# A Peaceman-Rachford Splitting Approach with Deep Equilibrium Network for Channel Estimation

Dingli Yuan[1], Shitong Wu[1], Lu Yang[2], Chenghui Peng[2], Haoran Tang[1], and Hao Wu[1†]

[1]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
[2]Wireless Technology Lab, Central Research Institute, 2012 Labs, Huawei Tech. Co. Ltd., China
Email: hwu@tsinghua.edu.cn

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. Multiple-input multiple-output (MIMO) is pivotal for wireless systems, yet its high-dimensional, stochastic channel poses significant challenges for accurate estimation, highlighting the critical need for robust channel estimation techniques. In this paper, we introduce a novel channel estimation method for the MIMO system. The main idea is to solve the dual form of the channel estimation problem, employing the Peaceman-Rachford (PR) splitting method with the deep equilibrium (DEQ) network. Specifically, the PR splitting method is applied to the Fenchel conjugate of the channel estimation functions to construct the non-expansive operators, which forms a fixed point equation of the optimal condition. The non-expansive operator is learned implicitly via the DEQ network and the fixed point equation is computed analytically using the learned operators, which further enables efficient iterative step with only a constant memory. Moreover, we provide a rigorous theoretical analysis using the monotone operator theory, offering the convergence proof of our proposed framework. Additionally, simulations of hybrid far- and near-field channels and channels generated by QuaDRiGa, demonstrate that our approach yields favorable results, indicating its ability to advance channel estimation in MIMO systems.**

## I. Introduction

Multiple-input multiple-output (MIMO) technology is pivotal for wireless systems [1]. Recognizing its paramount importance, the channel estimation algorithm should have high adaptability and performance for different types of channels in the MIMO systems, including the hybrid far- and near-field channels [2], [3] and the realistic channels generated by QuaDRiGa (QUAsi Deterministic RadIo channel GenerAtor [4]), etc. Hence, there is a strong motivation for improvements in channel estimation algorithms with broad applicability and high accuracy.

Traditional channel estimation schemes with closed-form expressions, such as least square (LS) and minimum mean-squared error (MMSE), have limitations in practical applications. The LS can not ensure a high estimation accuracy for different scenarios, while the MMSE has a high computation and deployment complexity [5], [6]. Guided by the pursuit of higher estimation accuracy and lower computation complexity, numerous iterative algorithms have been proposed, and can be classified into different categories. One of them is to add the regularization terms after least squares and utilize the optimization methods such as proximal gradient descent (PGD) or alternating direction method of multipliers (ADMM) [7]. Another one employs probabilistic models aimed at minimizing the mean square error (MSE), combining the channel prior information for algorithm design, including the classical approximate message passing (AMP) [8], the Orthogonal AMP (OAMP) [9], and the algorithms extended based on them.

Despite using different design principles, these algorithms share a similar iteration update rule, that each iteration is composed of a linear estimator (LE) and a non-linear estimator (NLE) [10]. The LE is explicit and can be designed with the low complexity. The bottleneck of a high performance estimator design lies in the NLE, since its own complexity is higher and its design requires enough channel prior information as accurate as possible, which is difficult to acquire [11]. A heuristic way is to replace NLE with neural network (NN) [12], [13], which has advantage in capturing data features. Additionally, deep neural networks, known as powerful denoisers [14], are well-suited to replace the NLE, which motivate the thriving of model-driven channel estimators.

Model-driven deep learning-based estimators, especially those employing deep unfolding [12], encounter several systematic challenges. These estimators are constructed by truncating a classical iterative algorithm into a predefined or fixed number of layers, denoted as $T$, and then replacing the NLE in layer $t$ with a deep neural network parameterized by $\Theta_t$ [12], [15]. However, this conventional deep unfolding formulation has critical issues [11], [16]. First, its scalability and generalization ability is poor. Second, its reliability is lack of theoretical guarantees, as truncating the algorithm into $T$ layers disrupts the convergence of a classical iterative algorithm. Third, its complexity is high, but its adaptiveness is low, as it requires full execution of $T$ layers with unreliable intermediate states. Considering these difficulties, a reconsideration of the feasibility of the deep unfolding framework is imperative. Recent works have consider the deep equilibrium (DEQ) Model with its application in inverse problem [10], image demonising [14], video snapshot compressive imaging [17] and hybrid far- and near-field channel estimation [3], [11]. In their work, they replace the NLE with NN and trained the model by computing the implicit gradient based

The first two authors contributed equally to this work and † marked the corresponding author. This work was partially supported by National Key Research and Development Program of China (2018YFA0701603) and National Natural Science Foundation of China (12271289 and 62231022).

on implicit function theorem [16], [18]. All of these work requires the estimation of the Lipschitz constant [19] to ensure a linear convergence rate. However, in real applications, the measurement matrix may be ill posed and these methods may lack theoretical foundations [3].

In this paper, we propose a novel approach for channel estimation in MIMO systems. Unlike the traditional approach of modeling channel estimation using the minimization problem with regularization terms, we take a different perspective by considering the dual problem. This choice ensures a stable convex model, enhancing robust convergence properties for different scenarios. Leveraging the advantageous convexity of the dual problem, we formulate the channel estimation algorithm using the Peaceman-Rachford (PR) splitting method. Specifically, we derive an alternative algorithm from the PR splitting algorithm for dual problem with explicit iteration and constructs a fixed point equation of an intermediate variable, ensuring the non-expansive property of the combined operator [20]. Due to the computational challenges associated with the nonlinear term in PR iteration, we employ NN to approximate this nonlinear term. We introduce a NN framework called Peaceman-Rachford (PR) splitting method implemented with Deep Equilibrium Network, (PR-DEN), which leverages the DEQ model for analytical computing the fixed point equation. This approach enables efficient iterative steps, with a same NN for each iteration, and thus the memory cost of our NN framework is a constant. Additionally, based on the monotone operator theory, our NN framework is supported by rigorous proofs of both convergence and optimality. Furthermore, extensive simulations including the hybrid far- and near-field channels in Terahertz (THz) ultra-massive MIMO (UM-MIMO) and the channels generated by QuaDRiGa validate the exceptional performance of our methodology. These promising results underscore the potential of our proposed approach to significantly advance channel estimation in MIMO systems.

## II. System Model and Problem Formulation

### A. General Problem Formulation

In this work, we consider the channel estimation problem in a MIMO system, where the received signal vector $\boldsymbol{y} \in \mathbb{C}^{M\times 1}$ is modeled as

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{h}\boldsymbol{s} + \mathbf{W}\bar{\boldsymbol{n}}, \tag{1}$$

Here, $\mathbf{W} \in \mathbb{C}^{M\times N}$ is the beamforming matrix. $\boldsymbol{h} \in \mathbb{C}^{N\times 1}$ represents the channel vector from the transmitter to the receiver. $\boldsymbol{s} \in \mathbb{C}^{1\times 1}$ is the transmitted signal, typically a pilot or a data symbol, and $\bar{\boldsymbol{n}} \in \mathbb{C}^{M\times 1}$ is the additive noise vector.

### B. Equivalent Formulation

In this subsection, we use regularization-based method to transform the channel estimation problem to a minimization problem. Taking $\mathbf{A} = \mathbf{W}\boldsymbol{s}$, $\boldsymbol{n} = \mathbf{W}\bar{\boldsymbol{n}}$, the received signal $\boldsymbol{y}$ can be transformed into the following real form:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{h} + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^{2M\times 1}$, $\mathbf{A} \in \mathbb{R}^{2M\times 2N}$, $\boldsymbol{h} \in \mathbb{R}^{2N}$, $\boldsymbol{n} \in \mathbb{R}^{2N}$ and $M$, $N$ are the length of the received signals and channels, respectively. Assuming that $\boldsymbol{h}$ is generated via a finite alphabet $S = \{\alpha_1, \cdots, \alpha_{|S|}\}$, our objective is to reconstruct $\boldsymbol{h}$ based on matrix $\mathbf{A}$ and received signal $\boldsymbol{y}$. Prior works such as [21], [22] have demonstrated that $\boldsymbol{h}$ is the unique solution to (2) if and only if it is the unique solution to the following problem:

$$\min_{\boldsymbol{h}\in\mathbb{R}^{2N}} \frac{1}{|S|}\sum_{i=1}^{|S|} \|\boldsymbol{h} - \alpha_i \mathbf{1}_{2N}\|_0, \quad \text{subject to} \quad \boldsymbol{y} = \mathbf{A}\boldsymbol{h} + \boldsymbol{n},$$

where $\mathbf{1}_{2N}$ denotes a $2N$-dimensional all-ones vector. This problem formulation is also equivalent to the regularized problem [22]:

$$\min_{\boldsymbol{h}\in\mathbb{R}^{2N_t}} \quad g(\boldsymbol{h}) + \frac{1}{2}\|\boldsymbol{y} - \mathbf{A}\boldsymbol{h}\|_2^2, \tag{3}$$

where the regularization function $g(\boldsymbol{h})$ encompasses the prior information of the channel characteristics [22], [23], such as the sparsity information $\alpha_i$ and the noise $\boldsymbol{n}$.

## III. A Peaceman-Rachford Splitting Approach with Deep Equilibrium Network

Solving (3) directly is difficult, since the regularization term $g(\boldsymbol{h})$ might face the following two challenges. First, this term might be non-convex, leading to no theoretical convergence guarantees of numerical algorithms. Second, this term might not have an explicit formulation, due to the limited information about the prior information of the channel characteristics, leading to no explicit iterative formulae by directly computation.

To overcome these difficulties, in this section, we introduce a framework employing the PR splitting approach for dual form to avoid the difficulty of the non-convexity, and implementing the DEQ network to learn prior information. Specifically, we consider the dual of the channel estimation problem (3) and apply the PR splitting method to find the solution of the derived dual problem [24], [25]. The convexity of the Fenchel conjugate ensures the convergence of the PR algorithm, which remains robust enough even when the regularization term $g(\boldsymbol{h})$ is non-convex.

Based on the PR splitting framework, we obtain the a fixed point equation consisting of only non-expansive operators, whose Lipschitz constant is not larger than one. This fixed point equation inspires us to utilize the DEQ model, which leverages a fixed point approach with a NN architecture where each layer is not changing across all the iterations. In this way, the proximal operator associated with the term $g(\boldsymbol{h})$ can be implicitly learned through the equilibrium formulation, and the learned operator is fixed for each iteration.

### A. Peaceman-Rachford Splitting Algorithm for Dual Problem

As presented in the previous discussion, the regularized term $g(\boldsymbol{h})$ in (3) encapsulates the prior information of channel characteristics. In order to solve the problem in a more stable and robust manner, we choose the dual form of (3) the above regularized problem. In the following, we denote $f(x) = \frac{1}{2}\|y - Ax\|_2^2$ for brevity, and denote $f^*$ and $g^*$ as the Fenchel conjugate of $f$ and $g$, respectively. Then, by

employing the Fenchel conjugate, we derive the dual of (3) as

$$\max_{\boldsymbol{x}\in\mathbb{R}^{2N_t}} \quad \{-g^*(-\boldsymbol{x}) - f^*(\boldsymbol{x})\}. \tag{4}$$

Here, the variable $\boldsymbol{x}$ is the corresponding dual variable and the derivation details are summarized in the appendix.

Traditional methodologies [3], [10] dealing with the channel estimation problem is based on the prime form (3), where the regularization term $g(\boldsymbol{h})$ may be non-convex, leading to noconvergence guarantees in theory. On the contrary, we aim to find the optimal solution of the dual problem (4). This choice can ensure a stable model, given that the Fenchel conjugate is always convex [7], thereby providing excellent robust convergence properties for different channel scenarios.

To solve the dual problem (4), the most important procedure is to construct efficient algorithm to compute the dual variable $\boldsymbol{x}$ under the formulation of Fenchel conjugate. Using first-order conditions, the optimal solution of (4) is equivalent to solving the following rooting problem [7], [26]:

$$0 \in \partial g^*(-\boldsymbol{x}) + \partial f^*(\boldsymbol{x}). \tag{5}$$

Taking $\mathbf{M}_1 = \partial(g^* \circ (-\mathbf{I}))$, $\mathbf{M}_2 = \partial(f^*)$, then $\mathbf{M}_1, \mathbf{M}_2$ are maximal monotone operators, since subgradients are maximal monotone operators [27]. Hence, the PR splitting method [24], [25] can be directly applied to solve the zeros of system (5), where the PR splitting method can be represented as:

$$\boldsymbol{x} = \mathbf{J}_{\sigma\mathbf{M}_2}\boldsymbol{\eta}, \quad \text{where } \boldsymbol{\eta} \text{ satisfies } \mathbf{R}_{\sigma\mathbf{M}_1}\mathbf{R}_{\sigma\mathbf{M}_2}\boldsymbol{\eta} = \boldsymbol{\eta}.$$

Here, $\sigma > 0$, $\mathbf{J}_{\sigma\mathbf{M}} = (\mathbf{I}+\sigma\mathbf{M})^{-1}$ is the resolvent of $\sigma\mathbf{M}$ and $\mathbf{R}_{\sigma\mathbf{M}} = 2\mathbf{J}_{\sigma\mathbf{M}} - \mathbf{I}$ is called the reflected resolvent [25], [26].

It is worth mentioning that the variable $\boldsymbol{\eta}$ in the fixed point equation (6) is an intermediate variable, which can usually be solved by fixed point iteration, i.e.,

$$\boldsymbol{\eta}^{k+1} = \mathbf{T}_{\mathbf{PR}}(\boldsymbol{\eta}^{\boldsymbol{k}}) = \mathbf{R}_{\sigma\mathbf{M}_1}\mathbf{R}_{\sigma\mathbf{M}_2}(\boldsymbol{\eta}^k) \tag{6}$$

In the following, we present our algorithm from the PR splitting algorithm for dual problem (4) with explicit iteration. The details of the derivation are provided in the appendix.

**Algorithm 1** PR Splitting for Dual Problem

1: **Input:** $\boldsymbol{x}^0 \in \mathbb{R}^n, \boldsymbol{p}^0 = \boldsymbol{x}^0$, and $\sigma > 0$.
2: **Initialize:** $\boldsymbol{\eta}^0 = \boldsymbol{x}^0 + \sigma\boldsymbol{p}^0$.
3: **for** $k = 0, 1, \ldots L-1$ **do**
4: $\quad \boldsymbol{q}^{k+1} = (\mathbf{A}^\top\mathbf{A} + \sigma\mathbf{I})^{-1}(\mathbf{A}^\top\boldsymbol{y} + \boldsymbol{\eta}^k)$
5: $\quad \boldsymbol{p}^{k+1} = \mathrm{Prox}_{\sigma^{-1}g}(\sigma^{-1}(2\sigma\boldsymbol{q}^{k+1} - \boldsymbol{\eta}^k))$.
6: $\quad \boldsymbol{x}^{k+1} = \boldsymbol{\eta}^k + \sigma\boldsymbol{p}^{k+1} - 2\sigma\boldsymbol{q}^{k+1}$.
7: $\quad \boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}^k + 2\sigma(\boldsymbol{p}^{k+1} - \boldsymbol{q}^{k+1})$.
8: **Output**: $\boldsymbol{x}^L$

Based on the derived algorithm, we prove the convergence of the sequences $\{\boldsymbol{x}^k\}$ and $\{\boldsymbol{h}^k\}$ to the corresponding optimal points of the dual and prime problem respectively.

**Theorem 1.** *The sequence $\{\boldsymbol{x}^k\}$ generated by Algorithm 1 converges to the optimal solution $\boldsymbol{x}^*$ of (4). Moreover, when the regularization term $g(\boldsymbol{h})$ in (3) is assumed to be convex, the sequence $\{\boldsymbol{p}^k\}$ generated by Algorithm 1 converges to the optimal solution $\boldsymbol{h}^*$ of (3) and the duality holds.*

*Proof.* The proof is based on the convergence property of the PR iteration [25], [28] for convex problem and the optimal condition of the system [27]. The detailed proof is moved to the appendix due to space limitation. □

### B. Deep Equilibrium Network

In the previous discussion, an unresolved issue was the specific form of the regularization term $g(\boldsymbol{h})$, which remains unknown, presenting challenges in directly applying the PR algorithm to solve the channel estimation problem. Therefore, our next step is utilizing NNs, which possess powerful feature extraction capabilities, to learn the regularization term.

Since neural networks process powerful abilities to learn statistical features, it is logical and achievable to learn this nonlinear term through neural networks. Similar to the deep unrolling approach of [12], [29], we consider replacing $\mathrm{prox}_{\sigma^{-1}g}$ with a trainable network $R_\theta$. Specifically, $R_\theta$ is constructed using a classical residual network [11], [30], which facilitates efficient training of deeper models by leveraging shortcut connections for learning residual mappings. The residual network is mainly composed of two residual blocks, each of which is activated by two convolution layers and two RELU functions.

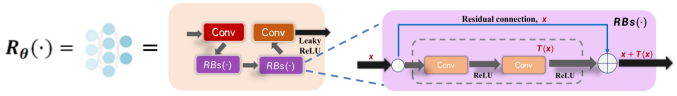



Fig. 1: The design of a CNN block $R_\theta$ to approximate the proximal operator, denoted by $R_\theta \to \mathrm{Prox}_{\sigma^{-1}g}$

We use the DEQ model to learn the network parameter $\theta$, since compared with traditional deep unfolding approaches that unfold distinct layers for multiple iterations [11], [17], the application of DEQ offers several advantages. First, the DEQ model uses a fixed NN for each iteration, corresponding to our fixed point iterative framework, while traditional approaches need different NNs for each layer of iteration, deviating from the fixed point framework and affecting the NN's stability due to varying depth. Second, the DEQ employs only a fixed single layer, which significantly reduces the network's memory requirements and the training overhead [16], [17].

The abstract iteration of $\boldsymbol{\eta}^k$ in DEQ can be represented as

$$\begin{aligned}\lim_{k\to+\infty} \boldsymbol{\eta}^{(k)} &= \lim_{k\to+\infty} f_\theta\left(\boldsymbol{\eta}^{(k)}; \boldsymbol{y}\right) \\ &\equiv \hat{\boldsymbol{\eta}} = f_\theta(\hat{\boldsymbol{\eta}}; \boldsymbol{y}),\end{aligned}$$

where $\hat{\boldsymbol{\eta}}$ denotes the fixed point in the network and $f_\theta$ denotes each iteration of the whole NNs with a fixed parameter $\theta$, which contains the the trainable network $R_\theta$.

To optimize network parameters $\theta$, stochastic gradient descent [3] is used to minimize a loss function as follows:

$$\theta^* = \arg\min_\theta \frac{1}{m}\sum_{i=1}^{m} \ell\left(f_\theta\left(\hat{\boldsymbol{x}}_i; \boldsymbol{y}_i, \right), \boldsymbol{h}_i^\star\right),$$

where $m$ is the number of training samples, $\hat{\boldsymbol{x}}_i$ denotes the fixed point generated by the network iteration, $\boldsymbol{h}_i^\star$ is the ground truth channel of the $i$-th training sample, and $\boldsymbol{y}_i$ is the paired measurement. $\ell(\cdot,\cdot)$ is the loss function, defined by the mean squared error (MSE), as

$$\ell\left(\hat{\boldsymbol{x}}, \boldsymbol{h}^\star\right) = \frac{1}{2}\left\|\hat{\boldsymbol{x}} - \boldsymbol{h}^\star\right\|_2^2.$$

Then, we calculate the loss gradient. Let $\ell$ be an abbreviation of $\ell\left(\hat{\boldsymbol{x}}, \boldsymbol{h}^\star\right)$, then the loss gradient is [16], [18]:

$$\frac{\partial \ell}{\partial \theta} = \left[\frac{\partial f_\theta(\hat{\boldsymbol{x}}; \boldsymbol{y})}{\partial \theta}\right]^\top \left[\boldsymbol{I} - \left.\frac{\partial f_\theta(\boldsymbol{x}; \boldsymbol{y})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}\right]^{-\top} (\hat{\boldsymbol{x}} - \boldsymbol{h}^\star),$$

where $^{-\top}$ denotes the inversion followed by transpose.

Using DEQ model, the memory complexity of our approach stays at $O(1)$ [17], independent of iteration count, and notably lower than deep unfolding's $O(T)$, enabling efficient gradient calculation for the loss term with minimal memory demand.

### C. Implementation and Convergence analysis

Based on the PR algorithm and the DEQ model, we utilize the fixed point equation from (6) to construct a fixed point network operator $f_\theta$. According to Algorithm 1, it can be seen that the algorithm consists of three linear iterations (steps 4,6,7) and one nonlinear iteration (step 4) within one loop. As the convolutional neural network (CNN) block $R_\theta$ to approximate the nonlinear proximal term, denoted as $R_\theta \rightarrow \mathrm{Prox}_{\sigma^{-1}g}$. Then, combined with other linear terms, the iteration of $\boldsymbol{\eta}^k$ in the DEQ network can be represented as

$$\boldsymbol{\eta}_R^{k+1} = f_\theta(\boldsymbol{\eta}_R^k) = f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ R_\theta \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R^k). \tag{7}$$

Here, $f_{\mathrm{LT}_i}$ denotes the $i$-th linear term (LT) in Algorithm 1, and $\boldsymbol{\eta}_R^k$ denotes the intermittent variable produced by the network. We call the whole framework as Peaceman-Rachford splitting method implemented with Deep Equilibrium Network (denoted as PR-DEN for short), and Fig.2 illustrates the iterative process of the PR-DEN approach.

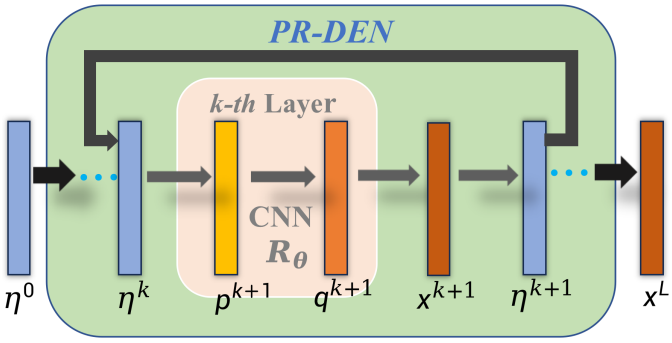



Fig. 2: Peaceman-Rachford splitting method implemented with Deep Equilibrium Network (PR-DEN) approach

Next, we focus on demonstrating the theoretical advantages of our network architecture. We can not only strictly guarantee its convergence theoretically, but also prove the optimality of the converged solution. Our proof is based on the observation that the reflected resolvent $\mathbf{R}_{\sigma\mathrm{M}}$ is nonexpansive [26, Corollary 23.11], i.e., the Lipschitz constant of $\mathbf{R}_{\sigma\mathrm{M}}$ is not larger one. Hence, its composition $\mathbf{T}_\sigma^{\mathbf{PR}}$ is also nonexpansive. Without loss of generality, we assume $\mathrm{Lip}(\mathbf{T}_\sigma^{\mathbf{PR}}) < 1$, since when $\mathrm{Lip}(\mathbf{T}_\sigma^{\mathbf{PR}}) = 1$, a perturbation parameter $\varepsilon$ can be added to construct a contraction mapping [26, Proposition 4.20] and the sequence of iteration converge to the same solution. Thus, we can claim that the intermediate variable $\boldsymbol{\eta}^k$ in the PR-DEN approach converges.

**Theorem 2.** *There exists a CNN $R_\theta(\cdot)$ which approximates the corresponding proximal operator, such that the sequence $\{\boldsymbol{\eta}_R^k\}$ produced by PR-DEN is strictly convergent, and a linear convergence rate is achieved.*

*Proof.* We utilize the Universal Approximation Theorem [31] to prove the existence of an network operator $R_\theta$ approximating the proximal term, and then employ the Banach Fixed Point Theorem to demonstrate convergence. The detailed proof is provided in the appendix. □

Up to now, we have proved that our PR-DEN approach is convergent, with a linear convergence rate. In the following, we will prove that the iterative solution produced by PR-DEN is convergent to the optimal solution of dual problem (4).

**Theorem 3.** *There exists a CNN $R_\theta$ sufficiently approximates $\mathrm{Prox}_{\sigma^{-1}g}$, such that $\boldsymbol{x}_R$ converges to $\boldsymbol{x}_*$, where $\boldsymbol{x}_R$ is the solution produced by PR-DEN, and $\boldsymbol{x}_*$ is the optimal solution of dual problem* (4).

*Proof.* We establish the convergence by analyzing the error between the solution given by PR-DEN and the optimal solution. The detailed proof can be found in the appendix. □

## IV. Numerical Results

This section presents the results of experimental simulations, to compare the performance of our PR-DEN with existing classical algorithms for channel estimation. Two distinct scenarios are considered: the hybrid far- and near-field channels in THz UM-MIMO framework and the QuaDRiGa-generated realistic channels in urban moblie environment.

Our performance metric is the Normalized Mean Square Error (NMSE), defined as

$$\mathrm{NMSE} = 10\log_{10}\frac{\|\mathbf{h} - \hat{\mathbf{h}}\|}{\|\mathbf{h}\|},$$

where $\mathbf{h}$ denotes the testing channel while $\hat{\mathbf{h}}$ is the estimated channel in the simulation.

We calculate the NMSE of comprehensive testing datasets comprising 5000 samples for both scenarios. The benchmarks include (a) the classical channel estimation methods with closed-form expression: **LS**, **MMSE** [6], (b) classical iterative algorithm including **OAMP** [9], **FISTA** [32], and (c) NN-based algorithms: **ISTA-Net+** [12] and **FPN-OAMP** [3].

The simulations of ISTA-Net+, FPN-OMAP and PR-DEN, implemented in Python, have been conducted on a Linux platform equipped with a pair of NVIDIA V100 Graphics Processing Units. LS, MMSE, OAMP, FISTA have been implemented by Matlab R2022a on a Linux platform with 128G RAM and one Intel(R) Xeon(R) Gold 5117 CPU@2.00GHz. The code is available on https://github.com/wushitong1234/PR-DEN.

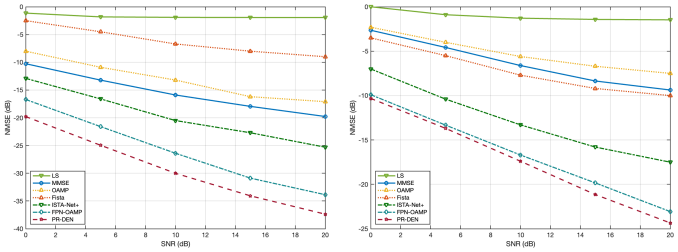
### A. Scenarios of MIMO with QuaDRiGa LOS Channels

In this scenario, the beamforming matrix is denoted by $\mathbf{W_{BB}}$, taken as a discrete Fourier transform (DFT) matrix in simulation. The concrete channel model is expressed as:

$$\boldsymbol{y} = \mathbf{W}_{\text{BB}}\boldsymbol{hs} + \mathbf{W}_{\text{BB}}\bar{\boldsymbol{n}},$$

where $\mathbf{h}$ is generated using QuaDRiGa [4].

In detail, we consider a Line-of-Sight (LOS) scenario that closely mirrors a realistic urban communication environment, with both circularly and linearly polarized antennas. These antennas are set to rotate through a range of angles from -150 to 150 degrees. The channel data is then randomly generated based on these angular positions.




(a) The QuaDRiGa Scenario. (b) The THz UM-MIMO Scenario.

Fig. 3: NMSE performance versus SNR in two scenarios, the curves represent: LS (Green), MMSE (Blue), OAMP (Orange), Fista (Dark Green), ISTA-Net+ (Red dashed), FPN-OAMP (Maroon dash-dot), and PR-DEN (Brown dash-dot).

Fig. 3 (a) illustrates the NMSE versus SNR of our PR-DEN and six existing algorithms in QuaDRiGa scenario. Compared to closed-form solutions such as LS and MMSE, and classical iterative algorithms like FISTA and OAMP, our method demonstrates significant advantages. Furthermore, when compared to the network-based iterative algorithm FPN-OAMP, our approach also shows a notable performance gain, with improvements ranging from 10% to 20% across various SNR levels.

### B. Scenarios of THz UM-MIMO with Hybrid Channels

In this scenario, the received signal $\mathbf{y}$ is modeled considering both digital and analog beamforming influences [3]:

$$\mathbf{y} = \mathbf{W}_{\text{BB}}\mathbf{W}_{\text{RF}}\mathbf{F}\tilde{\mathbf{h}}s + \mathbf{W}_{\text{BB}}\mathbf{W}_{\text{RF}}\bar{\mathbf{n}}.$$

Here, $\mathbf{W}_{\text{BB}}$ and $\mathbf{W}_{\text{RF}}$ are the digital and analog beamforming matrices, respectively, and $\mathbf{F}$ is a DFT matrix transforming the channel vector $\tilde{\mathbf{h}}$ into angular domain.

The hybrid near-field and far-field channels are generated according to the following formula, with the Rayleigh distance ($D_{Rayleigh}$) serving as the demarcation line:

$$a(\varphi_l, \xi_l, \rho_l) = \begin{cases} a_{far}(\varphi_l, \xi_l), & \text{if } \rho_l > D_{Rayleigh}, \\ a_{near}(\varphi_l, \xi_l, \rho_l), & \text{otherwise,} \end{cases}$$

where $a_{far}(\varphi_l, \xi_l)$ and $a_{near}(\varphi_l, \xi_l, \rho_l)$ correspond to the far- and near-field array responses, respectively.

More detailed simulation setting, including antenna configurations, signal propagation models, and hybrid far- and near-field field channel characteristics, are defined in [11].

Fig.3 (b) illustrates the NMSE performance versus SNR of our PR-DEN and six benchmarks in THz UM-MIMO scenario with hybrid channels. The results clearly illustrate that PR-DEN significantly outperforms LS and MMSE for different SNR values. Moreover, in comparison to network-based unfolding algorithms such as ISTA-Net+ and FPN-OAMP, we observe a notable improvement over ISTA-Net+, and approximately a 5% performance gain compared to FPN-OAMP. These results indicate that PR-DEN has superior accuracy on channel estimation under different wireless environments.

TABLE I: Running time of different algorithms

| | OAMP | FISTA | ISTA-Net+ | FPN-OAMP | FPDE-Net |
|---|---|---|---|---|---|
| UM-MIMO | 0.907 | 3.828 | 0.339 | 0.128 | 0.142 |
| QuaDRiGa | 0.0823 | 0.238 | 0.0125 | 0.00373 | 0.00418 |

Notes: a) The computation time (in seconds) for all algorithms is averaged over five instances at different SNR levels (0:5:20 dB). b) We set a uniform convergence criterion across all algorithms, i.e., $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| < \varepsilon$, pacifically $\varepsilon = 10^{-4}$ in QuaDRiGa and $\varepsilon = 10^{-2}$ in UM-MIMO

Table I illustrates a comparison on the computational times between six benchmarks and PR-DEN. Compared to traditional iterative algorithms such as OAMP and FISTA, PR-DEN demonstrates significantly lower computational time. In comparison to network unfolding algorithms, our approach exhibits lower execution time than ISTA-Net+ and is comparable to FPN-OAMP. These experimental findings validate the efficiency of our algorithm in handling the channel estimation problem.

## V. Conclusion

This paper introduces a novel channel estimation methodology for MIMO systems, addressing the dual problem through the Peaceman-Rachford (PR) splitting approach with the deep equilibrium (DEQ) network. This approach constructs non-expansive operators using PR splitting on the Fenchel conjugate of the channel estimation problem, forming a fixed point equation. The DEQ model implicitly learns the proximal operators, enabling efficient iterative steps with constant memory. Theoretical analysis based on monotone operator theory provides a convergence proof. The simulations demonstrate the approach's effectiveness, that our approach has high accuracy and efficiency for channel estimation in different scenarios, compared with six benchmarks.

## REFERENCES


[1] H. Sarieddeen, M.-S. Alouini, and T. Y. Al-Naffouri, "An overview of signal processing techniques for terahertz communications," *Proceedings of the IEEE*, vol. 109, no. 10, pp. 1628–1665, 2021.

[2] X. Wei and L. Dai, "Channel estimation for extremely large-scale massive mimo: Far-field, near-field, or hybrid-field?" *IEEE Communications Letters*, vol. 26, no. 1, pp. 177–181, 2021.

[3] W. Yu, Y. Shen, H. He, X. Yu, J. Zhang, and K. B. Letaief, "Hybrid far- and near-field channel estimation for thz ultra-massive mimo via fixed point networks," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 5384–5389.

[4] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "Quadriga: A 3-d multi-cell channel model with time evolution for enabling virtual field trials," *IEEE transactions on antennas and propagation*, vol. 62, no. 6, pp. 3242–3256, 2014.

[5] M. Tuchler, A. C. Singer, and R. Koetter, "Minimum mean squared error equalization using a priori information," *IEEE Transactions on Signal processing*, vol. 50, no. 3, pp. 673–683, 2002.

[6] Y. S. Cho, J. Kim, W. Y. Yang, and C. G. Kang, *MIMO-OFDM wireless communications with MATLAB*. John Wiley & Sons, 2010.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[8] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[9] J. Ma and L. Ping, "Orthogonal amp," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[10] D. Gilton, G. Ongie, and R. Willett, "Deep equilibrium architectures for inverse problems in imaging," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1123–1133, 2021.

[11] W. Yu, Y. Shen, H. He, X. Yu, S. Song, J. Zhang, and K. B. Letaief, "An adaptive and robust deep learning framework for thz ultra-massive mimo channel estimation," *IEEE Journal of Selected Topics in Signal Processing*, 2023.

[12] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1828–1837.

[13] 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.

[14] J. Liu, X. Xu, W. Gan, U. Kamilov *et al.*, "Online deep equilibrium learning for regularization by denoising," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 363–25 376, 2022.

[15] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "A model-driven deep learning network for mimo detection," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2018, pp. 584–588.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "Deep equilibrium models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] Y. Zhao, S. Zheng, and X. Yuan, "Deep equilibrium models for video snapshot compressive imaging," *arXiv preprint arXiv:2201.06931*, 2022.

[18] S. W. Fung, H. Heaton, Q. Li, D. McKenzie, S. Osher, and W. Yin, "Jfb: Jacobian-free backpropagation for implicit networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6648–6656.

[19] C. Pabbaraju, E. Winston, and J. Z. Kolter, "Estimating lipschitz constants of monotone deep equilibrium models," in *International Conference on Learning Representations*, 2020.

[20] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[21] A. Aissa-El-Bey, D. Pastor, S. M. A. Sbai, and Y. Fadlallah, "Sparsity-based recovery of finite alphabet solutions to underdetermined linear systems," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 2008–2018, 2015.

[22] R. Hayakawa and K. Hayashi, "Convex optimization-based signal detection for massive overloaded mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7080–7091, 2017.

[23] R. Sun, Y. Zhang, H. Zheng, J. Guo, J. Sun, and J. Xue, "A douglas-rachford splitting approach based deep network for mimo signal detection," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2023, pp. 1–6.

[24] G. Zhang, Y. Yuan, and D. Sun, "An efficient hpr algorithm for the wasserstein barycenter problem with $O(\mathrm{Dim}(P)/\varepsilon)$ computational complexity," *arXiv preprint arXiv:2211.14881*, 2022.

[25] P. L. Combettes, "The geometry of monotone operator splitting methods," *arXiv preprint arXiv:2310.08443*, 2023.

[26] H. H. Bauschke, P. L. Combettes, and et al., *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011, vol. 408.

[27] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970, vol. 18.

[28] A. Davydov, S. Jafarpour, A. V. Proskurnikov, and F. Bullo, "Non-euclidean monotone operator theory with applications to recurrent neural networks," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 6332–6337.

[29] M. Mardani, Q. Sun, D. Donoho, V. Papyan, H. Monajemi, S. Vasanawala, and J. Pauly, "Neural proximal gradient descent for compressive imaging," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] P. Kidger and T. Lyons, "Universal approximation with deep narrow networks," in *Conference on learning theory*. PMLR, 2020, pp. 2306–2327.

[32] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

## APPENDIX

### DERIVATION FOR THE DUAL

the problem (3) is equivalently expressed by introducing slackness variable $\boldsymbol{p}, \boldsymbol{q}$:

$$\begin{aligned} \min_{\boldsymbol{p},\boldsymbol{q}} \quad & g(\boldsymbol{p}) + f(\boldsymbol{q}) \\ \text{s.t.} \quad & \boldsymbol{p} - \boldsymbol{q} = 0. \end{aligned}$$

The Lagrangian $\mathcal{L}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{x})$ for this problem with the Lagrange multiplier $\boldsymbol{x}$ (also known as the dual variable) is given by:

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{x}) = g(\boldsymbol{p}) + f(\boldsymbol{q}) + \boldsymbol{x}^\top(\boldsymbol{p} - \boldsymbol{q}).$$

Minimizing the Lagrangian over $\boldsymbol{p}$ and $\boldsymbol{q}$ gives:

$$\inf_{\boldsymbol{p},\boldsymbol{q}} \mathcal{L}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{x}) = \inf_{\boldsymbol{p}}(g(\boldsymbol{p}) + \boldsymbol{x}^\top\boldsymbol{p}) + \inf_{\boldsymbol{q}}(f(\boldsymbol{q}) - \boldsymbol{x}^\top\boldsymbol{q}).$$

The Fenchel conjugates $f^*$ and $g^*$ are defined as:

$$f^*(\boldsymbol{x}) = \sup_{\boldsymbol{y}}(\boldsymbol{x}^\top\boldsymbol{y} - f(\boldsymbol{y})), \quad g^*(\boldsymbol{x}) = \sup_{\boldsymbol{y}}(\boldsymbol{x}^\top\boldsymbol{y} - g(\boldsymbol{y})).$$

Thus, the dual problem is:

$$\max_{\boldsymbol{x}\in\mathbb{R}^{2N_t}} \left\{-g^*(-\boldsymbol{x}) - f^*(\boldsymbol{x})\right\}.$$

### DERIVATION OF ALGORITHM 1

We prove this proposition by induction. For $k = 0$, we use the definition of $\boldsymbol{q}^1$ in Algorithm 1, yielding the condition $0 \in \partial f(\boldsymbol{q}^1) - \boldsymbol{\eta}^0 + \sigma\boldsymbol{q}^1$. Thus, $\boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1 \in \partial f(\boldsymbol{q}^1)$. Invoking Theorem 23.5 in [27], we deduce that

$$\boldsymbol{q}^1 \in \partial f^*(\boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1).$$

Hence, $\sigma\boldsymbol{q}^1 = \sigma\partial f^*(\boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1)$. This means that

$$\boldsymbol{\eta}^0 \in \boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1 + \sigma\partial f^*(\boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1).$$

Denote that $\mathbf{M}_2 = \partial f^*$, it follows that

$$\boldsymbol{\eta}^0 \in \boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1 + \sigma\mathbf{M}_2(\boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1),$$

Let $\boldsymbol{w}^k \triangleq \mathbf{J}_{\sigma\mathbf{M}_2}(\boldsymbol{\eta}^{k-1})$, which implies

$$\boldsymbol{w}^1 = J_{\sigma\mathbf{M}_2}(\boldsymbol{\eta}^0) = \boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1.$$

In parallel, by the stipulation of $\boldsymbol{p}^1$, we observe that

$$0 \in \partial g(\boldsymbol{p}^1) + (\boldsymbol{\eta}^0 - 2\sigma\boldsymbol{q}^1 + \sigma\boldsymbol{p}^1).$$

It follows from Theorem 23.5 in [27] that

$$\boldsymbol{p}^1 \in \partial g^*(-(\boldsymbol{\eta}^0 - 2\sigma\boldsymbol{q}^1 + \sigma\boldsymbol{p}^1)).$$

Hence, $-\sigma\boldsymbol{p}^1 = -\sigma\partial g^*(-(\boldsymbol{\eta}^0 - 2\sigma\boldsymbol{q}^1 + \sigma\boldsymbol{p}^1)$, implying that $2(\boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1) - \boldsymbol{\eta}^0 \in 2(\boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1) - \boldsymbol{\eta}^0 + \sigma\boldsymbol{p}^1 - \sigma\partial g^*(-(\boldsymbol{\eta}^0 - 2\sigma\boldsymbol{q}^1 + \sigma\boldsymbol{p}^1))$.

Since $\mathbf{M}_1 = \partial(g^* \circ (-\mathbf{I}))$, we have

$$2(\boldsymbol{\eta}^0 - \sigma\boldsymbol{q}^1) - \boldsymbol{\eta}^0 \in \boldsymbol{\eta}^0 - 2\sigma\boldsymbol{q}^1 + \sigma\boldsymbol{p}^1 + \sigma M_1(\boldsymbol{\eta}^0 - 2\sigma\boldsymbol{q}^1 + \sigma\boldsymbol{p}^1),$$

which implies $\boldsymbol{x}^1 := \mathbf{J}_{\sigma\mathbf{M}_1}(2\boldsymbol{w}^1 - \boldsymbol{\eta}^0) = \boldsymbol{\eta}^0 - 2\sigma\boldsymbol{q}^1 + \sigma\boldsymbol{p}^1$. Hence, we have $\boldsymbol{\eta}^1 := \boldsymbol{\eta}^0 + 2(\boldsymbol{x}^1 - \boldsymbol{w}^1) = \boldsymbol{\eta}^0 + 2\sigma(\boldsymbol{p}^1 - \boldsymbol{q}^1)$.

Then, it follows that

$$\begin{aligned} \boldsymbol{\eta}^1 &= \boldsymbol{\eta}^0 + 2(\boldsymbol{x}^1 - \boldsymbol{w}^1) = \boldsymbol{\eta}^0 + 2(\mathbf{J}_{\sigma\mathbf{M}_1}(2\boldsymbol{w}^1 - \boldsymbol{\eta}^0) - \mathbf{J}_{\sigma\mathbf{M}_2}(\boldsymbol{\eta}^0)) \\ &= \boldsymbol{\eta}^0 + 2(\mathbf{J}_{\sigma\mathbf{M}_1}(2\mathbf{J}_{\sigma\mathbf{M}_2}(\boldsymbol{\eta}^0)) - \mathbf{J}_{\sigma\mathbf{M}_1}(-\boldsymbol{\eta}^0) - \mathbf{J}_{\sigma\mathbf{M}_2}(\boldsymbol{\eta}^0)) \\ &= (2\mathbf{J}_{\sigma\mathbf{M}_1} - \mathbf{I})(2\mathbf{J}_{\sigma\mathbf{M}_2} - \mathbf{I})(\boldsymbol{\eta}^0) = \mathbf{R}_{\sigma\mathbf{M}_1}\mathbf{R}_{\sigma\mathbf{M}_2}\boldsymbol{\eta}^0. \end{aligned}$$

It follows that the update of $\boldsymbol{\eta}^1$ is the same as our Algorithm. Hence, we prove the statement for $k = 0$. Assume that the statement holds for some $k \geq 1$. For $k := k+1$, we can prove that the statement holds similarly to the case $k = 0$. Thus, we prove the statement holds for any $k \geq 0$ by induction.

### PROOF OF THEOREM 1

Since $\mathrm{Fix}\left(\mathbf{T}_\sigma^{\mathrm{PR}}\right)$ is a closed convex set [26, Corollary 4.24], the fixed point is unique. Using the convergence result of [28, Theorem 22], we can directly obtain that the sequence $\{\boldsymbol{\eta}^k\}$ converges to the fixed point $\boldsymbol{\eta}^*$ of the PR iteration. Since the resolvent $\boldsymbol{J}_{\sigma M_2}$ is nonexpansive and $\boldsymbol{w}^{k+1} = \boldsymbol{J}_{\sigma M_2}\left(\eta^k\right)$ as derived in the proof of Proposition 1, we directly obtain

$$\lim_k \boldsymbol{x}^k = \lim_k \boldsymbol{J}_{\sigma M_2}\left(\boldsymbol{\eta}^k\right) = \boldsymbol{J}_{\sigma M_2}\left(\boldsymbol{\eta}^*\right).$$

As show in [20], if $\boldsymbol{\eta}^*$ is the fixed point of the PR iteration, then $\boldsymbol{x}^* = \boldsymbol{J}_{\sigma M_2}\left(\boldsymbol{\eta}^*\right)$. Hence, we prove the convergence of sequence $\{\boldsymbol{x}^k\}$ to the optimal solution $\boldsymbol{x}^*$ of (4).

Next, we prove the convergence of the sequence $\left\{\boldsymbol{q}^k\right\}$. According to Algorithm 1, we have for all $k \geq 0$,

$$\boldsymbol{q}^{k+1} = (\mathbf{A}^\top\mathbf{A} + \sigma\mathbf{I})^{-1}(\mathbf{A}^\top\boldsymbol{y} + \boldsymbol{\eta}^k).$$

Denote $\hat{f}(\boldsymbol{q}) := f(\boldsymbol{q}) + \frac{\sigma}{2}\left\|\boldsymbol{q}\right\|^2$, which is a strongly convex function. Thus, $\hat{f}^*$ is essentially smooth [27, Theorem 26.3]. The first-order optimality condition of the above iteration of $\boldsymbol{q}^{k+1}$ implies $0 \in \partial\hat{f}\left(\boldsymbol{q}^{k+1}\right) - \boldsymbol{\eta}^k$. Since $\hat{f}$ is a proper closed convex function, by [27, Theorem 23.5], the first-order optimality condition is equivalent to

$$\boldsymbol{q}^{k+1} = \nabla\hat{f}^*\left(\boldsymbol{\eta}^k\right).$$

It follows from the convergence of $\left\{\boldsymbol{\eta}^k\right\}$ and the continuity of $\nabla\hat{f}^*$ [27, Theorem 25.5] that $\left\{\boldsymbol{q}^k\right\}$ is convergent. Note that

$$\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}^k + 2\sigma(\boldsymbol{p}^{k+1} - \boldsymbol{q}^{k+1})$$

from Algorithm 1, and the sequence $\{\boldsymbol{\eta}^k\}$ is convergent, we obtain $\lim_k(\boldsymbol{p}^k - \boldsymbol{q}^k) = 0$. Then, by taking the limit, we obtain $\lim_k \boldsymbol{p}^k = \lim_k \boldsymbol{q}^k = \boldsymbol{q}^*$. Hence, $\{\boldsymbol{p}_k\}$ is convergent.

Assume that $(\boldsymbol{p}^*, \boldsymbol{q}^*, \boldsymbol{x}^*)$ is the limit point of the sequence $\left\{\boldsymbol{p}^k, \boldsymbol{q}^k, \boldsymbol{x}^k\right\}$. Since $\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}^k + 2\sigma(\boldsymbol{p}^{k+1} - \boldsymbol{q}^{k+1})$ from Algorithm 1, we can obtain $\boldsymbol{p}^* - \boldsymbol{q}^* = \mathbf{0}$ by taking the limit. It follows that

$$\lim_k \boldsymbol{w}^k = \lim_k[\boldsymbol{w}^k + \sigma(\boldsymbol{p}^k - \boldsymbol{q}^k)] = \lim_k \boldsymbol{x}^k = \boldsymbol{x}^*.$$

By Algorithm 1 and the derivation in Proposition 1, we have

$$0 \in \partial f(\boldsymbol{q}^{k+1}) - \boldsymbol{\eta}^k + \sigma\boldsymbol{q}^{k+1}.$$

Since $\boldsymbol{w}^{k+1} = \boldsymbol{\eta}^k - \sigma\boldsymbol{q}^{k+1}$, we have

$$\boldsymbol{w}^{k+1} \in \partial f\left(\boldsymbol{q}^{k+1}\right).$$

Similarly, by Algorithm 1 and the derivation details in Proposition 1, we have

$$0 \in \partial g(\boldsymbol{p}^{k+1}) + (\boldsymbol{\eta}^k - 2\sigma\boldsymbol{q}^{k+1} + \sigma\boldsymbol{p}^{k+1}).$$

Since $\boldsymbol{x}^{k+1} = \boldsymbol{\eta}^k - 2\sigma\boldsymbol{q}^{k+1} + \sigma\boldsymbol{p}^{k+1}$, we have

$$-\boldsymbol{x}^{k+1} \in \partial g(\boldsymbol{p}^{k+1}).$$

Hence, we have

$$\boldsymbol{w}^{k+1} \in \partial f\left(\boldsymbol{q}^{k+1}\right), \quad -\boldsymbol{x}^{k+1} \in \partial g\left(\boldsymbol{p}^{k+1}\right),$$

Together with $\boldsymbol{p}^* - \boldsymbol{q}^* = \mathbf{0}$ and taking limit, we have

$$\boldsymbol{x}^* \in \partial f\left(\boldsymbol{q}^*\right), \quad -\boldsymbol{x}^* \in \partial g\left(\boldsymbol{p}^*\right), \quad \boldsymbol{p}^* - \boldsymbol{q}^* = 0.$$

This completes the proof [27, Corollary 28.3.1].

## PROOF OF THEOREM 2

As indicated above, the iteration formula for $\boldsymbol{\eta}^k$ in the proposed PR-DEN can be succinctly represented as:

$$\boldsymbol{\eta}_R^{k+1} = f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ R_\theta \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R^k).$$

That is,

$$\begin{aligned}\boldsymbol{\eta}_R^{k+1} =& f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ (R_\theta - \mathrm{Prox}_{\sigma^{-1}g} + \mathrm{Prox}_{\sigma^{-1}g}) \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R^k) \\ =& f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ (R_\theta - \mathrm{Prox}_{\sigma^{-1}g}) \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R^k) \\ &+ f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ \mathrm{Prox}_{\sigma^{-1}g} \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R^k).\end{aligned}$$

From the proof of proposition 1, we obtain

$$\mathbf{T}_{\mathbf{PR}}(\boldsymbol{\eta}_R^k) = f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ \mathrm{Prox}_{\sigma^{-1}g} \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R^k).$$

Thus, we derive

$$\boldsymbol{\eta}_R^{k+1} = f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ (R_\theta - \mathrm{Prox}_{\sigma^{-1}g}) \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R^k) + \mathbf{T}_{\mathbf{PR}}(\boldsymbol{\eta}_R^k).$$

By the Universal approximation theorem [31], there exists a network $R_\theta(\cdot)$ approximating the proximal operator $\mathrm{Prox}_{\sigma^{-1}g}$ with arbitrary precision. Suppose $\max\{\|f_{\mathrm{LT}_1}\|, \|f_{\mathrm{LT}_2}\|, \|f_{\mathrm{LT}_3}\|\} = \alpha$, $\mathrm{Lip}(\mathbf{T}_{\mathbf{PR}}) = \beta$. There exists $R_\theta$, such that

$$\|(R_\theta - \mathrm{Prox}_{\sigma^{-1}g})(\boldsymbol{\eta}_R^k)\| < \frac{1-\beta}{\alpha^3}\|\boldsymbol{\eta}_R^k\|.$$

Thus, combined (A) with (A), we derive

$$\|\boldsymbol{\eta}_R^{k+1}\| < (1-\beta)\|\boldsymbol{\eta}_R^k\| + \beta\|\boldsymbol{\eta}_R^k\| = \|\boldsymbol{\eta}_R^k\|.$$

From Banach fix point theorem, $\{\boldsymbol{\eta}_R^k\}$ is convergent.

## PROOF OF THEOREM 3

From Theorem 2, we derive the convergence of $\{\boldsymbol{\eta}_R^k\}$, which denotes the sequence produced by PR-DEN. That is

$$\lim_{k\to\infty} \boldsymbol{\eta}_R^k = \boldsymbol{\eta}_R,$$

where $\boldsymbol{\eta}_R$ is the solution of the following fix point equation:

$$\boldsymbol{\eta}_R = f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ R_\theta \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R).$$

Additionally, since the dual problem (4) is convex, the optimal solution $\mathbf{x}_*$ is the equivalent to the solution of the following problem [20]:

$$\boldsymbol{x}_* = \mathbf{J}_{\sigma\mathbf{M_2}}(\boldsymbol{\eta}_*) \quad \text{s.t.} \quad \boldsymbol{\eta}_* = \mathbf{R}_{\sigma\mathbf{M}_1}\mathbf{R}_{\sigma\mathbf{M}_2}(\boldsymbol{\eta}_*).$$

Thus, subtracting (A) from (A), we obtain

$$\boldsymbol{\eta}_R - \boldsymbol{\eta}_* = \mathbf{T}_{\mathbf{PR}}(\boldsymbol{\eta}_R - \boldsymbol{\eta}_*) + f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ (R_\theta - \mathrm{Prox}_{\sigma^{-1}g}) \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R).$$

Suppose $\max\{\|f_{\mathrm{LT}_1}\|, \|f_{\mathrm{LT}_2}\|, \|f_{\mathrm{LT}_3}\|\} = \alpha$, $\mathrm{Lip}(\mathbf{T}_{\mathbf{PR}}) = \beta$, then

$$\|\boldsymbol{\eta}_R - \boldsymbol{\eta}_*\| \le \beta\|\boldsymbol{\eta}_R - \boldsymbol{\eta}_*\| + \alpha^3\|(R_\theta - \mathrm{Prox}_{\sigma^{-1}g}\|\|\boldsymbol{\eta}_R)\|.$$

Hence, we have

$$\|\boldsymbol{\eta}_R - \boldsymbol{\eta}_*\| \le \frac{\alpha^3}{1-\beta}\|(R_\theta - \mathrm{Prox}_{\sigma^{-1}g})\|\|\boldsymbol{\eta}_R\|.$$

We claim there exists a consistent upper bound $M$ on $\boldsymbol{\eta}_R$.

$$\|\boldsymbol{\eta}_R\| \le \|f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ R_\theta \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta}_R)\| + 1. \tag{8}$$

By the Universal approximation theorem [31], there exists a network $R_\theta(\cdot)$ approximating the proximal operator $\mathrm{Prox}_{\sigma^{-1}g}$ with arbitrary precision. This indicates that for any $\varepsilon > 0, \boldsymbol{\eta} \in \mathbb{X}$, we have

$$\|f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ (R_\theta - \mathrm{Prox}_{\sigma^{-1}g}) \circ f_{\mathrm{LT}_1}(\boldsymbol{\eta})\| < \epsilon$$

when $R_\theta$ sufficiently approximates $\mathrm{Prox}_{\sigma^{-1}g}$.

It demonstrates that

$$\mathrm{Lip}(f_{\mathrm{LT}_3} \circ f_{\mathrm{LT}_2} \circ R_\theta \circ f_{\mathrm{LT}_1}) < 1.$$

Combined with (8), it convinces the fact that $\|\boldsymbol{\eta}_R\|$ is bounded. Here, we denote the upper bound as $M$. Thus,

$$\|\boldsymbol{\eta}_R - \boldsymbol{\eta}_*\| \le \frac{\alpha^3 M}{1-\beta}\|(R_\theta - \mathrm{Prox}_{\sigma^{-1}g})\|.$$

Since $R_\theta \to \mathrm{Prox}_{\sigma^{-1}g}$, it can be indicated that

$$\boldsymbol{\eta}_R \to \boldsymbol{\eta}_*.$$

Therefore,

$$\boldsymbol{x}_R = \mathbf{J}_{\sigma\mathbf{M}_1}(\boldsymbol{\eta}_R) \to \mathbf{J}_{\sigma\mathbf{M}_1}(\boldsymbol{\eta}_*) = \boldsymbol{x}_*.$$