



ParamNet: A Multi-Layer Parametric Network for Joint Channel Estimation and Symbol Detection

Vincent Choqueuse, Alexandru Frunza, Adel Belouchrani, Stéphane Azou,
Pascal Morel

► To cite this version:

Vincent Choqueuse, Alexandru Frunza, Adel Belouchrani, Stéphane Azou, Pascal Morel. ParamNet: A Multi-Layer Parametric Network for Joint Channel Estimation and Symbol Detection. 2022 30th European Signal Processing Conference (EUSIPCO), Aug 2022, Belgrade, Serbia. pp.1616-1620, 10.23919/EUSIPCO55093.2022.9909961 . hal-04145035

HAL Id: hal-04145035

<https://hal.science/hal-04145035v1>

Submitted on 28 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ParamNet: A Multi-Layer Parametric Network for Joint Channel Estimation and Symbol Detection

Vincent Choqueuse
Lab-STICC, UMR CNRS 6285
ENIB, Technopole Brest-Iroise
29238 Brest Cedex 3, France
choqueuse@enib.fr

Alexandru Frunza
Lab-STICC, UMR CNRS 6285
ENIB, Technopole Brest-Iroise
29238 Brest Cedex 3, France
frunza@enib.fr

Adel Belouchrani
Electrical Engineering Department
Ecole Nationale Polytechnique
Algiers, Algeria
adel.belouchrani@enp.edu.dz

Stéphane Azou
Lab-STICC, UMR CNRS 6285
ENIB, Technopole Brest-Iroise
29238 Brest Cedex 3, France
azou@enib.fr

Pascal Morel
Lab-STICC, UMR CNRS 6285
ENIB, Technopole Brest-Iroise
29238 Brest Cedex 3, France
morel@enib.fr

Abstract—This paper proposes a parametric-based network architecture for joint channel estimation and data detection in communications systems with hardware impairments. This architecture is composed of a data-augmented layer, a custom soft thresholding function, and several linear layers modeling the effect of channel effects and hardware impairments. In the proposed network, the soft thresholding function softly constrains the detected data to be within the considered constellation. The latter depends only on one parameter that is optimized during training. The benefit of the proposed approach is illustrated through a communications chain corrupted by multiple impairments and noises.

Index Terms—Communications Systems, Parametric Estimation, Machine Learning.

I. INTRODUCTION

In communications systems, the use of advanced coding schemes and modulation formats makes the performance of the system quite sensible to signal distortion occurring at the physical layer. These distortions include the effect of the propagation channel, non-ideal synchronization, IQ imbalance, and carrier impairments. In the literature, a large number of digital algorithms have been developed for hardware impairment compensation and data detection [1]–[11]. These algorithms are mainly based on parametric signal processing approaches. Recently, several studies have investigated the use of Deep Learning (DL) techniques to address some physical-layer problems [12], [13]. In particular, many studies have focused on the use of model-based DL techniques for Multiple-Input Multiple-Output (MIMO) detection [14]–[17]. The main drawback of these approaches is that they usually require perfect channel state information. To address this issue, a DL network addressing the joint channel estimation and data detection problem has been proposed in [18]. As this technique does not assume any apriori model for the channel matrix, it often requires a large training database and is not suited for time-varying channels.

In this paper, we focus on the joint channel estimation and data detection problem in frequency-selective Single-Input Single-Output (SISO) communications. By modeling the channel as a cascade of parametric models, we propose a new parametric multi-layer network for channel estimation and data detection which is well suited for time-varying channels. This network can be trained using a small number of pilot symbols. Following the principle of Iterative thresholding algorithms, we propose to include a non-linear soft thresholding layer in the structure of the parametric network. This layer is based on the optimal denoiser for Gaussian noise that has been introduced in [16].

The paper is organized as follows. Section II describes the problem statement, Section III presents the proposed network architecture, and Section IV illustrates the benefit of our approach together with the obtained simulation results.

II. PROBLEM STATEMENT

A. Signal Model

Let us consider a communications channel with N complex-valued transmitted symbols $\mathbf{s}_c = [s_c[0], \dots, s_c[N-1]]^T$ and N complex-valued received samples $\mathbf{y}_c = [y_c[0], \dots, y_c[N-1]]^T$, where T corresponds to the vector transpose. Let us denote the augmented real-valued transmitted symbols and received samples by $\mathbf{s} = [\Re(\mathbf{s}_c^T), \Im(\mathbf{s}_c^T)]^T$ and $\mathbf{y} = [\Re(\mathbf{y}_c^T), \Im(\mathbf{y}_c^T)]^T$, respectively. By assuming a linear channel, the augmented received samples can be expressed by the following linear model:

$$\mathbf{y} = \mathbf{H}(\boldsymbol{\theta})\mathbf{s} + \mathbf{b}, \quad (1)$$

where $\mathbf{H}(\boldsymbol{\theta})$ is a $2N \times 2N$ global transfer matrix, $\boldsymbol{\theta}$ corresponds to the unknown channel parameters, and $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma_b^2 \mathbf{I})$ is an additive noise component with zero mean and variance σ_b^2 . In communications systems, the transmitted data usually belong to a finite alphabet. This property implies that each augmented symbol $s[n]$ belongs to a set

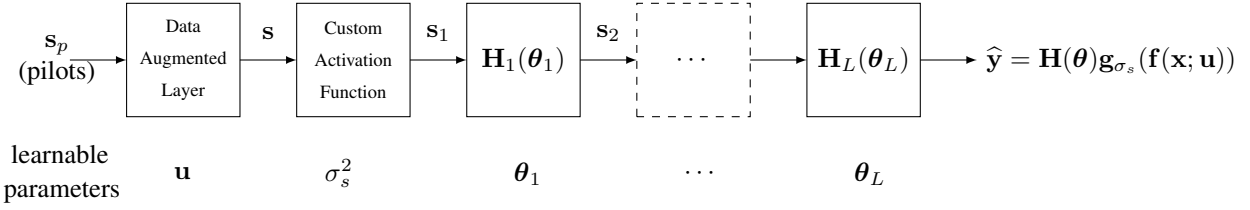


Fig. 1: Proposed ParamNet network architecture. The network is composed of a data augmented layer, $\mathbf{f}(\cdot; \mathbf{u})$, a non-linear activation function, $\mathbf{g}_{\sigma_s}(\cdot)$, and L linear layers $\mathbf{H}_l(\theta_l)$ modeling the hardware impairments and channel effects.

$\mathcal{M} = \{M_1, \dots, M_K\}$ composed of K real-valued symbols. Finally, we consider in this study that the global transfer matrix can be decomposed as a multi-layer channel as follows

$$\mathbf{H}(\theta) = \mathbf{H}_L(\theta_L) \times \dots \times \mathbf{H}_1(\theta_1), \quad (2)$$

where $\mathbf{H}_l(\theta_l)$ corresponds to the transfer matrix of the l^{th} layer and depends on the unknown parameters θ_l . Note that this multi-layer model arises naturally in communications systems corrupted by multiple linear impairments or effects such as transmitter or receiver IQ imbalance, Carrier Frequency Offset, Phase Noise, frequency selective channels, *etc.*

B. Joint Estimation and Detection Problem

The objective of this paper is to jointly estimate θ and $\mathbf{s} \in \mathcal{M}^{2N}$ from \mathbf{y} . For Gaussian noise, this optimization problem can be formulated as $\min_{\theta, \mathbf{s} \in \mathcal{M}^{2N}} \|\mathbf{y} - \mathbf{H}(\theta)\mathbf{s}\|_2^2$. This problem is challenging since θ is unknown and \mathbf{s} belongs to a discrete alphabet. To simplify this optimization problem, a solution is to consider that N_x complex-valued pilot symbols, \mathbf{s}_c , are known at the receiver. By exploiting the knowledge of $\mathbf{s}_p = [\Re(\mathbf{s}_c^T), \Im(\mathbf{s}_c^T)]^T$, the optimization problem can be reformulated as

$$\min_{\theta, \mathbf{u} \in \mathcal{M}^{2N_u}} \|\mathbf{y} - \mathbf{H}(\theta)\mathbf{f}(\mathbf{s}_p; \mathbf{u})\|_2^2, \quad (3)$$

where \mathbf{u} is an unknown vector composed of $2N_u$ data symbols, and $\mathbf{s} = \mathbf{f}(\mathbf{s}_p; \mathbf{u})$ is a data-augmented function that takes $2N_x$ real-valued symbols and returns $2N = 2(N_x + N_u)$ real-valued symbols.

To estimate θ and \mathbf{u} , two approaches can be employed. In the first approach, the alphabet of \mathbf{u} can be simply neglected by assuming $\mathbf{u} \in \mathbb{R}^{2N_u}$. In the second approach, the alphabet can be enforced using a projected gradient algorithm by mapping periodically \mathbf{u} to the closest point in \mathcal{M}^{2N_u} . Even if this second approach has the distinct advantage of exploiting the data constellation, the influence of projection errors in the medium-SNR region cannot be completely excluded. Furthermore, during training, the use of a hard-thresholding function makes it impossible to backpropagate the gradient to the data augmented layer since the function derivative is 0 almost everywhere. To address these two issues, we propose a new network architecture composed of a soft projection layer.

III. PROPOSED MULTI-LAYER NETWORK

The structure of the proposed Multi-Layer parametric network is presented in Fig. 1. This section describes the mathematical models of each layer and presents the associated learning algorithm.

A. Layer Models

In this section, the real-valued inputs and outputs of each layer are denoted \mathbf{x}_i and \mathbf{x}_o , respectively.

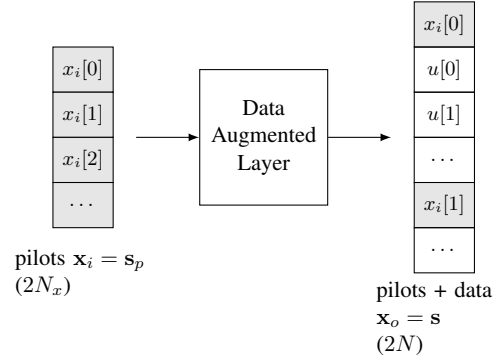


Fig. 2: Illustration of the Data Augmented Layer. The learnable parameters correspond to the data $\mathbf{u} = [u[0], u[1], \dots]^T$.

1) *Data Augmented Layer*: The purpose of the data augmented layer is to insert some data \mathbf{u} into the input vector \mathbf{x}_i (see Fig. 2). Mathematically, the input-output relationship of this layer can be expressed by

$$\mathbf{x}_o = \mathbf{f}(\mathbf{x}_i; \mathbf{u}) = \mathbf{P}_0 \mathbf{x}_i + \mathbf{P}_1 \mathbf{u}, \quad (4)$$

where the layer trainable parameters correspond to the unknown real-valued vector \mathbf{u} . The matrices \mathbf{P}_0 and \mathbf{P}_1 correspond to *allocation matrices* of size $2N \times 2N_x$ and $2N \times 2N_u$, respectively. An allocation matrix is a matrix with entries from $\{0, 1\}$ that only contains a single one in each row.

2) *Non-linear Activation Function*: The non-linear activation layer is based on the soft thresholding operator

$$\mathbf{x}_o = \mathcal{S}_{\sigma_s}(\mathbf{x}_i). \quad (5)$$

The purpose of the function $\mathcal{S}_{\sigma_s}(\cdot)$ is to mimic a soft projection into the data constellation. To this end, we propose to use the optimal denoiser for Gaussian noise, which has been recently used in MIMO detection problems [16]. By assuming an i.i.d.

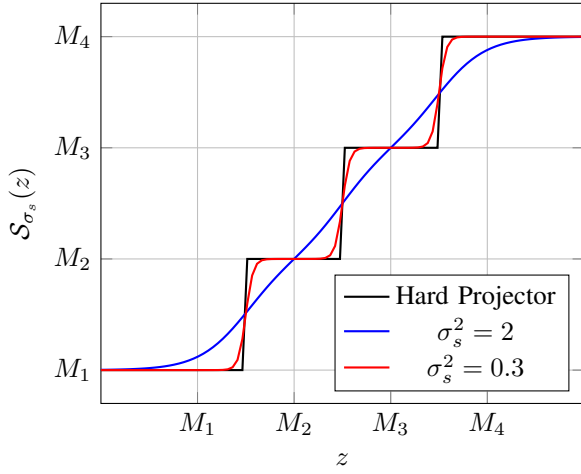


Fig. 3: Soft projection into $\mathcal{M} = \{M_1, M_2, M_3, M_4\}$.

Gaussian distribution with diagonal covariance matrix $(\sigma_s^2/2)\mathbf{I}$ for the error $\mathbf{x}_i - \mathbf{s}$ and an uniform distribution over \mathcal{M} for the data symbols \mathbf{s} , the n^{th} output of the optimal denoiser can be expressed as

$$[\mathcal{S}_{\sigma_s}(\mathbf{x}_i)]_n = \frac{\sum_{s \in \mathcal{M}} s e^{-\frac{1}{\sigma_s^2}(x_i[n]-s)^2}}{\sum_{s \in \mathcal{M}} e^{-\frac{1}{\sigma_s^2}(x_i[n]-s)^2}} \quad (6)$$

This layer contains only one learnable parameter called the *data noise variance* σ_s^2 .

For illustration purpose, Fig. 3 depicts the function $\mathcal{S}_{\sigma_s}(\cdot)$ for different values of σ_s^2 . Note that the data noise variance σ_s^2 is different from the additive noise variance σ_b^2 in (1). The data noise variance σ_s^2 allows to alternate between the exploration and exploitation of the data vector \mathbf{u} during training. For example, the exploration is encouraged when $\sigma_s^2 = 2$ since $\mathcal{S}_{\sigma_s}(\cdot)$ is nearly linear between M_1 and M_4 , while the exploitation is encouraged when $\sigma_s^2 = 0.3$ since this function is close to a threshold detector.

3) *Parametric SISO Channel Layers*: This section presents a list of linear layers commonly encountered in communications systems. While this list is not exhaustive and only focuses on SISO layers, it can be easily adapted to more specific problems or MIMO systems. For SISO layers, the augmented outputs can be expressed as $\mathbf{x}_o = \mathbf{H}_l(\boldsymbol{\theta}_l)\mathbf{x}_i$, where $\mathbf{x}_i = [\Re(\mathbf{x}_{c,i}^T), \Im(\mathbf{x}_{c,i}^T)]^T$ depends on the vector $\mathbf{x}_{c,i} = [x_{c,i}[0], \dots, x_{c,i}[N-1]]^T$. Herein, the mathematical models are expressed in scalar form for sake of simplicity.

- **IQ Imbalance Layer**: The output of an IQ imbalance layer is given by

$$x_{c,o}[n] = \mu x_{c,i}[n] + \nu x_{c,i}^*[n], \quad (7)$$

where $(\cdot)^*$ corresponds to the complex conjugate, and $(\mu, \nu) \in \mathbb{C}^2$ are the IQ imbalance parameters [4]. This layer depends on the 4 real-valued parameters $\boldsymbol{\theta}_l = [\Re(\mu), \Re(\nu), \Im(\mu), \Im(\nu)]^T$.

- **FIR Layer**: The output of a FIR layer is given by

$$x_{c,o}[n] = \sum_{k=0}^{D-1} h[k]x_{c,i}[n-k]. \quad (8)$$

This layer depends on the real-valued vector $\boldsymbol{\theta}_l = [\Re(\mathbf{h}^T), \Im(\mathbf{h}^T)]^T$ where $\mathbf{h} = [h[0], \dots, h[D-1]]^T$ corresponds to the channel impulse response.

- **Piecewise-Constant Phase Layer**: The output of this layer is given by

$$x_{c,o}[n] = e^{j\varphi[n \bmod N_s]} x_{c,i}[n] \quad (9)$$

where *mod* corresponds to the modulo operator [19]. This layer depends on N/N_s real-valued parameters that are given by $\boldsymbol{\theta}_l = [\varphi[0], \dots, \varphi[L-1]]^T$.

B. Network training/operating

To train the network, we propose to minimize the custom regularized loss function given by

$$\min_{\boldsymbol{\theta}, \mathbf{u}, \sigma_s^2} \frac{1}{2N} \|\mathbf{y} - \mathbf{H}(\boldsymbol{\theta})\mathcal{S}_{\sigma_s}(\mathbf{f}(\mathbf{x}; \mathbf{u}))\|_2^2 + \lambda \sigma_s^2, \quad (10)$$

where the first term corresponds to the Mean Squared Error (MSE), which gives an estimate of the *observed noise variance*, and where the second term is a regularization term that depends on the *data noise variance*. Note that the network training using the pilot symbols available at the receiver, \mathbf{x} , leads to the joint estimation of the channel parameters and symbol detection through the obtained vector $\hat{\mathbf{u}}$.

The regularization parameter λ plays an important role to prevent the overfitting problem. Indeed, for small number of pilot samples \mathbf{x} , only minimizing the estimated observed noise variance tends to spread the additive noise into the vector \mathbf{u} . This noise contribution could have a negative effect on the Symbol Error Rate (SER), which is our main objective. Increasing $\lambda > 0$ limits this phenomenon by encouraging a small *data noise variance*.

IV. SIMULATION RESULTS

This section illustrates the advantages of our approach using a SISO communications system composed of 5 layers: a transmitter IQ imbalance, a FIR channel with $\mathbf{h} = [0.9 + 0.1j, 0.1 + 0.1j, 0.01 + 0.05j, 0.02 - 0.003j, 0.004 + 0.012j]$, a receiver IQ imbalance, a time-varying Wiener phase noise with instantaneous phase $\varphi[n] = \sum_{k=0}^n \psi[k]$ where $\psi[n] \sim \mathcal{N}(0, \sigma_p^2)$, and a Gaussian additive noise $b[n] \sim \mathcal{N}(0, \frac{\sigma_b^2}{2})$. The simulation parameters are provided in Table I. The transmitted signal \mathbf{s} is composed of $N = 200$ symbols generated from a 16QAM constellation.

To jointly estimate $\boldsymbol{\theta}$ and detect \mathbf{u} , we consider a multi-layer parametric network composed of 6 layers: a data-augmented layer with $N_u = 180$ complex-valued data symbols, a soft projector with initial parameter $\sigma_s^2 = 1$, a transmitter IQ imbalance layer initialized with $\boldsymbol{\theta} = [1, 0, 0, 0]$, a FIR layer initialized with $\mathbf{h} = [1, 0, 0, 0, 0]$, a receiver IQ imbalance layer initialized with $\boldsymbol{\theta} = [1, 0, 0, 0]$, and a Piecewise-Constant Phase Layer initialized with $N/N_s = 20$ zeros. Therefore, the

Layer	Type	Parameters
1	IQ Tx	$(\mu, \nu) = (0.9 - 0.4j, 0.4 + 0.1j)$
2	FIR Channel	\mathbf{h}
3	IQ Rx	$(\mu, \nu) = (1.8 + 0.13j, 0.1 + 0.2j)$
4	Phase Noise	$\sigma_p^2 = 0.000125$
5	Noise	σ_b^2

TABLE I: Communication Chain Parameters.

Method	Activation Function	N_{it}
Simple	\times	\times
PG_500	\times	500 iterations
PG_1000	\times	1000 iterations
PG_2000	\times	2000 iterations
Proposed	soft thresholding	\times

TABLE II: Considered Techniques

network is composed of $1 + 2 \times 180 + 4 + 2 \times 5 + 4 + 20 = 399$ real-valued trainable parameters. The proposed network has been implemented using PyTorch¹. The input vector \mathbf{x} is composed of $N_x = 20$ complex-valued symbols transmitted periodically (one pilot symbol followed by 9 unknown data symbols). The loss function is given by (10) and the regularization parameter is set to $\lambda = 0.001$. In each experiment, the training stage uses the ADAM optimizer with a learning rate of 10^{-3} [20], and the optimization is stopped after $N_{it} = 10000$ iterations. The performance of the network is evaluated by computing the SER of the complex-valued estimated symbols $\hat{u}_c[n] = \hat{u}[n] + j\hat{u}[n + N_u]$, where $\hat{u}[n]$ corresponds to the n^{th} trainable parameter of the first layer after training.

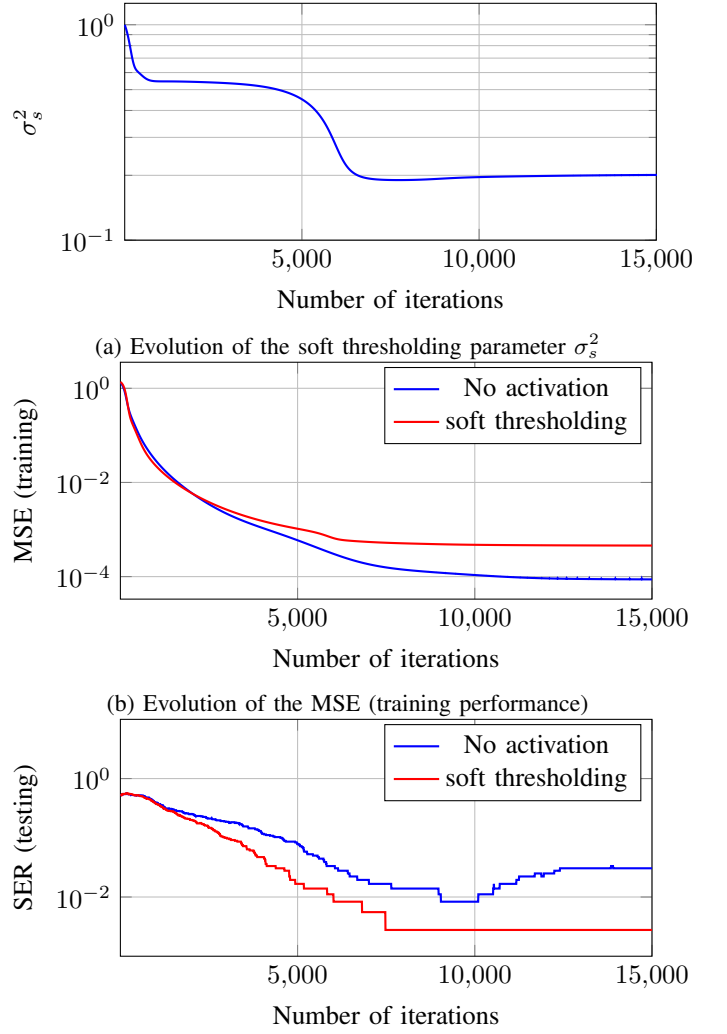
A. Single Trial Analysis

Fig. 4 shows the evolution of σ_s^2 , the MSE and the SER during the training stage. The performance of the proposed technique is compared with the performance obtained with the same parametric network without activation layer. While the MSE is clearly lower for the simple parametric network, the SER shows that this network strongly overfits the data due to the small number of pilots. After 20000 iterations, the SER is equal to 0.0361 for the parametric network, while the proposed network leads to a SER equal to 0.0027.

B. Monte Carlo Simulations

Fig. 5 presents the average SER versus SNR using Monte Carlo simulations. In each Monte Carlo trial, we generate a new realization of the Wiener phase and additive noises. The performance of the proposed technique is compared with the performances of the techniques reported in Table II. The Projected Gradient (PG) techniques are obtained by periodically projecting the trainable parameters \mathbf{u} into the set \mathcal{M} after N_{it} iterations [14]. As the projection rate has a significant impact on the SER performance, we have considered several PG detectors with different values of N_{it} . We observe that the simple technique requires a large SNR to perform well. Furthermore, we also note that the performance of the PG techniques critically depends on the projection rate and SNR.

¹The source code is available at <https://github.com/vincentchoqueuse/ParamNET>



(c) Evolution of the Symbol Error Rate (SER) (testing performance)

Fig. 4: Evolution of σ and SER performance (SNR=20dB).

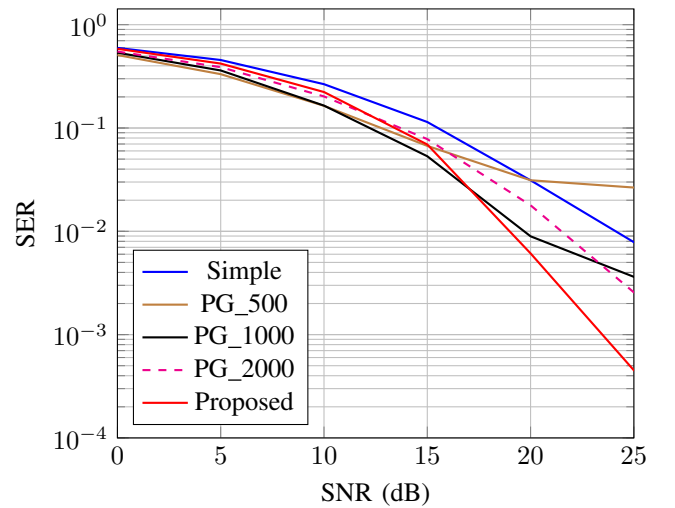


Fig. 5: Testing performance versus SNR.

In particular, projecting frequently \mathbf{u} into the set \mathcal{M} leads to poor performance for medium or high SNR. Regarding the proposed method, it leads to the best performance at medium and high SNR.

V. CONCLUSION

This paper proposes a new network architecture called ParamNet for joint estimation and data detection in communications systems. ParamNet is composed of a data augmented layer and a soft thresholding layer that encourages the data constellation. This network can be trained using a small number of pilot symbols and a custom regularized loss function. Simulation results have shown that this strategy outperforms projected gradient techniques. In particular, the proposed strategy provides a solution to avoid the overfitting problem in the medium and high SNR regions. Future works will investigate the application of the proposed network in more complex communications chains, including MIMO systems.

REFERENCES

- [1] J.-J. Van de Beek, M. Sandell, and P. O. Borjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Transactions on Signal Processing*, vol. 45, no. 7, pp. 1800–1805, 1997.
- [2] J. K. Tugnait, L. Tong *et al.*, "Single-user channel estimation and equalization," *IEEE Signal Processing Magazine*, vol. 17, no. 3, pp. 17–28, 2000.
- [3] M. Valkama, M. Renfors, and V. Koivunen, "Advanced methods for I/Q imbalance compensation in communication receivers," *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2335–2344, 2001.
- [4] A. Tarighat, R. Bagheri, and A. H. Sayed, "Compensation schemes and performance analysis of IQ imbalances in OFDM receivers," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3257–3268, 2005.
- [5] J. Tubbax, B. Come, L. Van der Perre, S. Donnay, M. Engels, H. De Man, and M. Moonen, "Compensation of IQ imbalance and phase noise in OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 872–877, 2005.
- [6] L. Anttila, M. Valkama, and M. Renfors, "Circularity-based I/Q imbalance compensation in wideband direct-conversion receivers," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2099–2113, 2008.
- [7] A. Spalvieri and L. Barletta, "Pilot-aided carrier recovery in the presence of phase noise," *IEEE Transactions on Communications*, vol. 59, no. 7, pp. 1966–1974, 2011.
- [8] H. Mehrpouyan, A. A. Nasir, S. D. Blostein, T. Eriksson, G. K. Karagiannidis, and T. Svensson, "Joint estimation of channel and oscillator phase noise in MIMO systems," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4790–4807, 2012.
- [9] O. H. Salim, A. A. Nasir, H. Mehrpouyan, W. Xiang, S. Durrani, and R. A. Kennedy, "Channel, phase noise, and frequency offset in ofdm systems: Joint estimation, data detection, and hybrid cramer-rao lower bound," *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3311–3325, 2014.
- [10] A. Leshem and M. Yemini, "Phase noise compensation for ofdm systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 21, pp. 5675–5686, 2017.
- [11] A. Frunza, V. Choqueuse, P. Morel, and S. Azou, "Global estimation and compensation of linear effects in coherent optical systems based on nonlinear least squares," *IEEE Systems Journal*, pp. 1–11, 2021.
- [12] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, ai-based, or both?" *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7331–7376, 2019.
- [13] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, 2017.
- [14] S. Takabe, M. Imanishi, T. Wadayama, R. Hayakawa, and K. Hayashi, "Trainable projected gradient detector for massive overloaded MIMO channels: Data-driven tuning approach," *IEEE Access*, vol. 7, pp. 93 326–93 338, 2019.
- [15] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2554–2564, 2019.
- [16] M. Khani, M. Alizadeh, J. Hoydis, and P. Fleming, "Adaptive neural signal detection for massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5635–5648, 2020.
- [17] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *arXiv preprint arXiv:2012.08405*, 2020.
- [18] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for MIMO detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1702–1715, 2020.
- [19] G. Durisi, "On the capacity of the block-memoryless phase-noise channel," *IEEE Communications Letters*, vol. 16, no. 8, pp. 1157–1160, 2012.
- [20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.