Optimal Nonlinear Digital Signal Processing: A Dynamical Systems Approach

by

Omer Tanovic

B.S., Automatic Control and Electronics, University of Sarajevo (2007)
S.M., Automatic Control and Electronics, University of Sarajevo (2011)

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
August 30, 2019

Signature redacted

Certified by .....
Alexandre Megretski
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Optimal Nonlinear Digital Signal Processing:

A Dynamical Systems Approach

by

Omer Tanovic

Submitted to the Department of Electrical Engineering and Computer Science
on August 30, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis addresses optimal nonlinear digital signal processing problems aimed to improve power efficiency of modern wireless transmission systems.

The first part of this thesis is motivated by peak-to-average power ratio reduction of communication signals. The problem is formulated as minimization of a frequency-weighted convex quadratic cost subject to time-domain output amplitude constraints. A new method for converting optimality conditions into finite-latency stable systems generating optimal outputs with arbitrary precision is proposed.

The second part contains analysis of the nonlinear distortion introduced into the baseband (discrete-time) input-output dynamics of the communication systems by the (continuous-time) power amplifier nonlinearity. It is shown that when the nonlinearity is represented by a Volterra series model the resulting baseband equivalent model is a series interconnection of a discrete-time Volterra series model, of the same degree and equivalent memory depth, and a linear system. The result suggests a new, analytically motivated, structure of digital pre-distortion (DPD) of power amplifier nonlinearities.

The third part of the thesis focuses on analysis and design of digitally implemented pulse-width modulators (DPWM) used as quantizers for power amplifiers in switched-mode operation. A time-domain input-output model of DPWM which offers new insight into nonlinear behavior of this system is developed. A modified Lloyd-Max quantization based algorithm for linearization of the baseband of a DPWM output is proposed.

Thesis Supervisor: Alexandre Megretski
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

The road to PhD is narrow, steep and thorny, which is a well-known fact to those who dare to step on that path. While it has brought tremendous joy to my scientific mind, craving for challenges, there were also many moments of self-doubt and reflection, deep and anxious, that are inevitable part of this long journey. I was privileged to have Professor Alexandre (Sasha) Megretski as my thesis advisor, whose invaluable support and precious mentorship have guided me throughout this process. His enormous energy, enthusiasm, and brilliance helped to shape this thesis, while his valuable feedback on my work helped my ideas blossom and come into life. I am very grateful to Sasha for generously helping me become a better researcher, a better engineer and a better human being.

I am also greatly thankful to my thesis committee, Professors George Verghese and Pablo Parrilo, for numerous thought-provoking discussions and insightful suggestions. Their generosity has helped me grow. I would also like to thank Professors Asu Ozdaglar, Munzer Dahleh, and Sanjoy Mitter for always having been available for my questions, for having provided me with a constant support and mentorship throughout my PhD.

I would like to thank the members (or the last lineup) of the ELASTx project team at MIT: Dr. Yehuda Avniel, Dr. Yan Li, Dr. Zhipeng Li, and Professor Vladimir Stojanovic, for all the work we have done together.

Results in Chapter 4 of this thesis have been obtained in collaboration with researchers from Mitsubishi Electric Research Laboratories (MERL) in Cambridge, MA, where I spent two summers as a research intern during my graduate studies. I would specifically like to thank Dr. Rui Ma, Dr. Koon Hoo Teo, and Dr. Philip Orlik for giving me the opportunity to work at MERL and for all the work we have done together.

I would like to thank Professor Russ Tedrake (and the Robot Locomotion Group) for his generosity, and for providing me with funding support and a research home in the critical years of my PhD.

I would like to thank all the great, both former and current, administrative staff at LIDS for their tremendous help at different stages, and in various aspects, of my student life: Lynne Dell, Jennifer Donovan, Gracie Gao, Lisa Gaumond, Francisco Jaimes, Brian Jones,

and Debbie Wright. Invaluable debt goes to Janet Fischer and Professor Leslie Kolodziejski of the EECS Graduate Student Office, who have been the bedrock for many generations of graduate students in the EECS Department. Thank you for that!

Over the years at MIT, and especially at LIDS, I have been blessed with many great friendships. We have been each others support throughout the years and I am grateful on that: my officemates (and incredible friends) Noele Norris, Hamza Fawzi, Jennifer Tang, Quan Li, Chenyang Chuan, and Suhas Kowshik; my lunchmates (and much more!) Christina Lee, Shreya Saxena, and Elie Adam; my LIDS and non-LIDS friends Ali Faghih, Ali Kazerani, Arman Rezaee, Elaheh Fata, Eren Kizildag, Ermin Wei, Flora Meng, Giancarlo Baldan, Hajir Roozbehani, Igor Kadota, Igor Spasojevic, Luis Castro, Mitra Osqui, Prince Singh, Rose Faghih, Kimon Drakopulos, Mark Tobenkin, and Yola Katsargyri. I hope that our friendships will continue for the many years to come.

I would like to thank my parents for their unconditional love and sacrifice. Knowledge is not to be taken for granted and they used every possible opportunity and fought hard to provide us with books and education even in times when it was unimaginable and impossible. They are those who planted a seed of love for science and learning in me and my sister, which we now cherish dearly and are very grateful for.

I would like to thank my sister Selma, for having been my best friend, an inspiration and a counselor during my whole life, and for taking care of my health and wellbeing at the times when I forgot to do so myself.

Above all, I would like to thank my lovely wife, Majda, for having been my stronghold, throughout all these years, in good and in bad. I am extremely thankful for her love, sacrifice and friendship. Without Majda, there would be no thesis. She is the one who made it possible. And I am extremely grateful to my three little boys, Bakir, Adem and Emin, for bringing me joy and happiness, for their warm hugs and beautiful smiles after long days spent in the LIDS office. You have made this work worthwhile!

The research in this thesis was funded in part by the following grants, which I gratefully acknowledge: DARPA, Award No. W911NF-10-1-0088; Lockheed Martin Corporation, Award No. RPP2016-002; National Science Foundation, Award No. 1743938; Office of Naval Research, Award No. N000141812210.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

System modeling is the process of developing abstract models of a system. The model we choose depends on the level of abstraction that we want to achieve and the perspective of the system that we wish to understand, that is, the nature of questions that we wish to answer about the system. On the other hand, system design is the process of defining the elements of a system such as the architecture, modules and components, and data that goes through the system, so that it satisfies specific needs and requirements. System modeling and design are ubiquitous in all branches of engineering and science. They range in the level of abstraction: from modeling and design of transistors at the physical level [1] to abstract modeling of bacteria as ergodic stochastic hybrid systems [2]. They range in scale: from designing simple operational amplifiers to designing national electric grids and nuclear power plants. The process of designing a system naturally starts with specifying the structure of the system (or some useful properties of the system) and specifying certain performance requirements that the system should adhere to. These requirements can be very different in nature: from specifying the maximal delay in a computer network to restricting the number of arithmetic operations that can be performed in a clock interval of a digital circuit or limiting the amount of power or area on a chip that a system can consume.

Very often, it is desirable to design systems that satisfy the imposed requirements in an *optimal way*, or, in other words, optimally with respect to some measure of quality or performance. Designing systems in such a way can be a very challenging task, depending on the corresponding abstract optimization problem (e.g., convex vs. non-convex) or the

assumed system properties (e.g., linear vs. nonlinear, time-invariant vs. time-varying, etc.). This is especially true if one aims at finding exact or suboptimal solutions of arbitrary accuracy. For that reason, in signal processing and control theory, optimal system design has been mainly restricted to linear systems (or, even more restricted, to linear time-invariant systems). There are many examples of such system design problems, and some of the most well known are the Wiener filter problem [3], Kalman filter problem [4], matched filter as a maximum likelihood estimator [5], optimal design of finite impulse response (FIR) filters [6], and H2 and H-infinity control [7] problems.

Adding seemingly simple constraints on the systems to be designed can very often lead to a significant increase in the complexity of the corresponding optimization problem. An example is the difference between an FIR, non-causal and general causal solution to the Wiener filter problem [3], where solving the latter becomes much more involved (though all three can still be solved explicitly). This difference becomes even more prominent as one drops assumptions of linearity of either the system to be designed or some of the subsystems involved in the optimization problem. For example, optimal filters for nonlinear systems are in general very difficult to derive or implement, despite the simplicity of their linear counterpart — the Kalman filter. The common approach is to use approximate solutions such as the extended Kalman filters, ensemble filters or particle filters (see, e.g., [8] or [9] and the references therein). However, no optimality properties can be guaranteed by these approximations. Moreover, even the stability of the estimation error often cannot be ensured [10].

Another example is the nonlinear H-infinity control problem [11]. It was shown that the solution of this problem in the case of state feedback can be determined from the solution of the corresponding Hamilton-Jacobi equation (HJE), which is the nonlinear version of the Riccati equation considered in the H-infinity control problem for linear systems [12]. In fact, the nonlinear output feedback H-infinity controller satisfies the separation property, and the necessary and sufficient conditions of optimality for such problems involve solving two HJE's [13], [14]. Though the theory of nonlinear H-infinity control has been well developed, actually solving the HJE still remains a challenge and is the major bottleneck for its practical application. See [15] and the references therein. Therefore, it is hard to

expect explicit solutions for optimal system design problems whenever some notion of nonlinearity is involved. Furthermore, even analysis of such problems sometimes becomes intractable.

In this thesis, we study a general problem on the intersection of signal processing, communications, and circuits, pertaining to the issue of efficiency of wireless transmission systems. This problem encompasses several modeling and optimal system design questions involving different nonlinear systems. We show (to one's surprise!) that each of these sub-problems either admits some form of explicit solution or enables rigorous mathematical analysis of the corresponding optimal system.

In the next section, we introduce the problem of energy efficiency in modern wireless communication systems and discuss several approaches to mitigating this problem. We show that each of these approaches leads to the modeling and design of a nonlinear system, each of which is optimal in a useful, specified way.

1.1 The Problem of Power Efficiency in Modern Wireless Transmission Systems

Wireless communications has been one of the most successful technological innovations in modern history. In the past two decades, cellular networks have developed significantly and the number of users has increased exponentially. This has lead to a rapid expansion of the related infrastructure. For example, more than 8 million cellular base stations (BS) have been deployed in the world today [16]. According to the often-cited Gartner report [17], the Information and Communications Technologies (ICT) sector contributes 2% of global Greenhouse Gases (CO_2) emissions [18], where, according to the SMART 2020 report [19] about 43% of emissions come from the mobile sector. This is predicted to rise to 51% by 2020 [19]. On the other hand, the operational expenditures of cellular networks were mostly on electricity bills. For example, in 2005, the ICT market in the US required 1% of the total energy produced [20]. Moreover, approximately 3% of the total worldwide electrical energy is consumed by ICT [21], [22]. Therefore, the cellular network

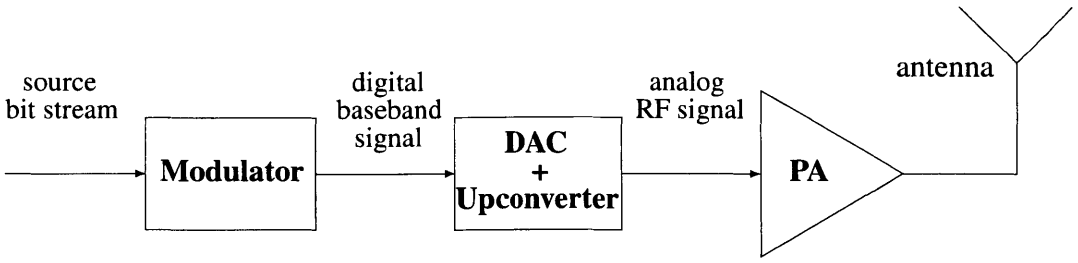


Figure 1-1: Simplified block diagram of an RF transmitter system.

operators aim to improve the energy efficiency of cellular networks not just for the negative environmental effects but to maintain their profitability as well.

The growing number of base stations has significantly increased the energy consumption of cellular networks because these stations account for around 57% of the total consumed energy in the network [18]. Out of the total supplied energy to the base station, between 50% and 80% gets consumed by the radio-frequency (RF) power amplifier (PA) and the corresponding transceiver circuitry [23]. Unfortunately, a tremendous quantity of the supplied energy to the PA is dissipated in terms of heat and only a small fraction corresponds to the useful output [24]. Therefore, the power amplifier is a component with the greatest impact on power efficiency of the whole transmitter system.

A simplified block diagram of a radio-frequency (RF) transmitter system is shown in Figure 1-1. Let P_i, P_o, and P_{DC} be the input, output, and the total DC power supplied to the PA, respectively. The ratio P_o/P_{DC} is called the *power efficiency* (PE) of the PA (sometimes also drain or collector efficiency, depending on the technology of the underlying transistors), and represents one of the most important characteristics of the transmitter circuit (we should note here that *power added efficiency* (PAE) is another popular metric and is defined as the ratio $(P_o - P_i)/P_{DC}$. For systems with high power gain, the PAE approaches the PE). A typical PA's input-output characteristic is shown in Figure 1-2. In order to achieve high PE, the PA must be operated close to its saturation level, that is, the PA should be driven in such a way that the average power of the input signal is as close to the compression region as possible. How close the operating point could be to the compression region depends on the peak-to-average power ratio (PAPR) of the input signal and the

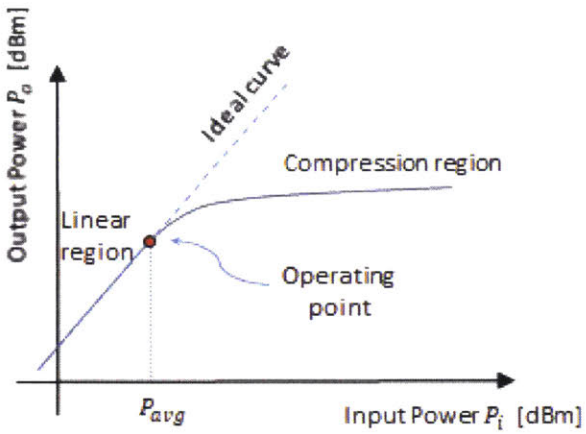


Figure 1-2: Typical PA input/output characteristic.

amount of distortion that can be tolerated at the output of the PA. Namely, if the operating point (i.e, the average power) is close to compression and the input signal PAPR is large, the input signal would be subject to a considerable nonlinear distortion when amplified by the PA. This distortion causes in-band and out-of-band spectral content which heavily degrades linearity measures of performance of a transmitter system. The conventional solution to this problem is to back-off the operating point of the PA, that is, to increase the DC supply to the PA in order to distance the operating point from the compression region (while keeping the average power at a needed level). This increases linearity but drastically decreases power efficiency of the transmitter. Unfortunately, modern wireless communication signals (LTE or LTE-A [25]) require orthogonal frequency division multiplexing (OFDM) modulation method, which generates signals with high PAPR (see Figure 1-3). Such signals suffer from significant distortion when passed through a nonlinear PA operating close to saturation and, therefore, significant back-off has to be applied. This forces conventional power amplifiers to operate over a large portion of their transfer curves which, in turn, causes low power efficiency of the transmitter circuit.

This motivates a search for methods or algorithms which would help increase power efficiency while keeping the system linearity measures in the prescribed limits. Various approaches for achieving this have been proposed in the literature in the past two decades. In the following sections, we describe three such approaches, review the corresponding

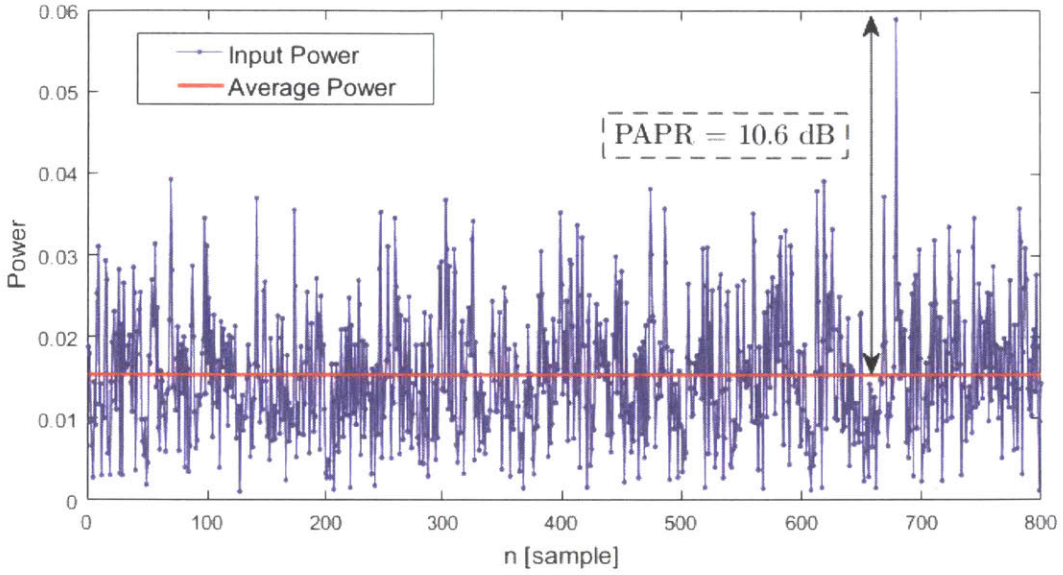


Figure 1-3: An example of power of a typical OFDM transmit waveform. It has undesirably high PAPR of 10.6 dB and high peaks that occur rarely.

state-of-the-art methods and discuss their strengths and shortcomings.

1.1.1 Peak-to-Average Power Ratio Reduction

From the above discussion, it is clear that the most obvious way for increasing efficiency of an RF PA is to reduce the peak-to-average power ratio of the communication signal being transmitted. PAPR of a signal is defined as the ratio of its peak power to the average power (as its name suggests). As noted earlier, modern communication signals suffer from undesirably high PAPR which, for LTE signals, is typically 10-12 dB [26]. For example, high value of PAPR represents the main drawback of orthogonal frequency division multiplexing (OFDM) signal generation for forthcoming wideband communication systems.

Over the last two decades, many PAPR reduction methods have been proposed in the literature. These methods can be broadly classified into three main categories ([27]), depending on where in the signaling chain one attempts to modify the transmit signal: *coding techniques* (modify the source bit stream), *multiple signaling and probabilistic techniques* (modify the way in which the OFDM signal is generated), and *signal distortion techniques* (modify the OFDM transmit signal). We now briefly describe and discuss the most popular

techniques in each of these categories (for more examples and detailed analysis see, e.g., [27, 28, 29], and the references therein).

- **Coding Techniques**:

  (a) *Linear Block Coding*: In this method, some bits of the codeword that are dedicated for error correction are used for PAPR reduction. The main goal is to choose the codewords with low PAPR. Various linear block coding, sub-block coding, and complement block coding schemes were proposed in [30, 31, 32]. The use of low-density parity-check codes, fountain codes, and Raptor codes were proposed in [33, 34, 35].

  (b) *Golay Complementary Sequences (GSC)*: GSCs have been used as codewords to modulate the subcarriers of the OFDM system and produce a signal with PAPR bounded by 2dB [36, 37]. Unfortunately, the usefulness of this technique is limited to OFDM systems with small number of subcarriers. In the case of large number of subcarriers, this technique suffers from transmission rate loss and increased computational complexity due to the exhaustive nature of the search required to find good codes.

  (c) *Turbo Coding*: In this method, turbo codes are used to generate various OFDM symbols by turbo encoding with different interleavers, and then choose the one with the lowest PAPR for transmission [38], Another approach is to use dual bose-ray-chaudhuri (BCH) codes, since it was shown that the IFFT of the codewords exhibits low PAPR [39].

- **Multiple Signaling and Probabilistic Techniques**:

  (a) *Selective Mapping (SLM)*: In this technique, a number of different candidate data blocks are generated from the original baseband data, all representing the same information. Then candidate block with the best PAPR is selected for transmission [40, 41]. This technique works with an arbitrary number of subcarriers and any modulation method. However, it requires information about

the modified data block to be transmitted to the receiver, which results in data rate loss.

(b) *Partial Transmit Sequence (PTS)*: In this method, an input data block is partitioned into disjoint sub-blocks, and the subcarriers in each subblock are weighted in phase by a carefully designed factor. These factors are chosen such that the PAPR of the combined signal is minimized [42, 43, 44]. Similar to SML, this technique works with any number of subcarriers and any modulation method. However, the PTS method suffers from high complexity in searching for the best phase factors. In addition, the phase factor information has to be transmitted to the receiver, which results in data rate loss.

(c) *Tone Reservation/Tone Injection (TR/TI)*: These methods are based on adding a data-block-dependent time domain signal to the original OFDM multicarrier signal to reduce its PAPR [45]. In TR, a certain subset of subcarriers (i.e., tones) is reserved to exclusively transmit carefully designed signals which, when added to the data bearing subcarriers, yield lower PAPR of the overall data block. Since the subcarriers are orthogonal, this method does not increase EVM. However, for good PAPR reduction performance, a non-negligible number of subcarriers has to be reserved resulting in bandwidth sacrifice. In addition, information about the reserved subcarriers has to be conveyed to the receiver further reducing the useful bandwidth. The basic idea of TI is to add a carefully generated signal to the true transmit signal, which would be equivalent to increasing the constellation size and would add redundancy (each point in the original constellation corresponds to several equivalent points in the new constellation). By optimizing the injection signal it is possible to reduce the PAPR of the transmit signal. However, the injected signal gets spectrally spread over the whole bandwidth and typically leads to an increase in EVM. In addition, the method may result in an increase in average power of the transmit signal potentially causing an increase in EVM.

(d) *Active Constellation Extension (ACE)*: In this method, the symbol constellation

(i.e., the modulation constellation points, e.g., QPSK) is modified such that the PAPR of the newly formed symbol sequence is lower than for the original data, while not degrading the EVM performance significantly [46]. This technique increases the average power of the transmitted signal and is not suitable for modulation methods with a large number of constellation points [28].

- **Signal Distortion Techniques**:

  (a) Companding Transform: The basic idea behind this method comes from the use of companding (*com*pressing and then ex*panding*) technique in speech processing which is used to reduce the number of bits needed to encode the signal. Since speech signals are similar to transmit OFDM signals, in that they exhibit high PAPR with large amplitude peaks that occur infrequently, similar companding techniques can be used [47, 48, 49, 50].

  (b) Peak Cancellation: In this technique, a pre-specified waveform is appropriately scaled, shifted and subtracted from the OFDM signal whenever a potential peak higher than a certain threshold is detected [51, 52]. Though very simple to implement this technique does not scale well with the number of peaks to cancel if they appear in a relatively short time-segment of the transmit signal. In addition, there are no guaranties on the amount of peak regrowth after performing the peak cancellation.

  (c) *Iterative Clipping and Filtering (ICF)*: By far the most popular technique among practitioners is (iterative) clipping and filtering [53]. Since large peaks occur with very low probability [54, 55], clipping seems to be an effective method for PAPR reduction. Unfortunately, nonlinear operation of clipping causes in-band and out-of-band distortion resulting in worsened EVM and ACLR performance. The out-of-band spectral regrowth caused by clipping is mitigated by post filtering, which in turn can generate significant peak regrowth. For that reason, clipping and filtering is commonly applied iteratively [56, 57]. In traditional ICF, the operation of filtering is performed by an ideal rectangular filter (this is feasible since, in practice, filtering is implemented through multiplication of

the DFT samples of the finite length signals involved). In [58] and [59], the post-clipping filter is optimized to ensure minimal in-band distortion (through minimizing EVM of the transmit signal). The filter coefficients are designed by solving a second-order cone (SOCP) program at each iteration. It was shown that this method achieves similar PAPR reduction performance as the traditional ICF methods but in significantly lower number of iterations.

The usefulness of coding techniques for PAPR reduction is practically limited to modulation schemes with a small number of constellation points and multicarrier systems with a small number of subcarriers. This is a result of the computational intractability of exhaustive search algorithms which are needed in order to look for good codes. Similarly, the signaling and probabilistic techniques for PAPR reduction result in significant spectral efficiency (i.e., data rate) losses due to the need for conveying important information about the signal transformation to the receiver. Furthermore, the above techniques for PAPR reduction rely on some form of signal preconditioning that is performed either on the source bit stream or the modulated constellation symbols, requiring access to the module/circuitry which generates the OFDM transmit signal. Though developing such techniques is of great academic interest, their utility for already deployed communication systems is somewhat restricted. Namely, in practice, the internal structure of the modules that generate 3GPP compliant signals is inaccessible due to hardware restrictions enforced by the circuit manufacturers and the functional limitations imposed by the communication standards. For that reason, the *signal distortion techniques* for PAPR reduction, which operate on the OFDM transmit signals, have been the most popular among practitioners. In particular, the iterative clipping and filtering is the method of choice due to its low hardware complexity. Unfortunately, in order to achieve good reduction in PAPR, the number of needed iterations can be prohibitively large. This is especially true when the number of subcarriers is large and the sample-rate of the baseband transmit signal is high. The number of required iterations can be reduced by employing ICF with optimized filters. However, due to the high computational complexity of the underlying SOCP this method also does not scale well with the number of subcarriers or the baseband sample-rate. Though no provable performance guarantees have been reported for the available ICF techniques, they remain the method of

choice among RF engineering practitioners.

Due to the various parameters that can be changed in multicarrier transmission systems, it is reasonable to expect that no specific PAPR reduction technique would be the best solution for all possible cases. Nevertheless, many of these methods have been successfully employed in industry. However, the theoretical results are limited and lacking. Furthermore, it is clear that the available techniques for PAPR reduction have some serious limitations, even in the most basic situations. Namely, for multicarrier systems with a large number of sabcarriers their practicality becomes questionable. Another example is PAPR reduction of transmit signals with multiple non-contiguous bands. Even more, no available technique provides any guaranties on system performance: this is true even in the case of optimized ICF where ACLR and EVM are minimized while there are no guaranties on the PAPR values after filtering. It is clear that having a technique which can resolve the above shortcomings, and possibly achieve that in an optimal way, can be very valuable.

1.1.2 Digital Predistortion and Power Amplifier Behavioral Modeling

Digital compensation offers an attractive approach to designing electronic devices with superior characteristics, and it is not a surprise that it has been used for PA linearization as well. When digital compensation is used to correct for analog nonlinearities in the PA, the operating point of the PA can be chosen closer to the saturation region of its transfer curve, which increases power efficiency of the transmission circuit. Nonlinear distortion in an analog system can be compensated with a pre-distorter or a post-compensator system. In particular, a pre-distorter inverts nonlinear behavior of the analog part, and is usually implemented as a digital system. Techniques which employ such systems are called digital predistortion (DPD) techniques, and they can produce highly linear transmission circuits [60]-[61].

Structure of a digital pre-distorter usually depends on the choice of a behavioral model used to describe the PA [62], and consequently on the corresponding equivalent baseband model. Based on this model, one choses a finite sequence (i.e., a basis) of subsystems of appropriate structure and restricts the actual compensator to be a linear combination of

these basis subsystems. Clearly, selecting a proper compensator structure is a major challenge in compensator design: a basis which is too simple will not be capable of canceling the distortions well, while a form that is too complex will consume excessive power and space in its hardware implementation. Therefore, having an insight into the compensator basis selection can be very valuable!

First attempts to mitigate PA's nonlinear effects by employing DPD used simple memoryless models in order to describe PA's behavior [63]. With the increase of signal bandwidth over time, it has been recognized that short and long memory effects play significant role in PA's behavior [64], and should be incorporated into the model. Since then several memory baseband models and corresponding predistorters have been proposed to compensate memory effects: memory polynomials [65], [66], Hammerstein and Wiener DPD models [67], [68], pruned Volterra series [69], generalized memory polynomials [70], dynamic deviation reduction-based Volterra models [71], [72], as well as the most recent neural networks and generalized rational functions based behavioral models [73], [74]. In behavioral modeling based on memory polynomials, it is commonly assumed that all branches in the model (i.e., all monomials of equal degree) are generated with the same memory requirements. Since memoryless nonlinear effects dominate in terms of degree, the above choice leads to unnecessary complexity of the PA model. Several models were proposed in which strong memoryless nonlinear effects and weak memory dynamic effects are decoupled in order to reduce model complexity while keeping good modeling/linearization performance, such as twin nonlinear two-box (TNTB) [75] and three-box (PLUME) models [76].

In the above works, the goal was to linearize the full (or more precisely, the non-negligible part of) spectral range of the PA output signal. This range is proportional to the order of nonlinearity of the PA, and is in practice taken to be about five times the input bandwidth [62]. Hence, for wideband input signals the linearization bandwidth would be very large, and would put a significant burden on the system design, e.g., it would require very high-speed data converters. This clearly represents a major drawback for the forthcoming wideband communication systems, and suggests that it is necessary to investigate transmission system dynamics when the PA's output is limited in bandwidth. In this case, DPD would ideally mitigate distortion in this limited frequency band (also called the ob-

servation band) while the rest of the damaging frequency content at the PA output could be taken care of by a carefully designed bandpass filter (feasibility and design of such bandpass filters is an important topic in itself but we will not attempt to tackle it in this thesis). Band-limited baseband models and the corresponding DPD were investigated in [77], and promising experimental results were shown. [77] follows the conventional baseband modeling approach, i.e., dynamic deviation reduction-based Volterra series modeling was used. However, restricting the observation bandwidth to be much smaller than the full bandwidth of the PA output introduces long (possibly infinite) memory dynamic behavior into the equivalent model. This implies that traditional modeling methods, such as memory polynomials or Volterra series modeling, might be too general to efficiently describe this new structure, and might not be well suited for practical implementations. Namely, long memory of a nonlinear model would require exponentially large number of coefficients to be implemented, which might be too expensive for hardware realization.

1.1.3 All-Digital Transmitters and Pulse-Width Modulation

In the previous sections, we saw how efficiency of the PA can be increased by an appropriate change in the PA driving signal. In this section, we discuss how to increase efficiency by changing the PA architecture.

A promising alternative to power inefficient traditional transmitter architectures that use conventional PAs (e.g., class A, B or AB [78]), are the so called *all-digital transmitters* (ADT) [79]. The RF power amplifiers in ADT architectures work in switch-mode operation (SMPA), like in, e.g., class-S mode [80, 81, 82]. Transistors in SMPA's are driven as switches, in an ON/OFF regime, and because of this specific operation principle, SMPA's can achieve drain efficiency of over 90%, while maintaining good linearity at the switching stage [83]. Namely, at every moment, either the output current or the output voltage of each transistor is equal to zero and, therefore, their product, i.e. the wasted power at the drain, is theoretically always equal to zero (in practice, due to finite switching times, the wasted power is non-zero but can still be negligible). Therefore, ADT's can achieve much higher power efficiency than traditional transmitters. However, the ON/OFF mode of operation

requires piece-wise constant SMPA driving signals (or RF signals with a piece-wise constant envelope like in the case of, e.g., burst-mode RF transmitters [84]). Such signals are commonly called *pulse trains* and are allowed to take only small number of discrete amplitude values (especially when the SMPA is part of a base station transceiver). Hence, the high resolution RF input signals, which drive PAs in traditional transmitter architectures, are not suitable as driving signals for SMPAs. Instead a reduction in amplitude resolution is needed at the PA input in order to adjust for the switch-mode regime. This operation of mapping of high resolution signals into pulse trains is commonly called *power encoding*, and is typically done in the form of delta-sigma modulation ($\Delta\Sigma$M) [85], pulse-width modulation (PWM) [86, 87, 88], or a combination thereof [89].

A block diagram of a typical ADT architecture is shown in Figure 1-4. In this example, power encoding is performed on in-phase and quadrature components of the baseband signal, and the encoded signal is up-converted to a carrier frequency by a digital modulator, before being amplified by the SMPA. In this case, the encoder system is usually called the *baseband power encoder* [87], as opposed to the *passband power encoder* which operates on an already up-converted signal [89]. Even though the electrical efficiency of the SMPA can be high (theoretically, close to 100%, and practically more than 90% [83]), the overall power efficiency of the transmitter, employing an SMPA, mainly depends on the utilized power encoding method. Namely, after an encoded pulse train is amplified by the SMPA, the passband output signal has to be reconstructed by a bandpass filter (Figure 1-4). However, the power encoded pulse trains contain significant amount of out-of-band spectral power (as a consequence of the quantization) which is dissipated at the filter, therefore reducing the overall power efficiency of the transmission system. For that reason, *coding efficiency (CE)* is introduced as a figure of merit to evaluate the encoder performance in terms of power efficiency. Coding efficiency of a power encoder is defined as a ratio of the desired in-band power to the total power of the encoder output. Another important performance metric of power encoders is linearity, which can be measured in terms of in-band signal-to-noise ratio (SNR) of the encoder output, or, in terms of the amount of error that is observed in the decoded constellation symbols at the receiver (i.e., EVM). We have seen previously that in the conventional transmitters the main source of signal distortion

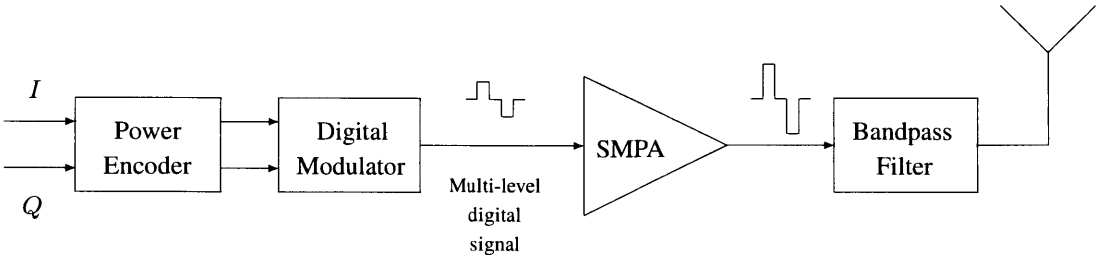


Figure 1-4: An example of all-digital transmitter architecture.

is the PA nonlinearity. However, in all-digital transmitters, power encoders are the major contributors of signal distortion. Therefore, the process of choosing a power encoder is a critical part of the all-digital transmitter design and the one that most significantly impacts its performance.

Pulse-width modulation (PWM) has been one of the most popular power encoding methods in microwave applications. A PWM system maps an input signal into a digital pulse train, where amplitude information of the input signal is encoded into time-varying width of the output pulses [90]. In RF applications, it is usually applied on I/Q components of the baseband signal, or the baseband envelope in the case of burst-mode transmitters. PWM has been extensively studied and reported in literature, starting with the seminal paper by Bennett [91]. It regained interest with the work of Raab [92] on radio-frequency (RF) PWM. Many different variants of pulse-width modulation have been studied and practically utilized in the past (see, e.g., [93], and the references therein). These are categorized based on various characteristics: number of output levels of PWM (2-level vs. multi-level), mode of comparison that is used to generated the output signal (reference waveform vs. constant signal, i.e., baseband vs. radio-frequency PWM), choice of reference signals (e.g., sinusoidal vs. sawtooth), the way the PWM input signal is sampled (trailing vs. double edge, natural vs. uniform), etc. The most notable characterization is based on time-domain in which the PWM is implemented: *analog PWM (APWM)* which is understood as a system mapping continuous-time input signals to continuous-time output signals, and *digital PWM (DPWM)*, which is, analogously, defined as mapping discrete-time inputs to discrete-time outputs. It should be noted that the term *digital PWM* is commonly used in power electron-

ics and digital audio applications to denote a system with DT input signal and CT output signal, so that it mimics analog PWM operation [94, 95, 96]. This term will not be used in this thesis (see Chapter 4 for further explanations).

Pulse-width modulator is clearly a nonlinear system and the main performance measure of PWM schemes, regardless of the variant and the choice of parameters, is the amount of spectral distortion (also called *harmonic noise*) that the input signal is subject to when modulated by PWM. Traditionally, harmonic noise analysis of analog PWM schemes has been done by simulation studies, followed by experiments on a practical implementation of PWM [97]. Comparing performance of different PWM schemes then becomes increasingly difficult, since it is hard to separate impact of the modulation scheme itself from second order effects of the hardware implementation. It is now clear that having an analytical description of the spectral distortion generated by the PWM modulation process would help in determining the best possible performance that can be achieved by a particular modulation scheme. In other words, it would help in setting a theoretical upper bounds on achievable performance for practical implementations.

Various exact (though very complicated) analytical expressions for spectral components of multi-level APWM were reported in [93, 97, 98, 99]. In these papers, PWM is studied from power electronics perspective, and hence only sinusoidal inputs are considered. Detailed analysis, both in time and frequency-domain, for various 2-level PWM schemes with arbitrary band-limited input signals was first reported in [100], and closed form expressions for the pulse train output of APWM have been derived. Similar expressions were reported in [82, 101], and additionally effects of input amplitude quantization and time sampling were briefly discussed. A compact input-output model for multi-level APWM, with arbitrary bounded input, was reported in [87, 102], and then further generalized in [103]. It has been shown that the APWM output signal has a quasi-harmonic structure with significant spectral content only at the integer multiples of the PWM frequency , when the input is band-limited [90]. These results imply that the in-band harmonic noise in APWM is negligible when the input signal bandwidth to the PWM frequency ratio is small.

Although many theoretical analysis results of different APWM schemes can be found in the literature, there are just a few reported results for its digital counterpart, even though the

interest in DPWM has significantly increased in the last decade [81]. Namely, the forthcoming communication standards envision transceivers implemented as software-defined radio in order to benefit from the reconfigurability and signal processing power that are available in digital domain. This naturally calls for the use of PWM fully implemented in discrete-time domain, that is, DPWM. As will be shown in detail in Chapter 4, DPWM can be easily described as just a time-sampled version of APWM. However, the APWM output signal has infinite bandwidth, which, when sampled, results in substantial spectral domain aliasing of the DPWM output signal. This introduces considerable amount of in-band distortion which is commonly referred to as the *DPWM aliasing noise* or just the *aliasing noise*. While not present in the APWM case, this additional harmonic distortion represents the main drawback of DPWM for practical utilization. An equivalent way of describing this harmonic noise can be developed by looking at the time-domain waveforms generated by DPWM. Namely, signal generation in APWM corresponds to the input signal being "sampled" per every PWM reference period. This signal value is encoded into the width of the corresponding pulse in the APWM output signal. However, in DPWM, finite-time resolution, which is a result of sampling, causes switching instants of the DPWM output pulse train to be rounded to the closest sample point. In addition, the output pulse can last only a finite number of sample periods resulting de facto in quantization of the input signal amplitude. This inherent quantization phenomenon has been reported for 2-level trailing edge PWM as early as 1983 (where it was called the *PWM hardware resolution*) [104].

The excessive aliasing noise in DPWM can be mitigated by increasing the number of achievable output levels of the system. Unfortunately, this number is usually limited to single digits by the complexity of hardware realization (especially in the case of high power PAs used in base stations) [105]. Another way to mitigate the aliasing distortion is by increasing the sampling rate of DPWM. In mobile communications, the underlying carrier frequency, denoted as f_{RF}, is usually on the GHz range. In order to achieve reasonable linearity performance in the system utilizing DPWM, very high sampling rates are needed ($50 \times f_{RF}$), which is clearly infeasible for practical purposes [89].

DPWM in-band harmonic distortion, and the spectral aliasing effects causing it, have been studied by simulations and reported in, e.g., [106, 107]. Various solutions have been

proposed in literature to mitigate this noise, with majority of those relying on some sort of spectral noise shaping, e.g., by delta-sigma modulation. In [108], a DPWM noise shaping through an additional $\Delta\Sigma$M type of system has been proposed. Similarly, in [109], reduction of aliasing effects in radio-frequency (RF) DPWM by employing feedback loop with $\Delta\Sigma$M, has been reported. In [87], authors propose to cancel distortion by realizing DPWM in a way which is analogous to applying an anti-aliasing low-pass filter to the APWM output before it is sampled, and hence the name: *aliasing-free digital PWM*. Filtering of the APWM output is done by truncating its double Fourier series, causing no spectral overlap in the band of interest after sampling is performed. Due to the Gibbs effect [110], the filtered output signal has amplitude ripples around the points of discontinuity, and hence the SMPA driving input is not a pulse train anymore. This introduces additional nonlinear distortion and decreases drain efficiency of the SMPA. In [111], $\Delta\Sigma$M is used to quantize amplitude of a RF-DPWM input signal and limit minimal DPWM duty-cycle, in order to prevent "pulse swallowing", which is a common issue observed in SMPAs (due to non-zero raising/falling times of clock signals, extremely short pulses get "swallowed" by hardware components) [82]. In [89], a power encoder based on a cascade of $\Delta\Sigma$M and many RF-DPWM blocks was used. The $\Delta\Sigma$M output signal of high resolution is outphased by *multidimensional power coding* scheme using many low resolution RF-DPWM blocks. Similarly, in [88], $\Delta\Sigma$M output of high resolution is processed by a mixed-domain FIR filter employing many low level DPWM blocks, having effects close to work in [89]. Similar mixed-domain filters were used in [112].

In the above references, various schemes for canceling in-band distortion caused by spectral aliasing in DPWM have been proposed. Typically, authors state a well known fact from sampling theory that when signal of infinite bandwidth is sampled aliasing will occur [110], and proceed to give general formulae for the spectra of signals involved in such an operation. However, none of the results attempt to utilize the time-domain characterization of the spectral aliasing: that it is a result of an underlying quantization that the input is subject to through finite resolution of the DPWM output pulse widths. This characterization, as it turns out, is crucial if noise-shaping techniques, like $\Delta\Sigma$M, are to be used to mitigate the distortion. It is, therefore, clear that having a good description of this quantization pro-

cess, i.e., having a detailed input/output model of DPWM, should help in better mitigating the aliasing noise.

1.2 Contributions of the Thesis

While the specific contributions of this thesis lie in addressing the particular problems outlined in sections 1.1.1-1.1.3., the general theme of the thesis is to provide a way to think about optimal digital signal processing algorithm design using the language of systems theory. This offers a different perspective on digital signal processing questions involving nonlinear systems and enables the use of sophisticated, and well developed, tools from systems theory to address seemingly 'hard' problems in this area.

In the following subsections, we give a detailed description of the contributions of this thesis, with each subsection pertaining to a specific way of mitigating the power efficiency problem, as described in sections 1.1.1-1.1.3. The last subsection provides a summary of the contributions in an itemized manner.

1.2.1 On-line Least Squares Optimization with Convex Sample-Wise Constraints and its Application to PAPR Reduction

In Chapter 2, we consider the task of minimizing PAPR of a baseband communication signal (i.e., a discrete-time signal with the significant low-frequency spectral content) by designing discrete-time systems which are optimal in frequency-weighted least squares sense subject to a maximal output amplitude constraint. The cost function is a linear combination of two terms, penalizing deviation from ACLR and EVM constraints, respectively. The PAPR is controlled by reducing value of the maximal output amplitude parameter. It is known for such problems that, in general, the optimality conditions do not provide an explicit way of generating the optimal output as a real-time implementable transformation of the input. We use tools from robust control to study these problems and prove that the optimal system has exponentially fading memory, which suggests existence of arbitrarily good receding horizon (i.e., finite latency) approximations. We show that, with adequate non-

linear stability analysis and careful structuring, the optimal system can be approximately realized in the form of a finite latency real-time signal processing algorithm. We propose a real-time realizable algorithm which, under an L1 dominance assumption about the equation coefficients, returns approximations to the optimal map of arbitrary accuracy , where the approximation quality is measured in terms of the L-infinity norm of the error signal. The algorithm exploits the optimality conditions and is realized as a causally stable nonlinear discrete-time system, which is allowed to look ahead at the input signal over a finite horizon (and is, therefore, of finite latency). Furthermore, we propose an extension of the well-known method of balanced truncation for linear systems to the class of nonlinear models with weakly contractive operators. This result is then used to derive a causally stable finite-latency nonlinear system which also returns approximations of arbitrary accuracy to the optimal map, where the approximation quality is now measured in terms of the L2 gain of the error system. In this case, the algorithm does not depend on any special assumptions about the optimization parameters (e.g., no need for the L1 dominance assumption).

The fading memory result and the L-infinity version of the approximate algorithm were published in [113]. The L2 version of the algorithm, as well as the generalized balanced truncation result, were published in [114].

1.2.2 Equivalent Baseband Modeling and Digital Compensation of Dynamic Passband Nonlinearities in Phase-Amplitude Modulation-Demodulation Schemes

In this thesis, we consider equivalent baseband representation of transmission circuits, in the form of a nonlinear dynamical system **S** in discrete time (DT) defined by a series interconnection of a phase-amplitude modulator, a nonlinear dynamical system **F** in continuous time (CT), and an ideal demodulator. We show that when **F** is a CT Volterra series model with fixed degree and memory depth, the resulting **S** is a series interconnection of a DT Volterra series model of same degree and equivalent memory depth, and a long memory discrete-time LTI system which can be viewed as a bank of *reconstruction filters*. This equivalent baseband model reveals that order of the nonlinearity and, more importantly,

memory of the underlying nonlinear system are preserved when passing from the passband to the baseband. Frequency responses of the above reconstruction filters exhibit discontinuities at frequency values $\pm\pi$, making their unit sample responses infinitely long. This discontinuity is mainly due to the lack of symmetry of the frequency response of the reconstruction filter with respect to the carrier frequency. Nevertheless, the frequency responses are shown to be smooth inside the interval $(-\pi, \pi)$, and thus approximable by low order FIR filters on compact subsets of $(-\pi, \pi)$. Length (i.e, number of taps) of these approximate FIR filters depends on the ratio of the signal to observation bandwidth (i.e., the amount of the observation bandwidth which is occupied by the useful signal). Relatively low memory/degree requirements of the nonlinear (Volterra) subsystem, as well as good approximability by FIR filters of the linear subsystem, allow for potentially efficient hardware implementation of the corresponding baseband model. The result suggests a new, non-obvious,analytically motivated structure of digital pre-compensation of analog nonlinear distortions such as those caused by power amplifiers in digital communication systems.

The two-box decomposition of the equivalent baseband model that we propose in this thesis is not motivated nor does it rely on decoupling strong from weak nonlinear effects. It rather relies on decoupling short memory nonlinear effects from long memory linear effects, and is therefore fundamentally different from those in [75, 76] or [115, 116]. Seemingly similar equivalent baseband model structures for discrete-time Volterra systems were derived in [117, 118, 119]. The work presented here differs from [117, 118] in that model dynamics are investigated over a limited bandwidth of the PA output, which, as described earlier, leads to significantly different dynamic behavior of the underlying system, and therefore to a distinct baseband equivalent model. Likewise, derivation of the model presented in [119] relies on the assumption that the PA input can only take amplitude values of +1 and -1, and leads to a somewhat restricted model, while in our work, there are no constraints on the RF signal driving PA, except that its envelope is piecewise constant. It should be noted that modeling and linearization of PAs using band-limited observation path was considered in [115, 116, 120], but the corresponding baseband equivalent models and DPDs were derived for the full-band PA output, and are therefore fundamentally different from the band-limited models investigated in this thesis.

The results on general equivalent baseband modeling were initially published in [121], and then extended in [122]. The results on equivalent baseband modeling of all-digital transmitters were initially reported in [123]. A patent describing hardware-efficient compensator architecture, theoretically based on the model presented in this thesis, has been issued [124].

1.2.3 Approximate Baseband Modeling and Digital Compensation of Digitally Implemented Pulse-Width Modulation (DPWM)

As mentioned in the previous sections, analysis of aliasing distortion in digital PWM has been mostly carried in spectral domain. In nonlinear systems analysis, it is usually more advantageous to look at time-domain rather than frequency-domain changes in order to get insight into system behavior. In this thesis, we follow that approach and derive compact closed-form time-domain model of both multi-level carrier-based and multi-level radio-frequency DPWM. The model is derived under assumptions of square summable excitation and arbitrary admissible reference signals (for carrier-based DPWM) and arbitrary threshold levels (for RF-DPWM). This model offers a complete description of the inherent quantization that the DPWM input is subject to; and which is an equivalent way of representing the aliasing harmonic noise. We call this inherent quantization process the *hidden quantization* in order to differentiate it from the quantization operation of the DPWM itself. This result reveals that, with sufficiently high DPWM sampling rate, the main cause of in-band harmonic distortion in DPWM is the hidden quantization. The presented analysis significantly improves understanding of the leading in-band distortion source in digitally implemented PWM, and represents a fundamental step in understanding behavior and limits of all-digital transmitter architectures employing DPWM.

The above result implies that aliasing-free DPWM is possible if and only if the input signal, before being fed into DPWM, is quantized exactly as defined by the hidden quantization. We exploit this specific time-domain relationship between the input and output signals of DPWM, and propose a delta-sigma modulator based pre-distortion of DPWM. When cascaded, $\Delta\Sigma$M and DPWM form a power encoding system that achieves linear-

ity of a delta-sigma modulator with high number of output levels using a SMPA driving signal of (relatively) low resolution (i.e. small number of DPWM output levels). Given a fixed input signal class (with i.i.d. time-samples according to some known distribution), the DPWM output levels are selected to minimize the mean squared error of the hidden quantization noise. Then the $\Delta\Sigma$M parameters are chosen so to match those of the optimal hidden quantization. Due to a mismatch between the optimized and actual hidden quantization levels in non-uniform DPWM, a compensation signal is generated and subtracted from the DPWM output, ensuring minimal harmonic noise in the output signal. In addition to providing high linearity, this power encoder scheme considerably reduces design requirements for the SMPA (which can be significant in case of GaN based architectures [105]). It also potentially offers much lower hardware complexity than the state of the art e.g. [89]-[88], which use high number of SMPAs, while maintaining similar level of performance (both in terms of linearity and coding efficiency). We show, by Matlab simulations, that, by the above co-design of $\Delta\Sigma$M and DPWM, it is possible to increase power efficiency of the DPWM encoded LTE test signals by, roughly, 15% to 20%, in comparison to that of the unoptimized (i.e., uniform) $\Delta\Sigma$M-DPWM encoding schemes. This is achieved while preserving similar in-band SNR quality.

The initial results on modeling of DPWM were published in [125] (baseband DPWM) and [126] (RF-DPWM). The results on optimal noise cancellation and corresponding ADT architectures were published in [127, 128, 129, 130] and a patent has been issued [131].

1.2.4 Itemized List of Contributions

In the previous subsections, the main contributions of the thesis were described in more detail. In the following, we summarize these contributions and present them in an itemized fashion.

Part I:

- A system for peak-to-average power ratio reduction of discrete-time signals is modeled as a time-invariant discrete-time system which is the optimal map of an infinite-

dimensional convex least-squares problem with box constraints.

- A characterization of the memory of such optimal maps (when viewed as dynamical systems) is given for the case of general sample-wise constraints (i.e., not necessarily the box constraints);

- An extension of the classical method of balanced truncation for linear systems to the class of nonlinear models with weakly contractive operators is presented;

- Two real-time finite-latency algorithms which return approximations of arbitrary accuracy to the optimal map of the above mentioned optimization problem were proposed. In one case, the approximation quality is measured in terms of the supremum norm of the error signal (subject to some L1-dominance assumption) and, in the other case, in terms of the L2 gain of the error system (under no additional assumptions on the problem);

Part II:

- An exact equivalent baseband model of transmission systems in the form of a nonlinear dynamical system $\mathbf{S}$ in discrete time (DT) defined by a series interconnection of a phase-amplitude modulator, a nonlinear dynamical system $\mathbf{F}$ in continuous time (CT), and an ideal demodulator was derived. It was show that when $\mathbf{F}$ is a CT Volterra series model, the resulting $\mathbf{S}$ is a series interconnection of a DT Volterra series model of the same degree and equivalent memory depth, and an LTI system with special properties;

- Based on the derived model, a novel, analytically motivated, structure of a digital predistorter for nonlinear RF power amplifiers was proposed. The equivalent baseband model was validated and the effectiveness of the proposed DPD was demonstrated by Matlab simulations with several common models of PA nonlinearity;

Part III:

- An exact closed-form input-output time-domain model of both carrier-based and threshold based digital pulse-width modulation (DPWM) systems is given. The

model is derived under assumptions of square summable excitation and arbitrary DPWM threshold levels, or arbitrary admissible DPWM reference signals in the case of carrier-based DPWM. The model gives a complete description of the inherent quantization that the DPWM input is subject to;

- A framework for design of optimal power encoding systems, realized as a series interconnection of a delta-sigma modulator and a digital pulse-width modulator, is proposed. Given a fixed input signal class (with i.i.d. time-samples according to some known distribution), the parameters of the $\Delta\Sigma$M-DPWM encoder are chosen so to minimize the mean squared error of the inherent DPWM quantization noise. Matlab simulations were carried to demonstrate the effectiveness of the proposed system in encoding several standardized LTE test signals;

1.3 Thesis Outline

The rest of the thesis is organized as follows. In Chapter 2, we formulate the problem of peak-to-average ratio reduction as an abstract convex optimization problem, study properties of its optimal map and propose several approximate algorithms to solve it. Equivalent baseband modeling of wireless transmission systems is considered in Chapter 3, and an efficient digital pre-distortion model is proposed. In Chapter 4, we derive a novel time-domain model of digital pulse-width modulation and propose several efficient and highly linear power encoding schemes that combine delta-sigma and digital pulse-width modulation.

1.4 Notation and Terminology

- $\mathbb{C}$, $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{Z}_+$ and $\mathbb{N}$ are the standard sets of complex, real, integer, non-negative integer and positive integer numbers. j is a fixed square root of -1.
- For $r > 0$, disk $\mathbb{D}_r \subset \mathbb{C}$ is defined as $\mathbb{D}_r = \{z \in \mathbb{C} : \ |z| \leq r\}$.
- $[1 : n]$, for $n \in \mathbb{N}$, is the set $\{1, \ldots, n\}$.

- X^d, for a set X, is the set of all d-tuples $(x_1, \ldots, x_d)$ with $x_i \in X$.
- For a finite set S, $|S|$ denotes the number of elements in S.
- For a set X, X-valued continuous-time (CT) signals and X-valued discrete-time (DT) signals are functions $\mathbb{R} \to X$ and $\mathbb{Z} \to X$, respectively.
- $\ell^2(X)$ and $\mathcal{L}^2(X)$, for a set X, are the sets of all X-valued square summable and X-valued uniformly bounded square integrable signals, respectively.
- For $w \in \ell^2(X)$, $w[n] \in X$ denotes the value of w at $n \in \mathbb{Z}$. In contrast, $v(t) \in X$ refers to the value of $v \in \mathcal{L}^2(X)$ at $t \in \mathbb{R}$.
- The Fourier transform $\mathcal{F}$ applies to both CT and DT signals. For $v \in \mathcal{L}^2(X)$, its Fourier transform $V = \mathcal{F}v$ is a square integrable function $V = V(\omega) : \mathbb{R} \to \mathbb{C}$. For $w \in \ell^2(X)$, the Fourier transform $W = \mathcal{F}w$ is a 2π-periodic function $W = W(\Omega) : \mathbb{R} \to \mathbb{C}$, square integrable on its period. Therefore, in this thesis, we use shorthand notation $X = X(\omega)$ (respectively $X = X(\Omega)$) to denote the Fourier transform of signal $x \in \mathcal{L}^2(\mathbb{R})$ or $x \in \mathcal{L}^2(\mathbb{C})$ (respectively $x \in \ell^2(\mathbb{R})$ or $x \in \ell^2(\mathbb{C})$), instead of a standard notation $X = X(j\omega)$ (respectively $X = X(e^{j\Omega})$).
- *Systems* are viewed as functions $\mathcal{L}^2(X) \to \mathcal{L}^2(Y)$, $\ell^2(X) \to \mathcal{L}^2(Y)$, $\mathcal{L}^2(X) \to \ell^2(Y)$, or $\ell^2(X) \to \ell^2(Y)$.
- $\mathbf{G}f$ denotes the response of system $\mathbf{G}$ to signal f (even when $\mathbf{G}$ is not linear), and the *series composition* $\mathbf{K} = \mathbf{QG}$ of systems $\mathbf{Q}$ and $\mathbf{G}$ is the system mapping f to $\mathbf{Q}(\mathbf{G}f)$.
- A system $\mathbf{G} : \mathcal{L}^2(X) \to \mathcal{L}^2(X)$ (or $\mathbf{G} : \ell^2(X) \to \ell^2(X)$) is said to be *linear and time invariant* (LTI) with *frequency response* $H : \mathbb{R} \to \mathbb{C}$ when $\mathcal{F}\mathbf{G}x = H \cdot \mathcal{F}x$ for all $x \in \mathcal{L}^(X)$ (respectively $x \in \ell^2(X)$). When $\mathbf{G} : \ell^2(X) \to \ell^2(X)$, its frequency response H is 2π-periodic and signal $h \in \ell^2(X)$ such that $H = \mathcal{F}h$ is called the unit sample response of $\mathbf{G}$.

- For $v, w \in \ell^2(\mathbb{C})$, the scalar product $(v, w) = w'v \in \mathbb{C}$ is defined by $w'v = \sum_n \overline{w[n]}v[n]$, where $\overline{w[n]}$ is the complex conjugate of $w[n]$.
- For $v \in \ell^2(\mathbb{C})$, the *L2 norm* $\|v\|_2 \in [0, \infty]$ is defined by $\|v\|_2^2 = v'v = \sum_n |v[n]|^2$.
- $E = \{e_i\}_{i=-\infty}^{\infty} \subset \ell^2(\mathbb{R})$ such that $e_i(t) = 1$ for $t = i$ and $e_i(t) = 0$ otherwise, is the standard orthonormal basis in $\ell^2(\mathbb{R})$.
- For a bounded linear operator $\mathbf{A} : \ell^2(X) \to \ell^2(Y)$, $\mathbf{A}' : \ell^2(Y) \to \ell^2(X)$ denotes the adjoint operator of $\mathbf{A}$.
- The matrix of $\mathbf{A}$, in the standard bases $\{e_i\}_{i=-\infty}^{\infty}$ and $\{\tilde{e}_i\}_{i=-\infty}^{\infty}$ of $\ell^2(X)$ and $\ell^2(Y)$, respectively, is denoted as $A = (A_{ij})_{i,j=-\infty}^{\infty}$.
- In this paper, $\mathbf{A}$ and A will be used interchangeably to denote the same operator.
- For any bounded (not necessarily linear) operator $\mathbf{T} : \ell^2(X) \to \ell^2(Y)$, the operator norm $\|\mathbf{T}\|$ of $\mathbf{T}$ is defined as $\|\mathbf{T}\| = \sup_{x \in \ell^2(X), x \neq 0} |\mathbf{T}x|/|x|$.
- For a positive real number r, function $\mathrm{sat}_r : \mathbb{C} \to \mathbb{D}_r$ is defined by

$$\mathrm{sat}_r(\xi) = \begin{cases} \xi, & |\xi| \leq r \\ r\xi/|\xi|, & |\xi| > r \end{cases}.$$

  Similarly, operator $\mathrm{Sat}_r : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ is defined by

$$y = \mathrm{Sat}_r(x) \Leftrightarrow y[n] = \mathrm{sat}_r(x[n]), \forall n \in \mathbb{Z}.$$

- $\lfloor \cdot \rfloor : \mathbb{R} \to \mathbb{Z}$ is the floor function, i.e., $\lfloor \xi \rfloor$ denotes the largest integer less than or equal to ξ.

Chapter 2

On-line Least Squares Optimization with Convex Sample-Wise Constraints and its Application to PAPR Reduction

This chapter is organized as follows. In section 2.1, we formulate the problem of optimal PAPR reduction system design and then introduce the general description of the underlying optimization problem. Main results are given in section 2.2. We first characterize the memory of the optimal map of the optimization problem under consideration, and then propose two algorithms for calculating arbitrarily accurate approximate solution to the optimal problem. In section 2.3, we present some numerical simulation results which verify the effectiveness of the proposed PAPR reduction algorithm.

2.1 Problem Formulation

We first introduce some preliminary notation and definitions in section 2.1.1 and then give a precise formulation of the framework for designing optimal PAPR reduction systems in section 2.1.2. A detailed analysis of the corresponding abstract optimization problem is given in section 2.1.3.

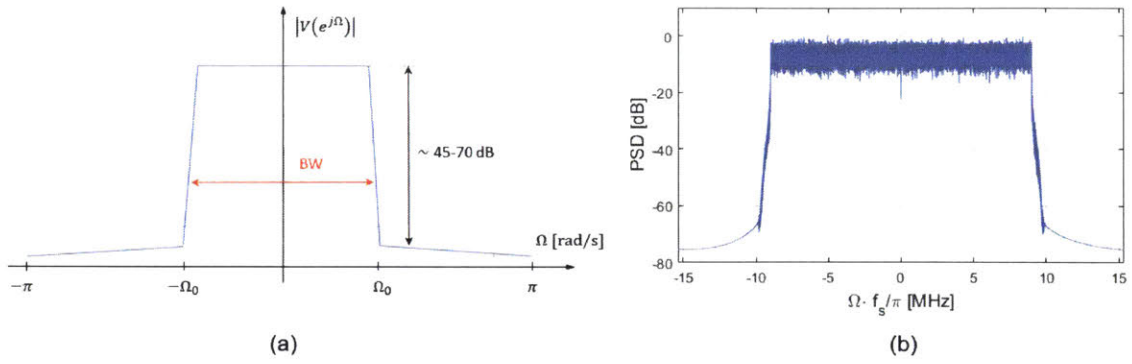


Figure 2-1: Power spectrum of a baseband signal: (a) a model spectral profile of a baseband signal considered in this thesis, (b) spectral profile of a typical LTE transmit signal with 20MHz-bandwidth at the rate of $f_s = 30.72$ mega samples per second (MS/s).

2.1.1 Preliminaries

In the area of mobile communications, it can often be ambiguous as to what is meant by a *baseband signal*, depending on the actual transmission system under consideration (e.g., baseband signals for mobile stations, WLAN, concurrent multi-band signal transmission, etc.). In this thesis, for simplicity, by a baseband signal we mean a discrete-time scalar-valued square summable signal which has 'significant' spectral content for low frequencies and 'negligible' spectral content for high frequencies. An approximate spectral profile of such a signal v is shown in Figure 2-1(a). We denote with $I = (-\Omega_0, \Omega_0)$ the smallest symmetric subinterval of $(-\pi, \pi)$ which contains all the frequencies at which v has significant power, where $\Omega_0 \in (0, \pi)$. In that case, $BW = 2\Omega_0$ is called the bandwidth of v. For baseband signals defined in these terms, the difference in average signal power between low and high frequencies is typically 45-70 dB, and heavily depends on the actual application. Magnitude spectrum of a 3GPP standard-compliant LTE signal, generated by Matlab's LTE Toolbox, is shown in Figure 2-1(b). As can be seen, the spectrum in Figure 2-1(a) is a good approximation of that in Figure 2-1(b) (modulo the bandwidth differences).

Now let $\mathbf{S} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be an arbitrary stable system mapping baseband input signal v to output signal y. The error vector magnitude (EVM) and adjacent channel leakage

ratio (ACLR) distortion measures for system S are defined as:

$$EVM(v,y) = \frac{\|V - Y\|_2}{\|V\|_2} \cdot 100 \quad [\%], \tag{2.1}$$

$$ACLR_{\Omega_0}(y) = 20\log_{10}\left(\frac{\|H_{id}Y\|_2}{\|(1 - H_{id})Y\|_2}\right) \quad [dB], \tag{2.2}$$

where $BW = 2\Omega_0$ is the bandwidth of v, V and Y are the Fourier transforms of v and y, respectively, and $H_{id} = H_{id}(\Omega)$ is the Fourier transform of an ideal low-pass filter with the critical frequency Ω_0, that is $H_{id}(\Omega) = 1$ for $|\Omega| \leq \Omega_0$ and $H_{id}(\Omega) = 0$ for $|\Omega| \in (\Omega_0, \pi)$. It is clear from (2.1) that the error vector magnitude measures the relative time-domain distortion introduced by S. The adjacent channel leakage ratio measures the spectral out-of-band distortion (or 'spectral re-growth', as it is commonly called) by comparing the in-band and out-of-band spectral content in y. It follows from (2.2) that ACLR is a function of both y and Ω_0, but, from now on, when denoting ACLR of y we will drop the subscript and write just $ACLR(y)$ assuming that it is clear from the context which Ω_0 is meant.
The peak-to-average power ratio (PAPR) of a discrete-time signal v is defined as:

$$PAPR(v) = 20\log_{10}\left(\frac{\max_n |v[n]|}{RMS(v)}\right) \quad [dB], \tag{2.3}$$

where $RMS(v)$ is the root-mean square value of v defined by

$$RMS(v)^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{n=-N}^{N} |v[n]|^2.$$

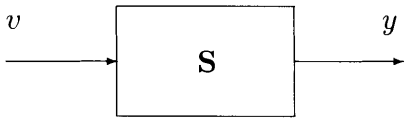
It should be noted here that the definitions of EVM, ALCR, and PAPR from (2.1),(2.2), and (2.3), are simplifications of the definitions introduced by the wireless communications 3GPP standard [132]. Namely, according to the latter definitions, the above distortion and signal quality measures are very complicated mathematical functions, which are often easy to practically measure but lead to very hard (even infeasible) theoretical analysis of the underlying systems. Whereas, the simplified definitions are easy to work with, capture the main characteristics of the above measures and, in most common setups, provide similar values as the standard ones.

For example, according to [132], calculating ACLR is meaningful only if we consider signals with $\Omega_0 \in (\pi/6, \pi/3)$ or, even more restricted, when $\Omega_0 \leq \pi/6$. In the latter case, [132] defines two ACLR measures, ACLR1 and ACLR2, each comparing the energy of the information bearing spectrum for $\Omega \in (-\Omega_0, \Omega_0)$ with that of the adjacent channels for $|\Omega| \in (\Omega_0, 2\Omega_0)$ and $|\Omega| \in (2\Omega_0, 3\Omega_0)$, respectively (when $\Omega_0 \in (\pi/6, \pi/3)$, only ACLR1 is considered).

2.1.2 Optimal Peak-to-Average Power Ratio Reduction

Before we start discussing our approach to reducing PAPR, let us observe that any transformation that maps DT signal into another DT signal with (non-negligibly) lower PAPR would, in most cases, lead to a change in EVM and ACLR (depending on the nature of this transformation, they can either increase or decrease). Therefore, any meaningful framework for designing algorithms for peak-to-average power ratio reduction should take this into account and include a trade-off between PAPR, ACLR, and EVM.

In this thesis, we aim to optimize and find efficient implementation of discrete-time signal processing systems with baseband input signal v and output signal y:



where the output $y = \mathbf{S}v$ is expected to be optimal, in the sense of minimizing a certain objective defined in terms of input v. We want to design S such that the following is true: $PAPR(y) < PAPR(v)$, the useful information in v is preserved (equivalent to $EVM(v, y)$ being small), and that there is no significant spectral re-growth at the output of S (equivalent to $ACLR(y)$ being small). We observe that limiting the maximal output amplitude of S to be smaller than $\max_n |v[n]|$ can potentially enforce decrease in PAPR of the output of S.

Let $r > 0$ be the maximal allowable output amplitude of system S which is pre-specified in accordance with the expected class of input signals. In order to make sure that S does

not just 'scale' down the input signal v (and therefore result in $PAPR(y) \approx PAPR(v)$) we would like to achieve $\|y\|_2 \approx \|v\|_2$, or equivalently, $\|y - v\|_2 \approx 0$.

Let $\Omega_0 \in (0, \pi)$, and let $\mathbf{L} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be an ideal high-pass filter whose frequency response $L = L(\Omega)$ satisfies $L(\Omega) = 1$ for $|\Omega| \in (\Omega_0, \pi)$, and $L(\Omega) = 0$ for $\Omega \in (-\Omega, \Omega_0)$. We observe that for every $x \in \ell^2(\mathbb{C})$, with Fourier transform $X = X(\Omega)$, the integral

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} L(\Omega)|X(\Omega)|^2 d\Omega$$

measures the spectral content of x in frequencies $|\Omega| \in (\Omega_0, \pi)$. It is clear that if x was a baseband signal with bandwidth $2\Omega_0$, the value of the above integral would be small relative to the energy of x. Likewise, if the above integral was small compared to the energy of x, then $ACLR(x)$ would be large.

Let $\gamma > 0$ and let functional $J_{L,\gamma,v} : \ell^2(\mathbb{C}) \to \mathbb{R}$ be defined by

$$J_{L,\gamma,v}(y) = \frac{1}{2\pi} \int_{-\pi}^{\pi} L(\Omega)|Y(\Omega)|^2 d\Omega + \frac{\gamma}{2\pi} \int_{-\pi}^{\pi} |Y(\Omega) - V(\Omega)|^2 d\Omega, \tag{2.4}$$

where $V = V(\Omega)$ and $Y = Y(\Omega)$ are the Fourier transforms of v and y, respectively. It is clear from the above discussion, that the two summands in $J_{L,\gamma,v}$ are the measures of ACLR and EVM of the output signal y, and γ is a parameter which controls the trade-off between the two.

The problem of designing a discrete-time system which reduces peak-to-average power ratio of an information bearing baseband signal v, of bandwidth equal to $2\Omega_0$, can now be stated as follows:

For every discrete-time signal $v \in \ell^2(\mathbb{C})$, the scalar signal $y = \mathbf{S}v \in \ell^2(\mathbb{C})$ should have samples $|y[n]| \leq r$, and minimize the quadratic functional $J_{L,\gamma,v}(y)$.

This can be formulated as the following optimization problem:

$$\min_y J_{L,\gamma,v}(y), \quad \text{subject to } |y[n]| \leq r, \ \text{ for all } n \in \mathbb{Z}. \tag{$\mathbb{P}_0$}$$

It should be noted that the ideal filter L in (2.4) can be replaced by some finite impulse

response (FIR) approximation $\tilde{L}$. In that case, we have to ensure that the approximate filter satisfies $\tilde{L}(\Omega) + \gamma > 0$, for all $\Omega \in [0, 2\pi)$, in order to preserve the convexity of $J_{L,\gamma,v}$.

2.1.3 Abstract Optimization Setup

The optimization problem presented in the previous chapter falls into the following general format.

Problem $\mathbb{P} = \mathbb{P}(\alpha, \beta, r, v)$: *given trigonometric polynomials* $\alpha : \mathbb{R} \to \mathbb{C}$ *and* $\beta : \mathbb{R} \to \mathbb{C}$, *such that* $\alpha(\Omega) \geq \epsilon > 0$ *for all* $\Omega \in \mathbb{R}$, *vector* $v \in \ell^2(\mathbb{C})$, *and a real number* $r > 0$, *minimize the functional* $J_{\alpha,\beta,v} : \ell^2(\mathbb{C}) \to \mathbb{R}$ *defined by*

$$J_{\alpha,\beta,v}(y) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \alpha(\Omega)|Y(\Omega)|^2 d\Omega - \frac{1}{\pi} \int_{-\pi}^{\pi} Re\{Y(\Omega)'\beta(\Omega)V(\Omega)\} d\Omega, \tag{2.5}$$

on the set

$$B_r = \{y \in \ell^2(\mathbb{C}) : |y[n]| \leq r, \ \forall n \in \mathbb{Z}\},$$

where $V = V(\Omega)$ *and* $Y = Y(\Omega)$ *are the Fourier transforms of* v *and* y, *respectively.*

Here the word "minimize" means "find the arguments of minimum", that is, vectors $y^* \in B_r$ such that $J_{\alpha,\beta,v}(y^*) \leq J_{\alpha,\beta,v}(y)$ for all $y \in B_r$, whenever such y^* do exist. Therefore, we are trying to solve the time-domain-value-constrained frequency-weighted least-squares optimization problem

$$\min_y J_{\alpha,\beta,v}(y), \quad \text{subject to } y \in B_r. \tag{2.6}$$

The system-theoretic interpretation of the above optimization problem is as follows: for every discrete-time signal $v \in \ell^2(\mathbb{C})$, the scalar signal $y = \mathbf{S}v \in \ell^2(\mathbb{C})$ should have samples $|y[n]| \leq r$, and minimize the quadratic functional $J_{\alpha,\beta,v}(y)$. It is easy to see that the original PAPR reduction problem $\mathbb{P}_0$ is equivalent to $\mathbb{P}$ when $\alpha(\Omega) = L(\Omega) + \rho$ and $\beta(\Omega) = -\rho$, for all $\Omega \in \mathbb{R}$.

The cost function $J_{\alpha,\beta,v}$, as defined in (2.5), clearly has a frequency-domain interpretation. In order to notationally simplify things, let us re-write $J_{\alpha,\beta,v}$ in a more abstract operator form. For that reason, let $\mathbf{T}_\alpha : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ and $\mathbf{T}_\beta : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be the finite unit-sample response LTI systems with frequency responses $\alpha(\Omega)$ and $\beta(\Omega)$, respec-

tively. The cost function $J_{\alpha,\beta,v}$ can now be expressed as

$$J_{\alpha,\beta,v}(y) = y'\mathbf{T}_\alpha y - y'\mathbf{T}_\beta v - (\mathbf{T}_\beta v)'y, \tag{2.7}$$

where $\mathbf{T}_\alpha$ and $\mathbf{T}_\beta$ should be understood as operators on $\ell^2(\mathbb{C})$. From now on, for notational simplicity, we will drop the subscript in $J_{\alpha,\beta,v}$, and write just J to denote the cost function. Moreover, let $w = \mathbf{T}_\beta v \in \ell^2(\mathbb{C})$ and $\mathbf{H} = \mathbf{I} - \mathbf{T}_\alpha$, where $\mathbf{I}$ is the identity operator on $\ell^2(\mathbb{C})$. Let $H = H(\Omega)$ and $h = h[n]$ be the frequency response and unit sample response of $\mathbf{H}$, respectively, and let N be the order of the trigonometric polynomial $\alpha = \alpha(\Omega)$. It follows that $h[-n] = h[n]$ for all $n \in \mathbb{Z}$, and $h[n] = 0$ for all $|n| > N$. The cost function J can now be expressed in the time-domain as follows

$$J(y) = \sum_{n=-\infty}^{\infty} |y[n]|^2 - \sum_{n=-\infty}^{\infty} y[n]' \sum_{k=-N}^{N} h[k]y[n-k] - \sum_{n=-\infty}^{\infty} y[n]'w[n] - \sum_{n=-\infty}^{\infty} w[n]'y[n]. \tag{2.8}$$

As is clear from the above expression, the time-domain description of J is very involved and is hard to intuitively interpret, and is, therefore, given only for completeness.

It is clear that $\mathbb{P}$ is a convex infinite-dimensional quadratic problem with box constraints, which is feasible and has a unique solution, due to strict positivity of α [133]. This is formally stated in the following lemma which is given without a proof.

Lemma 2.1.1. *An argument of minimum of problem $\mathbb{P}$ exists and is unique.*

Let us denote the optimal solution of $\mathbb{P}$ as y^*, that is

$$y^* = \arg\min_{y \in B_r} J(y). \tag{2.9}$$

The necessary and sufficient conditions of optimality of $\mathbb{P}$ are given in the following lemma

Lemma 2.1.2. *For every $y^* \in B_r$ the following conditions are equivalent*

(a) y^* is an argument of minimum in problem $\mathbb{P}$.

(b) The following is true

$$y^* = \mathrm{Sat}_r(\mathbf{H}y^* + w), \tag{2.10}$$

which can be written in sample-wise form as

$$y^*[n] = \mathrm{sat}_r\left(\sum_{k=-N}^{N} h[k]y^*[n-k] + w[n]\right). \tag{2.11}$$

The result in the above lemma is a direct consequence of the following general result in convex optimization, which is the basis for the well known Goldstein-Levitin-Polyak gradient projection method [134, 135, 136]:

Theorem 2.1.3. *Let $\mathcal{H}$ be a Hilbert space and $K \subset \mathcal{H}$ closed and convex. Let $P_K : \mathcal{H} \to K$ be the projection operator for K. Let f be a real-valued strictly convex and Fréchet differentiable functional on $\mathcal{H}$, with $\nabla f(x) \in \mathcal{H}$ denoting the Fréchet derivative of f at $x \in \mathcal{H}$. Then an argument of minimum of $\min_{x\in K} f(x)$ exists and is unique, and the following conditions are equivalent*

(a) $x^* = \arg\min_{x\in K} f(x)$

(b) The following is true for all $\rho > 0$

$$x^* = P_K(x^* - \rho\nabla f(x^*)). \tag{2.12}$$

The above theorem can be found in, e.g., [134], where it is stated in a more general form and given without a proof due to its straightforwardness.

Assumption: In the rest of this chapter, we assume, without loss of generality, that operator $\mathbf{H}$ is a contraction, that is, $||\mathbf{H}||_2 < 1$ or, equivalently, $|H(\Omega)| < 1$ for all $\omega \in [0, 2\pi)$ (even more, we can assume that $0 < H(\Omega) < 1$). Indeed, let $\alpha_0 > 0$ such that $\alpha(\Omega) \leq \alpha_0$ for all $\Omega \in [0, 2\pi)$ (such α_0 exists since α is a continuous function of Ω). Let $\tilde{J}_{\alpha,\beta,v} = \frac{1}{\alpha_0+\epsilon} J_{\alpha,\beta,v}$. Optimization problem $\mathbb{P}$ is now equivalent to the one of minimizing $\tilde{J}_{\alpha,\beta,v}$ subject to $\|y\|_\infty \leq r$. We denote this problem as $\tilde{\mathbb{P}}$. The necessary and sufficient condition of optimality of $\tilde{\mathbb{P}}$ is now given as

$$y = \mathrm{Sat}_r(\tilde{\mathbf{H}}y + \tilde{w}),$$

where $\tilde{\mathbf{H}} = \mathbf{I} - \mathbf{T}_{\alpha/(\alpha_0+\epsilon)}$ and $\tilde{w} = \mathbf{T}_{\beta/(\alpha_0+\epsilon)}v$. This implies that $\tilde{H}(\Omega) = 1 - \frac{1}{\alpha_0+\epsilon}\alpha(\Omega) \in (0,1)$, and, therefore, optimal problem $\mathbb{P}$ is equivalent to the one for which operator $\mathbf{H}$ is a contraction.

As a consequence of lemma 2.1.1, a dynamical equation (2.11) defines a system which maps input signal w into unique output signal y^*. We denote this system as $\mathbf{S}^*$ and refer to it, in the rest of this paper, as the optimal system or the optimal map. Some important properties of the optimal system $\mathbf{S}^*$ are summarized in the following lemma.

Lemma 2.1.4. *The optimal system* $\mathbf{S}^*$ *is*

(a) BIBO stable,

(b) time-invariant,

(c) non-causal,

(d) nonlinear.

Proof. See the section 2.5.1. □

As shown in lemma 2.1.4, in general, the optimal system $\mathbf{S}^*$ is nonlinear and non-causal. Moreover, the optimality condition (2.11) is an implicit equation in terms of w and y^*, and is, therefore, not attractive as a description of the optimal system $\mathbf{S}^*$ mapping input signal w to the optimal signal y^*. Intuitively, it is clear that a necessary condition for the existence of a finite-latency system, which is a good approximation of the optimal system $\mathbf{S}^*$ (in, e.g., L-infinity or L2 sense), is that $\mathbf{S}^*$ possesses some type of 'near-finite' memory. That is, one hopes that system $\mathbf{S}^*$ for any two input signals that are close in the recent past and future, but not necessarily close in the remote past and future, yields present outputs which are close. In that case, one can ignore the effect that samples of w from far-away into the future have on the current output sample, and consider approximately that $\mathbf{S}^*$ is of finite latency. Indeed, in the following sections, we show that with careful memory truncation and adequate nonlinear stability analysis, the optimal system $\mathbf{S}^*$ can be approximated with arbitrary precision by a carefully designed finite-latency system.

2.2 Main Results

In section 2.2.1, we show that the optimal dynamical system $\mathbf{S}^*$ has some favorable memory characteristic. We state and prove an extension of the classical method of balanced truncation to a class of nonlinear models with weakly contractive operators in section 2.2.2. We use this result in section 2.2.3 to show that certain nonlinear models yields arbitrarily precise approximations to the optimal system $\mathbf{S}^*$.

2.2.1 Fading Memory Property of the Optimal System

The following definition of fading memory is an intuitive one: a time-invariant system has fading memory if two input signals that are close in the recent past and future, but not necessarily close in the remote past and future, yield present outputs which are close. This is a generalization of the standard definition of fading memory for causal systems, see, e.g., [137]. Before we formally state this definition, let us first introduce one useful operator. For an arbitrary integer $M > 0$ let $P_M : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be the 'windowing' operator defined by

$$(P_M w)[n] = \begin{cases} w[n], & |n| \leq M \\ 0, & |n| > M \end{cases}.$$

The fading memory property, and specifically exponentially fading memory property, of non-causal time-invariant discrete-time systems is defined as follows.

Definition 2.2.1. *A non-causal time-invariant discrete-time system* $\mathbf{S}$ *has fading memory on* $\ell^2(\mathbb{C})$ *if for all* $r > 0$

$$\lim_{M\to\infty} \sup_{\substack{w_1,w_2\in\ell^2(\mathbb{C}),\\ P_M w_1 = P_M w_2,\\ ||w_1-w_2||_\infty \leq r}} |y_1[0] - y_2[0]| = 0, \tag{2.13}$$

where $y_1 = \mathbf{S}w_1$ *and* $y_2 = \mathbf{S}w_2$. *Furthermore, if there exist* $\gamma > 0$ *and* $\rho \in (0,1)$ *such that*

$$|y_1[0] - y_2[0]| \leq \gamma ||w_1 - w_2||_\infty \rho^M, \tag{2.14}$$

for any $w_1, w_2 \in \ell^2(\mathbb{C})$ *such that* $P_M w_1 = P_M w_2$, *we say that* $\mathbf{S}$ *has exponentially fading memory.*

We are now ready to state the theorem which establishes fading memory property of the optimal system $\mathbf{S}^*$.

Theorem 2.2.2. *Suppose that* $\mathbf{H}$ *has frequency response which is strictly less than 1, that is, there exists* $\epsilon > 0$ *such that* $H(\Omega) = h[0] + 2\sum_{n=1}^{N} h[n]\cos(n\Omega) \leq 1 - \epsilon < 1$ *for all* $\Omega \in \mathbb{R}$. *Then the optimal system* $\mathbf{S}^*$ *has exponentially fading memory.*

Proof. See the Appendix 2.5.2. □

It should be noted that the condition of theorem 2.2.2 is automatically satisfied since we assumed that the cost function of problem $\mathbb{P}$, defined in (2.5), is strictly convex, that is, there exists $\epsilon > 0$ such that $\alpha(\Omega) = 1 - H(\Omega) \geq \epsilon$, and therefore $H(\Omega) \leq 1 - \epsilon$, for all $\Omega \in \mathbb{R}$.

The proof of theorem 2.2.2 does not provide explicit formulas for the constants γ and ρ, that govern the rate of decay of memory of $\mathbf{S}^*$, but provides a constructive way of calculating upper bounds for the two. The result of the above theorem suggest that the 'memory into future' of system $\mathbf{S}^*$ could be truncated and $\mathbf{S}^*$ possibly approximated with arbitrary precision by some finite-latency system. In the following sections, we show that this approximation is possible, and present two such finite-latency approximate systems.

2.2.2 Generalized Balanced Truncation Lemma

We now state and prove a result on upper bounds on error of approximating a certain class of nonlinear systems by appropriately chosen reduced order models, similar to those of the classical balanced truncation algorithm for linear systems.

Let X, W and Y be Hilbert spaces. In general, X can be infinite dimensional. Consider systems $\mathbf{G} : \ell^2(W) \to \ell^2(Y)$ and $\tilde{\mathbf{G}} : \ell^2(W) \to \ell^2(Y)$ described by the following state space models

$$\mathbf{G} : x[n+1] = \varphi(Ax[n] + Bw[n]), \quad y[n] = Cx[n], \tag{2.15}$$

$$\tilde{\mathbf{G}} : \tilde{x}[n+1] = \Pi\varphi(A\tilde{x}[n] + Bw[n]), \quad \tilde{y}[n] = C\tilde{x}[n], \tag{2.16}$$

where $A : X \to X$, $B : W \to X$, $C : X \to Y$ and $\Pi : X \to X$ are bounded linear operators, Π is a projection, i.e., $\Pi^2 = \Pi$, and $\varphi : X \to X$ is a diagonal operator in the standard basis in X. The following theorem gives an upper bound on error of approximating $\mathbf{G}$ with $\tilde{\mathbf{G}}$.

Theorem 2.2.3. *Let $\sigma_1, \sigma_2 > 0$ be positive real numbers and $P = P' > 0$, $Q = Q' > 0$ be positive definite self-adjoint operators on X, satisfying the following Lyapunov inequalities*

$$P - APA' \geq \frac{1}{\sigma_1^2} BB', \qquad Q - A'QA \geq \frac{1}{\sigma_2^2} C'C. \tag{2.17}$$

Let Π and φ satisfy the following conditions

$$(P^{-1} - Q)(I - \Pi) = 0, \qquad P^{-1} + Q \geq \Pi(P^{-1} + Q)\Pi, \tag{2.18}$$

$$(\varphi(u) + \varphi(v))' P^{-1} (\varphi(u) + \varphi(v)) \leq (u + v)' P^{-1} (u + v), \quad \forall u, v \in X, \tag{2.19}$$

$$(\varphi(u) - \varphi(v))' Q (\varphi(u) - \varphi(v)) \leq (u - v)' Q (u - v), \quad \forall u, v \in X. \tag{2.20}$$

Then

$$||\mathbf{G} - \tilde{\mathbf{G}}|| \leq 2\sigma_1\sigma_2.$$

Proof. See the Appendix 2.5.3. □

Theorem 2.2.3 is a generalization of the well known result on upper bounds of H-infinity error for the exact implementation of the balanced truncation algorithm for linear systems [7]. Indeed, let $\varphi = I$ and let (A, B, C) be the balanced realization of system $\mathbf{G}$, where the controllability and observability gramians W_c and W_o, respectively, satisfy $W_c = W_o = \Sigma > 0$ for a block diagonal balanced gramian Σ. Let σ be the smallest Hankel singular value of $\mathbf{G}$, and let $\Sigma = \mathrm{diag}(\Sigma_0, \sigma I)$ with block diagonal Σ_0. In the classical balanced truncation method, one aims at truncating states of $\mathbf{G}$ that correspond to the lower-right block σI of Σ. Therefore, the projection matrix Π is defined as $\Pi = \mathrm{diag}(I, \mathbf{0})$, where the dimension of the zero matrix $\mathbf{0}$ corresponds to that of the σI submatrix. It now follows that the dissipation inequalities (2.17) are satisfied (with equality) for $\sigma_1 = \sigma_2 = \sqrt{\sigma}$

and $P = Q = \frac{1}{\sigma}\Sigma$. Expressions in (2.67)-(2.20) hold by the definitions of P, Q and Π. Therefore, the balanced truncation error bound follows from Theorem 2.2.3, i.e., the upper bound on H-infinity error of approximating $\mathbf{G}$ with $\tilde{\mathbf{G}}$ is 2σ.

2.2.3 Real-Time Suboptimal Algorithms

In this section we propose two algorithms for sequentially obtaining approximations of arbitrary accuracy to the optimal solution of problem (4.29). In the first case, the approximation error of modeling the optimal system with the proposed ones is measured in terms of the L2 induced gain of the error system, while in the second case the supremum or L-infinity norm of the error system is used.

As before, the optimal system $\mathbf{S}^*$ maps signal $w \in \ell^2(\mathbb{C})$ to $y \in \ell^2(\mathbb{C})$, as defined by (2.11), that is:

$$y[n] = \mathrm{sat}_r \left(\sum_{k=-N}^{N} h[k]y[n-k] + w[n] \right), \forall n \in \mathbb{Z}. \tag{2.21}$$

A. The L2 Case

We first introduce two fairly standard definitions of banded and Laurant operators on $\ell^2(\mathbb{C})$.

Definition 2.2.4. *Let $\mathbf{S} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be a bounded linear operator with matrix (in the standard basis) $S = (s_{ij})_{i,j=-\infty}^{\infty}$. We say that $\mathbf{S}$ is a banded operator if there exists a positive integer T such that $s_{ij} = 0$ for all $|i-j| > T$. Minimal integer T for which this is true is called the bandwidth of $\mathbf{S}$, in which case we say that $\mathbf{S}$ is an T-banded operator.*

Definition 2.2.5. *Let $\mathbf{S} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be a bounded linear operator with matrix (in the standard basis) $S = (s_{ij})_{i,j=-\infty}^{\infty}$. We say that $\mathbf{S}$ is a Laurent operator if there exists $f \in \ell^2(\mathbb{C})$ such that $s_{ij} = f[i-j]$ for all $i, j \in \mathbb{Z}$. Such f is called the symbol of S.*

It is clear that $\mathbf{H}$ is an N-banded Laurent operator with symbol $h = h[n]$. This fact will be used in the proof of theorem (2.2.6).

We now define a finite-latency nonlinear discrete-time system that approximates the optimal solution of (4.29) with arbitrary precision, and give an upper bound on the L2

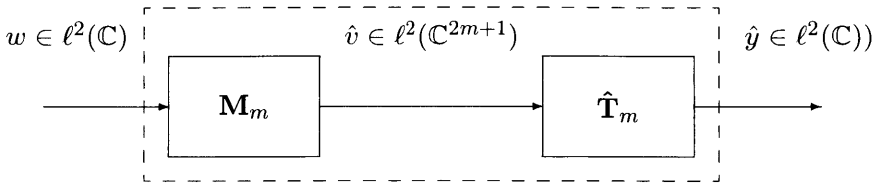


Figure 2-2: Equivalent representation of the approximate system $\hat{\mathbf{S}}_m$ as a series interconnection $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_m \mathbf{M}_m$ of the finite latency system $\mathbf{M}_m$ and the finite-dimensional state-space model $\hat{\mathbf{T}}_m$.

induced gain of the model approximation error.

For a given integer $m > N$, let system $\mathbf{M}_m : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C}^{2m+1})$, mapping w to $\hat{v}$, be defined by

$$\hat{v}[n] = \begin{bmatrix} w[n-m+1] & w[n-m+2] & \dots & w[n+m+1] \end{bmatrix}^T. \tag{2.22}$$

Let system $\hat{\mathbf{T}}_m : \ell^2(\mathbb{C}^{2m+1}) \to \ell^2(\mathbb{C})$ be defined by the following state space model

$$\hat{x}[n+1] = \mathrm{sat}_r(\hat{A}\hat{x}[n] + \hat{v}[n]), \quad \hat{y}[n] = \hat{C}\hat{x}[n], \tag{2.23}$$

where $\hat{x}[n], \hat{v}[n] \in \mathbb{C}^{2m+1}$, for all $n \in \mathbb{Z}$, and matrices $\hat{A} = (\hat{a}_{ij})_{i,j=1}^{2m+1} \in \mathbb{R}^{(2m+1)\times(2m+1)}$ and $\hat{C} = (\hat{c}_j)_{j=1}^{2m+1} \in \mathbb{R}^{1\times(2m+1)}$ are defined by

$$\hat{a}_{ij} = h[i-j+1], \quad \forall i,j \in \{1,\dots,2m+1\},$$

$$\hat{c}_k = 1 \text{ for } k = m, \text{ and } \hat{c}_k = 0 \text{ otherwise.}$$

Matrix $\hat{A}$ is clearly a Toeplitz matrix. Moreover, it can be understood as a truncation of the matrix of the infinite dimensional Laurent operator $\mathbf{H}$.

Systems $\hat{\mathbf{T}}_m$ and $\mathbf{M}_m$ are clearly time-invariant systems, where the former is nonlinear and causally stable while the latter is linear and non-causal but of finite latency (equal to $m+1$). Let system $\hat{\mathbf{S}}_m : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ mapping w to $\hat{y} = \hat{\mathbf{S}}_m w$ be defined as the series interconnection $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_m \mathbf{M}_m$ of $\mathbf{M}_m$ and $\hat{\mathbf{T}}_m$ (see Figure 2-2).

Let $\mathbf{E}_m \equiv \mathbf{S}^* - \hat{\mathbf{S}}_m$ be the 'error' system, mapping $w \in \ell^2(\mathbb{C})$ to $e = \mathbf{E}_m w =$

$\mathbf{S}^* w - \hat{\mathbf{S}}_m w \in \ell^2(\mathbb{C})$. System $\mathbf{E}_m$ models the error of approximating the optimal system $\mathbf{S}^*$ with the finite-latency system $\hat{\mathbf{S}}_m$. The following theorem establishes that the L2-induced gain of the error system $\mathbf{E}_m$ is upper bounded by $\epsilon = c\rho^m$ for some $c > 0$ and $\rho \in (0, 1)$.

Theorem 2.2.6. *If* $\mathbf{H}$ *is a strict contraction in H-infinity sense, that is,* $|H(\Omega)| = |h[0] + \sum_{n=1}^{N} h[n] \cos(n\Omega)| \leq 1 - \delta < 1$ *for some* $\delta \in (0, 1)$*, then there exist* $\rho \in (0, 1)$ *and* $c > 0$ *such that* $\|\mathbf{S}^* - \hat{\mathbf{S}}_m\|_2 \leq c\rho^m$ *for all* $m \in \mathbb{Z}, m > N$.

Proof. See the Appendix 2.5.5. □

It should be noted that the condition of theorem 2.2.6 is defacto not a condition at all, since we argued in earlier sections that it can be achieved by proper re-scaling of the cost function in (2.5). Therefore, regardless of the choice of the problem parameters γ and r, and the spectral weighting filter $L = L(\Omega)$, the above model can be used to find arbitrarily precise approximations to the optimal solution of $\mathbb{P}$.

The L-infinity Case

We first define what is meant by an 'approximate' system in this case. Let $\epsilon > 0$. System $\mathbf{T} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ is an ϵ-approximation to system $\mathbf{S} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ if

$$|\mathbf{S}w[n] - \mathbf{T}w[n]| < \epsilon ||w||_\infty, \quad \forall n \in \mathbb{Z}, \forall w \in \ell^2(\mathbb{C}).$$

In addition to the standard assumptions of positivity and finite unit-sample response (FIR) of system $\mathbf{H}$, we will also assume that $\mathbf{H}$ has an "L1 dominance" property in the sense that its unit-sample response h satisfies $\sum_{n=-N}^{N} |h[n]| < 1$.

For a given integer $m > 0$, let matrices $\hat{A} \in \mathbb{R}^{(N+m)\times(N+m)}$ and $\hat{C} \in \mathbb{R}^{1\times(N+m)}$ be defined as

$$\hat{A} = \begin{bmatrix} 0 & 1 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ h_N & h_{N-1} & \dots & h_1 & h_0 & \dots & h_N & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & h_N & \dots & h_2 & h_1 & \dots & h_{N-1} & h_N & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & h_N & h_{N-1} & \dots & h_1 & h_2 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & h_N & \dots & h_0 & h_1 & \dots & h_N \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & h_N & h_{N-1} & \dots & h_0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & h_N & \dots & h_1 \end{bmatrix}, \tag{2.24}$$

$$\hat{C} = \Big[\underbrace{0 \ \cdots \ 0}_{N-1} \ 1 \ \underbrace{0 \ \cdots \ 0}_{m}\Big]. \tag{2.25}$$

In (2.24), for simplicity, we use the shorthand notation $h[n] = h_n$. Let the system $\hat{\mathbf{S}}_m : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$, mapping $w \in \ell^2(\mathbb{C})$ to $\hat{y} = \hat{\mathbf{S}}_m w \in \ell^2(\mathbb{C})$, be defined by the following state space model

$$\begin{aligned} x[n+1] &= \mathrm{sat}_r(\hat{A}x[n] + \hat{w}[n]), \\ \hat{y}[n] &= \hat{C}x[n], \end{aligned}$$

where $x[n], \hat{w}[n] \in \mathbb{C}^{N+m}$, for all $n \in \mathbb{Z}$, and

$$\hat{w}[n] = \begin{bmatrix} 0 & \dots & 0 & w[n+1] & w[n+2] & \dots & w[n+m+1] \end{bmatrix}^T$$

for all $n \in \mathbb{Z}$.

Let $P : \mathbb{R} \to \mathbb{R}$ be defined as

$$P(x) = \left(1 - \sum_{k=1}^{N} |h[k]|\right) x^{N+1} - \sum_{k=-N}^{0} |h[k]| x^{N+k}. \tag{2.26}$$

Since $\sum_{k=-N}^{N} |h[k]| < 1$ and $|h[N]| = |h[-N]| > 0$, it is obvious that P has at least one root root on the interval $(0, 1)$. Indeed, the continuity of P and the fact that $P(0) = -|h[-N]| < 0$ and $P(1) = 1 - \sum_{k=-N}^{N} |h[k]| > 0$ implies the existence of $x_0 \in (0, 1)$ such that $P(x_0) = 0$. The following theorem establishes that the finite-latency system $\hat{\mathbf{S}}_m$ is an ϵ-approximation to the optimal system $\mathbf{S}^*$, where $\epsilon = c\rho^m$ for some $c > 0$, $\rho \in (0, 1)$ and any integer $m > N$.

Theorem 2.2.7. *If* $\|h\|_1 = \sum_{k=-N}^{N} |h[k]| \leq 1 - \delta$ *for some* $\delta \in (0, 1)$, *then for any* $m \in \mathbb{Z}$, $m > N$, *the following is true*

$$|\mathbf{S}^* w[n] - \hat{\mathbf{S}}_m w[n]| \leq c\|w\|_\infty \rho^m, \quad \forall n \in \mathbb{Z}, \forall w \in \ell^2(\mathbb{C}),$$

where ρ *is the largest root of* P, *defined in* (2.26), *on the interval* $(0, 1)$, *and* $c > 0$ *is defined by*

$$c = \max_{0 \leq i \leq N} \frac{\sum_{k=-N}^{-i} |h[k]|}{\rho^i - \sum_{k=1-i}^{N} |h[k]| \rho^{i+k-1}}.$$

Proof. See the Appendix 2.5.4. □

It is clear that condition $\sum_{k=-N}^{N} |h[k]| < 1$ is much stronger than condition $|H(\Omega)| < 1$, for all $\Omega \in \mathbb{R}$. In most PAPR reduction problems, the spectral weighting filter L is such that the L1 condition cannot be achieved (even with re-scaling of the cost function). However, the derived approximation bound in the case of L1 condition is much stronger than that in the L2 case (i.e., bounding of the worst error vs. bounding of the average error). Though model presented in this section does not yield useful results in the PAPR reduction problem, it can be utilized in other applications which can be formulated using the same abstract optimization setup (e.g., envelope tracking of rapidly changing DT signals).

2.3 Numerical Example: PAPR Reduction Algorithm

In the previous sections, we formulated the problem of designing discrete-time systems for peak-to-average power ratio reduction as an abstract convex optimization problem. We developed two algorithms which return arbitrarily precise approximations to the optimal solution of the above problem. In this section, aided by MATLAB simulations, we verify effectiveness of the optimal PAPR reduction method and confirm the quality of approximations generated by the algorithm derived under assumption $|H(\Omega)| < 1$, for all $\Omega \in \mathbb{R}$. Simulations were performed for standard-compliant LTE signals generated by MATLAB's LTE Toolbox, as explained below.

A. Simulation Setup

The input signals are downlink reference measurement channel (RMC) waveforms R.9 [132]. The input signal parameters are as follows: 64QAM modulation method, bandwidth of 20MHz (i.e., 100 resource blocks), frequency division duplexing (FDD) mode, and each signal contains 10 subframes. In order to get a signal appropriate for wireless transmission, the true LTE signal is filtered with a lowpass filter to satisfy the spectral mask conditions [132] and up-sampled by the factor of 2. The frequency-weighting filter $\mathbf{L}$, in the cost function, is modeled as an FIR filter with 101 nonzero coefficients (i.e., taps), passband ripple of 1dB and passband-to-stopband magnitude drop-off of 120dB. The EVM vs. ACLR trade-off parameter γ, saturation parameter r, and sliding window length m were varied depending on the type of performed simulation, as explained below. The quality of the proposed PAPR reduction method is measured in terms of PAPR, EVM and ACLR of the output signal. It should be noted that ACLR and EVM are measured using MATLAB's (LTE Toolbox) built-in functions, which calculate ACLR and EVM according to the 3GPP standard specifications [132]. All the results were obtained by Monte Carlo simulation with 100 simulation runs.

B. Simulation Results

We first report results for the truly (sub)optimal solution of problem ($\mathbb{P}_0$), that is, we report the results obtained by applying the optimal system $\mathbf{S}^*$ to the input LTE signals. This will serve as a benchmark for the results achieved by the approximate model. The (sub)optimal solution was generated by running the following fixed point iteration on the whole history of the input signal w (which, in simulations, is a vector of finite length):

$$y_{k+1} = \mathrm{sat}_r(\tilde{H}y_k + w),$$

where $\tilde{H}$ is a truncated (finite-dimensional) version of the matrix of operator $\mathbf{H}$. The input signals that were used in the Monte Carlo simulation have PAPR of approximately 11.75dB.

Figure 2-3 shows the EVM versus ACLR Pareto curve, for the cost function defined in ($\mathbb{P}_0$), parametrized by the trade-off parameter γ. For each point on the curve, the achieved output signal PAPR was kept at approximately 6.77dB, while parameter γ (and very slightly parameter r) was varied. In general, the limits on admissible EVM and ACLR values of a signal generated by a real communication system are set by the standard [132] and/or the regulatory agency, and depend on the type of transmission (e.g., mobile vs. WLAN) and parameters of the underlying communication signal. For our simulation setup, the admissible values of (EVM,ACLR) pairs are enclosed in the dashed rectangle in Figure 2-3. As can be seen, in the case of small γ the ACLR cost is more heavily penalized, which leads to good ACLR but inadmissible EVM performance of the optimal output signal. Likewise, taking large γ yields good EVM performance but inadmissible ACLR performance.

As the results of theorems (2.2.7) and (2.2.6) suggest, performance of the approximate algorithms proposed in this thesis primarily depend on the memory window length m. For that reason, we run the approximate algorithm for various values of parameter m while keeping all other parameters fixed. In figure 2-4, we report the achieved PAPR, EVM, and ACLR values relative to the optimal ones obtained by finding the true optimal solution. As expected, performance of the approximate algorithm converges to the optimal one as m increases. We also tested performance sensitivity of the approximate algorithm to the change in parameter γ. As can be seen from figure 2-4, as γ increases, the approximate

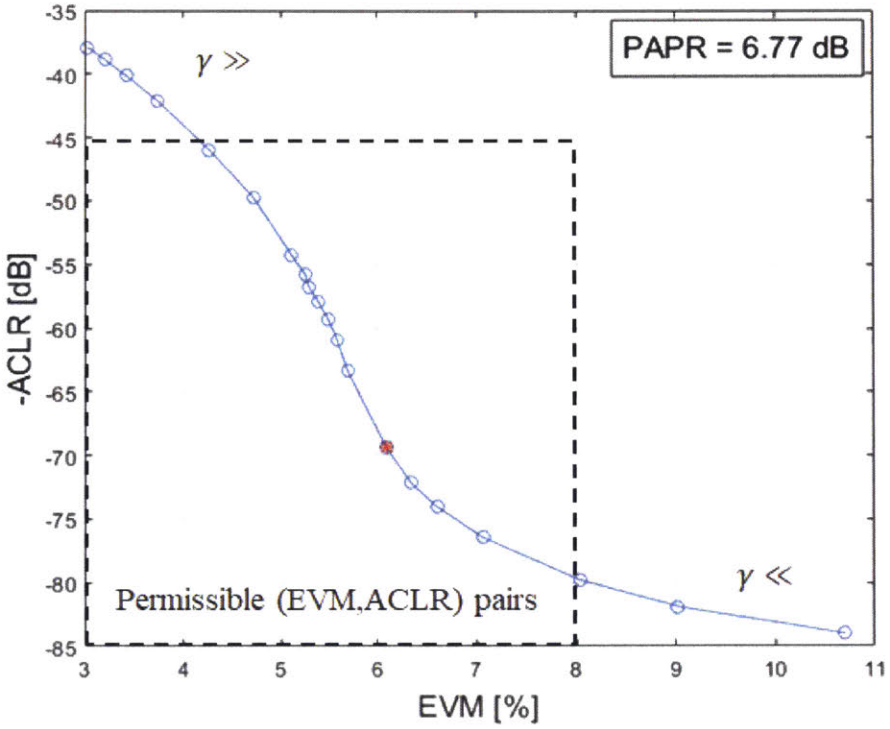


Figure 2-3: EVM vs. ACLR Pareto curve, parametrized by γ, for an optimal solution of $(\mathbb{P}_0)$ with PAPR of 6.77dB.

solution converges faster to the optimal one. This is not surprising, as increasing parameter γ causes the maximal value of the frequency response $H(\Omega)$ of $\mathbf{H}$ to back-off from 1, decreasing the Lipschitz constant of the state evolution map in the approximate model (2.23), therefore, increasing the speed of convergence of the approximate algorithm.

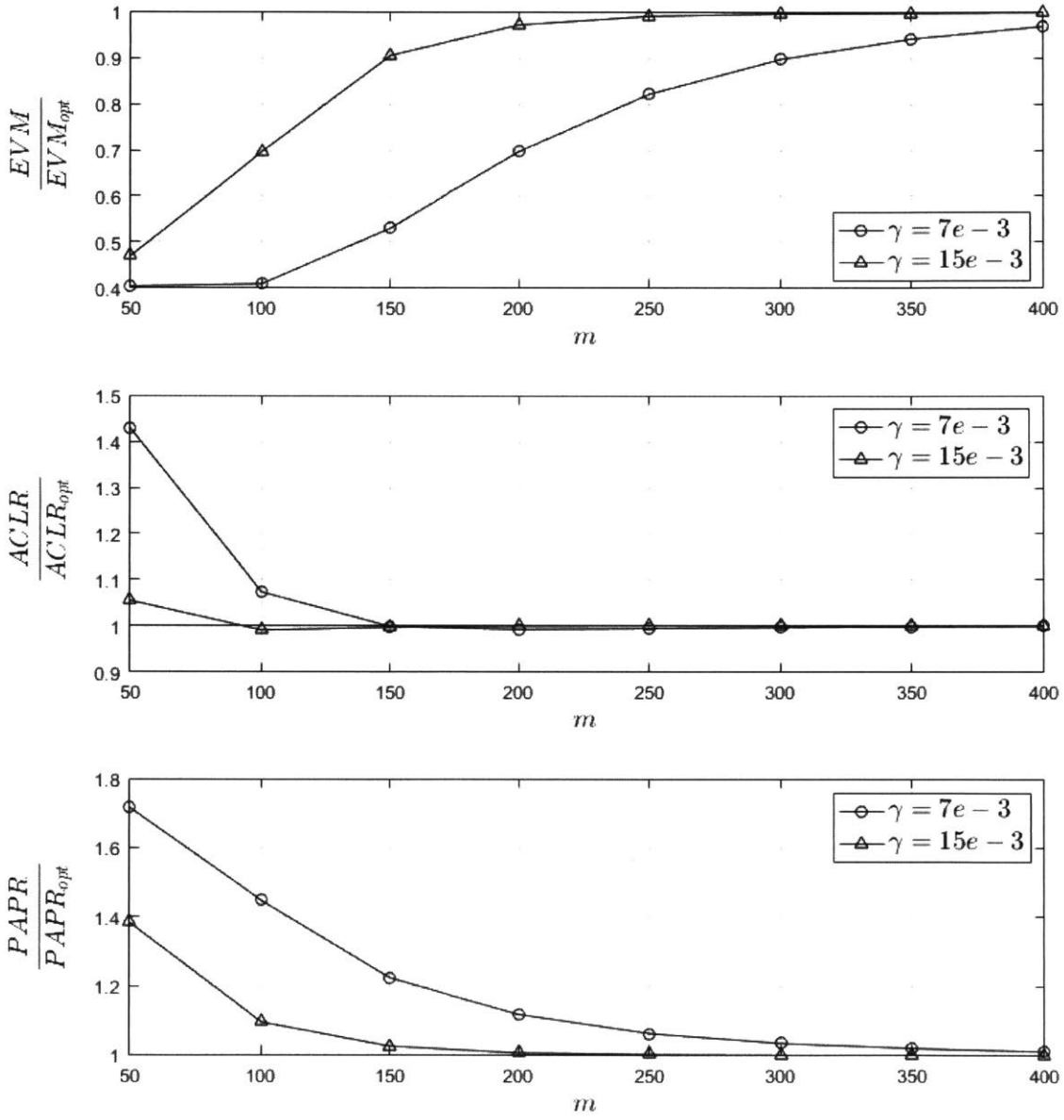


Figure 2-4: Relative values of EVM, ACLR, and PAPR achieved with the approximate model, as functions of the memory window m, for different values of the parameter γ.

2.4 Summary

In this chapter, we considered a problem of designing discrete-time systems for peak-to-average-power ratio reduction of communication signals. The problem was formulated as minimization of a frequency-weighted convex quadratic cost subject to time-domain output amplitude constraints. For such problems, in general, the optimality conditions do not provide an explicit way of generating the optimal output as a real-time implementable transformation of the input. We showed that the optimal system, corresponding to the optimal solution of such problem, has exponentially fading memory. Two algorithms, based on time and value iterations of carefully chosen stable finite-latency nonlinear systems, which return approximations of arbitrary precision to the optimal map, were proposed. In one case, the result holds under an L1 dominance assumption about some parameters of the cost function. In the other case, no special assumptions on cost function parameters are needed. The approximate system was obtained by a careful truncation of an infinite dimensional state space representation of the optimal system, were an upper bound on the error of approximation was derived by extending the method of balanced truncation for linear systems to a certain class of nonlinear models. Numerical simulations in Matlab were used to verify effectiveness of the proposed PAPR reduction method. The algorithm was applied on the realistic LTE baseband communication signals, and significant reduction in PAPR was achieved.

2.5 Proofs of Chapter 2

2.5.1 Proof of Lemma 2.1.4

(a) BIBO stability follows trivially from the maximal-value constraint of the optimal problem.

(b) This follows directly from the formulation of problem $\mathbb{P}$. The feasible set B_r is clearly time-invariant. The cost function $J_{\alpha,\beta}(v, y)$ is also time-invariant. Indeed, if (v_1, y_1) is an input output pair of S^*, then (v_2, y_2), where $v_2[n] = v_1[n-k], y_2[n] = y_1[n-k]$, for all n,

is an input-output pair as well, for all $k \in \mathbb{Z}$. This follows from

$$J_{\alpha,\beta}(v_1, y_1) = J_{\alpha,\beta}(v_2, y_2).$$

(c) Let $\mathbf{G} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be a BIBO LTI system with frequency response $G = G(\Omega)$ such that $G(\Omega) = \frac{1}{1-H(\Omega)}$ for all $\Omega \in \mathbb{R}$. It now follows from the BIBO stability assumption and the definition of $\mathbf{H}$ that system $\mathbf{G}$ is non-causal. Let $g = g[n]$ be the unit-sample response of $\mathbf{G}$, and let $\sum_{n=-\infty}^{\infty} |g[n]| = \gamma$. Let $w_0 \in \ell^2(\mathbb{C})$ such that $||w_0||_2 > 0$, $|w_0[n]| \leq \frac{r}{\gamma}$ for all n, and let $y_0 = \mathbf{G}w_0$. Then $y_0 = \mathbf{S}^* w_0$. Indeed, from the definition of $\mathbf{G}$ and w_0 it follows that $y_0[n] = w_0[n] + \sum_{k=-N}^{N} h[k] y_0[n-k]$ and $|y_0[n]| \leq \gamma \sup_n |w_0[n]| \leq r$ for all n. Hence, the pair (w_0, y_0) satisfies the optimal equation (2.11). Non-causality of system $\mathbf{S}^*$ now directly follows from the non-causality of $\mathbf{G}$.

(d) Let $(w_0, y_0) \in \ell^2(\mathbb{C}) \times \ell^2(\mathbb{C})$ be the input-output signal pair of $\mathbf{S}^*$, as defined in the proof of part (c). Let $c \in \mathbb{R}$ such that $c > \frac{r}{\sup_n |y_0[n]|}$. Then the signal pair $(w_1, y_2) \in \ell^2(\mathbb{C}) \times \ell^2(\mathbb{C})$ such that $w_1[n] = cw_0[n]$, $y_1[n] = cy_0[n]$ for all n, cannot satisfy $y_2 = \mathbf{S}^* w_1$ since $\sup_n |y_2[n]| > r$. Therefore, system $\mathbf{S}^*$ is nonlinear.

2.5.2 Proof of Theorem 2.2.2

Let us first state and prove the following lemma which is a direct consequence of the necessary and sufficient conditions of the minimum distance to a convex set problem (see, e.g., theorem 1 in chapter 3.12 of [138]).

Lemma 2.5.1. *Let $\mathcal{H}$ be a real Hilbert space with norm $\|\cdot\|$, and let $K \subset \mathcal{H}$ be closed and convex. Let $P_K : \mathcal{H} \to \mathcal{H}$ be the projection-onto-K map, that is, $x \mapsto P_K x = \arg\min_{z \in K} \|x - z\|$, for all $x \in \mathcal{H}$. The following quadratic inequality is satisfied for all $x_1, x_2 \in \mathcal{H}$:*

$$\|P_K x_1 - P_K x_2\|^2 \leq (x_1 - x_2)'(P_K x_1 - P_K x_2). \tag{2.27}$$

Proof. We consider the following three cases:

(i) When $x_1, x_2 \in K$, (2.27) becomes an equality and is trivially true.

(ii) Let $x_1, x_2 \notin K$. Since K is closed and convex, and $P_K x_1$ and $P_K x_2$ are projections of x_1 and x_2 onto K, respectively, the following two inequalities hold

$$(x_1 - P_K x_1)'(P_K x_2 - P_K x_1) \leq 0,$$

$$(x_2 - P_K x_2)'(P_K x_1 - P_K x_2) \leq 0.$$

Inequality (2.27) now directly follows by combining the above two inequalities.

(iii) Without loss of generality, we assume that $x_1 \notin K$ and $x_2 \in K$. It follows from $x_2 = P_K x_2$ and the necessary and sufficient conditions of optimality of the projection $P_K x_1$ that

$$(x_1 - P_K x_1)'(x_2 - P_K x_1) \leq 0,$$

and therefore,

$$(x_1 - x_2 + P_K x_2 - P_K x_1)'(P_K x_2 - P_K x_1) \leq 0,$$

which is equivalent to (2.27).

□

The following corollary, which will be used in the proof of the main theorem, is a direct consequence of the above lemma for $\mathcal{H} = \mathbb{C}$.

Corollary 2.5.2. *For any $z_1, z_2 \in \mathbb{C}$, the following inequality is true*

$$|\operatorname{sat}_r(z_1) - \operatorname{sat}_r(z_2)|^2 \leq \mathit{Re}\{(z_1 - z_2)'(\operatorname{sat}_r(z_1) - \operatorname{sat}_r(z_2))\}. \tag{2.28}$$

Now we proceed with the proof of the main theorem, where the proof consists of three major steps.

Step 1: For $w_1, w_2 \in \ell^2(\mathbb{C})$, let $y_1 = \mathbf{S}w_1$ and $y_2 = \mathbf{S}w_2$. Moreover, let $\delta = y_1 - y_2$ and $\tilde{w} = w_1 - w_2$. It is now sufficient to show that $\sup |\delta[0]|$ converges exponentially to 0 as $N \to \infty$, where the supremum is taken over all pairs of signals $w_1, w_2 \in \ell^2(\mathbb{C})$ such that

$w_1[n] = w_2[n]$ for all $|n| \leq N$.

Let $w_1, w_2 \in \ell^2(\mathbb{C})$ such that $w_1[n] = w_2[n]$ for all $|n| \leq N$, and let $\tilde{w} = w_1 - w_2$. It follows from

$$y_i[n] = \text{sat}_r \left(\sum_{\tau=-N}^{N} h[k] y_i[n-k] + w_i[n] \right), \forall n \in \mathbb{Z}, i \in \{1, 2\}, \tag{2.29}$$

and from corollary 2.5.2, that the following inequality is satisfied for all $n \in \mathbb{Z}$:

$$|y_1[n] - y_2[n]|^2 \leq \text{Re} \left\{ (y_1[n] - y_2[n])' \left(\tilde{w}[n] + \sum_{k=-N}^{N} h[k](y_1[n-k] - y_2[n-k]) \right) \right\}. \tag{2.30}$$

Since $\tilde{w}[n] = 0$ for $|n| \leq N$, the following inequality holds for all $n \in \{-N, \ldots, N\}$:

$$|\delta[n]|^2 \leq \text{Re} \left\{ \delta[n]' \sum_{k=-N}^{N} h[k] \delta[n-k] \right\}. \tag{2.31}$$

Step 2: Show that there exists $P = P' \in \mathbb{R}^{2N \times 2N}$ such that the quadratic storage function $V(x) = x'Px$ satisfies the following dissipation inequality for all $n \in \mathbb{Z}$:

$$|\delta[n]|^2 - \text{Re} \left\{ \delta[n]' \sum_{k=-N}^{N} h[k] \delta[n-k] \right\} - \epsilon \sum_{k=-N}^{N} |\delta[n-k]|^2 \geq V(x[n+1]) - V(x[n]). \tag{2.32}$$

where $x[n] = \begin{bmatrix} \delta[n-N] & \delta[n-N+1] & \ldots & \delta[n+N-1] \end{bmatrix}^T$.

To show this, consider the following state space representation of the LTI system $\mathbf{H}$, mapping $\delta = \delta[n]$ to $y = y[n] = \sum_{k=-N}^{N} h[k] \delta[n-k]$:

$$x[n+1] = Ax[n] + Bu[n], \quad x[n_0] = x_0, \tag{2.33}$$

$$y[n] = Cx[n] + Du[n], \tag{2.34}$$

where $u[n] = \delta[n+N]$, and matrices $A \in \mathbb{R}^{2N \times 2N}$, $B \in \mathbb{R}^{2N \times 1}$, $C \in \mathbb{R}^{1 \times 2N}$, $D \in \mathbb{R}$, are

defined as

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad C^T = \begin{bmatrix} h[N] \\ h(N-1) \\ \vdots \\ h(-N+1) \end{bmatrix}, \quad D = h(-N).$$

Suppose $\epsilon > 0$ such that $H(\Omega) \leq 1 - (2N+1)\epsilon$ for all $\Omega \in \mathbb{R}$. Let $L \in \mathbb{R}^{2N \times 1}$ such that $L_k = 1$ for $k = N+1$ and $L_k = 0$ otherwise. Let the quadratic form $\sigma : \mathbb{C}^{2N} \times \mathbb{C} \to \mathbb{R}$ be defined as

$$\sigma(X,U) = X'LL'X - \epsilon(X'X + U'U) - \frac{1}{2}\text{Re}\left\{X'L(CX+DU)\right\}, \tag{2.35}$$

that is

$$\sigma(X,U) = \begin{bmatrix} X \\ U \end{bmatrix}' \begin{bmatrix} LL' - \epsilon I - \frac{1}{2}LC - \frac{1}{2}C'L' & -\frac{1}{2}LD \\ -\frac{1}{2}D'L' & -\epsilon I \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix}. \tag{2.36}$$

For the above defined x and u, this is equivalent to

$$\sigma(x[n], u[n]) = |\delta[n]|^2 - \epsilon \sum_{k=-N}^{N} |\delta[n-k]|^2 - \text{Re}\left\{\delta[n]' \sum_{k=-N}^{N} h[k]\delta[n-k]\right\}.$$

For every $\Omega \in (0, 2\pi]$, let $\mathcal{M}_\Omega$ be defined as

$$\mathcal{M}_\Omega = \{(X,U) \in \mathbb{C}^{2N} \times \mathbb{C} : e^{j\Omega}X = AX + BU, \Omega \in (0, 2\pi]\}.$$

By relatively straightforward, but tedious, algebraic manipulations, it can be shown that

$$\sigma(X,U) = (1 - (2N+1)\epsilon - H(\Omega))|U|^2 \geq 0, \quad \forall (X,U) \in \mathcal{M}_\Omega.$$

The pair (A, B) is controllable since the controllability matrix

$$\mathcal{C} = \begin{bmatrix} B & BA & \dots & BA^{2N-1} \end{bmatrix} = I$$

is of maximal rank. Therefore, the KYP lemma (see e.g. [139] or [140]) guarantees existence of a (sign-indefinite) real matrix $P = P'$ such that the quadratic storage function $V(x) = x'Px$ satisfies the following dissipation inequality for all $n \in \mathbb{Z}$:

$$\sigma(x[n], u[n]) \geq V(x[n+1]) - V(x[n]), \tag{2.37}$$

that is,

$$|\delta[n]|^2 - \mathrm{Re}\left\{\delta[n]' \sum_{k=-N}^{N} h[k]\delta[n-k]\right\} - \epsilon \sum_{k=-N}^{N} |\delta[n-k]|^2 \geq V(x[n+1]) - V(x[n]). \tag{2.38}$$

Step 3: Establish an upper bound on $|\delta(0)|$.

Let us recall that $x[n] = \begin{bmatrix}\delta[n-N] & \delta[n-N+1] & \dots & \delta[n+N-1]\end{bmatrix}^T$. By expanding $|x[n]|^2$ and $|x[t+1]|^2$, we have the following sequence of inequalities, that hold for all $|n| \leq N$:

$$\begin{aligned}
\frac{\epsilon}{2}|x[n]|^2 + \frac{\epsilon}{2}|x[n+1]|^2 = \frac{\epsilon}{2}\sum_{k=-N+1}^{N} |\delta[n-k]|^2 + \frac{\epsilon}{2}\sum_{k=-N}^{N-1} |\delta[n-k]|^2 \leq \\
\leq \epsilon \sum_{k=-N}^{N} |\delta[n-k]|^2 \leq \\
\leq \mathrm{Re}\left\{\delta[n]' \sum_{k=-N}^{N} h[k]\delta[n-k]\right\} - |\delta[n]|^2 + \epsilon \sum_{k=-N}^{N} |\delta[n-k]|^2 \leq \\
\leq V(x[n]) - V(x[n+1]).
\end{aligned}$$

The first inequality holds trivially, the second inequality used (2.31), and the last inequality used (2.38). Therefore, for all $|n| \leq N$, we have

$$\frac{\epsilon}{2}|x[n]|^2 + \frac{\epsilon}{2}|x[n+1]|^2 \leq V(x[n]) - V(x[n+1]). \tag{2.39}$$

Summing inequalities (2.39) for $n \in \{-k, \dots, k-1\}$, where $0 \leq k \leq N$, and denoting

$S_M = \sum_{n=-M}^{M} |x[n]|^2$, it follows that

$$V(x[-k]) - V(x[k]) \geq \frac{\epsilon}{2}\left(\sum_{n=-k}^{k-1} |x[n]|^2 + \sum_{n=-k+1}^{k} |x[n]|^2\right) \geq \epsilon S_{k-1}. \tag{2.40}$$

Since $V = V(x)$ is a quadratic form, there exists $c > 0$ such that $V(x) \leq c|x|^2$ for all $x \in \mathbb{C}^{2N\times 1}$. Therefore

$$V(x[-k]) - V(x[k]) \leq c(|x[-k]|^2 + |x[k]|^2). \tag{2.41}$$

Since $|x[-k]|^2 + |x[k]|^2 = S_k - S_{k-1}$, it follows from (2.40) and (2.41) that

$$S_{k-1} \leq \frac{c}{c+\epsilon} S_k, \quad \forall k \leq N. \tag{2.42}$$

Since $S_N = \sum_{n=-N}^{N} |x[n]|^2$ and $||\delta||_\infty \leq r$, there exists $\gamma > 0$ such that $S_N \leq \gamma^2 ||\tilde{w}||_\infty^2$, and hence

$$|\delta[0]| \leq |x[0]| = \sqrt{S_0} \leq \gamma ||\tilde{w}||_\infty \left(\sqrt{\frac{c}{c+\epsilon}}\right)^T. \tag{2.43}$$

This concludes the proof.

2.5.3 Proof of Lemma 2.2.3

Step 1:

Show that P and Q satisfy the following dissipation inequalities, for all $w \in W$ and all $x \in X$:

$$\sigma_1^2 |w|^2 \geq (Ax + Bw)' P^{-1} (Ax + Bw) - x' P^{-1} x, \tag{2.44}$$

$$-\frac{1}{\sigma_2^2}|Cx|^2 \geq (Ax)' Q (Ax) - x' Q x, \tag{2.45}$$

It is not hard to see that dissipation inequality (2.45) immediately follows from the second inequality in (2.17). The first inequality in (2.17), after a congruence transformation by

$P^{-1/2}$, and some algebraic manipulation, is equivalent to

$$\begin{bmatrix} P^{-\frac{1}{2}}AP^{\frac{1}{2}} & \frac{1}{\sigma_1}P^{-\frac{1}{2}}B \end{bmatrix} \begin{bmatrix} P^{\frac{1}{2}}A'P^{-\frac{1}{2}} \\ \frac{1}{\sigma_1}B'P^{-\frac{1}{2}} \end{bmatrix} \leq I. \tag{2.46}$$

The above inequality implies that

$$\left\| \begin{bmatrix} P^{\frac{1}{2}}A'P^{-\frac{1}{2}} \\ \frac{1}{\sigma_1}B'P^{-\frac{1}{2}} \end{bmatrix} \right\| \leq 1. \tag{2.47}$$

Therefore, inequality

$$\left\| \begin{bmatrix} P^{-\frac{1}{2}}AP^{\frac{1}{2}} & \frac{1}{\sigma_1}P^{-\frac{1}{2}}B \end{bmatrix} \right\| \leq 1, \tag{2.48}$$

holds as well. After some straightforward algebraic manipulation, (2.48) is shown to be equivalent to

$$\begin{bmatrix} P^{\frac{1}{2}}A'P^{-1}AP^{\frac{1}{2}} & \frac{1}{\sigma_1}P^{\frac{1}{2}}A'P^{-1}B \\ \frac{1}{\sigma_1}B'P^{-1}AP^{\frac{1}{2}} & \frac{1}{\sigma_1^2}B'P^{-1}B \end{bmatrix} \leq \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \tag{2.49}$$

Identity matrices on the right-hand side of (2.49) are assumed to be of appropriate (possibly different) dimensions. Inequality (2.49), after a congruence transformation by the block diagonal matrix $D = \text{diag}(P^{-\frac{1}{2}}, \sigma_1 I)$, is equivalent to the dissipation inequality (2.44).

Step 2:

Show that the following dissipation inequality holds:

$$\sigma(w, Cx - C\tilde{x}) \geq V(x^+, \tilde{x}^+) - V(x, \tilde{x}), \tag{2.50}$$

where

$$\sigma(w, e) = 4\sigma_1^2|w|^2 - \sigma_2^{-2}|e|^2$$
$$V(x, \tilde{x}) = (x + \tilde{x})'P^{-1}(x + \tilde{x}) + (x - \tilde{x})'Q(x - \tilde{x})$$

To show this, consider the following state space model of the error system $\mathbf{G} - \tilde{\mathbf{G}}$ mapping

w to $e = y - \tilde{y}$:

$$\begin{aligned} x^+ &= \varphi(Ax + Bw), \\ \tilde{x}^+ &= \Theta\varphi(A\tilde{x} + Bw), \\ e &= Cx - C\tilde{x}. \end{aligned} \tag{2.51}$$

The positive definite quadratic form $V = V(x, \tilde{x})$ can be re-written as follows:

$$V(x, \tilde{x}) = x'(P^{-1} + Q)x + 2x'(P^{-1} - Q)\tilde{x} + \tilde{x}'(P^{-1} + Q)\tilde{x}. \tag{2.52}$$

By expanding $V(x^+, \tilde{x}^+)$, we have the sequence of inequalities as shown in (2.53), where shorthand notation $z = Ax + Bw$ and $\tilde{z} = A\tilde{x} + Bw$ was used.

$$\begin{aligned} V(x^+, \tilde{x}^+) &= \varphi(z)'(P^{-1} + Q)\varphi(z) + 2\varphi(z)'(P^{-1} - Q)\Theta\varphi(\tilde{z}) + \varphi(\tilde{z})'\Theta(P^{-1} + Q)\Theta\varphi(\tilde{z}) \\ &\leq \varphi(z)'(P^{-1} + Q)\varphi(z) + 2\varphi(z)'(P^{-1} - Q)\varphi(\tilde{z}) + \varphi(\tilde{z})'(P^{-1} + Q)\varphi(\tilde{z}) \\ &= (\varphi(z) + \varphi(\tilde{z}))'P^{-1}(\varphi(z) + \varphi(\tilde{z})) + (\varphi(z) - \varphi(\tilde{z}))'Q(\varphi(z) - \varphi(\tilde{z})) \\ &\leq (z + \tilde{z})'P^{-1}(z + \tilde{z}) + (z - \tilde{z})'Q(z - \tilde{z}) \\ &\leq 4\sigma_1^2|w| - \sigma_2^{-1}|Cx - C\tilde{x}| - V(x, \tilde{x}) \end{aligned} \tag{2.53}$$

The first inequality used (2.67), the second inequality used (2.19)-(2.20), and the last inequality used (2.44)-(2.45). This implies that (2.50) holds and, furthermore, that $\|\mathbf{G} - \tilde{\mathbf{G}}\| \leq 2\sigma_1\sigma_2$. This concludes the proof.

2.5.4 Proof of Theorem 2.2.7

From the definition of system $\hat{\mathbf{S}}_m$ we have that, for all n:

$$x_k[n+1] = \begin{cases} x_{k+1}[n], & k \in \{-N+1, \ldots, -1\} \\ \text{sat}_r\left(\sum_{i=-N}^{N} h[i]x_{k+1-i}[n] + w[n+k+1]\right), & k \in \{0, \ldots, m-N-1\}, , \\ \text{sat}_r\left(\sum_{i=k-m+1}^{N} h[i]x_{k+1-i}[n] + w[n+k+1]\right), & k \in \{m-N, \ldots, m\} \end{cases} \tag{2.54}$$

For the sake of clarity, we repeat the dynamic equation (2.11) that defines system $\mathbf{S}^*$, and gives explicit relationship between w and y. For the reasons that will become clear shortly, we write (2.11) for time $\tilde{n} = n + k + 1$ as follows:

$$y[n+k+1] = \mathrm{sat}_r \left(\sum_{i=-N}^{N} h[i] y[n+k=1-i] + w[n+k+1] \right). \tag{2.55}$$

Let $\delta \in \ell^2(\mathbb{R}^{m+N})$ be such that

$$\delta_k[n] = |y[n+k] - x_k[n]|, \quad k \in \{-N+1, \ldots, m-1, m\}$$

In order to prove the theorem, it is now sufficient to show that there exist $c > 0$ and $\rho \in (0,1)$ such that

$$\sup \|\delta_0\|_\infty \leq c \cdot \rho^m.$$

where $\delta_0[n] = |y[n] - \hat{y}[n]| = |y[n] - x_0[n]|$ for all $n \in \mathbb{Z}$, and the supremum is taken over all possible input signals w.

To do that, we first observe the the following is true for all $k \in \{-N+1, \ldots, -1\}$:

$$\delta_k[n+1] = |y[n+1+k] - x_k[n+1]| = |y[n+(k+1)] - x_{k+1}[n]| = \delta_{k+1}[n]. \tag{2.56}$$

Also, from (2.54)-(2.55) and the Lipschitz continuity of the saturation function, the following systems of inequalities are true:

$$\delta_k[n+1] \leq \sum_{i=-N}^{N} |h[i]| \delta_{k+1-i}[n], \quad \forall k \in \{0, \ldots, m-N-1\} \tag{2.57}$$

$$\delta_k[n+1] \leq \sum_{i=k-m+1}^{N} |h[i]| \delta_{k+1-i}[n] + r \sum_{i=-N}^{k-m} |h[i]|, \quad \forall k \in \{m-N, \ldots, m\} \tag{2.58}$$

It follows that

$$\delta[n+1] \leq A_p \delta[n] + b, \tag{2.59}$$

where the above inequality should be understood component-wise, and matrices $A_p =$

$\{a_{i,j}^p\} \in \mathbb{R}^{(m+N)\times(m+N)}$, and $b \in \mathbb{R}^{m+N}$ are defined by

$$a_{i,j}^p = |a_{i,j}|, \forall i,j, \qquad b_k = \begin{cases} 0, & k \in \{-N+1, \ldots, m-N-1\} \\ r\sum_{i=-N}^{k-m} |h[i]|, & k \in \{m-N, \ldots, m\} \end{cases}.$$

As a consequence of the Frobenius-Perron theorem (see e.g. chapter 6.3 in [141]) and the L1 dominance of $\mathbf{H}$, there exists a globally asymptotically stable positive equilibrium Δ of (2.59) such that $\Delta = \tilde{A}\Delta + b$ and $\delta[n] \leq \Delta$ for all sufficiently large $n \in \mathbb{Z}$. In order to bound $\delta[n]$ we now find an upper bound on Δ. We first state and prove the following lemma.

Lemma 2.5.3. *If there exists $\Delta^* \in \mathbb{R}^{m+N}$ such that $\Delta^* \geq \mathbf{0}$ and $\Delta^* \geq A_p\Delta^* + b$, then $\Delta \leq \Delta^*$.*

Proof. Let $F : \mathbb{R}^{m+N} \to \mathbb{R}^{m+N}$ be defined by $F\xi = A_p\xi + b$. Map F is clearly a monotonic map due to the positivity of A_p and b. It is also easy to see that $F^K\mathbf{0} = \sum_{i=0}^{K-1} A_p^i b$ converges to Δ as $i \to \infty$, since $\Delta = F\Delta$.

Let $\Delta^* \geq 0$ such that $\Delta^* \geq A_p\Delta^* + b$, and let $\Delta > \Delta^*$. By applying F successively K times on $0 \leq \Delta^*$ and taking the limit $K \to \infty$, it follows that $\Delta \leq \Delta^*$ which is a contradiction. Hence, $\Delta \leq \Delta^*$. □

Let $\Delta^* = \begin{bmatrix} \Delta_{-N+1}^* & \Delta_{-N+2}^* & \ldots & \Delta_m^* \end{bmatrix}^T \in \mathbb{R}^{m+N}$, such that $\Delta_k^* = rc\rho^{m-k}$ for $k \in \{0, \ldots, m\}$ and $\Delta_k^* = rc\rho^m$ otherwise, for some positive scalars c and ρ. We now show that there exist $c > 0$ and $\rho \in (0,1)$ such that the conditions of Lemma 2.5.3 are satisfied. The positivity of Δ^* trivially follows from the positivity of r, c, and ρ. Condition $\Delta^* \geq A_p\Delta^* + b$ is equivalent to the following system of inequalities:

$$\Delta_k^* \geq \Delta_{k+1}^*, \quad k \in \{-N+1, \ldots, -1\}, \tag{2.60}$$

$$\Delta_k^* \geq \sum_{i=-N}^{N} |h[i]|\Delta_{k-i+1}^*, \quad k \in \{0, \ldots, m-N-1\}, \tag{2.61}$$

$$\Delta_k^* \geq \sum_{i=k-m+1}^{N} |h[i]|\Delta_{k-i+1}^* + r\sum_{i=-N}^{k-m} |h[i]|, \quad k \in \{m-N, \ldots, m\}. \tag{2.62}$$

Inequalities (2.60) are true by the definition of Δ^*. When substituting $\Delta_k^* = rc\rho^{m-k}$ in (2.61), we have to consider two cases: $k \in \{0, \ldots, N-1\}$ and $k \in \{N, \ldots, m-N-1\}$. For $k \in \{0, \ldots, N-1\}$, (2.61) is, after some simple algebraic manipulations, equivalent to the following system of inequalities:

$$\rho^{N+1} - \rho^{N+k+1} \sum_{i=k+1}^{N} |h[i]| - \sum_{i=-N}^{k} |h[i]| \rho^{N+i} \geq 0. \tag{2.63}$$

Inequalities (2.61), for $k \in \{N, \ldots, m-N-1\}$, are all equivalent to the following inequality:

$$\rho^{N+1} - \sum_{i=-N}^{N} |h[i]| \rho^{N+i} \geq 0. \tag{2.64}$$

Finally, by substituting $\Delta_k^* = rc\rho^{m-k}$ in (2.62), and after some simple algebraic manipulations, we get the following system of inequalities:

$$c\rho^{m-k} - c \sum_{i=k-m+1}^{N} |h[i]| \rho^{m-k+i-1} - \sum_{i=-N}^{k-m} |h[i]|, \quad k \in \{m-N, \ldots, m\}. \tag{2.65}$$

Let polynomial functions $P_k : \mathbb{R} \to \mathbb{R}$, $k \in \{0, \ldots, m\}$ be defined by

$$P_k(x) = x^{N+1} - \left(\sum_{i=k+1}^{N} |h[i]| \right) x^{N+k+1} - \sum_{i=-N}^{k} |h[i]| x^{N+i}, \quad k \in \{0, \ldots, N-1\}, \tag{2.66}$$

$$P_k(x) = x^{N+1} - \sum_{i=-N}^{N} |h[i]| x^{N+i}, \quad k \in \{N, \ldots, m-N-1\}, \tag{2.67}$$

$$P_k(x) = cx^{m-k} - c \sum_{i=k-m+1}^{N} |h[i]| x^{m-k+i-1} - \sum_{i=-N}^{k-m} |h[i]|, \quad k \in \{m-N, \ldots, m\}, \tag{2.68}$$

It is clear that inequalities (2.63)-(2.65) are equivalent to $P_k(\rho) \geq 0$ for all $k \in \{0, \ldots, m\}$. The following lemma holds.

Lemma 2.5.4. *If* $\sum_{i=-N}^{N} |h[i]| < 1$ *and* P_k *is defined by* (2.66)-(2.68)*, then for each* $k \in \{0, \ldots, m-N-1\}$

(i) there exists $\rho_k \in (0,1)$ *such that* $P_k(\rho_k) = 0$ *and* $P_k(x) > 0$ *for all* $x \in (\rho_k, 1]$,

(ii) $\rho_k \geq \rho_{k+1}$ *for all* $k \in \{0, \ldots, m-N-2\}$.

Proof. Part (i) readily follows since $P_k(0) < 0$ and $P_k(1) > 0$ for all $k \in \{0, \ldots, m-N-1\}$. Assume now that $\rho_k < \rho_{k+1}$. Hence $P_{k+1}(\rho_{k+1}) = 0 < P_k(\rho_{k+1})$, which, by (2.66)-(2.67), implies $P_k(\rho_{k+1}) = (\rho_{k+1} - 1)\rho_{k+1}^{N+k+1} \sum_{i=k+2}^{T} |h[i]| > 0$ and, therefore, $\rho_k > 1$, a contradiction. Hence, (ii) is true. □

Let $\rho_0 \in (0, 1)$ be as defined in Lemma 2.5.4. It immediately follows from Lemma 2.5.4 that $P_k(\rho_0) \geq 0$ for all $k \in \{0, \ldots, m-N-1\}$, and therefore (2.63) and (2.64) are satisfied for all $\rho \in [\rho_0, 1]$, i.e., ρ_0 is an upper bound for ρ.

For $\rho = \rho_0$, the inequalities in (2.65) are equivalent to

$$c \geq \frac{\sum_{i=-N}^{k-m} |h[i]|}{\rho_0^{m-k} - \sum_{i=k-m+1}^{N} |h[i]| \rho_0^{m-k+i-1}}, \quad \forall k \in \{m-N, \ldots, m\}, \tag{2.69}$$

where

$$\rho_0^{m-k} - \sum_{i=k-m+1}^{N} |h[i]| \rho_0^{m-k+i-1} = \rho_0^{m-k-1-N} \left(P_N(\rho_0) + \sum_{i=-N}^{k-m} |h[i]| \rho_0^{N+i} \right) > 0, \tag{2.70}$$

for all $k \in \{m-N, \ldots, m\}$. Let us apply a change of variables $l = m - k$, for all $k \in \{m-N, \ldots, m\}$ in (2.69). If we denote the right-hand side of (2.69) as c_l, the inequalities (2.69) are equivalent to

$$c \geq c_l = \frac{\sum_{i=-N}^{-l} |h[i]|}{\rho_0^l - \sum_{i=1-l}^{N} |h[i]| \rho_0^{l+i-1}}, \tag{2.71}$$

for all $l \in \{0, \ldots, N\}$. Let us denote $c^* = \max_{l \in \{0, \ldots, N\}} c_l$. It follows that (2.65) will be satisfied for $\rho = \rho_0$ and $c = c^*$.

It now follows that $\delta_k[n] \leq \Delta_k \leq \Delta_k^* = r c^* \rho_0^{m-k}$ for all $k \in \{0, \ldots, m\}$, and therefore $|y[n] - \hat{y}[n]| = \delta_0[n] \leq r c^* \rho_0^m$ for all $n \in \mathbb{Z}$. This concludes the proof.

2.5.5 Proof of Theorem 2.2.6

Let us first state and prove the following lemma which will be used in the proof of the main theorem.

Lemma 2.5.5. *Let* $\mathbf{S} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ *be an* N*-banded operator, with symbol* $f \in \ell^2(\mathbb{R})$*, such that* $\|\mathbf{S}\| < 1$*. For* $\gamma \in (0, 1]$*, let* $\mathbf{D}_\gamma \in \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ *be a bounded linear operator such that* $(\mathbf{D}_\gamma w)[n] = \gamma^{|t|} w[n]$ *for all* $w \in \ell^2(\mathbb{R})$*. The bounded linear operator* $\mathbf{S}_\gamma : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ *satisfying* $\mathbf{S}_\gamma \mathbf{D}_\gamma = \mathbf{D}_\gamma \mathbf{S}$ *is well defined and there exists* $\gamma_0 \in (0, 1)$ *such that* $\|\mathbf{S}_\gamma\| < 1$ *for all* $\gamma \in (\gamma_0, 1]$*.*

Proof. It immediately follows from the definition of $\mathbf{S}_\gamma$ that it is an N-banded linear operator. Moreover, it can be easily shown that the matrix $S_\gamma = (s_\gamma^{i,j})$ of $\mathbf{S}_\gamma$, in the standard basis of $\ell^2(\mathbb{R})$, is defined by

$$s_\gamma^{i,j} = \begin{cases} \gamma^{|i|-|j|} f[i-j], & |i-j| \leq N, \\ 0, & \text{otherwise} \end{cases}. \tag{2.72}$$

Boundedness of $\mathbf{S}_\gamma$ immediately follows from (2.72) and the Hölder's inequality:

$$\|\mathbf{S}_\gamma\|^2 \leq \|\mathbf{S}_\gamma\|_1 \|\mathbf{S}_\gamma\|_\infty \leq \left(\gamma^{-N} \sum_{t=-N}^{N} |f[n]| \right)^2 < \infty.$$

Let $g : (0, 1] \to (0, \infty)$ be defined by $\gamma \mapsto g(\gamma) = \|\mathbf{S}_\gamma\|$. Clearly, $g(1) < 1$, since $\|\mathbf{S}\| < 1$. Let $F = F(\Omega)$ be the Fourier transform of f. Since $\|\mathbf{S}\| < 1$ then $|F(\Omega)| < 1$ for all $\Omega \in \mathbb{R}$, and hence $|f[n]| < 1$ for all n. From the definition of the operator norm, we have that

$$\|\mathbf{S}_\gamma\| = \sup_{|u|=|v|=1} |(u, \mathbf{S}_\gamma v)| \geq \sup_{u,v \in E} |(u, \mathbf{S}_\gamma v)| \geq \sup_{i,j \in \mathbb{Z}} |s_\gamma^{i,j}|.$$

Hence, for all $|n| \leq N$, there exist, large enough, $|i|$ and $|j|$ such that $s_\gamma^{i,j} = \gamma^{-|n|} f[n]$. Therefore,

$$\|\mathbf{S}_\gamma\| \geq \max_{|n| \leq N} \gamma^{-|n|} |f[n]|.$$

Let $c = \min\{|f[0]|, \min_{0<|n|\leq N} |f[n]|^{\frac{1}{|n|}}\}$. Then $c \leq \epsilon$ and $g(c) > 1$, so, by the continuity of g, there exists $\gamma_0 \in (c, 1)$ such that $g(\gamma_0) = 1$ and $g(\gamma) < 1$ for all $\gamma \in (\gamma_0, 1)$. This concludes the proof. □

Now we proceed with the proof of the main theorem, where the proof consists of four major steps.

Step 1: Represent system $\mathbf{S}^*$ as a series interconnection of two stable systems: a finite-latency system and a system represented by an infinite-dimensional state space model.

To show this, let $S : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be a forward-shift operator defined by $(Sw)[n] = w[n+1]$ for all $w \in \ell^2(\mathbb{C})$. Let $\mathbf{M}_\infty : \ell^2(\mathbb{C}) \to \ell(\ell^2(\mathbb{C}))$ be the unbounded operator mapping w to $v = \mathbf{M}_\infty w$ such that $v[n] = S^n w$. Let operators $A : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$, $B : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ and $C : \ell^2(\mathbb{C}) \to \mathbb{C}$ be defined by

$$(A\xi)[n] = \sum_{k=-N}^{N} h[k]\xi[n-k+1], \quad B\xi = \xi, \quad C\xi = \xi[0],$$

for all $\xi \in \ell^2(\mathbb{C})$. Clearly, operator A is a $(N+1)$-banded Laurent operator whose symbol $f = f[n]$ is defined by $f[n] = h[n-1]$, for all $n \in \mathbb{Z}$. This implies that A is a contraction, due to $|H(\Omega)| < 1$ for all $\Omega \in \mathbb{R}$ [142]. Let now system $\mathbf{T}_\infty : \ell(\ell^2(\mathbb{C})) \to \ell^2(\mathbb{C})$, mapping v to $y = \mathbf{T}_\infty v$, be defined by the following infinite-dimensional state space model

$$\mathbf{T}_\infty : x[n+1] = \mathrm{Sat}_r(Ax[n] + Bv[n]), \;\; y[n] = Cx[n]. \tag{2.73}$$

It immediately follows, from the definition (2.11) of the optimal map $\mathbf{S}^*$ and the above construction of $\mathbf{M}_\infty$ and $\mathbf{T}_\infty$, that $\mathbf{S}^* = \mathbf{T}_\infty \mathbf{M}_\infty$ (see Figure 2-5(a)).

A necessary assumption for the generalized balanced truncation algorithm from Lemma 2.2.3 is that the system to be approximated is driven by square summable signals. Due to unboundedness of $\mathbf{M}_\infty$, the input of $\mathbf{T}_\infty$ is not square summable, and Lemma 2.2.3 cannot be directly used to establish useful bounds on approximation error. In order to mitigate this, we introduce a suitable coordinate re-scaling as follows.

Let $D : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be a bounded linear operator such that $(Dw)[n] = \gamma_0^{|n|} w[n]$,

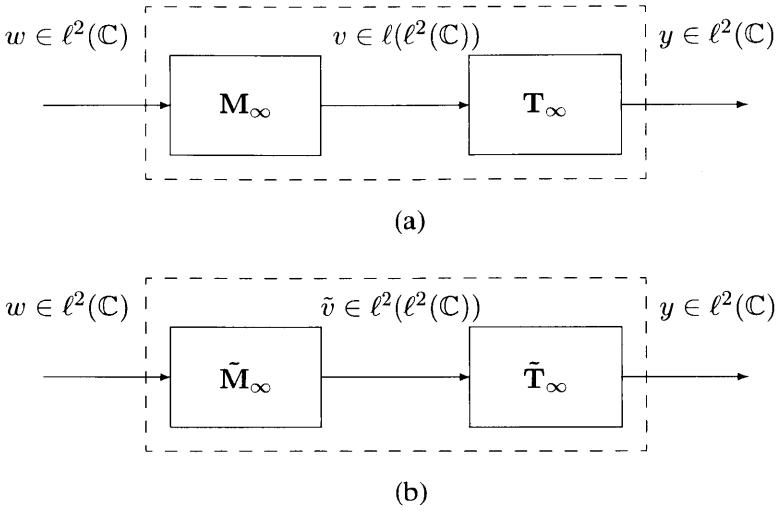


Figure 2-5: Equivalent representations of the optimal system $\mathbf{S}^*$: a) $\mathbf{S}^* = \mathbf{T}_\infty \mathbf{M}_\infty$ (subsystem $\mathbf{M}_\infty$ is unbounded) and b) $\mathbf{S}^* = \tilde{\mathbf{T}}_\infty \tilde{\mathbf{M}}_\infty$ (subsystem $\tilde{\mathbf{M}}_\infty$ is bounded)

$\forall w \in \ell^2(\mathbb{C})$, where $\gamma_0 \in (0,1)$ and $\|DAD^{-1}\| < 1$. The existence of such γ_0 is guaranteed by Lemma 2.5.5. If we apply coordinate transformation $\tilde{x} = Dx$ to (2.73), we get

$$\mathbf{T}_\infty : \tilde{x}[n+1] = \varphi(\tilde{A}\tilde{x}[n] + \tilde{B}Dv[n]), \;\; y[n] = \tilde{C}\tilde{x}[n], \tag{2.74}$$

where $\varphi = D\,\mathrm{Sat}_r\, D^{-1}$, $\tilde{A} = DAD^{-1}$, $\tilde{B} = DBD^{-1} = I$ and $\tilde{C} = CD^{-1} = C$. Since operators D and Sat_r are both diagonal operators, and Sat_r has Lipschitz constant equal to 1, it follows that the Lipschitz constant of the operator φ is also equal to 1. In the rest of this proof, we assume that $\delta \in (0,1)$ is such that $\|\tilde{A}\|^2 \leq 1 - \delta < 1$.

Let $\tilde{\mathbf{M}}_\infty : \ell^2(\mathbb{C}) \to \ell^2(\ell^2(\mathbb{C}))$, mapping w to $\tilde{v}$, be such that $\tilde{v}[n] = DS^n w$. Consider a system $\tilde{\mathbf{T}}_\infty$ described by the following state space model

$$\tilde{\mathbf{T}}_\infty : \tilde{x}[n+1] = \varphi(\tilde{A}\tilde{x}[n] + \tilde{B}\tilde{v}[n]), \;\; y[n] = \tilde{C}\tilde{x}[n], \tag{2.75}$$

It now clearly follows that system $\mathbf{S}^*$ can be represented as a series interconnection $\mathbf{S}^* = \tilde{\mathbf{T}}_\infty \tilde{\mathbf{M}}_\infty$ of $\tilde{\mathbf{M}}_\infty$ and $\tilde{\mathbf{S}}_\infty$ (see Figure 2-5(b)). It is not hard to show that $\tilde{\mathbf{M}}_\infty$ is bounded

and $||\tilde{\mathbf{M}}_\infty|| = \left(\frac{1+\gamma_0^2}{1-\gamma_0^2}\right)^{\frac{1}{2}}$ Indeed, for $\tilde{v} = \tilde{\mathbf{M}}_\infty w$, we have that

$$|\tilde{v}|^2 = \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \gamma_0^{2|n-k|} |w[k]|^2 = \frac{1+\gamma_0^2}{1-\gamma_0^2} |w|^2.$$

Step 2: Represent system $\hat{\mathbf{S}}_m$ as a series interconnection of two stable systems: a finite-latency system and a system represented by an infinite-dimensional state space model.

Let $\Theta_m : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be a projection operator such that $(\Theta_m w)[n] = w[n]$ for $-m+1 \leq n \leq m+1$, and $(\Theta_m w)[n] = 0$ otherwise. Let $\hat{X} = \{\Theta_m u : u \in \ell^2(\mathbb{C})\}$. Consider a system $\hat{\mathbf{T}}_{m,\infty} : \ell^2(\ell^2(\mathbb{C})) \to \ell^2(\mathbb{C})$, mapping $\tilde{v}$ to $\hat{y}$, defined by the following infinite-dimensional state space model

$$\hat{\mathbf{T}}_{m,\infty} : \hat{x}[n+1] = \Theta_m \varphi(\tilde{A}\hat{x}[n] + \tilde{B}\tilde{v}[n]),\ \hat{y}[n] = \tilde{C}\hat{x}[n]. \tag{2.76}$$

where $\hat{x}[n] \in \hat{X}$ for all $nS \in \mathbb{Z}$.
It is not hard to see that $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_{m,\infty}\tilde{\mathbf{M}}_\infty$ (see Figure 2-6(b)). Indeed, the state space model of $\hat{\mathbf{T}}_{m,\infty}$ is formally infinite-dimensional but in fact only $2m+1$ state components are nonzero, and those exactly correspond to the state variables of the subsystem $\hat{\mathbf{T}}_m$ of $\hat{\mathbf{S}}_m$.

Step 3: Find $\sigma_1 > 0$, $\sigma_2 > 0$, $P : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$, and $Q : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ that satisfy the assumptions of Lemma 2.2.3 for $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, Θ_m, and φ as given above.

To find this, let us first assume that $\delta \in (0,1)$, as chosen in Step 1, is such that $\|\tilde{A}\|^2 \leq 1-\delta$. It immediately follows that inequality $\tilde{A}\tilde{A}' \leq I - \delta I$ holds. Moreover, since $\tilde{B} = I$, the inequality $I - \tilde{A}I\tilde{A}' \geq \delta \tilde{B}\tilde{B}'$ holds as well. It now follows that $\sigma_1 = \frac{1}{\sqrt{\delta}}$ and $P = I$ satisfy the first inequality in (2.17).

For an arbitrary, but fixed, $\rho_0 \in (0,1)$, let $Q : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be defined as $(Qw)[n] = \rho_0^{|n|-m} w[n]$ for $|n| \leq m-1$, and $(Qw)[n] = w[n]$ otherwise, for all $w \in \ell^2(\mathbb{C})$. Similar to the proof of Lemma 2.5.5, it can be shown that there exist $\delta_0 \in (0,1)$ and $\rho_0 \in (0,1)$ (ρ_0 does not depend on m) such that $\|Q^{1/2}\tilde{A}Q^{-1/2}\|^2 \leq 1 - \delta_0 < 1$. This implies that

$$I - Q^{-1/2}\tilde{A}'Q\tilde{A}Q^{-1/2} \geq \delta_0 I,$$

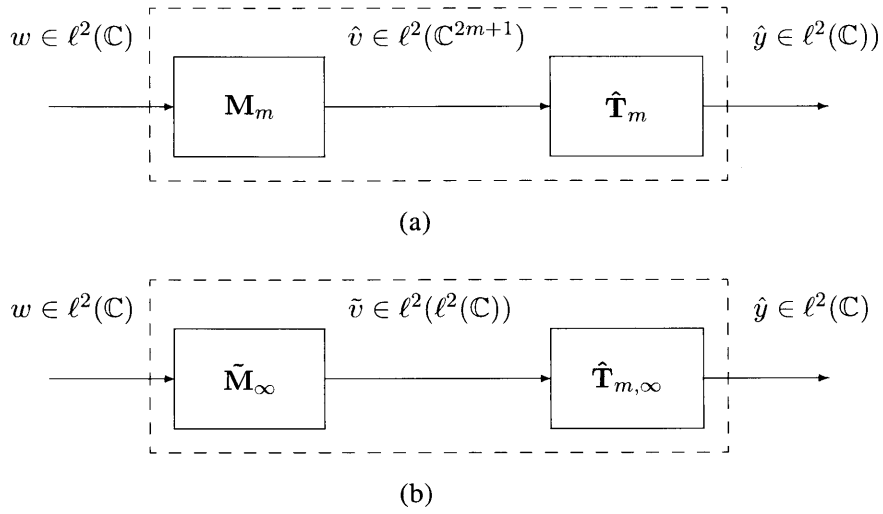


Figure 2-6: Equivalent representations of the approximate system $\hat{\mathbf{S}}_m$: a) $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_m \mathbf{M}_m$ (state-space model $\hat{\mathbf{T}}_m$ is finite dimensional) and b) $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_{m,\infty}\tilde{\mathbf{M}}_\infty$ (state-space model $\hat{\mathbf{T}}_{m,\infty}$ is infinite dimensional).

and, moreover,

$$Q - \tilde{A}'Q\tilde{A} \geq \delta_0 Q \geq \frac{\delta_0}{\rho_0^m}\tilde{C}'\tilde{C},$$

Therefore, the above defined Q and $\sigma_2 = \sqrt{\frac{\rho_0^m}{\delta_0}}$ satisfy the second inequality in (2.17).

From the definition of P, Q, and Θ_m it immediately follows that (2.67) is true, while (2.19) and (2.20) follow from the fact that φ is diagonal and has Lipschitz constant equal to 1, and P and Q are positive definite diagonal operators.

Step 4: The following series of inequalities hold

$$\begin{aligned}
\|\mathbf{S}^* - \hat{\mathbf{S}}_m\| &= \|(\tilde{\mathbf{T}}_\infty - \hat{\mathbf{T}}_{m,\infty})\tilde{\mathbf{M}}_\infty\| \\
&\leq \|\tilde{\mathbf{T}}_\infty - \hat{\mathbf{T}}_{m,\infty}\|\|\tilde{\mathbf{M}}_\infty\| \\
&\leq \left(\frac{4}{\delta\delta_0} \cdot \frac{1+\gamma_0^2}{1-\gamma_0^2}\right)^{\frac{1}{2}} \rho_0^{m/2}.
\end{aligned}$$

Therefore, $\rho = \sqrt{\rho_0}$ and $c = \left(\frac{4}{\delta\delta_0} \cdot \frac{1+\gamma_0^2}{1-\gamma_0^2}\right)^{\frac{1}{2}}$ satisfy the condition of Theorem 2.2.6. This concludes the proof.

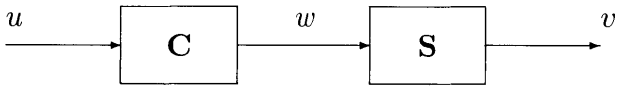
Chapter 3

Equivalent Baseband Modeling and Digital Compensation of Dynamic Passband Nonlinearities in Phase-Amplitude Modulation-Demodulation Schemes

This chapter is organized as follows. In section 3.1 we give motivation for the problem of equivalent baseband modeling and digital predistortion, and give full mathematical description of the system under consideration. Main result is stated in section 3.2, in which an explicit expression of the equivalent baseband model is given. In section 3.3 we provide some additional discussion on advantages and disadvantages of a DPD based on this model. We also show that the analysis readily extends to OFDM modulation. System model validation, as well as a digital predistorter design and its performance are demonstrated by MATLAB simulation results presented in section 3.4. We summarize the results of this chapter in section 3.5. The proof of the main result (theorem 3.2.1) is presented in section 3.6.

3.1 Problem Formulation

In this thesis, a digital compensator is viewed as a system $\mathbf{C} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$. More specifically, a pre-compensator $\mathbf{C} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ designed for a device modeled by a system $\mathbf{S} : \ell^2(\mathbb{C}) \to \mathcal{L}^2(\mathbb{R})$ (or $\mathbf{S} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$) aims to make the composition $\mathbf{SC}$, as shown on the block diagram below,



conform to a set of desired specifications (in the simplest case, the objective is to make $\mathbf{SC}$ as close to the identity map as possible, in order to cancel the distortions introduced by $\mathbf{S}$).

A common element in digital compensator design algorithms is selection of *compensator structure*, which usually means specifying a finite sequence $\tilde{\mathbf{C}} = (\mathbf{C}_1, \ldots \mathbf{C}_N)$ of systems $\mathbf{C}_k : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$, and restricting the actual compensator $\mathbf{C}$ to have the form

$$\mathbf{C} = \sum_{k=1}^{N} a_k \mathbf{C}_k, \qquad a_k \in \mathbb{C},$$

i.e., to be a linear combination of the elements of $\tilde{\mathbf{C}}$. Once the *basis* sequence $\tilde{\mathbf{C}}$ is fixed, the design usually reduces to a straightforward *least squares optimization* of the coefficients $a_k \in \mathbb{C}$. A popular choice is for the systems $\mathbf{C}_k$ to be some *Volterra monomials*, i.e., to map their input $u = u[n]$ to the outputs $w_k = w_k[n]$ according to the polynomial formulae

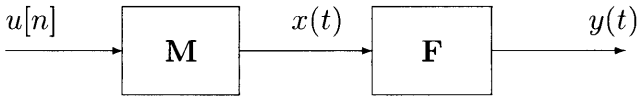
$$w_k[n] = \prod_{j=1}^{d_r(k)} \mathrm{Re}\, u[n - n^r_{k,j}] \prod_{j=1}^{d_i(k)} \mathrm{Im}\, u[n - n^i_{k,j}],$$

where the integers $d_r(k)$ and $d_i(k)$ (respectively, $n^r_{k,j}$ and $n^i_{k,j}$) will be referred to as *degrees* (respectively, *delays*). In this case, every linear combination $\mathbf{C}$ of $\mathbf{C}_k$ is a *DT Volterra series* [143], i.e., a DT system mapping signal inputs $u \in \ell^2(\mathbb{C})$ to outputs $w \in \ell^2(\mathbb{C})$ according to the polynomial expression

$$w[n] = \sum_{k=1}^{N} a_k \prod_{j=1}^{d_r(k)} \mathrm{Re}\, u[n - n^r_{k,j}] \prod_{j=1}^{d_i(k)} \mathrm{Im}\, u[n - n^i_{k,j}].$$

As previously mentioned, selecting a proper compensator structure is a major challenge in compensator design: a basis which is too simple will not be capable of canceling the distortions well, while a form that is too complex will consume excessive power and space in its hardware implementation. Having an insight into the compensator basis selection can be very valuable.

In this thesis, we establish that a certain special structure is good enough to compensate for imperfect power amplification. We consider modulation systems represented by the block diagram



where $\mathbf{M} : \ell^2(\mathbb{C}) \to \mathcal{L}^2(\mathbb{R})$ is the ideal modulator, and $\mathbf{F} : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ is a CT dynamical system used to represent linear and nonlinear distortion in the modulator and power amplifier circuits. We consider the ideal modulator of the form $\mathbf{M} = \mathbf{XZ}$, where $\mathbf{Z} : \ell^2(\mathbb{C}) \to \mathcal{L}^2(\mathbb{C})$ is the *zero order hold* (ZOH) map $u[\cdot] \mapsto x_0(\cdot)$:

$$x_0(t) = \sum_n p(t - nT)u[n], \quad p(t) = \begin{cases} 1, & t \in [0, T), \\ 0, & t \notin [0, T) \end{cases} \tag{3.1}$$

with fixed sampling interval length $T > 0$ and $\mathbf{X} : \mathcal{L}^2(\mathbb{C}) \to \mathcal{L}^2(\mathbb{R})$ is the *mixer* map

$$x_0(\cdot) \mapsto x(\cdot) : \quad x(t) = 2\mathrm{Re}[\exp(j\omega_c t)x_0(t)] \tag{3.2}$$

with modulation-to-sampling frequency ratio $M \in \mathbb{N}$, i.e., with $\omega_c = 2\pi M/T$. It should be noted that in real applications, there is no constraint on M to be integer valued. We assume $M \in \mathbb{N}$ for convenience, since it simplifies, already complex, derivations as will be seen in section 3.2. For non-integer valued M, time-invariance of certain subsystems is lost, but, in most applications, the introduced time-dependence is not significant enough to impact the overall system structure.

It is commonly assumed that the modulation/digital-to-analog conversion (DAC) is realized in a way which includes low-pass filtering of the DAC output. As has been described in

section 1, there has been a significant body of work recently focusing on the so called "all-digital transmitters" (ADT) that employ ZOH modulation and have no specific low-pass filtering of the DAC output. ADT utilizes power amplifiers in switched-mode operation (SMPA) which, due to their specific mode of operation (transistors work as switches), require piece-wise constant input signals. In such a setup, ZOH modulation is a natural method for converting digital RF/baseband into analog signals suitable for driving SMPAs. In Section 3.3, we discuss how results similar to the ones stated and proven under assumption of a ZOH modulator, hold for more general system models which include low-pass filtering after ZOH.

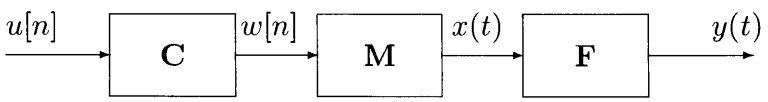
We are particularly interested in the case when $\mathbf{F}$ is described by the *CT Volterra series model*

$$y(t) = b_0 + \sum_{k=1}^{N_b} b_k \prod_{i=1}^{\beta_k} x(t - t_{k,i}), \tag{3.3}$$

where $N_b \in \mathbb{N}$, $b_k \in \mathbb{R}$, $\beta_k \in \mathbb{N}$, $t_{k,i} \geq 0$ are parameters. (In a similar fashion, it is possible to consider input-output relations in which the finite sum in (3.3) is replaced by an integral/infinite sum). One expects that the memory of $\mathbf{F}$ is not long, compared to T, i.e., that $\max t_{k,i}/T$ is not much larger than 1.

As a rule, the spectrum of the DT input $u \in \ell^2(\mathbb{C})$ of the modulator is carefully shaped at a pre-processing stage to guarantee desired characteristics of the modulated signal $x = \mathbf{M}u$. However, when the distortion $\mathbf{F}$ is not linear, the spectrum of the $y = \mathbf{F}x$ could be damaged substantially, leading to violations of error vector magnitude (EVM) and adjacent channel leakage ratio (ACLR) specifications, as defined in the previous chapter [72].

We consider the possibility of repairing the spectrum of y by pre-distorting the digital input $u \in \ell^2(\mathbb{C})$ by a compensator $\mathbf{C} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$, as shown on the block diagram below:



The desired effect of inserting $\mathbf{C}$ is cancellation of the distortion caused by $\mathbf{F}$. Naturally, since $\mathbf{C}$ acts in the baseband (i.e., in discrete time), there is no chance that $\mathbf{C}$ will achieve

a complete correction, i.e., that the series composition **FMC** of **F**, **M**, and **C** will be identical to **M**. However, in principle, it is sometimes possible to make the frequency contents of $\mathbf{M}u$ and $\mathbf{FMC}u$ to be identical within the CT frequency band $(\omega_c - \omega_b, \omega_c + \omega_b)$, where $\omega_b = \pi/T$ is the Nyquist frequency [144], [145]. To this end, let $\mathbf{H} : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ denote the ideal band-pass filter with frequency response $H(\omega) = 1$ for $|\omega_c - |\omega|| < \omega_b$ and $H(\omega) = 0$ otherwise. Let $\mathbf{D} : \mathcal{L}^2(\mathbb{R}) \to \ell^2(\mathbb{C})$ be the ideal de-modulator operating on the band selected by **H**, i.e., the linear system for which the series composition **DHM** is the identity function. Let $\mathbf{S} = \mathbf{DHFM}$ be the series composition of **D**, **H**, **F**, and **M**, i.e., the DT system with input $w = w[n]$ and output $v = v[n]$ shown on the block diagram in Fig. 3-1.

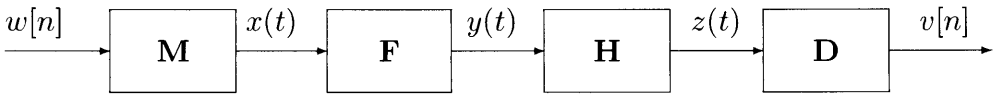


Figure 3-1: Block diagram of $\mathbf{S} = \mathbf{DHFM}$.

By construction, the ideal compensator **C** should be the inverse $\mathbf{C} = \mathbf{S}^{-1}$ of **S**, as long as the inverse does exist. A key question answered in this thesis is "what to expect from system **S**?". If one assumes that the continuous-time distortion subsystem **F** is simple enough, what does this say about **S**?

In the following section, we provide an explicit expression for **S** in the case when **F** is given in the CT Volterra series form (3.3) with degree $d = \max \beta_k$ and depth $t_{max} = \max t_{k,i}$. The result reveals that, even though **S** tends to have infinitely long memory (due to the ideal band-pass filter **H** being involved in the construction of **S**), it can be represented as a series interconnection $\mathbf{S} = \mathbf{LV}$, where $\mathbf{V} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{R}^N)$ maps complex scalar-valued input $w \in \ell^2(\mathbb{C})$ to real vector-valued output $s \in \ell^2(\mathbb{R}^N)$ in such a way that the k-th scalar component $s_k[n]$ of $s[n] \in \mathbb{R}^N$ is given by

$$s_k[n] = \prod_{i=0}^{m} (\mathrm{Re}\, w[n-i])^{\alpha_i} \prod_{i=0}^{m} (\mathrm{Im}\, w[n-i])^{\beta_i}, \quad \alpha_i, \beta_i \in \mathbb{Z}_+, \quad \sum_{i=0}^{m} \alpha_i + \sum_{i=0}^{m} \beta_i \leq d,$$

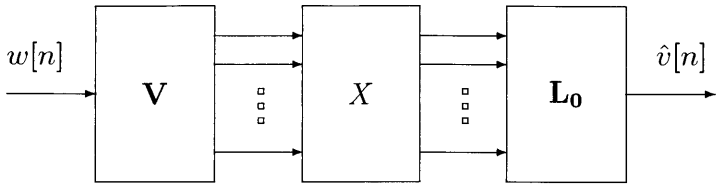


Figure 3-2: System $\mathbf{S}$ can be well approximated by the model shown in this block diagram.

where m is the minimal integer not smaller than t_{max}/T, and $\mathbf{L} : \ell^2(\mathbb{R}^N) \to \ell^2(\mathbb{C})$ is an LTI system. Moreover, $\mathbf{L}$ can be shown to have a good approximation of the form $\mathbf{L} \approx \mathbf{L}_0 X$, where X is a static gain matrix, and $\mathbf{L}_0$ is an LTI model which does not depend on b_k and $t_{k,i}$ (see Fig. 3-2). In other words, $\mathbf{S}$ can be well approximated by combining a Volterra series model with short memory, and a *fixed* (long memory) LTI system, as long as the memory depth t_{max} of $\mathbf{F}$ is short, relative to the sampling time T.

In most applications, with an appropriate scaling and time delay, the system $\mathbf{S}$ to be inverted can be viewed as a small perturbation of identity, i.e., $\mathbf{S} = \mathbf{I} + \mathbf{\Delta}$. When $\mathbf{\Delta}$ is "small" in an appropriate sense (e.g., has small incremental L2 gain[1] $\|\mathbf{\Delta}\| \ll 1$), the inverse of $\mathbf{S}$ can be well approximated by $\mathbf{S}^{-1} \approx \mathbf{I} - \mathbf{\Delta} = 2\mathbf{I} - \mathbf{S}$. Hence the result of this chapter suggests a specific structure of the compensator (pre-distorter) $\mathbf{C} \approx \mathbf{I} - \mathbf{\Delta} = 2\mathbf{I} - \mathbf{S}$. In other words, a plain Volterra monomials structure is, in general, not good enough for $\mathbf{C}$, as it lacks the capacity to implement the long-memory LTI post-filter $\mathbf{L}$. Instead, $\mathbf{C}$ should be sought in the form $\mathbf{C} = \mathbf{I} - \mathbf{L}_0 X \mathbf{V}$, where $\mathbf{V}$ is the system generating all Volterra series monomials of a limited depth and limited degree, $\mathbf{L}_0$ is a *fixed* LTI system with a very long time constant, and X is a matrix of coefficients to be optimized to fit the data available.

[1]Incremental L2 gain (as well as other similarly defined system "gains") can be viewed as a measure of "size" of a system, or more precisely as a measure of sensitivity of a system response to its input. With Euclidean norm (i.e., L2-norm) $\|\xi\|$ of an element $\xi = [\xi_1 \quad \dots \quad \xi_k] \in \mathbb{C}^k$ defined by $\|\xi\|^2 = \sum_{i=1}^k |\xi_i|^2$, the incremental L2 gain $\|\mathbf{G}\|$ of a DT system $\mathbf{G} : \ell^2(\mathbb{C}^n) \to \ell^2(\mathbb{C}^m)$ can be defined as the maximal upper bound on $\gamma \geq 0$ such that

$$\sum_{n\in\mathbb{Z}} \|y_1[n] - y_2[n]\|^2 \leq \gamma^2 \sum_{n\in\mathbb{Z}} \|w_1[n] - w_2[n]\|^2,$$

for all $w_1, w_2 \in \ell^2(\mathbb{C}^n)$, where $y_1 = \mathbf{G}w_1$ and $y_2 = \mathbf{G}w_2$.

3.2 Main Result

In this section, we first describe the process of ideal demodulation that we consider in this thesis. We then present the main result, that is, we derive the equivalent baseband model of the system S described in the previous section.

3.2.1 Ideal Demodulator

By definition, demodulator D should "invert" the operation of M, i.e., $\mathbf{DM} = \mathbf{I}$ should hold. In communications literature, demodulation is usually described as downconversion of the passband signal, followed by low-pass filtering (anti-aliasing filter) and sampling [146]. When modulator system M produces output signal with significant side-lobes (as is the case with ZOH) the above described demodulation process leads to $\mathbf{DM} \neq \mathbf{I}$. If the ratio $M = f_c/f_s$ is very large, the in-band spectral distortion that results from the folding of the ZOH side-lobes would be negligible compared to the signal in-band energy, which implies $\mathbf{DM} \approx \mathbf{I}$. As M decreases, this distortion becomes more notable and might match that caused by the PA nonlinearity. This implies that distortions introduced by the non-ideal demodulation could mask good performance of a digital pre-distorter. For that reason, in this thesis, we apply demodulation which recovers the input signal, without introducing distortion. We call this operation ideal demodulation, and in the following derive its mathematical model.

The most commonly known expression for the ideal demodulator inverts not $\mathbf{M} = \mathbf{XZ}$ but $\mathbf{M}_0 = \mathbf{XH}_0\mathbf{Z}$, i.e., the modulator which inserts $\mathbf{H}_0$, the *ideal low-pass filter* for the baseband, between zero-order hold $\mathbf{Z}$ and mixer $\mathbf{X}$, where $\mathbf{H}_0$ is the CT LTI system with frequency response $H_0(\omega) = 1$ for $|\omega| < \omega_b$ and $H_0(\omega) = 0$ otherwise. Specifically, let $\mathbf{X}_c : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{C})$ be the *dual mixer* mapping $x(\cdot)$ to $e(t) = \exp(-j\omega_c t)x(t)$. Let $\mathbf{E} : \mathcal{L}^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ be the *sampler* map $g(\cdot) \mapsto w[\cdot]$ such that $w[n] = g(nT)$. Finally, let $\mathbf{A}_0$ be the DT LTI system with frequency response A_0 defined by $A_0(\Omega) = P(\Omega/T)^{-1}$ for $|\Omega| < \pi$, where P is the Fourier transform of $p = p(t)$ (3.1). Then the composition $\mathbf{A}_0\mathbf{E}\mathbf{H}_0\mathbf{X}_c\mathbf{H}\mathbf{M}_0$ is an identity map. Equivalently, $\mathbf{A}_0\mathbf{E}\mathbf{H}_0\mathbf{X}_c$ is the ideal demodulator for $\mathbf{M}_0$.

For the modulation map $\mathbf{M} = \mathbf{XZ}$ considered in this thesis, the ideal demodulator has the form $\mathbf{AEH}_0\mathbf{X}_c$, where $\mathbf{A} : \ell^2(\mathbb{C}) \to \ell^2(\mathbb{C})$ is the linear system mapping $w \in \ell^2(\mathbb{C})$ to $s \in \ell^2(\mathbb{C})$ according to

$$\begin{aligned} \mathrm{Re}(s) &= \mathbf{A}_{rr}\mathrm{Re}(w) + \mathbf{A}_{ri}\mathrm{Im}(w), \\ \mathrm{Im}(s) &= \mathbf{A}_{ir}\mathrm{Re}(w) + \mathbf{A}_{ii}\mathrm{Im}(w), \end{aligned} \tag{3.4}$$

and $\mathbf{A}_{rr}$, $\mathbf{A}_{ri}$, $\mathbf{A}_{ir}$, $\mathbf{A}_{ii}$ are LTI systems with frequency responses $A_{rr} = (P_0 - P_i)Q$, $A_{ir} = A_{ri} = -P_qQ$, $A_{ii} = (P_0 + P_i)Q$, where $Q = (P_0^2 - P_i^2 - P_q^2)^{-1}$, $P_i = (P^+ + P^-)/2$, $P_q = (P^+ - P^-)/2j$, and 2π-periodic functions $P_0, P^+, P^- : \mathbb{R} \to \mathbb{C}$ are defined for $|\Omega| < \pi$ by

$$P_0(\Omega) = P(\Omega/T), \qquad P^+(\Omega) = P_0(\Omega + \theta), \qquad P^-(\Omega) = P_0(\Omega - \theta)$$

with $\theta = 4\pi M$.

3.2.2 Equivalent Baseband Model

Before stating the main result of this chapter, let us introduce some additional notation. For $d \in \mathbb{N}$ and $\tau = (\tau_1, \ldots, \tau_d) \in [0, \infty)^d$ let $\mathbf{F}_\tau : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ be the CT system mapping input $x \in \mathcal{L}^2(\mathbb{R})$ to the output $y \in \mathcal{L}^2(\mathbb{R})$ according to

$$y(t) = x(t - \tau_1)x(t - \tau_2) \ldots x(t - \tau_d). \tag{3.5}$$

In the rest of this section, many expressions will contain products of the above type, where the complex-valued signal x can be written as $x = I + jQ$, with I and Q representing its real and imaginary part, respectively. Therefore, such terms, as in (3.5), would correspond to products of delayed real and imaginary parts of x. As will be shown later (e.g., in (3.15)), the factors in these products can be classified into four groups: real and imaginary parts of two differently delayed versions of x. This explains appearance of the index set $[1 : 4]$ which will be used to encode these four groups of signals.

For every tuple $\mathbf{m} = (m_1, \ldots, m_d) \in [1 : 4]^d$ and integer $l \in [1 : 4]$ let $S^l_{\mathbf{m}}$ be the set of

all indices i for which $m_i = l$, i.e., $S^l_{\mathbf{m}} = \{i \in [1:d] : \ m_i = l\}$. Furthermore, define

$$N^1_{\mathbf{m}} = |S^1_{\mathbf{m}} \cup S^2_{\mathbf{m}}|, \quad N^2_{\mathbf{m}} = |S^3_{\mathbf{m}} \cup S^4_{\mathbf{m}}|.$$

Clearly $N^1_{\mathbf{m}} + N^2_{\mathbf{m}} = d$ for every $\mathbf{m} \in [1:4]^d$. Let $R^c_{\mathbf{m}} = \{-1,1\}^{N^1_{\mathbf{m}}}$ and $R^s_{\mathbf{m}} = \{-1,1\}^{N^2_{\mathbf{m}}}$. Let $(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ denote the standard scalar product in $\mathbb{R}^d$. Define the maps $\tilde{\sigma}, \sigma : \mathbb{R}^d \to \mathbb{R}$ by $\tilde{\sigma}(x) = \sum_{i=1}^d x_i$ and $\sigma(x) = \tilde{\sigma}(x) - 1$. For a given $\mathbf{m} \in [1:4]^d$ let $\Pi : \mathbb{R}^{N^2_{\mathbf{m}}} \to \mathbb{R}$ be defined as $\Pi(x) = \prod_{i=1}^{N^2_{\mathbf{m}}} x_i$. For $i \in \{1,2\}$, define projection operators $\mathcal{P}^i_{\mathbf{m}} : \mathbb{R}^d \to \mathbb{R}^{N^i_{\mathbf{m}}}$ by $\mathcal{P}^i_{\mathbf{m}} x = \begin{bmatrix} x_{n_1} & \dots & x_{n_{N^i_{\mathbf{m}}}} \end{bmatrix}^T$ where $\{n_1, \dots, n_{N^i_{\mathbf{m}}}\} = S^{2i-1}_{\mathbf{m}} \cup S^{2i}_{\mathbf{m}}, \quad n_1 < \dots < n_{N^i_{\mathbf{m}}}$. The following example should elucidate the above, somewhat involved, notation. Let $d = 7$, $\mathbf{m} = (3,1,4,2,1,3,1) \in [1:4]^7$. Then

$$S^1_{\mathbf{m}} = \{2,5,7\}, \quad S^2_{\mathbf{m}} = \{4\}, \quad S^3_{\mathbf{m}} = \{1,6\}, \quad S^4_{\mathbf{m}} = \{3\},$$

$$N^1_{\mathbf{m}} = |S^1_{\mathbf{m}} \cup S^2_{\mathbf{m}}| = 4, \quad N^2_{\mathbf{m}} = |S^3_{\mathbf{m}} \cup S^4_{\mathbf{m}}| = 3,$$

$$R^c_{\mathbf{m}} = \{-1,1\}^4, \quad R^2_{\mathbf{m}} = \{-1,1\}^3,$$

$$\mathcal{P}^1_{\mathbf{m}} x = \begin{bmatrix} x_2 & x_4 & x_5 & x_7 \end{bmatrix}^T, \quad \mathcal{P}^2_{\mathbf{m}} x = \begin{bmatrix} x_1 & x_3 & x_6 \end{bmatrix}^T.$$

Given a vector $\tau \in [0,\infty)^d$ let $\mathbf{k}$ be the unique vector in $(\mathbb{N} \cup \{0\})^d$ such that $\tau = \mathbf{k}T + \tau'$ and $\tau' \in [0,T)^d$.

Let $\theta : \mathbb{R} \to \{0,1\}$ denote the Heaviside step function $\theta(t) = 0$ for $t < 0$, $\theta(t) = 1$ for $t \geq 0$. For $T \in (0,\infty)$ let $p(t) = \theta(t) - \theta(t-T)$ denote the basic pulse shape of the zero-order hold (ZOH) system with sampling time T. Given $\mathbf{m} \in [1:4]^d$ and $\tau' \in [0,T)^d$ define

$$\tau^{\mathbf{m}}_{min} = \begin{cases} \max_{i \in S^2_{\mathbf{m}} \cup S^4_{\mathbf{m}}} \tau'_i, & |S^2_{\mathbf{m}} \cup S^4_{\mathbf{m}}| > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{3.6}$$

and

$$\tau^{\mathbf{m}}_{max} = \begin{cases} \min_{i \in S^1_{\mathbf{m}} \cup S^3_{\mathbf{m}}} \tau'_i, & |S^1_{\mathbf{m}} \cup S^3_{\mathbf{m}}| > 0, \\ T, & \text{otherwise.} \end{cases} \tag{3.7}$$

Let $p_{\mathbf{m},\tau} : \mathbb{R} \to \mathbb{R}$ be the continuous time signal defined by

$$p_{\mathbf{m},\tau}(t) = \begin{cases} \theta(t - \tau^{\mathbf{m}}_{min}) - \theta(t - \tau^{\mathbf{m}}_{max}), & \tau^{\mathbf{m}}_{min} < \tau^{\mathbf{m}}_{max} \\ 0, & \text{otherwise,} \end{cases} . \tag{3.8}$$

We denote its Fourier transform by $P_{\mathbf{m},\tau}(\omega)$.

As can be seen from (3.3), the general CT Volterra model is a linear combination of subsystems $\mathbf{F}_\tau$, with different values of τ. Therefore, in order to establish the desired decomposition $\mathbf{S} = \mathbf{LV}$ it is sufficient to consider the case $\mathbf{S}_\tau = \mathbf{DHF}_\tau\mathbf{M}$ with an arbitrary, but fixed, τ. The following theorem provides description of the underlying subsystems $\mathbf{V}$ and $\mathbf{L}$ of the aforementioned decomposition.

Theorem 3.2.1. *For $\tau \in [0,\infty)^d$, the system $\mathbf{DHF}_\tau\mathbf{M}$ maps $w \in \ell^2(\mathbb{C})$ to*

$$v = \mathbf{A}u \in \ell^2(\mathbb{C}), \quad \textit{with} \quad u = \sum_{\mathbf{m}\in[1:4]^d} s_{\mathbf{m},\mathbf{k}} * g_{\mathbf{m}},$$

where

$$s_{\mathbf{m},\mathbf{k}}[n] = \prod_{i\in S^1_{\mathbf{m}}} I[n-k_i-1] \cdot \prod_{i\in S^2_{\mathbf{m}}} I[n-k_i] \cdot \prod_{i\in S^3_{\mathbf{m}}} Q[n-k_i-1] \cdot \prod_{i\in S^4_{\mathbf{m}}} Q[n-k_i],$$

$$I[n] = \mathit{Re}(w[n]), \qquad Q[n] = \mathit{Im}(w[n]),$$

and the sequences (unit sample responses) $g_{\mathbf{m}} = g_{\mathbf{m}}[n]$ are defined by their Fourier transforms

$$G_{\mathbf{m}}(\Omega) = \frac{(j)^{N^2_{\mathbf{m}}}}{2^d} \sum_{r_c\in R^c_{\mathbf{m}}} \sum_{r_s\in R^s_{\mathbf{m}}} \Pi(r_c) P_{\mathbf{m},\tau}\left(\tilde{\Omega}\right) e^{-j\omega_c\tilde{\tau}}, \tag{3.9}$$

$$\tilde{\tau} = (r_c, \mathcal{P}^1_{\mathbf{m}}\tau') + (r_s, \mathcal{P}^2_{\mathbf{m}}\tau'), \qquad \tilde{\Omega} = \frac{\Omega}{T} - \omega_c \sum_{i=1}^{N^1_{\mathbf{m}}} r_c(i) - \omega_c \sum_{l=1}^{N^2_{\mathbf{m}}} r_s(l) + \omega_c.$$

Proof. See section 3.6. □

Block diagram of the proposed equivalent baseband model of $\mathbf{S}_\tau = \mathbf{DHF}_\tau\mathbf{M}$, as suggested by the statement of Theorem 3.2.1, is shown in Fig. 3-3. Therefore, system $\mathbf{S}_\tau$

can be represented as a series interconnection $\mathbf{S}_\tau = \mathbf{LV}$, of systems $\mathbf{V}$ and $\mathbf{L}$, where the components $\mathbf{V}_i$ of $\mathbf{V}$ map complex scalar-valued input $w \in \ell^2(\mathbb{C})$ to real scalar-valued outputs $s_{\mathbf{m}_i,\mathbf{k}_i} \in \ell^2(\mathbb{R})$ as defined in (3.31), and $\mathbf{L}$ maps real vector-valued input $s = (s_{\mathbf{m}_i,\mathbf{k}_i})_{i=1}^N \in \ell^2(\mathbb{R}^N)$ to complex scalar-valued output $v \in \ell^2(\mathbb{C})$ as defined in (3.28) and (3.4).

It is not hard to see, from (3.28) and (3.4), that frequency responses of the reconstruction filters $\mathbf{L}_i = \mathbf{AG}_i$ are discontinuous at frequencies $\Omega = \pm\pi$. Indeed, this is to be expected, since $\mathbf{L}$ represents a baseband equivalent of the linear part of the response of system $\mathbf{F}$ over the frequency interval $(\omega_c - \pi/T, \omega_c + \pi/T)$, which is, in general, not symmetric with respect to ω_c. Therefore, the frequency responses $L_i(\Omega)$ of $\mathbf{L}_i$ should, in general, be discontinuous at $\Omega = \pm\pi$. This further implies infinite memory of $\mathbf{L}_i$ in time-domain. That is, the unit sample responses of $\mathbf{L}_i$ are of infinite length, which is impracticable for hardware implementation. Nevertheless, from (3.28) and (3.4), it follows that $L_i(\omega)$ are smooth functions of Ω on the interval $(-\pi, \pi)$, and hence can be well approximated by polynomials (or some other appropriate basis) on compact subintervals of $(-\pi, \pi)$. This suggests an approximate model $\hat{\mathbf{S}} = \mathbf{L}_0 X \mathbf{V}$ of $\mathbf{S}$, as shown in Fig. 3-4. In this case, frequency responses of components of $\mathbf{L}_0$ are elements of a polynomial basis (1, $j\Omega$, Ω^2, etc). and X is a matrix of coefficients. By increasing degree of the polynomial approximating $\mathbf{L}$ (i.e., increasing the number of components in $\mathbf{L}_0$), the modeling error $||\mathbf{S} - \hat{\mathbf{S}}||$ can be made arbitrarily small.

3.3 Discussion

In this section, we give detailed discussions about some potentially limiting assumptions that were made in the derivation of the main result.

3.3.1 Effects of oversampling

The analytical result of this chapter suggests a special structure of a digital pre-distortion compensator which appears to be, in first approximation, both necessary and sufficient to match the discrete time dynamics resulting from combining modulation and demodulation

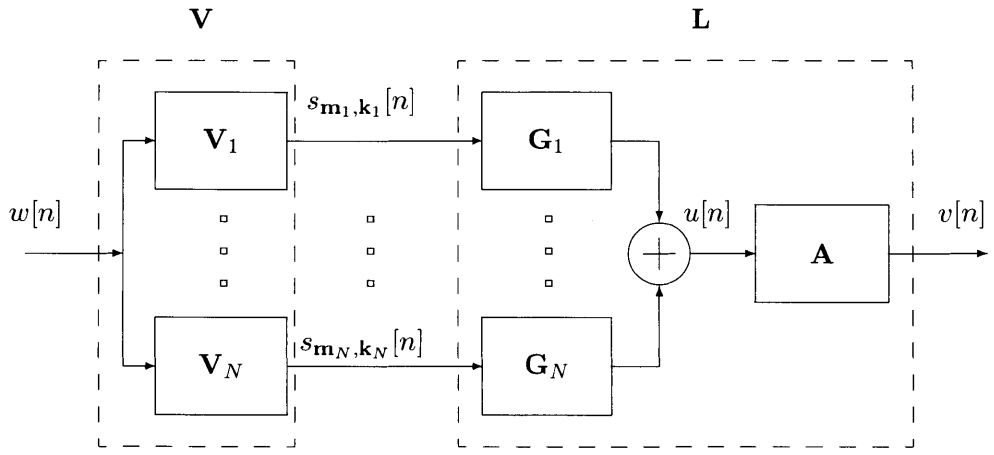


Figure 3-3: Block diagram depicting the novel equivalent baseband model structure as defined in Theorem 3.2.1.

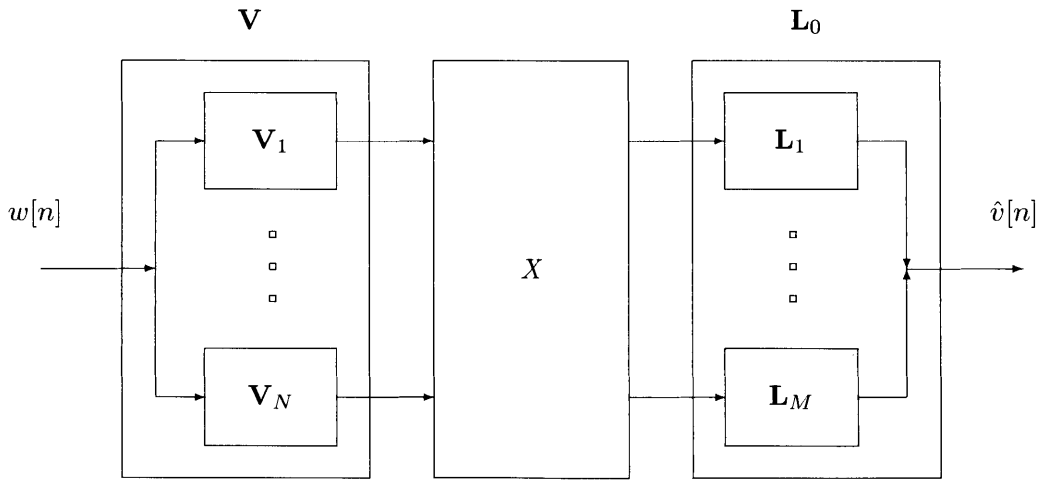


Figure 3-4: Detailed block diagram of an approximate model $\hat{\mathbf{S}} = \mathbf{L}_0 X \mathbf{V}$ of $\mathbf{S}$.

with a dynamic nonlinearity in continuous time. The "necessity" somewhat relies on the input signal u having "full" spectrum. Let us describe this in more detail. As reported in the previous section, frequency responses $L_i(\Omega)$ of $\mathbf{L_i}$, $i = \{1, \ldots, M\}$, are discontinuous at frequencies $\Omega = \pm\pi$, but are smooth on the interval $(-\pi, \pi)$ and therefore approximable by some appropriate basis (e.g., polynomials in Ω) on compact subintervals of $(-\pi, \pi)$. Moreover, these basis elements (e.g. frequency responses are monomials in Ω, in the case of polynomial basis) can be well approximated by FIR filters. When desired approximation accuracy is fixed, order of such an FIR filter mainly depends on the ratio of the baseband signal bandwidth B_w to the baseband sampling rate f_{dac} (i.e., rate of the DAC). Let us denote this ratio as $\xi = B_w/f_{dac} = B_wT$ (e.g., $\xi = 1$ corresponds to the case when spectrum of the baseband input signal occupies the whole Nyquist band). For ξ close to 1, if one wants to achieve good approximation of $\mathbf{L_i}$, the frequency responses of the approximate FIR filters must have sharp transition region (see Fig. 3-5a), which leads to high order (large number of taps) of the corresponding filters. Therefore, memory of the approximate reconstruction filter $\mathbf{L_0}$, and correspondingly approximate system $\hat{\mathbf{S}}$, will be long. Contrary to that, if ξ is relatively small compared to 1 (e.g., $1/2$ as in Fig. 3-5b), in order to transmit symbol information without distortion, the reconstruction filter $\mathbf{L_0}$ has to match the frequency response of the ideal baseband model LTI filter $\mathbf{L}$ only on the effective band defined by ξ, as shown in Fig. 3-5b. Therefore, reconstruction filters $\mathbf{L_0}$ can have smooth transitions and are realizable with low order FIR filters, which suggest that $\hat{\mathbf{S}}$ will have short memory as well.

Standard practice in transceiver design is to oversample baseband signal (symbols), and shape its spectrum (samples), before it is modulated onto a carrier [146]. In the case of large oversampling ratios (OSR), from symbol to sample space, the effective band of the signal containing symbol information is small compared to the rate of a transceiver DAC, i.e., ratio ξ is much smaller than 1. This implies that a plain Volterra structure with relatively short memory can capture dynamics of such a system well enough. A common implementation of amplitude-phase modulation will frequently employ a signal component separation approach (also known as out-phasing), such as LINC [147], where the low-pass signal u is decomposed into two components of constant amplitude, $u = u_1 + u_2$, $|u_1[n]| \equiv$

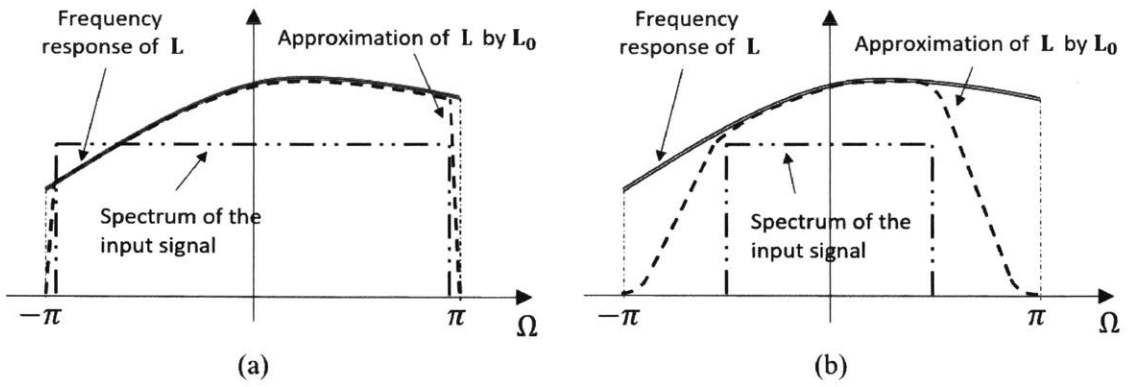


Figure 3-5: Simplified spectral diagrams that show how memory of the approximate reconstruction filter $\mathbf{L_0}$ depends on the ratio ξ: (a) $\xi \approx 1$, (b) $\xi = 1/2$.

$|u_2[n]| = \text{const}$, after which the components u_i are fed into two separate modulators, to produce continuous time outputs y_1, y_2, to be combined into a single output $y = y_1 + y_2$. Even when u is band-limited, the resulting components u_1, u_2 are not, and the full range of modulator's nonlinearity is likely to be engaged when producing y_1 and y_2. Furthermore, in forthcoming wideband communication systems, OSR is limited by the speed that digital baseband and DAC are able to sustain, and low OSR is more likely to be encountered, therefore emphasizing significance of a baseband model derived in this chapter.

3.3.2 Impact of low-pass filtering after zero order hold DAC

In section 3.1, we assume that digital-to-analog conversion (DAC) at the modulator $\mathbf{M}$ is performed using a zero order hold system. However, a common assumption in many communication systems is that DAC is modeled as a series interconnection of a ZOH and a low-pass filter, in order to suppress the high frequency harmonics caused by discontinuous nature of the ZOH pulse shape. Now, let this new modulator $\mathbf{M_0}$ be defined as a series interconnection $\mathbf{M_0} = \mathbf{XH_{dac}Z}$, where $\mathbf{X}$ and $\mathbf{Z}$ are the mixing and zero order hold systems, as defined in (3.1)-(3.2), and $\mathbf{H_{dac}}$ is an LTI system with memory equal to $m_0 > 0$. Let the memory and degree of the passband nonlinearity $\mathbf{F}$ be equal to m and d, respectively. Block diagram of a modified system $\mathbf{S_0} = \mathbf{DHFM_0}$ is shown in Fig. 3-6. We argue that system $\mathbf{S_0}$ has an equivalent baseband model decomposition similar to that from Theorem 3.2.1, where the nonlinear part $\mathbf{V}$ has the same maximal degree but its memory is equal

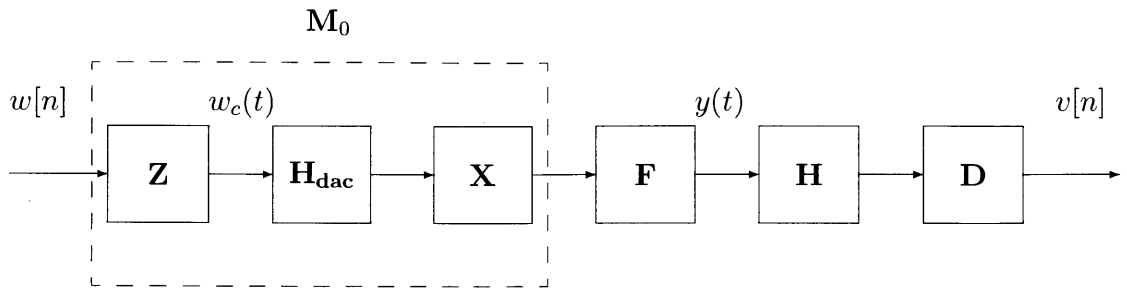
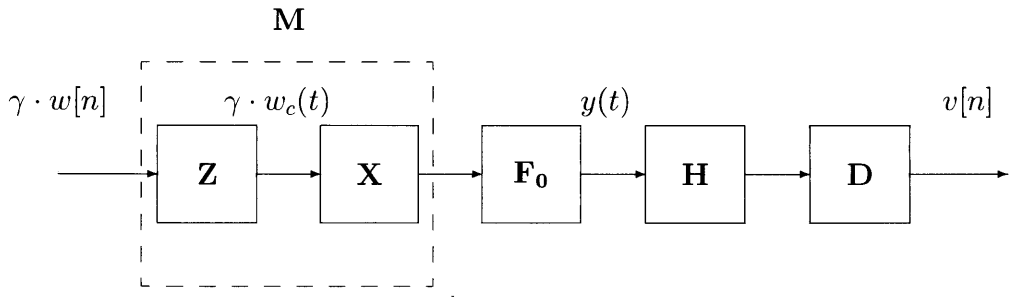


Figure 3-6: Block diagram of a modified system $\mathbf{S_0}$.

to $\lceil (m+m_0)/T \rceil$ (as opposed to $\lceil m/T \rceil$ in Theorem 3.2.1). Due to space constraints, we give only a simplified argument on why this should hold. Let us assume that the impulse response of $\mathbf{H_{dac}}$ is given by $h_{dac}(t) = c \cdot \delta(t - m_0)$ where $m_0 > 0, c \in \mathbb{R}$ and $\delta = \delta(t)$ is the Dirac delta function. Since

$$w_c(t-m_0)e^{j\omega_c t} = e^{j\omega_c m_0} w_c(t-m_0)e^{j\omega_c(t-m_0)},$$

it follows that $\mathbf{H_{dac}X} = \gamma \mathbf{X H_{dac}}$ where $\gamma = c \cdot e^{j\omega_c m_0}$, that is, systems $\mathbf{H_{dac}}$ and $\mathbf{X}$ commute up to a scaling factor. Hence, the system in Fig. 3-6 is equivalent to the following



where $\mathbf{F_0}$ is a series interconnection of a delay-by-m_0 and system $\mathbf{F}$. As a result, the memory and degree of $\mathbf{F_0}$ are equal to $m + m_0$ and d, respectively. This implies that $\mathbf{S_0}$ is equivalent to $\mathbf{S} = \mathbf{DHFM}$ from Section III, where $\mathbf{F}$ is replaced by $\mathbf{F_0}$ (with an additional scaling in baseband), and, according to Theorem 3.2.1, can be decomposed as $\mathbf{LV}$ where the maximal memory and degree of Volterra monomials in the nonlinear DT system $\mathbf{V}$ are equal to $\lceil (m+m_0)/T \rceil$ and d, respectively. It should be noted here that in the case of an

arbitrary low-pass filter $\mathbf{H_{dac}}$ the particular formulas from Theorem 3.2.1 would not hold anymore (i.e. nonlinear terms in $\mathbf{V}$ and expressions for reconstruction filters in $\mathbf{L}$ would change), but the general system structure should be preserved.

3.3.3 Extension to OFDM

Orthogonal frequency-division multiplexing (OFDM) is a multicarrier digital modulation scheme that has been the dominant technology for broadband multicarrier communications in the last decade. Compared with single-carrier digital modulation, by increasing the effective symbol length and employing many carriers for transmission, OFDM theoretically eliminates the problem of multi-path channel fading, which is the main type of disturbance on a terrestrial transmission path. It also mitigates low spectrum efficiency, impulse noise, and frequency selective fading [26]. One of the major drawbacks of OFDM is the relatively large Peak-to-Average Power Ratio (PAPR) [28]. This makes OFDM very sensitive to the nonlinear distortion introduced by high PA, which causes in-band as well as out-of-band (i.e., adjacent channel) radiation, decreasing spectral efficiency [148]. For that reason, linearization techniques play very important role in OFDM and have been studied extensively (see e.g., [149], [150], [151]).

Fig. 3-7 shows a block diagram of the typical implementation of an N-carrier OFDM system. Input stream of symbols $u[n]$, with bandwidth B, is first converted into blocks of length N by serial-to-parallel conversion, which are then fed to an N-point inverse FFT block. Output of this block is then transformed with a parallel-to-serial converter into a stream of N samples $w[n]$, and further converted to analog domain and used to modulate a single carrier. As can be seen from Figure 3-7, sequence $w[n]$ is fed into a system which can be modeled as $\mathbf{S} = \mathbf{DHFM}$, i.e., the model investigated in the previous section. Since the choice of input symbols' values (e.g., QPSK, QAM, etc.), was not relevant to the derivation of the baseband model from section 3.2, and hence input symbols can be arbitrary bounded complex numbers. Therefore, $w[n]$ can be considered as a legitimate input sequence to a system modeled as $\mathbf{DHFM}$. This implies that the baseband model derived in 3.2, and its corresponding DPD structure, can be also applied for distortion reduction in the case when

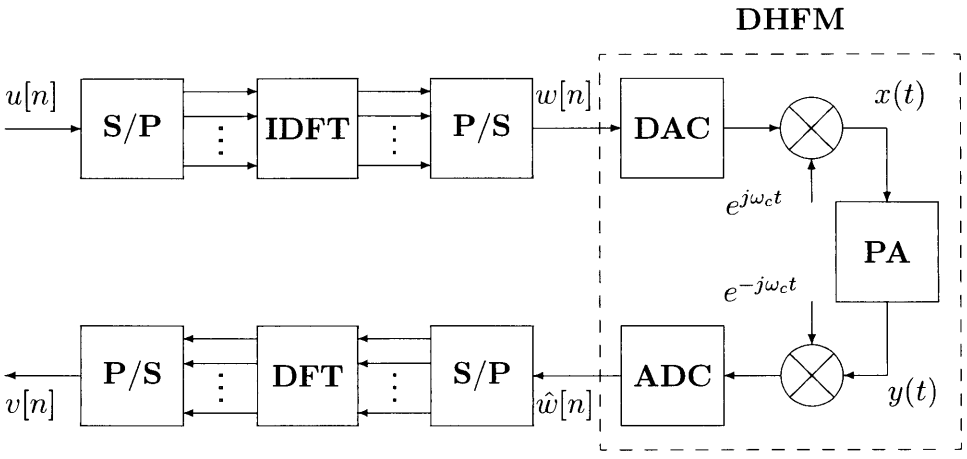


Figure 3-7: Block diagram of a typical implementation of OFDM.

OFDM modulation is used.

3.4 Simulation Results

In this section, aided by MATLAB simulations, we validate the proposed (approximate) baseband model and illustrate performance of a simple DPD structure based on this model. Simulations were performed for various models of passband nonlinearity, as explained below.

The general simulation parameters, used to obtain all results presented in this section, were chosen as follows. The input symbol sequence was generated as a 64QAM signal at 20MS/sec rate. The choice of 64QAM encoded baseband data is solely due to the ease of generating and processing such sequences in Matlab (as opposed to e.g., LTE signals). For the purpose of producing training and validation data, system **S** is simulated for two cases of baseband (or sampling) frequency of $f_{dac_1} = 20$MS/sec and $f_{dac_2} = 25$MS/sec. Therefore, the input signal bandwidth occupies 100% and 80% of observed bandwidth, respectively. After digital-to-analog conversion, the baseband input signal is up-converted to a passband carrier frequency of $f_c = 1$GHz, by a standard quadrature modulation (i.e., multiplication of I and Q components with $\cos(2\pi f_c t)$ and $\sin(2\pi f_c t)$, respectively, where t is the time variable). Digital simulation of the passband part of system **S** was done by

representing continuous-time signals by their discrete-time counterparts, obtained by sampling with very high sampling rate of $f_{ct} = 100 f_c$, so that spectral aliasing effects do not significantly affect modeling or linearization performance. Two cases of passband nonlinearity $\mathbf{F}$ were considered, and are described in detail in the next section. In order to capture the PA's output, bandpass filter $\mathbf{H}$, as well as the low-pass filter used in demodulation system $\mathbf{D}$, were realized as ideal rectangular filters. This was done in simulation by zeroing out the frequency response of the input signal outside the frequency band of interest. The length of the input symbol sequences used for training, validation and DPD performance evaluation was $N_{symb} = 16,384$. Both model validation and DPD performance evaluation were done by Monte Carlo simulation with 100 simulation runs.

3.4.1 Passband Nonlinearity Model

Two models of passband nonlinearity were used to confirm the proposed baseband model, and evaluate performance of a compensator based on this model. In the first case, a cubic term, similar to those from Theorem 3.2.1, was added to the identity function in order to model passband nonlinearity $\mathbf{F}$, which is in this case defined as:

$$(\mathbf{F}x)(t) = x(t) - \delta \cdot x(t-\tau_1)x(t-\tau_2)x(t-\tau_3), \tag{3.10}$$

where, without loss of generality, $0 \leq \tau_1 \leq \tau_2 \leq \tau_3$ and $\delta > 0$ is a parameter controlling magnitude of distortion $\mathbf{\Delta}$ in $\mathbf{S} = \mathbf{I} + \mathbf{\Delta}$.

In the second case, subsystem $\mathbf{F}$ is modeled by a series interconnection of a linear combination of the identity and simple delay, and Cann's model [152], which is frequently used for behavioral modeling of power amplifiers:

$$(\mathbf{F}x)(t) = f\left((1-\rho)x(t) + \rho x(t-\tau)\right), \tag{3.11}$$

where

$$f(a) = \frac{ga}{\left[1 + \left(\frac{ga}{L}\right)^s\right]^{\frac{1}{s}}}, \tag{3.12}$$

is Cann's model with fixed parameters $L, s, g > 0$. In this case, parameter $\rho \in (0,1)$ controls the proportion of memoryless dependence in this passband nonlinearity.

3.4.2 Model Selection

Results from Section 3.2 suggest that the approximate baseband model should be searched for within a family of models $\hat{\mathbf{S}} = \mathbf{L}_0 X \mathbf{V}$, as shown on the block diagram in Fig 3-4. Abbreviation L_0XV will be used in all figures and tables to denote our model. Reconstruction filters $\mathbf{L}$ are approximated by the basis $(\mathbf{L_{01}}, \mathbf{L_{02}}, \mathbf{L_{03}})$, where LTI subsystems $\mathbf{L}_{0i}, i \in \{1,2,3\}$, have frequency responses L_{0i} defined by

$$L_{01}(\Omega) = 1, \;\; L_{02}(\Omega) = j\Omega, \;\; L_{03}(\Omega) = \Omega^2, \;\; \forall \Omega \in [-\pi, \pi),$$

and 2π-periodically extended for other values of Ω. In an actual simulation, each basis element $\mathbf{L}_{0i}$ is realized (or, more precisely, approximated) as an FIR filter with 30 taps. Nonlinear system $\mathbf{V}$ is comprised of all Volterra monomials up to order d, with backward and forward memory m_b and m_f, respectively, i.e.,

$$(\mathbf{V}_i w)[n] = \prod_{k=m_b}^{m_f} I[n-k]^{\alpha_i(k)} \prod_{k=m_b}^{m_f} Q[n-k]^{\beta_i(k)},$$

for all

$$\alpha_i(k), \beta_i(k) \in \mathbb{Z}_+, \qquad \sum_{k=m_b}^{m_f} \alpha_i(k) + \sum_{k=m_b}^{m_f} \beta_i(k) \leq d,$$

where $I[n] = \mathrm{Re}\; w[n]$ and $Q[n] = \mathrm{Im}\; w[n]$.

The ability of our model to approximate system $\mathbf{S}$ is compared to that of a widely used model obtained by employing general Volterra series structure [61]:

$$(\hat{\mathbf{S}}_{\mathbf{v}} w)[n] = \sum_{(\alpha_i, \beta_i)} c_i \prod_{k=-m_b}^{m_f} I[n-k]^{\alpha_i(k)} \prod_{k=-m_b}^{m_f} Q[n-k]^{\beta_i(k)},$$

$$\alpha_i(k), \beta_i(k) \in \mathbb{Z}_+, \qquad \sum_{k=-m_b}^{m_f} \alpha_i(k) + \sum_{k=-m_b}^{m_f} \beta_i(k) \leq d,$$

Table 3.1: Parameter selection for different approximation models

Cubic nonlinearity model					
Model	L_0XV	Volt. 1	Volt. 2	Volt. 3	Volt. 4
d	3	3	5	5	3
m_b	1	1	4	2	4
m_f	0	0	0	2	4
Modified Cann's model					
Model	L_0XV	Volt. 1	Volt. 2	Volt. 3	Volt. 4
d	7	7	7	5	5
m_b	1	1	2	4	2
m_f	0	0	0	0	2

and $(\alpha_i, \beta_i) = (\alpha_i(m_b), \ldots, \alpha_i(m_f), \beta_i(m_b), \ldots, \beta_i(m_f))$. We should remark here that the standard practice in literature on equivalent baseband modeling, and corresponding digital predistortion (see e.g., [66] or [70]), is to assume much simpler approximate Volterra model, which is comprised only of monomials with odd degree $\sum_{k=-m_b}^{m_f} \alpha_i(k) + \sum_{k=-m_b}^{m_f} \beta_i(k)$. This is justified by the assumed low-pass filtering (LPF) operation both after digital-to-analog conversion (DAC) at the transmitter side, and before demodulation and analog-to-digital conversion (ADC) at the receiver side. In this work, an LPF after DAC is not assumed and therefore a full Volterra series model has to be taken into consideration if we hope to successfully approximate system **S**.

For each passband nonlinearity, we consider four different Volterra based models, by varying degree d, and forward and backward memory depths m_f and m_b, respectively, as given in Table 4.2. For each model, the corresponding coefficients (matrix X for our model, and vector (c_i) for Volterra models) are found by simple least squares optimization to fit the input-output data available. It should be noted that fitting has to be done for both real and imaginary part of $w[n]$.

3.4.3 Performance Evaluation

As a measure of quality of the approximate model, normalized mean square error (NMSE) metric is used, which, for a given input w, is defined as

$$\text{NMSE}(\hat{\mathbf{S}}, w) = 20\log_{10}\left(\frac{||\mathbf{S}w - \hat{\mathbf{S}}w||_2}{||\mathbf{S}w||_2}\right),$$

where $\mathbf{S}$ is the true system, and $\hat{\mathbf{S}}$ is the approximate model. We also evaluate performance of a simple compensator based on the proposed model. We assume that parameters δ (case 1) and ρ (case 2) are relatively small, so that the inverse $\mathbf{S}^{-1}$ of $\mathbf{S}$ can be well approximated by $2\mathbf{I} - \mathbf{S}$, as discussed in the previous sections. Then our goal is to build a compensator $\mathbf{C} = \widehat{\mathbf{S}^{-1}} = \mathbf{2I} - \hat{\mathbf{S}}$, where $\hat{\mathbf{S}}$ is again sought in the family of models as described above (i.e., $\hat{\mathbf{S}}$ is fit according to the above procedure and substituted into $\mathbf{2I} - \hat{\mathbf{S}}$). Compensator performance is measured in terms of output Error Vector Magnitude (EVM) and Adjacent-Channel-Leakage-Ratio (ACLR) [146], which are defined, for a given input w, as

$$\text{EVM}(\mathbf{C}, w) = 20\log_{10}\left(\frac{||w - \mathbf{S}\mathbf{C}w||_2}{||w||_2}\right),$$

$$\text{ACLR}(\mathbf{C}, w) = 10\log_{10}\left(\frac{\frac{1}{\xi}\int_{I_1}|W(\Omega)|^2\Omega}{\frac{1}{1-\xi}\int_{I_2}|W(\Omega)|^2\Omega}\right),$$

where W is the Fourier transform of w, intervals $I_1 = (-\frac{\pi\xi}{2}, \frac{\pi\xi}{2})$ and $I_2 = (-\pi, -\frac{\pi\xi}{2}) \cup (\frac{\pi\xi}{2}, \pi)$ with $\xi = B_w T_s$ where B_w is the input signal bandwidth.

In order to evaluate performance of the proposed model we have performed various simulations, varying different model and simulation parameters, as described in corresponding subsections. We compare approximation capability of the proposed model, as well as performance of a DPD based on it, with those obtained by employing pure Volterra series models, where the corresponding model parameter values are given in Table 4.2.

Effects of changing parameters δ and ρ

For the cubic passband nonlinearity model (3.10) we fix delays to $\tau = \begin{bmatrix} 0.2T_s & 0.3T_s & 0.4T_s \end{bmatrix}$, where $f_{dac} = 1/T_s$ is the DAC rate, and vary parameter δ to simulate various degrees of distortion. In the case of Cann's model defined in (3.12), we take arbitrary, but fixed, parameter values $\tau = 0.5T_s$, $L = 1$, $s = 8$ and $g = 1$. The level of passband distortion introduced by $\mathbf{F}$ is controlled by varying parameter ρ. Simulations are run for two values of sampling/DAC frequency $f_{dac1} = 20$MHz and $f_{dac1} = 25$MHz. This is equivalent to setting parameter ξ to $\xi = 1$ and $\xi = 0.8$, respectively.

In case of $f_{dac1} = 20$MHz (i.e. when information bearing signal occupies the whole observation band), approximation results for different models are shown in Figs. 3-8 and 3-9. It can be seen that our model significantly outperforms Volterra models, and results in approximation error of 0.1% or better in all cases. This result was to be expected, since model shown in Fig. 3-4 can approximate arbitrarily close the original system $\mathbf{S}$. This is not the case with Volterra series models, due to inherently long (or more precisely infinite) memory introduced by the LTI part of $\mathbf{S}$. Even if we use a non-causal Volterra series model (i.e., $m_f \neq 0$), which is expected to better capture the true dynamics of $\mathbf{S}$, we are still unable to get good fitting of $\mathbf{S}$ (though NMSE decreases with the increase of m_f). We get similar performance results for the proposed model in the case of $f_{dac2} = 25$MHz (though we do not report those to avoid repetition), with slightly better performance of Volterra based models than in the case $\xi = 1$. This is in accordance with presented theoretical results and discussions given in sections 3.2 and 3.3: as ratio ξ decreases, memory of $\mathbf{L_0}$ decreases, and Volterra models approach performance of the ideal model proposed in this thesis.

Comparison of different DPD structures in terms of output EVM, for the case of $f_{dac1} = 20$MHz, is shown in Figs. 3-10 and 3-11. As a baseline, we also plot output EVM for the case when no compensation is used. As can be seen from Figs. 3-10 and 3-11, the proposed DPD model outperforms other compensators for a wide range of parameter values. In the case of a cubic passband nonlinearity, as parameter δ increases, the DPD performance decreases and approaches that of Volterra series based models. This is caused by the relatively

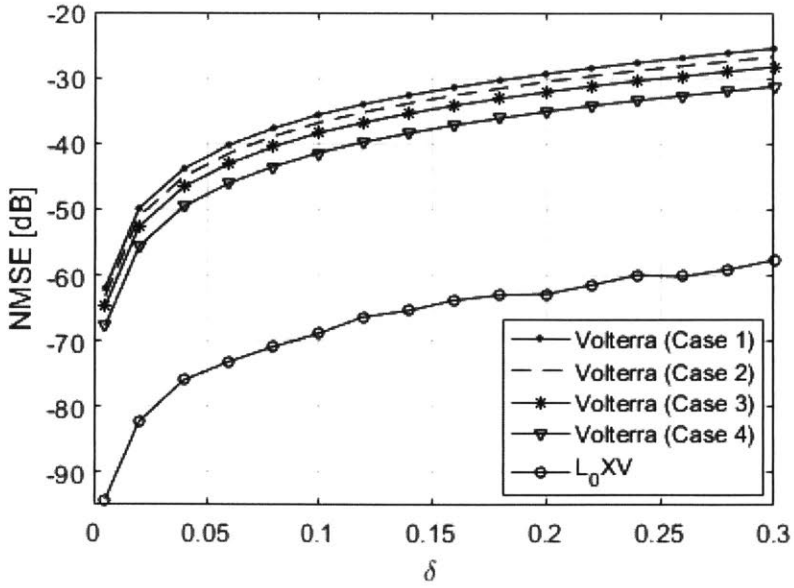


Figure 3-8: NMSE of approximation for different models in the case of cubic passband nonlinearity.

simple structure of the compensator: as δ increases, the inverse $(\mathbf{I}+\mathbf{\Delta})^{-1}$ is not approximated close enough by the linear term $\mathbf{I}-\mathbf{\Delta}$. Hence, simple compensator $\mathbf{C}=\mathbf{2I}-\hat{\mathbf{S}}$ is not capable of approximating well the inverse $\mathbf{S}^{-1}$ (even though $\hat{\mathbf{S}}$ still approximates $\mathbf{S}$ very well). When $\xi=1$, the observed (i.e., linearized) bandwidth of the PA output signal is equal to the baseband signal bandwidth, and therefore ACLR is not a meaningful metric of performance since there is no adjacent channel. Results in terms of NMSE and EVM for the case of $\xi=0.8$, are similar to those for $\xi=1$, and are omitted. Again, a slight increase in modeling/linearization capability of Volterra models is noticeable when going from $\xi=1$ to $\xi=0.8$. ACLR results for both cases of passband nonlinearity are shown in Figs. 3-12 and 3-13. Comparison of output signal power spectral density for various DPD models is shown in Fig. 3-14. As can be seen, about 10dB in ACLR improvement is achieved with the proposed DPD, while Volterra based DPDs were unable to clear much distortion. Again, this is expected since simulated Volterra models do not have enough memory capability to approximate well the frequency response of $\mathbf{L}$ close to the points $\Omega=\pm\pi$ of discontinuity.

In the next two subsections, due to space constraints and without loss of generality, we

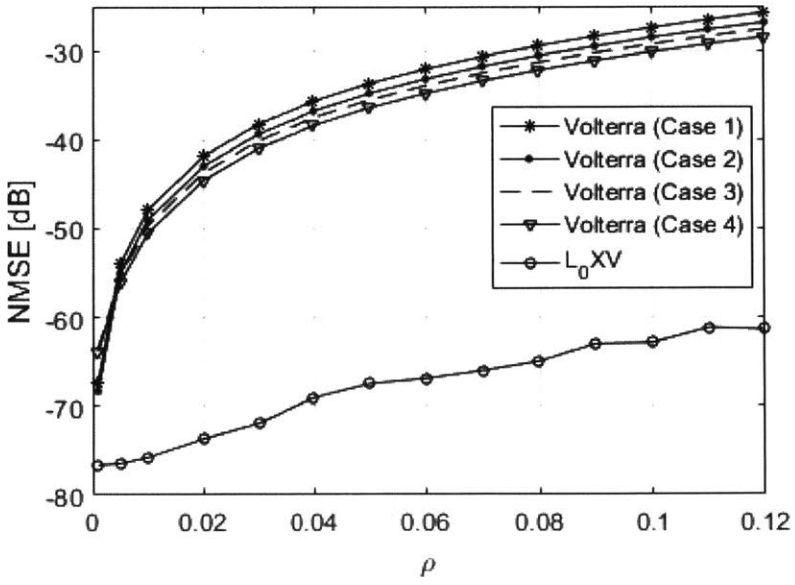


Figure 3-9: NMSE of approximation for different models in the case of modified Cann's nonlinearity model.

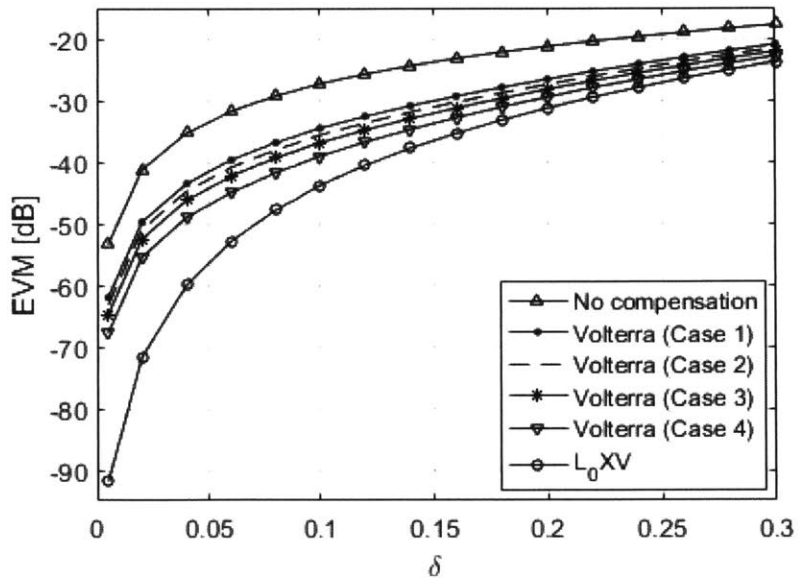


Figure 3-10: Output EVM, for different DPD structures, as a function of parameter δ (in the case of cubic passband nonlinearity).

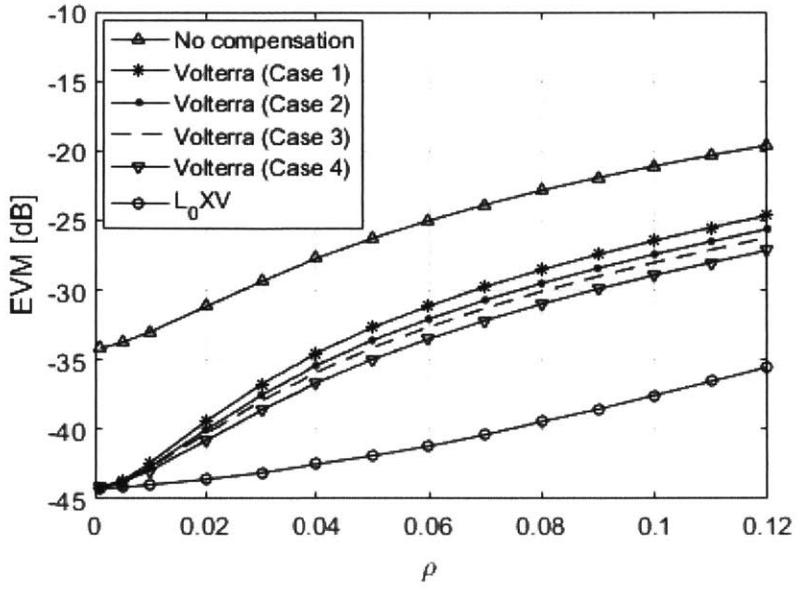


Figure 3-11: Output EVM, for different DPD structures, as a function of parameter ρ (in the case of modified Cann's nonlinearity model).

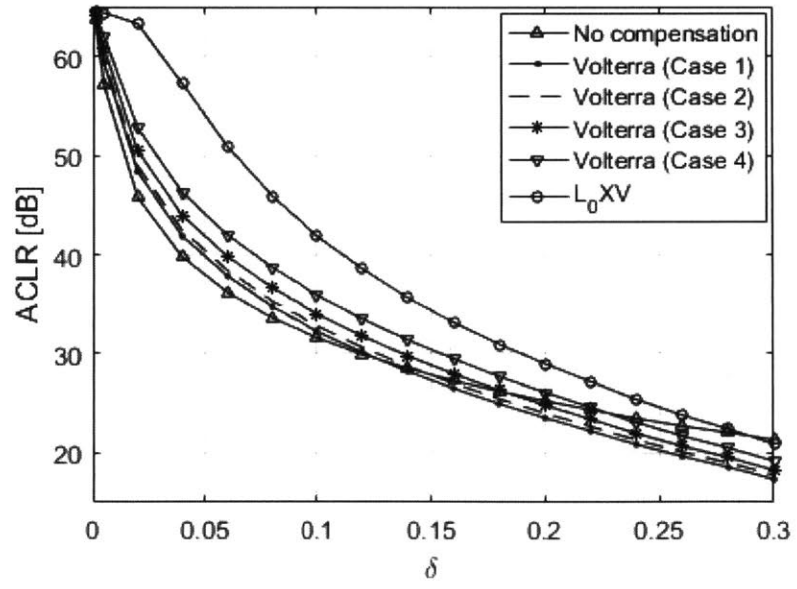


Figure 3-12: Output ACLR, for different DPD structures, as a function of parameter δ (in the case of cubic passband nonlinearity).

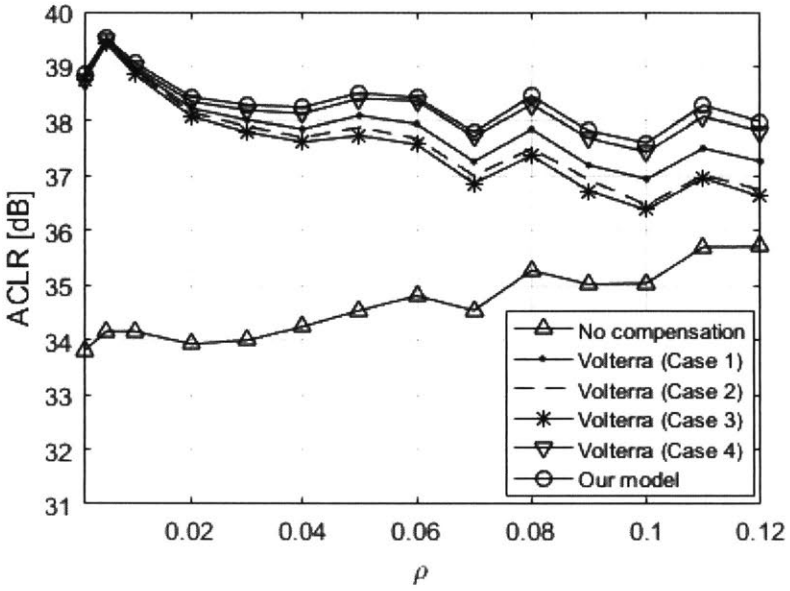


Figure 3-13: Output ACLR, for different DPD structures, as a function of parameter ρ (in the case of modified Cann's nonlinearity model).

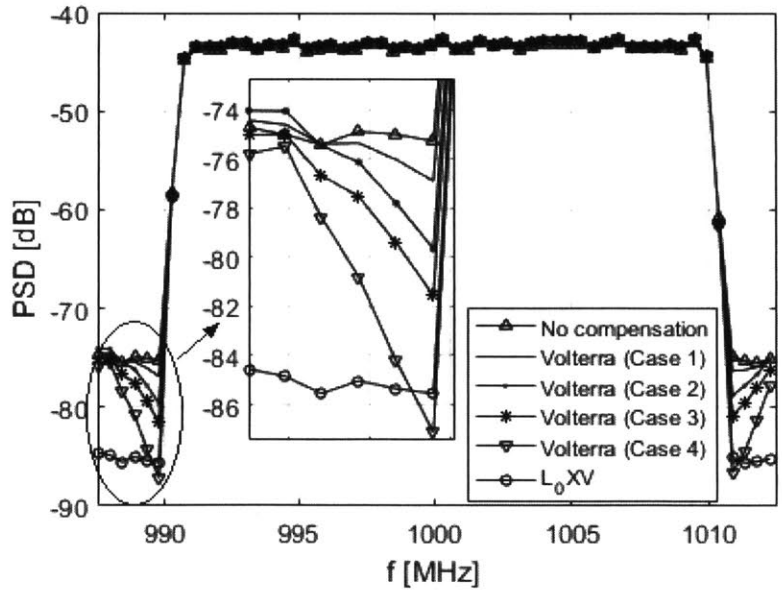


Figure 3-14: PSD of the PA output, for different DPD structures (in the case of cubic passband nonlinearity).

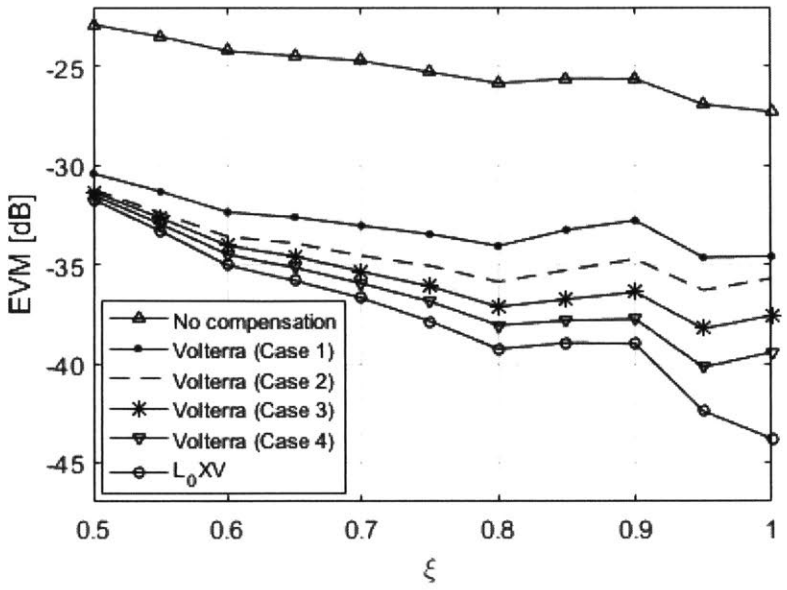


Figure 3-15: Output EVM, for different DPD structures, as a function of the ratio $\xi = B_w/f_{dac}$ (in the case of cubic passband nonlinearity).

report only results for cubic nonlinearity. Results for the other type of nonlinearity follow the same trends and do not bring any new information and can therefore be omitted.

Effects of changing parameter ξ

Now we fix model parameters to $\delta = 0.1$ and $\tau = \begin{bmatrix} 0.2T_s & 0.3T_s & 0.4T_s \end{bmatrix}$, and vary ratio ξ from 0.5 to 1. The results in terms of EVM and ACLR are shown in Figs. 3-15 and 3-16. As is expected, for small values of ξ, all models perform similarly, since the memory of $\mathbf{L_0}$ is not large. As ξ increases, our model achieves the best performance since it is the only one capable of approximating long memory effects.

Effects of increasing memory of F

In this case, we fix parameters $\delta = 0.1$ and $\xi = 1$, and vary memory of the passband nonlinearity $\mathbf{F}$ (and accordingly memory of our model). Maximal delay $\tau_{max} = \max_{i\in\{1,2,3\}} \tau_i$ is varied from 1 sample to 7 samples, i.e. from $\tau_{max} \in (0, T_s]$ to $\tau_{max} \in (6T_s, 7T_s]$. The results in terms of NMSE and EVM are shown in Figs. 3-17 and 3-18 (there is no reported ACLR since $\xi = 1$). Results are again as expected, i.e., for large values of τ_{max} Volterra

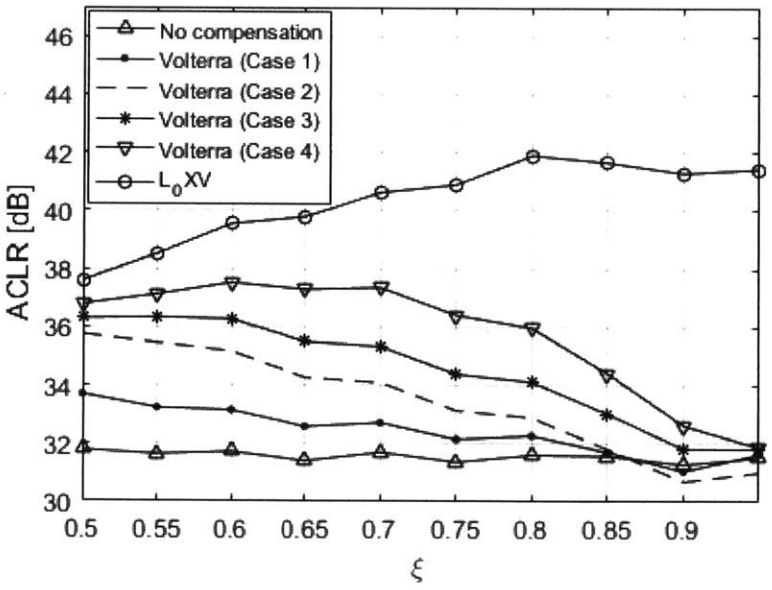


Figure 3-16: Output ACLR, for different DPD structures, as a function of the ratio $\xi = B_w/f_{dac}$ (in the case of cubic passband nonlinearity).

models struggle with approximating well the nonlinearity $\mathbf{F}$ due to their limited memory capability. Our model is again capable of successfully modeling and linearizing the PA output for all values of parameter τ_{max}.

Model Complexity

Advantage of the proposed compensator structure is not only in better compensation performance, but also in that it achieves this performance in a more efficient way. That is, significantly lower number of Volterra monomials (basis elements) is needed in order to represent the nonlinear part of the compensator. Table 3.2 shows a comparison in the number of basis elements for different compensator structures (parameter values for the corresponding passband nonlinearities were fixed to $\delta = 0.04$ and $\rho = 0.06$). For each passband nonlinearity case, numbers in the first row in Table 3.2 represent the total number of basis elements of the corresponding model. The second row shows the actual number of basis elements used to build the compensator, that is, least squares optimization yields many nonzero coefficients, but only a subset of those are considered significant and thus used in an actual compensator implementation. In this case coefficient is considered significant if

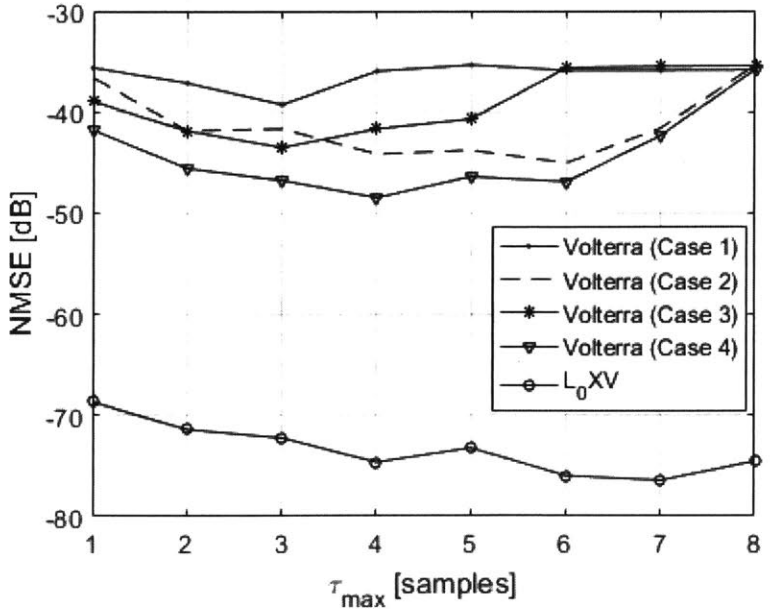


Figure 3-17: Output NMSE, for different DPD structures, as a function of the maximal delay τ_{max} (in the case of cubic passband nonlinearity).

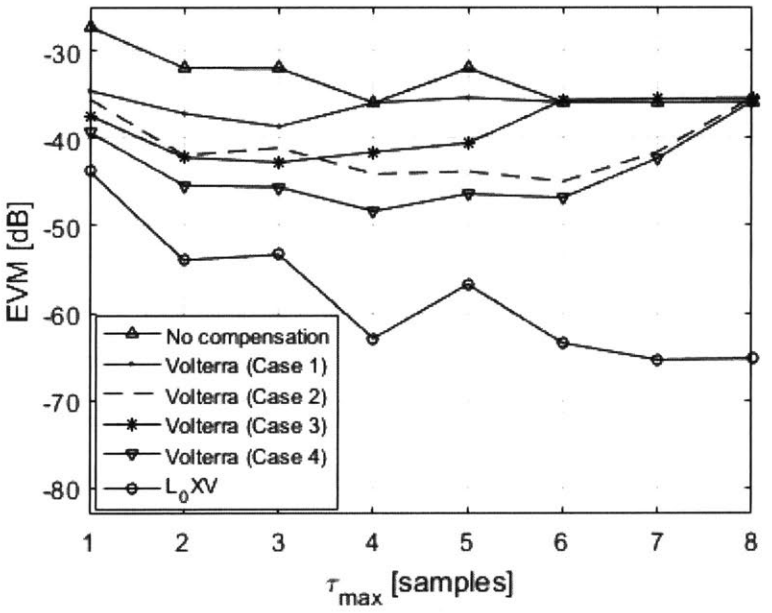


Figure 3-18: Output EVM, for different DPD structures, as a function of the maximal delay τ_{max} (in the case of cubic passband nonlinearity).

Table 3.2: Complexity of different compensator models in terms of the number of coefficients that are needed for hardware implementation of the nonlinear part.

Cubic nonlinearity model					
Model	L_0XV	Volt. 1	Volt. 2	Volt. 3	Volt. 4
# of basis elements	70	70	6006	6006	2660
# of significant basis elements	42	14	2058	1190	48
Modified Cann's model					
Model	L_0XV	Volt. 1	Volt. 2	Volt. 3	Volt. 4
# of basis elements	660	660	3432	6006	6006
# of significant basis elements	253	528	2914	1881	449

its value falls above a certain threshold t_0, where t_0 is chosen such that increase in EVM after zeroing non-significant coefficients is not larger than 1% of the best achievable EVM (i.e., when all basis elements are used). From Table 3.2 we can see that, even for the best (in terms of EVM) Volterra structure, more basis elements are needed in order to implement the compensator (and its performance is still below the one achieved by our model). It should be noted though that the low complexity advantage of the proposed model gets lost as memory of the passband nonlinearity increases. This is due to an increase in the number of nonlinear basis elements in $\mathbf{V}$, which then approaches that of a plain Volterra model.

We should mention here that in [153] decomposition to a short memory nonlinear system and a long memory linear system, for a system architecture similar to the one proposed in this thesis, but with a different modulator model, has been verified by simulations which combine Cadence Spectre Circuit Simulator and Matlab (PA is simulated in Spectre, a phase modulator is realized with verilog-A model in Spectre, and other processing blocks/subsystems are realized in Matlab). Similarly, in [154], a DPD based on the model proposed in this thesis has been implemented on an FPGA and successfully tested for an outphasing transmitter at Q-band (45GHz) and for an RF PA with 1.97GHz carrier.

3.5 Summary

In this chapter, an exact equivalent baseband model of transmission systems in the form of a nonlinear dynamical system $\mathbf{S}$ in discrete time (DT) defined by a series interconnection of a phase-amplitude modulator, a nonlinear dynamical system $\mathbf{F}$ in continuous time (CT), and an ideal demodulator was derived. It was show that when $\mathbf{F}$ is a CT Volterra series model, the resulting $\mathbf{S}$ is a series interconnection of a DT Volterra series model of the same degree and equivalent memory depth, and an LTI system with special properties. Based on the derived model, a novel, analytically motivated, structure of a digital predistorter for nonlinear RF power amplifiers was proposed. The equivalent baseband model was validated and the effectiveness of the proposed DPD was demonstrated by MATLAB simulations with several common models of PA nonlinearity.

3.6 Proof of Theorem 3.2.1

We first state and prove the following Lemma, that is a special case of Theorem 3.2.1, in which τ ranges over $[0,T)^d$ (instead of $\tau \in [0,\infty)^d$), and hence $\mathbf{k} = \mathbf{0}$. The proof of Theorem 3.2.1 then readily follows from this Lemma.

Lemma 3.6.1. *The DT system* $\mathbf{DHF}_\tau\mathbf{M}$ *with* $\tau \in [0,T)^d$ *maps* $w \in \ell$ *to*

$$v = \mathbf{A}u \in \ell, \quad \textit{with} \quad u = \sum_{\mathbf{m}\in[1:4]^d} s_{\mathbf{m}} * g_{\mathbf{m}},$$

where

$$s_{\mathbf{m}}[n] = (I[n-1])^{|S^1_{\mathbf{m}}|}(I[n])^{|S^2_{\mathbf{m}}|}(Q[n-1])^{|S^3_{\mathbf{m}}|}(Q[n])^{|S^4_{\mathbf{m}}|},$$

$$I[n] = \mathit{Re}(w[n]), \qquad Q[n] = \mathit{Im}(w[n]),$$

and the sequences $g_{\mathbf{m}}$ *are as defined in Theorem 3.2.1.*

Proof. A block diagram of system $\mathbf{DHF}_\tau\mathbf{M}$ is shown in Fig. 3-19, where $\mathbf{M}$ and $\mathbf{D}$ are decomposed into elementary subsystems as defined in the previous section. The proof of Lemma 3.6.1 consists of two steps. First, we express signal y (see Fig. 3-19) as a function

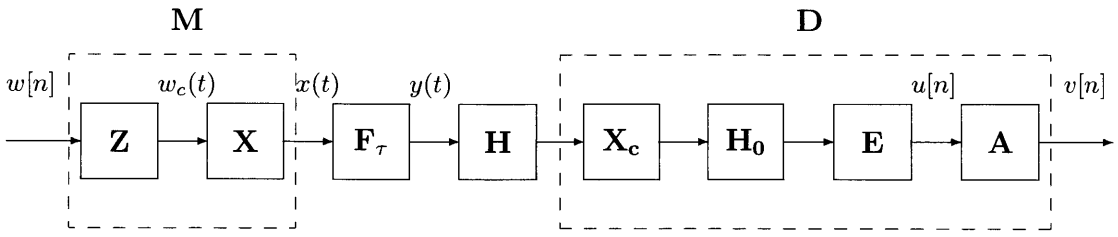


Figure 3-19: Block diagram of system $\mathbf{S}_\tau = \mathbf{DHF}_\tau\mathbf{M}$, with all corresponding subsystems.

of I, Q, ω_c, T and τ, which would imply description of $s_{\mathbf{m}}$ as given in the statement of Lemma 3.6.1. Then, by finding relationship between the Fourier transforms of signals y and u, we determine frequency responses $G_{\mathbf{m}}$, which concludes the proof of Lemma 3.6.1.

Consider first the case $d = 1$ (i.e., $\mathbf{F}_\tau$ is just a delay by $\tau \in [0, T)$). By definition, the outputs w_c, x and y of $\mathbf{Z}, \mathbf{X}$ and $\mathbf{F}$, respectively, are given by

$$w_c(t) = \frac{1}{T}\sum_n w[n]p(t-nT) = \underbrace{\frac{1}{T}\sum_n I[n]p(t-nT)}_{I_c(t)} + \underbrace{\frac{j}{T}\sum_n Q[n]p(t-nT)}_{jQ_c(t)}.$$

$$x(t) = (\mathbf{X}w_c)(t) = \mathrm{Re}\{\exp(j\omega_c t)w_c(t)\},$$

$$y(t) = I_c(t-\tau)\cos(\omega_c t - \omega_c\tau) - Q_c(t-\tau)\sin(\omega_c t - \omega_c\tau). \tag{3.13}$$

Consider the representation $p(t) = p_{1,\tau}(t) + p_{2,\tau}(t)$, where

$$p_{1,\tau}(t) = \theta(t) - \theta(t-\tau), \qquad p_{2,\tau}(t) = \theta(t-\tau) - \theta(t-T).$$

Let $\mathbf{Z_1} : \ell \to \mathcal{L}^2(\mathbb{C})$ and $\mathbf{Z_2} : \ell \to \mathcal{L}^2(\mathbb{C})$ be the pulse amplitude modulators with pulse shapes $p_{1,\tau}$ and $p_{2,\tau}$, respectively. Let $\mathbf{B}$ denote the backshift function mapping $x \in \ell$ to $y = \mathbf{B}x \in \ell$, defined by $y[n] = x[n-1]$. Then

$$I_c(t-\tau) = e_{1,\tau}(t) + e_{2,\tau}(t), \qquad Q_c(t-\tau) = e_{3,\tau}(t) + e_{4,\tau}(t), \tag{3.14}$$

where

$$
\begin{aligned}
e_{1,\tau} &= \mathbf{Z_1 B} I, \quad \text{i.e.,} \quad e_{1,\tau}(t) = \frac{1}{T}\sum_n I[n-1]p_{1,\tau}(t-nT), \\
e_{2,\tau} &= \mathbf{Z_2} I, \quad \text{i.e.,} \quad e_{2,\tau}(t) = \frac{1}{T}\sum_n I[n]p_{2,\tau}(t-nT), \\
e_{3,\tau} &= \mathbf{Z_1 B} Q, \quad \text{i.e.,} \quad e_{3,\tau}(t) = \frac{1}{T}\sum_n Q[n-1]p_{1,\tau}(t-nT), \\
e_{4,\tau} &= \mathbf{Z_2} Q, \quad \text{i.e.,} \quad e_{4,\tau}(t) = \frac{1}{T}\sum_n Q[n]p_{2,\tau}(t-nT).
\end{aligned}
\tag{3.15}
$$

According to (4.5)-(3.15), the output $y(t)$ of $\mathbf{F}_\tau$ can be expressed as:

$$y(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t),$$

where

$$f_i(t) = \begin{cases} e_{i,\tau}(t)\cos(\omega_c t - \omega_c \tau), & i = 1,2 \\ -e_{i,\tau}(t)\sin(\omega_c t - \omega_c \tau), & i = 3,4 \end{cases}. \tag{3.16}$$

Therefore, subsystem $\mathbf{F}_\tau\mathbf{M}$, mapping $w[n]$ to $y(t)$, can be represented as a parallel interconnection of amplitude modulated delayed and undelayed in-phase and quadrature components of $w[n]$. This is shown in Fig. 3-20, where $\tilde{\mathbf{D}} = \mathbf{E H_0 X_c H}$.

Suppose now that order d of $\mathbf{F}_\tau$ is an arbitrary positive integer larger than 1, i.e., that $\mathbf{F}_\tau : x \mapsto y$ defined by $y(t) = x(t-\tau_1)\cdot\cdots\cdot x(t-\tau_d)$. Then the output y of $\mathbf{F}_\tau$ can be written as a product

$$y(t) = y_1(t)\cdot y_2(t)\cdot\ldots\cdot y_d(t), \tag{3.17}$$

where, for all $i \in [1:d]$,

$$y_i(t) = I_c(t-\tau_i)\cos(\omega_c t - \omega_c \tau_i) - Q_c(t-\tau_i)\sin(\omega_c t - \omega_c \tau_i).$$

This implies that, for each i, signal $y_i(t)$ can be represented as the output of subsystem $\mathbf{F}_{\tau_i}\mathbf{M}$, where $\mathbf{F}_{\tau_i}$ is just a simple delay, as discussed in the case $d = 1$. Therefore, system $\mathbf{F}_\tau\mathbf{M}$, mapping w to y, can be described as shown in Fig. 3-21. Hence, by using the same

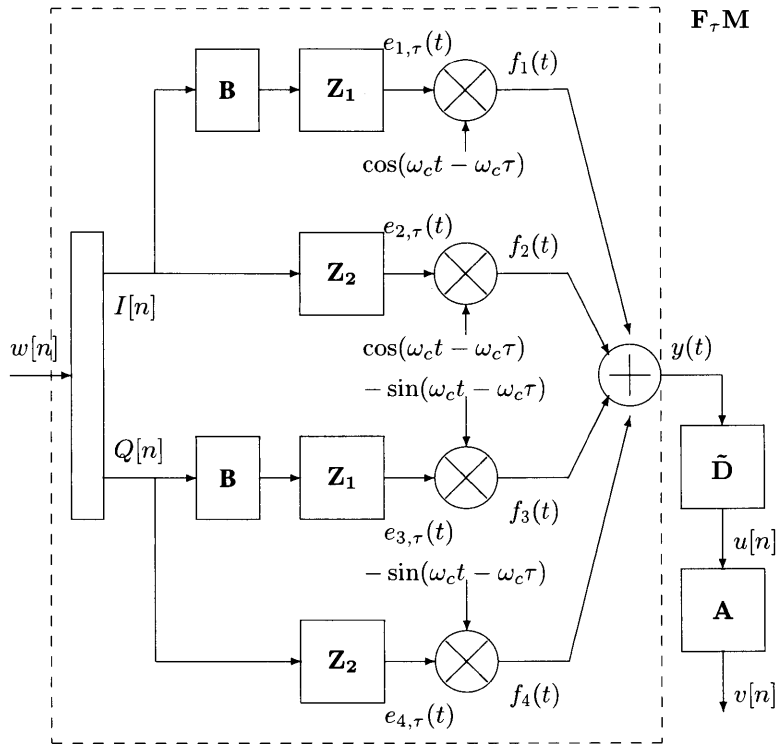


Figure 3-20: Equivalent representation of system $\mathbf{DHF}_\tau\mathbf{M}$.

notation as in Figs. 3-20 and 3-21, signal $y(t)$ can be expressed as

$$y(t) = \prod_{i=1}^{d} [f_1^i(t) + f_2^i(t) + f_3^i(t) + f_4^i(t)] = \sum_{\mathbf{m} \in [1:4]^d} f_{\mathbf{m}}(t), \tag{3.18}$$

where

$$f_{\mathbf{m}}(t) = f_{m_1}^1(t) \cdot \ldots \cdot f_{m_d}^d(t), \quad \forall \mathbf{m} \in [1:4]^d.$$

Here components m_i of $\mathbf{m} = (m_1, m_2, \ldots, m_d) \in [1:4]^d$, determine which signal f_j^i, $j \in [1:4]$ from (3.16) participates as a product factor in $f_{\mathbf{m}}(t)$. With signals $e_{m_i,\tau_i}(t)$ as defined in (3.15), it follows that summands in (3.18) can be written as

$$f_{\mathbf{m}}(t) = (-1)^{N_{\mathbf{m}}^2} \prod_{i=1}^{d} e_{m_i,\tau_i}(t) \cdot \prod_{k \in S_{\mathbf{m}}^1 \cup S_{\mathbf{m}}^2} \cos(\omega_c t - \omega_c \tau_k) \cdot \prod_{l \in S_{\mathbf{m}}^3 \cup S_{\mathbf{m}}^4} \sin(\omega_c t - \omega_c \tau_l). \tag{3.19}$$

Products of cosines and sines in (3.19), denoted Π_c and Π_s, can be expressed as sums of

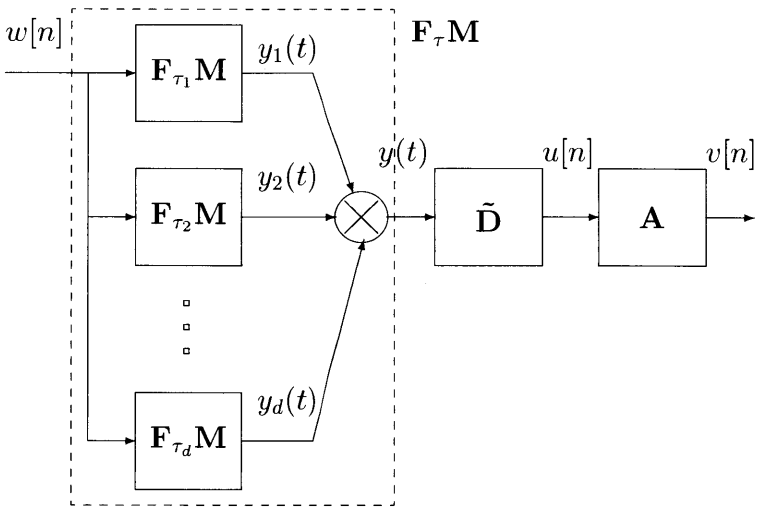


Figure 3-21: System $\mathbf{F}_\tau\mathbf{M}$ as an interconnection of subsystems $\mathbf{F}_{\tau_i}\mathbf{M}$.

complex exponents:

$$\Pi_c = \frac{1}{2^{N_{\mathbf{m}}^1}} \sum_{r \in R_{\mathbf{m}}^c} e^{j\omega_c \bar{\sigma}(r)t} \cdot e^{-j\omega_c(r, \mathcal{P}_{\mathbf{m}}^1 \tau)}, \tag{3.20}$$

$$\Pi_s = \frac{1}{(2j)^{N_{\mathbf{m}}^2}} \sum_{r \in R_{\mathbf{m}}^s} \Pi(r) \cdot e^{j\omega_c \bar{\sigma}(r)t} \cdot e^{-j\omega_c(r, \mathcal{P}_{\mathbf{m}}^2 \tau)}. \tag{3.21}$$

Recall that signals $e_{m_i,\tau_i}(t)$ are obtained by applying pulse amplitude modulation with pulse signals $p_{1,\tau_i}(t)$ or $p_{2,\tau_i}(t)$ on in-phase or quadrature components I and Q of the input signal (or their delayed counterparts $\mathbf{B}I$ and $\mathbf{B}Q$). Let $e_{\mathbf{m},\tau}(t)$ be the product of signals $e_{m_i,\tau_i}(t)$ (as given in (3.19)). We now derive an expression for $e_{\mathbf{m},\tau}(t)$ as a function of signals I, $Q, \mathbf{B}I$ and $\mathbf{B}Q$. We first investigate $e_{\mathbf{m},\tau}(t)$ for $t \in [nT, (n+1)T)$, with $n > 1$ an integer. For a fixed $\mathbf{m} \in [1:4]^d$ and $\tau \in (0,T)^d$ let $S_{\mathbf{m}}^1 \cup S_{\mathbf{m}}^3$, $S_{\mathbf{m}}^2 \cup S_{\mathbf{m}}^4$, $\tau_{min}^{\mathbf{m}}$ and $\tau_{max}^{\mathbf{m}}$ be as defined at the beginning of this section. It follows that $e_{\mathbf{m},\tau}(t) = 0$ for all $t \in [nT, (n+1)T)$ if $\tau_{min}^{\mathbf{m}} > \tau_{max}^{\mathbf{m}}$. Otherwise it is nonzero for $t \in [nT+\tau_{min}^{\mathbf{m}}, nT+\tau_{max}^{\mathbf{m}})$. This is depicted in Fig. 3-22 (for the sake of simplicity, only in-phase components I and $\mathbf{B}I$ are considered, but in general signals Q and $\mathbf{B}Q$ would appear as well). It follows from the above discussion that signal $e_{\mathbf{m},\tau}(t)$ can be expressed as

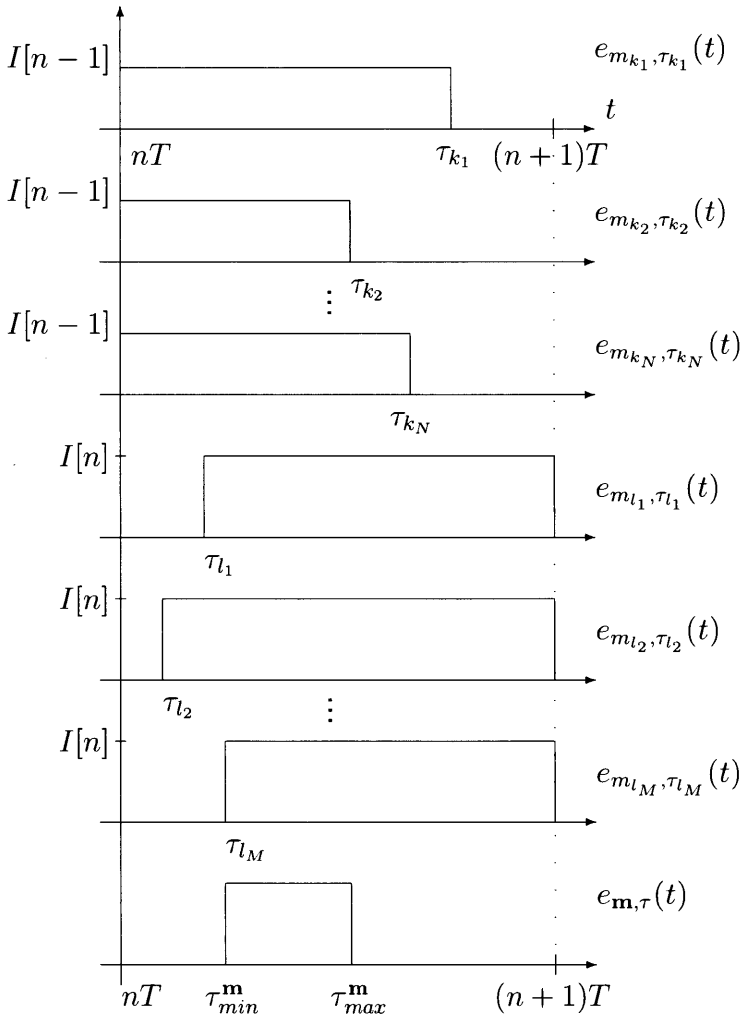


Figure 3-22: Signal $e_{\mathbf{m},\tau}$ for $S^1_{\mathbf{m}} \cup S^3_{\mathbf{m}} = \{k_1, k_2, \ldots, k_N\}$ and $S^2_{\mathbf{m}} \cup S^4_{\mathbf{m}} = \{l_1, l_2, \ldots, l_M\}$, where $N + M = d$.

$$e_{\mathbf{m},\tau}(t) = \sum_{n=-\infty}^{\infty} s_{\mathbf{m}}[n] p_{\mathbf{m},\tau}(t - nT), \tag{3.22}$$

where $p_{\mathbf{m},\tau}(t)$ was defined in (3.8), and DT signal $s_{\mathbf{m}} = s_{\mathbf{m}}[n]$ is defined as

$$s_{\mathbf{m}}[n] = I[n]^{|S^1_{\mathbf{m}}|} \cdot I[n-1]^{|S^2_{\mathbf{m}}|} \cdot Q[n]^{|S^3_{\mathbf{m}}|} \cdot Q[n-1]^{|S^4_{\mathbf{m}}|}.$$

From (3.19)-(3.22), it follows that $f_{\mathbf{m}}(t)$ can be written as

$$f_{\mathbf{m}}(t) = \sum_{r_c \in R^c_{\mathbf{m}}} \sum_{r_s \in R^s_{\mathbf{m}}} C_{r_c,r_s} \cdot e^{j\sigma(r_c,r_s)\omega_c t} \cdot \sum_{n=-\infty}^{\infty} s_{\mathbf{m}}[n] p_{\mathbf{m},\tau}(t - nT), \tag{3.23}$$

where $\sigma(r_c, r_s) = \sigma(r) = \sum_{i=1}^{N^1_{\mathbf{m}}} r_c(i) + \sum_{l=1}^{N^2_{\mathbf{m}}} r_s(l)$, and

$$C_{r_c,r_s} = \frac{(j)^{N^2_{\mathbf{m}}}}{2^d} \cdot e^{-j\omega_c[(r_c,\mathcal{P}^1_{\mathbf{m}}\tau)+(r_s,\mathcal{P}^2_{\mathbf{m}}\tau)]} \cdot \Pi(r_s), \tag{3.24}$$

depends only on $\mathbf{m}$. Therefore, the output signal y of system $\mathbf{F}_\tau\mathbf{M}$, can be expressed in terms of I, Q, ω_c. T and τ by plugging expression (3.23), for $f_{\mathbf{m}}(t)$, into (3.18). Thus we have found an explicit input-output description of system $\mathbf{F}_\tau\mathbf{M}$, which concludes the first part of the proof.

In order to find explicit expressions for the frequency responses $G_{\mathbf{m}}$, we first determine frequency domain relationship between signals u and y (see Fig. 3-19). Recall that $u = \tilde{\mathbf{D}}y = \mathbf{E}\mathbf{H_0}\mathbf{X_c}\mathbf{H}y$. Let $U(\Omega)$ and $Y(\omega)$ denote the Fourier transforms of signals $u[n]$ and $y(t)$ respectively. Also let $H(\omega)$ and $H_0(\omega)$ be the frequency responses of ideal band-pass and low-pass filters $\mathbf{H}$ and $\mathbf{H_0}$, given by

$$H(\omega) = \begin{cases} 1, & |\omega_c - |\omega|| \leq \pi/T \\ 0, & \text{otherwise} \end{cases}, H_0(\omega) = \begin{cases} 1, & |\omega| \leq \pi/T \\ 0, & \text{otherwise} \end{cases}. \tag{3.25}$$

The following sequence of equalities hold

$$\mathcal{F}\{\mathbf{H}y\} = Y(\omega)H(\omega),$$

$$\mathcal{F}\{\mathbf{X_c H} y\} = Y(\omega + \omega_c)H(\omega + \omega_c),$$
$$\mathcal{F}\{\mathbf{H_0 X_c H} y\} = Y(\omega + \omega_c)H(\omega + \omega_c)H_o(\omega),$$
$$U(\Omega) = Y\left(\frac{\Omega}{T} + \omega_c\right) H\left(\frac{\Omega}{T} + \omega_c\right) H_0\left(\frac{\Omega}{T}\right).$$

From the definition of $H(\omega)$ and $H_0(\omega)$, $U(\Omega)$ simplifies to

$$U(\Omega) = Y\left(\frac{\Omega}{T} + \omega_c\right), \tag{3.26}$$

which gives frequency domain relationship between y and u.

Next we express $Y(\omega)$ in terms of $S_{\mathbf{m}}(\Omega) = \mathcal{F}\{s_{\mathbf{m}}[n]\}$. For the sake of simplicity, we assume that $y(t)$ is equal to $f_{\mathbf{m}}(t)$ for some fixed $\mathbf{m}$, i.e., we omit the sum in (3.18). Since $\sigma(r) \in \mathbb{Z}$ and $\omega_c T = 2\pi n$, $n \in \mathbb{Z}$, it follows from (3.23) that

$$Y(\omega) = S_{\mathbf{m}}(\omega T) \cdot \sum_{r_c \in R^c_{\mathbf{m}}} \sum_{r_s \in R^s_{\mathbf{m}}} C_{r_c, r_s} P_{\mathbf{m},\tau}\left(\omega - \sigma(r)\omega_c\right). \tag{3.27}$$

It now follows from (3.26) and (3.27) that

$$U(\Omega) = S_{\mathbf{m}}(\Omega) \sum_{r_c \in R^c_{\mathbf{m}}} \sum_{r_s \in R^s_{\mathbf{m}}} C_{r_c, r_s} P_{\mathbf{m},\tau}\left(\tilde{\Omega}\right),$$

where $\tilde{\Omega} = \frac{\Omega}{T} - \sigma(r)\omega_c + \omega_c$ and C_{r_c,r_s} as defined in (3.24).
Therefore, the frequency response $G_{\mathbf{m}}(\Omega)$ of a LTI system mapping $s_{\mathbf{m}}$ to u is given by

$$G_{\mathbf{m}}(\Omega) = \sum_{r_c \in R^c_{\mathbf{m}}} \sum_{r_s \in R^s_{\mathbf{m}}} C_{r_c, r_s} P_{\mathbf{m},\tau}\left(\frac{\Omega}{T} - \sigma(r)\omega_c + \omega_c\right). \tag{3.28}$$

This concludes the proof of Lemma 3.6.1. □

In Lemma 3.6.1, it was assumed that $\tau_i \in [0, T)$, $\forall i \in [1 : d]$, but in general τ_i can take any positive real value depending on the depth of (2), i.e., vector $\mathbf{k}$ associated with τ is not necessarily the zero vector. Suppose now that $\tau = \mathbf{k}T + \tau'$, where $\tau' \in [0, T)^d$, and $\mathbf{k} \neq \mathbf{0}$. In the rest of this proof we adopt the same notation for corresponding signals and

systems as in the proof of Lemma 3.6.1. It is clear, in this case, that mapping from y to u is identical to the one derived for $\tau \in [0, T)$. Therefore, in order to prove the statement of Theorem 3.2.1 it suffices to determine a new relationship between signals w and y.

Let first $d = 1$, i.e., $\tau = kT + \tau'$, with $k \in \mathbb{N}$ and $\tau' \in [0, T)$. Analogously to the case in the proof of Lemma 3.6.1, it follows that signal y can be expressed as

$$y(t) = [e_{1,\tau}(t) + e_{2,\tau}(t)]\cos(\omega_c t - \omega_c \tau) - [e_{3,\tau}(t) + e_{4,\tau}(t)]\sin(\omega_c t - \omega_c \tau),$$

where

$$\begin{aligned}
e_{1,\tau} &= \mathbf{Z_1}\mathbf{B}^{k+1}I, \quad \text{i.e.} \quad e_{1,\tau}(t) = \frac{1}{T}\sum_{n=-\infty}^{\infty} I[n-k-1]p_{1,\tau'}(t-nT), \\
e_{2,\tau} &= \mathbf{Z_2}\mathbf{B}^{k}I, \quad \text{i.e.} \quad e_{2,\tau}(t) = \frac{1}{T}\sum_{n=-\infty}^{\infty} I[n-k]p_{2,\tau'}(t-nT), \\
e_{3,\tau} &= \mathbf{Z_1}\mathbf{B}^{k+1}Q, \quad \text{i.e.} \quad e_{3,\tau}(t) = \frac{1}{T}\sum_{n=-\infty}^{\infty} Q[n-k-1]p_{1,\tau'}(t-nT), \\
e_{4,\tau} &= \mathbf{Z_2}\mathbf{B}^{k}Q, \quad \text{i.e.} \quad e_{4,\tau}(t) = \frac{1}{T}\sum_{n=-\infty}^{\infty} Q[n-k]p_{2,\tau'}(t-nT).
\end{aligned} \tag{3.29}$$

Here $\mathbf{B}^k$ denotes the composition of $\mathbf{B}$ with itself k times, i.e., $\mathbf{B}^k : x \mapsto y$ such that $y[n] = x[n-k]$.

For $d > 1$, reasoning similar to that in the proof of Lemma 3.6.1 (see (3.17)-(3.22)), leads to the following expression for $e_{\mathbf{m},\tau}$:

$$e_{\mathbf{m},\tau}(t) = \sum_{n=-\infty}^{\infty} s_{\mathbf{m},\mathbf{k}}[n]p_{\mathbf{m},\tau'}(t-nT), \tag{3.30}$$

where

$$s_{\mathbf{m},\mathbf{k}}[n] = \prod_{i\in S^1_{\mathbf{m}}} I[n-k_i-1]\cdot\prod_{i\in S^2_{\mathbf{m}}} I[n-k_i]\cdot\prod_{i\in S^3_{\mathbf{m}}} Q[n-k_i-1]\cdot\prod_{i\in S^4_{\mathbf{m}}} Q[n-k_i], \tag{3.31}$$

and $p_{\mathbf{m},\tau'}(t)$ as defined in (3.8). Let $S_{\mathbf{m},\mathbf{k}} = S_{\mathbf{m},\mathbf{k}}(\Omega)$ be the Fourier transform of $s_{\mathbf{m},\mathbf{k}}$. With (3.30) at hand, it is straightforward to find the analytic expression for $U = \mathcal{F}u$, in

terms of $S_{\mathbf{m},\mathbf{k}}$. Similarly to (3.23)-(3.27), the Fourier transform $Y = \mathcal{F}y$, can be written as

$$Y(\omega) = S_{\mathbf{m},\mathbf{k}}(\omega T) \cdot \sum_{r_c \in R^c_{\mathbf{m}}} \sum_{r_s \in R^s_{\mathbf{m}}} C_{r_c,r_s} P_{\mathbf{m},\tau} \left(\omega - \sigma(r)\omega_c\right), \tag{3.32}$$

where C_{r_c,r_s} as defined in (3.24), with τ replaced by τ'. It follows from (3.26) and (3.32) that

$$U(\Omega) = S_{\mathbf{m},\mathbf{k}}(\Omega) G_{\mathbf{m}}(\Omega), \tag{3.33}$$

with $G_{\mathbf{m}}(\Omega)$ as defined in (3.28). Statement of the theorem now immediately follows from the above equality. This concludes the proof.

Chapter 4

Approximate Baseband Modeling and Digital Compensation of Digitally Implemented Pulse-Width Modulation (DPWM)

[1]

This chapter is organized as follows. In this section 4.2, we describe the principle of operation of a carrier-based PWM system and give simple mathematical descriptions of the system in both CT and DT cases. A novel closed-form input-output model for a multi-level DPWM is derived in section 4.3. In section 4.4, we propose a method for compensating in-band distortion in DPWM systems by preprocessing its input with a carefully designed delta-sigma modulator. MATLAB simulation and measurement results that show performance of the proposed method are presented in section 4.5. The chapter summarized in section 4.6 and the main statements are proven in section 4.7.

[1]This work was initiated while Omer Tanovic was an intern at Mitsubishi Electric Research Laboratories (MERL).

4.1 Terminology

The main classification of pulse-width modulation (PWM) systems is based on the nature of the time-domain (continuous-time or discrete-time) of its input and output signals. In this paper, we assume that input and output signals of a PWM system are either both continuous-time (CT) or both discrete-time (DT) signals. That is, the pulse-width modulator is considered as a system mapping CT signals to CT signals, or DT signals to DT signals. The former is commonly called the analog PWM (or more accurately continuous-time PWM), while the latter is called digital PWM (or discrete-time PWM). It should be noted here that there is some ambiguity in the literature as to what is meant by digital PWM. In the power electronics community, digital PWM is commonly regarded as a system mapping DT signals to CT signals. In this context, it is implicitly assumed that the input DT signal undergoes some form of digital-to-analog conversion, and its CT version is compared to a CT carrier to produce the PWM output signal. We should emphasize that such PWM schemes have different underlying mechanisms and are not considered in this thesis. In the rest of this chapter, analog PWM systems are denoted by APWM, while digital PWM systems are denoted by DPWM, in order to emphasize the time-domain in which the system is defined.

4.2 Background and Problem Formulation

4.2.1 Principle of Operation of APWM

In carrier-based pulse-width modulation schemes, a PWM output signal is generated by comparing a PWM input to fixed periodic signals called the PWM carrier or reference signals (e.g., sawtooths or sinusoids). A simplified block diagram of a 2-level (i.e., single-carrier) PWM system, with a trailing-edge sawtooth reference, is depicted in Figure 4-1. As can be seen in this example, at every time instant, the output signal is equal to 0 when the input is smaller than the reference, and is equal to 1 otherwise.

In this thesis we consider somewhat restricted family of reference signals, but large

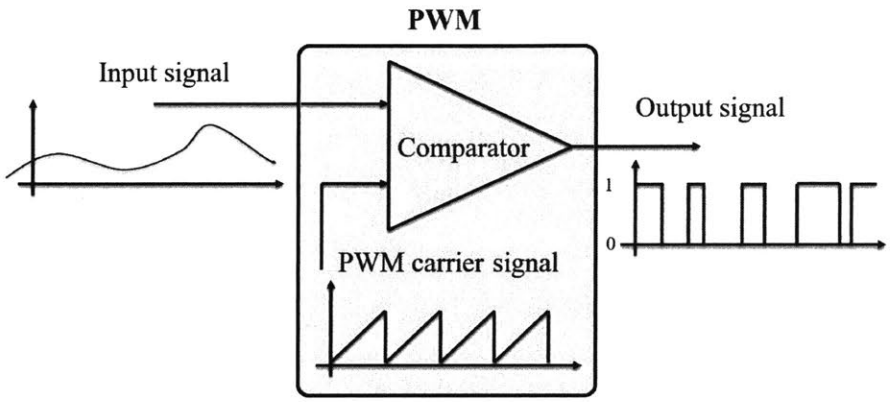


Figure 4-1: Block diagram describing basic operation of APWM.

enough that it includes the popular choice of sinusoid and sawtooth reference signals. Without loss of generality, we assume that the reference signals are normalized in amplitude to values in the interval $(0, 1]$ (it is straightforward to extend the results to other amplitude ranges).

Definition 4.2.1. *For a fixed $T_p > 0$, let $c : \mathbb{R} \to \mathbb{R}$ be a continuous function such that $c(t + T_p) = c(t)$ for all $t \in \mathbb{R}$, $\max_t c(t) = 1$, $\min_t c(t) = \epsilon$ for some $\epsilon \in (0, 1)$, and c has exactly one minimum and one maximum on each period $[kT_p, (k+1)T_p)$, for all $k \in \mathbb{Z}$. Let $\mathcal{C}_{T_p}$ be the set of all such signals c. Elements of $\mathcal{C}_{T_p}$ are called admissible APWM carrier or reference signals.*

For a given admissible reference signal $c \in \mathcal{C}_{T_p}$ with period T_p, a 2-level APWM can now be defined as the system $\mathbf{P}_c : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ which maps input signal $a \in \mathcal{L}^2(\mathbb{R})$ to output signal $y = \mathbf{P}_c a \in \mathcal{L}^2(\mathbb{R})$ such that

$$y(t) = \begin{cases} 0, & a(t) < c(t) \\ 1, & a(t) \geq c(t), \end{cases}, \quad \forall t \in \mathbb{R}. \tag{4.1}$$

T_p is called the carrier (or reference) period, and $f_p = 1/T_p$ (or, equivalently, $\omega_p = 2\pi/T_p$) the carrier (or reference) frequency of APWM (sometimes also called the pulse frequency of APWM). It follows from (4.1) that the output signal y is generated by comparing the input signal a to the carrier signal c at every time instant t.

Remark: In the literature, traditionally, the reference signals are functions $c : \mathbb{R} \to [0,1]$ with $\min_t c(t) = 0$. In that case, the APWM, when driven by finite energy signals, would not, in general, produce finite energy signals. Since our goal is to analyze spectral properties of PWM output signals (both APWM and DPWM), it is meaningful to restrict the set of admissible references in order to ensure that APWM acts as a map from $\mathcal{L}^2(\mathbb{R})$ to $\mathcal{L}^2(\mathbb{R})$. One can assume that ϵ is arbitrarily small in which case APWM with admissible references as defined in this thesis, acts approximately as the one with traditional references with $\min_t c(t) = 0$. We emphasize that our choice is for pure mathematical convenience since in reality all input signals are of finite length and therefore APWM systems would generate signals in $\mathcal{L}^2(\mathbb{R})$ regardless of the value of ϵ.

In the case of multi-level APWM, the output signal is generated by comparing the input signal to multiple reference signals which are scaled and shifted in amplitude to disjointly cover the $(0,1]$ interval. We now define this formally. Let M be a positive integer, and let $\mathcal{A} = \{\alpha_0, \ldots, \alpha_M\} \subset [0,1]$ be such that

$$0 = \alpha_0 < \cdots < \alpha_{m-1} < \alpha_m < \cdots < \alpha_M = 1. \tag{4.2}$$

Let $C = \{c'_1, \ldots, c'_M\} \subset \mathcal{C}_{T_p}$ and let signals $c_m = c_m(t)$, for $m \in \{1, \ldots, M\}$, be defined by

$$c_m(t) = (\alpha_m - \alpha_{m-1})c'_m(t) + \alpha_{m-1}. \tag{4.3}$$

The $(M+1)$-level APWM system, with reference signals $c_1, \ldots, c_M$ as in (4.3), can now be defined as the operator $\mathbf{P}_{c_1,\ldots,c_M} : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ which maps input signal $a \in \mathcal{L}^2(\mathbb{R})$ to output signal $y \in \mathcal{L}^2(\mathbb{R})$ as defined by

$$y(t) = \begin{cases} 0, & a(t) < c_1(t) \\ \alpha_m, & c_m(t) \le a(t) < c_{m+1}(t), \quad 1 \le m \le M-1 \\ 1, & c_M(t) \le a(t), \end{cases} \cdot \tag{4.4}$$

An example of output signal generation in a 3-level PWM scheme with sawtooth reference

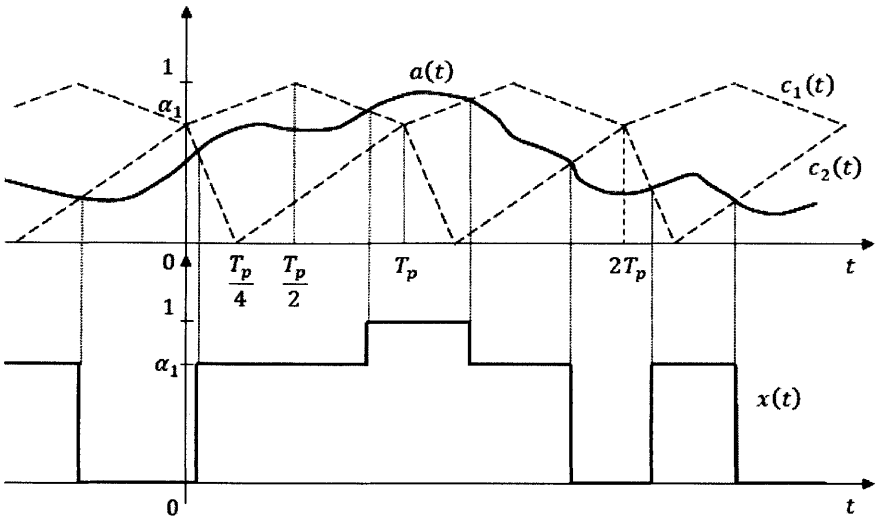


Figure 4-2: An example of output signal generation in 3-level PWM with sawtooth carrier signals.

signals is illustrated in Fig.4-2. In this case, $M = 2$ and $\alpha_1 = 3/4$, and, therefore, the output signal $y(t)$ assumes values in $\left\{0, \frac{3}{4}, 1\right\}$.

Definitions of APWM in (4.1) and (4.4) are very simple and easy to understand, but they clearly have limited value with regard to analysis of the spectral properties of y. A more useful model, with respect to analysis of the APWM output spectra, is presented in the next section.

4.2.2 Time-Domain Analysis of APWM

In this section, we present a generalization of the input-output models of APWM found in [87] (for APWM with double-edge symmetric sawtooth signals) or in [103] (for general sawtooth signals). We extend the above results to model APWM's with arbitrary admissible reference signals given in definition 4.2.4. The generalized model serves as a starting point for the derivation of the DPWM model presented in the next section.

In this section, we present a non-trivial time-domain description of APWM system, which is then used, in the following sections, to derive an equivalent model for the DPWM system. First we introduce some necessary notation.

Let functions $f_1 : \mathcal{C}_{T_p} \times \mathbb{R} \to [0, 1]$ and $f_0 : \mathcal{C}_{T_p} \times \mathbb{R} \to [-1, 1]$ be defined as follows:

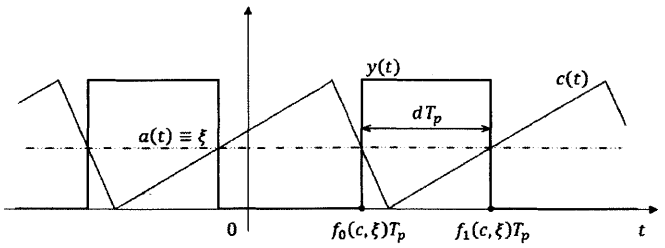


Figure 4-3: An example showing geometric interpretation of f_0, f_1 and d, for an arbitrary sawtooth reference signal.

(1) $f_1(c,\xi) = f_0(c,\xi) = 0$ for all $c \in \mathcal{C}_{T_p}$ when $\xi < \min_t c(t)$.

(2) For any $\xi \in (0,1)$ let $f_1 = f_1(c,\xi)$ be the smallest positive real number such that $c(f_1 T_p) = \xi$ and there exists $\delta > 0$ such that $c(uT_p) \leq c(vT_p)$ for all $u, v \in (f_1 - \delta, f_1 + \delta)$ and $u \leq v$. Similarly, let $f_0 = f_0(c,\xi)$ be the largest real number such that $f_0(c,\xi) < f_1(c,\xi)$, $c(f_0 T_p) = \xi$ and there exists $\delta > 0$ such that $c(uT_p) \geq c(vT_p)$ for all $u, v \in (f_0 - \delta, f_0 + \delta)$ and $u \leq v$.

(3) $f_1(c,\xi) = 1, f_0(c,\xi) = 0$ for all $c \in \mathcal{C}_{T_p}$ when $\xi \geq 1$.

The interpretation of f_0 and f_1 for an arbitrary admissible baseline reference signal $c = c(t)$, with period $T_p > 0$, is given in Figure 4-3. As can be seen, f_0 and f_1 are the relative (to T_p) times when constant signal $a(t) \equiv \xi$ crosses the reference on the trailing and rising edge, respectively. It follows that $d = f_1(c,\xi) - f_0(c,\xi)$ is the (relative) width of the pulse starting at time $t_0 = f_0(c,\xi)T_p$ and ending at time $t_1 = f_1(c,\xi)T_p$ (see Figure 4-3).

The following theorem gives an input-output model of a 2-level APWM, which depicts harmonic nature of the PWM output signal and is more convenient for spectral analysis of PWM signals than the model given by (4.1).

Theorem 4.2.2. *Let $c \in \mathcal{C}_{T_p}$ with period T_p. APWM system $\mathbf{P}_c$ maps $a \in \mathcal{L}^2(\mathbb{R})$ to $y = \mathbf{P}_c a \in \mathcal{L}^2(\mathbb{R})$ such that, for all $t \in \mathbb{R}$ with $a(t) \neq c(t)$,*

$$y(t) = \sum_{k=-\infty}^{\infty} C_k(a(t)) e^{jk\omega_c t}, \tag{4.5}$$

where

$$C_k(a(t)) = \frac{\sin(\pi k d(t))}{\pi k} e^{-j\pi k D(t)}, \tag{4.6}$$

with

$$d(t) = f_1(c, a(t)) - f_0(c, a(t)), \quad D(t) = f_1(c, a(t)) + f_0(c, a(t)), \quad \forall t \in \mathbb{R}. \tag{4.7}$$

When $a(t) = c(t)$, *then* $y(t) = 1/2$.

Proof. See the section 4.7.1. □

Remark 1: The convergence of the series on the right-hand side of (4.5) should be understood in the sense of its principal value, that is the infinite sum in (4.5) is equal to $\lim_{K\to\infty} \sum_{k=-K}^{K} C_k(a(t)) e^{jk\omega_p t}$. Equality in (4.5) holds point-wise, and is true for every $t \in \mathbb{R}$ for which $c(t) \neq a(t)$. At the points of discontinuity of y, that is, for t such that $c(t) = a(t)$, we have that

$$y(t) = \frac{1}{2}(y(t_-) + y(t_+)) = \frac{1}{2},$$

where $y(t_-)$ and $y(t_+)$ are the directional limits of y at t.

Remark 2: For the popular choice of trailing edge and symmetric double-edge sawtooth references, expressions for d and D are very simple and (4.5) significantly simplifies. For all sawtooth signals we have the following: $d(t) = 0$ when $a(t) \leq \min_t c(t)$, $d(t) = a(t)$ when $\min_t c(t) < a(t) \leq 1$, and $d(t) = 1$ otherwise. Expression for D now simplifies to: $D(t) = d(t)$ for the trailing-edge sawtooth, $D(t) = 0$ for the in-phase double-edge symmetric sawtooth, and $D(t) = 1$ for the out-of-phase double-edge symmetric sawtooth.

Let now $M \in \mathbb{N}$. For $C = \{c'_1, c'_2, \ldots, c'_M\} \subset \mathcal{C}_{T_p}$ and $\mathcal{A} = \{\alpha_0, \ldots, \alpha_M\} \subset [0, 1]$ satisfying (4.2), let c_m, for $m \in \{1, \ldots, M\}$, be the M contiguous reference signals as defined in (4.3). The following theorem gives an input-output model for the $(M + 1)$-level APWM, similar to that for the 2-level case from theorem 4.2.2.

Theorem 4.2.3. *APWM system* $\mathbf{P}_{c_1,\ldots,c_M}$ *maps* $a \in \mathcal{L}^2(\mathbb{R})$ *to* $y = \mathbf{P}_{c_1,\ldots,c_M} a \in \mathcal{L}^2(\mathbb{R})$ *such*

that

$$y(t) = \alpha(a(t)) + \sum_{k=-\infty}^{\infty} C_k(a(t)) e^{jk\omega_c t}, \tag{4.8}$$

where

$$\alpha(a(t)) = \begin{cases} 0, & a(t) < \alpha_0 \\ \alpha_{m-1}, & a(t) \in [\alpha_{m-1}, \alpha_m), \forall m \in \{1, \ldots, M\} \\ 1, & a(t) > \alpha_M \end{cases}, \tag{4.9}$$

$$C_k(a(t)) = \frac{\gamma(t) \sin(\pi k d(t))}{\pi k} e^{-j\pi k D(t)}, \tag{4.10}$$

with $\gamma(t) = d(t) = D(t) = 0$ *when* $a(t) \leq \alpha_0$, $\gamma(t) = d(t) = D(t) = 1$ *when* $a(t) > \alpha_M$, *and*

$$\gamma(t) = \alpha_m - \alpha_{m-1},$$

$$d(t) = f_1\left(c'_m, \frac{a(t) - \alpha_{m-1}}{\alpha_m - \alpha_{m-1}}\right) - f_0\left(c'_m, \frac{a(t) - \alpha_{m-1}}{\alpha_m - \alpha_{m-1}}\right),$$

$$D(t) = f_1\left(c'_m, \frac{a(t) - \alpha_{m-1}}{\alpha_m - \alpha_{m-1}}\right) + f_0\left(c'_m, \frac{a(t) - \alpha_{m-1}}{\alpha_m - \alpha_{m-1}}\right),$$

when $a(t) \in (\alpha_{m-1}, \alpha_m]$, *for all* $m \in \{1, \ldots, M\}$.
When $a(t) = c_m(t)$, *then* $y(t) = \alpha_m$, *for all* $m \in \{1, \ldots, M\}$.

Proof. See the section 4.7.2. □

The popularity of APWM stems from the fact that when the admissible references c'_m are sawtooth signals, and $a(t) \in (0, 1)$, the expression (4.8) for $y(t)$ simplifies to

$$y(t) = y_0(t) + \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} y_k(t) e^{jk\omega_c t}, \tag{4.11}$$

where

$$\begin{aligned} y_0(t) &= a(t), \\ y_k(t) &= C_k(a(t)) = \frac{\gamma(t) \sin(\pi k d(t))}{\pi k} e^{-j\pi k D(t)}, \quad \forall k \neq 0. \end{aligned} \tag{4.12}$$

In that case, it follows from theorem 4.2.3 that the output $y(t)$ of the APWM system

can be described as a sum of the baseband component $y_0(t)$, equal to the input signal $a(t)$, and phase-amplitude modulated harmonics $y_k(t)\exp(jk\omega_c t)$ at non-zero integer multiples of the PWM carrier frequency ω_p. This characteristic of the APWM output signal is of great importance when input a is approximately band-limited with its bandwidth B_w much smaller than the PWM carrier frequency f_p. In this case, it is possible to recover signal a by low-pass filtering signal y, where the amount of spectral distortion introduced by the APWM system can be made arbitrarily small by increasing the ratio f_p/B_w. This property of APWM has been the main driving force behind its popularity in various engineering applications and especially in communications [92], where f_p is often on the order of the RF carrier frequency and B_w is commonly much smaller than f_p (at least an order of magnitude).

In the next section, we describe the operation principle of DPWM and show that it can be understood as a sampled version of APWM. This characterization will be of great importance in time-domain analysis of DPWM.

4.2.3 Principle of Operation of DPWM

The operation of DPWM is defined analogous to the continuous-time case. Let us first define admissible reference signals in the discrete-time case.

Definition 4.2.4. *For a fixed $N \in \mathbb{Z}$, $N > 1$, let $\tilde{c} : \mathbb{Z} \to (0, 1]$ be a periodic discrete-time signal with fundamental period N, such that for any $T > 0$, continuous-time signal $c = c(t)$ defined by $c(nT) = \tilde{c}[n]$ and $c(t) = \frac{\tilde{c}[n+1]-\tilde{c}[n]}{T}(t - nT) + \tilde{c}[n]$ for $t \in (nT, (n+1)T)$, is an admissible reference signal with period $T_p = NT$, i.e., $c \in \mathcal{C}_{T_p}$. Let $\tilde{\mathcal{C}}_N$ be the set of all such discrete-time signals $\tilde{c}$. Elements of $\tilde{\mathcal{C}}_N$ are called admissible DPWM carrier or reference signals.*

Remark: Signal c is a linear interpolation of $\tilde{c}'$ with respect to time sample T. Likewise, $\tilde{c}$ can be understood as generated from c by sampling at rate $1/T$.

For given $N, M \in \mathbb{Z}$, $N > 1$, $M > 0$, let $\mathcal{A} = \{\alpha_0, \ldots, \alpha_M\} \subset [0, 1]$ be such that (4.2) is satisfied and let $\tilde{C} = \{\tilde{c}'_1, \ldots, \tilde{c}'_M\} \subset \tilde{\mathcal{C}}_N$ be a collection of admissible DPWM reference

signals. Let DT signals $\tilde{c}_m = \tilde{c}_m[n]$, for all $m \in \{1, \ldots, M\}$, be defined by

$$\tilde{c}_m[n] = (\alpha_m - \alpha_{m-1})\tilde{c}'_m[n] + \alpha_{m-1}. \tag{4.13}$$

An $(M+1)$-level digital pulse-width modulator with reference signals $\tilde{c}_m$, for $m \in \{1, \ldots, M\}$, is defined as a system $\tilde{\mathbf{P}}_{\tilde{c}_1,\ldots,\tilde{c}_M} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ that maps input signal $\tilde{a} \in \ell^2(\mathbb{R})$ to output signal $\tilde{y} = \tilde{\mathbf{P}}_{\tilde{c}_1,\ldots,\tilde{c}_M}\tilde{a} \in \ell^2(\mathbb{R})$, as defined by

$$\tilde{y}[n] = \begin{cases} 0, & \tilde{a}[n] < \tilde{c}_1[n], \\ \alpha_m, & \tilde{c}_m[n] \le \tilde{a}[n] < \tilde{c}_{m+1}[n], \quad m \in \{1 \ldots, M-1\}, \\ 1, & \tilde{c}_M[n] \le \tilde{a}[n]. \end{cases} \tag{4.14}$$

Parameter N is called the oversampling ratio (OSR) of DPWM.

Let us now give an alternative definition of DPWM which illuminates its relationship with APWM.

For $T > 0$, let $\mathbf{R}_T : \ell^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ be *the perfect recovery system* that maps $\tilde{w} \in \ell^2(\mathbb{R})$ to $w = \mathbf{R}_T w \in \mathcal{L}^2(\mathbb{R})$ as defined by

$$w(t) = (\mathbf{R}_T\tilde{w})(t) = \sum_{n=-\infty}^{\infty} \tilde{w}[n] \cdot \frac{\sin\left(\frac{\pi}{T} \cdot (t - nT)\right)}{\frac{\pi}{T}(t - nT)}. \tag{4.15}$$

Let $\mathbf{S}_T : \mathcal{L}^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be *the sampling system* that maps $v \in \mathcal{L}^2(\mathbb{R})$ to $\tilde{v} = \mathbf{S}_T v \in \ell^2(\mathbb{R})$ as defined by

$$\tilde{v}[n] = (\mathbf{S}_T v)[n] = v(nT). \tag{4.16}$$

It is now clear that, for any fixed $T > 0$, the following equality is true

$$\tilde{\mathbf{P}}_{\tilde{c}_1,\ldots,\tilde{c}_M} = \mathbf{S}_T \mathbf{P}_{c_1,\ldots,c_M} \mathbf{R}_T,$$

where $c_1, \ldots, c_M$ are the linear interpolation of $\tilde{c}_1, \ldots, \tilde{c}_M$ with respect to the time-sample T. That is, DPWM system is equivalent to a series interconnection of the perfect recovery system, the APWM and the sampler, as depicted in Figure 4-4. This alternative definition

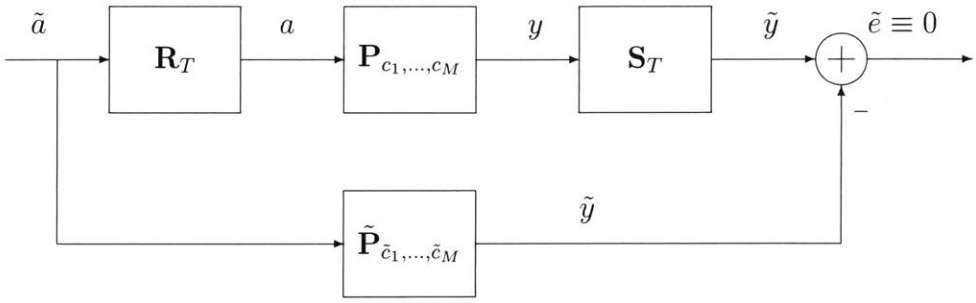


Figure 4-4: Equivalent block diagram representation of DPWM.

of DPWM suggests that it can be understood as a 'sampled version' of APWM. It should be noted here that the above characterization of DPMW would still hold if the perfect recovery system in (4.15) is replaced by any perfect interpolation system $\mathbf{R}_T : \ell^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ that maps $w \in \ell^2(\mathbb{R})$ to $w_c = \mathbf{R}_T w \in \mathcal{L}^2(\mathbb{R})$ such that $w_c(nT) = w[n]$ for all n.

As noted in the previous section, signal $y_c = \mathbf{P}_{c_1,\ldots,c_M}\mathbf{R}_T a$ has, in general, an infinite bandwidth. Sampling such a signal (independent of how large the sampling frequency) would produce an infinite amount of spectral aliasing in the baseband range of frequencies, since every copy of the original spectrum contributes to the overall level of aliasing [110]. This aliasing noise represents the major distortion source in DPWM systems and is one of the main limitations for their practical deployment.

As previously noted, theorem 4.2.3 describes the APWM output $y = \mathbf{P}_c a$ as a sum of the baseband component y_0 (which, is in the case of sawtooth references, equal to a) and the amplitude modulated harmonics $y_k(t)\exp(j2\pi f_p t)$. It is natural to ask the following question: does there exist a harmonic series description of the signal $\tilde{y} = \tilde{\mathbf{P}}_{\tilde{c}}\tilde{a}$ similar to that of y from theorem 4.2.3? If yes, then how do harmonics of $\tilde{y}$ depend on $\tilde{a}$? In the next section we use the above characterization of DPWM to answer in affirmative the former question, and give an explicit input-output model which describes dependence of the harmonics of $\tilde{y}$ on the DPWM input signal $\tilde{a}$. This model also gives time-domain description of the aliasing noise, and suggests a method for mitigating it.

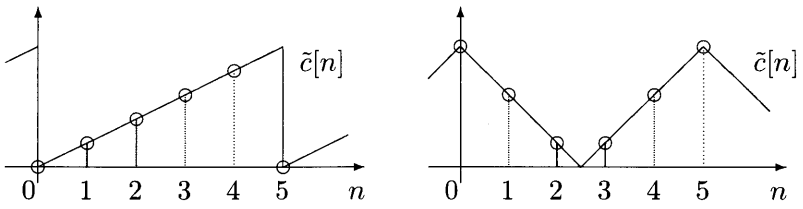


Figure 4-5: Examples of reference signal amplitude levels: (a) trailing-edge sawtooth, (b) symmetric double-edge sawtooth.

4.3 Time-Domain Analysis of Carrier-Based DPWM

4.3.1 2-Level DPWM

In this section, we consider the case when $M = 1$, and for simplicity write $c(t)$ (equivalently, $\tilde{c}[n]$) instead of $c_1(t)$ (equivalently, $\tilde{c}_1[n]$). We first introduce some additional notation in order to make the main theorem statement more compact and easier to parse.

Let $p = (k_0, k_1, \ldots, k_{N-1})$ be the permutation of $\{0, 1, \ldots, N-1\}$, such that $0 \leq \tilde{c}[k_0] \leq \tilde{c}[k_1] \leq \cdots \leq \tilde{c}[k_{N-1}] \leq 1$. That is, $\tilde{c}[k_0], \tilde{c}[k_1], \ldots, \tilde{c}[k_{N-1}]$ is a sorted sequence of amplitude values over one period of the reference signal $\tilde{c}$. For any $i \in \{0, 1, \ldots, N-2\}$, let l_i be defined as

$$l_i = \begin{cases} 0, & \tilde{c}[k_i] < \tilde{c}[k_{i+1}], \\ 1, & \tilde{c}[k_i] = \tilde{c}[k_{i+1}]. \end{cases}$$

Let $L_{\tilde{c}}$ be the number of unique amplitude values of $\tilde{c}$, that is, $L_{\tilde{c}}$ is the cardinality of the set $\{\tilde{c}[0], \ldots, \tilde{c}[N-1]\}$. It is then easy to show that $L_{\tilde{c}} = N - \sum_{i=0}^{N-2} l_i$. For example, if $N = 5$ and $\tilde{c}$ is a trailing-edge sawtooth signal (see Fig. 4-5a), then $(k_0, \ldots, k_4) = (0, 1, 2, 3, 4)$ and $\tilde{c}[k_0] < \tilde{c}[k_1] < \tilde{c}[k_2] < \tilde{c}[k_3] < \tilde{c}[k_4]$, so $l_i = 0$ for all i and, hence, $L_{\tilde{c}} = 5$. On the other hand, if $\tilde{c}$ is a symmetric double-edge sawtooth signal (see Fig. 4-5b), then $(k_0, \ldots, k_4) = (2, 3, 1, 4, 0)$ with $\tilde{c}[k_0] = \tilde{c}[k_1] < \tilde{c}[k_2] = \tilde{c}[k_3] < \tilde{c}[k_4]$, so $l_0 = l_2 = 1$, $l_1 = l_3 = 0$ and, hence, $L_{\tilde{c}} = 3$.

Let $\mathbf{Q}_{N,\tilde{c}} : \mathbb{R} \to \{0, 1/N, \ldots, (N-1)/N, 1\}$ and $\mathbf{Q}_N : \mathbb{R} \to \{(2i+1)/2N\}_{i \in \mathbb{Z}}$ be

quantizers such that

$$\mathbf{Q}_{N,\tilde{c}}(\xi) = \begin{cases} 0, & \xi < \tilde{c}[k_0], \\ \frac{i+l_i+1}{N}, & \xi \in [\tilde{c}[k_i], \tilde{c}[k_{i+1}]), 0 \leq i \leq N-2, \\ 1, & \xi \geq \tilde{c}[k_{N-1}], \end{cases} \tag{4.17}$$

$$\mathbf{Q}_N(\xi) = \frac{2i+1}{2N}, \quad \xi \in \left[\frac{i}{N}, \frac{i+1}{N}\right), \forall i \in \mathbb{Z}. \tag{4.18}$$

It is not hard to see that quantizer $\mathbf{Q}_{N,\tilde{c}}$ has exactly $L_{\tilde{c}}$ quantization levels, while $\mathbf{Q}_N$ has N output levels.

In the following theorem, we give an input-output model of DPWM, which depicts harmonic nature of the DPWM output signal, similar to the result from theorem 4.2.2 for APWM.

Theorem 4.3.1. *2-level DPWM system $\tilde{\mathbf{P}}_{\tilde{c}} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ maps input signal $\tilde{a}$ to output signal $\tilde{y} = \tilde{\mathbf{P}}_{\tilde{c}}\tilde{a}$ such that, for all $n \in \mathbb{Z}$ with $\tilde{a}[n] \neq \tilde{c}[n]$,*

$$\tilde{y}[n] = \sum_{k=-\lfloor \frac{N-1}{2} \rfloor}^{\lfloor \frac{N}{2} \rfloor} \tilde{y}_k[n] e^{j\frac{2\pi}{N}kn}, \tag{4.19}$$

where

$$\begin{aligned} \tilde{y}_0[n] &= d_q[n], \\ \tilde{y}_k[n] &= \frac{\sin(\pi k d_q[n])}{N \sin\left(\frac{\pi k}{N}\right)} e^{-j\pi k(f_{1q}[n]+f_{0q}[n])}, k \neq 0, \end{aligned} \tag{4.20}$$

$$d_q[n] = \mathbf{Q}_{N,\tilde{c}}(d(\tilde{a}[n])), \quad f_{iq}[n] = \mathbf{Q}_N(f_i(c, \tilde{a}[n])), \ i \in \{0, 1\}.$$

When $\tilde{a}[n] = \tilde{c}[n]$ then $\tilde{y}[n] = \frac{1}{2}$.

Proof. See the section 4.7.3. □

If the reference signals are sawtooths, signal d_q satisfies $d_q[n] = \mathbf{Q}_{N,\tilde{c}}(\tilde{a}[n])$ for all n. In other words, the baseband component $\tilde{y}_0$ becomes equal to the quantized version of the

input $\tilde{a}$, that is $\tilde{y}_0[n] = \mathbf{Q}_{N,\tilde{c}}(\tilde{a}[n])$ for all n. The results from theorem 4.3.1 suggest that the baseband of the DPWM output signal mainly depends on the quantized version of the high-resolution baseband signal $\tilde{a}$, that is, not on $\tilde{a}$ directly. We call this effect a *hidden quantization*, to make distinction from the quantization (i.e., signal resolution reduction) which is achieved by the very operation of PWM. Moreover, in the rest of this chapter, we will refer to $\mathbf{Q}_{N,\tilde{c}}$ as the 'hidden quantizer'. Hidden quantization is the result of a displacement of switching instants in the DPWM output caused by the sampling of an APWM output signal. It follows from theorem 4.3.1 that spectral aliasing effects, which are caused by the sampling of an infinite bandwidth APWM output, can be described in time-domain as this additional (hidden) quantization of the input signal. Since hidden quantization is a time-domain manifestation of spectral aliasing, it is, therefore, an inherent characteristic of all digital PWM schemes.

4.3.2 Model Validation

In order to verify the closed-form expressions derived in the previous section, numerical simulations in MATLAB have been carried. In-phase component of a randomly generated 64QAM signal, upsampled and rescaled in amplitude to the $(0, 1)$ interval, has been used to drive the DPWM. Input signal bandwidth is set to $B = 20$MHz, and the PWM carrier frequency is $f_p = 0.5$GHz. Both the oversampling ratio N and the number M of reference signals of DPWM have been varied. Relative root mean square error between the DPWM output signals obtained by simulation and derived analytic formulas, was used as a measure of error. That is, if $\tilde{y}_{true}$ and $\tilde{y}$ denote the outputs generated by numerical simulation and analytical formulas, respectively, the error is given as

$$d(\tilde{y}_{true}, \tilde{y}) = \frac{\|\tilde{y}_{true} - \tilde{y}\|_2}{\|\tilde{y}_{true}\|_2} \cdot 100 \ [\%]. \tag{4.21}$$

The modeling error results, calculated for various combinations of N and M, are shown in Table 4.1. As can been seen, the error is negligible (and is non-zero only due to the finite precision of numerical calculations in MATLAB), which confirms validity of the derived expressions.

Table 4.1: Model validation error for various values of the oversampling ratio N and the number of carriers M (consequently, $M+1$ is the number of output levels). All values are in parts-per-trillion.

64QAM input signal							
N \ M+1	2	3	4	5	6	7	8
3	19.1	7.69	8.31	9.56	7.19	4.32	8.79
4	19.1	7.69	8.31	9.56	7.19	4.32	8.79
5	19.8	7.65	7.02	9.92	5.83	4.3	7.19
6	19.1	7.69	8.31	9.56	7.19	4.32	8.79
7	20.2	7.74	6.59	10.1	4.77	4.15	5.51
8	19.1	7.69	8.31	9.56	7.19	4.32	8.79
9	20.2	7.69	6.53	10.09	4.71	4.11	5.23
10	20.2	7.69	6.53	10.09	4.71	4.11	5.23
LTE input signal							
N \ M+1	2	3	4	5	6	7	8
3	38.22	15.36	16.81	19.11	14.53	8.6	17.76
4	38.22	15.36	16.81	19.11	14.53	8.6	17.76
5	38.75	15.29	13.9	19.37	10.58	8.55	13.07
6	38.22	15.36	16.81	19.11	14.53	8.6	17.76
7	40,11	15.41	13.21	20.06	9.74	8.21	11.08
8	38.22	15.36	16.81	19.11	14.53	8.6	17.76
9	40.63	15.49	13.46	20.31	9.64	8.26	10.52
10	40.63	15.49	13.46	20.31	9.64	8.26	10.52

4.4 Compensation of In-Band Noise in DPWM

4.4.1 Delta-Sigma Modulator Based Noise-Shaping

Intuitively, it is clear that increasing the oversampling ratio N and/or the number M of reference signals in DPWM, yields better linearity performance, due to either less aliasing effects (higher sampling rate) or better amplitude resolution of the output signal (more output levels in DPWM). Unfortunately, the upper limits on M and N are specified by hardware constraints and it is not possible to satisfactorily decrease the hidden quantization error while using feasible values for parameters M and N. A common conclusion in the literature is that in order to use DPWM for power encoding in all-digital transmitters (ADT), the oversampling ratio has to be orders of magnitude larger than the carrier frequency of PWM, in order to satisfy strict linearity constraints of digital communication standards [89]. Generally, this is true if signals taking arbitrary amplitude values in the interval (0,1) are to be fed into DPWM.

As noted in the previous section, the DPWM input-output model of theorem **??** suggests that the leading in-band harmonic distortion in the DPWM output comes from the hidden quantization $\mathbf{Q}_{N,\tilde{c}}$ that the input signal is subject to. It other words, the aliasing effects in the frequency-domain (due to time-sampling), can be equivalently explained as hidden quantization in the time-domain. This implies that in order to have no hidden quantization noise in the baseband component $\tilde{y}_0$ of the DPWM output (see (4.19)-(4.20)), the input signal has to be quantized (or pre-distorted) before being fed into DPWM, so that no additional distortion is introduced by the hidden quantizer $\mathbf{Q}_{N,\tilde{c}}$. In general, such pre-distortion of the true high resolution input signal would introduce its own quantization noise, with or without DPWM, therefore, diminishing possible gains since the total distortion (pre-distortion + hidden quantization) would still be significant.

It is possible to avoid the above problem by using a noise-shaping quantization system, e.g., delta-sigma modulator ($\Delta\Sigma$M), as a pre-distortion quantizer. By applying $\Delta\Sigma$M on the original high-resolution DPWM input signal it is possible to shape noise to out-of-band frequencies and control the level of in-band distortion. In order to have no in-band spectral regrowth once such pre-quantized signal is passed through the DPWM, it is necessary that

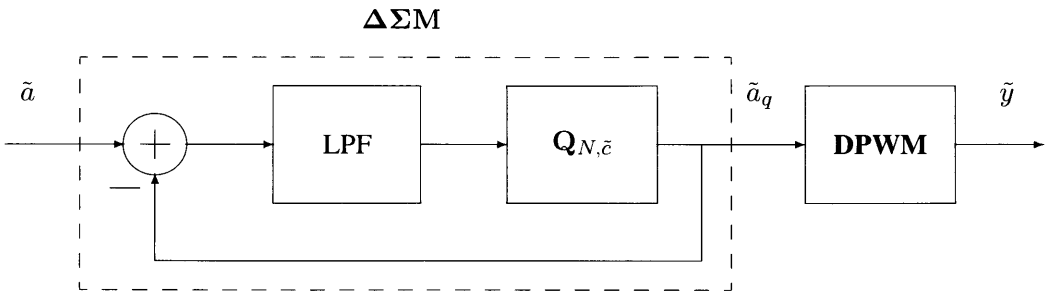


Figure 4-6: An abstract block diagram of a $\Delta\Sigma$M-DPWM power encoder.

parameters of the underlying quantizer in $\Delta\Sigma$M match those of the hidden quantizer $\mathbf{Q}_{N,\tilde{c}}$. That is, the output signal, denoted $\tilde{a}_q$, of the delta-sigma modulator driven by the input signal $\tilde{a}$, should assume the same amplitude levels as signal $\mathbf{Q}_{N,\tilde{c}}(\tilde{a})$. When driven by the signal $\tilde{a}_q$, the DPWM system would generate an output signal $\tilde{y}$ such that its baseband component $\tilde{y}_0$ satisfies $\tilde{y}_0[n] = \mathbf{Q}_{N,\tilde{c}}(\tilde{a}_q[n]) = \tilde{a}_q[n]$. Therefore, the total quantization noise in the baseband signal $\tilde{y}_0$ will correspond to that of the $\Delta\Sigma$M output $\tilde{a}_q$, which is shaped to out-of-band frequencies, ensuring acceptable levels of in-band noise in the DPWM output $\tilde{y}$. A block diagram of such a $\Delta\Sigma$M-DPWM power encoder is shown in Fig. 4-6. Indeed, if there is a mismatch, additional in-band distortion will be generated by DPWM, as confirmed by MATLAB simulation results shown in Figure 4-7. This plot depicts baseband spectrum of the output of a DPWM with $M = 2$ symmetric double-edge sawtooth reference signals and oversampling ratio $N = 8$. The DPWM input signal is pre-distorted by a $\Delta\Sigma$M with four different quantizer parameters: uniformly quantized with the number of quantization levels L equal to 100, 15, 8 and 3. Configuration with $L = 8$ matches the one of the hidden quantizer $\mathbf{Q}_{N,\tilde{c}}$, and gives minimal in-band SNR.

4.4.2 Optimal Co-Design of $\Delta\Sigma$M and DPWM

It is not hard to see, from the definition of $\mathbf{Q}_{N,\tilde{c}}$ given in (4.17), that in the case of sawtooth references and equidistant DPWM output levels, i.e., when $\alpha_m = m/M$ for all m, the hidden quantizer $\mathbf{Q}_{N,\tilde{c}}$ becomes a uniform quantizer. Clearly, DPWM power encoding schemes with uniform hidden quantization (i.e., with equidistant output levels) have

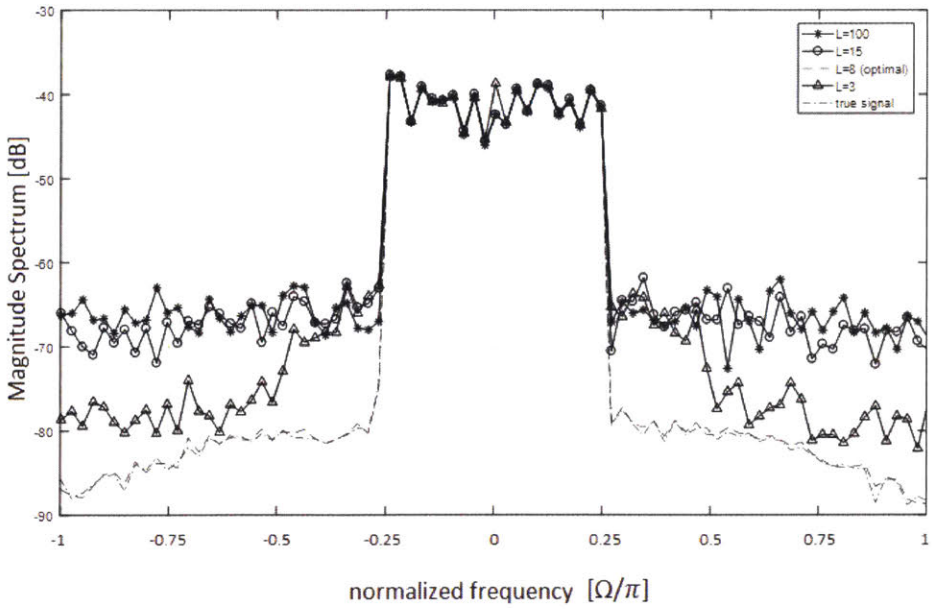


Figure 4-7: In-band magnitude spectra of the DPWM input signal (dash-dotted line), and DPWM output signals for a different number L of $\Delta\Sigma$M pre-distortion quantizer levels: $L = 100$ (asterisk), $L = 15$ (circle), $L = 8$ (dash), and $L = 3$ (triangle).

simplistic appeal from a hardware perspective. On the other hand, modern baseband communication signals (e.g., LTE) have amplitude distributions which are highly non-uniform. For example, the I/Q components of an OFDM signal have approximately Gaussian distribution [155]. It follows that power encoders with uniform quantization do not fully exploit the available information about the underlying communication signals and might lead to sub-optimal performance of the transmitters employing them. Namely, the number L of the hidden quantization levels of $\mathbf{Q}_{N,\tilde{c}}$ is commonly relatively small leading to uniform hidden quantizers of low resolution which cannot achieve satisfying out-of-band noise rejection for signals with highly non-uniform amplitude distribution. In the next section we show how $\Delta\Sigma$M-DPWM power encoder can be optimally designed to incorporate information about the input signal statistics in order to not just reduce the hidden quantization (i.e., in-band) noise but also minimize the out-of-band harmonic noise and, therefore, improve coding efficiency. In the following, for simplicity, we assume that the reference signals are symmetric double-edge sawtooth signals. It follows from (4.17) that hidden quantizer $\mathbf{Q}_{N,\tilde{c}}$ has $L = NM/2$ levels. The optimal power encoder design framework extends to arbitrary

references as well.

Let a be a real-valued scalar discrete-time signal whose amplitude samples are i.i.d. according to the probability density function $p_a = p_a(x)$, where $p_a : \mathbb{R} \to [0, \infty)$. The mean squared error of quantizing a by an arbitrary quantizer $\mathbf{Q} : \mathbb{R} \to \mathbb{R}$ is given by

$$J_p(\mathbf{Q}) = \int_{-\infty}^{\infty} (x - \mathbf{Q}(x))^2 \, p_a(x) dx. \tag{4.22}$$

In practice, the utilized baseband communication signals have finite dynamic range due to hardware or some other constraints (e.g., amplitude clipping due to PAPR reduction). For that reason, in the rest of this paper, we assume that $p_a(x) = 0$ for all $x \notin [a_0, a_M]$, for some fixed $a_0, a_M \in \mathbb{R}$.

For all $m \in \{1, \ldots, M\}$, let b_m^n and q_m^n be the decision boundaries and quantization levels, respectively, of the m-th sub-quantizer $\mathbf{Q}_{N,\tilde{c}_m}$. Dependence of b_m^n and q_m^n on the elements of $\mathcal{A} = \{\alpha_0, \ldots, \alpha_M\}$ is given by the following expressions

$$b_m^n = \alpha_{m-1} + \frac{2n}{N}(\alpha_m - \alpha_{m-1}), \quad 0 \le n \le N/2, \tag{4.23}$$

$$q_m^n = b_m^n, \quad 1 \le n \le N/2. \tag{4.24}$$

For $L = MN/2$, let $\mathcal{B} = \{b_0, \ldots, b_L\}$ and $\mathcal{Q} = \{q_1, \ldots, q_L\}$ be the decision boundaries and quantization levels of $\mathbf{Q}_{N,\tilde{c}_m}$, respectively. By the definition of $\mathbf{Q}_{N,\tilde{c}_m}$ we have

$$b_k = b_m^n, \text{ when } k = (m-1)\frac{N}{2} + n, \tag{4.25}$$

for all $n \in \{0, \ldots, N/2\}$ and $m \in \{1, \ldots, M\}$, and

$$q_k = b_k, \text{ for all } k \in \{1, \ldots, L\}. \tag{4.26}$$

The mean-squared error cost (4.22) for $\mathbf{Q} = \mathbf{Q}_{N,\tilde{c}_m}$ is, therefore, a function of $\mathcal{A}$ and can be written as

$$J_p(\mathbf{Q}_{N,\tilde{c}_m}) = \sum_{m=1}^{M} \sum_{n=1}^{N/2} \int_{b_m^{n-1}}^{b_m^n} (x - q_m^n)^2 p_a(x) dx. \tag{4.27}$$

Assume now that DPWM is driven by the above described signal a with amplitude distribution p_a. It follows that the out-of-band harmonic noise in the DPWM output can be decreased by choosing the output levels α_m such that the hidden quantizer $\mathbf{Q}_{N,\tilde{c}}$ minimizes the above defined mean squared quantization error (MSQE). Unfortunately, the structure of $\mathbf{Q}_{N,\tilde{c}}$ is significantly restricted: decision boundaries b_k are piece-wise uniformly distributed and quantization levels q_k are at the boundary of each individual decision interval $(b_{k-1}, b_k]$. This implies that, in general, one can expect very poor MSQE performance of such a quantizer (even in the case of optimal MSQE!). This problem can be mitigated in the following way.

Let $\tilde{\mathbf{Q}} : [\alpha_0, \alpha_M] \to \mathbb{R}$ be a quantizer whose decision boundaries $\tilde{b}_k$, for $k \in \{0, \ldots, L\}$, and quantization levels $\tilde{q}_k$, for $k \in \{1, \ldots, L\}$, satisfy the following

$$\tilde{b}_k = b_k \qquad \tilde{q}_k \in (\tilde{b}_{k-1}, \tilde{b}_k]. \tag{4.28}$$

Clearly, quantizers $\tilde{\mathbf{Q}}$ and $\mathbf{Q}_{N,\tilde{c}}$ have identical decision boundaries (for a fixed choice of $\mathcal{A}$), while the quantization levels of $\tilde{\mathbf{Q}}$ are unrestricted unlike those of $\mathbf{Q}_{N,\tilde{c}}$. Now we want to find $\tilde{b}_k$ and $\tilde{q}_k$ such that the quantizer $\tilde{\mathbf{Q}}$ minimizes mean squared error (4.22). This problem can be formulated as follows

$$\begin{aligned} &\min_{\alpha_1,\ldots,\alpha_M,\tilde{q}_1,\ldots,\tilde{q}_L} J_p(\tilde{\mathbf{Q}}) \\ \text{s.t.} \quad & a_0 = \alpha_0 < \alpha_1 < \cdots < \alpha_M = a_M, \\ & b_{(m-1)N/2+n} = \alpha_{m-1} + \frac{2n}{N}(\alpha_m - \alpha_{m-1}), \\ & \forall m \in \{1, \ldots, M\}, \quad \forall n \in \{0, \ldots, N/2\}, \\ & \tilde{q}_k \in (b_{k-1}, b_k], \quad \forall k \in \{1, \ldots, L\}. \end{aligned} \tag{4.29}$$

Let $\tilde{\mathbf{Q}}^*$ be the argument of minimum of (4.29) (more precisely, let $\tilde{\mathbf{Q}}^*$ be a quantizer whose parameters are the arguments of minimum of (4.29)).

Remark 1: It should be noted that, though similar, in general, the optimal solution $\tilde{\mathbf{Q}}^*$ of

the above optimization problem is not a Lloyd-Max quantizer [156], since the decision boundaries of $\tilde{\mathbf{Q}}^*$ are fixed to be uniformly distributed on sub-intervals $(\alpha_m, \alpha_{m+1}]$.

Remark 2: It is easy to see that, in general, $J_p = J_p(\tilde{\mathbf{Q}})$ is a non-convex function of $\alpha_0, \alpha_1, \ldots, \alpha_M, \tilde{q}_1, \ldots, \tilde{q}_L$, and the optimal problem (4.29) cannot be solved explicitly. Furthermore, there is no guarantee that a global optimum would be achieved by applying any of the standard non-convex optimization algorithms. In practice, one calculates off-line the parameters of $\tilde{\mathbf{Q}}^*$ and then programs digital hardware that the power encoder should be implemented on. For that reason, we find an approximate optimal solution of (4.29) by performing a grid search. The number M of output levels of DPWM is commonly low (for high power amplifiers to be used in base stations it is typically not larger than 5 [105]), and performing a grid search to find an approximate optimal solution of (4.29) commonly imposes just a mild computational burden.

The optimal quantization parameters, as defined above, cannot, in general, match the quantization parameters of $\mathbf{Q}_{N,\tilde{c}}$ for any DPWM. That is, quantizer $\tilde{\mathbf{Q}}^*$ is, in general, not equivalent to $\mathbf{Q}_{N,\tilde{c}}$. Hence, if the optimal quantizer $\tilde{\mathbf{Q}}^*$ was used in the $\Delta\Sigma$M pre-quantizer system of a $\Delta\Sigma$M-DPWM encoder, then amplitude values of the $\Delta\Sigma$M output would not match those of the hidden quantization. This implies that the DPWM system would then generate significant in-band harmonic noise in the output and any benefit of the $\Delta\Sigma$M pre-quantization would be lost. This problem can be mitigated by introducing a *compensation signal* that should be added to the DPWM output to compensate for the difference between the quantization levels of $\tilde{\mathbf{Q}}^*$ and $\mathbf{Q}_{N,\tilde{c}}$. A block diagram of such an optimal power encoder is shown in Figure 4-8. The $\Delta\Sigma$M subsystem utilizes the optimal quantizer $\tilde{\mathbf{Q}}^*$ and maps baseband input signal a to a quantized signal a_q. Since $\tilde{\mathbf{Q}}^* \neq \mathbf{Q}_{N,\tilde{c}}$, the output $\tilde{y}$ of the DPWM subsystem will be equal to $\tilde{y} = a_q + e + harmonics$, where e is the error signal $e = \mathbf{Q}_{N,\tilde{c}}(a_q) - a_q$. Compensator $\mathbf{C}$ takes input signal a and generates the error signal e, as defined above. More precisely, the compensator system $\mathbf{C}$, mapping signal a into signal e is defined as follows:

$$e[n] = q_k - \tilde{q}_k^*, \text{when } a[n] \in (\tilde{b}_{k-1}^*, \tilde{b}_k^*] \equiv (b_{k-1}, b_k], \tag{4.30}$$

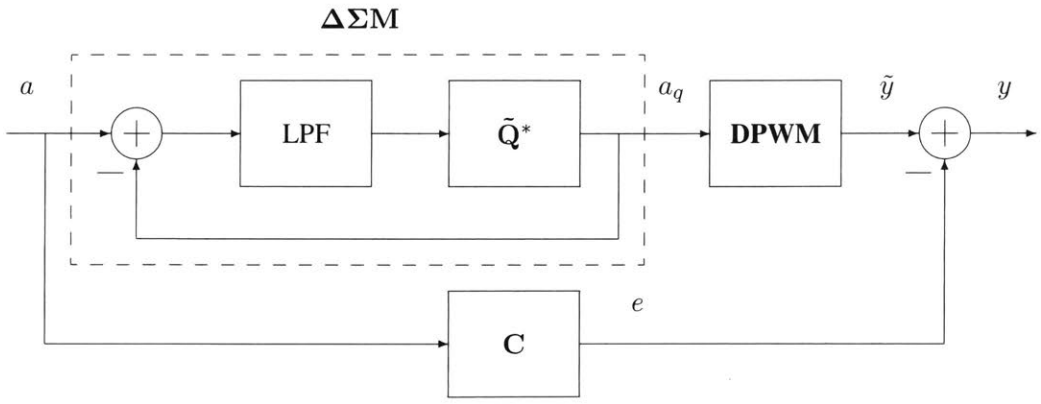


Figure 4-8: Block diagram of the proposed optimal power encoder.

for all $k \in \{1, \ldots, L\}$ and all $n \in \mathbb{Z}$. The compensator $\mathbf{C}$ is such that

$$e[n] = (\mathbf{C}a)[n] = \mathbf{Q}_{N,\tilde{c}}(a_q[n]) - a_q[n].$$

In other words, signal e should cancel the hidden quantization error that is caused by the difference in the quantization level values of $\tilde{\mathbf{Q}}^*$ and $\mathbf{Q}_{N,\tilde{c}}$.

Remark: It should be noted that the compensation signal takes at most L amplitude values and has much smaller power than the main signal. Therefore, its power added cost to the overall system design is minimal since a low-power linear PA can be used to amplify this signal.

4.5 Performance Evaluation

4.5.1 Experimental Results

In order to evaluate performance of a $\Delta\Sigma$M-DPWM power coding scheme, measurements have been carried at the Mitsubishi Electric Research Labs, Cambridge, MA. Performance is measured in terms of achieved coding efficiency and in-band signal-to-noise ratio (SNR), and proposed $\Delta\Sigma$M-DPWM power encoder is compared to the one employing only DPWM, while keeping parameters of DPWM fixed in both cases.

Envelope of a 64QAM signal with 20MHz bandwidth was used as the input into power

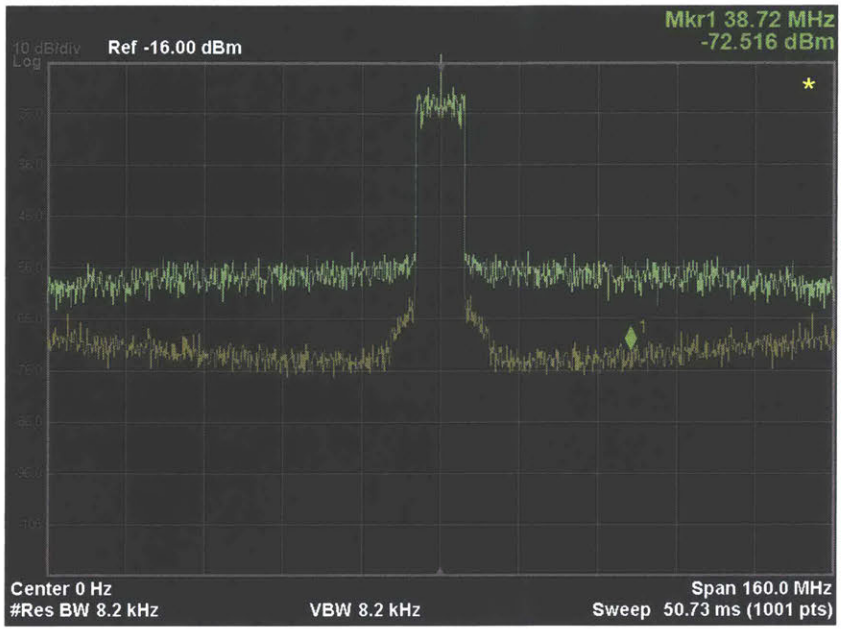


Figure 4-9: Measured baseband spectrum of the DPWM (green) and $\Delta\Sigma$M-DPWM (yellow) outputs.

encoders. DPWM reference signals are symmetric double-edge sawtooth signals. PWM carrier frequency is set to 1GHz, and the oversampling ratio is $N = 8$, thus giving the sampling frequency of 8GS/s. Number of output levels of DPWM is set to 5, which implies the number of hidden quantization levels equal to $L = 8 \cdot (5 - 1)/2 = 16$. We therefore employ $\Delta\Sigma$M with 16 output levels. The order of $\Delta\Sigma$M is 1, so to simplify the power encoder design.

MATLAB simulated DPWM output signal file is loaded into an arbitrary waveform generator (AWG-34G from Micram Instruments) to generate the signals. Keysight EXA 9010A signal analyzer is used to measure the spectrum. Fig. 4-9 compares the measured spectrum for DPWM (green) and $\Delta\Sigma$M-DPWM (yellow) output signals. Measured coding efficiencies are 75.95% and 72.7% respectively. It is expected that DPWM alone can achieve better coding efficiency than hybrid $\Delta\Sigma$M-DPWM, due to noise shaping property of $\Delta\Sigma$M as shown in Fig. 4-10, but the effective number of quantization levels is large, and thus these coding efficiencies differ by insignificant amount (3%). Meanwhile, it can be seen that the latter one offers considerably improved in-band SNR of -54.8dB, compared to the in-band SNR of -33.4dB for DPWM.

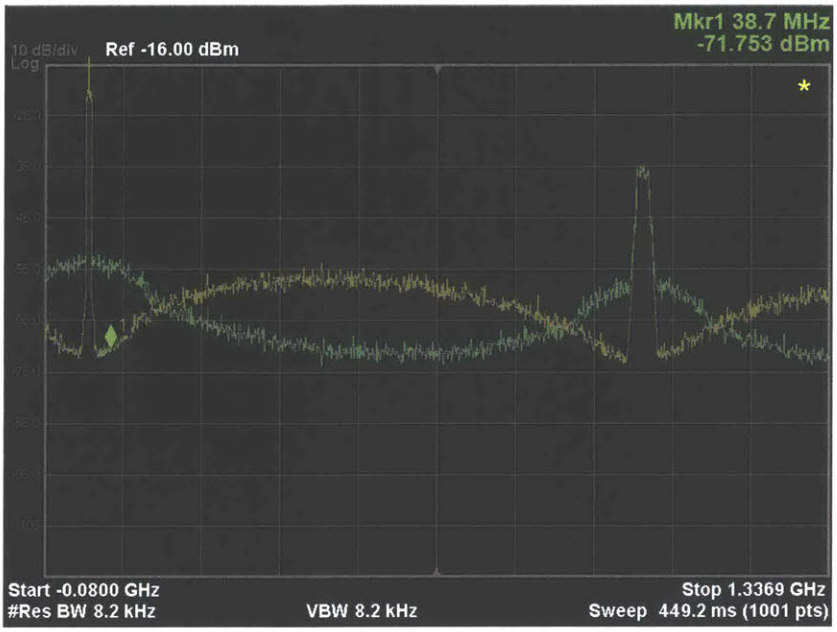


Figure 4-10: Measured wideband spectrum of the DPWM (green) and $\Delta\Sigma$M-DPWM (yellow) outputs.

Table 4.2: Simulation parameters and test signals

	Test Signal	Band-width	Modulation	PAPR	f_{samp}	N	M	L
Case 1	E-TM 3.1	20 MHz (100 RB)	64QAM	11.1 dB	6.14 GS/s	4	3	6
Case 2	E-TM 3.1a	20 MHz (100 RB)	256QAM	11 dB	9.21 GS/s	6	4	12

4.5.2 Simulation Results

Performance of the proposed optimal $\Delta\Sigma$M-DPWM power encoding scheme was evaluated by MATLAB simulations, where CE and in-band SNR of the encoder output were again used as measures of performance. The proposed encoder was compared to the non-optimized $\Delta\Sigma$M-DPWM power encoder (i.e, the one with uniform quantization in $\Delta\Sigma$M, as in the previous section) and the non-compensated DPWM encoder (i.e., without $\Delta\Sigma$M pre-distortion of the input).

Simulations were performed for two types of input signals, both E-UTRA test models as specified in [157]. The simulation parameters, as well as the test signals' parameters,

Table 4.3: Performance Comparison for E-TM 3.1 Test Signal

	DPWM	$\Delta\Sigma$M-DPWM	Optimized Model
CE	30.33%	29.94%	45.26%
SNR	19.62 dB	43.2 dB	46.32 dB

for each case, are presented in Table 4.2. As can be seen, the DPWM parameters N and M take relatively small values and, as a consequence, the number L of the hidden quantization levels is also relatively small (see Table 4.2). In both cases, test signals in FDD mode were used, and the order of the $\Delta\Sigma$M was again set to 1, so to simplify the power encoder design.

The signal flow of the simulation is as follows: the I and Q components of the input LTE signal, generated through MATLAB's LTE System Toolbox, are fed into the above described power encoders, and their outputs combined to get the complex baseband output signal. For simplicity, the input signals are normalized so that their amplitude values fully span the DPWM dynamic range which was set to $(-1, 1)$ (i.e., $a_0 = -1$ and $a_M = 1$). For the optimal $\Delta\Sigma$M-DPWM encoder, the input signal amplitude pdf parameters are first estimated and then used to calculate the optimal decision boundaries and quantization levels (both of these tasks are done off-line).

Performance results of the tested encoding methods are reported in Tables 4.3 and 4.4. The wideband and in-band output spectra, for each power encoding method, are depicted in Figs. 4-11, 4-13 and 4-12, 4-14, respectively. As can be seen, the in-band harmonic noise of both $\Delta\Sigma$M-DPWM encoders is significantly lower than that of the regular DPWM encoder, which was to be expected from [125]. It should be noted that the optimal encoder is slightly better than the uniform one in terms of SNR (by 1-3 dB). On the other hand, in terms of coding efficiency, the proposed optimal power encoder significantly outperforms (by 15%-20%) the other two methods. This can also be inferred from the wideband output spectra plots in Figs. 4-11 and 4-13.

Table 4.4: Performance Comparison for E-TM 3.1a Test Signal

	DPWM	ΔΣM-DPWM	Optimized Model
CE	46.47%	46.1%	66.13%
SNR	28.63 dB	42.73 dB	43.15 dB

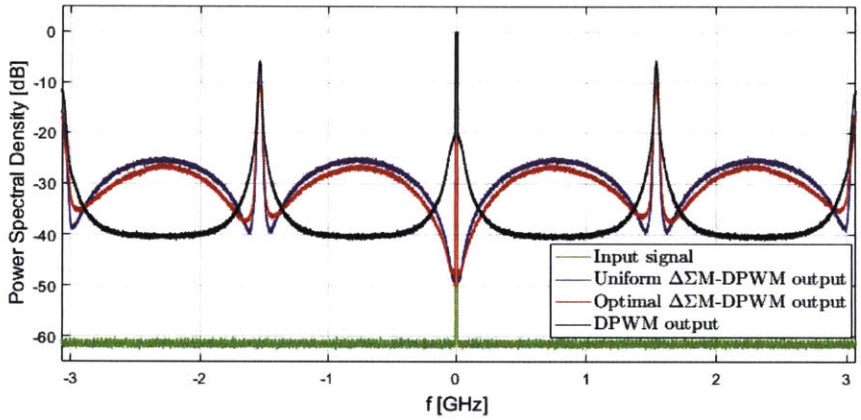


Figure 4-11: Wideband output spectra for E-TM3.1 test signal.

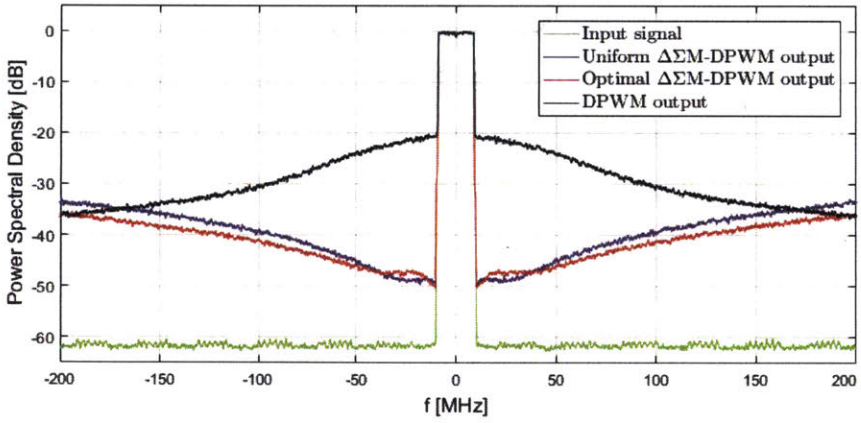


Figure 4-12: Baseband (zoom-in) output spectra for E-TM3.1 test signal.

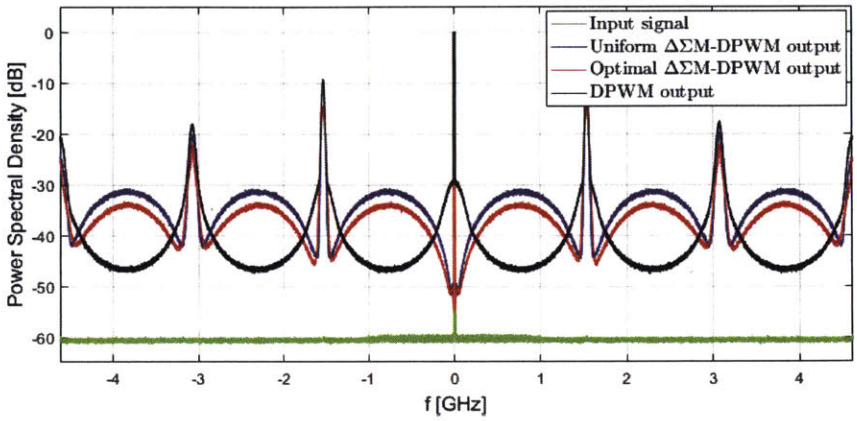


Figure 4-13: Wideband output spectra for E-TM3.1a test signal.

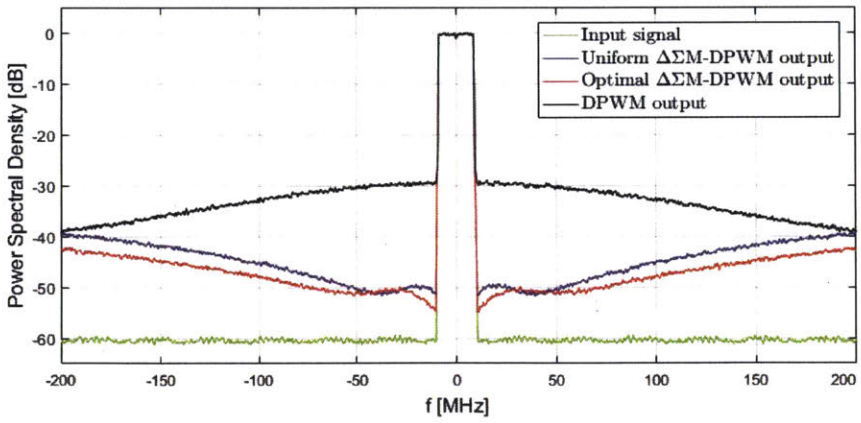


Figure 4-14: Baseband (zoom-in) output spectra for E-TM3.1a test signal.

4.6 Summary

In this chapter, we studied digital pulse-width modulation systems with arbitrary reference or carrier signals, and arbitrary square summable excitation. We derived a novel input-output time-domain model of DPWM which fully characterizes the nonlinear behavior of this system. We further proposed a modified Lloyd-Max quantization based algorithm for linearization of the baseband of a DPWM output. Simulation results show that the proposed model can achieve significantly better coding efficiency at the expense of negligible increase in in-band signal to noise ratio in comparison to the DPWM system with uniform hidden quantization.

4.7 Proofs for Chapter 4

4.7.1 Proof of Theorem 4.2.2

Let $c = c(t)$ be an arbitrary admissible reference signal with period equal to T_p, for some $T_p > 0$ (as shown in Fig. 4-15). E.g., in the case of a sawtooth reference, $c(t) = c_\delta(t+\rho T_p)$, for some $\delta, \rho \in (0,1)$. As before, let $\mathbf{P}_c : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ denote a APWM operator with reference c, which maps $a \in \mathcal{L}^2(\mathbb{R})$ to $y \in \mathcal{L}^2(\mathbb{R})$ such that

$$y(t) = (\mathbf{P}_c a)(t) = \begin{cases} 0, & a(t) < c(t), \\ 1, & a(t) \geq c(t). \end{cases}$$

Let $\mathcal{P} : \mathbb{R} \times [0,1] \to \{0,1\}$ be a comparator map such that

$$\mathcal{P}(\xi, \mu) = \begin{cases} 0, & \xi < \mu, \\ 1, & \xi \geq \mu, \end{cases} \quad \text{for all } \xi \in \mathbb{R}, \mu \in [0,1]. \tag{4.31}$$

For a fixed $a \in \mathcal{L}^2(\mathbb{R})$ let $v = v(t_1, t_2) = \mathcal{P}(a(t_1), c(t_2))$. Function v is periodic in t_2, with period T_p, since c is periodic with period T_p. Therefore, for any fixed value t_1^* of t_1 (i.e., for a fixed value $a(t_1^*)$ of $a(t_1)$), signal $v(t_1, t_2)$ can be expanded into its Fourier series in

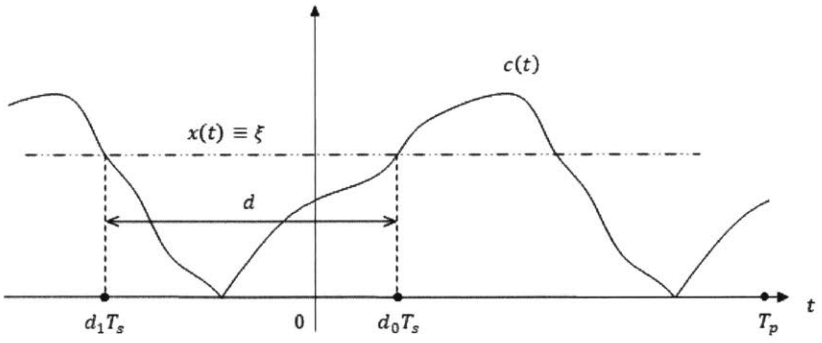


Figure 4-15: An example showing geometric interpretation of f_0, f_1 and d, for an arbitrary admissible reference signal.

t_2 as follows

$$v(t_1^*, t_2) = \sum_{k=-\infty}^{\infty} C_k(a(t_1^*))e^{jk\omega_p t_2}, \tag{4.32}$$

where coefficients $C_k(a(t_1^*))$ are defined as

$$C_k(a(t_1^*)) = \frac{1}{T_p}\int_0^{T_p} \mathcal{P}(a(t_1^*), c(\tau))e^{-jk\omega_p \tau} d\tau.$$

The convergence of the series on the right-hand side of (4.32) should be understood in the sense of its principal value, that is the infinite sum is equal to $\lim_{K\to\infty} \sum_{k=-K}^{K} C_k(a(t_1))e^{jk\omega_p t_2}$. Equality in (4.32) holds point-wise, and is true for every $t_2 \in \mathbb{R}$ for which $c(t_2) \neq a(t_1^*)$. At the points of discontinuity of v, that is, for t_2 such that $c(t_2) = a(t_1^*)$, we have that

$$v(t_1^*, t_2) = \frac{1}{2}(v(t_1^*, t_{2-}) + v(t_1^*, t_{2+})) = \frac{1}{2},$$

where $v(t_1^*, t_{2-})$ and $v(t_1^*, t_{2+})$ are the directional limits of $v(t_1^*, \cdot)$ at t_2. It follows from (4.31) that $v(t_1^*, t_2)$ is a periodic squared wave for which (see, e.g., [110]) the following Fourier series expansion holds:

$$v(t_1^*, t_2) = \sum_{k=-\infty}^{\infty} \frac{\sin(k\pi d)}{k\pi} e^{-jk\pi D} e^{jk\omega_p t_2}, \tag{4.33}$$

where (see Figure 4-15)

$$d = f_1(c,\xi) - f_0(c,\xi), \quad D = f_1(c,\xi) + f_0(c,\xi). \tag{4.34}$$

Again, convergence of the series and equality in (4.33) should be understood as described above. Indeed, $C_0(\xi) = f_1(c,\xi) - f_0(c,\xi) = d$, and for $k \neq 0$:

$$C_k(\xi) = \frac{1}{T_p}\int_{T_p f_0}^{T_p f_1} e^{-jk\omega_p \tau} d\tau = \frac{e^{-j2\pi k f_0} - e^{-j2\pi k f_1}}{j2\pi k} = \frac{\sin(k\pi d)}{k\pi} e^{-jk\pi D},$$

where we used the shorthand notation $f_1 = f_1(c,\xi)$ and $f_0 = f_0(c,\xi)$.

Since (4.33) holds point-wise whenever $c(t_2) \neq a(t_1^*)$, by setting $t_1 = t_2 = t$ we get that the APWM output signal y at every time t such that $c(t) \neq a(t)$, is given by the formula

$$y(t) = v(t,t) = \sum_{k=-\infty}^{\infty} \frac{\sin(k\pi d(t))}{k\pi} e^{-jk\pi D(t)} e^{jk\omega_c t}, \tag{4.35}$$

where $d(t)$ and $D(t)$ are defined by

$$d(t) = f_1(c,a(t)) - f_0(c,a(t)), \quad D(t) = f_1(c,a(t)) + f_0(c,a(t)), \quad \forall t \in \mathbb{R}. \tag{4.36}$$

This concludes the proof of Theorem 4.2.2.

4.7.2 Proof of Theorem 4.2.3

Let $\alpha, \beta \in \mathbb{R}$ such that $\alpha < \beta$, and let $\mathcal{P}$ be the comparator map as defined in (4.31). It is easy to see that for any $\xi, \mu \in \mathbb{R}$ the following expression holds:

$$\mathcal{P}(\xi, (\beta-\alpha)\mu + \alpha) = \mathcal{P}\left(\frac{\xi-\alpha}{\beta-\alpha}, \mu\right), \tag{4.37}$$

Hence, for any admissible reference signal c and any $a \in \mathcal{L}^2(\mathbb{R})$, it follows that $\mathbf{P}_{\tilde{c}} a = \mathbf{P}_c \tilde{a}$, where $\tilde{c}(t) = (\beta-\alpha)c(t) + \alpha$ and $\tilde{a}(t) = \frac{a(t)-\alpha}{\beta-\alpha}$, for all $t \in \mathbb{R}$.

For the given α_m and c'_m let signals $a_m \in \mathcal{L}^2(\mathbb{R})$ and $y_m \in \mathcal{L}^2(\mathbb{R})$ be defined by $a_m(t) = \frac{a(t)-\alpha_{m-1}}{\alpha_m - \alpha_{m-1}}$, for all $t \in \mathbb{R}$ and $y_m = \mathbf{P}_{c'_m} a_m$, for all $m \in \{1, \ldots, M\}$. From (4.37)

and the definition of $(M+1)$-level APWM system, given in (4.4), it follows that the output signal $y = \mathbf{P}_{c_1,\ldots,c_M} a$ of APWM can be expressed as

$$\begin{aligned} y(t) &= \sum_{m=1}^{M} (\alpha_m - \alpha_{m-1}) \mathcal{P}(a(t), c_m(t)) = \\ &= \sum_{m=1}^{M} (\alpha_m - \alpha_{m-1}) \mathcal{P}\left(a_m(t), c'_m(t)\right), \\ &= \sum_{m=1}^{M} (\alpha_m - \alpha_{m-1}) y_m(t), \end{aligned} \tag{4.38}$$

Let signals $d_m = d_m(t)$ and $D_m = D_M(t)$, for all $m \in \{1, \ldots, M\}$, be defined by

$$d_m(t) = f_1(c'_m, a_m(t)) - f_0(c'_m, a_m(t)), \tag{4.39}$$

$$D_m(t) = f_1(c'_m, a_m(t)) + f_0(c'_m, a_m(t)). \tag{4.40}$$

Remark: Its is clear from Appendix 4.7.1 that in the case of sawtooth carriers we have

$$d_m(t) = \begin{cases} 0, & a(t) < \alpha_{m-1} \\ \frac{a(t)-\alpha_{m-1}}{\alpha_m - \alpha_{m-1}}, & \alpha_{m-1} \leq a(t) < \alpha_m \\ 1, & a(t) \geq \alpha_m \end{cases} \tag{4.41}$$

$$D_m(t) = (2\rho_m - 1) d_m(t) - 2\tau_m, \tag{4.42}$$

for all $m \in \{1, \ldots, M\}$.

It now immediately follows from (4.38) and Theorem 4.2.2 that

$$y(t) = \sum_{k=-\infty}^{\infty} \sum_{m=1}^{M} \frac{\alpha_m - \alpha_{m-1}}{\pi k} \sin(\pi k d_m(t)) e^{-j\pi k D_m(t)} e^{jk\omega_c t},$$

where $d_m(t)$ and $D_m(t)$ are as defined in (4.39) and (4.40). When $a(t) \in [\alpha_{m-1}, \alpha_m)$ then

$y_i(t) = 1$ for all $i < m$, and $y_i(t) = 0$ for all $i > m$. Hence,

$$\begin{aligned} y(t) &= \alpha_m + \sum_{k=-\infty}^{\infty} \frac{\alpha_m - \alpha_{m-1}}{\pi k} \sin(\pi k d_m(t)) e^{-j\pi k D_m(t)} e^{jk\omega_c t}, \\ &= a(t) + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{\alpha_m - \alpha_{m-1}}{\pi k} \sin(\pi k d_m(t)) e^{-j\pi k D_m(t)} e^{jk\omega_c t}. \end{aligned} \tag{4.43}$$

By recalling definitions of $d = d(t)$ and $D = D(t)$ from the statement of Theorem **??**, the expression for $y(t)$ in (4.5) now immediately follows from (4.43).

This concludes the proof of Theorem 4.7.2.

Remark: By noting that $a(t) = (\alpha_m - \alpha_{m-1})d(t) + \alpha_{m-1}$ when $a(t) \in [\alpha_{m-1}, \alpha_m)$ for all $m \in \{1, \ldots, M\}$, the expression for $y(t)$ can be rewritten as

$$y(t) = \alpha(a(t)) + \sum_{k=-\infty}^{\infty} C_k(a(t)) e^{jk\omega_c t}, \tag{4.44}$$

where $\alpha : \mathbb{R} \to \{\alpha_0, \ldots, \alpha_{M-1}\}$ is defined by

$$\alpha(\xi) = \begin{cases} 0, & \xi < \alpha_1, \\ \alpha_{m-1}, & \xi \in [\alpha_{m-1}, \alpha_m), \forall m \in \{2, \ldots, M\}. \\ 1, & \xi \geq \alpha_M. \end{cases}$$

Expression (4.44) will be used in the proof of Theorem **??**.

4.7.3 Proof of Theorem 4.3.1

Let $T \in \mathbb{R}$ be an arbitrary positive real number, and let $T_p = NT$. Let c be a T_p-periodic signal such that $c(nT) = \tilde{c}[n]$, and $c(t) = \tilde{c}[n] + \frac{\tilde{c}[n+1] - \tilde{c}[n]}{T}(t - nT)$ when $t \in (nT, nT + T)$, for all $n \in \mathbb{Z}$. Let $a \in \mathcal{L}^2(\mathbb{R})$ be such that $a(nT) = \tilde{a}[n]$, and let $y = \mathbf{P}_c a$, where $\mathbf{P}_c$ is the APWM system with reference signal c. From theorem 4.2.2 and the definition of DPWM,

it follows that signal $\tilde{y}$ can be expressed as

$$\tilde{y}[n] = y(nT) = \sum_{k=-\infty}^{\infty} C_k(a(nT))e^{jk\omega_p nT} = \sum_{k=-\infty}^{\infty} \frac{\sin(\pi k\tilde{d}[n])}{\pi k} e^{-j\pi k\tilde{D}[n]} e^{j\frac{2\pi}{N}kn}, \quad (4.45)$$

where $\tilde{d}, \tilde{D} \in \ell^2(\mathbb{R})$ such that $\tilde{d}[n] = d(nT)$, $\tilde{D}[n] = D(nT)$. The number of possibly different harmonics in (4.45) is finite, and equal to N [110]. Therefore, (4.45) can be rewritten as

$$\tilde{y}[n] = \tilde{y}_0[n] + \sum_{\substack{k=-\lfloor\frac{N-1}{2}\rfloor \\ k\neq 0}}^{\lfloor\frac{N}{2}\rfloor} \tilde{y}_k[n] e^{j\frac{2\pi}{N}kn}. \quad (4.46)$$

where

$$\tilde{y}_0[n] = \sum_{m=-\infty}^{\infty} \frac{\sin(\pi mN\tilde{d}[n])}{\pi mN} e^{-j\pi mN\tilde{D}[n]}, \quad (4.47a)$$

$$\tilde{y}_k[n] = \sum_{m=-\infty}^{\infty} \frac{\sin(\pi(mN+k)\tilde{d}[n])}{\pi(mN+k)} e^{-j\pi(mN+k)\tilde{D}[n]}. \quad (4.47b)$$

Here $\tilde{y}_0[n]$ denotes the baseband component of $\tilde{y}[n]$, and $\tilde{y}_k[n]$ are the higher order harmonics, similar to the description of the APWM output, given in (4.11). In the following analysis, we assume that N is odd, in order to avoid unnecessary repetition of cumbersome expressions. Derivation for the case when N is even, differs only in additional multiplication factor of 1/2 in the expression for $\tilde{x}_{N/2}$.

It can be observed from (4.47a) and (4.47b) that expressions for harmonics of $\tilde{y}$ involve only infinite sums of discrete sinc functions. These sums can be computed analytically, details of which are given in the following Lemma.

Lemma 4.7.1. *Let $a \in (0,1)$, $b \in \mathbb{R}$, $\tau \in \mathbb{R}$, and $T > 0$. Then the following equality holds*

$$\sum_{n=-\infty}^{\infty} \frac{\sin(\pi(Tn+\tau)a)}{\pi(Tn+\tau)} e^{j\pi(Tn+\tau)b} = \frac{\sin(\pi\tau\frac{q-p}{T})}{T\sin(\frac{\pi\tau}{T})} e^{j\pi\tau\frac{p+q+1}{T}}, \quad (4.48)$$

when $b-a \in \left(\frac{2p}{T}, \frac{2p+2}{T}\right]$, $b+a \in \left[\frac{2q}{T}, \frac{2q+2}{T}\right)$, $\forall p, q \in \mathbb{Z}$.

Proof. Let us denote the infinite sum in (4.48) as $S(T, \tau, a, b)$. We will prove formula

(4.48) by using Fourier transform to compute the DC component of a certain discrete-time signal obtained by sampling a carefully chosen continuous-time signal.

For arbitrary, but fixed, $a \in (0,1)$, $t_0 \in \mathbb{R}$, and $\omega_0 \in \mathbb{R}$, let continuous-time signals $f = f(t)$, $g = g(t)$ and $h = h(t)$ be defined by

$$f(t) = \begin{cases} \frac{\sin(\pi a t)}{\pi t}, & t \neq 0 \\ a, & t = 0 \end{cases}, \tag{4.49}$$

$$g(t) = f(t) \cdot e^{j\omega_0 t}, \qquad h(t) = g(t - t_0). \tag{4.50}$$

The Fourier transforms $F = F(\omega)$, $G = G(\omega)$ and $H = H(\omega)$ of $f(t)$, $g(t)$ and $h(t)$, respectively, are then given by [110]

$$F(\omega) = \begin{cases} 1, & |\omega| < \pi a \\ 1/2, & |\omega| = \pi a \\ 0, & \text{otherwise} \end{cases}. \tag{4.51}$$

$$G(\omega) = F(\omega - \omega_0), \qquad H(\omega) = F(\omega - \omega_0)e^{-j\omega t_0}. \tag{4.52}$$

For $T > 0$, let a discrete-time signal $h_T = h_T[n]$ be defined by

$$h_T[n] = h(nT) = \frac{\sin(\pi(nT - t_0)a)}{\pi(nT - t_0)} e^{j\omega_0(nT - t_0)}, \quad \forall n \in \mathbb{Z}. \tag{4.53}$$

The sum $S(T, \tau, a, b)$ can now be rewritten in terms of h_T as

$$S(T, \tau, a, b) = \sum_{n=-\infty}^{\infty} h_T[n] \Big|_{\substack{t_0 = -\tau \\ \omega_0 = \pi b}} = H_T(0)|_{\substack{t_0 = -\tau \\ \omega_0 = \pi b}}, \tag{4.54}$$

where $H_T = H_T(\Omega)$ denotes the Fourier transform of h_T, and is defined by

$$H_T(\Omega) = \frac{1}{T} \sum_{m=-\infty}^{\infty} H\left(\frac{\Omega}{T} + \frac{2\pi m}{T}\right). \tag{4.55}$$

Therefore, computing $S(T,\tau,a,b)$ amounts to finding the value of $H_T(\Omega)$ at $\Omega = 0$, for $t_0 = -\tau$ and $\omega_0 = \pi b$. From (4.52), (4.54) and (4.55) it follows that

$$S(T,\tau,a,b) = \frac{1}{T}\sum_{m=-\infty}^{\infty} F\left(\frac{2\pi m}{T} - \pi b\right) e^{j\frac{2\pi\tau m}{T}}. \tag{4.56}$$

For fixed values of parameters T, a and b, the number of summands in (4.56) is finite, since

$$F\left(\frac{2\pi m}{T} - \pi b\right) > 0 \quad \text{for} \quad \frac{b-a}{2} \leq \frac{m}{T} \leq \frac{b+a}{2}.$$

Let $p, q \in \mathbb{Z}$ such that $\frac{b-a}{2} \in \left[\frac{p}{T}, \frac{p+1}{T}\right)$ and $\frac{b+a}{2} \in \left[\frac{q}{T}, \frac{q+1}{T}\right)$. Now (4.56) simplifies to

$$S(T,\tau,a,b) = \frac{1}{T}\sum_{m=p+1}^{q} e^{j\frac{2\pi\tau m}{T}}. \tag{4.57}$$

After some simple algebraic manipulation we get

$$S(T,\tau,a,b) = \frac{\sin(\pi\tau\frac{q-p}{T})}{T\sin(\frac{\pi\tau}{T})} e^{j\pi\tau\frac{p+q+1}{T}}, \tag{4.58}$$

which concludes the proof of Lemma 4.7.1. □

Remark: It should be noted that the right-hand side of (4.48) can be rewritten as

$$\frac{\sin\left(\pi\tau\left(Q\left(\frac{b+a}{2}\right) - Q\left(\frac{b-a}{2}\right)\right)\right)}{T\sin(\frac{\pi\tau}{T})} e^{j\pi\tau\left(Q\left(\frac{b+a}{2}\right)+Q\left(\frac{b-a}{2}\right)\right)}, \tag{4.59}$$

where $Q : \mathbb{R} \to \mathbb{R}$ is a uniform quantizer such that $Q(\xi) = \frac{2p+1}{2T}$ when $\xi \in \left[\frac{p}{T}, \frac{p+1}{T}\right)$, for all $p \in \mathbb{Z}$.

The statement of the theorem 4.3.1 now follows from the above lemma and (4.47a) and (4.47b).

Chapter 5

Conclusions and Future Directions

Conclusions

In this thesis, we studied several optimal digital signal processing problems, involving different classes of nonlinear dynamical systems, to improve the power efficiency of modern wireless transmission systems.

In chapter 2, we considered the problem of designing discrete-time systems for peak-to-average-power ratio reduction of communication signals. The problem was formulated as the minimization of a frequency-weighted convex quadratic cost subject to time-domain output amplitude constraints. For such problems, in general, the optimality conditions do not provide an explicit way of generating the optimal output as a real-time implementable transformation of the input, and showed that the optimal system has exponentially fading memory. We proposed two algorithms, based on time and value iterations of carefully chosen stable finite-latency nonlinear systems, which return approximations of arbitrary precision to the optimal map. In one case, the result holds under an L1 dominance assumption regarding parameters of the cost function. In the other case, we presumed no special assumptions on cost function parameters. We obtained approximate system by a careful truncation of an infinite dimensional state space representation of the optimal system, where an upper bound on the error of approximation was derived by extending the method of balanced truncation for linear systems to a certain class of nonlinear models. We used numerical simulations in MATLAB to verify effectiveness of the proposed PAPR reduction

method. Numerical results show that a significant (about 50%) reduction in PAPR can be achieved for certain standard-compliant LTE communication signals.

In chapter 3, we analyzed effects of the nonlinear distortion introduced into the baseband (discrete-time) input-output dynamics of the communication systems by the (continuous-time) power amplifier, on the structure of the equivalent baseband model of the communication system. We showed that when the nonlinearity is represented by a Volterra series model the resulting baseband equivalent model is a series interconnection of a discrete-time Volterra series model - of the same degree and equivalent memory depth - and a linear system. It has been shown that the order of nonlinearity and, more importantly, memory of the underlying nonlinear system are preserved when passing from passband to baseband. This results in a novel equivalent baseband model, which is a series interconnection of a fixed degree/short memory Volterra model and a long memory LTI system. The result suggests a new, non-obvious, analytically motivated structure of digital pre-compensation of passband nonlinear distortions caused by power amplifiers. We showed that the memory, and therefore complexity, of an approximate baseband model and its corresponding digital pre-distorter increases as the ratio of the baseband signal bandwidth to the observed bandwidth increases. This suggests that the proposed model is best utilized under the assumption of a full input signal bandwidth (that is, no oversampling of the baseband signal). Unlike conventional digital pre-distorter (DPD) implementations which have long memory specifications in the corresponding nonlinear subsystem, the DPD model proposed in this thesis exploits the underlying system structure to model long memory requirements in terms of high order FIR filters, which are relatively simple for digital implementation. Numerical simulation results show that the proposed model outperforms various general Volterra series based models in its ability to approximate the true communication system. Furthermore, the results show that a DPD modeled according to the proposed structure outperforms DPDs modeled using general Volterra series models, in tested performance metrics (ACLR and EVM).

In chapter 4, we analyzed and designed digitally implemented pulse-width modulators (DPWM) used as quantizers for power amplifiers in switched-mode operation. We derived a novel closed-form input-output time-domain model of digital PWM systems, for both

carrier-based and threshold based PWM schemes. We derived model under assumptions of square summable excitation and arbitrary admissible DPWM reference signals. This result suggests that the in-band distortion in the DPWM output signal, which has been traditionally explained by spectral aliasing, can be understood as a consequence of an additional, hidden, quantization that the DPWM input is subject to (with primary quantization being the operation of PWM itself). The presented analysis significantly improves understanding of the leading in-band distortion source in digitally implemented PWM, and will perhaps enable analytical performance evaluation of all-digital transmitters employing it.

We proposed framework for designing power encoding systems, realized as a series interconnection of a delta-sigma modulator and a digital pulse-width modulator. Under assumptions that the DPWM input signal has time-samples i.i.d. according to some known distribution, we provided framework for designing a modified Lloyd-Max quantization based algorithm for linearization of the baseband of a DPWM output with a carefully designed delta-sigma modulator. We chose parameters of the $\Delta\Sigma$M-DPWM encoder so to minimize the mean squared error of the hidden quantization noise in DPWM. Results of the MATLAB simulations and experiments with an arbitrary signal generator show the effectiveness of the proposed system in encoding several standardized LTE test signals.

Future Directions

Regarding the work presented in chapter 2, there are several directions for future work. In our PAPR reduction problem, the abstract optimization is carried over a relatively simple convex set K: an infinite-dimensional hypercube. In this case, the projection operator P_K is the sat_r operator and can be evaluated component-wise, which simplifies the optimality conditions and, correspondingly, the design of an approximate algorithm. The most natural extension of the abstract optimization problem considered in this chapter is to optimization over convex sets K with more complicated description: e.g., $K = \{x \in \ell^2(\mathbb{C}) : |x[n]|^2 + |x[n-1]|^2 \leq 1, \forall n \in \mathbb{Z}\}$. The next step would then be to extend the abstract problem formulation (and the corresponding analysis techniques) to model predictive control (MPC) problems (or, in the simplest case, to the infinite-dimensional linear quadratic regulator problem). This is a promising direction that we briefly looked at but have not explored in

detail.

There are also several directions for future work related to the analysis and synthesis of DPWM systems presented in chapter 4. Given the derived input-output model that describes nonlinear behavior of a DPWM system, it would be interesting to find analytical bounds on in-band and out-of-band harmonic distortion. It would also be interesting to test the optimal formulation for delta-sigma and pulse-width modulation co-design for DPWM schemes with non-sawtooth reference signals. It seems that under this assumption, there would be no need for the compensation signal to cancel the mismatch between the modified Max-Loyd quantizer and the true hidden quantizer. Another future direction could be to try to extend the optimal delta-sigma and pulse-width modulation co-design to radio-frequency DPWM.

Another aspect of the problems analyzed in the thesis that should not be forgotten, is the analysis of hardware complexity as well as implementation of the proposed signal processing algorithms. For example, in the PAPR reduction problem, it seems that the presence of feedback in the proposed system models presents the biggest challenge for their hardware implementation, especially when the latency parameter m is large and one operates on baseband signals with high sampling rates. We believe that any efficient implementation should probably rely on some special-purpose hardware, but more research should be done to reach a conclusion.

Appendix A

Digital Radio-Frequency PWM

A.0.1 Background and Problem Formulation

The method of operation of analog radio-frequency PWM (RF-APWM) is described in the first subsection. In this case, RF-PWM is understood as a CT system, i.e., it maps CT signals to CT signals. In the second subsection, an input-output model of RF-APWM is presented. Digital RF-PWM system (RF-DPWM) is defined in the third subsection.

A. Principle of Operation of RF-APWM

The main idea of radio-frequency PWM is to modulate an RF signal into a stream of unipolar or bipolar pulses, where the envelope and phase are represented by the width and timing of the pulses [92]. In 3-level RF-PWM schemes [158], a baseband input signal is first up-converted to RF frequency, and the modulated carrier is then compared to a fixed threshold value to produce the output pulse train [80], as shown in Fig. A-1. Let $a \in \mathcal{L}^2(\mathbb{R})$ be the baseband input signal and let $x = x(t) = a(t)\cos(2\pi f_c t)$ be the modulated RF carrier, where $f_c > 0$ is the RF or carrier frequency. Let $\gamma > 0$. The analog radio-frequency PWM system $\mathbf{P} : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$, mapping input signal a into output signal $y = \mathbf{P}a$, is defined

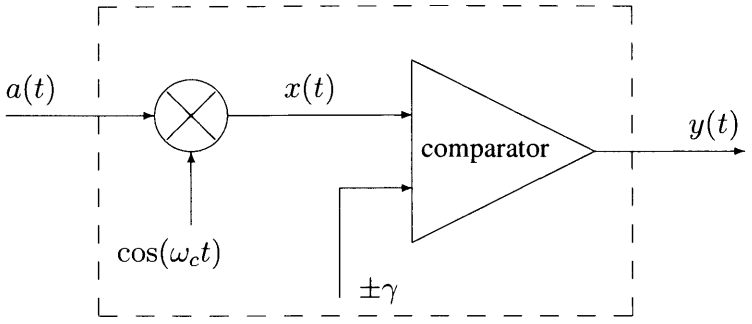


Figure A-1: Principle block diagram of CT RF-PWM describing the output signal generation.

by

$$y(t) = \begin{cases} 1, & x(t) > \gamma \\ 0, & |x(t)| \leq \gamma \\ -1, & x(t) < \gamma \end{cases}, \tag{A.1}$$

where γ is called the RF-APWM threshold. In practical applications, the threshold parameter γ is commonly chosen such that $\gamma \in (0, \sup \|a\|_\infty)$, where the supremum is taken over the 'expected' class of input signals.

In the above expression, non-zero values of the output signal $y(t)$ are normalized to 1 and -1, but can, in general, be set to any desired values. An example of output signal generation is shown in Fig. A-2, for an arbitrary choice of the baseband input signal $a(t)$. As can be seen from Fig. A-2, signal $y(t)$ is a sequence of alternating pulses, where the width of each pulse depends on the value of the envelope $a(t)$ of the RF carrier signal $x(t)$.

In general, an RF-APWM scheme with $2M + 1$ levels (for an arbitrary integer $M \geq 1$) is defined in a similar way, where instead of one threshold, there are M different threshold values. Let $\Gamma = \{\gamma_1, \ldots, \gamma_M\} \subset \mathbb{R}$ such that

$$\gamma_i > 0, \text{ and } \gamma_i < \gamma_j, \forall i, j \in \{1, \ldots, M\} \text{ such that } i < j. \tag{A.2}$$

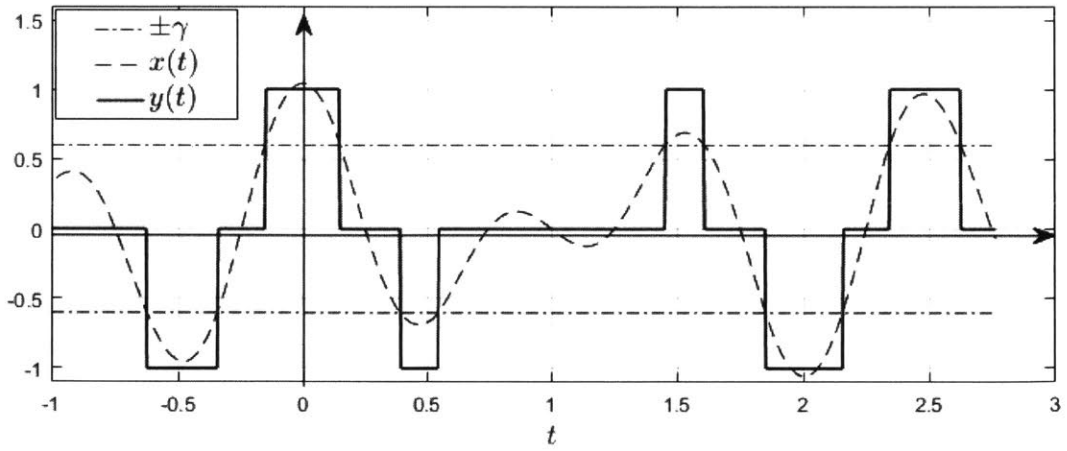


Figure A-2: An example of output signal generation in a 3-level RF-APWM.

Let $\mathcal{A} = \{\alpha_0, \alpha_1, \ldots, \alpha_M\} \subset \mathbb{R}$ such that

$$0 = \alpha_0 < \alpha_1 < \alpha_2 < \cdots < \alpha_{M-1} < \alpha_M = 1. \tag{A.3}$$

The $(2M+1)$-level RF-APWM is defined as a system $\mathbf{P}_{\Gamma,\mathcal{A}} : \mathcal{L}^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$, mapping input signal a into output signal $y = \mathbf{P}a$ such that

$$y(t) = \begin{cases} 0, & |x(t)| \le \gamma_1 \\ \alpha_m \cdot \operatorname{sign}(x(t)), & \gamma_m < |x(t)| \le \gamma_{m+1}, \quad \forall m \in \{1, \ldots, M-1\}, \\ 1, & |x(t)| > \gamma_M \end{cases} \tag{A.4}$$

where $x(t) = a(t)\cos(2\pi f_c t)$ for all $t \in \mathbb{R}$.

Let $\mathbf{P}_{\gamma_m}$ be the 3-level RF-APMW system with threshold γ_m, and let $y_m = \mathbf{P}_{\gamma_m} a$, for all $m \in \{1, \ldots, M\}$. It is easy to see from (A.4), that signal y can be defined as $y(t) = \sum_{i=1}^{M} (\alpha_m - \alpha_{m-1}) \cdot y_m(t)$. That is, the RF signal $x(t)$ is compared to each threshold γ_m to generate the corresponding signal $y_m(t)$, and the RF-APWM output $y(t)$ is formed by taking a linear combination of all $y_m's$.

Similar to the case of DPWM, the simple to understand definitions of RF-APMW given in this section are not valuable for analysis of spectral properties of the RF-APWM output signal. As before, a more useful input-output model is given in the next section.

B. Time-Domain Analysis of RF-APWM

In this section, for completness, we present the RF-APWM model that can be found in, e.g., [92].

Let $\gamma > 0$, and let function $g_\gamma : \mathbb{R} \to \left(-\frac{1}{2}, \frac{1}{2}\right)$ be defined by

$$g_\gamma(\xi) = \begin{cases} -\frac{1}{\pi} \arccos \frac{\gamma}{|\xi|}, & \xi < -\gamma \\ 0, & |\xi| \leq \gamma \\ \frac{1}{\pi} \arccos \frac{\gamma}{|\xi|}, & \xi > \gamma \end{cases} . \tag{A.5}$$

The following theorem gives an input-output model of a 3-level RF-APWM, which depicts harmonic nature of the output signal and is more convenient for spectral analysis of RF-APWM signals than the model given in the previous section.

Theorem A.0.1. *3-level RF-APWM system* $\mathbf{P}_\gamma$ *maps* $a \in \mathcal{L}^2(\mathbb{R})$ *to* $y = \mathbf{P}_\gamma a \in \mathcal{L}^2(\mathbb{R})$ *such that, for all* $t \in \mathbb{R}$ *with* $|a(t)\cos(2\pi f_c t)| \neq \gamma$,

$$y(t) = \sum_{k=1}^{\infty} C_k(a(t)) \cos(2\pi(2k-1) f_c t), \tag{A.6}$$

where

$$C_k(a(t)) = \frac{4 \sin(\pi(2k-1)d(t))}{\pi(2k-1)}, \tag{A.7}$$

with

$$d(t) = g_\gamma(a(t)). \tag{A.8}$$

For $t \in \mathbb{R}$ *such that* $|a(t)\cos(2\pi f_c t)| = \gamma$, *we have that* $y(t) = \frac{1}{2}\operatorname{sign}(a(t)\cos(2\pi f_c t))$.

Proof. See the section A.0.3. □

Signal $d = d(t)$, from theorem A.0.1, is commonly called the RF-PWM width or (time-varying) duty-cycle signal, since for a constant baseband input signal $a(t) = const, \forall t$, signal d would be identically equal to a half of the width of the pulse in the periodic RF-APWM output signal $y(t)$.

Let M be an arbitrary positive integer, and let $\Gamma = \{\gamma_1, \ldots, \gamma_M\}$ and $\mathcal{A} = \{\alpha_0, \alpha_1, \ldots, \alpha_M\}$ be as defined in (A.2)-(A.3). The following theorem gives an input-output model for the $(2M+1)$-level RF-APWM.

Theorem A.0.2. *$(2M+1)$-level RF-APWM system $\mathbf{P}_{\Gamma,\mathcal{A}}$ maps $a \in \mathcal{L}^2(\mathbb{R})$ to $y = \mathbf{P}_{\Gamma,\mathcal{A}} a \in \mathcal{L}^2(\mathbb{R})$ such that, for all $t \in \mathbb{R}$ with $|a(t)\cos(2\pi f_c t)| \neq \gamma_m$ for all $m \in \{1, \ldots, M\}$,*

$$y(t) = \sum_{k=1}^{\infty} C_k(a(t)) \cos(2\pi(2k-1) f_c t), \tag{A.9}$$

where

$$C_k(a(t)) = \frac{4}{\pi(2k-1)} \sum_{m=1}^{M} (\alpha_m - \alpha_{m-1}) \sin(\pi(2k-1) d_m(t)), \tag{A.10}$$

with

$$d_m(t) = g_{\gamma_m}(a(t)). \tag{A.11}$$

For $t \in \mathbb{R}$ such that $|a(t)\cos(2\pi f_c t)| = \gamma_m$, for all $m \in \{1, \ldots, M\}$, we have that $y(t) = \frac{1}{2}(\alpha_m + \alpha_{m-1}) \cdot \operatorname{sign}(a(t)\cos(2\pi f_c t))$.

Proof. The proof of theorem A.0.2 immediately follows from the definition of multi-level RF-APWM and the results of theorem A.0.1. □

It is clear from (A.6) (equivalently, (A.9)) that the RF-APWM output $y(t)$ is a quasi-harmonic signal, with the fundamental frequency equal to the RF carrier f_c. Therefore, the first harmonic of y can be understood as an amplitude modulated carrier, bearing input signal information, that is, as a conventional wireless communication signal. Moreover, since $a(t)$ is modulating a cosine, due to symmetry, only odd harmonics are present. This is very advantageous, since spectral regrowth coming from higher order harmonics is not going to affect much the RF signal band (i.e. frequency band around f_c). This lowers the specification constraints for the analog band-pass filter that is used to filter the signal after amplification by the switch-mode power amplifier (SMPA). It follows from the previous theorems, that envelope y_1 of the first harmonic of signal y can be written as

$$y_1(t) = const \cdot \sin(\pi d(t)). \tag{A.12}$$

Let function $f_\gamma : \mathbb{R} \to (-1, 1)$ be defined by $f(\xi) = \sin(g_\gamma(\xi))$ for all $\xi \in \mathbb{R}$. It is not hard to show, that f_γ is a monotonic function (but not strictly monotonic) and can be simplified to

$$f_\gamma(\xi) = \begin{cases} \text{sign}(\xi) \cdot \sqrt{1 - \left(\frac{\gamma}{\xi}\right)^2}, & |\xi| \geq \gamma \\ 0, & |\xi| < \gamma \end{cases}. \tag{A.13}$$

Static nonlinear function f_γ is the nonlinear characteristic from the input baseband signal a to the envelope y_1 and is commonly called the AM-AM characteristic. By appropriately pre-distorting baseband signal $a(t)$, it is possible to (partially) invert the nonlinear AM-AM characteristic and achieve highly linear mode of operation of the RF-APWM [158]. Indeed, let function $\phi_\gamma : (-1, 1) \to \mathbb{R}$ be defined by

$$\phi_\gamma(\xi) = \begin{cases} \text{sign}(\xi) \cdot \frac{\gamma}{\sqrt{1-\xi^2}}, & |\xi| \geq \gamma \\ 0, & |\xi| < \gamma \end{cases}, \tag{A.14}$$

By taking series interconnection of the static nonlinearity $\phi_\gamma(\cdot)$ and the system RF-APWM, the envelope y_1 of the first harmonic of the output signal y becomes equal to

$$y_1(t) = const \cdot \begin{cases} a(t), & |a(t)| \geq \gamma \\ 0, & |a(t)| < \gamma \end{cases}. \tag{A.15}$$

The values of $a(t)$ from the interval $(-\gamma, \gamma)$ get mapped into 0 by the $a(t)$-to-$d(t)$ conversion, and this accounts for most of the in-band distortion in RF-APWM. The other major contributor of spectral noise are higher order harmonics, as defined in (A.6). Similar to the case of carrier-based APWM, if the ratio f_c/B_w is sufficiently high, were B_w is the input signal bandwidth, this noise becomes negligible. This is depicted in Fig. A-3, for a 3-level RF-APWM with $f_c = 80$MHz and driven by an input signal with $B_w = 10$MHz (this implies $f_c/B_w = 8$). The upper plot shows magnitude spectrum of the baseband input signal $a(t)$, and the lower plot shows magnitude spectrum of the pre-distorted RF-APWM output $y(t)$. As can be seen, the in-band spectral noise is indeed negligible.

Similar reasoning applies to the multi-level RF-APWM, with an exception in that the

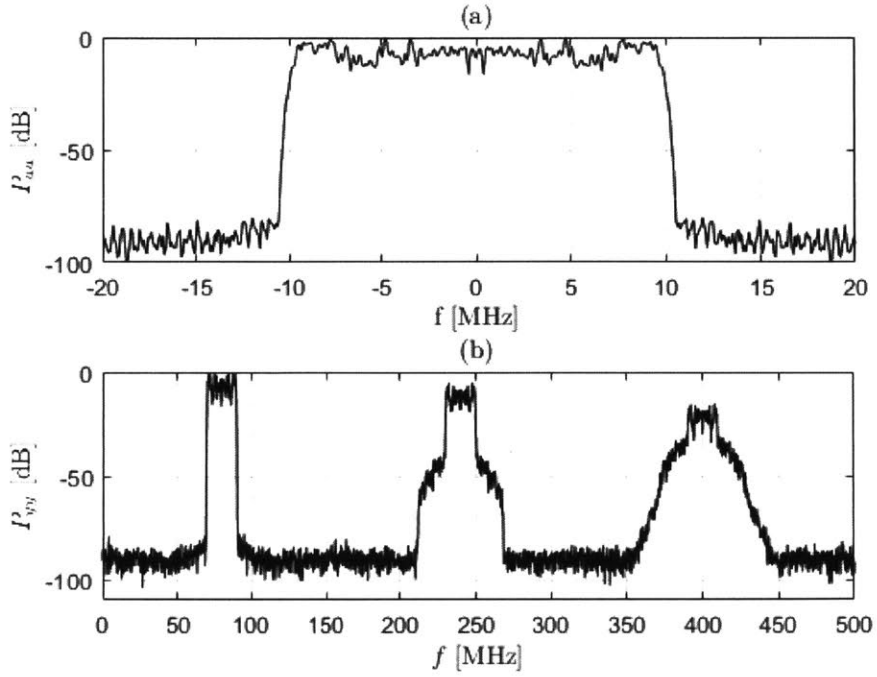


Figure A-3: An example of input and output spectra of a 3-level analog RF-PWM with $f_c = 80$MHz: (a) PSD of a 20MHz input signal (b) PSD of the pre-distorted RF-PWM output: only odd harmonics are present with negligible in-band noise level.

AM-AM characteristic is more complicated and there does not exist a simplified explicit formula like that given in (A.13). Indeed, the AM-AM characteristic of a $(2M+1)$-level RF-APWM is a function $f_\Gamma : \mathbb{R} \to (-1, 1)$ defined by

$$f_\Gamma(\xi) = \sum_{m=1}^{M} \sin(\pi g_{\gamma_m}(\xi)). \tag{A.16}$$

C. Principle of Operation of RF-DPWM

The operation of RF-DPWM can be defined in a way similar to what was described in the previous section for RF-APWM.

For given $N, M \in \mathbb{Z}, N > 1, M > 0$, let $\Gamma = \{\gamma_1, \ldots, \gamma_M\} \subset \mathbb{R}$ and $\mathcal{A} = \{\alpha_0, \ldots, \alpha_M\} \subset \mathbb{R}$ satisfy (A.2)-(A.3). The $(2M+1)$-level RF-DPWM is defined as a system $\tilde{\mathbf{P}}_{\Gamma,\mathcal{A}} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$, mapping input signal $\tilde{a} \in \ell^2(\mathbb{R})$ into output signal $\tilde{y} = \tilde{\mathbf{P}}_{\Gamma,\mathcal{A}}\tilde{a} \in \ell^2(\mathbb{R})$

such that

$$\tilde{y}[n] = \begin{cases} 0, & |\tilde{x}[n]| \leq \gamma_1 \\ \alpha_m \cdot \text{sign}(\tilde{x}[n]), & \gamma_m < |\tilde{x}[n]| \leq \gamma_{m+1}, \quad \forall m \in \{1, \ldots, M-1\} \\ 1, & |\tilde{x}[n]| > \gamma_M \end{cases}, \tag{A.17}$$

where $\tilde{x}[n] = \tilde{a}[n] \cos\left(\frac{2\pi n}{N}\right)$ for all $n \in \mathbb{Z}$. Parameter N is called the oversampling ratio of RF-DPWM.

For a fixed $T > 0$, let $\mathbf{R}_T : \ell^2(\mathbb{R}) \to \mathcal{L}^2(\mathbb{R})$ and $\mathbf{S}_T : \mathcal{L}^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be the perfect recovery and sampling systems, as defined in (4.15) and (4.16), respectively. Let $f_c = N/T$, and let $\mathbf{P}_{\Gamma,\mathcal{A}}$ be the RF-APWM with carrier frequency f_c. It now follows that, for any $T > 0$, the following equality is true

$$\tilde{\mathbf{P}}_{\Gamma,\mathcal{A}} = \mathbf{S}_T \mathbf{P}_{\Gamma,\mathcal{A}} \mathbf{R}_T. \tag{A.18}$$

In other words, RF-DPWM system is equivalent to a series interconnection of the perfect recovery system, RF-APWM, and sampler, as depicted in Figure **??**. Therefore, digital RF-PWM can be seen as a sampled version of its analog counterpart if the carrier frequency f_c is a kN-multiple of the sampling frequency, for any positive integer k. In this case, the continuous-time carrier frequency f_c corresponds to the digital carrier frequency $1/N$.

It was shown in the previous section that the RF-APWM output has infinite spectrum. Therefore, when this signal is sampled, regardless of how large the sampling frequency f_s is (or equivalently how large N is), aliasing will occur in the spectrum of the RF-DPWM output signal [110]. This aliasing is the main source of in-band distortion in digital RF-PWM, and its main drawback in practical applications. This is depicted in Fig. A-4 for two values of the oversampling ratio: $N = 5$ and $N = 50$. In the case of low OSR ($N = 5$), even though a pre-distortion has been used to mitigate the static AM-AM nonlinearity, the amount of remaining in-band spectral noise is so significant that the signal is completely buried in noise. Clearly, when a sufficiently high OSR ($N = 50$) is used, the level of noise is

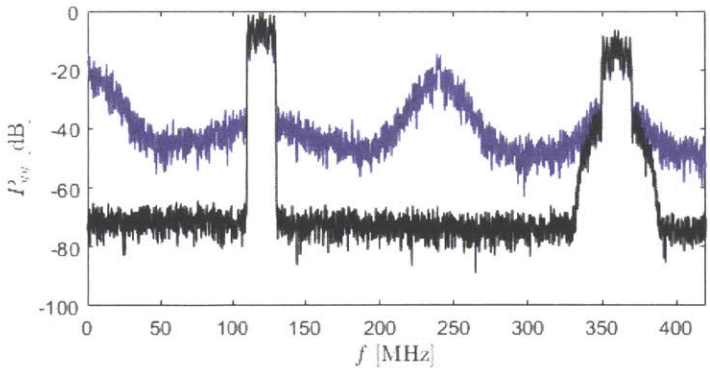


Figure A-4: An example of output spectra of a 3-level RF-DPWM with $f_c = 120$MHz and a 20MS/sec input signal. PSD of a pre-distorted RF-DPWM output for $N = 50$ (black) and $N = 7$ (blue): significant amount of noise is present in the case of low OSR.

relatively low. It therefore follows that one way of mitigating aliasing noise is by increasing the oversampling ratio N. From a hardware point of view, this is not advantageous since OSR values needed to guarantee sufficient linearity for wireless applications (i.e. f_c is on the order of GHz) become either too expensive or infeasible with current technology.

As in the case of carrier-based DPWM, the aliasing noise represents the main drawback of digital implementation of radio-frequency PWM. It is hence clear that better understanding of the aliasing noise might lead to the more efficient ways of reducing it and, consequently, making RF-DPWM feasible for practical purposes. In the next section, we present a novel input-output model for RF-DPWM, similar to the one for DPWM shown in section 4.3.

A.0.2 Time-Domain Analysis of Radio-Frequency DPWM

In this subsection we derive a novel analytical model of a multilevel RF-DPWM system. This result suggests an alternative, time-domain, description for aliasing noise. Moreover, the result suggests a natural way for mitigating the noise, which does not rely on significant increase in the oversampling ratio of RF-DPWM.

A. 3-Level RF-DPWM

Before we state the main result, let us introduce some additional notation.

Let $N \in \mathbb{Z}, N > 1$ and let $\gamma > 0$. Let $\mathbf{Q}_o : \mathbb{R} \to (-\frac{1}{2}, \frac{1}{2})$ and $\mathbf{Q}_e : \mathbb{R} \to (-\frac{1}{2}, \frac{1}{2})$ be quantizers with $L_o = N + 2$ and $L_e = 2\lfloor \frac{N+2}{4} \rfloor + 1$ output levels, respectively, defined by

$$\mathbf{Q}_o(\xi) = \begin{cases} \frac{2i+1}{2N}, & \left|\xi - \frac{2i+1}{2N}\right| < \frac{1}{2N}, \lfloor -\frac{N}{2} \rfloor \leq i \leq \lfloor \frac{2N-1}{4} \rfloor, \\ 0, & \xi = 0. \end{cases} \tag{A.19}$$

$$\mathbf{Q}_e(\xi) = \begin{cases} \frac{2i+1}{N}, & \left|\xi - \frac{2i+1}{N}\right| < \frac{1}{N}, \lfloor -\frac{N}{4} \rfloor \leq i \leq \lfloor \frac{N-1}{4} \rfloor, \\ 0, & \xi = 0. \end{cases} \tag{A.20}$$

It is easy to see that $\mathbf{Q}_o$ and $\mathbf{Q}_e$ are 'almost' uniform quantizers in the sense that if there was no fixed point at 0 they would be uniform. Let $\mathbf{Q}_{dc}$ be a modulo quantizer [161], with 3 output levels, defined by

$$\mathbf{Q}_{dc}(\xi) = \begin{cases} \frac{(-1)^i}{N}, & \left|\xi - \frac{2i+1}{2N}\right| < \frac{1}{2N}, \lfloor -\frac{N}{2} \rfloor \leq i \leq \lfloor \frac{2N-1}{4} \rfloor, \\ 0, & \xi = 0. \end{cases} \tag{A.21}$$

Clearly, quantizer $\mathbf{Q}_{dc}$ has a fixed point at 0 as well.

The input-output model of a 3-level RF-DPWM system is given in the following theorem.

Theorem A.0.3. *3-level RF-DPWM system $\tilde{\mathbf{P}}_\gamma : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ maps input signal $\tilde{a} \in \ell^2(\mathbb{R})$ into output signal $\tilde{y} = \tilde{\mathbf{P}}_\gamma \tilde{a} \in \ell^2(\mathbb{R})$ such that, for all $n \in \mathbb{Z}$ with $\left|\tilde{a}[n] \cos\left(\frac{2\pi n}{N}\right)\right| \neq \gamma$, the following expressions hold*

(1) if $\mod(N, 2) = 1$ *and* $N_0 = \frac{N-1}{2}$, *then*

$$\tilde{y}[n] = \tilde{y}_0[n] + \sum_{k=1}^{N_0} \tilde{y}_k[n] \cos\left(\frac{2\pi k n}{N}\right), \tag{A.22}$$

where $\tilde{d}[n] = g_\gamma(\tilde{a}[n])$ *and*

$$\begin{aligned} \tilde{y}_0[n] &= \mathbf{Q}_{dc}(\tilde{d}[n]), \\ \tilde{y}_k[n] &= \frac{2 \sin\left(\pi k \mathbf{Q}_o(\tilde{d}[n])\right)}{N \sin\left(\frac{\pi k}{2N}\right)} \quad \forall k \in \{1, \ldots, N_0\}, \end{aligned} \tag{A.23}$$

(2) if $\mod(N,4)=0$ *then*

$$\tilde{y}[n]=\sum_{k=1}^{N_0}\tilde{y}_k[n]\cos\left(\frac{2\pi(2k-1)n}{N}\right), \tag{A.24}$$

where $\tilde{d}[n]=g_\gamma(\tilde{a}[n])$ *and*

(i) if $\mod(N,4)=0$ *then* $N_0=\frac{N}{4}$ *and*

$$\tilde{y}_k[n]=\frac{4\sin\left(\pi(2k-1)\mathbf{Q}_e(\tilde{d}[n])\right)}{N\sin\left(\frac{\pi(2k-1)}{N}\right)},\quad \forall k\in\{1,\ldots,N_0\} \tag{A.25}$$

(ii) if $\mod(N,4)=2$ *then* $N_0=\frac{N+2}{4}$ *and*

$$\begin{aligned}\tilde{y}_k[n]&=\frac{4\sin\left(\pi(2k-1)\mathbf{Q}_e(\tilde{d}[n])\right)}{N\sin\left(\frac{\pi(2k-1)}{N}\right)},\quad \forall k\in\{1,\ldots,N_0-1\},\\ \tilde{y}_{\frac{N+2}{4}}[n]&=\frac{2}{N}\sin\left(\frac{N\pi}{2}\mathbf{Q}_e(\tilde{d}[n])\right).\end{aligned} \tag{A.26}$$

Quantizers $\mathbf{Q}_o$, $\mathbf{Q}_e$, $\mathbf{Q}_{dc}$ *are as defined in* (A.19),(A.20), *and* (A.21), *respectively. For* $n\in\mathbb{Z}$ *such that* $\left|\tilde{a}[n]\cos\left(\frac{2\pi n}{N}\right)\right|=\gamma$, *we have* $\tilde{y}[n]=\frac{1}{2}\operatorname{sign}\left(\tilde{a}[n]\cos\left(\frac{2\pi n}{N}\right)\right)$.

Proof. See the section A.0.4. □

It now follows from the theorem that the fundamental component $\tilde{y}_1$ of signal $\tilde{y}$ can be written as

$$\tilde{y}_1[n]=const\cdot\sin\left(\pi\mathbf{Q}(\tilde{d}[n])\right), \tag{A.27}$$

where the constant in the above expression depends only on N and quantizer $\mathbf{Q}$ is defined either by (A.19) or by (A.20), depending on the parity of the oversampling ratio N. Quantization of $\tilde{d}$ can be equivalently understood as quantization of signal $\tilde{a}$, since $\tilde{d}[n]=g_\gamma(\tilde{a}[n])$.

B. Multi-Level RF-DPWM

It is easy to see that a similar model for a general $(2M+1)$-level digital RF-PWM now readily follows from Theorem A.0.3. Let $\Gamma = \{\gamma_1, \ldots, \gamma_M\} \subset \mathbb{R}$ and $\mathcal{A} = \{\alpha_0, \ldots, \alpha_M\} \subset \mathbb{R}$ satisfy (A.2)-(A.3).

Theorem A.0.4. *M-level RF-DPWM system $\tilde{\mathbf{P}}_{\Gamma,\mathcal{A}} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ maps input signal $\tilde{a} \in \ell^2(\mathbb{R})$ into output signal $\tilde{y} = \tilde{\mathbf{P}}_{\Gamma,\mathcal{A}}\tilde{a} \in \ell^2(\mathbb{R})$ such that, for all $n \in \mathbb{Z}$ with $\left|\tilde{a}[n]\cos\left(\frac{2\pi n}{N}\right)\right| \neq \gamma$, the following expressions hold*

(1) if $\mod(N,2) = 1$ *and* $N_0 = \frac{N-1}{2}$, *then*

$$\tilde{y}[n] = \tilde{y}_0[n] + \sum_{k=1}^{N_0} \tilde{y}_k[n]\cos\left(\frac{2\pi kn}{N}\right), \tag{A.28}$$

where $\tilde{d}_m[n] = g_{\gamma_m}(\tilde{a}[n])$, *for all* $m \in \{1, \ldots, M\}$ *and*

$$\begin{aligned} \tilde{y}_0[n] &= \sum_{m=1}^{M} (\alpha_m - \alpha_{m-1})\mathbf{Q}_{dc}(\tilde{d}_m[n]), \\ \tilde{y}_k[n] &= \sum_{m=1}^{M} \frac{2(\alpha_m - \alpha_{m-1})}{N\sin\left(\frac{\pi k}{2N}\right)} \sin\left(\pi k \mathbf{Q}_o(\tilde{d}_m[n])\right) \quad \forall k \in \{1, \ldots, N_0\}, \end{aligned} \tag{A.29}$$

(2) if $\mod(N,4) = 0$ *then*

$$\tilde{y}[n] = \sum_{k=1}^{N_0} \tilde{y}_k[n]\cos\left(\frac{2\pi(2k-1)n}{N}\right), \tag{A.30}$$

where $\tilde{d}_m[n] = g_{\gamma_m}(\tilde{a}[n])$, *for all* $m \in \{1, \ldots, M\}$ *and*

(i) if $\mod(N,4) = 0$ *then* $N_0 = \frac{N}{4}$ *and*

$$\tilde{y}_k[n] = \sum_{m=1}^{M} \frac{4(\alpha_m - \alpha_{m-1})}{N\sin\left(\frac{\pi(2k-1)}{N}\right)} \sin\left(\pi(2k-1)\mathbf{Q}_e(\tilde{d}_m[n])\right), \quad \forall k \in \{1, \ldots, N_0\} \tag{A.31}$$

(ii) if $\mod(N,4)=2$ *then* $N_0=\frac{N+2}{4}$ *and*

$$\tilde{y}_k[n]=\sum_{m=1}^{M}\frac{4(\alpha_m-\alpha_{m-1})}{N\sin\left(\frac{\pi(2k-1)}{N}\right)}\sin\left(\pi(2k-1)\mathbf{Q}_e(\tilde{d}_m[n])\right),\quad \forall k\in\{1,\ldots,N_0-1\},$$
$$\tilde{y}_{\frac{N+2}{4}}[n]=\frac{2}{N}\sum_{m=1}^{M}(\alpha_m-\alpha_{m-1})\sin\left(\frac{N\pi}{2}\mathbf{Q}_e(\tilde{d}_m[n])\right). \tag{A.32}$$

Quantizers $\mathbf{Q}_o$, $\mathbf{Q}_e$, $\mathbf{Q}_{dc}$ *are as defined in* (A.19),(A.20), *and* (A.21), *respectively. For* $n\in\mathbb{Z}$ *such that* $\left|\tilde{a}[n]\cos\left(\frac{2\pi n}{N}\right)\right|=\gamma_m$, *for any* $m\in\{1,\ldots,M\}$, *we have* $\tilde{y}[n]=\frac{1}{2}(\alpha_m+\alpha_{m-1})\operatorname{sign}\left(\tilde{a}[n]\cos\left(\frac{2\pi n}{N}\right)\right)$.

Proof. The proof of theorem A.0.4 immediately follows from the definition of multi-level RF-DPWM and the results of theorem A.0.3. □

A.0.3 Proof of Theorem A.0.1

Let $\mathcal{P}:\mathbb{R}\times\mathbb{R}^+\to\{-1,0,1\}$ be a comparator map defined by

$$\mathcal{P}(\xi,\mu)=\begin{cases}1, & \xi\geq\mu,\\ 0, & |\xi|<\mu,\\ -1, & \xi\leq-\mu,\end{cases}\qquad \text{for all } \xi\in\mathbb{R},\mu\in\mathbb{R}^+. \tag{A.33}$$

For arbitrary $a\in\mathcal{L}^2(\mathbb{R})$, let $v=v(t_1,t_2)=\mathcal{P}(a(t_1)\cos(2\pi f_c t_2),\gamma)$. Clearly, function v is periodic in t_2, with period $1/f_c$. Therefore, for any fixed value t_1^* of t_1 (i.e., for a fixed value $a(t_1^*)$ of $a(t_1)$), signal $v(t_1^*,t_2)$ can be expanded into its Fourier series in t_2 as follows

$$v(t_1^*,t_2)=\sum_{k=-\infty}^{\infty}C_k(a(t_1^*))e^{jk\omega_p t_2}, \tag{A.34}$$

where coefficients $C_k(a(t_1^*))$ are defined as

$$C_k(a(t_1^*)) = \frac{1}{T_p}\int_0^{T_p} \mathcal{P}(a(t_1^*)\cos(2\pi f_c\tau), \gamma)e^{-jk\omega_p\tau}d\tau.$$

Convergence of the series on the right-hand side of (A.34) should be understood in the sense of its principal value, that is the infinite sum is equal to $\lim_{K\to\infty}\sum_{k=-K}^{K} C_k(a(t_1^*))e^{jk\omega_p t_2}$. Equality in (A.34) holds point-wise for every $t_2 \in \mathbb{R}$ such that $a(t_1^*)\cos(2\pi f_c t_2) \neq \gamma$. At the points of discontinuity of v, that is, for t_2 such that $a(t_1)\cos(2\pi f_c t_2) = \gamma$, we have that

$$v(t_1^*, t_2) = \frac{1}{2}(v(t_1^*, t_{2-}) + v(t_1^*, t_{2+})) = \frac{1}{2},$$

where $v(t_1^*, t_{2-})$ and $v(t_1^*, t_{2+})$ are the directional limits of $v(t_1^*, \cdot)$ at t_2. It follows from (A.33) that $v(t_1^*, t_2)$ is a periodic square wave for which the following Fourier series expansion holds (see, e.g., [110]):

$$v(t_1^*, t_2) = \sum_{k=1}^{\infty} \frac{4\sin((2k-1)\pi d(t_1^*))}{(2k-1)\pi}\cos(2(2k-1)\pi f_c t_2), \tag{A.35}$$

where

$$d(t_1^*) = g_\gamma(a(t_1^*)). \tag{A.36}$$

Again, the equality sign in (A.35) should be understood as described above.

Since (A.35) holds point-wise whenever $a(t_1)\cos(2\pi f_c t_2) \neq \gamma$, by setting $t_1 = t_2 = t$ we get that the RF-APWM output signal y at every time t such that $a(t)\cos(2\pi f_c t) \neq \gamma$, is given by the formula

$$y(t) = v(t,t) = \sum_{k=1}^{\infty} \frac{4\sin((2k-1)\pi d(t))}{(2k-1)\pi}\cos(2(2k-1)\pi f_c t)e^{jk\omega_c t}, \tag{A.37}$$

where $d = d(t)$ is defined by

$$d(t) = g_\gamma(a(t)). \tag{A.38}$$

This concludes the proof of Theorem A.0.1.

A.0.4 Proof of Theorem A.0.3

Let us first state the following corollary of lemma 4.7.1.

Corollary A.0.5. *Let* $M \in \mathbb{N}$, $M > 1$, $d \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, *and* $\mathcal{I} = \left\{\left\lfloor -\frac{M}{4} \right\rfloor, \ldots, \left\lfloor \frac{M}{4} \right\rfloor\right\}$. *Then for all* $m \in \left\{1, \ldots, \left\lfloor \frac{M-1}{2} \right\rfloor\right\}$ *the following equalities hold*

$$\sum_{n=-\infty}^{\infty} \frac{\sin(\pi M n d)}{\pi M n} = \begin{cases} \frac{2i+1}{M}, & d \in \left(\frac{2i}{M}, \frac{2i+2}{M}\right], i \in \mathcal{I}, d \neq 0 \\ 0, & d = 0 \end{cases}, \tag{A.39}$$

$$\sum_{n=-\infty}^{\infty} \frac{\sin(\pi(Mn+m)d)}{\pi(Mn+m)} == \begin{cases} \frac{\sin\left(\pi m \frac{2i+1}{M}\right)}{M \sin\left(\frac{\pi m}{M}\right)}, & d \in \left(\frac{2i}{M}, \frac{2i+2}{M}\right], i \in \mathcal{I}, d \neq 0 \\ 0, & d = 0 \end{cases}. \tag{A.40}$$

As previously noted, the output $\tilde{y}[n]$ of RF-DPWM can be thought of as a result of sampling of the output $y(t)$ of the equivalent RF-APWM system at sampling frequency $f_s = 1/T$, for some $T > 0$, where the corresponding carrier frequency of RF-APWM is equal to $f_c = f_s/N$. From (A-2) it follows that $\tilde{y}[n]$ can be expressed as

$$\tilde{y}[n] = y(nT) = \sum_{k=1}^{\infty} \frac{4 \sin((2k-1)\pi \tilde{d}[n])}{\pi(2k-1)} \cos\left(\frac{2\pi(2k-1)n}{N}\right), \tag{A.41}$$

where $\tilde{d}[n] = d(nT)$. Integer period of the discrete cosine in (A.41) implies existence of only a finitely many independent harmonics, and the infinite sum in (A.41) can be simplified to a finite sum, where the number of summands depends on the parity of N as follows.

(1) $\mathrm{mod}(N, 2) = 1$

By careful examination of discrete frequencies in $(A.41)$, it is easy to see that the total number of independent harmonics is equal to $(N+1)/2$. It follows that signal $\tilde{y}[n]$ can be expressed as

$$\tilde{y}[n] = \tilde{y}_0[n] + \sum_{k=1}^{\frac{N-1}{2}} \tilde{y}_k[n] \cos\left(\frac{2\pi k n}{N}\right), \tag{A.42}$$

where

$$\tilde{y}_0[n] = \sum_{l=1}^{\infty} \frac{4\sin(\pi N(2l-1)\tilde{d}[n])}{\pi N(2l-1)}, \tag{A.43}$$

and

$$\tilde{y}_k[n] = \sum_{l=0}^{\infty} \frac{4\sin(\pi(2Nl+k)\tilde{d}[n])}{\pi(2Nl+k)} + \sum_{l=0}^{\infty} \frac{4\sin(\pi(2Nl+2N-k)\tilde{d}[n])}{\pi(2Nl+2N-k)}, \tag{A.44}$$

for $k \in \{1, \ldots, (N-1)/2\}$. Signal $\tilde{y}_0$ can be understood as a baseband component of $\tilde{y}$, while $\tilde{y}_k$ are higher order harmonics, similar to the description of APWM output, given in (4.5). It is now clear that spectral aliasing effects manifest in (A.43)-(A.44) in terms of an infinite number of additional summands.

Let us now simplify expressions in (A.43)-(A.44). Equation (A.43) for $\tilde{y}_0[n]$ can be rewritten as

$$\tilde{y}_0[n] = \sum_{l=-\infty}^{\infty} \frac{2\sin(\pi N(2l-1)\tilde{d}[n])}{\pi N(2l-1)} = \; = 2\sum_{l=-\infty}^{\infty} \frac{\sin(\pi Nl\tilde{d}[n])}{\pi Nl} - 2\sum_{l=-\infty}^{\infty} \frac{\sin(2\pi Nl\tilde{d}[n])}{2\pi Nl}, \tag{A.45}$$

while $\tilde{y}_k[n]$ can be expressed as

$$\tilde{y}_k[n] = 4 \cdot \sum_{l=-\infty}^{\infty} \frac{\sin(\pi(2Nl+k)\tilde{d}[n])}{\pi(2Nl+k)}. \tag{A.46}$$

It follows from corollary A.0.5 that the infinite sums in (A.45) simplify to

$$\sum_{l=-\infty}^{\infty} \frac{\sin(\pi Nl\tilde{d}[n])}{\pi Nl} = \frac{2i+1}{N}, \tag{A.47}$$

when $\tilde{d}[n] \in \left(\frac{2i}{N}, \frac{2i+2}{N}\right]$, for all $i \in \mathcal{I} = \left\{\left\lfloor -\frac{N}{4}\right\rfloor, \ldots, \left\lfloor \frac{N}{4}\right\rfloor\right\}$, and

$$\sum_{l=-\infty}^{\infty} \frac{\sin(2\pi Nl\tilde{d}[n])}{2\pi Nl} = \frac{2m+1}{2N}, \tag{A.48}$$

when $\tilde{d}[n] \in \left(\frac{m}{N}, \frac{m+1}{N}\right]$, for all $m \in \mathcal{M} = \left\{\left\lfloor -\frac{N}{2}\right\rfloor, \ldots, \left\lfloor \frac{N}{2}\right\rfloor\right\}$. Therefore, $\tilde{y}_0[n]$ can

be expressed as

$$\tilde{y}_0[n] = \frac{4i - 2m + 1}{N}, \tag{A.49}$$

where i and m depend on the value of $\tilde{d}[n]$, as given above. It is not hard to see that for each even value of $m \in \mathcal{M}$ there exists an $i \in \mathcal{I}$ such that $i = m/2$. Similarly, for each odd value of $m \in \mathcal{M}$ there exists an $i \in \mathcal{I}$ such that $i = (m-1)/2$. It now follows from (A.49) that for all $m \in \mathcal{M}$ when $\tilde{d}[n] \in \left(\frac{m}{N}, \frac{m+1}{N}\right]$ we have

$$\tilde{y}_0[n] = \begin{cases} \frac{1}{N}, & m - \text{even} \\ -\frac{1}{N}, & m - \text{odd} \end{cases}. \tag{A.50}$$

A closed-form expression for $\tilde{y}_k[n]$ is obtained by applying corollary A.0.5, for $M = 2N$ and $m = k$, to (A.46):

$$\tilde{y}_k[n] = \frac{2\sin\left(\pi k \frac{2i+1}{2N}\right)}{N\sin\left(\frac{\pi k}{2N}\right)} \tag{A.51}$$

when $\tilde{d}[n] \in \left(\frac{i}{N}, \frac{i+1}{N}\right]$, $\tilde{d}[n] \neq 0$ for all $i \in \left\{\left\lfloor -\frac{N}{2} \right\rfloor, \ldots, \left\lfloor \frac{2N-1}{4} \right\rfloor\right\}$, and $\tilde{y}_k[n] = 0$ when $\tilde{d}[n] = 0$.

(2) $\mod(N, 2) = 0$

Let $\mathcal{I} = \left\{\left\lfloor -\frac{N}{4} \right\rfloor, \ldots, \left\lfloor \frac{N-1}{4} \right\rfloor\right\}$. We consider the following two sub-cases.

(i) $\mod(N, 4) = 0$

The number of independent harmonics is equal to $N/4$ and signal $\tilde{y}[n]$, from (A.41), can be rewritten as

$$\tilde{y}[n] = \sum_{k=1}^{\frac{N}{4}} \tilde{y}_k[n] \cos\left(\frac{2\pi(2k-1)n}{N}\right), \tag{A.52}$$

where

$$\tilde{y}_k[n] = \sum_{l=0}^{\infty} \frac{4\sin(\pi(Nl + 2k - 1)\tilde{d}[n])}{\pi(Nl + 2k - 1)} + \sum_{l=1}^{\infty} \frac{4\sin(\pi(Nl - 2k + 1)\tilde{d}[n])}{\pi(Nl - 2k + 1)}. \tag{A.53}$$

Similar to simplifications done in (A.45), it can be shown that $\tilde{y}_k[n]$ can be rewritten as

$$\tilde{y}_k[n] = \sum_{l=-\infty}^{\infty} \frac{4\sin(\pi(Nl+2k-1)\tilde{d}[n])}{\pi(Nl+2k-1)}, \tag{A.54}$$

It follows from corollary A.0.5, for $M = N$ and $m = 2k-1$, that (A.54) can be simplified to

$$\tilde{y}_k[n] = \frac{4\sin\left(\pi(2k-1)\frac{2i+1}{N}\right)}{N\sin\left(\frac{\pi(2k-1)}{N}\right)}, \tag{A.55}$$

when $\tilde{d}[n] \in \left(\frac{2i}{N}, \frac{2i+2}{N}\right]$, for all $i \in \mathcal{I}$, with $\tilde{d}[n] \neq 0$, and $\tilde{y}_k[n] = 0$ for $\tilde{d}[n] = 0$.

(ii) $\mod(N, 4) = 2$

In this case, we have

$$\tilde{y}[n] = \sum_{k=1}^{\frac{N+2}{4}} \tilde{y}_k[n]\cos\left(\frac{2\pi(2k-1)n}{N}\right), \tag{A.56}$$

where $\tilde{y}_k[n]$ is as given in (A.55), for all $1 \leq k \leq \frac{N-2}{4}$. For $k_0 = \frac{N+2}{4}$ we have

$$\tilde{y}_{k_0}[n] = \sum_{l=0}^{\infty} \frac{4\sin(\pi(Nl+N/2)\tilde{d}[n])}{\pi(Nl+N/2)} = \sum_{l=-\infty}^{\infty} \frac{2\sin(\pi(Nl+N/2)\tilde{d}[n])}{\pi(Nl+N/2)}. \tag{A.57}$$

It follows from corollary A.0.5, for $M = N$ and $m = 2k-1$, that the above expression for $\tilde{y}_{k_0}$ simplifies to

$$\tilde{y}_{k_0}[n] = \begin{cases} \frac{2}{N}\sin\left(\frac{\pi(2i+1)}{2}\right), & \tilde{d}[n] \in \left(\frac{2i}{N}, \frac{2i+2}{N}\right], \forall i \in \mathcal{I} \\ 0, & \tilde{d}[n] = 0. \end{cases} \tag{A.58}$$

This concludes the proof of theorem A.0.3.

Bibliography

[1] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits, A Design Perspective*. Prentice Hall, 2003.

[2] N. R. Norris, "Exploring the optimality of various bacterial motility strategies: a stochastic hybrid systems approach," M.S. Dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. of Technology, Cambridge, MA, Sep 2013.

[3] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: Wiley, 1949.

[4] T. A. S. C. Technical Staff, *Applied Optimal Estimation*, ser. MIT Press, A. Gelb, Ed. MIT Press, 1974.

[5] G. Turin, "An introduction to matched filters," *IRE Transactions on Information Theory*, vol. 6, no. 3, pp. 311–329, June 1960.

[6] Shao-Po Wu, S. Boyd, and L. Vandenberghe, *FIR Filter Design via Spectral Factorization and Convex Optimization*, ser. Applied and Computational Control, Signals and Circuits: Volume 1. Boston, MA: Birkhäuser Boston, 1999, pp. 215–245.

[7] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, ser. Feher/Prentice Hall Digital and. Prentice Hall, 1996.

[8] J. FISHER, "Optimal nonlinear filtering," ser. Advances in Control Systems, C. LEONDES, Ed. Elsevier, 1967, vol. 5, pp. 197 – 300.

[9] F. Daum, "Nonlinear filters: beyond the kalman filter," *IEEE Aerospace and Electronic Systems Magazine*, vol. 20, no. 8, pp. 57–69, Aug 2005.

[10] C. Novara, F. Ruiz, and M. Milanese, "A new approach to optimal filter design for nonlinear systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 11 453 – 11 458, 2011, 18th IFAC World Congress.

[11] P. P. Khargonekar, I. R. Petersen, and M. A. Rotea, "H/sub infinity /-optimal control with state-feedback," *IEEE Transactions on Automatic Control*, vol. 33, no. 8, pp. 786–788, Aug 1988.

[12] A. J. van der Schaft, "L/sub 2/-gain analysis of nonlinear systems and nonlinear state-feedback h/sub infinity / control," *IEEE Transactions on Automatic Control*, vol. 37, no. 6, pp. 770–784, June 1992.

[13] A. Isidori and A. Astolfi, "Disturbance attenuation and h/sub infinity /-control via measurement feedback in nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 37, no. 9, pp. 1283–1293, Sep. 1992.

[14] J. A. Ball, J. W. Helton, and M. L. Walker, "H/sup infinity / control for nonlinear systems with output feedback," *IEEE Transactions on Automatic Control*, vol. 38, no. 4, pp. 546–559, April 1993.

[15] M. D. S. Aliyu, *Nonlinear $\mathcal{H}_\infty$-Control, Hamiltonian Systems and Hamilton-Jacobi Equations*. CRC Press, Taylor & Francis Group, 2011.

[16] A. Arbi, "Spectral and energy efficiency in cellular mobile radio access networks," Ph.D. Dissertation, Dept. Elect. Eng. Comput. Sci., University of Sheffield, Sheffield, UK, June 2017.

[17] S. M. G. Research), "Green it: The new industry shock wave," December 2007.

[18] L. Suarez, L. Nuaymi, and J.-M. Bonnin, "An overview and classification of research approaches in green wireless networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, p. 142, Apr 2012.

[19] Global e-sustainability initiative (GeSI), "SMART 2020: Enabling the low carboneconomy in the information age," 2008.

[20] "Us. residential information technology energy consumption in 2005, final report," March 2006.

[21] C. Forster, I. DIckie, G. Maile, H. Smith, and M. Crisp, "Understanding the environmental impact of communication systems," April 2009.

[22] I. Humar, X. Ge, L. Xiang, M. Jo, M. Chen, and J. Zhang, "Rethinking energy efficiency models of cellular networks with embodied energy," *IEEE Network*, vol. 25, no. 2, pp. 40–49, March 2011.

[23] O. Blume, D. Zeller, and U. Barth, "Approaches to energy efficient wireless access networks," in *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, March 2010, pp. 1–5.

[24] A. He, A. E. Amanna, T. Tsou, X. Chen, D. Datla, J. D. Gaeddert, T. R. Newman, S. M. S. Hasan, H. Volos, J. H. Reed, and T. Bose, "Green communications: A call for power efficient wireless systems," *Journal of Communications*, vol. 6, pp. 340–351, 2011.

[25] 3GPP, 3rd Generation Partnership Project (3GPP), Tech. Rep., version 16.

[26] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[27] Y. Rahmatallah and S. Mohan, "Peak-to-average power ratio reduction in ofdm systems: A survey and taxonomy," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1567–1592, Fourth 2013.

[28] Seung Hee Han and Jae Hong Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," *IEEE Wireless Communications*, vol. 12, no. 2, pp. 56–65, April 2005.

[29] G. Wunder, R. F. H. Fischer, H. Boche, S. Litsyn, and J. No, "The papr problem in ofdm transmission: New directions for a long-lasting problem," *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 130–144, Nov 2013.

[30] Kyeongcheol Yang and Seok-Il Chang, "Peak-to-average power control in ofdm using standard arrays of linear block codes," *IEEE Communications Letters*, vol. 7, no. 4, pp. 174–176, April 2003.

[31] S. B. Slimane, "Reducing the peak-to-average power ratio of ofdm signals through precoding," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 686–695, March 2007.

[32] M. Hao and C. Lai, "Precoding for papr reduction of ofdm signals with minimum error probability," *IEEE Transactions on Broadcasting*, vol. 56, no. 1, pp. 120–128, March 2010.

[33] O. Daoud and O. Alani, "Reducing the papr by utilisation of the ldpc code," *IET Communications*, vol. 3, no. 4, pp. 520–529, April 2009.

[34] T. Jiang and X. Li, "Using fountain codes to control the peak-to-average power ratio of ofdm signals," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 3779–3785, Oct 2010.

[35] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, June 2006.

[36] B. M. Popovic, "Synthesis of power efficient multitone signals with flat amplitude spectrum," *IEEE Transactions on Communications*, vol. 39, no. 7, pp. 1031–1033, July 1991.

[37] J. A. Davis and J. Jedwab, "Peak-to-mean power control in ofdm, golay complementary sequences, and reed-muller codes," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2397–2417, Nov 1999.

[38] Y. Tsai, S. Deng, K. Chen, and M. Lin, "Turbo coded ofdm for reducing papr and error rates," *IEEE Transactions on Wireless Communications*, vol. 7, no. 1, pp. 84–89, Jan 2008.

[39] K. G. Paterson and V. Tarokh, "On the existence and construction of good codes with low peak-to-average power ratios," *IEEE Transactions on Information Theory*, vol. 46, no. 6, pp. 1974–1987, Sep. 2000.

[40] R. W. Bauml, R. F. H. Fischer, and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *Electronics Letters*, vol. 32, no. 22, pp. 2056–2057, Oct 1996.

[41] H. Breiling, S. H. Muller-Weinfurtner, and J. B. Huber, "Slm peak-power reduction without explicit side information," *IEEE Communications Letters*, vol. 5, no. 6, pp. 239–241, June 2001.

[42] S. H. Muller and J. B. Huber, "Ofdm with reduced peak-to-average power ratio by optimum combination of partial transmit sequences," *Electronics Letters*, vol. 33, no. 5, pp. 368–369, Feb 1997.

[43] L. J. Cimini and N. R. Sollenberger, "Peak-to-average power ratio reduction of an ofdm signal using partial transmit sequences," *IEEE Communications Letters*, vol. 4, no. 3, pp. 86–88, March 2000.

[44] Seung Hee Han and Jae Hong Lee, "Papr reduction of ofdm signals using a reduced complexity pts technique," *IEEE Signal Processing Letters*, vol. 11, no. 11, pp. 887–890, Nov 2004.

[45] J. Tellado, "Peak-to-average power reduction for multicarrier modulation," PhD Dissertation, Stanford University, Sept 1999.

[46] B. S. Krongold and D. L. Jones, "Par reduction in ofdm via active constellation extension," *IEEE Transactions on Broadcasting*, vol. 49, no. 3, pp. 258–268, Sep. 2003.

[47] Xianbin Wang, T. T. Tjhung, and C. S. Ng, "Reduction of peak-to-average power ratio of ofdm system using a companding technique," *IEEE Transactions on Broadcasting*, vol. 45, no. 3, pp. 303–307, Sep. 1999.

[48] Tao Jiang and Guangxi Zhu, "Nonlinear companding transform for reducing peak-to-average power ratio of ofdm signals," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 342–346, Sep. 2004.

[49] Tao Jiang, Yang Yang, and Yong-Hua Song, "Exponential companding technique for papr reduction in ofdm systems," *IEEE Transactions on Broadcasting*, vol. 51, no. 2, pp. 244–248, June 2005.

[50] J. Hou, J. Ge, D. Zhai, and J. Li, "Peak-to-average power ratio reduction of ofdm signals with nonlinear companding scheme," *IEEE Transactions on Broadcasting*, vol. 56, no. 2, pp. 258–262, June 2010.

[51] P. Börjesson, H. G. Feichtinger, N. Grip, M. Isaksson, N. Kaiblinger, P. Ödling, and L. Persson, "A low-complexity PAR-reduction method for DMT-VDSL," in *Proceedings. 5th International Symposium on Digital Signal Processing for Communications Systems (DSPCS)*, Feb 1999, p. 164âĂŞ169.

[52] ——, "DMT PAR-reduction by weighted cancellation waveforms," in *Proceedings of Radiovetenskap och kommunikation 99 (RVK 99)*, June 1999, p. 303âĂŞ307.

[53] J. Armstrong, "Peak-to-average power reduction for ofdm by repeated clipping and frequency domain filtering," *Electronics Letters*, vol. 38, no. 5, pp. 246–247, Feb 2002.

[54] Xiaodong Li and L. J. Cimini, "Effects of clipping and filtering on the performance of ofdm," *IEEE Communications Letters*, vol. 2, no. 5, pp. 131–133, May 1998.

[55] H. Ochiai and H. Imai, "Performance analysis of deliberately clipped ofdm signals," *IEEE Transactions on Communications*, vol. 50, no. 1, pp. 89–101, Jan 2002.

[56] Luqing Wang and C. Tellambura, "A simplified clipping and filtering technique for par reduction in ofdm systems," *IEEE Signal Processing Letters*, vol. 12, no. 6, pp. 453–456, June 2005.

[57] K. Anoh, C. Tanriover, B. Adebisi, and M. Hammoudeh, "A new approach to iterative clipping and filtering papr reduction scheme for ofdm systems," *IEEE Access*, vol. 6, pp. 17 533–17 544, April 2018.

[58] Y. C. Wang and Z. Q. Luo, "Optimized iterative clipping and filtering for papr reduction of ofdm signals," *IEEE Transactions on Communications*, vol. 59, no. 1, pp. 33–37, January 2011.

[59] X. Zhu, W. Pan, H. Li, and Y. Tang, "Simplified approach to optimized iterative clipping and filtering for papr reduction of ofdm signals," *IEEE Transactions on Communications*, vol. 61, no. 5, pp. 1891–1901, May 2013.

[60] P. B. Kennington, *High linearity RF amplifier design*. Norwood, MA: Artech House, 2000.

[61] J. Vuolevi and T. Rahkonen, *Distortion in RF Power Amplifiers*. Norwood, MA: Artech House, 2003.

[62] J. C. Pedro and S. A. Mass, "A comparative overview of microwave and wireless power-amplifier behavioral modeling approaches," *IEEE Trans. Microw. Theory Techn.*, vol. 53, no. 4, pp. 1150–1163, Apr. 2005.

[63] A. A. M. Saleh and J. Salz, "Adaptive linearization of power amplifiers in digital radio systems," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1019–1033, Apr. 1983.

[64] W. Bösch and G. Gatti, "Measurement and simulation of memory effects in predistortion linearizers," *IEEE Trans. Microw. Theory Techn.*, vol. 37, no. 12, pp. 1885–1890, Nov. 1989.

[65] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electron. Lett.*, vol. 37, no. 23, pp. 1417–1418, Nov. 2001.

[66] L. Ding, G. T. Zhou, D. R. Morgan, Z. Ma, J. S. Kenney, J. Kim, and C. R. Giardina, "A robust digital baseband predistorter constructed using memory polynomials," *IEEE Trans. Commun.*, vol. 52, no. 1, pp. 159–165, Jan. 2004.

[67] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*. New York: Wiley, 2000.

[68] T. Liu, S.Boumaiza, and F.M.Ghannouchi, "Augmented hammerstein predistorter for linearization of broad-band wireless transmitters," *IEEE Trans. Microw. Theory Techn.*, vol. 54, no. 4, pp. 1340–1349, Jun. 2006.

[69] A. Zhu and T. Brazil, "Behavioral modeling of rf power amplifiers based on pruned volterra series," *IEEE Microw. Wireless Compon. Lett.*, vol. 14, no. 12, pp. 563–565, Dec. 2004.

[70] D. R. Morgan, Z. Ma, J. Kim, M. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of rf power amplifiers," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3852–3860, Oct. 2006.

[71] A. Zhu, J. C. Pedro, and T. J. Brazil, "Dynamic deviation reduction-based volterra behavioral modeling of rf power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 54, no. 12, pp. 4323–4332, Dec. 2006.

[72] A. Zhu, P. J. Draxler, J. J. Yan, T. J. Brazil, D. F. Kimball, and P. M. Asbeck, "Open-loop digital predistorter for rf power amplifiers using dynamic deviation reduction-based volterra series," *IEEE Trans. Microw. Theory Techn.*, vol. 56, no. 7, pp. 1524–1534, Jul. 2008.

[73] M. Rawat, K. Rawat, and F. M. Ghannouchi, "Adaptive digital predistortion of wireless power amplifiers/transmitters using dynamic real-valued focused time-delay line neural networks," *IEEE Trans. Microw. Theory Techn.*, vol. 58, no. 1, pp. 95–104, Jan. 2010.

[74] M. Rawat, K. Rawat, F. M. Ghannouchi, S. Bhattacharjee, and H. Leung, "Generalized rational functions for reduced-complexity behavioral modeling and digital predistortion of broadband wireless transmitters," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 2, pp. 485–498, Feb. 2014.

[75] O. Hammi and F. M. Ghannouchi, "Twin nonlinear two-box models for power amplifiers and transmitters exhibiting memory effects with application to digital predistortion," *IEEE Microw. Wireless Compon. Lett.*, vol. 19, no. 8, pp. 530–532, Aug. 2009.

[76] M. Younes, O. Hammi, A. Kwan, and F. M. Ghannouchi, "An accurate complexity-reduced âĂIJplumeâĂİ model for behavioral modeling and digital predistortion of rf power amplifiers," *IEEE Trans. Indust. Electron.*, vol. 58, no. 4, pp. 1397–1405, Apr. 2011.

[77] C. Yu, L. Guan, and A. Zhu, "Band-limited volterra series-based digital predistortion for wideband rf power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 12, pp. 4198–4208, Dec. 2012.

[78] F. H. Raab, P. Asbeck, S. Cripps, P. B. Kenington, Z. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "Power amplifiers and transmitters for rf and microwave," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 3, pp. 814–826, March 2002.

[79] M. Eron, B. Kim, F. Raab, R. Caverly, and J. Staudinger, "The head of the class," *IEEE Microwave Magazine*, vol. 12, no. 7, pp. S16–S33, Dec 2011.

[80] F. H. Raab, "Class-d power amplifier with rf pulse-width modulation," in *2010 IEEE MTT-S International Microwave Symposium*, May 2010, pp. 924–927.

[81] F. M. Ghannouchi, "Power amplifier and transmitter architectures for software defined radio systems," *IEEE Circuits and Systems Magazine*, vol. 10, no. 4, pp. 56–63, Fourthquarter 2010.

[82] P.-J. Nuyts, P. Reynaert, and W. Dehaene, *Continuous-Time Digital Front-Ends for Multistandard Wireless Transmission*, ser. Analog Circuits and Signal Processing. Springer, 2014.

[83] A. Wentzel, C. Meliani, and W. Heinrich, "Rf class-s power amplifiers: State-of-the-art results and potential," in *2010 IEEE MTT-S International Microwave Symposium*, May 2010, pp. 812–815.

[84] B. Francois, E. Kaymaksút, and P. Reynaert, "Burst mode operation as an efficiency enhancement technique for RF power amplifiers," in *2011 XXXth URSI General Assembly and Scientific Symposium*, Istanbul, 2011, pp. 1–4.

[85] T. Johnson and S. P. Stapleton, "Rf class-d amplification with bandpass sigma-delta modulator drive signals," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 12, pp. 2507–2520, Dec 2006.

[86] H. Ruotsalainen, H. Arthaber, and G. Magerl, "A new quadrature pwm modulator with tunable center frequency for digital rf transmitters," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 11, pp. 756–760, Nov 2012.

[87] K. Hausmair, S. Chi, P. Singerl, and C. Vogel, "Aliasing-free digital pulse-width modulation for burst-mode rf transmitters," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 2, pp. 415–427, Feb 2013.

[88] R. Hezar, L. Ding, A. Banerjee, J. Hur, and B. Haroun, "A pwm based fully integrated digital transmitter/pa for wlan and lte applications," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 5, pp. 1117–1125, May 2015.

[89] S. Chung, R. Ma, S. Shinjo, H. Nakamizo, K. Parsons, and K. H. Teo, "Concurrent multiband digital outphasing transmitter architecture using multidimensional power coding," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 2, pp. 598–613, Feb 2015.

[90] H. S. Black, *Modulation theory*, ser. Bell Telephone Laboratories series. Van Nostrand, 1953.

[91] W. R. Bennett, "New results in the calculation of modulation products," *The Bell System Technical Journal*, vol. 12, no. 2, pp. 228–243, April 1933.

[92] F. Raab, "Radio frequency pulsewidth modulation," *IEEE Transactions on Communications*, vol. 21, no. 8, pp. 958–966, August 1973.

[93] J. Sun, *Dynamics and Control of Switched Electronic Systems*, ser. 2nd ed. Springer-Verlag, 2012.

[94] J. S. Chang, Meng-Tong Tan, Zhihong Cheng, and Yit-Chow Tong, "Analysis and design of power efficient class d amplifier output stages," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 6, pp. 897–902, June 2000.

[95] Bah-Hwee Gwee, J. S. Chang, and Huiyun Li, "A micropower low-distortion digital pulsewidth modulator for a digital class d amplifier," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 49, no. 4, pp. 245–256, April 2002.

[96] M. Norris, L. M. Platon, E. Alarcon, and D. Maksimovic, "Quantization noise shaping in digital pwm converters," in *2008 IEEE Power Electronics Specialists Conference*, June 2008, pp. 127–133.

[97] B. P. McGrath and D. G. Holmes, "An analytical technique for the determination of spectral components of multilevel carrier-based pwm methods," *IEEE Transactions on Industrial Electronics*, vol. 49, no. 4, pp. 847–857, Aug 2002.

[98] G. Carrara, S. Gardella, M. Marchesoni, R. Salutari, and G. Sciutto, "A new multilevel pwm method: a theoretical analysis," *IEEE Transactions on Power Electronics*, vol. 7, no. 3, pp. 497–505, July 1992.

[99] W. H. Lau, Bin Zhou, and H. S. H. Chung, "Compact analytical solutions for determining the spectral characteristics of multicarrier-based multilevel pwm," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 8, pp. 1577–1585, Aug 2004.

[100] Z. Song and D. V. Sarwate, "The frequency spectrum of pulse width modulated signals," *Signal Processing*, vol. 83, no. 10, pp. 2227 – 2258, Oct 2003.

[101] P. A. J. Nuyts, P. Reynaert, and W. Dehaene, "Frequency-domain analysis of digital pwm-based rf modulators for flexible wireless transmitters," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 1, pp. 238–246, Jan 2014.

[102] K. Hausmair, S. Chi, and C. Vogel, "How to reach 100% coding efficiency in multilevel burst-mode rf transmitters," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, May 2013, pp. 2255–2258.

[103] H. Enzinger and C. Vogel, "Analytical description of multilevel carrier-based pwm of arbitrary bounded input signals," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, June 2014, pp. 1030–1033.

[104] M. B. Sandler, "Investigation by simulation of a digitally addressed audio amplifier," PhD Dissertation, University of Essex, 1983.

[105] R. Ma, "A review of recent development on digital transmitters with integrated gan switch-mode amplifiers," in *2015 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT)*, Aug 2015.

[106] S. Santi, R. Rovatti, and G. Setti, "Spectral aliasing effects of pwm signals with time-quantized switching instants," in *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512)*, vol. 4, May 2004, pp. IV–689.

[107] A. M. A. Amin, M. I. El-Korfolly, and S. A. Mohammed, "Exploring aliasing distortion effects on regularly-sampled pwm signals," in *2008 3rd IEEE Conference on Industrial Electronics and Applications*, June 2008, pp. 2036–2041.

[108] H. Gheidi and P. M. Asbeck, "An improved algorithm for waveform generation for digitally-driven switching-mode power amplifiers," in *2016 IEEE Radio and Wireless Symposium (RWS)*, Jan 2016, pp. 187–189.

[109] D. Seebacher, P. Singerl, C. Schuberth, F. Dielacher, P. Reynaert, and W. BÃűsch, "Reduction of aliasing effects of rf pwm modulated signals by cross point estimation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 11, pp. 3184–3192, Nov 2014.

[110] A. V. Oppenheim, A. S. Willsky, and S. H. Nawad, *Signals and Systems*, ser. 2nd ed. Prentice-Hall, Inc., 1996.

[111] D. Markert, C. Haslach, H. Heimpel, A. Pascht, and G. Fischer, "Phase-modulated dsm-pwm hybrids with pulse length restriction for switch-mode power amplifiers," in *2014 44th European Microwave Conference*, Oct 2014, pp. 1364–1367.

[112] D. C. Dinis, R. F. Cordeiro, A. S. R. Oliveira, J. Vieira, and T. O. Silva, "A fully parallel architecture for designing frequency-agile and real-time reconfigurable fpga-based rf digital transmitters," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 3, pp. 1489–1499, March 2018.

[113] O. Tanovic and A. Megretski, "Real-time realization of a family of optimal infinite-memory non-causal systems," in *6th IFAC Conference on Nonlinear Model Predictive Control (NMPC)*, Madison, WI, Aug 2018, pp. 168–173.

[114] ——, "Causally stable approximation of optimal maps in maximal value constrained least-squares optimization," in *2019 European Control Conference (ECC)*, Naples, Italy, June 2019, pp. 1–6.

[115] Y. Liu, W. Pan, S. Shao, and Y. Tang, "A general digital predistortion architecture using constrained feedback bandwidth for wideband power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 5, pp. 1544–1555, May 2015.

[116] O. Hammi, A. Kwan, S. Bensmida, K. A. Morris, and F. M. Ghannouchi, "A digital predistortion system with extended correction bandwidth with application to lte-a nonlinear power amplifiers," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 12, pp. 3487–3495, Dec. 2014.

[117] G. M. Raz and B. D. V. Veen, "Baseband volterra filters for implementing carrier based nonlinearities," *IEEE Trans. Signal Process.*, vol. 46, no. 1, pp. 103–114, Jan. 1998.

[118] M. Morhac, "A fast algorithm of nonlinear volterra filtering," *IEEE Trans. Signal Process.*, vol. 39, no. 10, pp. 2353–2356, Oct. 1991.

[119] H. Ruotsalainen, N. Leder, B. Pichler, H. Arthaber, and G. Magerl, "Equivalent complex baseband model for digital transmitters based on 1-bit quadrature pulse encoding," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 11, pp. 2739–2747, Nov. 2015.

[120] L. Ding, F. Mujica, and Z. Yang, "Digital predistortion using direct learning with reduced bandwidth feedback," in *2013 IEEE MTT-S International Microwave Symposium Digest (MTT)*, Seattle, WA, June 2013, pp. 1–3.

[121] O. Tanovic, A. Megretski, Y. Li, V. M. Stojanovic, and M. Osqui, "Discrete-time models resulting from dynamic continuous-time perturbations in phase-amplitude modulation-demodulation schemes," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Las Vegas, NV, Dec 2016, pp. 6619–6624.

[122] O. Tanovic, A. Megretski, Y. Li, V. Stojanovic, and M. Osqui, "Equivalent baseband models and corresponding digital predistortion for compensating dynamic passband nonlinearities in phase-amplitude modulation-demodulation schemes," *IEEE Transactions on Signal Processing*, vol. 66, no. 22, pp. 5972–5987, Nov 2018.

[123] O. Tanovic, R. Ma, and K. H. Teo, "Novel baseband equivalent models of quadrature modulated all-digital transmitters," in *Radio Wireless Symposium (RWS) 2017*, Phoenix, AZ, Jan 2017, pp. 211–214.

[124] Patent.

[125] O. Tanovic, R. Ma, and K. H. Teo, "Theoretical bounds on time-domain resolution of multilevel carrier-based digital pwm signals used in all-digital transmitters," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boston, MA, pp. 1146–1149.

[126] O. Tanovic, R. Ma, and H. Sun, "Compact analytical description of digital radio-frequency pulse-width modulated signals," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[127] O. Tanovic, R. Ma, and K. H. Teo, "Optimal delta-sigma modulation based noise shaping for truly aliasing-free digital pwm," in *2017 47th European Microwave Conference (EuMC)*, Nuremberg, Germany, pp. 731–734.

[128] ——, "Simultaneous power encoding and upconversion for all-digital transmitters using digital pwm," in *2017 IEEE Asia Pacific Microwave Conference (APMC)*, Kuala Lumpar, Malaysia, pp. 837–840.

[129] O. Tanovic and R. Ma, "Truly aliasing-free digital rf-pwm power coding scheme for switched-mode power amplifiers," in *2018 IEEE Radio and Wireless Symposium (RWS)*, Anaheim, CA, pp. 68–71.

[130] O. Tanovic, R. Ma, P. Orlik, and A. Megretski, "Optimal power encoding of ofdm signals in all-digital transmitters," in *2019 IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, December 2019, pp. 1–5.

[131] R. Ma and O. Tanovic, "Noise mitigating quantizer for reducing nonlinear distortion in digital signal transmission," U.S. Patent 10 177 776, Jan, 2019.

[132] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.104, 06, version 14.2.2.

[133] T. R. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[134] A. A. Goldstein, "Convex programming in hilbert space," *Bull. Amer. Math. Soc.*, vol. 70, no. 5, pp. 709–710, 09 1964.

[135] E. Levitin and B. Polyak, "Constrained minimization methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 6, no. 5, pp. 1 – 50, 1966.

[136] D. Bertsekas, "On the Goldstein-Levitin-Polyak gradient projection method," *IEEE Transactions on Automatic Control*, vol. 21, no. 2, pp. 174–184, Apr 1976.

[137] S. Boyd and L. Chua, "Fading memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Transactions on Circuits and Systems*, vol. 32, no. 11, pp. 1150–1161, Nov 1985.

[138] D. Luenberger, *Optimization by Vector Space Methods*. New York, NY, USA: John Wiley & Sons, 1969.

[139] V. A. Yakubovich, "A frequency theorem in control theory," *Siberian Mathematical Journal*, vol. 14, no. 2, pp. 265–289, 1973.

[140] J. C. Willems, "Dissipative dynamical systems Part II: Linear systems with quadratic supply rates," *Archive for Rational Mechanics and Analysis*, vol. 45, no. 5, pp. 352–393, Jan 1972.

[141] D. Luenberger, *Introduction to Dynamic Systems: Theory, Models, and Applications*. New York, NY, USA: John Wiley & Sons, 1979.

[142] A. E. Frazho and W. Bhosri, *An Operator Perspective on Signals and Systems, Chapter 2 "Toeplitz and Laurent Operators"*. Basel: Birkhäuser, 2010.

[143] M. Schetzen, *The Volterra and Wiener theories of nonlinear systems*. reprint ed. Malabar, FL: Krieger, 2006.

[144] W. Frank, "Sampling requirements for volterra system identification," *IEEE Signal Process. Lett.*, vol. 3, no. 9, pp. 266–268, Sep. 1996.

[145] J. Tsimbinos and K. V. Lever, "Input nyquist sampling suffices to identify and compensate nonlinear systems," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2833–2837, Oct. 1998.

[146] J. G. Proakis and M. Salehi, *Digital Communications*. McGraw-Hill, 2007.

[147] D. C. Cox, "Linear amplification with nonlinear components," *IEEE Trans. Commun.*, vol. 22, no. 12, pp. 1942–1945, Dec. 1974.

[148] Q. Shi, "OFDM in bandpass nonlinearity," *IEEE Trans. Consumer Electron.*, vol. 42, pp. 253–258, Aug. 1996.

[149] A. N. D'Andrea, V. Lottici, , and R. Reggiannini, "Nonlinear predistortion of OFDM signals over frequency-selective fading channels," *IEEE Trans. Commun.*, vol. 49, no. 5, pp. 837–843, May 2001.

[150] F. Wang, D. Kimball, D. Lie, P. Asbeck, and L. E. Larson, "A monolithic high-efficiency 2.4-GHz 20-dbm SiGe BiCMOS envelope-tracking OFDM power amplifier," *IEEE J. Solid-State Circuits*, vol. 42, no. 6, pp. 1271–1281, Jun. 2007.

[151] J. Reina-Tosina, M. Allegue-Martinez, C. Crespo-Cadenas, C. Yu, and S. Cruces, "Behavioral modeling and predistortion of power amplifiers under sparsity hypothesis," *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 2, pp. 745–753, Feb. 2015.

[152] A. J. Cann, "Nonlinearity model with variable knee sharpness," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 16, no. 6, pp. 874–877, Nov. 1980.

[153] Y. Li, "Digital assistance design for analog systems: Digital baseband for outphasing power amplifiers," PhD Dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. of Technology, Cambridge, MA, Jun. 2013.

[154] Z. Li, "Efficient baseband design and implementation for high-throughput transmitters," PhD Dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. of Technology, Cambridge, MA, Sep. 2015.

[155] S. Wei, D. L. Goeckel, and P. A. Kelly, "Convergence of the complex envelope of bandlimited ofdm signals," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 4893–4904, Oct 2010.

[156] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, Oct 1998.

[157] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) conformance testing," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.104, 06, version 14.2.2.

[158] M. Nielsen and T. Larsen, "The head of the class," *IEEE Trans. Microw. Theory Techn.*, vol. 56, no. 2, pp. 300–304, Feb 2008.

[159] D. Markert, C. Haslach, G. Fischer, and A. Pascht, "Coding efficiency of rf pulse-width-modulation for mobile communications," in *2012 International Symposium on Signals, Systems, and Electronics (ISSSE)*, Oct 2012, pp. 1–5.

[160] T. Koike-Akino, Z. Qiuyao, R. Ma, and K. H. Teo, "System and method for linearizing power amplifiers," U.S. Patent 9 749 163, Aug, 2017.

[161] W. Chou and R. M. Gray, "Modulo sigma-delta modulation," *IEEE Transactions on Communications*, vol. 40, no. 8, pp. 1388–1395, Aug 1992.