

Online unsupervised deep unfolding for massive MIMO channel estimation

Luc Le Magoarou, Stéphane Paquelet

Abstract—Massive MIMO communication systems have a huge potential both in terms of data rate and energy efficiency, although channel estimation becomes challenging for a large number of antennas. Using a physical model allows to ease the problem by injecting a priori information based on the physics of propagation. However, such a model rests on simplifying assumptions and requires to know precisely the configuration of the system, which is unrealistic in practice. In this letter, we propose to perform online learning for channel estimation in a massive MIMO context, adding flexibility to physical channel models by unfolding a channel estimation algorithm (matching pursuit) as a neural network. This leads to a computationally efficient neural network structure that can be trained online when initialized with an imperfect model. The method allows a base station to automatically correct its channel estimation algorithm based on incoming data, without the need for a separate offline training phase. It is applied to realistic millimeter wave channels and shows great performance, achieving a channel estimation error almost as low as one would get with a perfectly calibrated system.

Index Terms—Autoencoders, deep unfolding, MIMO channel estimation, online learning.

I. INTRODUCTION

DATA processing techniques are often based on the *manifold assumption*: Meaningful data (signals) lie near a low dimensional manifold, although their apparent dimension is much larger [1], [2] [3, Section 5.11.3] [4, Section 9.3].

In MIMO channel estimation, using a physical model amounts to parameterize a manifold by physical parameters such as the directions, delays and gains of the propagation paths, the dimension of the manifold being equal to the number of real parameters considered in the model. Physical models allow to inject strong a priori knowledge based on solid principles [5], [6], [7], but necessarily make simplifying assumptions (e.g. the plane wave assumption [8]) and require to know exactly the system configuration (positions of the antennas, gains, etc.).

On the other hand, machine learning techniques have recently led to tremendous successes in various domains [9], [3]. Their main feature is to learn the data representation (manifold) directly on training data, without requiring any specific a priori knowledge. This flexibility in the manifold construction comes at the price of computationally heavy learning and difficulties to inject a priori knowledge on the problem at hand.

Recently, it has been proposed to unfold iterative inference algorithms so as to express them as neural networks that can be optimized [10], [11], [12]. This has the advantage of adding flexibility to algorithms based on classical models, and

amounts to constrain the search for the appropriate manifold with a priori knowledge on the problem at hand. Moreover, this leads to inference algorithms of controlled complexity [13].

Contributions. In this letter, we propose to perform *online learning* for channel estimation in a massive MIMO context. Starting from an imperfect physical channel model, our method allows a base station to automatically correct its channel estimation algorithm based on incoming data, without the need for a separate offline training phase. It is based on the unfolding of the matching pursuit algorithm, which is simple and computationally efficient. The obtained neural network is trained in an *unsupervised* way. The overall complexity of the forward and backward passes in the network is of the same order as performing channel estimation only (without any learning), which makes online learning feasible. Such a method is particularly suited to imperfectly known or non-calibrated systems.

Related work. Machine learning holds promise for wireless communications (see [14], [15] for exhaustive surveys). It has recently been proposed to use adaptive data representations for MIMO channel estimation using dictionary learning techniques [16]. However, dictionary learning with algorithms such as K-SVD [17] as proposed in [16] is very computationally heavy, and thus not suited to online learning.

Deep unfolding has also been considered by communication researchers (see [18] and references therein). It has been proposed in [19] to perform channel estimation in a massive MIMO context, based on the unfolding of a sparse recovery algorithm (namely denoising-based approximate message passing [20]). However, the method is directly adapted from image processing [21] and does not make use of a physical channel model as initialization. A recent work also proposes to use deep unfolding for channel estimation [22], but using a physical model to optimize the shrinkage function. However, previously proposed methods based on unfolding all require an offline training phase and are of high complexity compared to classical methods [23].

The main novelty of this letter is the online nature of the method, that does not require a separate offline learning phase, since learning is done while using the channel estimation algorithm that corrects itself over time. This is made possible by the very low complexity of the considered estimation algorithm.

II. PROBLEM FORMULATION

A. System settings

We consider in this letter a massive MIMO system, also known as multi-user MIMO (MU-MIMO) system [24], [25], [26], in which a base station equipped with N antennas

Luc Le Magoarou and Stéphane Paquelet are both with bcom, Rennes, France. Contact addresses: luc.lemagoarou@b-com.com, stephane.paquelet@b-com.com.

communicates with K single antenna users ($K < N$). The system operates in time division duplex (TDD) mode, so that channel reciprocity holds and the channel is estimated in the uplink: each user sends a pilot sequence \mathbf{p}_k (orthogonal to the sequences of the other users, $\mathbf{p}_k^H \mathbf{p}_l = \delta_{kl}$) for the base station to estimate the channel. The received signal is thus expressed $\mathbf{R} = \sum_{k=1}^K \mathbf{h}_k \mathbf{p}_k^H + \mathbf{N}$, where \mathbf{N} is noise. After correlating the received signal with the pilot sequences, and assuming no pilot contamination from adjacent cells for simplicity, the base station gets noisy measurements of the channels of all users, each taking the canonical form

$$\mathbf{x} = \mathbf{h} + \mathbf{n}, \quad (1)$$

where \mathbf{h} is the channel of the considered user and \mathbf{n} is the noise, with $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_d)$. We drop the user index k here and in the following, since our approach treats the channels of all users the same way. Note that \mathbf{x} is already an unbiased estimator of the channel, and we call it the least squares (LS) estimator in the sequel. Its performance can be assessed by the signal to noise ratio (SNR)

$$\text{SNR}_{\text{in}} \triangleq \frac{\|\mathbf{h}\|_2^2}{N\sigma^2}.$$

However, one can get better channel estimates using a physical model, as is explained in the next subsection.

B. Physical model

Let us denote $\{g_1, \dots, g_N\}$ the complex gains of the base stations antennas and $\{\vec{a}_1, \dots, \vec{a}_N\}$ their positions with respect to the centroid of the antenna array. Then, under the plane wave assumption and assuming omnidirectional antennas (isotropic radiation patterns), the channel resulting from a single propagation path with direction of arrival (DoA) \vec{u} is proportional to the *steering vector*

$$\mathbf{e}(\vec{u}) \triangleq (g_1 e^{-j\frac{2\pi}{\lambda} \vec{a}_1 \cdot \vec{u}}, \dots, g_N e^{-j\frac{2\pi}{\lambda} \vec{a}_N \cdot \vec{u}})^T$$

which reads $\mathbf{h} = \beta \mathbf{e}(\vec{u})$, with $\beta \in \mathbb{C}$. In that case, a sensible estimation strategy [5], [6], [7] is to build a dictionary of steering vectors corresponding to A potential DoAs: $\mathbf{E} \triangleq (\mathbf{e}(\vec{u}_1), \dots, \mathbf{e}(\vec{u}_A))$ and to compute a channel estimate with the procedure

$$\begin{aligned} \vec{v} &= \underset{\vec{u}_i}{\text{argmax}} |\mathbf{e}(\vec{u}_i)^H \mathbf{x}|, \\ \hat{\mathbf{h}} &= \mathbf{e}(\vec{v}) \mathbf{e}(\vec{v})^H \mathbf{x}. \end{aligned} \quad (2)$$

The first step of this procedure amounts to find the column of the dictionary the most correlated with the observation to estimate the DoA and the second step amounts to project the observation on the corresponding steering vector. The SNR at the output of this procedure reads

$$\text{SNR}_{\text{out}} \triangleq \frac{\|\mathbf{h}\|_2^2}{\mathbb{E}[\|\mathbf{h} - \hat{\mathbf{h}}\|_2^2]},$$

and we have at best $\text{SNR}_{\text{out}} = N \text{SNR}_{\text{in}}$ (neglecting the discretization error).

Note that the evoked strategy can be generalized to multipath channels of the form $\mathbf{h} = \sum_{p=1}^P \beta_p \mathbf{e}(\vec{u}_p)$, using greedy sparse recovery algorithms such as matching pursuit (MP) or orthogonal matching pursuit (OMP) [27], [28].

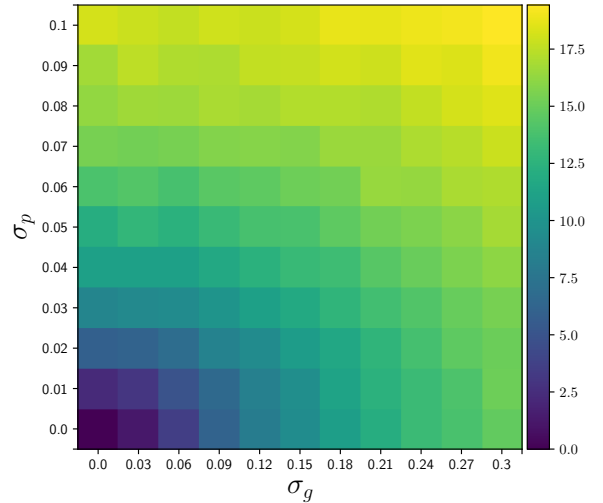


Fig. 1: SNR loss in decibels (dB) due to imperfect knowledge of the system.

III. IMPACT OF IMPERFECT MODELS

The estimation strategy based on a physical model requires knowing the system configuration (antenna gains and positions) and necessarily relies on hypotheses. What happens if the configuration is imperfectly known or if some hypotheses are not valid? In order to answer this simple question, let us perform an experiment. Consider an antenna array of $N = 64$ antennas at the base station, whose known *nominal* configuration is an uniform linear array (ULA) of unit gain antennas separated by half-wavelengths and aligned with the x -axis. This nominal configuration corresponds to gains and positions $\{\tilde{g}_i, \vec{a}_i\}_{i=1}^N$. Now, suppose the knowledge of the system configuration is imperfect, meaning that the unknown *true* configuration of the system is given by the gains and positions $\{g_i, \vec{a}_i\}_{i=1}^N$, with

$$\begin{aligned} g_i &= \tilde{g}_i + n_{g,i}, \quad n_{g,i} \sim \mathcal{CN}(0, \sigma_g^2), \\ \vec{a}_i &= \vec{a}_i + \lambda \mathbf{n}_{p,i}, \quad \mathbf{n}_{p,i} = (e_{p,i}, 0, 0)^T, \quad e_{p,i} \sim \mathcal{N}(0, \sigma_p^2). \end{aligned} \quad (3)$$

This way, σ_g (resp. σ_p) quantifies the uncertainty about the antenna gains (resp. spacings). Moreover, let

$$\tilde{\mathbf{e}}(\vec{u}) \triangleq (\tilde{g}_1 e^{-j\frac{2\pi}{\lambda} \vec{a}_1 \cdot \vec{u}}, \dots, \tilde{g}_N e^{-j\frac{2\pi}{\lambda} \vec{a}_N \cdot \vec{u}})^T$$

be the nominal steering vector and $\tilde{\mathbf{E}} \triangleq (\tilde{\mathbf{e}}(\vec{u}_1), \dots, \tilde{\mathbf{e}}(\vec{u}_A))$ be a dictionary of nominal steering vectors. The experiment consists in comparing the estimation strategy of (2) using the true (perfect but unknown) dictionary \mathbf{E} with the exact same strategy using the nominal (imperfect but known) dictionary $\tilde{\mathbf{E}}$. To do so, we generate measurements according to (1) with channels of the form $\mathbf{h} = \mathbf{e}(\vec{u})$ where \vec{u} corresponds to azimuth angles chosen uniformly at random, and SNR_{in} is set to 10 dB. Then, the dictionaries \mathbf{E} and $\tilde{\mathbf{E}}$ are built by choosing $A = 32N$ directions corresponding to evenly spaced azimuth angles. Let $\hat{\mathbf{h}}_{\mathbf{E}}$ be the estimate obtained using \mathbf{E} in (2), and $\hat{\mathbf{h}}_{\tilde{\mathbf{E}}}$ the estimate obtained using $\tilde{\mathbf{E}}$. The SNR loss caused by using $\tilde{\mathbf{E}}$ instead of \mathbf{E} is measured by the quantity $\|\hat{\mathbf{h}}_{\tilde{\mathbf{E}}} - \mathbf{h}\|_2^2 / \|\hat{\mathbf{h}}_{\mathbf{E}} - \mathbf{h}\|_2^2$. Results in terms of SNR loss, in average over 10 antenna array realizations and 1000 channel realizations per antenna

array realization are shown on figure 1. From the figure, it is obvious that even a relatively small uncertainty about the system configuration can cause a great SNR loss. For example, an uncertainty of 0.03λ on the antenna spacings and of 0.09 on the antenna gains leads to an SNR loss of more than 10 dB, which means that the mean squared error is increased more than ten times. This experiment highlights the fact that using imperfect models can severely harm estimation performance. The main contribution of this letter is to propose a way to correct imperfect physical models using machine learning.

IV. ONLINE DEEP UNFOLDING FOR CHANNEL ESTIMATION

Let us now propose a strategy based on deep unfolding allowing to correct a channel estimation algorithm based on an imperfect physical model incrementally, via online learning.

A. Basic principle

Unfolding. The estimation strategy of (2) can be unfolded as a neural network taking the observation \mathbf{x} as input and outputting a channel estimate $\hat{\mathbf{h}}$. Indeed, the first step amounts to perform a linear transformation (multiplying the input by the matrix \mathbf{E}^H) followed by a nonlinear one (finding the inner product of maximum amplitude and setting all the others to zero) and the second step corresponds to a linear transformation (multiplying by the matrix \mathbf{E}). Such a strategy is parameterized by the dictionary of steering vectors \mathbf{E} . In the case where the dictionary \mathbf{E} is unknown (or imperfectly known), we propose to learn the matrix used in (2) directly on data via backpropagation [29], using as initialization the matrix $\tilde{\mathbf{E}}$ corresponding to the imperfect physical model.

Neural network structure. Such a neural network structure corresponds to the k -sparse autoencoder [30], which has originally been introduced for image classification. The deep unfolding of channel estimation using a physical model as in (2) corresponds to use a k -sparse autoencoder, setting the sparsity parameter to $k = 1$. This neural network structure is shown on figure 2, where HT_1 refers to the hard thresholding operator which keeps only the entry of greatest modulus of its input and sets all the others to zero. The parameters of this neural network are the weights $\mathbf{W} \in \mathbb{C}^{N \times A}$. Note that complex weights and inputs are handled classically by stacking the real and imaginary parts for vectors and using the real representation for matrices.

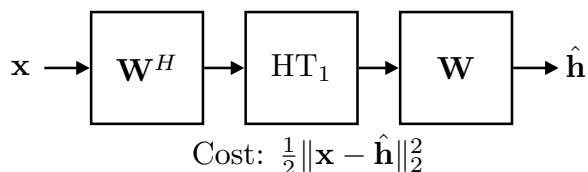


Fig. 2: Deep unfolding for single path channel estimation.

Training. The method we propose to jointly estimate channels while simultaneously correcting an imperfect physical model amounts to initialize the network of figure 2 with a dictionary of nominal steering vectors $\tilde{\mathbf{E}}$ and then to perform a minibatch gradient descent [31] on the cost function $\frac{1}{2} \|\mathbf{x} - \hat{\mathbf{h}}\|_2^2$ to update the weights \mathbf{W} in order to correct the model. It

operates *online*, on streaming observations $\mathbf{x}_i, i = 1, \dots, \infty$ of the form (1) acquired over time (coming from all users simultaneously). Note that, as opposed to the classical unfolding strategies [10], [11], [12], the proposed method is totally *unsupervised*, meaning that it requires only noisy channel observations and no clean channels to run.

Implementation details. In all the experiments performed in this letter, we use minibatches of 200 observations and Nesterov accelerated gradient descent [32], [33] with exponentially decreasing learning rates starting at 1 and being multiplied by 0.9 every 200 minibatches and momentum coefficient starting at 0.9 and being divided by 2 every 200 minibatches. Moreover, the method was found to perform better with the input data being normalized and expressed in the Fourier domain. To simplify notation, we denote abusively also $\mathbf{x}_i, i = 1, \dots, \infty$ the data after normalization and Fourier transform and $\mathbf{F} \in \mathbb{C}^{N \times N}$ the discrete Fourier transform (DFT) matrix. The initial weights are thus actually set to $\mathbf{F}\tilde{\mathbf{E}}$ to take the Fourier transform into account.

B. Generalization to multipath channels

Real channels are often not made of a single path, in which case the proposed method becomes suboptimal. Indeed, it uses a k -sparse autoencoder with $k = 1$, implicitly assuming a single path. However, real world channels are often sparse (well approximated by only a few paths). This is particularly true at millimeter wave frequencies [34]. In order to adapt the unfolding strategy to such channels, we propose to apply recursively the structure of figure 2, subtracting at each step the current output from the observation, exactly mimicking the matching pursuit (MP) algorithm [27]. The number K of times the structure is replicated (depth of the network) corresponds to the number of estimated paths. The neural network corresponding to such a strategy is schematized on figure 3, we call it **mpNet** (for matching pursuit network). It is trained exactly as the network of figure 2, with tied weights across iterations (we tried to untie the weights but observed no improvement) and cost function $\frac{1}{2} \|\mathbf{x} - \hat{\mathbf{h}}\|_2^2 = \frac{1}{2} \|\mathbf{r}_K\|_2^2$.

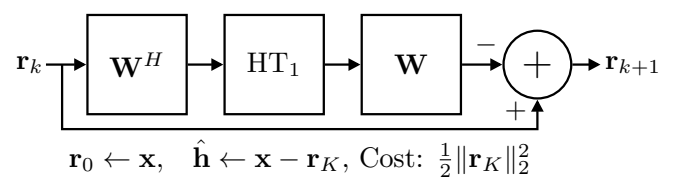


Fig. 3: mpNet: Unfolding for multipath channel estimation.

Computational complexity. Note that the forward pass in mpNet costs $\mathcal{O}(KNA)$ arithmetic operations and the backpropagation step costs $\mathcal{O}(KN)$ arithmetic operations (A times less). This means that jointly learning the model and estimating the channel (computing the forward and backward passes) is done at a cost that is the same order as the one of simply estimating the channel with a greedy algorithm (MP or OMP) without adapting the model at all to data (which corresponds to computing only the forward pass). This very light computational cost makes the method adapted to online learning, as opposed to previously proposed channel estimation strategies based on deep unfolding [19], [22], [23].

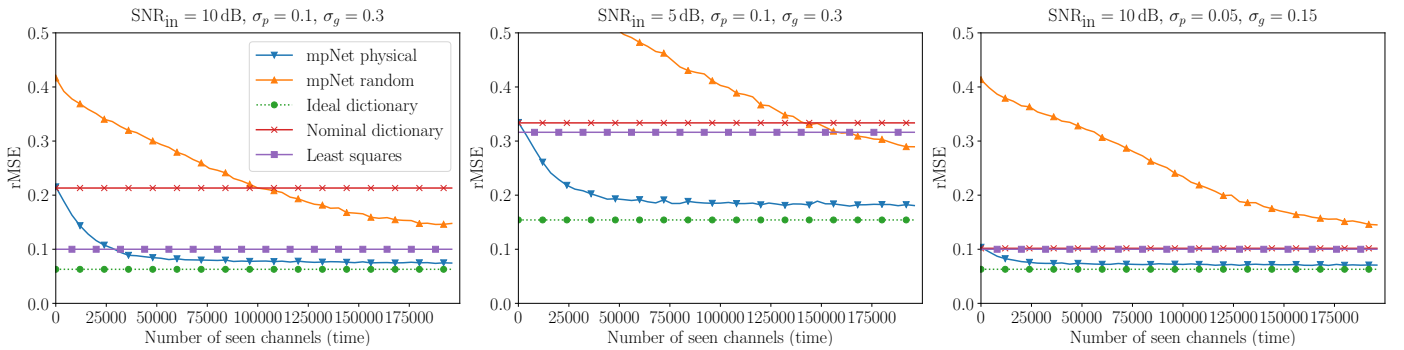


Fig. 4: Channel estimation performance on synthetic realistic channels for various SNRs and model imperfections.

C. Experiment

Setting. Let us now assess mpNet on realistic channels. To do so, we consider the SSCM channel model [34] in order to generate non-line-of-sight (NLOS) channels at 28 GHz (see [34, table IV]) corresponding to all users. We consider the same setting as in section III, namely a base station equipped with an ULA of 64 antennas, with an half-wavelength nominal spacing and unit nominal gains used to build the imperfect nominal dictionary $\tilde{\mathbf{E}}$ (with $A = 8N$) which serves as an initialization for mpNet. The actual antenna arrays are generated the same way as in section III, using (3), and are kept fixed for the whole experiment. We consider two model imperfections: $\sigma_p = 0.05$, $\sigma_g = 0.15$ (small uncertainty) and $\sigma_p = 0.1$, $\sigma_g = 0.3$ (large uncertainty) to build the unknown ideal dictionary \mathbf{E} . The input SNR takes the values $\{5, 10\}$ dB while the parameter K (controlling the depth of mpNet) is set to $\{8, 12\}$ respectively (determined by cross validation). The proposed method is compared to the least squares estimator and to the OMP algorithm with K iterations using either the imperfect nominal dictionary $\tilde{\mathbf{E}}$ or the unknown ideal dictionary \mathbf{E} . In order to show the interest of the imperfect model initialization, we also compare the proposed method to mpNet using a random (Gaussian) initialization.

Results. The results of this experiment are shown on figure 4 as a function of the number of channels of the form (1) seen by the base station over time. The performance measure is the relative mean squared error ($\text{rMSE} = \|\hat{\mathbf{h}} - \mathbf{h}\|_2^2 / \|\mathbf{h}\|_2^2$) averaged over minibatches of 200 channels. First of all, the imperfect model is shown to be well corrected by mpNet, the blue curve being very close to the green one (ideal unknown dictionary) after a certain amount of time. This is true both for a small uncertainty and for a large one and at all tested SNRs. Note that using the nominal dictionary (initialization of mpNet) may be even worse than the least squares method, showing the interest of correcting the model, since with learning mpNet always ends up outperforming the least squares. Second, comparing the leftmost and center figures, it is interesting to notice that learning is faster and the attained performance is better with a large SNR (the blue and green curves get closer, faster), which can be explained by the better quality of data used to train the model. Third, comparing the leftmost and rightmost figures, it is apparent that a smaller

uncertainty, which means a better initialization since the nominal dictionary is closer to the ideal unknown dictionary, leads to a faster convergence, but obviously also to a smaller improvement. Finally, comparing the blue and orange curves on all figures, it is apparent that initialization matters. Indeed, the random initialization performs much worse than the initialization with the nominal dictionary and takes longer to converge. These conclusions are very promising and highlight the applicability of the proposed method.

V. CONCLUSION AND PERSPECTIVES

In this letter, we proposed a method to add flexibility to physical models used for MIMO channel estimation. It is based on the deep unfolding strategy that views classical algorithms as neural networks that can be optimized. The proposed method was shown to correct incrementally (via online learning) an imperfect or imperfectly known physical model in order to make channel estimation as efficient as if the unknown ideal model were known. This claim was empirically validated on realistic millimeter wave outdoor channels, for various SNRs and model imperfections.

We used here uncertainty on the antenna gains and positions to illustrate physical models imperfection, but the presented method applies in principle to any imperfection (be it linear or not). For example, it could correct models in cases where the radiation pattern of the antennas differs from the nominal one, or if the plane wave assumption is not perfectly valid. Moreover, we chose to unfold the MP algorithm (since it is simple and often used in MIMO channel estimation), but more sophisticated sparse recovery algorithm could be unfolded the same way (such as iterative soft thresholding [35] or approximate message passing [36]).

In the future, an improvement of the method could be to allow it to adapt to the SNR level, for example by adjusting automatically the depth of the network, taking advantage of the optimal stopping criteria designed for greedy sparse recovery algorithms [37], [38].

Finally, note that the deep unfolding strategy used here is not limited to classical massive MIMO systems and can in principle be applied to any system that uses a physical model (for example to downlink channel estimation in a frequency division duplex (FDD) system).

REFERENCES

- [1] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.
- [2] Gabriel Peyré. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):249–260, 2009.
- [3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [4] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.
- [5] Akbar M Sayeed. Deconstructing multiantenna fading channels. *IEEE Transactions on Signal Processing*, 50(10):2563–2579, 2002.
- [6] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak. Compressed channel sensing: A new approach to estimating sparse multipath channels. *Proceedings of the IEEE*, 98(6):1058–1076, June 2010.
- [7] Luc Le Magoarou and Stéphane Paquelet. Parametric channel estimation for massive MIMO. In *IEEE Statistical Signal Processing Workshop (SSP)*, 2018.
- [8] Luc Le Magoarou, Antoine Le Calvez, and Stéphane Paquelet. Massive mimo channel estimation taking into account spherical waves. In *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5. IEEE, 2019.
- [9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [10] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 399–406. Omnipress, 2010.
- [11] John R Hershey, Jonathan Le Roux, and Felix Weninger. Deep unfolding: Model-based inspiration of novel deep architectures. *arXiv preprint arXiv:1409.2574*, 2014.
- [12] Ulugbek S Kamilov and Hassan Mansour. Learning optimal nonlinearities for iterative thresholding algorithms. *IEEE Signal Processing Letters*, 23(5):747–751, 2016.
- [13] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *arXiv preprint arXiv:1912.10557*, 2019.
- [14] Timothy OShea and Jakob Hoydis. An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, 3(4):563–575, 2017.
- [15] Tianqi Wang, Chao-Kai Wen, Hanqing Wang, Feifei Gao, Tao Jiang, and Shi Jin. Deep learning for wireless physical layer: Opportunities and challenges. *China Communications*, 14(11):92–111, 2017.
- [16] Yacong Ding and Bhaskar D Rao. Dictionary learning-based sparse channel representation and estimation for fdd massive mimo systems. *IEEE Transactions on Wireless Communications*, 17(8):5437–5451, 2018.
- [17] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, Nov 2006.
- [18] Alexios Balatsoukas-Stimming and Christoph Studer. Deep unfolding for communications systems: A survey and some new directions. *arXiv preprint arXiv:1906.05774*, 2019.
- [19] Hengtao He, Chao-Kai Wen, Shi Jin, and Geoffrey Ye Li. Deep learning-based channel estimation for beamspace mmwave massive mimo systems. *IEEE Wireless Communications Letters*, 7(5):852–855, 2018.
- [20] Christopher A Metzler, Arian Maleki, and Richard G Baraniuk. From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144, 2016.
- [21] Chris Metzler, Ali Mousavi, and Richard Baraniuk. Learned d-amp: Principled neural network based compressive image recovery. In *Advances in Neural Information Processing Systems*, pages 1772–1783, 2017.
- [22] Xiuhong Wei, Chen Hu, and Linglong Dai. Knowledge-aided deep learning for beamspace channel estimation in millimeter-wave massive mimo systems. *arXiv preprint arXiv:1910.12455*, 2019.
- [23] Michel van Lier, Alexios Balatsoukas-Stimming, Henk Corporaal, and Zoran Zivkovic. Optcomnet: Optimized neural networks for low-complexity channel estimation. *arXiv preprint arXiv:2002.10493*, 2020.
- [24] Fredrik Rusek, Daniel Persson, Buon Kiong Lau, Erik G Larsson, Thomas L Marzetta, Ove Edfors, and Fredrik Tufvesson. Scaling up mimo: Opportunities and challenges with very large arrays. *IEEE Signal Processing Magazine*, 30(1):40–60, 2013.
- [25] Erik G Larsson, Ove Edfors, Fredrik Tufvesson, and Thomas L Marzetta. Massive mimo for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, 2014.
- [26] Lu Lu, Geoffrey Ye Li, A Lee Swindlehurst, Alexei Ashikhmin, and Rui Zhang. An overview of massive mimo: Benefits and challenges. *IEEE journal of selected topics in signal processing*, 8(5):742–758, 2014.
- [27] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, Dec 1993.
- [28] J.A. Tropp and S.J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, June 2010.
- [29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*, page 318362. MIT Press, Cambridge, MA, USA, 1986.
- [30] Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.
- [31] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.
- [32] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(k^{-2})$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.
- [33] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- [34] Mathew K Samimi and Theodore S Rappaport. 3-d millimeter-wave statistical channel model for 5g wireless system design. *IEEE Transactions on Microwave Theory and Techniques*, 64(7):2207–2225, 2016.
- [35] Ingrid Daubechies, Michel DeFRise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.
- [36] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.
- [37] T Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information theory*, 57(7):4680–4688, 2011.
- [38] Rui Wu, Wei Huang, and Di-Rong Chen. The exact support recovery of sparse signals with noise via orthogonal matching pursuit. *IEEE signal processing letters*, 20(4):403–406, 2012.