HUMBERTO FIORAVANTE FERRO

HYBRID CONVEX COMBINATIONS FOR IIR SYSTEMS IDENTIFICATION

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Engenharia Elétrica.

São Paulo
2016

HUMBERTO FIORAVANTE FERRO

HYBRID CONVEX COMBINATIONS FOR IIR SYSTEMS IDENTIFICATION

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Engenharia Elétrica.

Área de Concentração:

Sistemas Eletrônicos

Orientador:

Prof. Dr. Cássio Guimarães Lopes

São Paulo
2016

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 15 de abril de 2016

Assinatura do autor: Humberto - Fioravante Ferro

Assinatura do orientador:

Catalogação-na-publicação

Ferro, Humberto Fioravante
Hybrid Convex Combinations for IIR Systems Identification / H. F. Ferro - versão corr. -- São Paulo, 2016.
167 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1.Adaptive Filters 2.Filters Combinations 3.Systems Identification 4.Mean Square Analysis 5.Energy Conservation Relation I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II.t.

"*La simplicité est la complexité résolue.*"

CONSTANTIN BRANCUSI

To Professor Fioravante Ferro, Dr., for teaching me so much about the virtue of living.

To Professor Lydia Semenow, Dra., for making me understand that things look really harder when they are being dealt with the wrong way.

To Mrs. Maria Elena, Mr. Paulo Lopes and their children Ana, Karina and Cassio for always supporting me and welcoming me in their family.

To Aline. For being around and supporting me unconditionally all the time, even suspecting that my adaptive filters were driving me irredeemably crazy.

ACKNOWLEDGEMENTS

I certainly owe so much to the University of São Paulo – USP, where I could not only learn a bunch of interesting things about adaptive filtering – and earn a doctor's degree for that – but also meet some of the most fascinating people I ever met in my life, including many bright minds I've had the privilege of getting acquainted with. Of course, that would be impressive by itself, but I was also blessed with so much more in the classes of Fundamentals of Neuroscience, where I met my wife Aline. Thank you so much for that hard time you gave us in your classes on membrane physiology, Professor Kohn!

I have to mention that nothing of that would happen without the influence of my friend and thesis supervisor, Professor Cassio Lopes, Dr., who was the one that encouraged me to challenge what I thought it was the natural order of the things and face a PhD program at USP. By his hands, I was capable of concluding successfully my thesis. Nothing bad, professor! Oliver Wendell once said that a mind that is stretched by a new experience can never go back to its old dimensions. Well, while I am still confused about the old dimensions part, I can definitely say I've faced a dramatic new experience by pursuing my PhD under your advices. Thank you, Cassio!

To live in a such great city as São Paulo can be traumatic for someone who, like me, lived his entire life in the small island of Florianópolis. So, I have also to mention that I probably would have not survived to that experience without the assistance of my physician friends: Rodrigo Massaud, M.D., whose rock-solid knowledge on Neurology not only allowed him to correctly diagnose my broken shoulder but also my gastroesophageal reflux, both caused by the healthy lifestyle I learned in São Paulo, his wife Ana Scalzaretto, M.D., who helped to put that shoulder in the place it was supposed to be and also taught me how to turn antihistamine into a reasonable anxiolytic and, the last but certainly not the least, the terrific Sabino, M.D., whose generosity I abused so many times to stay healthy enough to lead my research (God, I wish I knew how the Paulistanos keep body and soul together in São Paulo!). Thank you so much, guys!

Regarding the thesis itself, I am thankful to the members of my qualifying committee, Professors Magno da Silva, Dr. and Phillip Burt, Dr. for the advices and corrections made in the original manuscript of my qualifying thesis.

I am also indebted to my bosses at Celesc, Mr. Maurílio Santos e Mrs. Cristiane Lacerda, who agreed to grant me a non-paid leave from my work, which allowed me to pursue my PhD. My detractors keep saying that both actually got happy to get rid of me for some time, but I know it is lie!

Speaking of Celesc, I also want to thank Miss Daniela Rosa and Mr. Gustavo de Souza, who gave me a hand – literally – when I left Florianópolis for São Paulo. I definitely owe you guys a couple of beers!

ABSTRACT

The low complexity of IIR adaptive filters (AFs) is specially appealing to real-time applications but some drawbacks have been preventing their widespread use so far. For gradient based IIR AFs, adverse operational conditions cause convergence problems in system identification scenarios: underdamped and clustered poles, undermodelling or non-white input signals lead to error surfaces where the adaptation nearly stops on large plateaus or get stuck at sub-optimal local minima that can not be identified as such *a priori*. Furthermore, the non-stationarity in the input regressor brought by the filter recursivity and the approximations made by the update rules of the stochastic gradient algorithms constrain the learning step size to small values, causing slow convergence. In this work, we propose IIR performance enhancement strategies based on hybrid combinations of AFs that achieve higher convergence rates than ordinary IIR AFs while keeping the stability.

RESUMO

A baixa complexidade dos filtros adaptativos (FAs) IIR é atrativa para aplicações em tempo real, mas certos inconvenientes têm impedido sua ampla utilização até agora. Para os FAs baseados no gradiente descendente, condições operacionais adversas suscitam problemas de convergência em cenários de identificação de sistemas: pólos subamortecidos ou agrupados, submodelagem ou sinais correlacionados originam superficies de erro onde a adaptação desacelera em grandes planícies ou para em mínimos locais sub-ótimos que não podem ser identificados como tais *a priori*. Além disso, a não-estacionaridade do regressor de entrada causada pela recursividade do filtro e as aproximações feitas pelas regras de atualização dos algoritmos de gradiente estocástico restringem o passo de aprendizado a valores pequenos, retardando a convergência. Neste trabalho, propomos estratégias de aprimoramento de desempenho baseadas em combinações híbridas e estáveis de FAs que alcançam taxas de convergências mais altas do que FAs IIR comuns.

CONTENTS

List of Figures **iv**

Acronyms **vii**

List of Symbols **8**

1 Introduction **11**

2 IIR Adaptive Filtering Principles **18**

2.1 Systems Modelling 18

2.2 Some Remarks on IIR Realizations 20

2.3 System Identification Setup 23

2.4 Systems Identification Models 28

2.4.1 Output Error (OE) and Mean Square Output Error (MSOE) 29

2.4.2 Equation Error (EE) and Mean Square Equation Error (MSEE) 32

2.4.3 Output Error Versus Equation Error 36

2.5 State Space Description 38

2.6 Stability of Time Varying Systems 41

2.7 Exponential Stability and the Direct Form 44

2.8 Adaptation 46

2.8.1 FIR LMS Recursion 46
2.8.2 IIR OE LMS Recursion (LMSOE) 47
2.8.3 Normalized IIR LMS (N-LMSOE) 51
2.9 Stability Bounds on the Step Size μ 53

3 Parallel Combinations of Adaptive Filters 55
3.1 Initial Considerations 55
3.2 Supervisor Design 57
3.3 Normalized Supervisor Design 59

4 Hybrid Convex Combinations with IIR Adaptive Filters 62
4.1 Initial Considerations 62
4.2 T-OE: Transversal Output-Error Combinations 63
4.2.1 Considerations on the Mapping Functions $\mathcal{F}$ 66
4.2.2 Balanced Model Reduction (BMR) 67
4.2.3 The Padè Approximants Method (PAM) 71
4.2.4 FIR $\rightarrow$ IIR Mapping Assessment 75
4.2.5 Cycle Length Estimate 83
4.2.6 Computational Complexity of the T-OE 86
4.2.7 Experimental Results on T-OE 88
4.3 IIR-IIR Adaptive Convex Combinations 97
4.3.1 Components for the IIR-IIR Adaptive Combinations . . . 98
4.3.2 E-OE: Equation Error–OE Combinations 99

4.3.3 P-OE: Pseudo Linear Regression–OE Combinations 102

4.3.4 Experimental Results on IIR-IIR Combinations 106

Simulations with the E-OE Combinations 106

Simulations with the P-OE Combinations 108

4.3.5 Limitations of the IIR-IIR Combinations 113

5 Performance Analysis **119**

6 Conclusions and Comments **131**

Appendices **134**

A Composite Filters **135**

A.1 The Composite Regressor Algorithm (CRA) 135

A.2 The Combined Square Error Algorithm (CSE) 137

A.3 Some Remarks on Composite Structures 139

B Considerations on Mismodelling **141**

B.1 Algebraic Perspective on Undermodelling 143

C Error Surface Topology and Performance Issues **148**

D Rational Approximation in the Hardy Space $\mathcal{H}^2$ **151**

References **157**

LIST OF FIGURES

1 An Adaptive Filter in a System Identification Scenario 12

2 An Adaptive Filter in a System Identification Scenario 23

3 3rd Order Direct Form IIR 24

4 Output Error Formulation 29

5 Equation Error Formulation (A) 33

6 Equation Error Formulation (B) 33

7 Filters Combination in the a Systems Identification Configuration 56

8 The Sigmoidal Activation Function for the Mixing Factor $\lambda(i)$ . . 58

9 The T-OE combination 64

10 Frequency Response of the Random Plant (Scenario 1) 79

11 Performance Assessment - Transfer Error (Scenario 1) 79

12 Frequency Response of the Random Plant (Scenario 2) 80

13 Performance Assessment - Transfer Error (Scenario 2) 80

14 Frequency Response of the Random Plant (Scenario 3) 81

15 Performance Assessment - Transfer Error (Scenario 3) 81

16 Convergence Time X Cycle Length for the Notch Scenario 83

17 Linear Approximation for The FIR Guide Identifying The Notch . 84

18 Frequency Responses of the Test Scenarios 88

19 Impulse Responses of the Test Scenarios 89

20 Poles-Zeros Diagram of the Test Scenarios 89

21 MSE - Notch (Stationary Scenario) 92

22 EMSE - Notch (Stationary Scenario) 92

23 EMSE - Notch (Stationary Scenario, zoom) 93

24 MSE - Bandpass (Stationary Scenario) 94

25 EMSE - Bandpass (Stationary Scenario) 94

26 EMSE - Bandpass (Stationary Scenario, zoom) 95

27 MSE - Notch (Non-Stationary Scenario) 96

28 EMSE - Notch (Non-Stationary Scenario) 96

29 MSE - Notch (Non-Stationary Scenario) 96

30 EMSE - Notch (Non-Stationary Scenario) 97

31 Hybrid Combination of IIR Adaptive Filters 98

32 EMSE - Butterworth SNR = 30 dB (Non-Stationary Scenario) . . 107

33 EMSE - Butterworth SNR = 40 dB (Non-Stationary Scenario) . . 108

34 EMSE - Notch SNR = 30 dB (Non-Stationary Scenario) 109

35 EMSE - Notch SNR = 30 dB (Non-Stationary Scenario) 110

36 EMSE - Butterworth SNR = 30 dB (Non-Stationary Scenario) . . 111

37 EMSE - Butterworth SNR = 30 dB (Non-Stationary Scenario) . . 111

38 EMSE - Butterworth SNR = 60 dB (Non-Stationary Scenario) . . 112

39 EMSE - Butterworth SNR = 60 dB (Non-Stationary Scenario) . . 113

40 Long Tailed Impulse Response 114

41 E-OE and P-OE Combinations Identifying a Long-Tailed Plant . . 115

42 T-OE Endowed with a Long Guide and BMR-Based Transfers Identifying a Long-Tailed Plant 117

43 The Separation Principle 125

44 The Separation Principle (zoom) 125

45 EMSE for the Butterworth Scenario 129

46 EMSE for the Notch Scenario 129

47 EMSE for the Butterworth Scenario 130

48 Multimodal Error Surface 143

49 Countour Plot of a Multimodal Error Surface 144

ACRONYMS

AF	Adaptive Filter.
BMR	Balanced Model Reduction.
EE	Equation Error.
EMSE	Excess Mean Square Error.
E-OE	Equation Error-Output Error (filter combination).
FIR	Finite Impulse Response.
IIR	Infinite Impulse Response.
LMS	Least Mean Square.
LMSOE	Least Mean Square - Output Error (Output Error version of the standard FIR LMS).
MSE	Mean Square Error.
N-LMS	Normalized Least Mean Square.
N-LMSOE	Normalized Least Mean Square - Output Error (Output Error version of the standard FIR N-LMS).
OE	Output Error.
PAM	Paddè Approximants Method (Method of Prony).
P-OE	PLR-Output Error (filter combination).
PLR	Pseudo-Linear Regression.
SPR	Strictly Positive Real (or Strictly Positive Realness).
T-OE	Transversal-Output Error (filter combination, with "transversal" denoting a regular transfersal FIR AF).

LIST OF SYMBOLS

^{o} Superscript that identifies measures belonging to or produced by an unknown plant, as opposed to those associated to adaptive filters. For example, $H^o(z)$ identifies the rational transfer function of an unknown plant in opposition to that of an AF, denoted by $H(z)$.

$a_k(i)$ k-th feedback coefficient of an AF. $A(z,i) = \sum_{k=1}^{M} a_k(i) z^k$.

$A(z)$ Polynomial that defines the denominator (i.e., the poles) of a rational transfer function $H(z) = \frac{B(z)}{1-A(z)}$.

A_i State transition matrix that describes the dynamics of a given system in the state space. $\mathcal{W}_{i+1} = \mathsf{A}_i \mathcal{W}_i + \mathsf{b}_i u(i)$.

$b_k(i)$ k-th feed-forward coefficient of an AF. $B(z,i) = \sum_{k=0}^{M} b_k(i) z^k$.

$B(z)$ Polynomial that defines the numerator (i.e., the zeros) of a rational transfer function $H(z) = \frac{B(z)}{1-A(z)}$.

b_i State space variable that describes the dynamics of a given system represented in the state space. $\mathcal{W}_{i+1} = \mathsf{A}_i \mathcal{W}_i + \mathsf{b}_i u(i)$.

c_i State space variable that describes the output of a given system represented in the state space. $y(i) = \mathsf{c}_i \mathcal{W}_i + \mathsf{d}(i) u(i)$.

$d(i)$ Output of an unknown plant as perceived by the outside environment in a system identification scenario; i.e., $d(i) = y^o(i) + v(i)$.

$\mathsf{d}(i)$ State space variable that describes the output of a given system represented in the state space. $y(i) = \mathsf{c}_i \mathcal{W}_i + \mathsf{d}(i) u(i)$.

$e(i)$ Estimation error committed by an AF at the instant i.
$e(i) = d(i) - y(i)$.

$e_a(i)$ *A priori* estimation error committed by an AF at the instant i.
$e_a(i) = y^o(i) - y(i) = x_i^o w^o - x_i w_{i-1}$.

$\bar{e}_a(i)$ Filtered *a priori* estimation error committed by an AF at the instant i. $\bar{e}_a(i) = \phi_i \tilde{w}_{i-1}$.

$e_p(i)$ *A posteriori* estimation error committed by an AF at the instant i.
$e_p(i) = x_i^o w^o - x_i w_i$.

$\bar{e}_p(i)$ Filtered *a posteriori* estimation error committed by an AF at the instant i. $\bar{e}_p(i) = \phi_i \tilde{w}_i$.

E Expectancy (expected value) .

ϕ_i Filtered regressor of an Output-Error AF.
$\phi_i = \frac{x_i}{1-A(z)} = x_i + \sum_{k=1}^{M} a_k \phi_{i-k}$.

J Cost function used to optimize the coefficients w_i with respect to w^o.
$J = J(w_{i-1}) = \mathrm{E}\ e^2(i)$.

M Order of a given AF as defined by the order of the underlying rational transfer function $H_n(z)$.

μ Learning step size (or adaptation constant) .

$u(i)$ Input signal that feeds a given unknown plant and/or an adaptive filter in a system identification scenario.

u_i Regressor of a FIR AF.
$u_i = [u(i)\ u(i-1) \cdots\ u(i-M-1)\ u(i-M)]$.

$v(i)$ Aditive noise, normally assumed to be zero-mean, white and Gaussian, that contaminates the output signal $y^o(i)$ produced by an unknown plant in a system identification scenario. $d(i) = y^o(i) + v(i)$.

w_i Weight vector that stores the coefficients of an AF at the instant i.
$w_i = [a_1(i)\ a_2(i) \cdots a_M(i)\ b_0(i)\ b_1(i)\ b_2(i)\ \cdots b_M(i)]$.

$\tilde{w}_i$ Weight error vector that gives the misalignment between w_i and w^o.
$\tilde{w}_i = w^o - w_i$.

$\mathcal{W}_i$ State vector of a given system represented in the state space .

x_i Regressor of an Output-Error AF.

$x_i = [y(i-1)\ y(i-2) \dots y(i-M)\ u(i)\ u(i-1) \cdots u(i-M)]$.

$x_{\mathrm{EE},i}$ Regressor of an Equation-Error AF.

$x_i = [d(i-1)\ d(i-2) \dots d(i-M)\ u(i)\ u(i-1) \cdots u(i-M)]$.

x_i^o Regressor of an unknown IIR plant.

$x_i^o = [y^o(i-1)\ y^o(i-2) \dots y^o(i-M)\ u(i)\ u(i-1) \cdots u(i-M)]$.

$y(i)$ Output signal produced by an adaptive filter in a system identification scenario.

z Operator of the Z-Transform, also understood as the unit delay operator; e.g., $H(z) = \sum_{k=0}^{\infty} h_k z^k$ (the Z-Transform) and $H(z)u(i) = \sum_{k=0}^{\infty} h_k z^k u(i) = \sum_{k=0}^{\infty} h_k u(i-k)$ (delay operator in the mixed notation).

1 INTRODUCTION

Adaptive filters (AFs) are versatile tools that have been used as real time solutions for several applications, such as [1–5]:

1. System Identification
2. Equalization and Inverse Modelling
3. Adaptive Linear Prediction
4. Adaptive Autoregressive Spectrum Analysis
5. Echo Cancellation
   (a) Network Echo Cancellers
   (b) Acoustic Echo Cancellers
6. Adaptive Interference Cancelling

Once the main concepts in Adaptive Filtering may be properly studied via a system identification perspective, it is focused in this work through scenarios like the one depicted in Fig. 1. In this figure, $u(i)$ is the input signal, $H^o(z)$ is the unknown plant [1] which produces the output $y^o(z)$, $v(i)$ is a Gaussian additive signal that models any external disturbance on $y^o(z)$ such as thermal noise or anything else that causes systematic error measurements and $d(i)$ is the output of the unknown plant as perceived by the outside environment.

[1] In this work, we employ the terms *system* and *plant* interchangeably

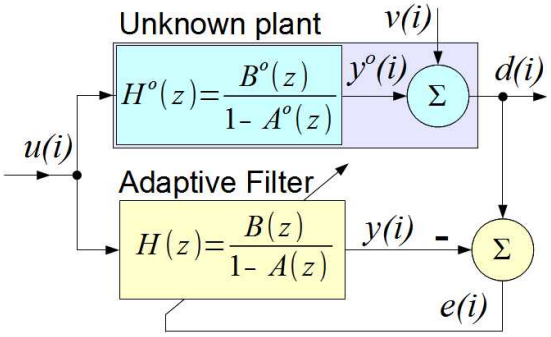

Figure 1: An Adaptive Filter in a System Identification Scenario

In order to identify the system $H^o(z)$, an adaptive algorithm adjusts the parameters of the AF $H(z)$ iteratively, making its estimates $y(i)$ approximate $d(i)$ as close as possible; i.e., minimizing the *estimation error* $e(i)$ defined as

$$e(i) = d(i) - y(i), \tag{1.1}$$

according to a methodical approach [1, 4, 5]. This way, the adaptation becomes an optimization problem that seeks to minimize some properly chosen function $J\big(e(i)\big)$ and then achieve identification [2] . Several cost functions are used to this end, but the most employed in the literature is the *Mean Squared Error* (MSE) given by

$$J_{MSE} = \mathrm{E}\, e^2(i), \tag{1.2}$$

which owes its popularity to some nice mathematical properties as continuous differentiability [6]. Given that the cost function drives the adaptation, the analysis of the MSE and its plots against the AF parameters (known as *error surfaces*) gives valuable insights to understand convergence issues.

As any digital filter, AFs are characterized by a transfer function that determines their impulse responses, so they can be either FIR (Finite Impulse Response) or IIR (Infinite Impulse Response) AFs [7–9]. Strictly speaking, IIR

[2] Here, we tacitly admit that the orders of both $H^o(z)$ and $H(z)$ are the same. If they are not, optimal estimation can still be possible, but identification is not (at least in a literal sense).

systems can be correctly identified only by an IIR AF because a FIR one can at the best yield a truncated version of the original impulse response. In fact, IIR AFs tend to require fewer coefficients than comparable FIR AFs, particularly when the underlying system they are trying to model has sharp resonances [8,10]. For example, whereas an FIR AF may need some thousands of taps in applications like channel equalization or echo cancellation, an equivalent IIR AF could require as few as 200 taps to achieve the same performance, which makes IIR adaptive filtering appealing [4, 11–16].

However, two major downsides have traditionally prevented the large scale usage of IIR AFs so far: stability and convergence rate [1,17–19]. Stability concerns materialize when the adaptive algorithm allocates some of its poles outside the unit circle, which could cause the AF to diverge [20]. Such concerns have already been addressed by the use of non-expensive checks as the Schurr recursion [12], which detect and handle unstable updates. Also, the normalized lattice first devised by Gray et al. [21] and later improved by Rodriguez-Fonollosa [22] is internally bounded by trigonometric functions and it is intrinsically stable [3] [12,23]. So, stability is no longer an issue for IIR adaptive filtering.

Speed of convergence remains an open problem though [1, 4, 11, 19, 24], with IIR AFs sometimes exceeding the typical convergence times of FIR AFs by orders of magnitude [10]. Unlike the stability issues, this can not be solved by merely changing the filter structure, although a given IIR realization could show improved performance against another under specific conditions [13] [4]. In op-

[3] In the adaptive filtering literature, to state that the normalized lattice is intrinsically stable is common sense. The fact is that the lattice does require stability tests to run smoothly; however, these tests turn out to be quite simple (unlike in the direct form case, where the stability tests depend upon the zeros of a polynomial). Indeed, all the normalized lattice requires to become exponentially stable is to constrain its rotation angles θ_k such that $0 < \theta_k < \frac{\pi}{2}$, allowing one to conclude that stability is *structurally induced* in this case [12, Theorem 6.6].

[4] Any constant gain IIR formulation such as the RPEM (Recursive Prediction Error Method), Steiglitz-McBride method, SHARF (Simple Hyperstable Adaptive Recursive Filter) or the PLR (Pseudo Linear Regression) is prone to very slow convergence under adverse conditions [10]. Even IIR Newton-based formulations, generally faster, show convergence times that may be

position to the FIR case where the only existing constraints for the parameters adaptation are those related to a proper choice of the step size, trivially derived from the eigenvalues of the input data correlation matrix, the IIR AFs parameters evolution must be further constrained for some good technical reasons as follows.

First, in the case of stochastic gradient based algorithms for the direct forms, the approximations made on the update rules become progressively coarse as the learning step size increases. As a matter of fact, the implementation of the stochastic gradient is possible only under the assumption that the denominator coefficients are slowly varying [4], as discussed in Section 2.8.2.

Second, once IIR filters are recursive, larger step sizes lead to a fast parameters evolution and, hence, increase the non-stationarity degree of the regressor signal to an extent that the AF fails to converge. In the case of the normalized lattice, this phenomenon can be managed by keeping the rotation angles into the interval $[0 \cdots \frac{\pi}{2}]$ instead of $[0 \cdots 2\pi]$ [12, Tables 7.1 to 7.5], which just emphasizes that the step size must be kept relatively small.

Third, it is possible that unstable updates become so frequent that the stability checks - if existent - prevent the parameters of the AF from evolving, which is known as *lockup* [23, 25]. Again, this could be avoided by decreasing the step size but, sometimes, the gradient magnitudes change so abruptly along the error surfaces [26] that a sufficiently small step size could make the convergence rate unacceptably slow anyway.

Finally, in algebraic terms the presence of poles amounts to decompose the transfer function of the plant onto a non-orthogonal basis defined by the poles of the filter, which may render a certain *information matrix* ill-conditioned and increase its eigenvalue spread (see Appendix C). If the spread is large enough,

much bigger than those typically found in the FIR case.

the adaptation will slow down regardless the numerical precision [5] [11], even for exact gradient algorithms with optimally tuned step sizes.

As discussed above, the benchmark we used to tell whether a given adaptive filter is *fast* or *slow* are the FIR adaptive filters, which can be significantly faster than their IIR correlates. It is that performance gap that motivates this work, where we try to answer the following question: *how to increase the adaptation speed of an IIR AF by one or more orders of magnitude while keeping the stability?*

The answer we found to that question gives this thesis its title – *Hybrid Convex Combinations for IIR Systems Identification.* In order to describe those hybrid combinations and understand their behaviour, some theoretical topics on adaptive filtering and related matters were gathered and summarized along this work. Chapter **2** shows some of the fundamentals of adaptive filtering, introduces the mixed notation, talks about filter realizations and makes a brief comment on stability aspects, explaining how all these things relate to the algorithms we developed. It also examines two mainstream IIR approaches for IIR adaptive filtering (Output Error and Equation Error) and adaptation rules, where the Output Error LMS is derived by exploiting the mixed notation within the direct-form realization.

Chapter **3** examines convex combinations of adaptive filters, covering the regular and normalized supervisor designs in the FIR adaptive filtering context.

The hybrid combinations themselves are described in Chapter **4**. Section 4.2 shows a FIR-IIR AFs combination (the "T-OE") for the system identification setup that we first introduced in [28], which shows improved performance over an ordinary IIR AF while keeping the computational complexity low by skipping

[5] This is a well known issue in linear algebra: the solution of ill-conditioned systems is extremely sensitive to small perturbations and no numerical trick can prevent that [27]. In case of stochastic gradient-based algorithms, the adaptation itself is "disturbing" due the gradient noise; so, the optimal solution will take longer to be achieved, regardless of the algorithm used.

stability checks. To achieve this goal, the FIR AF (*guide filter*) is designed to be fast and robust while the IIR AF has an arbitrarily small step size, therefore achieving both accuracy and exponential stability (refer to Sections 2.6 and 2.7).

Inspired by the cyclic feedbacks of weights originally intended for FIR combinations (see [29]), the T-OE is also endowed with conditional FIR→IIR weights transfers to improve the overall performance. We suggested in [28] that these transfers were carried via the Balanced Model Reduction algorithm (BMR) proposed by Beliczynski et. al [30], but here we also propose an alternative scheme based on Prony's method and Paddè Approximants (PAM). Both approaches have their own merits as we discuss in Section 4.2.1, but the latest is considerably simpler.

By recognizing that the weights transfer procedure may be too involving (either BMR or PAM handle matrices), we also exploited IIR-IIR filters combinations in Section 4.3. This idea resembles the *composite filters* introduced in several prior works (e.g., [18,31–36]) but it offers new possibilities such as a greater control over the switching between the components, improved performance in the transient due to a cyclic feedback mechanism and enhanced tracking capabilities.

Conceptually, the IIR-IIR filters combinations are an extension to the T-OE in which the guide filter is no longer a FIR AF, but an Equation Error (EE) or Pseudo-Linear Regression (PLR) AF (in the former case, they were called "E-OE" and in the second, "P-OE"). Compared to the T-OE, these new combinations have the obvious advantage of facilitating the weights transfers procedure as the components are structurally alike; however, they lack the possibility of changing the order of the guide filter as mismodelling can seriously deteriorate the performance of IIR AFs. Besides, both Equation Error and Pseudo-Linear Regression algorithms have known convergence issues, as detailed in Sections 4.3.2 and 4.3.3. As indicated by the experiments shown in Section 4.3.4, the best choice

is application-dependent.

Chapter **4** also contains a brief consideration on the computational complexity of the combinations in Section 4.2.6. This study had to be somewhat heterodox because the weights transfers are *cyclic* and *conditional*, which made us consider the amount of operations performed among the transfers instead of the amount of operations per iteration. Of course, this applies to the T-OE only, because the FIR→IIR weights transfers are computationally demanding; in the E-OE or the P-OE, the transfers yield a negligible overhead. We also mention some limitations of the IIR-IIR combinations when identifying plants endowed with long-tailed impulse responses in Section 4.3.5. In those cases, the E-OE and P-OE behave anomalously and can be ineffective, although the T-OE can still be used successfully by increasing the length of the FIR guides as necessary.

In Chapter **5**, we propose a theoretical contribution to the study of IIR adaptive filters: a mean-square analysis that covers the steady-state of the combinations introduced in Chapter **4** and can also be applied to stand-alone IIR AFs (normalized Output Error LMS, to be more specific). This analysis is based on the energy conservation arguments originally developed by Sayed [5] in the context of FIR adaptive filtering. Although recursive systems as the IIR AFs are inherently more difficult to analyze than their FIR counterparts, the energy conservation principles proved to fit well into the IIR adaptive filtering framework. In fact, some of the independence assumptions made by Sayed to derive the energy conservation relation are still more reasonable for the IIR AFs than for the FIR ones given that, usually, IIR AFs constrain the learning step-size to small values for stability reasons.

Finally, this work ends in Chapter **6**, where we make some final remarks and draw our conclusions about the IIR combinations introduced in Chapter **4** and the corresponding steady-state mean square analysis in Chapter **5**.

2 IIR ADAPTIVE FILTERING PRINCIPLES

2.1 Systems Modelling

The goal of a system identification application like that in Fig. 1 is to build a model able to reproduce accurately the input-output relationship of an unknown system [37]. While the very nature of this system can be entirely concealed - in fact, that relationship could not even exist in a physical sense but hold in a statistical sense only [1] - the identification is generally attainable when some reasonable premises are adopted.

In adaptive filtering, a set of parameters evolves along time according to an update strategy that seeks to optimize some predefined figure of merit as the mean squared error (MSE) (defined in Eq. 1.2) [3–5]. In order to gain some insight of how this goal may be accomplished, let us assume the existence of an externally observable, vector-valued quantity x_i that is related to $y^o(i)$ in some unknown way. Herewith, under a Bayesian perspective, the identification problem takes the form of determining the expected value of $y^o(i)$ given x_i; i.e., [1]

$$y(i) = \mathrm{E}\left(\mathbf{y^o} \mid \mathbf{x} = x_i\right) \tag{2.1}$$

In the recursive scenario of Fig. 1, the random vector $\mathbf{x}$ has to collect the

[1] Rigorously, the estimator $y(i) = \mathrm{E}(\mathbf{y^o} \mid \mathbf{x})$ does not comply with the scenario of Fig. 1 in which $y(i) = \mathrm{E}(\mathbf{d} \mid \mathbf{x})$. However, we are tacitly admitting that $v(i)$ is zero-mean and statistically independent of x_i and, hence, $\mathrm{E}(\mathbf{y^o} \mid \mathbf{x}) = \mathrm{E}(\mathbf{d} \mid \mathbf{x})$. For notational convenience, we kept the former to the detriment of the latter.

prior samples of the filter's output $\mathbf{y}$ and the current input signal $\mathbf{u}$; i.e.,

$$x_i = [y(i-1)\cdots y(i-M)\ u(i)\ u(i-1)\cdots u(i-M)], \tag{2.2}$$

where M is the order of the filter. Now, from the Bayesian estimation theory, the optimal estimator that minimizes the MSE is the *least-mean-squares estimator* (*LMSE*) [5] given by

$$y(i) = \mathrm{E}\left(\mathbf{y^o} \mid \mathbf{x} = x_i\right) = \int_{S_{y^o}} y^o(i) f_{\mathbf{y^o}|\mathbf{x}}\big(y^o(i) \mid x_i\big)\mathrm{d}y^o, \tag{2.3}$$

where S_{y^o} is the domain of the random variable $\mathbf{y^o}$ and $f_{\mathbf{y^o}|\mathbf{x}}$ is the conditional probability density function (pdf) of $\mathbf{y^o}$ given the occurrence of $\mathbf{x}$. Although this integral is not trivial in the general case, it can be shown that for a Gaussian x_i, the probability $P(\mathbf{y^o} \mid \mathbf{x})$ is a *linear combination of* x_i *with the filters coefficients* as in [5,38]

$$\begin{aligned} y(i) &= \sum_{k=0}^{M} b_k u(i-k) + \sum_{k=1}^{M} a_k y(i-k) \\ &= \sum_{k=0}^{M} b_k u(i) z^k + \sum_{k=1}^{M} a_k y(i) z^k \\ &= x_i w, \end{aligned} \tag{2.4}$$

where $w = [a_1\ a_2 \cdots a_M \cdots b_0\ b_1 \cdots b_M]^T$. [2]

As Gaussian signals rule a wide variety of physical processes, Eq. 2.4 works in several practical applications [5]. In addition, this simple model is also robust and can handle unanticipated factors as unmodeled dynamics, modeling errors, measurement noise and quantization errors to some extent [39, Chapter 20].

[2] Broadly speaking, the expansion orders does not have to be the same and $y(i) = \sum_{k=0}^{M1} b_k(i)u(i-k) + \sum_{k=1}^{M2} a_k(i)y(i-k)$ with $M_1 = M_2$ is only a possibility. However, both are often assumed to be equal to M; as Regalia [12] remarks, actually the adoption of *two* distinct orders is meaningless because the AF transfer function is characterized by only *one* degree (the so-called *McMillan degree*). The seemingly resulting incongruence may be easily solved by taking $M = max(M_1, M_2)$ and then canceling some higher order parameters accordingly.

2.2 Some Remarks on IIR Realizations

If the inner physical structure of an adaptive filter is a literal implementation of the model in Eq. 2.4, it is said that it was *realized* in the direct form. However, AFs may also be realized in different ways (lattice, normalized lattice and so on), according to the existing design requirements.

A particular realization (or *form*) defines the flow of the signal samples within the filter and what are the operations performed along that flow [9]. As different realizations perform distinct operations when implementing a given transfer function, they exhibit distinct numerical properties. This is even more noticeable with time varying systems like adaptive filters, particularly with the IIR ones.

One desirable characteristic for a given realization is stability, usually defined in the BIBO (bounded-input bounded-output) sense. As a rule, any AF is capable of operating in a stable manner if its learning step size [3] respects certain *stability bounds*. There are theoretical criteria to determine a proper step size either for IIR or FIR AFs, although the limits of the former are quite inaccurate [4] [40]. Such criteria are required for a smooth adaptation, but in the IIR case they are not sufficient (although necessary) due to the existence of poles that could be moved outside the unit border and possibly crash the AF [41].

In this context, the normalized lattice first devised by Gray [21] is surely an advance. As mentioned previously, its internal signals are bounded by trigonometric functions, which restricts the range of the location of the poles and enforces stability. Also, although the original gradient descent algorithms for the normal-

[3] The learning step size is a real constant that sets the rate at which the coefficients of the adaptive filter are updated along time. See Eq. 2.61 in Section 2.8 for more information.

[4] While the stability bounds of the step size of a FIR AF can be trivially determined through the eigenvalues of the input signal correlation matrix [4, 5], in the IIR case the bounds on the step size are a function of the impulse response of the unknown plant, the input spectral density and the parameters of the AF itself. Hence, accurate step sizes can not be determined but only roughly approximated and it is often advisable to decrease them by a factor of 10 or more [12].

ized lattice were somewhat complex, featuring a complexity of $O(M^2)$ [4], more recently Rodriguez-Fonollosa [22] provided an efficient algorithm with a complexity of $O(M)$, making it *asymptotically* comparable to the direct form [12, 22].

Moreover, the lattice endures the finite precision effects found on fixed-point arithmetics or short word length hardwares [42], whereas the direct form suffers from roundoff errors accumulation in the state vector loop [5] [23, 44, 45] and may experience quantization induced oscillations known as *limit cycles* [46, 47] [6] [7].

While all the arguments above seem to promote the normalized lattice to the detriment of the direct form, the fact is that the latter continues preponderant. This could simply indicate that designers tend to be conservative by refusing to adhere in large scale to the lattice as claimed by Regalia [12], but there are practical reasons that make the direct form more appealing as well.

First, the 32 and 64 bits wordlength standards for current CPUs allow the implementation of recursive filters in the direct form with no limit-cycles [47]. As a matter of fact, the limit cycle behavior can be *completely avoided even with much shorter wordlengths* by using floating-point arithmetics [48], which virtually encompasses all modern CPUs, including low-profile embedded systems.

Second, there are no consistent reports in the literature about the superiority of the normalized lattice in terms of steady state error or speed of convergence [1, 11, 19]. For instance, Burt et al [15] showed that the convergence rate

[5] This is highlighted with bigger values of M, which explains why the numerical sensitivity of higher order structures decreases when they are built by associating low-order sections (e.g., second order sections or *biquads*) [12]. Such an issue is recurrent in numerical analysis, with the polynomial roots being affected by the polynomial-coefficient round-offs (e.g., the Wilkinson's polynomial [43, *The Perfidious Polynomial*]). Such sensitivity tends to increase when the roots are clustered and to decrease when they are spread out in the complex plane, justifying the option for factoring the filter transfer function into series and/or parallel second-order sections.

[6] Limit cycles or *multiplier roundoff limit cycles* require recursion to exist and result from the nonlinearity associated with rounding (or truncating) internal filter calculations [1].

[7] Although both floating-point and fixed-point binary arithmetics feature finite word length, the former has a very large dynamic range even for small exponent wordlengths which makes the signal-to-quantization-noise ratio (SQNR) nearly constant over the entire dynamic range [48].

of the lattice could outperform that of the direct form when identifying all-pass functions with clustered poles, but this superiority does not extend to different configurations. As Fan [11] remarks, locally most lattice algorithms do not offer much improvement over the direct form ones and underdamped low frequency poles are known to prevent some algorithms from converging or make the convergence extremely slow in spite of their theoretical properties under ideal conditions. In these cases, the slow convergence of the direct form is not necessarily related to its alleged poor numerical properties, but to the large eigenvalue spread of its information matrix [8] *even in infinite precision* [11].

Third, some caveat is needed when analyzing the complexity of an algorithm. The big O notation defines an *asymptotic* bound for the amount of resources required by an algorithm to run when $M \to \infty$. As that bound does not have to have a practical meaning, it could lead to a coarse perception of the resources actually involved and should not be taken literally. Hence, it is somewhat abusive to state that both direct and lattice forms feature *asymptotically* the same complexity once the later executes about four times more operations per iteration than the former.

Finally, if the direct form is powered by an algorithm whose step size is carefully designed (i.e., if it is *small enough*), the need for stability checks fades [9] . In the general case, they remain crucial for a direct form IIR AF once its poles may be allocated outside the unit border whatever the step size is. Note, however, that the concern is not to casually place the poles outside the unit border once

[8] The information matrix is the covariance matrix associated to the input vector (sometimes called *information regressor* [3]) that feeds the adaptive filter. For FIR filters, it is simply the covariance of the input signal but, for IIR filters, it is the covariance of a vector that gathers both the input and prior output signals.

[9] Unlike in the FIR adaptive filtering, where the step size trades convergence speed for accuracy at the steady state, in IIR adaptive filtering a perverse effect may arise: the feedback loop injects a certain non-stationarity degree into the input so that bigger steps could inadvertently slowdown the convergence in some cases. Therefore, smaller step sizes could not only enforce stability but also improve the steady state *and incidentally increase the convergence rate* [12].

the gradient tend to attract them back to the stability region, but to keep them there enough time to make the AF diverge [49] as detailed in Section 2.6.

2.3 System Identification Setup

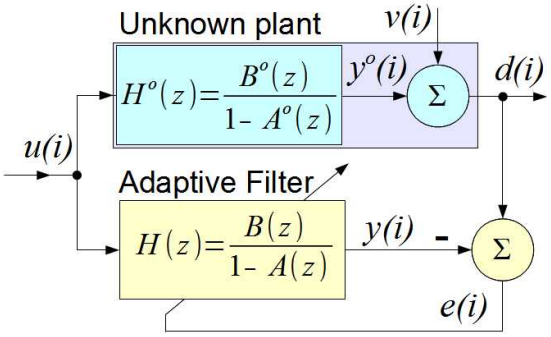

Figure 2: An Adaptive Filter in a System Identification Scenario

Let us consider again the IIR system identification scenario, shown in Fig. 2. The unknown system denoted by $H^o(z)$ is described by the difference equation [8]

$$y^o(i) = \sum_{k=0}^{M} b_k^o u(i-k) + \sum_{k=1}^{M} a_k^o y^o(i-k), \tag{2.5}$$

where $y^o(i)$ is the output, M is the system order, $u(i)$ is the input signal, $\{b_k^o\}_{k=0}^{M}$ and $\{a_k^o\}_{k=1}^{M}$ are respectively the feedforward and feedback coefficients sets ($a_0^o \triangleq 1$), and the sign ^{o} (*naught*) identifies the measures produced by or belonging to the unknown plant.

Usually, the output $y^o(i)$ is measured in the presence of an additive background noise $v(i)$, so that the actual measured system output $d(i)$ is given by

$$d(i) = y^o(i) + v(i), \tag{2.6}$$

where $v(i)$ is normally assumed to be zero-mean and statistically independent of both $y^o(i)$ and $u(i)$. $u(i)$ feeds the unknown system $H^o(z)$ as well as the adaptive filter and is often thought as a white process but, in some applications like active

noise control, can be an arbitrary process.

The adaptive filter seen in Fig. 2 ($H(z)$) implements the difference equation

$$y(i) = \sum_{k=0}^{M} b_k(i-1)u(i-k) + \sum_{k=1}^{M} a_k(i-1)y(i-k), \tag{2.7}$$

where $\{b_k(i-1)\}_{k=0}^{M}$ and $\{a_k(i-1)\}_{k=1}^{M}$ are respectively its feedforward and feedback coefficients sets at instant $i-1$ ($a_0(i-1) \triangleq 1 \; \forall \; i$), which are iteratively adjusted by a training rule in order to make the output $y(i)$ as close as possible to $d(i)$. For simplicity, it is assumed that the order M holds for both AF and plant (*sufficient modelling*) but in the general case the orders could be different.

A literal interpretation of Eq. 2.7 yields the *direct form* filtering structures as that of Fig. 3 for $M = 3$. Some authors remark that the direct form is not so resilient to the finite-precision effects as other alternative realizations and may be too sensitive to round-off errors in the coefficients [8, 12]. However, as discussed in Section 2.2, those drawbacks become a matter of concern on fixed point arithmetics only and will not be taken into the consideration in this work.

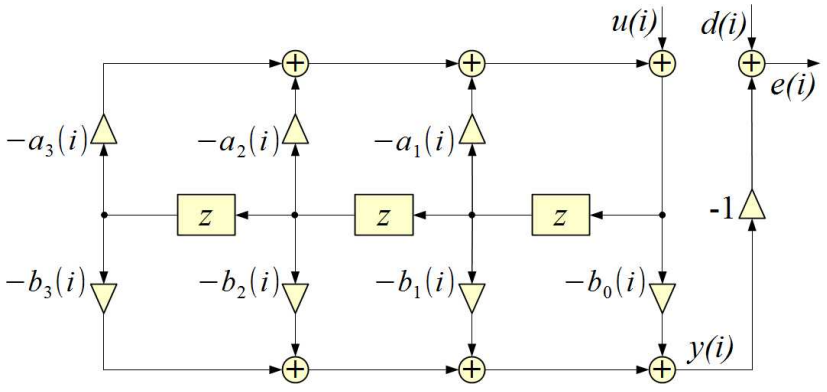

Figure 3: 3rd Order Direct Form IIR

Depending on the realization or the specific parametrization, the adaptive filter is subject to some convergence problems caused by multimodal performance surfaces (with several local minima) and saddle points (which deteriorates the

convergence rate) [12] [10] . In this work though, such issues do not appear because it considers scenarios with sufficiently modelled adaptive filters and broadband input signals $u(i)$, setups in which the performance surfaces are unimodal as demonstrated by Soderstorm and Stoica [50] and Fan and Nayeri [51] (refer to Appendix B to more details on mismodelling aspects).

There are two useful ways to represent and handle the filter structures and the related equations, the *mixed notation* [12] (also *delay operator* [20, 25, 50, 52, 53], *backward delay operator* [54] or *polynomial operator* [4] notation) and the *vector notation* [4, 5]. Consonant to some works in which these notations are used interchangeably to emphasize different aspects of certain analyses (e.g., [25]), here we keep both [11].

In the mixed notation, the key point is to mix the time domain i and the Z domain; for instance,

$$y(i) = \sum_{k=-\infty}^{+\infty} h_k u(i-k) = \sum_{k=-\infty}^{+\infty} h_k z^k u(i) \rightarrow y(i) = H(z)u(i), \tag{2.8}$$

which allows to denote a convolution in a compact mode. Here we adopt the notation suggested by Regalia in [12], where the unit delay operator is represented by z instead of z^{-1} as this choice simplifies the derivation of the mean square measures used to analyze the IIR AFs. Herewith, the signs of the powers of the Z-Transform are the opposite of the conventionally adopted; i.e., we use

$$H(z,i) = \sum_{k=-\infty}^{+\infty} h_k(i) z^k$$

[10] An example of this situation is an IIR AF trying to model a FIR plant, because several parameters sets could make the impulse response of the AF close to the original one. Another example is the grouping of low-order sections in cascade or in parallel: upon swapping the coefficients of a section with another, the resulting transfer function does not change. Even if the cost function of the equivalent direct form has an unique minimum, the allowable permutations between the sections of the cascade/parallel form will lead to multiple equally good minima.

[11] Although the mixed notation of Regalia [12] and the aforementioned correlates are equivalent, sometimes different authors use them in slightly distinct manners. In any case, note that such a notation is just a convenient algebraic trick that requires some caveat to be used.

instead of

$$H(z,i) = \sum_{k=-\infty}^{+\infty} h_k(i) z^{-k}$$

Besides being the operator of the Z Transform, z is also understood as the standard delay operator and, as such, a signal $u(i)$ delayed by k samples can be represented either as $u(i)z^k$ or $u(i-k)$. Such a flexibility is extremely convenient when deriving several results along this work.

By using the mixed notation, Eq. 2.7 would evolve to

$$\begin{aligned} y(i) &= \sum_{k=0}^{M} b_k(i-1)u(i)z^k + \sum_{k=1}^{M} a_k(i-1)y(i)z^k \\ &= u(i)\sum_{k=0}^{M} b_k(i-1)z^k + y(i)\sum_{k=1}^{M} a_k(i-1)z^k \end{aligned} \tag{2.9}$$

Now, let $B(z,i-1)$ and $A(z,i-1)$ the Z transforms of the parameters sets $\{b_k(i-1)\}_{k=0}^{M}$ and $\{a_k(i-1)\}_{k=1}^{M}$, defined as

$$B(z,i-1) = \sum_{k=0}^{M} b_k(i-1)z^k \tag{2.10}$$

$$A(z,i-1) = \sum_{k=1}^{M} a_k(i-1)z^k, \tag{2.11}$$

By replacing $B(z,i-1)$ and $A(z,i-1)$ in Eq. 2.9, the output $y(i)$ can be represented as

$$y(i) = u(i)B(z,i-1) + y(i)A(z,i-1) \tag{2.12}$$

$$y(i)\big(1 - A(z,i-1)\big) = B(z,i-1)u(i) \tag{2.13}$$

$$y(i) = \frac{B(z,i-1)}{1-A(z,i-1)}u(i) \triangleq H(z,i-1)u(i) \tag{2.14}$$

Accordingly, the plant output $y^o(i)$ given in Eq. 2.5 can be rewritten as

$$y^o(i) = H^o(z)u(i) = \sum_{k=0}^{M} b_k^o u(i-k) + \sum_{k=1}^{M} a_k^o y^o(i-k)$$

$$= \sum_{k=0}^{M} b_k^o u(i) z^k + \sum_{k=1}^{M} a_k^o y^o(i) z^k = u(i) \sum_{k=0}^{M} b_k^o z^k + y^o(i) \sum_{k=1}^{M} a_k^o z^k$$
$$= u(i) B^o(z) + y^o(i) A^o(z) = \frac{B^o(z)}{1 - A^o(z)} u(i) \tag{2.15}$$

The vector notation works in the time domain only and collects the signals into vectors [12] such that

$$u_i = [u(i)\ u(i-1)\ u(i-2)\ \cdots\ u(i-M)] \tag{2.16}$$

$$y_i = [y(i-1)\ y(i-2)\ \cdots\ y(i-M)], \tag{2.17}$$

whereas its coefficients are represented by the column vectors

$$b_{i-1} = [b_0(i-1)\ b_{1-1}(i-1)\ \cdots\ b_M(i-1)]^T \tag{2.18}$$

$$a_{i-1} = [a_1(i-1)\ a_{2-1}(i-1)\ \cdots\ a_M(i-1)]^T \tag{2.19}$$

Now, by collecting the coefficients vectors b_i and a_i into a single column vector, the *parameters vector* w_i emerges as

$$w_{i-1} = \begin{bmatrix} a_{i-1} \\ b_{i-1} \end{bmatrix} \tag{2.20}$$

Similarly, the *regressor* x_i is defined by grouping the signal vectors u_i and y_i in a single row vector; i.e.,

$$x_i = [y(i-1)\ \cdots\ y(i-M)\ u(i)\ \cdots\ u(i-M)], \tag{2.21}$$

so that the output $y(i)$ can be rewritten as a simple inner product; i.e,

$$y(i) = u_i b_{i-1} + y_i a_{i-1} = x_i w_{i-1} \tag{2.22}$$

[12] Bultheel [55] calls the vector made by the polynomial or series coefficients a *stacking vector*, alluding to the fact that the convolution can be conveniently written as the product of a Toeplitz or Hankel matrix (i.e., stacked vectors shifted by one element at a time) by a vector. That explains why the factorization of such matrices is a recurrent matter in adaptive filtering.

The vector notation makes the similarities between FIR and IIR filtering more evident, posing the former as a particular case of the last when the feedbacks coefficients $\{a_k\}_{k=1}^{M}$ are zero.

Note that the *regressor* x_i defined in Eq. 2.21 contains the filter outputs $y(i-k),\ k = 1 \cdots M$, which means that a different regressor should be defined for the plant as

$$x_i^o = [y^o(i-1)\ \cdots\ y^o(i-M+1)\ u(i)\ \cdots\ u(i-M+1)] \tag{2.23}$$

In the same vein, the plant coefficients $\{b_k^o\}$ and $\{a_k^o\}$ are collected into a single plant coefficients column vector w^o ; i.e.,

$$w^o = \begin{bmatrix} a^o \\ b^o \end{bmatrix}, \tag{2.24}$$

where $b^o = [b_0^o\ b_1^o\ \cdots b_M^o]^T$ and $a^o = [a_1^o\ a_2^o\ \cdots a_M^o]^T$. So, the measured output $d(i)$ becomes

$$d(i) = y^o(i) + v(i) = x_i^o w^o + v(i) \tag{2.25}$$

2.4 Systems Identification Models

Although adaptive filters generally fit into the model of Eq. 2.14 when *operating*, they may follow a different framework when *adapting* according to the figure of merit used to optimize its coefficients. Often, the figure of merit depends upon the *estimation error* $e(i)$ defined in terms of the filter output $y(i)$ as

$$e(i) = d(i) - y(i), \tag{2.26}$$

situation in which the AF is said to follow the *Output-Error* (OE) approach. Different definitions for $e(i)$ lead to new cost functions and new algorithmic proper-

ties as in the *Equation-Error* (EE) or the *Steiglitz-McBride* (SM) approaches [13] . Once the peculiar figure of merit associated to the SM algorithm does not lead to a well-defined (or, at least, not well-understood) optimality criterion surrounding the stationary points of some cost function [12, Section 8.12] [1, Chapter 23], this work focuses the OE and EE approaches only.

2.4.1 Output Error (OE) and Mean Square Output Error (MSOE)

The output error (OE) approach relies on the recursive differences equations given in Eq. 2.14 either to run or adapt its coefficients as shown in Fig. 4.

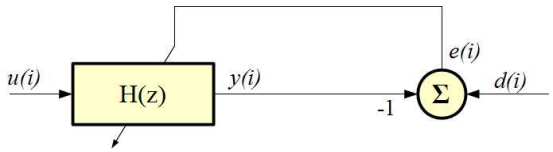

Figure 4: Output Error Formulation

Eq. 2.27 shows that in this approach the recursivity of $H(z)$ makes $e(i)$ a non-linear function of the parameters set w_{i-1} which, in the general case, can raise multimodal performance surfaces and trap the gradient descent algorithms into local minima (see Appendix B). However, as aforementioned, such issues are not faced in this work once we assume sufficient modelling and broadband input signals.

$$\begin{aligned} e(i) &= d(i) - y(i) \\ &= H^o(z)u(i) + v(i) - H(z, i-1)u(i) \\ &= \left(H^o(z) - H(z, i-1)\right) u(i) + v(i) \\ &= \tilde{H}(z, i-1)u(i) + v(i) \end{aligned}$$

[13] The literature of systems identification (e.g., [45]) considers other approaches as well, such as ARMAX, Generalized Least Squares and Box-Jenkins. However, these models are not pervasive in adaptive filtering and are generally intended to provide a more complete description about the additive noise $v(i)$ [20, 45].

$$= x_i^o w^o + v(i) - x_i w_{i-1} \tag{2.27}$$

In the literature of adaptive filtering, the name *Output Error* comes from the fact that the OE estimation error $e(i)$ is the subtraction of the "actual" output $y(i)$ from the desired output $d(i)$, as seen in Eq. 2.27 [20]. The corresponding minimization criterion for the gradient descent algorithm (refer to Section 2.8.2) is the *Mean Squared Output Error* (MSOE) $\mathrm{E}\, e^2(i)$ defined in Eqs. 2.28-2.29, where the additive noise $v(i)$ is assumed to be zero-mean (such that its variance σ_v^2 equals $\mathrm{E}\, v^2(i)$) and statistically independent from the other signals.

$$\begin{aligned}
\mathrm{E}\, e^2(i) &= \mathrm{E}\left(\left(H^o(z) - H(z, i-1)\right) u(i) \left(H^o(z) - H(z, i-1)\right) u(i) + v^2(i) \right) \\
&= \mathrm{E}\, \tilde{H}(z, i-1) u(i) \tilde{H}(z, i-1) u(i) + \sigma_v^2 \qquad (2.28) \\
&= \mathrm{E} \sum_{k=0}^{\infty} \tilde{h}_k(i-1) z^k u(i) \sum_{l=0}^{\infty} \tilde{h}_l(i-1) z^l u(i) + \sigma_v^2 \\
&= \mathrm{E} \sum_{k=0}^{\infty} \tilde{h}_k(i-1) u(i-k) \sum_{l=0}^{\infty} \tilde{h}_l(i-1) u(i-l) + \sigma_v^2 \\
&= \mathrm{E} \sum_{k=0}^{\infty} \tilde{h}_k(i-1) \sum_{l=0}^{\infty} \tilde{h}_l(i-1) u(i-l) u(i-k) + \sigma_v^2 \qquad (2.29)
\end{aligned}$$

If $\tilde{H}(z, i-1)$ is thought as a constant $\tilde{H}(z)$, then it becomes statistically independent of the input $u(i)$. As such, by taking into account that the auto-correlation function of $u(i)$ is $r_u(k-l) = r_u(l-k) = \mathrm{E}\, u(i-l)u(i-k)$, the so defined MSOE may be further elaborated as in Equations 2.30-2.32, where $*$ denotes a convolution and $\langle ., . \rangle$ is the usual inner product, whose definition is implied in Equations 2.30 and 2.31.

$$\mathrm{E}\, e^2(i) = \sum_{k=0}^{\infty} \tilde{h}_k \sum_{l=0}^{\infty} \tilde{h}_l r_u(l-k) + \sigma_v^2 \tag{2.30}$$

$$\begin{aligned}
\therefore Ee^2(i) &= \langle \tilde{h}, \tilde{h} * r_u \rangle + \sigma_v^2 \equiv \langle \tilde{H}(z), \tilde{H}(z) S_u(z) \rangle + \sigma_v^2 \qquad (2.31) \\
&= \frac{1}{j2\pi} \oint_{|z|=1} S_u(z) \tilde{H}(z) \tilde{H}(z^{-1}) \frac{1}{z} dz + \sigma_v^2
\end{aligned}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_u(e^{j\omega}) |\tilde{H}(e^{j\omega})|^2 d\omega + \sigma_v^2 \tag{2.32}$$

Besides consolidating the intuitive notion that the smaller the difference $\tilde{H}(z) = H^o(z) - H(z)$ the smaller the estimation error, Eq. 2.32 reveals that the MSOE is related to $S_u(e^{j\omega})$, the spectral density of $u(i)$ defined as the Fourier Transform of $r_u(i)$. Now, for a white $u(i)$ with power p_u, $S_u(z) = p_u$ because $r_u(k) = 0 \; \forall \; k \neq 0$. In these cases, Fan and Nayeri [51] showed that the surface defined by $\mathrm{E}\, e^2(i) = p_u \langle \tilde{H}(z), \tilde{H}(z) \rangle$ is unimodal if the modelling of the AF is sufficient (see Appendix B), which had been formerly conjectured by Stearns [26].

It is important to note that an unimodal cost $\mathrm{E}\, e^2(i)$ does not avoid convergence issues, because the feedback of IIR AFs makes $e(i)$ a non linear function of the coefficients [4, 20, 56]. A perfectly symmetrical (i.e., paraboloidal) error surface arises only when all poles are at the origin of the unit circle [14] – as the poles magnitude grow, $e(i)$ becomes increasingly non-linear and $\mathrm{E}\, e^2(i)$ distorts, assuming a nonquadratic shape on which the gradient changes abruptly [26]. As such, issues as low convergence rates and high steady state errors may arise even in the absence of multiple local minima [10] (See Appendices B and C) [15] .

Further Considerations Regarding OE

- Due to their recursive nature, algorithms based on the OE approach (as the LMSOE, defined in Section 2.8.2) are sometimes called Recursive Prediction Error Algorithm or RPE [20, 45];
- The Output Error algorithms have a degenerate form called Pseudolinear

[14] If all poles are at the origin, the filter is all-zeros and it is actually FIR. This way, as mentioned in Section 2.1, the estimation error $e(i)$ will be a strictly linear function of the filter coefficients, which leads to a paraboloidal error surface and predictable convergence properties.

[15] Underdamped poles distort significantly the error surface (particularly if they are clustered), often making it near flat right before the minima, which decreases the magnitude of the gradient descent. As a result, the adaptation may be ended prematurely to avoid huge convergence times or just because the AF looks stagnated. In both cases, decent error levels could not be reached sooner by enlarging the step size either because that increases the non-stationarity of the regressor x_i (due to the recursivity) or simply because bigger steps worsen the steady state.

Regression or PLR [45, 54] in which the non-linear dependence between $y(i)$ and $A(z)$ is neglected when the stochastic gradient is computed. As a result, the adaptation of the parameters w_i becomes an iterative linear regression problem that has to be optimized in the mean square sense [57, Section 3.3] [1, Section 23.3.2]. However, there is no way of guaranteeing that it converges to a minimum (actually, it could stabilize far from it) unless the so-called SPR (Strictly Positive Real) condition is met (refer to Section 4.3.3) [20, 25, 58];

- To make PLR converge is a goal of several works. As the SPR condition is sufficient but not necessary, these approaches try to overcome it by selecting carefully the input sequence, adopting larger adaptation gains, using system overmodelling or extra-filtering the input [20, 25, 59];
- the hyperstable algorithms can be derived from an extension of the PLR approach called *filtered error* [20, 54] [60, Chapter 7, *Analysis and Synthesis Tools for Robust SPR Discrete Systems*].

2.4.2 Equation Error (EE) and Mean Square Equation Error (MSEE)

In the Equation Error approaches, the AF is not recursive during the adaptation, implementing the difference equation shown in Eq. 2.33. It can be noted that the feedback coefficients $\{a_k\}$ are actually associated to the outputs $d(i)$ of the unknown plant, not to $y(i)$. As a result, the AF is perceived as a dual-channel FIR AF in which $B(z)$ and $A(z)$ are adapted independently like seen in Fig. 5.

$$\begin{aligned} y_e(i) &= \sum_{k=0}^{M} b_k(i-1)u(i-k) + \sum_{k=1}^{M} a_k(i-1)d(i-k) \\ &= \sum_{k=0}^{\infty} b_k(i-1)z^k u(i) + \sum_{k=1}^{M} a_k(i-1)z^k d(i) \\ &= B(z, i-1)u(i) + A(z, i-1)d(i) \end{aligned} \tag{2.33}$$

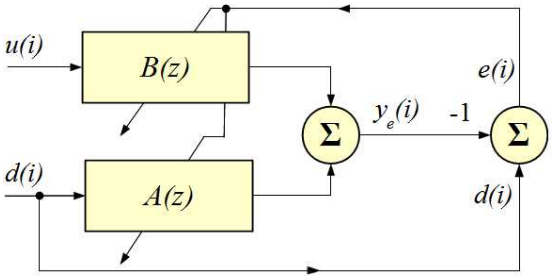

Figure 5: Equation Error Formulation (A)

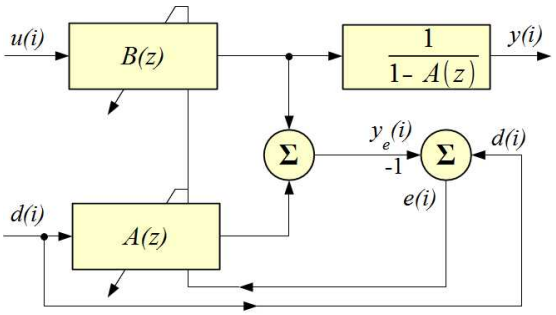

Figure 6: Equation Error Formulation (B)

The term *Equation Error* results from the fact that the EE estimation error is produced by subtracting two difference equations, $\big(1 - A(z, i-1)\big)d(i) - B(z, i-1)u(i)$, as Eq. 2.34 shows. In contrast with $y_e(i)$, which is plain linear, $y(i)$ is yielded by the rational expression $y(i) = \frac{B(z,i-1)}{1-A(z,i-1)}u(i) = H(z, i-1)u(i)$ and thus differs from $y_e(i)$ as seen in Fig. 6. This picture highlights the fact that the actual AF output $y(i)$ is available to the application whereas $y_e(i)$ is used for the adaptation only.

$$\begin{aligned} e(i) &= d(i) - y_e(i) \\ &= d(i) - B(z, i-1)u(i) - A(z, i-1)d(i) \\ &= \big(1 - A(z, i-1)\big)d(i) - B(z, i-1)u(i) \end{aligned} \tag{2.34}$$

Unlike the OE case, in the EE approaches the additive noise $v(i)$ impacts the adaptation and causes a *bias* in the AF coefficients w_i that is proportional to $\mathrm{E}\, v^2(i) = \sigma_v^2$ [20, 35, 61]. In order to better understand the causal link between the bias and the noise, we first consider the case in which $v(i) = 0$ and only then

it is shown how the Equation Error is biased by the presence of noise.

The EE Approach in The Noiseless Case

In case $v(i) = 0$, $d(i) = y^o(i)$, which means that $d(i) = H^o(z)u(i)$ (see Eq. 2.6). Herewith, Eq. 2.34 becomes

$$\begin{aligned} e(i) &= \big(1 - A(z, i-1)\big)d(i) - B(z, i-1)u(i) \\ &= \Big(\big(1 - A(z, i-1)\big)H^o(z) - B(z, i-1)\Big)u(i) \end{aligned} \tag{2.35}$$

Accordingly, the *Mean Squared Equation Error* (MSEE) is defined as

$$\mathrm{E}\, e^2(i) = \mathrm{E}\left(\Big(\big(1 - A(z, i-1)\big)H^o(z) - B(z, i-1)\Big)u(i)\right)^2 \tag{2.36}$$

Similarly to what was made with the MSOE in Equations 2.30 and 2.31, let us treat $H(z, i-1)$ as a constant $H(z)$ so that $B(z, i-1) = B(z)$ and $A(z, i-1) = A(z)$. This way, by defining $C(z) \triangleq \big(1 - A(z)\big)H^o(z) - B(z)$, Eq. 2.36 becomes

$$\begin{aligned} \mathrm{E}\, e^2(i) &= \mathrm{E}\left(C(z)u(i)\right)^2 \\ \mathrm{E}\, e^2(i) &= \mathrm{E}\left(C(z)u(i)\right)\left(C(z)u(i)\right) \\ &= \mathrm{E} \sum_{k=0}^{\infty} c_k z^k u(i) \sum_{l=0}^{\infty} c_l z^l u(i) \\ &= \sum_{k=0}^{\infty} c_k \sum_{l=0}^{\infty} c_l \,\mathrm{E}\, u(i-k)u(i-l), \end{aligned} \tag{2.37}$$

and we note that the EE estimation error $e(i)$ nullifies if $H(z)$ equals $H^o(z)$.

As in the case of the derivation of the output error in Eq. 2.30, here it is possible to apply the definition of the auto-correlation of $u(i)$ in Eq. 2.37 to further elaborate it, as shown in Eqs. 2.38-2.40.

$$\mathrm{E}\, e^2(i) = \sum_{k=0}^{\infty} c_k \sum_{l=0}^{\infty} c_l(i) r_u(k-l) \tag{2.38}$$

$$\therefore Ee^2(i) = \langle c, c * r_u \rangle \equiv \langle C(z), C(z)S_u(z) \rangle \tag{2.39}$$

$$
\begin{aligned}
&= \frac{1}{j2\pi} \oint_{|z|=1} S_u(z)C(z)C(z^{-1})\frac{1}{z}dz \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_u(e^{j\omega})|C(e^{j\omega})|^2 d\omega \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_u(e^{j\omega})|1 - A(e^{j\omega})H^o(e^{j\omega}) - B(e^{j\omega})|^2 d\omega
\end{aligned} \tag{2.40}
$$

Once $y_e(i)$ does not depend on the prior filter outputs $y(i)$ (Eq. 2.33), it becomes a *strictly linear function* of the parameters and the signal input (made of u_i and the prior values of $d(i)$). Hence, $\mathrm{E}\, e^2(i)$ will be a quadratic function of the coefficients sets $\{b_k(i)\}_{k=0}^{M}$ and $\{a_k(i)\}_{k=1}^{M}$, leading to unimodal performance surfaces, predictable convergence (no poles to concern about) and intrinsic stability. However, as aforementioned, in the presence of noise this comes at the price of biased estimates, which may lead to a poor performance if the SNR is low.

The EE Approach in The Noisy Case

If $v(i)$ is not zero, then the EE estimation error becomes as in Eq. 2.41, where $e_u(i)$ is the unbiased estimation error as defined in Eq. 2.35 for the noiseless case and $v(i)$ is filtered by $1 - A(z, i-1)$. The bias results from the two conflicting goals the AF has to cope with, namely, identify the system poles (determined by $A(z)$) and, at the same time, minimize the MSEE [20].

$$
\begin{aligned}
e(i) &= d(i) - y_e(i) \\
&= d(i) - B(z, i-1)u(i) - A(z, i-1)d(i) \\
&= \big(1 - A(z, i-1)\big)d(i) - B(z, i-1)u(i) \\
&= \big(1 - A(z, i-1)\big)\big(H^o(z)u(i) + v(i)\big) - B(z, i-1)u(i) \\
&= \big(1 - A(z, i-1)\big)H^o(z)u(i) + \big(1 - A(z, i-1)\big)v(i) - B(z, i-1)u(i) \\
&= \Big(\big(1 - A(z, i-1)\big)H^o(z) - B(z, i-1)\Big)u(i) + \big(1 - A(z, i-1)\big)v(i) \\
&= e_u(i) + \big(1 - A(z, i-1)\big)v(i)
\end{aligned} \tag{2.41}
$$

Once $v(i)$ is assumed to be independent of the input signal and zero-mean, the MSEE becomes as in Eq. 2.42. It can be seen that the power of the bias is a function of the feedback coefficients and the power σ_v^2 of the noise $v(i)$. Also, unlike in the noiseless case (Eq. 2.37), even if $H(z) = H^o(z)$, the MSEE $\mathrm{E}\, e^2(i)$ will not be zero because of the bias term $\sigma_v^2 \,\mathrm{E} \sum_{k=0}^{M} a_k^2$.

$$\begin{aligned} \mathrm{E}\, e^2(i) &= \mathrm{E}\left(e_u(i) + \big(1 - A(z)\big)v(i)\right)^2 \\ &= \mathrm{E}\, e_u^2(i) + \sigma_v^2 \,\mathrm{E} \sum_{k=0}^{M} a_k^2 \end{aligned} \tag{2.42}$$

Some works cope with the bias by using additional filtering, noise suppressing or, more recently, constrained minimization techniques but all debiasing techniques have drawbacks. An interesting and simple approach to get rid of the bias is the *monic normalization* described by Kim et al. [62], which is grounded on the perception that *the bias impacts every feedback coefficient roughly the same way.* Once a_0 (the filter coefficient related to the current output $y_e(i)$) is constant and unitary, the update rule is supposed to adapt a_k for $0 < k \leq M$ only. Now, by the time $y_e(i)$ is calculated the plant output $d(i)$ is available already, so a simple change in the update rule allows to adapt a_0 along with the remaining feedback coefficients. If those coefficients are normalized by a_0 right afterwards to enforce the monic constraint $a_0 = 1$, the bias shall be corrected as the AF adapts. Despite promising practical results though, Soderstrom [63] observed that the method is not always convergent and relies on independence assumptions that are unrealistic unless $v(i)$ is white.

2.4.3 Output Error Versus Equation Error

By comparing the OE and EE estimation error (Eqs. 2.27 and 2.41), it becomes clear that the last is a filtered version of the prior, as shown in Eq. 2.43

where, for convenience, the errors are identified by the subscripts *ee* and *oe*.

$$\begin{aligned} e_{ee}(i) &= \Big(\big(1 - A(z, i-1)\big)H^o(z) - B(z, i-1)\Big)u(i) + \big(1 - A(z, i-1)\big)v(i) \\ &= \Big(\big(1 - A(z, i-1)\big)\big(H^o(z) - H(z, i-1)\big)\Big)u(i) + \big(1 - A(z, i-1)\big)v(i) \\ &= e_{oe}(i)\big(1 - A(z, i-1)\big) \end{aligned} \tag{2.43}$$

As Section 2.8.2 discusses, the LMS derivation for the OE approach (LMSOE) assumes that the step size μ is small, otherwise the error between the actual update formula and its stochastic approximation (refer to Equations 2.72 and 2.73) is emphasized and may prevent the algorithm from converging [12]. Furthermore, a large μ makes the adapting $A(z)$ highly variable, which may disturb the recursive OE regressor to an extent that the AF stagnates. On the other hand, in the EE approach there are no further restrictions other than the stability bounds on the step size. Also, as a dual FIR configuration, it requires no stability checks.

In a nutshell, there is a clear trade-off between EE and OE approaches: while the former is generally faster though biased [16] , the latter is more accurate albeit prone to slowness. This fact inspired the *composite* IIR AFs discussed in Appendix A, which exploit the good characteristics of both approaches by merging them into a single structure (e.g., [18,31–34,36]). It should be also noted that the EE bias may not be an issue whenever the application can deal with estimation errors within certain bounds or the SNR is good, as unveiled by Eq. 2.42.

From a mathematical point of view, the linearity in the parameters offered by the EE approach is relevant. In fact, the use of linearization techniques in systems identification is recommended wherever possible [50,64,65] [17] . As afore-

[16] Note that adapting both $B(z, i)$ and $A(z, i)$ through separate FIR AFs does not necessarily make the EE approaches immune to poor performance. In fact, even with $u(i)$ white, $y^o(i)$ (and, hence, $d(i)$) will be correlated due to the recursivity of the IIR plant. If the correlation is too strong, the adaptation rate at which $A^o(z)$ is identified by the EE AF may drop considerably [5].

[17] Algebraically, linearity in parameters is a matter of choosing an appropriate basis $\{F_m(z)\}_{m=0}^{M-1}$ so that a linear parameters set $\{\theta_m^o(z)\}_{m=0}^{M-1}$ is attainable with $y(i) =$

mentioned, in situations where the bias can not be tolerated some debiasing schemes could be applied, producing good results at the expense of computational power (e.g., [66]), additional degrees of freedom (e.g., [67]) or somewhat strict assumptions (e.g., [59, 62]). However, formal analyses of performance and convergence of these methods are not always reported.

Rigorously, only the OE approaches are really IIR filtering techniques because (as previously noted) the EE methods are actually a dual-FIR configuration whose coefficients happen to fit in an IIR structure. For this reason, some authors (e.g., [12] and [20]) are reluctant to classify the EE approaches as IIR adaptive filtering. In this work, unless explicitly stated otherwise, *IIR adaptive filtering* will always refer to some OE approach.

2.5 State Space Description

Strictly speaking, IIR systems can not be represented directly through their impulse responses because that would require infinite storage. For this reason, IIR systems are usually conceived as a recursive difference equation in the time domain, a rational function in the Z domain or, alternatively, as a matrix system in the state space form. In the latter option, the recursivity is dealt with by creating a dependency between the current state of the system and the prior one.

All state space representations are equivalent in the sense that their input/output properties are alike. Notwithstanding, a particular representation may have some advantages over others for a given task. In particular, the *controller canonical* and *observer canonical* forms are best suited to digital filters [68, 69].

$(\sum_{m=0}^{M-1} \theta_m^o F_m(z))u(i)$. Herewith, as the parameters are linear in spite of the original very nature of $H^o(z)$, an unbiased least squares estimate $\{\theta_m(z)\}_{m=0}^{M-1}$ can be found in the closed form. However, off-the-shelf linearization schemes as the Laguerre and Krautz models will always require some a priori knowledge of the system $H^o(z)$ if it is endowed with resonances [13, 16, 65].

A digital filter can be converted to a state-space canonical form by inspection given the strictly proper transfer-function coefficients [69, 70]. Let $\{\mathsf{A}, \mathsf{b}, \mathsf{c}, \mathsf{d}\}$ be the state-space parameters that represent the IIR AF in the observable canonical form. The evolution of the filter's state $\mathcal{W}_i$ in terms of the input is then given in Eq. 2.44, which relies on the definitions given by Eqs. 2.45-2.49.

$$\begin{bmatrix} \mathcal{W}_{i+1} \\ y(i) \end{bmatrix} = \begin{bmatrix} \mathsf{A} & \mathsf{b} \\ \mathsf{c} & \mathsf{d} \end{bmatrix} \begin{bmatrix} \mathcal{W}_i \\ u(i) \end{bmatrix} \tag{2.44}$$

$$\mathcal{W}_{i+1} = \begin{bmatrix} \mathcal{W}_1(i+1) \\ \mathcal{W}_2(i+1) \\ \vdots \\ \mathcal{W}_M(i+1) \end{bmatrix} \tag{2.45}$$

$$\mathsf{A} = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_M \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \tag{2.46}$$

$$\mathsf{b} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.47}$$

$$\mathsf{c} = \begin{bmatrix} b_1 - b_0 a_1 & b_2 - b_0 a_2 & \cdots & b_M - b_0 a_M \end{bmatrix} \tag{2.48}$$

$$\mathsf{d} = b_0 \tag{2.49}$$

In all these equations, the measures $\{b_k\}$ and $\{a_k\}$ are the feedforward and feedback coefficients of the adaptive filter.

The matrix A, defined in Eq. 2.46, is the *state transition matrix* and determines the dynamics of the system (its poles or resonant modes). In the particular case of the state space description of Equations 2.44-2.49, the system under consideration is known as *direct form IIR filter realization type I*. An alternate description of this system is given by Eqs. 2.50-2.52, which may be far more intuitive and convenient for algorithm design as the filter coefficients $\{b_k\}$ and $\{a_k\}$ are dealt with individually [12].

$$\begin{bmatrix} \mathcal{W}_{i+1} \\ w(i) \end{bmatrix} = Q \begin{bmatrix} \mathcal{W}_i \\ u(i) \end{bmatrix} \tag{2.50}$$

$$y(i) = \begin{bmatrix} b_0 & b_1 & \cdots & b_{M-1} & b_M \end{bmatrix} \begin{bmatrix} \mathcal{W}_{i+1} \\ w(i) \end{bmatrix} \tag{2.51}$$

$$Q = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_M & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & 1 & \vdots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \tag{2.52}$$

It is possible to build a state space model $\{\mathsf{A}, \mathsf{b}, \mathsf{c}, \mathsf{d}\}$ whose order is greater than M, the order of the difference equation of the corresponding system. However, any state space model that matches the behaviour of that system and is both controllable and observable has to have order M [71, Theorem 8.9], the so called *McMillan degree* [18] . In this case, the model is said to be a *minimal realization* because it describes the system with the minimum number of states $\mathcal{W}_k(i)$ with $k = 1, 2, \cdots, M$ [55].

[18] From the Linear Systems Theory, if a single realization of a given system $H(z)$ exists, then actually $H(z)$ admits an infinite number of realizations. There may be no upper bound to the order of these realizations, but it can be shown that there is a lower bound known as the McMillan degree. A realization of such lowest order is called *minimal* (or *irreducible*) [68, 71].

2.6 Stability of Time Varying Systems

The notion that keeping the poles inside the unit circle is both a necessary and a sufficient condition for stability is grounded on a coarse approximation of the truth that holds for fixed filters only. For time varying systems as adaptive filters, BIBO stability requires *exponential stability* [71], which is usually defined in the state space and describes the dynamics of the AF under conditions of no input. So, as for the study of the exponential stability aspects, the complete time-varying state space description of an AF given by

$$\begin{bmatrix} \mathcal{W}_{i+1} \\ y(i) \end{bmatrix} = \begin{bmatrix} \mathsf{A}_i & \mathsf{b}_i \\ \mathsf{c}_i & \mathsf{d}(i) \end{bmatrix} \begin{bmatrix} \mathcal{W}_i \\ u(i) \end{bmatrix} \tag{2.53}$$

boils down to the underlying homogeneous system

$$\mathcal{W}_{i+1} = \mathsf{A}_i \mathcal{W}_i, \tag{2.54}$$

where $\mathcal{W}_i$ is the state vector, A_i is the state transition matrix drawn from $A(z,i)$ (i.e., the polynomial $A(z)$ at the instant i), b_i and c_i are vectors, $u(i)$ is the input signal, $y(i)$ is the output and $\mathsf{d}(i)$ is a scalar. Note that differently of Eq. 2.44, in Eq. 2.53 the state space parameters $\{\mathsf{A}_i, \mathsf{b}_i, \mathsf{c}_i, \mathsf{d}(i)\}$ are time-varying as implied by the time index i.

Along the adaptation, $A(z,i)$ defines A_i in a way that the eigenvalues of the later give the roots of the former (i.e., the poles). That being said, it follows that the AF will be (exponentially) stable in the interval $\{i-k\}_{k=1}^{n}$ if and only if there exists a matrix norm $\|.\|$ such that the sequence $\{\mathsf{A}_{i-1}, \mathsf{A}_{i-2} \cdots \mathsf{A}_{i-n+1}, \mathsf{A}_{i-n}\}$ is *contractive*; i.e.,

$$\|\mathsf{A}_{i-1}\mathsf{A}_{i-2} \cdots \mathsf{A}_{i-n+1}\mathsf{A}_{i-n}\| \leq \beta\alpha^{n-1}, \tag{2.55}$$

where $0 \leq \alpha < 1$ and $\beta \geq 1$ are real constants. Thus, under conditions of no input,

the state space vector norm $\|\mathcal{W}_i\|$ of a exponentially stable system is supposed to decay *faster* than the exponential envelope defined by Eq. 2.55 [12, Section 6.2].

For a fixed IIR filter, A_{i-k} is constant and exponential stability will be achieved provided that all of its eigenvalues are bounded by the unit, but in the adaptive case this is not necessarily true: even if all poles are kept into the unit circle during the adaptation, the AF may diverge. On the other hand, even if some eigenvalues of A_{i-k} casually slip to the outside of the unit border for some k, the system may remain stable if the condition of Eq. 2.55 is met.

To clarify this matter, Carini et al. [40] refer to a simple 2nd order system whose impulse response grows with no control in spite of its poles be coincident at $\pm 1/2$ along time, described as

$$y(i) = (-1)^{i-1} y(i-1) - \frac{1}{4} y(i-2) + u(i) \tag{2.56}$$

where a_1 (the coefficient associated to $y(i-1)$) switches from -1 to $+1$ every time i changes. In this example,

$$\mathsf{A}_i = \begin{bmatrix} -1 & -\frac{1}{4} \\ 1 & 0 \end{bmatrix} \text{ for } i \text{ even and } \mathsf{A}_i = \begin{bmatrix} 1 & -\frac{1}{4} \\ 1 & 0 \end{bmatrix} \text{ for } i \text{ odd,} \tag{2.57}$$

whose eigenvalues are $\lambda_1 = \lambda_2 = -1/2$ for i even and $\lambda_1 = \lambda_2 = 1/2$ for i odd; i.e., the system keeps its poles inside the unit border all the time. Notwithstanding, it is unstable, which looks counterintuitive at the first sight but makes sense because even if A_i and A_{i-1} are stable feedback matrices, their product may not be (likewise, even if they were unstable, their product could be stable).

In fact, by inspection, it is possible to tell that Eq. 2.56 is periodic and A_i alternates between the values shown in Eq. 2.57. Therefore, the system will be stable if the product of these matrices is stable, which is not the case: the eigenvalues of $\mathsf{A}_{i-1}\mathsf{A}_i$ are -0.43 and -1.46 [19] and the system is clearly unstable.

[19] The eigenvalues of $\mathsf{A}_{i-1}\mathsf{A}_i$ (supposed to be square matrices) will be the same as those of

Technically, the problem here is not the original eigenvalues *per se* as their moduli are always smaller than 1, but *the abrupt variation* of a_1 in Eq. 2.56, that amounts to $\pi/2$ radians per iteration (the highest frequency in discrete-time signal analysis).

For a similar system in which a_1 changed slower from -1 to $+1$; e.g.,

$$y(i) = \sin\left(i/100\right) y(i-1) - \frac{1}{4}y(i-2) + u(i) \implies \mathsf{A}_i = \begin{bmatrix} \sin\left(i/100\right) & -\frac{1}{4} \\ 1 & 0 \end{bmatrix}, \tag{2.58}$$

the stability will be preserved because the condition shown in Eq. 2.55 is respected. Likewise, even if the moduli of a few eigenvalues are greater than 1, the system keeps exponentially stable provided that it changes slowly enough; e.g.,

$$y(i) = e^{-(i-5)}y(i-1) - \frac{1}{4}y(i-2) + u(i) \implies \mathsf{A}_i = \begin{bmatrix} e^{-(i-5)} & -\frac{1}{4} \\ 1 & 0 \end{bmatrix}, \tag{2.59}$$

which is a stable system despite of some of its eigenvalues have moduli greater than 1 for $i < 5$.

To derive the maximum rate of variation that a given IIR AF endures without becoming unstable is a rich field of study by itself (e.g., [40,73]), where one defines an "energy" function which is contractive for the homogeneous part of the system, normally via Lyapunov methods. In practice though, often the step size is kept small as mentioned in Chapter 1, which causes the parameters of the AF to vary slowly and, therefore, it is common that IIR AFs be exponentially stable by construction (see [74, Condition 2.6]). However, even if the condition in Eq. 2.55 is obeyed for a certain AF, actually it could require an unreasonably high β so that an overflow would be raised - even for a system that is theoretically stable. For this reason, to restrict the poles to the unit circle during the adaptation became a common sense among the designers [3, 4, 12].

$\mathsf{A}_i\mathsf{A}_{i-1}$ [72, Theorem 1.3.20], so the order in which the product is computed does not matter.

2.7 Exponential Stability and the Direct Form

All discussions led in Section 2.6 tackle the notion that keeping the poles inside the unit circle is necessary and sufficient to assure stability, except for fixed filters. As Wu points out [75], the individual matrices $\{\mathsf{A}_{i-k}\}_{k=1}^{n}$ can not tell much about the stability of a linear time-varying system and, in the general case, *stability criteria of linear time-invariant systems do not apply to linear time-varying systems* and *instability criteria of linear time-invariant systems do not apply to linear time-varying systems.*

Within the scope of this work, it is important to understand the implications of the exponential stability, which explains why we adopted the direct form instead of the intrinsically stable normalized lattice. The point of interest here is to understand that it is possible to keep a direct form IIR AF stable without resorting to stability tests despite occasional unstable updates. Theoretically, this can be achieved by making it operate far below the exponential stability upper bounds as told by the greatest adaptation rate this AF can cope with (e.g., [49]; also, [40,73]). In practice, *this amounts to adopt an arbitrarily small learning step size*, such that occasional unstable poles are brought back into the unit circle by the gradient in the next iterations, enforcing the condition in Eq. 2.55 [20] . Under these circumstances, the IIR AF is going to be intrinsically stable anyway and the option for the normalized lattice becomes a matter of choice [21] .

Now, in this work we aim to improve the convergence rates of IIR AFs by employing parallel hybrid filters combinations in which a fast but inaccurate AF

[20] This stabilizing property, in which the poles of a system often drift towards instability and recover next, is known as *bursting* in the context of adaptive echo cancellation and adaptive control systems. The same concept is used to explain the stabilizing features of the pseudo-linear regression (PLR) algorithm, also an IIR adaptive strategy [2, Section 2.9.4] [76].

[21] Several works exploit the concept of exponential stability to determine the maximum learning step size that a given IIR AF endures (e.g., [40, 77]). Normally, the goal of these works is related to improve the convergence speed, not avoid the stability tests, although exponential stability criteria can be used to set the upper bounds under which the AF could be operated safely without resorting to such tests as implied in [49].

is associated to an accurate but slow AF. Once we are addressing IIR system identification scenarios in this work, the accurate component has to be IIR and to keep its learning step size arbitrarily small is a strict requirement to enforce low estimation errors at the steady state (refer to Chapter 4) [78]. As a result, in practice the IIR AF is expected to operate far below the exponential stability upper bounds in one way or another, and then the stability tests can be discarded.

We should also mention an additional aspect that encourages the use of the direct form to the detriment of the normalized lattice: the mathematical tools for performance analysis. As a literal interpretation of difference equations like

$$y(i) = \sum_{k=0}^{M} b_k(i-1)u(i-k) + \sum_{k=1}^{M} a_k(i-1)y(i-k), \tag{2.60}$$

where M is the order of the AF, $y(i)$ is its output and $u(i)$ is the input signal, the direct form manipulates the parameters $\{b_k\}$ and $\{a_k\}$ directly in a *linear fashion* [22], which is far more intuitive for the designer [4, 20] and makes the mean square analyses of the adaptive algorithms mathematically more tractable. In turn, the normalized lattice features an input/output relationship that is strongly nonlinear due to the rotation parameters [12, 19, 23], rendering the analysis quite complex.

For those reasons and by taking into account that the direct form is convenient when mixing differently designed AFs as noted by Lopes in [79, 80], the experiments and analysis led in this work consider the direct form realization for IIR AFs. In spite of that, we point out that our option for the direct form does not prevent the use of the normalized lattice in the algorithms we propose along this work because, as noted previously, they are functionally equivalent in the sense both can approximate a given transfer function $H^o(z)$.

[22] By linear *fashion*, we mean that difference equations are *represented by linear operations* such as the sums shown in the text or an inner product of space vectors, not that the AF itself is linear. As a matter of fact, the recursive nature of IIR AFs make them intrinsically non-linear.

2.8 Adaptation

In the identification scenario of Fig. 2, the adaptation of an adaptive filter is driven by an iterative optimization criterion to adjust its coefficients towards the *optimal values* which makes the filter match the plant behaviour as close as possible in some statistical sense.

Different algorithms and optimization criteria are used to adapt the coefficients of an AF, but in many cases a common structure is observed [4,5,45]. Such structure is given by Eq. 2.61, where $e(i)$ is the estimation error [23] , w_i is the parameters vector of the AF, $\mu(i)$ is a positive (possibly constant) *learning step size*, B_i is a positive definite matrix computed as a function of x_i, η_i is a vector also derived from x_i, typically related to the gradient of $e^2(i)$ with respect to w_{i-1} and ^{T} denotes the Hermitian conjugation.

$$w_i = w_{i-1} + \mu(i)\mathsf{B}_i\eta_i^T e(i) \tag{2.61}$$

The main advantage of conceiving a general updating rule like Eq. 2.61 is that all major adaptive algorithms arise via proper choices for B_i and η_i. By forgoing this approach, these algorithms must be derived individually from scratch, which ultimately unveils that the rule of Eq. 2.61 underlies all of them.

2.8.1 FIR LMS Recursion

In the FIR LMS case, $\mathsf{B}_i = \mathbb{I}$ (the identity) and the regressor is made of the current and delayed versions of the input $u(i)$ only; i.e., $\eta_i = x_i = u_i = [u(i)\ u(i-1)\ \cdots\ u(i-M)]$ and w_i is changed accordingly, making the recursion

[23] As noted in Section 2.4, the estimation error depends upon the formulation employed to adapt the coefficients of the IIR AF (OE or EE), which defines an output $y_e(i)$ that can differ from the regular AF output $y(i)$. However, as emphasized in Section 2.4.3, in this work we consider primarily the OE approaches.

of Eq. 2.61 become [4, 5]

$$w_i = w_{i-1} + \mu u_i^T e(i), \tag{2.62}$$

which turns out to be the same updating rule of the OE LMS with the feedback coefficients $\{a_k\}_{k=1}^{M}$ set to 0, as shown in the sequel.

2.8.2 IIR OE LMS Recursion (LMSOE)

Once the EE approach consists in a dual-channel FIR configuration as mentioned in Section 2.4.2, the LMS version for the EE (LMSEE) can be derived naturally from that of the FIR LMS which, in turn, is equal to the OE LMS (LMSOE) when the feedback coefficients are 0. For this reason, here we pursue the formal derivation of the LMSOE only, treating it as an optimization problem whose iterative solution will be shown to follow the rule given by Eq. 2.61 [24] .

As mentioned in Section 2.4, the adaptation resorts to a *figure of merit* used as a performance criterion which has to be optimized in relation to the adjustable parameters set of the AF. Several figures of merit are used according to the application and mathematical tractability, but in the case of the Output Error methods it is normally the *mean square output error* (MSOE) defined in Section 2.4.1 as

$$\mathrm{E}\, e^2(i) = \mathrm{E}\left(d(i) - y(i)\right)^2, \tag{2.63}$$

which is also denoted as the *cost* $J = J(w_{i-1})$, given as a function of the parameters set w_i once $y(i) = x_i w_{i-1}$.

The MSOE is endowed with a well-defined global minimum in case the plant is sufficiently modelled by the AF (refer to Appendix B) and the input $u(i)$ is

[24] The LMSEE is described in Section 4.3.2 (Eq. 4.53), where it is evident that the EE gradient vector is proportional to minus the regression vector, just like in the LMS FIR case seen in Eq. 2.62.

white, as it is the case of the scenarios dealt with in this work. So, by considering that $\mathrm{E}\, e^2(i)$ is continually differentiable with respect to each one of the elements of the parameter vector w_{i-1} [1], the lowest cost is achieved when the derivative

$$\frac{\partial \mathrm{E}\, e^2(i)}{\partial w_{i-1}} = -2\,\mathrm{E}\, e(i)\frac{\partial y(i)}{\partial w_{i-1}} \tag{2.64}$$

is zero. By using the unconstrained optimization method, this leads to the opposite of the *update term* p used to update iteratively the parameters set as

$$w_i = w_{i-1} - \mu p, \tag{2.65}$$

in which μ is the step size.

In Eq. 2.64, $\frac{\partial y(i)}{\partial w_{i-1}}$ can be determined resorting to the mixed notation as in Eqs. 2.66-2.71.

$$y(i) = H(z, i-1)u(i) = \frac{B(z, i-1)}{1 - A(z, i-1)}u(i) \tag{2.66}$$

$$\frac{\partial y(i)}{\partial B(z, i-1)} = \frac{1}{1 - A(z, i-1)}u(i) \tag{2.67}$$

$$\begin{aligned}\frac{\partial y(i)}{\partial b_m(i-1)} &= \frac{1}{1 - A(z, i-1)}u(i-m) \\ &= u(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial b_m(i-1)}\end{aligned} \tag{2.68}$$

$$\frac{\partial y(i)}{\partial A(z, i-1)} = \frac{B(z, i-1)}{1 - A(z, i-1)}\frac{1}{1 - A(z, i-1)}u(i) \tag{2.69}$$

$$\frac{\partial y(i)}{\partial A(z, i-1)} = \frac{1}{1 - A(z, i-1)}y(i) \tag{2.70}$$

$$\begin{aligned}\frac{\partial y(i)}{\partial a_m(i-1)} &= \frac{1}{1 - A(z, i-1)}y(i-m) \\ &= y(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial a_m(i-1)}\end{aligned} \tag{2.71}$$

Note that the partial derivatives of these equations can not be easily determined in a closed form [2, 3]. However, *if the step-size is kept sufficiently small*,

they can be approximated as in Eqs. 2.72 and 2.73.

$$\begin{aligned}\frac{\partial y(i)}{\partial b_m(i-1)} &= u(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial b_m(i-1)} \\ &\approx u(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial b_m(i-1-k)} \end{aligned} \tag{2.72}$$

$$\begin{aligned}\frac{\partial y(i)}{\partial a_m(i-1)} &= y(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial a_m(i-1)} \\ &\approx y(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial a_m(i-1-k)} \end{aligned} \tag{2.73}$$

These approximations can be used in Eq. 2.64, leading to Eqs. 2.74 and 2.75.

$$\begin{aligned}\frac{\partial \mathrm{E}\, e^2(i)}{\partial b_m(i-1)} &\approx -2\,\mathrm{E}\, e(i)\frac{\partial y(i)}{\partial b_m(i-1)} \\ &= -2\,\mathrm{E}\, e(i)\left(u(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial b_m(i-1-k)}\right) \end{aligned} \tag{2.74}$$

$$\begin{aligned}\frac{\partial \mathrm{E}\, e^2(i)}{\partial a_m(i)} &\approx -2\,\mathrm{E}\, e(i)\frac{\partial y(i)}{\partial a_m(i-1)} \\ &= -2\,\mathrm{E}\, e(i)\left(y(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial a_m(i-1-k)}\right) \end{aligned} \tag{2.75}$$

By dropping the expectancies from these equations, they become the stochastic approximations for $\frac{\partial \mathrm{E}\, e^2(i)}{\partial b_m(i)}$ and $\frac{\partial \mathrm{E}\, e^2(i)}{\partial a_m(i)}$; i.e.,

$$\begin{aligned}\frac{\partial \mathrm{E}\, e^2(i)}{\partial b_m(i-1)} &\approx \frac{\partial e^2(i)}{\partial b_m(i-1)} \\ &= -2e(i)\left(u(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial b_m(i-1-k)}\right) \end{aligned} \tag{2.76}$$

$$\begin{aligned}\frac{\partial \mathrm{E}\, e^2(i)}{\partial a_m(i-1)} &\approx \frac{\partial e^2(i)}{\partial a_m(i-1)} \\ &= -2e(i)\left(y(i-m) + \sum_{k=1}^{M} a_k(i-1)\frac{\partial y(i-k)}{\partial a_m(i-1-k)}\right) \end{aligned} \tag{2.77}$$

Eqs. 2.76 and 2.77 show that both the regressor signals $u(i-m)$ and $y(i-m)$

are filtered by $A(z)$. So, a *filtered regressor* ϕ_i could be defined as [3]

$$\phi_i = x_i + \sum_{k=1}^{M} a_k(i-1)\phi_{i-k} = \frac{x_i}{1 - A(z, i-1)}, \tag{2.78}$$

allowing the derivative $\frac{\partial \mathrm{E}\, e^2(i)}{\partial w_{i-1}}$ of Eq. 2.64 be stochastically approximated as

$$\frac{\partial \mathrm{E}\, e^2(i)}{\partial w_{i-1}} \approx \frac{\partial e^2(i)}{\partial w_{i-1}} = -2\phi(i)e(i), \tag{2.79}$$

whose negative replaces the update term p in Eq. 2.65 as shown in Eq. 2.80 [25].

$$w_i = w_{i-1} + \mu\phi_i^T e(i) \tag{2.80}$$

Note that Eq. 2.80 follows the general updating rule of Eq. 2.61 with $B_i = \mathbb{I}$ and $\eta_i = \phi_i$. Also, as previously mentioned, the FIR LMS recursion (and, by extension, the LMSEE) can be retrieved from Eq. 2.80 if $a_k = 0$ for $0 < k \leq M$.

Once the regressor x_i is actually made of delayed versions of the input and output signals, the small step-size assumption allows the filtered regressor shown in Eq. 2.78 be rewritten in terms of a computationally simpler approximation for $\frac{\partial y(i)}{\partial w_{i-1}}$ as in Eqs. 2.81- 2.83 [3, 4].

$$y_F(i) = y(i) + \sum_{k=1}^{M} a_k(i-1)y_F(i-k) \tag{2.81}$$

$$u_F(i) = u(i) + \sum_{k=1}^{M} a_k(i-1)u_F(i-k) \tag{2.82}$$

$$\phi_i = [y_F(i-1) \ \cdots \ y_F(i-M) \ u_F(i) \ \cdots \ u_F(i-M)] \tag{2.83}$$

By using this formulation, the computation of ϕ_i is reduced from M autoregressive operations of order $2M+1$ in Eq. 2.78 to only two operations of the same order in Eq. 2.83 .

[25] As the constant "2" does not affect the direction of the gradient, it may be conveniently dropped in order to simplify the expression.

2.8.3 Normalized IIR LMS (N-LMSOE)

In spite of its good general properties, the LMSOE algorithm may be inefficient in the presence of underdamped or clustered poles [10, 11, 19], particularly when the parameters are approaching the global minimum (see Appendix C). Besides, experience and reports from several works indicate that the magnitude of the instant gradient vectors can change appreciably over the error surfaces (e.g., [10, 26]). Under these circumstances, the normalization of the regressor ϕ_i mitigates the influence of the input signal amplitude on the gradient noise and simplify the choice of μ.

This normalization strategy applied to the LMSOE yields the algorithm known as normalized LMSOE (N-LMSOE). In this algorithm, the minimization of the objective function $J(w_{i-1}) = \mathrm{E}\, e^2(i)$ is performed by using the LMSOE stochastic gradient seen in Eq. 2.79 and the inverse of an stochastic approximation for the Hessian matrix $\nabla_w^2 J(w_{i-1})$.

The Hessian is given by

$$\nabla_w^2 J(w_{i-1}) = 2\,\mathrm{E}\left(\phi_i^T \phi_i + \frac{\partial^2 e(i)}{\partial w_{i-1}^2} e(i)\right), \tag{2.84}$$

where the rightmost term is approximately 0 since $e(i)$ is zero mean and becomes independent of $\frac{\partial^2 e(i)}{\partial w_{i-1}^2}$ as $i \to 0$ [4, Section 10.4.2]. Thus, $\nabla_w^2 J(w_{i-1})$ may be stochastically approximated as

$$\nabla_w^2 J(w_{i-1}) \approx \phi_i^T \phi_i \tag{2.85}$$

Once the term $\phi_i^T \phi_i$ is a rank-one approximation for the Hessian, it can not be inverted directly. However, by using a small regularization term ϵ and applying the matrix inversion lemma [5, Section 5.1], it can be shown that

$$\frac{1}{\epsilon \mathbb{I} + \phi_i^T \phi_i} \phi_i^T = \frac{\phi_i^T}{\epsilon + \|\phi_i\|^2} \tag{2.86}$$

Therefore, by making $\mathsf{B}_i = \left(\epsilon\mathbb{I} + \phi_i^T\phi_i\right)^{-1}$ and using the stochastic gradient $\eta_i = \phi_i e(i)$ given in Eq. 2.79, Eq. 2.61 can then be restated as

$$\begin{aligned} w_i &= w_{i-1} + \mu\mathsf{B}_i\eta_i^T e(i) \\ &= w_{i-1} + \mu\frac{1}{\epsilon\mathbb{I} + \phi_i^T\phi_i}\phi_i^T e(i) \\ &= w_{i-1} + \mu\frac{\phi_i^T}{\epsilon + \|\phi_i\|^2}e(i) \end{aligned} \tag{2.87}$$

which is the update rule for the N-LMSOE [81].

Extensive simulations corroborate the claims found in the literature (e.g., [45, 82]) about the superior performance of the N-LMSOE against the LMSOE when identifying systems with clustered poles of large magnitude. Although the presence of poles bring unique features to the IIR AFs in terms of modelling capabilities and stability issues, the motivation for a normalization procedure in the N-LMSOE resembles pretty much that in the FIR adaptive filtering: to make the AF (IIR or not) more resilient to varying statistics in the input regressor [83].

Similar benefits are observed if the normalization is applied to the regressor x_i instead of to its filtered version ϕ_i, as described in Eqs. 2.88-2.90. However, in this work we employed the regular N-LMSOE as given by Eq. 2.87.

$$x_i^N = \frac{x_i}{\epsilon + \|x_i\|} \tag{2.88}$$

$$\phi_i^N = x_i^N + \sum_{k=1}^{M} a_k\phi_{i-k}^N = \frac{x_i^N}{A(z)} \tag{2.89}$$

$$w_i = w_{i-1} + \mu\phi_i^N e(i) \tag{2.90}$$

2.9 Stability Bounds on the Step Size μ

In the FIR LMS case, the stability bounds on the step size μ can be computed directly from the statistical properties of the input signal $u(i)$ as [5]

$$0 < \mu < \frac{2}{\lambda_{max}}, \tag{2.91}$$

where λ_{max} denotes the largest eigenvalue of the correlation matrix

$$R_u = \mathrm{E}\, u_i^T u_i \tag{2.92}$$

In the LMSOE case, these bounds can not be determined so directly because the recursivity poses them in terms of the time-varying polynomial $A(z)$, which means that μ can not be determined in the closed form but only coarsely approximated. In fact, if the adaptation is slow enough, $A(z)$ may be considered as a constant *locally*; i.e., within an interval $\{i,\ i+1,\ i+2, \ldots,\ i+\mathbb{L}\}$ with $\mathbb{L}$ arbitrary. By exploiting a similar reasoning, Regalia derived the following bound for the LMSOE [12, Section 7.9]

$$0 < \mu < \frac{2}{\sum_{k=0}^{M_2} \left(\frac{\partial y(i)}{\partial b_k(i-1)}\right)^2 + \sum_{k=1}^{M_2} \left(\frac{\partial y(i)}{\partial a_k(i-1)}\right)^2} \tag{2.93}$$

$$0 < \mu < \frac{2}{\|\phi_i\|^2}, \tag{2.94}$$

where $\frac{\partial y(i)}{\partial b_k(i-1)}$ and $\frac{\partial y(i)}{\partial a_k(i-1)}$ are the elements of the stochastic gradient $\frac{\partial y(i)}{\partial w_{i-1}}$ (see Equations 2.74 and 2.75).

Given that $A(z)$ is time-varying, Equations 2.93 and 2.94 can be understood as local estimations for the stability bounds and no global bound can be determined. Since those approximations are coarse, in practice these values need to be decreased by an order of magnitude or more in a trial-and-error approach. In our case, as aforementioned, the IIR AFs are designed deliberately to operate far

below the exponential stability bounds, such that μ becomes naturally a fraction of its maximum theoretical value.

3 PARALLEL COMBINATIONS OF ADAPTIVE FILTERS

3.1 Initial Considerations

In IIR adaptive filtering, the magnitude of the stochastic gradient is likely to vary significantly over the error surfaces depending on the configuration of the poles (Appendix C). As a result of the gradient noise, such variations could set the parameters of the AF to locations far away from the global minimum [10, 26]. Whereas this phenomenon decreases with the normalized form of the LMSOE, the speed of convergence is still an issue due to the small step size requirement. Clearly, there is room for improvement in the IIR gradient descent algorithms.

One way of getting around those limitations is by using *filters combinations*, whose concept is a paradigm shift in the design of adaptive filters. Instead of using a single AF designed under limited or no knowledge about the filtering scenario, a combination allows the use of a pool of differently designed AFs (the *experts*). These experts present their individual outputs to a *supervisor* in charge of delivering a global estimate at least as good as the best expert in the pool, normally in terms of some figure of merit as the mean-square error.

The design of a filter combination involves some fundamental choices: the number of AFs (that balances the relation between performance and computational complexity), the individual experts design (comprehending filtering structure and learning rule), the supervisor design (discussed in Section 3.2) and a

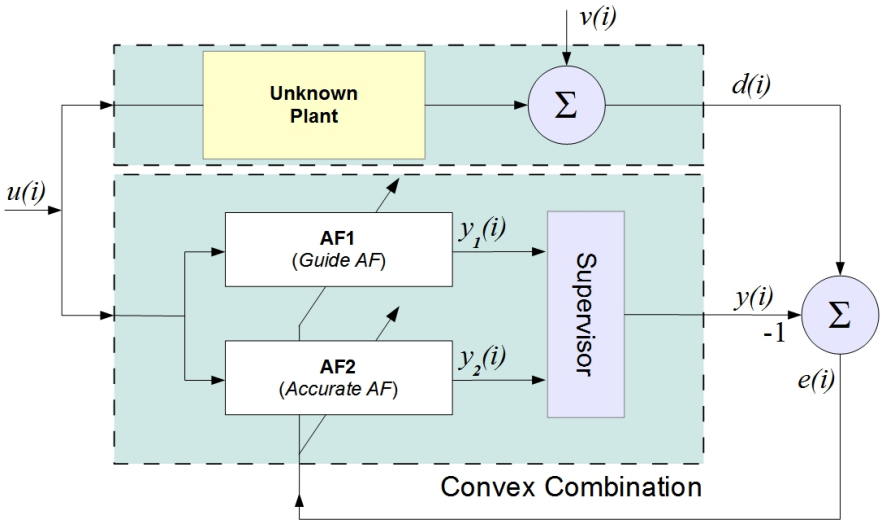

Figure 7: Filters Combination in the a Systems Identification Configuration

topology.

The combination topology define how the AFs are structurally organized. The first works on this matter introduced the parallel topology for the FIR case, in which every component evolves independently from its peers (Fig. 7) [79, 84, 85]. Herewith, a combination of two experts is able to deliver the convergence of a fast μ_1-LMS AF allied with the low steady-state error of an accurate μ_2-LMS AF where $\mu_1 >> \mu_2$. This leads to a better overall performance, albeit the combination experiments a stagnation effect that persists until the accurate AF outperforms the fast one (compare the different curves in Figures 21 or 22). However, this effect may be combated via ad-hoc weights transfers [78, 86] or, more recently, with cyclic coefficients feedbacks [29].

As the number of experts increases, the hierarchical topology may be employed as an alternative to the standard parallel structure where the components are combined in parallel and in layers, which tends to accelerate convergence [79].

The incremental topology was proposed as a way to combat the stagnation effect. This topology is a cascade in which every expert receives the weights estimates, update them and pass them to the next expert, therefore implementing

a serial structure as opposed to parallel. Although efficient for combating the stagnation effect, the supervisor design is challenging, as the experts no longer evolve independently. However, it can be efficiently explored as a technique to reduce complexity: an incremental combination of several LMS AFs was shown to outperform the Affine Projection Algorithm (APA) with less computational complexity in the FIR system identification scenario [87].

During the development of this thesis, several studies involving the incremental combination of FIR AFs and IIR AFs were conducted with very promising results (e.g., [87]). However, in the techniques developed in the next chapters, we focus on parallel structures in which at least one of the components is IIR.

3.2 Supervisor Design

As aforementioned, the supervisor aggregates the individual outputs yielded by the experts to deliver an output as good as the best one available at every iteration [29]. In the case of a parallel topology with two AFs combined through a convex supervisor as seen in Fig. 7, the outputs are weighted by a mixing parameter $\lambda(i)$ such that the combined output $y(i)$ is [78]

$$y(i) = \lambda(i)y_1(i) + \big(1 - \lambda(i)\big)y_2(i),\ 0 < \lambda(i) < 1, \tag{3.1}$$

where $y_k(i)$ denotes the output of the k-th AF.

Usually, the mixing factor $\lambda(i)$ is defined by a sigmoidal activation function such that [78, 88, 89]

$$\lambda(i) = \frac{1}{1 + e^{-a(i-1)}}, \tag{3.2}$$

which means that $\lambda(i)$ is not adapted directly, but is mapped from the instantaneous value of the auxiliary variable $a(i)$. Fig. 8 shows that the sigmoid constrains $\lambda(i)$ to the asymptotic interval $(0, 1)$, ensuring that the combination is indeed con-

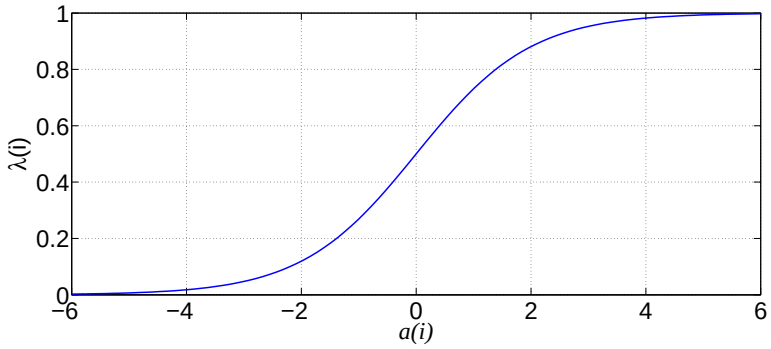

Figure 8: The Sigmoidal Activation Function for the Mixing Factor $\lambda(i)$

vex [90] and switches smoothly between the component filters according to their relative performance.

Once the combined output is given by $y(i)$, then the global quadratic error is defined as

$$e^2(i) = \big(d(i) - y(i)\big)^2 = \Big(d(i) - \big(\lambda(i)y_1(i) + \big(1 - \lambda(i)\big)y_2(i)\big)\Big)^2 \tag{3.3}$$

In order to minimize $e^2(i)$, $a(i)$ is continuously adapted via a gradient descent procedure, leading to the following update rule [78]

$$\begin{aligned} a(i) &= a(i-1) - \frac{\mu_a}{2}\frac{\partial e^2(i)}{\partial a(i)} \\ &= a(i-1) - \frac{\mu_a}{2}\frac{\partial e^2(i)}{\partial \lambda(i)}\frac{\partial \lambda(i)}{\partial a(i)} \\ &= a(i-1) + \mu_a e(i)\big(y_1(i) - y_2(i)\big)\lambda(i)\big(1 - \lambda(i)\big), \end{aligned} \tag{3.4}$$

where μ_a is the learning step for $a(i)$.

Fig. 8 shows that the slope of the sigmoid (i.e., its derivative) becomes progressively smoother when approaching its limits. Herewith, not only the adaptation accelerates with the intermediate values of $\lambda(i)$ but also the stochastic gradient noise decreases with the extreme values [78]. On the other hand, this also means that $a(i)$ virtually stops adapting as $\lambda(i)\big(1 - \lambda(i)\big)$ approaches 0. So, to ensure a minimal level of adaptation, $a(i)$ may be constrained to a symmetrical

interval such that [78, 91]

$$a(i) = \Big(a(i-1) + \mu_a e(i)\big(y_1(i) - y_2(i)\big)\lambda(i)\big(1-\lambda(i)\big)\Big)_{-A^+}^{A^+}, \tag{3.5}$$

where $(.)_{-A^+}^{A^+}$ denotes that the argument ranges within the interval $(-A^+, A^+)$ and A^+ is a real positive constant chosen in a way that if $a(i) \geq A^+$ then $\lambda(i)$ is arbitrarily close to 1 (Eq. 3.2).

Now, by realizing that the scheme of Eq. 3.5 prevents $\lambda(i)$ from reaching its limit values, Arenas-Garcia [78] suggests a slight modification for the rule of Eq. 3.1, such that

$$y_u(i) = \lambda_u(i) y_1(i) + \big(1 - \lambda_u(i)\big) y_2(i) \tag{3.6}$$

$$\lambda_u = \begin{cases} 1, & a(i) \geq A^+ \\ \lambda(i), & -A^+ < a(i) < A^+ \\ 0, & a(i) \leq -A^+ \end{cases}$$

With this modified rule, λ_u is able to switch entirely to one of the component filters, which is impossible with the rule of Eq. 3.1. This ensures *universality*, with $y_u(i)$ being as good as the best available components outputs in the mean square sense. In fact, under certain circumstances $y_u(i)$ may even outperform the component filters.

3.3 Normalized Supervisor Design

Albeit the scheme ruled by Eq. 3.4 succeeds in merging adaptively the complementary capabilities of differently designed AFs, in non-stationary scenarios the correct setup of the learning step μ_a may be somewhat tricky. An incorrect value for μ_a makes the supervisor switches mistakenly (too early or late) and compromises the universality of the combination, especially if the statistics of the

input signals are varying [92].

By noticing that, Azpicueta-Ruiz et al. [88] suggested a normalization strategy that simplifies the proper selection of μ_a and makes the combination more robust to non-stationary inputs, particularly the signal-to-noise ratio (SNR).

Let the update rule of Eq. 3.4 restated as

$$\begin{aligned} a(i) &= a(i-1) + \mu_a e(i)\big(y_1(i) - y_2(i)\big)\lambda(i)\big(1-\lambda(i)\big) \\ &= a(i-1) + \mu_a e(i)\big(e_2(i) - e_1(i)\big)\lambda(i)\big(1-\lambda(i)\big), \end{aligned} \tag{3.7}$$

where $e_k(i) = d(i) - y_k(i)$, $k = 1, 2$. Now, it is possible to notice by inspection that Eq. 3.7 closely resembles the LMS updating rule for an AF endowed with a variable step size $\mu_a\lambda(i)\big(1-\lambda(i)\big)$ and fed by an input signal $e_2(i) - e_1(i)$. As Azpicueta-Ruiz et al. [88] properly argue, such equivalence becomes clearer by restating the combined output as

$$\begin{aligned} y(i) &= y_2(i) + \lambda(i)\big(y_1(i) - y_2(i)\big) \\ &= y_2(i) + \lambda(i)\big(e_1(i) - e_2(i)\big), \end{aligned} \tag{3.8}$$

making the combination a two-layer adaptive arrangement in which the first layer is made of the component filters performing independently and the second is a single filter that processes the signal $e_2(i) - e_1(i)$ to minimize the overall error. So, in the same fashion the Normalized LMS (NLMS) mitigate the influence of the (possibly varying) statistics of the input signal, it is tempting to normalize the updating rule of Eq. 3.7 as

$$a(i) = a(i-1) + \frac{\mu_a\lambda(i)\big(1-\lambda(i)\big)}{\big(e_2(i) - e_1(i)\big)^2} e(i)\big(e_2(i) - e_1(i)\big) \tag{3.9}$$

Notwithstanding, experience indicates that merely squaring the instantaneous value $\big(e_2(i) - e_1(i)\big)$ gives a poor indicative of the power of the signal, which renders the rule of Eq. 3.9 unsatisfactory in practice. Better results arise by

employing the NLMS with power normalization algorithm to Eq.3.7, yielding

$$a(i) = a(i-1) + \frac{\mu_a}{p(n)}\lambda(i)\big(1-\lambda(i)\big)e(i)\big(e_2(i)-e_1(i)\big), \tag{3.10}$$

where $p(n)$ is a filtered version of $\big(e_2(i)-e_1(i)\big)^2$ defined as

$$p(i) = \gamma p(i-1) + (1-\gamma)\big(e_2(i)-e_1(i)\big)^2, \tag{3.11}$$

with the forgetting factor γ chosen somewhat arbitrarily close to 1, such as $\gamma = 0.9$

Eq. 3.10 shows a greater robustness when compared to the original adaptive scheme of Eq. 3.4. Also, this scheme facilitates the choice of the step size μ_a, in the same fashion that the design of the step size in the NLMS algorithms is less tricky than in the ordinary LMS.

4 HYBRID CONVEX COMBINATIONS WITH IIR ADAPTIVE FILTERS

4.1 Initial Considerations

A discrete-time system is stable if and only if its impulse response is absolutely summable [7,93]. For this reason, the impulse response $H^o = \{h_m^o\}_{m=0}^{\infty}$ of a given IIR system $H^o(z)$ has to decay along time, otherwise it would not be stable. As such, in practice it is clear that H^o is made of a finite amount of *significant* samples, say $0 < \mathcal{M} \ll \infty$, such that $|h_m^o| \approx 0$ for $m > \mathcal{M}$. In this work, we will refer to the truncated set $H^{o\prime} = \{h_m^o\}_{m=0}^{\mathcal{M}}$ as the *useful impulse response* of $H^o(z)$ and, intuitively, assume that it stores most of the energy of H^o; i.e.,

$$\sum_{m=0}^{\mathcal{M}} |h_m^o|^2 - \sum_{m=0}^{\infty} |h_m^o|^2 \approx 0 \tag{4.1}$$

Once $\mathcal{M} \ll \infty$, it is possible to store $H^{o\prime}$ into the taps of any FIR AF $H_1(z)$ with order $M_1 \geq \mathcal{M}$, which means that $H_1(z)$ is capable of identifying $H^o(z)$. Therefore, in principle one can take advantage of the robustness of the FIR AFs in the IIR systems identification scenarios, avoiding the severe performance issues related to the adaptation of the poles (see Chapter 1 and Appendix B).

However, there is no way of telling a priori how large $\mathcal{M}$ should be. Relatively short FIR AFs may not accommodate a significant amount of the energy of the original infinite impulse response, converging prematurely to a high steady state

level [1]. On the other hand, really large FIR AFs with order $M_1 \gg \mathcal{M}$ are not only a waste of computational resources *per se* as an equivalent IIR AF is likely to have much less coefficients, but may also be inefficient simply because $|h_m| \approx 0$ for $\mathcal{M} < m \leq M_1$.

In this context, parallel combinations of LMSOE AFs would be of little help: the stability issues remain and the trade-offs between convergence rate and accuracy are less dramatic than in the FIR case. In fact, experience shows that in case the unknown plant has a large Hankel singular value spread [2] , the LMSOE may stagnate for a large range of values for the step size even in the absence of local minima (see Appendix C) [82]. As a result, the convergence speed of the IIR AFs is likely to be poor exactly in the scenarios they exhibit the greatest advantage over their FIR counterparts in terms of computational complexity [10].

4.2 T-OE: Transversal Output-Error Combinations

Motivated by these issues, the hybrid FIR–IIR adaptive combination of Fig. 9 is proposed, featuring the high convergence speed of a low complexity FIR (AF1, whose transfer function is $H_1(z)$) and the low steady-state error of an accurate Output-Error IIR (AF2, transfer function $H_2(z)$). Since the FIR component is transversal, this arrangement was called a *Transversal Output Error* combination or *T-OE*.

Once the plant (denoted by $H^o(z)$) is assumed to be IIR in nature, a suitable

[1] The relation between the order of an AF and its convergence is not often exploited in the literature, but practitioners are aware that the longer an AF, the lower is its convergence rate. In the FIR case, to increase the order has an effect roughly similar to decrease the step size but, for IIR AFs, the convergence rate may drop significantly as the order of the filter grows and often stagnates for AFs of 3rd order and above [12, 82].

[2] A Hankel matrix Γ of the system $H^o(z)$ stores the impulse response of $H^o(z)$ in its first row and the subsequent rows are left-shifted by a sample regarding the previous one. If the ratio of the maximum to the minimum singular values of Γ (i.e., *its singular value spread*) is large, then it is ill-conditioned and identification by any IIR AF becomes harder (see Appendix C). Examples of Hankel matrices are seen in Eq. 4.4 (Section 4.2.2) for a FIR AF and in Eq. B.8 (Section B.1 from Appendix B) for an IIR AF.

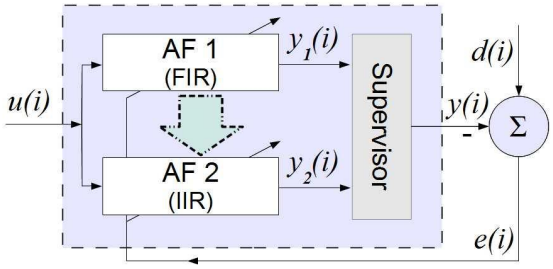

Figure 9: The T-OE combination

candidate for AF2 is a direct form N-LMSOE [3] equipped with a small step size. As for AF1 (also called the *guide filter*), good choices are the LMS or N-LMS due to their robustness, good tracking capabilities and computational simplicity [4,5,83].

In this scheme, AF1 is designed to adapt fast and reach the vicinity of the global minimum rapidly while AF2 is slower but manages to identify the system accurately. Albeit these complementary capabilities lead to a superior overall performance, the gap between the relative performances of the component filters may still be an issue: depending on the pole-zero configuration of $H^o(z)$, AF1 may feature a too high steady state error while AF2 may converge considerably later. As a result, the combination will show no performance improvement until AF2 outperforms AF1, which could be unacceptable for some applications.

As aforementioned, this performance gap also occurs in parallel combinations of FIR AFs, where it is tackled to a large extent through weights transferring techniques [29,78,86,89]. In the T-OE context, a similar idea is used as indicated by the thick arrow in Fig. 9. However, here the transfers are performed *unidirectionally* from AF1 to AF2 which means that their coefficients are not aggregated. This is necessary because AF2 is IIR (therefore, subjected to instabilities) and could destabilize the whole combination in case its coefficients were propagated to

[3] Section 2.2 provides the technical justifications for the choice of the direct form IIR AF to the detriment of the intrinsically stable normalized lattice. Notwithstanding, it is possible to employ other realizations as the lattice in practical implementations of the T-OE.

AF1 somehow, particularly in the presence of abrupt non-stationarities or burst noise.

Once the transfers are unidirectional, they have to cease whenever the later outperform the former, otherwise the overall performance is worsened; i.e., the transfers are *conditional*. As such, the supervisor compares the relative performance of both components by checking the mixing factor $\lambda(i)$ every L iterations and enable the transfers only when necessary. Succinctly, the T-OE is described in Algorithm 1 [28].

Algorithm 1 Transversal FIR–Output Error Filters Combination (T-OE)

for $i = 0$, $i{+}{+}$ **do**

$$
\begin{aligned}
y_{1,i} &= u_i w_{1,i-1} \\
e_{1,i} &= d(i) - y_1(i) \\
w_{1,i} &= w_{1,i-1} + \mu_1 u_i^T e_1(i) \\
w_{2,i-1} &= \begin{cases} \delta\mathcal{F}(w_{1,i-1}) + \big(1-\delta\big) w_{2,i-1}, & \lambda(i) \geq \lambda \\ w_{2,i-1}, & \lambda(i) < \lambda \end{cases} \\
y_{2,i} &= x_i w_{2,i-1} \\
\phi_i &= \frac{x_i}{1 - A(z, i-1)} \\
e_{2,i} &= d(i) - y_2(i) \\
w_{2,i} &= w_{2,i-1} + \frac{\mu_2}{\|\phi_i\|^2 + \epsilon} \phi_i^T e_2(i) \\
y(i) &= \lambda(i) y_1(i) + \big(1 - \lambda(i)\big) y_2(i)
\end{aligned}
$$

end for

In Algorithm 1, the subscript $n = 1, 2$ identifies a measure related to the AF n, $u_i = [u(i)\ u(i-1)\ \ldots u(i-M_1)]$ is the regressor of AF1, M_n is the order of the AF n, $x_i = [y_2(i-1)\ y_2(i-2)\ \ldots\ y_2(i-M_2)\ u(i)\ u(i-1)\ \ldots u(i-M_2)]$ and ϕ_i are the ordinary and filtered regressors of AF2, μ_n is the step size, $w_{n,i}$ is the parameters vector, δ denotes $\delta(i-kL)$ (i.e., the discrete impulse shifted by kL samples), $k \in \mathbb{N}$, L is the cycle length, $\lambda(i)$ is the adaptive mixing factor (Eq. 3.2), $0 \ll \lambda < 1$ is a real constant that establishes for which range of $\lambda(i)$ AF1 is

considered better than AF2 (0.7 is a typical value), and $\mathcal{F}$ is a *mapping function* that converts the FIR coefficients into the IIR domain, such that

$$H_2(z) \leftarrow \mathcal{F}\big(H_1(z)\big) \tag{4.2}$$

4.2.1 Considerations on the Mapping Functions $\mathcal{F}$

The implementation of $\mathcal{F}$ considers $M_2 < M_1$ and, therefore, can be seen as an order reduction procedure. Typically, the need for compact descriptions like those offered by $\mathcal{F}$ arises from the increasingly complex problems often found in several mathematical and engineering processes. In many applications as optimal control or simulation, the goal is related to the creation of simple models in which accuracy is deliberately traded for numerical simplicity: as long as the properties of interest of the original processes are retained with sufficient realism for analysis purposes, it does not really matter if the reduced model is not fully accurate [94].

The motivation for $\mathcal{F}$ in the T-OE context is different because the overall complexity will be *constant* as both components operate all the time – clearly, the goal is to accelerate the convergence of AF2 to *gain* accuracy instead of trading it for a lower complexity as in the regular order reduction applications. However, if necessary, the implementation of $\mathcal{F}$ can be simplified at the expense of the accuracy because the mapping errors of simpler algorithms can be compensated to some extent by the adaptive capabilities of AF2 [4] .

In this work, $\mathcal{F}$ is implemented via the Balanced Model Reduction (computationally more elaborated, intended to the design of fixed digital filters) or the Padè approximants built through the method of Prony (simpler, normally used to expand a function as a ratio of two power series), described respectively in

[4] Hankel optimal model reduction is perhaps one of the most powerful model reduction options available [95], but it is computationally demanding and requires $O(M^3)$ flops, which makes it impractical for real-time identification of systems of a fair order. It also requires a perfect description of the system as given by its parameters or frequency response.

Sections 4.2.2 and 4.2.3.

4.2.2 Balanced Model Reduction (BMR)

The Balanced Model Reduction (or BMR) is a class of state-space based methods that apply certain linear transformations to a system $H(z)$ in order to make it suitable for model reduction. As originally conceived, BMR defines a balanced realization for $H(z)$ by building diagonal observability and controllability Grammians made by the singular values of Γ_H, the Hankel form of $H(z)$ [5] [97]. This amounts to project $H(z)$ onto a basis where the states related to the smallest singular values are also harder to reach and observe. A reduced system is then produced by truncating these states, therefore assuming that the largest singular values matter the most [98, 99]. For this reason, BMR is also called Truncated Balanced Realization (TBR) [94,99] or Balanced Model Truncation (BMT) [39].

This idea of reducing a high-dimensional space by exploiting the decomposition of some symmetric matrix is at the core of many statistical tools too. Principal Component Analysis (PCA), for example, takes the covariance matrix of a large dataset, perform its eigen-decomposition and defines a reduced subspace by truncating the eigen-vectors related to the smallest eigenvalues [100]. Either with PCA or BMR, the bound for the approximation error is given by

$$\zeta_{BMR} = \frac{\sum_{k_2=M_2+1}^{M_1} \sigma_{k_2}}{\sum_{k_1=0}^{M_1} \sigma_{k_1}}, \tag{4.3}$$

where σ_k is the k-th largest singular value of that matrix and M_1 and M_2 are the orders of the original and the reduced systems (in our case, $H_1(z)$ and $H_2(z)$).

[5] Let $H(z)$ be an stable linear system described in the space-state by $\{\mathsf{A}, \mathsf{b}, \mathsf{c}, \mathsf{d}\}$. Then, its Observability and Controllability Grammians $\mathcal{G}_\mathcal{O}$ and $\mathcal{G}_\mathcal{C}$ are such that they will be positive definite if and only if the pair $\{\mathsf{A}, \mathsf{c}\}$ (resp. $\{\mathsf{A}, \mathsf{b}\}$) is completely observable (resp. completely controllable) [96]. Moreover, the singular values of the Hankel matrix Γ_H are the square roots of the eigenvalues of $\mathcal{G}_\mathcal{O}\mathcal{G}_\mathcal{C}$ and, if $\mathcal{G}_\mathcal{O}$ and $\mathcal{G}_\mathcal{C}$ are diagonal and identical, the realization becomes *internally balanced* [12] and well-behaved numerically, which is convenient for model reduction.

In this work, we use the BMR algorithm developed by Beliczynski et al [30], which avoids the explicit handling of any observability and controllability Grammians by decomposing directly the Hankel form $\Gamma_{H_1} \in \mathbb{R}^{M_1 \times M_1}$ defined as

$$\Gamma_{H_1} = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \cdots & \cdots & h_{1,M1-1} & h_{1,M1} \\ h_{1,2} & h_{1,3} & h_{1,4} & \cdots & h_{1,M1-1} & h_{1,M1} & 0 \\ h_{1,3} & h_{1,4} & h_{1,5} & \cdots & h_{1,M1} & 0 & 0 \\ \vdots & \vdots & \vdots & \iddots & 0 & 0 & 0 \\ \vdots & \vdots & h_{1,M1} & \cdots & 0 & 0 & 0 \\ h_{1,M1-1} & h_{1,M1} & 0 & \cdots & 0 & 0 & 0 \\ h_{1,M1} & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}, \tag{4.4}$$

where $\{h_{1,k}\}_{k=1}^{M1}$ are the samples of the impulse response of a FIR filter ($H_1(z)$) at the instant $i-1$ (note that $h_{1,0}(i-1)$ is omitted in Γ_{H_1}). While this straightforward strategy circumvents the ill-conditioning of the matrices normally used in the balancing methods by successfully avoiding any matrix inversions, it is not general and works *only* when $H_1(z)$ is FIR as demonstrated by the authors. On the other hand, the resulting low-dimensional system $H_2(z)$ will be always IIR.

The algorithm starts with the eigen-decomposition of Γ_{H_1} (Eq. 4.4) as

$$\Gamma_{H_1} = U \Lambda U^T \tag{4.5}$$

Next, the parameters set $\{\mathsf{A}_{i-1}, \mathsf{b}_{i-1}, \mathsf{c}_{i-1}, \mathsf{d}(i-1)\}$ that describes $H_2(z)$ in the state-space at the instant $i-1$ is given by

$$\mathsf{A}_{i-1} = [U]_{k_1,k_2}^T [U]_{k_3,k_2} \tag{4.6}$$

$$\mathsf{b}_{i-1} = [U]_{1,k_2}^T \tag{4.7}$$

$$\mathsf{c}_{i-1} = w_1' [\Lambda]_{k_4,k_2} \tag{4.8}$$

$$\mathsf{d}(i-1) = h_{1,0} \tag{4.9}$$

where $w'_{1,i-1} = [h_{1,1}(i-1)\ h_{1,2}(i-1)\ ...\ h_{1,M1}(i-1)]$ (i.e., $w_{1,i-1}$ with $h_{1,0}$ ommited),

$k_1 = \{2\ 3\ \cdots M_1\}$, $k_2 = \{1\ 2\ \cdots\ M_2\}$, $k_3 = \{1\ 2\ \cdots\ M_1 - 1\}$, $k_4 = \{1\ 2\ \cdots\ M_1\}$ and $[.]_{r,s}$ is a submatrix of the argument, made of the rows and columns indicated by r and s, respectively.

In case no order reduction is performed (i.e., $M_2 = M_1$), the resulting parameters $\{\mathsf{A}_{i-1}, \mathsf{b}_{i-1}, \mathsf{c}_{i-1}\}$ can be partitioned as

$$\mathsf{A}_{i-1} = \begin{bmatrix} \mathsf{A}_{1,1} & \mathsf{A}_{1,2} \\ \mathsf{A}_{2,1} & \mathsf{A}_{2,2} \end{bmatrix} \tag{4.10}$$

$$\mathsf{b}_{i-1} = \begin{bmatrix} \mathsf{b}_1 \\ \mathsf{b}_2 \end{bmatrix} \tag{4.11}$$

$$\mathsf{c}_{i-1} = \begin{bmatrix} \mathsf{c}_1 & \mathsf{c}_2 \end{bmatrix}, \tag{4.12}$$

such that the parameters sets $\{\mathsf{A}_{1,1}, \mathsf{b}_1, \mathsf{c}_1\}$ and $\{\mathsf{A}_{2,2}, \mathsf{b}_2, \mathsf{c}_2\}$ (the index $i-1$ was omitted in the partitions for clarity) give rise to the subsystems

$$\begin{bmatrix} \mathcal{W}_{R,i+1} \\ y_R(i) \end{bmatrix} = \begin{bmatrix} \mathsf{A}_{1,1} & \mathsf{b}_1 \\ \mathsf{c}_1 & \mathsf{d} \end{bmatrix} \begin{bmatrix} \mathcal{W}_{R,i} \\ u(i) \end{bmatrix} \tag{4.13}$$

$$\begin{bmatrix} \mathcal{W}_{D,i+1} \\ y_D(i) \end{bmatrix} = \begin{bmatrix} \mathsf{A}_{2,2} & \mathsf{b}_2 \\ \mathsf{c}_2 & \mathsf{d} \end{bmatrix} \begin{bmatrix} \mathcal{W}_{D,i} \\ u(i) \end{bmatrix}, \tag{4.14}$$

In the general case though, $M_2 < M_1$ and clearly the reduction procedure results into the system in Eq. 4.13 whose order is M_2. Also, it is straightforward to notice that the discarded system given by Eq. 4.14 is related to the smallest singular values of Γ_{H_1}. It can be shown that both systems are stable and balanced.

Eq. 4.13 can be converted into the transfer function form by applying the Z-Transform as [101]

$$\mathcal{W}_{R,i+1} = \mathsf{A}_{1,1}\mathcal{W}_{R,i} + \mathsf{b}_1 u(i) \implies z\mathsf{W}_R(z) = \mathsf{A}_{1,1}\mathsf{W}_R(z) + \mathsf{b}_1\mathsf{U}(z) \tag{4.15}$$

$$y_R(i) = \mathsf{c}_1\mathcal{W}_{R,i} + \mathsf{d}(1)u(i) \implies \mathsf{Y}_R(z) = \mathsf{c}_1\mathsf{W}_R(z) + \mathsf{d}(1)\mathsf{U}(z) \tag{4.16}$$

Developing Eq. 4.15 yields

$$(z\mathbb{I} - \mathsf{A}_{1,1})\mathsf{W}_R(z) = \mathsf{b}_1\mathsf{U}(z) \tag{4.17}$$

$$\mathsf{W}_R(z) = (z\mathbb{I} - \mathsf{A}_{1,1})^{-1}\mathsf{b}_1\mathsf{U}(z) \tag{4.18}$$

If Eq. 4.18 is replaced into Eq. 4.16,

$$\mathsf{Y}_R(z) = \mathsf{c}_1(z\mathbb{I} - \mathsf{A}_{1,1})^{-1}\mathsf{b}_1\mathsf{U}(z) + \mathsf{d}(1)\mathsf{U}(z) \tag{4.19}$$

and, therefore, we conclude that

$$H_2(z) = \frac{\mathsf{Y}_R(z)}{\mathsf{U}(z)} = \mathsf{c}_1(z\mathbb{I} - \mathsf{A}_{1,1})^{-1}\mathsf{b}_1 + \mathsf{d}(1) \tag{4.20}$$

Succinctly, the BMR-based mapping $\mathcal{F} : H_1(z) \rightarrow H_2(z)$ can be described as

1 Build the Hankel form Γ_{H_1} of $H_1(z)$ (Eq. 4.4);

2 Factorize Γ_{H_1} into the factors U and Λ (Eq. 4.5);

3 Find the parameters set $\{\mathsf{A}_{1,1}, \mathsf{b}_1, \mathsf{c}_1, \mathsf{d}\}$ from U and Λ (Eq. 4.13);

4 Determine $H_2(z)$ from the set $\{\mathsf{A}_{1,1}, \mathsf{b}_1, \mathsf{c}_1, \mathsf{d}\}$ (Eq. 4.20).

Note that this BMR method avoided any matrix inversions only in the state-space (steps **1-3**). To yield the corresponding transfer function, step **4** requires either the inversion of $(z\mathbb{I} - \mathsf{A}_{i-1}) \in \mathbb{R}^{M_2 \times M_2}$ or, alternatively, the computation of its determinant and adjoint matrix since Eq. 4.20 may be restated as [102] [6]

$$\begin{aligned} H_2(z) &= \mathsf{c}_1(z\mathbb{I} - \mathsf{A}_{1,1})^{-1}\mathsf{b}_1 + \mathsf{d}(1) \\ &= \frac{\mathsf{c}_1 \; adj \; (z\mathbb{I} - \mathsf{A}_{1,1})\mathsf{b}_1 + \mathsf{d}(1)|z\mathbb{I} - \mathsf{A}_{1,1}|}{|z\mathbb{I} - \mathsf{A}_{1,1}|} \end{aligned} \tag{4.21}$$

[6] In the same fashion as the characteristic polynomial $\lambda\mathbb{I} - \mathbf{A}$ in Linear Algebra, the polynomial matrix $(z\mathbb{I} - \mathsf{A}_{1,1})$ is recurrent in several problems. Leverrier's algorithm succeeds in producing a transfer function without inverting $(z\mathbb{I} - \mathsf{A}_{1,1})$ [103,104]; however, this is achieved at the expense of a great computational burden, being impractical for real-time systems.

Also, the method resorts to an eigen-decomposition in step **2**. However, given the particular conformation of the Hankel matrix specified in Eq. 4.4, one could employ the fast eigen-decomposition algorithm proposed by [105], which manages to reduce the complexity of the task from $O\big((M_1+1)^3\big)$ to $O\big((M_1+1)^2\log(M_1+1)\big)$. Notwithstanding, there are other techniques that could take advantage of the *triangularity* besides the symmetry of the Hankel matrix, as discussed in [106] and [107]. A more efficient eigen-decomposition solution is an open problem in our research and it is currently under investigation.

4.2.3 The Padè Approximants Method (PAM)

Padè approximants (PAs) are rational functions whose power series expansions match a given power series to the highest possible order. Since they are formed by a ratio of two polynomials, they are far more powerful than the Taylor series expansions and may converge even when the latter does not, particularly when the function being approximated is endowed with poles [108, 109].

In order to derive the approximants, the transfer functions of the adaptive filters are thought in terms of their Z Transforms and the mixed notation is not used. This way, the transfer function of AF2 is denoted by

$$\mathbb{H}_2(z) = \frac{B(z)}{1+A(z)} \tag{4.22}$$

and not by

$$H_2(z) = \frac{B(z)}{1-A(z)}, \tag{4.23}$$

which will allow a better understanding of how the approximants are produced.

Let an arbitrary (stable) rational form as $\mathbb{H}_2(z)$; i.e.,

$$\mathbb{H}_2(z) = \frac{B(z)}{1+A(z)} = \frac{\sum_{k=0}^{M_2} b_k z^k}{\sum_{k=0}^{M_2} a_k z^k} = \sum_{k=0}^{\infty} h_{2,k} z^k, \tag{4.24}$$

be an approximation for a polynomial like the transfer function of AF1; i.e.,

$$\mathbb{H}_1(z) = \sum_{k=0}^{M_1} h_{1,k} z^k, \tag{4.25}$$

where M_1 is infinite in the general case.

Then, $\mathbb{H}_2(z)$ is said a *Padè approximant* of $\mathbb{H}_1(z)$ if and only if [108]

$$\mathbb{H}_2(z) - \mathbb{H}_1(z) = O\big(z^{2M_2+1}\big) \implies \begin{cases} \mathbb{H}_2(0) = \mathbb{H}_1(0) \\ \frac{d^k}{dz^k}\mathbb{H}_2(z)\Big|_{z=0} = \frac{d^k}{dz^k}\mathbb{H}_1(z)\Big|_{z=0} \end{cases} \quad 0 < k \leq 2M_2, \tag{4.26}$$

where $O\big(z^{2M_2+1}\big)$ denotes that the monomial of lowest order in the difference $\mathbb{H}_2(z) - \mathbb{H}_1(z)$ has order equal or greater than $2M_2 + 1$ and, therefore,

$$h_{2,k} = h_{1,k} \text{ for } 0 \leq k \leq 2M_2 \tag{4.27}$$

Since Eq. 4.26 states that $\mathbb{H}_2(z) - \mathbb{H}_1(z) = \frac{B(z)}{1+A(z)} - \mathbb{H}_1(z) = O\big(z^{2M_2+1}\big)$, then

$$\begin{aligned} O\big(z^{2M_2+1}\big) &= B(z) - \mathbb{H}_1(z)\big(1 + A(z)\big) \\ &= \sum_{k=0}^{M_2} b_k z^k - \left(\sum_{k_1=0}^{M_1} h_{1,k_1} z^{k_1}\right)\left(\sum_{k_2=0}^{M_2} a_{k_2} z^{k_2}\right) \end{aligned} \tag{4.28}$$

By its definition, Eq. 4.28 states nothing about the monomials with powers higher than $2M_2$ and can be used only to solve the lower order monomials. These can be determined by considering the terms of the rightmost polynomial whose powers $k_1 + k_2$ are smaller or equal to $2M_2$, such that the corresponding coefficients are

either equal to 0 (when $M_2 < k_1 + k_2 \leq 2M_2$) or $b_{k_1+k_2}$ ($0 \leq k_1 + k_2 \leq M_2$); i.e.,

$$\sum_{j_1=1}^{M_2} \sum_{j_2=0}^{M_2} a_{j_2} h_{1,M_2-j_2+j_1} = 0 \tag{4.29}$$

$$\sum_{j_1=0}^{M_2} \sum_{j_2=0}^{j_1} a_{j_2} h_{1,j_1-j_2} = b_{j_1} \tag{4.30}$$

Equations 4.29 and 4.30 define an undetermined linear system with $2M_2 + 1$ equations and $2M_2 + 2$ unknowns ($\{b_k\}_{k=0}^{M2}$ and $\{a_k\}_{k=0}^{M2}$), which can not be solved unless an extra equation as the usual monic constraint $a_0 = 1$ is provided.

By using the monic constraint, this system can be handled by first solving the top M_2 equations for $\{a_k\}_{k=1}^{M_2}$ in terms of $\{h_k\}_{k=1}^{2M_2}$ (Eq. 4.29), from which $\{b_k\}_{k=0}^{M_2}$ can then be computed by using the second set of $M_2 + 1$ equations (Eq. 4.30). In a matrix form, all these equations can be expressed as

$$\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{M2-1} \\ b_{M2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} h_{1,0} & 0 & 0 & 0 & 0 \\ h_{1,1} & h_{1,0} & 0 & 0 & 0 \\ h_{1,2} & h_{1,1} & h_{1,0} & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ h_{1,M_2} & h_{1,M_2-1} & \dots & h_{1,1} & h_{1,0} \\ h_{1,M_2+1} & h_{1,M_2} & \dots & h_{1,2} & h_{1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{1,M_1} & h_{1,M_1-1} & \dots & \dots & h_{1,M_1-M_2} \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ \vdots \\ \vdots \\ a_{M_2-1} \\ a_{M_2} \end{bmatrix}, \tag{4.31}$$

where $M_1 = 2M_2$ (although $M_1 > 2M_2$ is useful in some cases [110]) and the samples $\{h_{1,k}\}_{k=0}^{M_1}$ are stored into a $\mathbb{R}^{M_1+1 \times M_2+1}$ Toeplitz matrix denoted by $\mathbb{T}$.

Once the parameters $\{b_k\}_{k=0}^{M_2}$ and $\{a_k\}_{k=1}^{M_2}$ are computed separately, Eq. 4.31 can be conveniently partitioned, splitting into two systems given by

$$\begin{bmatrix} b \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbb{T}_1 \\ \mathbb{T}_2 \end{bmatrix} \begin{bmatrix} 1 \\ a \end{bmatrix}, \tag{4.32}$$

where $b = [b_0 \ \dots \ b_{M_2}]^T \in \mathbb{R}^{M_2+1\times 1}$, $a = [a_1 \ \dots \ a_{M_2}]^T \in \mathbb{R}^{M_2\times 1}$, $\mathbb{T}_1 \in \mathbb{R}^{M_2+1\times M_2+1}$ stores the first M_2+1 rows of $\mathbb{T}$ and $\mathbb{T}_2 \in \mathbb{R}^{M_2\times M_2+1}$ stores the remainder ones. This procedure, known as the *Prony's method*, is exact and yields the Padè Approximants referred in Eq. 4.26 as long as $M_1 \geq 2M_2$ [97, 110, 111].

By using $\mathbb{T}_2$, the parameters a_k can be determined via the system

$$-\begin{bmatrix} h_{1,M_2+1} \\ h_{1,M_2+2} \\ \vdots \\ h_{1,2M_2-1} \\ h_{1,2M_2} \end{bmatrix} = \begin{bmatrix} h_{1,M_2} & h_{1,M2-1} & \cdots & h_{1,2} & h_{1,1} \\ h_{1,M_2+1} & h_{1,M2} & \cdots & h_{1,3} & h_{1,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{1,2M_2-2} & h_{1,2M_2-3} & \cdots & h_{1,M_2} & h_{1,M_2-1} \\ h_{1,2M_2-1} & h_{1,2M_2-2} & \cdots & h_{1,M_2+1} & h_{1,M_2} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{M_2-1} \\ a_{M_2} \end{bmatrix} \tag{4.33}$$

where the constant terms are the first column of $\mathbb{T}_2$ and the remainder columns are stored into a $\mathbb{R}^{M_2\times M_2}$ coefficients matrix denoted by $\mathbb{T}_3$ [7] . Once this system is solved, the parameters b_k are found via backward substitution using $\mathbb{T}_1$ as

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_{M_2-1} \\ b_{M_2} \end{bmatrix} = \begin{bmatrix} h_{1,0} & 0 & 0 & \cdots & 0 & 0 \\ h_{1,1} & h_{1,0} & 0 & \cdots & 0 & 0 \\ h_{1,2} & h_{1,1} & h_{1,0} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ h_{1,M_2-1} & h_{1,M_2-2} & h_{1,M_2-3} & \cdots & h_{1,0} & 0 \\ h_{1,M_2} & h_{1,M_2-1} & h_{1,M_2-2} & \cdots & h_{1,1} & h_{1,0} \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_{M_2-1} \\ a_{M_2} \end{bmatrix} \tag{4.34}$$

The system in Eq. 4.33 yields a single, exact solution as long as $\mathbb{T}_3$ is non-singular; i.e., its rank must be M_2. In case $\mathbb{T}_3$ is singular, the system is undetermined and will lead to multiple solutions, indicating that $H_1(z)$ can actually be

[7] As $\mathbb{T}_3$ is Toeplitz, it becomes Hankel if its columns are flipped from left-to-right. So, by handling Eq. 4.33 conveniently, it can be shown that the parameters a_k (gathered in an upside-down version of a) lie in the null space of that Hankel matrix. As mentioned in Appendix B.1, this is expected because *virtually all system identification algorithms boil down to fit a certain parametric vector into the null space of a Hankel form built from the system to identify* [12].

reproduced by a lower order IIR filter. Hence, assuming sufficient modelling scenarios, in the FIR→IIR weights transfers context $\mathbb{T}_3$ will not be singular provided that AF1 reproduces the impulse response of $H^o(z)$ with a sufficient accuracy *and* $M_1 \geq 2M_2$, such that the dimensions of $\mathbb{T}$ fit the system in Eq. 4.31.

It can be seen that the order of the systems involved in this procedure is M_2, not M_1 as it is the case of BMR (refer to Section 4.2.2). As the mapping $\mathcal{F} : H_1(z) \rightarrow H_2(z)$ assumes that $M_2 < M_1$, the method of Prony is clearly simpler. In fact, the solution for Toeplitz systems can be found in $O\big((M_2+1)^2\big)$ with the *look ahead* versions of the Levinson, Schur or Bareiss algorithms (e.g., see [112, 113]). Moreover, some divide-and-conquer strategies intended for large Toeplitz or Hankel systems (known as *superfast methods*) can feature a complexity as low as $O\big((M_2+1)\log(M_2+1)\big)$ (see [114–116]).

4.2.4 FIR → IIR Mapping Assessment

As described in Sections 4.2.2 and 4.2.3, the BMR and PAM are functionally equivalent and in principle can be used interchangeably to implement the mapping $\mathcal{F}$. Once certain environmental aspects that could explain the preference for a method to the detriment of the other (such as the existence of sharp resonances or the SNR of the system) are generally unavailable in advance, there is no further information to support that choice but the intrinsic properties of each method. In another context [8] , Gurgecin and Antoulas [97] point out three major criteria that an order reduction algorithm should meet in order to yield a suitable model:

1 preserve fundamental properties of the system as the stability;

2 be computationally efficient;

[8] In opposition to the context of the traditional order reduction applications, where the goal is to create a compact model for a system whose description is usually available, here the mapping $\mathcal{F}$ resorts to an *outline* of the unknown plant given by the impulse response of AF1.

3 keep the approximation errors low in some sense, such that if $H_2(z) = \mathcal{F}\big(H_1(z)\big)$ then ideally $\big(H_1(z) - H_2(z)\big)u(i) \approx 0$.

Criterion **1** is debatable in the context of the T-OE because AF2 is adaptive and continues to evolve after the weights transfers, improving its estimates with respect to the input-output behaviour of the plant, not AF1. Considering that AF1 undermodels $H^o(z)$, the impulse response of the former is at best a "sketch" of the later and hence to make AF2 reproduce $H_1(z)$ accurately may be pointless because what really matters is to update the parameters vector $w_2(i)$ towards w^o, even if some properties of $H_1(z)$ are lost. This rule of thumb still holds when the approximation is unstable, as long as AF2 keeps exponentially stable and the error $||H^o(z) - H_2(z)||$ decreases somehow for some conveniently defined norm.

Considerations about computational efficiency (Criterion **2**) are quite relative here because, typically, just a few transfers are performed in a stationary environment as seen in Section 4.2.7. Regarding the complexity of BMR and PAM themselves, specific measures related to the orders of AF1 and AF2 are supplied in Sections 4.2.2 and 4.2.3, where it is seen that the latter is considerably simpler.

As of the approximation errors associated to the mapping methods (Criterion **3**), evaluation is somewhat tricky. For example, BMR can be quite sensitive to the order of AF1 and, as a general rule, the larger the order the better. However, this principle opposes the idea of bounding the overall complexity and ignores that the performance of the FIR AFs may worsen with overmodelling [9] [56], not to mention that the Hankel singular values are decaying [117] and, therefore, a large M_1 may be meaningless. On the other hand, PAM requires only the first $2M_2+1$ coefficients of the FIR filter and simply neglects the remaining. Notwithstanding,

[9] In principle, no realizable FIR filter can overmodel the infinite impulse response of an IIR plant $H^o(z)$. However, if that response is fast-decaying (e.g., if $H^o(z)$ is critically damped), in practice a truly large FIR filter can overmodel the useful impulse response of $H^o(z)$. See the definition of *useful impulse response* in opposition to the actual impulse response in Section 4.1.

unlike BMR, it may lead to an unstable realization and is more sensitive to noise.

Additionally, any SVD based methods such as BMR depends on the M_2 largest Hankel singular values of $H_1(z)$, which means that the sum of the neglected singular values provides a bound for the approximation error (see Eq. 4.3) [30, 117]. In contrast, if the linear systems used to compute the Padè approximants are not fed with at least the first $2M_2+1$ samples of the impulse response of $H_1(z)$, there will be no error bound at all and the only information available is that the approximation is flawed. Nonetheless, as the exact impulse response of $H^o(z)$ is not known beforehand, such considerations are meaningless in the context of the mapping $\mathcal{F}$. Clearly, the choice between these methods is not trivial.

In order to get an insight of how the order reduction techniques behave within the T-OE context, some simulations were performed in the sufficient modelling case (i.e., considering identical orders for both $H_2(z)$ and $H^o(z)$). First, three stable plants generated at random were identified by several FIR LMS AFs of different orders, identical to the FIR components actually used in the T-OE associations; i.e., $H_1(z)$. Each one of those AFs ran multiple rounds of simulations that, thanks to the stochastic nature of the LMS, yielded slightly distinct values for w_i. These values were averaged and then transferred to an IIR N-LMSOE AF ($H_2(z)$) by using both BMR and PAM. Right after the transfers, the impulse responses of the IIR AFs were compared with that of the plant being identified.

In all cases, the learning step of AF1 was $\mu_1 = 20 \times 10^{-3}$, $M_2 = 6$, the input signal $u(i)$ was zero-mean Gaussian with power $\sigma_u^2 = 1$ and the additive noise $v(i)$ was also zero-mean Gaussian with power $\sigma_v^2 = 1 \times 10^{-3}$. Once the accuracy of the order reduction depends on the order of AF1, M_1 ranged from 8 to 50, so that the performance of the BMR could be fairly evaluated against the PAM.

The figure of merit created for the assessment was the *frequency response mismatch* $\tilde{H}(\omega)$, defined as the RMS (*Root Mean Square*) error measure

$$\tilde{H}(\omega) = 10 \log \left(\| H^o(\omega) - H_2(\omega) \| \right), \tag{4.35}$$

which is the logarithmic norm of the difference of the frequency responses of the plant $H^o(\omega)$ and the adaptive filter $H_2(\omega)$ right after a transfer from $H_1(\omega)$ (supposed at the steady state) is made. The lower the mismatch, the best the approximation quality. As a benchmark, the simulations included the frequency response mismatch of the FIR AF (i.e., AF1) given by

$$\tilde{H}(\omega) = 10 \log \left(\| H^o(\omega) - H_1(\omega) \| \right) \tag{4.36}$$

The frequency response of the random plants and the results of the assessment are shown in the charts of Figures 10-15, where *FIR App*, *IIR App (BRM)* and *IIR App (PAM)* are, respectively, the frequency response mismatches of AF1, AF2 fed via BMR and AF2 fed via PAM.

It is clear from the Figures 11, 13 and 15 that no technique is always superior to the other and they perform quite differently in different situations. However, PAM transfers present a remarkable behaviour that stays the same: its performance tends to improve consistently along the length of AF1 until $M_1 = 2M_2$ and remains unchanged from that point on, consonant to the definition of Padè approximants in Equations 4.26, 4.33 and 4.34.

The performance of the BMR transfers, in contrast, is likely to improve beyond this threshold as seen in Figures 11 and 13, but this is not a consistent pattern as Figure 15 shows. This is related to the way the Hankel singular values $\{\sigma_k\}_{k=0}^{M_1}$ decrease: the faster they decay, the more accurately AF2 can be approximated by the first components of U and Λ [30] (see Equations 4.5–4.9). Assuming sufficient modelling, in the limit the approximation error bound given

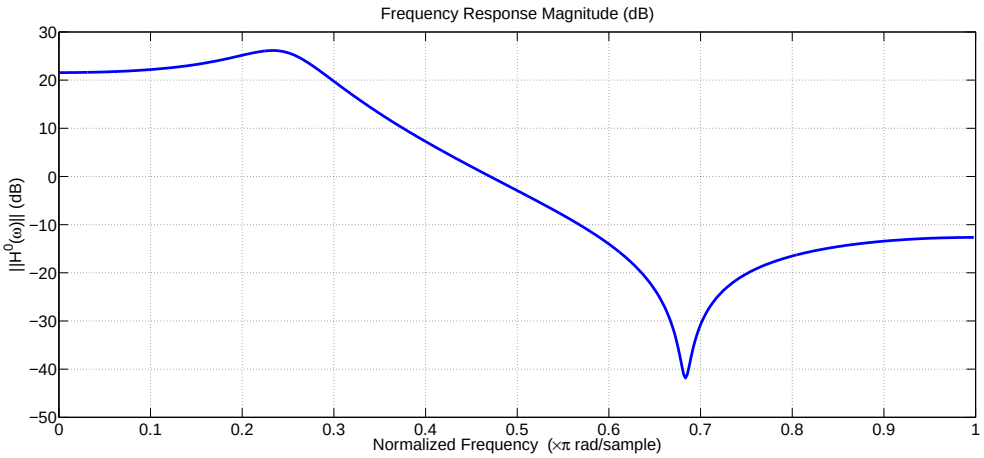

Figure 10: Frequency Response of the Random Plant (Scenario 1)

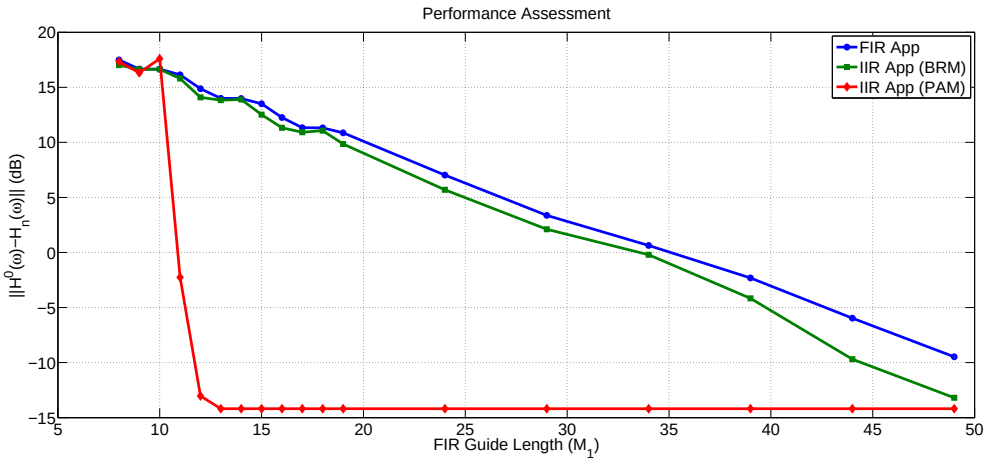

Figure 11: Performance Assessment - Transfer Error (Scenario 1)

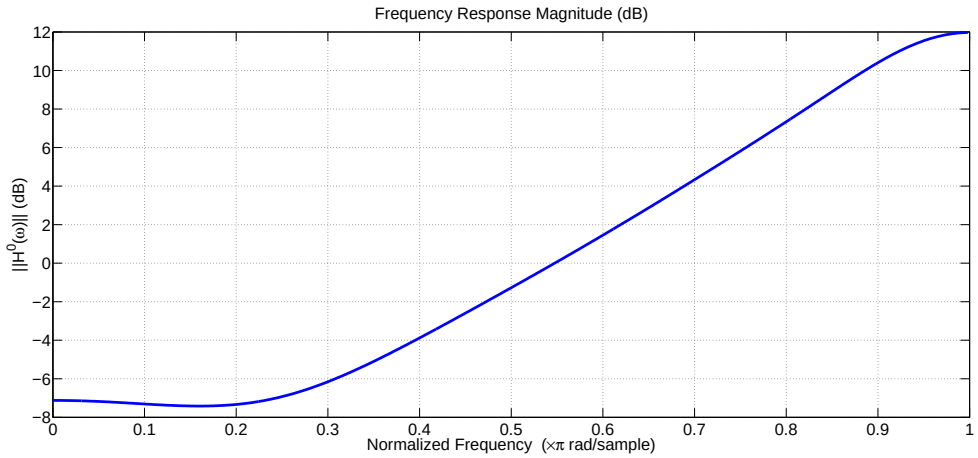

Figure 12: Frequency Response of the Random Plant (Scenario 2)

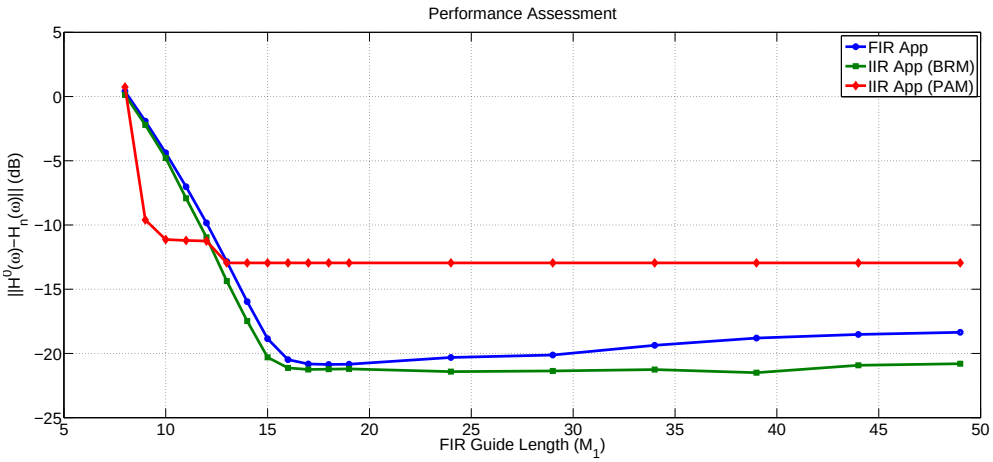

Figure 13: Performance Assessment - Transfer Error (Scenario 2)

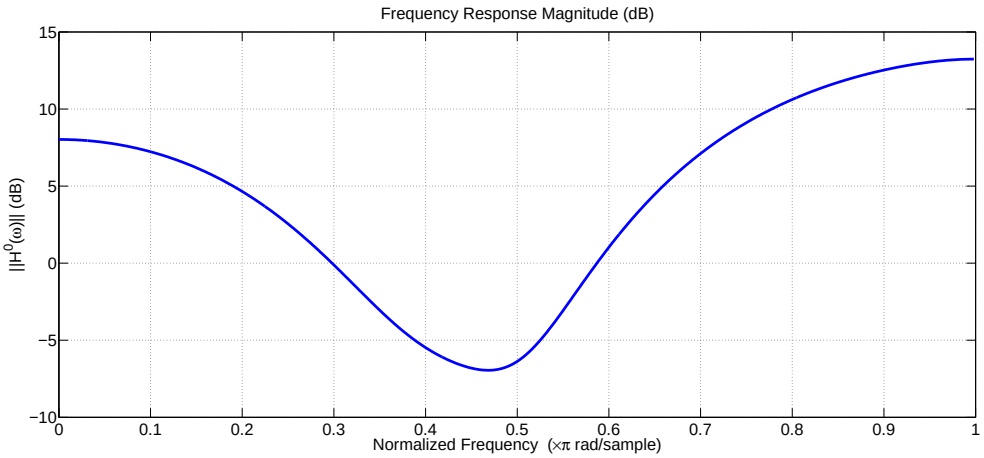

Figure 14: Frequency Response of the Random Plant (Scenario 3)

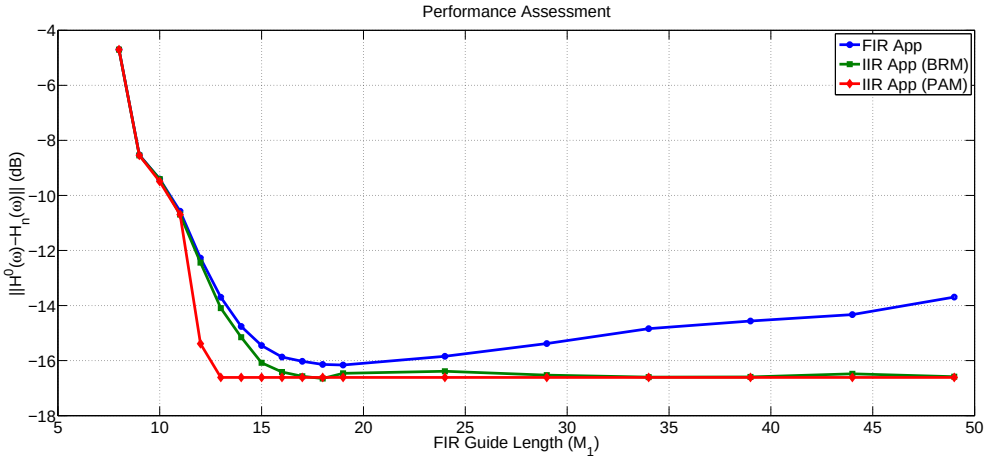

Figure 15: Performance Assessment - Transfer Error (Scenario 3)

by Eq. 4.3 approaches 0 for any value of M_2, indicating that the smallest singular values becomes negligible and, therefore, AF1 could be shorter. As a result, the frequency response mismatch becomes nearly flat after a while as seen in Fig. 15.

It is interesting to note that the frequency response mismatch of the FIR component does not improve consistently with its order M_1, as seen in Figures 13 and 15. In these cases, the frequency response mismatch of AF1 increases only slightly with M_1 but experience shows that in some particular cases it may deteriorate rapidly. Although $H^o(z)$ is IIR and AF1 is FIR, such an issue seems to be caused by overmodelling: if the impulse response of $H^o(z)$ decays considerably fast, a large AF1 could overmodel the *useful* impulse response of $H^o(z)$ (see Section 4.1). However, there is not a correspondingly drastic deterioration in the BMR transfers once the Hankel singular values that rule the construction of the balanced realizations are very often fast-decaying as noted by Antoulas et al [117].

Also, often the frequency response mismatches of AF2 are smaller than those of AF1 either by using PAM or BMR, which sounds counter intuitive at a first sight because the mapping makes AF2 model AF1, not the plant. However, in spite of being FIR and therefore subject to a large mismatch when identifying $H^o(z)$, AF1 induces an *infinite* impulse response in AF2 via the mapping $\mathcal{F}$. As such, AF2 can be expected to fit $H^o(z)$ better than AF1 very often (particularly when the impulse response of $H^o(z)$ is long tailed) as Figures 11, 13 and 15 show.

Ultimately, distinct frequency response mismatches do not necessarily lead to different performances in the T-OE because, as aforementioned, AF2 is adaptive and could compensate such gaps. PAM transfers tend to be more accurate when M_1 is small (i.e., $M_1 \leq 2M_2 + 1$) while BMR transfers could benefit from bigger values for M_1. Hence, if computational cost is an issue, PAM is the best option not only because it requires shorter guide filters but also because its complexity can be much smaller than that of BMR as discussed in Sections 4.2.2 and 4.2.3.

4.2.5 Cycle Length Estimate

Fig. 16 shows how the cycle length L impacts the convergence time of the T-OE used to identify the notch whose frequency response is shown in Fig. 18(a) (see Section 4.2.7).

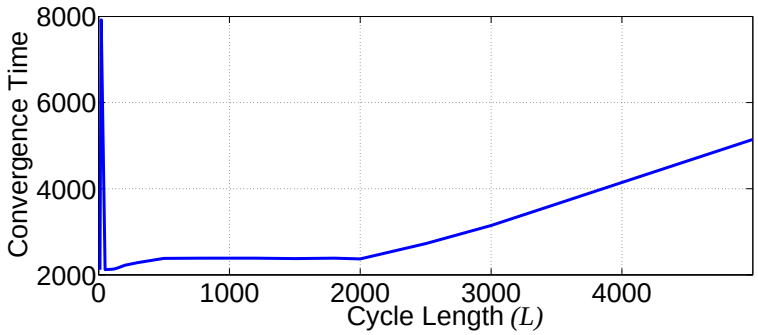

Figure 16: Convergence Time X Cycle Length for the Notch Scenario

In the plot of Fig. 16, the abscissae axis represents L and the ordinates axis stands for the convergence time, defined as the time (in iterations) required to make the T-OE reach 85% of the EMSE reached by an standalone IIR AF configured exactly as AF2 and processing the same plant with similar datasets.

Extensive simulations with several distinct scenarios have corroborated the pattern shown in Fig. 16, in which too small values for L delay the convergence. Experimental evidences suggest this is a result of the representation errors committed by the order reduction procedure added to the high estimation errors suffered by AF1 when its adaptation is just starting. On the other hand, excessively large values postpone unnecessarily the transfers, also delaying the convergence.

Fig. 16 also shows that the convergence time is about the same for an appreciably wide range of values for L ($100 < L < 2000$), suggesting that L has not an optimal *value* but an optimal *range*. Although less evident, it is also possible to infer that, in the stationary case, a single weights transfer will be sufficient if it is made *after* AF1 converges.

Given that AF1 is FIR LMS, an analytical framework can be derived to describe its adaptation in terms of σ_d^2, σ_u^2 and σ_v^2 (the powers of the signals $d(i)$, $u(i)$ and $v(i)$, respectively), which is accomplished via ECR (Energy Conservation Relation) and VR (Variance Relation) arguments [5]. Herewith, simple linear models are generated for the transient and steady state of AF1, so that L is computed by comparing both as seen in Fig. 17.

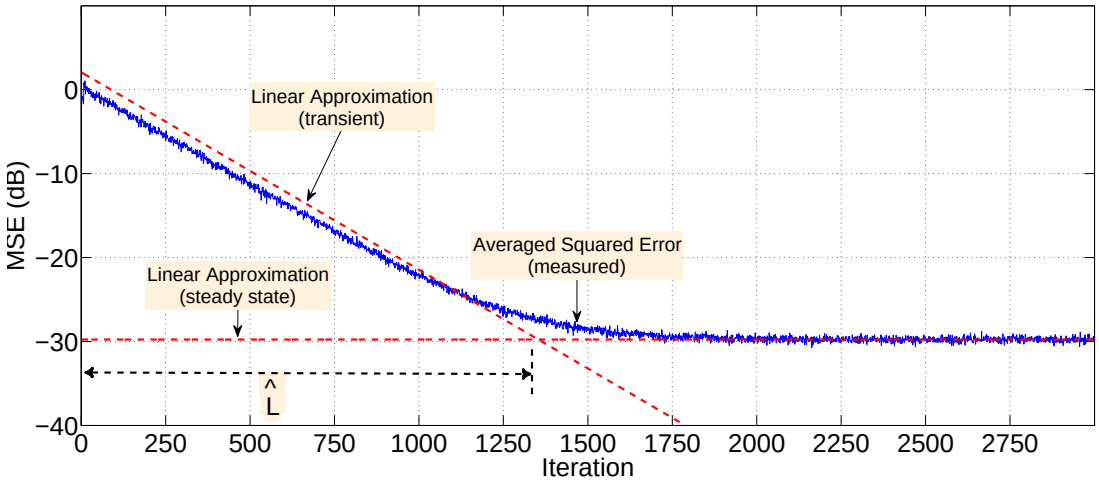

Figure 17: Linear Approximation for The FIR Guide Identifying The Notch

Let w_1^o be the optimal value for the parameters of AF1 (given by the first M_1+1 samples of the impulse response of the plant) and the weight error vector $\tilde{w}_{1,i}$ defined as $\tilde{w}_{1,i} = w_1^o - w_{1,i}$. By exploiting the weighted Variance Relation for adaptive algorithms of the form $w_{1,i} = w_{1,i-1} + \mu_1 u_i \big(g(u_i)\big)^{-1} e_1(i)$ with $g(u_i) = 1$ in the case of a regular FIR LMS AF, it can be shown that [5, Theorem 22.4]

$$\mathrm{E}\,\|\tilde{w}_i\|_W^2 = \mathrm{E}\,\|\tilde{w}_{i-1}\|_{W'}^2 + \mu_1\sigma_v^2\,\mathrm{E}\,u_i W u_i^T \tag{4.37}$$

$$W' = W - \mu_1 W(\mathrm{E}\,U_i) - \mu_1(\mathrm{E}\,U_i)W + \mu_1^2(\mathrm{E}\,U_i W U_i), \tag{4.38}$$

where $U_i = u_i^T u_i$, W is a positive semi-definite weighting matrix and $\|.\|_W^2$ is the L2 norm of the argument weighted by W (e.g., $\|w_i\|_W^2 = w_i^T W w_i$). As u_i is white, then $\mathrm{E}\,U_i = \mathrm{E}\,u_i^T u_i = \sigma_u^2\mathbb{I}$ ($\mathbb{I}$ is the identity matrix). So, by making $W = \mathbb{I}$, $\mathrm{E}\,u_i W u_i^T = \sigma_u^2(M_1+1)$ and $\mathrm{E}\,U_i W U_i = \sigma_u^2\mathrm{Tr}(\sigma_u^2\mathbb{I})\mathbb{I} + 2\sigma_u^4\mathbb{I} = \sigma_u^4(M_1+3)\mathbb{I}$ [29]. If

γ is defined as

$$\gamma \triangleq 1 - 2\mu\sigma_u^2 + \mu_1^2\sigma_u^4(M_1 + 3) \tag{4.39}$$

then, by using mathematical induction on Eq. 4.37, it is possible to conclude that

$$\mathrm{E}\,\|\tilde{w}_{1,i-1}\|^2 = \gamma^i\,\mathrm{E}\,\|w_1^o\|^2 + \mu_1\sigma_v^2\sigma_u^2(M_1+1)\sum_{k=0}^{i-1}\gamma^k, \tag{4.40}$$

where $\tilde{w}_{1,i} = w_1^o - w_{1,i}$. In case the SNR is high and the step size μ_1 is moderate, the rightmost term of Eq. 4.40 can be truncated, leading to

$$\mathrm{E}\,\|\tilde{w}_{1,i-1}\|^2 \approx \gamma^i\,\mathrm{E}\,\|w_1^o\|^2 \tag{4.41}$$

Given that the estimation error for AF1 is $e_1(i) = u_i\tilde{w}_{1,i-1} + v(i)$ with $u_i = [u(i)\ u(i-1)\ldots u(i-M_1)]$, then $\mathrm{E}\,e_1^2(i) = \sigma_u^2\,\mathrm{E}\,\|\tilde{w}_{i-1}\|^2 + \sigma_v^2$ and, for a high SNR, $\mathrm{E}\,e_1^2(i) \approx \sigma_u^2\,\mathrm{E}\,\|\tilde{w}_{1,i-1}\|^2$. Therefore, by considering that $\sigma_u^2\|w_1^o\| = \sigma_d^2 - \sigma_v^2$ [5],

$$\mathrm{E}\,e_1^2(i) \approx \gamma^i(\sigma_d^2 - \sigma_v^2) \tag{4.42}$$

and then the transient of AF1 can be linearized in a logarithmic scale as

$$MSE_{dB}(i) \cong 10i\log\gamma + 10\log(\sigma_d^2 - \sigma_v^2) \tag{4.43}$$

For a white Gausian $u(i)$, a linear approximation for the steady state of AF1 is

$$S_{AF1} = 10\log\left(\frac{2\sigma_v^2(1-\mu\sigma_u^2)}{2-\mu(M_1+3)\sigma_u^2}\right) \approx 10\log(\sigma_v^2) \tag{4.44}$$

Finally, the optimal value for L can be found by determining where the transient and steady-state models of Equations 4.44 and 4.43 cross each other; i.e.,

$$L = i|MSE_{dB}(i) = S_{AF1} \tag{4.45}$$

It is worth noting that, unlike in the FIR scenarios exploited in [29], these models may not be accurate for the T-OE weights transfers because here AF1

undermodels the unknown plant. Hence, by keeping in mind that a mildly large L is unlikely to do any harm as shown in Fig. 16, practical applications may consider the theoretical L as a lower bound for the actual L as computed by Eq. 4.45, as we pointed out in [28].

4.2.6 Computational Complexity of the T-OE

In general, the resources considered in the studies on computational complexity are the time (number of operations executed) and space (memory utilization) consumed in a single iteration of a given algorithm. In the T-OE filter combinations though, this measure is not suitable since the most demanding task in the algorithm - the weights transfers - are intermittent and should ideally occur once if the cycle length L is chosen correctly.

However, for some types of embedded systems as mobile computing, energy is also a very valuable computational resource and should be considered: the more steps an algorithm performs, the higher its energy consumption. Under this perspective, it makes sense to "distribute" the clock cycles spent with the transfers among L iterations. Herewith, a pessimist estimate for the complexity is determined by the number of operations executed within the cycle L (including the transfers) and the theoretical worst case scenario is reached when $L = 1$.

Once the complexity of both FIR and IIR versions of LMS is linearly proportional to the order of the filters [3–5], every single iteration of the T-OE features a complexity proportional to the sum of the orders of AF1 and AF2; so that we can define an "overall length" as

$$M \triangleq (M_1 + 1) + (2M_2 + 1), \tag{4.46}$$

where we remark that, following the nomenclature usually found in the IIR filtering literature (e.g., [23]), in this work the order of a filter is given by the highest

power of the polynomials $B(z)$ and $A(z)$, not by its length.

As for the transfers, Sections 4.2.2 and 4.2.3 mention that the complexity is upper bounded either by the inversion of a $\mathbb{R}^{M_2+1\times M_2+1}$ matrix (Eq. 4.21) plus the eigen-decomposition of a $\mathbb{R}^{M_1+1\times M_1+1}$ matrix (Eq. 4.5) for BMR transfers or by the solution of a $\mathbb{R}^{M_2+1\times M_2+1}$ Toeplitz system (Eq. 4.31) for PAM.

In the case of the T-OE equipped with BMR transfers, the overall complexity in a single cycle L is given by

$$\begin{aligned} Complexity &= \frac{(L-1)\ O(M) + O\big((M_2+1)^3 + (M_1+1)^2\log(M_1+1)\big)}{L} \\ &\approx O(M) + \frac{O\big((M_2+1)^3 + (M_1+1)^2\log(M_1+1)\big)}{L}, \end{aligned} \tag{4.47}$$

where $L \gg 1$ ($L > 100$ is typical) and the complexities related to the inversion of the matrix in Eq. 4.20 and the eigen-decomposition referred in Eq. 4.5 are respectively $O\big((M_2+1)^3\big)$ and $O\big((M_1+1)^2\log(M_1+1)\big)$ (assuming that the eigen-decomposition is made via the algorithm in [105]).

Similarly to the BMR transfers case, the overall complexity of the T-OE equipped with PAM transfers in a single cycle L is given by

$$\begin{aligned} Complexity &= \frac{(L-1)\ O(M) + O(M_2+1)^2}{L} \\ &\approx O(M) + \frac{O(M_2+1)^2}{L}, \end{aligned} \tag{4.48}$$

where $O(M_2+1)^2$ is the PAM transfers complexity (Section 4.2.3).

In any event, the cycle L is likely to exceed M by a large amount, leading to an overall complexity close to $O(M)$. It is also important to take into account that the weights transfers cease when AF1 converges, so that in stationary or slowly varying environments the complexity will be defined as $O(M)$ in the long run.

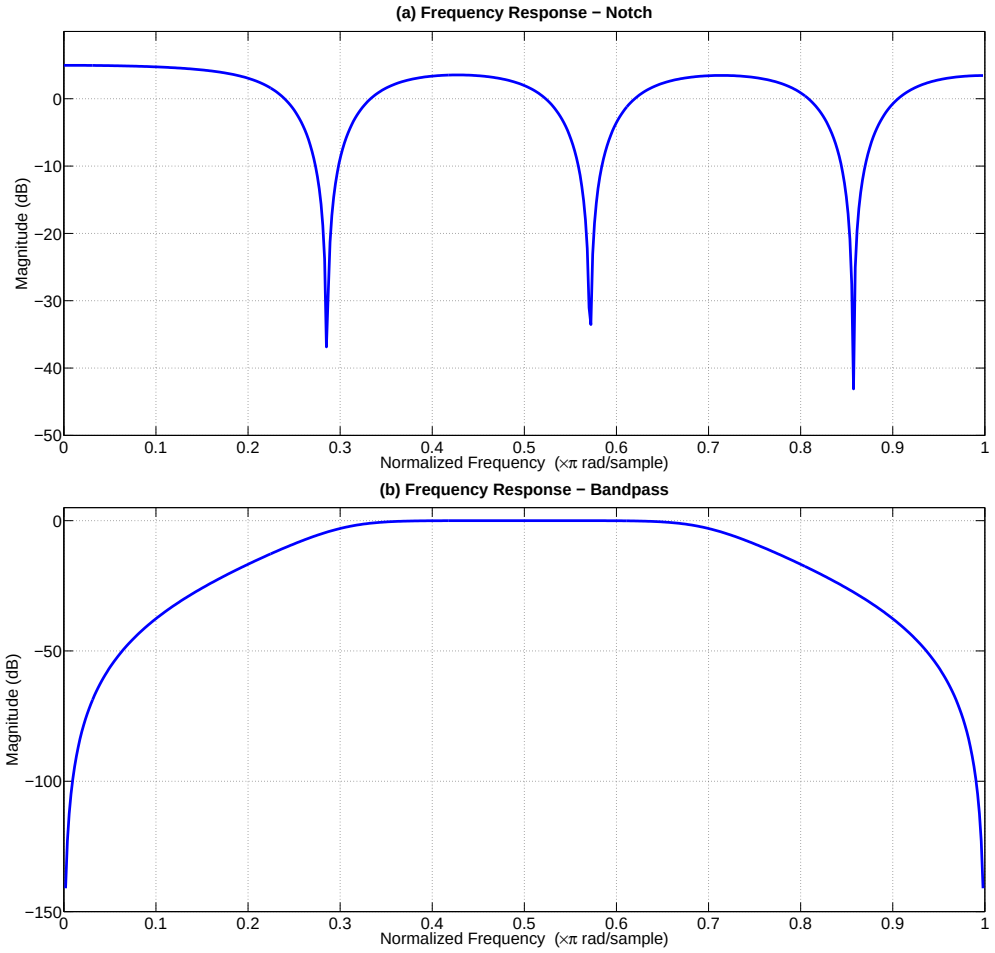

Figure 18: Frequency Responses of the Test Scenarios

4.2.7 Experimental Results on T-OE

Fig. 18 shows the frequency responses of the unknown plants of the simulated identification scenarios, both with order $M = 6$. In *(a)*, the unknown plant has $w^o = [0.2949 \ -0.7709 \ 0.2375 \ -0.4934 \ 0.1208 \ -0.2621 \ 1.0 \ -0.3686 \ 1.2045 \ -0.4639 \ 1.2045 \ -0.3686 \ 1.0]$, which is a notch filter that attenuates the first three harmonics of 50 Hz of a given input sampled at a rate of 420 Hz. Scenario *(b)* is a Butterworth bandpass system where $w^o = [0 - .5772 \ 0 - .4218 \ 0 - .0563 \ .0985 \ 0 - .2956 \ 0 \ .2956 \ 0 - .0985]$.

Fig. 19 shows the impulse responses of both scenarios, which decay fast in both cases (in less than 40 samples, they vanish). However, such relatively short responses do not lead to an easy identification, as the simulations show.

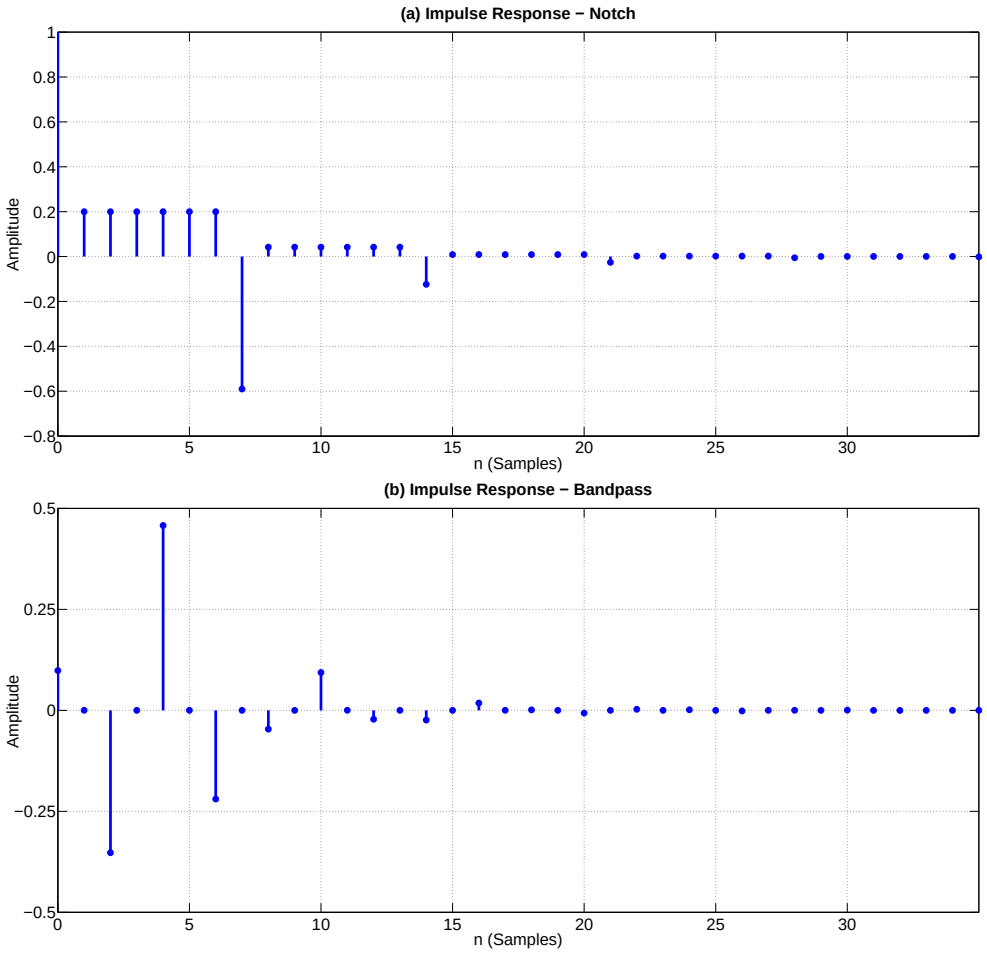

Figure 19: Impulse Responses of the Test Scenarios

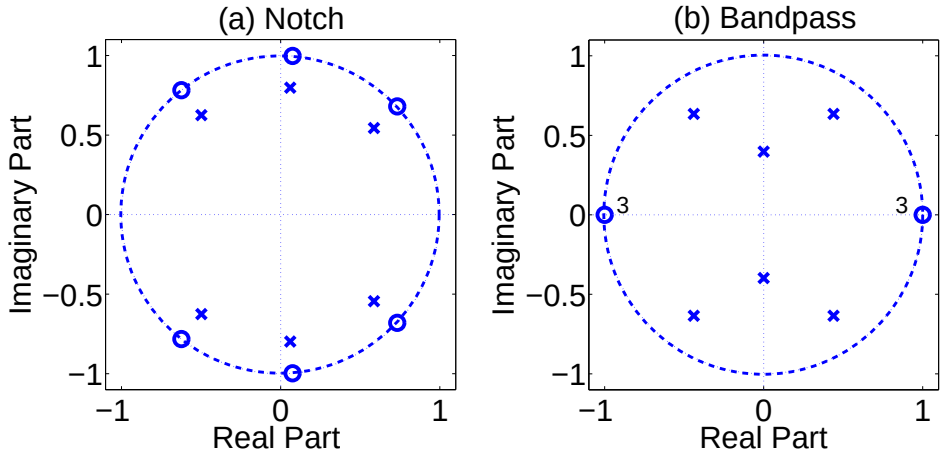

Figure 20: Poles-Zeros Diagram of the Test Scenarios

Fig. 20 shows the poles-zeros diagrams of the scenarios, where the presence of underdamped and clustered poles are known to make the identification harder for a standard standalone IIR AF, leading to long convergence times [10, 19]. From the designer perspective, these diagrams are far more valuable than the corresponding frequency or impulse responses, as the poles-zeros configuration of the plant gives more insights about the filtering process.

Given that nearby poles and zeros tend to undermine each other effects, their relative proximity in the notch case intuitively suggests that the convergence problems are attenuated. Notwithstanding, experiments show the opposite, with the convergence taking longer as the poles magnitude increases; i.e., unless actual zeros-poles cancellations occur, the singular value spread of the underlying Hankel matrix increases with high magnitudes poles, slowing down the convergence rate (refer to Appendix C).

Every simulation consists in an ensemble of 400 realizations. The input signal $u(i)$ is zero-mean, Gaussian white with power $\sigma_u^2 = 1$ and the additive noise $v(i)$ is zero-mean, Gaussian white with power $\sigma_v^2 = 1 \times 10^{-3}$. Since this work addresses sufficient modelling cases only, in both scenarios the order of AF2 is $M_2 = 6$, matching the unknown plants order. In the notch case, the order of AF1 is $M_1 = 15$, its step size is $\mu_1 = 100 \times 10^{-3}$ and the step size of AF2 is $\mu_2 = 33.33 \times 10^{-3}$. For the bandpass, $M_1 = 15$, $\mu_1 = 200 \times 10^{-3}$ and $\mu_2 = 5 \times 10^{-3}$. In all simulations AF2 is implemented in the direct form and no stability checks are performed because μ_2 is designed to be accurate; as such, it has to be relatively small, which makes the filter stable as seen in Section 2.2.

In order to give a clear picture of how the weights transfers affect the accurate AF (AF2) in practice, L was set to values smaller than necessary. This was made arbitrarily for illustrative purposes only; in real-world applications, the same scenarios could use larger values. As a matter of fact, L can range substantially

without perceptible effects on the overall performance as seen in Fig. 16. For the notch, L was set to 100 (for BMR transfers) and 390 (for PAM transfers) whereas with the bandpass it was set to 50 (BMR) and 270 (PAM). For comparison purposes, the same simulations were also run without weights transfers.

When the weights transfers were enabled, the convex supervisor was configured with $\mu_a = 500$ and $A^+_{max} = 5.00$ for the notch and $\mu_a = 1200$ and $A^+_{max} = 4.00$ for the bandpass (see Section 3.2). With the transfers disabled, the bandpass supervisor changes with $\mu_a = 400$ and $A^+_{max} = 5.00$, whereas the configuration for the notch remains the same. In all cases, AF1 is considered better whenever $\lambda(i) \geq 0.7$ (*i.e.*;$\lambda = 0.7$). For reference, the T-OE algorithm is summarized in Section 4.2 (Algorithm 1).

In the plots, the learning curves of the FIR guide are identified by the label "AF1", the T-OE combinations by "T-OE BMR" and "T-OE PAM" (when BMR or PAM weights transfers are enabled) or "T-OE"(weights transfers disabled). AF2 is not shown (except when the figures are zoomed) because the curves of the combinations with weights transfers overlaps those of AF2 in the scale used in the plots. As a benchmark, the learning curves of regular, standalone IIR AFs ("N-LMSOE") configured exactly as AF2 were also plotted to allow an easy performance comparison.

T-OE - Stationary Case

Figs. 21 and 22 show the MSE and the EMSE of the T-OE when identifying the notch. Note that the convex combination with weights transfers adapts smoothly and faster than the standalone IIR AF. A superior performance is also achieved when the weights transfers are not activated but, as already mentioned, in this case the combination stagnates for a long period until the accurate filter (AF2) reaches the error level of the FIR guide (AF1).

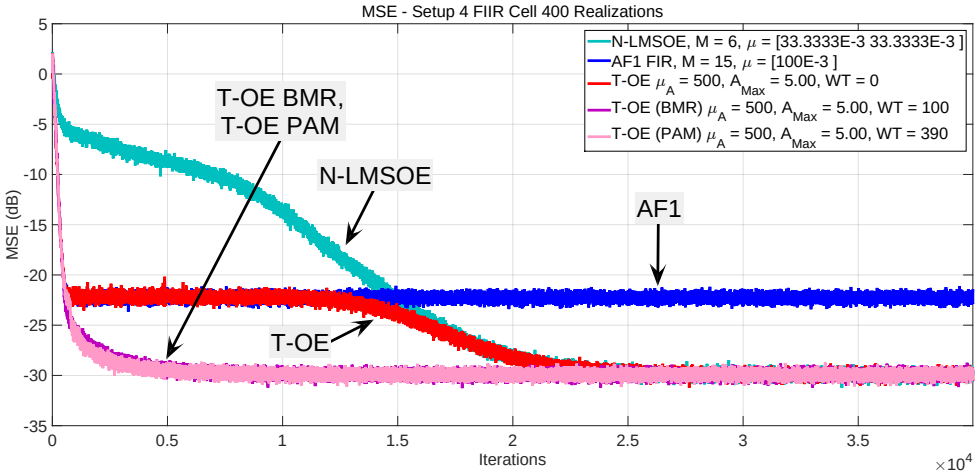

Figure 21: MSE - Notch (Stationary Scenario)

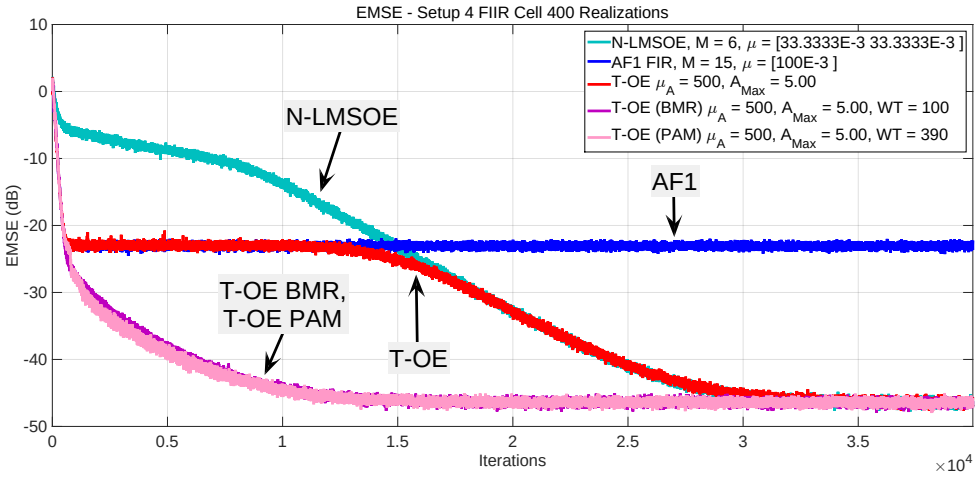

Figure 22: EMSE - Notch (Stationary Scenario)

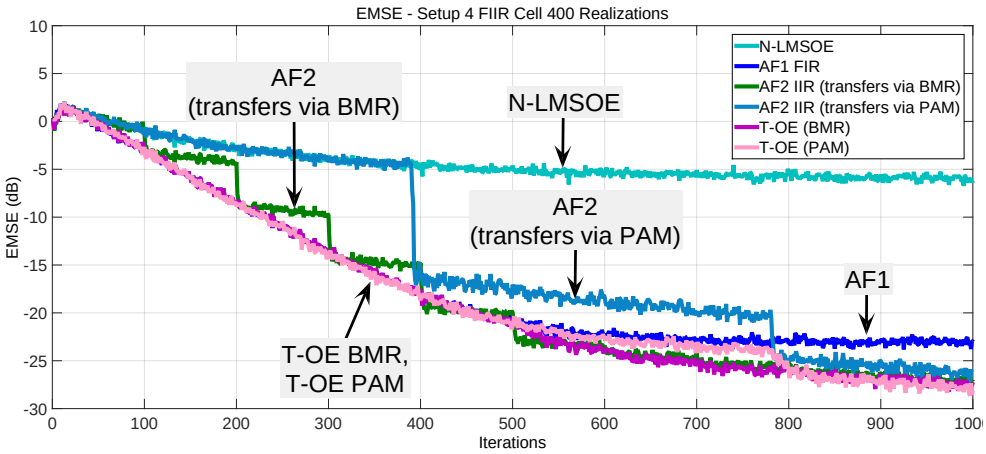

Figure 23: EMSE - Notch (Stationary Scenario, zoom)

In the zoom provided in Fig. 23, the effect of the weights transfers on AF2 are perceptible - every time a transfer occurs, its EMSE drops about 5 dB, forming "steps". As aforementioned, in these simulations L is clearly smaller than necessary in order to highlight how the performance of AF2 changes with the transfers. Also, *although the stability checks are disabled*, both AF2 and the standalone IIR kept exponentially stable as evidenced by the uniform pattern of their learning curves.

Figs. 24 and 25 show the MSE and the EMSE of the T-OE when identifying the bandpass. This scenario is significantly harder than the notch because the poles cluster – under these conditions, the adaptation of a standalone IIR AF becomes slower and the convergence could take too long. While this also affects the performance of the T-OE, it does adapt faster than the standalone filter, even when no weights transfers are made. Notwithstanding, like in the notch case, the weights transfers improve considerably the convergence rate (for clarity, only the first 30000 iterations are shown in the plot).

Fig. 26 shows a zoom of the very first iterations of the EMSE of the components of the T-OE and the standalone IIR AF. As in the prior case, the effect of the weights transfers are emphasized, with every transfer leading to an expressive

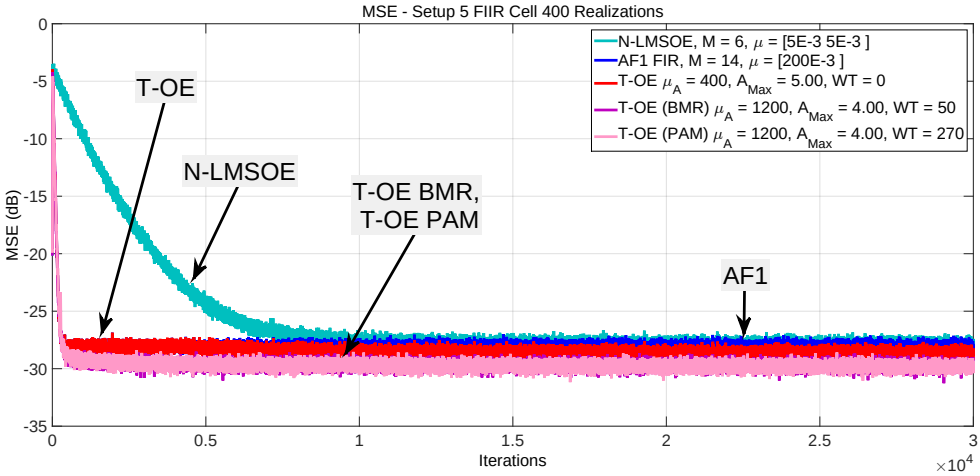

Figure 24: MSE - Bandpass (Stationary Scenario)

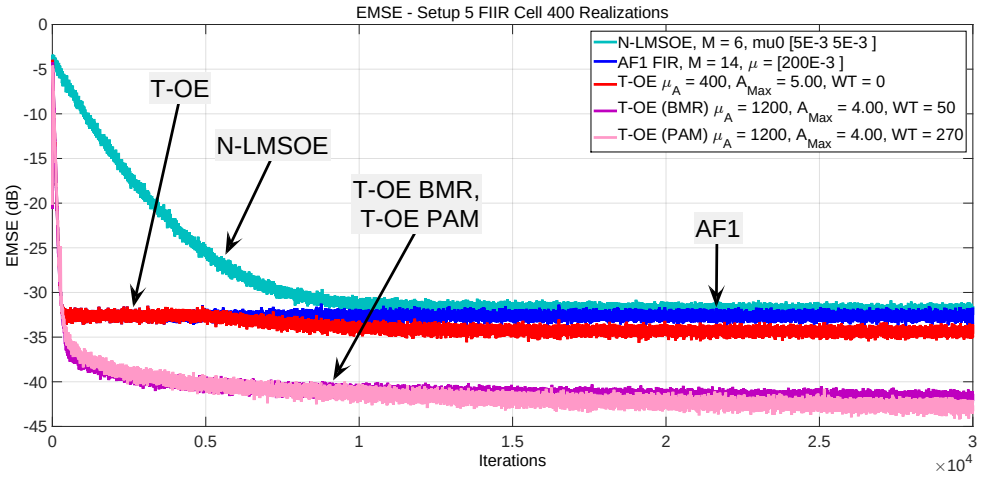

Figure 25: EMSE - Bandpass (Stationary Scenario)

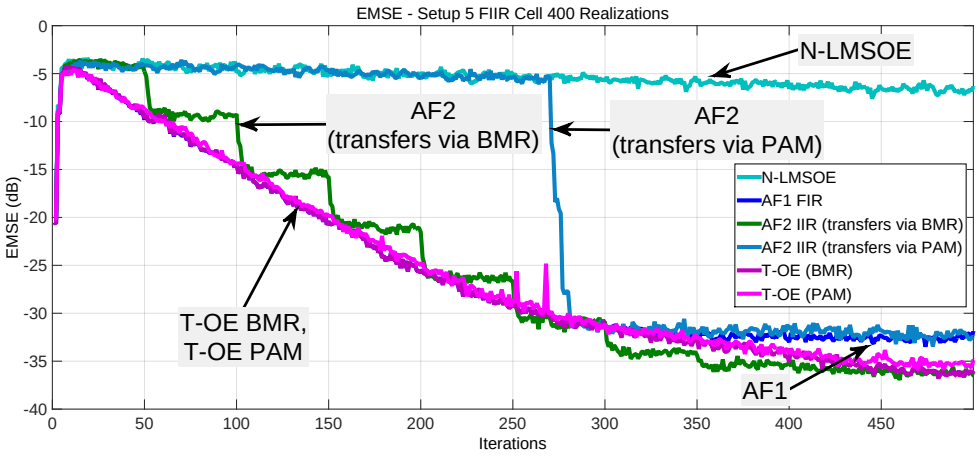

Figure 26: EMSE - Bandpass (Stationary Scenario, zoom)

drop (between 5 and 10 dB) in the AF2 error levels.

T-OE - Non-Stationary Case

As the benefits of the weights transfers became evident in the stationary case scenarios (Figures 21, 22, 24 and 25), here we omitted the combinations without weights transfers to improve the clarity of the plots. Figs. 27 and 28 show the same scenario than Figs. 21 and 22, but considering an abrupt non-stationarity caused when the notch frequency of the plant changes suddenly from 50 Hz to 55 Hz. All the AFs (the standalone IIR included) manage to recover from the non-stationarity, but the T-OE resumes much faster.

In the same fashion, Figs. 29 and 30 show how the AFs behave when the non-stationarity is more abrupt and the notch frequency changes from 50 Hz to 60 Hz. In this case, the T-OE was able to track the change whereas the standalone IIR AF was not and stagnated too far from its steady state. Incidentally, a major flaw that plagues OE adaptive filters is highlighted here: frequently, a standalone IIR AF is not capable of delivering an acceptable performance in reasonable time frames when the plant being identified is endowed with sharp resonances as it is the case of the notch.

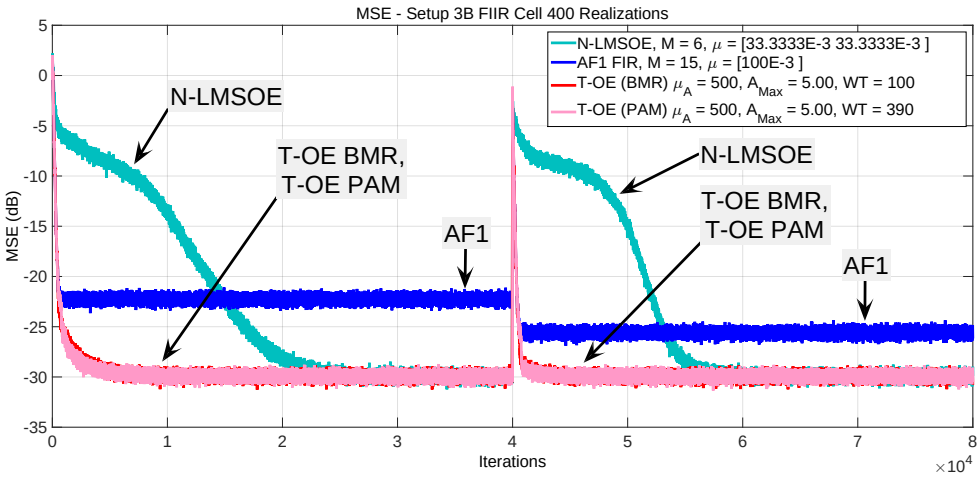

Figure 27: MSE - Notch (Non-Stationary Scenario)

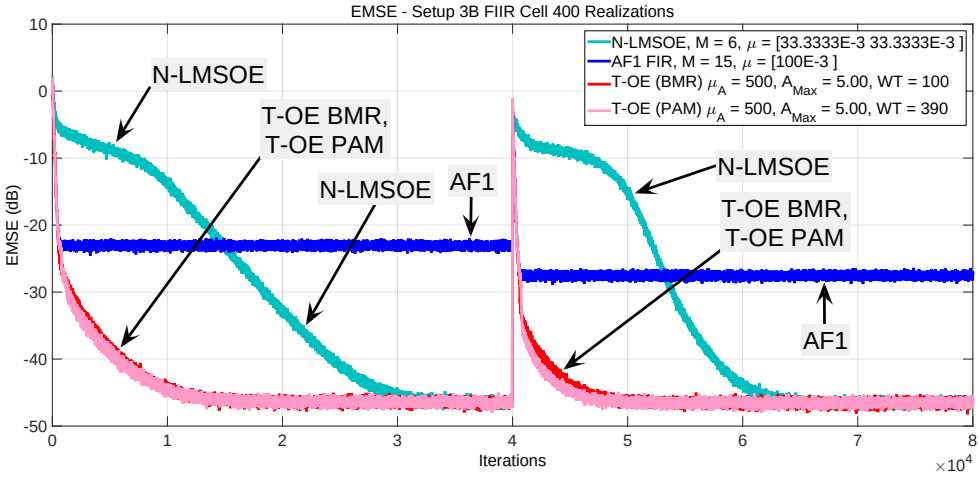

Figure 28: EMSE - Notch (Non-Stationary Scenario)

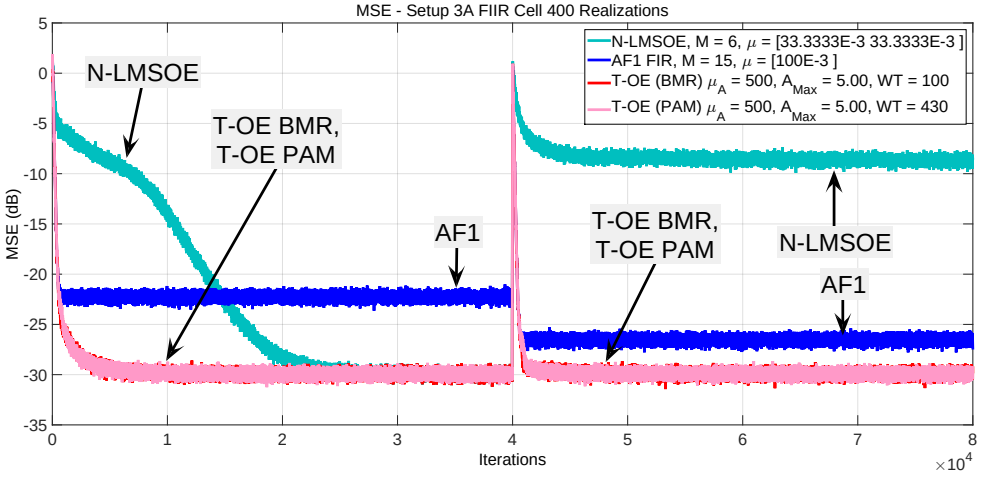

Figure 29: MSE - Notch (Non-Stationary Scenario)

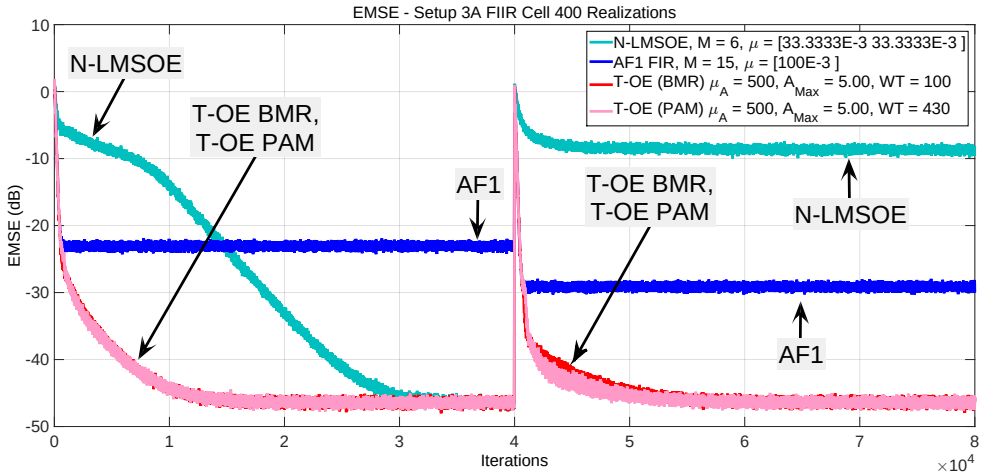

Figure 30: EMSE - Notch (Non-Stationary Scenario)

4.3 IIR-IIR Adaptive Convex Combinations

Despite the superior performance of the T-OE over single standalone IIR AFs, the FIR $\rightarrow$ IIR weights transfers require some extra computations. Albeit the number of times that the mapping $\mathcal{F}$ is executed can be dramatically reduced with a proper choice for the cycle transfer L and the overall convergence rate is quite immune to large variations in its range (see Section 4.2.5 and Fig. 16), state-space based algorithms as BMR deal with matrices and vectors whose dimensions are proportional to the FIR filter order. PAM-based weights transfers are simpler as they are related to the order of the IIR filter; nonetheless, they need a linear system to be solved. So, the transfers can be an issue in terms of computational cost.

Additionally, any SVD-based order reduction as the BMR is a kind of dimension truncation that can lead to a variable loss of accuracy (refer to Section 4.2.2) [30, 94, 118]. In turn, the PAM-based order reduction assumes that the impulse response of the plant is approximated accurately by AF1 under the penalty of yielding unstable transfers (Section 4.2.3).

Clearly, such issues could be avoided by eliminating the order reduction pro-

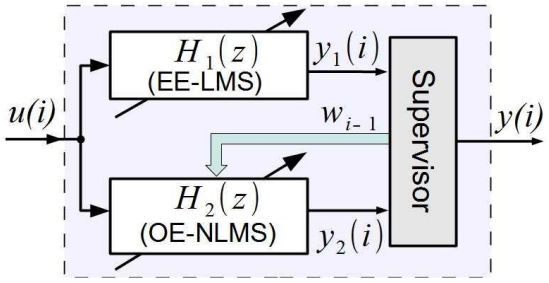

Figure 31: Hybrid Combination of IIR Adaptive Filters

cedure, which requires a guide filter that is compatible with the accurate IIR AF *not only in the transfer function space as the FIR guide is, but in the parameter space as well*; i.e., by using IIR guides of the same order as AF2. In different contexts, some works already exploited an analogous concept through the *composite* IIR AFs (e.g., [18, 31–36]). Although no actual AFs combination have been used in those works, they do merge the properties of different IIR filtering formulations to improve the overall convergence rate as briefly mentioned in Appendix A.

4.3.1 Components for the IIR-IIR Adaptive Combinations

Fig. 31 shows the architecture of a generic, purely IIR convex association we introduce in this work, in which both components share compatibility in the parameters space besides pursuing the same transfer function $H^o(z)$ in a system identification setup. Hence, unlike the T-OE referred into Section 4.2, here the supervisor is able to cyclically feedback the *global weights* w_i defined as

$$w_{i-1} = \lambda(i)w_{1,i-1} + \big(1 - \lambda(i)\big)w_{2,i-1} \tag{4.49}$$

to the accurate component (i.e., AF2) without resorting to any kind of mapping. Differently from the purely FIR convex associations (refer to [29]), in principle the global weights should not feedback both components as AF2 is subject to unstable updates in case of non-stationarities like burst noise, which could be inadvertently propagated to the supervisor and then destabilize AF1 as well.

In the same vein as in the T-OE, the accurate component (AF2) is implemented as a N-LMSOE adaptive filter endowed with an arbitrarily small learning step μ, so that exponential stability is enforced even if stability checks are not performed (Section 2.7). Hence, the guide filter (AF1) must be fast to lead AF2 rapidly towards the neighborhood of the global minimum and be stable as well because the cyclic feedbacks could compromise the combination otherwise.

The LMSOE or N-LMOSE are not an option for AF1 because their step sizes $\mu(i)$ are constrained to a small range of values in order to bound the non-stationarity caused by the recursivity and keep the stochastic gradient approximations accurate (see Chapter 1). Clearly, a superior performance can be achieved by IIR algorithms that either *avoid the recursivity during the adaptation* or *approximate the stochastic gradient differently.* The former strategy is implemented through the Equation Error methods (Section 4.3.2) and the latter comes in the form of the Pseudo Linear Regression (Section 4.3.3) or hyperstable methods [12].

Both approaches are robust with respect to the stability and can be much faster than the N-LMSOE; however, such a performance is achieved to the detriment of accuracy: biased convergence for the Equation Error or stagnation into an arbitrary point onto the surface error for the PLR (even with white inputs and correctly modeled AFs). Such drawbacks are expected to be compensated to some extent by the combination of Fig. 31.

4.3.2 E-OE: Equation Error–OE Combinations

Within the context of systems identification, the Output Error was defined with the mixed notation in Section 2.4.1 as

$$\begin{aligned} e(i) &= d(i) - y(i) \\ &= H^o(z)u(i) + v(i) - H(z, i-1)u(i) \end{aligned}$$

$$= \left(\frac{B^o(z)}{1 - A^o(z)} - \frac{B(z, i-1)}{1 - A(z, i-1)} \right) u(i) + v(i), \tag{4.50}$$

where the dependency on $A(z)$ is clearly non linear. By changing the definition of the estimation error and defining the "Equation Error" output $y_e(i)$ as a function of the prior plant's outputs $d(i)$ in lieu of the own AF's outputs $y(i)$, the Equation Error was defined in Section 2.4.2 as

$$\begin{aligned} e(i) &= d(i) - y_e(i) \\ &= d(i) - \big(B(z, i-1)u(i) + A(z, i-1)d(i)\big) \\ &= \big(1 - A(z, i-1)\big)d(i) - B(z, i-1)u(i) + \big(1 - A(z, i-1)\big)v(i), \end{aligned} \tag{4.51}$$

and the dependency on $A(z)$ now becomes plain linear. This means that the stochastic gradient can be properly approximated without enforcing the parameters to change slowly as is the case of the LMSOE derived in Section 2.8.2.

Structurally, the Equation Error realization is a dual-channel filter in which the coefficients $B(z)$ and $A(z)$ are adapted through separated FIR AFs (that is why EE approaches are not IIR filtering strictly speaking). As Eq. 4.51 shows, in this setup $v(i)$ is filtered by $A(z)$, which makes the algorithm cope with two conflicting goals, namely, minimize the MSEE and identify the poles of the plant. This becomes clear by analyzing the Mean Square Equation Error given by

$$\mathrm{E}\, e^2(i) = \mathrm{E}\, e_u^2(i) + \sigma_v^2\, \mathrm{E} \sum_{k=0}^{M} a_k^2(i-1), \tag{4.52}$$

where $e_u(i)$ is the unbiased Equation Error and, basically speaking, $\sigma_v^2\, \mathrm{E} \sum_{k=0}^{M} a_k^2$ acts as an undesirable regularization term. Hence, convergence to the minimum error is not achieved but in the unlikely scenarios where $v(i) = 0\ \forall\ i$. On the other hand, without any further constraints on the parameters evolution, the convergence rates may be far higher than those observed in LMSOE AFs, making them suitable as the fast component of Fig. 31.

The update rule of the LMS version for the Equation Error formulation (LMSEE) is given by [1, 4, 20]

$$w_i = w_{i-1} + \mu x_{\mathrm{EE},i}^T e(i), \tag{4.53}$$

where $x_{\mathrm{EE},i}$ is the EE regressor defined as

$$x_{\mathrm{EE},i} = [d(i-1)\ d(i-1)\ \cdots\ d(i-M)\ u(i)\ u(i-1)\ \cdots\ u(i-M)] \tag{4.54}$$

Eq. 4.54 unveils that the input regressor $x_{\mathrm{EE},i}$ is not disturbed by the time-varying $A(z,i)$ like in a truly recursive AF; i.e., no matter how $A(z,i)$ changes, $x_{\mathrm{EE},i}$ will remain statistically the same because $d(i)$ is yielded by the plant. Considering that the non-stationarity injected by the feedback loop (i.e., $A(z,i)$) into the input regressor is a major performance bottleneck for the OE AFs [12], the EE AFs tend to feature better convergence rates even if $d(i)$ is highly correlated.

Similarly to what happens with the LMSOE, often it is advantageous to normalize the EE regressor, yielding the Normalized EE LMS (N-LMSEE) by modifying Eq. 4.53 as

$$w_i = w_{i-1} + \mu_1 \frac{x_{\mathrm{EE},i}^T}{\epsilon + \|x_{\mathrm{EE},i}\|^2} e(i), \tag{4.55}$$

where ϵ is a small regularization factor.

Putting all together, the algorithm originally developed for the T-OE (Section 4.2) modifies to accommodate the Equation Error–Output Error combination (E-OE) as shown in Algorithm 2, with the mixing factor $\lambda(i)$ defined as in Section 3.2 (Equations 3.2, 3.5 and 3.10).

Algorithm 2 Equation Error–Output Error Filters Combination (E-OE)

for $i = 0$, $i{+}{+}$ **do**

$$
\begin{aligned}
y_{1,i} &= x_{\mathrm{EE},i} w_{1,i-1} \\
e_{1,i} &= d(i) - y_1(i) \\
w_{1,i} &= w_{1,i-1} + \mu_1 x_{\mathrm{EE},i}^T e_1(i) \\
w_{2,i-1} &= \delta w_{i-1} + (1-\delta)(w_{2,i-1}) \\
y_{2,i} &= x_i w_{2,i-1} \\
\phi_i &= \frac{x_i}{1 - A(z, i-1)} \\
e_{2,i} &= d(i) - y_2(i) \\
w_{2,i} &= w_{2,i-1} + \frac{\mu_2}{\|\phi_i\|^2 + \epsilon} \phi_i^T e_2(i) \\
y(i) &= \lambda(i) y_1(i) + \big(1 - \lambda(i)\big) y_2(i) \\
w_i &= \lambda(i) w_{1,i} + \big(1 - \lambda(i)\big) w_{2,i}
\end{aligned}
$$

end for

4.3.3 P-OE: Pseudo Linear Regression–OE Combinations

As previously discussed, the dependency between the current output and the prior ones in the LMSOE estimates, given by

$$y(i) = H(z)u(i) = x_i w_{i-1}, \tag{4.56}$$

results from the definition of the OE regressor

$$x_i = [y(i-1)\ y(i-1)\ \cdots\ y(i-M)\ u(i)\ u(i-1)\ \cdots\ u(i-M)], \tag{4.57}$$

which renders the corresponding update rule non-linear. However, by neglecting this dependency and taking $y(i)$ as it was linear during the adaptation, the update rule simplifies. In fact, this was originally done by Feintuch [119] [10] and, regardless of the criticisms of great names of the systems identification community as Johnson [120] and Widrow [121], its fundamentals were shown to be consistent under certain constraints by Rupp and Sayed [122], Landau [24] and others.

[10] Due to the pioneer work in [119], this strategy is also known as *the Feintuch's algorithm*.

Let the update rule of Eq. 2.61 (Section 2.8) with $B_i = \mathbb{I}$ and $\eta_i = x_i$; i.e., the gradient $\nabla_w J(w_{i-1})$ is (roughly) approximated by $x_i^T e(i)$, producing

$$w_i = w_{i-1} + \mu_1(i) x_i^T e(i), \tag{4.58}$$

From Section 2.8.2, the approximation in Eq. 4.58 is readily recognized as an abrupt truncation of the true gradient formulas (Eqs. 2.74 and 2.75) such that

$$\begin{aligned}
\frac{\partial E e^2(i)}{\partial b_m(i)} &= -2\,\mathrm{E}\, e(i) \left(u(i-m) + \sum_{k=1}^{M} a_k(i) \frac{\partial y(i-k)}{\partial b_m(i-k)} \right) \\
&\cong -2\,\mathrm{E}\, e(i) u(i-m) \qquad (4.59) \\
\frac{\partial E e^2(i)}{\partial a_m(i)} &= -2\,\mathrm{E}\, e(i) \left(y(i-m) + \sum_{k=1}^{M} a_k(i) \frac{\partial y(i-k)}{\partial a_m(i-k)} \right) \\
&\cong -2\,\mathrm{E}\, e(i) y(i-m) \qquad (4.60) \\
\therefore \frac{\partial E e^2(i)}{\partial w_{i-1}} &\cong -2\,\mathrm{E}\, x_i^T e(i) \qquad (4.61)
\end{aligned}$$

Because it casts a model that is not really linear, this strategy is called *Pseudo Linear Regression* (PLR) [20, 45, 52, 53] or, given its similarity to the FIR LMS algorithm, *Recursive LMS* (RLMS) [2, Section 2.9.4]. In the same vein as the EE approaches, PLR can also be thought as a dual FIR setup [119, 121] except that the regressor and error signal of the Equation Error formulation are replaced by their Output Error correlates (i.e., unlike EE, PLR is genuinely recursive and does have poles while adapting).

Once it does not follow the true gradient, PLR can not be expected to always achieve the global minimum but it can be shown to be upper-bounded (i.e., *stable*) in the sufficient modelling case [45, 121]. In other words, if the order of the adaptive filter is sufficient, its MSE will decrease along time although it is uncertain to what it may converge; otherwise, stability can not be assured.

If, on one hand, the gradient truncation leads to an optimization crite-

rion other than the MSE, on the other hand it also makes PLR self-stabilizing [20, 45, 121] [2, Section 2.9.4], such that unstable poles tend to migrate into the unit circle along the adaptation. At an intuitive level, Widrow and McCool [121] described this property by stating that *the feedback of the adaptive process interacts favorably with the feedback of the filter itself to produce a "superstability" that will pull the poles back from beyond the brink of instability* [11] . In fact, it has been suggested that the PLR adaptation rule should be applied to ordinary LMSOE AFs whenever their updates are unstable, which may lead to a better performance than simply skipping them, as practitioners usually do [20].

However, this self-stabilizing feature is not completely understood, and so it is hard to tell categorically if it holds for any kind of input signals such as rational inputs. Anyway, as Elliott [2], Eriksson and Greiner [41] and others noted, PLR shows its strong stability virtues in many real applications like active sound attenuation in factories or inside ducts and, for this reason, it is widely used in the industry.

Given its stability features, PLR copes with larger learning step sizes than LMSOE or N-LMSOE and therefore can achieve higher convergence rates. Such an improvement is welcome but, as aforementioned, the update rule of Eq. 4.58 optimizes w_i regarding to a figure of merit other than $J(w_i)$, possibly making the AF converge far from the actual global minimum. However, this side effect vanishes if the plant meets a *strictly positive real* (*SPR*) condition [123] given by

$$\Re\left(\frac{C(z)}{A^o(z)}\right) - K > 0 \ \forall |z| = 1, \tag{4.62}$$

where $\Re(.)$ returns the real part of the argument, $\frac{1}{A^o(z)}$ with $z = 1$ is the frequency response of $\frac{1}{A^o(z)}$, $C(z)$ (the so-called *compensator*) is usually set to 1 [45] but it

[11] These occasional drifts that cause a temporary instability in the AF is called *bursting*, which is quite common in adaptive echo cancellation applications and in adaptive control systems [76]. A formal analysis on the stability and convergence of PLR is led by Rupp and Sayed in [122].

may assume other values as is the case of the hyperstable approaches [12, 58] and K is a slack variable normally set to $1/2$ as Landau did in the pioneer work where the hyperstability ideas were first applied to the identification problems [24], [60, Chapter 7] (although $K = 0$ is also possible in other approaches).

Note that Eq. 4.62 is a sufficient, not necessary, condition for convergence [20, 45]. In fact, in spite of the SPR condition, the PLR may converge correctly depending on the initialization and on the input sequence $u(i)$ [12] .

By using the update rule of Eq. 4.58, the P-OE is summarized in Algorithm 3, with $\lambda(i)$ defined as in Section 3.2 and the AF1 and AF2 regressors given by $x_{n,i} = [y_n(i-1)\ y_n(i-1)\ \cdots\ y_n(i-M_2)\ u(i)\ u(i-1)\ \cdots\ u(i-M_2)]$ $(n = 1, 2)$.

Algorithm 3 PLR–Output Error Filters Combination (P-OE)

for $i = 0$, $i{+}{+}$ **do**

$$
\begin{aligned}
y_{1,i} &= x_{1,i} w_{1,i-1} \\
e_{1,i} &= d(i) - y_1(i) \\
w_{1,i} &= w_{1,i-1} + \mu_1 x_{1,i}^T e_1(i) \\
w_{2,i-1} &= \delta w_{i-1} + (1-\delta)(w_{2,i-1}) \\
y_{2,i} &= x_{2,i} w_{2,i-1} \\
\phi_i &= \frac{x_{2,i}}{1 - A(z, i-1)} \\
e_{2,i} &= d(i) - y_2(i) \\
w_{2,i} &= w_{2,i-1} + \frac{\mu_2}{\|\phi_i\|^2 + \epsilon} \phi_i^T e_2(i) \\
y(i) &= \lambda(i) y_1(i) + \big(1 - \lambda(i)\big) y_2(i) \\
w_i &= \lambda(i) w_{1,i} + \big(1 - \lambda(i)\big) w_{2,i}
\end{aligned}
$$

end for

[12] A persistently high σ_x^2 (regressor power) essentially compensates for the SPRless of the plant, which would require some *a priori* knowledge about the original plant though [25, 58].

4.3.4 Experimental Results on IIR-IIR Combinations

The same scenarios of Section 4.2.7 were used here to examine the performance of the E-OE and P-OE combinations of Fig. 31 against the T-OE. *All algorithms were designed to match each other learning rates as much as possible, so that a comparison makes sense.* Also, since the superiority induced by the weights transfers became clear when studying the T-OE, the comparison will consider only combinations with the transfers enabled.

In order to improve the readability, only the EMSE plots are shown here as they tend to highlight the existing differences among the algorithms. They exhibit the learning curves of the IIR combinations and the T-OE labeled, respectively, as "T-OE BMR" and "T-OE PAM" as well the guide filter ("AF1") of both. All AFs and combinations are identified by arrows with labels which refer to the accompanying legend.

In all cases, every simulation consists in an ensemble of 400 realizations, in which the input signal $u(i)$ was a zero-mean white process with power $\sigma_u^2 = 1$, the additive noise $v(i)$ is Gaussian white and the power σ_v^2 is indicated specifically along each individual simulation.

Simulations with the E-OE Combinations

The simulations considered the butterworth case with $w^o = [0 \; -.5772 \; 0 \; -.4218 \; 0 \; -.0563 \; .0985 \; 0 \; -.2956 \; 0 \; .2956 \; 0 \; -.0985]$, fed with a zero-mean, Gaussian white signal $u(i)$ with power $\sigma_u^2 = 1$. The background noise $v(i)$ is Gaussian and zero-mean and two cases are explored, with power $\sigma_v^2 = 10^{-3}$ and $\sigma_v^2 = 10^{-4}$. When $i = 80 \times 10^3$, the plant changes abruptly to $w^o = [0 \; .5772 \; 0 \; -.4218 \; 0 \; .0563 \; .2569 \; 0 \; -.7707 \; 0 \; .7707 \; 0 \; -.2569]$. The transfers cycle is $L = 500$, which combats the stagnation effect and enables the combinations to react upon

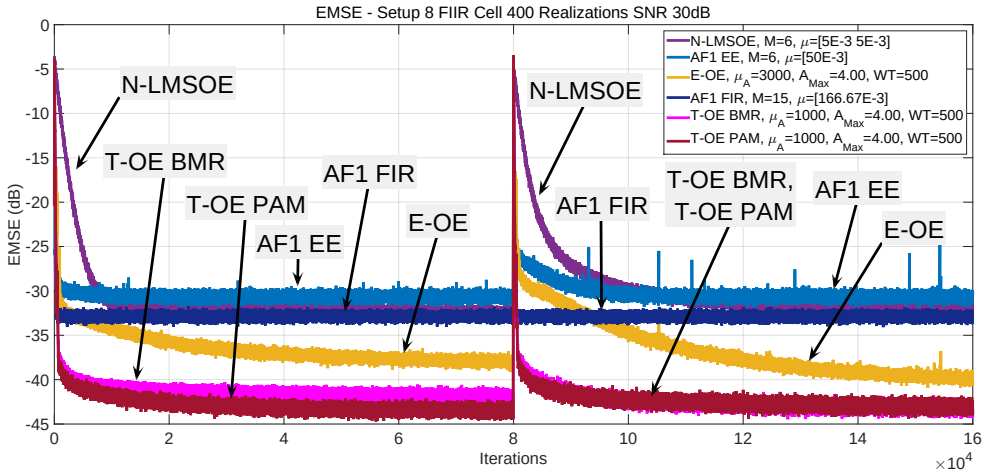

Figure 32: EMSE - Butterworth SNR = 30 dB (Non-Stationary Scenario)

the abrupt non-stationarity at $i = 80 \times 10^3$.

In the simulation depicted in Fig. 32 $\sigma_v^2 = 1 \times 10^{-3}$. The accurate component of all combinations (i.e., AF2) is an N-LMSOE with order $M_2 = 6$ and $\mu_2 = 5 \times 10^{-3}$. The fast component of the E-OE (AF1 EE) is an EE-LMS with order $M_1 = 6$ and $\mu_1 = 50 \times 10^{-3}$ and we note that the corresponding combination outperforms AF1 without the stagnation effect, even after the abrupt change in w^o. The fast component of the T-OE combinations (AF1 FIR) has order $M_1 = 15$ and $\mu_1 = 166.7 \times 10^{-3}$ and the corresponding combinations (T-OE BMR and T-OE PAM) expressively outperform the E-OE, with the T-OE PAM performing slightly better than the T-OE BMR.

In the second example, depicted in Fig. 33, the SNR increases by 10 dB, i.e., $\sigma_v^2 = 10^{-4}$. Since the EE-LMS is biased by $v(i)$ (see Eq. 4.52 in Section 4.3.2), thus improving over a better SNR, the E-OE also improves considerably and reaches an error level almost 20 dB lower than the prior case. Note that the E-OE starts at a lower convergence rate, but reaches and outperforms both T-OE combinations later on. In other simulated cenarios, with SNR at 50 dB and 60 dB, the improvement for EE-OE is more expressive, reflecting the gain in SNR; however, still slower than the T-OE.

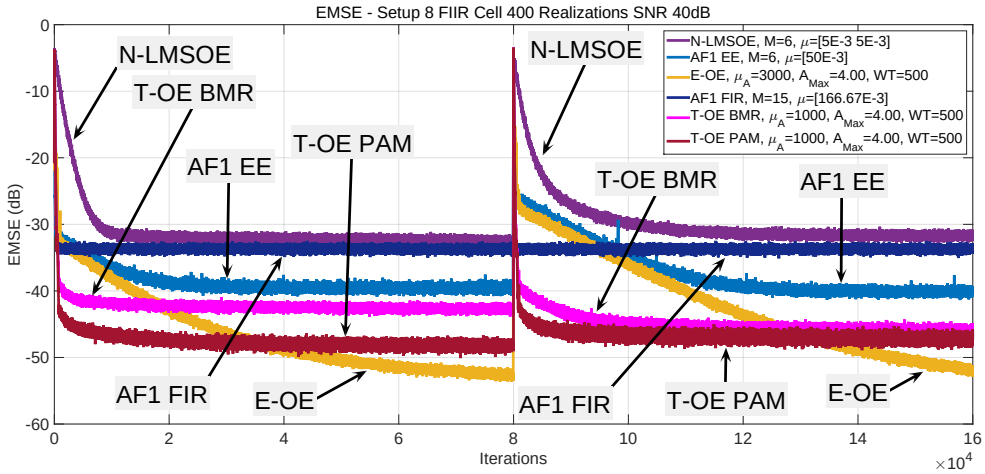

Figure 33: EMSE - Butterworth SNR = 40 dB (Non-Stationary Scenario)

In both examples, a major drawback of IIR AFs becomes evident: the N-LMSOE and both T-OE combinations seem to have reached the steady-state in the plots of Figures 32 and 33, while actually they keep evolving for nearly hundreds of thousands of iterations (not shown in the pictures for clarity). In any event, the combinations are clearly effective at learning acceleration.

Simulations with the P-OE Combinations

The first plant considered here is the non-stationary notch examined in Section 4.2.7, fed with a zero-mean, Gaussian white signal $u(i)$ with power $\sigma_u^2 = 1$. The background noise $v(i)$ is Gaussian and zero-mean with power $\sigma_v^2 = 10^{-3}$.

The accurate component of all combinations (AF2) is an N-LMSOE with order $M_2 = 6$ and $\mu_2 = 33.3 \times 10^{-3}$. The fast component of the E-OE (AF1 EE) is an EE-LMS with order $M_1 = 6$ and $\mu_1 = 30 \times 10^{-3}$ and the fast component of the P-OE combinations (AF1 PLR) has also order $M_1 = 6$ and $\mu_1 = 30 \times 10^{-3}$. At the beginning, the parameters of the notch are $w^o = [0.2949\ -0.7709\ 0.2375\ -0.4934\ 0.1208\ -0.2621\ 1.0\ -0.3686\ 1.2045\ -0.4639\ 1.2045\ -0.3686\ 1.0]$, which makes it attenuate the first three harmonics of 50 Hz in $u(i)$ (sampled at a rate of 420 Hz). For $i > 40 \times 10^3$, it modifies its poles-zeros configuration such that

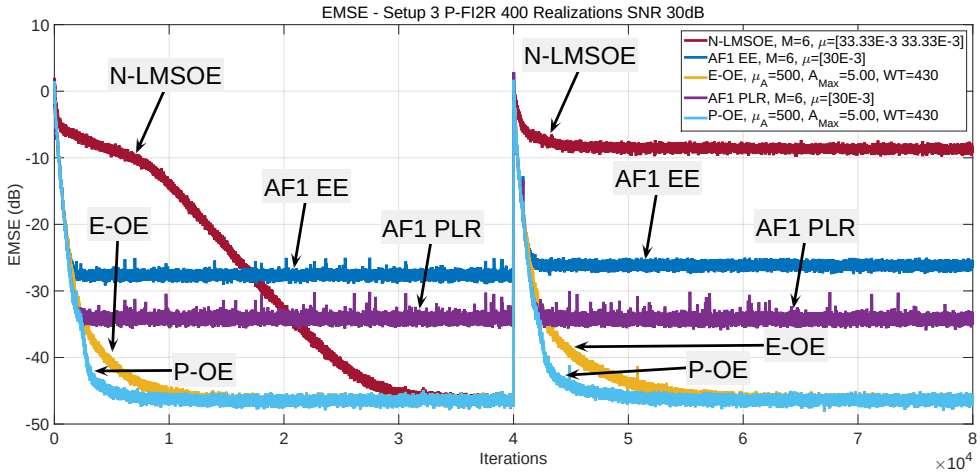

Figure 34: EMSE - Notch SNR = 30 dB (Non-Stationary Scenario)

$w^o = [-0.8\ -0.64\ -0.512\ -0.4096\ -0.3277\ -0.2621\ 1.0\ 1.0\ 1.0\ 1.0\ 1.0\ 1.0\ 1.0]$, starting to attenuate the harmonics of 60 Hz.

The plot of Fig. 34 shows how the P-OE behaves compared to the E-OE in this scenario, where it is seen that the performance of both combinations is quite similar. Considering that the algorithms were tuned to match each other transient as much as possible as remarked in Section 4.3.4, it is worth noting that although the PLR guide is nearly 10 dB better than the EE guide at steady state, the P-OE exhibits only a modest gain over the E-OE in the transient.

Fig. 35 includes the EMSE of the T-OE combinations in the same plot for comparison, but omits all guide filters (AF1) to improve readability. It can be seen that P-OE features the faster learning, reaching nearly 10 dB below the T-OE equipped with BMR transfers at $i = 5 \times 10^3$. In this same interval, E-OE virtually matches the T-OE with PAM transfers, although it is a bit superior to T-OE with BMR. For all mentioned combinations (P-OE, E-OE and T-OE), the transfers cycle is $L = 430$ and, for the T-OE combinations, AF1 has $M_1 = 15$ and $\mu_1 = 100 \times 10^{-3}$. Both E-OE and T-OE have the accurate component (AF2) configured as in the P-OE.

Next, the performance of the P-OE was measured when identifying the but-

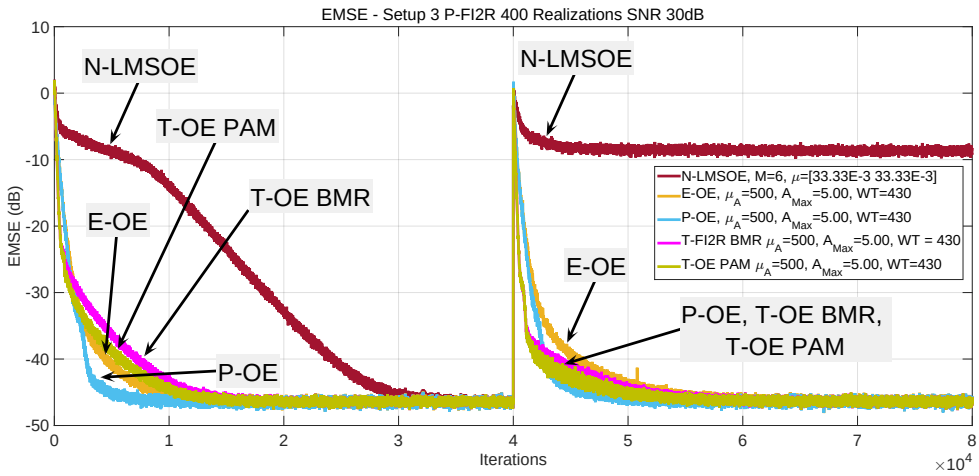

Figure 35: EMSE - Notch SNR = 30 dB (Non-Stationary Scenario)

terworth plant used to test the E-OE combinations (Section 4.3.4), fed by a zero-mean, Gaussian white input $u(i)$ with power $\sigma_u^2 = 1$. The background noise $v(i)$ is Gaussian and zero-mean and two cases are explored, with power $\sigma_v^2 = 10^{-3}$ and $\sigma_v^2 = 10^{-6}$. Like in Section 4.3.4, the plant changes abruptly when $i = 80 \times 10^3$. The transfers cycle used in the combinations is $L = 500$.

Fig. 36 depicts the EMSE of the P-OE compared with that of E-OE when the SNR=30dB. The accurate component of both combinations (AF2) is an N-LMSOE with order $M_2 = 6$ and $\mu_2 = 5 \times 10^{-3}$. The fast component of the E-OE (AF1 EE) is an EE-LMS with order $M_1 = 6$ and $\mu_1 = 50 \times 10^{-3}$ and the fast component of the P-OE combinations (AF1 PLR) has also order $M_1 = 6$ and $\mu_1 = 50 \times 10^{-3}$. Both guides (PLR and EE) are equalized at the transient, but the PLR guide outperforms the EE guide by nearly 5 dB at steady state (indicating that the former is closer to the global minimum since the error surface is monomodal). However, the difference between the combinations is bigger when $i = 80 \times 10^3$ (12 dB) and, considering that the E-OE is almost stagnated at this point, it would increase if the non-stationarity didn't take place. This tendency keeps after the non-stationarity as well.

Fig. 37 depicts the same setup, comparing the P-OE and E-OE against to the

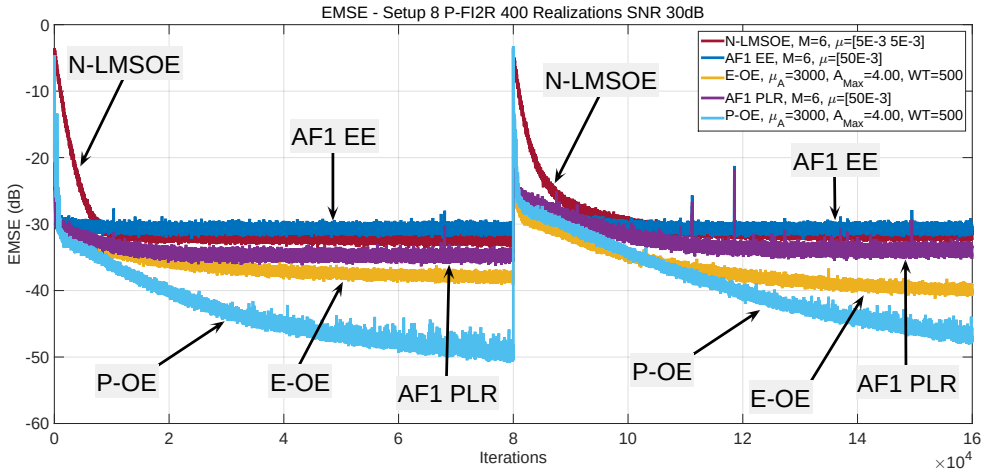

Figure 36: EMSE - Butterworth SNR = 30 dB (Non-Stationary Scenario)

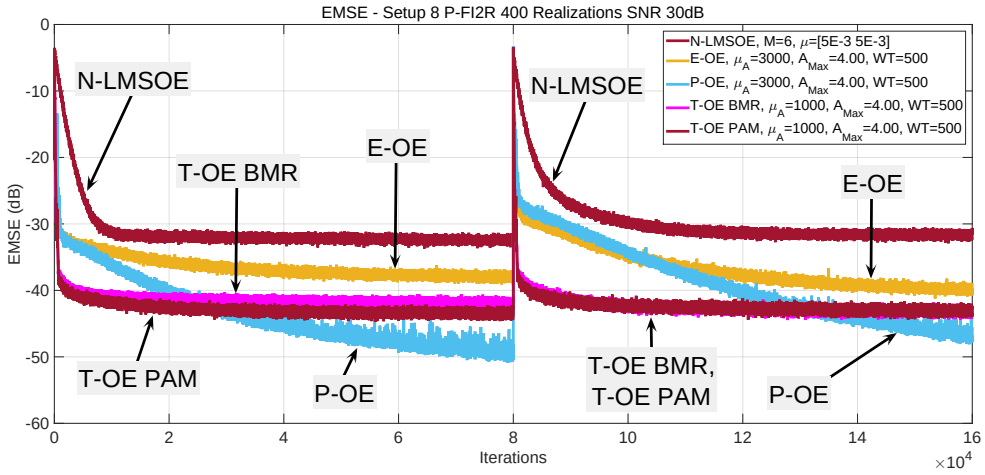

Figure 37: EMSE - Butterworth SNR = 30 dB (Non-Stationary Scenario)

T-OE combinations equipped with BMR and PAM weights transfers. Both T-OE combinations perform alike (mainly after the non-stationarity) and converge initially faster than P-OE or E-OE, but stagnate with an error level of approximately -43dB. For the T-OE combinations, AF1 has $M_1 = 15$ and $\mu_1 = 166.67 \times 10^{-3}$ and AF2 is configured as in the P-OE and E-OE in the setup of Fig. 36.

Fig. 38 shows how a higher SNR (60dB) change the behaviour of P-OE and E-OE, and we note that the performance of both improve considerably (25dB for the P-OE and 35dB for the E-OE). In particular, the E-OE benefits the most from the new SNR, indicating that its noise-induced bias decreased. Indeed, unlike in the prior case where the lower SNR harmed the E-OE far more than the P-OE,

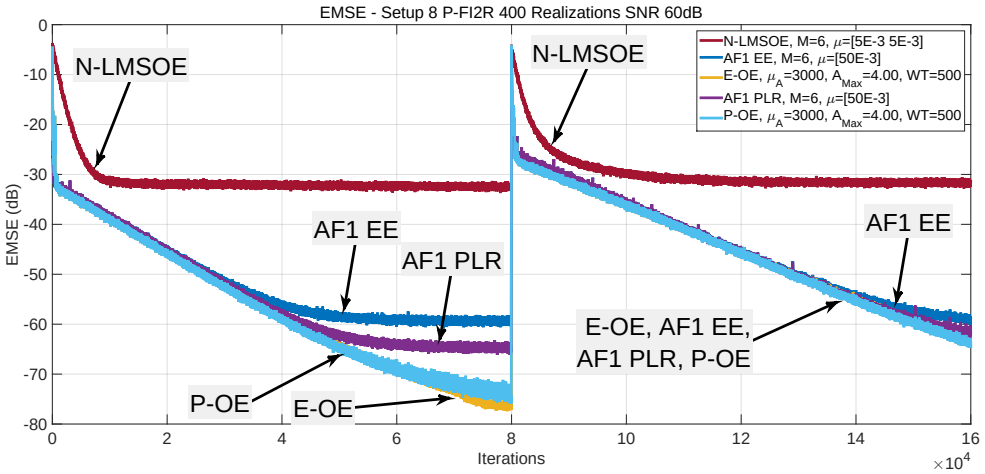

Figure 38: EMSE - Butterworth SNR = 60 dB (Non-Stationary Scenario)

now both are virtually indistinguishable. However, such a high SNR is not often found in real-world applications.

The same experiments are seen in Fig. 39, but with the inclusion of the T-OE combinations to allow an easy performance comparison among all algorithms. Although the higher SNR also contributed to improve the performance of the T-OE combinations (see Figures 32 and 33), their gain was much smaller than with the P-OE or E-OE.

Compared to Fig. 32 (that depicted the T-OE performance in the butterworth scenario with SNR=30dB), the T-OE equipped with the PAM transfers gained 5dB while the T-OE with BMR did not change appreciably. Compared to Fig. 33 (SNR=40dB), *the performance of the* T-OE *combinations look virtually the same* as in Fig. 39 (SNR=60dB). Once the SNR increased two orders of magnitude without any noteworthy impact on the combinations, clearly the additive noise is no longer a major cause for the estimation errors committed by the T-OE. Considering that AF2 is the same for all combinations, this indicates that as the SNR increases other performance limiters such as mismodelling become more evident for the FIR guides.

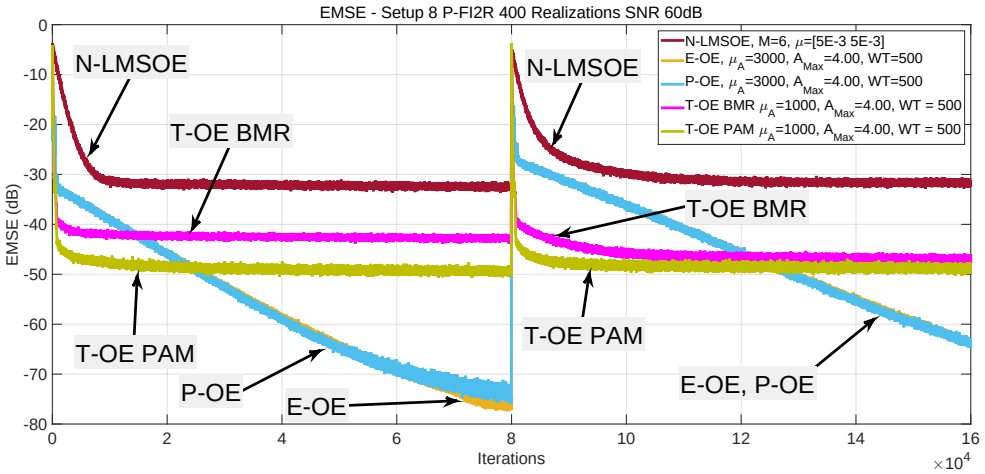

Figure 39: EMSE - Butterworth SNR = 60 dB (Non-Stationary Scenario)

4.3.5 Limitations of the IIR-IIR Combinations

In comparison with the T-OE, the purely IIR combinations (E-OE and P-OE) show a considerable gain in terms of computational complexity as they implement the weights transfers with no mapping at all. However, there is an issue we did not face in the scenarios explored in Section 4.3.4: in case the impulse response of the plant being identified is slowly-decaying (i.e., it is "long tailed"), *the weights transfers may not work.*

In order to illustrate how the EE or PLR weights transfer may fail, let us consider a simple second-order plant given by

$$H^o(z) = \frac{1}{1 - 1.7z + 0.7225z^2}, \tag{4.63}$$

which is an all-poles, non-SPR system whose poles are coincident at 0.85 [60, Section 7.4, Eq. 7.47]. The corresponding impulse response is plotted in Fig. 40, where we note its tail is thicker and takes longer to fade than those seen in the scenarios used in Sections 4.2.7 and 4.3.4 (refer to Fig. 19).

When identifying a system like that, the accurate component of the E-OE and P-OE combinations do not benefit from the weights transfers (at least, not as much as it did in the simulations scenarios used in Section 4.3.4). This can be

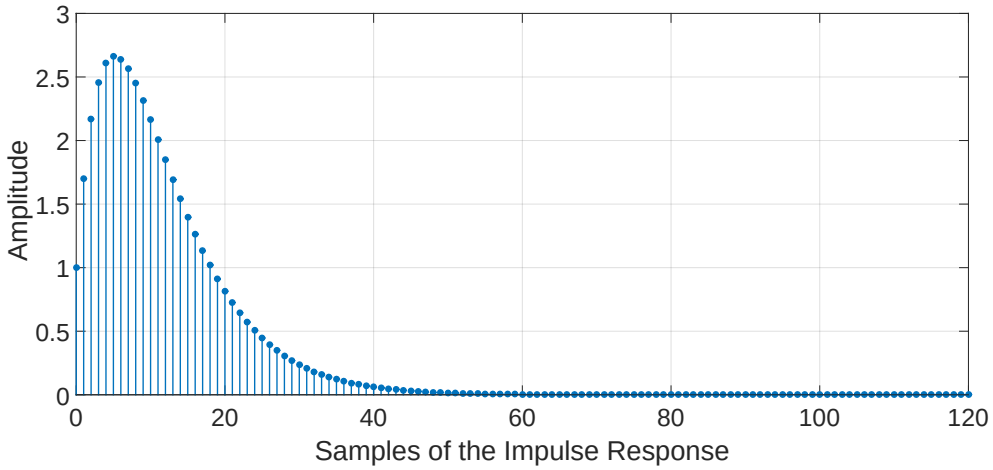

Figure 40: Long Tailed Impulse Response

seen in Fig. 41, which depicts how the EMSE of these associations behave when the plant being identified is described by Eq. 4.63. In these simulations, the input is white Gaussian with unitary power (i.e., $\sigma_u^2 = 1$), $\sigma_v^2 = 1{\times}10^{-2}$ with v also white Gaussian, and $L = 10,000$ within a set of 100,000 iterations. Unlike the scenarios of Section 4.3.4, where the EMSE of both components equalized instantaneously whenever a transfer occurred, here the transfers are quite ineffective or even harmful to the performance.

In the E-OE case, only the first two transfers yielded some gain; in the 3rd, there was no gain at all and the remainder transfers *worsened* AF2. For the P-OE, some gain was observed in the first transfer only: the 2nd and 3rd transfers yielded no gain and the remainder ones also worsened AF2. As a result, both combinations are forced to always track the fastest component (i.e., AF1).

At a first sight, this looks odd because (unlike the FIR guide) both the EE and PLR guides are compatible with the accurate OE component AF2 in the parameters space and, therefore, one could expect that the transfers lead AF2 to the exactly same point onto the performance surface as the guides. However, in the E-OE combination this reasoning is plainly wrong: the components are indeed compatible in the parameters space, but their performance surfaces are

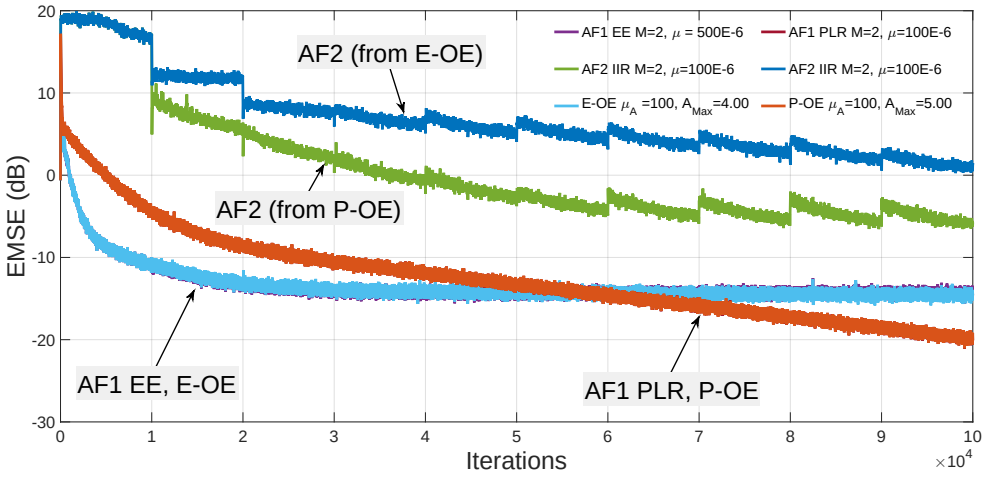

Figure 41: E-OE and P-OE Combinations Identifying a Long-Tailed Plant

not the same.

As a matter of fact, not only the noise-induced bias moves the minima of the EE guide away from that of AF2 but *the shape of the respective performance surfaces is different*, so the same set of parameters makes the components yield different costs. With clustered or underdamped poles (that yield long-tailed impulse responses), the performance surface of AF2 may distort in a way that the minimum will be enclosed into a narrow valley surrounded by nearly-flat plateaus on which the squared error is almost constant. In turn, the EE guide exhibits steeper surfaces on which the error changes more intensely. As a result, when the parameters of the EE guide are transferred to AF2, it might be moved around onto the aforementioned plateaus with little gains in terms of error level - if any.

Even worse than moving around the parameters of AF2 onto a nearly flat surface without any practical effect, the transfers from an EE guide may actually *increase* its error level. This happens when AF2 is closer to the minimum than AF1 but the latter is outperforming the former (which is possible because the surfaces are different), signaling the convex supervisor to enable the transfers and inadvertently preventing AF2 from converging. In the simulations of Section 4.3.4, this was not seen because the convex supervisor commuted to AF2 right

after the first transfers took place.

Although the differences between the error surfaces explain why the transfers are inefficient in the E-OE combinations, there is also another performance limiter. Indeed, given that the PLR guide is an OE AF, its error surface is identical to that of AF2; therefore, the latter should be instantaneously moved to the same point than the former every time a transfer occurs in the P-OE. However, this is not seen in the simulation of Fig. 41: the learning curves of the PLR guide and AF2 do not equalize and may actually divert even more with the transfers. While the exact reasons for this phenomenon are not known, they seem related to one of the more fascinating properties of the IIR systems: their *memory*.

As AF2 is IIR and therefore recursive, it can reach the state $\mathcal{W}_{2,i+1}$ depending upon its current state $\mathcal{W}_{2,i}$, the parameters set $\{\mathsf{A}_{2,i}, \mathsf{b}_{2,i}\}$ and the input signal $u(i)$ according to (see Section 2.5)

$$\mathcal{W}_{2,i+1} = \mathsf{A}_{2,i}\mathcal{W}_{2,i} + \mathsf{b}_{2,i}u(i) \tag{4.64}$$

It is clear that unless the parameters set of AF2 matches that of AF1 for every i, a given state reachable by AF1 (say, $\mathcal{W}_{1,i+1}$) can not be reached simultaneously by AF2. This holds *even if both parameters sets coincide once in a while* (i.e., right after the transfers) because their current states $\mathcal{W}_{1,i}$ and $\mathcal{W}_{2,i}$ are likely to be distinct. Moreover, this means that the outputs $y_1(i)$ and $y_2(i)$ may be quite different right after the transfers if AF2 still "remembers" in which state it was before the transfers. By considering that, it seems reasonable to hypothesize that AF2 has not a strong memory in the scenarios of Section 4.3.4 (which is corroborated by the short IR of the plants), making the transfers effective. This conjecture is supported by experimental evidences and it is currently under investigation [13] .

[13]This issue is not evident because, more or less implicitly, the designer assumes that the signals generated by the IIR AF will be close enough to those of a fixed filter – and this is one

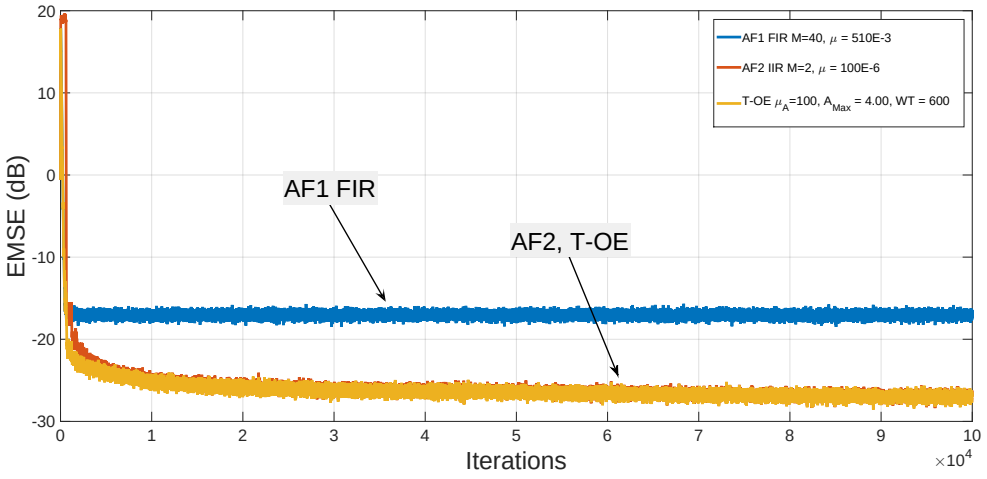

Figure 42: T-OE Endowed with a Long Guide and BMR-Based Transfers Identifying a Long-Tailed Plant

Thus, whenever the impulse response of $H^o(z)$ is endowed with a long tail, the stagnation effect that the transfers are intended to tackle worsens. So, we have to give up on the E-OE and P-OE combinations and switch back to the T-OE, which has an additional parameter to be set: the *order* M_1 of the FIR guide. If, on one hand, it is true that the PAM transfers need only the first $2M_2+1$ samples of the impulse response of the FIR guide to work properly (Section 4.2.3), on the other hand the BMR transfers can benefit if more samples are available as discussed in Section 4.2.4 [14] , outperforming the transfers made by an EE or PLR guides if M_1 is large enough. Intuitively, this suggests that longer FIR guides get closer to the minimum of AF2 than EE or PLR guides whenever the original impulse response is long-tailed.

In contrast with Fig. 41, Fig. 42 shows how a T-OE with a long FIR guide

of the reasons the step size μ must be kept "small enough". Rigorously speaking, every time the AF parameters change, its current state must be computed by reprocessing the whole sequence of inputs $\{u(1)\ u(2)\ u(3)\ \ldots\ u(i)\}$ and only then the error $e(i)$ can be determined together with the other signals that will be used in the next update round. Given that such a procedure is unfeasible, the adaptive algorithm computes all signals from the AF states and outputs like if they were yielded by a fixed filter with coefficients equal to those of the AF at the instant i.

[14] PAM may be modified to exploit more than $2M_2+1$ coefficients, but this will no longer lead to exact solutions as Parks and Burrus remark in [111, Section 7.5, Eq. 7.141]. It is unclear whether such an approach is effective when identifying plants endowed with long-tailed impulse responses but, in our simulations with this kind of plants, BMR-based transfers always worked for a sufficiently large M_1 whereas the PAM transfers as defined in Section 4.2.3 did not.

($M_1 = 40$) behaves under the same circumstances, using $L = 600$. It can be seen that the transfers are now effective, allowing AF2 to reach EMSE levels far below AF1 (about 10 dB) – and this difference can be still greater with longer FIR guides. As a drawback, the number of taps of AF1 (41 coefficients) is large when compared to that of AF2 (3 coefficients) and this extra cost could be prohibitive in some situations. On the other hand, if AF1 is too short, the weights transfers may be ineffective.

As such, we believe that practical, general purpose solutions could take advantage of a strategy to vary the length of the FIR guides like suggested in [124–126]. Probably, the best trade-off between performance and complexity can be reached by reducing the length of AF1 as the supervisor switches to AF2 and increasing it otherwise although it is currently unclear how such a control could be efficiently implemented.

5 PERFORMANCE ANALYSIS

The guide filters (AF1) of the combinations seen in Sections 4.2 and 4.3 (T-OE, P-OE and E-OE) are designed to adapt fast and therefore are expected to reach a steady-state error higher than the accurate component (AF2). As the performances of both AFs are tracked by the mixing factor $\lambda(i)$, clearly the transient of the combination is dominated by AF1 until it stops adapting. From that point on, the combined performance is ruled by AF2 until its convergence. Hence, the steady state mean-square analysis of the overall combination reduces to that of AF2, which we describe in terms of the Energy Conservation Relation (ECR).

The ECR is a theory that describes the FIR adaptive filters via an energy balance that is carried out over the adaptation [5,127,128]. This balance is defined in terms of the *a priori* and *a posteriori* weights error vectors $\tilde{w}_{i-1}$ and $\tilde{w}_i$ given by

$$\tilde{w}_{i-1} = w^o - w_{i-1} \tag{5.1}$$

$$\tilde{w}_i = w^o - w_i, \tag{5.2}$$

the *a priori* and *a posteriori* estimation errors

$$e_a(i) = u_i \tilde{w}_{i-1} \tag{5.3}$$

$$e_p(i) = u_i \tilde{w}_i, \tag{5.4}$$

and the output estimation error $e(i) = d(i) - y(i)$.

From the ECR, it is possible to arrive at a Variance Relation (VR) that allows to compute the desired EMSE and MSE in steady-state. In this work, we show that the same theory holds for LMS-based IIR adaptive filters too and may be used to describe their steady state performance . Transient performance is beyond the scope here as it is strongly influenced by the poles/zeros allocation, making the errors surfaces distorted even with white inputs and correct modelling. However, as aforementioned, in the T-OE context the overall transient is dominated by the FIR component, which was modeled already in Section 4.2.5.

Unlike in the FIR case, the LMSOE mean square analysis is somewhat tricky due the recursivity and an exact solution can not be determined in a closed-form. However, under reasonable assumptions, an approximated relation can be achieved, leading to quite accurate results even with underdamped systems.

We start from the generic IIR update recursion

$$w_i = w_{i-1} + \mu \phi_i^T g\big(e(i)\big), \tag{5.5}$$

where $g\big(e(i)\big)$ cover different adaptive algorithms and, similarly to the FIR case, define the weights error vector

$$\tilde{w}_i = w^o - w_i \tag{5.6}$$

Now, let the *a priori* and *a posteriori* errors $e_p(i)$ and $e_a(i)$ defined as

$$e_a(i) = x_i^o w^o - x_i w_{i-1} \tag{5.7}$$

$$e_p(i) = x_i^o w^o - x_i w_i, \tag{5.8}$$

which, differently from their correlates in the FIR adaptive filtering (Equations 5.3 and 5.4), are not defined in terms of $\tilde{w}_i$ because the regressor vectors x_i^o and x_i are different. As a result, the derivation of an ECR can be difficult and lead

to algebraic pitfalls, which we shall address soon.

By subtracting the update rule of Eq. 5.5 from w^o and then left multiplying by ϕ_i yields

$$w^o - w_i = w^o - w_{i-1} - \mu\phi_i^T g\big(e(i)\big) \tag{5.9}$$

$$\tilde{w}_i = \tilde{w}_{i-1} - \mu\phi_i^T g\big(e(i)\big) \tag{5.10}$$

$$\phi_i\tilde{w}_i = \phi_i\tilde{w}_{i-1} - \mu\phi_i\phi_i^T g\big(e(i)\big) \tag{5.11}$$

$$\bar{e}_p(i) = \bar{e}_a(i) - \mu\|\phi_i\|^2 g\big(e(i)\big), \tag{5.12}$$

where the *filtered a posteriori error* $\bar{e}_p(i)$, defined as

$$\bar{e}_p(i) \triangleq \phi_i\tilde{w}_i \tag{5.13}$$

and the *filtered a priori error* $\bar{e}_a(i)$ defined as

$$\bar{e}_a(i) \triangleq \phi_i\tilde{w}_{i-1} \tag{5.14}$$

are actually placeholders for the *a posteriori* and *a priori* errors $e_p(i)$ and $e_a(i)$ (see Eqs. 5.8 and 5.7) at steady-state, as we show later on.

Next, $g\big(e(i)\big)$ can be isolated in Eq. 5.12 as

$$g\big(e(i)\big) = \frac{\bar{e}_a(i) - \bar{e}_p(i)}{\mu\|\phi_i\|^2} \tag{5.15}$$

and then be replaced into Eq. 5.10, leading to

$$\begin{aligned}\tilde{w}_i &= \tilde{w}_{i-1} - \mu\phi_i^T g\big(e(i)\big) \\ &= \tilde{w}_{i-1} - \phi_i^T \frac{\bar{e}_a(i) - \bar{e}_p(i)}{\|\phi_i\|^2}\end{aligned} \tag{5.16}$$

and

$$\tilde{w}_i + \frac{\phi_i^T}{\|\phi_i\|^2}\bar{e}_a(i) = \tilde{w}_{i-1} + \frac{\phi_i^T}{\|\phi_i\|^2}\bar{e}_p(i) \tag{5.17}$$

By squaring both sides of Eq. 5.17 and applying the definition of the filtered

errors $\bar{e}_p(i)$ and $\bar{e}_a(i)$ (Eqs. 5.13 and 5.14), an exact Energy Conservation Relation can be derived straightforwardly as

$$\|\tilde{w}_i\|^2 + \frac{1}{\|\phi_i\|^2}\bar{e}_a^2(i) = \|\tilde{w}_{i-1}\|^2 + \frac{1}{\|\phi_i\|^2}\bar{e}_p^2(i), \tag{5.18}$$

and then a Variance Relation can be pursued as shown in the sequel.

Variance Relation (VR)

Assumption 1 (Steady State Expectancies)**.** *At the steady state, the following identities hold*

$$\mathrm{E}\,\tilde{w}_i = s \qquad i \to \infty \tag{5.19}$$

$$\mathrm{E}\,\tilde{w}_i\tilde{w}_i^T = C \qquad i \to \infty, \tag{5.20}$$

and usually $s \to 0$.

Assumption 1 states that both w_i and $\tilde{w}_i\tilde{w}_i^T$ do not change in the mean sense as $i \to \infty$ [1] . Hence, assuming that the AF is at steady state, Eq. 5.18 becomes

$$\mathrm{E}\,\frac{1}{\|\phi_i\|^2}\bar{e}_a^2(i) = \mathrm{E}\,\frac{1}{\|\phi_i\|^2}\bar{e}_p^2(i) \text{ as } i \to \infty \tag{5.21}$$

By taking $\bar{e}_p(i)$ from Eq. 5.12 and replacing it in Eq. 5.21,

$$\begin{aligned}
\mathrm{E}\,\frac{1}{\|\phi_i\|^2}\bar{e}_a^2(i) &= \mathrm{E}\,\frac{1}{\|\phi_i\|^2}\Big(\bar{e}_a(i) - \mu\|\phi_i\|^2 g\big(e(i)\big)\Big)^2 \\
&= \mathrm{E}\,\frac{1}{\|\phi_i\|^2}\Big(\bar{e}_a^2(i) + \mu^2\|\phi_i\|^4 g^2\big(e(i)\big) - 2\mu\bar{e}_a(i)\|\phi_i\|^2 g\big(e(i)\big)\Big) \\
&= \mathrm{E}\,\frac{\bar{e}_a^2(i)}{\|\phi_i\|^2} + \mu^2\|\phi_i\|^2 g^2\big(e(i)\big) - 2\mu\bar{e}_a(i)g\big(e(i)\big),
\end{aligned} \tag{5.22}$$

[1] Given that the OE-based algorithms are unbiased, the expected value $\mathrm{E}\,\tilde{w}_i$ is 0 (i.e., strictly speaking $s = 0$ in Assumption 1). However, in practice, the adaptation can be so hard that the steady state might be declared before the AF reaches the global minimum simply because the convergence rate is too low and the adaptation looks stagnated. Herewith, s could be incorrectly perceived as non-zero although such a fact does not invalidate the Variance Relation in Eq. 5.23.

and, by making the appropriate manipulations,

$$2\,\mathrm{E}\,\bar{e}_a(i)g\big(e(i)\big) = \mu\,\mathrm{E}\,\|\phi_i\|^2 g^2\big(e(i)\big), \tag{5.23}$$

which is the Variation Relation of an IIR AF. For the N-LMSOE, the error is normalized by the power of ϕ_i (Section 2.8.2, Eq. 2.87) such that

$$g\big(e(i)\big) = \frac{e(i)}{\epsilon + \|\phi_i\|^2}, \tag{5.24}$$

where ϵ is a regularization factor and a new assumption is required.

Assumption 2. *The regularization factor ϵ is small compared to $\|\phi_i\|^2$.*

If the regularization factor ϵ is small enough when compared to $\|\phi_i\|^2$ (which is usually the case) then, under persistent excitation, the N-LMSOE may be retrieved by neglecting ϵ. Therefore, $g\big(e(i)\big) = {}^{e(i)}/_{\|\phi_i\|^2}$ and Eq. 5.23 becomes

$$2\,\mathrm{E}\,\bar{e}_a(i)g\big(e(i)\big) = \mu\,\mathrm{E}\,\|\phi_i\|^2 g^2\big(e(i)\big) \implies 2\,\mathrm{E}\,\bar{e}_a(i)\frac{e(i)}{\|\phi_i\|^2} = \mu\,\mathrm{E}\,\frac{e^2(i)}{\|\phi_i\|^2} \tag{5.25}$$

Once the data model establishes that $d(i) = x_i w^o + v(i)$ and by considering the definition of $e_a(i)$ in Eq. 5.7, it is straightforward to show that

$$e(i) = e_a(i) + v(i) \tag{5.26}$$

which, replaced in Eq. 5.25, yields

$$2\,\mathrm{E}\,\bar{e}_a(i)\frac{e_a(i) + v(i)}{\|\phi_i\|^2} = \mu\,\mathrm{E}\,\frac{\big(e_a(i) + v(i)\big)^2}{\|\phi_i\|^2} \tag{5.27}$$

$$2\,\mathrm{E}\,\frac{\bar{e}_a(i)e_a(i)}{\|\phi_i\|^2} = \mu\,\mathrm{E}\,\frac{e_a^2(i) + v^2(i)}{\|\phi_i\|^2}$$

$$2\,\mathrm{E}\,\frac{\bar{e}_a(i)e_a(i)}{\|\phi_i\|^2} = \mu\,\mathrm{E}\,\frac{e_a^2(i)}{\|\phi_i\|^2} + \mu\sigma_v^2\,\mathrm{E}\,\frac{1}{\|\phi_i\|^2}, \tag{5.28}$$

where we remember that, by definition, the additive noise $v(i)$ is zero-mean (i.e., $\mathrm{E}\,v(i) = 0$) and independent from $\bar{e}_a(i)$; therefore, $\mathrm{E}\,\bar{e}_a(i)v(i) = 0$.

Proceeding, we note that $\|\phi_i\|^2$ is statistically dependent on both $e_a^2(i)$ and

$\bar{e}_a(i)e_a(i)$, which means that the terms of Eq. 5.28 can not be further developed and a new assumption is required to move on.

Assumption 3 (Separation Principle)**.** $\|\phi_i\|^2$ *is independent of either* $e_a^2(i)$ *or* $\bar{e}_a(i)e_a(i)$ *as* $i \to \infty$.

This assumption is an extension of the commonly adopted independence principle in the FIR case [5, Section 16.3] [127], which states that the regressor power $\mathrm{E}\,\|u_i\|^2$ and the *a priori error* $e_a^2(i)$ becomes statistically independent as $i \to \infty$. As such, it can be justified through the same arguments here: the parameters updates decrease in magnitude at steady state because the gradient fades when w_i approaches the global minimum; hence, in the limit, the behavior of $\tilde{w}_i$ is less sensitive to the input data and the statistical dependency upon x_i and ϕ_i becomes weaker. Hence, at steady state, $e_a(i)$ and $\bar{e}_a(i)$ become less sensitive to x_i and ϕ_i as well.

For the sake of this argument, consider the stand-alone N-LMSOE AF ($H_2(z)$) used in the notch scenario of Sections 4.2.7 and 4.3.4. Fig. 43 shows the sequences $\mathrm{E}\,\|\phi_i\|^2 e_a^2(i)$ and $\mathrm{E}\,\|\phi_i\|^2\,\mathrm{E}\,e_a^2(i)$ yielded by $H_2(z)$ in this scenario and, in accordance to the Separation Principle, both curves overlap soon after the AF starts to adapt and become indistinguishable as the adaptation evolves (and the same effect occurs if $e_a^2(i)$ is replaced by $\bar{e}_a(i)e_a(i)$). To allow a better visualization, Fig. 44 provides a zoom of the same curves in the beginning of the adaptation, where its clearer that the curves tend to each other.

Once the Separation Principle was shown to be reasonable, it can be applied to Eq. 5.28, yielding

$$\begin{aligned} 2\,\mathrm{E}\,\frac{\bar{e}_a(i)e_a(i)}{\|\phi_i\|^2} &= \mu\,\mathrm{E}\,\frac{e_a^2(i)}{\|\phi_i\|^2} + \mu\sigma_v^2\,\mathrm{E}\,\frac{1}{\|\phi_i\|^2} \\ 2\,\mathrm{E}\,\bar{e}_a(i)e_a(i)\,\mathrm{E}\,\frac{1}{\|\phi_i\|^2} &= \mu\,\mathrm{E}\,e_a^2(i)\,\mathrm{E}\,\frac{1}{\|\phi_i\|^2} + \mu\sigma_v^2\,\mathrm{E}\,\frac{1}{\|\phi_i\|^2}, \end{aligned} \tag{5.29}$$

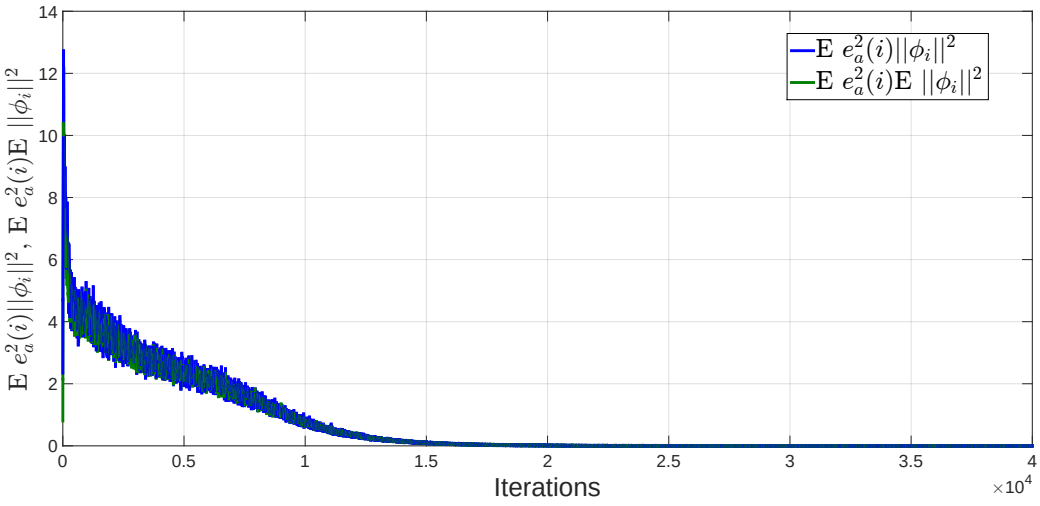

Figure 43: The Separation Principle

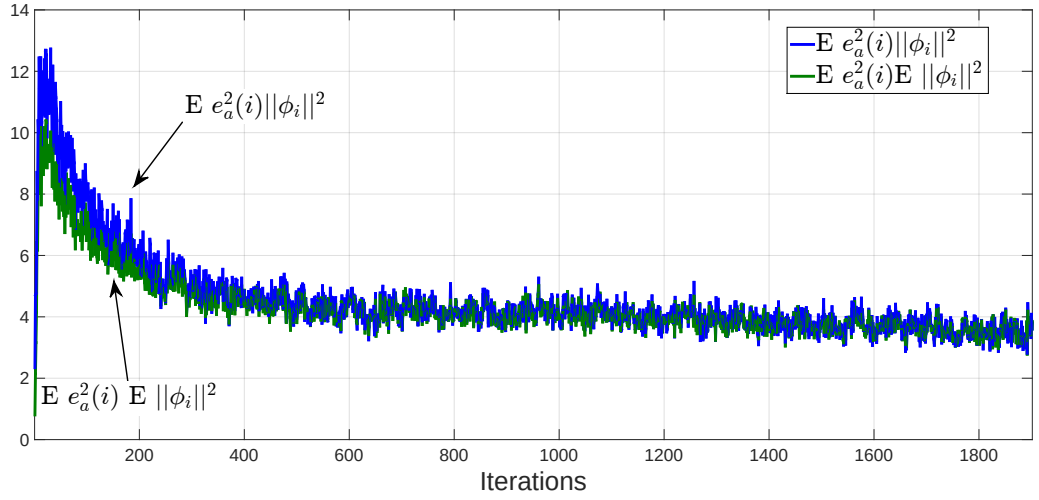

Figure 44: The Separation Principle (zoom)

The next step is to relate $\bar{e}_a(i)e_a(i)$ with $e_a^2(i)$ in order to make Eq. 5.29 tractable. This can be made by restating $e_a(i)$ from Eq. 5.7 as

$$\begin{aligned} e_a(i) &= y^o(i) - y(i) = x_i^o w^o - x_i w_{i-1} \\ &= B^o(z)u(i) + A^o(z)y^o(i) - B(z)u(i) - A(z)y(i) \\ &= \big(B^o(z) - B(z)\big)u(i) \\ &\big(A^o(z) - A(z)\big)y^o(i) + A(z)\big(y^o(i) - y(i)\big) \\ &= x_i^o \tilde{w}_{i-1} + A(z)e_a(i) \qquad (5.30) \\ \therefore\, e_a(i)\big(1 - A(z)\big) &= x_i^o \tilde{w}_{i-1} \\ e_a(i) &= \frac{1}{1 - A(z)} x_i^o \tilde{w}_{i-1} \qquad (5.31) \end{aligned}$$

where to improve clarity, $B(z, i-1)$ and $A(z, i-1)$ are replaced by $B(z)$ and $A(z)$. This manipulation makes $e_a(i)$ a function of the regressor x_i^o with no explicit mention to $A^o(z)$, which is convenient for analysis purposes.

Furthermore, $\bar{e}_a(i)$ (Eq.5.14) can be expanded by using the definition of ϕ_i given in Eq. 2.78, so that

$$\bar{e}_a(i) = \phi_i \tilde{w}_{i-1} = \left(x_i + \sum_{k=1}^{M_2} a_k(i-1)\phi_{i-k} \right) \tilde{w}_{i-1} \qquad (5.32)$$

Now, due to their recursivity, IIR adaptive filters are endowed with relatively small step sizes (refer to Chapter 1). In particular, the LMSOE and the N-LMSOE algorithms are derived by implicitly assuming that $w_{i-1-k} = w_{i-1}$ for $1 \leq k \leq M_2$ as discussed in Section 2.8.2 (Equations 2.72–2.80). Therefore, with no loss of generality w_{i-1} may be replaced by w_{i-1-k} so that $\tilde{w}_{i-1} = \tilde{w}_{i-1-k}$ for $1 \leq k \leq M_2$. As a result, $\bar{e}_a(i)$ from Eq. 5.32 may be restated as

$$\begin{aligned} \bar{e}_a(i) = \phi_i \tilde{w}_{i-1} &= \left(x_i + \sum_{k=1}^{M_2} a_k(i-1)\phi_{i-k} \right) \tilde{w}_{i-1} \\ &= x_i \tilde{w}_{i-1} + \sum_{k=1}^{M_2} a_k(i-1)\phi_{i-k}\tilde{w}_{i-1-k} \end{aligned}$$

$$
\begin{aligned}
&= x_i\tilde{w}_{i-1} + A(z)\bar{e}_a(i) \\
&= \frac{1}{1-A(z)}x_i\tilde{w}_{i-1},
\end{aligned} \tag{5.33}
$$

where the similarity between $\bar{e}_a(i)$ and $e_a(i)$ (Eq. 5.31) now becomes apparent.

Assumption 4 (Equivalence of the cross product)**.** *Under the small step-size assumption, the cross product* $\mathrm{E}\, e_a(i)\bar{e}_a(i)$ *equals* $\mathrm{E}\, e_a^2(i)$.

By using the definitions of $e_a(i)$ and $\bar{e}_a(i)$ in Eqs. 5.31 and 5.33,

$$
\begin{aligned}
e_a(i) - \bar{e}_a(i) &= \frac{1}{1-A(z)}x_i^o\tilde{w}_{i-1} - \frac{1}{1-A(z)}x_i\tilde{w}_{i-1} \\
&= \frac{1}{1-A(z)}\big(x_i^o - x_i\big)\tilde{w}_{i-1},
\end{aligned} \tag{5.34}
$$

where $\big(x_i^o - x_i\big)$ and $\tilde{w}_{i-1}$ are readily recognized as, respectively, the vectors $[e_a(i-1)\ e_a(i-2)\ \ldots\ e_a(i-M_2+1)\ 0\ 0 \ldots\ 0\ 0]$ and $[a_1^o - a_1\ a_2^o - a_2 \ldots\ a_{M_2}^o - a_{M2}\ b_1^o - b_1\ b_2^o - b_2 \ldots\ b_{M_2}^o - b_{M_2}]$. Hence, by defining $\tilde{A}(z)$ as

$$
\tilde{A}(z) = A^o(z) - A(z) = \sum_{k=1}^{M_2}\big(a^o - a_k(i-1)\big)z^k, \tag{5.35}
$$

then $\big(x_i^o - x_i\big)\tilde{w}_{i-1} = \tilde{A}(z)e_a(i)$ and Eq. 5.34 can be restated as

$$
\begin{aligned}
e_a(i) - \bar{e}_a(i) &= \frac{\tilde{A}(z)}{1-A(z)}e_a(i) \\
&= \frac{\big(1-A(z)\big) - \big(1-A^o(z)\big)}{1-A(z)}e_a(i) \\
&= e_a(i) - \frac{1-A^o(z)}{1-A(z)}e_a(i),
\end{aligned} \tag{5.36}
$$

By multiplying by $e_a(i)$ and taking expectations,

$$
\mathrm{E}\, e_a(i)\bar{e}_a(i) = \mathrm{E}\, e_a(i)\left(\frac{1-A^o(z)}{1-A(z)}e_a(i)\right) \tag{5.37}
$$

Under white inputs and sufficient modeling, OE-NLMS is unbiased, i.e., $\mathrm{E}\, A(z) = A^o(i)$ and $\mathrm{E}\, \tilde{A}(z) = 0$. Herewith, if μ is small enough as it is usually the case with IIR AFs as we discussed in Chapter 1, $\tilde{A}(z)$ becomes a residual sequence whose

power tends to 0 at steady state [2] . So,

$$\mathrm{E}\, e_a(i) \left(\frac{1 - A^o(z)}{1 - A(z)} e_a(i) \right) \rightarrow \mathrm{E}\, e_a^2(i), \tag{5.38}$$

and the following approximation becomes plausible for Eq. 5.37

$$\mathrm{E}\, e_a(i)\bar{e}_a(i) = \mathrm{E}\, e_a(i) \left(\frac{1 - A^o(z)}{1 - A(z)} e_a(i) \right) = \mathrm{E}\, e_a^2(i) \tag{5.39}$$

By using this result, Eq. 5.29 becomes

$$2\, \mathrm{E}\, e_a^2(i)\, \mathrm{E}\, \frac{1}{\|\phi_i\|^2} = \mu\, \mathrm{E}\, e_a^2(i)\, \mathrm{E}\, \frac{1}{\|\phi_i\|^2} + \mu\sigma_v^2\, \mathrm{E}\, \frac{1}{\|\phi_i\|^2}, \tag{5.40}$$

and the final expression for the EMSE results

$$\zeta \triangleq \mathrm{E}\, e_a^2(i) = \frac{\mu\sigma_v^2}{2 - \mu} \quad \text{as } i \rightarrow \infty \tag{5.41}$$

It is interesting to note that Eq. 5.41 holds also for a FIR N-LMS AF, with no references to the 1st and 2nd moments of the regressors x_i and ϕ_i. This corroborates the notion that the normalized LMS steady state is independent of the statistics of the input signal, suggesting it is more robust to correlations or changes in the power of the input.

The result is quite accurate as there is a good agreement between Eq. 5.41 and the actual measured results for a wide range of μ, as shown in Figs. 45 and 46. These charts exhibit the measured EMSE for the butterworth and notch cases against the theory given by Eq. 5.41. Also, the Figures show the cross-products $\mathrm{E}\,|\bar{e}_a(i) \cdot e_a(i)|$ for both cases, and we note that they are virtually indistinguishable from $\mathrm{E}\, e_a^2(i)$ within the scale of the plots, corroborating the conclusion seen in Eq. 5.40. Fig. 47 shows more accurately the mismatch between $\bar{e}_a(i)$ and $e_a(i)$ as measured by $\mathrm{E}\left(\bar{e}_a(i) - e_a(i)\right)^2$, and it is seen that the fit improves as μ decreases.

[2] Once $A^o(z) = A(z) + \tilde{A}(z)$, $1 - A^o(z) / 1 - A(z) \rightarrow 1$ in case $\tilde{A}(z) \rightarrow 0$ or diverts from 1 otherwise. However, as IIR AFs have to have small values for μ by construction, usually $\tilde{A}(z) \rightarrow 0$.

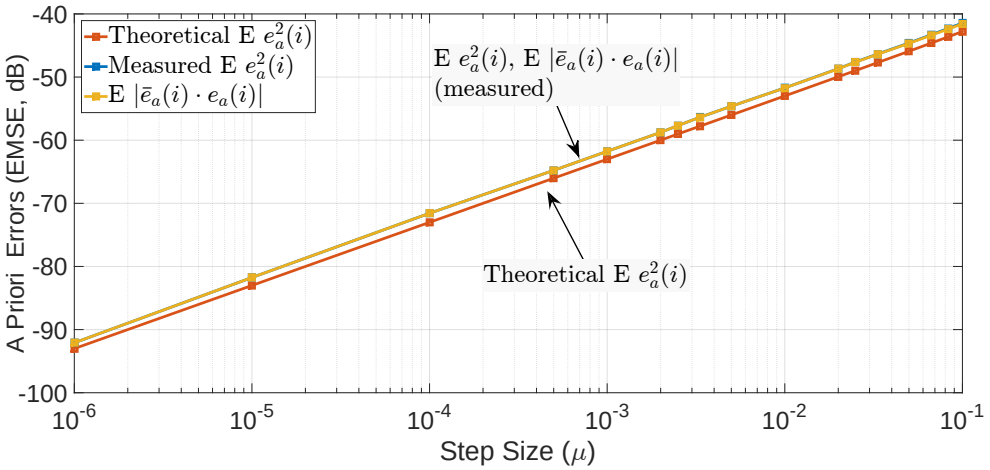

Figure 45: EMSE for the Butterworth Scenario

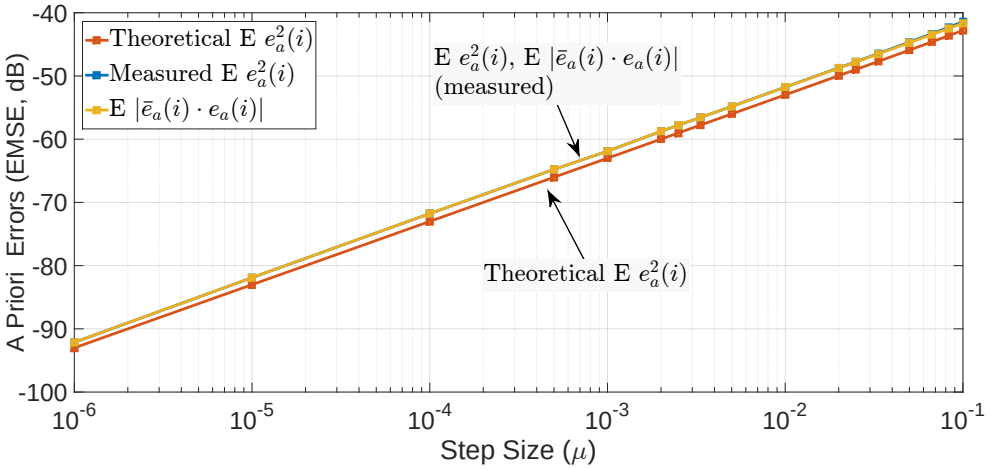

Figure 46: EMSE for the Notch Scenario

The match between the theoretical EMSE and the actual values are not perfect as there is always a narrow margin between the curves seen in Figs. 45 and 46, with the theory providing a lower bound for the actual steady state level. However, this error margin is acceptable in most of the cases since it is smaller than 3dB in a scale that ranges from -90dB to -40dB, *increasing only in extreme cases where the plant is endowed with very high-magnitude poles and/or the AF is equipped with large step sizes.*

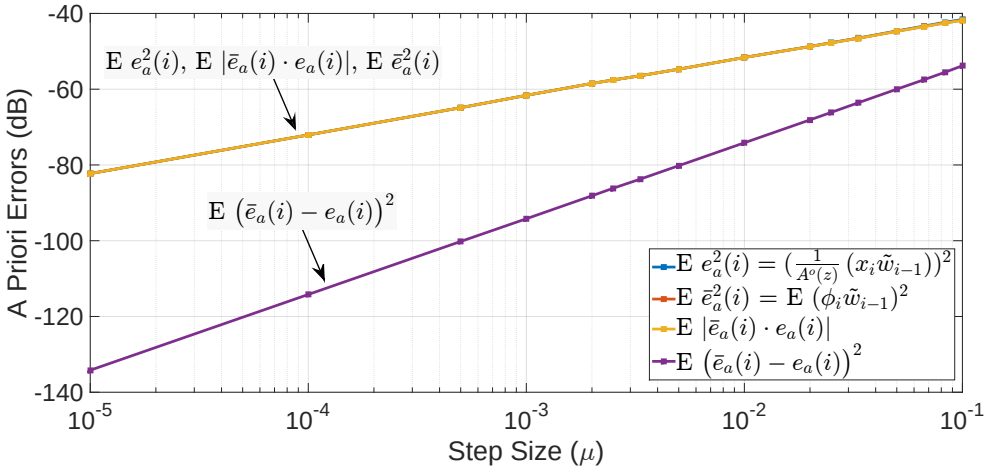

Figure 47: EMSE for the Butterworth Scenario

6 CONCLUSIONS AND COMMENTS

When compared with other approaches in which different filtering structures are used to improve the performance of IIR AFs (e.g., [33,34,129], among others), the T-OE we first introduced in [28] offers some advantages: adaptive switching between the components, universality [1] , enhanced tracking capabilities and improved transient thanks to the weights transfers. In addition, the current work exploited the simpler PAM-based weights transfers and purely IIR adaptive filters combinations (the E-OE and P-OE).

In theory, the BMR-based transfers look more appealing as they always lead to stable realizations. However, in practice this is often innocuous *provided that* $H_2(z)$ is endowed with a sufficiently small step size (such that it operates far bellow the exponential stability bounds) and the cycle transfer L is not too small. In turn, PAM provides an *exact* mapping provided that $2M_2 \leq M_1$.

Anyway, the accuracy requirements for $H_2(z)$ makes its step size arbitrarily small and a lower-bound for L is provided in Section 4.2. On top of that, the transfers are *unidirectional* and *periodical*, which means that the accurate component can be always recovered since the fast component $H_1(z)$ is robust by design. These facts indicate that the PAM transfers should be chosen as the preferred method to implement the mapping unless the operational conditions explicitly

[1] Here, *universality* denotes that a given combination performs at least as well as its best component. In the context of FIR combinations, Arenas-García et al. [78] managed to show that a correctly configured convex supervisor can lead to universal performance or even *outperform* its components in case they meet certain conditions.

encourages the use of BMR (for instance, if long FIR guides are used).

The purely IIR-IIR combinations E-OE and P-OE are simpler alternatives to T-OE that implement an efficient weights transfers mechanism without resorting to any kind of mapping as seen in Section 4.3. In a sense, these combinations look like the composite filters mentioned in Appendix A as they merge different IIR filtering formulations into a single structure. Notwithstanding, unlike the composite filters, both E-OE and P-OE are truly adaptive and therefore have a finer control over the switching between the filtering formulations, which results in better steady state error levels and advanced tracking capabilities. In comparison with the T-OE though, either the E-OE or P-OE (and, by extension, also the composite filters) can not identify properly plants endowed with long-tailed impulse responses as seen in Section 4.3.5.

The Mean Square Steady State Analysis seen in Chapter **5** applied the theory of the Energy Conservation introduced by Sayed [5, 127, 128] to the IIR case. Although the results are not exact, they are quite accurate as Figures 45, 46 and 47 show. Considering that the recursivity of the IIR AFs results into a non-linear dynamics, this is a very useful result. Indeed, to the best of our knowledge, there were not comprehensive analytical methods to determine the steady state performance of IIR AFs so far. As such, the only tool available to study the behaviour of arbitrary LMSOE-based filters were analyses made via computer simulations.

A major virtue of the analysis developed in Chapter **5** is that it is *comprehensive* in the sense it does not rely on any strong assumption other than correct modelling and the usual statistical independences as stated by the *Separation Principle* (Fig. 44). In fact, although it resorted to the commonly used assumption that the coefficients of the AF are changing slowly, the analysis matched closely the measured EMSE levels for a wide range of μ as unveiled by Figures

45, 46 and 47. Also, there are no further constraints on the transfer function of the plant being identified except those related to numerical difficulties caused by pole-zero cancellations or underdamped poles. This is remarkable once the analytical results found in the literature normally assume constrained poles and zeros (typical in the design of adaptive IIR notch filters) or reduced orders (1st and 2nd orders).

Appendices

A COMPOSITE FILTERS

A.1 The Composite Regressor Algorithm (CRA)

The *Composite Regressor Algorithm* (CRA), introduced by Kenney and Rohrs [33], seeks to merge the OE (Output Error) and EE (Equation Error) formulations into a single framework. The filter output is yielded by means of a *composite regressor* whose feedback related elements are a convex combination of the AF and plant outputs according to the scheme described by Eqs. A.1–A.3.

$$\bar{x}_i = [\bar{y}(i-1)\ \bar{y}(i-2)\cdots\ \bar{y}(i-M)\ u(i)\ u(i-1)\ \cdots\ u(i-M)] \tag{A.1}$$

$$y(i) = \bar{x}_i w_{i-1} \tag{A.2}$$

$$\bar{y}(i) = (1-\lambda(i))y(i) + \lambda(i)d(i) \tag{A.3}$$

In the CRA, $\lambda(i)$ is a weighting parameter that makes the AF to feature properties from both the EE and OE realizations, $\bar{y}(i)$ is the *composite output*, $y(i)$ is the filter output, $d(i)$ is the measurable plant output (see Fig. 1 in Chapter 1) and $\bar{x}_i$ is the *composite regressor* for which the algorithm is named after. $\lambda(i)$ is either fixed or, in *a posteriori output error* fashion, time-varying according to

$$\lambda(i) = \frac{\mu \bar{x}'_i {\bar{x}'_i}^T}{1 + \mu \bar{x}'_i {\bar{x}'_i}^T}, \tag{A.4}$$

where $\bar{x}'_i = [\bar{y}'(i)\ \bar{y}(i-1)\ \bar{y}(i-2)\cdots\ \bar{y}(i-M)\ u(i)\ u(i-1)\ \cdots\ u(i-M)]$ with $\bar{y}'(i) = \bar{x}_i w_i$.

As the outputs collected by the vector $\bar{x}_i$ (i.e., $\{\bar{y}(i-k)\}_{k=1}^{M}$) are a combination of the plant outputs like in the EE approach and the filter outputs like in the OE approach, the resulting AF can behave as either of them according to the current value of $\lambda(i)$. This way, the AF is expected to initially adapt in a stable, predictable fashion (like a LMSEE AF) to finally converge to the global minimum with no bias (like a LMSOE AF).

However, it is not clear if the convergence is actually unbiased in the general case since the update rule of the algorithm, given by

$$e(i) = d(i) - y(i) \tag{A.5}$$

$$w_i = w_{i-1} + \frac{\mu\ \bar{x}_i^T e(i)}{1 + \mu \bar{x}_i \bar{x}_i^T}, \tag{A.6}$$

does not truly avoid bias or follow the gradient descent for all values of $\lambda(i)$. In fact, by making $\lambda(i) = 1$, $\bar{y}(i) = d(i)$ and then the CRA becomes an EE realization, hence, subjected to bias whenever the additive noise $v(i)$ is not zero as discussed in Section 2.4.2.

On the other hand, if $\lambda(i) = 0$, then $\bar{y}(i) = y(i)$, which surely leads to an OE realization. However, that would *not* be the bias-proof LMSOE once the regressor is not filtered by the polynomial that defines the poles of the AF (Eqs. 2.78 and 2.80), but a degenerate form of it known as Pseudo Linear Regression or PLR (Section 4.3.3). Now, although PLR is stable [130], it may not converge unless the unknown plant meets the so-called *Strictly Positive Real* (SPR) condition [20, 25, 58](Eq. 4.62).

In a nutshell, if $v(i)$ is not zero and $\lambda(i)$ is fixed, unbiased convergence can not be assured even when the plant is SPR because the composite algorithm will never commute entirely to OE mode (except if $\lambda(i) = 0$, of course, but in this case the fast EE mode will not be used). If $\lambda(i)$ is time-decaying and becomes 0, CRA could fail converging whenever the plant is not SPR.

A.2 The Combined Square Error Algorithm (CSE)

The *Combined Square Error* (CSE) algorithm introduced by Netto and Agathoklis [34] is also a composite solution that merges the good characteristics of the EE and OE into a single structure. Instead of combining regressors though, the CSE combines directly squared errors as shown in Eqs. A.7-A.11, where $x_{EE,i}$ and $e_{EE}(i)$ are the EE regressor and estimation error, $x_{OE,i}$ and $e_{OE}(i)$ are the OE regressor and estimation error and $e^2_{CSE}(i)$ is the *combined square error*.

$$x_{EE,i} = [d(i-1)\ d(i-2)\ \cdots\ d(i-M)\ u(i)\ u(i-1)\ \cdots\ u(i-M)] \tag{A.7}$$

$$e_{EE}(i) = d(i) - x_{EE,i}w_{i-1} \tag{A.8}$$

$$x_{OE,i} = [y(i-1)\ y(i-2)\ \cdots\ y(i-M)\ u(i)\ u(i-1)\ \cdots\ u(i-M)] \tag{A.9}$$

$$e_{OE}(i) = d(i) - x_{OE,i}w_{i-1} \tag{A.10}$$

$$e^2_{CSE}(i) = \lambda(i)e^2_{EE}(i) + \big(1-\lambda(i)\big)e^2_{OE}(i) \tag{A.11}$$

Unlike in CRA, $\lambda(i)$ is continuously adapted by the gradient-based strategy shown in Eq. A.12, where μ_λ is a constant learning step. This way, the algorithm itself finds the optimal composition for $e^2_{CSE}(i)$ and $\lambda(i)$ automatically ranges within the interval [0 1] (in fact, it has to be *constrained* to such an interval).

$$\lambda(i) = \lambda(i-1) - \mu_\lambda|e^2_{EE}(i) - e^2_{OE}(i)| \tag{A.12}$$

A remarkable advantage of the CSE update rule over that of CRA is that the former, shown in Eq. A.13, commutes to LMSOE if $\lambda(i) = 0$ while the latest does to PLR as Eq. A.6 omits the filtered regressor $\phi(i)$. Hence, in this case Eq. A.13 boils down to the ordinary LMSOE update rule (Eq. 2.80) and no bias will be noticed even if the additive noise $v(i)$ is not zero.

$$w_i = w_{i-1} + \mu\Big(\lambda(i)x_i^T e_{EE}(i) + \big(1-\lambda(i)\big)\phi_i^T e_{OE}(i)\Big) \tag{A.13}$$

All that being said, note that the modulus in Eq. A.12 makes $\lambda(i)$ a decreasing sequence, which forces the commutation to the OE mode in the long run despite the relative values of $e_{EE}(i)$ and $e_{OE}(i)$. Considering that the EE formulation could outperform the OE in some situations (e.g., when tracking a time varying plant), this is inconvenient.

Interesting enough, the original gradient-based update rule for $\lambda(i)$ was explicitly modified to include the modulus of the difference $|e_{EE}^2(i) - e_{OE}^2(i)|$ in the lieu of the difference. It is not clear why the possibility of switching back to the EE mode was abandoned, but it could be related to the lack of an elaborated saturation procedure similar to that used for convex combinations of AFs.

Indeed, the normalized supervisor [88] employs a sigmoidal function along with a power normalization to adapt $\lambda(i)$ and deal with the random nature of the instantaneous estimation errors, preventing the combination from behaving erratically (see Eq. 3.9 in Section 3.3). Now, while the problem was formulated differently in the CSE, the same erratic behaviour tends to show up as well. In order to tackle that, the modulus in Eq. A.12 smooths the switching from the fast mode (EE) to the slow one (OE) but, at the same time, it prevents the composition from switching back as it ignores the actual direction of the gradient. Hence, CSE has no tracking capabilities other than those of the slow mode.

Also, depending on how μ is chosen, Eq. A.13 is subject to unstable updates as $\lambda(i) \to 0$ because so is LMSOE. This means that $e_{OE}(i)$ could grow with no control and destabilize the composite structure unless specific measures are taken to control the poles of the AF.

A.3 Some Remarks on Composite Structures

Besides the CRA and CSE, the rationale of merging the complementary properties of different IIR filtering formulations into a single adaptive structure has been exploited by other works as well, like the Master-Slave [31], Composite Error (CE) [18, 32], the Switched Regressor [35] and the EEOE/MEEOE [36] algorithms, among some others.

Whereas the aforementioned algorithms are distinct and present different convergence behaviors, they actually exhibit common design principles. In all cases, the overall adaptive structure will rely on a stable formulation able to reach the vicinity of the global minimum rapidly but not necessarily the minimum itself, a slower formulation capable of reaching the global minimum and a weighting factor that trades off the convergence speed and accuracy of both.

The LMSOE is prevalent as the accurate formulation albeit some works use PLR and, for this reason, could stagnate instead of converging if the plant is not SPR (e.g., the CRA discussed in Section A.1). For the rapid formulation, the EE is often chosen because it is fast, stable and consistently monomodal even in undermodelled scenarios, but it is not an universal choice. For example, the Switched Regressor proposed by Burt [35] uses the Steiglitz-Mcbride method (SMM) instead. As the SMM itself can be considered a way of combining the characteristics of the EE and OE approaches [1, 12], the arrangement could be debatable in cases of correct modelling and a white additive noise $v(i)$, but it features a superior performance whenever the additive noise is coloured and the AF undermodels the unknown plant.

As for $\lambda(i)$, the comparison between the CRA and CSE algorithms unveils that a time varying strategy tends to perform better, at least in steady state. Indeed, any arbitrary decaying sequence should outperform a fixed mixing factor,

although an adaptive criterion that takes into account the instantaneous errors is expected to perform better. In this sense, the time-varying weighting that equips the CSE suffers from one major weakness: in spite of being adaptive, it invariably decays along time regardless the relative values of the errors.

B CONSIDERATIONS ON MISMODELLING

In a system identification scenario, consider an unknown plant and a given AF whose transfer functions are given respectively by

$$\mathbb{H}^o(z) = \frac{B^o(z)}{1 + A^o(z)} \qquad \text{with } B^o(z) = \sum_{k=0}^{M^o} b_k^o z^k \text{ and } A^o(z) = \sum_{k=1}^{M^o} a_k^o z^k \tag{B.1}$$

$$\mathbb{H}(z) = \frac{B(z)}{1 + A(z)} \qquad \text{with } B(z) = \sum_{k=0}^{M} b_k z^k \text{ and } A(z) = \sum_{k=1}^{M} a_k z^k \tag{B.2}$$

If the unknown plant $\mathbb{H}^o(z)$ and the AF $\mathbb{H}(z)$ match each other orders ($M = M^o$, *exact modelling*), then identification is achieved when the learning algorithm successfully adjusts the parameters of $\mathbb{H}(z)$ towards those of $\mathbb{H}^o(z)$. Nevertheless, if the orders are not the same (*mismodelling*), coefficient matching looses any physical relevance and additional issues arise.

If the transfer function $\mathbb{H}^o(z)$ has finite zeros, it can be properly approximated by a model without any zeros or, conversely, if $\mathbb{H}^o(z)$ has finite poles it can be closely approximated by a model without any poles [56, 131]. Often, there is no objective criteria available and the order has to be chosen based on considerations of goodness of fit and mathematical tractability [132].

As a rule, exact modelling yields the sharpest results as the AF complies exactly with the structure of the unknown plant in Fig. 1, but this is not always possible or desirable [54]. In the cases where the unknown system order is finite [1] , intuition dictates that making $M > M^o$ (the so called over-parametrization

[1] The order of some physical systems can not be estimated accurately, but experience suggests

or *overmodelling*) sounds like an option because the adaptation could produce pole-zero cancellations or nullify the higher order coefficients [53]. This insight leads naturally to the definition of *sufficient modelling*, in which an ideal $\mathbb{H}(z)$ has to have *at least* as many zeros and poles as $\mathbb{H}^o(z)$; i.e., $M \geq M^o$ [26] [134].

Besides trying to provide identification, overmodelling seeks to avoid multimodal error surfaces, on which gradient-based algorithms may slow down or get stuck into sub-optimal local minima. However, in practice it could deteriorate the performance due to the finite precision effects and the intrinsic variation of the AF internal signals. Even worse, it has been shown that arbitrary over-parametrization could in reality cause multimodality instead of avoiding it (e.g., [50, Eqs. 11 and 12] and [51]). Clearly, to rely on overmodelling and assume that the learning algorithm will take care of the "extra" parameters is misleading.

The intuitive idea that sufficient modelling enforces unimodal error surfaces was formally proposed by Stearns [26] and it is often referred to as the *Stearns's conjecture*. Indeed, the conjecture holds for broadband input signals and short filters (1st and 2nd orders) but it may fail for higher order filters as shown by Soderstrom and Stoica [50] soon afterwards. Furthermore, later Fan and Nayeri [51] proved that a high order filter (3rd order and above) raises unimodality whenever it is excited by a white process and the degree of $B(z)$ plus 1 equals or exceeds that of $A^o(z)$, establishing a theoretical constraint under which the sufficient modelling is expected to work as originally hypothesized by Stearns.

In order to illustrate the risks of the arbitrary overmodelling as a way to pursue unimodality, Fan and Nayeri cite a practical example in which the error surface of a 2nd-order AF is unimodal whereas that of a 3rd order AF turns out to be multimodal [51, Section III.C *Overparametrization*]. Of course, this case alone suffices to establish that the AF fitting does not improve by simply

that they can rarely be modelled exactly by finite order rational functions [12]. Note that order estimation is a fruitful field of study by itself (e.g., [133] or [132]).

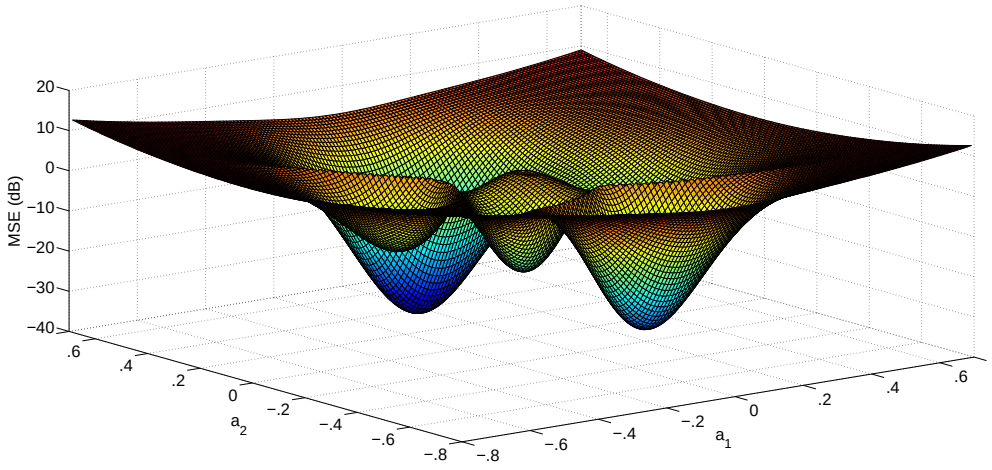

Figure 48: Multimodal Error Surface

increasing M in the general case. However, experience shows that overmodelling could indeed lead to unimodality in a large class of real-world applications, which was incidentally noted by Stearns and motivated his aforementioned conjecture.

Conversely, albeit it could not be a major limiting factor for FIR AFs [2], undermodelling may seriously compromise IIR AFs. If the input signal does not happen to be an ARMA process of a certain order or white in the sufficient modelling case, then there is no way of assuring that all stationary points of the surface error are global minima [4, 50]. This means that local minima and saddle points become a serious possibility and the error surface is potentially multimodal, like those shown in Figures 48 and 49 in which a hypothetical 2nd order IIR AF tries to identify a system whose order is presumably higher.

B.1 Algebraic Perspective on Undermodelling

The stationary points of the error surfaces can be characterized algebraically in order to describe the form of the error function with the interpolation conditions provided by the Beurling-Lax Theorem as described by Regalia [12]. In

[2] In the FIR AFs case, normally the AF steady state improves as the order increases until some point from which the fit starts to worsen slowly. In a nutshell, the performance suffers slightly as the order changes its optimal value in either direction [56].

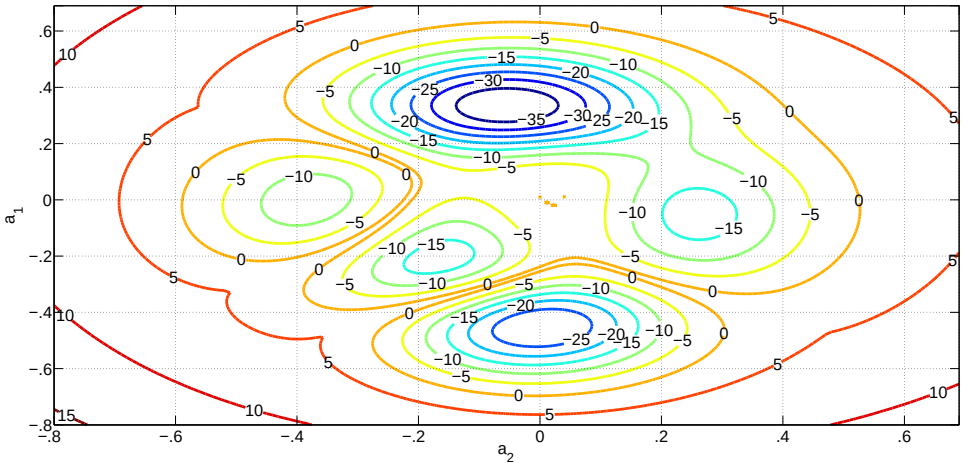

Figure 49: Countour Plot of a Multimodal Error Surface

FIR adaptive filtering, a similar goal is pursued through the energy conservation and variance relation arguments developed by Sayed [5], but his studies refers to another context and do not comprise the peculiarities of rational systems.

Although not so straightforward as the energy conservation and variance relation arguments, classical approaches are far more comprehensive, encompassing naturally IIR adaptive filtering and undermodelling aspects. By establishing connections among these seemingly distinct methods, Regalia [12] managed to show that essentially all of them can be described as attempts to fit a vector into the null space of the Hankel form of $\mathbb{H}^o(z)$. Here, this perspective is exploited to illustrate why undermodelling issues arise in the Output Error context (see Section 2.4), but the same rationale can be extended *mutatis mutandis* to the Padè Approximants, Equation Error, Steiglitz-Mcbride and hyperstable methods.

Let the estimation error and the squared error, respectively, be defined in the transfer function space as

$$\tilde{\mathbb{H}}(z) = \mathbb{H}^o(z) - \mathbb{H}(z) \tag{B.3}$$

$$\|\tilde{\mathbb{H}}(z)\|_2^2 = \langle \tilde{\mathbb{H}}(z), \tilde{\mathbb{H}}(z) \rangle \tag{B.4}$$

So, the stationary points of the error surface of Eq. B.4 will satisfy the condition

$$\frac{\partial}{\partial w_k}\langle\tilde{\mathbb{H}}(z),\tilde{\mathbb{H}}(z)\rangle = 0 \quad k = 0,\ 1,\ \cdots\ 2M+1, \tag{B.5}$$

where $w = [a_1\ a_2\ \cdots\ a_M\ b_0\ b_1\ \cdots\ b_M]$ collects the terms of the polynomials $A(z)$ and $B(z)$. In the case of the feedback parameters $\{a_k\}_{k=1}^{M}$, Eq. B.5 becomes

$$\frac{\partial}{\partial a_k}\langle\tilde{\mathbb{H}}(z),\tilde{\mathbb{H}}(z)\rangle = 0 \quad k = 1,\ \cdots\ M \tag{B.6}$$

However, $\tilde{\mathbb{H}}(z)$ is not a linear function of $A(z)$ [3] and, therefore, stationary points in Eq. B.6 are not necessarily global minima, but may be local minima, saddles or maxima as well. Although this matter is discussed in Section 2.8.2 with greater details, it considers correctly modeled scenarios and no clue is given about the AF behaviour outside this context. Here, it is provided a glimpse of why undermodelling may be awkward within a strictly algebraic perspective.

Assume that the set $\{h_k^o\}_{k=0}^{\infty}$ are the components of the impulse response of $\mathbb{H}^o(z)$ such that

$$\mathbb{H}^o(z) = \sum_{k=0}^{\infty} h_k^o z^k \tag{B.7}$$

and define the doubly infinite Hankel matrix $\Gamma_{\mathbb{H}^o}$ as

$$\Gamma_{\mathbb{H}^o} = \begin{bmatrix} h_1^o & h_2^o & h_3^o & h_4^o & \dots \\ h_2^o & h_3^o & h_4^o & h_5^o & \dots \\ h_3^o & h_4^o & h_5^o & h_6^o & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{B.8}$$

which deliberately excludes the term h_0^o. If the rank of $\Gamma_{\mathbb{H}^o}$ is finite, then it can be shown that its null space $\mathcal{N}(\Gamma_{\mathbb{H}^o})$ allows to determine the poles of $\mathbb{H}^o(z)$ [12,

[3] $A(z)$ is the polynomial that defines the feedback loop whose effect is to relate the current output $y(i)$ to the prior outputs $y(i-k)\ \forall k = 1, 2, \dots, M$. Therefore, although $y(i) = \mathbb{H}(z)U(z)$ is a *structurally linear* operation (because it is described by a difference equation; i.e., linear operations), the recursivity caused by the feedback loop makes $\mathbb{H}(z)$ a non-linear operator.

Theorem 3.5] as follows.

Let $V(z)$ be the *all-pass complement* of the polynomial $A(z)$ defined as

$$V(z) = \sum_{k=0}^{\infty} v_k z^k = \frac{z^M\big(1 + A(z^{-1})\big)}{1 + A(z)} \tag{B.9}$$

Then, the optimal values for $A(z)$ meet the condition

$$\min_{\{a_k\}} \|\Gamma_{\mathbb{H}^o} v\| \quad k = 1, 2, \cdots, M, \tag{B.10}$$

where v is the infinite vector that stores the elements of $V(z)$ as

$$v = [v_0 \; v_2 \; v_3 \; v_4 \; v_5 \; \cdots]^T \rightarrow \|v\| = 1 \tag{B.11}$$

In case of exact modelling, coefficient matching is attainable and $A(z)$ can be adjusted towards the ideal $A^o(z)$ so that $v \in \mathcal{N}(\Gamma_{\mathbb{H}^o})$. In the undermodelled case though, the best solution is to fit v into an *approximate* nullspace of $\Gamma_{\mathbb{H}^o}$ by making $\|\Gamma_{\mathbb{H}^o} v\|$ as small as possible and then some further concerns have to be dealt with: such an approximation often makes the problem numerically ill-posed and distorts the search space (i.e., makes the error surface multimodal), with no way of classifying the stationary points into local/global minima and saddles.

Once $A(z)$ is determined from Eq. B.10, the Beurling-Lax Theorem states that $B(z)$ can be find through the interpolation condition (see Appendix D)

$$\tilde{\mathbb{H}}(z) = V(z)g(z), \; g(z) \in \mathcal{H}^2 \tag{B.12}$$

which, by its turn, is equivalent to solve the differential equation

$$\frac{\partial}{\partial b_k} \langle \tilde{\mathbb{H}}(z), \tilde{\mathbb{H}}(z) \rangle = 0 \quad k = 0, \; 1, \; \cdots \; M, \tag{B.13}$$

where $\mathcal{H}^2$ denotes the Hardy Space and $\{b_k\}_{k=0}^M$ are the components of $B(z) = \sum_{k=0}^{M} b_k z^k$. This is a conceptually easier problem once the estimation error of Eq. B.3 is a linear function of $B(z)$, which makes the squared error of Eq. B.4 a

quadratic function. Consequently, this function will admit a single minimum even in the case of undermodelling.

From the above, it is clear that Eq. B.10 may be restated as

$$\min_{\deg \mathbb{H}(z)=M} \|\tilde{\mathbb{H}}(z)\|_2, \tag{B.14}$$

and consequently there will be a bijection among the local minima of Eq. B.14 and those of Eq. B.10, on which the following condition holds [12, Theorem 3.14]

$$\tilde{\mathbb{H}}(z) = zV^2(z)Q(z),\ Q(z) \in \mathcal{H}^2, \tag{B.15}$$

which is known as the Walsh's Theorem. Whenever $\deg\big(\mathbb{H}^o(z)\big) = \deg\big(\mathbb{H}(z)\big)$, $\mathbb{H}(z) = \mathbb{H}^o(z)$ is the only solution but if $\deg\big(\mathbb{H}^o(z)\big) > \deg\big(\mathbb{H}(z)\big)$, multiple solutions are admissible. Unfortunately, whatever the case, the theorem will hold for *any* stationary point, including local minima and saddles.

C ERROR SURFACE TOPOLOGY AND PERFORMANCE ISSUES

As aforementioned, multimodality is a severe performance limiter as a global minimum is known to yield a good solution whilst local minima could give poor approximations [4, 18, 26, 122], and there are no practical means to distinguish them in real time [12].

Local maxima and saddles may also adversely affect the performance of the filter parameters adaptation because gradient based algorithms search for stationary points on the error surfaces. Whereas the former are intrinsically unstable and naturally circumvented by the stochastic noise, the latest render gradient search techniques less efficient and slow down the adaptation [17, 23], as it is the case of the surfaces shown in Figures 48 and 49.

At this point, it is important to remark that a monomodal error surface is not a synonym of fast convergence. In the Output Error approaches, the estimation error of the filter is not a linear function of its feedback coefficients, which may distort the quadratic error surface in a way that, beyond the vicinity of the minimum, the gradient fades and the convergence rate drops significantly [35]. As it was not enough, the performance relies heavily on the pole-zeros configuration of the plant as well as the parameters initialization, tending to deteriorate in the presence of underdamped or clustered poles [10, 11].

In particular, underdamped poles make the performance surface highly non-quadratic with the minimum enclosed in narrow valleys preceded by nearly flat

plateaus [26], establishing a paradox: smaller step sizes would lead to huge convergence times due the almost-null gradients while bigger steps could result in unacceptable steady state errors due to such a peculiar topology. As Burt et. al [82] remark, if the AF overall convergence is slow enough, the adaptation of $B(z)$ behaves locally as driven by a prefiltered input such that

$$y(i) = B(z)u'(i), \tag{C.1}$$

where $u'(i)$ is straightforwardly seen as being

$$u'(i) \approx \frac{1}{1 - A(z)}u(i), \tag{C.2}$$

which is associated to a correlation matrix

$$R_{u'} = \mathrm{E}\left[u'(i)\ u'(i-1)\ \cdots\ u'(i-M)\right]^T\left[u'(i)\ u'(i-1)\ \cdots\ u'(i-M)\right], \tag{C.3}$$

so that with a white $u(i)$ and poles fairly distant from the unit border, $R_{u'}$ becomes well-conditioned with a low eigenvalue spread, making the zeros converge rapidly. Otherwise, the eigen-spread grows and $B(z)$ adapts slower [1] [11, 13].

An algebraic perspective of this phenomenon arises by realizing that, under the assumption that $A(z)$ is either fixed or slow-changing, the AF transfer function $H(z)$ could be approximated through the decomposition

$$H'(z) = \sum_{k=0}^{M} b_k \frac{z^k}{1 - A(z)}, \tag{C.4}$$

where clearly $H'(z)$ is spanned from the basis $\left\{\frac{z^k}{1-A(z)}\right\}_{k=0}^{M}$ [2]. If, by chance, these basis functions happen not to be orthogonal, the estimation could become

[1] Such a conditioning is related to the polynomial $A^o(z)$ once the AF pursues the plant parameters. So, if the Hankel matrix of the plant (given by its impulse response) has a large singular value spread, the AF convergence will necessarily slow down in spite of its design or modelling issues (unless a different filtering structure is used to realize $H^o(z)$). Note, however, that the convergence rate is also related to factors like parameters initialization or local minima.

[2] In the literature, this is known as the *Fixed Denominator Model* [16] and can be related to the concept of reduced error surface. See Appendix D for further considerations on this matter.

numerically ill-conditioned and decrease the convergence rate significantly [16].

None of the foregoing issues affect the FIR AFs: their performance surfaces are quadratic and they are known to always converge to the global minimum. Also, as FIR AFs are not recursive and have no poles to deal with, they exhibit more relaxed stability bounds and tend to be much faster than IIR AFs. Algebraically, this means that $1 - A(z) = 1$ so that the decomposition in Eq. C.4 will be orthogonal and no further numerical difficulties arise but those intrinsic to $u(i)$.

D RATIONAL APPROXIMATION IN THE HARDY SPACE $\mathcal{H}^2$

The definition of *controllability vector*, denoted by $\mathcal{C}(z)$, is pervasive throughout classical system identification and rational approximation. Along with the Beurling-Lax Theorem, $\mathcal{C}(z)$ plays a role in the definition of a basis for the Hardy space $\mathcal{H}^2$ that conveniently spans the mean squared error

$$\mathrm{E}\,||\mathbb{H}^o(z) - \mathbb{H}(z)||_2^2 = \mathrm{E}\left|\left|\tilde{\mathbb{H}}(z)\right|\right|_2^2, \tag{D.1}$$

where $\mathbb{H}(z)$ is a given AF that is trying to identify the unknown system $\mathbb{H}^o(z)$. Additionally, the Beurling-Lax Theorem shows that the condition

$$\langle \mathcal{C}(z), f(z)\rangle = 0, \tag{D.2}$$

with $\mathcal{C}(z), f(z) \in \mathcal{H}^2$, defines a right shift invariant subspace in $\mathcal{H}^2$, which is useful to unveil the stationary points of the mean squared error [1]

$$\mathrm{E}\,e^2(i) = \mathrm{E}\left(\big(\mathbb{H}^o(z) - \mathbb{H}(z)\big)U(z)\right)^2 = \mathrm{E}\left(\tilde{\mathbb{H}}(z)U(z)\right)^2, \tag{D.3}$$

where $U(z)$ is the Z Transform of $u(i)$. Eq. D.3 leads to the Walsh's Theorem as briefly mentioned in Appendix B.1. Some highlights from this theoretical framework are adapted from [12, 135] as follows.

[1] While Eq.D.1 measures the mismatch between $\mathbb{H}^o(z)$ and $\mathbb{H}(z)$ directly in the transfer function space, Eq. D.3 provides the mean squared *estimation* error. As shown in Section 2.4.1, both relate to each other as the latter equals the former weighted by the spectral density $S_u(z)$ of the input $u(i)$. In case of an unitary white process $u(i)$, $S_u(z) = 1$ and Equations D.1 and D.3 become identical.

Let a M-th order adaptive filter $\mathbb{H}(z)$ whose state space representation corresponds to the Eq. 2.44. Then,

$$\mathcal{W}_{i+1} = \mathsf{A}\mathcal{W}_i + \mathsf{b}u(i) \tag{D.4}$$

$$\mathcal{W}_{i+1} = \mathsf{A}\mathcal{W}_i + \mathsf{b}\mathsf{U}(z) \tag{D.5}$$

$$= z\mathsf{A}\mathcal{W}_{i+1} + \mathsf{b}\mathsf{U}(z) \tag{D.6}$$

$$\mathcal{W}_{i+1}(\mathbb{I} - z\mathsf{A}) = \mathsf{b}\mathsf{U}(z) \tag{D.7}$$

$$\mathcal{W}_{i+1} = (\mathbb{I} - z\mathsf{A})^{-1}\mathsf{b}\mathsf{U}(z) \tag{D.8}$$

By understanding that *state controllability* is the ability of a given system to change from a given state to another according to the input $u(i)$, define the *controllability vector* $\mathcal{C}(z)$ as

$$\mathcal{C}(z) = (\mathbb{I} - z\mathsf{A})^{-1}\mathsf{b} \tag{D.9}$$

So, clearly $u(i)$ drives the system among different states through $\mathcal{C}(z)$; i.e.,

$$\mathcal{W}_{i+1} = \mathcal{C}(z)\mathsf{U}(z), \tag{D.10}$$

where can be noted that $\mathcal{C}(z)$ is a vector-valued quantity such as

$$\mathcal{C}(z) = [\mathcal{C}_1(z)\ \mathcal{C}_2(z)\ \mathcal{C}_3(z)\ \cdots\ \mathcal{C}_{M-1}(z)\ \mathcal{C}_M(z)], \tag{D.11}$$

whose components have the form

$$\mathcal{C}_{k+1}(z) = \frac{P_k}{1 + A(z)} \quad k = 0, 1, 2, \cdots, M-1, \tag{D.12}$$

with $\deg(P_k) < M$. If $\mathbb{H}(z)$ is implemented in the direct form, A will be as in Eq.

2.46 and

$$\mathcal{C}_{k+1}(z) = \frac{P_k}{1+A(z)} = \frac{z^k}{1+A(z)} \rightarrow \mathcal{C}(z) = \begin{bmatrix} \frac{1}{1+A(z)} \\ \frac{z}{1+A(z)} \\ \frac{z^2}{1+A(z)} \\ \vdots \\ \frac{z^{M-1}}{1+A(z)} \end{bmatrix} \tag{D.13}$$

By noting that both $\mathbb{H}(z)$ and the components of $\mathcal{C}(z)$ have the same poles, then $\mathbb{H}(z)$ can be opportunely decomposed as

$$\mathbb{H}(z) = \sum_{k=0}^{M} \nu_k \frac{P_k}{1+A(z)} \tag{D.14}$$

$$\nu_k = \langle \mathbb{H}(z), \frac{P_k}{1+A(z)} \rangle, \tag{D.15}$$

which works to our advantage when characterizing the stationary points of the mean squared error function of Eq. D.1 as show in the sequel.

Let $V(z)$ be the *all-pass complement* of $A(z)$ defined as

$$V(z) = \frac{z^M \big(1 + A(z^{-1})\big)}{1+A(z)} \tag{D.16}$$

According to the Beurling-Lax Theorem, in the Hardy space $\mathcal{H}^2$ the subspace $\mathcal{S}^\perp$ lying orthogonal to the subspace $\mathcal{S}$ spanned by $\mathcal{C}(z)$ is causally divided by $V(z)$. So, $\mathcal{S}^\perp$ is right shift-invariant and the following identities hold

$$\langle \mathcal{C}(z), f(z) \rangle = 0 \longleftrightarrow f(z) = V(z)g(z),\ g(z) \in \mathcal{H}^2 \tag{D.17}$$

$$\langle \mathcal{C}(z), f(z) \rangle = 0 \longleftrightarrow \langle \mathcal{C}(z), zf(z) \rangle = 0 \tag{D.18}$$

These remarkable results lead naturally to the conclusion that $\mathcal{H}^2$ itself can be spanned as

$$\mathcal{H}^2 = \mathcal{R}\Big(\{z^k V(z)\}_{k=0}^{\infty} \Big) \tag{D.19}$$

$$\mathcal{H}^2 = \mathcal{R}\left(\left\{\frac{P_k(z)}{1+A(z)}\right\}_{k=0}^{M} \cup \left\{z^k V(z)\right\}_{k=1}^{\infty}\right) \tag{D.20}$$

By inspection, the similarity of the term $\left\{\frac{P_k(z)}{1+A(z)}\right\}_{k=0}^{M}$ in Eq. D.20 with the controllability vector $\mathcal{C}(z)$ of Eq. D.13 is readily verified, suggesting that an "augmented" controllability vector $\mathcal{C}_a(z)$ could be defined as

$$\mathcal{C}_a(z) \triangleq \begin{bmatrix} \frac{1}{1+A(z)} \\ \frac{z}{1+A(z)} \\ \frac{z^2}{1+A(z)} \\ \vdots \\ \frac{z^{M-1}}{1+A(z)} \\ \frac{z^M}{1+A(z)} \end{bmatrix}, \tag{D.21}$$

and we note from Eq. D.14 that $\mathcal{C}_a(z)$ spans the subspace on which $\mathbb{H}(z)$ and, hence, $V(z)$ lie onto. In case the poles of $\mathbb{H}(z)$ match those of $\mathbb{H}^o(z)$, it can be shown that $V(z) \in \mathcal{N}\big(\Gamma_{H^o}\big)$ [12, Theorem 3.5], where Γ_{H^o} is the Hankel form of $\mathbb{H}^o(z)$ (refer to Section B.1), so that $\mathcal{C}_a(z)$ is expected to span $\mathcal{N}(\Gamma_{H^o})$ as well.

With the basis referred in Eq. D.20, $\mathbb{H}^o(z)$ (whose poles allegedly do not match those of $\mathbb{H}(z)$), may be appropriately decomposed as

$$\mathbb{H}^o(z) = \sum_{k=0}^{M} \nu_k^o \frac{P_k}{1+A(z)} + \sum_{k=1}^{\infty} \rho_k z^k V(z) \tag{D.22}$$

$$\nu_k^o = \langle \mathbb{H}^o(z), \frac{P_k}{1+A(z)} \rangle \tag{D.23}$$

$$\rho_k = \langle \mathbb{H}^o(z), z^k V(z) \rangle, \tag{D.24}$$

which allow us to express the mean squared error of Eq. D.1 as

$$\mathrm{E}\,\|\tilde{\mathbb{H}}(z)\|_2^2 = \mathrm{E}\sum_{k=0}^{M}(\nu_k^o - \nu_k)^2 + \sum_{k=1}^{\infty}\rho_k^2 \tag{D.25}$$

Clearly, in case $\mathbb{H}^o(z)$ lies on the space spanned by $\mathcal{C}_a(z)$, then identification is feasible. Now, once only $[\nu_k]_{k=1}^M$ depend upon $B(z)$ and $A(z)$ while both $[\nu_k^o]_{k=1}^M$

and $[\rho_k]_{k=1}^{M}$ depend upon $A(z)$ only, it is straightforward to check that Eq. D.25 boils down to

$$\mathrm{E}\,\|\tilde{\mathbb{H}}(z)\|_2^2 = \mathrm{E}\sum_{k=1}^{\infty}\rho_k^2 \tag{D.26}$$

whenever the zeros of $\mathbb{H}(z)$ are optimized with respect to a fixed, non-optimized poles set. Herewith, $\tilde{\mathbb{H}}(z)$ will be said to be on the *reduced error surface*, which means that $\mathbb{H}(z)$ will fit the projection of $\mathbb{H}^o(z)$ onto the space spanned by $\mathcal{C}_a(z)$. So, Eq. D.26 becomes a particular interpretation of the orthogonality principle since the (optimal) error $\tilde{\mathbb{H}}(z)$ is perpendicular to the estimator $\mathbb{H}(z)$. Indeed, by applying the Beurling-Lax Theorem (Equations D.17 and D.18) to Eq. D.26, the following condition arises

$$\langle \mathcal{C}_a(z), \tilde{\mathbb{H}}(z)\rangle = 0 \tag{D.27}$$

which, not by accident, also arises when setting the derivative of the squared error $\|\tilde{\mathbb{H}}(z)\|_2^2$ in relation to $B(z)$ to 0; i.e.,

$$\frac{\partial}{\partial b_k}\langle \tilde{\mathbb{H}}(z), \tilde{\mathbb{H}}(z)\rangle = \langle \mathcal{C}_a(z), \tilde{\mathbb{H}}(z)\rangle = 0 \quad k = 0,\ 1,\ \cdots\ M \tag{D.28}$$

As a result of the condition of Eq. D.28, the Beurling-Lax Theorem leads to

$$\langle \mathcal{C}_a(z), \tilde{\mathbb{H}}(z)\rangle = 0 \rightarrow \tilde{\mathbb{H}}(z) = V(z)g(z),\ g(z) \in \mathcal{H}^2, \tag{D.29}$$

where $g(z)$ can be determined as

$$\begin{aligned}\tilde{\mathbb{H}}(z) &= \sum_{k=1}^{\infty}\rho_k z^k V(z) = V(z)\sum_{k=1}^{\infty}\rho_k z^k \\ \therefore\ g(z) &= \sum_{k=1}^{\infty}\rho_k z^k\end{aligned} \tag{D.30}$$

Alternatively, a straightforward algebraic manipulation on Eq. D.29 gives

$$\tilde{\mathbb{H}}(z) = V(z)g(z) \rightarrow g(z) = [\mathbb{H}^o(z)V(z^{-1})]_+ - [\mathbb{H}(z)^{-1}V(z^{-1})]_+ \tag{D.31}$$

$$g(z) = [\mathbb{H}^o(z)V(z^{-1})]_+, \tag{D.32}$$

where $[.]_+$ is the causal projection operator and the term $[\mathbb{H}(z)^{-1}V(z^{-1})]_+$ can be shown to be anti-causal (thus, vanishes in $\mathcal{H}^2$).

In a nutshell, whenever the error is on the reduced error surface, the interpolation condition $\tilde{\mathbb{H}}(z) = V(z)g(z)$ (Eq. D.29) will hold. It can be shown that the minima of the reduced error surface and the "complete" error surface $\|\tilde{\mathbb{H}}(z)\|_2^2$ (in which $A(z)$ is allowed to vary w.r.t. the optimized $B(z)$) are the same whereas the saddles and maxima of the reduced surface correspond to saddles onto the complete surface [12, Property 5.5]. As mentioned in Appendix B.1, there is no way of classifying *a priori* those stationary points into saddles and minima once the Walsh's Theorem will hold at all of them indistinguishably, making $\tilde{\mathbb{H}}(z)$ become

$$\tilde{\mathbb{H}}(z) = zV^2(z)Q(z),\ Q(z) \in \mathcal{H}^2 \tag{D.33}$$

REFERENCES

[1] Vijay K. Madisetti and Douglas B. Williams, Eds., *Digital Signal Processing Handbook*, Electrical Engineering Handbook. CRC Press, December 1999.

[2] S. Elliott, *Signal Processing for Active Control*, Signal Processing and its Applications. Academic Press, 2000.

[3] Michael G. Larimore, C. Richard Johnson, and John R. Treichler, *Theory and Design of Adaptive Filters*, Prentice Hall, 2001.

[4] Paulo S. R Diniz, *Adaptive Filtering: Algorithms and Practical Implementation*, Springer, 2008.

[5] Ali H. Sayed, *Adaptive Filters*, Wiley-IEEE Press, 2008.

[6] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan 2009.

[7] S.S. Haykin and B. Van Veen, *Signals and systems*, Wiley, 1999.

[8] John G. Proakis and Dimitris K Manolakis, *Digital Signal Processing*, Prentice Hall, 2006.

[9] Richard G. Lyons, *Understanding Digital Signal Processing*, Prentice Hall 3 edition, November 2010.

[10] P.M.S. Burt and M. Gerken, "A polyphase iir adaptive filter: error surface analysis and application," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, apr 1997, vol. 3, pp. 2285 –2288 vol.3.

[11] H. Fan, "A structural view of asymptotic convergence speed of adaptive iir filtering algorithms. i. infinite precision implementation," *Signal Processing, IEEE Transactions on*, vol. 41, no. 4, pp. 1493 –1517, apr 1993.

[12] Phillip A. Regalia, *Adaptive IIR Filtering in Signal Processing and Control*, CRC Press, 1994.

[13] J. Cousseau, P. S. R. Diniz, G. Sentoni, and O. Agamennoni, "On orthogonal realizations for adaptive iir filters," *International Journal of Circuit Theory and Applications*, vol. 28, no. 5, pp. 481–500, 2000.

[14] Yegui Xiao, Y. Takeshita, and K. Shida, "Steady-state analysis of a plain gradient algorithm for a second-order adaptive iir notch filter with constrained poles and zeros," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 48, no. 7, pp. 733–740, Jul 2001.

[15] P.M.S. Burt and P.A. Regalia, "A new framework for convergence analysis and algorithm development of adaptive iir filters," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, May 2004, vol. 2, pp. 441–444.

[16] Peter S.C. Heuberger, Paul M.J. van den van den Hof, and Bo Wahlberg, *Modelling and Identification with Rational Orthogonal Basis Functions*, Springer-Verlag London, June 30 2005.

[17] M. Nayeri and W. Jenkins, "Alternate realizations to adaptive iir filters and properties of their performance surfaces," *Circuits and Systems, IEEE Transactions on*, vol. 36, no. 4, pp. 485–496, 1989.

[18] S.L. Netto, P.S.R. Diniz, and P. Agathoklis, "Adaptive iir filtering algorithms for system identification: a general framework," *Education, IEEE Transactions on*, vol. 38, no. 1, pp. 54 –66, feb 1995.

[19] Juan E. Cousseau, "Adaptive iir filtering: Available results," *IEEE Circuits and Systems Newsletter*, vol. 10, pp. 2–10, 1999.

[20] J.J. Shynk, "Adaptive iir filtering," *ASSP Magazine, IEEE*, vol. 6, no. 2, pp. 4 –21, april 1989.

[21] Jr. Gray, A. and J. Markel, "A normalized digital filter structure," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 3, pp. 268 – 277, jun 1975.

[22] J.A. Rodriguez-Fonollosa and E. Masgrau, "Simplified gradient calculation in adaptive iir lattice filters," *Signal Processing, IEEE Transactions on*, vol. 39, no. 7, pp. 1702 –1705, jul 1991.

[23] P.A. Regalia, "Stable and efficient lattice algorithms for adaptive iir filtering," *Signal Processing, IEEE Transactions on*, vol. 40, no. 2, pp. 375 –388, feb 1992.

[24] I. Landau, "Unbiased recursive identification using model reference adaptive techniques," *Automatic Control, IEEE Transactions on*, vol. 21, no. 2, pp. 194 – 202, apr 1976.

[25] Jr. Johnson, C., "Adaptive iir filtering: Current results and open issues," *Information Theory, IEEE Transactions on*, vol. 30, no. 2, pp. 237 – 250, mar 1984.

[26] S. Stearns, "Error surfaces of recursive adaptive filters," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 3, pp. 763 – 766, jun 1981.

[27] Carl D. Meyer, *Matrix Analysis and Applied Linear Algebra Book*, SIAM - Society for Industrial and Applied Mathematics, February 2001.

[28] H.F. Ferro, L.F.O. Chamon, and C.G. Lopes, "Fir-iir adaptive filters hybrid combination," *Electronics Letters*, vol. 50, no. 7, pp. 501–503, March 2014.

[29] L.F.O. Chamon, W.B. Lopes, and C.G. Lopes, "Combination of adaptive filters with coefficients feedback," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, march 2012, pp. 3785 –3788.

[30] B. Beliczynski, I. Kale, and G.D. Cain, "Approximation of fir by iir digital filters: an algorithm based on balanced model reduction," *Signal Processing, IEEE Transactions on*, vol. 40, no. 3, pp. 532 –542, mar 1992.

[31] M. C. Hall and P.M. Hughes, "The master-slave iir filter adaptation algorithm," in *Circuits and Systems, 1988., IEEE International Symposium on*, June 1988, pp. 2145–2148 vol.3.

[32] S. Lima Netto and P.S.R. Diniz, "Composite algorithms for adaptive iir filtering," *Electronics Letters*, vol. 28, no. 9, pp. 886–888, April 1992.

[33] J.B. Kenney and C.E. Rohrs, "The composite regressor algorithm for iir adaptive systems," *Signal Processing, IEEE Transactions on*, vol. 41, no. 2, pp. 617–628, Feb 1993.

[34] S.L. Netto and P. Agathoklis, "A new composite adaptive iir algorithm," in *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, Oct 1994, vol. 2, pp. 1506–1510 vol.2.

[35] P.M.S. Burt, "A switched regressor algorithm for adaptive iir filtering," in *Circuits and Systems, 1995., Proceedings., Proceedings of the 38th Midwest Symposium on*, aug 1995, vol. 2, pp. 990 –993 vol.2.

[36] N. Avessta and T. Aboulnasr, "Combined regressor methods and adaptive iir filtering," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 51, no. 11, pp. 2222–2234, Nov 2004.

[37] Lennart Ljung, "Perspectives on system identification," in *In Plenary talk at the proceedings of the 17th IFAC World Congress, Seoul, South Korea*, 2008.

[38] Atanu Biswas and Peter X.-K. Song, "Discrete-valued arma processes," *Statistics & Probability Letters*, vol. 79, no. 17, pp. 1884 – 1889, 2009.

[39] V. Madisetti, *Digital Signal Processing Fundamentals*, The Digital Signal Processing Handbook, Second Edition. CRC Press, 2010.

[40] A. Carini, V.J. Mathews, and G.L. Sicuranza, "Sufficient stability bounds for slowly varying direct-form recursive linear filters and their applications in adaptive iir filters," *Signal Processing, IEEE Transactions on*, vol. 47, no. 9, pp. 2561 –2567, sep 1999.

[41] L.J. Eriksson, M.C. Allie, and R. Greiner, "The selection and application of an iir adaptive filter for use in active sound attenuation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 4, pp. 433–437, Apr 1987.

[42] W. Mills, C.T. Mullis, and R.A. Roberts, "Digital filter realizations without overflow oscillations," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 4, pp. 334–338, 1978.

[43] Gene H. Golub, Ed., *Studies in Numerical Analysis*, vol. 24 of *Studies in Mathematics*, Mathematical Assn of Amer, 1985.

[44] C.T. Mullis and R.A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *Circuits and Systems, IEEE Transactions on*, vol. 23, no. 9, pp. 551–562, 1976.

[45] Lennart Ljung and Torsten Soderstrom, *Theory and Practice of Recursive Identification (Signal Processing, Optimization, and Control)*, The MIT Press, 1983.

[46] L.E. Turner, "Second-order recursive digital filter that is free from all constant-input limit cycles," *Electronics Letters*, vol. 18, no. 17, pp. 743–745, 1982.

[47] HJ Butterweck, JHF Ritzerfeld, and MJ Werter, *Finite Wordlength Effects in Digital Filters: A Review*, Eindhoven University of Technology, 1988.

[48] T. Laakso, B. Zeng, I. Hartimo, and Y. Neuvo, "Elimination of limit cycles in floating-point implementations of direct-form recursive digital filters," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 41, no. 4, pp. 308–313, 1994.

[49] Y.S. Wu, "Fast converging adaptive iir lms filter in direct form using the suboptimal step-size," in *Circuits and Systems, 1991, IEEE International Symposium on*, June 1991, vol. 1, pp. 448–451.

[50] T. Soderstrom and P. Stoica, "Some properties of the output error method," *Automatica*, vol. 18, no. 1, pp. 93 – 99, 1982.

[51] Hong Fan and M. Nayeri, "On error surfaces of sufficient order adaptive iir filters: Proofs and counterexamples to a unimodality conjecture," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 9, pp. 1436–1442, Sep 1989.

[52] Petre Stoica, Torsten Soderstrom, Anders Ahlen, and Gote Solbrand, "On the asymptotic accuracy of pseudo-linear regression algorithms," *International Journal of Control*, vol. 39, no. 1, pp. 115–126, 1984.

[53] Petre Stoica, Torsten Söderstrom, Anders Ahlén, and Góte Solbrand, "On the convergence of pseudo-linear regression algorithms," *International Journal of Control*, vol. 41, no. 6, pp. 1429–1444, 1985.

[54] P.A. Regalia, M. Mboup, and M. Ashari, "Existence of stationary points for pseudo-linear regression identification algorithms," *Automatic Control, IEEE Transactions on*, vol. 44, no. 5, pp. 994–998, May 1999.

[55] A. Bultheel and M. Van Barel, *Linear Algebra, Rational Approximation and Orthogonal Polynomials*, North Holland, 1997.

[56] Anthony Zaknich, *Principles of Adaptive Filters and Self-learning Systems*, Advanced Textbooks in Control and Signal Processing. Springer-Verlag London Limited, 2005.

[57] M. Hazewinkel and J.C. Williams, *Stochastic Systems: The Mathematics of Filtering and Identification and Applications: Proceedings of the NATO Advanced Study Institute held at Les Arcs, Savoie, France, June 22 – July 5, 1980*, Nato Science Series C:. Springer Netherlands, 2012.

[58] M. Tomizuka, "Parallel mras without compensation block," *Automatic Control, IEEE Transactions on*, vol. 27, no. 2, pp. 505 – 506, apr 1982.

[59] P.A. Regalia, "An unbiased equation error identifier and reduced-order approximations," *Signal Processing, IEEE Transactions on*, vol. 42, no. 6, pp. 1397 –1412, jun 1994.

[60] D. DoCampo, A. Figueiras-Vidal, and Perez-Gonzalez, *Intelligent Methods in Signal Processing and Communications*, Birkhäuser, 1996.

[61] Hong Fan and W. Jenkins, "A new adaptive iir filter," *Circuits and Systems, IEEE Transactions on*, vol. 33, no. 10, pp. 939–947, Oct 1986.

[62] H. N. Kim and Woo-Jin Song, "Unbiased equation-error adaptive iir filtering based on monic normalization," *Signal Processing Letters, IEEE*, vol. 6, no. 2, pp. 35–37, Feb 1999.

[63] T. Soderstrom, "Comments on adaptive iir filtering with monic normalization," *Signal Processing, IEEE Transactions on*, vol. 48, no. 3, pp. 892–894, March 2000.

[64] K.J. Astrom and P. Eykhoff, "System identification a survey," *Automatica*, vol. 7, no. 2, pp. 123 – 162, 1971.

[65] B.M. Ninness and F. Gustafsson, "A unifying construction of orthonormal bases for system identification," in *Decision and Control, 1994., Proceedings of the 33rd IEEE Conference on*, Dec 1994, vol. 4, pp. 3388–3393 vol.4.

[66] Jun-Chan Kwon, Young-Seok Choi, and Woo-Jin Song, "Equation-error adaptive iir filtering based on data reuse," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 54, no. 8, pp. 695 –699, aug. 2007.

[67] Woo-Jin Song and Hyun-Chool Shin, "Bias-free adaptive iir filtering," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 1, pp. 364 –367 vol.1.

[68] Katsuhiko Ogata, *Discrete-Time Control Systems*, Prentice Hall, 1995.

[69] Julius O. Smith, *Introduction to Digital Filters with Audio Applications*, W3K Publishing, http://www.w3k.org/books/, 2007.

[70] D. S.G. Pollock, Richard C. Green, and Truong Nguyen, *Handbook of Time Series Analysis, Signal Processing, and Dynamics*, Signal Processing and its Applications. Academic Press, November 1999.

[71] Panos J Antsaklis and Anthony N. Michel, *A Linear Systems Primer*, Birkhäuser Basel, 2007.

[72] Roger A. Horn and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[73] F. Amato, G. Celentano, and F. Garofalo, "New sufficient conditions for the stability of slowly varying linear systems," *Automatic Control, IEEE Transactions on*, vol. 38, no. 9, pp. 1409 –1411, sep 1993.

[74] A. Ilchmann, D.H. Owens, and D. Prätzel-Wolters, "Sufficient conditions for stability of linear time-varying systems," *Systems & Control Letters*, vol. 9, no. 2, pp. 157 – 163, 1987.

[75] M. Wu, "A note on stability of linear time-varying systems," *Automatic Control, IEEE Transactions on*, vol. 19, no. 2, pp. 162, apr 1974.

[76] C.R. Johnson, "On the interaction of adaptive filtering, identification, and control," *Signal Processing Magazine, IEEE*, vol. 12, no. 2, pp. 22–37, Mar 1995.

[77] A. Carini, V.J. Mathews, and G.L. Sicuranza, "Sufficient stability bounds for slowly varying discrete-time recursive linear filters," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, apr 1997, vol. 3, pp. 1877 –1880 vol.3.

[78] J. Arenas-Garcia, A.R. Figueiras-Vidal, and A.H. Sayed, "Mean-square performance of a convex combination of two adaptive filters," *Signal Processing, IEEE Transactions on*, vol. 54, no. 3, pp. 1078 – 1090, march 2006.

[79] C. G. Lopes, E. Satorius, P. Estabrook, and A. H. Sayed, "Efficient adaptive carrier tracking for mars to earth communications during entry, descent and landing," in *Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on*, aug. 2007, pp. 517 –521.

[80] C.G. Lopes, E.H. Satorius, P. Estabrook, and A.H. Sayed, "Adaptive carrier tracking for mars to earth communications during entry, descent, and landing," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 46, no. 4, pp. 1865 –1879, oct. 2010.

[81] Ching-An Lai, "Nlms algorithm with decreasing step size for adaptive iir filters," *Signal Processing*, vol. 82, no. 10, pp. 1305 – 1316, 2002.

[82] P.M.S. Burt and P.A. Regalia, "A new framework for convergence analysis and algorithm development of adaptive iir filters," *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 3129 – 3140, aug. 2005.

[83] Rama Chellappa and Sergios Theodoridis, Eds., *Signal Processing Theory and Machine Learning*, vol. 1 of *Academic Press Library in Signal Processing*, Elsevier Academic Press, 2013.

[84] J. Arenas-Garcia, V. Gomez-Verdejo, and AR. Figueiras-Vidal, "New algorithms for improved adaptive convex combination of lms transversal filters," *Instrumentation and Measurement, IEEE Transactions on*, vol. 54, no. 6, pp. 2239–2249, Dec 2005.

[85] M. T M Silva and V.H. Nascimento, "Convex combination of adaptive filters with different tracking capabilities," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, April 2007, vol. 3, pp. III–925–III–928.

[86] M.T.M. Silva, V.H. Nascimento, and J. Arenas-Garcia, "A transient analysis for the convex combination of two adaptive filters with transfer of coefficients," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, march 2010, pp. 3842 –3845.

[87] L.F.O. Chamon, H.F. Ferro, and C.G. Lopes, "A data reusage algorithm based on incremental combination of lms filters," in *Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on*, Nov 2012, pp. 406–410.

[88] L.A. Azpicueta-Ruiz, A.R. Figueiras-Vidal, and J. Arenas-Garcia, "A normalized adaptation scheme for the convex combination of two adaptive filters," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 3301–3304.

[89] V.H. Nascimento, M. Silva, R. Candido, and J. Arenas-Garcia, "A transient analysis for the convex combination of adaptive filters," in *Statistical Signal Processing, 2009. SSP '09. IEEE/SP 15th Workshop on*, 31 2009-sept. 3 2009, pp. 53 –56.

[90] M. Martinez-Ramon, J. Arenas-Garcia, A. Navia-Vazquez, and A.R. Figueiras-Vidal, "An adaptive combination of adaptive filters for plant identification," in *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, 2002, vol. 2, pp. 1195–1198 vol.2.

[91] M. Lazaro-Gredilla, L.A. Azpicueta-Ruiz, A.R. Figueiras-Vidal, and J. Arenas-Garcia, "Adaptively biasing the weights of adaptive filters," *Signal Processing, IEEE Transactions on*, vol. 58, no. 7, pp. 3890–3895, 2010.

[92] R. Candido, M. T M Silva, and V.H. Nascimento, "Affine combinations of adaptive filters," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, 2008, pp. 236–240.

[93] A.V. Oppenheim, A.S. Willsky, and I.T. Young, *Signals and systems*, Prentice-Hall signal processing series. Prentice-Hall, 1983.

[94] Wilhelmus H. Schilders, Henk A. Van der Vorst, and Joost Rommes, *Model Order Reduction: Theory, Research Aspects and Applications*, vol. 13 of *Mathematics in Industry - The European Consortium for Mathematics in Industry*, Springer-Verlag Berlin Heidelberg, August 2008.

[95] A. Megretski, *Lecture Notes on Fundamentals of Model Order Reduction*, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-245-multivariable-control-systems-spring-2004/lecture-notes/lec8_6245_2004.pdf, March 2004.

[96] Tatjana Stykel, "Gramian-based model reduction for descriptor systems," *Mathematics of Control, Signals and Systems*, vol. 16, no. 4, pp. 297–319, 2004.

[97] Serkan Gugercin and Athanasios C. Antoulas, "Model reduction of large-scale systems by least squares," *Linear Algebra and its Applications*, vol. 415, no. 2-3, pp. 290–321, 2006, Special Issue on Order Reduction of Large-Scale Systems.

[98] Serkan Gugercin and Athanasios C. Antoulas, "A survey of balancing methods for model reduction," in *In Proc. European Control Conference ECC 2003*. Citeseer, 2003.

[99] J.R. Phillips and L.M. Silveira, "Poor man's tbr: A simple model reduction scheme," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 24, no. 1, pp. 43–55, Jan 2005.

[100] I.T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics – Statistical Theory and Methods. Springer-Verlag New York, 2nd edition, 2002.

[101] B. Wayne Bequette, *Process Control: Modeling, Design and Simulation*, Prentice Hall International Series in the Physical and Chemical Engineering Sciences. Prentice Hall, 1st edition, December 2003.

[102] B.L. Stevens, F.L. Lewis, and E.N. Johnson, *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, Wiley, 2015.

[103] B. G. Mertzios and F. L. Lewis, "An algorithm for the computation of the transfer function matrix of generalized two-dimensional systems," *Circuits, Systems and Signal Processing*, vol. 7, no. 4, pp. 459–466, 1988.

[104] A.G. Phadke, *Handbook of Electrical Engineering Calculations*, Electrical and Computer Engineering. Taylor & Francis, 1999.

[105] Franklin T. Luk and Sanzheng Qiao, "A fast eigenvalue algorithm for hankel matrices," *Linear Algebra and its Applications*, vol. 316, no. 1, pp. 171–182, 2000, Special Issue: Conference celebrating the 60th birthday of Robert J. Plemmons.

[106] V.Y. Pan, Z. Chen, A. Zheng, et al., "The complexity of the algebraic eigenproblem," *Preprint*, vol. 71, 1998.

[107] Victor Y. Pan and Zhao Q. Chen, "The complexity of the matrix eigenproblem," in *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, New York, NY, USA, 1999, STOC '99, pp. 507–516, ACM.

[108] J Hagel, *Nonlinear Perturbation Methods with Emphasis to Celestial Mechanics*, Madeira Univesity Press, 1995.

[109] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, 3rd edition, September 2007.

[110] Charles S. Burrus and T.W. Parks, "Time domain design of recursive digital filters," *Audio and Electroacoustics, IEEE Transactions on*, vol. 18, no. 2, pp. 137–141, Jun 1970.

[111] T W Parks and C S Burrus, *Digital Filter Design*, Topics in Digital Signal Processing. Wiley-Interscience, August 1987.

[112] T.F. Chan and P.C. Hansen, "A look-ahead levinson algorithm for general toeplitz systems," *Signal Processing, IEEE Transactions on*, vol. 40, no. 5, pp. 1079–1090, May 1992.

[113] Roland W. Freund, "A look-ahead bareiss algorithm for general toeplitz matrices," *Numerische Mathematik*, vol. 68, no. 1, pp. 35–69, 1994.

[114] T. Kailath and A. H. Sayed, *Fast Reliable Algorithms for Matrices with Structure*, Advances in Design and Control. Society for Industrial and Applied Mathematics, January 1987.

[115] Bernhard Beckermann and George Labahn, "A uniform approach for the fast computation of matrix-type pade approximants," *SIAM Journal on Matrix Analysis and Applications*, vol. 15, no. 3, pp. 804–823, July 1994.

[116] Vadim Olshevsky, *Fast Algorithms for Structured Matrices: Theory and Applications*, Contemporary Mathematics. Amer Mathematical Society, June 2003.

[117] A.C. Antoulas, D.C. Sorensen, and Y. Zhou, "On the decay rate of hankel singular values and related issues," *Systems & Control Letters*, vol. 46, no. 5, pp. 323–342, 2002.

[118] Hon Keung Kwan and Aimin Jiang, "Recent advances in fir approximation by iir digital filters," in *Communications, Circuits and Systems Proceedings, 2006 International Conference on*, june 2006, vol. 1, pp. 185 –190.

[119] P.L. Feintuch, "An adaptive recursive lms filter," *Proceedings of the IEEE*, vol. 64, no. 11, pp. 1622 – 1624, nov. 1976.

[120] Jr. Johnson, C.R., M.G. Larimore, P.L. Feintuch, and N.J. Bershad, "Comments on and additions to an adaptive recursive lms filter," *Proceedings of the IEEE*, vol. 65, no. 9, pp. 1399 –1402, sept. 1977.

[121] B. Widrow, J.M. McCool, and P.L. Feintuch, "Comments on an adaptive recursive lms filter," *Proceedings of the IEEE*, vol. 65, no. 9, pp. 1402 – 1404, sept. 1977.

[122] M. Rupp and AH. Sayed, "On the stability and convergence of feintuch's algorithm for adaptive iir filtering," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, May 1995, vol. 2, pp. 1388–1391 vol.2.

[123] Ali H. Sayed and Markus Rupp, "An l2-stable feedback structure for nonlinear adaptive filtering and identification," *Automatica*, vol. 33, no. 1, pp. 13 – 30, 1997.

[124] Z. Pritzker and A. Feuer, "Variable length stochastic gradient algorithm," *Signal Processing, IEEE Transactions on*, vol. 39, no. 4, pp. 997–1001, Apr 1991.

[125] Yu Gong and C.F.N. Cowan, "An lms style variable tap-length algorithm for structure adaptation," *Signal Processing, IEEE Transactions on*, vol. 53, no. 7, pp. 2400–2407, July 2005.

[126] Yonggang Zhang and J.A. Chambers, "Convex combination of adaptive filters for a variable tap-length lms algorithm," *Signal Processing Letters, IEEE*, vol. 13, no. 10, pp. 628–631, Oct 2006.

[127] N.R. Yousef and A.H. Sayed, "An unified approach to the steady-state and tracking analyses of adaptive filters," *Signal Processing, IEEE Transactions on*, vol. 49, no. 2, pp. 314–324, Feb 2001.

[128] Kenneth E. Barner and Gonzalo R. Arce, Eds., *Nonlinear Signal and Image Processing: Theory, Methods, and Applications*, Electrical Engineering & Applied Signal Processing Series. CRC Press, 2003.

[129] L. Pasquato and Z. Kale, "Adaptive iir filter initialization via hybrid fir/iir adaptive filter combination," *Instrumentation and Measurement, IEEE Transactions on*, vol. 50, no. 6, pp. 1830 –1835, dec 2001.

[130] Lennart Ljung, *System Identification: Theory for the User*, Prentice-Hall Information and System Sciences Series, 1987.

[131] Robert H. Shumway and David S. Stoffer, *Time Series Analysis and Its Applications*, Springer Texts in Statistics. Springer, 3rd edition, November 2010.

[132] T. Cassar, K.P. Camilleri, and S.G. Fabri, "Order estimation of multivariate arma models," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 3, pp. 494–503, 2010.

[133] J. Botts, J. Escolano, and Ning Xiang, "Design of iir filters with bayesian model selection and parameter estimation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 669–674, 2013.

[134] M. Nayeri, Hong Fan, and W.K. Jenkins, "Some characteristics of error surfaces for insufficient order adaptive iir filters," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 7, pp. 1222–1227, Jul 1990.

[135] Thomas Edson Filgueiras Filho, *Sobre a velocidade de convergência da filtragem adaptativa IIR*, Ph.D. thesis, Escola Politécnica, Universidade de São Paulo, São Paulo, 2008.