ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096658

DEEP UNFOLDING-ENABLED HYBRID BEAMFORMING DESIGN FOR MMWAVE MASSIVE MIMO SYSTEMS

Nhan Nguyen, Mengyuan Ma*, Nir Shlezinger†, Yonina C. Eldar‡, A. L. Swindlehurst§, and Markku Juntti**

*Centre for Wireless Communications, University of Oulu, P.O.Box 4500, FI-90014, Finland

†School of ECE, Ben-Gurion University of the Negev, Beer-Sheva, Israel

‡Faculty of Math and CS, Weizmann Institute of Science, Rehovot, Israel

§Department of EECS, University of California, Irvine, CA, USA

Emails: {nhan.nguyen, mengyuan.ma, markku.juntti}@oulu.fi; nirshl@bgu.ac.il; yonina.eldar@weizmann.ac.il; swindle@uci.edu

ABSTRACT

Hybrid beamforming (HBF) is a key enabler for millimeter-wave (mmWave) communications systems, but HBF optimizations are often non-convex and of large dimension. In this paper, we propose an efficient deep unfolding-based HBF scheme, referred to as *ManNet-HBF*, that approximately maximizes the system spectral efficiency (SE). It first factorizes the optimal digital beamformer into analog and digital terms, and then reformulates the resultant matrix factorization problem as an equivalent maximum-likelihood problem, whose analog beamforming solution is vectorized and estimated efficiently with ManNet, a lightweight deep neural network. Numerical results verify that the proposed ManNet-HBF approach has near-optimal performance comparable to or better than conventional model-based counterparts, with very low complexity and a fast run time. For example, in a simulation with 128 transmit antennas, it attains 98.62% the SE of the Riemannian manifold scheme but 13250 times faster.

***Index Terms*—** mmWave, hybrid beamforming, massive MIMO, deep learning, AI, deep unfolding.

1. INTRODUCTION

Millimeter-wave (mmWave) massive multiple-input multiple-output (mMIMO) systems have emerged as a key enabler for 5G wireless networks with substantial improvements in the system spectral and energy efficiency (SE/EE) [1]. In such systems, hybrid beamforming (HBF) transceivers can maintain significant multiplexing gains with reduced numbers of power-hungry radio frequency (RF) chains [2–5]. However, their design and optimization are challenging due to the constant modulus constraints and the strongly coupled high-dimensional variables. Conventional optimization techniques such as Riemannian manifold minimization (MO-AltMin) [6] and alternating optimization (AO) [7] show good performance but are highly complex. Recently, the applications of deep learning (DL) in wireless communications have attracted much attention [8–11], ranging from signal detection, channel estimation [12–16] to HBF designs [16–26]. Two typical DL techniques, including purely data-driven DL and deep unfolding, are generally applied. The former relies mainly on the learning capability of deep neural networks (DNNs) [16–18] or convolutional neural networks (CNNs) [19–22] to generate HBF beamformers. This approach exhibits major limitations due to its resource-constraints, high complexity, and black-box nature [9, 12–14, 27]. Alternatively, in the deep unfolding approach, both domain knowledge and DL capabilities are leveraged to build explainable DL models that achieve performance gains and are easier to implement [27–29]. Based on this advantage, deep unfolding models have been proposed [23–26] for HBF designs with reduced feedback and improved convergence speed. However, these schemes are still complex due to the operations of highly-parameterized DNNs [23], multiple CNNs [25], or conventional projected gradient ascent/descent with learned step sizes [24, 26].

Because a deep unfolding model is constructed by unrolling a principled mathematical-oriented algorithm into layers of a DNN, its efficiency significantly depends on the conventional algorithm. Motivated by this fact, we herein propose a near-optimal low-complexity deep unfolded HBF design based on Riemannian *man*ifold optimization [6], referred to as *ManNet-HBF*. Unlike most of the existing DL-aided HBF designs, ManNet-HBF is developed based on investigating the matrix factorization problem for HBF design rather than the original SE maximization. This is efficient in the sense that the complicated log-det objective function is transformed into a simpler norm-squared form that admits a maximum-likelihood (ML) type least squares (LS) solution. We first develop a lightweight DNN architecture called ManNet to efficiently estimate the ML solution to the analog beamformer. Then, the digital beamformer is obtained using a closed-form solution. Our simulation results demonstrate that with only several layers composed of element-wise multiplications/additions, the ManNet-HBF scheme performs comparably to conventional near-optimal complex algorithms such as the MO-AltMin [6] and AO [7] schemes, in much less time and with much lower computational complexity.

2. SYSTEM MODEL AND DESIGN PROBLEM

We consider the downlink of a point-to-point mmWave mMIMO system, where the base station (BS) and the mobile station (MS) are equipped with N_{t} and N_{r} antennas,

respectively. The BS sends signal vector $\mathbf{s} \in \mathbb{C}^{N_\mathrm{s}\times 1}$ of N_s data streams to the MS, with $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \mathbf{I}_{N_\mathrm{s}}$. An analog precoder $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_\mathrm{t}\times N_{\mathrm{RF}}}$ and a digital baseband precoder $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{RF}}\times N_\mathrm{s}}$ are employed at the BS. Here, N_{RF} is the number of RF chains at the BS, $N_\mathrm{s} \le N_{\mathrm{RF}} \le N_\mathrm{t}$, and the normalized transmit power constraint at the BS is given as $\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2 = N_\mathrm{s}$. We focus on the design of hybrid precoders and assume that N_r is relatively small so that a fully digital combiner $\mathbf{V} \in \mathbb{C}^{N_\mathrm{r}\times N_\mathrm{s}}$ is employed at the MS. The post-processed signal at the BS is expressed as

$$\mathbf{y} = \sqrt{\rho}\mathbf{V}^H\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{V}^H\mathbf{n}, \tag{1}$$

where ρ denotes the average received power, $\mathbf{n}$ is an additive white Gaussian noise (AWGN) vector at the MS with elements distributed as $\mathcal{CN}(0,\sigma_\mathrm{n}^2)$, and $\mathbf{H}$ is the channel matrix.

Based on (1), the achievable SE for Gaussian symbols is given by [6]

$$R = \log_2\det\left(\mathbf{I}_{N_\mathrm{s}} + \frac{\rho}{\sigma_\mathrm{n}^2 N_\mathrm{s}}\mathbf{V}^\dagger\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^H\mathbf{F}_{\mathrm{RF}}^H\mathbf{H}^H\mathbf{V}\right),$$

where $(\cdot)^\dagger$ denotes the matrix pseudo-inverse. We aim at designing $\{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}, \mathbf{V}\}$ to maximize R, which is challenging due to the strong coupling among the variables. However, given $\{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}\}$, the optimal solution for $\mathbf{V}$ is the set of N_s left singular vectors corresponding to the N_s largest singular values of $\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$. Therefore, we focus on the designs of the hybrid precoders $\{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}\}$ in the sequel.

The SE maximization can be approximately achieved using the following design [6, 30]

$$\underset{\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}}{\text{minimize}} \quad \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_{\mathcal{F}} \tag{2a}$$

$$\text{subject to} \quad \left|f_{mn}^{\mathrm{RF}}\right| = 1,\ \forall m,n, \tag{2b}$$

$$\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_{\mathcal{F}}^2 = N_\mathrm{s}, \tag{2c}$$

where f_{mn}^{RF} is the (m,n)-th entry of $\mathbf{F}_{\mathrm{RF}}$, and $\mathbf{F}_{\text{opt}}$ is the unconstrained optimal digital precoder, given as $\mathbf{F}_{\text{opt}} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}$. Here, $\mathbf{U}$ contains columns as the N_s right singular vectors corresponding to the N_s largest singular values of $\mathbf{H}$, and $\mathbf{\Sigma}$ is a diagonal matrix with N_s water-filling power allocation factors on the diagonal. Eq. (2b) enforces the unit modulus constraints of the analog precoding coefficients, and (2c) ensures the transmit power constraint at the BS. Problem (2) is a nonconvex matrix factorization problem, and joint optimization of $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}$ is complicated due to the element-wise unit-modulus constraint (2b). The MO-AltMin [6] and orthogonal matching pursuit (OMP) [30] algorithms are two conventional model-based approaches to solving (2). In the former, $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}$ are solved by alternating between a Riemannian manifold optimization and a LS problem. Such a nested loop procedure is relatively complex and converges slowly when the system dimensions are large. In contrast, the OMP scheme requires only N_{RF} iterations to construct $\mathbf{F}_{\mathrm{RF}}$, which has low complexity but unsatisfactory performance. We overcome these challenges by proposing an efficient deep unfolding approach next.

3. PROPOSED MANNET-HBF SCHEME

3.1. Main Idea

In the proposed approach we apply the decoupling method of [7]. Specifically, we first optimize $\mathbf{F}_{\mathrm{RF}}$ with $\mathbf{F}_{\mathrm{BB}}$ given and constraint (2c) omitted. Then we design $\mathbf{F}_{\mathrm{BB}}$ to meet the constraint given the optimized $\mathbf{F}_{\mathrm{RF}}$. Thus, we first consider the following problem:

$$\underset{\mathbf{F}_{\mathrm{RF}}}{\text{minimize}}\ \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_{\mathcal{F}}^2,\ \text{subject to (2b)}, \tag{3}$$

where the quadratic form of the objective function is introduced without affecting the solution. Let $\tilde{\mathbf{z}} \triangleq \mathrm{vec}(\mathbf{F}_{\text{opt}}) \in \mathbb{C}^{N_\mathrm{t}N_\mathrm{s}\times 1}$, $\tilde{\mathbf{x}} \triangleq \mathrm{vec}(\mathbf{F}_{\mathrm{RF}}) \in \mathbb{C}^{N_\mathrm{t}N_{\mathrm{RF}}\times 1}$, and $\tilde{\mathbf{B}} \triangleq \mathbf{F}_{\mathrm{BB}} \otimes \mathbf{I}_{N_\mathrm{t}} \in \mathbb{C}^{N_\mathrm{t}N_\mathrm{s}\times N_\mathrm{t}N_{\mathrm{RF}}}$ with $\otimes$ denoting the Kronecker product. Then, the objective function in (3) can be re-expressed as $\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_{\mathcal{F}}^2 = \|\tilde{\mathbf{z}} - \tilde{\mathbf{B}}\tilde{\mathbf{x}}\|^2$. By denoting

$$\mathbf{z} \triangleq \begin{bmatrix}\Re(\tilde{\mathbf{z}})\\ \Im(\tilde{\mathbf{z}})\end{bmatrix}, \mathbf{x} \triangleq \begin{bmatrix}\Re(\tilde{\mathbf{x}})\\ \Im(\tilde{\mathbf{x}})\end{bmatrix}, \mathbf{B} \triangleq \begin{bmatrix}\Re\left(\tilde{\mathbf{B}}\right) & -\Im\left(\tilde{\mathbf{B}}\right)\\ \Im\left(\tilde{\mathbf{B}}\right) & \Re\left(\tilde{\mathbf{B}}\right)\end{bmatrix}, \tag{4}$$

with $\Re(\cdot)$ and $\Im(\cdot)$ representing the real and imaginary parts of a complex vector/matrix, respectively, we can write $\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_{\mathcal{F}}^2 = \|\mathbf{z} - \mathbf{B}\mathbf{x}\|^2$. Let a_i be the i-th element of a real-valued vector $\mathbf{a}$, and let $\mathcal{S} \triangleq \{\mathbf{x} \in \mathbb{C}^{2N_\mathrm{t}N_{\mathrm{RF}}\times 1} : x_i + jx_{N_\mathrm{t}N_{\mathrm{RF}}+i} = \tilde{x}_i, |\tilde{x}_i| = 1, i = 1,\ldots,N_\mathrm{t}N_{\mathrm{RF}}\}$. Then, $\mathcal{S}$ consists of real-valued vectors whose corresponding complex representations have unit-modulus elements, which are feasible for problem (3). With the newly introduced variables and feasible set, the optimal solution to problem (3) admits the LS problem similar to ML estimation in Gaussian noise as

$$\mathbf{x}_{\mathrm{ML}} = \underset{\mathbf{x}\in\mathcal{S}}{\operatorname{argmin}} \|\mathbf{z} - \mathbf{B}\mathbf{x}\|^2. \tag{5}$$

In the deep unfolding technique, a DNN of L layers is designed to mimic the projected gradient descent algorithm to approximate $\mathbf{x}_{\mathrm{ML}}$. Specifically, let $\mathbf{x}_\ell$ be the output of the ℓ-th layer of the DNN. From (5), $\mathbf{x}_\ell$ can be produced as [31]

$$\begin{aligned}\mathbf{x}_\ell &= \Pi_\ell\left(\mathbf{x} - \delta_\ell\frac{\partial\|\mathbf{z}-\mathbf{B}\mathbf{x}\|^2}{\partial\mathbf{x}}\right)_{\mathbf{x}=\mathbf{x}_{\ell-1}}\\ &= \Pi_\ell\left(\mathbf{x}_{\ell-1} - \delta_\ell\mathbf{B}^T\mathbf{z} + \delta_\ell\mathbf{B}^T\mathbf{B}\mathbf{x}_{\ell-1}\right),\end{aligned} \tag{6}$$

where δ_ℓ denotes a step size, and $\Pi_\ell(\cdot)$ represents a nonlinear projection operator mapping $\mathbf{x}_{\ell-1}$ to $\mathbf{x}_\ell$. The relationship in (6) motivates a DNN model to learn $\mathbf{x}_{\mathrm{ML}}$ wherein the output of a given layer (i.e., $\mathbf{x}_\ell$ in the ℓ-th layer) results from a nonlinear projection applied to the output of the previous layer (i.e., $\mathbf{x}_{\ell-1}$ in the $(\ell-1)$-th layer) and other given information, including $\mathbf{B}$ and $\mathbf{z}$. The nonlinear projection is performed with trainable parameters, including the weights and biases of the DNN, and the activation function. In this regard, the DNN can efficiently learn the projection and the step size of the projected gradient descent algorithm. Applied over multiple layers, the final output, i.e., $\mathbf{x}_L$, will be a good estimate of $\mathbf{x}_{\mathrm{ML}}$ as long as the DNN is well structured and trained. Next, we develop such an efficient DNN architecture refered to as ManNet.

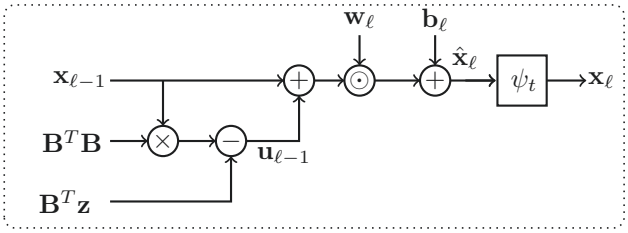


Fig. 1. Detailed operation of sparse layer ℓ of ManNet. Here, $\odot$ represents the Hadamard product of two vectors.

3.2. Proposed ManNet Approach

To configure ManNet, we denote $\mathbf{u}_{\ell-1} \triangleq \mathbf{B}^T\mathbf{B}\mathbf{x}_{\ell-1} - \mathbf{B}^T\mathbf{z}$ and expand (6) as

$$\mathbf{x}_\ell = \Pi_\ell \left(\mathbf{x}_{\ell-1} + \delta_\ell \mathbf{u}_{\ell-1}\right). \tag{7}$$

In this approach the classical gradient descent optimization is learned by a DNN that performs nonlinear transformations, avoiding computationally intensive tasks (e.g., line search of the step size, computing the gradient) as required in conventional Riemannian manifold optimization. As such, we propose ManNet as a network of L layers defined by (7) whose goal is to learn $\mathbf{x}_{\text{ML}}$. Its implementation is detailed next.

Remark 1 *ManNet takes $\mathbf{x}_{\ell-1}$ and $\mathbf{u}_{\ell-1}$ as the input of the ℓ-th layer, and outputs $\mathbf{x}_\ell$ as the result of the nonlinear transformation Π_ℓ, as indicated in* (7)*. Importantly, the i-th element of $\mathbf{x}_\ell$ only depends on the i-th elements of $\mathbf{x}_{\ell-1}$ and $\mathbf{u}_{\ell-1}$. Thus, only the nodes (or neurons) at the same vertical level between layers are connected making ManNet a sparsely connected DNN. We employ activation function [12]*

$$\psi_t(x) = -1 + \frac{1}{|t|}\left(\sigma(x+t) - \sigma(x-t)\right), \tag{8}$$

where $\sigma(\cdot)$ is the rectified linear unit (ReLU) activation function, and t is a training parameter. This guarantees that the amplitudes of the elements of $\mathbf{x}_\ell$ are in the range $[-1, 1]$, i.e., $|x_i| \le 1, i = 1, \ldots, 2N_tN_{RF}$. As a result, its corresponding complex-valued representation $\tilde{\mathbf{x}}_\ell$ has elements $\tilde{x}_{\ell,i}$ with $|\tilde{x}_{\ell,i}| \le \sqrt{2}$. The final output of the DNN is then normalized to produce a feasible solution satisfying constraint (2b)*.*

Let $\mathbf{w}_\ell$ and $\mathbf{b}_\ell$ denote the weight and bias vectors of the ℓ-th layer of ManNet. A detailed network architecture illustrating the operation of each layer is shown in Fig. 1. We employ the loss function

$$\mathcal{L} = \sum_{\ell=1}^{L} \log(\ell) \left\|\mathbf{z} - \mathbf{B}\mathbf{x}_\ell\right\|^2, \tag{9}$$

which sums the total objective values of all L layers. The DNN is trained to optimize the parameter set $\left\{\{\mathbf{w}_\ell, \mathbf{b}_\ell\}_{\ell=1}^L, t\right\}$ such that $\mathcal{L}$ is minimized, which also directly minimizes the objective function in (5) at the network output $\ell = L$. It is seen from the loss function (9) that training labels for $\mathbf{F}_{\text{RF}}$ are not required. Thus, the training method is unsupervised. Note that if supervised training were used, it would require implementation of a conventional HBF scheme to obtain the training labels, which would dramatically increase the training complexity.

Algorithm 1 ManNet-HBF

Input: $\mathbf{H}, \mathbf{F}_{\text{opt}}$, ManNet's trained parameters $\left\{\{\mathbf{w}_\ell, \mathbf{b}_\ell\}_{\ell=1}^L, t\right\}$.
Output: $\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}$.
1: Initialize $\mathbf{F}_{\text{RF}}^{(0)}$ and $\mathbf{F}_{\text{BB}}^{(0)}$ based on the OMP scheme.
2: Obtain $\mathbf{z}$, $\mathbf{x}$, and $\mathbf{B}$ based on (4).
3: **for** $\ell = 1 \to L$ **do**
4: Construct the input: $\mathbf{u}_{\ell-1} \triangleq \mathbf{B}^T\mathbf{B}\mathbf{x}_{\ell-1} - \mathbf{B}^T\mathbf{z}$.
5: Apply weights: $\hat{\mathbf{x}}_\ell = \mathbf{w}_\ell \odot \mathbf{x}_{\ell-1} + \mathbf{b}_\ell$.
6: Apply the activation function: $\mathbf{x}_\ell = \psi_t(\hat{\mathbf{x}}_\ell)$.
7: **end for**
8: Reconstruct the complex RF precoding matrix $\mathbf{F}_{\text{RF}}$ from $\mathbf{x}_L$.
9: Obtain $\mathbf{F}_{\text{BB}}$ based on (11).

Table 1. Computational complexity of the proposed ManNet-HBF scheme compared with conventional MO-AltMin, AO, and OMP approaches.

HBF schemes	Complexity per iter.	No. iter.
ManNet-HBF	$\mathcal{O}(8N_{\text{t}}^2N_{\text{RF}}^2)$ (real)	L
MO-AltMin	$\mathcal{O}(2N_{\text{t}}^2N_{\text{RF}}N_{\text{s}}I_{\text{MO}}^{\text{in}})$ (complex)	$I_{\text{MO}}^{\text{in}}I_{\text{MO}}^{\text{out}}$
AO	$\mathcal{O}(2N_{\text{t}}^3N_{\text{RF}})$ (complex)	$N_{\text{t}}N_{\text{RF}}I_{\text{AO}}$
OMP	$\mathcal{O}(N_{\text{t}}^2N_{\text{RF}}N_{\text{s}})$ (complex)	N_{RF}

3.3. Proposed ManNet-HBF Algorithm

Once the offline training process is completed, ManNet is readily applied for online HBF design. The overall deep unfolding-enabled HBF scheme is summarized in Algorithm 1. Steps 1–2 are used to initialize the algorithm, wherein the low-complexity OMP scheme is applied to generate the initial analog and digital precoders. After that, ManNet executes steps 3–7 to construct the outputs of each layer. Note that only element-wise multiplications between the weight and input vectors are required, as seen in step 5 and Fig. 1. The final output of ManNet, i.e., $\mathbf{x}_L$, is reconstructed as the feasible solution to $\mathbf{F}_{\text{RF}}$ in step 8. More specifically, let

$$x_i^\star = \frac{x_{L,i} + jx_{L,i+N_{\text{t}}N_{\text{RF}}}}{|x_{L,i} + jx_{L,i+N_{\text{t}}N_{\text{RF}}}|}, i = 1, \ldots, N_{\text{t}}N_{\text{RF}},$$

which satisfies (2b), with $x_{L,i}$ being the i-th element of $\mathbf{x}_L$. Then, $\mathbf{F}_{\text{RF}}$ is obtained as $\mathbf{F}_{\text{RF}} = \text{vec}^{-1}([x_1^\star, \ldots, x_{N_{\text{t}}N_{\text{RF}}}^\star]^T)$, where $\text{vec}^{-1}(\cdot)$ reshapes a vector of size $N_{\text{t}}N_{\text{RF}} \times 1$ to form a matrix of size $N_{\text{t}} \times N_{\text{RF}}$.

With $\mathbf{F}_{\text{RF}}$ obtained, define $\tilde{\mathbf{H}} = \mathbf{H}\mathbf{F}_{\text{RF}}$ and $\mathbf{Q} = \mathbf{F}_{\text{RF}}^H\mathbf{F}_{\text{RF}}$. Then, the digital precoder design problem can be written as

$$\underset{\mathbf{F}_{\text{BB}}}{\text{maximize}} \quad \log_2 \det\left(\mathbf{I}_{N_{\text{s}}} + \frac{\rho}{\sigma_{\text{n}}^2 N_{\text{s}}}\tilde{\mathbf{H}}\mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^H\tilde{\mathbf{H}}^H\right) \tag{10a}$$

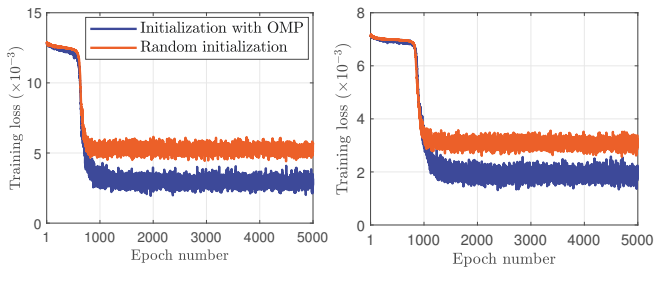
$$\text{subject to} \quad \text{trace}\left(\mathbf{Q}\mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^H\right) = N_{\text{s}}, \tag{10b}$$

which has the well-known water-filling solution:

$$\mathbf{F}_{\text{BB}} = \mathbf{Q}^{-\frac{1}{2}}\tilde{\mathbf{U}}\tilde{\mathbf{\Gamma}}, \tag{11}$$

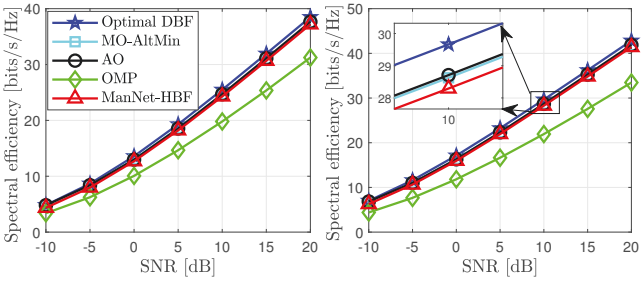
where the columns of $\tilde{\mathbf{U}}$ are taken from the right singular vectors corresponding to the N_{s} largest singular values of $\tilde{\mathbf{H}}\mathbf{Q}^{-\frac{1}{2}}$, and $\tilde{\mathbf{\Gamma}}$ is a diagonal matrix whose elements are defined by the power allocated to the N_{s} data streams [7]. In Algorithm 1, $\mathbf{F}_{\text{BB}}$ is obtained in step 9.

Table 1 presents the per-iteration complexity and the number of iterations of Algorithm 1 compared with those of MO-AltMin [6], AO [7], and OMP [30]. First, these com-



(a) $N_t = 64$ (b) $N_t = 128$

Fig. 2. Normalized training loss of ManNet with $N_r = N_{RF} = N_s = 4$, $N_t = \{64, 128\}$, and $L = \{6, 7\}$.



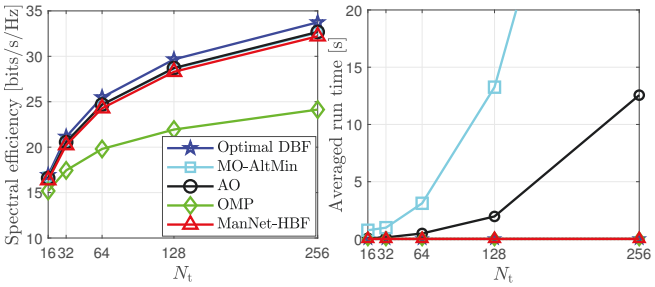
(a) $N_t = 64$ (b) $N_t = 128$

Fig. 3. SE performance of ManNet-HBF with $N_r = N_{RF} = N_s = 4$, $N_t = \{64, 128\}$, and $L = \{6, 7\}$.

pared schemes require $\mathcal{C}_{\text{MO-AltMin}} = I_{\text{MO}}^{\text{in}} I_{\text{MO}}^{\text{out}} \mathcal{O}(2N_t^2 N_{RF} N_s)$, $\mathcal{C}_{\text{AO}} = N_t N_{RF} I_{\text{AO}} \mathcal{O}(2N_t^3 N_{RF})$, and $\mathcal{C}_{\text{OMP}} = N_{RF} \mathcal{O}(N_t^2 N_{RF} N_s)$ complex operations, respectively. Here, $I_{\text{MO}}^{\text{in}}$, $I_{\text{MO}}^{\text{out}}$, and I_{AO} denote the number of inner and outer iterations for MO-AltMin and the number of iterations for AO, respectively. Algorithm 1 has a total complexity of $\mathcal{C}_{\text{ManNet-HBF}} = \mathcal{C}_{\text{OMP}} + \mathcal{C}_{\text{ManNet}}$, where $\mathcal{C}_{\text{ManNet}} = L\mathcal{O}(8N_t^2 N_{RF}^2)$ real operations, dominated by the computation in step 4. The required number of iterations is fixed as L, the number of network layers. Note that $\mathbf{B}^T\mathbf{B}$ and $\mathbf{B}^T\mathbf{z}$ need to be computed only once and do not change over the layers, and that ManNet requires only element-wise vector multiplications/additions (see step 5), which explains its low complexity. In general, $L \ll N_t$ and $L \ll I_{\text{MO}}^{\text{in}} I_{\text{MO}}^{\text{out}}$, while L is of the same order as N_{RF}. For example, with $N_t = 128, N_r = N_s = N_{RF} = 4$, ManNet needs only $L = 7$ layers, whereas our simulations show that $I_{\text{MO}}^{\text{in}} I_{\text{MO}}^{\text{out}} = 648$ to achieve a convergence tolerance of 10^{-3}. Thus, it is clear that $\mathcal{C}_{\text{ManNet-HBF}} \ll \mathcal{C}_{\text{MO-AltMin}}$, $\mathcal{C}_{\text{ManNet-HBF}} \ll \mathcal{C}_{\text{AO}}$, and $\mathcal{C}_{\text{ManNet-HBF}} \approx 2\mathcal{C}_{\text{OMP}}$.

4. SIMULATION RESULTS

Here we provide numerical results to demonstrate the performance of ManNet-HBF. We assume scenarios with $N_r = N_{RF} = N_s = 4$, $N_t = \{16, 64, 128, 256\}$, and various numbers of layers for ManNet: $L = \{4, 6, 7, 10\}$. The channel realizations are generated as in [6]. Specifically, we assume the Saleh-Valenzuela model for the channel $\mathbf{H}$, with the numbers of clusters and paths and the average power of each cluster being set as 5, 10, and 1, respectively, and we assume that the azimuth/elevation angles of departure/arrival follow a Laplacian distribution with a uniformly distributed mean over $[0, 360^\circ)$ and an angular spread of 10°. ManNet is implemented using Python with the Pytorch library and a Tesla V100-SXM2 processor. For the training phase, a decaying learning rate of 0.97, an initial learning rate of 0.0001, and $t = 0.1$ are used. For comparison, we consider optimal fully digital beamforming (DBF), MO-AltMin [6], AO [7], and OMP [30].



(a) SE performance (b) Time complexity

Fig. 4. SE performance and run time of ManNet-HBF with $N_t \in [16, 256]$, $N_r = N_{RF} = N_s = 4$, and SNR = 10 dB.

We first show the loss obtained in (9) during training ManNet with $N_t = \{64, 128\}$ in Fig. 2. It is seen for both cases that the loss decreases and essentially converges after about 1500 epochs. Furthermore, OMP allows a better convergence compared with the random initialization. As the loss function (9) also measures the objective in (3), the convergence of the training loss reflects the ability of ManNet to solve (3).

In Figs. 3 and 4, we show the SE and run time of ManNet-HBF. While the AO and MO-AltMin methods are near-optimal, the OMP approach exhibits a significant performance loss in all the considered scenarios. On the other hand, ManNet-HBF achieves almost the same performance as MO-AltMin for all SNR and N_t. For example, at 10 dB SNR and $N_t = 128$, it attains 98.51%, 98.62%, and 129.29% of the SE achieved by AO, MO-AltMin, and OMP, respectively. ManNet-HBF is further shown to be the fastest approach among the near-optimal schemes in Fig. 4(b) with a run time of only 0.001 s, which is about 196 and 13250 times faster than AO (1.96 s) and MO-AltMin (13.25 s), respectively, at $N_t = 128$. In particular, its time complexity gain is more significant when N_t increases. This show that ManNet-HBF achieves a remarkable complexity reduction with only a marginal loss in performance compared to the conventional approaches.

5. CONCLUSION

The nonconvexity and high-dimensional variables have imposed significant challenges to HBF designs in the literature, which have usually required cumbersome iterative procedures. We have overcome these difficulties by proposing the efficient ManNet-HBF scheme based on unfolding Riemannian manifold minimization. In this scheme, the lightweight ManNet produces the analog precoder with only several layers and sparse connections in each, which explains the computational and time efficiency of the ManNet-HBF scheme. Our extensive simulation results have demonstrated that the ManNet-HBF has superior performance with lightweight implementation, low complexity, and fast execution.

6. REFERENCES

[1] A. L. Swindlehurst *et al.*, "Millimeter-wave massive MIMO: The next wireless revolution?," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 56–62, 2014.

[2] X. Gao, L. Dai, and A. M. Sayeed, "Low RF-complexity technologies to enable millimeter-wave MIMO with large antenna array for 5G wireless communications," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 211–217, 2018.

[3] N. T. Nguyen and K. Lee, "Unequally sub-connected architecture for hybrid beamforming in massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1127–1140, 2019.

[4] G. M. Gadiel, N. T. Nguyen, and K. Lee, "Dynamic unequally sub-connected hybrid beamforming architecture for massive MIMO systems," *IEEE Trans. Veh. Technol.*

[5] S. S. Ioushua and Y. C. Eldar, "A family of hybrid analog–digital beamforming methods for massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3243–3257, 2019.

[6] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, 2016.

[7] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, 2016.

[8] Q.-V. Pham *et al.*, "Intelligent radio signal processing: A survey," *IEEE Access*, vol. 9, pp. 83818–83850, 2021.

[9] A. Jagannath, J. Jagannath, and T. Melodia, "Redefining wireless communication for 6G: Signal processing meets deep learning with deep unfolding," *IEEE Trans. Artificial Intelligence*, vol. 2, no. 6, pp. 528–536, 2021.

[10] L. Dai, R. Jiao, F. Adachi, H. V. Poor, and L. Hanzo, "Deep learning for wireless communications: An emerging interdisciplinary paradigm," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 133–139, 2020.

[11] G. Zhu *et al.*, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, 2020.

[12] N. T. Nguyen and K. Lee, "Deep learning-aided tabu search detection for large MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4262–4275, 2020.

[13] N. T. Nguyen *et al.*, "Application of Deep Learning to Sphere Decoding for Large MIMO Systems," *IEEE Trans. Wireless Commun.*, 2021, early access.

[14] L. V. Nguyen *et al.*, "Leveraging deep neural networks for massive MIMO data detection," *IEEE Wireless Commun.*, 2022.

[15] J. He *et al.*, "Learning to estimate RIS-aided mmwave channels," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 841–845, 2022.

[16] X. Li and A. Alkhateeb, "Deep learning for direct hybrid precoding in millimeter wave massive MIMO systems," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, 2019, pp. 800–805.

[17] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, 2019.

[18] Q. Hu *et al.*, "Two-timescale end-to-end learning for channel acquisition and hybrid precoding," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 163–181, 2021.

[19] A. M. Elbir and K. V. Mishra, "Joint antenna selection and hybrid beamformer design using unquantized and quantized deep learning networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1677–1688, 2019.

[20] A. M. Elbir and A. K. Papazafeiropoulos, "Hybrid precoding for multiuser millimeter wave massive MIMO systems: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 552–563, 2019.

[21] T. Peken *et al.*, "Deep learning for SVD and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6621–6642, 2020.

[22] H. Hojatian, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, "Unsupervised deep learning for massive MIMO hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7086–7099, 2021.

[23] E. Balevi and J. G. Andrews, "Unfolded hybrid beamforming with GAN compressed ultra-low feedback overhead," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8381–8392, 2021.

[24] O. Agiv and N. Shlezinger, "Learn to rapidly optimize hybrid precoding," in *Proc. IEEE Works. on Sign. Proc. Adv. in Wirel. Comms.*, 2022, pp. 1–5.

[25] S. Shi, Y. Cai, Q. Hu, B. Champagne, and L. Hanzo, "Deep-unfolding neural-network aided hybrid beamforming based on symbol-error probability minimization," *IEEE Trans. Veh. Technol.*, 2022.

[26] K.-Y. Chen *et al.*, "Hybrid beamforming in mmwave MIMO-OFDM systems via deep unfolding," in *Proc. IEEE Veh. Technol. Conf.*, 2022, pp. 1–7.

[27] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?," *IEEE Trans. Wireless Commun.*, vol. 67, no. 10, pp. 7331–7376, 2019.

[28] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *arXiv*, 2022.

[29] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, 2021.

[30] O. El Ayach *et al.*, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, 2014.

[31] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2554–2564, 2019.