LEARNING FROM NOISY DATA WITH APPLICATIONS TO
FILTERING AND DENOISING

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL
ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Taesup Moon
September 2008
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Tsachy Weissman)  Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Thomas M. Cover)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Abbas El Gamal)

Approved for the University Committee on Graduate Studies.

Peter J. O'Neil
Abstract

A conventional approach to estimation problems is the so-called Bayesian inference, in which the estimator employs an optimum strategy with respect to a pre-assumed probabilistic model on the data. Although the Bayesian approach has resulted in many efficient and practical estimation schemes, one of its main drawbacks is the fact that it assumes a pre-specified prior model on the data, which does not always correspond with practical scenarios, such as image denoising and target tracking, that give rise to model uncertainty and mismatches.

In this thesis, inspired by much of the work in information theory, we present an alternative universality approach that alleviates this drawback by constructing schemes that have guaranteed performances regardless of the statistical characteristics of the data. In particular, we focus on the case of noisy data, emanating from an unknown source corrupted by noise whose statistical characteristics are known to various degrees. We consider two types of estimators: the causal estimator and the non-causal estimator. The first is also referred to as a filter, and the latter a denoiser. For several different scenarios, we devise universal estimation schemes that achieve the performance of the optimum scheme that would have been designed with complete knowledge of the source and noise statistics. Our results demonstrate that as the data observation length increases, knowledge of the noisy channel suffices to learn about the source well enough to optimally estimate it from the noisy data.

More specifically, we consider three different problem settings. First, we devise a universal filter that attains the performance of the optimum filter for any stationary and ergodic, finite-alphabet source data corrupted by discrete memoryless channels. We utilize a rich class of hidden Markov processes to model the noisy data and learn
the true posterior probability required for the filtering. Next, we again consider a universal filtering problem with real-valued source and noise. With known noise variance, we devise a filter that universally attains the performance of the best FIR filter for any bounded source data, under the mean-squared error (MSE) criterion. We combine the techniques from information theory and online learning to design and analyze the family of filters that we devise. Finally, we consider the problem of discrete denoising and generalize the recently introduced Discrete Universal DEnoiser (DUDE) to obtain a practical scheme that can compete with the best switching between sliding window denoisers. To accomplish this, we generalize the tools developed for the DUDE framework and employ dynamic programming to devise an efficiently implementable algorithm.
When I first arrived at Stanford six years ago, I had no idea about what kind of people
I would meet and what kind of life it would be as a graduate student at Stanford.
Now, in hindsight, I have so many people to thank for helping me stay afloat and
finish.

First, I am deeply indebted to Professor Tsachy Weissman, my advisor, for giving
me the chance to be his first PhD student. I still remember the moment I knocked on
his office door when he just arrived at Stanford as a new faculty. Since then, I have
learned tremendously from him not only about information theory, but also about
what it means to be a genuine researcher. I feel truly fortunate to have a great mentor
as him as my advisor. Without his endless patience, guidance, encouragement, and
support, I could not have survived at Stanford. I hope I can continue the sincere
relationship with him even after I graduate.

I also would like to thank Professor Thomas M. Cover, my associate advisor,
for his constant inspiration. I have attended his weekly research meetings for three
years, and the various topics discussed in the meeting were the sources of stimuli for
my PhD research. I admire his great passion, curiosity, and humility. I also would
like to thank Professor Abbas El Gamal for his care and concern about my studies.
Moreover, his excellent courses and lecture notes on statistical signal processing and
network information theory were true gems that I was able to enjoy as a Stanford
graduate student. Thanks also to Professor Joseph Kahn for generously agreeing to
serve as a university chair for my oral defense.

My thanks also goes to Kelly Yilmaz, Denise Murphy, and Rashmi Shah who
stood between me and the university bureaucracy to make every administrative issue
work out smoothly. I also thank The Korean Foundation for Advanced Studies and The Samsung Scholarship Foundation for their financial supports throughout my six years of graduates studies, which enabled me to freely explore the academia without worrying about finances.

My office mates, Haim Permuter, Paul Cuff, Shirin Jalali, Lei Zhao, Stephanie Pereira, and Shahriyar Matloub, were great friends and colleagues to share our lives and ideas. I sincerely thank them for making stimulating and fun office atmosphere. I also want to thank many Korean friends I met at Stanford: Seong Taek Chung, Young-Han Kim, Jung-Yong Lee, Sangwook Ha, Jung Woo Lee, Daeho Lee, Jungho Ahn, Jungwon Lee, Sunhwan Lee, YeoMyoung Cho, Younggeun Cho, Kahye Song, Wooyul Lee, Kwangmoo Koh, Hyunok Lee, Seungbum Rim, and much more. Many thanks also go to my Bible study group: Jungmin Ha, Junghoon Chun, Jieun Oh, Sehui Kim, Sooyeon Lee, Yongsuk Moon, Haekang Lee, and Donghyun Kim. My high school friends Chulhan Lee, Joonsoo Lee, Sehwan Park, Seunghoon Nam, Doh Chang Lee, Joonhyung Park, Sukwon Hong, Dae-Seok Nam, and Sihwa Jung are my true buddies that I can always count on.

Finally, I owe much to my family. I sincerely thank my parents and my younger brother, Jangsup, for their constant support and love throughout my life. The free and academic atmosphere in our family was what enabled me to persist in learning and studying till today.

Words are not enough to thank my beloved wife, Jeeyoung, for being a life companion with me. I am confident that the memory of our first encounter at Stanford and the joint struggle to survive our graduate studies together will be the firm foundation of our married life. I am excited and looking forward to facing the new chapter of our life journey.

For the last but not at all the least, I thank God for always being there even when I was not aware of and for making everything happen.
Contents

Abstract iv

Acknowledgments vi

1 Introduction 1
  1.1 Basic setting ................................................. 3
  1.2 Results and organization .................................... 5
    1.2.1 Universal filtering via hidden Markov modeling .......... 5
    1.2.2 Universal FIR MMSE filtering .......................... 5
    1.2.3 Discrete denoising with shifts ........................ 6

2 Universal filtering via hidden Markov modeling 8
  2.1 Introduction ............................................... 8
  2.2 Notation and preliminaries ................................ 11
    2.2.1 General notation ..................................... 11
    2.2.2 Hidden Markov processes (HMPs) ........................ 12
  2.3 The universal filtering problem ............................ 15
  2.4 Universal filtering based on hidden Markov modeling ...... 17
    2.4.1 Description of the filter .............................. 17
    2.4.2 Intuition behind the scheme and proof sketch .......... 19
    2.4.3 Proof of the theorem ................................. 20
  2.5 Extension: Universal filtering for channel with memory ... 35
  2.6 Discussion ................................................. 38
  2.7 Appendix ................................................... 39

viii
2.7.1 Three lemmas .................................................. 39
2.7.2 Proof of Lemma 3 .............................................. 44
2.7.3 Proof of Corollary 1 .......................................... 54

3 Universal FIR MMSE filtering .................................. 57
   3.1 Introduction .................................................. 57
   3.2 Problem formulation, Filter description, and Main result ........................................ 60
      3.2.1 Problem formulation .................................. 60
      3.2.2 Description of our filter .............................. 62
      3.2.3 Main result ............................................. 64
   3.3 Derivation of the universal filter ......................... 65
   3.4 Analysis ....................................................... 69
   3.5 Stochastic setting ............................................ 76
   3.6 Discussion .................................................... 77
      3.6.1 Algorithmic description ............................... 77
      3.6.2 Requirement of the knowledge on bounds of signal and noise ........... 78
      3.6.3 Comments on the expectation result .................. 78
      3.6.4 Competing with any finite order filter .......... 79
   3.7 Simulation results ........................................... 80
      3.7.1 Linear, stochastic signal ................................ 80
      3.7.2 Nonlinear, stochastic signal ......................... 82
      3.7.3 Universality of our filter ............................. 83
      3.7.4 Filtering deterministic signal ....................... 84
      3.7.5 Effect of constants on the convergence rate of regret .................. 85
   3.8 Appendix ....................................................... 86
      3.8.1 Proof of Lemma 10 .................................... 86
      3.8.2 Proof of Lemma 11 .................................... 87
      3.8.3 Proof of Lemma 12 .................................... 89
      3.8.4 Proof of Lemma 13 .................................... 91
      3.8.5 Proof of Lemma 14 .................................... 92
5 Contributions and future work

5.1 Contributions of the thesis ........................................ 144
5.2 Future work .......................................................... 145

Bibliography .............................................................. 147
List of Figures

1.1 Basic setting of the universal estimation problem .......................... 4

2.1 The time line ................................................................................. 55

3.1 MSEs for AR(1) signal (3.40). Figure 3.1(a) is for a single sample path, and Figure 3.1(b) is for the average of 100 experiments. ............ 94

3.2 MSEs for nonlinear signal (3.43). Figure 3.2(a) is for a single sample path, and Figure 3.2(b) is for the average of 100 experiments. .... 95

3.3 Average MSEs for signals (3.44) and (3.45). Kalman filter and the extended Kalman filter used here are matched to wrong signals (3.40) and (3.43), respectively. ................................................. 96

3.4 MSE results averaged over 100 experiments for Henon map (3.46). . 97

3.5 Regrets averaged over 100 experiments for nonlinear signal (3.43) with varying parameters. ......................................................... 98

4.1 400 x 400 binary images. .............................................................. 127

4.2 Clean and noisy images, the bit error rate plot for (k, m)-S-DUDE, and three denoised outputs for (k, m) = (4, 0), (4, 2), (6, 1), respectively. 128

4.3 BER for switching binary hidden Markov process (δ = 0.1, n = 10^6). The switch of the underlying binary Markov chain occurs when t = 5 \times 10^5, from the transition probability p = 0.01 to p = 0.2. ............. 130
Chapter 1

Introduction

Estimating clean source data from noisy observations is one of the central problems encountered in many science and engineering disciplines including information theory, signal processing, statistics, communications, computer science, and biology. The goodness of the estimation is usually measured by a given loss criterion per application, and a conventional approach to this estimation problem is to assume a certain model of the source and the noise that is represented by their joint probability distribution. In this approach, which is also known as the Bayesian approach, an optimum estimation strategy with respect to the assumed model and the loss criterion can readily be developed, e.g., the maximum a posteriori (MAP) estimator, to minimize the estimation loss. Moreover, when the assumed joint distribution of the source and the noise has a sufficiently simple structure, both optimal and computationally efficient schemes can be devised, for instance, the state estimation via dynamic programming for hidden Markov chains and Kalman filter for Gaussian signal and noise processes.

The Bayesian approach clearly makes sense when the source and the noise do coincide with the assumed model, or equivalently, when the estimator has clear knowledge on their joint distribution. As an extreme example, in communication systems, in which the clean source is designed via channel coding techniques, the optimum MAP estimator at the receiver can often lead to a perfect recovery of the clean source. However, the obvious drawback of the Bayesian approach occurs when there is an uncertainty in the assumed joint distribution. That is, in some applications, such as
image denoising or target tracking, where obtaining the source models is not straightforward, Bayesian approach-based estimators can assume mismatched source models rather than the true ones, and the optimum schemes with respect to the mismatched models can cause significant performance degradations.

This drawback stemming from the uncertainty of the model occurs not only in estimation problems, but also in many other problems, such as statistical modeling, data compression, prediction, investment, and learning problems. As a result, there have been several attempts to combat such uncertainties and drawbacks. Examples are the empirical Bayes method [1] and the nonparametric statistical methods [2] developed in statistics to estimate the probability distributions (or models) from the observations when there is no knowledge of the underlying models. Other examples include universal compression [3] and prediction [4] schemes in information theory that compress and predict data asymptotically optimally without assuming anything about the data model. The ideas behind the construction of these universal schemes have been extended to stock market investment [5] to devise an investment strategy that performs almost as well as the best constant rebalanced portfolio regardless of the stock sequence. Online learning techniques [6] that carry out typical goals of machine learning, e.g., classification or regression, in a sequential manner are additional examples of efforts to circumvent the shortcomings of mismatched model assumptions.

The recent work on universal filtering and denoising problems [7] [8], where the terms filtering and denoising will be concretely defined in the next subsection, are also in line with the efforts of circumventing the above mentioned drawback in estimation problems. In this thesis, we extend the frameworks developed in [7] and [8], and present novel and refined results for several different types of estimation problems. The resulting schemes do not assume any specific model about the underlying clean source data, and yet, as the data size grows, they are proven to achieve the optimum estimation performances, which will be specified concretely later for each problem context. The main gist of the schemes is that since the estimators do not have any prior knowledge about the source, they have to both learn about what kind of source data they are dealing with and estimate the actual source data at the same time, merely from the noisy data - the rationale behind the title of this thesis.
In the next subsection, we concretely describe the basic problem settings that will be dealt with in this thesis. Then, the exact definition of the terms *filtering* and *denoising* that we refer to will be presented. The following subsection contains the summary of the main results and the organization of the rest of the thesis.

### 1.1 Basic setting

One assumption that we make throughout this thesis is that the uncertainty lies only in the noise-free source characteristics. In other words, the noise mechanism, or the noisy channel model, is assumed to be completely known to the estimator as in [7]. One of the motivations for this assumption is to at least make the mission of achieving the optimum estimation performance seem not impossible. That is, when there are uncertainties in both the source and the noise models, it would be prohibitively difficult to achieve the performance of the optimum estimator that knows both the source and the noise models. The reason for the difficulty is because two different pairs of the source and the noise models can lead to the identical noisy data model, therefore, if only observing the noisy data, it would be impossible to ascertain which pair is indeed the correct one. Not only for the sanity checking purpose, another motivation for this assumption originates from practical scenarios where the noise tends to be more stationary than the source and is relatively easy to accurately estimate the model behind it before the estimation process begins. For example, suppose we want to make the noisy image obtained from a scanner as clean as possible. Since the noise injected by a scanner may be stationary over time, we can obtain the noise model by passing a blank image (or a training sequence) to the scanner several times and get a good estimate of the noise model before the actual estimation process begins. For these reasons, we assume that the estimator knows about the noisy channel model, but knows nothing about the model that generated the source.

With this assumption, Figure 1 depicts the basic setting of the $n$-th order universal estimation problem that is considered throughout the thesis. As shown in the figure, the $n$-th order estimation problem is to design the reconstruction sequence...
(\hat{X}_1, \cdots, \hat{X}_n) based on the noisy data \((Z_1, \cdots, Z_n)\) as close as possible to the clean source sequence \((X_1, \cdots, X_n)\), where the closeness is specified by the given loss criterion.\(^1\) Since the estimator is ignorant of the source model, the challenge of this problem is to learn the source from noisy data and to universally achieve the performance of the optimum estimator for any source data as the data size \(n\) becomes large. A more specific setting for each problem considered in this thesis will be presented in each subsequent chapter.

In designing the reconstruction sequence, we consider two types of estimation: causal and non-causal. In causal estimation, which we refer to as filtering, the reconstruction at \(t\) should strictly rely on the observations of the noisy data with no delay, i.e.,

\[ \hat{X}_t = \hat{X}_t(Z_1, \cdots, Z_t) \text{ for } 1 \leq t \leq n. \]

On the other hand, in non-causal estimation, which we refer to as denoising, the reconstruction at \(t\) can depend on the whole noisy data, i.e.,

\[ \hat{X}_t = \hat{X}_t(Z_1, \cdots, Z_n) \text{ for } 1 \leq t \leq n. \]

The next subsection summarizes the main results of this thesis, which will be presented in the following chapters.

--

\(^1\)In Chapter 3, we denote \(\{Y_t\}_{t \geq 1}\) as the noisy data instead of \(\{Z_t\}_{t \geq 1}\) to comply with the conventional notations of the signal processing literatures.
CHAPTER 1. INTRODUCTION

1.2 Results and organization

The main results of this thesis consist of three parts. The following three chapters will focus on each topic in an independent manner. Each chapter will contain its own introduction of the problem, preliminaries and notation, and presentation and proofs of the results. The notation will be consistent within each chapter. A brief summary of each chapter is given in the following subsections. Chapter 5 will conclude with a summary of the contributions of the thesis and some future directions.

1.2.1 Universal filtering via hidden Markov modeling

In Chapter 2, the problem of universal filtering for finite-alphabet signals, in which the known noisy channel is assumed to be memoryless, is considered. A family of filters are derived, and are shown to be universally asymptotically optimal in the sense of achieving the optimum filtering performance whenever the clean signal is stationary, ergodic, and satisfies an additional mild positivity condition. Our schemes are comprised of approximating the noisy signal using a hidden Markov process (HMP) via maximum-likelihood (ML) estimation, followed by the use of the forward recursions for HMP state estimation. It is demonstrated that as the data length increases, and as the number of states in the HMP approximation increases, our family of filters attains the performance of the optimal distribution-dependent filter. An extension to the case of channels with memory is also established.

The result in Chapter 2 has primarily the theoretical value since the applications of HMPs in filtering problems are already popular in practice, but the heuristics lack any theoretical justifications. The chapter presents the first theoretical justification of the heuristic schemes, and the result also appears in a paper [9].

1.2.2 Universal FIR MMSE filtering

Chapter 3 considers a different universal filtering problem, namely, the case in which the source is real-valued signal and is corrupted by a memoryless, zero mean, real-valued additive noise channel. We assume only that the variance of the noise is
known to the filter. When the estimation loss is measured by the mean-squared error (MSE) criterion, we build a computationally efficient filtering scheme whose per-symbol squared error is essentially as small as that of the best finite-duration impulse response (FIR) filter of a given order, for every bounded underlying signal. In contrast to in Chapter 2, we do not assume any stochastic mechanism generating the underlying source and treat it as an individual sequence. The regret of the expected MSE of our scheme is shown to decay as $O\left(\frac{\log n}{n}\right)$, where $n$ is the length of the signal. Moreover, we present a stronger concentration result which guarantees the performance of our scheme not only in expectation, but also with high probability. Our result implies a conventional stochastic setting result, i.e., when the underlying signal is a stationary process, our filter achieves the performance of the optimal FIR filter, also known as Wiener filter.

The explicit decay rate of the regret cannot be attained in a straightforward manner from previous results in the literature; as a result, we combine techniques developed in information theory and online learning to prove it. Moreover, our experimental results showcase the potential merits of our universal filter in practice.

The result of Chapter 3 also appears in a forthcoming paper [10].

1.2.3 Discrete denoising with shifts

We then shift gears to denoising problem in Chapter 4. We again consider a finite-alphabet setting as in Chapter 2 and generalize the recently introduced DUDE (Discrete Universal DEnoiser) [7]. We introduce a new algorithm S-DUDE, which aims to compete with a genie that has access, in addition to the noisy data, to the underlying clean data, and can choose to switch, up to $m$ times, between sliding window denoisers in a way that minimizes the overall loss. When the underlying data form an individual sequence, we show that the S-DUDE performs essentially as well as this genie, provided that $m$ is sub-linear in the size of the data. When the clean data are emitted by a piecewise stationary process, we show that the S-DUDE achieves the optimum distribution-dependent performance, provided that the same sub-linearity
CHAPTER 1. INTRODUCTION

condition is imposed on the number of switches. To further substantiate the universal optimality of the S-DUDE, we demonstrate that when the number of switches is allowed to grow linearly with the size of the data, any (sequence of) scheme(s) fails to compete in the above senses.

We not only offer above theoretical performance guarantees, but also derive an efficient implementation algorithm of S-DUDE. That is, by using dynamic programming, our algorithm has complexity (time and memory) growing only linearly with the data size and the number of switches $m$. We also present preliminary experimental results, suggesting that S-DUDE has the capacity to significantly improve on the performance attained by the original DUDE in applications where the nature of the data abruptly changes in time (or space), as is often the case in practice.

The result of Chapter 4 also appears in a forthcoming paper [11].
Chapter 2

Universal filtering via hidden Markov modeling

2.1 Introduction

In this chapter, we focus on the universal filtering problem of estimating an unknown discrete-time, finite-alphabet source \( \{X_t\}_{t=1}^n \) from the causal observation of a noisy signal \( \{Z_t\}_{t=1}^n \), which has been corrupted by a known discrete memoryless channel (DMC). This problem has been considered in [12][8], but we revisit this case here by taking a different approach.

We will only focus on the stochastic setting, where we assume \( \{X_t\} \) is a stationary and ergodic stochastic process. With the stochastic setting assumption, and under the same performance criterion of [7], i.e., minimizing the expected normalized cumulative loss, knowledge of the conditional distribution of \( X_t \) given \( Z_t \) at each time \( t \) is required to achieve the optimal performance. Also, by the same argument as in [7, Section III], this conditional distribution can be obtained by the conditional distribution of \( Z_t \) given \( Z^{t-1} \) when the invertible DMC is known. (We call a channel is "invertible" if its transition probability matrix is of full row rank.)

However, for the universal filtering setting, where the probability distribution of the source is unknown, the conditional distribution of \( Z_t \) given \( Z^{t-1} \) is also not known and need be learned from the observed noisy signal. Therefore, if we can learn
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

this conditional distribution accurately as the observation length increases, we can hope to build the universal filtering scheme that achieves the asymptotically optimal performance from the estimated conditional distribution. To pursue this goal, [12][8] adopt the universal prediction [13] approach. That is, they first get an estimate of the conditional distribution of $Z_t$ given $Z_t^{-1}$ by employing a universal predictor for the observed noisy signal, and then by inverting the known DMC, obtain an estimate of the conditional distribution of $X_t$ given $Z_t$.

Unlike the approach of [12][8], in this work, we turn our attention to the rich theory of hidden Markov process (HMP) models to directly obtain a different kind of estimate of the conditional distribution of $X_t$ given $Z_t$, without going through the channel inversion stage.

Generally, HMPs are defined as a family of stochastic processes that are outputs of a memoryless channel whose inputs are finite state Markov chains. As can be seen in [14], these HMP models arise in many areas, such as information theory, communications, statistics, learning, and speech recognition. Among these applications of HMPs, there are many situations where the state of the underlying Markov chain need be estimated based on the observed hidden Markov process. If the exact parameters of the HMP, namely, the state transition probability of the Markov chain and the channel transition density, as well as the order, the number of states, of the Markov chain are known, then this problem can be easily solved via well-known forward-backward recursions which were discovered by [15] and [16]. Especially, when we are estimating the state based on the causal observation of the HMP, we only need the forward recursion formula. In addition, much work has been done for the state estimation, where the order is known, but the parameters of the HMP are unknown. In this case, the parameters are first estimated via maximum likelihood (ML) estimation or the EM algorithm, then the state is estimated by using the estimated parameters in the recursion formula. A detailed explanation of this approach and the property of the ML parameter estimation can be found in [16][17][18][19]. Furthermore, this was extended to the case where the order of the Markov chain is also not known, but the upper bound on the order is known. In this case, the order estimation is first performed before the parameter and state estimation, and the above process is
repeated. The references for the order estimation are given in [20][21][22]. There also has been work for the case where even the knowledge of the upper bound on the order of the Markov chain is not required [19][23].

From these rich theories for the state, parameter, and order estimation of HMPs, we can see that it is possible to build a universal filtering scheme if the underlying source process is known to be a Markov process. That is, since the channel is memoryless and fixed in our setting, if our source \( \{X_t\} \) is a finite order Markov process, then obviously, \( \{Z_t\} \) is a HMP, and we can first estimate the order of the Markov process, then estimate the parameter, and finally perform forward recursion to learn the conditional distribution of \( X_t \) given \( Z_t \). From the consistency results of order estimation and parameter estimation, this conditional distribution will be an accurate estimate of the true one, and we can use it to build the universal filtering scheme.

Now, in our work, we extend this approach to the case where our source \( \{X_t\} \) is a general stationary and ergodic process (with some benign conditions), which need not be a Markov process at all, and show that we can still build a universal filtering scheme that achieves asymptotically optimal performance. The skeleton of our scheme is the following: We first "model" our source as a Markov process with a certain order, or equivalently, model the noisy observed signal \( \{Z_t\} \) as a HMP in a certain class. Then, we estimate the parameters of the HMP that "approximates" the noisy signal best in that class. We will show that from the consistency result about the ML parameter estimation for the mismatched model [19], these estimated parameters will give an accurate estimation of the conditional distribution of \( X_t \) given \( Z_t \), as the observation length increases and the HMP class gets richer. Then, this result will guarantee that our universal filter using this conditional distribution will attain the asymptotically optimal performance. In practice, this approach of HMP modeling has been heuristically employed in many applications, such as speech recognition [24], target tracking [25], and DNA sequence analysis [26], without theoretical justification. Additional samples of these practical applications can be found in [27]. Therefore,

\[1\] We slightly abuse the term 'order' here. Generally, the order of a 'finite state Markov chain' stands for the number of states, but we also refer by the 'order of a Markov process' to the length of the memory of the process. Hence, once we know the order of a Markov process, we also know the order of the associated finite state Markov chain induced from the Markov process.
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

we focus on pursuing the rationale of the existing practical methodologies, and the main contribution of this work is in providing theoretical justification of the HMP modeling based approach to universal filtering.

The remainder of the chapter is organized as follows. Section 2.2 introduces some notation and preliminaries that are needed for setting up the problem. In Section 2.3, the universal filtering problem is defined explicitly. In Section 2.4, our universal filtering scheme is devised, the main theorem is stated, and proved. Section 2.5 extends our approach to the case where the channel has memory, and Section 2.6 gives discussions on our filter. Detailed technical proofs that are needed in the course of proving our main results are given in Section 2.7, the Appendix.

2.2 Notation and preliminaries

2.2.1 General notation

We assume that the clean, noisy and reconstruction signal components take their values in the same finite \( M \)-ary alphabet \( \mathcal{A} = \{0, \cdots, M-1\} \). The simplex of \( M \)-dimensional column probability vectors will be denoted as \( \mathcal{M} \).

The DMC is known to the filter and is denoted by its transition probability matrix \( \Pi = \{\Pi(i,j)\}_{i,j \in \mathcal{A}} \). Here, \( \Pi(i,j) \) denotes the probability of channel output symbol \( j \) when the input is \( i \). We assume \( \Pi(i,j) > 0 \ \forall i, j \), and let \( \Pi_{\text{min}} = \min_{i,j} \Pi(i,j) \). We assume this channel matrix is invertible and denote the inverse as \( \Pi^{-1} \). Let \( \Pi^{-1}_i \) denote the \( i \)-th column of \( \Pi^{-1} \). We also assume a given loss function (fidelity criterion) \( \Lambda : \mathcal{A}^2 \rightarrow [0, \infty) \), represented by the loss matrix \( \Lambda = \{\Lambda(i,j)\}_{i,j \in \mathcal{A}} \), where \( \Lambda(i,j) \) denotes the loss incurred when estimating the symbol \( i \) with the symbol \( j \). The maximum single-letter loss will be denoted by \( \Lambda_{\text{max}} = \max_{i,j \in \mathcal{A}} \Lambda(i,j) \), and \( \lambda_j \) will denote the \( j \)-th column of \( \Lambda \).

As in [7], we define the extended Bayes response associated with the loss matrix \( \Lambda \) to any column vector \( V \in \mathbb{R}^M \) as

\[
B(V) = \arg \min_{x \in \mathcal{A}} \lambda_x^T V,
\]
where \( \text{arg min}_{a \in A} \) denotes the minimizing argument, resolving ties by taking the letter in the alphabet with the lowest index.

We let \( P \) denote the true joint probability law of the clean and noisy signal, and \( E(\cdot) \) denote expectation with respect to \( P \). Throughout the chapter, every almost sure convergence is with respect to \( P \), and any equalities or inequalities between random variables should be understood in almost sure sense. If we need to refer to the probability law of clean or noisy signal induced by \( P \), we denote \( P_X \) and \( P_Z \), respectively. If \( P \) is written in a bold face, \( \mathbf{P} \), with a subscript, it stands for a simplex vector in \( \mathcal{M} \) for the corresponding distribution of the subscript. For example, \( \mathbf{P}_{X_t|Z^t} \) is a column \( M \)-vector whose \( i \)-th component is \( P(X_t = i|Z^t = z^t) \).

When we have some other probability law denoted as \( Q \), and want to measure its difference from \( P \), a natural choice of such a measure is the relative entropy rate. First, denote the \( n \)-th order relative entropy between \( P \) and \( Q \) as

\[
D_n(P||Q) = \sum_{z^n} P(z^n) \log \frac{P(z^n)}{Q(z^n)} = E\left( \log \frac{P(Z^n)}{Q(Z^n)} \right).
\]

Then, the relative entropy rate (also known as Kullback-Leibler divergence rate) is defined as

\[
\mathbf{D}(P||Q) \triangleq \lim_{n \to \infty} \frac{1}{n} D_n(P||Q),
\]

if the limit exists. When \( Q \) is a probability law in a certain class of HMPs, this limit always exists and the relative entropy rate is well defined. A more detailed discussion about this limit will be given in Lemma 2. This relative entropy rate will play a central role in analyzing our universal filtering scheme.

### 2.2.2 Hidden Markov processes (HMPs)

**Definition**

As stated in Section 2.1, the HMPs are defined as a family of stochastic processes that are outputs of a memoryless channel whose inputs are finite state Markov chains. Let us denote a general HMP as \( \{Y_t\} \) and the underlying finite state Markov chain as \( \{S_t\} \). The corresponding alphabet sizes of each component are denoted as \( |Y| \) and \( |S| \),
respectively. Then, there are three parameters that determine the probability laws
of \( \{Y_t\} \): \( \pi \in \mathbb{R}^{1 \times |S|} \), the initial distribution of finite state Markov chain; \( A \in \mathbb{R}^{|S| \times |S|} \),
the probability transition matrix of finite state Markov chain, and \( C \in \mathbb{R}^{|S| \times |Y|} \),
the probability transition matrix of the memoryless channel. The triplet \( \{\pi, A, C\} \) is
referred to as the parameter of HMP. Let \( \Theta \) be a set of all \( \theta \)'s where \( \theta := \{\pi_\theta, A_\theta, C_\theta\} \).
For each \( \theta \) and each realization \( y^n \), we can calculate the likelihood function
\[
Q_\theta(y^n) = \pi_\theta \prod_{t=1}^{n} (C_\theta y_t A_\theta) 1,
\]
where \( C_\theta y_t \) is \( |S| \times |S| \) diagonal matrix whose \( (j, j) \)-th entry is the \( (j, y_t) \)-th entry of
\( C_\theta \), and \( 1 \) is the \( |S| \times 1 \) vector with all entries equal to 1.

Now, suppose \( \{Z_t\} \) is an output of a DMC \( \Pi \), when the input is a stationary
Markov process \( \{X_t\} \). For simplicity, we assume that the alphabet of each component
of \( \{Z_t\} \) and \( \{X_t\} \) are finite and equal, i.e., \( Z = X = A \). When the order of the
underlying Markov process is \( k \), we can associate the state \( S_t \) of underlying finite
state Markov chain with \( X_{t-k}^{t-1} \), which has alphabet size \( M^k \). Then, clearly, \( \{Z_t\} \) is
also a stationary HMP, and the parameter set of such HMP is denoted as \( \Theta_k \subset \Theta \).
Note that when \( \theta \in \Theta_k, \pi_\theta \in \mathbb{R}^{1 \times M^k}, A_\theta \in \mathbb{R}^{M^k \times M^k}, \) and \( C_\theta \in \mathbb{R}^{M^k \times M} \).
Furthermore, for some \( \delta > 0 \), we define a set \( \Theta_k^\delta \subset \Theta_k \) as the set of \( \theta \in \Theta_k \) that has following
properties: for the \( i \)-th \( k \)-tuple state \( x_k^i(i) \) and the \( j \)-th \( k \)-tuple state \( x_k^j(j) \),

- \( a_{i,j,\theta} \geq \delta, \) if \( x_k^i(i) \) is equal to \( x_k^{i-1}(j) \)
- \( a_{i,j,\theta} = 0, \) otherwise
- \( c_{i,r,\theta} = \Pi(x_k(i), r), \) for all \( 1 \leq i \leq M^k \) and \( 1 \leq r \leq M, \)

where \( a_{i,j,\theta} \) is \( (i, j) \)-th entry of \( A_\theta \), and \( c_{i,r,\theta} \) is \( (i, r) \)-th entry of \( C_\theta \). In particular, if
\( \theta \in \Theta_k^\delta \) then: 1) the stochastic matrix \( A_\theta \) is irreducible and aperiodic; thus, since
the Markov chain is stationary, \( \pi_\theta \) is the stationary distribution of the Markov chain,
and is uniquely determined from \( A_\theta \), 2) \( C_\theta \) is the same for all \( \theta \), and, therefore, \( \theta \)
is completely specified by \( A_\theta \). For notational brevity, we omit the subscript \( \theta \) and
denote the probability law \( Q \in \Theta_k^\delta \), if \( Q = Q_\theta \), and \( \theta \in \Theta_k^\delta \).
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

Maximum likelihood (ML) estimation

Generally, suppose a probability law $Q$ is in a certain class $\Omega$. Then, the $n$-th order maximum likelihood (ML) estimator in $\Omega$ for the observed sequence $z^n$, is defined as

$$\hat{Q}[z^n] = \arg\max_{Q \in \Omega} Q(z^n),$$

resolving ties arbitrarily. Now, if $Q \in \Theta^d_k$, then there is an algorithm called expectation-maximization (EM) [4] that iteratively updates the parameter estimates to maximize the likelihood. Thus, when $Q$ is in the class of probability laws of a HMP, the maximum likelihood estimate can be efficiently attained.\footnote{We neglect issues of convergence of the EM algorithm and assume that the ML estimation is performed perfectly.} We denote the ML estimator in $\Theta^d_k$ based on $z^n$ by

$$\hat{Q}_{k,\delta}[z^n] = \arg\max_{Q \in \Theta^d_k} Q(z^n).$$

Obviously, when the $n$-tuple $Z^n$ is random, $\hat{Q}_{k,\delta}[Z^n]$ is also a random probability law that is a function of $Z^n$.

Consistency of ML estimator

When $P_Z \in \Theta^d_k$, an ML estimator $\hat{Q}_{k,\delta}[z^n]$ is said to be strongly consistent if

$$\lim_{n \to \infty} \hat{Q}_{k,\delta}[Z^n] = P_Z \quad a.s.$$ 

The strong consistency of the ML estimator $\hat{Q}_{k,\delta}[Z^n]$ of the parameter of a finite-alphabet stationary ergodic HMP was proved in [28]. For the case of a general stationary ergodic HMP, the strong consistency was proved in [18].

We also have a sense of strong consistency for the case where $P_Z$ is a general stationary and ergodic process. By the similar argument as in [19, Theorem 2.2.1], we have the consistency in the sense that if the observed noisy signal is not necessarily
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

a HMP, and we still perform the ML estimation in $\Theta_k$, then we get

$$\lim_{n \to \infty} \hat{Q}_{k,n}[Z^n] \in \mathcal{N} \quad \text{a.s.,}$$

where $\mathcal{N} \triangleq \{Q \in \Theta_k : D(P\|Q) = \min_{Q' \in \Theta_k} D(P\|Q')\}$.\(^3\) This second consistency result is the key result that we will use in devising and analyzing our universal filtering scheme.

2.3 The universal filtering problem

Here, we describe more concretely about the universal filtering problem that is outlined in Chapter 1. As mentioned in Section 2.1, in this chapter, we focus on a stochastic setting, that is, the underlying clean signal is an output of some stationary and ergodic process whose probability law is $P_X$. From $P_X$ and $\Pi$, we can get the true joint probability law $P$ and corresponding probability law of noisy observed signal, $P_Z$. That is,

$$P(X^n = x^n, Z^n = z^n) = P_X(X^n = x^n) \prod_{i=1}^{n} \Pi(x_i, z_i), \quad \text{and} \quad P_Z(Z^n = z^n) = \sum_{x^n} P(X^n = x^n, Z^n = z^n).$$

A filter is a sequence of probability distributions $\hat{X} = \{\hat{X}_t\}$, where $\hat{X}_t : \mathcal{A}^t \to \mathcal{M}$. The interpretation is that, upon observing $z^t$, the reconstruction for the underlying, unobserved $x_t$ is represented by the symbol $\hat{x}$ with probability $\hat{X}_t(z^t)[\hat{x}]$. The notation $\hat{X}_t(z^t)[\hat{x}]$ stands for the $\hat{x}$-th element of a vector $\hat{X}_t(z^t)$. A filter is called deterministic if $\hat{X}_t(z^t)$ is a standard basis vector in $\mathbb{R}^M$ for all $t$ and $z^t$, and randomized if $\hat{X}_t(z^t)$ can be any vectors in $\mathcal{M}$ other than standard basis vectors for some $t$ and $z^t$. The normalized cumulative loss of the scheme $\hat{X}$ on the individual pair $(x^n, z^n)$ is defined

\(^3\)Just as in [19, Theorem 2.2.1], the notion of a.s. set convergence is used. For any subset $\mathcal{E} \subset \Theta$, define $\mathcal{E}_\epsilon \triangleq \{Q \in \Theta : d(Q, \mathcal{E}) < \epsilon\}$, where $d$ is the Euclidean distance. Then, $\lim_{n \to \infty} \hat{Q}[Z^n] \in \mathcal{E}$ a.s. if for all $\epsilon > 0$, exists $N(\epsilon, \omega)$ such that for all $n \geq N(\epsilon, \omega), \hat{Q}_n[Z^n] \in \mathcal{E}_\epsilon$. 


by
\[ L_X(x^n, z^n) = \frac{1}{n} \sum_{t=1}^{n} \ell(x_t, \hat{X}_t(z^t)), \]
where \( \ell(x_t, \hat{X}_t(z^t)) = \sum_{\hat{x} \in \mathcal{X}} \Lambda(x_t, \hat{x}) \hat{X}_t(z^t)[\hat{x}] \). Then, the goal of a filter is to minimize the expected normalized cumulative loss \( E(L_X(X^n, Z^n)) \).

The optimal performance of the \( n \)-th order filter is defined as
\[ \phi_n(P_X, \Pi) = \min_{X \in \mathcal{F}} E\left( L_X(X^n, Z^n) \right), \]
where \( \mathcal{F} \) denotes the class of all filters. Sub-additivity arguments similar to those in [7] imply
\[ \lim_{n \to \infty} \phi_n(P_X, \Pi) = \inf_{n \in \mathbb{N}} \phi_n(P_X, \Pi) = \Phi(P_X, \Pi). \]

By definition, \( \Phi(P_X, \Pi) \) is the (distribution-dependent) optimal asymptotic filtering performance attainable when the clean signal is generated by the law \( P_X \) and corrupted by \( \Pi \). This \( \Phi(P_X, \Pi) \) can be achieved by the optimal filter \( \hat{X}_P = \{\hat{X}_{P,t}\} \)
where
\[ \hat{X}_{P,t}(z^t)[\hat{x}] = \Pr(B(P_{X_t|z^t}) = \hat{x}). \]
For brevity of notation, we denote \( \hat{X}_P(z^t) = \hat{X}_{P,t}(z^t) \). Note that this is a deterministic filter, i.e., for a given \( z^t \), the filter is a standard basis vector in \( \mathbb{R}^M \) for all \( t \). We can easily see that this filter is optimal since it minimizes \( E(\ell(X_t, \hat{X}(Z^t))) \) for all \( t \), and thus, it minimizes \( E(L_X(X^n, Z^n)) \) for all \( n \).

As can be seen, \( \hat{X}_P(z^t) \) needs the exact knowledge of \( P_{X_t|z^t} \), and thus, is dependent on the distribution of the underlying clean signal. The universal filtering problem is to construct (possibly a sequence of) filter(s), \( \hat{X}_{\text{univ}} \), that is independent of the distribution of underlying clean signal, \( P_X \), and yet asymptotically achieving \( \Phi(P_X, \Pi) \). We describe our sequence of universal filters in the next section.
2.4 Universal filtering based on hidden Markov modeling

2.4.1 Description of the filter

Before describing our sequence of universal filters, we make the following assumption on the source.

**Assumption 1** There exists a sequence of positive reals \( \{\delta_k\} \), such that \( \delta_k \downarrow 0 \) as \( k \to \infty \), and \( P_X \) satisfies

\[
P_X(X_0|X_{-k}^{-1}) \geq \delta_k \quad \text{a.s.} \quad \forall k \in \mathbb{N}.
\] (2.2)

For any probability law \( Q \), we construct a randomized filter as follows: For \( \epsilon > 0 \), denote \( L_2 \epsilon \)-ball in \( \mathbb{R}^M \) as 

\[
B_\epsilon = \{ V \in \mathbb{R}^M : \|V\|_2 \leq \epsilon \}
\]

Then, we define a filter for fixed \( \epsilon \) as

\[
\hat{X}_{Q,t}^\epsilon(z^t)[\hat{x}] = \Pr(B(Q_{X_t|z^t} + U) = \hat{x}),
\] (2.3)

where \( U \in \mathbb{R}^M \) is a random vector, uniformly distributed in \( B_\epsilon \). For brevity of notation, we denote \( X_{t}^Q(z^t) = \hat{X}_{Q,t}^\epsilon(z^t) \). This filter is randomized since depending on \( Q \) and \( z^t \), \( \hat{X}_{Q}^\epsilon(z^t) \) can be a probability simplex vector in \( \mathcal{M} \) that is not a standard basis vector. The reason we needed this randomization will be explained in proving Lemma 3.

To devise our filter, let us first consider an increasing sequence of positive integers, \( \{m_i\}_{i \geq 1} \), that satisfies following conditions:

\[
\lim_{i \to \infty} \frac{m_{i-1}}{m_i} = 0, \quad \lim_{i \to \infty} m_i = \infty.
\] (2.4)

Now, define

\[
i(t) \triangleq \max\{i : m_i \leq t\}.
\]
Then, given that our source distribution satisfies (2.2), and for fixed $k$, define a random probability law

$$Q_k^t \triangleq \hat{Q}_{k,t},\delta_k[Z] = \arg \max_{Q \in \Theta_k^t} Q(Z).$$

That is, $Q_k^t$ is the ML estimator in $\Theta_k^t$ based on $Z$. As discussed in Section 2.2.2, we only need to estimate the state transition probabilities of the underlying Markov chain to obtain this ML estimator, and this can be done by the Expectation-Maximization (EM) algorithm. Performing the EM algorithm requires iterative forward-backward recursions, and is the most expensive part of our scheme in terms of the complexity. However, since the recursions are efficiently implemented by linear complexity dynamic programming (a.k.a. BCJR algorithm), which is described in detail in [4], the overall complexity of our scheme is still linear in data length. Once we get $Q_k^t$, we can then calculate $Q_{k,t}^t$, which stands for the simplex vector in $\mathcal{M}$ whose $i$-th component is $Q_k^t(X_t = i|Z^t = z^t)$. This vector can be obtained again from using the forward-recursion formula. Note that we get this conditional distribution directly, not by first estimating the output distribution, and then inverting the channel, as was done in [12][8][7].

Finally, we take as our sequence of universal filtering schemes, indexed by $k$ and $\epsilon$,

$$\hat{X}_{\text{univ},k}^\epsilon = \{\hat{X}_{Q_k^t,t}^\epsilon\}.$$  

The following theorem states the main result of this chapter.

**Theorem 1** Let $X^\infty \in \mathcal{A}^\infty$ be a stationary, ergodic process emitted by the source $P_X$ which satisfies Assumption 1. Let $Z^\infty \in \mathcal{A}^\infty$ be the output of the DMC, $\Pi$, whose input is $X^\infty$. Then:

(a) \quad \lim_{\epsilon \to 0} \lim_{k \to \infty} \limsup_{n \to \infty} L_{\hat{X}_{\text{univ},k}^\epsilon} (X^n, Z^n) \leq \Phi(P_X, \Pi) \quad \text{a.s.}

(b) \quad \lim_{\epsilon \to 0} \lim_{k \to \infty} \limsup_{n \to \infty} E\left(L_{\hat{X}_{\text{univ},k}^\epsilon} (X^n, Z^n)\right) = \Phi(P_X, \Pi)
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

Remark: In defining our universal filter $\hat{X}_{\text{uni},k}^*$, one might intuitively think that it would be better to use the ML estimator that updates every point of time, i.e., to define $Q_k^* = \hat{Q}_{k,\delta_k}[Z^n]$. However, our definition of $Q_k^*$, namely, using the same ML estimator throughout each block, is crucial for our proof of the above theorem, especially for proving Corollary 1 that follows below. Besides this technical reason, updating the ML estimator every time would require higher complexity than our scheme requires.

2.4.2 Intuition behind the scheme and proof sketch

The intuition behind our scheme parallels that of the universal compression and universal prediction problems in the stochastic setting. In the $n$-th order problem of both cases [29][4], the excess expected codeword length per symbol and the excess expected normalized cumulative loss incurred by using the wrong probability law $Q$ in place of the true probability law $P$ could be upper bounded by the normalized $n$-th order relative entropy $\frac{1}{n}D_n(P\|Q)$. Then, to achieve the asymptotically optimum performance, the compressor and the predictor try to find and use some data-dependent $Q$ that makes $\frac{1}{n}D_n(P\|Q) \to 0$ as $n \to \infty$, that is, makes $D(P\|Q)$ zero.

We follow the same intuition in our universal filtering problem. For fixed $k$ and $\epsilon$, our scheme, as can be seen from (2.5), divides the noisy observed signal into sub-blocks of length $(m_i - m_{i-1})$. Since $\frac{m_{i-1}}{m_i}$ tends to zero as $i \to \infty$, the length of each sub-block grows faster than exponential. Now, to filter each sub-block, it plugs the ML estimator in $\Theta_k^{\delta_k}$ obtained from the entire observation of noisy signal up to the previous sub-block. From (2.1), we know that as the observation length $n$ increases, this ML estimator will converge to the parameter that minimizes the relative entropy rate between the true output probability law $P_Z$. Then, to show that this scheme achieves the asymptotically optimum performance, we bound the excess expected normalized cumulative loss with this relative entropy rate, and show that the bound goes to zero as the HMP parameter set becomes richer, that is, $k$ increases.

To be more specific, we briefly sketch the proof of our main theorem. Part (b) of Theorem 1 states that our scheme is asymptotically optimal. As described in the proof
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

of Theorem 1, it is not hard to show that Part (b) follows directly from Part (a) and Fatou's Lemma. Therefore, proving Part (a) is the key in proving the theorem. Part (a) states that in the limit, the normalized cumulative loss of our scheme, for almost every realization, is less than or equal to the asymptotically optimum performance.

To prove Part (a), we first fix $k$ and $\epsilon$, and get the following inequality

$$
\limsup_{n \to \infty} \left( L^{X_{\text{univ}}, k}_{n} (X^n, Z^n) - \phi_n(P_X, \Pi) \right) 
\leq F \left( \limsup_{t \to \infty} D(P_Z || Q^t_k), \epsilon \right) \quad \text{a.s.,}
$$

where $F(x, y)$ is some function such that $F(x, y) \to 0$ as $x \downarrow 0$, and then $y \downarrow 0$.\footnote{Note that $Q^t_k$ in $D(P_Z || Q^t_k)$ is a function of $Z^{m(t)}$, and thus, is random. A more formal definition of relative entropy rate between true and the random probability law like this case will be given after Lemma 4.} There are two keys in getting this inequality. The first one is to show the concentration of $L^{X_{\text{univ}}, k}_{n} (X^n, Z^n)$ to its expectation which will be shown in Lemma 3 and Corollary 1. The second is to get the explicit upper bound function $F(x, y)$ which will be based on Lemma 4. Once establishing this inequality, we show that

$$
\lim_{k \to \infty} \limsup_{t \to \infty} D(P_Z || Q^t_k) = 0 \quad \text{a.s.,}
$$

from Lemma 5 and then send $\epsilon \downarrow 0$ to get Part (a). Keeping this proof sketch in mind, let us move on to the detailed proof in the next section.

2.4.3 Proof of the theorem

Before proving the theorem, we introduce several lemmas as building blocks. Lemma 1 and Lemma 2 below give some general results for the HMPs that we are considering. Our lemmas are similar to [19, Lemma 2.3.4] and [19, Theorem 2.3.3]. The latter assumed that all the parameters are lower bounded by $\delta > 0$, whereas in $\Theta^t_k$, some parameters can be zero. We take this into account in proving Lemma 1 and Lemma 2. Lemma 3 shows the uniform concentration property of the normalized cumulative loss on $\Theta^t_k$, which is an important property that we need to prove the main theorem.
Lemma 1 Suppose \( Q \in \Theta_k^\delta \) and fix \( \delta > 0 \). Then, for all \( \omega \), \( Q(Z_0|Z_{-1}^{-1}) \) converges to a limit \( Q(Z_0|Z_{-\infty}^{-1}) \) uniformly on \( \Theta_k^\delta \).

Proof: To prove this lemma, we need three more lemmas in Appendix 1, which are variations on those found in [28]. Let us denote \( f_t := Q(Z_0|Z_{-1}^{-1}) \), and \( f_0 = 0 \). Then, the sequence \( \{f_t\} \) uniformly converges on \( \Theta_k^\delta \), if following \( k \) subsequences,

\[
\{f_{jk+l}, j = 0, 1, 2, \ldots \}, \quad 0 \leq l \leq k - 1,
\]

uniformly converge on \( \Theta_k^\delta \), and have the same limit.

First, the uniform convergence of each subsequence \( \{f_{jk+l}\} \) can be shown by showing the series \( \sum_{j=0}^t (f_{(j+1)k+l} - f_{jk+l}) \) converges uniformly. From Lemma 8 in Appendix 1, and setting \( m = k \),

\[
\sum_{j=0}^t |f_{(j+1)k+l} - f_{jk+l}|
\]

\[
= \sum_{x_0} Q(Z_0|x_0) \sum_{j=1}^t |Q(x_0|Z_{-(j+1)k-l}) - Q(x_0|Z_{-jk-l})|
\]

\[
\leq M \sum_{j=1}^t (\rho_{\delta,k,k})^{j+1},
\]

where \( \rho_{\delta,k,k} < 1, M < \infty \) and \( \rho_{\delta,k,k} \) does not depend on \( Q, \omega, \) and \( l \). Therefore, the series \( \sum_{j=0}^t (f_{(j+1)k+l} - f_{jk+l}) \) converges absolutely regardless of \( Q \), and, hence, we conclude that each subsequences \( \{f_{jk+l}\} \) converges uniformly on \( \Theta_k^\delta \).

Now, to show that the \( k \) subsequences have the same limit, construct another subsequence, \( \{f_{j(k+1)+l}, j = 0, 1, 2, \ldots \} \). Since this subsequence contains infinitely many terms from all \( k \) subsequences, if this subsequence converges uniformly on \( \Theta_k^\delta \), we can conclude that the \( k \) subsequences have the same limit. The derivation of the
uniform convergence of this subsequence is the same as that described above, but setting \( m = k + 1 \) in Lemma 8. Therefore, the original sequence \( \{f_i\} \) converges to its limit uniformly on \( \Theta_k^\delta \). ■

The remarkable fact of this lemma is that the convergence is not only uniform on \( \Theta_k^\delta \), but also in \( \omega \). That is, the convergence holds uniformly on every realization of \( z_0^{\infty} \).

**Lemma 2** For the distribution of the observed noisy process \( \{Z_t\} \), \( P_Z \), and every \( Q \in \Theta_k^\delta \),

\[
D(P_Z||Q) \triangleq \lim_{n \to \infty} \frac{1}{n} D_n(P_Z||Q) = E\left( \log \frac{P_Z(Z_0|Z_{-1}^{\infty})}{Q(Z_0|Z_{-1}^{\infty})} \right).
\]

Moreover, uniformly on \( \Theta_k^\delta \),

\[
\lim_{n \to \infty} \frac{1}{n} \log \frac{P_Z(Z^n)}{Q(Z^n)} = D(P_Z||Q) \quad \text{a.s.}
\]

*Proof:* This lemma consists of three parts. The first part is to show the existence of the first limit in the lemma so that the definition of \( D(P_Z||Q) \) is valid. The second part is to show that the value of the limit is indeed \( E\left( \log \frac{P_Z(Z_0|Z_{-1}^{\infty})}{Q(Z_0|Z_{-1}^{\infty})} \right) \). Finally, the last part is to show the uniform convergence of normalized log-likelihood ratio to the relative entropy rate. The first two parts and the pointwise convergence of the third part is a generalization of the Shannon-McMillan-Breiman theorem. The proof of these parts is identical to those in [19, Theorem 2.3.3] even for the case where some parameters in \( \Theta_k^\delta \) can be zero.

The uniform convergence in the third part of the lemma is crucial in that it enables to obtain the second consistency result (2.1) as in [19, Theorem 2.2.1]. We take into account our parameter set, and repeat the argument of [19, Lemma 2.4.1]. To show the uniform convergence, we need to show

\[
\lim_{n \to \infty} \frac{1}{n} \log Q(Z^n) = E\left( \log Q(Z_0|Z_{-\infty}^{0}) \right) \quad \text{a.s.,}
\]

uniformly on \( \Theta_k^\delta \). Since the pointwise convergence can be shown and the parameter set \( \Theta_k^\delta \) is compact, it is enough to show that \( \frac{1}{n} \log Q(Z^n) \) is an equicontinuous sequence
by Ascoli’s Theorem. That is, we need to show that for all \( \varepsilon > 0 \), there exists \( \delta(\varepsilon) > 0 \) such that if \( \|Q - Q'\|_1 < \delta(\varepsilon) \), then

\[
\frac{1}{n} \log Q(Z^n) - \frac{1}{n} \log Q'(Z^n) \leq \varepsilon \quad \text{for all } n,
\]

(2.8)

where \( \|Q - Q'\|_1 = \sum_{i,j} |a_{ij} - a'_{ij}| \) is defined to be the \( L_1 \) distance between the two parameters defining \( Q \) and \( Q' \). This equicontinuity can be proved by observing that a process \( \{S_t = (X_t^{(k-1)}, Z_t)\} \) is a Markov process under any \( Q \in \Theta_k \), where \( \{S_t\} \) has a state space \( S = A^k \times A \). This is true since

\[
Q(S_{t+1}|S^t) = Q(X_{t+1}, Z_{t+1}|X^t, Z^t)
= Q(X_{t+1}|X^t, Z^t)Q(Z_{t+1}|X_{t+1}, Z^t)
= Q(X_{t+1}|X^t_{(k-1)})\Pi(X_{t+1}, Z_{t+1})
= Q(S_{t+1}|S_t).
\]

Let \( \{x^k_i(i) : i = 1, \cdots, M^k\} \) denote the set of all possible \( k \)-tuples of \( \{X_t\} \), and let \( s = (x^k_1(i), z), \bar{s} = (x^k_1(j), \bar{z}) \). Then, the transition matrix \( T \) of \( \{S_t\} \) has elements 

\[
t_{s\bar{s}} \triangleq Q(S_{t+1} = \bar{s}|S_t = s) = a_{ij}\Pi(x_k(j), \bar{z}). \quad \text{Since all } A \quad \text{that are in } \Theta_k \quad \text{are irreducible}
\]

and aperiodic, and \( \Pi(x_k(j), \bar{z}) > 0 \) for all \( x_k(j) \) and \( \bar{z} \), \( T \) is also irreducible and aperiodic. Hence, \( T \) has the unique stationary distribution \( \tau \). In addition, from Assumption 1, we can observe that for all \( Q \in \Theta_k \), \( Q(S^n) > 0 \) a.s. (with respect to \( P \)).

Since \( \{S_t\} \) is also stationary, we have

\[
Q(S^n) = \tau_{S_1} \prod_{t=k}^{n-1} t_{S_tS_{t+1}} = \tau_{S_1} \prod_{(s,\bar{s})} t_{ss}^N
\]

where

\[
N_{s\bar{s}} \triangleq \sum_{t=k}^{n-1} 1(S_t = s, S_{t+1} = \bar{s}).
\]
For another probability law $Q' \in \Theta_k^f$, we have

$$\frac{1}{n} \log Q(S^n) - \frac{1}{n} \log Q'(S^n)$$

$$\leq \frac{1}{n} \log \tau_{S_1} - \frac{1}{n} \log \tau'_{S_1} + \frac{1}{n} \sum_{(s,\bar{s})} N_{s\bar{s}} \log t_{s\bar{s}} - \frac{1}{n} \sum_{(s,\bar{s})} N_{s\bar{s}} \log t'_{s\bar{s}}$$

$$\leq |\log \tau_{S_1} - \log \tau'_{S_1}| + \sum_{(s,\bar{s})} |\log t_{s\bar{s}} - \log t'_{s\bar{s}}|$$

$$= |\log \tau_{S_1} - \log \tau'_{S_1}| + \sum_{(i,j)} |\log a_{ij} - \log a'_{ij}|,$$

where (2.9) follows from the fact that $\frac{1}{n} \leq 1, \frac{N_{s\bar{s}}}{n} \leq 1$, with probability 1, and (2.10) follows from the fact that DMC, II, is equal for $Q$ and $Q'$. The summations are over the pairs that have nonzero transition probabilities.

Since the function $f(x) = \log x$ is a uniformly continuous function for $\delta \leq x < 1$, and $a_{ij} \geq \delta$ that occur in the summation, we have for $\epsilon > 0$,

$$\sum_{(i,j)} |\log a_{ij} - \log a'_{ij}| < \frac{\epsilon}{2} \quad \text{if} \quad \|Q - Q'\|_1 < \delta_1(\epsilon).$$

In addition, we know that all the elements of the stationary distribution of $T$ are bounded away from zero, since the largest element of the stationary distribution of $T$ is lower bounded by $\frac{1}{M+1}$, and any state can be reached by finite number of steps whose transition probabilities are bounded away from zero. Therefore, for some $C_1 < \infty$,

$$|\log \tau_{S_2} - \log \tau'_{S_2}| < C_1 |\tau_{S_1} - \tau'_{S_1}|.$$

Then, from the result of the sensitivity of the stationary distribution of a Markov chain [30], for some $C_2 < \infty$,

$$|\tau_{S_1} - \tau'_{S_1}| \leq C_2 \sum_{(s,\bar{s})} |t_{s\bar{s}} - t'_{s\bar{s}}| = C_2 \sum_{(i,j)} |a_{ij} - a'_{ij}|.$$

Hence, for $\epsilon > 0$, we obtain,

$$|\log \tau_{S_1} - \log \tau'_{S_1}| < \frac{\epsilon}{2} \quad \text{if} \quad \|Q - Q'\|_1 < \delta_2(\epsilon).$$
Therefore, by letting \( \delta(\epsilon) = \min(\delta_1(\epsilon), \delta_2(\epsilon)) \), we have

\[
\left| \frac{1}{n} \log Q(S^n) - \frac{1}{n} \log Q'(S^n) \right| < \epsilon \quad \text{if} \quad \|Q - Q'\|_1 < \delta(\epsilon).
\]

Let us now go back to the original process \( Z \). From

\[
\left| \frac{1}{n} \log Q(S^n) - \frac{1}{n} \log Q'(S^n) \right| < \epsilon,
\]

we have

\[
Q'(X^n, Z^n) < \exp(n\epsilon)Q(X^n, Z^n),
\]

thus,

\[
Q'(Z^n) = \sum_{x^n} Q'(x^n, Z^n) < \exp(n\epsilon) \sum_{x^n} Q(x^n, Z^n) = \exp(n\epsilon)Q(Z^n),
\]

where the summations are again over the sequences that have nonzero probabilities. By changing the role of \( Q \), and \( Q' \), we get the result (2.8), namely, \( \frac{1}{n} \log Q(Z^n) \) is an equicontinuous sequence. Therefore, we have the uniform convergence of the lemma.

**Lemma 3 (Uniform Concentration)** Suppose \( Q \in \Theta^\delta_k \) for some fixed \( \delta > 0 \). Let \( \hat{X}_Q \) be the randomized filter defined in (2.3). Then,

\[
\lim_{n \to \infty} \left( L_{\hat{X}_Q}(X^n, Z^n) - E\left(L_{\hat{X}_Q}(X^n, Z^n)\right) \right) = 0 \quad \text{a.s.}
\]

uniformly on \( \Theta^\delta_k \). Proof: This lemma shows the uniform concentration property of \( L_{\hat{X}_Q}(X^n, Z^n) \). The randomization of the filter is needed to deal with ties occur in deciding the Bayes response. A detailed proof of this lemma is given in Appendix 2.

**Lemma 4 (Continuity)** Consider a single letter filtering setting. Suppose \( Q \) is some other joint probability law of \( X \) and \( Z \). Define single letter filters \( \hat{X}_P(z) \) and...
\[ \hat{X}_Q^e(z) \] as

\[ \hat{X}_P(z)[\hat{x}] = \Pr(B(P_{X|z}) = \hat{x}) \]
\[ \hat{X}_Q^e(z)[\hat{x}] = \Pr(B(Q_{X|z} + U) = \hat{x}), \]

where \( U \in \mathbb{R}^M \) is a uniform random vector in \( B \) as before. Then,

\[ E\left( \ell(X, \hat{X}_Q^e(Z)) \right) - E\left( \ell(X, \hat{X}_P(Z)) \right) \leq \Lambda_{\max} K_1 \cdot \|P_z - Q_z\|_1 + C_A \cdot \epsilon, \]

where the expectations on the left hand side of the inequality are under \( P \) and \( K_1 = \sum_{i=1}^M \|\Pi_i^{-1}\|_2 \), and \( C_A = \max_{a,b\in A} \|\lambda_a - \lambda_b\|_2 \).

Remark: This lemma states that the excess expected loss of a randomized filter optimized for a mismatched probability law can be upper bounded by the \( L_1 \) difference between the true and the mismatched probability laws of output symbol, plus a small constant term which diminishes with the randomization probability. This is somewhat analogous to a result for the prediction problem which was derived in [4, (20)].
Proof of Lemma 4: Define $\hat{X}_Q(z)[\hat{x}] = \Pr(B(Q_x|z) = \hat{x})$. Then,

$$
E\left(\ell(X, \hat{X}_Q(Z))\right) - E\left(\ell(X, \hat{X}_P(Z))\right)
= \sum_{x,z} P(x, z) \left(\ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_P(z))\right)
= \sum_{x,z} \left(P(x, z) - Q(x, z)\right) \left(\ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_P(z))\right)
+ \sum_{x,z} Q(x, z) \left(\ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_P(z))\right)
\leq \Lambda_{\max} \sum_{x,z} |P(x, z) - Q(x, z)|
+ \sum_{x,z} Q(x, z) \left(\ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_P(z))\right)
\leq \Lambda_{\max} \sum_{x,z} |P(x, z) - Q(x, z)|
+ \sum_{x,z} Q(x, z) \left(\ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_P(z))\right),
$$

(2.11)

where (2.11) follows from Hölder’s inequality, and (2.12) follows from the fact that $\sum_{x,z} Q(x, z)(\ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_P(z))) \leq 0$. Now, let us bound the first term in (2.12).

$$
\Lambda_{\max} \sum_{x,z} |P(x, z) - Q(x, z)|
= \Lambda_{\max} \sum_{x} |P(x) - Q(x)| \left(\sum_{z} \Pi(x, z)\right)
= \Lambda_{\max} \sum_{x} |P(x) - Q(x)|
= \Lambda_{\max} \sum_{i} \left|(P_Z - Q_Z)^T \Pi_i^{-1}\right|
\leq \Lambda_{\max} \sum_{i} \|\Pi_i^{-1}\|_2 \cdot \|P_Z - Q_Z\|_2
\leq \Lambda_{\max} K_{\Pi} \cdot \|P_Z - Q_Z\|_1,
$$

(2.13)

(2.14)

(2.15)

where (2.13) follows from the fact that $\sum_{z} \Pi(x, z) = 1$; (2.14) follows from Cauchy-Schwartz inequality, and (2.15) follows from the fact that $L_2$-norm is less than or
equal to $L_1$-norm.

The second term in (2.12) becomes

$$
\sum_{x,z} Q(x, z) \left( \ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_Q(z)) \right)
\leq \sum_{x \in X} \sum_{z} Q(x | z) \left( \hat{X}_Q(z)[\hat{x}] - \hat{X}_Q(z)[\hat{x}] \right) \sum_{x} \Lambda(x, \hat{x}) Q(x | z)
\leq \sum_{x} Q(z) \sum_{x} \left( \hat{X}_Q(z)[\hat{x}] - \hat{X}_Q(z)[\hat{x}] \right) \sum_{x} \Lambda(x, \hat{x}) Q(x | z)
\leq \sum_{x} Q(z) \sum_{z} \left( \hat{X}_Q(z)[\hat{x}] - \hat{X}_Q(z)[\hat{x}] \right) \cdot \lambda_x^T Q_{X|z}. \tag{2.16}
$$

It is easy to see that the inner summation in (2.16) is always nonnegative since by definition, $\hat{X}_Q(z)$ assigns probability 1 to $B(Q(x | z))$. Now, for a given $Q$, define

$$
U_{\text{max}} = \arg \max_{U \in B_+} \left( \lambda_{B(QX|z)+U} - \lambda_{B(QX|z)} \right) Q_{X|z}, \tag{2.17}
$$

resolving ties arbitrarily. Then, we have,

$$
\sum_{\hat{x}} \left( \hat{X}_Q(z)[\hat{x}] - \hat{X}_Q(z)[\hat{x}] \right) \cdot \lambda_x^T Q_{X|z}
\leq \left( \lambda_{B(QX|z)+U_{\text{max}}} - \lambda_{B(QX|z)} \right) Q_{X|z} \tag{2.18}
\leq \left( \lambda_{B(QX|z)} - \lambda_{B(QX|z)+U_{\text{max}}} \right) U_{\text{max}} \tag{2.19}
\leq \max_{a, b \in A} \| \lambda_a - \lambda_b \|_2 \cdot \| U_{\text{max}} \|_2 \tag{2.20}
\leq C_A \cdot \epsilon,
$$

where (2.18) follows from (2.17); (2.19) follows from the fact

$$
\lambda_{B(QX|z)+U_{\text{max}}}(Q_{X|z} + U_{\text{max}}) \leq \lambda_{B(QX|z)}(Q_{X|z} + U_{\text{max}}).$

and (2.20) follows from the Cauchy-Schwartz inequality. Note that depending on $Q$ and $z$, (2.18) and (2.19) can be both zero and hold with equality. Together with (2.15), the lemma is proved.

Before moving on to Lemma 5, we need following three definitions. In Lemma 2, we have seen that for $Q \in \Theta^k$, $D(P_Z \| Q)$ is well-defined. Now, let us consider the case where $Q \in \Theta^k$ is some function of the noisy observation $Z^n$ (denoted as $Q[Z^n]$). As mentioned in the footnote of Section 2.4.2, the notion of the relative entropy rate between $P_Z$ and that random $Q[Z^n]$ is defined in Definition 2 using Definition 1. Definition 3 is also needed for the inequality in Lemma 5.

\textbf{Definition 1} Suppose $Q[Z^n] \in \Theta^k$ and $f$ is some function of $(X^\infty, Z^\infty, Q[Z^n])$ such that the expectation

$$E\left( f(X^\infty, Z^\infty, Q[Z^n]) \right) = \int f(x^\infty, z^\infty, Q[z^n])dP(x^\infty, z^\infty)$$

exists. Then, the notation $\hat{E}(\cdot)$ is defined as

$$\hat{E}\left( f(X^\infty, Z^\infty, Q[Z^n]) \right) \triangleq \int f(x^\infty, z^\infty, Q[z^n])dP(x^\infty, z^\infty).$$

That is, in $\hat{E}\left( f(X^\infty, Z^\infty, Q[Z^n]) \right)$, the Lebesgue integration with respect to the randomness of $Q[Z^n]$ is excluded. Moreover, suppose $T$ is a set of some time indices, and $z_T$ denotes a subsequence of $z^\infty$ with the time indices in $T$. Then, the conditioning on $z_T$ with respect to $\hat{E}$ is defined as

$$\hat{E}\left( f(X^\infty, Z^\infty, Q[Z^n]) \bigg| Z_T = z_T \right) \triangleq \int f(x^\infty, z^\infty, Q[z^n])dP(x^\infty, z^\infty | z_T),$$

where $P(\cdot | z_T)$ denotes the conditional probability measure on $(x^\infty, z^\infty)$ given $z_T$. Again, the randomness of $Q[Z^n]$ is excluded in the Lebesgue integration.

\textbf{Remark}: The above definitions of $\hat{E}(\cdot)$ and $\hat{E}(\cdot | \cdot)$ may seem subtle. However, the main point of two definitions is simple, namely, they exclude the randomness of $Q[Z^n]$ in calculating Lebesgue integrations.
Definition 2 Suppose \( Q[Z^n] \in \Theta_k \). Then, the relative entropy rate between \( P_Z \) and \( Q[Z^n] \) is defined as,
\[
D(P_Z \| Q[Z^n]) = \mathbb{E}\left( \log \frac{P_Z(Z_0|Z_{-1}^{n-1})}{Q(Z^n)(Z_0|Z_{-\infty}^{n-1})} \right).
\]

Remark: Note that \( D(P_Z \| Q[Z^n]) \) is a function of \( Z^n \), and still is a random variable.

Definition 3 Define the \( k \)-th order Markov approximation of \( P_X \) for \( n \geq k \) as
\[
P_X^{(k)}(X^n) \triangleq P_X(X^k) \prod_{i=k+1}^n P_X(X_i|X_{i-1}^{i-1}).
\]

Furthermore, denote \( P_Z \) and \( P_Z^{(k)} \) as the probability law of the output of DMC, \( \Pi \), when the probability law of input is \( P_X \) and \( P_X^{(k)} \), respectively.

Remark: Note that \( P_Z^{(k)} \) is not the \( k \)-th order Markov approximation of \( P_Z \), but is the distribution of the channel output whose input is \( P_X^{(k)} \), the \( k \)-th order Markov approximation of the original input distribution \( P_X \).

Now, we give following lemma that upper bounds the relative entropy rate between \( P_Z \) and the ML estimator.

Lemma 5 For the given sequence \( \{\delta_k\} \) defined in Section 2.4.1 and for fixed \( k \), we have
\[
\lim_{n \to \infty} D(P_Z \| \hat{Q}_{k,\delta_k}[Z^n]) \leq D(P_X \| P_X^{(k)}) \quad a.s.
\]

Proof: Recall that \( \hat{Q}_{k,\delta_k}[Z^n] \) is an ML estimator in \( \Theta_k^{\delta_k} \) based on the observation \( Z^n \). From (2.1), we know that
\[
\lim_{n \to \infty} D(P_Z \| \hat{Q}_{k,\delta_k}[Z^n]) = \min_{Q \in \Theta_k^{\delta_k}} D(P_Z \| Q) \quad a.s.
\]

Also, (2.2) and Definition 3 assures that \( P_Z^{(k)} \in \Theta_k^{\delta_k} \). Therefore, we have
\[
\lim_{n \to \infty} D(P_Z \| \hat{Q}_{k,\delta_k}[Z^n]) \leq D(P_Z \| P_Z^{(k)}) \quad a.s.
\]
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

This is the link where we needed Assumption 1. Now, let us denote \( P^{(k)} \) as the joint probability law of \((X^n, Z^n)\) when the probability law of input process is \( P_X^{(k)} \). Then, by the chain rule of relative entropy [29, (2.67)], we have

\[
E\left( \log \frac{P(X^n, Z^n)}{P^{(k)}(X^n, Z^n)} \right) = D_n(P_X \| P_X^{(k)}) + E\left( \log \frac{P(Z^n \| X^n)}{P^{(k)}(Z^n \| X^n)} \right) = D_n(P_Z \| P_Z^{(k)}) + E\left( \log \frac{P(X^n \| Z^n)}{P^{(k)}(X^n \| Z^n)} \right).
\]

Since the DMC is fixed, we have \( E\left( \log \frac{P(Z^n \| X^n)}{P^{(k)}(Z^n \| X^n)} \right) = 0 \). Moreover, by the nonnegativity of relative entropy, \( E\left( \log \frac{P(X^n \| Z^n)}{P^{(k)}(X^n \| Z^n)} \right) \geq 0 \). Therefore, we get \( D_n(P_Z \| P_Z^{(k)}) \leq D_n(P_X \| P_X^{(k)}) \). Since \( D(P_X \| P_X^{(k)}) = \lim_{n \to \infty} \frac{1}{n} D_n(P_X \| P_X^{(k)}) \) always exists by ergodicity, we have

\[
D(P_Z \| P_Z^{(k)}) \leq D(P_X \| P_X^{(k)}),
\]

and the lemma is proved. ■

We are now finally in a position to prove our main theorem.

Proof of Theorem 1: As mentioned in Section 2.4.2, we first fix \( k \) and \( \epsilon \), and try to get the inequality in the form of (2.6) to prove Part (a). To refresh, (2.6) is given again here.

\[
\limsup_{n \to \infty} \left( L_{X_{\text{univ}},k} (X^n, Z^n) - \phi_n(P_X, \Pi) \right) \leq F\left( \limsup_{t \to \infty} D(P_Z \| Q_k^t), \epsilon \right) \quad \text{a.s.}
\]

From the definition of \( L_{X_{\text{univ}},k} (X^n, Z^n) \),

\[
L_{X_{\text{univ}},k} (X^n, Z^n) = \frac{1}{n} \sum_{t=1}^{n} \ell(X_t, \hat{X}_{Q_k^t}^c(Z_t')),
\]

where from (2.5), we know that \( Q_k^t \) is a function of \( Z^{m_k(t)} \). Since \( \ell(X_t, \hat{X}_{Q_k^t}^c(Z_t')) \) is a function of \( (X_t, Z', Q[Z^{m_k(t)}]) \), we can define a quantity \( \hat{E}(\ell(X_t, \hat{X}_{Q_k^t}^c(Z_t'))) \) from
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

Definition 1. From this, we also define

\[ \hat{E}\left(L_{\hat{X}_{n}^{\text{univ},k}}(X^n, Z^n)\right) = \frac{1}{n} \sum_{i=1}^{n} \hat{E}\left(\ell(X_i, \hat{X}_{Q_k}(Z_i))\right). \]

Now, we have following Corollary 1 from Lemma 3, whose proof is given in Appendix 3. This corollary is a key step in proving the main theorem, since it provides a crucial link that enables to get the inequality in (2.6).

**Corollary 1** For fixed \( k \) and \( \epsilon \), we have

\[
\lim_{n \to \infty} \left( L_{\hat{X}_{n}^{\text{univ},k}}(X^n, Z^n) - \hat{E}\left(L_{\hat{X}_{n}^{\text{univ},k}}(X^n, Z^n)\right) \right) = 0 \quad \text{a.s.}
\]

From Corollary 1, we have following equality

\[
\limsup_{n \to \infty} \left( L_{\hat{X}_{n}^{\text{univ},k}}(X^n, Z^n) - \phi_n(P_X, \Pi) \right) = \limsup_{n \to \infty} \left( \hat{E}\left(L_{\hat{X}_{n}^{\text{univ},k}}(X^n, Z^n)\right) - \phi_n(P_X, \Pi) \right) \quad \text{a.s.}
\]

Therefore, to get the inequality of the form of (2.6), we can equivalently show

\[
\limsup_{n \to \infty} \left( \hat{E}\left(L_{\hat{X}_{n}^{\text{univ},k}}(X^n, Z^n)\right) - \phi_n(P_X, \Pi) \right) \leq F\left( \limsup_{t \to \infty} \mathcal{D}(P_Z \| Q_k^t), \epsilon \right).
\]
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

Now, let us consider following chain of inequalities:

\[
\hat{E}\left( L_{X^{n}}^{\text{univ},k} (X^{n}, Z^{n}) \right) - \phi_{n}(P_{X}, \Pi) \\
= \frac{1}{n} \sum_{t=1}^{n} \left( \hat{E} \left( \ell(X_{t}, \hat{X}_{Q_{k}}^{t} (Z^{t})) \right) - \hat{E} \left( \ell(X_{t}, \hat{X}_{P} (Z^{t})) \right) \right) \\
= \frac{1}{n} \sum_{t=1}^{n} \hat{E} \left( \hat{E} \left( \ell(X_{t}, \hat{X}_{Q_{k}}^{t} (Z_{t}, Z_{t-1}^{t-1})) \right) | Z_{t-1}^{t-1} \right) \\
- \hat{E} \left( \ell(X_{t}, \hat{X}_{P} (Z_{t}, Z_{t-1}^{t-1})) | Z_{t-1}^{t-1} \right) \\
\leq \frac{K_{\Pi} \Lambda_{\text{max}}}{n} \sum_{t=1}^{n} \hat{E} \left( |P_{Z_{t}|Z_{t-1}^{t-1}} - Q_{k}^{t} (Z_{t}|Z_{t-1}^{t-1})| \right) + C_{A} \cdot \epsilon \\
\leq \sqrt{2 \ln 2 K_{\Pi} \Lambda_{\text{max}}} \sum_{t=1}^{n} \hat{E} \left( \log \frac{P(Z_{t}|Z_{t-1}^{t-1})}{Q_{k}^{t} (Z_{t}|Z_{t-1}^{t-1})} \right) + C_{A} \cdot \epsilon. \tag{2.21}
\]

The notation \( Q_{k}^{t} (Z_{t}|Z_{t-1}^{t-1}) \) in (2.21) stands for the simplex vector in \( M \) whose \( i \)-th component stands for \( Q_{k}^{t} (Z_{t} = i|Z_{t-1}^{t-1}) \). The inequality in (2.21) is obtained from Lemma 4, since \( \Pi \) does not vary with \( t \), and given \( Z_{t}^{t-1} \), estimating \( X_{t} \) based on \( Z^{t} \) is equivalent to the single letter setting as in Lemma 4 with the corresponding conditional distribution. Furthermore, (2.22) follows from Pinsker’s inequality [29, Lemma 11.6.1], and (2.23) follows from Jensen’s inequality. By taking \( \limsup \) on both sides, we have

\[
\limsup_{n \to \infty} \left( \hat{E}\left( L_{X^{n}}^{\text{univ},k} (X^{n}, Z^{n}) \right) - \phi_{n}(P_{X}, \Pi) \right) \\
\leq \sqrt{2 \ln 2 K_{\Pi} \Lambda_{\text{max}}} \limsup_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \hat{E} \left( \log \frac{P(Z_{t}|Z_{t-1}^{t-1})}{Q_{k}^{t} (Z_{t}|Z_{t-1}^{t-1})} \right) \\
+ C_{A} \cdot \epsilon \quad \text{a.s.},
\]

\(^{5}\)All the equalities and inequalities between random variables in this proof should be understood in almost sure sense.
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

since the square root function is a continuous function. For the expression inside the square root of the right-hand side of the inequality,

\[
\limsup_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \hat{E} \left( \log \frac{P(Z_t|Z_{t-1})}{Q_k(Z_t|Z_{t-1})} \right) = \limsup_{t \to \infty} \hat{E} \left( \log \frac{P(Z_t|Z_{t-1})}{Q_k(Z_t|Z_{t-1})} \right) \quad \text{a.s.} \quad (2.24)
\]

\[
= \limsup_{t \to \infty} \hat{E} \left( \log \frac{P(Z_0|Z_{-\infty}^t)}{Q_k(Z_0|Z_{-\infty}^t)} \right) \quad \text{a.s.} \quad (2.25)
\]

\[
= \limsup_{t \to \infty} D(P_Z || Q_k) \quad \text{a.s.,} \quad (2.26)
\]

where (2.24) follows from Cesáro’s mean convergence theorem; the numerator of (2.25) follows from the fact that \( P \) is stationary and \( P(Z_0|Z_{-1}^t) \to P(Z_0|Z_{-1}^\infty) \) almost surely by martingale convergence theorem; and the denominator of (2.25) follows from the fact that \( Q_k \) is also a stationary law, and with probability 1, for any \( \epsilon > 0 \), there exists \( N \) such that for all \( t \geq N \), \( |Q_k(Z_0|Z_{-1}^t) - Q_k(Z_0|Z_{-1}^\infty)| < \epsilon \), which is guaranteed by the uniform convergence result of Lemma 1. Finally, (2.26) follows from Definition 2. Therefore,

\[
\limsup_{n \to \infty} \left( \hat{E} \left( L_{X_{\text{univ},k}}(X^n, Z^n) - \phi_n(P_X, \Pi) \right) \right) \leq \sqrt{2 \ln 2K_\Pi \lambda_{\text{max}} } \limsup_{t \to \infty} D(P_Z || Q_k) + C_\lambda \cdot \epsilon \quad \text{a.s.,} \quad (2.27)
\]

which finally is in the form of (2.6). Now, we need to check if the right-hand side of (2.27) goes to zero if we let \( k \to \infty \) and \( \epsilon \downarrow 0 \). To see this, consider following further upper bounds.

\[
\limsup_{t \to \infty} D(P_Z || Q_k')
\]

\[
= \limsup_{t \to \infty} D(P_Z || \hat{Q}_k, \delta_h[Z_t]) \quad (2.28)
\]

\[
\leq D(P_X || P_{X_k}), \quad (2.29)
\]
where (2.28) follows from the fact that $m_{it(t)} \to \infty$ as $t \to \infty$, and (2.29) follows from Lemma 5. The inequality (2.29) holds for every $k$, and by Shannon-McMillan-Breiman Theorem [29, Chapter 16.8], we know $D(P_x \| P_{X_k}) \to 0$ as $k \to \infty$. Therefore,

$$\lim_{k \to \infty} \limsup_{t \to \infty} D(P_x \| Q_t^t) = 0,$$

and thus,

$$\lim_{k \to \infty} \limsup_{n \to \infty} \left( L_{\hat{X}_{univ, k}}(X^n, Z^n) - \phi_n(P_X, \Pi) \right) \leq C_A \cdot \epsilon \quad a.s.$$

Finally, sending $\epsilon$ to zero, Part (a) of the theorem is proved. Part (b) follows directly from (a), and Fatou's Lemma. That is,

$$\lim_{k \to \infty} \limsup_{n \to \infty} \left( E\left( L_{\hat{X}_{univ, k}}(X^n, Z^n) - \phi_n(P_X, \Pi) \right) \right) \leq C_A \cdot \epsilon.$$

Note that the expectation here is with respect to the randomness of probability law within the parenthesis, too. By sending $\epsilon$ to zero, Part (b) is proved. $\blacksquare$

### 2.5 Extension: Universal filtering for channel with memory

Now, let us extend our result to the case where channel has memory. With the identical assumption on \{X_t\}, now suppose \{Z_t\} is expressed as

$$Z_t = X_t \oplus N_t,$$  \hspace{1cm} (2.30)
where \( \oplus \) denotes modulo-\( M \) addition, and \( \{N_t\} \) is an \( \mathcal{A} \)-valued noise process which is not necessarily memoryless. We assume we have a complete knowledge of the probability law of \( \{N_t\} \). Specifically, let us consider the case where \( \{N_t\} \) is a HMP, that is, it is an output of an invertible memoryless channel \( \Gamma = \{\Gamma(i,j)\}_{i,j \in \mathcal{A}} \) whose input is irreducible, aperiodic \( \ell \)-th order Markov process, \( \{S_t\} \), which is independent of \( \{X_t\} \). Let \( \Gamma_{\min} = \min_{i,j \in \mathcal{A}} \{\Gamma(i,j)\} \), and suppose \( \Gamma_{\min} > 0 \). For simplicity, assume that the alphabet size of \( \{S_t\} \) is also \( \mathcal{A} \).

In this model, the channel between \( X_t \) and \( Z_t \) at time \( t \) is an \( M \)-ary symmetric channel, which is specified by the \( S_t \)-th row of \( \Gamma \). Define an \( M \times M \) matrix \( \Pi_t \) whose \( (x_t, z_t) \)-th element is

\[
\Pi_t(x_t, z_t) = P_{N_t}(z_t \oplus x_t) = \Pr(Z_t = z_t | X_t = x_t) = \sum_{s_t} \Pr(Z_t = z_t | X_t = x_t, S_t = s_t) \Pr(S_t = s_t),
\]

where \( \oplus \) denotes modulo-\( M \) subtraction. Now, let us make following assumptions on the noise process.

- \( \{N_t\} \) is stationary, i.e., \( \Pi_t \) is identical for all \( t \)
- \( \Pi_t \) is invertible
- for all \( S_{t-\ell}^t(\omega) \), there exists an \( \alpha \) such that \( \Pr(S_t | S_{t-\ell}^{t-1}) \geq \alpha > 0 \).

As stated in [31, 2-A], the first and the second assumptions are rather benign. Especially, for the second assumption, it can be shown that under benign conditions on the parametrization, almost all parameter values except for those in a set of Lebesgue measure zero, give rise to a process satisfying this assumption. In addition, since this only corresponds to the case when \( k = 0 \) in [31, Assumption 1], it is a much weaker assumption. The third assumption is a similar positivity assumption as Assumption 1, which enables our universal filtering scheme.

Under these assumptions on the noise process, we can extend our scheme to do the universal filtering for this channel. First, we can convert this channel to the
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

An equivalent memoryless channel, \( \Xi = \{ \xi((i,j), h) \}_{i,j,h \in A} \), where the input process is \( \{(X_t, S_t)\} \) and the output is \( \{Z_t\} \). Here, \( \Xi \) is a \( M^2 \times M \) matrix, and the channel transition probability is

\[
\xi((i,j), h) = \Gamma(j, h \oplus i),
\]

for all \( i, j, \) and \( k \). To do the filtering, we apply our scheme to this equivalent memoryless channel. For fixed \( k \geq \ell \), as in Section 2.2.2, define a parameter set of HMPs, \( \Theta_k \), whose Markov chain has \( M^{k+\ell} \) states, and the memoryless channel has dimension \( \mathbb{R}^{M^{(k+\ell)} \times M} \). The \( k \)-th order conditional probability of our new input process is

\[
\Pr(X_t, S_t | X_{t-k}^{t-1}, S_{t-k}^{t-1}) = \Pr(X_t | X_{t-k}^{t-1}) \cdot \Pr(S_t | S_{t-k}^{t-1}) \geq \delta_k \cdot \alpha,
\]

(2.31)

where (2.31) follows from Assumption 1 and the third condition on the noise process. Let \( \gamma_k = \delta_k \cdot \alpha \). Then, we can model \( \{Z_t\} \) in \( \Theta_k^\gamma \), or equivalently, model \( (X_t, S_t) \) as \( k \)-th order Markov chain, and obtain \( Q_k^\gamma \), the ML estimator in \( \Theta_k^\gamma \) based on \( Z_{m_i(t)} \). By forward recursion, we can get \( Q_k^\gamma(X_t, S_t | Z^t) \), and by summing over \( S_t \)'s we can calculate \( Q_k^\gamma(X_t | X_{t-i} | Z^t) \), the simplex vector in \( \mathcal{M} \) whose \( i \)-th component is \( Q_k^\gamma(X_t = i | Z^t) \). Then, finally we define our sequence of universal filtering schemes as,

\[
\hat{X}_{\text{univ},k}^\epsilon = \{ \hat{X}_{Q_k^\gamma, t}^\epsilon \},
\]

exactly the same as we proposed in Section 2.4.1.

The analysis of this scheme is identical to the one given in the proof of the main theorem. The equation (2.21), which is the only place where the invertibility of the \( \Pi \) is used, can also be obtained in this case due to the second assumption of the noise process. Thus, we again get

\[
\lim sup_{n \to \infty} \left( L_{\hat{X}_{\text{univ},k}^\epsilon} \left( X^n, Z^n \right) - \phi_n(P_X, \Pi) \right) \leq 2\sqrt{2 \ln 2} K_{\Pi} \Lambda_{\max} \sqrt{\lim sup_{t \to \infty} D(P_Z \| Q_k^\gamma)} + C \Lambda \cdot \epsilon \quad \text{a.s.}
\]
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

Since
\[
\limsup_{t \to \infty} D(P_z \| Q_k^t) = \limsup_{t \to \infty} D(P_z \| \hat{Q}_{k,t}[Z]) \leq D(P_X \| P_{X_k})
\]
by the same argument as Lemma 5, we have the same result as Theorem 1. Thus, we can successfully extend our scheme to the case where the channel noise is a HMP with some mild assumptions.

2.6 Discussion

Throughout the chapter, we have only considered the case where the input, output, and reconstruction alphabets are equal. However, we can easily extend our result to the case where the alphabet sizes are different (but still finite). In that case, the condition on the channel, parallel to the invertibility condition, is that the channel transition matrix should have full row-rank. Since the argument of the extension would be rather straightforward, we omit the details in this chapter.

The result that we attain in Theorem 1 can also be attained by the schemes devised in [12][8]. Therefore, we have shown that a completely different approach can achieve the same goal in the universal filtering problem when the underlying signal is a stationary and ergodic process. In addition, our work gives the first theoretical justification of using HMP models for filtering, which is a prevalent approach in practice where the underlying signal need not be a Markov process. Furthermore, it is not clear how to extend the schemes in [12][8] to the case of channels with memory, whereas the extension of our scheme to such cases is quite simple in some settings (e.g., when the noise is a HMP), as in Section 2.5.

As described in Section 2.4.1, our filter is a randomized filter. The randomization is necessary in obtaining the continuity result of Lemma 9(a), which we use for proving our main theorem. Whether a deterministic version of our filter, i.e., a filter that is defined without \( U \) in (2.3), is universally optimal, remains an open question. The filter devised in [12][8] is also a randomized scheme, and a parallel discussion regarding
the randomization is also given in [8, Section VI]. In contrast, a filter that appears in
[32], which is equivalent to the scheme in [7] that only utilizes a one-sided context, is a
deterministic scheme that can indeed achieve the asymptotically optimal performance
in Theorem 1. Therefore, when the channel is memoryless, the randomization of a
universal filter is not necessary in general to achieve the performance goal of Theorem
1. However, when the channel has memory as in Section 2.5, we are not aware of any
deterministic filter that can be universally optimal for any underlying stationary and
ergodic process.

2.7 Appendix

2.7.1 Three lemmas

Here, we revise three lemmas from [28] regarding probability law of HMP. These
are needed to prove Lemma 1. For the following three lemmas, fix k and δ, and
suppose \( Q \in \Theta_k^\delta \). Also, fix some \( m \in \mathbb{N} \), such that \( m \geq k \). Proofs are similar to [28,
Appendix]. Note that \( \{X_t\} \) is still our clean signal and \( \{Z_t\} \) is the noisy observed
signal (not necessarily a HMP).

**Lemma 6** We have

\[
Q(X_{t+m} = j | X_t = i, Z_\infty) \geq \mu_{\delta,k,m},
\]

where \( \mu_{\delta,m,k} = \frac{1 + \frac{M-1}{(\delta \Pi_{\min})^{m+\epsilon}}}{} \) is independent of \( Q, Z_\infty, i, j \).
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

**Proof:** From Markovity and conditioning,

\[
\frac{Q(X_{t+m} = j, Z_{t+m+k+1} = i) \div Q(X_{t+m} = j', Z_{t+m+k+1} = i)}{Q(X_{t+m} = j, Z_{t+m+k+1} = i) \div Q(X_{t+m} = j', Z_{t+m+k+1} = i)}
\]

\[
= \frac{Q(X_{t+m} = j, Z_{t+m+k+1}^\infty | X_t = i)}{Q(X_{t+m} = j', Z_{t+m+k+1}^\infty | X_t = i)}
\]

\[
= \frac{Q(Z_{t+m+k}^\infty | X_t = i, X_{t+m} = j)}{Q(Z_{t+m+k}^\infty | X_t = i, X_{t+m} = j')}
\]  \hspace{1in} (2.32)

Now, let us bound the terms in (2.32). First,

\[
\frac{Q(X_{t+m} = j, Z_{t+m+k+1}^\infty | X_t = i)}{Q(X_{t+m} = j', Z_{t+m+k+1}^\infty | X_t = i)}
\]

\[
= \sum_{j_0} \frac{Q(X_{t+m+k} = j_0, X_{t+m} = j, Z_{t+m+k+1}^\infty | X_t = i)}{Q(X_{t+m+k} = j_0, X_{t+m} = j', Z_{t+m+k+1}^\infty | X_t = i)}
\]

\[
= \sum_{j_0} a_{ij_0}^m a_{j_0}^k Q(Z_{t+m+k+1}^\infty | X_{t+m+k} = j_0)
\]

\[
\sum_{j_0} a_{ij_0}^m a_{j_0}^k Q(Z_{t+m+k+1}^\infty | X_{t+m+k} = j_0),
\]

where \(a_{rs}^p\) stands for the \((r, s)\)-th element of \(A^p\), the \(p\)-th power of the transition matrix \(A\).

Note that \(a_{ij_0}^m \geq \delta^m\) and \(a_{j_0}^k \geq \delta^k\), \forall i, j_0 from the assumption of \(\Theta_k\). Let \(Q(Z_{t+m+k+1}^\infty | X_{t+m+k} = j_0) = \alpha_{j_0}\). Then, the last expression is

\[
\frac{a_{ij}^m \sum_{j_0} a_{ij_0}^k \alpha_{j_0}}{a_{ij}^m \sum_{j_0} a_{j_0}^k \alpha_{j_0}}.
\]  \hspace{1in} (2.33)

Since

\[
\frac{\sum_{j_0} a_{ij_0}^k \alpha_{j_0}}{\sum_{j_0} a_{j_0}^k \alpha_{j_0}} = \frac{\sum_{j_0} \alpha_{j_0} a_{ij_0}^k a_{j_0}^k}{\sum_{j_0} \alpha_{j_0} a_{j_0}^k \alpha_{j_0} \alpha_{j_0}^k} \leq \max_{j_0} \left( \frac{a_{ij_0}^k}{a_{j_0}^k} \right),
\]

we have

\[
(2.33) \leq \frac{a_{ij}^m}{a_{ij}^m} \max_{j_0} \left( \frac{a_{ij_0}^k}{a_{j_0}^k} \right) \leq \max_{i,j,j_0} \left( \frac{a_{ij}^m a_{ij_0}^k}{a_{j_0}^k a_{j_0}^k} \right) \leq \frac{1}{\delta^{m+k}}. \hspace{1in} (2.34)
\]
Now, let us look at the second term in (2.32). Suppose $\mathcal{T} = \{t + 1, \ldots, t + m + k\}\setminus \{t + m\}$, and let $x_\mathcal{T}$ stand for the sequence of $x_t$'s where $t' \in \mathcal{T}$. Then,

$$Q(Z_{t+1+m+k} | X_t = i, X_{t+m} = j)$$

$$= \frac{\sum_{x_\mathcal{T}} Q(Z_{t+1+m+k} | X_t = i, X_{t+m} = j, X_\mathcal{T} = x_\mathcal{T}) \cdot Q(x_\mathcal{T} \mid i, j)}{\sum_{x_\mathcal{T}} Q(Z_{t+1+m+k} | X_t = i, X_{t+m} = j', X_\mathcal{T} = x_\mathcal{T}) \cdot Q(x_\mathcal{T} \mid i, j')}$$

$$\leq \frac{1}{(\Pi_{\text{min}})^{m+k}}, \tag{2.35}$$

where $Q(x_\mathcal{T} \mid i, j)$ stands for the conditional probability $Q(X_\mathcal{T} = x_\mathcal{T} | X_t = i, X_{t+m} = j)$. Thus, from (2.34) and (2.35),

$$\text{(2.32)} \leq \frac{1}{(\delta \cdot \Pi_{\text{min}})^{m+k}}.$$

Let now $\rho_j \triangleq Q(X_{t+m} = j | X_t = i, Z_{\infty}),$ then $1 = \rho_j + \sum_{j' \neq j} \rho_{j'} \leq \rho_j + (M - 1) \frac{\rho_j}{(\delta \cdot \Pi_{\text{min}})^{m+k}},$ and thus, $\rho_j \geq (1 + \frac{M-1}{(\delta \cdot \Pi_{\text{min}})^{m+k}})^{-1},$ which proves the lemma. \[\square\]

**Lemma 7** Suppose when $\mathcal{T}$ is a set of time indices, $x_\mathcal{T}$ and $z_\mathcal{T}$ stand for the sequences of $x_t$'s and $z_t$'s where $t' \in \mathcal{T}$. Now, consider following two arbitrarily given sets.

$$C_t \triangleq x_t^\infty \triangleq \left\{ x_\mathcal{T} : \mathcal{T} \subseteq \mathbb{Z}_2 \cup \{\infty\} \right\} \quad \text{and}$$

$$D \triangleq Z_{\infty}^\infty \triangleq \left\{ z_\mathcal{T} : \mathcal{T} \subseteq \mathbb{Z} \cup \{\infty, -\infty\} \right\}.$$

For $d \in \mathbb{N}$, define

$$M_d^+ \triangleq \max_i Q(C_t \mid X_{t-dm} = i, D),$$

$$M_d^- \triangleq \min_i Q(C_t \mid X_{t-dm} = i, D).$$

Then,

$$M_d^+ - M_d^- \leq (\rho_{5,k,m})^{d-1},$$

where $\rho_{5,k,m} = 1 - 2\mu_{5,k,m}$. 

CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

Proof: From the argument of Lemma 6, it is easy to see that

\[ Q(X_{t+m} = j | X_t = i, D) \geq \mu_{\delta,k,m}, \]

independent of \( D \), too. Now, define

\[ \gamma_i(d) \triangleq Q(C_t | X_{t-dm} = i, D) \]
\[ \beta_{ij}(d) \triangleq Q(X_{t-dm} = j | X_{t-(d+1)m} = i, D) \]
\[ i^+(d) \triangleq \arg\max_i Q(C_t | X_{t-(d+1)m} = i, D) \]
\[ i^-(d) \triangleq \arg\min_i Q(C_t | X_{t-dm} = i, D). \]

Since \( \delta, k \) and \( m \) are fixed, let us simply denote \( \mu = \mu_{\delta,k,m} \). Also, let us omit \( d \) and the parenthesis for above four quantities to simplify notation. Then,

\[ M_{d+1}^+ = Q(C_t | X_{t-(d+1)m} = i^+, D) = \sum_j \gamma_j \beta_{i+j} \]
\[ = \mu M_d^- + (\beta_{i^+} - \mu) M_d^- + \sum_{j \neq i^-} \gamma_j \beta_{i+j} \]
\[ \leq \mu M_d^- + (\beta_{i^+} - \mu) M_d^+ + \sum_{j \neq i^-} \beta_{i+j} M_d^+ \]
\[ = \mu M_d^- + (1 - \mu) M_d^+, \quad (2.37) \]

where (2.36) is possible from Lemma 6, since \( \beta_{ij} \geq \mu \) for all \( i, j \). By the similar argument, we get

\[ M_{d+1}^- \geq \mu M_d^+ + (1 - \mu) M_d^- + (2.38) \]

By subtracting (2.38) from (2.37), we get

\[ M_{d+1}^+ - M_{d+1}^- \leq (1 - 2\mu)(M_d^+ - M_d^-) \leq \cdots \leq (1 - 2\mu)^d, \]

and, thus, proves the lemma. Note that since \( \mu = \mu_{\delta,k,m} < \frac{1}{2} \), and thus, \( 0 < \rho_{\delta,k,m} < 1 \). Also, the result does not depend on \( Q \). \( \blacksquare \)
Lemma 8 For all $p, d \geq 1$, and $0 \leq l \leq m - 1$,

$$|Q(C_{t} | Z_{t-1}^{p} - dm - l) - Q(C_{t} | Z_{t-(d+1)m-l}^{p})| \leq (\rho_{d,k,m})^{d+1}. $$

Proof: By conditioning,

$$Q(C_{t} | Z_{t-(d+1)m-l}^{p}) = \sum_{j} Q(C_{t} | Z_{t-(d+1)m-l}^{p}, X_{t-(d+2)m} = j) \cdot Q(X_{t-(d+2)m} = j | Z_{t-(d+1)m-l}^{p}),$$

and, therefore,

$$M_{d+2}^{+} \leq Q(C_{t} | Z_{t-(d+1)m-l}^{p}) \leq M_{d+2}^{-}.$$ 

On the other hand,

$$Q(C_{t} | Z_{t-1}^{p} - dm - l) = \sum_{z_{t-dm-l}^{l-1}} Q(C_{t} | Z_{t-(d+1)m-l}^{p}) \cdot Q(Z_{t-(d+1)m-l-1}^{l-1} = z_{t-(d+2)m-l}^{l-1} | Z_{t-dm-l}^{p}),$$

and, thus,

$$M_{d+2}^{+} \leq Q(C_{t} | Z_{t-dm-l}^{p}) \leq M_{d+2}^{-}.$$ 

Therefore, from Lemma 7, we have

$$|Q(C_{t} | Z_{t-dm-l}^{p}) - Q(C_{t} | Z_{t-(d+1)m-l}^{p})| \leq M_{d+2}^{+} - M_{d+2}^{-} \leq (\rho_{d,k,m})^{d+1}. $$

Note that the result does not depend on either $Q$ or $l$.  ■
2.7.2 Proof of Lemma 3

Before proving Lemma 3 we need the following lemma first. Part (b), (c), and (d) are crucial for Lemma 3, and Part (a) enables Part (b). The continuity result of $\hat{X}_Q(\varepsilon_{t-1}^0)$ in Part (a) is the key reason why we need the randomization of the filter.

**Lemma 9** Suppose $Q \in \Theta_k^\delta$ and fix $\delta > 0$.

(a) We have

$$\|\hat{X}_Q(\varepsilon_{t-1}^0) - \hat{X}_Q(\varepsilon_{t-2}^0)\|_1 \leq M^2 \cdot \|Q_{X_0|z_0} - Q_{X_0|z_2}^0\|_1,$$

where $t_1, t_2 > 0$ are arbitrary integers. That is, for any integer $t > 0$ and any individual sequence $\varepsilon_{t-1}^0$, $\hat{X}_Q(\varepsilon_{t-1}^0)$ is a Lipschitz continuous function in $Q_{X_0|z_0}^0$.

(b) $\ell(X_0, \hat{X}_Q(\varepsilon_{t-1}^0)) \rightarrow \ell(X_0, \hat{X}_Q(\varepsilon_{\infty}^0))$ a.s. uniformly on $\Theta_k^\delta$.

(c) For all $Q \in \Theta_k^\delta$, and for all $\omega$, there exist $0 < \gamma < 1$, $\beta > 0$, such that $|Q(X_0|Z_{\infty}^0) - Q(X_0|Z_{\infty}^0)| < \beta \gamma$.

(d) For fixed $t, \eta > 0$, there exists some finite set $\mathcal{F}_k(t, \eta) \subset \Theta_k^\delta$, such that

$$\max_{Q \in \Theta_k^\delta} \min_{Q' \in \mathcal{F}_k(t, \eta)} \max_{x_0, z_{t-1}^0} |Q(x_0|z_{t-1}^0) - Q'(x_0|z_{t-1}^0)| \leq \eta.$$

**Proof:**

(a) For given simplex vector $Q$, fixed $\hat{x}$, and $B_c$ defined as in Section 2.4.1, we define followings.

- $S_{\hat{x}}(Q) \triangleq \{W \in B_c : B(Q + W) = \hat{x}\}$
- $DP(\hat{x}) \triangleq \{c : c = \lambda_{\hat{x}} - \lambda_a, \text{ for all } a \in \mathcal{A}\setminus\{\hat{x}\}\}$
- $\text{dist}(Q, c^T y = 0) \triangleq \min_{y \in \mathcal{Y}} \|Q - y\|_2$

In words, $S_{\hat{x}}(Q)$ is a set of vectors in $c$-ball, $B_c$, that makes the Bayes response $B(Q + W)$ equal to $\hat{x}$; $DP(\hat{x})$ is a set of normal vectors that define the decision
planes \( \{ \mathbf{y} \in \mathbb{R}^M : \mathbf{c}^T \mathbf{y} = 0, \mathbf{c} \in DP(\hat{x}) \} \) which separate the reconstruction alphabet \( \hat{x} \) and other alphabets, and \( \text{dist}(\mathbf{Q}, \mathbf{c}^T \mathbf{y} = 0) \) is the shortest \( L_2 \) distance from a simplex vector \( \mathbf{Q} \) to the plane \( \{ \mathbf{y} \in \mathbb{R}^M : \mathbf{c}^T \mathbf{y} = 0 \} \). Then, for some fixed \( t \), by definition,

\[
\hat{X}_t^e(z_{-t}^0)[\hat{x}] = \frac{\text{Vol}(S_x(Q_{X_0|z_{-t}^0}))}{\text{Vol}(B_e)},
\]

where \( \text{Vol}(\cdot) \) is a volume of a set. Since \( \text{Vol}(B_e) \) is a constant, for any \( t_1 \) and \( t_2 \), we have

\[
\left| \hat{X}_t^e(z_{-t_1}^0)[\hat{x}] - \hat{X}_t^e(z_{-t_2}^0)[\hat{x}] \right| = \frac{|\text{Vol}(S_x(Q_{X_0|z_{-t_1}^0})) - \text{Vol}(S_x(Q_{X_0|z_{-t_2}^0}))|}{\text{Vol}(B_e)}. \tag{2.39}
\]

For the numerator, as a crude bound, we obtain

\[
|\text{Vol}(S_x(Q_{X_0|z_{-t_1}^0})) - \text{Vol}(S_x(Q_{X_0|z_{-t_2}^0}))| \
\leq \text{Vol}(B_e^{M-1}) \sum_{\mathbf{c} \in DP(\hat{x})} \left| \text{dist}(Q_{X_0|z_{-t_1}^0}, \mathbf{c}^T \mathbf{y} = 0) \right| \
- \text{dist}(Q_{X_0|z_{-t_2}^0}, \mathbf{c}^T \mathbf{y} = 0), \tag{2.40}
\]

where \( B_e^{M-1} = \{ \mathbf{U} \in \mathbb{R}^{M-1} : \| \mathbf{U} \|_2 \leq e \} \). Since

\[
\text{dist}(\mathbf{Q}, \mathbf{c}^T \mathbf{y} = 0) = \frac{|\mathbf{c}^T \mathbf{Q}|}{\| \mathbf{c} \|_2},
\]
we have

\[ \text{dist}(Q_{X_0}|z_{-t_1}^0, c^T y = 0) - \text{dist}(Q_{X_0}|z_{-t_2}^0, c^T y = 0) \]
\[ \leq \frac{|c^T Q_{X_0}|z_{-t_1}^0| - |c^T Q_{X_0}|z_{-t_2}^0|}{\|c\|_2} \]
\[ \leq \frac{|c^T (Q_{X_0}|z_{-t_1}^0 - Q_{X_0}|z_{-t_2}^0)|}{\|c\|_2} \]
\[ \leq \|Q_{X_0}|z_{-t_1}^0 - Q_{X_0}|z_{-t_2}^0\|_2 \]
\[ \leq \|Q_{X_0}|z_{-t_1}^0 - Q_{X_0}|z_{-t_2}^0\|_1, \]

where (2.41) follows from the triangular inequality; (2.42) follows from Cauchy-Schwartz inequality, and (2.43) follows from the fact that \( L_2 \)-norm is less than or equal to \( L_1 \)-norm. Therefore, (2.40) becomes

\[ |\text{Vol}(S_\hat{X}(Q_{X_0}|z_{-t_1}^0)) - \text{Vol}(S_\hat{X}(Q_{X_0}|z_{-t_2}^0))| \]
\[ \leq M \cdot \text{Vol}(B_\epsilon^{M-1}) \cdot \|Q_{X_0}|z_{-t_1}^0 - Q_{X_0}|z_{-t_2}^0\|_1, \]

and, thus, (2.39) becomes

\[ |\hat{X}_Q(z_{-t_1}^0)[\hat{x}] - \hat{X}_Q(z_{-t_2}^0)[\hat{x}]| \]
\[ \leq M \cdot \frac{\text{Vol}(B_\epsilon^{M-1})}{\text{Vol}(B_\epsilon)} \cdot \|Q(X_0|z_{-t_1}^0) - Q(X_0|z_{-t_2}^0)\|_1. \]
\[ \leq M \cdot \|Q_{X_0}|z_{-t_1}^0 - Q_{X_0}|z_{-t_2}^0\|_1. \]

Therefore, we have

\[ \|\hat{X}_Q(z_{-t_1}^0) - \hat{X}_Q(z_{-t_2}^0)\|_1 \leq M^2 \cdot \|Q_{X_0}|z_{-t_1}^0 - Q_{X_0}|z_{-t_2}^0\|_1, \]

and Part (a) is proved.

(b) By the exact same argument as in proving Lemma 1, we can easily know that
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

$Q(X_0|Z^0_{-t}) \rightarrow Q(X_0|Z^0_{-\infty})$ for all $\omega$, uniformly on $\Theta_k$. Since we have

$$
\left| \ell(X_0, \hat{X}_Q(Z^0_{-t})) - \ell(X_0, \hat{X}_Q(Z^0_{-\infty})) \right|
= \left| \sum_{\bar{x}} A(X_0, \bar{x}) \left( \hat{X}_Q(Z^0_{-t})[\bar{x}] - \hat{X}_Q(Z^0_{-\infty})[\bar{x}] \right) \right|
\leq A_{\max} \| \hat{X}_Q(Z^0_{-t}) - \hat{X}_Q(Z^0_{-\infty}) \|_1
\leq A_{\max} M^2 \cdot \| Q(X_0|Z^0_{-t}) - Q(X_0|Z^0_{-\infty}) \|_1,
$$

we get the uniform convergence.

(c) Again, let us follow the argument in the proof of Lemma 1. Suppose $t = jk + l$, where $j = [t/k]$, and $l = t \mod k$. Then,

$$
\left| Q(X_0|Z^0_{-t}) - Q(X_0|Z^0_{-\infty}) \right|
= \left| Q(X_0|Z^0_{-jk-l}) - Q(X_0|Z^0_{-\infty}) \right|
\leq \sum_{i=j}^{\infty} \left| Q(X_0|Z^0_{-ik-l}) - Q(X_0|Z^0_{-(i+1)k-l}) \right|
\leq \sum_{i=j}^{\infty} \rho^{i+1}
= \frac{\rho^{j+1}}{1 - \rho} = \frac{\rho}{1 - \rho} \rho^{[t/k]} = \rho^{1-k} \left( \rho^1 \rho^{-k} \right)^t
\leq \frac{1}{1 - \rho} \left( \rho^1 \rho^{-k} \right)^t,
$$

where $\rho = \rho_{\delta,k,k}$ as defined in Lemma 7, and (2.44) follows from Lemma 8. By letting $\beta = \frac{1}{1 - \rho}$, and $\gamma = \rho^1 / k$, we have proved Part (c).
(d) We know that for the individual sequence pair \((x_0, z^0_t)\),

\[
Q(x_0 | z^0_{-t}) = \frac{\sum_{x_{-t}^0} Q(x_{-t}^0, z^0_{-t})}{Q(x_0^0)} = \frac{\sum_{x_{-t}^0} Q(x_{-t}^0, z^0_{-t})}{\sum_{x_{-t}^0} Q(x_{-t}^0, z^0_{-t})} = \frac{\sum_{x_{-t}^0} Q(x_{-t}^0)Q(z^0_{-t}|x^0_{-t})}{\sum_{x_{-t}^0} Q(x_{-t}^0)Q(z^0_{-t}|x^0_{-t})} = \frac{\sum_{x_{-t}^0} \left( Q(x_{-t}^0) \prod_{i=-t}^0 \Pi(x_i, z_i) \right)}{\sum_{x_{-t}^0} \left( Q(x_{-t}^0) \prod_{i=-t}^0 \Pi(x_i, z_i) \right)}.
\]

For \(Q \in \Theta_k^d\), \(\Pi\) is fixed and we can think of \(\prod_{i=-t}^0 \Pi(x_i, z_i)\) as a constant for the individual sequence pair \((x_0^0, z^0_{-t})\). Since

\[
Q(x_0^0) = Q(x_{-t}^0) \prod_{j=-k}^{-t} a^0_{j-k} x^0_{j+k+1},
\]

\(Q(x_0 | z^0_{-t})\) is the ratio of two finite order polynomials of \(\{a_{ij}\}\), and as \(\Theta_k^d\) is closed and bounded, \(Q(x_0 | z^0_{-t})\) is a uniformly continuous function of \(\{a_{ij}\}\). Therefore, for given \(\eta\), \(\exists \epsilon(\eta)\) such that \(\|Q - Q'\|_1 < \epsilon(\eta)\) implies

\[
\max_{x_0, z^0_{-t}} |Q(x_0 | z^0_{-t}) - Q'(x_0 | z^0_{-t})| \leq \eta,
\]

since there are only finite number of possible \((x_0, z^0_{-t})\) pairs. Also, since \(\Theta_k^d\) is compact, we can always find a finite set, \(\mathcal{F}_k(t, \eta)\) that for any \(Q \in \Theta_k^d\), there exists at least one \(Q' \in \mathcal{F}_k(t, \eta)\), that satisfies \(\|Q - Q'\|_1 < \epsilon(\eta)\). Therefore, Part (d) is proved.
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

Proof of Lemma 3: To prove Lemma 3, first consider following limit.

\[ \lim_{n \to \infty} E\left( L_{\hat{X}_Q}(X^n, Z^n) \right) \]
\[ = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} E\left( \ell(X_t, \hat{X}_Q(Z')) \right) \]
\[ = \lim_{t \to \infty} E\left( \ell(X_t, \hat{X}_Q(Z')) \right) \quad (2.47) \]
\[ = \lim_{t \to \infty} E\left( \ell(X_0, \hat{X}_Q(Z_0^{(t-1)})) \right) \quad (2.48) \]
\[ = E\left( \ell(X_0, \hat{X}_Q(Z_{-\infty}^{0})) \right) \text{ uniformly on } \Theta_k, \quad (2.49) \]

where (2.47) is from Cesáro’s mean convergence theorem, (2.48) is from stationarity, and (2.49) is from Lemma 9(b) and bounded convergence theorem. Thus, to complete the proof, we need to show that

\[ \lim_{n \to \infty} L_{\hat{X}_Q}(X^n, Z^n) = E\left( \ell(X_0, \hat{X}_Q(Z_{-\infty}^{0})) \right) \quad \text{a.s.} \quad (2.50) \]

uniformly on \( \Theta_k \). Now, let us show the pointwise convergence in (2.50) without the uniformity by using ergodic theorem. For given \( Q \), define

\[ g_{t,Q}(X, Z) \triangleq \ell(X_0, \hat{X}_Q(Z_{(t-1)}^{0})) \]
\[ g_Q(X, Z) \triangleq \ell(X_0, \hat{X}_Q(Z_{-\infty}^{0})), \]

and denote by \( T \) the shift operator. Then, what we should prove becomes

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} g_{t,Q}(T^t(X, Z)) = E\left( g_Q(X, Z) \right) \quad \text{a.s.,} \]

while the ergodic theorem gives

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} g_Q(T^t(X, Z)) = E\left( g_Q(X, Z) \right) \quad \text{a.s.} \]
Observe that
\[
\left| \frac{1}{n} \sum_{t=1}^{n} g_{t,Q}(T_t(X,Z)) - \frac{1}{n} \sum_{t=1}^{n} g_Q(T_t(X,Z)) \right|
\leq \frac{1}{n} \sum_{t=1}^{n} \left| g_{t,Q}(T_t(X,Z)) - g_Q(T_t(X,Z)) \right|
= \frac{1}{n} \sum_{t=1}^{n} \left| \ell(X_t, \hat{X}_Q^t(Z_t^t)) - \ell(X_t, \hat{X}_Q^t(Z_{t-\infty}^t)) \right|.
\]

Since Lemma 9(c) holds for all \( \omega \), we can think that the lemma holds for all individual sequence pair \((x_0, z_{-\infty}^0)\). Thus, it holds for all individual pair \((x_t, z_{-\infty}^t)\), too, and we can conclude that \( Q(X_t|Z_t^t) \to Q(X_t|Z_{-\infty}^t) \) for all \( \omega \) as \( t \to \infty \). Hence, by exactly the same argument as Lemma 9(a) and Lemma 9(b), we conclude that \( \ell(X_t, \hat{X}_Q^t(Z_t^t)) \to \ell(X_t, \hat{X}_Q^t(Z_{-\infty}^t)) \) almost surely as \( t \to \infty \). Now, by Cesáro’s mean convergence theorem, we obtain
\[
\frac{1}{n} \sum_{t=1}^{n} \left| \ell(X_t, \hat{X}_Q^t(Z_t^t)) - \ell(X_t, \hat{X}_Q^t(Z_{-\infty}^t)) \right| \to 0 \quad \text{a.s.}
\]

Therefore, we get
\[
L_{\hat{X}_Q}(X^n, Z^n) \to E\left( \ell(X_0, \hat{X}_Q^0(Z_{-\infty}^0)) \right) \quad \text{a.s.}
\]

Note that up to this point we cannot guarantee the uniformity of the convergence, since the ergodic theorem only gives the individual convergence for each \( Q \). To show the uniformity of the convergence in (2.50), first define the following quantity for some fixed integer \( t \in [1, n - 1] \),
\[
L_{\hat{X}_Q^t}(X^n, Z^n) = \frac{1}{n} \left( \sum_{i=1}^{t} \ell(X_i, \hat{X}_Q^t(Z_i^t)) + \sum_{i=t+1}^{n} \ell(X_i, \hat{X}_Q^t(Z_{i-1}^t)) \right).
\]
From Lemma 9(d), for any $Q \in \Theta^k_\epsilon$ and fixed $t, \eta > 0$, we can pick some $Q' \in \mathcal{F}_k(t, \eta)$ such that $\|Q - Q'\|_1 < \epsilon(\eta)$, and thus,

$$\max_{x_0, z_{-t}^0} |Q(x_0|z_{-t}^0) - Q'(x_0|z_{-t}^0)| \leq \eta.$$ 

By adding and subtracting some common terms involving such $Q'$, and from the triangle inequality, we have,

\[
\begin{align*}
|L_{X^Q_t}(X^n, Z^n) - E(\ell(X_0, \hat{X}^c_Q(Z_{-\infty})))| &
\leq |L_{X^Q_t}(X^n, Z^n) - L_{\hat{X}^Q_t}(X^n, Z^n)| \\
&+ |L_{\hat{X}^Q_t}(X^n, Z^n) - L_{\hat{X}^Q'_{Q,t}}(X^n, Z^n)| \\
&+ |L_{\hat{X}^Q'_{Q,t}}(X^n, Z^n) - L_{\hat{X}^Q'_{Q'}}(X^n, Z^n)| \\
&+ |L_{\hat{X}^Q_{Q'}}(X^n, Z^n) - E(\ell(X_0, \hat{X}^c_{Q'}(Z_{-\infty})))| \\
&+ E(\ell(X_0, \hat{X}^c_{Q'}(Z_{-\infty}))) - E(\ell(X_0, \hat{X}^c_Q(Z_{-\infty})))|.
\end{align*}
\] 

(2.51)

Now, the goal becomes to show that the terms in the righthand side of the inequality converges to zero independent of $Q$ as $n$, $t$, and $\eta$ varies. First, we will bound each term, and send $n \to \infty$. 

CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

(i)

\[
|L_{X_Q^i}(X^n, Z^n) - L_{X_{Q',i}}(X^n, Z^n)|
\leq \frac{1}{n} \sum_{i=t+1}^{n} \left| \ell(X_i, \hat{X}_Q^i(Z^n)) - \ell(X_i, \hat{X}_{Q'}^i(Z_{i-t}^i)) \right|
\leq \Lambda_{\text{max}} \cdot \frac{1}{n} \sum_{i=t+1}^{n} \left\| \hat{X}_Q^i(Z^n) - \hat{X}_{Q'}^i(Z_{i-t}^i) \right\|_1
\leq \Lambda_{\text{max}} M^2 \cdot \frac{1}{n} \sum_{i=t+1}^{n} \left\| Q_{X_0|Z_{i-t}^i} - Q_{X_0|Z_{i-t}^i} \right\|_1
\leq \Lambda_{\text{max}} M^3 \cdot \frac{1}{n} \sum_{i=t+1}^{n} (\beta \gamma^t + \beta \gamma^i)
\rightarrow \Lambda_{\text{max}} M^3 \beta \gamma^t \quad \text{a.s. uniformly on } \Theta_k^3,
\]

where (2.52) follows from stationarity and Lemma 9(a); (2.53) follows from Lemma 9(c), and (2.54) follows from the Cesàro's mean convergence theorem. Since (2.53) does not depend on \(Q\), the limit is uniform on \(\Theta_k^3\).

(ii)

\[
|L_{X_{Q',i}}(X^n, Z^n) - L_{X_{Q',i}}(X^n, Z^n)|
\leq \frac{1}{n} \sum_{i=t+1}^{n} \left| \ell(X_i, \hat{X}_{Q'}^i(Z_{i-t}^i)) - \ell(X_i, \hat{X}_{Q'}^i(Z_{i-t}^i)) \right| + \frac{t \Lambda_{\text{max}}}{n}
\leq \Lambda_{\text{max}} \cdot \frac{1}{n} \sum_{i=t+1}^{n} \left\| \hat{X}_Q^i(Z_{i-t}^i) - \hat{X}_{Q'}^i(Z_{i-t}^i) \right\|_1 + \frac{t \Lambda_{\text{max}}}{n}
\leq \Lambda_{\text{max}} M^2 \cdot \frac{1}{n} \sum_{i=t+1}^{n} \left\| Q_{X_0|Z_{i-t}^i} - Q_{X_0|Z_{i-t}^i} \right\|_1 + \frac{t \Lambda_{\text{max}}}{n}
\leq \Lambda_{\text{max}} M^3 \frac{n-t}{n} \cdot \eta + \frac{t \Lambda_{\text{max}}}{n}
\rightarrow \Lambda_{\text{max}} M^3 \eta \quad \text{a.s. uniformly on } \Theta_k^3,
\]

where (2.55) follows from Lemma 9(a), and (2.56) follows from Lemma 9(d). Since (2.56) does not depend on \(Q\), the limit is also uniform on \(\Theta_k^3\).
(iii) \[
\left| L_{x_{Q,t}}^e (X^n, Z^n) - L_{x_{Q}^e} (X^n, Z^n) \right| \to \Lambda_{\max} M^3 \beta \gamma^{t} \ a.s.,
\]
by following the same argument as (i). Since \( \mathcal{F}_k(t, \eta) \) is finite, this convergence is uniform on \( \mathcal{F}_k(t, \eta) \).

(iv) \[
\left| L_{x_{Q}^e} (X^n, Z^n) - E \left( \ell(X_0, \hat{X}_{Q}^e (Z_{-\infty}^0)) \right) \right| \to 0 \ a.s.,
\]
from the proof of pointwise convergence above. As in (iii), this convergence is also uniform on \( \mathcal{F}_k(t, \eta) \).

(v) \[
\left| E \left( \ell(X_0, \hat{X}_{Q}^e (Z_{-\infty}^0)) \right) - E \left( \ell(X_0, \hat{X}_{Q}^e (Z_{-\infty}^0)) \right) \right|
\leq \left| E \left( \ell(X_0, \hat{X}_{Q}^e (Z_{-\infty}^0)) \right) - E \left( \ell(X_0, \hat{X}_{Q}^e (Z_{-l}^0)) \right) \right|
+ \left| E \left( \ell(X_0, \hat{X}_{Q}^e (Z_{-l}^0)) \right) - E \left( \ell(X_0, \hat{X}_{Q}^e (Z_{-\infty}^0)) \right) \right|
\leq \sum_{x_0, z_{-\infty}^0} P(x_0, z_{-\infty}^0) \left| \ell(x_0, \hat{X}_{Q}^e (z_{-\infty}^0)) - \ell(x_0, \hat{X}_{Q}^e (z_{-l}^0)) \right|
+ \sum_{x_0, z_{-l}^0} P(x_0, z_{-l}^0) \left| \ell(x_0, \hat{X}_{Q}^e (z_{-l}^0)) - \ell(x_0, \hat{X}_{Q}^e (z_{-\infty}^0)) \right|
+ \sum_{x_0, z_{-\infty}^0} P(x_0, z_{-\infty}^0) \left| \ell(x_0, \hat{X}_{Q}^e (z_{-\infty}^0)) - \ell(x_0, \hat{X}_{Q}^e (z_{-l}^0)) \right|
\leq \Lambda_{\max} M^3 \left( 2 \beta \gamma^{t} + \eta \right),
\]
by similar argument as in (i) and (ii).

Therefore, by taking limit supremum on both side of (2.51), we get
\[
\limsup_{n \to \infty} \left| L_{x_{Q}^e} (X^n, Z^n) - E \left( \ell(X_0, \hat{X}_{Q}^e (Z_{-\infty}^0)) \right) \right|
\leq \Lambda_{\max} M^3 \left( 4 \beta \gamma^{t} + 2 \eta \right) \ a.s.
uniformly on $\Theta^t_k$. Since $t$ and $\eta$ are arbitrary, by sending $t \to \infty$ and $\eta \downarrow 0$, we have
\[
\limsup_{n \to \infty} \left| L_{X_q^t}(X^n, Z^n) - E\left( \ell(X_0, \hat{X}^t_q(Z^n)) \right) \right| \leq 0 \quad \text{a.s.}
\]
uniformly on $\Theta^t_k$. Thus, the lemma is proved. 

2.7.3 Proof of Corollary 1

Proof of Corollary 1: First note the subtle point that Corollary 1 does not directly follow from Lemma 3. Since the probability law $Q^t_k$ that we are using to filter each block is changing every block, whereas the uniform convergence in Lemma 3 is for the fixed $Q \in \Theta^t_k$ for all $t$, it is not enough to guarantee the Corollary. However, since $Q^t_k$ remains the same within each block, we can still use the result of Lemma 3 if the block length gets long enough. Keeping this in mind, let us take a more careful look at each block. In the proof, for the brevity of notation, let us denote
\[
\ell_t(Q) \triangleq \ell(X_t, \hat{X}^t_q(Z^n)),
\]
since we are always dealing with the randomized filter, and there is no possibility of confusion. Now, fix any $\delta > 0$. Then, from (2.4), there exists some $I$, such that
\[
\frac{m_{I-1}}{m_I} < \frac{\delta}{8\Lambda_{\max}},
\]
and from Lemma 3, there exists some $N$ such that for all $n \geq N$,
\[
\max_{Q \in \Theta^t_k} \left| L_{X_q^t}(X^n, Z^n) - E L_{X_q^t}(X^n, Z^n) \right| < \delta/4. \tag{2.57}
\]
CHAPTER 2. UNIVERSAL FILTERING VIA HIDDEN MARKOV MODELING

Recalling the definition \( i(t) \triangleq \max\{i : m_i \leq t\} \), we let \( I_0 = \max(I, i(N) + 1) \). Then, for any \( n \geq m_{i_0} \), and \( m_{i(n)} \leq n < m_{i(n)+1} \),

\[
\left| L_{X_{\text{univ}}^k}(X^n, Z^n) - \tilde{E} L_{X_{\text{univ}}^k}(X^n, Z^n) \right| \\
\leq \frac{1}{n} \sum_{t=1}^{m_{i(n)}-1} \left( \ell_t(Q_k^t) - \tilde{E}(\ell_t(Q_k^t)) \right) \\
+ \frac{1}{n} \sum_{t=m_{i(n)}-1+1}^{m_{i(n)}} \left( \ell_t(\tilde{Q}[Z^{m_{i(n)}-1}]) - \tilde{E}(\ell_t(\tilde{Q}[Z^{m_{i(n)}-1}])) \right) \\
+ \frac{1}{n} \sum_{t=m_{i(n)}+1}^{n} \left( \ell_t(\tilde{Q}[Z^{m_{i(n)}-1}]) - \tilde{E}(\ell_t(\tilde{Q}[Z^{m_{i(n)}-1}])) \right). \tag{2.58}
\]

Note that in the second and third term, \( Q_k^t \) is fixed to \( \tilde{Q}[Z^{m_{i(n)}-1}] \) and \( \tilde{Q}[Z^{m_{i(n)}-1}] \) from the definition of our filter. Figure 2.1 summarizes the time line with above notations.

![Figure 2.1: The time line](image)

Now, we can bound each term in (2.58). For the first term, since \( n \geq m_{i(n)} \geq m_I \), we know that \( \frac{m_{i(n)} - 1}{n} \leq \frac{m_{i(n)} - 1}{m_{i(n)}} \leq \frac{\delta}{8\Lambda_{\text{max}}} \). Therefore,

\[
\frac{1}{n} \left| \sum_{t=1}^{m_{i(n)}-1} \left( \ell_t(Q_k^t) - \tilde{E}(\ell_t(Q_k^t)) \right) \right| \leq \frac{\delta}{8\Lambda_{\text{max}}} \cdot \Lambda_{\text{max}} = \frac{\delta}{8}.
\]
For the second term, since \( n \geq m_{i(n)} \geq N \), and from (2.57),

\[
\frac{1}{n} \left| \sum_{t=m_{i(n)}+1}^{m_{i(n)}} \left( \ell_t(\hat{Q}[Z^{m_{i(n)}-1}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}-1}])) \right) \right|
\leq \frac{m_{i(n)}}{n} \frac{1}{m_{i(n)}} \left| \sum_{t=1}^{m_{i(n)}} \left( \ell_t(\hat{Q}[Z^{m_{i(n)}-1}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}-1}])) \right) \right|
+ \frac{1}{n} \left| \sum_{t=1}^{m_{i(n)}-1} \left( \ell_t(\hat{Q}[Z^{m_{i(n)}-1}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}-1}])) \right) \right|
\leq \frac{\delta}{4} + \frac{\delta}{8\Lambda_{\text{max}}} \cdot \Lambda_{\text{max}} = \frac{3\delta}{8}.
\]

Finally, for the last term,

\[
\frac{1}{n} \left| \sum_{t=m_{i(n)}+1}^{n} \left( \ell_t(\hat{Q}[Z^{m_{i(n)}-1}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}-1}])) \right) \right|
\leq \frac{1}{n} \left| \sum_{t=1}^{n} \left( \ell_t(\hat{Q}[Z^{m_{i(n)}-1}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}-1}])) \right) \right|
+ \frac{1}{n} \left| \sum_{t=1}^{m_{i(n)}-1} \left( \ell_t(\hat{Q}[Z^{m_{i(n)}-1}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}-1}])) \right) \right|
\leq \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2}.
\]

Therefore, for any \( n \geq m_{i_0} \), and \( m_{i(n)} \leq n \leq m_{i(n)+1} \), we have

\[
\left| L_{X_{\text{univ},k}}(X^n, Z^n) - \hat{E} L_{X_{\text{univ},k}}(X^n, Z^n) \right| < \delta,
\]

and since \( \delta \) was arbitrary, we have the corollary. \( \blacksquare \)
Chapter 3

Universal FIR MMSE filtering

3.1 Introduction

We now switch our focus to the real-valued signal setting as opposed to the finite-alphabet setting in Chapter 2. Estimating the real-valued components of a signal corrupted by zero mean real-valued additive noise is a fundamental problem in signal processing and estimation theory. When the underlying signal is a stationary process, the usual criterion for the estimation is the mean square error (MSE), and much work on minimum MSE (MMSE) estimation has been done since Wiener [33]. Moreover, due to the ease of implementation, linear MMSE estimation has been popular for many decades [34]. There are noncausal and causal versions of linear MMSE estimation, and in the signal processing literature, the term filtering is used for both cases. However, in this chapter, we will only use that term for causal estimation and again refer to a filter as the causal estimator. The most common form of the linear MMSE filter is the finite-duration impulse response (FIR) filter, since stability is not an issue and it is easy to implement.

In practice, there are two limitations in building the linear MMSE estimators. One is that we need prior knowledge of the first and second moment of the signal which we usually do not have. The other, which may be more severe, is that we need stationarity assumptions on the underlying signal, whereas in practice the signal may be nonstationary, or even non-stochastic in many cases. In this chapter, we will focus
on FIR MMSE filters, and try to tackle these limitations jointly.

Robust minimax [35][36][37] and adaptive filtering [38] are approaches that have been taken to deal with the above limitations. The former aims to optimize for the worst case in the signal uncertainty set, to get a robust estimator. However, this approach ignores the fact that we can learn about the signal, and most of them allow large delay in estimation, i.e., noncausal estimation, which is not applicable in filtering problems that have strict causality constraints. On the other hand, adaptive filtering tries to build an FIR filter that sequentially updates its filter coefficients by learning from the noisy observation and a desired response signal, which the filter output aims to approach. However, this is also not directly applicable to our setting of filtering the underlying signal, since the desired response signal, which is the underlying signal itself, is not available to the filter. Unsupervised adaptive filtering [39] considered the case where the desired response signal is not available, but certain statistical assumptions on the underlying signal were needed. Hence, when there is no knowledge about the statistical property of the underlying signal, or when the underlying signal is not a stochastic process, it is not clear how we can apply the above approaches.

Instead, we take an on-line learning approach, whereby we do not assume any stochastic mechanism in generating the underlying signal. Unlike the underlying signal, we do make assumptions on the noise, i.e., we assume that the noise is additive zero mean, time-independent, bounded, and the variance of the noise is known to the filter. Note that the known noisy channel assumption is even weaker here due to the special problem setting of linear filtering and MSE criterion. As described in Section 1.1, the assumption of known noise variance is not too stringent in practice given that the noise is time-independent. That is, by sending some training sequence before the filtering process begins, we can have a good estimate on the noise variance by taking the sample variance of the noise and assuming that the noise variance is known. Given above assumptions, we build a filter that performs essentially as well as the best FIR filter which is tuned to the actual underlying sequence, as the length of the observation sequence increases, regardless of what that underlying sequence may be. We obtain performance guarantees pertaining both to the expected and the actual MSEs. By doing so, we overcome the two limitations mentioned above, guaranteeing uniformly
CHAPTER 3. UNIVERSAL FIR MMSE FILTERING

good performance for every possible underlying individual signal. This individual sequence setting result is strong enough to imply the conventional stochastic setting result as well, namely, when the underlying signal is assumed to be stationary, the performance of our filter achieves the performance of the optimal FIR filter. A more precise problem formulation will be given in Section 3.2.

Our on-line learning approach for FIR MMSE filtering is intimately related to two lines of research in information theory and learning theory. One is the universal filtering problem, also known as sequential compound decision problem, which is the problem of causally estimating the finite alphabet individual sequence based on the Discrete Memoryless Channel (DMC) corrupted noisy observation. This problem has been initiated and was the focus of much attention in 1950's and 1960's [40][41][42]. Recently, there has been resurgent interest in this area. For example, [8] establishes a connection between universal filtering and universal prediction [4]. The other related problem area is the competitive on-line linear regression problem for real-valued data, which is the problem of estimating the signal components based on past side information-signal pairs and current side information. [43] has developed on-line linear regressors for square error loss that compete with finite order linear regressors, and [44] extended this to the universal linear least squares prediction problem for real-valued data. Our work is an extension of both problems, i.e., an extension of the universal filtering problem to the case of real-valued individual sequences with squared error loss and linear experts, and an extension of the competitive on-line linear regression problem to the case where the clean signal is not available for learning. Naturally, we try to merge the methods of [8] and [43] in developing our universal FIR MMSE filter.

The rest of the chapter is organized as follows. The formulation of the problem and the main result are given in Section 3.2. We derive our universal filter in Section 3.3, and prove the main theorem of this chapter in Section 3.4. The stochastic setting result follows in Section 3.5, and several discussions are given in Section 3.6. Section 3.7 presents five different experiment sets that showcases the potential merits of our universal filter in practice. Proofs of lemmas are moved to Section 3.8, the Appendix, to allow for a smooth flow of the arguments.
CHAPTER 3. UNIVERSAL FIR MMSE FILTERING

3.2 Problem formulation, Filter description, and Main result

3.2.1 Problem formulation

Let \( \{x_t\}_{t \geq 1} \) denote the real-valued signal that we want to estimate, and assume that for all \( t \), \( x_t \) takes value in \( D = [-B_X, B_X] \subset \mathbb{R} \), for some \( B_X < \infty \). We denote the signal with lower case, since we do not make any probabilistic assumption on the generation of \( x_t \). Hence, \( \{x_t\}_{t \geq 1} \) can be any arbitrary bounded individual sequence, even chaotic and adversarial. Suppose this signal goes through an additive channel, where the noise \( \{N_t\}_{t \geq 1} \) is independent over \( t \), and \( E(N_t) = 0, E(N_t^2) = \sigma^2 \) for all \( t \). Thus, the noise at each time \( t \) is not necessarily identically distributed, but we require the variance to be equal for all time \(^1\). Additionally, we assume that the noise is bounded almost surely, i.e., there exists a \( B_N < \infty \), such that \( |N_t| \leq B_N \) for all \( t \), with probability one. The bounded noise assumption simplifies our analysis but is not essential. We denote \( \{Y_t\}_{t \geq 1} \) as the output of the additive noisy channel whose input is \( \{x_t\}_{t \geq 1} \), i.e.,

\[
Y_t = x_t + N_t, \quad t = 1, 2, \ldots
\]  

(3.1)

Note that the noisy signal was denoted by \( \{Y_t\}_{t \geq 1} \) in this chapter, instead of \( \{Z_t\}_{t \geq 1} \) as in Chapter 2, to comply with the conventional signal processing notations.

The boldface notations will denote the \( d \)-dimensional column vector of \( d \) recent symbols, i.e., \( x_t = [x_t, x_{t-1}, \ldots, x_{t-(d-1)}]^T \), \( N_t = [N_t, N_{t-1}, \ldots, N_{t-(d-1)}]^T \), and \( Y_t = [Y_t, Y_{t-1}, \ldots, Y_{t-(d-1)}]^T \), where \( (\cdot)^T \) is a transposition operator. For completeness, we assign zeros to the elements of vectors whose indices are less than or equal to zero. We denote \( x^n = (x_1, \ldots, x_n) \) and \( Y^n = (Y_1, \ldots, Y_n) \). Also, we denote \( c = [\sigma^2, 0, \ldots, 0]^T \in \mathbb{R}^d \). \( \| \cdot \| \) denotes the Euclidean norm if it is used for vectors, and

\(^1\)In fact, the equal variance assumption for the noise components is not crucial, but it was assumed for the simplicity of the argument. Our scheme and results would naturally generalize to the case of \( E(N_t^2) = \sigma_t^2 \), where \( \sigma_t^2 \) is bounded away from zero for all \( t \), provided that the variance sequence \( \{\sigma_t^2\}_{t \geq 1} \) is known to the filter.
operator norm (i.e., maximum singular value) if used for matrices. Also, for matrices, \( \| \cdot \|_1 \) denotes \( \ell_1 \)-norm, i.e., \( \| A \|_1 = \sum_{i,j} |a_{ij}| \).

In the real-valued setting, a filter \( \hat{X} \) is a sequence of mappings \( \{ \hat{X}_t(\cdot) \}_{t \geq 1} \), where \( \hat{X}_t(\cdot) : \mathbb{R}^t \to \mathbb{R} \) and \( \hat{X}_t(Y^t) \) is again the causal estimator of \( x_t \) based on the noisy observation \( Y^t = (Y_1, Y_2, \cdots, Y_t) \). The performance of a filter for \( x^n \) is measured by the normalized cumulative squared error or, equivalently, the mean-squared error (MSE):

\[
\frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t(Y^t))^2.
\]

Now, an FIR filter of order \( d \), the focus of this chapter, can be denoted as \( \hat{X}_{u,t}(Y^t) = u^T Y_t \), where \( u \in \mathbb{R}^d \) is a vector of filter coefficients. Then, for each individual sequence \( x^n \) and noisy sequence realization \( Y^n \), the best FIR filter coefficients \( u^* \) that achieves

\[
\min_{u \in \mathbb{R}^d} \frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2
\]

is given as

\[
u^* = \left( \frac{1}{n} \sum_{t=1}^{n} Y_t Y_t^T \right)^{-1} \left( \frac{1}{n} \sum_{t=1}^{n} x_t Y_t \right).\]

Therefore, for given clean and noisy signal realization of length \( n \), the best FIR filter of order \( d \) is obtained from the complete knowledge of \( (x^n, Y^n) \).

In this chapter, we devise a filter \( \hat{X}_* \triangleq \{ \hat{X}_t^*(Y^t) \}_{t \geq 1} \) that only depends on the noisy signal \( \{ Y_t \}_{t \geq 1} \) and the noise variance \( \sigma^2 \), whose MSE asymptotically achieves (3.3) for every underlying signal \( \{ x \}_{t \geq 1} \), as \( n \) becomes large. A more precise description of the performance guarantee will be presented in the main theorem of this chapter.

As mentioned in the Introduction, this universal FIR MMSE filtering problem is more challenging than the online linear least-squares regression problem [43][44], since

---

\( ^2 \)In conventional signal processing literature where the underlying signal is usually a stationary stochastic process, MSE means the expected squared-error at certain time \( t \), where the expectation is with respect to the signal and the noise stationary distributions. In our setting, however, since we do not make any assumption on the distribution of the underlying signal, we use the empirical average of squared-errors and refer MSE to that quantity. As shown in Section 3.5, this performance measure is more general than the conventional one, since our result implies a result for stochastic setting with conventional MSE.
the filter cannot observe the clean signal, but only observes its noisy observation. Therefore, the filter needs to combat not only the arbitrariness of the underlying signal, but also the randomness of the noise. A similar setting of the linear least-squares prediction with noisy observations has been considered in [45]. The difference between our filter and the noisy predictor in [45] is that, by definition, our filter utilizes the noisy observation $Y_t$ for estimating $x_t$, whereas the noisy predictor does not have the access to $Y_t$. This difference is a crucial one since $Y_t$ is the most important observation for estimating $x_t$, and it will result in a significant performance gap between the two schemes. Several experiments in Section 3.7 will stress this point. Furthermore, the result in [45] was obtained directly from the prediction result in [44] and a concentration of the sum of noise symbols, namely, the noisy predictor for $x_t$ simply tries to predict $Y_t$ based on $Y_t^{-1}$, whereas our result is attained by adopting a more involved prediction-filtering association developed in [8] and applying probabilistic arguments. Hence, similarly as [45] is an extension of [46] from finite-alphabet to the continuous-valued setting in the prediction context, our work can be considered an extension of [8] in the same direction for the filtering context.

### 3.2.2 Description of our filter

Here, we describe our filter. A detailed derivation of the filter will be given in Section 3.3. First, we define a positive definite matrix $A_t \triangleq \left( I + \sum_{i=1}^{t} Y_i Y_t^T \right) \in \mathbb{R}^{d \times d}$, and a preliminary filter coefficient vector

$$w_{t-1}^* \triangleq A_{t-1}^{-1} \left( \sum_{i=1}^{t-1} \{Y_i Y_i - c\} \right) \tag{3.4}$$

for each $t$. We also define a ball of filter coefficients

$$\mathcal{U} \triangleq \{ u \in \mathbb{R}^d : \| u \| \leq R \}, \tag{3.5}$$
where \( R = \max\{2^2\sqrt{dB_X(B_X + B_N)}, 4\} \), and a projection to the ball

\[
\Pi_U(u) = \begin{cases} 
    u, & \text{if } u \in U \\
    \frac{u}{\|u\|}, & \text{if } u \notin U
\end{cases}
\]

for any \( u \in \mathbb{R}^d \). The value of \( R \) will be justified later. Then, our filter at time \( t \) is given as

\[
\hat{X}_t^*(\hat{Y}^t) = \hat{w}_{t-1}^T Y_t,
\]

where \( \hat{w}_{t-1}^* = \Pi_U(w_{t-1}^*) \), a projection of \( w_{t-1}^* \) to \( U \). Note that this filter is not linear in the noise sequence \( \{Y_t\}_{t \geq 1} \), but, for given \( \hat{w}_{t-1}^* \), it linearly combines the noisy components \( Y_t \) to estimate \( x_t \). A discussion of an algorithmic aspect of our filter will be given in Section 3.6. The definition of our filter (3.7) also requires the knowledge of signal and noise bounds, \( B_X \) and \( B_N \), in addition to the noise variance \( \sigma^2 \). This is a requirement to bound our filter coefficient \( \hat{w}_{t-1}^* \) for all \( t \), and is needed for proving our high probability results below. However, in Section 3.6.2, we argue that this requirement is not necessary in any meaningful practical scenarios, and only the knowledge about \( \{Y_t\}_{t \geq 1} \) and \( \sigma^2 \) are enough in building our universal filter. Furthermore, one may be intrigued by the exclusion of \( Y_t \) in determining the filter coefficients at time \( t \) since \( Y_t \) is the most important observation in estimating \( x_t \). Although this may seem counterintuitive, it is a necessary requirement for our analysis that will become clear in Section 3.3. Nonetheless, a reader should not be confused with the fact that our scheme indeed is a filter since it does use \( Y_t \) by combining components of \( Y_t \) in estimating \( x_t \). It will also become clear that the exclusion of \( Y_t \) in determining the filter coefficients does not affect the filter performance much as we present our simulation results in Section 3.7.

Following subsection presents our main result of this chapter.
CHAPTER 3. UNIVERSAL FIR MMSE FILTERING

3.2.3 Main result

Theorem 2 Consider a filter $\hat{X}^* = \{\hat{X}_t^*(Y_t^t)\}_{t \geq 1}$ as defined in (3.7). Then, we have the following two theorems.

(a) For all $x^n \in D^n$ and all $n$,
\[
E\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t^*(Y_t))^2\right) - \min_{u \in \mathbb{R}^d} E\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2\right) \leq \Theta\left(\frac{\log n}{n}\right).
\]

(b) For all $x^n \in D^n$, all $\epsilon > 0$, and sufficiently large $n$,
\[
P\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t^*(Y_t))^2 - \min_{u \in \mathbb{R}^d} \frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2 > \epsilon + \Theta\left(\frac{\log n}{n}\right)\right) 
\leq \exp \left(-\Theta\left(n^{1/3}\right)\right).
\]

Remark: Note that we have suppressed all the constants in the bound with $\Theta(\cdot)$ notation. To state the dependencies on constants qualitatively, the bound in Part (a) depends polynomially on $B_X$, $B_N$, $\sigma^2$, and $d$, and the bound in Part (b) depends polynomially on $B_X$, $B_N$, $\sigma^2$, and $\frac{1}{\epsilon}$, and exponentially on $d$. However, we omit these dependencies on constants in stating the theorem to avoid unnecessarily complicated expression of the theorem and to highlight the dependence of the bound on the sequence length $n$. Instead, we examine the effect of constants on the convergence rate via various experimentations given in Section 3.7, which will show that the effects are not as severe as we see on the complicated upper bound expressions. Part (a) of the theorem asserts the logarithmic decay rate of the regret of the expected MSE of our filter, where the expectation is with respect to the noise distribution. Note that this logarithmic decay rate parallels that of the results in [43],[44]. Part (b) gives a much stronger result than Part (a), i.e., it shows that, as $n$ grows, not only the expected MSE of our filter gets close to the minimum expected MSE $\min_{u \in \mathbb{R}^d} E\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2\right)$, but also the actual MSE of our filter is guaranteed to be no larger than the minimum actual MSE $\min_{u \in \mathbb{R}^d} \frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2$, with high probability. It is worth noting that, while in most statistical signal processing contexts with a stochastic setting, it is usually satisfactory and informative enough to make statements regarding
the expected performance of a filter, this is not the case in the individual sequence
setting considered here. The whole point of the individual sequence setting is to have
a complete picture of what is really happening (actual rather than expected MSE)
for every possible sequence. This is why we obtain the high probability result, which
guarantees the actual performance of the filter, in addition to Part (a). Finally, note
that from Part (b), we easily obtain the almost sure convergence

\[
\limsup_{n \to \infty} \left[ \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t^*(Y^t))^2 - \min_{u \in \mathbb{R}^d} \frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2 \right] \leq 0 \quad \text{a.s.}
\]

by the fact that \(\exp(-\Theta(n^{1/3}))\) is summable and applying the Borel-Cantelli lemma.²

3.3 Derivation of the universal filter

In this section, we derive our universal filter based on a similar argument as in [8].
We first introduce the following definition to further simplify our notation.

Definition 4 For any \(u \in \mathbb{R}^d\), define

(a) \(\Lambda_t(u) \triangleq (x_t - u^TY_t)^2\)

(b) \(\ell_t(u) \triangleq (Y_t - u^TY_t)^2 + 2u^Tc\)

Remark: When we think of \(u\) as a filter coefficient of order \(d\), \(\Lambda_t(u)\) denotes the
squared loss incurred by a filter \(\hat{X}_t(Y^t) = u^TY_t\). Note that, although suppressed
in the notation, \(\Lambda_t(u)\) depends not only on \(Y_t\), but also on \(x_t\). In contrast, \(\ell_t(u)\)
denotes the estimated loss of \(\hat{X}_t(Y^t) = u^TY_t\) based on \(\{Y_t\}_{t \leq t}\), the meaning of which
will become clear in what follows. Unlike \(\Lambda_t(u)\), \(\ell_t(u)\) does not depend on \(x_t\) and
hence is observable.

Equipped with this notation, we have the following martingale lemma, which is
inspired by [8].

²A part of this result was presented in [47].
Lemma 10 Consider a sequence of random vectors \( \{w_{t-1}\}_{t \geq 1} \), where each \( w_{t-1} \in \mathbb{R}^d \). Suppose \( w_{t-1} \) is \( \sigma(Y^{t-1}) \)-measurable for all \( t \). Then, for all \( x^n \in D^n \),

\[
\left\{ \left[ \sum_{t=1}^{n} \Lambda_t(w_{t-1}) - \sum_{t=1}^{n} \{ \ell_t(w_{t-1}) - \sigma^2 \} \right] \right\}_{n \geq 1}
\]

is a \( \{Y_n\} \)-martingale

Proof: See Appendix 3.8.1. ■

Now, consider a class of filters of the form \( \hat{X}_t(Y^t) = w_{t-1}^T Y_t \), where \( w_{t-1} \in \sigma(Y^{t-1}) \). Then, since Lemma 10 also holds for any constant weight vector \( u \in \mathbb{R}^d \), we have

\[
E\left( \sum_{t=1}^{n} (x_t - \hat{X}_t(Y^t))^2 \right) - E\left( \sum_{t=1}^{n} (x_t - u^T Y_t)^2 \right) = E\left( \sum_{t=1}^{n} (\Lambda_t(w_{t-1}) - \Lambda_t(u)) \right) = E\left( \sum_{t=1}^{n} (\ell_t(w_{t-1}) - \ell_t(u)) \right)
\]

for all \( u \in \mathbb{R}^d \), where (3.8) is from the martingale result established in Lemma 10. Hence, the observable \( \sum_{t=1}^{n} \{ \ell_t(w_{t-1}) - \ell_t(u) \} \) is an unbiased estimate of \( \sum_{t=1}^{n} \{ \Lambda_t(w_{t-1}) - \Lambda_t(u) \} \). This is the reason why we referred to \( \ell_t(u) \) as an estimated loss in Definition 4. One important thing to note is that, from the relationship in (3.8), we can replace the sum \( \sum_{t=1}^{n} \{ \Lambda_t(w_{t-1}) - \Lambda_t(u) \} \), that depends both on \( x^n \) and \( Y^n \) with its unbiased estimate that only depends on \( Y^n \). We attempt to build our universal filter that, by definition, should only depend on the noisy observation causally, based on these unbiased estimates of the squared-error losses. This approach of working with an unbiased estimate to circumvent the difficulty of not observing the underlying clean signal has also been utilized in various previous research papers such as wavelet-based denoising [48], parameter estimation [49], discrete denoising [7][11][50], and universal filtering of finite-alphabet signals [42][8][32].

To derive our universal filter, we follow the perspective of prediction-filtering association developed in [8]. Namely, we can think of the filter coefficient \( w_{t-1} \), which
is based on \( Y^{t-1} \), as a prediction of a linear mapping for time \( t \) that maps a vector \( Y_t \) into \( \mathbb{R} \). Then, \( \ell_t(w_{t-1}) \) can be thought of as the corresponding loss incurred at time \( t \) by that prediction. Conversely, whenever we have a sequence of predictors \( \{w_{t-1}\}_{t \geq 1} \) in the above sense, we can associate an FIR filter by merely defining \( \hat{X}_t(Y^t) = w_{t-1}^T Y_t \). As in [8], we continue to adhere to the prediction viewpoint in further development of our filter. Note the difference that we are trying to predict a linear mapping to apply at time \( t \), unlike the scheme in [45] which tries to predict \( Y_t \). The sum \( \sum_{t=1}^n \{\ell_t(w_{t-1}) - \ell_t(u)\} \) can then be interpreted as a difference between the cumulative loss incurred by the sequence of predictors \( \{w_{t-1}\}_{t \geq 1} \) and that of a constant predictor \( u \). Our approach is to come up with a sequence of predictors \( \{w_{t-1}\}_{t \geq 1} \) that makes the cumulative loss of the predictors close to that of the best constant predictor, and then show that the associated filter indeed is defined as (3.7) and has the properties presented in Theorem 2.

In solving the above prediction problem, by recognizing \( \ell_t(u) \) as a convex function in \( u \), one may be tempted to use algorithms that are developed in the context of online convex optimization [51] in the learning theory community. That is, to obtain the logarithmic decay rate of Part (a), we can proceed as in [8] by treating \( \{Y_t\}_{t \geq 1} \) as an individual sequence and simply apply the algorithms in [51] to the prediction problem inside the expectation in (3.8), and get the logarithmic regret even before taking the expectation. A slower rate than the logarithmic rate, e.g., \( O(\frac{1}{\sqrt{n}}) \), can indeed be attained this way by applying general online gradient descent algorithms as in [52]. However, for the logarithmic rate, the subtle point is that, due to \( Y_t \) being random, the induced loss function \( \ell_t(u) \) does not satisfy the conditions required by the algorithms in [51]: being exp-concave\(^4\) with some constant \( \alpha > 0 \) for all \( t \). Therefore, we cannot directly apply the algorithms developed in [51]. Instead, we derive our predictor in a rather intuitive way, and carefully analyze the behavior of our associated filter's performance by taking into account the randomness of \( \{Y_t\}_{t \geq 1} \). A detailed analysis will follow in the next section.

\(^4\)A convex function \( f(x) \) is an exp-concave function with parameter \( \alpha > 0 \), if \( \exp(-\alpha f(x)) \) is a concave function in \( x \).
Before obtaining our filter, we consider our estimator for the (regularized) cumulative loss up to time $t$, which we define to be

$$L_t(u) \triangleq \|u\|^2 + \sum_{i=1}^{t} \ell_i(u)$$

$$= u^T \left( I + \sum_{i=1}^{t} Y_i Y_i^T \right) u - 2u^T \left( \sum_{i=1}^{t} \{Y_i Y_i - c\} \right) + \sum_{i=1}^{t} Y_i^2,$$

where $I$ is the $d$-by-$d$ identity matrix. Note that $A_t \triangleq \left( I + \sum_{i=1}^{t} Y_i Y_i^T \right) \in \mathbb{R}^{d \times d}$, defined in Section 3.2.2, is the Hessian of $L_t(u)$ and is positive definite for all $t$. Then, it is clear to realize that $w_{t-1}^*$ defined in (3.4) is a unique minimizer of $L_{t-1}(u)$, the cumulative estimated losses up to time $t-1$. Note that depending on $Y_{t-1}$, $\|w_{t-1}^*\|$ can grow without bound as $t$ becomes large. However, as shown in the next section, the best FIR filter coefficient $u^*$ that achieves (3.3) is bounded with high probability, and we would only need to consider the filter coefficients that are bounded, i.e., coefficients in $U$. Therefore, by projecting $w_{t-1}^*$ onto $U$, we obtain our prediction $\hat{w}_{t-1}^*$ for time $t$ which is always in $U$ and $\sigma(Y_{t-1})$-measurable. This predictor can be thought of as a follow-the-leader type predictor in [41][42] except for the ridge term $I$ in $A_t$ that prevents $A_t^{-1}$ from diverging. Finally, following the prediction-filtering association mentioned above, we define our filter at time $t$ as

$$\hat{X}_t^*(Y^t) = \hat{w}_{t-1}^T Y_t,$$

which is also given in (3.7). Since $\hat{w}_{t-1}^*$ is $\sigma(Y_{t-1})$-measurable, (3.8) remains valid with $\hat{w}_{t-1}^*$ replacing $w_{t-1}$. The form of our filter resembles that of the Recursive Least Square (RLS) adaptive filter [38, Chapter 9] or the on-line ridge regressor [43]. The difference is that (3.7) is solely expressed with the noisy signals and the noise variance, whereas the other two need to know a desired response or the clean past signal components. We now move on to prove that our filter (3.7) satisfies the properties stated in Theorem 2.
3.4 Analysis

We first present two lemmas needed for the proof of Part (a) of our theorem. Lemma 11, which resembles the steps in [43] and [6, Chapter 11.7], collects properties of $w^*_t$ and $\hat{w}^*_{t-1}$. Lemma 12 asserts a key concentration result and borrows a law of large numbers argument from [42].

Lemma 11 Consider $w^*_t$ and $\hat{w}^*_{t}$ defined in Section 3.2.2.

(a) $w^*_t$ satisfies

$$w^*_t = w^*_{t-1} - A_t^{-1}\{(w^*_{t-1}Y_t - Y_t)Y_t + c\},$$

and

$$\|w^*_t\| \leq 1 + (1 + t\sigma^2)\lambda_{\text{max}}(A_t^{-1}),$$

(3.10)

where $\lambda_{\text{max}}(A_t^{-1})$ is the maximum eigenvalue of $A_t^{-1}$.

(b) Let $R_t = w^*_{t-1}Y_t - Y_t$. Then,

$$|R_t| \leq (1 + (t - 1)\sigma^2)\lambda_{\text{max}}(A_t^{-1})\|Y_t\|.$$  

(c) For all $u \in \mathbb{R}^d$,

$$\sum_{t=1}^{\infty} \{\ell_t(\hat{w}^*_{t-1}) - \ell_t(u)\} \leq \|u\|^2 + \sum_{t=1}^{\infty} (\hat{w}^*_{t-1} - w^*_t)^T A_t(\hat{w}^*_{t-1} - w^*_t).$$

Proof: Part (a) and (b) follow from manipulations of the definition of $w^*_t$. Part (c) builds a telescoping sum and uses the convexity of $L_t(u)$. See Appendix 3.8.2 for a detailed proof. ■

Lemma 12 Denote $K_t = \sum_{i=1}^{t} Y_iY_i^T$. Then,

(a) For any $\epsilon > 0$,

$$P\left(\frac{1}{t}K_t - \left(\sigma^2 I + \frac{1}{t} \sum_{i=1}^{t} x_ix_i^T\right)\|_1 > \epsilon\right) \leq 2d^2 \exp\left(-\frac{2\epsilon^2}{Ct}\right),$$

where $C = B_N^2(B_N + 2B_X)^2d^4$.

\(^5\)Here and throughout, equalities and inequalities between random variables, when not explicitly mentioned, are to be understood in the almost sure sense.
(b) Let \( \lambda_{\min}(K_t) \) be the minimum eigenvalue of the random matrix \( K_t \). Then,

\[
P(\lambda_{\min}(K_t) \geq \frac{\sigma^2 t}{2}) \geq 1 - 2d^2 \exp\left(-\frac{\sigma^4}{2CF^2 t}\right),
\]

where \( F = \frac{d+2}{d}(B_X + B_N)^2(1-1/d) \).

**Proof:** Part (a) is based on the concentration of the sum of bounded martingale differences. Part (b) uses the fact that the minimum eigenvalue of a matrix is a continuous function of the elements of the matrix. See Appendix 3.8.3 for a detailed proof.

**Remark:** Part (b) of the lemma shows that, as \( t \) grows, the minimum eigenvalue of \( K_t \) will grow linearly in \( t \) with high probability. This property plays a central role in the proof of our theorem.

Equipped with the above two lemmas, we now prove Part (a) of our theorem.

**Proof of Theorem 2(a):** First, note that

\[
\hat{u} = \left(\sigma^2 + \frac{1}{n} \sum_{t=1}^{n} x_t x_t^T\right)^{-1} \left(\frac{1}{n} \sum_{t=1}^{n} x_t\right)
\]

achieves \( \min_{u \in \mathbb{R}^d} E\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2\right) \) and \( \|\hat{u}\| \leq \frac{\sqrt{dR^2}}{\sigma} \leq R \). Hence, it is enough to only consider the filter coefficients in \( \mathcal{U} \) and show

\[
E\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t)^2\right) - \min_{u \in \mathcal{U}} E\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2\right) \leq \Theta\left(\frac{\log n}{n}\right) \tag{3.11}
\]
to prove Part (a) of our theorem. To show this, for our filter \( \hat{X}_t^* (Y_t) \) defined in (3.7) and for all \( u \in U \), we begin with the following inequality:

\[
E \left( \sum_{t=1}^{n} (x_t - \hat{X}_t^* (Y_t))^2 \right) - E \left( \sum_{t=1}^{n} (x_t - u^T Y_t)^2 \right) - ||u||^2
\]

\[
= E \left( \sum_{t=1}^{n} \Lambda_t(\hat{w}_{t-1}^*) - \sum_{t=1}^{n} \Lambda_t(u) \right) - ||u||^2
\]

\[
= E \left( \sum_{t=1}^{n} (\ell_t(\hat{w}_{t-1}^*) - \ell_t(u)) \right) - ||u||^2
\]

where (3.12) follows from (3.8) and definition of \( \hat{w}_{t-1}^* \), and (3.13) follows from Lemma 11(b). To proceed, consider

\[
(\hat{w}_{t-1}^* - w_t^*)^T A_t (\hat{w}_{t-1}^* - w_t^*)
\]

\[
= (w_{t-1}^* - w_t^*)^T A_t (w_{t-1}^* - w_t^*)
\]

\[
+ (\hat{w}_{t-1}^* - w_{t-1}^*)^T A_t (\hat{w}_{t-1}^* - w_{t-1}^*) + 2(\hat{w}_{t-1}^* - w_{t-1}^*)^T A_t (w_{t-1}^* - w_t^*)
\]

\[
\leq (R_t Y_t + c)^T A_t^{-1} (R_t Y_t + c) + \| A_t \| \| \hat{w}_{t-1}^* - w_{t-1}^* \|^2 + 2\| R_t Y_t + c \| \| \hat{w}_{t-1}^* - w_{t-1}^* \|.
\]

where (3.14) follows from applying \( w_{t-1}^* - w_t^* = A_t^{-1} (R_t Y_t + c) \) obtained in Lemma 11(a) and the Cauchy-Schwartz inequality. We now continue (3.13) separately on
each term of (3.14). The expected sum of the first term in (3.14) becomes

\[
E\left(\sum_{t=1}^{n} (R_t Y_t + c)^T A_t^{-1}(R_t Y_t + c)\right)
\]

\[
\leq E\left(\sum_{t=1}^{n} (\|R_t\|\|Y_t\| + \sigma^2)^2 \lambda_{\max}(A_t^{-1})\right)
\]

\[
\leq E\left(\sum_{t=0}^{n} \left\{b_1(1 + (t - 1)\sigma^2)\lambda_{\max}(A_{t-1}^{-1}) + \sigma^2\right\}^2 \lambda_{\max}(A_{t-1}^{-1})\right)
\]

\[
\leq E\left(\sum_{t=0}^{n} \left\{b_1(1 + t\sigma^2)\lambda_{\max}(A_{t-1}^{-1}) + \sigma^2\right\}^2 \lambda_{\max}(A_{t-1}^{-1})\right)
\]

\[
= E\left(\sum_{t=0}^{n} \left(\sigma^2 + b_1 \frac{1 + t\sigma^2}{1 + \lambda_{\min}(K_t)}\right)^2 \cdot \frac{1}{1 + \lambda_{\min}(K_t)}\right)
\]

(3.15)

(3.16)

(3.17)

(3.18)

where (3.15) follows from the fact that \(A_t^{-1}\) is symmetric and \(\|A_t^{-1}\| = \lambda_{\max}(A_t^{-1})\); (3.16) follows from Lemma 11(b) and setting \(b_1 = \max \|Y_t\|^2 = d(B_X + B_N)^2\); (3.17) follows from \(\lambda_{\max}(A_{t-1}^{-1}) \geq \lambda_{\max}(A_{t-1}^{-1})\) by interlacing inequality [53, Theorem 4.3.1] and adding the \((n + 1)\)-th term in the end, and (3.18) follows from the fact \(A_t = I + K_t\).

Now, by applying Lemma 12(b) again, we know that with probability of at least \(1 - 2d^2 \exp(-\frac{\sigma^4}{2CF^2} t)\), the event

\[
\left\{\sigma^2 + b_1 \frac{1 + \sigma^2 t}{1 + \lambda_{\min}(K_t)}\right\}^2 \cdot \frac{1}{1 + \lambda_{\min}(K_t)} \leq \left\{\sigma^2 + b_1 \frac{1 + \sigma^2 t}{1 + (\sigma^2/2)t}\right\}^2 \cdot \frac{1}{1 + (\sigma^2/2)t}
\]

(3.19)

hold. Therefore, by conditioning on this event and its complement, we can continue to upper bound (3.18) as

\[
(3.18) \leq \sum_{t=0}^{n} (\sigma^2 + b_1(1 + \sigma^2 t))^2 \cdot 2d^2 \exp\left(-\frac{\sigma^4}{2CF^2} t\right)
\]

\[
+ \sum_{t=0}^{n} \left\{\sigma^2 + b_1 \frac{1 + \sigma^2 t}{1 + (\sigma^2/2)t}\right\}^2 \cdot \frac{1}{1 + (\sigma^2/2)t}.
\]

(3.20)

(3.21)

Since \(\sum_{t=0}^{\infty} \alpha_1 t^2 e^{-\alpha_2 t} < \infty\) for any \(\alpha_1, \alpha_2 > 0\), we know that (3.20) is upper bounded by a constant. Furthermore, since the bound \(\sum_{t=0}^{n} \frac{(b+c)^2}{(1+a)^3} \leq b^2 + \int_{1}^{n} \frac{(b+c)^2}{(ax)^3} dx \leq b^2 + \frac{(b+c)^2}{a^3} \cdot \log n\) holds for any \(a, b > 0\), we conclude that (3.21) is \(O(\log n)\).
We can apply a similar technique to bound the expected sum of the second and the third term in (3.14). From Lemma 11(a) and Lemma 12(b), we can see that for \( t > 1 + \frac{2}{\sigma^2} \), with probability of at least \( 1 - 2d^2\exp\left(-\frac{\sigma^4}{2CF^2}(t-1)\right) \), we have \( \|w_{t-1}^*\| \leq 4 \leq R \) and thus, \( \|\hat{w}_{t-1} - w_{t-1}^*\| = 0 \). Therefore, by conditioning on this event and its complement, we have

\[
E\left( \sum_{t=1}^{n} \|A_t\| \|\hat{w}_{t-1} - w_{t-1}^*\|^2 \right)
\]
\[
\leq \sum_{t=1}^{n} (1 + t(B_X + B_N)^2)(2 + \sigma^2)^2 \exp\left(-\frac{\sigma^4}{2CF^2}(t-1)\right),
\]

(3.22)

where (3.22) follows from \( \|A_t\| \leq t(B_X + B_N)^2 \) and \( \|\hat{w}_{t-1} - w_{t-1}^*\| \leq \|w_{t-1}^*\| \leq 2 + t\sigma^2 \).

Thus, we conclude that (3.22) is upper bounded by a constant. Similarly, we have

\[
E\left( \sum_{t=1}^{n} 2\|R_t Y_t + c\| \|\hat{w}_{t-1} - w_{t-1}^*\| \right)
\]
\[
\leq E\left( \sum_{t=1}^{n} 2 \left( b_1(1 + (t-1)\sigma^2)\lambda_{\text{max}}(A_{t-1}^{-1}) + \sigma^2 \right) \|\hat{w}_{t-1} - w_{t-1}^*\| \right)
\]
\[
\leq \sum_{t=1}^{n} 2 \left( b_1(1 + (t-1)\sigma^2) + \sigma^2 \right) (2 + t\sigma^2) \exp\left(-\frac{\sigma^4}{2CF^2}(t-1)\right),
\]

(3.23)

and see that (3.23) is again upper bounded by a constant. Therefore, by combining the bounds on (3.20),(3.21),(3.22), and (3.23), we continue from (3.13) and obtain

\[
E\left( \sum_{t=1}^{n} (x_t - \hat{X}_t^*(Y^t))^2 - \sum_{t=1}^{n} (x_t - u^T Y_t)^2 \right) \leq \|u\|^2 + \Theta(\log n).
\]

(3.24)

for all \( u \in \mathcal{U} \). Since \( \mathcal{U} \) is a bounded set, we have proved (3.11) and Part (a) of the theorem.

To prove Part(b) of Theorem 2, we need two additional lemmas. Lemma 13 below shows that when the probability of each random variable indexed by \( t \) being positive has an upper bound that exponentially decreases in \( t \), the probability of the average being positive also has a bound that is summable in \( n \). Lemma 14 gives a result
parallelizing (3.24) for the high probability setting.

**Lemma 13** Let \( \{Z_t\}_{t \geq 1} \) be a sequence of random variables satisfying \( 0 \leq Z_t \leq c_1(c_2 + t)^2 \) a.s. for some positive constants \( c_1 \) and \( c_2 \) and, for each \( t \), \( P(Z_t > 0) \leq 2d^2 \exp(-tC) \) for some positive constant \( C \). Then,

\[
P\left( \frac{1}{n} \sum_{t=1}^{n} Z_t > \epsilon \right) \leq \frac{2d^2}{1-\exp(-C)} \exp \left( - \left( \frac{3nC}{c_1} \right)^{1/3} - c_2 \right) C \tag{3.25}
\]

**Proof:** The lemma follows from successive applications of the union bound. See Appendix 3.8.4 for a detailed proof.

**Remark:** As mentioned above, the key point of this lemma is that the right hand side of (3.25) decays fast enough with \( n \) so that it ensures \( \sum_{n=1}^{\infty} P\left( \frac{1}{n} \sum_{t=1}^{n} Z_t > \epsilon \right) < \infty \).

**Lemma 14** Fix \( \epsilon > 0 \). Then, for all \( n \), all \( x^n \in \mathcal{D}^n \), and for any fixed \( u \in \mathcal{U} \), our filter \( \hat{X}_t(Y_t) \) defined in (3.7) satisfies

\[
P\left( \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t(Y_t))^2 - \frac{1}{n} \left\{ \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right\} > \epsilon + \Theta\left( \frac{\log n}{n} \right) \right)
\leq \exp \left( - \Theta\left( n^{1/3} \right) \right).
\]

**Proof:** The proof follows from the martingale result in Lemma 10 and the result of Lemma 13. See Appendix 3.8.5 for a detailed proof.

Now, we can prove the second part of our theorem.

**Proof of Theorem 2(b):** Recall from Section 3.2.1 that the best FIR filter coefficients that achieves (3.3) is given as

\[
u^* = \arg \min_{u \in \mathbb{R}^d} \sum_{t=1}^{n} \Lambda_t(u) = \left( \frac{1}{n} \sum_{t=1}^{n} Y_t Y_t^T \right)^{-1} \left( \frac{1}{n} \sum_{t=1}^{n} x_t Y_t \right).
\]

Lemma 12(b) shows that with probability of at least \( 1 - 2d^2 \exp\left(-\frac{\sigma_x^2}{2C_B n}\right) \), the maximum eigenvalue of \( \left( \frac{1}{n} \sum_{t=1}^{n} Y_t Y_t^T \right)^{-1} \) is less than or equal to \( \frac{3}{\sigma_x^2} \), hence, \( \|u^*\| \leq \frac{3}{\sigma_x^2} \sqrt{d} B_X (B_X + B_N) \) and \( u^* \in \mathcal{U} \). This shows the reason why we set the value of
From this observation, we know that
\[
P\left( \min_{u \in U} \left\{ \frac{1}{n} \sum_{t=1}^{n} \Lambda_t(u) \right\} - \min_{u \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{t=1}^{n} \Lambda_t(u) \right\} > 0 \right) \leq 2d^2 \exp\left( - \frac{\sigma^4}{2CF^2 n} \right) = \exp\left( - \Theta(n) \right),
\] (3.26)
and similarly as in Part (a), it suffices to only consider the filter coefficients in $U$ and prove
\[
P\left( \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t(Y^t))^2 - \min_{u \in U} \left\{ \frac{1}{n} \sum_{t=1}^{n} \Lambda_t(u) \right\} > \epsilon + \Theta\left( \frac{\log n}{n} \right) \right) \leq \exp\left( - \Theta\left( n^{1/3} \right) \right).
\] (3.27)

Since $U$ is compact and $\{\Lambda_t(u)\}_{t \geq 1}$ are bounded for all $u \in U$, we can easily verify that $\frac{1}{n} \left\{ \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right\}$ is a Lipschitz continuous function on $U$. Now, let $U_\delta$ be a finite set that is obtained by uniformly quantizing $U$ with resolution $\delta$. Then, from Lipschitz continuity, we can find a constant $G$ such that
\[
\min_{u \in U_\delta} \left\{ \frac{1}{n} \left( \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right) \right\} \leq \min_{u \in U} \left\{ \frac{1}{n} \left( \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right) \right\} + G\delta,
\] (3.28)
where $G$ is a constant independent of $\delta$. Note that $|U_\delta| \leq (\frac{R}{\delta})^d$. Furthermore, for given $\epsilon > 0$, there exists some sufficiently large $N_\epsilon$ such that for all $n \geq N_\epsilon$,
\[
\min_{u \in U_\delta} \left\{ \frac{1}{n} \left( \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right) \right\} \leq \min_{u \in U} \left\{ \frac{1}{n} \sum_{t=1}^{n} \Lambda_t(u) \right\} + \frac{\epsilon}{4},
\] (3.29)
CHAPTER 3. UNIVERSAL FIR MMSE FILTERING

since \(\|u\| \leq R\) for all \(u \in U\). Now, fix \(\epsilon > 0\), and let \(\delta = \frac{\epsilon}{4G}\). Then, we have

\[
P\left(\frac{1}{n} \sum_{i=1}^{n} \Lambda_t(\hat{w}_{t-1}^*) - \min_{u \in U} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Lambda_t(u) \right\} > \epsilon + \Theta\left(\frac{\log n}{n}\right)\right)
\]

\[
\leq P\left(\frac{1}{n} \sum_{t=1}^{n} \Lambda_t(\hat{w}_{t-1}) - \min_{u \in U} \left\{ \frac{1}{n} \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right\} > \frac{\epsilon}{2} + \Theta\left(\frac{\log n}{n}\right)\right) + G\delta + \frac{\epsilon}{4} > \epsilon + \Theta\left(\frac{\log n}{n}\right)
\]

(3.30)

\[
\leq P\left(\frac{1}{n} \sum_{t=1}^{n} \Lambda_t(\hat{w}_{t-1}) - \min_{u \in U} \left\{ \frac{1}{n} \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right\} > \frac{\epsilon}{2} + \Theta\left(\frac{\log n}{n}\right)\right)
\]

(3.31)

\[
\leq \sum_{u \in U} P\left(\frac{1}{n} \sum_{t=1}^{n} \Lambda_t(\hat{w}_{t-1}) - \min_{u \in U} \left\{ \frac{1}{n} \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right\} > \frac{\epsilon}{2} + \Theta\left(\frac{\log n}{n}\right)\right)\]

(3.32)

\[
\leq \left(\frac{4RG}{\epsilon}\right)^d \cdot P\left(\frac{1}{n} \sum_{t=1}^{n} \Lambda_t(\hat{w}_{t-1}) - \min_{u \in U} \left\{ \frac{1}{n} \sum_{t=1}^{n} \Lambda_t(u) + \|u\|^2 \right\} > \frac{\epsilon}{2} + \Theta\left(\frac{\log n}{n}\right)\right),
\]

(3.33)

where (3.30) follows from (3.29); (3.31) follows from \(G\delta < \frac{\epsilon}{4}\); (3.32) follows from the union bound, and (3.33) follows from \(|U| \leq \left(\frac{R}{\delta}\right)^d\). Now, applying Lemma 14 asserts that (3.33) is upper bounded by \(\exp(-\Theta(n^{1/3}))\). Therefore, (3.27) is proved and Part (b) of the theorem follows.

3.5 Stochastic setting

The individual sequence setting result of Theorem 2 ensures a conventional stochastic setting result as well. Namely, when the underlying signal is a bounded, real-valued, stationary stochastic process, our universal filter achieves the performance of the optimal FIR filter, or the Wiener FIR filter. This result is analogous to the stochastic setting result for the finite-alphabet underlying signals in Chapter 2.

Suppose the underlying signal \({X_t}_{t \geq 1}\) is now a stationary stochastic process, independent of the noise process, and denote \(P_X\) as its probability distribution. Without loss of generality, we assume \(E(X_t) = 0\) for all \(t\). Then, we can denote the minimum MSE (MMSE) attained by the Wiener FIR filter as \(D_d(P_X, \sigma^2) = \min_{u \in \mathbb{R}^d} E(X_t - u^TY_t)^2 = \sigma_X^2 - \Sigma_{X Y} \Sigma_{Y}^{-1} \Sigma_{X Y}\), where \(\sigma_X^2 = \text{Var}(X_t), \Sigma_{XY} = E[X_t Y_t], \)
and $\Sigma_Y = E[Y_t Y_t^T]$. The following corollary asserts our stochastic setting result.

**Corollary 2** Suppose the underlying signal $\{X_t\}_{t \geq 1}$ is a stationary stochastic process. Then the filter $\hat{X}^* = \{\hat{X}_t^*(Y^t)\}$ defined in (3.7) satisfies

$$E\left(\frac{1}{n} \sum_{t=1}^{n} (X_t - \hat{X}_t^*(Y^t))^2\right) - d_d(P_X, \sigma^2) \leq \Theta\left(\frac{\log n}{n}\right).$$

**Proof:** The proof follows from applying Part (a) of Theorem 2. Note that from the stationarity,

$$d_d(P_X, \sigma^2) = \min_{u \in \mathbb{R}^d} E(X_t - u^T Y_t)^2 = \min_{u \in \mathbb{R}^d} E\left(\frac{1}{n} \sum_{t=1}^{n} (X_t - u^T Y_t)^2\right). \quad (3.34)$$

Therefore, we have

$$E\left(\frac{1}{n} \sum_{t=1}^{n} (X_t - \hat{X}_t^*(Y^t))^2\right) - d_d(P_X, \sigma^2)$$

$$= E\left(\frac{1}{n} \sum_{t=1}^{n} (X_t - \hat{X}_t^*(Y^t))^2\right) - \min_{u \in \mathbb{R}^d} E\left(\frac{1}{n} \sum_{t=1}^{n} (X_t - u^T Y_t)^2\right) \quad (3.35)$$

$$\leq E\left(E\left(\frac{1}{n} \sum_{t=1}^{n} (X_t - \hat{X}_t^*(Y^t))^2 \mid X^n\right) - \min_{u \in \mathcal{U}} E\left(\frac{1}{n} \sum_{t=1}^{n} (X_t - u^T Y_t)^2 \mid X^n\right)\right) \quad (3.36)$$

$$\leq \Theta\left(\frac{\log n}{n}\right), \quad (3.37)$$

where (3.35) follows from (3.34), (3.36) follows from exchanging expectation with minimum, and (3.37) follows from applying Part (a) of Theorem 2 for each conditioned sequence $X^n = x^n$. ■

### 3.6 Discussion

#### 3.6.1 Algorithmic description

As shown in the definition of our filter, the main requirement in implementing our filter is to calculate the preliminary filter coefficient $w_{t-1}$ for each time $t$. Lemma 11(a) and the matrix inversion lemma $A_t^{-1} = A_{t-1}^{-1} - \frac{(A_{t-1}^{-1} Y_t)(A_{t-1}^{-1} Y_t)^T}{1 + Y_t^T A_{t-1}^{-1} Y_t}$ shows that $w_{t-1}$
CHAPTER 3. UNIVERSAL FIR MMSE FILTERING

78

can be recursively updated with complexity of $O(d^2)$, instead of $O(d^3)$, which a naive
inversion of $A_t$ will require. Therefore, the total complexity of our filter for given
sequence length $n$ and filter order $d$ is $O(nd^2)$.

3.6.2 Requirement of the knowledge on bounds of signal and
noise

As we mentioned in Section 3.2.2, implementation of our filter coefficient $\hat{w}_{t-1}^*$ requires
the knowledge of the signal and noise bounds, $B_X$ and $B_N$. This was necessary
in proving Lemma 14 where we needed to make sure that the martingale differences
$\{ \Lambda_t(\hat{w}_{t-1}^*) - \ell_t(\hat{w}_{t-1}^*) + \sigma^2 \}_{t \geq 1}$ are bounded for all $t$. However, in any practical scenar­
ios, we claim that this requirement is not necessary since all possible implementable
filter coefficients that we are competing with, including the best implementable FIR
filter coefficient, should be bounded anyway. More specifically, when we build the
FIR filters with Digital Signal Processor (DSP) chips, any possible filter coefficients
should have bounded norms due to the memory limits of the processors. Suppose
$B_{DSP}$, which is independent of $B_X$ and $B_N$, is the maximum bound on the coefficients
that a DSP chip can support. Then, it is clear that the norm of the best FIR
filter coefficient that is implementable with the DSP is less than or equal to $B_{DSP}$.
Therefore, when we set the bound of $U$ in (3.5) as $R = \max\{B_{DSP}, 4\}$, all the analysis
that we gave will still hold. Hence, in most practical scenarios, we would not need to
know the bounds $B_X$ and $B_N$ explicitly. Instead, the knowledge of the predetermined
parameter of a DSP chip $B_{DSP}$, which we know from the specification of the DSP
chips, the noise variance $\sigma^2$, and the noisy signal $\{Y_t\}_{t \geq 1}$ would suffice to implement
our universal filter $\{\hat{X}_t(Y^t)\}_{t \geq 1}$.

3.6.3 Comments on the expectation result

In Theorem 2(a), we focused on the regret of the expected MSEs

$$E\left( \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t^* (Y^t))^2 \right) - \min_{u \in \mathcal{U}} E\left( \frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2 \right)$$

(3.38)
and showed that this regret goes to zero at rate $\Theta(\frac{\log n}{n})$. In fact, we can consider an even stronger notion, the expectation of the actual regret,

$$E\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t^*(Y'_t))^2 - \min_{u \in U} \frac{1}{n} \sum_{t=1}^{n} (x_t - u^TY_t)^2\right).$$  \hspace{1cm} (3.39)

Clearly, (3.39) is an upper bound on (3.38), and we do not know how to attain the logarithmic decay rate for (3.39). However, with additional complexity of the filtering scheme, we can upper bound (3.39) by $\Theta(\frac{(\log n)^2}{n})$. The trick would be to consider the noisy signal components with blocks (of length $k$) so that concentration of the block-sum of the estimated losses can happen to ensure the exp-concavity of the loss functions with sufficiently high probability, and use the result of [51]. This trick would lose additional $\log n$ factor due to treating estimated losses with blocks. Although this gives a meaningful bound for the stronger measure (3.39), we omit a detailed analysis.

3.6.4 Competing with any finite order filter

Although the dependency of the bounds on $d$ was suppressed in our result, we can increase $d$ with $n$ with sufficiently slow speed and still guarantee the asymptotically optimal performance. We omit a detailed mathematical argument here, but, for example, for each sequence length $n$, if we set the order of our filter as $d_n = O(\log n)$, the regret of our filter to any FIR filter would still go to zero as $n$ goes to infinity. This scheme resembles the schemes devised for the universal compression [3], prediction [13], and filtering [8] problems for finite-alphabet signals, that successfully compete with any order of Markov schemes. Our above scheme assumes the knowledge of $n$ at the beginning of the filtering process to determine $d_n$, which implies that the scheme is not strongly sequential and depends on the horizon. However, it is straightforward to construct a strongly sequential scheme from above scheme by using techniques that are by now standard, e.g., doubling tricks in [6, Chapter 2.3]. That is, we can divide the sequence into blocks with exponentially growing length, and apply above scheme separately by blocks to make the regret to any FIR filter vanish as the sequence length.
3.7 Simulation results

In this section, we demonstrate the performance of our universal filter with several experiments. The figures are moved to the end of this chapter.

3.7.1 Linear, stochastic signal

Our first example considers the case where the underlying signal is a stationary, first order autoregressive signal. More specifically, the clean signal \( \{X_t\}_{t \geq 1} \) evolves as

\[
X_t = \alpha X_{t-1} + Z_t, \quad \alpha = 0.9, \quad t = 1, 2, \ldots ,
\]

where \( \{Z_t\}_{t \geq 1} \) is iid ~ \( \mathcal{N}(0,1) \), and \( X_0 \sim \mathcal{N}(0, \frac{1}{1-\alpha^2}) \) to assure the stationarity of \( \{X_t\}_{t \geq 1} \). The noisy signal \( \{Y_t\}_{t \geq 1} \) is obtained from passing the clean signal through the additive channel (3.1), where \( \{N_t\}_{t \geq 1} \) is iid ~ \( \mathcal{N}(0,1) \), independent of \( \{X_t\}_{t \geq 1} \).

Note that we assumed the signal and the noise are Gaussian processes, although we required them to be bounded in the analysis of the theorem. However, for any finite \( n \), the signal and the noise are bounded by \( B_X = \max_{1 \leq t \leq n} |X_t| \) and \( B_N = \max_{1 \leq t \leq n} |N_t| \), which are both finite. Therefore, the analysis of our theorem still holds. Moreover, in the practical scenario as discussed in Section 3.6.2, our universal filter \( \{\hat{X}_t^{(Y)}(Y^t)\}_{t \geq 1} \) in (3.7) can still be implemented without any knowledge of \( B_X \) or \( B_N \). That is, we assumed that the limit of a DSP chip \( D_{DSP} \) is sufficiently large, and we used the raw \( \mathbf{w}_{t-1}^* \) for our filter coefficients.

We implemented our universal filter of order \( d = 5 \), and experimented with the sequence length \( n = 10^4 \). For comparison purpose, we implemented the noisy predictor in [45] and a filter that can be induced by applying the online gradient descent algorithm in [52], both with the same order. The noisy predictor in [45] is given as
\[ \hat{X}_{t,ZS}(Y^{t-1}) = w_{t-1,ZS}^T Y_{t-1}, \] where

\[ w_{t-1,ZS} \triangleq A_{t-1}^{-1} \left( \sum_{i=1}^{t-1} Y_i Y_{t-1} \right), \quad (3.41) \]

and \( A_t \), \( Y_t \) are defined in the same way as our filter. (3.41) looks very similar to (3.4), but \( \hat{X}_{t,ZS}(Y^{t-1}) \) is clearly a predictor and not a filter, since it does not utilize \( Y_t \) in estimating \( X_t \). The gradient-descent filter obtained by applying the online gradient descent algorithm in [52], as described in Section 3.3, is given as 
\[ X_t^{GD}(Y^t) = w_{t-1,GD}^T Y_t, \]
where

\[ \text{Proj}(\cdot) \text{ is again the projection function in (3.6), and } \eta_t \text{ is the learning rate. Since the estimated loss function } \ell_t(u) \text{ is convex for all } t \text{ and } u, \text{ when } \eta_t = \frac{1}{\sqrt{t}}, [52] \text{ assures the same asymptotical optimality of } \{ \hat{X}_{t,GD}(Y^t) \}_{t \geq 1} \text{ as our filter, but with a slower convergence rate } O(\frac{1}{\sqrt{n}}). \text{ As an ultimate comparison scheme, we also implemented Kalman filter, which is the optimal filter for above Gaussian signal and noise. Note that although Kalman filter is also a linear filter, the order is not finite, but is increasing with } t. \]

Figure 3.1(a) shows the MSE results of our universal filter \( \hat{X}_t^*(Y^t) \), the noisy predictor \( \hat{X}_{t,ZS}(Y^{t-1}) \), the gradient-descent filter \( \hat{X}_{t,GD}(Y^t) \), the best FIR filter \( u^T Y_t \) that achieves (3.3), and Kalman filter, for a single realization of the signal and the noise. Since the convergence of \( \hat{X}_{t,GD}(Y^t) \) with \( \eta_t = \frac{1}{\sqrt{t}} \) was extremely slow in our experiment, we instead plot with a faster learning rate \( \eta_t = \frac{1}{t} \). First thing to note in Figure 3.1(a) is that the performance of the best FIR filter of order \( d = 5 \) nearly overlaps with the performance of the optimal Kalman filter. This may be due to the diminishing dependency of noisy signal on the past and enhances justification of our focus on the finite-order filters. Now, from the MSE curve of our universal filter, we can clearly see that our filter, which only observes the noisy signals causally together with the knowledge of the noise variance, successfully attains the performance of the
best FIR filter with the same order, which is determined by a complete knowledge on \((X^n, Y^n)\), as guaranteed by our Theorem 2. In addition, from above observation, we notice that our filter nearly attains the optimal performance of Kalman filter with almost negligible margin as the sequence length increases. Moreover, we observe that the convergence rate of our filter is much faster than the gradient-descent filter \(\hat{X}_{t, GD}(Y^t)\), which is again predicted by our theorem. Thus, although the gradient-descent filter may have the same asymptotically optimal performance as our filter, it performs poorly in practice with finite-length signal. It is also obvious from the figure that the noisy predictor \(\hat{X}_{t, BS}(Y^{t-1})\) is not able to achieve the performance of the best FIR filter. This is because the noisy predictor does not have an access to \(Y_t\) in estimating \(X_t\), whereas \(Y_t\) is the most important observation for \(X_t\). Therefore, this experiment demonstrates that our universal filter successfully generalizes the noisy predictor in [45] to the filtering setting. Figure 3.1(b) presents the result of an average performance of 100 different sample paths of the signal and the noise. We observe that the performance and convergence rate of each scheme for a single sample path is consistent with the average performance. This asserts the high probability result of our theorem.

### 3.7.2 Nonlinear, stochastic signal

Our next example considers the case where the clean signal involves nonlinear terms. That is, we consider the following underlying nonlinear signal

\[
X_t = 0.1X_{t-1} - 0.5\cos(3X_{t-1}) + 0.4\sin(X_{t-2}) + 0.1X_{t-2} + Z_t, \quad t = 2, 3, \cdots (3.43)
\]

where \(\{Z_t\}_{t \geq 0}\) is iid \(\mathcal{N}(0,1)\) and \(X_t = Z_t\) for \(t = 0,1\), which also appears in [54, Section VI]. We pass this signal again through the additive channel (3.1) with \(\{N_t\}_{t \geq 1}\) iid \(\mathcal{N}(0,1)\), independent of \(\{X_t\}_{t \geq 1}\). We again experimented with \(d = 5\) for our filter and \(n = 10^4\). Unlike the autoregressive signal case, Kalman filter is neither optimal nor implementable for this signal. Instead, we compare our filter with the extended Kalman filter [55], which is commonly used in practice for filtering nonlinear signals of known statistics. Note that, however, the extended Kalman
filter is not an optimal filter, but just one heuristic that approximates the nonlinear terms with the first order Taylor expansions. Therefore, the extended Kalman filter would not necessarily perform better than our universal filter. Figure 3.2 shows the MSE results of our filter, the noisy predictor, the gradient-descent filter, the best FIR filter, and the extended Kalman filter for the nonlinear signal (3.43). The single sample path result in Figure 3.2(a) again shows the similar result as the autoregressive signal case in Figure 3.1(a). The most notable point of this experiment is that our filter outperforms the performance of the extended Kalman filter. That is, although our filter only competes with linear filters with finite order, since the performance target of our filter is the best FIR filter that is determined by the actual realization of the signal and the noise, it can outperform the extended Kalman filter which is nonlinear and knows the signal model (3.43). Again, Figure 3.2(b) shows the average performance which is consistent with the single sample path result.

3.7.3 Universality of our filter

The above two examples show that our filter, which does not know about the underlying signal model, can learn about the signal and perform as well as or better than the schemes that rely on the exact knowledge of the signal model. The third example stresses this powerful universality of our filter. We again experiment with the first order autoregressive signal and the nonlinear signal, but with different models. That is,

\[ X_t = \alpha X_{t-1} + Z_t, \quad \alpha = 0.1, \quad t = 1, 2, \ldots, \quad (3.44) \]

and

\[ X_t = 0.1X_{t-1} - 2\cos(5X_{t-1}) + \sin(0.1X_{t-2}) + 0.1X_{t-2} + Z_t, \quad t = 2, 3, \ldots \quad (3.45) \]

with the same initial conditions as (3.40) and (3.43), respectively, are now the inputs to the additive channel (3.1) with \( \{N_t\}_{t \geq 1} \) iid \( \sim \mathcal{N}(0,1) \), independent of \( \{X_t\}_{t \geq 1} \). Since our filter does not depend on the signal model, the exact same scheme as
what we used for the above two experiments is again applied for filtering both (3.44) and (3.45). For comparison schemes, we use the Kalman filter that is matched to (3.40) for (3.44) and the extended Kalman filter that is matched to (3.43) for (3.45) to see the sensitivity of those schemes to the underlying signal models. Figure 3.3 shows the average MSE results of 100 experiments with \( d = 5 \) for our filter and sequence length \( n = 10^4 \). We observe that our filter outperforms the mismatched Kalman and extended Kalman filter for both cases with significant margins. These experiments plainly show that our filter *universally* attains the performance of the best FIR filter regardless of the signal models, whereas schemes that heavily depend on the knowledge of the signal models are very sensitive to the assumed models. Therefore, when there are uncertainties in the signal model, which is usually the case in practice, our universal filter clearly has a potential in improving on conventional filtering schemes that require knowledge of signal models.

3.7.4 Filtering deterministic signal

We next consider the case where the underlying signal is the Henon map,

\[
X_t = 1 - 1.4X_{t-1}^2 + 0.3X_{t-2}, \quad t = 2, 3, \cdots \tag{3.46}
\]

with \( X_0 = X_1 = 0 \), which is deterministic but known to exhibit chaotic behavior. For the demonstration of the chaotic behavior of Henon map, refer to [54, Section VI, Figure 8]. Again, this signal is corrupted by the channel (3.1) with \( \{N_t\}_{t \geq 1} \) iid \( \sim \mathcal{N}(0,1) \). Now, since the underlying signal is deterministic, a filtering scheme that relies on the knowledge of the signal model does not make sense in this case, because knowing the model is equivalent to knowing the signal completely. Therefore, it is not clear what conventional schemes to apply for filtering the above Henon map-generated signal. However, we can still apply our universal filter since it does not depend on the underlying signal. Figure 3.4 again shows the average MSE results of our filter, the noisy predictor, the gradient-descent filter, and the best FIR filter with \( d = 5 \) and \( n = 10^4 \). We observe that our filter reduces MSE significantly from the noise variance 1, which is the MSE of *saying-what-you-see* filter \( \hat{X}_t(Y^t) = Y_t \), and outperforms both
the noisy predictor and the gradient-descent filter significantly.

### 3.7.5 Effect of constants on the convergence rate of regret

Finally, our next example illustrates the effect of constants to the convergence rate of our filter, which was suppressed in the presentation of our theorem. We set the nonlinear signal (3.43) as the underlying signal and measured the impact of three constants: the signal bound $B_X$, the noise variance $\sigma^2$, and the filter order $d$. We omitted to vary the bound on the noisy signal $B_N$ since it is closely related to $\sigma^2$, and varying $B_N$ would show the similar behavior as varying $\sigma^2$. Moreover, instead of varying $B_X$ of the signal directly, we varied the variance of the innovation $\{Z_t\}_{t \geq 1}$ denoted as $\sigma^2_Z$ for the sake of simple simulation. Clearly, the effect of varying $B_X$ is tied up with that of varying $\sigma^2_Z$. Figure 3.5 summarizes the results of the experiments. First, note that instead of MSEs, the regrets,

$$
\frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{X}_t(Y^t))^2 - \min_{u \in \mathcal{U}} \frac{1}{n} \sum_{t=1}^{n} (x_t - u^T Y_t)^2
$$

are plotted, and the scale of y-axis of the plots are slightly different. The plots are again averages of 100 realizations with sequence length $n = 10^4$. Figure 3.5(a) shows the effect of the bound on the signal by experimenting with varying $\sigma^2_Z$. The noise variance $\sigma^2$ was fixed to 1, and the filter order was $d = 5$. We can observe that as signal amplitude becomes large, the convergence of regret gets attenuated, but not so severely. On the contrary, Figure 3.5(b) shows the effect of the noise variance and the bound on the noise by experimenting with varying $\sigma^2$. In this case, the innovation variance $\sigma^2_Z$ was fixed to 1, and the filter order was again $d = 5$. The figure shows that larger noise variance, or smaller signal-to-noise ratio have an impact in attenuating the convergence rate of the regret more severely than the vice versa case. Figure 3.5(c) shows the effect of the filter order $d$ on the convergence rate of regret. The innovation variance $\sigma^2_Z$ and the noise variance $\sigma^2$ were all set to 1 in this experiment. We observe that, although the dependency on $d$ was exponential in our upper bound (3.33), the slowdown of the convergence rate is not so severe in this case. Overall, although
a qualitative statement, we state that despite the complex constant expressions in our analysis, the effect of those constants are not as severe as we got in our bound. Indeed, we believe this tendency of the dependency on the constants would mostly be the case in practice, since many of the constant bounds are obtained from the worst case scenario, i.e., signal or noise having always the maximum amplitudes.

From this representative set of simulations, we observe that our simple universal filter provides considerable performance gains in filtering noisy signals, especially when there are uncertainties in the underlying signal models.

3.8 Appendix

3.8.1 Proof of Lemma 10

Proof: Fix $x^n \in \mathcal{D}^n$. Consider

$$E\left[(x_t - w^T_{t-1} Y_t)^2 | Y^{t-1}\right]$$

$$= E\left[(x_t^2 - 2x_t w^T_{t-1} Y_t + w^T_{t-1} Y_t Y^T_t w_{t-1}) | Y^{t-1}\right]$$

$$= E\left[\left\{(Y_t^2 - \sigma^2) - (2Y_t w^T_{t-1} Y_t - 2w^T_{t-1} c) + w^T_{t-1} Y_t Y^T_t w_{t-1}\right\} | Y^{t-1}\right]$$

$$= E\left[\left\{(Y_t - w^T_{t-1} Y_t)^2 + 2w^T_{t-1} c - \sigma^2\right\} | Y^{t-1}\right],$$

where (3.47) follows since $E(x_t^2 | Y^{t-1}) = x_t^2 = E(Y_t^2 - \sigma^2) = E(Y_t^2 - \sigma^2 | Y^{t-1})$ and $E(x_t w^T_{t-1} Y_t | Y^{t-1}) = w^T_{t-1} E(x_t Y_t | Y^{t-1}) = w^T_{t-1} E(Y_t Y_t - c | Y^{t-1}) = E(Y_t w^T_{t-1} Y_t - w^T_{t-1} c | Y^{t-1})$. Therefore, for all $t \geq 1$,

$$E\left[\Lambda_t(w_{t-1}) - \{\ell_t(w_{t-1}) - \sigma^2\} | Y^{t-1}\right] = 0$$

(3.48)

Note that the condition $w_{t-1} \in \sigma(Y^{t-1})$ is crucial to have the above equality. Hence, $\{\Lambda_t(w_{t-1}) - \{\ell_t(w_{t-1}) - \sigma^2\}\}_{t \geq 1}$ is a martingale difference, and therefore, $\{\sum_{t=1}^n \Lambda_t(w_{t-1}) - \sum_{t=1}^n \{\ell_t(w_{t-1}) - \sigma^2\}\}_{n \geq 1}$ is a martingale. ■
3.8.2 Proof of Lemma 11

Proof: The argument for Part (a) and Part (b) almost coincides with that of [6, Chapter 11.7] except for the constant vector \( c \) in the definition of \( \ell_t(u) \). But, that difference hardly affects the argument.

(a) From (3.4),

\[
\mathbf{w}_t^* = A_t^{-1}\left(\sum_{i=1}^{t} \{\mathbf{Y}_i \mathbf{Y}_i^T - \mathbf{c}\}\right) = A_t^{-1}\left(\mathbf{A}_{t-1} \mathbf{w}_{t-1}^* + \mathbf{Y}_t \mathbf{Y}_t^T - \mathbf{c}\right)
\]

\[
= A_t^{-1}\left(\mathbf{A}_t \mathbf{w}_{t-1}^* - \mathbf{Y}_t \mathbf{Y}_t^T \mathbf{w}_{t-1}^* + \mathbf{Y}_t \mathbf{Y}_t^T - \mathbf{c}\right)
\]

\[
= \mathbf{w}_{t-1}^* - A_t^{-1}\{\left(\mathbf{w}_{t-1}^T \mathbf{Y}_t - \mathbf{Y}_t\right) \mathbf{Y}_t + \mathbf{c}\}
\]

Also,

\[
\|\mathbf{w}_t^*\| = \|A_t^{-1}\left(\sum_{i=1}^{t} \{\mathbf{Y}_i \mathbf{Y}_i^T - \sigma^2 \mathbf{I}\}\mathbf{e}_1\right)\|
\]

\[
= \|A_t^{-1}\left(\sum_{i=1}^{t} \{\mathbf{Y}_i \mathbf{Y}_i^T - \sigma^2 \mathbf{I}\}\mathbf{e}_1\right)\|
\]

\[
= \|A_t^{-1}\left(\mathbf{A}_t - \mathbf{I} - 2\sigma^2 \mathbf{I}\right)\mathbf{e}_1\|
\]

\[
\leq \|\mathbf{I} - (1 + 2\sigma^2)\mathbf{A}_t^{-1}\| \leq 1 + (1 + 2\sigma^2)\lambda_{\max}(A_t^{-1})
\]

(b) The bound can be obtained by following inequalities.

\[
|\mathbf{w}_{t-1}^T \mathbf{Y}_t - \mathbf{Y}_t| = |(\mathbf{w}_{t-1}^* - \mathbf{e}_1)^T \mathbf{Y}_t|
\]

\[
= \left|(A_t^{-1}\left(\sum_{i=1}^{t-1} \{\mathbf{Y}_i \mathbf{Y}_i^T - \sigma^2 \mathbf{I}\} - \mathbf{A}_{t-1}^{-1}\right)\mathbf{e}_1)^T \mathbf{Y}_t\right|
\]

\[
\leq \|(1 + (t - 1)\sigma^2)\mathbf{A}_{t-1}^{-1}\mathbf{e}_1\| \cdot \|\mathbf{Y}_t\|
\]

\[
\leq (1 + (t - 1)\sigma^2)\|\mathbf{A}_{t-1}^{-1}\| \cdot \|\mathbf{Y}_t\|
\]

\[
= (1 + (t - 1)\sigma^2)\lambda_{\max}(A_{t-1}^{-1}) \cdot \|\mathbf{Y}_t\|.
\]
where (3.49) is from the definition (3.4), (3.50) is from Cauchy-Schwartz inequality, (3.51) is from the definition of matrix norm, and (3.52) is from the fact that $A_{t-1}^{-1}$ is a symmetric matrix.

(c) From Definition 4, $\ell_t(u) = \{L_t(u) - L_t(w_t^*)\} + \{L_t(w_t^*) - L_{t-1}(u)\}$ and $\ell_t(\hat{w}_{t-1}^*) = \{L_t(\hat{w}_{t-1}^*) - L_t(w_t^*)\} + \{L_t(w_t^*) - L_{t-1}(\hat{w}_{t-1}^*)\}$. Hence,

$$\ell_t(\hat{w}_{t-1}^*) - \ell_t(u) = \{L_t(\hat{w}_{t-1}^*) - L_t(w_t^*)\} - \{L_t(u) - L_t(w_t^*)\} + \{L_t(w_t^*) - L_{t-1}(\hat{w}_{t-1}^*)\} \leq \{L_t(\hat{w}_{t-1}^*) - L_t(w_t^*)\} - \{L_t(u) - L_t(w_t^*)\} + \{L_t(w_t^*) - L_{t-1}(w_{t-1}^*)\},$$

(3.53)

where (3.53) holds since $L_{t-1}(w_{t-1}^*) \leq L_{t-1}(\hat{w}_{t-1}^*)$ from definition in (3.4). Therefore, summing over $t$ leads to

$$\sum_{t=1}^{n} \{\ell_t(\hat{w}_{t-1}^*) - \ell_t(u)\} \leq \{L_0(u) - L_0(w_0^*)\} - \{L_n(u) - L_n(w_n^*)\} + \sum_{t=1}^{n} \{L_t(\hat{w}_{t-1}^*) - L_t(w_t^*)\} \leq \|u\|^2 + \sum_{t=1}^{n} \{L_t(\hat{w}_{t-1}^*) - L_t(w_t^*)\}.$$

(3.54)

The inequality in (3.54) holds since $L_n(w_n^*) \leq L_n(u)$ for all $u \in \mathbb{R}^d$. Now, since $L_t(u)$ is convex, and $w_t^*$ is its minimizing argument, $\nabla L_t(w_t^*) = 0$. Following some algebra, we obtain

$$L_t(\hat{w}_{t-1}^*) - L_t(w_t^*) = L_t(\hat{w}_{t-1}^*) - L_t(w_t^*) - (\hat{w}_{t-1}^* - w_t^*)^T \nabla L_t(w_t^*) = (\hat{w}_{t-1}^* - w_t^*)^T A_t(\hat{w}_{t-1}^* - w_t^*),$$

which proves the lemma. \(\blacksquare\)
3.8.3 Proof of Lemma 12

Proof:

(a) From the union bound,

\[
P\left(\left\| \frac{1}{t} \sum_{i=1}^{t} Y_i Y_i^T - \left( \sigma^2 I + \frac{1}{t} \sum_{i=1}^{t} x_i x_i^T \right) \right\|_1 > \epsilon \right) \leq \sum_{1 \leq a, b \leq d} P \left( \left\| \left( \frac{1}{t} \sum_{i=1}^{t} Y_i Y_i^T \right)_{ab} - \left( \sigma^2 I + \frac{1}{t} \sum_{i=1}^{t} x_i x_i^T \right)_{ab} \right\| > \frac{\epsilon}{t^2} \right),
\]

(3.55)
(3.56)

where \((A)_{ab}\) denotes the \(ab\)-th entry of the matrix \(A\). Since \(\sum_{i=1}^{t} Y_i Y_i^T = \sum_{i=1}^{t} x_i x_i^T + 2 \sum_{i=1}^{t} x_i N_i^T + \sum_{i=1}^{t} N_i N_i^T\), we consider the concentration of

\[
\frac{1}{t} \left( 2 \sum_{i=1}^{t} x_i N_i^T + \sum_{i=1}^{t} N_i N_i^T \right).
\]

We note that

\[
\left( 2 \sum_{i=1}^{t} x_i N_i^T + \sum_{i=1}^{t} N_i N_i^T \right)_{ab} = 2 \sum_{i=1}^{t} x_{i+1-a} N_{i+1-b} + \sum_{i=1}^{t} N_{i+1-a} N_{i+1-b},
\]

and consider the cases when \(a = b\) and \(a \neq b\), separately. When \(a = b\), we can verify that the sequence \(\{2x_{i+1-a} N_{i+1-b} + N_{i+1-b}^2 - \sigma^2\}_{i \geq 1}\) is a martingale difference with respect to \(\{N_{i+1-b}\}_{i \geq 1}\), since \(\{N_i\}\) are assumed to be independent with \(EN_i = 0, EN_i^2 = \sigma^2\) for all \(i\). When \(a \neq b\), without loss of generality, we can assume \(a > b\). Then, we can again verify that \(\{2x_{i+1-a} N_{i+1-b} - N_{i+1-a} N_{i+1-b}\}_{i \geq 1}\) is a martingale difference with respect to \(\{N_{i+1-b}\}_{i \geq 1}\), since \(N_{i+1-a}\) and \(N_{i+1-b}\) are zero mean and independent. Therefore, since we assumed that \(\{x_t\}\)'s and \(\{N_t\}\)'s are all bounded, we can apply Hoeffding-Azuma inequality [6, Section A.1.3] to get the bound,

\[
P\left( \left\| \frac{1}{t} \sum_{i=1}^{t} \{N_i N_i^T + 2x_i N_i^T\} \right\|_1 > \epsilon \right) 
\leq 2 \exp \left( -\frac{2\epsilon^2}{B^2_N (B_N + 2B_X)^2 d^4 t} \right),
\]

(3.57)
which, combined with (3.56), proves part (a).

(b) In [56, (2.2)], we find the inequality

\[
\max_i \min_j |\lambda_j - \mu_i| \leq \frac{d + 2}{d} G_{AB}^{1 - 1/d} \|A - B\|_1, \tag{3.58}
\]

where \(\{\lambda_j\}_{1 \leq j \leq d}\) and \(\{\mu_i\}_{1 \leq i \leq d}\) are the eigenvalues of \(d\)-by-\(d\) matrix \(A\) and \(B\), respectively, and \(G_{AB} = \max_{i,j}(|a_{ij}|, |b_{ij}|)\). Let us denote \(F_{AB} = \frac{d + 2}{d} G_{AB}^{1 - 1/d}\).

Then, since (3.58) is symmetric in \(A\) and \(B\), the inequality \(\max_i \min_j |\mu_i - \lambda_j| \leq F_{AB} \|A - B\|_1\) is also true. Now, we observe that

\[
|\lambda_{\min}(A) - \lambda_{\min}(B)| \leq \max_i \min_j |\lambda_j - \mu_i|, \max_j \min_i |\mu_i - \lambda_j|,
\]

due to the symmetry of \(|\lambda_{\min}(A) - \lambda_{\min}(B)|\) in \(A\) and \(B\), and, thus, deduce that

\[
|\lambda_{\min}(A) - \lambda_{\min}(B)| \leq F_{AB} \|A - B\|_1, \tag{3.59}
\]

i.e., the minimum eigenvalue is a Lipschitz continuous function of the elements of the matrix. Now, denote the event \(E_t = \{\omega : \frac{1}{t} \sum_{i=1}^{t} x_i x_i^T \leq \epsilon\}\). Then, if \(\omega \in E_t\), we have

\[
\frac{1}{t} \lambda_{\min}(K_t) \geq \lambda_{\min}(\sigma^2 I + \frac{1}{t} \sum_{i=1}^{t} x_i x_i^T) - \epsilon F \tag{3.60}
\]

\[
\geq \sigma^2 - \epsilon F. \tag{3.61}
\]

where \(F = \frac{d + 2}{d} (B_X + B_N)^{2(1 - 1/d)}\), (3.60) is from (3.59), and (3.61) is from the fact that \(1/t \sum_{i=1}^{t} x_i x_i^T\) is positive semidefinite. Since \(F > 0\), by choosing \(\epsilon = \frac{\sigma^2}{2F}\), part (b) is proven by applying the result of part (a).  \(\square\)
3.8.4 Proof of Lemma 13

Proof: Note first that, by the union bound, for any \( m \)

\[
P\left( \bigcup_{t=m}^{\infty} \{ Y_t > 0 \} \right) \leq 2d^2 \sum_{t=m}^{\infty} \exp(-tC) = \frac{2d^2}{1 - \exp(-C)} \exp(-mC). \tag{3.62}
\]

Fix \( m < \left( \frac{3\nu}{c_1} \right)^{1/3} - c_2 - 1 \) for sufficiently large \( n \). Then, consider

\[
P\left( \frac{1}{n} \sum_{t=1}^{n} Y_t > \epsilon \right) = P\left( \sum_{t=1}^{n} Y_t > n\epsilon \right) = P\left( \sum_{t=1}^{m} Y_t + \sum_{t=m+1}^{n} Y_t > n\epsilon \right)
\]

\[
\leq P\left( \sum_{t=m+1}^{n} Y_t > n\epsilon - c_1 \sum_{t=1}^{m} (c_2 + t)^2 \right) \tag{3.63}
\]

\[
\leq P\left( \sum_{t=m+1}^{n} Y_t > n\epsilon - c_1 \frac{(m + c_2 + 1)^3}{3} \right) \tag{3.64}
\]

\[
\leq P\left( \sum_{t=m+1}^{n} Y_t > 0 \right) \tag{3.65}
\]

\[
\leq P\left( \bigcup_{t=m+1}^{n} \{ Y_t > 0 \} \right) \tag{3.66}
\]

\[
\leq P\left( \bigcup_{t=m+1}^{\infty} \{ Y_t > 0 \} \right) \tag{3.67}
\]

where (3.63) follows from the given bound \( Y_t \leq c_1 (c_2 + t)^2 \); (3.64) follows from \( \sum_{t=1}^{m} (c_2 + t)^2 \leq \sum_{t=1}^{m+c_2} t^2 \leq \sum_{t=1}^{m+c_2} t^2 \leq \frac{(m+c_2+1)^3}{3} \); (3.65) follows from the condition on \( m \); (3.66) follows from the union of events, and (3.67) follows from (3.62). In particular, taking \( m = \lfloor \left( \frac{3\nu}{c_1} \right)^{1/3} - c_2 \rfloor - 1 \) gives

\[
P\left( \frac{1}{n} \sum_{t=1}^{n} Y_t > \epsilon \right) \leq \frac{2d^2}{1 - \exp(-C)} \exp\left( - \left( \frac{3\nu}{c_1} \right)^{1/3} - c_2 \right) C
\]

and proves the lemma. ■
3.8.5 Proof of Lemma 14

To simplify the notation, we will use the $\Lambda_t(u)$ notation in Definition 4. First, note that we have following decomposition:

$$\sum_{t=1}^{n} \Lambda_t(\hat{w}_{t-1}^*) - \left\{\sum_{t=1}^{n} \Lambda_t(u) + ||u||^2\right\}$$

$$= \sum_{t=1}^{n} \left\{\Lambda_t(\hat{w}_{t-1}^*) - \left\{\ell_t(\hat{w}_{t-1}^*) - \sigma^2\right\}\right\} + \sum_{t=1}^{n} \left\{\ell_t(\hat{w}_{t-1}^*) - \ell_t(u)\right\} - ||u||^2$$

$$- \sum_{t=1}^{n} \left\{\Lambda_t(u) - \left\{\ell_t(u) - \sigma^2\right\}\right\}. \quad (3.68)$$

Then, from the union bound, we have

$$P\left(\frac{1}{n} \sum_{t=1}^{n} \Lambda_t(\hat{w}_{t-1}^*) - \frac{1}{n} \left\{\sum_{t=1}^{n} \Lambda_t(u) + ||u||^2\right\} > \epsilon + \Theta\left(\frac{\log n}{n}\right)\right)$$

$$\leq P\left(\frac{1}{n} (a) \geq \frac{\epsilon}{3}\right) + P\left(\frac{1}{n} (b) \geq \frac{\epsilon}{3} + \Theta\left(\frac{\log n}{n}\right)\right) + P\left(-\frac{1}{n} (c) \geq \frac{\epsilon}{3}\right). \quad (3.69)$$

Since $\{\hat{w}_{t-1}^*\}_{t \geq 1}$ and $u$ are bounded, and (a) and (c) are bounded martingale from Lemma 10. Thus, we can use the Hoeffding-Azuma inequality [6, Lemma A.7] to bound the first and third term of (3.69) as

$$P\left(\frac{1}{n} (a) \geq \frac{\epsilon}{3}\right) \leq \exp\left(-n\frac{2\epsilon^2}{9L_{\max}^2}\right) \quad \text{and} \quad (3.70)$$

$$P\left(\frac{1}{n} (c) \leq -\frac{\epsilon}{3}\right) \leq \exp\left(-n\frac{2\epsilon^2}{9L_{\max}^2}\right), \quad (3.71)$$

where

$$L_{\max} \triangleq \max_{x_t \in D, y_t \in [-(B_X + B_N), (B_X + B_N)]^d, u \in U} \left\{\Lambda_t(u) - \ell_t(u) + \sigma^2\right\}.$$
It is obvious that (3.70) and (3.71) vanish much faster than \(\exp(-\Theta(n^{1/3}))\), and thus, the remaining property we need is

\[
P\left(\frac{1}{n}(b) > \frac{\epsilon}{3} + \Theta\left(\frac{\log n}{n}\right)\right) \leq \exp\left(-\Theta\left(n^{1/3}\right)\right).
\] (3.72)

To show this, recall (3.14) and (3.19) and define

\[
Z_{1t} = \left\{\sigma^2 + b_1 \frac{1 + \sigma^2 t}{1 + \lambda_{\min}(K_t)}\right\}^2 \cdot \frac{1}{1 + \lambda_{\min}(K_t)}
+ \|A_t\|\|\hat{w}_{t-1} - w_{t-1}^*\|^2 + 2\|R_t Y_t + c\|\|\hat{w}_{t-1} - w_{t-1}^*\|,
\]

\[
Z_{2t} = \left\{\sigma^2 + b_1 \frac{1 + \sigma^2 t}{1 + (\sigma^2/2)t}\right\}^2 \cdot \frac{1}{1 + (\sigma^2/2)t}.
\]

Then, by denoting \(Z_t = Z_{1t} - Z_{2t}\), again from Lemma 11(a) and Lemma 12(b), we have \(P(Z_t > 0) \leq 2d^2 \exp\left(-\frac{\sigma^4}{2}\epsilon\frac{t}{\epsilon}\right)\). Since \(K_t\) is positive semi-definite and \(Z_{2t} \geq 0\), \(Z_t \leq Z_{1t} \leq (\sigma^2 + b_1(1 + \sigma^2 t))^2 = (b_1\sigma^2)^2(\frac{1}{b_1} + \frac{1}{\sigma^2} + t)^2\). Hence, we can apply Lemma 13 and show

\[
\exp\left(-\Theta\left(n^{1/3}\right)\right) \geq P\left(\frac{1}{n} \sum_{t=1}^{n} Z_t > \frac{\epsilon}{3}\right) \geq P\left(\frac{1}{n} \sum_{t=1}^{n} Z_{1t} > \frac{\epsilon}{3} + \frac{1}{n} \sum_{t=1}^{n} Z_{2t}\right)
\geq P\left(\frac{1}{n} \sum_{t=1}^{n} \{\ell_t(w_{t-1}^*) - \ell_t(u)\} - \|u\|^2\right) > \frac{\epsilon}{3} + \Theta\left(\frac{\log n}{n}\right),
\] (3.73)

(3.74)

(3.75)

where (3.73) follows from Lemma 13; (3.74) follows from \(Z_t = Z_{1t} - Z_{2t}\), and (3.75) follows from identical steps as in (3.12)-(3.18), the fact that \(Z_{10} = Z_{20}\), and \(w_{*1}^* \triangleq 0\). Therefore, (3.72) and the lemma are proved. \(\blacksquare\)
Figure 3.1: MSEs for AR(1) signal (3.40). Figure 3.1(a) is for a single sample path, and Figure 3.1(b) is for the average of 100 experiments.
Figure 3.2: MSEs for nonlinear signal (3.43). Figure 3.2(a) is for a single sample path, and Figure 3.2(b) is for the average of 100 experiments.
Figure 3.3: Average MSEs for signals (3.44) and (3.45). Kalman filter and the extended Kalman filter used here are matched to wrong signals (3.40) and (3.43), respectively.
Figure 3.4: MSE results averaged over 100 experiments for Henon map (3.46).
Figure 3.5: Regrets averaged over 100 experiments for nonlinear signal (3.43) with varying parameters.
Chapter 4

Discrete denoising with shifts

4.1 Introduction

We now shift our gear to the discrete denoising problem, the denoising problem described in Chapter 1 with finite-alphabet source and Discrete Memoryless Channel (DMC). Clearly, a denoiser is much more powerful than a filter in estimating the source since it can wait until it observes the whole noisy data. Universal discrete denoising, in which no statistical or other properties are known a priori about the underlying clean data and the goal is to attain optimum performance, was considered and solved in [7]. The main problem setting there is the semi-stochastic one as in Chapter 3, in which the underlying signal is assumed to be an individual sequence, and the randomness is due solely to the channel noise. In this setting, it is unreasonable to expect to attain the best performance among all the denoisers in the world, since for every given sequence, there exists a denoiser that recovers all the sequence components perfectly. Thus, [7] limits the comparison class, also known as expert class, and uses the competitive analysis approach. Specifically, it is shown that regardless of what the underlying individual sequence may be, the Discrete Universal DEnoiser (DUDE) essentially attains the performance of the best sliding window denoiser that would be chosen by a genie with access to the underlying clean sequence, in addition to the observed noisy sequence. This semi-stochastic setting result is shown in [7] to imply the stochastic setting result, i.e., that for any underlying stationary signal, the
DUDE attains the optimal distribution-dependent performance.

The setting of an arbitrary individual sequence, combined with competitive analysis, has been very popular in many other research areas, especially for problems of sequential decision-making. Examples include universal compression [3], universal prediction [4], universal filtering [8], repeated game playing [57, 58, 41], universal portfolios [5], online learning [59, 60], zero-delay coding [61, 62], and much more. A comprehensive account of this line of research can be found in [6]. The beauty of this approach is the fact that it leads to the construction of schemes that perform, on every individual sequence, essentially as well as the best in a class of experts, which is the performance of a genie that had hindsight on the entire sequence before selecting his actions. Moreover, if the expert class is judiciously chosen, the relative sense of such a performance guarantees can, in many cases, imply optimum performance in absolute senses as well.

One extension to this approach is competition with an expert class and a genie that has the freedom to form a compound action, which breaks the sequence into a certain (limited) number of segments, applies different experts in each segment, and achieves an even better performance overall. Note that the optimal segmentation of the sequence and the choice of the best expert in each segment is also determined by hindsight. Clearly, competing with the best compound action is more challenging, since the number of possible compound actions is exponential in the sequence length $n$, and the brute-force vanilla implementation of the ordinary universal scheme requires prohibitive complexity. However, clever schemes with linear complexity that successfully track the best segments and experts have been devised in many different areas, such as online learning, universal prediction [63, 64], universal compression [65, 66], online linear regression [67], universal portfolios [68], and zero-delay lossy source coding [69].

In this chapter, we expand the idea of compound actions and apply it to the discrete denoising problem. The motivation of this expansion is natural: the characteristics of the underlying data in the denoising problem often tend to be time- or space-varying. In this case, determining the best segmentation and the best expert for each segment requires complete knowledge of both clean and noisy sequences.
Therefore, whereas the challenge in sequential decision-making problems is to track
the shift of the best expert based on the past, true observation, the challenge in the
denoising problem is to learn the shift based on the entire, but noisy, observation. We
extend DUDE to meet this challenge and provide results that parallel and strengthen
those of [7].

Specifically, we introduce the S-DUDE and show first that, for every underlying
noiseless sequence, it attains the performance of the best compound finite-order slid­
ing window denoiser (concretely defined later), both in expectation and in a high
probability sense. We develop our scheme in the semi-stochastic setting as in [7].
The toolbox for the construction and analysis of our scheme draws on ideas devel­
oped in [8]. We circumvent the difficulty of not knowing the exact true loss by using
an observable unbiased estimate of it. This kind of an estimate has proved to be very
useful in [8] and [70] to devise schemes for filtering and for denoising with dynamic
contexts. Building on this semi-stochastic setting result, we also establish a stochas­
tic setting result, which can be thought of as a generalization and strengthening of
the stochastic setting results of [7], from the world of stationary processes to that of
piecewise stationary processes.

Our stochastic setting has connections to other areas, such as change-point detec­
tion problems in statistics [71, 72] and switching linear dynamical systems in machine
learning and signal processing [73, 74]. Both of these lines of research share a com­
mon approach with S-DUDE, in that they try to learn the change of the underlying
time-varying parameter or state of stochastic models, based on noisy observations
of the parameter or state. One difference is that, whereas our goal is the noncausal
estimation, i.e., denoising, of the general underlying piecewise stationary process,
the change-point detection problems mainly focus on sequentially detecting the time
point where the change of model happened. Another difference is that the switching
linear dynamical systems focus on a special class of underlying processes, the linear
dynamical system. In addition, they deal with continuous-valued signals, whereas our
focus is the discrete case, with finite-alphabet signals.

As we explain in detail, the S-DUDE can be practically implemented using a
two-pass algorithm with complexity (both space and time) linear in the sequence
length and the number of switches. We also present initial experimental results that demonstrate the S-DUDE's potential to outperform the DUDE on both simulated and real data.

The remainder of the chapter is organized as follows. Section 4.2 provides the notation, preliminaries and background for the chapter; in Section 4.3 we present our scheme and establish its strong universality properties via an analysis of its performance in the semi-stochastic setting. Section 4.4 establishes the universality of our scheme in a fully stochastic setting, where the underlying noiseless sequence is emitted by a piecewise stationary process. Algorithmic aspects and complexity of the actual implementation of the scheme is considered in Section 4.5, and some experimental results are displayed in Section 4.6. The detailed proofs are again moved to the Appendix section, Section 4.7.

4.2 Notation, Preliminaries, and Motivation

4.2.1 Notation

We use a combination of notation of [7] and [8]. Let $\mathcal{X}, \mathcal{Z}, \hat{\mathcal{X}}$ denote, respectively, the alphabet of the clean, noisy, and reconstructed sources, which are assumed to be finite. As in [7] and [8], the noisy sequence is a DMC-corrupted version of the clean one, where the channel matrix $\Pi = \{\Pi(x, z)\}_{z \in \mathcal{Z}, x \in \mathcal{X}}$, $\Pi(x, z)$ denoting the probability of a noisy symbol $z$ when the underlying clean symbol is $x$, is assumed to be known and fixed throughout the chapter, and of full row rank. The $z$-th column of $\Pi$ will be denoted as $\pi_z$. Upper case letters will denote random variables as usual; lower case letters will denote either individual deterministic quantities or specific realizations of random variables.

Without loss of generality, the elements of any finite set $\mathcal{V}$ will be identified with $\{0, 1, \cdots, |\mathcal{V}| - 1\}$. We let $\mathcal{V}^\infty$ denote the set of one-sided infinite sequences with $\mathcal{V}$-valued components, i.e., $\mathbf{v} \in \mathcal{V}^\infty$ is of the form $\mathbf{v} = (v_1, v_2, \cdots), v_i \in \mathcal{V}, i \geq 1$. For $\mathbf{v} \in \mathcal{V}^\infty$, let $v^n = (v_1, \cdots, v_n)$ and $v^n_m = (v_m, \cdots, v_n)$. Furthermore, we let $v^{n|l}$ denote the sequence $v^{l-1}v^n_{l+1}$. $\mathbb{R}^\mathcal{V}$ is a space of $|\mathcal{V}|$-dimensional column vectors.
CHAPTER 4. DISCRETE DENOISING WITH SHIFTS

with real-valued components indexed by the elements of $\mathcal{V}$. The $a$-th component of $q \in \mathbb{R}^\mathcal{V}$ will be denoted by either $q_a$ or $q[a]$. Subscripting a vector or a matrix by "max" will represent the difference between the maximum and minimum of all its components. Thus, for example, if $\Gamma$ is a $|\mathcal{Z}| \times |\mathcal{X}|$ matrix, then $\Gamma_{\text{max}}$ stands for $\max_{z \in \mathcal{Z}, x \in \mathcal{Z}} \Gamma(z, x) - \min_{z \in \mathcal{Z}, x \in \mathcal{Z}} \Gamma(z, x)$ (in particular, if the components of $\Gamma$ are nonnegative and $\Gamma(z, x) = 0$ for some $z$ and $x$, then $\Gamma_{\text{max}} = \max_{z \in \mathcal{Z}, x \in \mathcal{X}} \Gamma(z, x)$. In addition, $1\{\cdot\}$ denotes an indicator of the event inside $\{\cdot\}$.

Generally, let the finite sets $\mathcal{Y}, \mathcal{A}$ be, respectively, a source alphabet and an action space. For a general loss function $l : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, a Bayes response for $\zeta \in \mathbb{R}^\mathcal{V}$ under the loss function $l$ is given as

$$b_l(\zeta) = \arg \min_{a \in \mathcal{A}} \zeta^T \cdot L_a,$$  \hspace{1cm} (4.1)

where $L_a$ denotes the column of the matrix of the loss function $l$ corresponding to the $a$-th action, and ties are resolved lexicographically. The corresponding Bayes envelope is denoted as

$$U_l(\zeta) = \min_{a \in \mathcal{A}} \zeta^T \cdot L_a.$$  \hspace{1cm} (4.2)

Note that when $\zeta$ is a probability, namely, it has non-negative components summing to one, $U_l(\zeta)$ is the minimum achievable expected loss (as measured under the loss function $l$) in guessing the value of $Y \in \mathcal{Y}$ which is distributed according to $\zeta$. The associated optimal guess is $b_l(\zeta)$.

An $n$-block denoiser is a collection of $n$ mappings $\hat{X}^n = \{\hat{X}_t\}_{1 \leq t \leq n}$, where $\hat{X}_t : \mathcal{Z}^n \to \hat{\mathcal{X}}$. We assume a given loss function $\Lambda : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$, where the maximum single-letter loss is denoted by $\Lambda_{\text{max}}$, and $\lambda_{\hat{x}}$ denotes the $\hat{x}$-th column of the loss matrix. The normalized cumulative loss of the denoiser $\hat{X}^n$ on the individual sequence pair $(x^n, z^n)$ is represented as

$$L_{\hat{X}^n}(x^n, z^n) = \frac{1}{n} \sum_{t=1}^{n} \Lambda(x_t, \hat{X}_t(z^n)).$$

In words, $L_{\hat{X}^n}(x^n, z^n)$ is the normalized (per-symbol) loss, as measured under the loss
function $\Lambda$, when using the denoiser $\hat{X}^n$ and when the observed noisy sequence is $z^n$ while the underlying clean one is $x^n$. The notation $L_{X^n}$ is extended for $1 \leq i \leq j \leq n$,
\[
L_{X^n}(x^j_i, z^n) = \frac{1}{j - i + 1} \sum_{t=i}^{j} \Lambda(x_t, \hat{X}_t(z^n))
\]
denoting the normalized (per-symbol) loss between (and including) locations $i$ and $j$.

Now, consider the set $\mathcal{S} = \{s : \mathcal{Z} \rightarrow \hat{X}\}$, which is the (finite) set of mappings that take $\mathcal{Z}$ into $\hat{X}$. We refer to elements of $\mathcal{S}$ as “single-symbol denoisers”, since each $s \in \mathcal{S}$ can be thought of as a rule for estimating $X \in \mathcal{X}$ on the basis of $Z \in \mathcal{Z}$. Now, for any $s \in \mathcal{S}$, an unbiased estimator for $\Lambda(x, s(Z))$ (based on $Z$ only), where $x$ is a deterministic symbol and $Z$ is the output of the DMC when the input is $x$, can be obtained as in [8]. First, pick a function $h : \mathcal{Z} \rightarrow \mathbb{R}^\mathcal{X}$ with the property that, for $a, b \in \mathcal{X}$,
\[
E_a h_b(Z) = \sum_{z \in \mathcal{Z}} h_b(z) \Pi(a, z)
\]
\[
= \delta(a, b) \Delta \left\{ \begin{array}{ll}
1, & \text{if } a = b \\
0, & \text{otherwise} \\
\end{array} \right.,
\] (4.3)
where $E_a$ denotes expectation over the channel output $Z$ when the underlying channel input is $a$, and $h_b(z)$ denotes the $b$-th component of $h(z)$. Let $H$ denote the $|\mathcal{Z}| \times |\mathcal{X}|$ matrix whose $z$-th row is $h^T(z)$, i.e., $H(z, b) = h_b(z)$. To see that our assumption of a channel matrix with full row rank guarantees the existence of such an $h$, note that (4.3) can equivalently be stated in matrix form as
\[
\Pi H = I,
\] (4.4)
where $I$ is the $|\mathcal{X}| \times |\mathcal{X}|$ identity matrix. Thus, e.g., any $H$ of the form $H = \Gamma^T (\Pi \Gamma^T)^{-1}$, for any $\Gamma$ such that $\Pi \Gamma^T$ is invertible, satisfies (4.4). In particular, $\Gamma = \Pi$ is a valid choice ($\Pi \Gamma^T$ is invertible, since $\Pi$ is of full row rank) corresponding to the Moore-Penrose generalized inverse [75]. Now, for any $s \in \mathcal{S}$, $\rho(s) \in \mathbb{R}^\mathcal{X}$ denotes the
column vector with $x$-th component

$$
\rho_x(s) = \sum_z \Lambda(x, s(z))\Pi(x, z) = E_x\Lambda(x, s(Z)).
$$

(4.5)

In words, $\rho_x(s)$ is the expected loss using the single-symbol denoiser $s$, while the underlying symbol is $x$. Considering $S$ as an action space alphabet, we define a loss function $\ell : \mathcal{Z} \times S \to \mathbb{R}$ as

$$
\ell(z, s) = h(z)^T \cdot \rho(s).
$$

(4.6)

We observe from (4.3) and (4.5) that $\ell(Z, s)$ is an unbiased estimate of $\Lambda(x, s(Z))$ since

$$
E_x\ell(Z, s) = E_x h(Z)^T \cdot \rho(s) = \sum_{x'} E_x h(x') \rho_x(s) = \rho_x(s) = E_x\Lambda(x, s(Z)) \quad \forall x \in \mathcal{X}.
$$

(4.7)

For $\xi \in \mathbb{R}^Z$, let $B_H(\xi, \cdot) \in S$ be defined by

$$
B_H(\xi, z) = \arg \min_{\xi'} \xi'^T \cdot H \cdot [\lambda_{\xi'} \odot \pi_z],
$$

(4.8)

where, for vectors $v_1$ and $v_2$ of equal dimensions, $v_1 \odot v_2$ denotes the vector obtained by component-wise multiplication. Note that, similarly as in [8, (88),(89)],

$$
B_H(\xi, \cdot) = \arg \min_{s \in S} \sum_z \xi^T \cdot H \cdot [\lambda_{s(z)} \odot \pi_z]
= \arg \min_{s \in S} \xi^T \cdot H \cdot \rho(s)
= \arg \min_{s \in S} \sum_z \xi_z \cdot [h^T(z) \cdot \rho(s)]
= \arg \min_{s \in S} \sum_z \xi_z \cdot \ell(z, s) = b_l(\xi).
$$

(4.9)

Thus, $B_H(\xi, \cdot)$ is a Bayes response for $\xi$ under the loss function $\ell$ defined in (4.6).
4.2.2 Preliminaries

In this section, we summarize the results from [7] and motivate the approach underlying the construction of our new class of denoisers. Analogously as in [8], the n-block denoiser \( \hat{X}^n = \{\hat{X}_t\}_{1 \leq t \leq n} \) can be associated with \( F^n = \{F_t\}_{1 \leq t \leq n} \), where \( F_t : Z^{n \setminus t} \to S \) is defined as follows: \( F_t(z^{n \setminus t}, \cdot) \) is the single-symbol denoiser in \( S \) satisfying

\[
\hat{X}_t(z^n) = F_t(z^{n \setminus t}, z_t) \quad \forall z_t.
\] (4.10)

Therefore, we can adopt the view that at each time \( t \), an n-block denoiser is choosing a single-symbol denoiser based on all the noisy sequence components but \( z_t \), and applying that single-symbol denoiser on \( z_t \) to yield the \( t \)-th reconstruction \( \hat{x}_t \). Conversely, any sequence of mappings into single-symbol denoisers \( F^n \) defines a denoiser \( \hat{X}^n \), again via (4.10). We will adhere to this viewpoint in what follows.

One special class of widely used n-block denoisers is that of k-th order “sliding window” denoisers, which we denote by \( \hat{X}^{nS_k} \). Such denoisers are of the form

\[
\hat{X}^{sk}_t(z^n) = s_k(z^{t+k}_{t-k}), \quad t = k + 1, \ldots, n - k,
\] (4.11)

where \( s_k \) is an element of \( S_k = \{s_k : Z^{2k+1} \to \hat{X}\} \), the (finite) set of mappings from \( Z^{2k+1} \) into \( \hat{X} \).\(^1\) We also refer to \( s_k \in S_k \) as a “k-th order denoiser”. Note that \( S_0 = S \).

From the definition (4.11), it follows that

\[
\hat{X}^{sk}_t(z^n) = \hat{X}^{sk}_j(z^n) \quad \text{whenever} \quad z_{i-k}^{t+k} = z_{j-k}^{t+k}.
\] (4.12)

Following the association in (4.10), we can adopt an alternative view that the k-th order sliding window denoiser chooses a single-symbol denoiser \( s_k(z^{t-1}_{t-k}, z_{t+1}^{t+k}, \cdot) \in S \) at time \( t \) on the basis of the context, and \( \hat{X}^{sk}_t(z^n) = s_k(z_{t-k}^{-1}, z_{t+1}^{t+k}, z_t) \).

We denote \( c_t \equiv (z_{t-k}^{-1}, z_{t+1}^{t+k}) \) as a (two-sided) context for \( z_t \), and define the set of all possible k-th order contexts, \( C_k \equiv \{(u_{-k}^{-1}, u_{t}^{k}) : (u_{-k}^{-1}, u_{t}^{k}) \in Z^{2k}\} \). Then, for given

\(^1\)The value of \( \hat{X}^{sk}_t(z^n) \) for \( t < k \) and \( t > n - k \) is defined, for concreteness and simplicity, as an arbitrary fixed symbol in \( \hat{X} \).
\(z^n\) and for each \(c \in \mathcal{C}_k\), we define

\[
T(c) \triangleq \{ t : c_t = c, \quad k + 1 \leq t \leq n - k \} = \{ t : (z_{t-k}^t, z_{t+1}^{t+k}) = c, \quad k + 1 \leq t \leq n - k \},
\]

the set of indices where the context equals \(c\). Now, an equivalent interpretation for (4.12) is that for each \(c \in \mathcal{C}_k\), the \(k\)-th order sliding window denoiser employs a time-invariant single-symbol denoiser, \(s_k(c, \cdot)\), at all points \(t \in T(c)\). In other words, the sequence \(z^n\) is partitioned into the subsequences associated with the various contexts, and on each such subsequence a time-invariant single-symbol scheme is employed.

In [7], for integers \(k \geq 0\) and \(n > 2k\), the \(k\)-th order minimum loss of \((x^n, z^n)\) is defined by

\[
D_k(x^n, z^n) \triangleq \min_{\hat{x}^n \in \hat{X}^n, s_k} L_{\hat{x}^n}(x_{k+1}^{n-k}, z^n)
\]

\[
= \min_{s_k \in \mathcal{S}_k} \frac{1}{n - 2k} \sum_{t=k+1}^{n-k} \Lambda(x_t, s_k(c_t, z_t)).
\]

(4.14)

The identity of the element \(s_k \in \mathcal{S}_k\) that achieves (4.14) depends not only on \(z^n\), but also on \(x^n\), since (4.14) can be expressed as

\[
\frac{1}{n - 2k} \sum_{c \in \mathcal{C}_k} \left[ \min_{s \in \mathcal{S}} \sum_{\tau \in \tau(c)} \Lambda(x_\tau, s(z_\tau)) \right],
\]

and at each time \(t\), the best \(k\)-th order sliding window denoiser that achieves (4.14) will employ the single-symbol denoiser

\[
\arg\min_{s \in \mathcal{S}} \sum_{\tau \in \tau(c_t)} \Lambda(x_\tau, s(z_\tau)),
\]

(4.15)

which is determined from the joint empirical distribution of pairs \(\{(x_\tau, z_\tau) : \tau \in T(c_t)\}\).

It was shown in [7] that, despite the lack of knowledge of \(x^n\), \(D_k(x^n, Z^n)\) is achievable in a sense made precise below, in the limit of growing \(n\), by a scheme that only has access to \(Z^n\). This scheme is dubbed in [7] as the Discrete Universal DEnoiser.
(DUDE), $\hat{X}_{\text{univ}}^{n,k}$. The algorithm is defined by

$$\hat{X}_{\text{univ},t}^{k}(z^n) = B_H(m(z^n, z_{t-k}^{t+k}), z_t),$$

(4.16)

where $m(z^n, c)$ is the vector of counts of the appearances of the various symbols within the context $c$ along the sequence $z^n$. That is, for all $\beta \in \mathcal{Z}$, $m(z^n, z_{-k}^{-1}, z_k)$ is the $|\mathcal{Z}|$-dimensional column vector whose $\beta$-th component is

$$m(z^n, z_{-k}^{-1}, z_k)[\beta] = \left| \{ t : k + 1 \leq t \leq n - k, z_{t-k} = z_{-k}^{-1} \beta z_k \} \right|,$$

namely, the number of appearances of $z_{-k}^{-1} \beta z_k$ along the sequence $z^n$.

The main result of [7] is the following theorem, pertaining to the semi-stochastic setting of an individual sequence $x = (x_1, x_2, \ldots)$ corrupted by a DMC that yields the stochastic noisy sequence $Z = (Z_1, Z_2, \ldots)$.

**Theorem 3 ([7, Theorem 1])** Take $k = k_n$ satisfying $k_n |\mathcal{Z}|^{2k_n} = o(n / \log n)$. Then, for all $x \in \mathcal{X}^{\infty}$, the sequence of denoisers $\{\hat{X}_{\text{univ}}^{n,k_n}\}$ defined in (4.16) satisfies:

a) $$\lim_{n \to \infty} \left[ L_{\hat{X}_{\text{univ}}^{n,k_n}}(x^n, Z^n) - D_{k_n}(x^n, Z^n) \right] = 0 \quad \text{a.s.}$$

b) $$E \left[ L_{\hat{X}_{\text{univ}}^{n,k_n}}(x^n, Z^n) - D_{k_n}(x^n, Z^n) \right] = O \left( \sqrt{\frac{k_n |\mathcal{Z}|^{2k_n}}{n}} \right).$$

Theorem 3 was further shown in [7] to imply the universality of the DUDE in the fully stochastic setting where the underlying sequence is emitted by a stationary source (and the goal is to attain the performance of the optimal distribution-dependent denoiser).

From (4.16), it is apparent that the DUDE ends up employing a $k$-th order sliding window denoiser (where the sliding window scheme the DUDE chooses depends on $z^n$). Moreover, (4.9) implies that, at each time $t$, DUDE is merely employing the single-symbol denoiser $B_H(m(z^n, z_{t-k}^{t+k}), \cdot) \in \mathcal{S}$, which can be obtained by finding
CHAPTER 4. DISCRETE DENOISING WITH SHIFTS

the Bayes response \( b_\ell \left( m(z^n, z_{i-k}^{+1}, z_{i+k}^{+1}) \right) \) or, equivalently, the mapping in \( S \) given by

\[
\arg\min_{s \in S} \sum_{\tau \in T(c_\tau)} \ell(z_\tau, s),
\]

where \( \ell(z, s) \) is the loss function defined in (4.6). By comparing (4.15) with (4.17), and from Theorem 3, we observe that working with the estimated loss \( \ell(z_\tau, s) \) in lieu of the genie-aided \( \Lambda(x_\tau, s(z_\tau)) \) allows us to essentially achieve the genie-aided performance in (4.14).

4.2.3 Motivation

Our motivation for this chapter is based on the observation that the \( k \)-th order sliding window denoisers ignore the time-varying nature of the underlying sequence \( x^n \). That is, as discussed above, for time instances with the same contexts, the single-symbol denoiser employed along the associated subsequence is time-invariant. In other words, for each \( t \), only the empirical distribution of the sequence \( \{(x_\tau, z_\tau) : \tau \in T(c_t)\} \) matters, but its order of composition, i.e., its time-varying nature, is not considered. It is clear, however, that when the characteristics of the underlying clean sequence \( x^n \) are changing, the (normalized) cumulative loss that is achieved by sliding window denoisers that can shift from one rule to another along the sequence may be strictly lower (better) than (4.14). We now devise and analyze our new scheme that achieves this more ambitious target performance.

4.3 The shifting denoiser (S-DUDE)

In this section, we derive our new class of denoisers and analyze their performance. In Subsection 4.3.1, we begin with the simplest case, competing with shifting symbol-by-symbol denoisers, or, in other words, shifting 0-th order denoisers. The argument is generalized to shifting \( k \)-th order denoisers in Subsection 4.3.2, and the framework and results include Subsection 4.3.1 as a special case. We will use the notation \( S_0 \), instead of \( S \), for consistency in denoting the class of single-symbol denoisers. Throughout
CHAPTER 4. DISCRETE DENOISING WITH SHIFTS

this section, we assume the semi-stochastic setting.

4.3.1 Switching between symbol-by-symbol (0-th order) denoisers

Consider an $n$-tuple of single-symbol denoisers $S = \{s_1, \ldots, s_n\} \in \mathcal{S}_0^n$. Then, as mentioned in Section 4.2.2, for such $S$, we can define the associated $n$-block denoiser $\hat{x}_n^S$ as

$$\hat{x}_n^S(z^n) = s_t(z_t). \quad (4.18)$$

Note that in this case, the single-symbol denoiser applied at each time may depend on the time $t$ (but not on $z^m$, as would be the case for a general denoiser). We also denote the estimated normalized cumulative loss as

$$\tilde{L}_S(z^n) \triangleq \frac{1}{n} \sum_{t=1}^{n} \ell(z_t, s_t), \quad (4.19)$$

whose property is given in the following lemma, which parallels [8, Theorem 4].

**Lemma 15** Fix $\epsilon > 0$. For fixed $S \in \mathcal{S}_0^n$, and all $x^n \in \mathcal{X}^n$,

$$P\left( L_{\hat{x}_n^S}(x^n, Z^n) - \tilde{L}_S(Z^n) > \epsilon \right) \leq \exp \left( -n \frac{2\epsilon^2}{L_{\max}^2} \right) \quad \text{and} \quad (4.20)$$

$$P\left( \tilde{L}_S(Z^n) - L_{\hat{x}_n^S}(x^n, Z^n) > \epsilon \right) \leq \exp \left( -n \frac{2\epsilon^2}{L_{\max}^2} \right), \quad (4.21)$$

where $L_{\max} = \Lambda_{\max} + \ell_{\max}$.

In words, the lemma shows that for every $S \in \mathcal{S}_0^n$, the estimated loss $\tilde{L}_S(Z^n)$ is concentrated around the true loss $L_{\hat{x}_n^S}(x^n, Z^n)$ with high probability, as $n$ becomes large, regardless of the underlying sequence $x^n$.

**Proof of Lemma 15:** See Appendix 4.7.1. ■
Now, let the integer $0 \leq m \leq \lfloor \frac{n}{2} \rfloor$ denote the maximum number of shifts allowed along the sequence. Then, define a set $S^n_{0,m} \subseteq S^n_0$ as

$$S^n_{0,m} = \{ S \in S^n_0 : \sum_{t=2}^{n} 1_{(s_{t-1} \neq s_t)} \leq m \},$$

(4.22)

namely, $S^n_{0,m}$ is the set of $n$-tuples of single-symbol denoisers with at most $m$ shifts from one mapping to another. Analogously to (4.14), for the class of $n$-block denoisers $\hat{X}^{n,S}$ with $S \in S^n_{0,m}$, we define

$$D_{0,m}(x^n, z^n) \triangleq \min_{S \in S^n_{0,m}} L_{\hat{X}^{n,S}}(x^n, z^n)$$

$$= \min_{S \in S^n_{0,m}} \frac{1}{n} \sum_{t=1}^{n} \Lambda(x_t, s_t(z_t)),$$

(4.23)

which is the minimum normalized cumulative loss that can be achieved for $(x^n, z^n)$ by the sequence of $n$ single-symbol denoisers that allow at most $m$ shifts. Our goal in this section is to build a universal scheme that only has access to $Z^n$, but still essentially achieves $D_{0,m}(x^n, Z^n)$.

As hinted by the DUDE, we build our universal scheme by working with the estimated loss. That is, define

$$\hat{S} = \hat{S}(z^n) \triangleq \arg \min_{S \in S^n_{0,m}} \hat{L}_S(z^n),$$

(4.24)

and our $(0, m)$-Shifting Discrete Universal Denoiser (S-DUDE), $\hat{X}^{n,0,m}_{\text{univ}}$, is defined as $\hat{X}^{n,\hat{S}}$. It is clear that, by definition, $L_{\hat{X}^{n,S}}(x^n, z^n) \geq D_{0,m}(x^n, z^n)$ for all $x^n$ and $z^n$, but we can also show that, with high probability, $L_{\hat{X}^{n,S}}(x^n, Z^n)$ does not exceed $D_{0,m}(x^n, Z^n)$ by much, as stated in the following theorem.

**Theorem 4** Let $\hat{X}^{n,0,m}_{\text{univ}}$ be defined as $\hat{X}^{n,\hat{S}}$, where $\hat{S}$ is given in (4.24). Then, for all

\textit{Note that, when }$m = 0$, $S^n_{0,0}$ \textit{is the set of constant }$n$-\textit{tuples consisting of the same single-symbol denoiser.}
$\epsilon > 0$ and $x^n \in X^n$,

\[
P\left( L_{X^n,0,m}(x^n, Z^n) - D_{0,m}(x^n, Z^n) > \epsilon \right) \\
\leq 2\exp \left( -n \left[ -\frac{\epsilon^2}{2L_{\max}^2} - 2 \left\{ h\left( \frac{m}{n} \right) + \frac{(m + 1)\ln N}{n} \right\} \right] \right),
\]

where $h(x) = -x\ln x - (1 - x)\ln(1 - x)$ for $0 \leq x \leq 1$, and $N = |S| = |Z|^{|X|}$. In particular, the right-hand side of the inequality is exponentially small, provided $m = o(n)$.

**Remark:** It is reasonable to expect this theorem to hold, given Lemma 15. That is, since, for fixed $S \in S^n_{0,m}$, $L_S(Z^n)$ is concentrated on $L_{X^n,S}(x^n, Z^n)$, it is plausible that $\hat{S}$ that achieves $\min_{S \in S^n_{0,m}} L_S(Z^n)$ will have a loss $L_{X^n,\hat{S}}(x^n, Z^n)$ close to $\min_{S \in S^n_{0,m}} L_{X^n,S}(x^n, Z^n)$, i.e., $D_{0,m}(x^n, Z^n)$.

**Proof of Theorem 4:** See Appendix 4.7.2. ■

### 4.3.2 Switching between $k$-th order denoisers

Now, we extend the result from Subsection 4.3.1 to the case of shifting between $k$-th order denoisers. The argument parallels that of Subsection 4.3.1. Let $\{s_{k,t}\}_{t=k+1}^{n-k}$ be an arbitrary sequence of the $k$-th order denoiser mappings, i.e., $s_{k,t} \in S_k$ for $k + 1 \leq t \leq n - k$. Now, for given $z^n$, define an $(n - 2k)$-tuple of (k-th order denoiser induced) single-symbol denoisers

\[
S_k(z^n) \triangleq \{s_{k,t}(c_t, \cdot)\}_{t=k+1}^{n-k} \in S_0^{n-2k},
\]

where, to recall, $c_t = (z_{t-k}, z_{t+k})$, and $s_{k,t}(c_t, \cdot)$ is the single-symbol denoiser induced from $s_{k,t} \in S_k$ and $c_t$. For brevity of notation, we will suppress the dependence on $z^n$ in $S_k(z^n)$ and denote it as $S_k$. Then, as in (4.18), we define the associated $n$-block
denoiser $\hat{X}^{n,S_k}$ as

$$
\hat{X}^{S_k^n}(z^n) = s_{k,t}(c_t, z_t).
$$

(4.26)

In addition, extending (4.19), the estimated normalized cumulative loss is given as

$$
\tilde{L}_{S_k}(z^n) = \frac{1}{n - 2k} \sum_{t=k+1}^{n-k} \ell(z_t, s_{k,t}(c_t, \cdot)).
$$

(4.27)

Then, we have the following lemma, which parallels Lemma 15.

**Lemma 16** Fix $\epsilon > 0$. For any fixed sequence $\{s_{k,t}\}_{t=k+1}^{n-k}$, and all $x^n \in X^n$,

$$
\Pr\left( L_{X^n,s_k}(x_{k+1}^{n-k}, Z^n) - \tilde{L}_{S_k}(Z^n) > \epsilon \right) \leq (k + 1) \exp \left( - \frac{2(n - 2k)\epsilon^2}{(k + 1)L_{\text{max}}^2} \right),
$$

(4.28)

$$
\Pr\left( \tilde{L}_{S_k}(Z^n) - L_{X^n,s_k}(x_{k+1}^{n-k}, Z^n) > \epsilon \right) \leq (k + 1) \exp \left( - \frac{2(n - 2k)\epsilon^2}{(k + 1)L_{\text{max}}^2} \right),
$$

(4.29)

where $L_{\text{max}} = \Lambda_{\text{max}} + \ell_{\text{max}}$.

**Remark:** Note that when $k = 0$, this lemma coincides with Lemma 15. The proof of this lemma combines Lemma 15 and the de-interleaving argument in the proof of [7, Theorem 2]. Namely, we de-interleave $Z^n$ into $(k + 1)$ subsequences consisting of symbols separated by blocks of $k$ symbols, and exploit the conditional independence of symbols in each subsequence, given all symbols not in that subsequence, to use Lemma 15.

**Proof of Lemma 16:** See Appendix 4.7.3.

Now, for an integer $0 \leq m \leq \left\lfloor \frac{n-2k}{2} \right\rfloor$ and given $z^n$, let $n(c) \triangleq |T(c)|$, and $m(c) \triangleq \min\{n(c), m\}$ for $c \in C_k$. Then, analogously as in (4.22), we define

$$
\mathcal{S}_{k,m}^n(z^n) = \left\{ S_k(z^n) \in \mathcal{S}_{0}^{n-2k} : \{s_{k,\tau}(c, \cdot)\}_{\tau \in T(c)} \in \mathcal{S}_{0,m(c)}^{n(c)} \text{ for all } c \in C_k \right\}.
$$

(4.30)

\(3\)Again, the value of $\hat{X}^{S_k^n}(z^n)$ for $t \leq k$ and $t > n - k$ can be defined as an arbitrary fixed symbol, since it will be inconsequential in subsequent development.
In words, $S_{k,m}^n(z^n)$ is the set of $(n-2k)$-tuples of $(k$-th order denoiser induced) single-symbol denoisers that allow at most $m(c)$ shifts within the subsequence $\{t : t \in T(c)\}$ for each context $c \in C_k$. Again, for brevity, the dependence on $z^n$ in $S_{k,m}^n(z^n)$ is suppressed, and we write simply $S_{k,m}^n$. It is worth noting that $S_{k,m}^n$ is a larger class than the class of $k$-th order ‘sliding window’ denoisers that are allowed to shift at most $m$ times. The reason is that in $S_{k,m}^n$, the shift within each subsequence associated with each context can occur at any time, regardless of the shifts in other subsequences, whereas in the latter class, the shifts in each subsequence occur together with other shifts in other subsequences.

For integers $k \geq 0$ and $n > 2k$, we now define, for the class of $n$-block denoisers $\hat{X}^{n,S}$ with $S \in S_{k,m}^n$,

$$D_{k,m}(x^n, z^n) = \min_{S \in S_{k,m}^n} L_{\hat{X}^{n,S}}(x_{k+1}^{n-k}, z^n)$$

$$= \min_{S \in S_{k,m}^n} \frac{1}{n-2k} \sum_{t=k+1}^{n-k} \Lambda(x_t, s_{k,t}(c_t, z_t)), \quad (4.31)$$

the minimum normalized cumulative loss of $(x^n, z^n)$ that can be achieved by the sequence of $k$-th order denoisers that allow at most $m$ shifts within each context. Now, to build a legitimate (non genie-aided) universal scheme achieving (4.31) on the basis of $z^n$ only, we define

$$\hat{S}_{k,m} = \arg \min_{S \in S_{k,m}^n} \hat{L}(z^n), \quad (4.32)$$

and the $(k,m)$-S-DUDE, $\hat{X}^{n,k,m}_{\text{univ}}$, is defined as $\hat{X}^{n,\hat{S}_{k,m}}$. Note that when $m = 0$, $\hat{X}^{n,\hat{S}_{k,m}}$ coincides with the DUDE in [7]. The following theorem generalizes Theorem 4 to the case of general $k \geq 0$.

**Theorem 5** Let $\hat{X}^{n,k,m}_{\text{univ}}$ be given by $\hat{X}^{n,\hat{S}_{k,m}}$, where $\hat{S}_{k,m}$ is defined in (4.32). Then,

---

When $m = 0$, $S_{k,0}^n(z^n)$ becomes the set of $n$-block $k$-th order ‘sliding window’ denoisers.
for all $\epsilon > 0$ and $x^n \in \mathcal{X}^n$,
\[
\Pr\left( L_{\mathbf{X}_{\text{uni}^n, k, m}}^n(x^n_{k+n}, Z^n) - D_{k, m}(x^n, Z^n) > \epsilon \right)
\leq 2(k + 1) \exp \left( -\frac{e^2}{2(k + 1) \max L^2} - 2|Z|^{2k} \cdot \left\{ h\left( \frac{m}{n - 2k} \right) + \frac{(m + 1) \ln N}{n - 2k} \right\} \right),
\]
where $h(x) = -x \ln x - (1 - x) \ln(1 - x)$ for $0 \leq x \leq 1$, and $N = |\mathcal{S}| = |Z|^{|x|}$.

Remark: Note that when $k = 0$, this theorem coincides with Theorem 4. Similarly to the way Theorem 4 was plausible given Lemma 15, Theorem 5 can be expected given Lemma 16, since $\hat{S}_{k, m}$ achieves $\min_{S \in S^n_{k, m}} L_S(Z^n)$, and we expect $L_{\mathbf{X}_{\text{uni}^n, \hat{S}_{k, m}}}(x^n_{k+n}, Z^n)$ to be close to $D_{k, m}(x^n, Z^n)$ from the concentration of $\hat{L}_S(Z^n)$ to $L_{\mathbf{X}_{\text{uni}^n, S^n_{k, m}}}(x^n_{k+n}, Z^n)$ for all $S \in S^n_{k, m}$.

Proof of Theorem 5: See Appendix 4.7.4. 

From Theorem 5, we now easily obtain one of the main results of the chapter, which extends Theorem 1 from the case $m = 0$ to the case of general $0 \leq m \leq \left\lfloor \frac{n - 2k}{2} \right\rfloor$. That is, the following theorem asserts that, for every underlying sequence $x \in \mathcal{X}^\infty$, our $(k, m)$-S-DUDE performs essentially as well as the best shifting $k$-th order denoiser that allows at most $m$ shifts within each context, both in high probability and expectation sense, provided a growth condition on $k$ and $m$ is satisfied.

**Theorem 6** Suppose $k = k_n$ and $m = m_n$ are such that the right-hand side of (4.34) is summable in $n$. Then, for all $x \in \mathcal{X}^\infty$, the sequence of denoisers $\{\mathbf{X}_{\text{uni}^n, k_n, m_n}\}$ satisfies

(a)
\[
\lim_{n \to \infty} \left[ L_{\mathbf{X}_{\text{uni}^n, k_n, m_n}}^n(x^n, Z^n) - D_{k_n, m_n}(x^n, Z^n) \right] = 0 \text{ a.s.} \tag{4.35}
\]

(b) For any $\delta > 0$,
\[
E\left[ L_{\mathbf{X}_{\text{uni}^n, k_n, m_n}}^n(x^n, Z^n) - D_{k, m}(x^n, Z^n) \right] = O\left( \sqrt{k_n |Z|^{2k_n} \left( \frac{m_n}{n} \right)^{1-\delta}} \right). \tag{4.36}
\]
Remark: It will be seen in Claim 1 below that the stipulation in the theorem implies
\[ \lim_{n \to \infty} k_n |Z|^{2k_n} \left( \frac{m_n}{n} \right)^{1-\delta} = 0, \]
which, when combined with (4.36), implies that the expected difference on the left hand side of (4.36) vanishes with increasing \( n \). That in itself, however, can easily be deduced from (4.35) and bounded convergence. The more significant value of (4.36) is in providing a rate of convergence result for the 'redundancy' in the S-DUDE's performance, as a function of both \( k \) and \( m \). In particular, note that for any \( \eta > 0 \), \( O(n^{-1/2+\eta}) \) is achievable provided \( k_n = c \log n \) and \( m_n = n^\xi \), for small enough positive constants \( c, \xi \).

In what follows, we specify the maximal growth rates for \( k = k_n \) and \( m = m_n \) under which the summability condition stipulated in Theorem 6 holds.

Claim 1  
\begin{itemize}
  \item[a)] Maximal growth rate for \( k \): The summability condition in Theorem 6 is satisfied provided \( k_n = c_1 \log n \) with \( c_1 < \frac{1}{2 \log |Z|} \) and \( m_n \) grows at any sub-polynomial rate. On the other hand, the condition is not satisfied for \( k_n = c_1 \log n \) with any \( c_1 \geq \frac{1}{2 \log |Z|} \), even when \( m \) is fixed (not growing with \( n \)).
  
  \item[b)] Maximal growth rate for \( m \): The summability condition in Theorem 6 is satisfied for any sub-linear growth rate of \( m_n \), provided \( k_n \) is taken to increase sufficiently slowly that \( k_n |Z|^{2k_n} = o((n/m_n)^{1-\delta}) \) for some \( \delta > 0 \). On the other hand, the condition is not satisfied whenever \( m_n \) grows linearly with \( n \), even when \( k \) is fixed.
\end{itemize}

Proof of Claim 1: See Appendix 4.7.5.■

Proof of Theorem 6: See Appendix 4.7.6.■

4.3.3 A “strong converse”

In Claim 1, we have shown the necessity of \( m = o(n) \) for the condition required in Theorem 6 to hold. However, we can prove the necessity of \( m = o(n) \) in a much stronger sense, described in the following theorem.
Theorem 7 Suppose that $X = \tilde{X}$, that $\Lambda(x, \tilde{x}) \geq 0$ for all $x, \tilde{x}$ with equality if and only if $x = \tilde{x}$, and that $\Pi(x, z) > 0$ for all $x, z$. If $m = \Theta(n)$, then for any sequence of denoisers $\{X^n\}$, there exists $x^\infty \in X^\infty$ such that

$$\limsup_{n \to \infty} E [L_{X^n}(x^n, Z^n) - D_{0,m}(x^n, Z^n)] > 0. \quad (4.37)$$

Remark: The theorem establishes the fact that when $m = o(n)$ does not hold, namely, when $m = \Theta(n)$, not only does the almost sure convergence in Theorem 6 not hold but, in fact, even the much weaker convergence in expectation would fail. Further, it shows that this would be the case for any sequence of denoisers, not necessarily the S-DUDE. Furthermore, (4.37) features $D_{0,m}(x^n, Z^n)$, pertaining to competition with a genie that shifts among single-symbol denoisers so, a fortiori, it implies that for any fixed $k > 0$ or $k$ that grows with $n$,

$$\limsup_{n \to \infty} E [L_{X^n}(x^n, Z^n) - D_{k,m}(x^n, Z^n)] > 0 \quad (4.38)$$

also holds since, by definition, $D_{0,m}(x^n, z^n) \geq D_{k,m}(x^n, z^n)$ for all $x^n, z^n$ and $k \geq 0$. Therefore, the theorem asserts that for any sequence of denoisers to compete with $D_{k,m}(x^n, Z^n)$, even in expectation sense, $m = o(n)$ is necessary. Finally, we mention that the conditions stipulated in the statement of the theorem regarding the loss function and the channel can be considerably relaxed without compromising the validity of the theorem. These conditions are made to allow for the simple proof that we give in Appendix 4.7.7.

### 4.4 Stochastic setting

In [7], the semi-stochastic setting result, [7, Theorem 1], was shown to imply the result for the stochastic setting as well. That is, when the underlying data form a stationary process, [7, Section VI] shows that the DUDE attains optimum distribution-dependent performance. Analogously, we can now use the results from the semi-stochastic setting of the previous section to generalize the results of [7, Section VI] and show that our S-DUDE attains optimum distribution-dependent performance when the underlying
data form a piecewise stationary process. We first define the precise notion of the class of piecewise stationary processes in Subsection 4.4.1, and discuss the richness of this class in Subsection 4.4.2. Subsection 4.4.3 gives the main result of this section: the stochastic setting optimality of the S-DUDE.

4.4.1 Definition of the class of processes \( \mathcal{P}\{m_n\} \)

Let \( P_{\mathcal{X}}(1), \ldots, P_{\mathcal{X}}(M) \) be a finite collection of \( M \) probability distributions of stationary processes, with components taking the values in \( \mathcal{X} \). Let \( A \) be a process with components taking the values in \( \{1, \ldots, M\} \). Then, a piecewise stationary process \( X \) is generated by shifting between the \( M \) processes in a way specified by the "switching process" \( A \), as we now describe.

First, denote \( r(A^n) \) as the number of shifts that have occurred along the \( n \)-tuple \( A^n \), i.e.,

\[
r(A^n) \triangleq \sum_{j=1}^{n-1} 1_{\{A_j \neq A_{j+1}\}}.
\]

Thus, there are \( r(A^n) + 1 \) "blocks" in \( A^n \), where each block is a tuple of constant values that are different from the values of adjacent blocks. Now, for each \( 1 \leq i \leq r(A^n) + 1 \), we define

\[
\tau_i(A^n) \triangleq \begin{cases} 
\inf\{t : \sum_{j=1}^{t} 1_{\{A_j \neq A_{j+1}\}} = i\} & \text{if } 1 \leq i \leq r(A^n) \\
n & \text{if } i = r(A^n) + 1
\end{cases}
\]

as the last time instance of the \( i \)-th block in \( A^n \). In addition, define \( \tau_0(A^n) \triangleq 0 \). Clearly, \( r(A^n) \) and \( \tau_i(A^n) \) depend on \( A^n \) and, thus, are random variables. However, for brevity, we suppress the dependence on \( A^n \) when there is no confusion, and write simply \( r \) and \( \tau_i \), respectively.

Using these definitions, and by denoting \( P_{A^n} \) as the \( n \)-th order marginal distribution of \( A \), we define a piecewise stationary process \( X \) by characterizing its \( n \)-th order
marginal distribution \( P_{X^n} \) as

\[
P_{X^n}(X^n = x^n) = \sum_{a^n} P_{A^n}(a^n) P(X^n = x^n | A^n = a^n) = \sum_{a^n} P_{A^n}(a^n) \prod_{i=1}^{r+1} P_{X}^{(a_{n+1}^{(i)})}(X_{n+1}^{i}), \tag{4.39}
\]

for each \( n \). The corresponding distribution of the process \( X \) is denoted as \( P_X \).\(^5\) In words, \( X \) is constructed by following one of the \( M \) probability distributions in each block, switching from one to another depending on \( A \). Furthermore, conditioned on the realization of \( A \), each stationary block is independent of other blocks, even if the distribution of distinct blocks is the same. This property of conditional independence is reasonable for modeling many types of data arising in practice, since we can think of the \( M \) distributions as different 'modes'; if the process returns to the same mode, it is reasonable to model the new block as a new independent realization of that same distribution. In other words, the 'mode' may represent the kind of 'texture' in a certain region of the data, but two different regions with the same 'texture' should have independent realizations from the texture-generating source. Our notion of a piecewise stationary process almost coincides with that developed in [76]. The main difference is that we allow an arbitrary distribution for the process \( A \).

Now, we define \( \mathcal{P}\{m_n\} \) to be the class of all process distributions that can be constructed as in (4.39) for some \( M \), some collection \( P_{X}^{(1)}, \ldots, P_{X}^{(M)} \) of stationary processes, and some switching process \( A \) whose number of shifts satisfies

\[
r(A^n) \leq m_n \text{ a.s. } \forall n. \tag{4.40}
\]

In words, a process \( X \) belongs to\(^6\) \( \mathcal{P}\{m_n\} \) if and only if it can be formed by switching between a finite collection of independent processes in which the number of switches by time \( n \) does not exceed \( m_n \).

\(^5\)\( \{P_{X^n}\}_{n \geq 1} \) is readily verified to be a consistent family of distributions and, thus, by Kolmogorov's extension theorem, uniquely defines the distribution of the process \( X \).

\(^6\)The phrase "the process \( X \) belongs to \( \mathcal{P}\{m_n\} \)" is shorthand for "the distribution of the process \( X \), \( P_X \), belongs to \( \mathcal{P}\{m_n\} \)."
4.4.2 Richness of $\mathcal{P}\{m_n\}$

In this subsection, we examine how rich the class $\mathcal{P}\{m_n\}$ is, in terms of the growth rate $m_n$ and the existence of denoising schemes that are universal with respect to $\mathcal{P}\{m_n\}$. First, given any distribution on a noiseless $n$-tuple, $P_{X^n}$, we define

$$D(P_{X^n}, \Pi) \triangleq \min_{X^n \in \mathcal{D}_n} E L_{X^n}(X^n, Z^n),$$

(4.41)

where $\mathcal{D}_n$ is the class of all $n$-block denoisers. The expectation on the right-hand side of (4.41) assumes that $X^n$ is generated from $P_{X^n}$ and that $Z^n$ is the output of the DMC, $\Pi$, whose input is $X^n$. Thus, $D(P_{X^n}, \Pi)$ is the optimum denoising performance (in the sense of expected per-symbol loss) attainable when the source distribution $P_{X^n}$ is known.

What happens when the source distribution is unknown? Theorem 3 of [7] established the fact that

$$\lim_{n \to \infty} \left[ E L_{X^n}^{\text{DUDE}}(X^n, Z^n) - D(P_{X^n}, \Pi) \right] = 0 \quad \text{for all stationary } P_X. \quad (4.42)$$

Note that our newly-defined class of processes, $\mathcal{P}\{m_n\}$, is simply the class of all stationary processes if one takes the sequence $m_n$ to be $m_n = 0$ for all $n$. Thus, assuming $m_n = 0$, (4.42) is equivalent to

$$\lim_{n \to \infty} \left[ E L_{X^n}^{\text{DUDE}}(X^n, Z^n) - D(P_{X^n}, \Pi) \right] = 0 \quad \text{for all } P_X \in \mathcal{P}\{m_n\}. \quad (4.43)$$

At the other extreme, when $m_n = n$, $\mathcal{P}\{m_n\}$ consists of all possible (not necessarily stationary) processes. We can observe this equivalence by having $M = |\mathcal{X}|$ processes each be a constant process at a different symbol in $\mathcal{X}$, and creating any process by switching to the appropriate symbol. In this case, not only does (4.43) not hold for the DUDE, but clearly (4.43) cannot hold under any sequence of denoisers. In other words, $\mathcal{P}\{m_n\}$ is far too rich to allow for the existence of schemes that are universal

---

7When $P_X$ is stationary, the limit $\lim_{n \to \infty} D(P_{X^n}, \Pi) \triangleq D(P_X, \Pi)$ was shown to exist in [7]. Thus, (4.42) was equivalently stated as $\lim_{n \to \infty} E L_{X^n}^{\text{DUDE}} = D(P_X, \Pi)$ in [7, Theorem 3].
with respect to it.

It is obvious then that \( \mathcal{P}\{m_n\} \) is significantly richer than the family of stationary processes whenever \( m_n \) grows with \( n \). It is of interest then to identify the maximal growth rate of \( m_n \) that allows for the existence of schemes that are universal with respect to \( \mathcal{P}\{m_n\} \), and to find such a universal scheme. In what follows, we offer a complete answer to these questions. Specifically, we show that if the growth rate of \( m_n \) allows for the existence of any scheme which is universal with respect to \( \mathcal{P}\{m_n\} \), the S-DUDE is universal, too.

### 4.4.3 Universality of S-DUDE

Here, we state our stochastic setting result, which establishes the universality of \((k, m)\)-S-DUDE with respect to the class \( \mathcal{P}\{m_n\} \).

**Theorem 8** Let \( k = k_n \) and \( m = m_n \) satisfy the growth rate condition stipulated in Theorem 6, in addition to \( \lim_{n \to \infty} k_n = \infty \). Then, the sequence of denoisers \( \{\hat{X}_{n, m_n}^{k_n, m_n}\} \) defined in Section 4.3 satisfy

\[
\lim_{n \to \infty} \mathbb{E}L_{X_{n, m_n}^{k_n, m_n}}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \Pi) = 0 \quad \text{for all} \quad P_X \in \mathcal{P}\{m_n\}. \tag{4.44}
\]

**Remark 1:** Recall that, as noted in Claim 1, \( m_n = o(n) \) together with appropriately slowly growing \( k = k_n \) is sufficient to guarantee the growth rate condition stipulated in Theorem 6. Hence, by Theorem 8, \( m = o(n) \) and the sufficiently slowly growing \( k = k_n \) suffices for (4.44) to hold. Therefore, Theorem 8 implies the existence of schemes that are universal with respect to \( \mathcal{P}\{m_n\} \) whenever \( m_n \) increases sublinearly in \( n \). Since, as discussed in Subsection 4.4.2, no universal scheme exists for \( \mathcal{P}\{m_n\} \) when \( m_n \) is linear in \( n \), we conclude that the sub-linearity of \( m_n \) is the necessary and sufficient condition for a universal scheme to exist with respect to \( \mathcal{P}\{m_n\} \). Moreover, Theorem 8 establishes the strong sense of optimality of the S-DUDE, as it shows that whenever \( \mathcal{P}\{m_n\} \) is universally "competable", the S-DUDE does the job. This fact is somewhat analogous to the situation in [76], where the optimality of the universal lossless coding scheme presented therein for piecewise stationary sources
was established under the condition that \( m = o(n) \).

Remark 2: A pointwise result

\[
\lim_{n \to \infty} \left[ L_{X \| Z}^n(X_n, Z^n) - D(P_{X^n}, \Pi) \right] = 0 \quad \text{a.s.}
\]

for all \( P_X \in \mathcal{P}(m_n) \), which is analogous to [7, Theorem 4], can also be derived. However, we omit such a result here since the details required for stating it rigorously would be convoluted, and its added value over the strong point-wise result we have already established in the semi-stochastic setting would be little.

Proof of Theorem 8: See Appendix 4.7.8. 

4.5 Algorithm and complexity

4.5.1 An efficient implementation of S-DUDE

In the preceding two sections, we gave strong asymptotic performance guarantees for the new class of schemes, the S-DUDE. However, the question regarding the practical implementation of (4.32), i.e., obtaining

\[
\hat{S}_{k,m} = \arg \min_{S \in \mathcal{S}_{k,m}^n} \tilde{L}_S(z^n),
\]

for fixed \( k, m \) and \( n \) remains and, at first glance, may seem to be a difficult combinatorial optimization problem. In this section, we devise an efficient two-pass algorithm, which yields (4.32) and performs denoising with linear complexity in the sequence length \( n \). A recursion similar to that in the first pass of the algorithm we present appears also in the study of tracking the best expert in on-line learning [63, 64].

From the definition of \( S_{k,m}^n \), (4.30), we can see that obtaining (4.32) is equivalent to obtaining the best combination of single-symbol denoisers with at most \( m(c) \) shifts that minimizes the cumulative estimated loss along \( \{ t : t \in T(c) \} \), for each \( c \in C_k \). Thus, our problem breaks down to \( |C_k| \) independent problems, each being a problem of competing with the best combination of single-symbol schemes allowing \( m \) switches.

To describe an algorithm that implements this parallelization efficiently, we first
CHAPTER 4. DISCRETE DENOISING WITH Shifts

define variables. For \( (k,m)\)-S-DUDE, let \( I = m + 1, J = N + 1 \), where \( N = |S| = |Z|^k \). Then, a matrix \( M_t \in \mathbb{R}^{I \times J} \) is defined for \( k + 1 \leq t \leq n - 2k \), where \( M_t(i,j) \) for \( 1 \leq i \leq I \) and \( 1 \leq j \leq J - 1 \) represents the minimum (un-normalized) cumulative estimated loss of the sequence of single-symbol denoisers along the time index \( \{ \tau : \tau \leq t, c_\tau = c_t \} \), allowing at most \((i - 1)\) shifts between single-symbol denoisers and applying \( s_t = j \). Moreover, \( M_t(i,J) \), for \( 1 \leq i \leq I \), is the symbol-by-symbol denoiser that attains the minimum value of the \( i \)-th row of \( M_t \), i.e., \( \arg \min_{1 \leq j \leq J - 1} M_t(i,j) \).

A time pointer \( T \in \mathbb{R}^D \), where \( D = |C_k| = |Z|^{2k} \), is defined to store the closest time index that has the same context as current time, during the first and second pass. That is,

\[
T(c_t) = \begin{cases} 
\max\{ \tau : \tau < t, c_\tau = c_t \}, & \text{when first pass} \\
\min\{ \tau : \tau > t, c_\tau = c_t \}, & \text{when second pass}
\end{cases}
\]  

We also define \( r \in \mathbb{R}^D \) and \( q \in \mathbb{R}^D \) as variables for storing the pointer enabling our scheme to follow the best combination of single-symbol denoisers during the second pass. Thus, the total memory size required is \( O(mNn + |Z|^{2k}) = O(mn) \) (assuming that \( k \) satisfies the growth rate stipulated in the previous sections, which implies \( |Z|^{2k} = o(n) \)).

Our two-pass algorithm has ingredients from both the DUDE and from the forward-backward recursions of hidden Markov models [14] and, in fact, the algorithm becomes equivalent to DUDE when \( m = 0 \). The first pass of the algorithm runs forward from \( t = k + 1 \) to \( t = n - k \), and updates the elements of \( M_t \) recursively. The recursions have a natural dynamic programming structure. For \( 2 \leq i \leq I, 1 \leq j \leq J - 1 \), \( M_t(i,j) \) is determined by

\[
M_t(i,j) = \ell(z_t, j) + \min \left\{ M_{T(c_t)}(i,j), M_{T(c_t)}(i-1, M_{T(c_t)}(i-1,j)) \right\},
\]

that is, adding the current loss to the best cumulative loss up to \( T(c_t) \) along \( \{ \tau : \tau < t, c_\tau = c_t \} \). When \( i = 1 \), the second term in the minimum of (4.46) is not defined, and \( M_t(i,j) \) just becomes \( \ell(z_t, j) + M_{T(c_t)}(i,j) \). The validity of (4.46) can be verified by observing that there are two possible cases in achieving \( M_t(i,j) \): either
the \((i-1)\)-th shift to the single-symbol denoiser \(j\) occurred before \(t\), or it occurred at time \(t\). We can see that the first term in the minimum of (4.46) corresponds to the former case; the second term corresponds to the latter. Obviously, the minimum of these two (where ties may be resolved arbitrarily), leads to the value of \(M_t(i,j)\) as in (4.46). After updating all \(M_t\)'s during the first pass, the second pass runs backwards from \(t = n - k\) to \(t = k + 1\), and extracts \(S_{k,m}\) from \(\{M_t\}_{t=k+1}^{n-2k}\) by following the best shifting between single-symbol denoisers. The actual denoising (i.e., assembling the reconstruction sequence \(\tilde{X}^n\)) is also performed in that pass. The pointers \(r(c_t)\) and \(q(c_t)\) are updated recursively, and they track the best shifting point and combination of single-symbol denoisers, respectively, for each of the subsequences associated with the various contexts. A succinct description of the algorithm is provided in Algorithm 1. The time complexity of the algorithm is readily seen to be \(O(mn)\) as well.
Algorithm 1 The \((k, m)\)-Shifting Discrete Denoising Algorithm

Require: \(M_t(i, j) \in \mathbb{R}^{I \times J}, \quad k + 1 \leq t \leq n - 2k, 1 \leq i \leq I, 1 \leq j \leq J, \quad T \in \mathbb{R}^D, r \in \mathbb{R}^D, q \in \mathbb{R}^D, L \in \mathbb{R}\)

Ensure: \(\hat{S}_k = \{s_{k,t}(c_t, \cdot)\}_{t=k+1}^{n-k}\) in (4.32) and the denoised output \(\{\hat{x}_t\}_{t=k+1}^{n-k}\)

\(\tau(c) \leftarrow \phi\) for all \(c \in C_k\)

for \(t = k + 1\) to \(n - 2k\) do
  if \(T(c_t) = \phi\) then
    \(M_t(i, j) \leftarrow \ell(z_t, j)\) for \(1 \leq i \leq I, \quad 1 \leq j \leq J - 1\)
    \(M_t(i, J) \leftarrow \arg \min_{1 \leq j \leq J - 1} M_t(i, j)\) for \(1 \leq i \leq I\)
  else
    for \(1 \leq j \leq J - 1\)
    \(M_{T(c_t)}(i, j) \leftarrow \begin{cases} M_{T(c_t)}(i, j), & i = 1 \\ \min \{M_{T(c_t)}(i, f), M_{T(c_t)}(i - 1, M_{T(c_t)}(i - 1, J))\}, & i \neq 1 \end{cases}\)
    \(M_t(i, j) \leftarrow M_{T(c_t)}(i, j) + \ell(z_t, j)\) for \(1 \leq i \leq I, 1 \leq j \leq J - 1\)
    for \(j = J\)
    \(M_t(i, J) \leftarrow \arg \min_{1 \leq j \leq J - 1} M_t(i, j)\) for \(1 \leq i \leq I\)
  end if
  \(T(c_t) \leftarrow t\)
end for

\(T(c) \leftarrow \phi\) for all \(c \in C_k\)

for \(t = n - 2k\) to \(k + 1\) do
  if \(T(c_t) = \phi\) then
    \(r(c_t) \leftarrow I, \quad q(c_t) \leftarrow M_t(r(c_t), J)\)
  else
    \(L \leftarrow M_{T(c_t)}(r(c_t), q(c_t)) - \ell(z_t, q(c_t))\)
    if \(L < M_t(r(c_t), q(c_t))\) then
      \(r(c_t) \leftarrow r(c_t) - 1, \quad q(c_t) \leftarrow M_t(r(c_t), J)\)
    end if
  end if
  \(T(c_t) \leftarrow t, \quad s_{k,t}(c_t, \cdot) \leftarrow q(c_t)\)
  \(\hat{x}_t \leftarrow s_{k,t}(c_t, z_t)\)
end for
4.5.2 Extending the S-DUDE to multi-dimensional data

As noted, our algorithm is essentially separately employing the same algorithm to compete with the best shifting single-symbol denoisers, on each subsequence associated with each context. The overall algorithm is the result of parallelizing the operations of the schemes for the different subsequences, which allows for a more efficient implementation than if these schemes were to be run completely independently of one another. This characteristic of running the same algorithm in parallel along each subsequence enables us to extend S-DUDE to the case of multi-dimensional data: run the same algorithm along each subsequence associated with each (this time multi-dimensional) context. It should be noted, however, that the extension of the S-DUDE to the multidimensional case is not as straightforward as the extension of the DUDE was, since, whereas the DUDE's output is independent of the ordering of the data within each context, this ordering may be very significant in its effect on the output and, hence, the performance of S-DUDE. Therefore, the choice of a scheme for scanning the data and capturing its local spatial stationarity, e.g., Peano-Hilbert scanning [77], is an important ingredient in extending S-DUDE to the denoising of multi-dimensional data. Findings from the recent study on universal scanning reported in [78, 79] can be brought to bear on such an extension.

4.6 Experimentation

In this section, we report some preliminary experimental results obtained by applying S-DUDE to several kinds of noise-corrupted data.

4.6.1 Image denoising

In this subsection, we report some experimental results of denoising a binary image under the Hamming loss function. The first and most simplistic experiment is with the 400 x 400 black-and-white binary image shown in Figure 1. The first figure is the clean underlying image. The image is passed through a binary symmetric channel (BSC) with crossover probability $\delta = 0.1$, to obtain the noisy image (second image in
Figure 4.1). Note that in this case, there are only four symbol-by-symbol denoisers, namely, \( S = \{0, 1, z, \bar{z}\} \), representing always-say-0, always-say-1, say-what-you-see, and flip-what-you-see, respectively. The third image in Figure 4.1 is the DUDE output with \( k = 0 \), and the last image is the output of our S-DUDE with \( k = 0, m = 1 \). The DUDE with \( k = 0 \) is competes with the best time-invariant symbol-by-symbol denoiser which, in this case, is the say-what-you-see denoiser, since the empirical distribution of the clean image is \((0.5, 0.5)\) and \( \delta = 0.1 \). Thus, the DUDE output is the same as the noisy image; hence, no denoising is performed. However, it is clear that, for this image, the best compound action of the symbol-by-symbol denoisers is always-say-0 for the first half and then a shift to always-say-1 for the remainder. We can see that our \((0,1)\)-S-DUDE successfully captures this shift from the noisy observations, and results in perfect denoising with zero bit errors.

Now, we move on to a more realistic example. The first image in Figure 4.2, a concatenation of a half-toned Einstein image \((300 \times 300)\) and scanned Shannon’s 1948 paper \((300 \times 300)\), is the clean image. We pass the image through a binary symmetric channel (BSC) with crossover probability \( \delta = 0.1 \), to obtain the second noisy image, which we raster scan and employ the S-DUDE on the resulting one-dimensional sequence. Since the two concatenated images are of a very different nature, we expect our S-DUDE to perform better than the DUDE, because it is designed to adapt to the possibility of employing different schemes in different regions of the data. The plot shows the performance of our \((k,m)\)-S-DUDE with various values of \( k \) and \( m \). The horizontal axis reflects \( k \), and the vertical axis represents the ratio of bit error per symbol (BER) to \( \delta = 0.1 \). Each curve represents the BER of schemes with different \( m = 0, 1, 2, 3 \). Note that \( m = 0 \) corresponds to the DUDE. We
can see that S-DUDE with $m > 0$ mostly dominates the DUDE, with an additional BER reduction of $\sim 11\%$, including when $k = 6$, the best $k$ value for the DUDE. The bottom three figures show the denoised images with $(k,m) = (4,0), (4,2), (6,1)$, achieving BERs of $\delta \times (0.744, 0.6630, 0.4991)$, respectively. Thus, in this example, (4, 2)-S-DUDE achieves an additional BER reduction of 11% over the DUDE with $k = 4$, and the overall best performance is achieved by (6, 1)-S-DUDE. Given the nature of the image, which is a concatenation of two completely different types of images, each reasonably uniform in texture, it is not surprising to find that the S-DUDE with $m = 1$ performs the best.

Figure 4.2: Clean and noisy images, the bit error rate plot for $(k,m)$-S-DUDE, and three denoised outputs for $(k,m) = (4,0), (4,2), (6,1)$, respectively.
4.6.2 State estimation for a switching binary hidden Markov process

Here, we give a stochastic setting experiment. A switching binary hidden Markov process in this example is defined as a binary symmetric Markov chain observed through a BSC, where the transition probabilities of the Markov chain switches over time. The goal of a denoiser here is to estimate the underlying Markov chain based on the noisy output.

In our example, we construct a simple switching binary hidden Markov process of length \( n = 10^6 \), in which the transition probability of the underlying binary symmetric Markov source switches from \( p = 0.01 \) to \( p = 0.2 \) at the midpoint of the sequence, and the crossover probability of BSC is \( \delta = 0.1 \). Then, we estimate the state of the underlying Markov chain based on the BSC output. The goodness of the estimation is again measured by the Hamming loss, i.e., the fraction of errors made. Slightly better than the optimal Bayesian distribution-dependent performance for this case can be obtained by employing the forward-backward recursion scheme, incorporating the varying transition probabilities with the help of a genie that knows the exact location of the change in the process distribution. Figure 4.3 plots the BER of \((k,m)\)-S-DUDE with various \( k \) and \( m \), compared to the genie-aided Bayes optimal BER. The horizontal axis represents \( k \), and the two curves refer to \( m = 0 \) (DUDE) and \( m = 1 \). The vertical axis is the ratio of BER to \( \delta = 0.1 \).

We can observe that the optimal Bayesian BER is (lower bounded by) \( 0.4865 \times \delta \). The best performance of the DUDE was achieved when \( k = 6 \) with a BER of \( 0.5738 \times \delta \), which is far above (18% more than) the optimal BER. It is clear that, despite the size of the data, the DUDE fails to converge to the optimum, as it is confined to be employing the same sliding-window scheme throughout the whole data. However, we can see that the \((4,1)\)-S-DUDE achieves a BER of \( 0.4979 \times \delta \), which is within 2.3% of the optimal BER. This example shows that our S-DUDE is competent in attaining the optimum performance for a class richer than that of the stationary processes. Specifically, it attains the optimum performance for piecewise stationary processes, on which the DUDE generally fails.
CHAPTER 4. DISCRETE DENOISING WITH-shifts

Figure 4.3: BER for switching binary hidden Markov process ($\delta = 0.1, n = 10^6$). The switch of the underlying binary Markov chain occurs when $t = 5 \times 10^5$, from the transition probability $p = 0.01$ to $p = 0.2$.

4.7 Appendix

4.7.1 Proof of Lemma 15

We first establish the fact that for all $x^n \in \mathcal{X}^n$, and for fixed $S \in \mathcal{S}^n$,

$$\left\{ n[L_{X^n,S}(x^n,Z^n) - \tilde{L}_S(Z^n)] \right\}_{n \geq 1}$$

is a $\{Z^n\}$-martingale. This is not hard to see by following:

$$E\left( n[L_{X^n,S}(x^n,Z^n) - \tilde{L}_S(Z^n)] | Z^{n-1} \right)$$

$$= E\left( \sum_{t=1}^{n} \Lambda(x_t, s_t(Z_t)) - \sum_{t=1}^{n} \ell(Z_t, s_t) | Z^{n-1} \right)$$

$$= (n - 1)[L_{X^{n-1},S}(x^{n-1},Z^{n-1}) - \tilde{L}_S(Z^{n-1})] + E\left( \Lambda(x_n, s_n(Z_n)) - \ell(Z_n, s_n) | Z^{n-1} \right)$$

$$= (n - 1)[L_{X^{n-1},S}(x^{n-1},Z^{n-1}) - \tilde{L}_S(Z^{n-1})]$$

(4.47)
where (4.47) follows from the fact that $Z_n$ is independent of $Z^{n-1}$, and $EA(x_n, s_n(Z_n)) = E\ell(Z_n, s_n)$. Therefore, $L_S(x^n, Z^n) - \tilde{L}_S(Z^n)$ is a normalized sum of bounded martingale differences; therefore the inequalities (4.20) and (4.21) follow directly from the Hoeffding-Azuma inequality [6, Lemma A.7].

4.7.2 Proof of Theorem 4

Consider following chain of inequalities:

\[
P(L_{\hat{X}_n,S}(x^n, Z^n) - D_{0,m}(x^n, Z^n) > \epsilon) = P\left(\max_{s \in S_{0,m}^n} \left\{ L_{\hat{X}_n,S}(x^n, Z^n) - L_{\hat{X}_n,S}(x^n, Z^n) \right\} > \epsilon\right) \\
\leq \sum_{s \in S_{0,m}^n} P\left(\left| L_{\hat{X}_n,S}(x^n, Z^n) - L_{\hat{X}_n,S}(x^n, Z^n) \right| > \epsilon\right) \\
\leq \sum_{s \in S_{0,m}^n} P\left(\left| L_{\hat{X}_n,S}(x^n, Z^n) - \tilde{L}_S(Z^n) \right| > \epsilon/2\right) + \sum_{s \in S_{0,m}^n} P\left(\left| \tilde{L}_S(Z^n) - L_{\hat{X}_n,S}(x^n, Z^n) \right| > \epsilon/2\right)
\]

(4.48)

where (4.48) follows from the union bound, and (4.49) follows from adding and subtracting $\tilde{L}_S(Z^n)$, and the union bound. For term (i) in (4.49),

\[
(i) \leq \sum_{s \in S_{0,m}^n} P\left(\max_{s \in S_{0,m}^n} \left\{ L_{\hat{X}_n,S}(x^n, Z^n) - \tilde{L}_S(Z^n) \right\} > \epsilon/2\right)
\]

(4.50)

\[
\leq \sum_{s \in S_{0,m}^n} \sum_{s \in S_{0,m}^n} \exp\left(-\frac{n\epsilon^2}{2L_{\max}^2}\right),
\]

(4.51)

where (4.50) follows from $L_{\hat{X}_n,S}(x^n, Z^n) - \tilde{L}_S(Z^n) \leq \max_{s \in S_{0,m}^n} \left\{ L_{\hat{X}_n,S}(x^n, Z^n) - \tilde{L}_S(Z^n) \right\}$, and (4.51) follows from the union bound and (4.20). Similarly, for term (ii)
in (4.49),

\[
(ii) \leq \sum_{s \in \mathcal{S}_{0,m}^n} P \left( \hat{L}_S(Z^n) - L_{X^n,s}(x^n, Z^n) > \epsilon/2 \right) \\
\leq \sum_{s \in \mathcal{S}_{0,m}^n} \exp \left( -n \frac{\epsilon^2}{2L_{\max}^2} \right),
\]

where (4.52) follows from \( \hat{L}_S(Z^n) \leq \hat{L}_S(Z^n) \) a.s., and (4.53) follows from (4.21). Therefore, continuing (4.49), we obtain

\[
(4.49) \leq 2 \sum_{s \in \mathcal{S}_{0,m}^n} \sum_{s \in \mathcal{S}_{0,m}^n} \exp \left( -n \frac{\epsilon^2}{2L_{\max}^2} \right) \\
= 2 \left( \sum_{k=0}^{m} \binom{n-1}{k} N(N-1)^k \right)^2 \exp \left( -n \frac{\epsilon^2}{2L_{\max}^2} \right) \\
\leq 2 \exp \left( -n \left[ \frac{\epsilon^2}{2L_{\max}^2} - 2h \left( \frac{m}{n} \right) - \frac{2(m+1) \ln N}{n} \right] \right),
\]

where (4.54) follows from \( |\mathcal{S}_{0,m}^n| = \sum_{k=0}^{m} \binom{n-1}{k} N(N-1)^k \), and (4.55) follows from \( |\mathcal{S}_{0,m}^n| \leq N^{m+1} \exp \left( n h \left( \frac{m}{n} \right) \right) \). Hence, the theorem is proved. 

\[\boxed{}\]

### 4.7.3 Proof of Lemma 16

We will prove (4.28) since the proof of (4.29) is essentially identical. As in [7], define

\[ I_d \triangleq \{ t : k + 1 \leq t \leq n - k, t \equiv d \mod (k + 1) \}, \]
CHAPTER 4. DISCRETE DENOISING WITH SHIFTS

whose cardinality is denoted \( n_d = [(n - d - k)/(k + 1)] \). Then, by denoting \( C_t = (Z_{t-k}^{t-1}, Z_{t+k}^{t+1}) \), we start the chain of inequalities,

\[
\Pr\left( L_{x_{n-k}^{n+1}}(x_{n-k+1}^n, Z^n) - \tilde{L}_{S_k}(Z^n) > \epsilon \right) \\
\leq \Pr\left( \sum_{d=0}^{k} \sum_{\tau \in \mathcal{I}_d} \left\{ \Lambda\left(x_{\tau}, s_{k,\tau}(C_{\tau}, Z_{\tau})\right) - \ell\left(Z_{\tau}, s_{k,\tau}(C_{\tau}, \cdot)\right) \right\} > (n - 2k)\epsilon \right) \\
\leq \sum_{d=0}^{k} \Pr\left( \sum_{\tau \in \mathcal{I}_d} \left\{ \Lambda\left(x_{\tau}, s_{k,\tau}(C_{\tau}, Z_{\tau})\right) - \ell\left(Z_{\tau}, s_{k,\tau}(C_{\tau}, \cdot)\right) \right\} > (n - 2k)\gamma_d\epsilon \right),
\]

where (4.56) follows from the triangle inequality, (4.57) follows from the union bound, and \( \{\gamma_d\} \) is a set of nonnegative constants (to be specified later) satisfying \( \sum_d \gamma_d = 1 \).

In the sequel, for simplicity, we will denote \( \Lambda\left(x_{\tau}, s_{k,\tau}(C_{\tau}, Z_{\tau})\right) \) and \( \ell\left(Z_{\tau}, s_{k,\tau}(C_{\tau}, \cdot)\right) \) in (4.48) as \( \Lambda_{\tau} \) and \( \ell_{\tau} \), respectively. Now, the collection of random variables \( Z(d) \) is defined to be

\[
Z(d) \triangleq \{Z_t : 1 \leq t \leq n, t \notin \mathcal{I}_d\},
\]

and \( z(d) \in Z^{n-n_d} \) denotes a particular realization of \( Z(d) \). Then, by conditioning, we have

\[
(4.57) \\
\leq \sum_{d=0}^{k} \sum_{z(d) \in Z^{n-n_d}} \Pr(Z(d) = z(d)) \Pr\left( \sum_{\tau \in \mathcal{I}_d} \left\{ \Lambda_{\tau} - \ell_{\tau} \right\} > (n - 2k)\gamma_d\epsilon \right | Z(d) = z(d),
\]

and let \( P_d \) denote the conditional probability of (4.58). Now, conditioned on \( Z(d) = z(d), \{Z_{\tau}\}_{\tau \in \mathcal{I}_d} \) are all independent, and the summation in \( P_d \) becomes

\[
\sum_{\tau \in \mathcal{I}_d} \left\{ \Lambda\left(x_{\tau}, s_{k,\tau}(c_{\tau}, Z_{\tau})\right) - \ell\left(Z_{\tau}, s_{k,\tau}(c_{\tau}, \cdot)\right) \right\},
\]

which is the sum of the absolute differences of the true and estimated losses of the symbol-by-symbol denoisers \( s_{k,\tau}(c_{\tau}, \cdot) \) over \( \tau \in \mathcal{I}_d \). Thus, we can apply (4.20), and
obtain

\[
P_d = \Pr \left( \sum_{\tau \in \ell_d} \{ \Lambda_\tau - \ell_\tau \} > n_d \cdot \frac{(n - 2k)\gamma_d \epsilon}{n_d} \mid Z(d) = z(d) \right)
\leq \exp \left( - \frac{2(n - 2k)^2 \gamma_d^2 \epsilon^2}{L_{\text{max}}^2 n_d} \right).
\] (4.59)

Following [7], we choose \( \gamma_d = \frac{\sqrt{n_d}}{\sum_j \sqrt{n_j}} \), and from the Cauchy-Schwartz inequality and \( \sum_d n_d = n - 2k \), we arrive at

\[
\frac{n_d}{\gamma_d^2} \leq (k + 1) \sum_{d=0}^k n_d = (k + 1)(n - 2k),
\]

and, hence,

\[
P_d \leq \exp \left( - \frac{2(n - 2k)\epsilon^2}{(k + 1)L_{\text{max}}^2} \right).
\] (4.60)

Therefore, plugging (4.60) into (4.58), we finally have

\[
(4.58) \leq (k + 1) \exp \left( - \frac{2(n - 2k)\epsilon^2}{(k + 1)L_{\text{max}}^2} \right),
\]

which proves the lemma. \( \blacksquare \)
4.7.4 Proof of Theorem 5

The proof resembles that of Theorem 4. Consider

\[
\Pr\left( L_{X^n, S_{k,m}}(x_{k+1}^{n-k}, Z^n) - D_{k,m}(x^n, Z^n) > \epsilon \right)
\]

\[
= P\left( \max_{S \in S_{k,m}^n} \left\{ L_{X^n, S_{k,m}}(x_{k+1}^{n-k}, Z^n) - L_{X^n, S}(x_{k+1}^{n-k}, Z^n) \right\} > \epsilon \right)
\]

\[
\leq \sum_{S \in S_{k,m}^n} P\left( L_{X^n, S_{k,m}}(x_{k+1}^{n-k}, Z^n) - L_{X^n, S}(x_{k+1}^{n-k}, Z^n) > \epsilon \right)
\]

\[
\leq \sum_{S \in S_{k,m}^n} \left\{ P\left( L_{X^n, S_{k,m}}(x_{k+1}^{n-k}, Z^n) - L_{S_{k,m}}(Z^n) > \frac{\epsilon}{2} \right) \right\} + P\left( L_{S_{k,m}}(Z^n) - L_{X^n, S}(x_{k+1}^{n-k}, Z^n) > \frac{\epsilon}{2} \right) \tag{4.61}
\]

\[
\leq 2(k + 1) \sum_{S \in S_{k,m}^n} \exp \left( -\frac{(n - 2k)\epsilon^2}{2(k + 1)L_{\max}^2} \right) \tag{4.62}
\]

\[
= 2(k + 1) \sum_{k=0}^{m(c)} \binom{n(c) - 1}{k} N(N - 1)^k \sum_{c \in C_k} \left| C_k \right| \exp \left( -\frac{(n - 2k)\epsilon^2}{2(k + 1)L_{\max}^2} \right) \tag{4.63}
\]

where (4.61)-(4.62) follow similarly as in (4.48)-(4.49); (4.63) follows from arguments similar to (4.50), (4.52), and Lemma 16 (which plays the role that Lemma 15 played there); and (4.64) follows from \( |S_{k,m}^n| = \left[ \sum_{k=0}^{m(c)} \binom{n(c) - 1}{k} N(N - 1)^k \right] |C_k| \). Now, for all \( c \in C_k \),

\[
\sum_{k=0}^{m(c)} \binom{n(c) - 1}{k} N(N - 1)^k \leq N^{m+1} \exp \left( n(c)h \left( \frac{m(c)}{n(c)} \right) \right)
\]

\[
\leq N^{m+1} \exp \left( (n - 2k)h \left( \frac{m(c)}{n - 2k} \right) \right) \tag{4.65}
\]

\[
\leq N^{m+1} \exp \left( (n - 2k)h \left( \frac{m}{n - 2k} \right) \right) \tag{4.66}
\]
where (4.65) is based on the fact that \( \exp(nh(m/n)) \) is an increasing function in \( n \), and (4.66) follows from \( m \leq \lfloor \frac{n-2k}{2} \rfloor \). Therefore, together with \( |C_k| = |Z|^{2k} \), we have

\[
(4.64) \leq 2(k + 1) \exp \left( - (n - 2k) \cdot \left( \frac{\epsilon^2}{2(k + 1)L_{\text{max}}^2} - 2|Z|^{2k} \cdot \left\{ h\left( \frac{m}{n - 2k} \right) + \frac{(m + 1) \ln N}{n - 2k} \right\} \right) \right),
\]

which proves the theorem. \( \blacksquare \)

### 4.7.5 Proof of Claim 1

For part a), to show the necessity first, suppose \( c_1 \geq \frac{1}{2 \log |Z|} \). Then, from \( |Z|^{2k} = n^{2k \log |Z|} \), we have \( 2|Z|^{2k} \cdot \left\{ h\left( \frac{m}{n - 2k} \right) + \frac{(m + 1) \ln N}{n - 2k} \right\} = \Omega \left( n^{2k \log |Z|} \right) \), which will grow to infinity as \( n \) grows, even when \( m \) is fixed. Therefore, the right-hand side of (4.34) is not summable. On the other hand, \( k = c_1 \log n \) with \( c_1 < \frac{1}{2 \log |Z|} \) is readily verified to suffice for the summability, provided that \( m = m_n \) grows at any sub-polynomial rate, i.e., grows more slowly than \( n^{\alpha} \) for any \( \alpha > 0 \) (e.g., \( c_2 \log n \)).

For part b), to show the necessity, suppose \( m = \Theta(n) \). Then, \( h\left( \frac{m}{n - 2k} \right) + \frac{(m + 1) \ln N}{n - 2k} = \Omega(1) \), and, thus, for sufficiently small \( \epsilon \), \( \frac{\epsilon^2}{2(k + 1)L_{\text{max}}^2} - 2|Z|^{2k} \cdot \left\{ h\left( \frac{m}{n - 2k} \right) + \frac{(m + 1) \ln N}{n - 2k} \right\} < 0 \) even for \( k \) fixed. Therefore, the right-hand side of (4.34) is not summable. Hence, \( m = o(n) \) is necessary for the summability. For sufficiency, suppose \( m = m_n \) is any rate, such that \( \lim_{n \to \infty} \frac{m_n}{n} = 0 \). Then,

\[
\frac{\epsilon^2}{2(k + 1)L_{\text{max}}^2} - 2|Z|^{2k} \cdot \left\{ h\left( \frac{m}{n - 2k} \right) + \frac{(m + 1) \ln N}{n - 2k} \right\} = \frac{1}{k} \left( \frac{\epsilon^2}{2(1 + \frac{1}{k} L_{\text{max}}^2)} - 2k|Z|^{2k} \cdot O\left( \left( \frac{m_n}{n} \right)^{1-\delta} \right) \right). \tag{4.68}
\]

Thus, if \( k \) grows sufficiently slowly that \( k|Z|^{2k} = o\left( \left( \frac{m}{m_n} \right)^{1-\delta} \right) \), then (4.68) becomes positive for sufficiently large \( n \), and the right-hand side of (4.34) becomes summable. \( \blacksquare \)
4.7.6 Proof of Theorem 6

First, denote the random variable $A_{k,m}^n = L_{X_{n,k,m}}(x_{n,k+1}^n, Z^n) - D_{k,m}(x^n, Z^n)$. Then, for part a), we have

$$L_{X_{n,k,m}}(x^n, Z^n) - D_{k,m}(x^n, Z^n) \leq \frac{2k\Lambda_{\text{max}}}{n} + A_{k,m}^n \ a.s.$$

Since the maximal rate for $k$ is $c_1 \log n$ as specified in Claim 1, $\lim_{n \to \infty} \frac{2k\Lambda_{\text{max}}}{n} = 0$.

Furthermore, from the summability condition on $k$ and $m$, Theorem 5, and the Borel-Cantelli lemma, we get $\lim_{n \to \infty} A_{k,m}^n = 0$ with probability 1, which proves part a).

To prove part b), note that, for any $\epsilon > 0$,

$$E[L_{X_{n,k,m}}(x^n, Z^n) - D_{k,m}(x^n, Z^n)]$$

$$\leq \frac{2k\Lambda_{\text{max}}}{n} + E(A_{k,m}^n)$$

$$= \frac{2k\Lambda_{\text{max}}}{n} + E(A_{k,m}^n | A_{k,m}^n \leq \epsilon)Pr(A_{k,m}^n \leq \epsilon) + E(A_{k,m}^n | A_{k,m}^n > \epsilon)Pr(A_{k,m}^n > \epsilon)$$

$$\leq \frac{2k\Lambda_{\text{max}}}{n} + \epsilon + \Lambda_{\text{max}} \cdot Pr(A_{k,m}^n > \epsilon)$$

$$\leq \frac{2k\Lambda_{\text{max}}}{n} + \epsilon + \Lambda_{\text{max}} \cdot (\text{right-hand side of (4.34)}). \quad (4.69)$$

From the proof of Claim 1, the condition of Theorem 6 requires $k = k_n$ and $m = m_n$ to satisfy

$$\lim_{n \to \infty} k_n |Z|^{2k_n} \left(\frac{m_n}{n}\right)^{1-\delta} = 0.$$ 

Therefore, if we set $\epsilon^2 = \Theta(k_n |Z|^{2k_n} \left(\frac{m_n}{n}\right)^{1-\delta})$ with sufficiently large constant then, from (4.68), we can see that the right-hand side of (4.34) will decay almost exponentially, which is much faster than $\Theta(k_n |Z|^{2k_n} \left(\frac{m_n}{n}\right)^{1-\delta})$. Hence, from (4.69), we conclude that $E(A_{k,m}^n) = O\left(\sqrt{k_n} |Z|^{2k_n} \left(\frac{m_n}{n}\right)^{1-\delta}\right)$, which results in part b). \(\blacksquare\)

4.7.7 Proof of Theorem 7

The fact that $m = \Theta(n)$ implies the existence of $\alpha > 0$, such that $m \geq \alpha n$ for all sufficiently large $n$. Let $X$ be the process formed by concatenating i.i.d. blocks of
length \([1/\alpha]\), each block consisting of the same repeated symbol chosen uniformly from \(\mathcal{X}\). The first observation to note is that, for all \(n\) large enough that \(m \geq n\alpha\),

\[
D_{0,m}(X^n, Z^n) = 0 \quad a.s. \tag{4.70}
\]

This is because, by construction, \(X^n\) is, with probability 1, piecewise constant with constancy sub-blocks of length, at least, \([1/\alpha]\). Thus, a genie with access to \(X^n\) can choose a sequence of symbol-by-symbol schemes (in fact, ignoring the noisy sequence), with less than \(n\alpha\) (and, therefore, less than \(m\)) switches, that perfectly recover \(X^n\) (and, therefore, by our assumption on the loss function, suffers zero loss). On the other hand, the assumptions on the loss function and the channel imply that, for the process \(X\) just constructed,

\[
\limsup_{n \to \infty} \min_{\mathcal{X}^n} E L_{\mathcal{X}^n}(X^n, Z^n) > 0, \tag{4.71}
\]

since even the Bayes-optimal scheme for this process incurs a positive loss, with a positive probability, on each \([1/\alpha]\) super-symbol. Thus, we get

\[
E \left\{ \limsup_{n \to \infty} E \left[ L_{\mathcal{X}^n}(X^n, Z^n) - D_{0,m}(X^n, Z^n) \mid X^n \right] \right\} \geq \limsup_{n \to \infty} E \left[ L_{\mathcal{X}^n}(X^n, Z^n) - D_{0,m}(X^n, Z^n) \right] \geq \limsup_{n \to \infty} E L_{\mathcal{X}^n}(X^n, Z^n) \geq \limsup_{n \to \infty} \min_{\mathcal{X}^n} E L_{\mathcal{X}^n}(X^n, Z^n) > 0, \tag{4.75}
\]

where (4.73) follows from Fatou's lemma; (4.74) follows from (4.70); and (4.75) follows from (4.71). In particular, there must be one particular individual sequence \(x \in \mathcal{X}^\infty\) for which the expression inside the curled brackets of (4.72) is positive, i.e.,

\[
\limsup_{n \to \infty} E \left[ L_{\mathcal{X}^n}(X^n, Z^n) - D_{0,m}(X^n, Z^n) \mid X^n = x^n \right] > 0, \tag{4.76}
\]

which is equivalent to (4.37).
4.7.8 Proof of Theorem 8

First, by adding and subtracting the same terms, we obtain

\[ \begin{align*}
& EL_{\hat{X}_{n,k,m}}(X^n, Z^n) - D(P_{\hat{X}^n}, \Pi) \\
& = EL_{\hat{X}_{n,k,m}}(X^n, Z^n) - \min_{S \in S^n_{k,m}} EL_{\hat{X}_{n,s}}(X^n, Z^n) \\
& \quad + \min_{S \in S^n_{k,m}} EL_{\hat{X}_{n,s}}(X^n, Z^n) - D(P_{\hat{X}^n}, \Pi). 
\end{align*} \] (4.77)

We will consider term (i) and term (ii) separately. For term (i),

(i) \[ EL_{\hat{X}_{n,k,m}}(X^n, Z^n) - \min_{S \in S^n_{k,m}} EL_{\hat{X}_{n,s}}(X^n, Z^n) \]
\[ \leq 2k\Lambda_{\text{max}} n \left( \frac{2k}{n} - \frac{k}{n} \right) \left[ EL_{\hat{X}_{n,k,m}}(X_{k+1}^{n-k}, Z^n) - \min_{S \in S^n_{k,m}} EL_{\hat{X}_{n,s}}(X_{k+1}^{n-k}, Z^n) \right] \] (4.78)
\[ \leq 2k\Lambda_{\text{max}} n \left( \frac{n - 2k}{n} - \frac{2k}{n} \right) E \left[ L_{\hat{X}_{n,k,m}}(X_{k+1}^{n-k}, Z^n) - \min_{S \in S^n_{k,m}} L_{\hat{X}_{n,s}}(X_{k+1}^{n-k}, Z^n) \right] \] (4.79)
\[ \leq 2k\Lambda_{\text{max}} n \left( \frac{n - 2k}{n} - \frac{n}{n} \right) E \left[ L_{\hat{X}_{n,k,m}}(X_{k+1}^{n-k}, Z^n) - D_{k,m}(X^n, Z^n) \right], \] (4.80)

where (4.78) follows from upper bounding and omitting the losses for time instances \( t \leq k \) and \( t > n - k \) in the first and second terms of (i), respectively; (4.79) follows from exchanging the minimum with the expectation, and (4.80) follows from the definition (4.31) and \( \frac{n - 2k}{n} \leq 1 \).

For term (ii), we bound the first term in (ii) as

\[ \min_{S \in S^n_{k,m}} EL_{\hat{X}_{n,s}}(X^n, Z^n) \]
\[ \leq \frac{2k(m + 1)\Lambda_{\text{max}}}{n} + \frac{1}{n} \min_{S \in S^n_{k,m}} E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i - k} \Lambda(X_j, s_{k,j}(Z_{j+k}^{j+k})) | A^n \right] \], (4.81)

by upper bounding the losses with \( \Lambda_{\text{max}} \) on the boundary of the shifting points. Now, let \( P_{X_j | Z_{i}^t, A^n} \in \mathbb{R}^{|\mathcal{X}|} \) denote the \( |\mathcal{X}| \)-dimensional probability vector whose \( x \)-th component is \( Pr(X_j = x | Z_{i}^t, A^n) \). Then, we can bound the second term in (4.81) by
the following chain of inequalities:

\[
\begin{align*}
\min_{s \in S_{k,m}} E \left[ E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+k+1}^{\tau_i-k} \Lambda(X_j, s_k, (Z^i_{j-k})) \mid A^n \right] \right] \\
= E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+k+1}^{\tau_i-k} \min_{s_k \in S_k} E \left[ \Lambda(X_j, s_k, (Z^i_{j-k})) \mid A^n \right] \right] \\
= E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+k+1}^{\tau_i-k} \sum_{z_k \in Z^{2k+1}} P(Z^i_{j-k} = z_k \mid A^n) \min_{\hat{x} \in X} E \left[ \Lambda(X_j, \hat{x}) \mid Z^i_{j-k} = z_k, A^n \right] \right] \\
= E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+k+1}^{\tau_i-k} \sum_{z_k \in Z^{2k+1}} P(Z^i_{j-k} = z_k \mid A^n) U_A(P_{X_j \mid Z^i_{j-k} = z_k, A^n}) \right] \\
= E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+k+1}^{\tau_i-k} E \left[ U_A(P_{X_j \mid Z^i_{j-k} = z_k, A^n}) \mid A^n \right] \right] \\
= E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+k+1}^{\tau_i-k} E \left[ U_A(P^{(A^n)}_{X_0 \mid Z^i_{j-k}}) \mid A^n \right] \right] \\
\leq E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E \left[ U_A(P^{(A^n)}_{X_0 \mid Z^i_{j-k}}) \mid A^n \right] \right],
\end{align*}
\]

where (4.83) follows from the stationarity of the distribution in each block as well as the fact that the combination of the best k-th order sliding window denoiser for each block is in \( S_{k,m} \) and achieves the minimum in (4.82); (4.84) follows from conditioning; (4.85) follows from the definition (4.2); (4.86) follows from the stationarity of the distribution in each \( i \)-th block; and (4.87) follows from adding more nonnegative terms.

For the second term in (ii), we first define

\[
n_i(A^n) \triangleq \tau_i(A^n) - \tau_{i-1}(A^n)
\]

as the length of the \( i \)-th block, for \( 1 \leq i \leq r(A^n) + 1 \). Obviously, \( n_i(A^n) \) also depends on \( A^n \), and, thus, is a random variable, but we again suppress \( A^n \) for brevity and
CHAPTER 4. DISCRETE DENOISING WITH SHIFTS

141

denote it as \( n_i \). Then, similar to the first term above, we obtain

\[
\mathbb{D}(P^{X^n}, \Pi) = \min_{\hat{X}_n \in \mathcal{D}_n} E L_{X^n}(X^n, Z^n)
\]

\[
= \frac{1}{n} \min_{\hat{X}_n \in \mathcal{D}_n} E \left[ E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} \lambda(X_j, \hat{X}_j(Z^n)) \big| A^n \right] \right]
\]

\[
= \frac{1}{n} E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} \min_{\hat{X}_n \in \mathcal{D}_n} E \left[ \lambda(X_j, \hat{X}_j(Z^n)) \big| A^n \right] \right]
\]

\[
= \frac{1}{n} E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} \min_{\hat{X}_n \in \mathcal{D}_n} E \left[ \lambda(X_j, \hat{X}_j(Z^n)) \big| A^n \right] \right] \tag{4.88}
\]

\[
= \frac{1}{n} E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E \left[ U_{\lambda}(P^{(A_{\tau_i})}_{X_{\tau_i}^n}) \big| A^n \right] \right]
\]

\[
= \frac{1}{n} E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E \left[ U_{\lambda}(P^{(A_{\tau_i})}_{X_{\tau_i}^n}) \big| A^n \right] \right] \tag{4.89}
\]

\[
\geq \frac{1}{n} E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E \left[ U_{\lambda}(P^{(A_{\tau_i})}_{X_{\tau_i}^n}) \big| A^n \right] \right], \tag{4.90}
\]

where (4.88) follows from the conditional independence between different blocks, given \( A^n \); (4.89) follows from the stationarity of the distribution in each block, and (4.90) follows from \( [7, \text{Lemma 4(1)}] \). Therefore, from (4.81), (4.87), and (4.90), we obtain

\[
(b) = \min_{S \in S_{k,m}} E L_{X^n,s}(X^n, Z^n) - \mathbb{D}(P^{X^n}, \Pi)
\]

\[
\leq \frac{2k(m+1) \Lambda_{\max}}{n} + \frac{1}{n} E \left[ \sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E \left[ U_{\lambda}(P^{(A_{\tau_i})}_{X_{\tau_i}^n}) \big| A^n \right] - E \left[ U_{\lambda}(P^{(A_{\tau_i})}_{X_{\tau_i}^n}) \big| A^n \right] \right]
\]

\[
= \frac{2k(m+1) \Lambda_{\max}}{n} + \frac{1}{n} E \left[ \sum_{i=1}^{r+1} \frac{n_i}{n} \cdot \left\{ E \left[ U_{\lambda}(P^{(A_{\tau_i})}_{X_{\tau_i}^n}) \big| A^n \right] - E \left[ U_{\lambda}(P^{(A_{\tau_i})}_{X_{\tau_i}^n}) \big| A^n \right] \right\} \right]. \tag{4.91}
\]

Now, observe that, regardless of \( A^n \), the sequence of numbers \( \left\{ \frac{n_i}{n} \right\}_{i=1}^{r+1} \) form a probability distribution, since \( \sum_{i=1}^{r+1} \frac{n_i}{n} = 1 \) and \( \frac{n_i}{n} \geq 0 \) for all \( i \), with probability 1. Then, based on the fact that the average is less than the maximum, we obtain the further
upper bound

\[
(4.91) \leq \frac{2k(m+1)\Lambda_{\text{max}}}{n} + E\left[\max_{i\in\{1,\ldots,M\}} \left\{ E\left[U_{\Lambda}(P^{(i)}_{X_0|Z_{k-1}^k})\right] - E\left[U_{\Lambda}(P^{(i)}_{X_0|Z_{k}^{\infty}})\right]\right\}\right].
\]

The remaining argument to prove the theorem is to show that the upper bounds (4.80) and (4.92) converge to 0 as \(n\) tends to infinity. First, from the given condition on \(k = k_n\) and \(m = m_n\), the maximal allowable growth rate for \(k\) is \(k = c_1 \log n\), which leads to \(\lim_{n \to \infty} \frac{2k\Lambda_{\text{max}}}{n} = 0\). In addition, the condition requires \(m = o(n)\), and \(k\) to be sufficiently slow, such that \(k|Z|^{2k} = o\left(\left(\frac{n}{m}\right)^{1-\delta}\right)\), which implies \(k = o\left(\frac{n}{m}\right)\). Therefore, \(\lim_{n \to \infty} \frac{2k(m+1)\Lambda_{\text{max}}}{n} = 0\). Furthermore, from conditioning on \(X^n\), bounded convergence theorem, and part b) of Theorem 6, we obtain \(\lim_{n \to \infty} E[L_{X_{\text{univ}}}^{n,k,m}(X_{k+1}^n, Z^n) - D_{k,m}(X^n, Z^n)] = 0\). Thus, we have

\[
\begin{align*}
\limsup_{n \to \infty} E[L_{X_{\text{univ}}}^{n,k,m}(X^n, Z^n) - D(P_X, \Pi)] &\leq \limsup_{n \to \infty} E\left[\max_{i\in\{1,\ldots,M\}} \left\{ E\left[U_{\Lambda}(P^{(i)}_{X_0|Z_{k-1}^k})\right] - E\left[U_{\Lambda}(P^{(i)}_{X_0|Z_{k}^{\infty}})\right]\right\}\right] \\
&\leq E\left[\limsup_{n \to \infty} \max_{i\in\{1,\ldots,M\}} \left\{ E\left[U_{\Lambda}(P^{(i)}_{X_0|Z_{k-1}^k})\right] - E\left[U_{\Lambda}(P^{(i)}_{X_0|Z_{k}^{\infty}})\right]\right\}\right] \\
&= 0, \\
\end{align*}
\]

where (4.93) follows from the reverse Fatou’s lemma, and (4.94) follows from [7, Lemma 4(2)] and \(M\) being finite. Since it is clear that \(\liminf_{n \to \infty} [E[L_{X_{\text{univ}}}^{n,k,m}(X^n, Z^n) - D(P_X, \Pi)] \geq 0\) by definition of \(D(P_X, \Pi)\), the theorem is proved.

Remark: As in [7, Theorem 3], the convergence rate in (4.44) may depend on \(P_X\), and there is no vanishing upper bound on this rate that holds for all \(P_X \in \mathcal{P}\{m_n\}\). However, we can glean some insight into the convergence rate from (i) and (ii): whereas the term (i) is uniformly upper bounded for all \(P_X \in \mathcal{P}\{m_n\}\), the rate at which

\[\text{Recall part b) of Theorem 6, where a uniform bound (uniform in the underlying individual sequence) on } E[L_{X_{\text{univ}}}^{n,k,m}(x^n, Z^n) - D_{k,m}(x^n, Z^n)] \text{ was provided in the semi-stochastic setting. Clearly,}\]

term (ii) vanishes depends on $P_X$. In general, we observe that the slower the rate of increase of $k = k_n$, the faster the convergence in (i), but the convergence in (ii) is slower. With respect to the rate of increase of $m_n$, the slower it is, the faster the convergence in (i), but whether or not the convergence in (ii) is accelerated by a slower rate of increase of $m_n$ may depend on the underlying process distribution $P_X$.

In the stochastic setting the same bound holds on $E[L_{X_{min},m}(X^n, Z^n) - D_{k,m}(X^n, Z^n)]$, regardless of the distribution of $X^n$. 

Chapter 5

Contributions and future work

5.1 Contributions of the thesis

Here, we summarize the contributions of this thesis.

In Chapter 2, the first theoretical justification for using hidden Markov modeling in the universal filtering problem was established. We proved that, for the known, invertible DMC, a family of filters based on HMPs is universally asymptotically optimal for any general stationary and ergodic \( \{X_t\} \) satisfying some mild positivity condition. That is, we showed that our sequence of schemes achieves optimum performance regardless of the clean source distribution. We also extend this scheme to the case in which the channel noise itself is an HMP.

A different take on the universal filtering problem was presented in Chapter 3. We devised a universal filter that, for every bounded underlying signal, performs essentially as well as the best FIR filter without any knowledge of the underlying signal and with only the knowledge of the first and second moments of the noise, under the MSE criterion. We showed that the regret vanishes in both expectation and high probability, and the decay rate of the regret of the expected MSE was shown to be logarithmic in \( n \). Obtaining the logarithmic regret was not straightforward because the estimated loss functions \( \{\ell_t(u)\}_{t \geq 1} \) are not always exp-concave functions. We also presented several simulation results that support our theoretical guarantees and show the potential merits of applying our filter.
Finally, in Chapter 4, inspired by the DUDE algorithm, we developed a generalization that accommodates switching between sliding window rules. We have shown a strong semi-stochastic setting result for our new scheme by competing with shifting $k$-th order denoisers. This result implied a stochastic setting result as well, asserting that the S-DUDE asymptotically attains the optimal distribution-dependent performance for the case in which the underlying data is piecewise stationary. We also described an efficient low-complexity implementation of the algorithm and presented some simple experiments to demonstrate the potential benefits of employing S-DUDE in practice.

In this thesis, we have used two different methods to learn from the noisy data and devise universal estimation schemes. In Chapter 2, we used the rich parametric model class, the HMP class, to learn the posterior probability of the true probability distribution for any stationary and ergodic source processes. Finding the HMP model that best approximates the noisy data enabled us to learn the true posterior probability required for the filtering. In the latter two chapters, we utilized the unbiased estimates of the loss and showed that minimizing the sum of estimated losses leads to the asymptotically optimum performance guarantees of the universal estimation schemes. The unbiased estimates were available due to the assumption of known and memoryless channels. Above two methods are not new, but we were able to rigorously analyze the first method and attain novel universal estimation schemes rooted in the second method. Our investigation asserts that with the known noisy channel assumption, we can efficiently learn the source data and attain optimum estimation performance, as the size of the data increases.

5.2 Future work

We now outline several possible extensions of the research in this thesis.

With respect to future work related to Chapter 3, we can extend our scheme to compete with reference classes that are larger than the class of FIR filters in order to
further minimize the MSE. One possible extension is to devise a scheme that competes with the class of switching FIR filters that parallels the switching predictors in [67] and switching denoisers in [11]. Again, obtaining the expected regret of rate $O(\frac{m \log n}{n})$ as in [67], where $m$ is the number of switches, may not be straightforward since the induced loss function is not always exp-concave, a condition required for the scheme in [67]. Another direction is to compete with the class of general non-linear schemes as has been accomplished for the denoising case in [80]. However, in this case, the procedure for obtaining an unbiased estimate of the true MSE would not be as simple as that in this paper, since the martingale relationship in Lemma 10 relied heavily on the linearity of the filter. Instead, the channel inversion process developed in [80] may be a necessary component.

There are several future research directions related to Chapter 4. The S-DUDE can be thought of as a generalization of the DUDE, with the introduction of a new component captured by the non-negative integer parameter $m$. Many previous extensions of the DUDE, such as the settings of channels with memory [31], channel uncertainty [81], applications to channel decoding [82], discrete-input, continuous-output data [83], denoising of analog data [80], and decoding in the Wyner-Ziv problem [84], may benefit from a revision that would incorporate the viewpoint of switching between time-invariant schemes. Particularly, extending S-DUDE to the case in which the data are analog as in [80] will be non-trivial and interesting from both a theoretical and a practical viewpoint. In addition, as mentioned in Section 4.5, an extension of the S-DUDE to the case of multi-dimensional data is not as straightforward as the extension of the DUDE. Such an extension should prove interesting and practically important. We have conducted some initial experimental work on this extension that uses the quadtree decomposition with some promising experimental results. The details of this extension will be provided in [50]. Finally, it would be useful to devise guidelines, in the spirit of those in [85, 70], for the choice of $k$ and $m$ based on $n$ and the noisy observation sequence $z^n$. 
Bibliography


