times\mathcal{Y}} \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma(s)|^{\frac{1}{2}}}\, e^{-(x-\kappa^y(s))^T\frac{\Sigma(s)^{-1}}{2}(x-\kappa^y(s))+\theta_1(s,\Phi,y)}dydx = \int_{\mathcal{Y}} e^{\theta_1(s,\Phi,y)}dy. \quad \text{(C.25)}$$

Therefore, the worst case measure, $dQ^*_{X,Y}(x,y)$, is given by

$$\begin{aligned} dQ^*_{X,Y}(x,y) &= \frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma(s)|^{\frac{1}{2}}}\, e^{-(x-\kappa^y(s))^T\frac{\Sigma(s)^{-1}}{2}(x-\kappa^y(s))+\theta_1(s,\Phi,y)}dydx}{\int_{\mathcal{Y}} e^{\theta_1(s,\Phi,y)}dy} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma(s)|^{\frac{1}{2}}}\, e^{-(x-\kappa^y(s))^T\frac{\Sigma(s)^{-1}}{2}(x-\kappa^y(s))}\frac{e^{\theta(s,\Phi,y)}}{\int_{\mathcal{Y}} e^{\theta(s,\Phi,y)}dy}dydx \quad \text{(C.26)} \end{aligned}$$

where

$$\theta(s,\Phi,y) = \kappa^y(s)^T\frac{\Sigma(s)^{-1}}{2}\kappa^y(s) + \Phi(y)^T\frac{U}{s}\Phi(y) - y^T\frac{\Sigma_W^{-1}}{2}y.$$

Next, the average pay-off is given by (3.42). Using another form of Bayes' formula, $dP_{X,Y}(y,x) = \eta(y,dx)dP_Y(y)$, results to,

$$L_4(Q^*_{X,Y},\lambda^*,s) = s\log\Big(\int_{\mathcal{Y}}\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{s}}\eta(y,dx)dP_Y(y)\Big) + sR. \quad \text{(C.27)}$$

The average pay-off (C.27) has to be differentiated and the derivative set to zero. The inner integral will be differentiated as follows

$$\frac{d}{d\Phi}\Big[\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{s}}\eta(y,dx)\Big]\Big|_{\Phi=\Phi^*} = 0. \quad \text{(C.28)}$$

First, the differentiation is performed and then the integration, therefore,

$$\int_{\mathcal{X}} \frac{d}{d\Phi}\Big[e^{\frac{\ell(x,\Phi(y))}{s}}\Big]\Big|_{\Phi=\Phi^*}\eta(y,dx) = 0. \quad \text{(C.29)}$$

Given (3.57), the above function results to

$$\begin{aligned} &\int_{\mathcal{X}}\Big(\frac{2U}{s}\Phi^*(y) - \frac{2U}{s}x\Big)e^{\frac{\ell(x,\Phi^*(y))}{s}}\eta(y,dx) = 0 \\ &\Rightarrow \Phi^*(y) = \frac{\int_{\mathcal{X}} xe^{\frac{\ell(x,\Phi^*(y))}{s}}\eta(y,dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s}}\eta(y,dx)} = \frac{\int_{\mathcal{X}} xe^{\frac{\ell(x,\Phi^*(y))}{s}}\mu(x,dy)dP_X(x)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s}}\mu(x,dy)dP_X(x)}. \quad \text{(C.30)} \end{aligned}$$

The function $e^{\frac{\ell(x,\Phi^*(y))}{s}}\mu(x,dy)dP_X(x)$ has already derived and is equivalent to (C.24) (note that $\Phi(y) = \Phi^*(y)$), therefore

$$\begin{aligned} \Phi^*(y) &= \frac{m^{*,y}(s)\; e^{\theta_1(s,\Phi^*,y)}dy}{e^{\theta_1(s,\Phi^*,y)}dy} \\ &= \Big(H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - \frac{2U}{s}\Big)^{-1}\Big(H^T\Sigma_W^{-1}y - \frac{2U}{s}\Phi^*(y)\Big). \quad \text{(C.31)} \end{aligned}$$

The above function can be further manipulated, before it reaches its final form. That is,

$$\begin{aligned}
&\Rightarrow \Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}-\frac{2U}{s}\Big)\Phi^*(y)=H^T\Sigma_W^{-1}y-\frac{2U}{s}\Phi^*(y)\\
&\Rightarrow \Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}\Big)\Phi^*(y)=H^T\Sigma_W^{-1}y\\
&\Rightarrow \Phi^*(y)=\Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}\Big)^{-1}H^T\Sigma_W^{-1}y. \qquad \text{(C.32)}
\end{aligned}$$

Finally, applying the PosDef Identity

$$\Phi^*(y)=\Sigma_X H^T\Big(H\Sigma_X H^T+\Sigma_W\Big)^{-1}y. \qquad \text{(C.33)}$$

Application 4

The worst case measure, $\nu^*(y,dx)$, is given by (3.37). By substituting Bayes' formula $\eta(y,dx)=\frac{\mu(x,dy)dP_X(x)}{\int_{\mathcal{X}}\mu(x,dy)dP_X(x)}$ into (3.37) the worst case measure becomes

$$\nu^*(y,dx)=\frac{e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x)}{\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x)}. \qquad \text{(C.34)}$$

Then the procedure is the same as the one in Application 1 above. Starting with the numerator, which is given by

$$\begin{aligned}
e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x) &= (2\pi)^{-\frac{d}{2}}(2\pi)^{-\frac{n}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\\
&\quad\times\exp\Big\{-(y-Hx)^T\frac{\Sigma_W^{-1}}{2}(y-Hx)-x^T\frac{\Sigma_X^{-1}}{2}x\\
&\quad+(x-\Phi(y))^T\frac{U}{s}(x-\Phi(y))\\
&\quad-(y-\bar{H}x)^T\tilde{U}(y-\bar{H}x)\Big\}dxdy\\
&= (2\pi)^{-\frac{n}{2}}\exp\Big\{-x^T\Big(\bar{H}^T\tilde{U}\bar{H}-\frac{U}{s}+H^T\frac{\Sigma_W^{-1}}{2}H\\
&\quad+\frac{\Sigma_X^{-1}}{2}\Big)x+x^T\Big(-2\frac{U}{s}\Phi(y)+2\bar{H}^T\tilde{U}y+H^T\Sigma_W^{-1}y\Big)\\
&\quad+\Phi(y)^T\frac{U}{s}\Phi(y)-y^T\tilde{U}y-y^T\frac{\Sigma_W^{-1}}{2}y\\
&\quad+\log\Big[(2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}|\Sigma_X|^{-\frac{1}{2}}\Big]\Big\}dxdy. \qquad \text{(C.35)}
\end{aligned}$$

The above function (C.35) is equivalent with the following function

$$(2\pi)^{-\frac{n}{2}}\exp\Big\{-(x-\tilde{m}^y(s))^T\frac{\tilde{\Sigma}(s)^{-1}}{2}(x-\tilde{m}^y(s))+g_1(s,\Phi,y)$$

$$+\log|\tilde{\Sigma}(s)|^{-\frac{1}{2}}\Big\}dxdy \tag{C.36}$$

where

$$\begin{aligned}\tilde{\Sigma}(s) &= \Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}+2\bar{H}^T\tilde{U}\bar{H}-\frac{2U}{s}\Big)^{-1},\\ \tilde{m}^y(s) &= \tilde{\Sigma}(s)\Big(H^T\Sigma_W^{-1}y+2\bar{H}^T\tilde{U}y-\frac{2U}{s}\Phi(y)\Big).\end{aligned}$$

The denominator of (C.34) is equivalent to

$$\begin{aligned}&\int_{\mathcal{X}}(2\pi)^{-\frac{n}{2}}\exp\Big\{-(x-\tilde{m}^y(s))^T\frac{\tilde{\Sigma}(s)^{-1}}{2}(x-\tilde{m}^y(s))+g_1(s,\Phi,y)\\ &+\log|\tilde{\Sigma}(s)|^{-\frac{1}{2}}\Big\}dxdy=e^{g_1(s,\Phi,y)}dy.\end{aligned} \tag{C.37}$$

Therefore, the worst case measure, (C.34), is finally given by

$$\begin{aligned}\nu^*(y,dx) &= \frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\tilde{\Sigma}(s)|^{\frac{1}{2}}}\,e^{-(x-\tilde{m}^y(s))^T\frac{\tilde{\Sigma}(s)^{-1}}{2}(x-\tilde{m}^y(s))+g_1(s,\Phi,y)}dxdy}{e^{g_1(s,\Phi,y)}dy}\\ &= \frac{1}{(2\pi)^{\frac{n}{2}}|\tilde{\Sigma}(s)|^{\frac{1}{2}}}\,e^{-(x-\tilde{m}^y(s))^T\frac{\tilde{\Sigma}(s)^{-1}}{2}(x-\tilde{m}^y(s))}dx.\end{aligned} \tag{C.38}$$

Next, the average pay-off, $L_3^R(\nu^*,\lambda^*,s^*)$ is given by (3.39)

$$L_3^R(\nu^*,\lambda^*,s^*)=\inf_{s\geq 0}\int_{\mathcal{Y}}s\log\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\eta(y,dx)dP_Y(y)+s\overline{R}.$$

The average pay-off (3.39) has to be differentiated and the derivative set to zero. Similar to Application 3 above, the inner integral will be differentiated as follows

$$\frac{d}{d\Phi}\Big[\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\eta(y,dx)\Big]\Big|_{\Phi=\Phi^*}=0. \tag{C.39}$$

First, the differentiation is performed and then the integration, therefore,

$$\int_{\mathcal{X}}\frac{d}{d\Phi}\Big[e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\Big]\Big|_{\Phi=\Phi^*}\eta(y,dx)=0. \tag{C.40}$$

Given (3.57), the above function results in

$$\int_{\mathcal{X}}\Big(\frac{2U}{s}\Phi^*(y)-\frac{2U}{s}x\Big)e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\eta(y,dx)=0$$

$$\begin{aligned}\Rightarrow \Phi^*(y) &= \frac{\int_{\mathcal{X}} x e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\eta(y,dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\eta(y,dx)} \\ &= \frac{\int_{\mathcal{X}} x e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x)}. \end{aligned} \tag{C.41}$$

The function $e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x)$ is derived and is equivalent to (C.36) (note that $\Phi(y)=\Phi^*(y)$). Therefore

$$\begin{aligned}\Phi^*(y) &= \frac{\tilde{m}^{*,y}(s)\; e^{g_1(s,\Phi^*,y)}dy}{e^{g_1(s,\Phi^*,y)}dy} \\ &= \Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}+2\bar{H}^T\tilde{U}\bar{H}-\frac{2U}{s}\Big)^{-1}\Big(H^T\Sigma_W^{-1}y+2\bar{H}^T\tilde{U}y-\frac{2U}{s}\Phi^*(y)\Big).\end{aligned} \tag{C.42}$$

The above function can be further manipulated, before it reaches its final form. That is,

$$\begin{aligned}&\Rightarrow \Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}+2\bar{H}^T\tilde{U}\bar{H}-\frac{2U}{s}\Big)\Phi^*(y)=H^T\Sigma_W^{-1}y+2\bar{H}^T\tilde{U}y-\frac{2U}{s}\Phi^*(y) \\ &\Rightarrow \Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}+2\bar{H}^T\tilde{U}\bar{H}\Big)\Phi^*(y)=H^T\Sigma_W^{-1}y+2\bar{H}^T\tilde{U}y \\ &\Rightarrow \Phi^*(y)=\Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}+2\bar{H}^T\tilde{U}\bar{H}\Big)^{-1}\Big(H^T\Sigma_W^{-1}+2\bar{H}^T\tilde{U}\Big)y.\end{aligned} \tag{C.43}$$

Derivation of Example of Section 3.3.2 - Estimation of a Sequence of Random Variables

Just like in Application 1 of Section 3.3.1, the worst case measure, $\nu^*(y,dx)$, is given by (C.1). Similarly, the computation begins from the numerator of (C.1). By substituting (3.78), (3.79) and (3.80) into (C.1) the numerator is given by

$$\begin{aligned}e^{\frac{\ell(x,\Phi(y^m))}{\tilde{s}(y)}}\mu(x,dy^m)dP_X(x) &= \Big((2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}\Big)^{m+1}(2\pi)^{-\frac{n}{2}}|\Sigma_X|^{-\frac{1}{2}} \\ &\quad\times\exp\Big\{-\sum_{i=0}^{m}(y_i-Hx)^T\frac{\Sigma_W^{-1}}{2}(y_i-Hx)-x^T\frac{\Sigma_X^{-1}}{2}x \\ &\quad+\sum_{i=0}^{m-1}(x-\Phi_i^*(y^i))^T\frac{U}{\tilde{s}(y)}(x-\Phi_i^*(y^i))\end{aligned}$$

$$
\begin{aligned}
&+(x-\Phi_m(y^m))^T\frac{U}{\tilde{s}(y)}(x-\Phi_m(y^m))\Big\}dxdy^m\\
= \;& (2\pi)^{-\frac{n}{2}}\exp\Big\{-x^T\Big((m+1)H^T\frac{\Sigma_W^{-1}}{2}H+\frac{\Sigma_X^{-1}}{2}\\
&-(m+1)\frac{U}{\tilde{s}(y)}\Big)x+x^T\Big(-2\frac{U}{\tilde{s}(y)}\sum_{i=0}^{(m-1)}\Phi_i^*(y^i)\\
&-\frac{2U\Phi_m(y^m)}{\tilde{s}(y)}+H^T\Sigma_W^{-1}\sum_{i=0}^{m}y_i\Big)-\sum_{i=1}^{N}y_i^T\frac{\Sigma_W^{-1}}{2}y_i\\
&+\sum_{i=0}^{(m-1)}\Phi_i^*(y^i)^T\frac{U}{\tilde{s}(y)}\Phi_i^*(y^i)+\Phi_m(y^m)^T\frac{U}{\tilde{s}(y)}\Phi_m(y^m)\\
&+\log\Big[(2\pi)^{-\frac{d(m+1)}{2}}|\Sigma_W|^{-\frac{m+1}{2}}|\Sigma_X|^{-\frac{1}{2}}\Big]\Big\}dxdy^m. \qquad \text{(C.44)}
\end{aligned}
$$

The above function (C.44) is equivalent with the following function

$$
\begin{aligned}
&(2\pi)^{-\frac{n}{2}}\exp\Big\{-(x-\widehat{m}^y(\tilde{s}))^T\frac{\widehat{\Sigma}^y(\tilde{s})^{-1}}{2}(x-\widehat{m}^y(\tilde{s}))+\tilde{g}_1(\tilde{s},\Phi,y)\\
&+\log|\widehat{\Sigma}^y(\tilde{s})|^{-\frac{1}{2}}\Big\}dxdy \qquad \text{(C.45)}
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\Sigma}^y(\tilde{s}) &= \left(NH^T\Sigma_W^{-1}H+\Sigma_X^{-1}-\frac{2(m+1)U}{\tilde{s}(y)}\right)^{-1},\\
\widehat{m}^y(\tilde{s}) &= \widehat{\Sigma}^y(\tilde{s})\left(H^T\Sigma_W^{-1}\sum_{i=0}^{m}y_i-\frac{2U}{\tilde{s}(y)}\sum_{i=0}^{m-1}\Phi_i^*(y^i)-\frac{2U\Phi_m(y^m)}{\tilde{s}(y)}\right).
\end{aligned}
$$

The denominator of (C.1) is equivalent to

$$
\begin{aligned}
&\int_{\mathcal{X}}(2\pi)^{-\frac{n}{2}}\exp\Big\{-(x-\widehat{m}^y(\tilde{s}))^T\frac{\widehat{\Sigma}^y(\tilde{s})^{-1}}{2}(x-\widehat{m}^y(\tilde{s}))+\tilde{g}_1(\tilde{s},\Phi,y)\\
&+\log|\widehat{\Sigma}^y(\tilde{s})|^{-\frac{1}{2}}\Big\}dxdy^m=e^{\tilde{g}_1(\tilde{s},\Phi,y)}dy^m. \qquad \text{(C.46)}
\end{aligned}
$$

Therefore, the worst case measure, (C.1), is finally given by

$$
\begin{aligned}
\nu^*(y,dx) &= \frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\widehat{\Sigma}^y(\tilde{s})|^{-\frac{1}{2}}}\;e^{-(x-\widehat{m}^y(\tilde{s}))^T\frac{\widehat{\Sigma}^y(\tilde{s})^{-1}}{2}(x-\widehat{m}^y(\tilde{s}))+\tilde{g}_1(\tilde{s},\Phi,y)}dxdy}{e^{\tilde{g}_1(\tilde{s},\Phi,y)}dy}\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\widehat{\Sigma}^y(\tilde{s})|^{-\frac{1}{2}}}\;e^{-(x-\widehat{m}^y(\tilde{s}))^T\frac{\widehat{\Sigma}^y(\tilde{s})^{-1}}{2}(x-\widehat{m}^y(\tilde{s}))}. \qquad \text{(C.47)}
\end{aligned}
$$

Next, the average pay-off function $L_2(\nu^*,\lambda^*,s)$ given by (3.27) has to be differentiated and the derivative set to zero. The inner integral will be differentiated as follows

$$
\frac{d}{d\Phi_m(y^m)}\Big[\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi(y^m))}{\tilde{s}(y)}}\eta(y^m,dx)\Big]\Big|_{\Phi_m(y^m)=\Phi_m^*(y^m)}=0. \qquad \text{(C.48)}
$$

First, the differentiation is performed and then the integration, therefore,

$$\int_{\mathcal{X}} \frac{d}{d\Phi_m(y^m)} \Big[e^{\frac{\ell(x,\Phi(y^m))}{\tilde{s}(y)}} \Big] \Big|_{\Phi_m(y^m)=\Phi_m^*(y^m)} \eta(y^m, dx) = 0. \tag{C.49}$$

Given (3.80), the above function results to

$$\begin{aligned} &\int_{\mathcal{X}} \Big(\frac{2U}{\tilde{s}(y)} \Phi_m^*(y^m) - \frac{2U}{\tilde{s}(y)} x \Big) e^{\frac{\ell(x,\Phi^*(y^m))}{\tilde{s}(y)}} \eta(y^m, dx) = 0 \\ &\Rightarrow \Phi_m^*(y^m) = \frac{\int_{\mathcal{X}} x e^{\frac{\ell(x,\Phi^*(y^m))}{\tilde{s}(y)}} \eta(y^m, dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y^m))}{\tilde{s}(y)}} \eta(y^m, dx)} = \frac{\int_{\mathcal{X}} x e^{\frac{\ell(x,\Phi^*(y^m))}{\tilde{s}(y)}} \mu(x, dy^m) dP_X(x)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y^m))}{\tilde{s}(y)}} \mu(x, dy^m) dP_X(x)}. \end{aligned} \tag{C.50}$$

The function $e^{\frac{\ell(x,\Phi^*(y^m))}{\tilde{s}(y)}} \mu(x, dy^m) dP_X(x)$ has already derived and is equivalent to (C.45) (note that $\Phi_m(y^m) = \Phi_m^*(y^m)$), therefore

$$\begin{aligned} \Phi_m^*(y^m) &= \frac{\widehat{m}^{*,y}(\tilde{s})\ e^{\tilde{g}_1(s,\Phi^*,y)} dy^m}{e^{\tilde{g}_1(s,\Phi^*,y)} dy^m} \\ &= \left((m+1) H^T \Sigma_W^{-1} H + \Sigma_X^{-1} - \frac{2(m+1)U}{\tilde{s}(y)} \right)^{-1} \\ &\quad \times \left(H^T \Sigma_W^{-1} \sum_{i=0}^{m} y_i - \frac{2U}{\tilde{s}(y)} \sum_{i=0}^{m-1} \Phi_i^*(y^i) - \frac{2U}{\tilde{s}(y)} \Phi_m^*(y^m) \right). \end{aligned} \tag{C.51}$$

The above function can be further manipulated, before it reaches its final form. That is,

$$\begin{aligned} \Rightarrow \quad & \left((m+1) H^T \Sigma_W^{-1} H + \Sigma_X^{-1} - \frac{2(m+1)U}{\tilde{s}(y)} \right) \Phi_m^*(y^m) \\ & = H^T \Sigma_W^{-1} \sum_{i=0}^{m} y_i - \frac{2U}{\tilde{s}(y)} \sum_{i=0}^{m-1} \Phi_i^*(y^i) - \frac{2U}{\tilde{s}(y)} \Phi_m^*(y^m) \end{aligned}$$

$$\begin{aligned} \Rightarrow \Phi_m^*(y^m) &= \left((m+1) H^T \Sigma_W^{-1} H + \Sigma_X^{-1} - \frac{2mU}{\tilde{s}(y)} \right)^{-1} \left(H^T \Sigma_W^{-1} \sum_{i=0}^{m} y_i \right. \\ &\quad \left. - \frac{2U}{\tilde{s}(y)} \sum_{i=0}^{m-1} \Phi_i^*(y^i) \right). \end{aligned} \tag{C.52}$$

Derivations of Examples of Section 3.4.2 - Estimation from MIMO communication Systems

Here, the derivation of the results of the examples of Section 3.4.2, are presented.

Application 1

The procedure followed here is very similar to the one applied in Application 4 of Section 3.3.1. The worst case measure, $\nu^*(y,dx)$, is given by (C.34). By substituting (3.87), (3.88), (3.89) and (3.90) into (C.1) the numerator is given by

$$\begin{aligned} e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x) &= (2\pi)^{-d}(2\pi)^{-n}|\Sigma_W|^{-1}|\Sigma_X|^{-1} \\ &\quad \times \exp\Big\{ -(y-Hx)^\dagger\Sigma_W^{-1}(y-Hx) - x^\dagger\Sigma_X^{-1}x \\ &\quad +(x-\Phi(y))^\dagger\frac{U}{s}(x-\Phi(y)) \\ &\quad -(y-\bar{H}x)^\dagger\tilde{U}(y-\bar{H}x)\Big\}dxdy \\ &= (2\pi)^{-n}\exp\Big\{ -x^\dagger\Big(\bar{H}^\dagger\tilde{U}\bar{H} - \frac{U}{s} + H^T\Sigma_W^{-1}H \\ &\quad +\Sigma_X^{-1}\Big)x + x^\dagger\Big(-2\bar{H}^\dagger\frac{U}{s}\Phi(y) + 2\tilde{U}y + 2H^T\Sigma_W^{-1}y\Big) \\ &\quad +\Phi(y)^\dagger\frac{U}{s}\Phi(y) - y^\dagger\tilde{U}y - y^\dagger\Sigma_W^{-1}y \\ &\quad +\log\Big[(2\pi)^{-d}|\Sigma_W|^{-1}|\Sigma_X|^{-1}\Big]\Big\}dxdy. \end{aligned} \tag{C.53}$$

The above function (C.53) is equivalent to the following function

$$\begin{aligned} &(2\pi)^{-n}\exp\Big\{ -(x-\tilde{m}_M(s))^\dagger\tilde{\Sigma}_M(s)^{-1}(x-\tilde{m}_M(s)) + q(s,\Phi,y) \\ &+\log|\tilde{\Sigma}_M(s)|^{-1}\Big\}dxdy \end{aligned} \tag{C.54}$$

where

$$\begin{aligned} \tilde{\Sigma}_M(s) &= \Big(H^\dagger\Sigma_W^{-1}H + \Sigma_X^{-1} + \bar{H}^\dagger\tilde{U}\bar{H} - \frac{U}{s}\Big)^{-1}, \\ \tilde{m}_M(s) &= \tilde{\Sigma}(s)\Big(H^\dagger\Sigma_W^{-1}y + \bar{H}^\dagger\tilde{U}y - \frac{U}{s}\Phi(y)\Big). \end{aligned}$$

The denominator of (C.34) is equivalent to

$$\int_{\mathcal{X}}(2\pi)^{-n}\exp\Big\{ -(x-\tilde{m}_M(s))^\dagger\tilde{\Sigma}_M(s)^{-1}(x-\tilde{m}_M(s)) + q(s,\Phi,y)$$

$$+\log|\tilde{\Sigma}_M(s)|^{-1}\Big\}dxdy = e^{q_1(s,\Phi,y)}dy. \tag{C.55}$$

Therefore, the worst case measure, (C.34), is finally given by

$$\begin{aligned}\nu^*(y,dx) &= \frac{(2\pi)^{-n}|\tilde{\Sigma}_M(s)|^{-1}\ e^{-(x-\tilde{m}_M(s))^\dagger\tilde{\Sigma}_M(s)^{-1}(x-\tilde{m}_M(s))+q(s,\Phi,y)}dxdy}{e^{q(s,\Phi,y)}dy}\\ &= (2\pi)^{-n}|\tilde{\Sigma}_M(s)|^{-1}\ e^{-(x-\tilde{m}_M(s))^\dagger\tilde{\Sigma}_M(s)^{-1}(x-\tilde{m}_M(s))}dx.\end{aligned} \tag{C.56}$$

Next, the average pay-off, $L_3^R(\nu^*,\lambda^*,s^*)$ is given by (3.39)

$$L_3^R(\nu^*,\lambda^*,s^*) = \inf_{s\geq 0}\int_{\mathcal{Y}} s\log\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\eta(y,dx)dP_Y(y) + s\overline{R}.$$

The average pay-off (3.39) has to be differentiated and the derivative set to zero. Similar to Application 4 of Section 3.3.1, the inner integral will be differentiated as follows

$$\frac{d}{d\Phi}\Big[\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\eta(y,dx)\Big]\Big|_{\Phi=\Phi^*} = 0. \tag{C.57}$$

First, the differentiation is performed and then the integration, therefore,

$$\int_{\mathcal{X}}\frac{d}{d\Phi}\Big[e^{\frac{\ell(x,\Phi(y))}{s}-R(x,y)}\Big]\Big|_{\Phi=\Phi^*}\eta(y,dx) = 0. \tag{C.58}$$

Given (3.89), the above function is

$$\int_{\mathcal{X}}\Big(\frac{2U}{s}\Phi^*(y)-\frac{2U}{s}x\Big)e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\eta(y,dx) = 0$$

$$\begin{aligned}\Rightarrow \Phi^*(y) &= \frac{\int_{\mathcal{X}} xe^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\eta(y,dx)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\eta(y,dx)}\\ &= \frac{\int_{\mathcal{X}} xe^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x)}{\int_{\mathcal{X}} e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x)}.\end{aligned} \tag{C.59}$$

The function $e^{\frac{\ell(x,\Phi^*(y))}{s}-R(x,y)}\mu(x,dy)dP_X(x)$ has already derived and is equivalent to (C.54) (note that $\Phi(y)=\Phi^*(y)$), therefore

$$\begin{aligned}\Phi^*(y) &= \frac{\tilde{m}_M^*(s)\ e^{q(s,\Phi^*,y)}dy}{e^{q(s,\Phi^*,y)}dy}\\ &= \Big(H^\dagger\Sigma_W^{-1}H+\Sigma_X^{-1}+\bar{H}^\dagger\tilde{U}\bar{H}-\frac{U}{s}\Big)^{-1}\Big(H^\dagger\Sigma_W^{-1}y+\bar{H}^\dagger\tilde{U}y-\frac{U}{s}\Phi^*(y)\Big).\end{aligned} \tag{C.60}$$

The above function can be further manipulated, before it reaches its final form. That is,

$$\begin{aligned}
&\Rightarrow \Big(H^\dagger\Sigma_W^{-1}H+\Sigma_X^{-1}+\bar{H}^\dagger\tilde{U}\bar{H}-\frac{U}{s}\Big)\Phi^*(y)=H^\dagger\Sigma_W^{-1}y+\bar{H}^\dagger\tilde{U}y-\frac{U}{s}\Phi^*(y)\\
&\Rightarrow \Big(H^\dagger\Sigma_W^{-1}H+\Sigma_X^{-1}+\bar{H}^\dagger\tilde{U}\bar{H}\Big)\Phi^*(y)=H^\dagger\Sigma_W^{-1}y+\bar{H}^\dagger\tilde{U}y\\
&\Rightarrow \Phi^*(y)=\Big(H^\dagger\Sigma_W^{-1}H+\Sigma_X^{-1}+\bar{H}^\dagger\tilde{U}\bar{H}\Big)^{-1}\Big(H^\dagger\Sigma_W^{-1}+\bar{H}^\dagger\tilde{U}\Big)y. \qquad \text{(C.61)}
\end{aligned}$$

Application 2

The procedure followed here is very similar to the one applied in Application 3 of Section 3.3.1. The worst case measure, $dQ^*_{X,Y}(x,y)$, is given by (C.23). By substituting (3.87), (3.88), (3.89) and (3.90) into (C.23) the numerator is given by

$$\begin{aligned}
e^{\frac{\ell(x,\Phi(y))}{s}}\mu(x,dy)dP_X(x) &= (2\pi)^{-d}(2\pi)^{-n}|\Sigma_W|^{-1}|\Sigma_X|^{-1}\\
&\quad\times\exp\Big\{-(y-Hx)^\dagger\Sigma_W^{-1}(y-Hx)-x^\dagger\Sigma_X^{-1}x\\
&\quad+(x-\Phi(y))^\dagger\frac{U}{s}(x-\Phi(y))\Big\}dxdy\\
&= (2\pi)^{-n}\exp\Big\{-x^\dagger\Big(H^T\Sigma_W^{-1}H+\Sigma_X^{-1}-\frac{U}{s}\Big)x\\
&\quad+x^\dagger\Big(-2\frac{U}{s}\Phi(y)+2H^T\Sigma_W^{-1}y\Big)\\
&\quad+\Phi(y)^\dagger\frac{U}{s}\Phi(y)-y^\dagger\Sigma_W^{-1}y\\
&\quad+\log\Big[(2\pi)^{-d}|\Sigma_W|^{-1}|\Sigma_X|^{-1}\Big]\Big\}dxdy. \qquad \text{(C.62)}
\end{aligned}$$

The above function (C.62) is equivalent with the following function

$$\begin{aligned}
&(2\pi)^{-n}\exp\Big\{-(x-\kappa_M^y(s))^\dagger\Sigma_M(s)^{-1}(x-\kappa_M^y(s))+\tilde{\theta}_1(s,\Phi,y)\\
&+\log|\Sigma_M(s)|^{-1}\Big\}dxdy \qquad \text{(C.63)}
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_M(s) &= \Big(H^\dagger\Sigma_W^{-1}H+\Sigma_X^{-1}-\frac{U}{s}\Big)^{-1},\\
\kappa_M^y(s) &= \Sigma_M(s)\Big(H^\dagger\Sigma_W^{-1}y-\frac{U}{s}\Phi(y)\Big),\\
\tilde{\theta}_1(s,\Phi,y) &= \kappa_M^y(s)^\dagger\Sigma_M(s)^{-1}\kappa_M^y(s)+\Phi(y)^\dagger\frac{U}{s}\Phi(y)-y^\dagger\Sigma_W^{-1}y\\
&\quad+\log\Big[(2\pi)^{-d}|\Sigma_W|^{-1}|\Sigma_X|^{-1}\Big].
\end{aligned}$$

The denominator of (C.23) is equivalent to

$$\begin{aligned}&\int_{\mathcal{X}\times\mathcal{Y}}(2\pi)^{-n}|\Sigma_M(s)|^{-1}\ e^{-(x-\kappa_M^y(s))^\dagger\Sigma_M(s)^{-1}(x-\kappa_M^y(s))+\tilde{\theta}_1(s,\Phi,y)}dxdy\\&=\int_{\mathcal{Y}}e^{\tilde{\theta}_1(s,\Phi,y)}dy.\end{aligned}\tag{C.64}$$

Therefore, the worst case measure, $dQ^*_{X,Y}(x,y)$, is given by

$$\begin{aligned}dQ^*_{X,Y}(x,y) &= \frac{(2\pi)^{-n}|\Sigma_M(s)|^{-1}\ e^{-(x-\kappa_M^y(s))^\dagger\Sigma_M(s)^{-1}(x-\kappa_M^y(s))+\tilde{\theta}_1(s,\Phi,y)}dxdy}{\int_{\mathcal{Y}}e^{\tilde{\theta}_1(s,\Phi,y)}dy}\\&= (2\pi)^{-n}|\Sigma_M(s)|^{-1}\ e^{-(x-\kappa_M^y(s))^\dagger\Sigma_M(s)^{-1}(x-\kappa_M^y(s))}\frac{e^{\tilde{\theta}(s,\Phi,y)}}{\int_{\mathcal{Y}}e^{\tilde{\theta}(s,\Phi,y)}dy}dydx\end{aligned}\tag{C.65}$$

where

$$\tilde{\theta}(s,\Phi,y)=\kappa_M^y(s)^\dagger\Sigma_M(s)^{-1}\kappa_M^y(s)+\Phi(y)^\dagger\frac{U}{s}\Phi(y)-y^\dagger\Sigma_W^{-1}y.$$

Next, the average pay-off, given by (C.27) has to be differentiated and the derivative set to zero. The inner integral will be differentiated as follows.

$$\frac{d}{d\Phi}\Big[\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi(y))}{s}}\eta(y,dx)\Big]\Big|_{\Phi=\Phi^*}=0.\tag{C.66}$$

First, the differentiation is performed and then the integration, therefore,

$$\int_{\mathcal{X}}\frac{d}{d\Phi}\Big[e^{\frac{\ell(x,\Phi(y))}{s}}\Big]\Big|_{\Phi=\Phi^*}\eta(y,dx)=0.\tag{C.67}$$

Given (3.89), the above function results to

$$\begin{aligned}&\int_{\mathcal{X}}\Big(\frac{2U}{s}\Phi^*(y)-\frac{2U}{s}x\Big)e^{\frac{\ell(x,\Phi^*(y))}{s}}\eta(y,dx)=0\\&\Rightarrow\Phi^*(y)=\frac{\int_{\mathcal{X}}xe^{\frac{\ell(x,\Phi^*(y))}{s}}\eta(y,dx)}{\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi^*(y))}{s}}\eta(y,dx)}=\frac{\int_{\mathcal{X}}xe^{\frac{\ell(x,\Phi^*(y))}{s}}\mu(x,dy)dP_X(x)}{\int_{\mathcal{X}}e^{\frac{\ell(x,\Phi^*(y))}{s}}\mu(x,dy)dP_X(x)}.\end{aligned}\tag{C.68}$$

The function $e^{\frac{\ell(x,\Phi^*(y))}{s}}\mu(x,dy)dP_X(x)$ has already derived and is equivalent to (C.63) (note that $\Phi(y)=\Phi^*(y)$). Therefore

$$\begin{aligned}\Phi^*(y) &= \frac{\kappa_M^{*,y}(s)\ e^{\tilde{\theta}_1(s,\Phi^*,y)}dy}{e^{\tilde{\theta}_1(s,\Phi^*,y)}dy}\\&= \Big(H^\dagger\Sigma_W^{-1}H+\Sigma_X^{-1}-\frac{U}{s}\Big)^{-1}\Big(H^\dagger\Sigma_W^{-1}y-\frac{U}{s}\Phi^*(y)\Big)\end{aligned}\tag{C.69}$$

The above function can be further manipulated, before it reaches its final form. That is,

$$\begin{aligned}
&\Rightarrow \Big(H^\dagger \Sigma_W^{-1} H + \Sigma_X^{-1} - \frac{U}{s}\Big)\Phi^*(y) = H^\dagger \Sigma_W^{-1} y - \frac{U}{s}\Phi^*(y) \\
&\Rightarrow \Big(H^\dagger \Sigma_W^{-1} H + \Sigma_X^{-1}\Big)\Phi^*(y) = H^\dagger \Sigma_W^{-1} y \\
&\Rightarrow \Phi^*(y) = \Big(H^\dagger \Sigma_W^{-1} H + \Sigma_X^{-1}\Big)^{-1} H^\dagger \Sigma_W^{-1} y. \qquad \text{(C.70)}
\end{aligned}$$

Finally, applying the PosDef Identity

$$\Phi^*(y) = \Sigma_X H^\dagger \Big(H \Sigma_X H^\dagger + \Sigma_W\Big)^{-1} y \qquad \text{(C.71)}$$

APPENDIX D

PROOF OF THEOREM 4.4.4

First the following parameters are defined, $\overline{\sigma}^{2,m} = \frac{\sigma^2}{1+2\sigma^2 K_i^m}$ and $\epsilon^m = V_i(y^m)\overline{\sigma}^{2,m}$ which are used throughout the derivation.

(a) When (4.63) is substituted into (4.64), the following is obtained.

$$\begin{aligned}\hat{\Lambda}(t_m) &= \prod_{i=1}^{N}\Big[\int_0^\infty \frac{r_i}{\sigma^2}\exp\Big(-\frac{r_i^2}{2\sigma^2}\Big)\exp\left(-r_i^2 K_i^m\right) \\ &\quad \times\frac{1}{2\pi}\int_0^{2\pi}\exp\Big(r_i V_i(y^m)\cos(\theta_i-\gamma_i(y^m))\Big)d\theta_i dr_i\Big] \\ &\quad \times\exp\Big(\frac{1}{2}\sum_{k=0}^{m}y^2(t_k)[1-D^{-2}(t_k)]-\sum_{k=0}^{m}\log|D(t_k)|\Big) \\ &= \prod_{i=1}^{N}\Big[\int_0^\infty \frac{r_i}{\sigma^2}\exp\Big(-\frac{r_i^2}{2(\frac{\sigma^2}{1+2\sigma^2K_i^m})}\Big)I_0\Big(r_i V_i(y^m)\Big)dr_i\Big] \\ &\quad \times\exp\Big(\frac{1}{2}\sum_{k=0}^{m}y^2(t_k)[1-D^{-2}(t_k)]-\sum_{k=0}^{m}\log|D(t_k)|\Big) \qquad \text{(D.1)}\end{aligned}$$

where $I_0(.)$ is the modified bessel function of the first kind and zeroth order defined by $I_0(x) \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi}\exp(x\cos\alpha)d\alpha$.

Next, using $\overline{\sigma}^{2,m}$ and ϵ^m the integral $\int_0^\infty \frac{r_i}{\sigma^2}\exp\Big(-\frac{r_i^2}{2(\frac{\sigma^2}{1+2\sigma^2K_i^m})}\Big)I_0\Big(r_i V_i(y^m)\Big)dr_i$ is computed as follows

$$\int_0^\infty \frac{r_i}{\sigma^2}\exp\Big(-\frac{r_i^2}{2(\frac{\sigma^2}{1+2\sigma^2K_i^m})}\Big)I_0\Big(r_i V_i(y^m)\Big)dr_i = \int_0^\infty \frac{r_i}{\sigma^2}\exp\Big(-\frac{r_i^2}{2\overline{\sigma}^{2,m}}\Big)I_0(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}})dr_i$$

$$= \frac{\overline{\sigma}^{2,m}}{\sigma^2}\exp\Big(\frac{\epsilon^{2,m}}{2\overline{\sigma}^{2,m}}\Big)\int_0^\infty \frac{r_i}{\overline{\sigma}^{2,m}}\exp\Big(-\frac{(r_i^2+\epsilon^{2,m})}{2\overline{\sigma}^{2,m}}\Big)I_0(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}})dr_i = \frac{\overline{\sigma}^{2,m}}{\sigma^2}\exp\Big(\frac{\epsilon^{2,m}}{2\overline{\sigma}^{2,m}}\Big)$$

$$= \frac{1}{1+2\sigma^2 K_i^m} \exp\Big(\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2 K_i^m}\Big). \tag{D.2}$$

Note that in (D.2) the fact that $\frac{r_i}{\overline{\sigma}^{2,m}} \exp\Big(-\frac{(r_i^2+\epsilon^{2,m})}{2\overline{\sigma}^{2,m}}\Big) I_0(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}})$ is a Rice probability distribution function which integrates to 1, is being used. By Substituting (D.2) into (D.1), equation (4.65) is obtained.

(b) The normalized conditional density $p_N(t_m, \boldsymbol{\theta}, \mathbf{r}|\mathcal{Y}_m)$ is derived by substituting (4.63) and (4.65) into (4.66).

(c) Using the definition of $h_i(t_m, \theta_i, r_i)$, and defining $\overline{K_i^m} \triangleq \frac{(1+2\sigma^2 K_i^m)}{2\pi\sigma^2}$, by substituting (4.67) in (4.68), the following is obtained

$$\begin{aligned}\hat{h}_i(t_m, \theta_i, r_i) &= \int_0^\infty \int_0^{2\pi} \Big[\cos(\omega_c(t_m - \tau_i(t_m)) + \theta_i) S(t_m - \tau_i(t_m)) r_i \overline{K_i^m} \exp\Big(-r_i^2 \overline{K_i^m}\Big) \\ &\quad \times \exp\Big(r_i V_i(y^m)\cos(\theta_i - \gamma_i(y^m))\Big) \exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2 K_i^m}\Big)\Big] d\theta_i dr_i.\end{aligned} \tag{D.3}$$

Now, by writing $\cos(\omega_c(t_m - \tau_i(t_m)) + \theta_i) = \cos(\omega_c(t_m - \tau_i(t_m)) + \gamma_i(y^m) + \theta_i - \gamma_i(y^m))$ and using the trigonometric identity $\cos(A+B) = \cos(A)\cos(B) - \sin(A)\sin(B)$, after some algebra, (D.3) transforms into

$$\begin{aligned}\hat{h}_i(t_m, \theta_i, r_i) &= S(t_m - \tau_i(t_m)) \cos(\omega_c(t_m - \tau_i(t_m)) + \gamma_i(y^m)) 2\pi \overline{K_i^m} \\ &\quad \times \exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2 K_i^m}\Big) \int_0^\infty r_i^2 \exp\Big(-2\pi r_i^2 \overline{K_i^m}\Big) I_1\Big(r_i V_i(y^m)\Big) dr_i\end{aligned} \tag{D.4}$$

where $I_1(.)$ is the modified first order Bessel function of the first order defined by $I_1(x) \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos(\alpha)\exp(x\cos\alpha) d\alpha$.

Next, by substituting $\overline{\sigma}^{2,m}$ and ϵ^m into (D.4), the following is obtained

$$\begin{aligned}\hat{h}_i(t_m, \theta_i, r_i) &= S(t_m - \tau_i(t_m)) \cos(\omega_c(t_m - \tau_i(t_m)) + \gamma_i(y^m)) \frac{1}{\overline{\sigma}^{2,m}} \exp\Big(-\frac{\epsilon^{2,m}}{2\overline{\sigma}^{2,m}}\Big) \\ &\quad \times \int_0^\infty r_i^2 \exp\Big(-\frac{r_i^2}{2\overline{\sigma}^{2,m}}\Big) I_1\Big(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}}\Big) dr_i \\ &= S(t_m - \tau_i(t_m)) \cos(\omega_c(t_m - \tau_i(t_m)) + \gamma_i(y^m)) \epsilon^m \\ &\quad \times \int_0^\infty \frac{r_i^2}{\overline{\sigma}^{2,m}\epsilon^m} \exp\Big(-\frac{(r_i^2+\epsilon^{2,m})}{2\overline{\sigma}^{2,m}}\Big) I_1\Big(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}}\Big) dr_i.\end{aligned} \tag{D.5}$$

Note that in (D.5) the fact that $\frac{r_i^2}{\overline{\sigma}^{2,m}\epsilon^m}\exp\Big(-\frac{(r_i^2+\epsilon^{2,m})}{2\overline{\sigma}^{2,m}}\Big)I_1\Big(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}}\Big)$ is the pdf of a noncentral chi-square distribution with 2 degrees of freedom and noncentrality parameter $\epsilon^{2,m}$, is being used, thus it integrates to 1. By taking this into consideration equation (4.69) is derived.

(d) The minimum least-square estimator $\tilde{\theta}_i^*(t_m)$ is derived by substituting (4.67) into (4.70).

(e) Substituting (4.67) in (4.72), the following function is obtained

$$\begin{aligned}\tilde{r}_i^*(t_m) &= \int_0^\infty\int_0^{2\pi} r_i\frac{r_i(1+2\sigma^2K_i^m)}{2\pi\sigma^2}\exp\Big(-\frac{r_i^2(1+2\sigma^2K_i^m)}{2\sigma^2}\Big)\\ &\quad\times\exp\Big(r_iV_i(y^m)\cos(\theta_i-\gamma_i(y^m))\Big)\exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2K_i^m}\Big)d\theta_i dr_i\\ &= \exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2K_i^m}\Big)\int_0^\infty r_i\frac{r_i(1+2\sigma^2K_i^m)}{\sigma^2}\\ &\quad\times\exp\Big(-\frac{r_i^2(1+2\sigma^2K_i^m)}{2\sigma^2}\Big)I_0\Big(r_iV_i(y^m)\Big)dr_i. \qquad \text{(D.6)}\end{aligned}$$

By substituting $\overline{\sigma}^{2,m}$ and ϵ^m into (D.6), then

$$\begin{aligned}\tilde{r}_i^*(t_m) &= \exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2K_i^m}\Big)\exp\Big(\frac{\epsilon^{2,m}}{2\overline{\sigma}^{2,m}}\Big)\\ &\quad\times\int_0^\infty r_i\frac{r_i}{\overline{\sigma}^{2,m}}\exp\Big(-\frac{(r_i^2+\epsilon^{2,m})}{2\overline{\sigma}^2}\Big)I_0\Big(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}}\Big)dr_i\\ &= \exp\Big(-\frac{V_i^2(y^m)\sigma^2}{2+4\sigma^2K_i^m}\Big)\exp\Big(\frac{\epsilon^{2,m}}{2\overline{\sigma}^{2,m}}\Big)E[r]. \qquad \text{(D.7)}\end{aligned}$$

Note that $\frac{r_i}{\overline{\sigma}^{2,m}}\exp\Big(-\frac{(r_i^2+\epsilon^{2,m})}{2\overline{\sigma}^{2,m}}\Big)I_0(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}})$ is the pdf of a Ricean distribution random variable, hence the integral $\int_0^\infty r_i\frac{r_i}{\overline{\sigma}^{2,m}}\exp\Big(-\frac{r_i^2}{2\overline{\sigma}^2}\Big)I_0\Big(\frac{r_i\epsilon^m}{\overline{\sigma}^{2,m}}\Big)dr_i$ is its fist moment. Using the expression of the moments of a Rice distribution which are given by

$$E[r_i^k]=(2\overline{\sigma}^{2,m})^{k/2}\exp\Big(-\frac{\epsilon^{2,m}}{2\overline{\sigma}^{2,m}}\Big)\Gamma(1+\frac{k}{2})_1F_1\Big(\frac{2+k}{2},1;\frac{\epsilon^{2,m}}{2\overline{\sigma}^{2,m}}\Big),$$

where ${}_1F_1\Big(\alpha,\beta;x\Big)$ is the confluent hypergeometric function equation (4.73) is obtained.

APPENDIX E

DERIVATIONS OF EXAMPLES OF CHAPTER 5

This Appendix presents the derivations of the solutions of the Examples of Chapter 5.

Derivations of Example of Section 5.4.1 - Generalized MAP Estimator

First the Generalized MAP estimator Φ_{MAP}, which is given by (5.27), is derived. Using (5.20), (5.21) and (5.22) the function $\overline{L}(x, y^m)$ is given by

$$\begin{aligned}
\overline{L}(x, y^m) &\triangleq \frac{\ell(x)}{s} + \log \frac{\mu(x, dy^m)}{dy^m} + \log \frac{dP_x(x)}{dx} \\
&= \frac{x^T U x}{s} + \log \left(\frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_W|^{\frac{1}{2}}} \right)^{m+1} \\
&\quad - \sum_{i=0}^{m} (y_i - Hx)^T \frac{\Sigma_W^{-1}}{2} (y_i - Hx) \\
&\quad + \log \left(\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_X|^{\frac{1}{2}}} \right) - x^T \frac{\Sigma_X^{-1}}{2} x \\
&= \frac{x^T U x}{s} + \log \left((2\pi)^{-\frac{d(m+1)}{2}} (2\pi)^{-\frac{n}{2}} |\Sigma_W|^{-\frac{1}{2}} |\Sigma_X|^{\frac{1}{2}} \right) \\
&\quad - \sum_{i=0}^{m} y_i^T \frac{\Sigma_W^{-1}}{2} y_i - N x^T H^T \frac{\Sigma_W^{-1}}{2} H x \\
&\quad + 2 x^T H^T \frac{\Sigma_W^{-1}}{2} \sum_{i=0}^{m} y_i - x^T \frac{\Sigma_X^{-1}}{2} x. \qquad \text{(E.1)}
\end{aligned}$$

Next, the MAP estimator is computed by differentiating the above pay-off function and setting the derivative to zero as shown in (5.26). That is

$$\frac{\partial}{\partial x}\overline{L}(x, y^m)|_{x=\Phi_{MAP}(y^m)} = 0 \tag{E.2}$$

$$\begin{aligned}
&\Rightarrow \frac{2(m+1)}{s}\Phi_{MAP}(y^m) - (m+1)H^T\Sigma_W^{-1}H\Phi_{MAP}(y^m) + H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i \\
&-\Sigma_X^{-1}\Phi_{MAP}(y^m) = 0 \\
&\Rightarrow \Big(\frac{2(m+1)}{s} - (m+1)H^T\Sigma_W^{-1}H - \Sigma_X^{-1}\Big)\Phi_{MAP}(y^m) + H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i = 0 \\
&\Rightarrow \Phi_{MAP}(y^m) = \Big((m+1)H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - \frac{2U}{s}\Big)^{-1} H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i.
\end{aligned} \tag{E.3}$$

Secondly, the mean square error, given by (5.30), is derived. The mean square error is given by (5.28), and can be written as

$$\begin{aligned}
&E\Big[(X - \Phi_{MAP}(Y^m))^T(X - \Phi_{MAP}(Y^m))\Big] = \\
&tr\int_{\Re^n}\int_{\Re^{(m+1)d}}(x - \Phi_{MAP}(y^m))(x - \Phi_{MAP}(y^m))^T\mu(x, dy^m)dP_X(x)
\end{aligned} \tag{E.4}$$

$$\begin{aligned}
&= tr\int_{\Re^n}\int_{\Re^{(m+1)d}}\Big(xx^T + \Phi_{MAP}(y^m)\Phi_{MAP}^T(y^m) - 2x\Phi_{MAP}^T(y^m)\Big)\mu(x, dy^m)dP_X(x) \\
&= tr\int_{\Re^n}\int_{\Re^{(m+1)d}} xx^T\mu(x, dy^m)dP_X(x) \\
&-tr\int_{\Re^n}\int_{\Re^{(m+1)d}} 2x\Phi_{MAP}^T(y^m)\mu(x, dy^m)dP_X(x) \\
&+tr\int_{\Re^n}\int_{\Re^{(m+1)d}} \Phi_{MAP}(y^m)\Phi_{MAP}^T(y^m)\mu(x, dy^m)dP_X(x).
\end{aligned} \tag{E.5}$$

Next, given (5.27) $\Phi_{MAP}(y^m) = \Sigma_{\Phi_{MAP}}\sum_{i=0}^{m} y_i$, where $\Sigma_{\Phi_{MAP}} = \Big((m+1)H^T\Sigma_W^{-1}H + \Sigma_X^{-1} - \frac{2U}{s}\Big)^{-1}H^T\Sigma_W^{-1}$, (E.5) can also be written as

$$\begin{aligned}
&tr\int_{\Re^n}\int_{\Re^{(m+1)d}} xx^T\mu(x, dy^m)dP_X(x) - tr\int_{\Re^n}\int_{\Re^{(m+1)d}} 2x\sum_{i=0}^{m} y_i^T\Sigma_{\Phi_{MAP}}^T\mu(x, dy^m)dP_X(x) \\
&+tr\int_{\Re^n}\int_{\Re^{(m+1)d}} \Sigma_{\Phi_{MAP}}\sum_{i=0}^{m} y_i\sum_{i=0}^{m} y_i^T\Sigma_{\Phi_{MAP}}^T\mu(x, dy^m)dP_X(x) \\
&= tr(\Sigma_X) - tr\Big(2\Sigma_{\Phi_{MAP}}^T\int_{\Re^n}\int_{\Re^{(m+1)d}} x\sum_{i=0}^{m}(Hx + w_i)^T\mu(x, dy^m)dP_X(x)\Big)
\end{aligned}$$

$$
\begin{aligned}
&+tr\Big(\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}\int_{\Re^n}\int_{\Re^{(m+1)d}}\sum_{i=0}^{m}\sum_{j=0}^{m}y_iy_j^T\mu(x,dy^m)dP_X(x)\Big)\\
&=tr(\Sigma_X)-tr\Big(2\Sigma^T_{\Phi_{MAP}}\int_{\Re^n}\int_{\Re^{(m+1)d}}(m+1)xx^TH^T\mu(x,dy^m)dP_X(x)\Big)\\
&+tr\Big(\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}\int_{\Re^n}\int_{\Re^{(m+1)d}}\sum_{i=0}^{m}\sum_{j=0}^{m}(Hx+w_i)(Hx+w_j)^T\mu(x,dy^m)dP_X(x)\Big)\\
&=tr(\Sigma_X)-tr(2(m+1)H^T\Sigma^T_{\Phi_{MAP}}\Sigma_X)\\
&+tr\Big(\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}\int_{\Re^n}\int_{\Re^{(m+1)d}}\sum_{i=0}^{m}\sum_{j=0}^{m}(Hxx^TH^T+xHw_j^T+w_ix^TH^T+w_iw_j^T)\\
&\times\mu(x,dy^m)dP_X(x)\Big)\\
&=tr(\Sigma_X-2(m+1)H^T\Sigma^T_{\Phi_{MAP}}\Sigma_X)\\
&+tr\Big(\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}\int_{\Re^n}\int_{\Re^{(m+1)d}}((m+1)^2Hxx^TH^T)\mu(x,dy^m)dP_X(x)\Big)\\
&+tr\Big(\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}\int_{\Re^n}\int_{\Re^{(m+1)d}}(m+1)w_iw_i^T\mu(x,dy^m)dP_X(x)\Big)\\
&=tr(\Sigma_X-2(m+1)H^T\Sigma^T_{\Phi_{MAP}}\Sigma_X)+tr\Big((m+1)^2H^T\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}H\Sigma_X\\
&+(m+1)\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}\Sigma_W)\\
&=tr\Big(\Sigma_X-2(m+1)H^T\Sigma^T_{\Phi_{MAP}}\Sigma_X+(m+1)^2H^T\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}H\Sigma_X\\
&+(m+1)\Sigma^T_{\Phi_{MAP}}\Sigma_{\Phi_{MAP}}\Sigma_W\Big).
\end{aligned}
\tag{E.6}
$$

Derivations of Example of Section 5.4.2 - Generalized ML Estimator

First the Generalized ML estimator Φ_{ML}, which is given by (5.35), is derived. The MAP estimator is computed by differentiating the pay-off function, $\overline{L}_2(x,y^m)$, and setting the derivative to zero as shown in (5.34). By expanding (5.33), $\overline{L}_2(x,y^m)$ is given by

$$
\overline{L}_2(x,y^m) \;=\; \frac{x^TUx}{s}+\log\left((2\pi)^{-\frac{d}{2}}|\Sigma_W|^{-\frac{1}{2}}\right)^{m+1}-\sum_{i=0}^{m}y_i^T\frac{\Sigma_W^{-1}}{2}y_i
$$

$$-(m+1)x^T H^T \frac{\Sigma_W^{-1}}{2} Hx + x^T H^T \Sigma_W^{-1} \sum_{i=0}^{m} y_i. \tag{E.7}$$

Next, the differentiation is performed.

$$\frac{\partial}{\partial x} \overline{L}_2(x, y^m)|_{x=\Phi_{ML(y^m)}} = 0$$

$$\Rightarrow \frac{2U}{s}\Phi_{ML}(y^m) - (m+1)H^T\Sigma_W^{-1}H\Phi_{ML}(y^m) + H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i = 0$$

$$\Rightarrow \Big(\frac{2U}{s} - (m+1)H^T\Sigma_W^{-1}H\Big)\Phi_{ML}(y^m) + H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i = 0$$

$$\Rightarrow \Phi_{ML}(y^m) = \Big((m+1)H^T\Sigma_W^{-1}H - \frac{2U}{s}\Big)^{-1} H^T\Sigma_W^{-1}\sum_{i=0}^{m} y_i. \tag{E.8}$$

Secondly the mean square error, given by (5.37), is derived. The mean square error is given by (5.36), and can be written as (for $X = x$ deterministic)

$$E\Big[x - \Phi_{ML}(Y^m))^T(x - \Phi_{ML}(Y^m))\Big] =$$

$$tr\int_{\Re^{(m+1)d}} (x - \Phi_{ML}(y^m))(x - \Phi_{ML}(y^m))^T \mu(x, dy^m) \tag{E.9}$$

$$= tr\int_{\Re^{(m+1)d}} \Big(xx^T + \Phi_{ML}(y^m)\Phi_{ML}^T(y^m) - 2x\Phi_{ML}^T(y^m)\Big)\mu(x, dy^m)$$

$$= tr\int_{\Re^{(m+1)d}} xx^T\mu(x, dy^m) - tr\int_{\Re^{(m+1)d}} 2x\Phi_{ML}^T(y^m)\mu(x, dy^m)$$

$$+tr\int_{\Re^{(m+1)d}} \Phi_{ML}(y^m)\Phi_{ML}^T(y^m)\mu(x, dy^m). \tag{E.10}$$

Given (5.35) $\Phi_{ML}(y^m) = \Sigma_{\Phi_{ML}} \sum_{i=0}^{m} y_i$, where $\Sigma_{\Phi_{ML}} = \Big((m+1)H^T\Sigma_W^{-1}H - \frac{2U}{s}\Big)^{-1} H^T\Sigma_W^{-1}$, (E.10) can also be written as

$$tr\int_{\Re^{(m+1)d}} xx^T\mu(x, dy^m) - tr\int_{\Re^{(m+1)d}} 2x\sum_{i=0}^{m} y_i^T\Sigma_{\Phi_{ML}}^T\mu(x, dy^m)$$

$$+tr\int_{\Re^{(m+1)d}} \Sigma_{\Phi_{ML}}\sum_{i=0}^{m} y_i \sum_{i=0}^{m} y_i^T\Sigma_{\Phi_{ML}}^T\mu(x, dy^m)$$

$$= tr(xx^T) - tr\Big(2\Sigma_{\Phi_{ML}}^T\int_{\Re^{(m+1)d}} x\sum_{i=0}^{m}(Hx + w_i)^T\mu(x, dy^m)\Big)$$

$$+tr\Big(\Sigma_{\Phi_{ML}}^T\Sigma_{\Phi_{ML}}\int_{\Re^{(m+1)d}} \sum_{i=0}^{m}\sum_{j=0}^{m} y_i y_j^T\mu(x, dy^m)\Big)$$

$$
\begin{aligned}
&= tr(xx^T) - tr\Big(2\Sigma_{\Phi_{ML}}^T \int_{\Re^{(m+1)d}} (m+1)xx^T H^T \mu(x, dy^m)\Big) \\
&+tr\Big(\Sigma_{\Phi_{ML}}^T \Sigma_{\Phi_{ML}} \int_{\Re^{(m+1)d}} \sum_{i=0}^{m}\sum_{j=0}^{m} (Hx + w_i)(Hx + w_j)^T \mu(x, dy^m)\Big) \\
&= tr(xx^T) - tr(2(m+1)H^T \Sigma_{\Phi_{ML}}^T xx^T) \\
&+tr\Big(\Sigma_{\Phi_{ML}}^T \Sigma_{\Phi_{ML}} \int_{\Re^{(m+1)d}} \sum_{i=0}^{m}\sum_{j=0}^{m} (Hxx^T H^T + xHw_j^T + w_i x^T H^T + w_i w_j^T)\mu(x, dy^m)\Big) \\
&= tr(xx^T - 2(m+1)H^T \Sigma_{\Phi_{ML}}^T xx^T) \\
&+tr\Big(\Sigma_{\Phi_{ML}}^T \Sigma_{\Phi_{ML}} \int_{\Re^{(m+1)d}} ((m+1)^2 Hxx^T H^T)\mu(x, dy^m)\Big) \\
&+tr\Big(\Sigma_{\Phi_{ML}}^T \Sigma_{\Phi_{ML}} \int_{\Re^{(m+1)d}} (m+1)w_i w_i^T \mu(x, dy^m)\Big) \\
&= tr(xx^T - 2(m+1)H^T \Sigma_{\Phi_{ML}}^T xx^T) + tr\Big((m+1)^2 H^T \Sigma_{\Phi_{ML}}^T \Sigma_{\Phi_{ML}} Hxx^T \\
&+(m+1)\Sigma_{\Phi_{ML}}^T \Sigma_{\Phi_{ML}} \Sigma_W) \\
&= tr\Big(xx^T - 2(m+1)H^T \Sigma_{\Phi_{ML}}^T xx^T + (m+1)^2 H^T \Sigma_{\Phi_{ML}}^T \Sigma_{\Phi_{ML}} Hxx^T \\
&+(m+1)\Sigma_{\Phi_{ML}}^T \Sigma_{\Phi_{ML}} \Sigma_W\Big).
\end{aligned}
\tag{E.11}
$$

Finally, the derivation of the $Cramér - Rao\ lower\ bound$ is presented. First, equation (5.42) is derived.

$$
\begin{aligned}
Var[\Phi_{ML}(Y^m)] &= E\Big[\Big(\Phi_{ML}(Y^m) - E[\Phi_{ML}(Y^m)]\Big)\Big(\Phi_{ML}(Y^m) - E[\Phi_{ML}(Y^m)]\Big)^T\Big] \\
&= E\Big[\Sigma_{\Phi_{ML}}\Big(\sum_{i=0}^{m} y_i - \sum_{i=0}^{m} E[y_i]\Big)\Big(\sum_{i=0}^{m} y_i - \sum_{i=0}^{m} E[y_i]\Big)^T \Sigma_{\Phi_{ML}}^T\Big] \\
&= \Sigma_{\Phi_{ML}} \sum_{i=0}^{m} Var[y_i] \Sigma_{\Phi_{ML}}^T \\
&= (m+1)\Sigma_{\Phi_{ML}} \Sigma_W \Sigma_{\Phi_{ML}}^T.
\end{aligned}
\tag{E.12}
$$

Now, given $E[\Phi_{ML}(y^m)] = \Sigma_{\Phi_{ML}}(m+1)Hx$, then

$$
V_X = \frac{\partial}{\partial x} E_X[\Phi_{ML}(y^m)] = \frac{\partial}{\partial x}\Sigma_{\Phi_{ML}}(m+1)Hx = \Sigma_{\Phi_{ML}}(m+1)H. \tag{E.13}
$$

Furthermore,

$$
\log \frac{\mu(x, dy^m)}{dy^m} = \log[(2\pi)^{\frac{d}{2}} |\Sigma_W|^{\frac{1}{2}}]^{-(m+1)} - \sum_{i=0}^{m} (y_i - Hx)^T \frac{\Sigma_W^{-1}}{2} (y_i - Hx)
$$

$$= \log[(2\pi)^{\frac{d}{2}}|\Sigma_W|^{\frac{1}{2}}]^{-(m+1)} - \sum_{i=0}^{m} y_i^2 \frac{\Sigma_W^{-1}}{2} y_i$$
$$-(m+1)x^T H^T \frac{\Sigma_W^{-1}}{2} Hx + x^T H^T \Sigma_W^{-1} \sum_{i=0}^{m} y_i. \quad \text{(E.14)}$$

Next, the first derivative of (E.14) is given by

$$\frac{\partial}{\partial x} \log \frac{\mu(x, dy^m)}{dy^m} = -(m+1)H^T \Sigma_w^{-1} Hx + H^T \Sigma_W^{-1} \sum_{i=0}^{m} y_i, \quad \text{(E.15)}$$

and the second derivative is given by

$$\frac{\partial^2}{\partial x^2} \log \frac{\mu(x, dy^m)}{dy^m} = -(m+1)H^T \Sigma_w^{-1} H. \quad \text{(E.16)}$$

Finally,

$$I_X = -E_X \left[\frac{\partial^2}{\partial x^2} \log \frac{\mu(x, dy^m)}{dy} \right] = (m+1)H^T \Sigma_w^{-1} H \quad \text{(E.17)}$$

List of Acronyms

1. a.s. - almost surely
2. AVC - Arbitrarily Varying Channel
3. BME - Blind Minimax Estimator
4. EKF - Extended Kalman Filter
5. iid - independent and identically distributed
6. KL - Kullback-Leibler
7. LS - Least-Square
8. MAP - Maximum $\acute{A}$ Posteriori
9. MIMO - Multiple-Input Multiple-Output
10. ML - Maximum Likelihood
11. MSE - Mean-Square-Error
12. MMAE - Minimum-Mean-Absolute-Error
13. MMSE - Minimum-Mean-Squared-Error
14. PDF - Probability Density Function
15. PSD - Power Spectral Density
16. RND -Radon Nikodym Derivative
17. RP - Random Process
18. RV - Random Variable
19. SNR - Signal-to-Noise Ratio

Notation and List of Symbols

1. $N \stackrel{\triangle}{=} \{1, 2, \ldots\}$.
2. $N_0 \stackrel{\triangle}{=} \{0, 1, 2, \ldots\}$.
3. $N_0^m \stackrel{\triangle}{=} \{0, 1, 2, \ldots\ m\}$.
4. $\mathbb{C}$: set of complex numbers.
5. $\mathbb{R}$: set of real numbers.
6. $\mathcal{Z}$: set.
7. B^T: transpose of a matrix B.
8. $|B|$: determinant of a square matrix B.
9. Ω: elementary outcomes of a random experiment.
10. $\mathcal{F}$: σ-algebra (or algebra) associated with a random experiment.
11. $\mathcal{X}$: Polish space.
12. $\mathcal{Y}$: Polish space.
13. $\Sigma_{\mathcal{X}}$ and $\Sigma_{\mathcal{Y}}$: σ-algebras generated by $\mathcal{X}$ and $\mathcal{Y}$, respectively.
14. $\mathcal{M}_1(\mathcal{X})$: space of probability measures on $\mathcal{X}$.
15. ℓ: sample pay-off.
16. $E[\cdot]$: expectation.
17. $E[X|Y]$: conditional expectation of X given Y.

18. $BC(\mathcal{Y})$: vector space of bounded continuous real valued functions defined on the Polish space $\mathcal{Y}$
19. $\Big(BC(\mathcal{Y})\Big)^*$: topological dual of $BC(\mathcal{Y})$.
20. $P_{X,Y}$: nominal joint distribution of X and Y.
21. $P_{X|Y}$: conditional distribution of X given Y.
22. $Q_{X,Y}$: true joint distribution of X and Y.
23. $Q_{X|Y}$: true conditional distribution of X given Y.
24. P_X: marginal probability distribution of X associated with joint probability of X, Y $P_{X,Y}$.
25. $\mu, \eta, \nu,$: stochastic kernels.
26. $\mathcal{P}$: class of all stochastic kernel.
27. $M_{rba}(\mathcal{Y})$: Banach space of finitely additive regular bounded signed measures on $\Sigma_{\mathcal{Y}}$.
28. $\Pi_{rba}(\mathcal{Y})$: set of regular bounded finitely additive probability measures on $\mathcal{Y}$.
29. $L_1(P_X, BC(\mathcal{Y}))$: space of all P_X integrable functions defined on $\mathcal{X}$ with values in $BC(\mathcal{Y})$.
30. $H(P|Q)$: relative entropy between the probability measures P with respect to Q.
31. $\hat{X} = \Phi(Y)$: estimate of X from the measurements of Y.
32. $\Phi^*(Y)$: best estimate of X from the measurements of Y.
33. $\|.\|_{\mathcal{X}}$: norm associated with elements in $\mathcal{X}$.
34. $\{\mathcal{G}_m^o\}$: σ-field generated by the complete data $\{x_0, x_1, \ldots, x_m, y_0, y_1, \ldots, y_m\}$.
35. $\{\mathcal{G}_m\}$: complete filtration of $\{\mathcal{G}_m^o\}$.
36. $\{\mathcal{Y}_m^o\}$: σ-field generated by the incomplete data $\{y_0, y_1, \ldots, y_m\}$.

37. $\{\mathcal{Y}_m\}$: complete filtration of $\{\mathcal{Y}_m^o\}$.

38. y^m: sequence $\{y_0, \ldots, y_m\}$ and similarly for other sequences.

39. $\tilde{x}_m$: estimate of the state x_m given $\{\mathcal{Y}_m\}$.

40. $\alpha(\cdot)$: unnormalized conditional probability distribution function.

41. $\overline{\alpha}(\cdot)$: unnormalized conditional probability density function.

42. ω_c: carrier frequency.

43. $\{\tau_i(t_k)\}$: propagation delay.

44. $\{r_i\}$, $\{\theta_i\}$: attenuation and phase, respectively, of the signal received associated with ith path.

45. $\pi_{r_0}(r_i)$: *á* priori probability density function of $\{r_i\}$, and similarly for other processes.

46. $\hat{\Lambda}(t_m)$: incomplete data likelihood ratio.

BIBLIOGRAPHY

[1] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 1994.

[2] H. L. VanTrees, *Detection, Estimation and Modulation Theory-Part1*. Wiley, 1968.

[3] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, fourth ed., 2002.

[4] T. Kailath and H. V. Poor, "Detection of stochastic processes," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2230–2259, October 1998.

[5] A. Jazwinski, *Stochastic Processes and Filtering Theory*. San Diego, CA: Academic, 1970.

[6] S. Haykin, *Kalman Filtering and Neural Networks*. New York: Wiley, 2001.

[7] B. Anderson and J. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentince-Hall, 1979.

[8] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2148–2177, October 1998.

[9] V. A. H. Sayed, H. Nascimento, and F. A. M. Cipparrone, "A regularized robust design criterion for uncertain data," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 4, pp. 1120–1142, 2002.

[10] A. H. Sayed, "A framework for state-space estimation with uncertain models," *IEEE Transactions on Automatic Control*, vol. 46, no. 7, pp. 998–1013, July 2001.

[11] Y. Guo and B. C. Levy, "Robust MSE equalizer design for MIMO communication systems in the presense of model uncertainties," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1840–1852, May 2006.

[12] A. G. Dabak and D. H. Johnson, "Geometrically based robust detection," in *Proc. Conf. Information Science and Systems (Baltimore, MD)*, pp. 73–77, March 1993.

[13] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and its Applications*. New York: Springer-Verlag, 2000.

[14] S. Amary and H. Nagaoka, *Methods of Information Geometry*. American Mathematical Society, 2001.

[15] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing," *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433–481, March 1985.

[16] B. C. Levy and R. Nikoukhah, "Robust least-squares estimation with relative entropy constraint," *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 89–104, January 2004.

[17] M. C. Yovits and J. L. Jackson, "Linear filter optimization with game theory considerations," *IRE National Convention Record, pt. 4*, pp. 193–199, 1955.

[18] A. G. Carlton and J. W. Follin, "Recent developments in fixed and adaptive filtering," in *Proc. 2nd AGARD Guided Missiles Seminar, AGARDograpk21, NATO Advanced Group for R & D*, pp. 285–300, 1956.

[19] P. J. Huber, "Robust estimation of a location parameter," *Math. Statist.*, vol. 35, pp. 73–101, 1964.

[20] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, pp. 1753–1758, 1965.

[21] W. L. Root, "Stability in signal detection problems," in *Proc. Symp. in Applied Mathematics*, vol. 16, pp. 247–263, 1964.

[22] F. R. Hampel, "A general qualitive definition of robustness," *Ann. Math. Stat.*, vol. 42, pp. 1887–1896, 1971.

[23] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimation," *IEEE Transactions on Information Theory*, vol. 53, no. 9, pp. 3145–3157, September 2007.

[24] Y. C. Eldar, "Comparing between estimation approaches: Admissible and dominating linear estimators," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1689–1702, May 2006.

[25] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean-squared error estimation in the presence of model uncertainties," *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 168–181, January 2005.

[26] Z. Ben-Haim and Y. C. Eldar, "Minimax estimators dominating the least-squares estimator," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2005)*, vol. IV, pp. 53–56, March 2005.

[27] J. H. Marton, V. Krishnamurthy, and H. V. Poor, "James-Stein state filtering algorithms," *IEEE Transactions on Signal Processing*, vol. 46, no. 9, pp. 2431–2447, September 1998.

[28] Y. C. Eldar and N. Merhav, "A competitive minimax approach to robust estimation of random parameters," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1931–1946, July 2004.

[29] R. K. Boel, M. R. James, and I. R. Petersen, "Robustness and risk-sensitive filtering," *IEEE Transactions on Automatic Control*, vol. 47, no. 3, pp. 451–461, March 2002.

[30] K. Vastola and H. V. Poor, "Robust Wiener-Kolmogorov theory," *IEEE Transactions on Information Theory*, vol. 30, no. 2, pp. 316–326, 1984.

[31] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261–1283, April 2005.

[32] Y. Wu and S. Verdú, "MMSE dimension," in *Proc. of 2010 IEEE International Symposium in Information Theory (ISIT 2010)*, pp. 1463–1467, June 2010.

[33] Y. Wu and S. Verdú, "Functional properties of MMSE," in *Proc. of 2010 IEEE International Symposium in Information Theory (ISIT 2010)*, pp. 1453–1457, June 2010.

[34] T. Feng, T. R. Field, and S. Haykin, "Stochastic differential equation theory applied to wireless channels," *IEEE Transactions on Communications*, vol. 55, no. 8, pp. 1478–1483, August 2007.

[35] K. E. Baddour and N. C. Beaulieu, "Autoregressive modeling for fading channel simulation," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1650–1662, July 2005.

[36] R. Elliott, L. Aggoun, and J. Moore, *Hidden Markov Models: Estimation and Control*. Springer-Verlag, Berlin-Heidelberg-New York 29, 1994.

[37] N. U. Ahmed, *Dynamic Systems and Control with Applications*. World Scientific, Singapore, 2006.

[38] F. Rezaei, C. D. Charalambous, and N. Ahmed, "Optimization of stochastic uncertain systems with variational norm constraints," in *46th IEEE Conference on Decision and Control*, pp. 2159–2163, December 2007.

[39] K. Fan, "Minmax theorems," in *Proc. Nat. Acad. Sci.*, pp. 42–47, 1953.

[40] D. G. Luenberger, *Optimization by Vector Space Method*. John Wiley, 1969.

[41] C. D. Charalambous, "Lecture notes on random processes." Dept. of ECE, University of Cyprus., 2008.

[42] E. Telatar, "Capacity of multi-antenna gaussian channels," *European Trans. Telecomm.*, vol. 10, no. 6, pp. 585–595, 1999.

[43] C. D. Charalambous and F. Rezaei, "Stochastic uncertain systems subject to relative entropy constraints: Induced norms and monotonicity properties of minimax games," *IEEE Transactions on Automatic Control*, vol. 52, no. 4, pp. 647–663, May 2007.

[44] P. D. Pra, L. Meneghini, and W. Runggaldier, "Some connections between stochastic control and dynamic games," *Mathematichs of Control Signals, and Systems*, vol. 9, pp. 303–326, 1996.

[45] N. Dunford and J. T. Schwartz, *Linear Operators, Part I: General Theory*. Interscience Publishers, Inc., New York, 1958.

[46] A. I. Tulcea and C. I. Tulcea, *Topics in the Theory of Lifting*. Springer Verlag,Berlin, Heidelberg, New York, 1969.

[47] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley, 1997.

[48] G. J. Foschini, "Layered space-time architecture for wireless communications in a fading environment," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[49] G. J. Foschini and M. J. Gans, "On the limit of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, no. 3, pp. 311–335, 1998.

[50] C. D. Charalambous and Y. Socratous, "Nonlinear estimation for a class of systems," in *Proc. of the 2006 IEEE Symposium in Information Theory*, pp. 841–845, July 2006.

[51] D. E. Asraf and M. G. Gustafsson, "An analytical series expansion to the problem of noncoherent detection," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3369–3375, December 2004.

[52] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 4th ed., 2001.

[53] C. D. Charalambous, A. Nejad, and D. Makrakis, "Coherent and noncoherent wireless channel estimation via ML/MAP and EM algorithm," in *Proc. of the 21st Biennial Symposium on Communications*, pp. 241–244, June 2002.

[54] T. T. Georgiou and A. Lindquist, "Kullback-Leibler approximations of spectral density functions," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 2910–2917, November 2003.

[55] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, 1964.

[56] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Uni. Press, 1985.

CURRICULUM VITAE

Yiannis Socratous

E-mail: yianniss@ucy.ac.cy, ysocratous@gmail.com

Education

2004 - 2010 PhD in Electrical Engineering,
Electrical and Computer Engineering Department,
University of Cyprus.

2001 - 2002 Master of Science in Mobile and Personal Communications,
King's College London, UK. Grade: with Distinction.

1998 - 2001 Bachelor of Engineering in Electrical and Electronic Engineering,
University of Plymouth, UK. Grade: First Class Honours.

Publications

Journals

1. Y. Socratous, F. Rezaei, and C. D. Charalambous, "Nonlinear estimation for a class of systems", *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1930-1938, April 2009.

Conferences

1. C. D. Charalambous and Y. Socratous, "Nonlinear estimation for a class of systems", *in Proc. of IEEE International Symposium on Information Theory*, pp. 841-845, 2006.

2. Y. Socratous, C. D. Charalambous and C. N. Georghiades, "Least-square estimation for nonlinear systems with applications to phase and envelope estimation in wireless fading channels", *in Proc. of the 47th IEEE Conference on Decision and Control*, pp. 4928-4932, 2008.

3. Y. Socratous, C. D. Charalambous and C. N. Georghiades, "Robust estimation with applications to phase and envelope estimation in frequency selective wireless fading channels", *in Proc. of IEEE International Symposium on Information Theory*, pp. 1238-1242, 2008.

Pending Publications

Journals

1. "Non-Coherent Detection and Estimation".

2. "Robust Least-square estimation for a class of systems".