KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Electrical Engineering

Machine Learning for Wireless Communications

Hybrid Data-Driven and Model-Based Approaches

LISSY PELLACO

Stockholm, Sweden 2022

Machine Learning for Wireless Communications

Hybrid Data-Driven and Model-Based Approaches

LISSY PELLACO

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Thursday, 15 December 2022, at 1:00 p.m. in F3, Lindstedtsvägen 26, KTH Campus, Stockholm.

Doctoral Thesis in Electrical Engineering
KTH Royal Institute of Technology
Stockholm, Sweden 2022

ISBN: 978-91-8040-356-6
TRITA-EECS-AVL-2022:58

Printed by: Universitetsservice US-AB, Sweden 2022

To Silvio and my family

Abstract

Machine learning has enabled extraordinary advancements in many fields and penetrates every aspect of our lives. Autonomous driving cars and automatic speech translators are just two examples of the numerous applications that have become a reality yet seemed so distant a few years ago. Motivated by this unprecedented success of machine learning, researchers have started investigating its potential within the field of wireless communications, and a plethora of outstanding data-driven solutions have appeared.

In this thesis, we acknowledge the success of machine learning, and we corroborate its role in shaping the future generation of cellular systems. However, we argue that machine learning should be combined with solid theoretical foundations and expert knowledge as the basis of wireless systems. Machine learning allows a substantial performance gain when traditional approaches fall short, e.g., when modeling assumptions fail to capture reality accurately or when conventional algorithms are computationally costly. Likewise, the injection of domain knowledge into data-driven solutions can compensate for typical machine learning shortcomings, such as a lack of interpretability and performance guarantees, poor scalability, and questionable robustness.

In this thesis, composed of five technical papers, we present novel hybrid model-based and data-driven approaches in three application areas: interference detection for satellite signals, channel prediction for link adaptation, and downlink beamforming in MU-MISO and MU-MIMO settings. We go beyond a mere application of machine learning and adopt a reasoned approach to integrate domain knowledge synergistically. As a result, the proposed approaches, on the one hand, achieve remarkable empirical performance and, on the other hand, are supported by theoretical analysis. Furthermore, we pay particular attention to the explainability of all our proposed approaches since the typical black-box nature of data-driven solutions constitutes one of the major obstacles to their actual deployment, especially in the wireless communications field.

Sammanfattning

Maskininlärning har möjliggjort stora framsteg inom många fält och genomsyrar alla aspekter av våra liv. Självkörande bilar och automatisk textöversättning är två exempel bland de talrika tillämpningar som har förverkligats trots att de verkade avlägsna för endast ett par år sedan. Framgångarna inom maskininlärning saknar motstycke, och forskare har sedermera börjat undersöka maskininlärningens potential inom trådlös kommunikation. I kölvattnet av denna forskning har otaliga och framgångsrika datadrivna tekniker uppstått.

Denna avhandling bygger på tidigare framgångar inom maskininlärning, och vi vidareutvecklar dess roll som tongivande teknik i framtida telekommunikationssystem. Dock argumenterar vi för att tekniker inom maskininlärning bör kombineras med goda teoretiska grunder och expertkunskap inom trådlösa system. Maskininlärning kan ge stora prestandavinster i fall där traditionella metoder är otillräckliga, till exempel när modelleringsantaganden visar sig vara inkorrekta eller när konventionella algoritmer är för kostsamma beräkningsmässigt. Dessutom kan domänspecifik kunskap kompensera för tillkortakommanden hos datadrivna metoder, exempelvis brist på tolkningsbarhet och prestandagarantier, dålig skalbarhet och bristfällig robusthet.

I denna avhandling, bestående av fem tekniska artiklar, presenterar vi nya hybrida modellbaserade och datadrivna tillvägagångssätt inom tre tillämpningsområden: interferensdetektion för satellitsignaler, kanalprediktion för länkanpassning, samt lobformning för nedlänk i MU-MISO- och MU-MIMO-tillämpningar. Vi går bortom enbart tillämpningar av maskininlärning, och använder istället ett balanserat tillvägagångssätt för att integrera domänspecifik kunskap på ett synergistiskt sätt. Dessa tillvägagångssätt resulterar å enda sidan i anmärkningsvärd prestanda i empiriska undersökningar, och grundas å andra sidan i teoretisk analys. Dessutom lägger vi stor vikt vid förklarbarheten hos de föreslagna tillvägagångssätten eftersom datadrivna metoder lider av brist på transparens, vilket är ett av de största hindren för metodernas praktiska användning, speciellt inom trådlösa kommunikation.

Acknowledgements

My first words of deepest gratitude go to my supervisor, Prof. Joakim Jaldén, without whom this thesis would not have been possible. His confident and passionate attitude towards research has been a true source of inspiration and motivation during these years. I am extremely thankful to him for his insightful and expert guidance, for his constant support, encouragement, and patience and, most importantly, for always believing in me especially when I doubted myself.

I am thankful to Prof. Santiago Segarra for serving as opponent, to Prof. Sergiy A. Vorobyov, Prof. Danyo Danev, and Dr. Eva Lagunas for serving as committee members and to Prof. Mikael Johansson for serving as substitute member. I would like to extend my gratitude to Prof. Saikat Chatterjee for reviewing this thesis and to Prof. Tobias Oechtering for serving as chair at the defense. I will forever be thankful to Mallikarjun Kande for believing in me and encouraging me to pursue a doctoral degree at KTH.

I am extremely thankful to my coauthors, from whom I have learnt so much. It has been a privilege working with them and I am very grateful to them for their time and patience and for sharing their invaluable knowledge. In particular, I would like to thank Prof. Mats Bengtsson for his expert advice and keen eye in our various collaborations. I would like to mention and thank also Nirankar Singh and his colleagues at the Swedish Space Corporation for our collaboration.

Next, I would like to express my appreciation to the faculty at ISE division for providing a great work and educational environment. At KTH, I have been very lucky to meet great colleagues and friends, including Alexander, Alireza, Amaury, Amirreza, Antoine, Antonios, Anubhab, Arun, Baptiste, Borja, Boules, Deyou, Dong, Ella, Fotios, Germán, Giulia, Hadi, Hamid, Hanwei, Hao, Hasan, Henrik, Håkan, Javier, Jaya, Jeannie, Jing, Leandro, Linghui, Manuel, Marie, Michail, Movitz, Ramana, Sandipan, Sara, Seyed, Sina, Vidit, Vishnu, Wanlu, Wendi, Xinyue, Yang, and Yusen. I am deeply thankful to all of them for their cheerful company during these years. I would like to express special thanks to Martin for his feedback on the abstract of this thesis and for translating it to Swedish. Special mention goes to my dear friends Peter, for all the laughs during our tea breaks and for making the WASP trips much more enjoyable, and Prakash, for the fun time spent together. Special mention also goes to my dear friend and very talented artist Xuechun and to my dear friends Mina and He, for our girls' lunches

and dinners. I was extremely lucky to enjoy Xuechun's company at the office too. With her support, positive energy, and jokes, she made my days much more fun and memorable and with her hard work she has certainly inspired me. My heartfelt thanks go to her for taking care of the GPU at the department and for her patience every time I asked for technical support. Another person whose presence during these years I am extremely thankful for, is my dear friend Sahar. I will always cherish our time together, our lunches and our talks. My gratitude goes to her for her kind attitude, for always being willing to help, and for involving me into the Female PhD Student Network at EECS school. I will be forever grateful to my academic brother and dear friend Pol del Aguila Pla for helping me in so many ways, professionally and personally, when I first moved to Sweden. I have learnt uncountable things from him and his passion for research and teaching are truly motivating. I am indebted to him for being such a great teacher and for always supporting me and motivating me regardless of the distance.

My warmest thanks go to all my friends in Italy for making me feel their support and friendship despite the distance. I am extremely thankful to my extended family, Annamaria, Roberto, Stella, and Gianluca, for encouraging me and always cheering for me throughout this journey.

My deepest gratitude goes to my parents and my grandmother Maria for always believing in me, for supporting me in uncountable ways and for their unconditional love. Their presence has certainly made a difference. Thank you all for being a firm point of reference all these years.

Finally, I will be forever indebted to my life partner Silvio. I have been extremely lucky to have him as a companion throughout this journey. Thank you for always encouraging me, unwaveringly cheering for me, and for being such a great listener. Thank you for proofreading most of my works, including this thesis, and for listening to the rehearsals of my presentations. Thank you for your precious suggestions and for always being there for me.

Lissy Pellaco
Stockholm, December 2022

List of Papers

This thesis is based on the following papers:

A **Lissy Pellaco**, Nirankar Singh and Joakim Jaldén, "Spectrum prediction and interference detection for satellite communications," *Advances in Communications Satellite Systems: Proceedings of the 37th International Communications Satellite Systems Conference (ICSSC-2019)*, 2019, pp. 1-18.

B **Lissy Pellaco**, Vidit Saxena, Mats Bengtsson and Joakim Jaldén, "Wireless link adaptation with outdated CSI - a hybrid data-driven and model-based approach," *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1-5.

C **Lissy Pellaco**, Mats Bengtsson and Joakim Jaldén, "Deep Weighted MMSE Downlink Beamforming," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4915-4919. (**Outstanding Student Paper Award**).

D **Lissy Pellaco**, Mats Bengtsson and Joakim Jaldén, "Matrix-Inverse-Free Deep Unfolding of the Weighted MMSE Beamforming Algorithm," in *IEEE Open Journal of the Communications Society*, vol. 3, pp. 65-81, 2022.

E **Lissy Pellaco** and Joakim Jaldén, "A matrix-inverse-free implementation of the MU-MIMO WMMSE beamforming algorithm," *arXiv:2205.08877*, May 2022. Submitted to *IEEE Transactions on Signal Processing*.

Other papers written during the doctoral studies but not included in this thesis:

I Pol del Aguila Pla, **Lissy Pellaco**, Satyam Dwivedi, Peter Händel and Joakim Jaldén, "Clock Synchronization Over Networks Using Sawtooth Models," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5945-5949.

II Pol del Aguila Pla, **Lissy Pellaco**, Satyam Dwivedi, Peter Händel and Joakim Jaldén, "Clock Synchronization Over Networks: Identifiability of the Sawtooth Model," in *IEEE Open Journal of Signal Processing*, vol. 1, pp. 14-27, 2020.

III Olof Engström, Sahar Tahvili, Auwn Muhammad, Forough Yaghoubi and **Lissy Pellaco**, "Performance Analysis of Deep Anomaly Detection Algorithms for Commercial Microwave Link Attenuation," *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2020, pp. 47-52.

Acronyms

AWGN	Additive white Gaussian noise
BICM	Bit-interleaved coded modulation
CNN	Convolutional neural network
CSI	Channel state information
E2E	End-to-end
FIR	Finite impulse response
FISTA	Fast iterative shrinkage-thresholding algorithm
GD	Gradient descent
GPU	Graphics processing unit
IAIDNN	Iterative algorithm induced deep-unfolding neural network
IRG	Interference reduction group
ITU	International telecommunication union
LSTM	Long short-term memory
MCS	Modulation and coding scheme
ML	Machine learning
MMSE	Minimum (Maximum) mean square error - in Papers B-D (Paper A)
MSE	Mean square error
MU-MIMO	Multi-user multiple-input multiple-output
MU-MISO	Multi-user multiple-input single-output
NN	Neural network
OFDM	Orthogonal frequency division multiplexing
PGD	Projected gradient descent
POCS	Projections onto convex set
PSD	Positive semidefinite
QoS	Quality of service
ReLu	Rectified linear unit
RF	Radio frequency
RIS	Reconfigurable intelligent surface
RNN	Recurrent neural network
RZF	Regularized zero forcing
SINR	Signal-to-interference-plus-noise ratio

SISO	Single-input single-output
SNR	Signal-to-noise ratio
WMMSE	Weighted minimum mean square error
WSR	Weighted sum rate
ZF	Zero forcing
3GPP	3rd Generation Partnership Project

Contents

Part I

Introduction and overview of contributions

Chapter 1

Introduction

With the ever-growing availability of data and access to computing power, machine learning has revolutionized many different areas unlocking new applications and achieving performance far beyond what was attainable with traditional approaches. This is true especially for application areas with limited theoretical foundations and in lack of good models, such as speech recognition and autonomous driving. This is not the case for wireless communications, a field deeply rooted in information theory and statistics and based on solid mathematical models. Therefore, at a first glance, one might conclude that the application of machine learning to wireless communications would only lead to marginal if not negligible gains. Yet, with the rapidly escalating traffic load, diversity of end-user applications, increasing network complexity and heterogeneity, machine learning is envisioned as essential for the next generation systems [1–4]. In fact, there are various cases in the field of wireless communications where machine learning can provide significant advances [5–8].

First, when the available models are inadequate. Wireless communication systems are traditionally based on modeling assumptions, typically Gaussianity, linearity, and stationarity, which make the models tractable and suitable for mathematical analysis, but only capable of representing reality to a limited extent. This discrepancy between the underlying theory and the practical systems, which are affected by impairments (e.g., due to the channel) and non-linearities (e.g., due to imperfect power amplifiers or low-resolution analog-to-digital converters), naturally penalizes the performance. Here, machine learning has true potential for improvement as it is not subject to any modeling assumption and can optimize the system for such physical phenomena hard to capture in a model. In addition, there are cases in which it is extremely challenging to model the entire communication system or a building block thereof with a mathematically tractable expression. Again, here machine learning can offer new design approaches by directly learning from the data.

Second, when human behavior is involved, e.g., for network tuning and resource optimization. If user mobility and traffic load could be accurately predicted, the

management of network resources would significantly improve and proactive actions to mitigate network congestion and service degradation could be taken. While manually designing a mathematical model that accurately characterizes human behavior is extremely hard, learning a mathematical model to predict user trajectories and traffic volumes directly from the data has shown promising results, leading to improved resource usage.

Third, when algorithms involve heuristics for parameters selection. Algorithms derived from solid theoretical foundations and with performance guarantees turn sub-optimal if the parameters are not properly chosen. Often it is not trivial, if not mathematically intractable, to optimize for such parameters. In lack of closed-form solutions, it is common to resort to experience, intuition, lookup tables, and a lot of hand-tuning effort. Machine learning instead constitutes a viable solution to effectively optimize such parameters without having to explicitly solve an optimization problem.

Fourth, when algorithms are computationally prohibitive. There are different scenarios for which provably optimal algorithms exist but are too complex for practical implementation, inefficient in terms of execution time and energy consumption. Machine learning constitutes a way to replace such algorithms with highly parallelizable structures that can be executed efficiently and at a lower energy. Neural networks have shown to be universal function approximators and as such have the potential to accurately approximate algorithms (supervised training). At the same time, neural networks architectures are extremely easy to parallelize and this can be leveraged to improve computational costs by orders of magnitude. In addition to problems for which well-structured yet complex algorithms exist, there are also common problems in the field of wireless communications which are NP-hard to address: multi-user scheduling, multiple-input multiple-output (MIMO) detection and sum rate maximization under power constraint, to name a few. In these cases, when no clear solution exists, machine learning can provide approximate solutions while maintaining a bounded computational complexity (unsupervised learning).

Fifth, when end-to-end performance optimization is the goal. Communications system are traditionally modelled as a chain of separate blocks, each performing a specific task (e.g., coding and modulation). This modeling approach makes the optimization of each individual block the only viable (although sub-optimal) option as joint block optimization is prohibitively complex. Here, machine learning offers instead a straightforward way to optimize the end-to-end performance by directly modeling the entire system without any modular structure.

1.1 Background

Motivated by the potential held by machine learning with respect to traditional solutions, researchers in the field of wireless communication have been actively exploring data-driven solutions in various areas. For a comprehensive survey of the plethora of machine-learning-based approaches that have recently appeared,

we refer the reader to [8–12]. Here, we list some applications that we deem most representative and most related to this thesis.

Power control and beamforming design. In [13–15], the authors focus on the sum rate of a multi-user single-antenna interference channel. Instead of addressing this non-convex NP-hard problem with traditional mathematical tools, which lead to computationally-heavy algorithms, the authors train a neural network to learn a mapping between the user channels and the optimal power control. In [13], the authors employ a neural network to approximate via supervised learning the well-established weighted minimum mean square error (WMMSE) algorithm [16], which attains satisfactory (although sub-optimal) performance, but exhibits excessive complexity. To overcome the implicit performance bound given by the WMMSE algorithm and the energy- and time-consuming task of creating a labelled dataset, an unsupervised learning approach is adopted in [14]. The authors train a neural network by directly maximizing the sum rate achieved with the power profile given as output. Supervised and unsupervised training are combined in [15]: the authors first pre-train a neural network with the output of the WMMSE algorithm as ground truth and then, to further boost performance, apply unsupervised training by directly maximizing the sum rate.

Researchers have extended the same data-driven ideas to the more challenging case of multiple-antenna transceivers, where the problem of power control becomes beamforming design. In [17–20], for the sum rate maximization under transmit power constraint, both supervised, unsupervised and combined approaches (analogous to the power control case) have been explored. In this case, however, the neural network gives as output the transmit beamformers, which have a higher dimensionality with respect to the scalar power coefficients. For outage probability minimization under transmit power constraint, a data-driven approach for beamforming design is adopted in [21]. Under the hypothesis that a set of past channel realizations (collected during measurement campaigns) is available, instead of making a modeling assumption about the channel distribution, the authors propose to learn the outage probability directly from the data.

With the increasing number of antennas at the base station, hybrid beamforming has recently garnered attention as it allows for a smaller number of costly radio frequency (RF) chains. With hybrid beamformig, in fact, different antenna elements share the same RF chain. This complicates the beamforming design, as extra constraints, such as the non-convex constant modulus constraint, come into play. Machine-learning-based approaches have been investigated to lower the computational burden and the execution time of traditional solutions. In [22], the authors consider a multiple-input single-output (MISO) system, where a base station is equipped with one large-scale antenna with a single RF chain. They train a neural network to learn the mapping from the channel to the optimal beamformer by directly maximizing the spectral efficiency under the constant modulus constraint. In [23, 24], the same idea is extended to the multiple-input multiple-output (MIMO) setting.

Channel state information (CSI) prediction. To address the problem of outdated

CSI at the base station, various works [25–29] have proposed machine-learning-based solution to accurately predict the future CSI from a collection of past CSIs. Different network functionalities, such as beamforming and link adaptation, rely on the availability of real-time CSI at the base station. While in time division duplex systems, thanks to channel reciprocity, the base station can directly estimate the downlink CSI, for frequency division duplex systems it has to rely on the CSI estimated by the receiver and fed back through signalling. This introduces a delay which affects the usefulness of the CSI fed back to the base station as the wireless channel can change rapidly, especially in high-mobility scenarios. Hence, the need to predict the CSI. Traditional approaches for prediction, e.g., Wiener and Kalman filters, are based on analytical channel models and correlation assumptions that might not always hold in practice. Machine learning instead has the ability to directly learn from the data without any prior assumption on the physical model underneath. In [25], the authors adopt a long short-term memory (LSTM), which is particularly suitable at learning temporal dependencies in time-series, to perform the CSI prediction in a MIMO system. For the same purpose, in [26], the LSTM structure is preceded with a 2D convolutional neural network (CNN) to better exploit the spatio-temporal correlation in the CSI sequence. In [27], to avoid the information loss caused by dimensionality reduction of the 2D CNN, the authors propose to adopt a 3D CNN instead. Researchers have also explored the relation between uplink CSI and donwlink CSI, which is extremely hard to characterize by a tractable mathematical expression. In [28], the authors use a neural network to approximate such uplink-to-downlink mapping function, while in [29] two-module convolutional LSTM is investigated. The first module learns the spatio-temporal correlation between the uplink CSI and downlink CSI and performs feature reduction. The second module performs the prediction from the compact representation and then expands it back to the original dimensionality.

Autoencoder-based end-to-end system. In [30], the authors first introduce the idea of modeling the full communication system composed of transmitter, channel, and receiver by an autoencoder. They cast the problem of designing the communication chain as a joint optimization problem over the transmitter and receiver and adopt the additive white Gaussian noise (AWGN) channel model. They treat communication as a classification task and choose the categorical cross-entropy as loss function. By doing so, the autoencoder is trained to learn a representation of the message which is robust enough to guarantee an accurate reconstruction of the message at the receiver. Experimental results show that the autoencoder performs on par with conventional solutions, yet without the injection of any prior knowledge. Despite the specific choice of loss function and channel model in [30], this approach is in principle applicable to any cost function and any channel model. This is where its convenience lies. It bypasses the classical modular structure to enable end-to-end optimization and offers the possibility to consider cost functions and channels for which no obvious solution exist. In the autoencoder structure in [30], the AWGN channel is modelled as a non trainable layer that adds fixed variance noise. Conversely, to be independent from any channel model, in [31] the authors

replace this layer with a generative adversarial network, trained to approximate the true channel from available data. In [32, 33], this autoencoder-based approach is extended to orthogonal frequency-division multiplexing (OFDM) systems and in [34] to MIMO systems. In [35], the authors consider a reconfigurable intelligent surface (RIS) as part of the communication system and adopt this autoencoder-based modeling approach to jointly optimize the receiver, the RIS beam index, and the transmitter.

Signal detection. In [36], the authors address the problem of symbol detection for molecular communications, where the information is encoded by chemical signals, such as pH concentration. Tractable mathematical models for this communication channel do not exist. Hence, instead of traditional detection methods, based on a crude model of the underlying channel, the authors train different network architectures to perform this task by using real data. Experimental results show that data-driven detectors (entirely channel agnostic) outperform conventional methods. In [37], the authors address the problem of detection in an OFDM system without CSI at the receiver. Conventional receivers first estimate the CSI from the training pilots and then use it to detect the transmitted data. The authors instead train a neural network to reconstruct the transmitted data directly from the pilot and the received data, bypassing the explicit CSI estimation step. While the conventional approach and the proposed approach perform similarly under ideal conditions, in case of few training pilots and hardware impairments, such as clipping distortions, the neural network offers a better solution.

Modulation classification. For this task, deep learning has shown great potential with respect to likelihood-based and feature-based approaches. Likelihood-based approaches can provide the optimal solution in the sense of minimum misclassification probability, but suffer from high computational complexity and are susceptible to model mismatch. Feature-based approaches strike a better complexity-performance trade-off but highly depend on the quality of the engineered features. Conversely, neural networks bypass the need to manually extract features and show excellent performance at a complexity significantly reduced with respect to likelihood-based methods. In [30], the authors employ a CNN for modulation classification and feed it with raw complex-valued basedband samples after transmission over a non-ideal wireless channel affected by fading and clock offset. Modulation classification can also be treated as image classification. In [38, 39], the CNN is fed with gray-scale constellation images and in [40] with spectrogram images. Next to CNNs, also recurrent neural networks have been investigated for modulation recognition because of their ability to leverage the temporal correlations in wireless communications signals [41].

Given the remarkable results, it is clear that machine learning will play a key role in the next generation cellular systems [1, 42]. This is also indicated by the interest shown by different standardization bodies, such 3GPP and ITU [2], and by the likely appearance of standardized hardware optimized for parallel computing in the base stations, as suggested by the recent partnership between Ericsson, one of the

major network equipment vendors, and NVIDIA, a leader in GPU-accelerated computing platforms [43]. These platforms specialized for advanced parallel computing can in fact enable a drastic speed up in execution time when running parallelizable algorithms. Neural network architectures in particular are straighforward to parallelize. As a result, although involving complex operations per se (e.g., matrix multiplications), neural networks can become extremely competitive and compatible with real-time requirements, overcoming the run-time drawbacks of traditional algorithms [8].

1.2 Limitations of machine learning

The advancements enabled by machine learning are impressive. However, they come with a cost as machine learning brings along different shortcomings which hinder its direct applicability in the field of wireless communications.

One major roadblock is represented by the typical black-box nature of machine learning and its consequent lack of explainability. There is in fact still very limited knowledge about the reasons behind the success and failure of machine learning approaches. Engineers in the field of wireless communications are used to rely on theoretical analysis, experience, and a deep understanding of the system. The introduction of machine learning clearly raises skepticism, but it is not only a matter of principle. The lack of explainabilty makes it extremely difficult for engineers to predict how the performance will be affected in case of changes in the system and in the environment and to address performance degradation.

In addition to this, also the generalizability of machine learning models causes concerns. It relates to the capability of a model trained on a given channel distribution to perform equally well on channel realizations drawn from unseen distributions. This is of great relevance as the wireless environment is non stationary and can change rapidly. This also makes evident how heavily machine learning solutions rely on data and, unfortunately, data is always limited in size and diversity.

Given the generalizability and interpretabilty issues, it is clear how difficult it is to provide any performance guarantee for machine-learning-based solutions. This significantly hurdles their adoption, as communication systems are built on error probability bounds, latency guarantees, and similar. As a result, businesses often choose to continue with traditional methods and renounce to potential performance benefits brought by machine learning.

To complicate the performance guarantee controversy, machine learning has been shown to be vulnerable to adversarial examples [44], which are malicious inputs carefully designed (typically by adding a moderate but specific perturbation to a legitimate input) in order to cause erroneous outputs. Thus, hackers might take advantage of such vulnerabilities and cause modulation misclassification [45] and performance degradation [46]. Clearly, developing machine learning models that are immune to adversarial attacks is crucial, but extremely challenging.

The issue of architecture selection represents another hurdle. This choice is extremely important as the architecture of a network influences its generalization capabilities. If the network size is too small, it might lead to underfitting, i.e., the network struggles to fit the training data. On the contrary, if the network size is too large, it might lead to overfitting, i.e., the network underperforms on the test data. As of now, there is no established rule to select the optimal network architecture for a given task. While research is still ongoing [47], it is common to rely on prior experience, intuition, and trial and error. These heuristic-based approaches are also required to set the hyperparameters for the training process, such as the specific optimizer to adopt, as well as its learning rate and the batch size, to name a few.

Finally, scalability must also be taken into account. Machine learning approaches typically do not scale well. In fact, the dimension of the learnable parameter space typically grows with the problem dimension, which could result in a prohibitive training complexity.

1.3 Hybrid model-based and data-driven solutions

Given the aforementioned shortcomings of machine learning solutions and given the expert knowledge available in the field of wireless communications, researchers have proposed to develop hybrid model-based and data-driven solutions to attain the most potential advantage by sourcing from the best of machine learning and communications domains [48]. Data-driven approaches and expert knowledge have been combined in various way.

For downlink beamforming design, the authors of [49] exploit the structure of the optimal solution to the sum rate maximization problem under power constraint [50]. They train a neural network, which, instead of directly predicting the beamformer, gives as output the key features (highly complex to compute via traditional algorithms) to plug into the optimal beamformer structure. This significantly facilitates the training, improves scalability (as the dimensionality of the network output is significantly reduced) and, further, promotes explainability. Analogously, downlink beamforming design for signal-to-interference-plus-noise ratio (SINR) balancing under power constraint is addressed in [49] by exploiting the expert knowledge about uplink-downlink duality. This duality allows to establish a mapping between the downlink beamformer matrix and the uplink power vector. Thus, the network can predict the uplink power vector, which has a lower dimensionality, and then leverage expert knowledge to map it to the beamformer matrix.

For codeword decoding, the naive approach would be to train a machine learning model to reconstruct the original codeword from a noisy version thereof, after transmission over the communication channel. However, the task of decoding cannot be treated as a typical classification problem because the number of classes grows exponentially with the block length and, further, the training set would also need to depict the channel variability over the exponentially growing number of codewords. Hence, to address this scalability issue, neural networks have been em-

bedded into traditional decoding structures. To improve the decoding performance for high density parity check codes, in [51], the authors assign trainable weights to the Tanner graph of the belief propagation algorithm. This approach scales very well for large block lengths and does not suffer from the curse of dimensionality. Also in [52] the authors adopt a hybrid model-based and data-driven approach. They replace sub-components of a typical iterative decoding algorithm for polar codes with neural-network-based units, each trained to decode a given codeword segment.

For symbol detection over finite-memory channels, the authors of [48] inject into the well-known Viterbi algorithm [53] dedicated neural networks to replace any model dependency. Virtebi algorithm can perform symbol detection under the assumption that the log-likelihood function is known. However, typical modeling assumptions might fail to accurately characterize the true log-likelihood. Hence, the authors insert a dedicated neural network to learn the true log-likelihood function directly from the data and make the resulting hybrid Viterbi algorithm model-agnostic and robust to model mismatch.

Among these different ways of integrating data-driven techniques into traditional algorithms, a structured approach tailored for iterative algorithms and specifically designed to optimize the performance within a fixed complexity budget has gained popularity. It was pioneered by Gregor and LeCun in 2010 [54] and is referred to as *deep unfolding*.

Deep unfolding

Deep unfolding is a deep learning technique aimed at addressing the computational complexity of iterative algorithms by injecting expert knowledge into the structure of a neural network. The key idea consists of building a neural network layer by replicating the operations in the iteration loop of the algorithm that we aim to optimize. By concatenating a finite number of such layers, passing through the resulting neural network is equivalent to running the same finite number of iterations of the considered algorithm. By setting the number of network layers, we fix the computational complexity and by inserting trainable parameters into the structure of the neural network, we add flexibility to it and enable performance improvement when training is carried out. The goal of the training is typically to attain the best possible performance within the fixed network architecture or to reach convergence within the fixed number of layers/iterations.

By building the network architecture in this principled way, deep unfolding mitigates several shortcomings of standard neural-network-based approaches [55]. First, the choice of network architecture is dramatically simplified as the architecture is mainly dictated by the reference algorithm. Naturally, the choice of the trainable parameters remains, but knowledge of the underlying algorithm significantly facilitates this choice. Second, the interpretability of the network is greatly improved as the behavior of the network closely follows the behavior of the original algorithm. Hence, it is extremely easier to tackle performance deterioration

and to reasonably estimate how changes in the environment can affect the output. Third, performance guarantees of the original algorithm could be transferred to the unfolded neural network, paving the way towards actual adoption of such techniques in real systems, as it would be possible to depict the worst-case behavior. Fourth, the number of trainable parameters is typically reduced with respect to traditional neural networks. This, in combination with the inherent built-in expert knowledge, enables the network to learn from smaller datasets and confers to it an improved generalizability and arguably a greater robustness to adversarial attacks. Moreover, the number of trainable parameters typically does not depend on the problem dimension (as in standard neural networks), making the overall approach more scalable.

These features make deep unfolding an attractive solution, and in fact, since 2010, it has been applied to tackle a variety of areas in the field of wireless communications. In [56], the authors address MIMO detection, which entails solving an NP-hard optimization problem over a discrete set. The exhaustive search yields the optimal solution but is computationally prohibitive. Thus, the authors propose to apply the iterative projected gradient descent (PGD) approach instead, which has an affordable computational complexity. The authors use the resulting algorithm as basis to build a neural network and enrich the gradient-based iterations by inserting trainable parameters and applying standard non linear operations. By training the neural network in a supervised manner, numerical results show that state-of-the-art performance can be achieved at a lower computational complexity.

The PGD is also used to address the beamforming problem of sum rate maximization under power constraint in the multi-user (MU)-MISO case [57]. The authors apply deep unfolding and augment the PGD steps by learning the gradient and the step sizes. The same beamforming design, but in the context of MU-MIMO in considered in [58, 59]. The authors unfold the well-known iterative WMMSE algorithm. Specifically, they unfold a variant thereof, in which matrix inverses are replaced by trainable modules structured according to the first-order Taylor approximation of the matrix inverse operations. In [60], the authors extend the application of deep unfolding to the problem of multicast beamforming and adopt a sequence of projections onto convex set (POCS) to address the NP-hard quality of service (QoS) problem. The proposed iterative algorithm is unfolded and trainable parameters are embedded in the structure of the network to enable convergence acceleration in the sequence of POCS, once trained.

Amenable to deep unfolding is also the task of compressive sensing, which in the context of wireless communications is used to the reduce the CSI feedback overhead by exploiting the channel sparsity in massive MIMO. In [61], the authors design a neural network by unfolding the fast iterative shrinkage-thresholding algorithm (FISTA) and treat the step sizes and the shrinkage coefficients as learnable parameters.

The application of deep unfolding to the field of wireless communications is quite extensive and not restricted to the examples mentioned here. For a comprehensive survey, we refer the reader to [62–64].

Chapter 2

Contributions

In this section, we summarize the key contributions of the papers that compose the thesis. We focus on wireless communications, both satellite and terrestrial and, urged by the tangible success of machine learning, we propose to adopt machine-learning-based approaches to address the problems at hand. However, we go beyond a blind and pure application of machine learning. We question it by comparing against alternative classical approaches and, instead of disregarding the vast domain knowledge in the application area, we incorporate it into the proposed machine-learning-based solutions. This leads to hybrid model-based and data-driven approaches that aim to combine the best of both worlds: theoretical guarantees and improved explainability from the domain-knowledge foundation, and excellent empirical results from the data-driven component.

We wish to mention that by addressing relevant issues within wireless communications, we significantly contribute towards the achievement of the UN sustainable development goals (SDGs). Wireless communications have in fact been recognized as a crucial enabler for equal (remote) access to healthcare (SDG 3) and education (SDG 4), for sustainable cities and for safe and resilient urban infrastructure (SGD 11), to name a few [65].

The code to reproduce the results of all our papers (except for Paper A because of proprietary issues due to industrial collaboration) is publicly available. The link to each repository can be found in the corresponding paper. The notation used in the following is coherent with the notation adopted in the corresponding paper each subsection refers to.

2.1 Interference detection for satellite links (Paper A)

In this paper, in collaboration with the Swedish Space Corporation, we address the issue of interference detection for satellite links. Interference is indeed one of the major causes for service degradation, poor operational efficiency, and ultimately revenue loss for SatCom operators, including Swedish Space Corporation. A timely

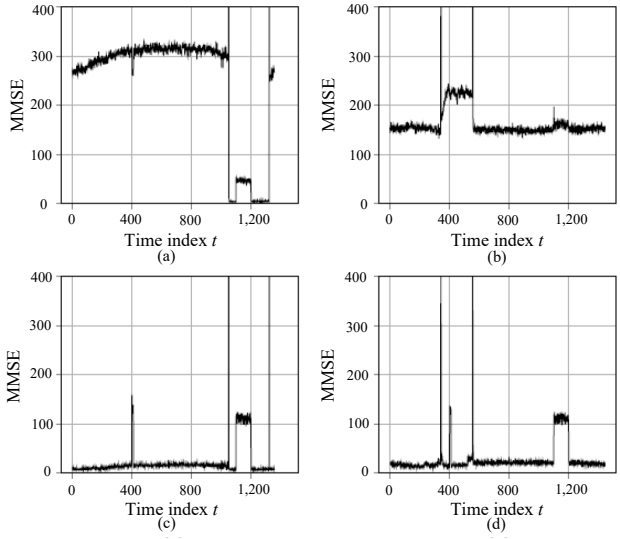

Figure 2.1: Maximum mean square error (MMSE) plots computed for two satellite passes from the prediction given by the baseline, (a) and (b), and from the prediction given by the LSTM, (c) and (d). The signal spikes at $t = 1048$ and $t = 1320$, in (a) and (c), and at $t = 343$ and $t = 555$, in (b) and (d), represent the beginning/end of the communication between the satellite and the base station. They clearly stand out and can be easily filtered out. Instead, the two rectangular pulses that start at $t = 400$ and $t = 1100$ (in all the four plots) denote the presence of interference. As can be seen such pulses can be straightforwardly detected in (c) and (d), while they are more difficult to identify in (a) and (b).

and effective interference detection is essential as it constitutes the first step towards interference mitigation and suppression. Our goal is to detect the interference in the signal spectrum received at the base station from the satellite. The key contribution of the paper consists of designing a real-time and automatic detection technique for both short-term and long-term interference, able to localize the interference both in time and frequency and universally applicable across a discrete set of different signal spectra. The basic idea consists of predicting the interference-free spectrum at each time instant and using it as a reference. If the actual received signal deviates sufficiently from the predicted interference-free spectrum, then the anomaly flag is raised. This approach bypasses the problem of new and unexpected anomalies. To perform the interference-free spectrum prediction, we propose to adopt the LSTM, which has proven successful in learning temporal correlations in many applications [25,26]. As a realistic term of comparison, we consider a baseline approach given by a least squares fit. To gain insights into the learning process of the LSTM, we also employ a single-layer neural network to perform modulation

classification from the compact spectrum representation given as output by the LSTM. Another significant contribution of the paper is represented by the empirical results, which are carried on real spectrum data provided by the Swedish Space Corporation. Both the LSTM-based approach and the baseline yield accurate predictions, but with an essential difference. The predictions given by the LSTM lead to a maximum mean square error (MMSE) plot that enables an extremely easy identification of short-term and long-term interference both in time and frequency, while the predictions given by the baseline yield a less robust interference detection (see Fig. 2.1).

Division of work: The use of a machine learning technique was a request from the Swedish Space Corporation. The doctoral student proposed to specifically adopt an LSTM and to compare its performance against a realistic baseline. The idea behind the proposed baseline was developed with the help of the main supervisor, while the numerical experiments were entirely carried out by the doctoral student. Nirankar Singh from the Swedish Space Corporation provided feedback and support with the real dataset used for the experimental results.

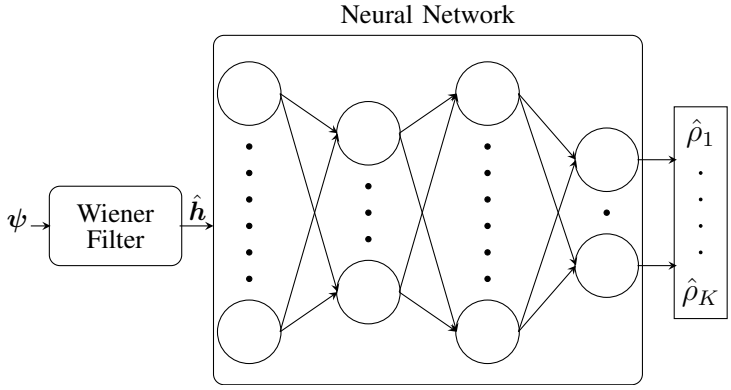

Figure 2.2: Overall hybrid approach. The past channel history $\boldsymbol{\psi}$ is used to optimally estimate, via Wiener filter, the channel in effect at transmission time. Then, the estimate $\hat{\boldsymbol{h}}$ is mapped by the neural network to the corresponding set of conditional error probabilities, i.e., $\hat{\rho}_1, ..., \hat{\rho}_K$, one for each MCS.

2.2 Wireless channel prediction (Paper B)

In this paper, we investigate the problem of outdated channel state information (CSI) in the context of link adapation. Link adaptation consists of tuning the modulation and coding scheme (MCS) parameters at the transmitter according to the instantaneous CSI in order to maximize the spectral efficiency. We consider and extend the data-driven MCS selection scheme proposed by Saxena, Jaldén, Bengtsson (co-authors of this paper), and Tullberg [66]. They propose to use a neural

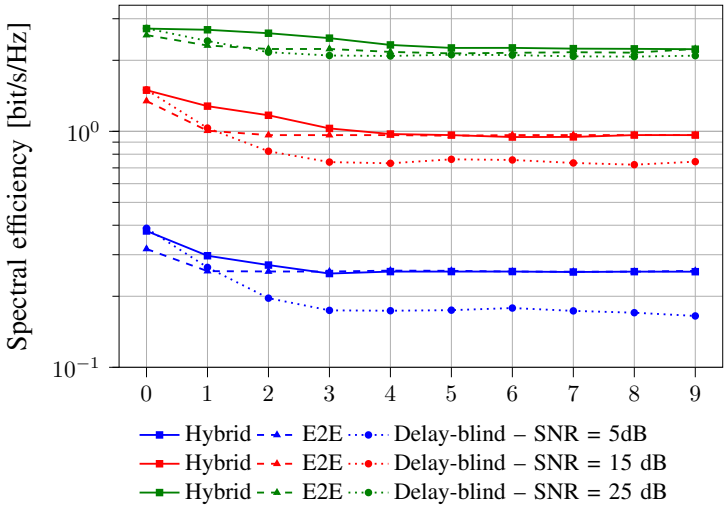

Figure 2.3: Spectral efficiency obtained with the hybrid approach, with the E2E approach, and with the delay-blind approach (used as baseline) at different SNR values, with relative user velocity of 60 km/h.

network to estimate the probability of unsuccessful frame decoding for each MCS, conditioned to the instantaneous CSI. However, the assumption that instantaneous CSI is available at the time of transmission is impractical due to the delay introduced by the feedback loop used to send back the estimated CSI to the transmitter. Therefore, the CSI available at the time of transmission is typically outdated and this naturally leads to a mismatch between the instantaneous CSI and the selected MCS, reducing the benefits of link adaptation. In this work, we adopt the more realistic assumption that some CSI history is available at the transmitter instead and we propose to leverage this history to compensate for the performance degradation due to the feedback delay. This problem was also addressed in [25–27]. As key contribution, we present two approaches: one fully data-driven and one hybrid data-driven and model-based. In the fully data-driven approach, we feed a neural network with the CSI history and train it end-to-end (E2E) to predict the error probability conditioned to the instantaneous CSI for all the MCSs. In the hybrid approach, we propose instead to feed the neural network with a lower-dimensionality sufficient statistic for the instantaneous CSI, computed from the CSI history, which in turn results into a smaller neural network. Under the (common) assumption that the channel evolves as a Gaussian random process, the sufficient statistic is given by the output of a finite impulse response (FIR) Wiener filter, which is a linear minimum mean square error estimator in this case (see Fig. 2.2). Another contribution of the paper consists of proving that replacing the CSI history with the sufficient statistics comes without loss of generality. We perform numerical ex-

periments and show that the E2E and the hybrid approaches are both successful at compensating for the feedback delay (see Fig. 2.3). However, considering that the hybrid approach i) exhibits improved explainability, as the CSI prediction step is clear and has theoretical justification, and ii) entails a smaller and hence easier to train neural network, we recommend the hybrid approach.

Division of work: The main idea of performing channel prediction and comparing the E2E and the hybrid approaches are attributed to the doctoral candidate, who also showed the optimality of the hybrid approach with the help of the main supervisor. The collaborators provided feedback and guidance. Specifically, Dr. Vidit Saxena contributed to the numerical results, providing the implementation of the neural network for MCS selection proposed in [66], and generated the dataset to perform the numerical evaluation.

2.3 Matrix-inverse-free deep unfolding of the WMMSE beamforming algorithm (Papers C, D, and E)

This body of work focuses on the well-known WMMSE algorithm [16], which addresses the NP-hard non-convex weighted sum rate (WSR) maximization problem under a total transmit power constraint. This problem was also addressed in [17–20, 49, 57–59]. The optimization variable is the transmit beamformer, which we denote by $\boldsymbol{V}$. The key idea behind the WMMSE algorithm is to work on an equivalent reformulation (i.e., with the same optimal $\boldsymbol{V}$) of the WSR maximization problem. This reformulation is more tractable as it introduces two extra optimization variables $\boldsymbol{U}$ and $\boldsymbol{W}$, which make the optimization problem amenable to block coordinate descent. Although still non-convex in ($\boldsymbol{U}$, $\boldsymbol{W}$, $\boldsymbol{V}$), the equivalent reformulation addressed by the WMMSE algorithm is convex in each individual optimization variable. By iteratively optimizing over each variable (while keeping the other two fixed), the WMMSE algorithm is guaranteed to converge to a stationary point of the WSR maximization problem [16, Theorem 3] and in practice attains satisfactory performance. Yet, it entails complex and hard-to-parallelize operations at each iteration (i.e., matrix inverses, eigedecompositions, and bisection searches), which, as such, make the WMMSE algorithm incompatible to real-time implementation. To address this, in this body of work we propose to replace such operations with operations efficiently implementable in parallel on modern hardware platforms, while i) retaining the same convergence guarantees and ii) achieving a comparable performance by facilitating the application of deep unfolding.

In papers C and D, we consider the MU-MISO downlink channel. In paper E, we consider the more general MU-MIMO downlink channel. Note that although paper C (conference paper) is largely superseded by paper D (journal paper), we include it in the thesis because it first introduced the key idea behind the proposed approach and because it contains numerical results not included in paper D.

Division of work: The basic idea of applying deep unfolding to the WMMSE algorithm is the result of various discussions between the doctoral candidate and

the main supervisor. The idea of addressing the MIMO case is attributed to the doctoral candidate, who also developed the proofs with support from the main supervisor, and carried out all the numerical evaluations. For papers C and D, Prof. Mats Bengtsson provided feedback and guidance.

MU-MISO scenario (Papers C and D)

In the MU-MISO scenario, optimization variables $\boldsymbol{U}$ and $\boldsymbol{W}$ are scalars and hence their iterative updates involve only lightweight scalar operations. Conversely, optimization variable $\boldsymbol{V}$ is a matrix and its iterative update entails an eigendecomposition, a bisection search, and a matrix inverse. This hard-to-parallelize update for $\boldsymbol{V}$ stems from the method of Lagrange multipliers which is used to solve the convex optimization on variable $\boldsymbol{V}$ (while keeping the other two fixed). We propose to tackle the optimization over $\boldsymbol{V}$ differently by resorting to a first-order method, i.e., the projected gradient descent (PGD), which only involves simple matrix-vector multiplications and summations. As a result, we replace the hard-to-parallelize operations in the original WMMSE algorithm with operations that can be efficiently executed in parallel and that are conformant with the structure of standard neural networks. Consequently, we facilitate the application of deep unfolding, which the authors of [13] openly avoided: "Unlike the algorithms with relatively simple structure that could be easily unfolded, SP [signal processing] algorithms often entail computationally heavy iterations involving operations such as matrix inversion, SVD [singular value decomposition], and/or bi-section. Therefore, their iterations are not amenable to approximate by a single layer of the network" [13, Section 1]. Thus, one key contribution of paper C consists of the proposed reformulation of the WMMSE algorithm that enables one of the first applications of deep unfolding to this popular beamforming algorithm. We refer to this reformulation as *unfoldable WMMSE* algorithm precisely to stress its suitability to deep unfolding. Concurrently to us, also Hu *et al.* [58] first apply deep unfolding to the WMMSE algorithm, albeit the resulting network (IAIDNN) still involves one matrix inverse operation (along with another one in the initialization step) and hence is not fully parallelizable. It must be mentioned, however, that IAIDNN is applicable to the MU-MIMO case, while the unfoldable WMMSE algorithm proposed in paper C is restricted to the MU-MISO scenario (in paper E we will extend it to the MU-MIMO case). Another key contribution is given by Theorem D.1 (in paper D, with full proof in the Appendix), in which, differently from Hu *et al.*, we provide theoretical guarantees for the unfoldable WMMSE algorithm. In particular, we establish that, provided that the step size of the PGD is appropriately chosen, our reformulation of the WMMSE algorithm retains the same convergence guarantees of the original WMMSE algorithm (convergence to a stationary point), although only a finite number of PGD steps is considered per iteration. The proof is essentially a combination of the proof of convergence of the block coordinate descent and the convergence property of the PGD in case of a convex L-smooth function, which the equivalent objective function of the WMMSE algorithm is (if treated as a function

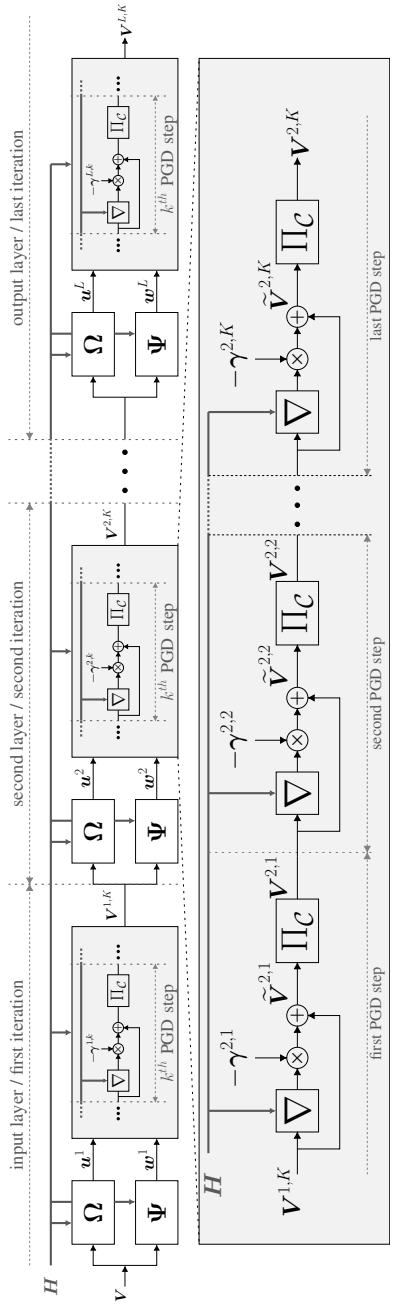

Figure 2.4: Network architecture given by L iterations of the unfoldable WMMSE algorithm in the MU-MISO case. The superscripts $(\cdot)^{l,k}$ indicate the k^{th} PGD step in the l^{th} layer/iteration. Each layer consists of the update equation of $\boldsymbol{U}$ (scalar), denoted by $\boldsymbol{\Omega}$, of the update equation of $\boldsymbol{W}$ (scalar), denoted by $\boldsymbol{\Psi}$, and of the update equation of $\boldsymbol{V}$, given by K PGD steps, as depicted in the gray box.

of $\boldsymbol{V}$), as we show in Lemma D.2.

As dictated by deep unfolding, we build a network by replicating a finite number of iterations of the unfoldable WMMSE algorithm, as depicted in Fig. 2.4. We treat the PGD step sizes as trainable parameters and train the network by maximizing the WSR achieved with the transmit beamformer given as output by the network. To further boost the performance and confer more flexibility to the network, we incorporate Nesterov acceleration and a generalization thereof (i.e., Super Nesterov) into the PGD steps and train the Nesterov momentum parameters jointly with the PGD step sizes. Numerical results indicate a performance on par with the original WMMSE algorithm truncated to the same number of iterations, even when considering only a small finite number of iterations (see Fig. 2.5), and show robustness to changes in the channel distribution. We highlight that our approach performs well both in the lightly loaded and in the fully loaded scenarios, while the IAIDNN struggles to compete in the fully loaded scenario.

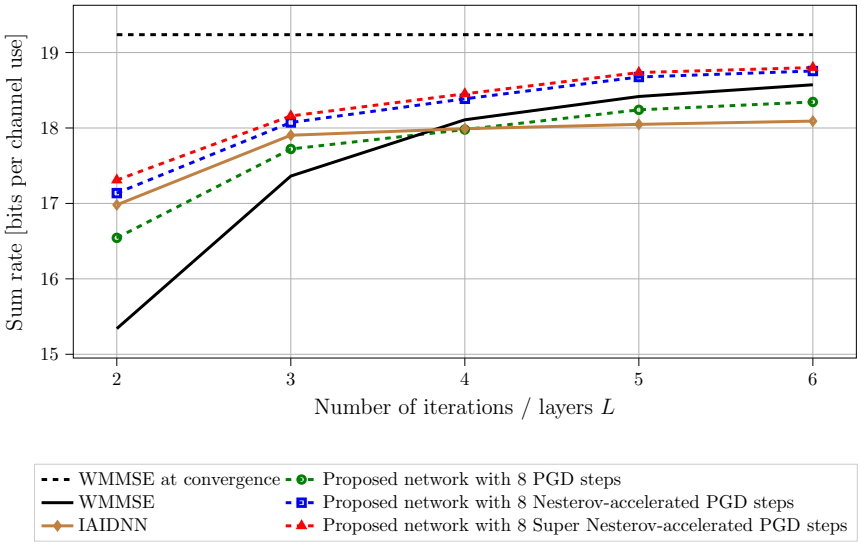

Figure 2.5: WSR obtained in a MU-MISO scenario with four antennas at the base station, four single-antenna users, and a ratio between transmit power and noise power of 20 dB.

MU-MIMO scenario (Paper E)

In the MU-MIMO scenario, all the three variables ($\boldsymbol{U}$,$\boldsymbol{W}$, and $\boldsymbol{V}$) are matrices and their iterative updates involve hard-to-parallelize operations. In particular, the updates of $\boldsymbol{U}$ and $\boldsymbol{W}$ involve matrix inverses, while the update of $\boldsymbol{V}$, in addition to matrix inverses, involves also eigendecompositions and bisection searches. Following the idea introduced by Hu *et al.*, we incorporate the transmit power constraint into

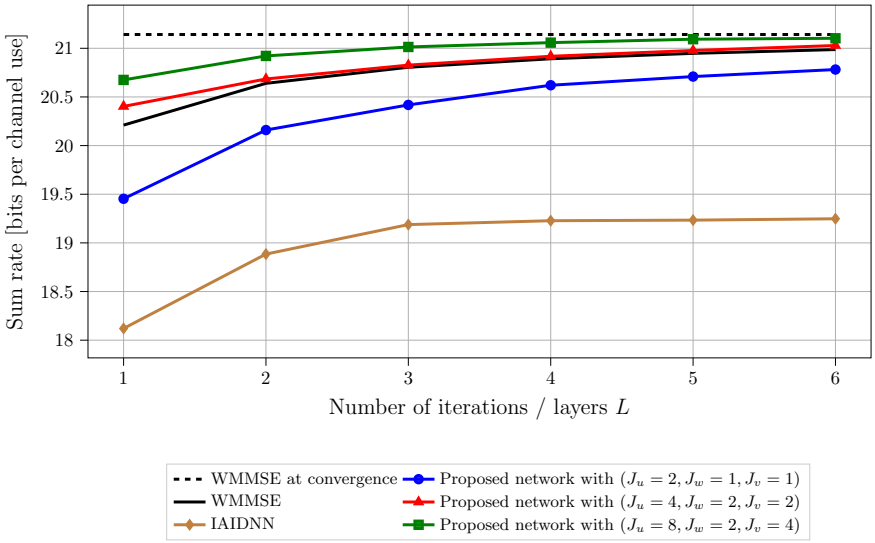

Figure 2.6: WSR obtained in a MU-MIMO scenario with eight antennas at the base station, two four-antenna users, four data streams, a ratio between transmit power, and noise power of 10 dB. J_u (J_v) indicates the number of GD steps in the update of $\boldsymbol{U}$ ($\boldsymbol{V}$) and J_w indicates the number of Schulz iterations in the update of $\boldsymbol{W}$.

the objective function of the WMMSE algorithm and by doing so we remove the eigendecompositions and the bisection searches in the update of $\boldsymbol{V}$. To replace the matrix inverses in the updates of $\boldsymbol{U}$ and $\boldsymbol{V}$, we extend the idea from papers C and D and resort to a first-order method (the gradient descent in this case) that involves only parallelizable operations amenable to deep unfolding. To replace the matrix inverse in the update of $\boldsymbol{W}$, however, we cannot resort to any first-order method. The gradient of the objective function of the WMMSE algorithm with respect to $\boldsymbol{W}$ involves a matrix inverse itself and furthermore with a first-order method it is not easy to guarantee the positive semidefiniteness of $\boldsymbol{W}$, which is an essential property. Specifically, the update of $\boldsymbol{W}$ in the WMMSE algorithm is given by $\boldsymbol{W} = \boldsymbol{E}^{-1}$, where $\boldsymbol{E}$ is the mean square error (MSE) matrix. We aim to bypass matrix inverses and therefore we propose to implicitly compute the inverse of $\boldsymbol{E}$ by applying (Newton) Schulz iterative approach, which is an established technique to approximate matrix inverses. It is immediate to show that Schulz iterations grant the positive semidefiniteness of $\boldsymbol{W}$ (see Lemma E.5), but it is far from trivial to guarantee that Schulz iterations do not alter the convergence properties of the original WMMSE algorithm. In Theorem E.1 and in Theorem E.2, we establish the (universal) conditions under which monotonicity and convergence to a stationary point hold. Thus, the key contribution of the paper consists of the first truly matrix-inverse-free (and hence parallelizable) formulation of the WMMSE algorithm that has the same theoretical guarantees of the original WMMSE algorithm. The proof

of Theorem E.1 is based on the following two key observations. First, the objective function in the WMMSE algorithm is convex and L_u (L_v)-smooth in $\boldsymbol{U}$ ($\boldsymbol{V}$) (Lemma E.3 and E.4). Thus, by properly bounding the step size of the gradient descent we can guarantee monotonicity over the update of $\boldsymbol{U}$ ($\boldsymbol{V}$). Second, the Schulz iterations grant a monotonic decrement if all the eigenvalues of $\boldsymbol{EW}$ are in the range $[0, \delta]$ where $1 < \delta < 2$ (Lemma E.5). Thus, by properly initializing the algorithm and imposing a bound on the perturbation of $\boldsymbol{E}$ given by the updates of $\boldsymbol{U}$ and $\boldsymbol{V}$, we can ensure that the condition on the eigenvalues of $\boldsymbol{EW}$ will be satisfied throughout the entire algorithm. The proof of Theorem E.2 is a combination of the convergence property of the gradient descent and the key observation that the only acceptable fixed point of the Schulz iterations is $\boldsymbol{E}^{-1}$.

To exemplify the relevance of our theoretical contribution, we apply deep unfolding to a more flexible variant of our matrix-inverse-free formulation of the WMMSE algorithm, which is more appropriate to capitalize on the benefits of deep unfolding, albeit missing some of the theoretical guarantees. Also in this case, numerical results i) show performance comparable to the original WMMSE algorithm truncated to the same number of iterations, even when considering a fixed low number of iterations (see Fig. 2.6), and ii) suggest superiority in terms of achieved WSR with respect to the IAIDNN proposed by Hu *et al.* in case of fully loaded scenarios.

Chapter 3

Future work

In this thesis, we proposed hybrid data-driven and model-based solutions for three different applications in the field of wireless communications: interference detection for satellite signals, channel prediction for link adaptation, and downlink beamforming in MU-MISO and MU-MIMO scenarios. The results, both from an empirical viewpoint and a theoretical viewpoint, are promising and encourage further research in these areas to pave the way towards actual deployment of such techniques.

In Paper A, we considered the task of interference detection. However, interference detection constitutes only the first step because the end goal is to reduce and suppress the interference. Therefore, it would be interesting to combine the proposed detector with mitigation techniques and assess the capability of effectively attenuating such anomalies. Further, in case of data-driven attenuation techniques, it would be interesting to train the overall approach (interference detection and mitigation) end-to-end and evaluate the performance gain given by replacing the modular structure. We tested the detection capabilities only on hand-crafted interference, manually added to the real spectra. This constitutes a clear limitation as synthetic data does not always accurately capture reality. Assuming availability of real spectrum data affected by interference, it would be relevant to assess the effectiveness of the proposed approach under these more realistic testing conditions. Moreover, although extremely challenging, it would be interesting to investigate the performance in case a given type of interference would not affect the spectrum shape in an evident manner, yet would make it similar to a legitimate modulation. Moreover, due to limited available data, only 16QAM and 8QPSK modulations were considered. However, it would be important to widen the discrete range of modulations and even expand the modulation set to a continuous set, where also slightly altered versions of the regular modulation spectra are included.

In Paper B, we addressed performance degradation due to outdated CSI when performing link adaptation. However, the delay introduced by the feedback loop used to signal back the estimated CSI is not the only cause of imperfect CSI at the transmitter. An interest research direction would therefore be to consider a

more realistic scenario and take into account also estimation errors, hardware non-linearities, low-resolution analog-to-digital converters, and carrier frequency offsets, that, together with feedback delay, contribute to impair link adaptation. Furthermore, we considered the sole objective of maximizing the spectral efficiency. In practice, however, many applications have additional QoS requirements which need to be taken into account when performing link adaptation. A relevant research direction would be to incorporate such requirements into the proposed data-driven MCS selection approach. Finally, in Paper B, the neural network training is performed off-line, over a finite and a-priori selected set of SNRs and user velocities. However, in practice, the system might face unseen scenarios. Thus, it would be interesting to envision an online learning approach to facilitate a rapid adaptation to new channel conditions, given the high variability of the wireless environment. Finally, assessing the performance on real data would be of great interest.

In Papers C, D, and E we considered the beamforming problem of sum rate maximization under a transmit power constraint at the base station. We addressed this problem under ideal conditions, i.e., we assumed perfect channel knowledge at the base station and we disregarded hardware impairments, such as power amplifier non-linearities, carrier frequency offsets, and phase noise, which in practice can severely impact the performance. Therefore, it would be relevant to replace our simplified assumptions and extend the approach to a more realistic scenario. Another relevant research direction would be the extension of our approach to RIS-aided downlink channels, which have garnered more and more attention as RIS technology promises to improve the channel propagation conditions for mobile users suffering from fading and shadowing. Specifically, it would be interesting to investigate the achieved performance when the WMMSE algorithm, which is used for the joint optimization of transmit beamformers and RIS phase shifts, is replaced with our proposed matrix-inverse-free variant. For the MU-MIMO case, an interesting research venue consists of replacing all the matrix inverses that appear in the WMMSE algorithm with a finite number of Schulz iterations instead of resorting to the gradient descent. It would be interesting to establish the convergence guarantees in this case as Schulz iterative approach considerably complicates the theoretical analysis. It would also be interesting to study more in depth the impact that the number of inner optimization steps has on the overall performance and design a strategy to select the number of inner steps per iteration in order to maximize the performance within a given complexity budget. Further, it would be significant to apply the idea of hypernetworks, namely building an auxiliary neural network whose goal is to output, for each given channel, the set of trainable parameters to use in the unfolded network. This would be an interesting solution to alleviate the need to retrain the unfolded network in case of different channel distributions. Additionally, evaluating the unfolded network on real channel data and investigating online training strategies would be relevant i) as simulated data is not always sufficiently representative of reality and ii) as the system will likely encounter scenarios not included in the training set. Also, exploring distributed training could be of interest. Finally, although we focused on the specific WMMSE beamforming

algorithm, we believe that the idea of replacing computationally heavy and hard-to-parallelize operations with matrix-inverse-free iterative first-order methods, is applicable to a wide range of algorithms that otherwise would be difficult to unfold and would not benefit from efficient parallel implementation. Therefore, an interesting research venue would be to uncover more algorithms that, like the WMMSE algorithm, could benefit from such an approach.

Part II

Included papers