ISEE

浙江大學 ZHEJIANG UNIVERSITY

Hybrid Precoding Design Based on Dual-Layer Deep-Unfolding Neural Network

Guangyi Zhang*, Xiao Fu*, Qiyu Hu, Yunlong Cai, Guanding Yu

(* denotes equal contribution)

浙江大學 信息与电子工程学院
College of Information Science & Electronic Engineering, Zhejiang University

Background

Dilemma of wireless communication

■ Most of the existing algorithms to solve the field of wireless communication are iterative optimization algorithms, although in convergence and convergence The performance is satisfactory, but it is difficult to be applied commercially due to the **high computational complexity**.

Goal

■ Use the properties of existing iterative algorithm to design the deep-unfolding network to reduce the computational complexity on the premise of ensuring that **the performance loss is within a certain range.**

Excellent performance and convergence
High time and space complexity

Iterative algorithm

Deep Unfolding

trade off between performance and complexity

High performance and convergence
Low time and space complexity

Deep-unfolding algorithm

Scenario：Massive MIMO

（Key technologies to improve system capacity and spectrum efficiency in 5G）

Core: Deep Unfolding

Traditional Methods

[1] Model-Based

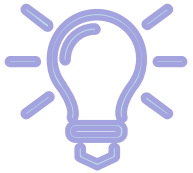
Pros：**Practical** and **accurate**

Cons：**Hard to model**, traditional modeling assumptions are **no longer valid**(e.g., instability and irregularity of interference)

[2] Machine-Learning

Pros：**Simple reasoning** and **general** architecture

Cons：**Only consider input and output**, generalization ability and interpretability are **poor**



Deep Unfolding

An efficient combination of [1] and [2].

Less iterations than [1].

Less learning parameters than [2].

It has better interpretability and convergence.

PDD algorithm on hybrid precoding design

(1) Select reasons

1. Pure digital coding: the application of landing equipment is **expensive**
2. WMMSE：only **single** layer

generalization

1. Hybrid precoding: the cost of landing equipment is **low**
2. PDD： The **double-layer** iterative network has not been explored

(2) Maximizing spectral efficiency of hybrid precoding[3]

$$\max_{\mathbf{V},\mathbf{U}} \sum_{k=1}^{K} \log det\left(\mathbf{I}+\mathbf{U}_{RF_k}^H\mathbf{H}_k\mathbf{V}_{RF}\mathbf{V}_{BB_k}\mathbf{V}_{BB_k}^H\mathbf{V}_{RF}^H\mathbf{H}_k^H\mathbf{U}_{RF_k}\tilde{\boldsymbol{\Upsilon}}_k^{-1}\right)$$

$$s.t.\sum_{k=1}^{K}\|\mathbf{V}_{RF}\mathbf{V}_{BB_k}\|^2 \le P$$

$$|\mathbf{V}_{RF}(i,j)| = 1, \forall i,j$$

$$|\mathbf{U}_{RF_k}(i,j)| = 1, \forall i,j,k$$

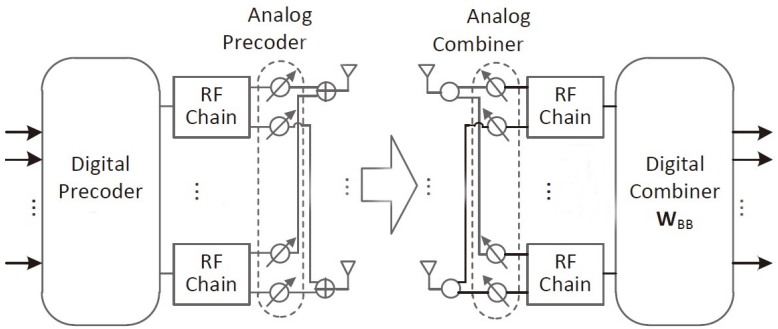
Constraints

$$\tilde{\boldsymbol{\Upsilon}}_k \triangleq \mathbf{U}_{RF_k}^H(\sigma^2\mathbf{I}+\sum_{j\neq k}\mathbf{H}_k\mathbf{V}_{RF}\mathbf{V}_{BB_j}\mathbf{V}_{BB_j}^H\mathbf{V}_{RF}^H\mathbf{H}_k^H\mathbf{U}_{RF_k}$$

$\mathbf{V_{BB}}$: digital precoder　　$\mathbf{V_{RF}}$: analog precoder

$\mathbf{U_{BB}}$: digital combiner　　$\mathbf{U_{RF}}$: analog combiner

$\mathbf{N_{RF}}$: transmit RF chains　　$\mathbf{M_{RF}}$: receive RF chains



Physical Diagram

[3] Shi Q, Hong M. Spectral efficiency optimization for millimeter wave multiuser MIMO systems[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2018, 12(3): 455-468.

A dual-layer penalty dual decomposition (PDD) algorithm

Algorithm 3 Penalty dual decomposition method

Input: Initialize $\mathbf{V}_{RF}$, $\mathbf{V}_{BB_k}$, $\mathbf{U}_{RF_k}$ and $\mathbf{X}_k$ to satisfy $\mathbf{X}_k = \mathbf{V}_{RF}\mathbf{V}_{BB_k}$, $|\mathbf{V}_{RF}(i,j)| = 1$, $|\mathbf{U}_{RF_k}(i,j)| = 1$, $\sum_{k=1}^{K} \|\mathbf{V}_{RF}\mathbf{V}_{BB_k}\|^2 \le P$. Set the constraint violation parameter$\{\epsilon, \eta_0\}$, the penalty parameter ρ, the control parameter c.

Output: optimal $\{\mathbf{V}_{RF}, \mathbf{V}_{BB_k}, \mathbf{U}_{RF_k}, \mathbf{U}_{BB_k}\}$

repeat
  repeat
    $\mathbf{U}_{BB_k} = \left(\mathbf{U}_{RF_k}^H \mathbf{A}_k \mathbf{U}_{RF_k}\right)^{\dagger} \left(\mathbf{U}_{RF_k}^H \mathbf{H}_k \mathbf{X}_k\right)$
    $\mathbf{W}_k = \left(\mathbf{I} - \mathbf{U}_{BB_k}^H \left(\mathbf{U}_{RF_k}^H \mathbf{H}_k \mathbf{X}_k\right)\right)^{-1}$
    $\mathbf{V}_{RF} = \mathrm{BCD}\left(\mathbf{I}, \mathbf{V}_{RF}, \sum_{k=1}^{K} \mathbf{V}_{BB_k}\mathbf{V}_{BB_k}^H, \sum_{k=1}^{K} \left(\mathbf{X}_k + \rho\mathbf{Y}_k\right)\mathbf{V}_{BB_k}^H\right)$
    $\mathbf{U}_{RF_k} = \mathrm{BCD}\left(\mathbf{A}_k, \mathbf{U}_{RF_k}, \mathbf{U}_{BB_k}\mathbf{W}_k\mathbf{U}_{BB_k}^H, \mathbf{H}_k\mathbf{X}_k\mathbf{W}_k\mathbf{U}_{BB_k}^H\right)$
    $\mathbf{V}_{BB_k} = \left(\mathbf{V}_{RF}\right)^{\dagger}\left(\mathbf{X}_k + \rho\mathbf{Y}_k\right)$
    $\mathbf{X}_k = \left(\mathbf{A}_\rho + \mu\mathbf{I}\right)^{-1}\mathbf{B}_{\rho,k}$
  until $\frac{|\mathcal{L}_k(\mathbf{x}^t) - \mathcal{L}_k(\mathbf{x}^{t-1})|}{|\mathcal{L}_k(\mathbf{x}^{t-1})|} \le \epsilon$
  if $\max_k \|\mathbf{X}_k - \mathbf{V}_{RF}\mathbf{V}_{BB_k}\|_\infty \le \eta_t$ then
    $\mathbf{Y}_k = \mathbf{Y}_k + \frac{1}{\rho}\left(\mathbf{X}_k - \mathbf{V}_{RF}\mathbf{V}_{BB_k}\right)$
  else
    $\rho = c\rho$
  end if $t = t + 1$
  η_t=0.9 $\max_k \|\mathbf{X}_k - \mathbf{V}_{RF}\mathbf{V}_{BB_k}\|_\infty$, $\epsilon_k = c\epsilon_k$
until $\max_k \|\mathbf{X}_k - \mathbf{V}_{RF}\mathbf{V}_{BB_k}\|_\infty \le \epsilon$

Use the **ADMM algorithm** of convex optimization→ The nonconvex coupled PDD problem is transformed into a left-hand bilevel algorithm, which iteratively updates 6 blocks to solve the global optimal solution

- coupling constraints: $\mathbf{X_k} = \mathbf{V_{RF}}\ \mathbf{V_{BB_k}}$
- dual variable: $\mathbf{Y_k}$
- penalty factor: ρ

$\mathbf{A}_k \triangleq \sigma^2\mathbf{I} + \sum_{j=1}^{K} \mathbf{H}_k\mathbf{X}_j\mathbf{X}_j^H\mathbf{H}_k^H$
BCD-type algorithm can be found in [4]
$\mathbf{A}_\rho \triangleq \sum_{j=1}^{K}(\mathbf{H}_j^H\mathbf{U}_{RF_j}\mathbf{U}_{BB_j}\mathbf{W}_j\mathbf{U}_{BB_j}^H\mathbf{U}_{RF_j}^H\mathbf{H}_j) + \frac{1}{2\rho}\mathbf{I}$
$\mathbf{B}_{\rho,k} \triangleq \mathbf{H}_k^H\mathbf{U}_{RF_k}\mathbf{U}_{BB_k}\mathbf{W}_k + \frac{1}{2}(\frac{1}{\rho}\mathbf{V}_{RF_k}\mathbf{V}_{BB_k} - \mathbf{Y}_k)$
μ satisfies $\sum_{k=1}^{K} \mathrm{Tr}(\mathbf{B}_{\rho,k}^H(\mathbf{A}_\rho + \mu\mathbf{I})^{-2}\mathbf{B}_{\rho,k}) = P$ and is obtained by applying bisection method.
$\mathcal{L}_k = \sum_{k=1}^{K}(\log\det(\mathbf{W_k}) - \mathrm{Tr}(\mathbf{W}_k\mathbf{E}_k(\mathbf{U},\mathbf{X})) + d) - \sum_{k=1}^{K}\frac{1}{2\rho}\|\mathbf{X}_k - \mathbf{V}_{RF}\mathbf{V}_{BB_k} + \rho\mathbf{Y}_k\|^2$

Supplementary Material

Pipeline：DLDUNN (Dual-layer deep-unfolding neural network

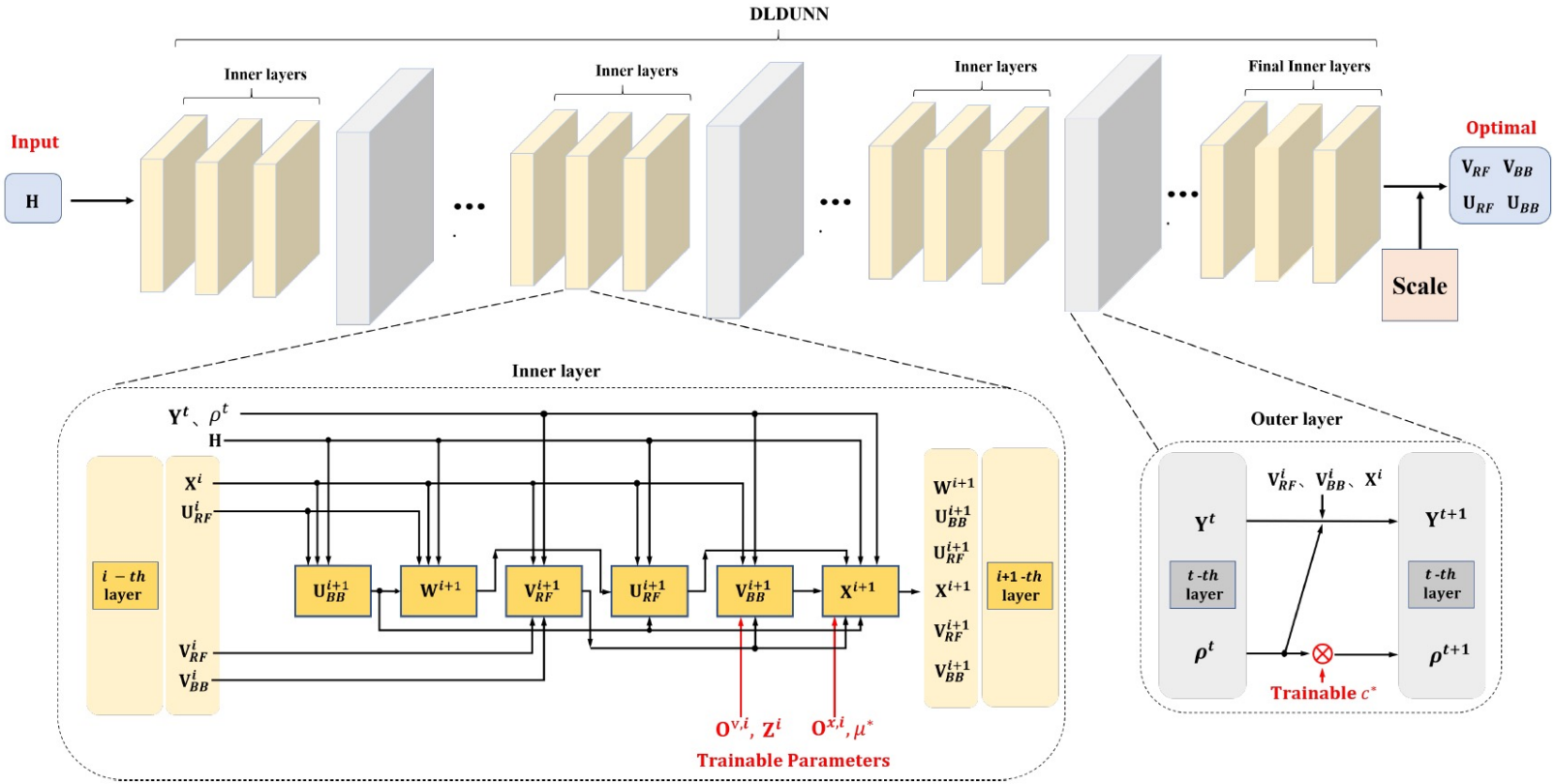


Fig. 1. The architecture of DLDUNN. The index i, $i+1$ denote the current inner layer and the next inner layer, repectively. The index t, $t+1$ denote the current outer layer and the next outer layer, repectively. The scale operation is placed after the final layer.

Step(1)：Fix the number of iterations for DLDUNN's inner and outer layers.

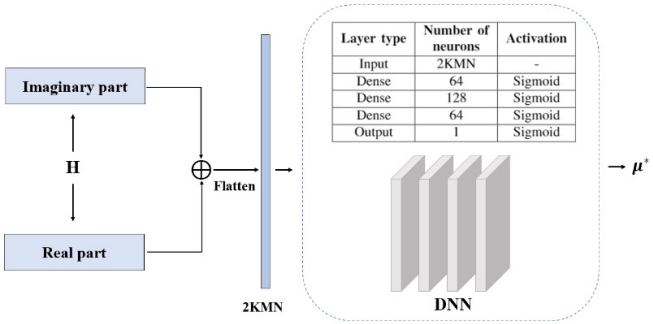
Step(2)：The conditional branch statements, i.e., if-else, may cause gradient interruption in the backpropagation. Thus, we keep the dual variable Y_k and the penalty parameter ρ updated in each outer layer. Fix the iteration number of the iterative BCD-type algorithm to 1.

Step(3)：To find the appropriate penalty parameter ρ for each inner layer, we introduce trainable parameter c* to replace the original scalar c in the outer layer as : $\rho^{t+1} = c^* \rho^t$

Step(4)：Introduce trainable parameters Z and O to address the performance loss caused by fixing $L_{BCD} = 1$.

$$\mathbf{V}_{BB_k}^{i+1} = \mathbf{Z}_k^i \left(\mathbf{V}_{RF}^i\right)^\dagger \left(\mathbf{X}_k^i + \rho^t \mathbf{Y}_k^t\right) + \mathbf{O}_k^{v,i}$$

Step(5)：Introduce DNN to replace the bisection method in X's layer to learn μ* and O to compensate for the performance loss.



Layer type	Number of neurons	Activation
Input	2KMN	-
Dense	64	Sigmoid
Dense	128	Sigmoid
Dense	64	Sigmoid
Output	1	Sigmoid

Flatten the real part and imaginary part of channel H into the DNN.

Activation function：sigmoid

Step(6)：The power constraint cannot be satisfied by the calculated μ*. Thus the scale operation is introduced at the end of the architecture,

$$\mathbf{V}_{BB_k} = \mathbf{V}_{BB_k}^* \frac{P}{\left\|\mathbf{V}_{RF}\mathbf{V}_{BB_k}^*\right\|^2}, \forall k$$

Experimental Result

(1) Settings :

Gaussian channel: $h_{i,j} \sim \mathcal{CN}(0,1)$

signal streams：d = 2

receive antennas: M

transmit antennas：N

transmit RF chains：$N_{RF} \geq Kd$

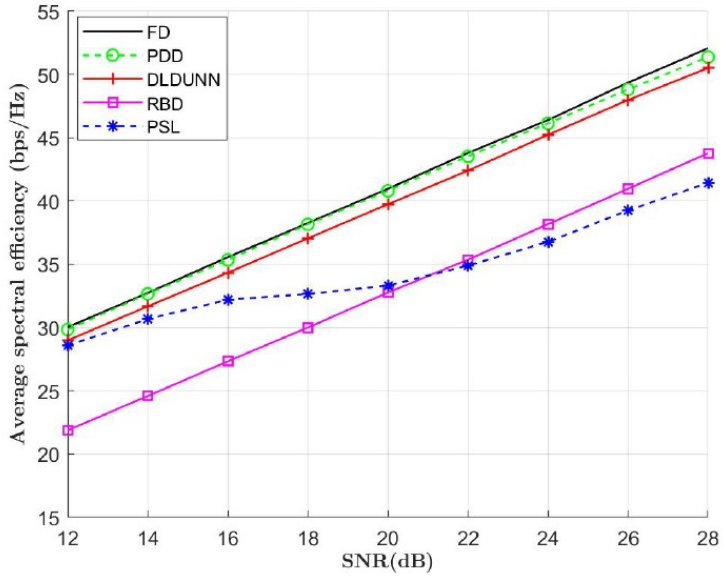
receive RF chains：$M_{RF} = d$

DLDUNN's inner layers：$D_{in} = 10$

DLDUNN's outer layers：$D_{out} = 7$

take PDD's average results of 100 randomly initialized channel

(2) Spectrum Efficiency



FD : fully-digital precoding

RBD : heuristic algorithm[4]

PSL : PDD of same layer

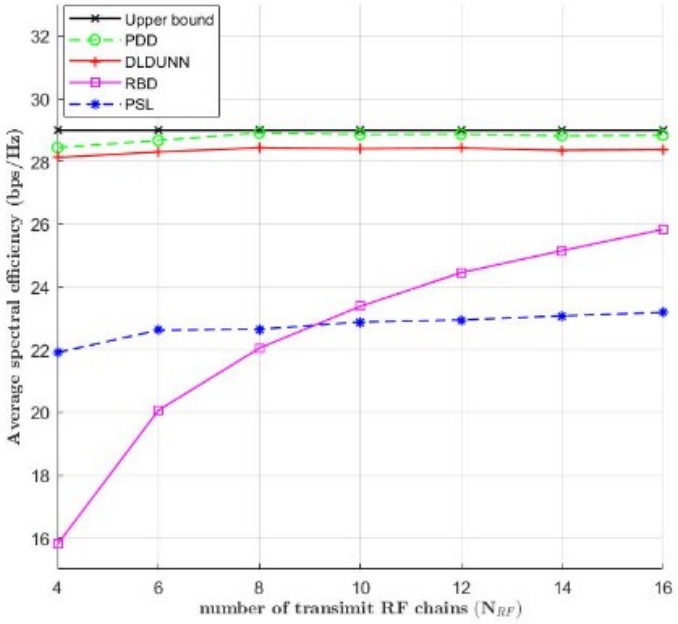
# of user (K)	2	3	4	5
FD(bps/Hz)	28.985	40.964	51.880	62.038
PDD(bps/Hz)	28.932	40.801	51.643	61.513
Achievable performance (PDD-based)				
DLDUNN	97.82%	97.49%	97.57%	97.35%
PSL	77.91%	81.64%	86.77%	91.63%
RBD	75.88%	79.90%	83.86%	87.56%

② Spectrum efficiency versus multi-users

① Spectrum efficiency achieved by different methods, K=3, d=2, N_{RF}=12

[4] J. Zhang, M. Haardt, I. Soloveychik, and A. Wiesel, "A channelmatching based hybrid analog-digital strategy for massive multi-userMIMO downlink systems," *2016 IEEE Sensor Array and MultichannelSignal Processing Workshop (SAM),* Jul 2016, pp. 1–5.

Experimental Result



③ Spectrum efficiency versus RF chains, K = 2, SNR = 20dB

(1) FD's result (28.985bps/Hz) is an upper bound. $N_{RF}=N$

(2) DLDUNN's result is close to the upper bound under $N_{RF}<2Kd$

(3) RBD method slowly approaches the upper bound as the number of the transmit RF chains increases, where system is gradually closer to the FD precoding structure.

(Successful signal detection: $N_{RF} \geq 2Kd$)

SNR/dB	12	14	16	18
DLDUNN	97.34%	97.77%	97.94%	97.78%
Black-box	84.50%	84.83%	87.23%	87.55%
SNR/dB	20	22	24	26
DLDUNN	97.84%	97.38%	97.62%	96.90%
Black-box	89.08%	89.02%	90.45%	90.64%

④ Performance comparison of DLDUNN and Black-box，K=1

(3) Generalization Analysis

(3.1) After being trained with a larger user number (K), the model can be directly applied to the system of smaller users without being retrained.
(3.2) Apply the trained model with SNR = 20dB to various SNR values for experiments.

SNR/dB	12	14	16	18
PDD (bps/Hz)	21.766	23.561	25.339	27.171
DLDUNN	97.41%	97.61%	97.43%	97.40%
SNR/dB	20	22	24	26
PDD (bps/Hz)	28.906	30.813	32.539	34.263
DLDUNN	97.82%	97.96%	97.77%	96.77%

⑤ Spectrum efficiency versus multi-users

(4) Computational Complexity Analysis

(4.1) The total number of iterations in DLDUNN is much smaller, i.e. $D_{in} \times D_{out}$(70 in this simulation) $\ll$ $P_{in} \times P_{out}$(generally larger than 10000).
(4.2) The main computations of the PDD algorithm and our proposed DLDUNN lie in each inner iteration and each inner layer, respectively. For PDD, the BCD-type algorithm requires to iterate for many times when optimizing V_{RF} and U_{RF} with complexity $O(N^2N_{RF}^2)$ and $O(KM^2M_{RF}^2)$, respectively, while in DLDUNN, we iterate it only once.
(4.3) DLDUNN replaces matrix inversion operations($O(N^3)$) with matrix multiplications ($O(N^{2.37})$). When N is large, it is efficient.
(4.4) No need to calculate the convergence variable $L_k(x)$

Contributions

(1) A general dual-layer deep-unfolding method is proposed. Its idea of reducing complexity and maintaining high performance is suitable for most high complexity communication algorithms (mostly double-layer loops). The “if-else” branch, bisection and BCD-type algorithm of PDD algorithm represent the highly complex situation of double-layer iterative algorithm, and DLDUNN has good robustness.

Algorithm 1 : General dual-layer iterative algorithm

Input: Initialize $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2\}$
Output: Optimized variables $\mathbf{X}^* = \{\mathbf{X}_1^*, \mathbf{X}_2^*\}$

repeat
  repeat
    $\mathbf{X}_1 = F_i(\mathbf{X}_1; \mathbf{X}_2)$
  until inner loop stopping criterion are met
  $\mathbf{X}_2 = F_t(\mathbf{X}_2; \mathbf{X}_1)$
until outer loop stopping criterion are met
return $\mathbf{X}_1, \mathbf{X}_2$

Dual-layer iterative algorithm framework

Algorithm 2 : General dual-layer deep-unfolding framework

Input: Initialize $\mathbf{X} = \{\mathbf{X}_1^{0,0}, \mathbf{X}_2^0\}$
Output: Optimized $\mathbf{X}^* = \{\mathbf{X}_1^{T,I}, \mathbf{X}_2^T\}$

for t from 0 to $T-1$
  for i from 0 to $I-1$
    $\mathbf{X}_1^{t,i+1} = \mathcal{F}_i(\mathbf{X}_1^{t,i}, \mathbf{X}_2^t; \Theta^{t,i})$
  end for
  $\mathbf{X}_2^{t+1} = \mathcal{F}_t(\mathbf{X}_1^{t,I}, \mathbf{X}_2^t; \Theta^t)$
end for
return $\mathbf{X}_1^{T,I}, \mathbf{X}_2^T$

Dual-layer deep-unfolding framework

(2) The transition from fully digital precoding to hybrid precoding is applicable to the 5G/6G trend of lower cost and commercialization.