
Optimization-Derived Learning with Essential Convergence Analysis of Training and Hyper-training

Risheng Liu^{1,2,3} Xuan Liu^{1,2} Shangzhi Zeng⁴ Jin Zhang^{5,6} Yixuan Zhang⁵

Abstract

Recently, Optimization-Derived Learning (ODL) has attracted attention from learning and vision areas, which designs learning models from the perspective of optimization. However, previous ODL approaches regard the training and hyper-training procedures as two separated stages, meaning that the hyper-training variables have to be fixed during the training process, and thus it is also impossible to simultaneously obtain the convergence of training and hyper-training variables. In this work, we design a Generalized Krasnoselskii-Mann (GKM) scheme based on fixed-point iterations as our fundamental ODL module, which unifies existing ODL methods as special cases. Under the GKM scheme, a Bilevel Meta Optimization (BMO) algorithmic framework is constructed to solve the optimal training and hyper-training variables together. We rigorously prove the essential joint convergence of the fixed-point iteration for training and the process of optimizing hyper-parameters for hyper-training, both on the approximation quality, and on the stationary analysis. Experiments demonstrate the efficiency of BMO with competitive performance on sparse coding and real-world applications such as image deconvolution and rain streak removal.

1. Introduction

There have been a number of methods to handle learning and vision tasks, and conventional ones utilize either classic optimization or machine learning schemes directly. Using earlier optimizers to solve manually designed objective functions arises two problems: the task-related objective functions might be hard to solve, and not able to accurately model actual tasks. In recent years, a series of approaches called Optimization-Derived Learning (ODL) have emerged, combining the ideas of optimization and learning, and leverage optimization techniques to establish learning methods (Chen et al., 2021; Monga et al., 2021; Shlezinger et al., 2020; He et al., 2021; Hutter et al., 2019; Yang et al., 2022). The fundamental idea of ODL is to incorporate trainable learning modules into an optimization process and then learn the corresponding parameterized models from collected training data. Therefore, ODL aims to not only possess the convergence guarantee of optimization methods, but also achieve satisfactory practical performance with the help of neural networks. Overall speaking, ODL has two goals in divergent directions, i.e., to train an algorithmic scheme to minimize the given objective faster or to minimize the reconstruction error between the established model and the actual task. These two goals correspond to two processes in ODL, called training and hyper-training. For training, we aim to solve the optimization model (minimize the objective and find the optimal training variables), while for hyper-training, we aim to find the optimal hyper-parameters (hyper-training variables) to characterize the task.

1.1. Related Works

Over the past years, ODL approaches have been established based on various numerical optimization schemes and parameterization strategies (e.g., numerical hyper-parameters and network architectures) (Feurer & Hutter, 2019; Thornton et al., 2013; Schuler et al., 2015; Chen & Pock, 2016). They have been widely applied to all kinds of learning and vision tasks, based on classic optimization schemes like Proximal Gradient or Iterative Shrinkage-Thresholding Algorithm (PG or ISTA) (Daubechies et al., 2004) and Alternating Direction Method of Multiplier (ADMM) (Boyd et al., 2011). For example, Learned ISTA (LISTA) (Gre-

¹DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, Liaoning, China. ²Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, Liaoning, China. ³Peng Cheng Laboratory, Shenzhen, Guangdong, China. ⁴Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada. ⁵Department of Mathematics, SUSTech International Center for Mathematics, Southern University of Science and Technology, Shenzhen, Guangdong, China. ⁶National Center for Applied Mathematics Shenzhen, Shenzhen, Guangdong, China. Correspondence to: Jin Zhang <zhangj9@sustech.edu.cn>.

gor & LeCun, 2010; Chen et al., 2018), Differentiable Linearized ADMM (DLADMM) (Xie et al., 2019), and DUBLID (Li et al., 2020) can be applied to image deconvolution; ISTA-Net (Zhang & Ghanem, 2018) can be applied to CS reconstruction; ADMM-Net (Yang et al., 2017) can be applied to CS-MRI; PADNet (Liu et al., 2019a) can be applied to image haze removal; FIMA (Liu et al., 2019b) can be applied to image restoration; and Plug-and-Play ADMM (Ryu et al., 2019; Venkatakrishnan et al., 2013; Chan et al., 2016) can be applied to image super resolution.

According to divergent starting points and corresponding types of employed schemes, we can roughly divide existing ODL into two main categories, called ODL based on Unrolling with Numerical Hyper-parameters (UNH) and ODL Embedded with Network Architectures (ENA), respectively. UNH starts from a traditional optimization process for training, and aims to unroll the iteration with learnable hyper-parameters. Thus, it utilizes parameterized numerical algorithms to optimize the determinate objective function. In the early stages, many classic optimization schemes have been designed based on theories and experiences, such as Gradient Descent (GD), ISTA, Augmented Lagrangian Method (ALM), ADMM, and Linearized ADMM (LADMM) (Lin et al., 2011). Parameters within these schemes can be regarded as hyper-parameters, and then learned via unrolling. There are also some works which introduce learnable modules into the optimization process to solve the determinate objective function faster, such as LBS (Liu et al., 2018a), GCM (Liu et al., 2018b), and TLF (Liu et al., 2020a). As for ENA, it starts from an optimization objective function, but it aims to design its network architecture to solve specific tasks. It replaces part of the original training iterations with (or just directly incorporates) trainable architectures to approach the ground truth more efficiently, so the relationship between the final and original model is unclear. In LISTA and DLADMM, they learn matrices as linear layers of networks. There are also some methods employing networks in the classical sense, such as ISTA-Net and Plug-and-Play ADMM. Very lately, pre-trained CNN-based modules such as DPSR and DPIR (Nguyen et al., 2017; Zhang et al., 2019; 2020; Yuan et al., 2020; Ahmad et al., 2020; Song et al., 2020) are implemented to achieve image restoration.

However, UNH cannot reduce the gap between the artificially designed determinate objective function and the actual task; while for ENA, the embedded complex network modules make the iterative trajectory and convergence property difficult to analyze. In addition, UNH and ENA have a common and inevitable flaw. As aforementioned, ODL aims to minimize the objective function for training and to minimize the reconstruction error for hyper-training. However, existing ODL methods can only achieve one of these two goals once, which means the training and hyper-training procedures have to be separated into two independent stages. In

the first stage, hyper-parameters as hyper-training variables are determined by hyper-training, while in the second stage, they are fixed and substituted into the training iterations to find the optimal training variables. Therefore, the optimal hyper-training and training variables cannot be obtained simultaneously. This feature makes existing ODL methods inflexible, and since they ignore the intrinsic relationship between hyper-training and training variables, the obtained solution may not be the true one (Liu et al., 2020b), especially when the optimal training variables are not unique.

From the viewpoint of theory, although the empirical efficiency of ODL has been witnessed in applications, research on solid convergence analysis is still in its infancy. This gap makes the broader usability of ODL questionable. Recently, some works have tried to provide convergence analysis on ODL methods through classic optimization tools. In particular, (Chan et al., 2016; Teodoro et al., 2018; Sun et al., 2019) analyze the non-expansiveness of incorporated trainable architectures under the boundedness assumption of the embedded network. (Ryu et al., 2019) requires that the network residual admits a Lipschitz constant strictly smaller than one. However, these methods can only guarantee the convergence towards some fixed points of the approximated model, instead of the solution to the original task. A naive strategy to ensure the convergence is to learn as few parameters (hyper-training variables) as possible, such as to learn nothing but the step size of ISTA (Ablin et al., 2019). Besides, when learning a neural network, we need additional artificially designed corrections. For example, (Liu et al., 2019b; Moeller et al., 2019; Heaton et al., 2020) use neural networks and optimization algorithms to generate temporary updates and manually design rules to select true updates. The restrictions seriously limit these methods. In addition, most previous ODL approaches are designed specially based on a specific optimization scheme, making their convergence theory hard to be extended to other methods.

1.2. Our Contributions

As mentioned above, existing ODL methods, containing UNH and ENA, can only handle learning and vision tasks by optimizing hyper-training variables and training variables separately, raising a series of problems. Besides, they only focus on some specific problems with special structures. In dealing with these issues, in this work, we construct the Generalized Krasnoselskii-Mann (GKM) scheme as a new and general ODL formulation from the perspective of fixed-point iterations. This implicitly defined scheme is more flexible than traditional optimization models, and includes more methods than existing ODL. After that we establish the Bilevel Meta Optimization (BMO) algorithmic framework to simultaneously solve the training and hyper-training tasks, inspired by the leader-follower game (Von Stackelberg, 2010; Liu et al., 2021). Then the process of training

under the fixed-point iterations is to solve the lower-level training variables, while the process of hyper-training is to find the optimal upper-level hyper-training variables. In this way, we can incorporate training and hyper-training together and obtain the true optimal solution even if the fixed points are not unique. A series of theoretical properties are strictly proved to guarantee the essential joint convergence of training and hyper-training variables, both on the approximation quality and on the stationary analysis. We also conduct experiments on various learning and vision tasks to verify the effectiveness of BMO. Our contributions are summarized as follows.

- We establish a new ODL formulation from the perspective of fixed-point iterations under the GKM scheme, which introduces learnable parameters as hyper-training variables. Serving as a general form of various ODL methods, the implicitly defined scheme not only contains existing ENA and UNH methodologies, but also produces combined models.
- Based on the GKM scheme, the BMO algorithm provides a leader-follower mechanism to incorporate the process of training and hyper-training. Unlike existing ODL methods separating these two sub-tasks, our method can optimize training and hyper-training variables simultaneously, making it possible to investigate their intrinsic relationship to obtain the true solution.
- To our best knowledge, this is the first work that provides strict essential convergence analysis of both training and hyper-training variables, containing the analysis on approximation quality and stationary convergence. This is what existing ODL methods cannot achieve since they regard training and hyper-training as independent procedures.

2. The Proposed Algorithmic Framework

In this section, we first present a general ODL platform by generalizing the classical fixed-point scheme (Edelstein & O'Brien, 1978), and existing ODL methods (UNH and ENA) can be regarded as special cases of our scheme. Then by considering the processes of training and hyper-training from meta optimization (Liu et al., 2021), we put forward our Bilevel Meta Optimization (BMO) algorithm.

2.1. Generalized Krasnoselskii-Mann Scheme for ODL

Here we propose the Generalized Krasnoselskii-Mann (GKM) learning scheme as a general form of ODL methods. To begin with, we consider the following optimization model as the training process:

$$\min_{\mathbf{u} \in U} f(\mathbf{u}), \text{ s.t. } \mathcal{A}(\mathbf{u}) = \mathbf{y}, \quad (1)$$

where $\mathbf{u} \in U$ is the training variable, $f(\mathbf{u})$ is the objective function related to the task, and $\mathcal{A}(\mathbf{u}) = \mathbf{y}$ is a necessary linear constraint (\mathcal{A} is a linear operator). Denote $\mathcal{D}(\cdot)$ as the corresponding non-expansive algorithmic operator. Then to solve Eq. (1) is to iterate for the fixed point of $\mathcal{D}(\cdot)$. Thus the model in Eq. (1) can be transformed into

$$\mathbf{u} \in \text{Fix}(\mathcal{D}(\cdot)), \quad (2)$$

where $\text{Fix}(\cdot)$ represents the set of fixed points. This serves as our fundamental training scheme, and is actually more general than Eq. (1), because it not only contains traditional optimization models, but also represents other implicitly defined ODL models and implicit networks (Fung et al., 2022), which are designed based on optimization but further added with learnable modules.

This process can be implemented via the following classical Krasnoselskii-Mann (KM) updating scheme (Reich & Zaslavski, 2000; Borwein et al., 1992), whose k -th iteration step is $\mathcal{T}(\mathbf{u}^k) = \mathbf{u}^k + \alpha(\mathcal{D}(\mathbf{u}^k) - \mathbf{u}^k)$, where $\alpha \in (0, 1)$, and $\mathcal{D}(\cdot)$ is the operator for some basic numerical methods such as GD, PG, and ALM. It can be observed that here \mathcal{T} is an α -averaged non-expansive operator.

In this work, we further generalize the classical KM schemes to the following Generalized KM (GKM) learning scheme:

$$\mathcal{T}(\mathbf{u}^k, \boldsymbol{\omega}) = \mathbf{u}^k + \alpha(\mathcal{D}(\mathbf{u}^k, \boldsymbol{\omega}) - \mathbf{u}^k), \quad (3)$$

where $\mathcal{D}(\cdot, \boldsymbol{\omega}) \in \{\mathcal{D}_{\text{num}}(\cdot, \boldsymbol{\omega}) \circ \mathcal{D}_{\text{net}}(\cdot, \boldsymbol{\omega})\}$. Here \circ represents compositions of operators, \mathcal{D}_{num} denotes numerical operators in traditional optimization schemes, and \mathcal{D}_{net} denotes iterative handcrafted network architectures. The variability of hyper-parameters $\boldsymbol{\omega}$ as hyper-training variables makes the scheme more flexible. These hyper-training variables correspond to different parameters for various ODL methods. For example, in classic optimization schemes, $\boldsymbol{\omega}$ denotes step size; in LISTA or DLADMM, $\boldsymbol{\omega}$ comes from differentiable proximal operators, thresholds, support selection or penalties; in other methods with network modules, such as LBS, GCM, TLF, ISTA-Net, Plug-and-Play ADMM, and pre-trained CNN, $\boldsymbol{\omega}$ denotes network parameters. Note that in order to guarantee the non-expansiveness of \mathcal{D} , we utilize some normalization techniques on these parameters, such as spectral normalization (Miyato et al., 2018). By specifying \mathcal{D} as different types of \mathcal{D}_{num} and \mathcal{D}_{net} , we can contain diverse ODL methods (e.g., UNH, ENA, and their combinations) within our GKM scheme. Examples of specific forms of operator \mathcal{D} can be found in Appendix B.

2.2. Bilevel Meta Optimization for Training and Hyper-training

As mentioned in Section 1.1, existing ODL methods separately consider training and hyper-training as two independent stages. Hence, they fail to investigate the intrinsic

sis relationship between training variables \mathbf{u} and hyper-training variables ω . In this work, we would like to utilize the perspective from meta optimization (Neumüller et al., 2011) to incorporate these two coupled processes of training and hyper-training. Thus, the optimal training and hyper-training variables can be obtained simultaneously. We formulate this meta optimization task within the leader-follower game framework. In the leader-follower (or Stackelberg) game, the leader commits to a strategy, while the follower observes the leader’s commitment and decides how to play after that.

Specifically in our task, by recognizing hyper-training variables ω and training variables \mathbf{u} as the leader and follower respectively, we have that with ω , the iteration module \mathcal{T} is determined, from which \mathbf{u} finds the fixed point of \mathcal{T} . Therefore, the leader ω optimizes its objective via the best response of the follower \mathbf{u} (Liu et al., 2021). In order to clearly describe this hierarchical relationship, we introduce the following bilevel formulation to model ODL:

$$\min_{\mathbf{u} \in U, \omega \in \Omega} \ell(\mathbf{u}; \omega), \text{ s.t. } \mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)), \quad (4)$$

where ℓ denotes the objective function for measuring the performance of the training process, which usually is set to be the loss function. Thus, the upper level corresponds to the hyper-training process. Hereafter, we call this formulation as Bilevel Meta Optimization (BMO). Note that this is more general than a traditional bilevel optimization problem, since the lower level for training is the solution mapping of a broader fixed-point iteration as mentioned for Eq. (2) in Section 2.1.

BMO can overcome several issues in existing ODL methods. When updating ω in the upper level for hyper-training, instead of only using the information from original data, BMO allows us to utilize the task-related priors by updating ω and \mathbf{u} simultaneously. On the other hand, in theory, BMO makes us able to analyze the essential joint convergence of ω and \mathbf{u} under their nested relationship, rather than only consider the convergence of \mathbf{u} . Thus we are able to obtain the true optimal solution of the problem, instead of just the fixed points of the lower iteration, especially when the fixed points are not unique. We will give the detailed convergence analysis in Section 3.

Now we establish the solution strategy to solve \mathbf{u} and ω simultaneously under BMO. Inspired by (Liu et al., 2020b; 2022), in order to make efficient use of the hierarchical information contained in Eq. (4), we update variables from divergent perspectives, aggregating the information from training and hyper-training.

First, we give the optimization direction \mathbf{v}_l from the perspective of training process in the lower level of Eq. (4). We use $\mathcal{T}(\cdot, \omega)$ to update \mathbf{u}^k at the k -th step, That is, to solve $\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega))$, we define $\mathbf{v}_l^k = \mathcal{T}(\mathbf{u}^{k-1}, \omega)$.

Subsequently, we further give the descent direction based on the hyper-training process in the upper level of Eq. (4). To make the updating direction contain the information of hyper-training variables ω , a simple idea is to directly use the gradient of loss function ℓ with respect to \mathbf{u} . However, the gradient descent of ℓ may destroy the non-expansive property. Hence, we add an additional positive-definite correction matrix \mathbf{H}_ω parameterized by ω , and set the step size as a decreasing sequence (i.e., $s_k \rightarrow 0$) to ensure the correctness of the descent direction, i.e.,

$$\mathbf{v}_u^k = \mathbf{u}^{k-1} - s_k \mathbf{H}_\omega^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^{k-1}, \omega). \quad (5)$$

Next, we aggregate the two iterative directions and introduce a projection operator to generate the final updating direction:

$$\mathbf{u}^k = \text{Proj}_{U, \mathbf{H}_\omega} (\mu \mathbf{v}_u^k + (1 - \mu) \mathbf{v}_l^k). \quad (6)$$

Here $\text{Proj}_{U, \mathbf{H}_\omega}(\cdot)$ is the projection operator associated to \mathbf{H}_ω and is defined as $\text{Proj}_{U, \mathbf{H}_\omega}(\mathbf{u}) = \text{argmin}_{\bar{\mathbf{u}} \in U} \|\bar{\mathbf{u}} - \mathbf{u}\|_{\mathbf{H}_\omega}$. Note that the projection $\text{Proj}_{U, \mathbf{H}_\omega}$ is only for ensuring the boundedness of \mathbf{u}^k in the theoretical analysis. In practice, U is usually chosen as a sufficiently large bounded set (e.g., \mathbb{R}^n), and thus the projection operator can be ignored. Finally, after K updates of the variable \mathbf{u} , we update the hyper-training variables ω by gradient descent. Note that $\frac{\partial}{\partial \omega} \ell(\mathbf{u}^K, \omega)$ can be obtained by automatic differential efficiently (Franceschi et al., 2017). Finally, we summarize the full BMO solution strategy in Algorithm 1.

Algorithm 1 The Solution Strategy of BMO

Require: Step sizes $\{s_k\}$, γ and parameter μ

- 1: Initialize ω^0 .
- 2: **for** $t = 1 \rightarrow T$ **do**
- 3: Initialize \mathbf{u}^0 .
- 4: **for** $k = 1 \rightarrow K$ **do**
- 5: $\mathbf{v}_l^k = \mathcal{T}(\mathbf{u}^{k-1}, \omega^{t-1})$.
- 6: $\mathbf{v}_u^k = \mathbf{u}^{k-1} - s_k \mathbf{H}_\omega^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^{k-1}, \omega^{t-1})$.
- 7: $\mathbf{u}^k = \text{Proj}_{U, \mathbf{H}_\omega} (\mu \mathbf{v}_u^k + (1 - \mu) \mathbf{v}_l^k)$.
- 8: **end for**
- 9: $\omega^t = \omega^{t-1} - \gamma \frac{\partial}{\partial \omega} \ell(\mathbf{u}^K, \omega^{t-1})$.
- 10: **end for**

3. Joint Convergence Analysis

In this section, we discuss the essential convergence analysis of our proposed BMO algorithm (Algorithm 1) for the GKM scheme, towards the optimal solution and stationary points of optimization problem in Eq. (4) with respect to both \mathbf{u} and ω . Note that this joint convergence of training and hyper-training also provides a unified theoretical guarantee for existing ODL methods containing UNH and ENA. Complete theoretical analysis is stated in Appendix A.

By introducing an auxiliary function, the problem in Eq. (4) can be equivalently rewritten as the following

$$\min_{\omega \in \Omega} \varphi(\omega), \quad \text{where } \varphi(\omega) := \inf_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)) \cap U} \ell(\mathbf{u}, \omega). \quad (7)$$

The sequence $\{\omega^t\}$ generated by BMO (Algorithm 1) actually solves the following approximation problem of Eq. (4)

$$\min_{\omega \in \Omega} \varphi_K(\omega) := \ell(\mathbf{u}^K(\omega), \omega), \quad (8)$$

where $\mathbf{u}^K(\omega)$ is derived by solving the problem $\inf_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)) \cap U} \ell(\mathbf{u}, \omega)$ and can be given by

$$\begin{cases} \mathbf{v}_l^k(\omega) = \mathcal{T}(\mathbf{u}^{k-1}(\omega), \omega), \\ \mathbf{v}_u^k(\omega) = \mathbf{u}^{k-1}(\omega) - s_k \mathbf{H}_\omega^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^{k-1}, \omega), \\ \mathbf{u}^k(\omega) = \text{Proj}_{U, \mathbf{H}_\omega}(\mu \mathbf{v}_u^k(\omega) + (1 - \mu) \mathbf{v}_l^k(\omega)), \end{cases} \quad (9)$$

where $k = 1, \dots, K$.

3.1. Approximation Quality and Convergence

In this part, we will show that Eq. (8) is actually an appropriate approximation to Eq. (4) in the sense that any limit point $(\bar{\mathbf{u}}, \bar{\omega})$ of the sequence $\{(\mathbf{u}^K(\omega^K), \omega^K)\}$ with $\omega^K \in \text{argmin}_{\omega \in \Omega} \varphi_K(\omega)$ is a solution to the bilevel problem in Eq. (4). Thus we can obtain the optimal solution of Eq. (4) by solving Eq. (8). We make the following standing assumption throughout this part.

Assumption 3.1 Ω is a compact set and U is a convex compact set. $\text{Fix}(\mathcal{T}(\cdot, \omega))$ is nonempty for any $\omega \in \Omega$. $\ell(\mathbf{u}, \omega)$ is continuous on $\mathbb{R}^n \times \Omega$. For any $\omega \in \Omega$, $\ell(\cdot, \omega) : \mathbb{R}^n \rightarrow \mathbb{R}$ is L_ℓ -smooth, convex and bounded below by M_0 .

Please notice that ℓ is usually defined to be the MSE loss, and thus Assumption 3.1 is quite standard for ODL (Ryu et al., 2019; Zhang et al., 2020). We first present some necessary preliminaries. For any two matrices $\mathbf{H}_1, \mathbf{H}_2 \in \mathbb{R}^{n \times n}$, we consider the following partial ordering relation:

$$\mathbf{H}_1 \succ \mathbf{H}_2 \Leftrightarrow \langle \mathbf{u}, \mathbf{H}_1 \mathbf{u} \rangle \geq \langle \mathbf{u}, \mathbf{H}_2 \mathbf{u} \rangle, \quad \forall \mathbf{u} \in \mathbb{R}^n. \quad (10)$$

If $\mathbf{H} \succ 0$, $\langle \mathbf{u}_1, \mathbf{H} \mathbf{u}_2 \rangle$ for $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n$ defines an inner product on \mathbb{R}^n . Denote the induced norm with $\|\cdot\|_{\mathbf{H}}$, i.e., $\|\mathbf{u}\|_{\mathbf{H}} := \sqrt{\langle \mathbf{u}, \mathbf{H} \mathbf{u} \rangle}$ for any $\mathbf{u} \in \mathbb{R}^n$. Denote the graph of operator $\mathcal{D}(\cdot, \omega)$ to be

$$\text{gph } \mathcal{D}(\cdot, \omega) := \{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \mathbf{v} = \mathcal{D}(\mathbf{u}, \omega)\}.$$

We assume $\mathcal{D}(\cdot, \omega)$ satisfies the following assumption throughout this part.

Assumption 3.2 There exist $\mathbf{H}_{ub} \succeq \mathbf{H}_{lb} \succ 0$, such that for each $\omega \in \Omega$, there exists $\mathbf{H}_{ub} \succeq \mathbf{H}_\omega \succeq \mathbf{H}_{lb}$ such that

(1) $\mathcal{D}(\cdot, \omega)$ is non-expansive with respect to $\|\cdot\|_{\mathbf{H}_\omega}$, i.e., for all $(\mathbf{u}_1, \mathbf{u}_2) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$\|\mathcal{D}(\mathbf{u}_1, \omega) - \mathcal{D}(\mathbf{u}_2, \omega)\|_{\mathbf{H}_\omega} \leq \|\mathbf{u}_1 - \mathbf{u}_2\|_{\mathbf{H}_\omega}.$$

(2) $\mathcal{D}(\cdot, \omega)$ is closed, i.e., $\text{gph } \mathcal{D}(\cdot, \omega)$ is closed.

Under Assumption 3.2, we obtain the non-expansive property of $\mathcal{T}(\cdot, \omega)$ defined in Eq. (3) from (Bauschke et al., 2011)[Proposition 4.25] immediately. Then we can show that the sequence $\{\mathbf{u}^k(\omega)\}$ generated by Eq. (9) not only converges to the solution set of $\inf_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)) \cap U} \ell(\mathbf{u}, \omega)$ but also admits a uniform convergence towards the fixed-point set $\text{Fix}(\mathcal{T}(\cdot, \omega))$ with respect to $\|\mathbf{u}^k(\omega) - \mathcal{T}(\mathbf{u}^k(\omega), \omega)\|_{\mathbf{H}_{lb}}^2$ for $\omega \in \Omega$.

Theorem 3.1 Let $\{\mathbf{u}^k(\omega)\}$ be the sequence generated by Eq. (9) with $\mu \in (0, 1)$ and $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{lb})}{L_\ell})$, where $\lambda_{\min}(\mathbf{H}_{lb})$ denotes the smallest eigenvalue of matrix \mathbf{H}_{lb} . Then, we have for any $\omega \in \Omega$,

$$\lim_{k \rightarrow \infty} \text{dist}(\mathbf{u}^k(\omega), \text{Fix}(\mathcal{T}(\cdot, \omega))) = 0,$$

$$\lim_{k \rightarrow \infty} \ell(\mathbf{u}^k(\omega), \omega) = \varphi(\omega).$$

Furthermore, there exists $C > 0$ such that for any $\omega \in \Omega$,

$$\|\mathbf{u}^k(\omega) - \mathcal{T}(\mathbf{u}^k(\omega), \omega)\|_{\mathbf{H}_{lb}}^2 \leq C \sqrt{\frac{1 + \ln(1+k)}{k^{\frac{1}{4}}}}.$$

Thanks to the uniform convergence property of the sequence $\{\mathbf{u}^k(\omega)\}$, inspired by the arguments used in (Liu et al., 2022), we can establish the convergence on both \mathbf{u} and ω of BMO (Algorithm 1) towards the solution of optimization problem in Eq. (4) in the following theorem.

Theorem 3.2 Let $\{\mathbf{u}^k(\omega)\}$ be the sequence generated by Eq. (9) with $\mu \in (0, 1)$ and $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{lb})}{L_\ell})$. Then, let $\omega^K \in \text{argmin}_{\omega \in \Omega} \varphi_K(\omega)$, and we have

(1) any limit point $(\bar{\mathbf{u}}, \bar{\omega})$ of the sequence $\{(\mathbf{u}^K(\omega^K), \omega^K)\}$ is a solution to the problem in Eq. (4), i.e., $\bar{\omega} \in \text{argmin}_{\omega \in \Omega} \varphi(\omega)$ and $\bar{\mathbf{u}} = \mathcal{T}(\bar{\mathbf{u}}, \bar{\omega})$.

(2) $\inf_{\omega \in \Omega} \varphi_K(\omega) \rightarrow \inf_{\omega \in \Omega} \varphi(\omega)$ as $K \rightarrow \infty$.

3.2. Stationary Analysis

Here we provide the convergence analysis of our algorithm with respect to stationary points, i.e., any limit point $\bar{\omega}$ of the sequence $\{\omega^K\}$ satisfies $\nabla \varphi(\bar{\omega}) = 0$, where $\varphi(\omega)$ is defined in Eq. (7). We consider the following assumptions. For $\mathcal{D}(\cdot, \omega)$ we request a stronger assumption than Assumption 3.2 on contractive property.

Assumption 3.3 Ω is compact and $U = \mathbb{R}^n$. $\text{Fix}(\mathcal{T}(\cdot, \omega))$ is nonempty for any $\omega \in \Omega$. $\ell(\mathbf{u}, \omega)$ is twice continuously differentiable on $\mathbb{R}^n \times \Omega$. For any $\omega \in \Omega$, $\ell(\cdot, \omega) : \mathbb{R}^n \rightarrow \mathbb{R}$ is L_ℓ -smooth, convex and bounded below by M_0 .

Assumption 3.4 $\mathcal{D}(\cdot, \omega)$ is contractive w.r.t. $\|\cdot\|_{\mathbf{H}, \omega}$.

Denote $\hat{S}(\omega) := \text{argmin}_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)) \cap U} \ell(\mathbf{u}, \omega)$, and we have the following stationary result.

Theorem 3.3 Suppose Assumptions 3.2, 3.3 and 3.4 are satisfied, $\frac{\partial}{\partial \mathbf{u}} \mathcal{T}(\mathbf{u}, \omega)$ and $\frac{\partial}{\partial \omega} \mathcal{T}(\mathbf{u}, \omega)$ are Lipschitz continuous with respect to \mathbf{u} , and $\hat{S}(\omega)$ is nonempty for all $\omega \in \Omega$. Let $\{\mathbf{u}^k(\omega)\}$ be the sequence generated by Eq. (9) with $\mu \in (0, 1)$ and $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{1b})}{L_\ell})$. Let ω^K be an ε_K -stationary point of $\varphi_K(\omega)$, i.e., $\|\nabla \varphi_K(\omega^K)\| = \varepsilon_K$. Then if $\varepsilon_K \rightarrow 0$, we have that any limit point $\bar{\omega}$ of the sequence $\{\omega^K\}$ is a stationary point of φ , i.e., $\nabla \varphi(\bar{\omega}) = 0$.

4. Applications for Learning and Vision

Here we demonstrate some examples on how to address various real-world learning and vision applications under our BMO framework by jointly optimizing the training and hyper-training variables. In sparse coding and image deconvolution, we model the task to be sparsification of coefficients as training, so the training variable \mathbf{u} is model parameters; in rain streak removal, the task is to decompose the background and rain streak layer as a generalized training process, so \mathbf{u} is the model output (a clear image). In all tasks, we embed learnable iterations to yield better results (closer to ground truth or target) than directly minimizing the original objective function. More detailed information of the operator \mathcal{D} can be found in Appendix B, containing the proof that \mathcal{D} in our applications satisfies the assumptions.

Sparse Coding. Sparse coding has become a popular technique to extract features from raw data (Zhang et al., 2015). The difficulty within is to recover the sparse vector from the noisy data and ill-posed transform matrix. To be specific, given the input dataset $\mathbf{b} \in \mathbb{R}^m$ and transform matrix $\mathbf{Q} \in \mathbb{R}^{m \times n}$, our goal is to find the model parameters representation $\mathbf{u}_1 \in \mathbb{R}^n$ and noise $\mathbf{u}_2 \in \mathbb{R}^n$ as training variables such that they satisfy the following model $\mathbf{Q}\mathbf{u}_1 + \mathbf{u}_2 = \mathbf{b}$, and the representation \mathbf{u}_1 and noise \mathbf{u}_2 are expected to be sparse enough, i.e., $\kappa\|\mathbf{u}_1\|_1 + \|\mathbf{u}_2\|_1$ is minimized. Hence, the optimization problem in Eq. (1) can be formulated as

$$\min_{\mathbf{u}} \kappa\|\mathbf{u}_1\|_1 + \|\mathbf{u}_2\|_1, \text{ s.t. } \mathbf{Q}\mathbf{u}_1 + \mathbf{u}_2 = \mathbf{b}. \quad (11)$$

Classic numerical optimization methods usually use ADMM with gradient descent and projection operations to solve Eq. (11). UNH attempts to learn descent directions that solve the problem faster and better than ADMM; ENA starts

from a traditional numerical algorithm for an optimization objective function and replaces the projection operator with a functionally similar network, thus solving the specific task (rather than solve the original objective) and maintaining interpretability. Notice that in comparison to some methods which intend to learn effective \mathbf{Q}^\top for each update of variables, BMO use the same learnable module each time to update variables, which decouples our parameter quantity from the number of updates. As for the descent operation, we consider $\mathcal{D}_{\text{DLADMM}}$ as \mathcal{D}_{net} .

Image Deconvolution For practical applications, we first consider image deconvolution (image deblurring) (Richardson, 1972; Andrews & Hunt, 1977), a typical low-level vision task as a branch of image restoration, whose purpose is to recover the clean image from the blurred one. The input image can be expressed as $\mathbf{b} = \mathbf{Q} * \mathbf{u} + \mathbf{n}$, where \mathbf{Q} , \mathbf{u} , and \mathbf{n} respectively represent the blur kernel, latent clean image, and additional noise, and $*$ denotes the two-dimensional convolution operator. Here we apply regularization methods based on Maximum A Posteriori (MAP) estimation. Then the problem can be expressed as $\min_{\mathbf{u} \in U} \|\mathbf{Q} * \mathbf{u} - \mathbf{b}\|_2^2 + \Phi(\mathbf{u})$, where $\Phi(\mathbf{u})$ is the prior function of the image. Since the image after wavelet transform is usually sparse, we consider $\Phi(\mathbf{u}) = \kappa\|\mathbf{W}\mathbf{u}\|_1$, where \mathbf{W} is the wavelet transform matrix. The objective function is

$$\min_{\mathbf{u}} \|\mathbf{Q} * \mathbf{u} - \mathbf{b}\|_2^2 + \kappa\|\mathbf{W}\mathbf{u}\|_1. \quad (12)$$

Hence, similar to sparse coding, based on the wavelet transform model, we learn a wavelet coefficients as the training variables for image deconvolution. For handling this problem, we composite \mathcal{D}_{PG} as \mathcal{D}_{num} with handcrafted network architectures DRUNet in DPIR (Zhang et al., 2020) as \mathcal{D}_{net} .

Rain Streak Removal. With signal-dependent or signal-independent noise, images captured under rainy conditions often suffer from weak visibility (Wang et al., 2019a; Li et al., 2019). The rain streak removal task aims to decompose an input rainy image \mathbf{b} into a rain-free background \mathbf{u}_b and a rain streak layer \mathbf{u}_r , i.e., $\mathbf{b} = \mathbf{u}_b + \mathbf{u}_r$, and hence enhances its visibility. This problem is ill-posed since the dimension of unknowns \mathbf{u}_b and \mathbf{u}_r to be recovered is twice as many as that of the input \mathbf{b} . The problem can be reformulated as the following optimization problem

$$\min_{\mathbf{u}_b, \mathbf{u}_r} \frac{1}{2} \|\mathbf{u}_b + \mathbf{u}_r - \mathbf{b}\|_2^2 + \psi_b(\mathbf{u}_b) + \psi_r(\mathbf{u}_r), \quad (13)$$

where $\psi_b(\mathbf{u}_b)$ denotes the priors on the background layer and $\psi_r(\mathbf{u}_r)$ represents the priors on rain streak layer. Here we set $\psi_b(\mathbf{u}_b) = \kappa_b\|\mathbf{u}_b\|_1, \psi_r(\mathbf{u}_r) = \kappa_r\|\nabla \mathbf{u}_r\|_1$. By introducing auxiliary variables \mathbf{v}_b and \mathbf{v}_r , the problem in Eq. (1)

is specified as

$$\begin{aligned} \min_{\mathbf{u}_b, \mathbf{u}_r, \mathbf{v}_b, \mathbf{v}_r} \quad & \frac{1}{2} \|\mathbf{u}_b + \mathbf{u}_r - \mathbf{b}\|_2^2 + \kappa_b \|\mathbf{v}_b\|_1 + \kappa_r \|\mathbf{v}_r\|_1, \\ \text{s.t.} \quad & \mathbf{v}_b = \mathbf{u}_b, \mathbf{v}_r = \nabla \mathbf{u}_r, \end{aligned} \quad (14)$$

where $\nabla = [\nabla_h; \nabla_v]$ denotes the gradient in horizontal and vertical directions. The training variables in this task are output images \mathbf{u}_b and \mathbf{u}_r based on the simple summation model $\mathbf{b} = \mathbf{u}_b + \mathbf{u}_r$. UNH usually uses ALM to solve Eq. (14). ENA starts from ALM and uses the network to approximate the solution of the sub-problems. Here we consider \mathcal{D}_{ALM} as \mathcal{D}_{num} and the designed iterable network architectures RCDNet (Wang et al., 2020) as \mathcal{D}_{net} .

5. Experimental Results

In this section, we illustrate the performance of BMO on sparse coding, image deconvolution and rain streak removal tasks. More detailed parameter setting and network architectures can be found in Appendix C.

5.1. Sparse Coding

We first investigate the performance in sparse coding, and compare BMO with LADMM and DLADMM, respectively as an example of UNH and ENA. Note that here we do not compare with pure networks, because we focus on convergence behaviors which pure networks cannot guarantee. Following the setting in (Chen et al., 2018), we experiment on the classic Set14 dataset, in which salt-and-pepper noise is added to 10% pixels of each image. Furthermore, the rectangle of each image is divided into non-overlapping patches of size 16×16 . We use the patchdictionary method (Xu & Yin, 2014) to learn a 256×512 dictionary \mathbf{Q} .

Table 1 shows the PSNR and SSIM results. It can be seen that the performance of our BMO on both PSNR and SSIM is superior than LADMM and DLADMM. In Figure 1, we compare the visual quality of denoised images on the Baboon image, and the quality of image recovered by BMO is visibly higher than that of LADMM and DLADMM. This is because LADMM can only learn few hyper-parameters (such as the step size) for not destroying convergence, and the structure of DLADMM is restricted by the number of layers, leading to a distance from the real fixed-point model. However, thanks to the hybrid strategy to incorporate training and hyper-training, BMO allows more hyper-training variables to improve the performance.

In Figure 2, we further analyze the convergence behavior of hyper-training variables ω in $\varphi_K(\omega)$ defined in Eq. (8). Although the final convergence of φ by the three methods is close, it can be seen from the gradient curve that BMO outperforms in convergence speed and stability. DLADMM can ensure the convergence of the upper loss function φ , but

Table 1. PSNR and SSIM results for sparse coding on Set14.

Methods	Layers	PSNR	SSIM
DLADMM	5	15.59±0.81	0.52±0.13
	25	15.64±0.87	0.52±0.13
LADMM	5	10.47±2.36	0.41±0.14
	25	11.31±2.29	0.41±0.15
BMO	5	18.82±1.59	0.63±0.16
	25	18.98±2.53	0.65±0.15

the convergence speed is slow. LADMM performs poorly in the convergence of the gradient of upper loss function φ .

Then, we verify the convergence of training variables \mathbf{u} in the lower fixed-point iteration with iterations of \mathbf{u} for testing in Figure 3. Note that for DLADMM, the number of iterations for training have to be more than those for testing, so in the right column we only show the curves of LADMM and BMO. It can be observed that BMO performs better in convergence stability and convergence speed with the increasing of iterations for testing. LADMM converges fast at first, but it cannot further improve the convergence performance due to too few hyper-training variables. DLADMM has slow convergence speed because its number of hyper-parameters and network structure are restricted.

In the right column of Figure 3, we show the convergence curve when the number of iterations for testing is more than those for training to further verify the stability and non-expansiveness of the learned lower iterative module. Still, as can be seen, BMO is superior to LADMM, and the mapping learned by BMO on testing data can indeed continue to converge in the iterations beyond training steps, implying that we have effectively learned a non-expansive mapping with convergence. Another interesting finding is that even when the convergence curves oscillate a little in the trained iterations, our method can still learn a stable non-expansive mapping and converge successfully in the untrained iterations. More investigations about the impact on the number of iterations for training are given in Appendix C.1.

Furthermore, we verify the influence of non-expansive property of \mathcal{D} on the convergence of ω in Figure 4. It can be seen that the non-expansive property reduces the gradient of φ by an order of magnitude, and provides a better convergence of the lower iteration. These verify the important influence of the non-expansive property on the convergence.

5.2. Image Deconvolution

For the practical application on image deconvolution, similar to (Zhang et al., 2020), we use a large dataset containing 400 images from Berkeley Segmentation Dataset, 4744 images from Waterloo Exploration Database, 900 images from DIV2K dataset, and 2750 images from Flick2K dataset. We test the performance of BMO on three classical testing im-

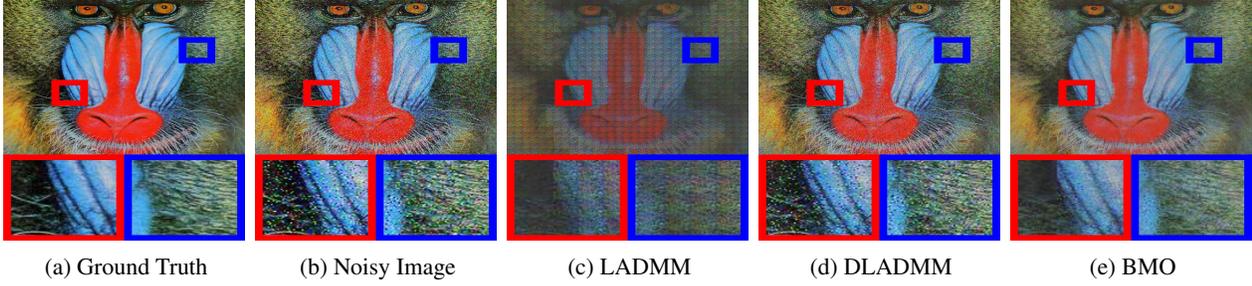


Figure 1. Denoising results of the baboon image. The larger red and blue boxes are the enlarged images of corresponding smaller boxes.

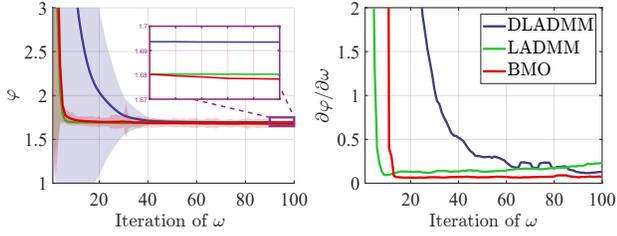


Figure 2. The convergence curves of φ and $\frac{\partial \varphi}{\partial \omega}$ by DLADMM, LADMM and our BMO. LADMM only learns the step size in ADMM, and DLADMM learns a matrix in ADMM. It can be seen that our method achieves the optimal convergence of loss function with a stationary gradient curve.

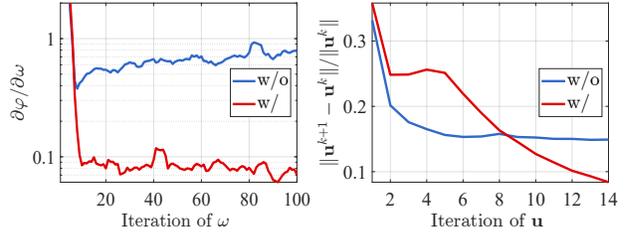
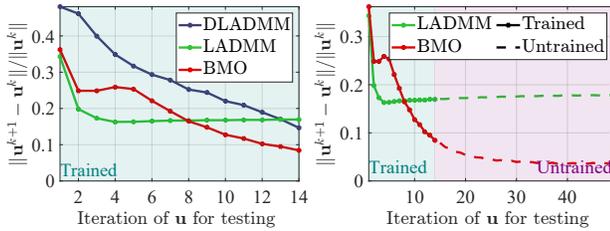
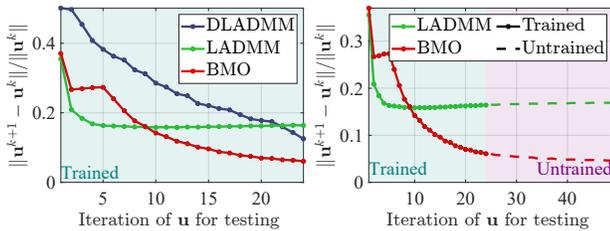


Figure 4. Convergence curves of $\frac{\partial \varphi}{\partial \omega}$ and $\|\mathbf{u}^{k+1} - \mathbf{u}^k\| / \|\mathbf{u}^k\|$ with or without non-expansive mapping.



(a) Iterations for training = 15



(b) Iterations for training = 25

Figure 3. Convergence curves of $\|\mathbf{u}^{k+1} - \mathbf{u}^k\| / \|\mathbf{u}^k\|$ with respect to k , the number of iterations of \mathbf{u} for testing, after (a) 15 and (b) 25 iterations for training. Solid lines on the right column represent the iterations for testing are less than those for training (trained iterations), while dotted lines represent the iterations for testing are more than those for training (untrained iterations).

ages in Table 2. and compare our method with representative methods including numerically designed method EPLL (Zoran & Weiss, 2011), learning-based method FDN (Kruse

Table 2. PSNR (dB) results compared with state-of-the-art methods on Set3c (Levin et al., 2009) for image deconvolution.

Noise level	$\sigma = 1\%$			$\sigma = 3\%$			
	Image	Butterfly	Leaves	Starfish	Butterfly	Leaves	Starfish
EPLL		20.55	19.22	24.84	18.64	17.54	22.47
FDN		27.40	26.51	27.48	24.27	23.53	24.71
IRCNN		32.74	33.22	33.53	28.53	28.45	28.42
IRCNN+		32.48	33.59	32.18	28.40	28.14	28.20
DPIR		34.18	35.12	33.91	29.45	30.27	29.46
BMO		33.67	35.39	33.98	29.46	30.69	29.64

et al., 2017), and plug-and-play methods including IRCNN, IRCNN+, and DPIR (Zhang et al., 2017; 2020). With embedded handcrafted network architectures \mathcal{D}_{net} and numerical schemes \mathcal{D}_{PG} to guarantee competitive performance, BMO performs best in the last five columns and achieves top two in the first column compared with state-of-the-art methods under different levels of noise on three testing images. Note that here we choose DRUNet in DPIR (Zhang et al., 2020) as \mathcal{D}_{net} , and the overall preferable results of BMO than directly using DPIR demonstrate the effect of using the composition of \mathcal{D}_{num} and \mathcal{D}_{net} .

5.3. Rain Streak Removal

As another example of applications, we carry out the rain streak removal experiment on synthesized rain datasets, including Rain100L and Rain100H. Rain100L contains 200 rainy/clean image pairs for training and another 100 pairs

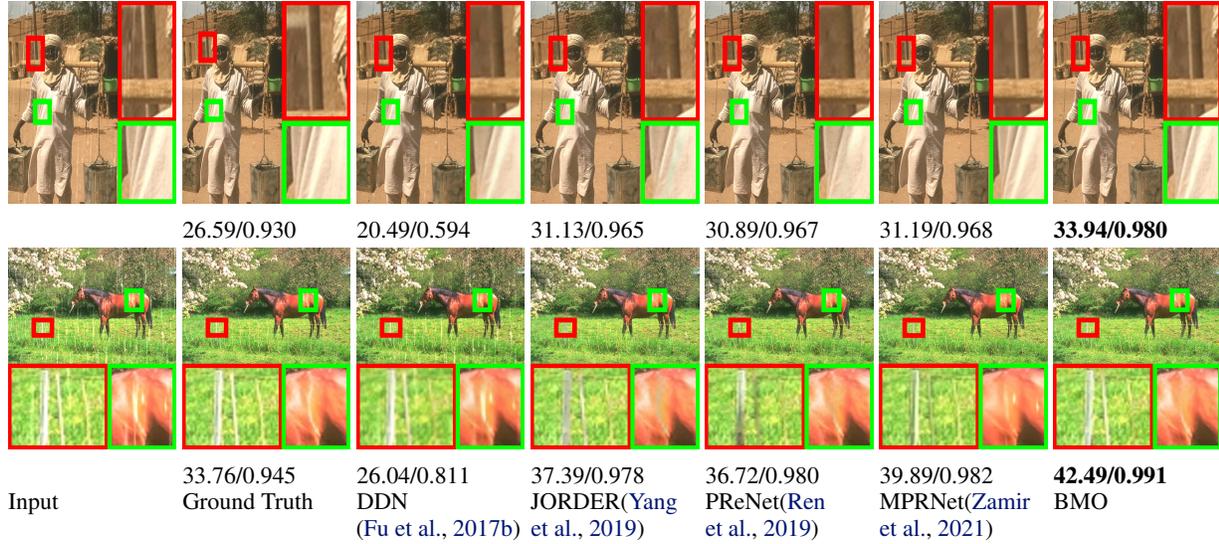


Figure 5. Performance of the rain streak removal task on two samples from Rain100L. The larger red and green boxes are the enlarged images of corresponding smaller boxes, and the numbers respectively correspond to PSNR and SSIM.

Table 3. Averaged PSNR and SSIM results among various methods for the single image rain removal task on two widely used synthesized datasets: Rain100L and Rain100H (Yang et al., 2019).

Datasets Metrics	Rain 100L		Rain 100H	
	PSNR	SSIM	PSNR	SSIM
Input	26.90	0.838	13.56	0.370
DSC	27.34	0.849	13.77	0.319
GMM	29.05	0.871	15.23	0.449
JCAS	28.54	0.852	14.62	0.451
Clear	30.24	0.934	15.33	0.742
DDN	32.38	0.925	22.85	0.725
RESCAN	38.52	0.981	29.62	0.872
PReNet	37.45	0.979	30.11	0.905
SPANet	35.33	0.969	25.11	0.833
JORDER_E	38.59	0.983	30.50	0.896
SIRR	32.37	0.925	22.47	0.716
MPRNet	36.40	0.965	30.41	0.890
RCDNet	40.00	0.986	31.28	0.903
BMO	40.07	0.986	30.96	0.905

for testing, and Rain100H has been updated to include 1800 images for training and 200 images for testing. We report the quantitative results of BMO in Table 3 with a series of state-of-the-art methods, containing DSC (Fu et al., 2017a), GMM (Li et al., 2016), JCAS (Gu et al., 2017), Clear (Fu et al., 2017a), DDN (Fu et al., 2017b), RESCAN (Li et al., 2018), PReNet (Ren et al., 2019), SPANet (Wang et al., 2019b), JORDER_E (Yang et al., 2019), SIRR (Wei et al., 2019), MPRNet (Zamir et al., 2021), and RCDNet (Wang et al., 2020). It can be seen that BMO achieves higher PSNR and SSIM on both benchmark datasets. Note that BMO has a competitive performance compared with RCDNet, but BMO also provides better theoretical property.

In Figure 5, we visually report the performance of rain

streak removal task on two images from Rain100L (Yang et al., 2019) compared with DDN, JORDER, PReNet and MPRNet. From the first row, one can observe that our BMO preserves the outline of the door in the background when removing the rain streak, compared with DDN, JORDER, and PReNet, which remain less details of the outline. In the second row, other learning-based methods tend to blur some image textures or leave distinct rain marks, while BMO remains the original color and outline, and gains better performance on PSNR and SSIM.

6. Conclusions

This paper first introduces the GKM scheme to unify various existing ODL approaches, and then proposes our BMO algorithmic framework to jointly solve the training and hyper-training task. We prove the essential convergence of training and hyper-training variables, from the perspective of both the approximation quality, and the stationary analysis. Experiments demonstrate our efficiency on sparse coding and real-world applications on image processing.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Nos. 61922019, 61733002, 62027826, and 11971220), the National Key R & D Program of China (No. 2020YFB1313503), the major key project of PCL (No. PCL2021A12), the Fundamental Research Funds for the Central Universities, Shenzhen Science and Technology Program (No. RCYX20200714114700072), and Guangdong Basic and Applied Basic Research Foundation (No. 2022B1515020082).

References

- Ablin, P., Moreau, T., Massias, M., and Gramfort, A. Learning step sizes for unfolded sparse coding. *arXiv preprint arXiv:1905.11071*, 2019.
- Ahmad, R., Bouman, C. A., Buzzard, G. T., Chan, S., Liu, S., Reehorst, E. T., and Schniter, P. Plug-and-play methods for magnetic resonance imaging: Using denoisers for image recovery. *IEEE SPM*, 37(1):105–116, 2020.
- Andrews, H. C. and Hunt, B. R. Digital image restoration. 1977.
- Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.
- Borwein, J., Reich, S., and Shafir, I. Krasnoselski-mann iterations in normed spaces. *Canadian Mathematical Bulletin*, 35(1):21–28, 1992.
- Boyd, S., Parikh, N., and Chu, E. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.
- Cabot, A. Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization. *SIAM Journal on Optimization*, 15(2):555–572, 2005.
- Chan, S. H., Wang, X., and Elgendy, O. A. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE TCI*, 3(1):84–98, 2016.
- Chen, T., Chen, X., Chen, W., Heaton, H., Liu, J., Wang, Z., and Yin, W. Learning to optimize: A primer and a benchmark. *arXiv preprint arXiv:2103.12828*, 2021.
- Chen, X., Liu, J., Wang, Z., and Yin, W. Theoretical linear convergence of unfolded ISTA and its practical weights and thresholds. In *NeurIPS*, 2018.
- Chen, Y. and Pock, T. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE TPAMI*, 39(6):1256–1272, 2016.
- Cui, F., Tang, Y., and Zhu, C. Convergence analysis of a variable metric forward–backward splitting algorithm with applications. *Journal of Inequalities and Applications*, 2019(1):1–27, 2019.
- Daubechies, I., Defrise, M., and De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.
- Edelstein, M. and O’Brien, R. C. Nonexpansive mappings, asymptotic regularity and successive approximations. *Journal of the London Mathematical Society*, 2(3): 547–554, 1978.
- Feurer, M. and Hutter, F. Hyperparameter optimization. In *Automated Machine Learning*, pp. 3–33. Springer, Cham, 2019.
- Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. Forward and reverse gradient-based hyperparameter optimization. In *ICML*, volume 70, pp. 1165–1173. JMLR, 2017.
- Fu, X., Huang, J., Ding, X., Liao, Y., and Paisley, J. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE TIP*, 26(6):2944–2956, 2017a.
- Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., and Paisley, J. Removing rain from single images via a deep detail network. In *CVPR*, 2017b.
- Fung, S. W., Heaton, H., Li, Q., McKenzie, D., Osher, S., and Yin, W. Jfb: Jacobian-free backpropagation for implicit networks. In *AAAI*, 2022.
- Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *ICML*, pp. 3748–3758. PMLR, 2020.
- Gregor, K. and LeCun, Y. Learning fast approximations of sparse coding. In *ICML*, 2010.
- Gu, S., Meng, D., Zuo, W., and Zhang, L. Joint convolutional analysis and synthesis sparse representation for single image layer separation. In *ICCV*, pp. 1708–1716, 2017.
- He, B., Ma, F., and Yuan, X. Optimal proximal augmented lagrangian method and its application to full Jacobian splitting for multi-block separable convex minimization problems. *IMA Journal of Numerical Analysis*, 40(2): 1188–1216, 2020.
- He, X., Zhao, K., and Chu, X. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.
- Heaton, H., Chen, X., Wang, Z., and Yin, W. Safeguarded learned convex optimization. *arXiv preprint arXiv:2003.01880*, 2020.
- Hutter, F., Kotthoff, L., and Vanschoren, J. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.
- Kruse, J., Rother, C., and Schmidt, U. Learning to push the limits of efficient fft-based image deconvolution. In *ICCV*, 2017.

- Levin, A., Weiss, Y., Durand, F., and Freeman, W. T. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, 2009.
- Li, S., Araujo, I. B., Ren, W., Wang, Z., Tokuda, E. K., Junior, R. H., Cesar-Junior, R., Zhang, J., Guo, X., and Cao, X. Single image deraining: A comprehensive benchmark analysis. In *CVPR*, pp. 3838–3847, 2019.
- Li, X., Wu, J., Lin, Z., Liu, H., and Zha, H. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, 2018.
- Li, Y., Tan, R. T., Guo, X., Lu, J., and Brown, M. S. Rain streak removal using layer priors. In *CVPR*, 2016.
- Li, Y., Tofighi, M., Geng, J., Monga, V., and Eldar, Y. C. Efficient and interpretable deep blind image deblurring via algorithm unrolling. *IEEE TCI*, 6:666–681, 2020.
- Lin, Z., Liu, R., and Su, Z. Linearized alternating direction method with adaptive penalty for low-rank representation. *NeurIPS*, 2011.
- Liu, R., Cheng, S., He, Y., Fan, X., and Luo, Z. Toward designing convergent deep operator splitting methods for task-specific nonconvex optimization. *arXiv preprint arXiv:1804.10798*, 2018a.
- Liu, R., He, Y., Cheng, S., Fan, X., and Luo, Z. Learning collaborative generation correction modules for blind image deblurring and beyond. In *ACM*, 2018b.
- Liu, R., Cheng, S., Ma, L., Fan, X., and Luo, Z. Deep proximal unrolling: Algorithmic framework, convergence analysis and applications. *IEEE TIP*, 28(10):5013–5026, 2019a.
- Liu, R., Mu, P., and Zhang, J. On the convergence of admm with task adaption and beyond. *CoRR*, 2019b.
- Liu, R., Mu, P., Chen, J., Fan, X., and Luo, Z. Investigating task-driven latent feasibility for nonconvex image modeling. *IEEE TIP*, 29:7629–7640, 2020a.
- Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *ICML*, 2020b.
- Liu, R., Gao, J., Zhang, J., Meng, D., and Lin, Z. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE TPAMI*, 2021.
- Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A general descent aggregation framework for gradient-based bi-level optimization. *IEEE TPAMI*, 2022.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.
- Moeller, M., Mollenhoff, T., and Cremers, D. Controlling neural networks via energy dissipation. In *ICCV*, pp. 3256–3265, 2019.
- Monga, V., Li, Y., and Eldar, Y. C. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE SPM*, 38(2):18–44, 2021.
- Neumüller, C., Wagner, S., Kronberger, G., and Affenzeller, M. Parameter meta-optimization of metaheuristic optimization algorithms. In *International Conference on Computer Aided Systems Theory*, pp. 367–374. Springer, 2011.
- Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., and Yosinski, J. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, 2017.
- Reich, S. and Zaslavski, A. Convergence of krasnoselskii-mann iterations of nonexpansive operators. *Mathematical and Computer Modelling*, 32(11-13):1423–1431, 2000.
- Ren, D., Zuo, W., Hu, Q., Zhu, P., and Meng, D. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, 2019.
- Richardson, W. H. Bayesian-based iterative method of image restoration. *JoSA*, 62(1):55–59, 1972.
- Rockafellar, R. T. Convex analysis princeton university press. *Princeton, NJ*, 1970.
- Ryu, E., Liu, J., Wang, S., Chen, X., Wang, Z., and Yin, W. Plug-and-play methods provably converge with properly trained denoisers. In *ICML*, 2019.
- Schuler, C. J., Hirsch, M., Harmeling, S., and Schölkopf, B. Learning to deblur. *IEEE TPAMI*, 38(7):1439–1451, 2015.
- Shlezinger, N., Whang, J., Eldar, Y. C., and Dimakis, A. G. Model-based deep learning. *arXiv preprint arXiv:2012.08405*, 2020.
- Song, G., Sun, Y., Liu, J., Wang, Z., and Kamilov, U. S. A new recurrent plug-and-play prior based on the multiple self-similarity network. *IEEE SPL*, 27:451–455, 2020.
- Sun, Y., Wohlberg, B., and Kamilov, U. S. An online plug-and-play algorithm for regularized image reconstruction. *IEEE TCI*, 5(3):395–408, 2019.

- Teodoro, A. M., Bioucas-Dias, J. M., and Figueiredo, M. A. A convergent image fusion algorithm using scene-adapted gaussian-mixture-based denoising. *IEEE TIP*, 28(1):451–463, 2018.
- Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *SIGKDD*, pp. 847–855, 2013.
- Venkatakrishnan, S. V., Bouman, C. A., and Wohlberg, B. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948. IEEE, 2013.
- Von Stackelberg, H. *Market structure and equilibrium*. Springer Science & Business Media, 2010.
- Wang, H., Wu, Y., Li, M., Zhao, Q., and Meng, D. A survey on rain removal from video and single image. *arXiv preprint arXiv:1909.08326*, 2019a.
- Wang, H., Xie, Q., Zhao, Q., and Meng, D. A model-driven deep neural network for single image rain removal. In *CVPR*, pp. 3103–3112, 2020.
- Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., and Lau, R. W. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, 2019b.
- Wei, W., Meng, D., Zhao, Q., Xu, Z., and Wu, Y. Semi-supervised transfer learning for image rain removal. In *CVPR*, pp. 3877–3886, 2019.
- Xie, X., Wu, J., Liu, G., Zhong, Z., and Lin, Z. Differentiable linearized admm. In *ICML*, 2019.
- Xu, Y. and Yin, W. A fast patch-dictionary method for whole image recovery. *arXiv preprint arXiv:1408.3740*, 2014.
- Yang, W., Tan, R. T., Feng, J., Guo, Z., Yan, S., and Liu, J. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE TPAMI*, 42(6): 1377–1393, 2019.
- Yang, Y., Sun, J., Li, H., and Xu, Z. Admm-net: A deep learning approach for compressive sensing mri. *arXiv preprint arXiv:1705.06869*, 2017.
- Yang, Y., Huang, Z., and Wipf, D. Transformers from an optimization perspective. *arXiv preprint arXiv:2205.13891*, 2022.
- Yuan, X., Liu, Y., Suo, J., and Dai, Q. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *CVPR*, 2020.
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. Multi-stage progressive image restoration. *arXiv preprint arXiv:2102.02808*, 2021.
- Zhang, J. and Ghanem, B. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *CVPR*, 2018.
- Zhang, K., Zuo, W., Gu, S., and Zhang, L. Learning deep cnn denoiser prior for image restoration. In *CVPR*, 2017.
- Zhang, K., Zuo, W., and Zhang, L. Deep plug-and-play super-resolution for arbitrary blur kernels. In *CVPR*, 2019.
- Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., and Timofte, R. Plug-and-play image restoration with deep denoiser prior. *arXiv preprint arXiv:2008.13751*, 2020.
- Zhang, Z., Xu, Y., Yang, J., Li, X., and Zhang, D. A survey of sparse representation: algorithms and applications. *IEEE Access*, 3:490–530, 2015.
- Zoran, D. and Weiss, Y. From learning models of natural image patches to whole image restoration. In *ICCV*, 2011.

Appendices

A. Detailed Proofs

In this section, we discuss the convergence analysis of our proposed BMO algorithm (Algorithm 1) for the GKM scheme, towards the optimal solution and stationary points of optimization problem in Eq. (4) with respect to both \mathbf{u} and ω . Note that this joint convergence of training and hyper-training also provides a unified theoretical guarantee for existing ODL methods containing UNH and ENA.

By introducing an auxiliary function, the problem in Eq. (4) can be equivalently rewritten as the following

$$\min_{\omega \in \Omega} \varphi(\omega), \quad \text{where} \quad \varphi(\omega) := \inf_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)) \cap U} \ell(\mathbf{u}, \omega). \quad (15)$$

The sequence $\{\omega^t\}$ generated by BMO (Algorithm 1) actually solves the following approximation problem of Eq. (4)

$$\min_{\omega \in \Omega} \varphi_K(\omega) := \ell(\mathbf{u}^K(\omega), \omega), \quad (16)$$

where $\mathbf{u}^K(\omega)$ is derived by solving the simple bilevel problem $\inf_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)) \cap U} \ell(\mathbf{u}, \omega)$ and can be given by

$$\begin{cases} \mathbf{v}_l^k(\omega) = \mathcal{T}(\mathbf{u}^{k-1}(\omega), \omega), \\ \mathbf{v}_u^k(\omega) = \mathbf{u}^{k-1}(\omega) - s_k \mathbf{H}_\omega^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^{k-1}, \omega), \\ \mathbf{u}^k(\omega) = \text{Proj}_{U, \mathbf{H}_\omega}(\mu \mathbf{v}_u^k(\omega) + (1 - \mu) \mathbf{v}_l^k(\omega)), \end{cases} \quad (17)$$

where $k = 1, \dots, K$.

A.1. Approximation Quality and Convergence

In this part, we will show that Eq. (16) is actually an appropriate approximation to Eq. (4) in the sense that any limit point $(\bar{\mathbf{u}}, \bar{\omega})$ of the sequence $\{(\mathbf{u}^K(\omega^K), \omega^K)\}$ with $\omega^K \in \text{argmin}_{\omega \in \Omega} \varphi_K(\omega)$ is a solution to the bilevel problem in Eq. (4). Thus we can obtain the optimal solution of Eq. (4) by solving Eq. (16). We make the following standing assumption throughout this part.

Assumption A.1 Ω is a compact set and U is a convex compact set. $\text{Fix}(\mathcal{T}(\cdot, \omega))$ is nonempty for any $\omega \in \Omega$. $\ell(\mathbf{u}, \omega)$ is continuous on $\mathbb{R}^n \times \Omega$. For any $\omega \in \Omega$, $\ell(\cdot, \omega) : \mathbb{R}^n \rightarrow \mathbb{R}$ is L_ℓ -smooth, convex and bounded below by M_0 .

Please notice that ℓ is usually defined to be the MSE loss, and thus Assumption A.1 is quite standard for ODL (Ryu et al., 2019; Zhang et al., 2020). We first present some necessary preliminaries. For any two matrices $\mathbf{H}_1, \mathbf{H}_2 \in \mathbb{R}^{n \times n}$, we consider the following partial ordering relation:

$$\mathbf{H}_1 \succeq \mathbf{H}_2 \quad \Leftrightarrow \quad \langle \mathbf{u}, \mathbf{H}_1 \mathbf{u} \rangle \geq \langle \mathbf{u}, \mathbf{H}_2 \mathbf{u} \rangle, \quad \forall \mathbf{u} \in \mathbb{R}^n.$$

If $\mathbf{H} \succ 0$, $\langle \mathbf{u}_1, \mathbf{H} \mathbf{u}_2 \rangle$ for $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n$ defines an inner product on \mathbb{R}^n . Denote the induced norm with $\|\cdot\|_{\mathbf{H}}$, i.e., $\|\mathbf{u}\|_{\mathbf{H}} := \sqrt{\langle \mathbf{u}, \mathbf{H} \mathbf{u} \rangle}$ for any $\mathbf{u} \in \mathbb{R}^n$. We assume $\mathcal{D}(\cdot, \omega)$ satisfies the following assumption throughout this part.

Assumption A.2 There exist $\mathbf{H}_{ub} \succeq \mathbf{H}_{lb} \succ 0$, such that for each $\omega \in \Omega$, there exists $\mathbf{H}_{ub} \succeq \mathbf{H}_\omega \succeq \mathbf{H}_{lb}$ such that

(1) $\mathcal{D}(\cdot, \omega)$ is non-expansive with respect to $\|\cdot\|_{\mathbf{H}_\omega}$, i.e., for all $(\mathbf{u}_1, \mathbf{u}_2) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$\|\mathcal{D}(\mathbf{u}_1, \omega) - \mathcal{D}(\mathbf{u}_2, \omega)\|_{\mathbf{H}_\omega} \leq \|\mathbf{u}_1 - \mathbf{u}_2\|_{\mathbf{H}_\omega}.$$

(2) $\mathcal{D}(\cdot, \omega)$ is closed, i.e.,

$$\text{gph } \mathcal{D}(\cdot, \omega) := \{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \mathbf{v} = \mathcal{D}(\mathbf{u}, \omega)\}$$

is closed.

Under Assumption A.2, we obtain the following non-expansive properties of $\mathcal{T}(\cdot, \omega)$ defined in Eq. (3) from (Bauschke et al., 2011)[Proposition 4.25] immediately.

Lemma A.1 *Given $\alpha \in (0, 1)$, $\omega \in \Omega$, let $\mathcal{T}(\cdot, \omega) := (1 - \alpha)\mathcal{I} + \alpha\mathcal{D}(\cdot, \omega)$, where \mathcal{I} denotes the identity operator, then $\mathcal{T}(\cdot, \omega)$ is closed and satisfies that for any $(\mathbf{u}_1, \mathbf{u}_2) \in \mathbb{R}^n \times \mathbb{R}^n$,*

$$\begin{aligned} & \|\mathbf{u}_1 - \mathbf{u}_2\|_{\mathbf{H}\omega}^2 - \|\mathcal{T}(\mathbf{u}_1, \omega) - \mathcal{T}(\mathbf{u}_2, \omega)\|_{\mathbf{H}\omega}^2 \\ & \geq \frac{1 - \alpha}{\alpha} \|(\mathbf{u}_1 - \mathcal{T}(\mathbf{u}_1, \omega)) - (\mathbf{u}_2 - \mathcal{T}(\mathbf{u}_2, \omega))\|_{\mathbf{H}\omega}^2 \geq 0. \end{aligned} \quad (18)$$

In this part, as the identity of ω is clear from the context, for succinctness we will write $\ell(\mathbf{u})$ instead of $\ell(\mathbf{u}, \omega)$, $\mathcal{T}(\mathbf{u})$ instead of $\mathcal{T}(\mathbf{u}, \omega)$, \mathbf{H} instead of $\mathbf{H}\omega$, ℓ^* instead of $\varphi(\omega)$, \mathcal{S} instead of $\text{Fix}(\mathcal{T}(\cdot, \omega))$, and $\hat{\mathcal{S}}$ instead of $\hat{\mathcal{S}}(\omega) := \text{argmin}_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)) \cap \mathcal{U}} \ell(\mathbf{u}, \omega)$. Moreover, we will omit the notation ω and use the notations \mathbf{u}^k , \mathbf{v}_u^k and \mathbf{v}_l^k instead of $\mathbf{u}^k(\omega)$, $\mathbf{v}_u^k(\omega)$ and $\mathbf{v}_l^k(\omega)$, respectively.

Before giving the proof of Theorem A.1, we present some helpful lemmas and propositions. The proof is inspired by (Liu et al., 2022).

Lemma A.2 *For any given $\omega \in \Omega$, let $\{\mathbf{u}^k\}$ be the sequence generated by Eq. (17) with $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{lb})}{L_\ell})$ and $\mu \in (0, 1)$, where $\lambda_{\min}(\mathbf{H}_{lb})$ denotes the smallest eigenvalue of matrix \mathbf{H}_{lb} in Assumption A.2. Then for any $\mathbf{u} \in \mathcal{S}$, we have*

$$\begin{aligned} \mu s \beta_k \ell(\mathbf{u}) & \geq \mu s \beta_k \ell(\mathbf{v}_u^{k+1}) - \frac{1}{2} \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{H}}^2 + \frac{1}{2} \|\mathbf{u} - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2 \\ & + \frac{1}{2} \|((1 - \mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1}) - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2 \\ & + \frac{(1 - \mu)\eta}{2} \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 + \frac{\mu}{2} (1 - s_k \frac{L_\ell}{\lambda_{\min}(\mathbf{H})}) \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2, \end{aligned} \quad (19)$$

where $\beta_k = \frac{1}{k+1}$ and $\eta = \frac{1-\alpha}{\alpha} > 0$.

Proof: We can obtain from the definition of \mathbf{v}_u^k that

$$0 = \beta_k \nabla \ell(\mathbf{u}^k) + \frac{1}{s} \mathbf{H}(\mathbf{v}_u^{k+1} - \mathbf{u}^k), \quad (20)$$

where $\beta_k = \frac{1}{k+1}$, and thus for any \mathbf{u} ,

$$0 = \beta_k \langle \nabla \ell(\mathbf{u}^k), \mathbf{u} - \mathbf{v}_u^{k+1} \rangle + \frac{1}{s} \langle \mathbf{v}_u^{k+1} - \mathbf{u}^k, \mathbf{H}(\mathbf{u} - \mathbf{v}_u^{k+1}) \rangle. \quad (21)$$

Since ℓ is convex and $\nabla \ell$ is L_ℓ -Lipschitz continuous, we have

$$\begin{aligned} & \langle \nabla \ell(\mathbf{u}^k), \mathbf{u} - \mathbf{v}_u^{k+1} \rangle \\ & = \langle \nabla \ell(\mathbf{u}^k), \mathbf{u} - \mathbf{u}^k \rangle + \langle \nabla \ell(\mathbf{u}^k), \mathbf{u}^k - \mathbf{v}_u^{k+1} \rangle \\ & \leq \ell(\mathbf{u}) - \ell(\mathbf{u}^k) + \ell(\mathbf{u}^k) - \ell(\mathbf{v}_u^{k+1}) + \frac{L_\ell}{2} \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|^2 \\ & \leq \ell(\mathbf{u}) - \ell(\mathbf{v}_u^{k+1}) + \frac{L_\ell}{2\lambda_{\min}(\mathbf{H})} \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2. \end{aligned} \quad (22)$$

Combining with $\langle \mathbf{v}_u^{k+1} - \mathbf{u}^k, \mathbf{H}(\mathbf{u} - \mathbf{v}_u^{k+1}) \rangle = \frac{1}{2} (\|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{H}}^2 - \|\mathbf{u} - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 - \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2)$ and Eq. (21) yields that for any \mathbf{u} ,

$$\begin{aligned} \beta_k \ell(\mathbf{u}) & \geq \beta_k \ell(\mathbf{v}_u^{k+1}) - \frac{1}{2s} \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{H}}^2 + \frac{1}{2s} \|\mathbf{u} - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\ & + \frac{1}{2s} (1 - s\beta_k \frac{L_\ell}{\lambda_{\min}(\mathbf{H})}) \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2. \end{aligned} \quad (23)$$

Next, since $\mathbf{v}_l^{k+1} = \mathcal{T}(\mathbf{u}^k)$ and \mathcal{T} satisfies the inequality in Lemma A.1, we have for any $\mathbf{u} \in \mathcal{S}$,

$$\|\mathbf{u}^k - \mathbf{u}\|_{\mathbf{H}}^2 - \|\mathbf{v}_l^{k+1} - \mathbf{u}\|_{\mathbf{H}}^2 \geq \eta \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2, \quad (24)$$

with $\eta = \frac{1-\alpha}{\alpha} > 0$. Multiplying Eq. (23) and Eq. (24) by μs and $\frac{1-\mu}{2}$, respectively, and then summing them up yields that for any $\mathbf{u} \in \mathcal{S}$,

$$\begin{aligned} \mu s \beta_k \ell(\mathbf{u}) &\geq \mu s \beta_k \ell(\mathbf{v}_u^{k+1}) - \frac{1}{2} \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{H}}^2 + \frac{1}{2} \left((1-\mu) \|\mathbf{u} - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 + \mu \|\mathbf{u} - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \right) \\ &\quad + \frac{(1-\mu)\eta}{2} \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 + \frac{\mu}{2} \left(1 - s \beta_k \frac{L_\ell}{\lambda_{\min}(\mathbf{H})} \right) \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2. \end{aligned} \quad (25)$$

The convexity of $\|\cdot\|_{\mathbf{H}}^2$ implies that

$$\begin{aligned} (1-\mu) \|\mathbf{u} - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 + \mu \|\mathbf{u} - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\ \geq \|\mathbf{u} - ((1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1})\|_{\mathbf{H}}^2. \end{aligned}$$

Next, as $\text{Proj}_{U, \mathbf{H}}$ is firmly non-expansive with respect to $\|\cdot\|_{\mathbf{H}}$ (see, e.g., (Bauschke et al., 2011)[Proposition 4.8]), for any $\mathbf{u} \in U$, we have

$$\begin{aligned} \|\mathbf{u} - ((1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1})\|_{\mathbf{H}}^2 \\ \geq \|\mathbf{u} - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2 + \|((1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1}) - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2. \end{aligned} \quad (26)$$

Then, we obtain from Eq. (25) that for any $\mathbf{u} \in \mathcal{S}$,

$$\begin{aligned} \mu s \beta_k \ell(\mathbf{u}) &\geq \mu s \beta_k \ell(\mathbf{v}_u^{k+1}) - \frac{1}{2} \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{H}}^2 + \frac{1}{2} \|\mathbf{u} - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2 \\ &\quad + \frac{1}{2} \|((1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1}) - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2 \\ &\quad + \frac{(1-\mu)\eta}{2} \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ &\quad + \frac{\mu}{2} \left(1 - s \beta_k \frac{L_\ell}{\lambda_{\min}(\mathbf{H})} \right) \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2. \end{aligned} \quad (27)$$

This completes the proof. \square

Lemma A.3 For any given $\omega \in \Omega$, let $\{\mathbf{u}^k\}$ be the sequence generated by Eq. (17) with $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}l_b)}{L_\ell})$ and $\mu \in (0, 1)$. Then for any $\bar{\mathbf{u}} \in \mathcal{S}$, we have

$$\|\mathbf{v}_l^{k+1} - \bar{\mathbf{u}}\|_{\mathbf{H}} \leq \|\mathbf{u}^k - \bar{\mathbf{u}}\|_{\mathbf{H}}, \quad (28)$$

and operator $\mathcal{I} - s_k \mathbf{H}^{-1} \nabla \ell$ is non-expansive (i.e., 1-Lipschitz continuous) with respect to $\|\cdot\|_{\mathbf{H}}$. Furthermore, when U is compact, sequences $\{\mathbf{u}^k\}$, $\{\mathbf{v}_l^k\}$, $\{\mathbf{v}_u^k\}$ are all bounded.

Proof: Since $\mathbf{v}_l^{k+1} = \mathcal{T}(\mathbf{u}^k)$ and \mathcal{T} satisfies the inequality in Lemma A.1, we have

$$\|\mathbf{v}_l^{k+1} - \bar{\mathbf{u}}\|_{\mathbf{H}} \leq \|\mathbf{u}^k - \bar{\mathbf{u}}\|_{\mathbf{H}}.$$

When U is compact, the desired boundedness of $\{\mathbf{u}^k\}$ follows directly from the iteration scheme given in Eq. (17). Since for any u_1, u_2 ,

$$\begin{aligned} \langle \mathbf{H}^{-1} \nabla \ell(u_1) - \mathbf{H}^{-1} \nabla \ell(u_2), u_1 - u_2 \rangle_{\mathbf{H}} &= \langle \nabla \ell(u_1) - \nabla \ell(u_2), u_1 - u_2 \rangle \\ &\geq \frac{1}{L_\ell} \|\nabla \ell(u_1) - \nabla \ell(u_2)\|^2 \geq \frac{\lambda_{\min}(\mathbf{H})}{L_\ell} \|\mathbf{H}^{-1} \nabla \ell(u_1) - \mathbf{H}^{-1} \nabla \ell(u_2)\|_{\mathbf{H}}^2, \end{aligned}$$

where the first inequality follows from (Bauschke et al., 2011)[Corollary 18.16]. This implies that $\mathbf{H}^{-1} \nabla \ell$ is $\frac{\lambda_{\min}(\mathbf{H})}{L_\ell}$ -cocoercive (see (Bauschke et al., 2011)[Definition 4.4]) with respect to $\langle \cdot, \cdot \rangle_{\mathbf{H}}$ and $\|\cdot\|_{\mathbf{H}}$. Then, according to (Bauschke

et al., 2011)[Proposition 4.33], we know that when $0 < s_k \leq \frac{\lambda_{\min}(\mathbf{H})}{L_\ell}$, operator $\mathcal{I} - s_k \mathbf{H}^{-1} \nabla \ell$ is non-expansive with respect to $\|\cdot\|_{\mathbf{H}}$. Then, since $\mathbf{v}_u^{k+1} = \mathbf{u}^k - \alpha_k s \mathbf{H}^{-1} \nabla \ell(\mathbf{u}^k)$, we have

$$\|\mathbf{v}_u^{k+1} - (\bar{\mathbf{u}} - s_k \mathbf{H}^{-1} \nabla \ell(\bar{\mathbf{u}}))\|_{\mathbf{H}} \leq \|\mathbf{u}^k - \bar{\mathbf{u}}\|_{\mathbf{H}},$$

and $s_k \in (0, \frac{\lambda_{\min}(\mathbf{H})}{L_\ell})$, we have that $\{\mathbf{v}_u^{k+1}\}$ is bounded. \square

Lemma A.4 (Liu et al., 2022)[Lemma 2] Let $\{a_k\}$ and $\{b_k\}$ be sequences of non-negative real numbers. Assume that there exists $n_0 \in \mathbb{N}$ such that

$$a_{k+1} + b_k - a_k \leq 0, \quad \forall k \geq n_0.$$

Then $\lim_{k \rightarrow \infty} a_k$ exists and $\sum_{k=1}^{\infty} b_k < \infty$.

Proposition A.1 For any given $\omega \in \Omega$, let $\{\mathbf{u}^k\}$ be the sequence generated by Eq. (17) with $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}l_b)}{L_\ell})$ and $\mu \in (0, 1)$. Suppose that U is compact and $\hat{\mathcal{S}}(\omega)$ is nonempty, we have

$$\lim_{k \rightarrow \infty} \text{dist}(\mathbf{u}^k, \hat{\mathcal{S}}(\omega)) = 0,$$

and then

$$\lim_{k \rightarrow \infty} \ell(\mathbf{u}^k, \omega) = \varphi(\omega).$$

Proof: Let $\delta > 0$ be a constant satisfying $\delta < \frac{1}{2} \min\{(1-\mu)\eta, \mu(1-sL_\ell/\lambda_{\min}(\mathbf{H}))\}$. We define a sequence of $\{\tau_n\}$ by

$$\begin{aligned} \tau_n := \max \left\{ k \in \mathbb{N} \mid k \leq n \text{ and } \delta \|\mathbf{u}^{k-1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2 + \delta \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 \right. \\ \left. + \frac{1}{4} \left\| ((1-\mu)\mathbf{v}_l^k + \mu\mathbf{v}_u^k) - \mathbf{u}^k \right\|_{\mathbf{H}}^2 + \mu s \beta_{k-1} (\ell(\mathbf{v}_u^k) - \ell^*) < 0 \right\}, \end{aligned}$$

where $\beta_k = \frac{1}{k+1}$ and $\eta = \frac{1-\alpha}{\alpha} > 0$. Inspired by (Cabot, 2005), we consider the following two cases:

(a) $\{\tau_n\}$ is finite, i.e., there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$\begin{aligned} \delta \|\mathbf{u}^{k-1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2 + \frac{1}{4} \left\| ((1-\mu)\mathbf{v}_l^k + \mu\mathbf{v}_u^k) - \mathbf{u}^k \right\|_{\mathbf{H}}^2 \\ + \delta \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 + \mu s \beta_{k-1} (\ell(\mathbf{v}_u^k) - \ell^*) \geq 0, \end{aligned}$$

(b) $\{\tau_n\}$ is not finite, i.e., for all $k_0 \in \mathbb{N}$, there exists $k \geq k_0$ such that

$$\begin{aligned} \delta \|\mathbf{u}^{k-1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2 + \frac{1}{4} \left\| ((1-\mu)\mathbf{v}_l^k + \mu\mathbf{v}_u^k) - \mathbf{u}^k \right\|_{\mathbf{H}}^2 \\ + \delta \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 + \mu s \beta_{k-1} (\ell(\mathbf{v}_u^k) - \ell^*) < 0. \end{aligned}$$

Case (a): We assume that $\{\tau_n\}$ is finite and there exists $k_0 \in \mathbb{N}$ such that

$$\begin{aligned} \delta \|\mathbf{u}^{k-1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2 + \frac{1}{4} \left\| ((1-\mu)\mathbf{v}_l^k + \mu\mathbf{v}_u^k) - \mathbf{u}^k \right\|_{\mathbf{H}}^2 + \\ \delta \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 + \mu s \beta_{k-1} (\ell(\mathbf{v}_u^k) - \ell^*) \geq 0, \end{aligned} \quad (29)$$

for all $k \geq k_0$. Let $\bar{\mathbf{u}}$ be any point in $\hat{\mathcal{S}}$, setting \mathbf{u} in Eq. (19) of Lemma A.2 to be $\bar{\mathbf{u}}$, as $\mu \in (0, 1)$ and $\beta_k \leq 1$, we have

$$\begin{aligned} & \frac{1}{2} \|\bar{\mathbf{u}} - \mathbf{u}^k\|_{\mathbf{H}}^2 \\ & \geq \frac{1}{2} \|\bar{\mathbf{u}} - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2 + \left(\frac{(1-\mu)\eta}{2} - \delta \right) \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ & \quad + \left(\frac{\mu(1-sL_\ell/\lambda_{\min}(\mathbf{H}))}{2} - \delta \right) \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\ & \quad + \frac{1}{4} \left\| ((1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1}) - \mathbf{u}^{k+1} \right\|_{\mathbf{H}}^2 + \delta \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ & \quad + \frac{1}{4} \left\| ((1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1}) - \mathbf{u}^{k+1} \right\|_{\mathbf{H}}^2 + \delta \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\ & \quad + \mu s \beta_k (\ell(\mathbf{v}_u^{k+1}) - \ell^*). \end{aligned} \quad (30)$$

For all $k \geq k_0$, combining $0 < \delta < \frac{1}{2} \min\{(1 - \mu)\eta, \mu(1 - sL\ell/\lambda_{\min}(\mathbf{H}))\}$ and Eq. (30), it follows from Lemma A.4 that

$$\begin{aligned} \sum_{k=0}^{\infty} \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 &< \infty, \\ \sum_{k=0}^{\infty} \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 &< \infty, \\ \sum_{k=0}^{\infty} \left\| \left((1 - \mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1} \right) - \mathbf{u}^{k+1} \right\|_{\mathbf{H}}^2 &< \infty, \\ \sum_{k=0}^{\infty} \beta_k (\ell(\mathbf{v}_u^{k+1}) - \ell^*) &< \infty, \end{aligned}$$

and $\lim_{k \rightarrow \infty} \|\bar{\mathbf{u}} - \mathbf{u}^k\|_{\mathbf{H}}^2$ exists.

Now, we show the existence of subsequence $\{\mathbf{u}^j\} \subseteq \{\mathbf{u}^k\}$ such that $\lim_{j \rightarrow \infty} \ell(\mathbf{u}^j) \leq \ell^*$. This obviously holds if for any $\hat{k} > 0$, there exists $k > \hat{k}$ such that $\ell(\mathbf{u}^k) \leq \ell^*$. Therefore, we only need to consider the case where there exists $\hat{k} > 0$ such that $\ell(\mathbf{u}^k) > \ell^*$ for all $k \geq \hat{k}$. If there does not exist subsequence $\{\mathbf{u}^j\} \subseteq \{\mathbf{u}^k\}$ such that $\lim_{j \rightarrow \infty} \ell(\mathbf{u}^j) \leq \ell^*$, there must exist $\epsilon > 0$ and $k_1 \geq \max\{\hat{k}, k_0\}$ such that $\ell(\mathbf{u}^k) - \ell^* \geq 2\epsilon$ for all $k \geq k_1$. As U is compact, it follows from Lemma A.3 that sequences $\{\mathbf{u}^k\}$ and $\{\mathbf{v}_u^{k+1}\}$ are both bounded. Then since ℓ is continuous and $\lim_{k \rightarrow \infty} \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}} = 0$ with $\mathbf{H} \succ 0$, there exists $k_2 \geq k_1$ such that $|\ell(\mathbf{u}^k) - \ell(\mathbf{v}_u^{k+1})| < \epsilon$ for all $k \geq k_2$ and thus $\ell(\mathbf{v}_u^{k+1}) - \ell^* \geq \epsilon$ for all $k \geq k_2$. Then we have

$$\epsilon \sum_{k=k_2}^{\infty} \beta_k \leq \sum_{k=k_2}^{\infty} \beta_k (\ell(\mathbf{v}_u^{k+1}) - \ell^*) < \infty,$$

where the last inequality follows from $\sum_{k=0}^{\infty} \beta_k (\ell(\mathbf{v}_u^{k+1}) - \ell^*) < \infty$. This result contradicts to the definition of β_k and the fact that $\sum_{k=0}^{\infty} \beta_k = \sum_{k=0}^{\infty} \frac{1}{k+1} = +\infty$. As $\{\mathbf{u}^j\}$ and $\{\mathbf{v}_l^{j+1}\}$ are bounded, and $\lim_{j \rightarrow \infty} \|\mathbf{u}^j - \mathbf{v}_l^{j+1}\|_{\mathbf{H}} = 0$ with $\mathbf{H} \succ 0$, we can assume without loss of generality that $\lim_{j \rightarrow \infty} \mathbf{v}_l^{j+1} = \mathbf{u}^j = \tilde{\mathbf{u}}$ by taking a subsequence. By the continuity of ℓ , we have $\ell(\tilde{\mathbf{u}}) = \lim_{j \rightarrow \infty} \ell(\mathbf{u}^j) \leq \ell^*$. Next, since $\mathbf{v}_l^{j+1} = \mathcal{T}(\mathbf{u}^j)$, let $\ell \rightarrow \infty$, by the closedness of \mathcal{T} , we have

$$\tilde{\mathbf{u}} = \mathcal{T}(\tilde{\mathbf{u}}),$$

and thus $\tilde{\mathbf{u}} \in \mathcal{S}$. Combining with $\ell(\tilde{\mathbf{u}}) \leq \ell^*$, we show that $\tilde{\mathbf{u}} \in \hat{\mathcal{S}}$. Then by taking $\bar{\mathbf{u}} = \tilde{\mathbf{u}}$ and since $\lim_{k \rightarrow \infty} \|\bar{\mathbf{u}} - \mathbf{u}^k\|_{\mathbf{H}}^2$ exists, we have $\lim_{k \rightarrow \infty} \|\bar{\mathbf{u}} - \mathbf{u}^k\|_{\mathbf{H}}^2 = 0$ with $\mathbf{H} \succ 0$ and thus $\lim_{k \rightarrow \infty} \text{dist}(\mathbf{u}^k, \hat{\mathcal{S}}) = 0$.

Case (b): We assume that $\{\tau_n\}$ is not finite and for any $k_0 \in \mathbb{N}$, there exists $k \geq k_0$ such that $\delta \|\mathbf{u}^{k-1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2 + \frac{1}{4} \left\| \left((1 - \mu)\mathbf{v}_l^k + \mu\mathbf{v}_u^k \right) - \mathbf{u}^k \right\|_{\mathbf{H}}^2 + \delta \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 + \mu s \beta_{k-1} (\ell(\mathbf{v}_u^k) - \ell^*) < 0$. It follows from the assumption that τ_n is well defined for n large enough and $\lim_{n \rightarrow \infty} \tau_n = +\infty$. We assume without loss of generality that τ_n is well defined for all n .

By setting \mathbf{u} in Eq. (19) of Lemma A.2 to be $\text{Proj}_{\hat{\mathcal{S}}}(\mathbf{u}^k)$, we have

$$\begin{aligned} &\frac{1}{2} \text{dist}_{\mathbf{H}}^2(\mathbf{u}^k, \hat{\mathcal{S}}) \\ &\geq \frac{1}{2} \text{dist}_{\mathbf{H}}^2(\mathbf{u}^{k+1}, \hat{\mathcal{S}}) \\ &\quad + \left(\frac{(1 - \mu)\eta}{2} - \delta \right) \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ &\quad + \left(\frac{\mu(1 - sL\ell/\lambda_{\min}(\mathbf{H}))}{2} - \delta \right) \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\ &\quad + \frac{1}{4} \left\| \left((1 - \mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1} \right) - \mathbf{u}^{k+1} \right\|_{\mathbf{H}}^2 + \delta \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ &\quad + \frac{1}{4} \left\| \left((1 - \mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1} \right) - \mathbf{u}^{k+1} \right\|_{\mathbf{H}}^2 + \delta \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\ &\quad + \mu s \beta_k (\ell(\mathbf{v}_u^{k+1}) - \ell^*), \end{aligned} \tag{31}$$

where $\text{dist}_{\mathbf{H}}^2(\mathbf{u}, \hat{\mathcal{S}}) := \inf_{\mathbf{u}' \in \hat{\mathcal{S}}} \|\mathbf{u} - \mathbf{u}'\|_{\mathbf{H}}$. Suppose $\tau_n \leq n - 1$, and by the definition of τ_n , we have

$$\begin{aligned} & \delta \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ & + \delta \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 + \frac{1}{4} \left\| \left((1 - \mu) \mathbf{v}_l^{k+1} + \mu \mathbf{v}_u^{k+1} \right) - \mathbf{u}^{k+1} \right\|_{\mathbf{H}}^2 \\ & + \mu s \beta_k \left(\ell(\mathbf{v}_u^{k+1}) - \ell^* \right) \geq 0, \end{aligned}$$

for all $\tau_n \leq k \leq n - 1$. Then

$$h_{k+1} - h_k \leq 0, \quad \tau_n \leq k \leq n - 1, \quad (32)$$

where $h_k := \frac{1}{2} \text{dist}_{\mathbf{H}}^2(\mathbf{u}^k, \hat{\mathcal{S}})$. Adding these $n - \tau_n$ inequalities, we have

$$h_n \leq h_{\tau_n}. \quad (33)$$

Eq. (33) is also true when $\tau_n = n$ because $h_{\tau_n} = h_n$. Then, once we are able to show that $\lim_{n \rightarrow \infty} h_{\tau_n} = 0$, we can obtain from Eq. (33) that $\lim_{n \rightarrow \infty} h_n = 0$.

By the definition of $\{\tau_n\}$, $\ell^* > \ell(\mathbf{v}_u^k)$ for all $k \in \{\tau_n\}$. Since U is compact, according to Lemma A.3, both $\{\mathbf{u}^{\tau_n}\}$ and $\{\mathbf{v}_u^{\tau_n}\}$ are bounded, and hence $\{h_{\tau_n}\}$ is bounded. As ℓ is assumed to be bounded below by M_0 , we have

$$0 \leq \ell^* - \ell(\mathbf{v}_u^k) \leq \ell^* - M_0.$$

According to the definition of τ_n , we have for all $k \in \{\tau_n\}$,

$$\begin{aligned} & \delta \|\mathbf{u}^{k-1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2 + \delta \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 + \frac{1}{4} \left\| \left((1 - \mu) \mathbf{v}_l^k + \mu \mathbf{v}_u^k \right) - \mathbf{u}^k \right\|_{\mathbf{H}}^2 \\ & < \mu s \beta_{k-1} \left(\ell^* - \ell(\mathbf{v}_u^k) \right) \\ & \leq \mu s \beta_{k-1} \left(\ell^* - M_0 \right). \end{aligned}$$

As $\lim_{n \rightarrow \infty} \tau_n = +\infty$, $\beta_k = \frac{1}{k+1} \rightarrow 0$, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \|\mathbf{u}^{\tau_n-1} - \mathbf{v}_l^{\tau_n}\|_{\mathbf{H}} & = 0, \\ \lim_{n \rightarrow \infty} \|\mathbf{u}^{\tau_n-1} - \mathbf{v}_u^{\tau_n}\|_{\mathbf{H}} & = 0, \\ \lim_{n \rightarrow \infty} \left\| \left((1 - \mu) \mathbf{v}_l^{\tau_n} + \mu \mathbf{v}_u^{\tau_n} \right) - \mathbf{u}^{\tau_n} \right\|_{\mathbf{H}} & = 0. \end{aligned}$$

Let $\tilde{\mathbf{u}}$ be any limit point of $\{\mathbf{u}^{\tau_n}\}$, and $\{\mathbf{u}^j\}$ be the subsequence of $\{\mathbf{u}^{\tau_n}\}$ such that

$$\lim_{j \rightarrow \infty} \mathbf{u}^j = \tilde{\mathbf{u}}.$$

As $\lim_{n \rightarrow \infty} \|\mathbf{u}^{\tau_n-1} - \mathbf{u}^{\tau_n}\|_{\mathbf{H}} \leq \lim_{n \rightarrow \infty} (\|\mathbf{u}^{\tau_n-1} - ((1 - \mu) \mathbf{v}_l^{\tau_n} + \mu \mathbf{v}_u^{\tau_n})\|_{\mathbf{H}} + \|((1 - \mu) \mathbf{v}_l^{\tau_n} + \mu \mathbf{v}_u^{\tau_n}) - \mathbf{u}^{\tau_n}\|_{\mathbf{H}}) = 0$ and $\mathbf{H} \succ 0$, we have $\lim_{j \rightarrow \infty} \mathbf{u}^{j-1} = \tilde{\mathbf{u}}$. Next, since $\lim_{\ell \rightarrow \infty} \|\mathbf{u}^{j-1} - \mathbf{v}_l^j\|_{\mathbf{H}} = 0$ and $\mathbf{H} \succ 0$, it holds that $\lim_{j \rightarrow \infty} \mathbf{v}_l^j = \tilde{\mathbf{u}}$. Then, it follows from $\mathbf{v}_l^j = \mathcal{T}(\mathbf{u}^{j-1})$, and the closedness of \mathcal{T} , we have

$$\tilde{\mathbf{u}} = \mathcal{T}(\tilde{\mathbf{u}}),$$

and thus $\tilde{\mathbf{u}} \in \mathcal{S}$. As $\ell^* > \ell(\mathbf{v}_u^k)$ for all $k \in \{\tau_n\}$ and hence $\ell^* > \ell(\mathbf{v}_u^j)$ for all j . Then it follows from the continuity of ℓ and $\lim_{n \rightarrow \infty} \|\mathbf{v}_u^{\tau_n} - \mathbf{u}^{\tau_n}\|_{\mathbf{H}} = 0$, $\mathbf{H} \succ 0$ that $\ell^* \geq \ell(\tilde{\mathbf{u}})$, which implies $\tilde{\mathbf{u}} \in \hat{\mathcal{S}}$ and $\lim_{j \rightarrow \infty} h_j = 0$. Now, as we have shown above that $\tilde{\mathbf{u}} \in \hat{\mathcal{S}}$ for any limit point $\tilde{\mathbf{u}}$ of $\{\mathbf{u}^{\tau_n}\}$, we can obtain from the boundness of $\{\mathbf{u}^{\tau_n}\}$ and $\{h_{\tau_n}\}$ that $\lim_{n \rightarrow \infty} h_{\tau_n} = 0$. Thus $\lim_{n \rightarrow \infty} h_n = 0$, and $\lim_{k \rightarrow \infty} \text{dist}(\mathbf{u}^k, \hat{\mathcal{S}}) = 0$. \square

We denote $D = \sup_{\mathbf{u}, \mathbf{u}' \in U} \|\mathbf{u} - \mathbf{u}'\|_{\mathbf{H}_{ub}}$ and $M_\ell := \sup_{\mathbf{u} \in U, \boldsymbol{\omega} \in \Omega} \left\| \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}, \boldsymbol{\omega}) \right\|$. And it should be noticed that both D and M_ℓ are finite when U and Ω are compact.

Lemma A.5 For any given $\boldsymbol{\omega} \in \Omega$, let $\{\mathbf{u}^k\}$ be the sequence generated by Eq. (17) with $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{lb})}{L_\ell})$ and $\mu \in (0, 1)$, then we have

$$\|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 \leq \|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 + \frac{\mu}{(k+1)^2} \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 + \frac{2\mu s D M_\ell}{\lambda_{\min}(\mathbf{H}) k (k+1)}.$$

Proof: Since

$$\mathbf{u}^{k+1} = \text{Proj}_{U, \mathbf{H}} (\mu \mathbf{v}_u^{k+1} + (1 - \mu) \mathbf{v}_l^{k+1}),$$

by denoting $\Delta_\beta^k := \beta_k - \beta_{k-1}$, we have the following inequality

$$\begin{aligned} & \|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 \\ & \leq \mu \|\mathbf{v}_u^{k+1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 + (1 - \mu) \|\mathbf{v}_l^{k+1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2, \\ & \leq \mu \left(\|(\mathcal{I} - s_k \mathbf{H}^{-1} \nabla \ell)(\mathbf{u}^k - \mathbf{u}^{k-1})\|_{\mathbf{H}}^2 + \frac{s^2}{k^2(k+1)^2} \|\mathbf{H}^{-1} \nabla \ell(\mathbf{u}^{k-1})\|_{\mathbf{H}}^2 \right. \\ & \quad \left. + \frac{2s}{k(k+1)} \|(\mathcal{I} - s_k \mathbf{H}^{-1} \nabla \ell)(\mathbf{u}^k - \mathbf{u}^{k-1})\|_{\mathbf{H}} \|\mathbf{H}^{-1} \nabla \ell(\mathbf{u}^{k-1})\|_{\mathbf{H}} \right) \\ & \quad + (1 - \mu) \|\mathbf{v}_l^{k+1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2 \\ & \leq \|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 + \frac{2\mu s}{k(k+1)} \|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}} \|\mathbf{H}^{-1} \nabla \ell(\mathbf{u}^{k-1})\|_{\mathbf{H}} \\ & \quad + \frac{\mu}{(k+1)^2} \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2, \end{aligned}$$

where the first inequality follows from the non-expansiveness of $\text{Proj}_{U, \mathbf{H}}$ with respect to $\|\cdot\|_{\mathbf{H}}$ and the convexity of $\|\cdot\|_{\mathbf{H}}^2$, the second inequality comes from the definition of \mathbf{v}_u^k and the last inequality follows from the definitions of \mathbf{v}_u^k , \mathbf{v}_l^k and the non-expansiveness of $\mathcal{I} - s_k \mathbf{H}^{-1} \nabla \ell$ and \mathcal{T} with respect to $\|\cdot\|_{\mathbf{H}}$ from Lemma A.3 and A.1. Then, since $\sup_{\mathbf{u}, \mathbf{u}' \in U} \|\mathbf{u} - \mathbf{u}'\|_{\mathbf{H}} \leq D$, $\sup_{\mathbf{u} \in U} \|\nabla \ell(\mathbf{u})\| \leq M_\ell$, we have the following result

$$\|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 \leq \|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 + \frac{\mu}{(k+1)^2} \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 + \frac{2\mu s D M_\ell}{\lambda_{\min}(\mathbf{H}) k(k+1)}.$$

□

Proposition A.2 For any given $\omega \in \Omega$, let $\{\mathbf{u}^k\}$ be the sequence generated by Eq. (17) with $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{lb})}{L_\ell})$ and $\mu \in (0, 1)$. Suppose $\hat{\mathcal{S}}(\omega)$ is nonempty, U is compact, $\ell(\cdot, \omega)$ is bounded below by M_0 , we have for $k \geq 2$,

$$\|\mathbf{v}_l^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 \leq (2C_2 + C_3) \frac{1 + \ln(1+k)}{k^{\frac{1}{4}}},$$

where $C_1 := (3(D^2 + 2\mu s(\ell^* - M_0)) + 2\mu s D M_\ell / \lambda_{\min}(\mathbf{H}_{lb})) / \min\{(1 - s L_\ell / \lambda_{\min}(\mathbf{H}_{lb})), \frac{1-\alpha}{\alpha}, 1\}$, $C_2 := 12D\sqrt{C_1}$ and $C_3 := \frac{\alpha(D^2 + 2\mu s(\ell^* - M_0))}{(1-\alpha)(1-\mu)}$.

Proof: Let $\bar{\mathbf{u}}$ be any point in $\hat{\mathcal{S}}$, and set \mathbf{u} in Eq. (19) of Lemma A.2 to be $\bar{\mathbf{u}}$. Then we have

$$\begin{aligned} & \frac{1}{2} \|\bar{\mathbf{u}} - \mathbf{u}^k\|_{\mathbf{H}}^2 + \frac{\mu s}{k+1} (\ell^* - \ell(\mathbf{v}_u^{k+1})) \\ & \geq \frac{1}{2} \|\bar{\mathbf{u}} - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2 + \frac{(1-\mu)(1-\alpha)}{2\alpha} \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ & \quad + \frac{\mu}{2} (1 - s_k L_\ell / \lambda_{\min}(\mathbf{H})) \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\ & \quad + \frac{1}{2} \|((1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1}) - \mathbf{u}^{k+1}\|_{\mathbf{H}}^2. \end{aligned} \tag{34}$$

Adding the Eq. (34) from $k = 0$ to $k = n - 1$ with $n \geq 1$ and since $s_k \leq s$, we have

$$\begin{aligned}
 & \frac{1}{2} \|\bar{\mathbf{u}} - \mathbf{u}^n\|_{\mathbf{H}}^2 + \frac{(1-\mu)(1-\alpha)}{2\alpha} \sum_{k=0}^{n-1} \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\
 & + \frac{\mu}{2} (1 - sL_\ell/\lambda_{\min}(\mathbf{H})) \sum_{k=0}^{n-1} \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\
 & + \frac{1}{2} \sum_{k=0}^{n-1} \left\| ((1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1}) - \mathbf{u}^{k+1} \right\|_{\mathbf{H}}^2 \\
 & \leq \frac{1}{2} \|\bar{\mathbf{u}} - \mathbf{u}^0\|_{\mathbf{H}}^2 + \sum_{k=0}^{n-1} \frac{\mu s}{k+1} (\ell^* - \ell(\mathbf{v}_u^{k+1})) \\
 & \leq \frac{1}{2} \|\bar{\mathbf{u}} - \mathbf{u}^0\|_{\mathbf{H}}^2 + \mu s (1 + \ln n) (\ell^* - M_0),
 \end{aligned} \tag{35}$$

where the last inequality follows from the assumption that $\inf \ell \geq M_0$. By Lemma A.5, we have

$$\begin{aligned}
 \|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 & \leq \|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 + \frac{\mu}{(k+1)^2} \|\mathbf{u}^{k-1} - \mathbf{v}_u^k\|_{\mathbf{H}}^2 \\
 & + \frac{2\mu s D M_\ell}{\lambda_{\min}(\mathbf{H}) k(k+1)},
 \end{aligned}$$

and thus

$$\begin{aligned}
 n \|\mathbf{u}^n - \mathbf{u}^{n-1}\|_{\mathbf{H}}^2 & \leq \sum_{k=0}^{n-1} \|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 \\
 & + \mu \sum_{k=0}^{n-2} \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 + 2\mu s D M_\ell / \lambda_{\min}(\mathbf{H}).
 \end{aligned} \tag{36}$$

Then it follows from Eq. (35) and Eq. (36) that

$$\begin{aligned}
 & \min \left\{ (1 - sL_\ell/\lambda_{\min}(\mathbf{H})), \frac{1-\alpha}{\alpha}, 1 \right\} n \|\mathbf{u}^n - \mathbf{u}^{n-1}\|_{\mathbf{H}}^2 \\
 & \leq \min \left\{ (1 - sL_\ell/\lambda_{\min}(\mathbf{H})), \frac{1-\alpha}{\alpha}, 1 \right\} \sum_{k=0}^{n-1} \|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 \\
 & + \mu (1 - sL_\ell/\lambda_{\min}(\mathbf{H})) \sum_{k=0}^{n-1} \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\
 & + 2\mu s D M_\ell / \lambda_{\min}(\mathbf{H}) \\
 & \leq \frac{2(1-\mu)(1-\alpha)}{\alpha} \sum_{k=0}^{n-1} \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\
 & + 3\mu (1 - sL_\ell/\lambda_{\min}(\mathbf{H})) \sum_{k=0}^{n-1} \|\mathbf{u}^k - \mathbf{v}_u^{k+1}\|_{\mathbf{H}}^2 \\
 & + 2 \sum_{k=0}^{n-1} \left\| (1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1} - \mathbf{u}^{k+1} \right\|_{\mathbf{H}}^2 \\
 & + 2\mu s D M_\ell / \lambda_{\min}(\mathbf{H}) \\
 & \leq 3 \left(\|\bar{\mathbf{u}} - \mathbf{u}^0\|_{\mathbf{H}}^2 + 2\mu s (1 + \ln n) (\ell^* - M_0) \right) \\
 & + 2\mu s D M_\ell / \lambda_{\min}(\mathbf{H}),
 \end{aligned}$$

where the second inequality comes from $\mathbf{u}^k - \mathbf{u}^{k+1} = (1-\mu)(\mathbf{u}^k - \mathbf{v}_l^{k+1}) + \mu(\mathbf{u}^k - \mathbf{v}_u^{k+1}) + (1-\mu)\mathbf{v}_l^{k+1} + \mu\mathbf{v}_u^{k+1} - \mathbf{u}^{k+1}$ and the convexity of $\|\cdot\|_{\mathbf{H}}^2$. Combining with the fact that $\|\bar{\mathbf{u}} - \mathbf{u}^0\|_{\mathbf{H}} \leq D$, we have

$$\|\mathbf{u}^n - \mathbf{u}^{n-1}\|_{\mathbf{H}}^2 \leq \frac{C_1(1 + \ln n)}{n}, \tag{37}$$

where $C_1 := (3(D^2 + 2\mu s(\ell^* - M_0)) + 2\mu s D M_\ell / \lambda_{\min}(\mathbf{H}_{lb})) / \min\{(1 - sL_\ell / \lambda_{\min}(\mathbf{H}_{lb})), \frac{1-\alpha}{\alpha}, 1\}$. Next, by Lemma A.3, we have for all k ,

$$\begin{aligned} \|\mathbf{v}_l^{k+1} - \mathbf{u}^k\|_{\mathbf{H}} &\leq \|\mathbf{v}_l^{k+1} - \bar{\mathbf{u}}\|_{\mathbf{H}} + \|\mathbf{u}^k - \bar{\mathbf{u}}\|_{\mathbf{H}} \\ &\leq 2\|\mathbf{u}^k - \bar{\mathbf{u}}\|_{\mathbf{H}} \leq 2D. \end{aligned}$$

Then, we have

$$\begin{aligned} &\|\mathbf{v}_l^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 \\ &\leq 2\|\mathbf{v}_l^k - \mathbf{u}^{k-1}\|_{\mathbf{H}} \|\mathbf{v}_l^{k+1} - \mathbf{u}^k - \mathbf{v}_l^k + \mathbf{u}^{k-1}\|_{\mathbf{H}} \\ &\quad + \|\mathbf{v}_l^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 + \|\mathbf{v}_l^{k+1} - \mathbf{u}^k - \mathbf{v}_l^k + \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 \\ &\leq 4D \left(\|\mathbf{v}_l^{k+1} - \mathbf{v}_l^k\|_{\mathbf{H}} + \|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}} \right) \\ &\quad + \|\mathbf{v}_l^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 + 2\|\mathbf{v}_l^{k+1} - \mathbf{v}_l^k\|_{\mathbf{H}}^2 + 2\|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 \\ &\leq \|\mathbf{v}_l^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 + 12D\|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}, \end{aligned} \tag{38}$$

where the last inequality comes from $\|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}} \leq D$, $\mathbf{v}_l^k \in \mathcal{T}(\mathbf{u}^{k-1})$ and the non-expansiveness of \mathcal{T} with respect to $\|\cdot\|_{\mathbf{H}}$. This implies that for any $n > n_0 \geq 0$,

$$\|\mathbf{v}_l^{n+1} - \mathbf{u}^n\|_{\mathbf{H}}^2 \leq 12D \sum_{k=n_0+1}^n \|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}} + \|\mathbf{v}_l^{n_0+1} - \mathbf{u}^{n_0}\|_{\mathbf{H}}^2.$$

Thus, for any $m \geq 2$ and $n_0 = n - m + 1$, the following holds

$$\begin{aligned} &m\|\mathbf{v}_l^{n+1} - \mathbf{u}^n\|_{\mathbf{H}}^2 \\ &\leq 12D \sum_{k=n_0+1}^n (k - n_0) \|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}} + \sum_{k=n_0}^n \|\mathbf{v}_l^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 \\ &\leq \sum_{k=n_0}^n \|\mathbf{v}_l^{k+1} - \mathbf{u}^k\|_{\mathbf{H}}^2 + 12D\sqrt{C_1} \frac{m(m-1)}{2} \frac{\sqrt{(1+\ln n_0)}}{\sqrt{n_0}}, \end{aligned} \tag{39}$$

where the last inequality follows from Eq. (37) that $\|\mathbf{u}^k - \mathbf{u}^{k-1}\|_{\mathbf{H}}^2 \leq \frac{C_1(1+\ln n_0)}{n_0}$ for all $k \geq n_0$, and it can be easily verified that the above inequality also holds when $m = 1$. According to Eq. (35), we have

$$\begin{aligned} &\frac{(1-\mu)(1-\alpha)}{2\alpha} \sum_{k=0}^{n-1} \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ &\leq \frac{1}{2} \|\bar{\mathbf{u}} - \mathbf{u}^0\|_{\mathbf{H}}^2 + \mu s(1 + \ln n) (\ell^* - M_0). \end{aligned}$$

Then, for any $n > 0$, let m be the smallest integer such that $m \geq n^{\frac{1}{4}}$ and let $n_0 = n - m + 1$, combining the above inequality with Eq. (39), we have

$$\begin{aligned} &\frac{\|\bar{\mathbf{u}} - \mathbf{u}^0\|_{\mathbf{H}}^2 + 2\mu s(1 + \ln(1+n))(\ell^* - M_0)}{(1-\mu)(1-\alpha)/\alpha} \\ &\geq \sum_{k=n_0}^n \|\mathbf{u}^k - \mathbf{v}_l^{k+1}\|_{\mathbf{H}}^2 \\ &\geq m\|\mathbf{v}_l^{n+1} - \mathbf{u}^n\|_{\mathbf{H}}^2 - C_2 \frac{m(m-1)}{2} \frac{\sqrt{(1+\ln n_0)}}{\sqrt{n_0}}, \end{aligned}$$

where $C_2 := 12D\sqrt{C_1}$. Next, as $n^{\frac{1}{4}} + 1 \geq m \geq n^{\frac{1}{4}}$, and hence $n_0 \geq (m-1)^4 - m + 1$. Then $16n_0 - m^2(m-1)^2 \geq (m-1)[(m-1)(3m-4)(5m-4) - 1] > 0$ when $m \geq 2$. Thus, when $n \geq 2$, we have $m \geq 2$, $m(m-1) \leq 4\sqrt{n_0}$ and thus $\frac{m(m-1)}{2} \frac{\sqrt{(1+\ln n_0)}}{\sqrt{n_0}} \leq 2\sqrt{(1+\ln n_0)}$. Then, let $C_3 := \frac{D^2 + 2\mu s(\ell^* - M_0)}{(1-\mu)(1-\alpha)/\alpha}$, we have for any $n \geq 2$,

$$\begin{aligned} \|\mathbf{v}_l^{n+1} - \mathbf{u}^n\|_{\mathbf{H}}^2 &\leq \frac{1}{m} \left(C_3(1 + \ln(1+n)) + 2C_2\sqrt{(1+\ln n_0)} \right) \\ &\leq (2C_2 + C_3) \frac{1 + \ln(1+n)}{n^{\frac{1}{4}}}, \end{aligned}$$

where the last inequality follows from $\sqrt{1 + \ln n_0} \leq 1 + \ln(1+n)$ and $m \geq n^{\frac{1}{4}}$. \square

Based on the discussion above, we can show that the sequence $\{\mathbf{u}^k(\boldsymbol{\omega})\}$ generated by Eq. (17) not only converges to the solution set of $\inf_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \boldsymbol{\omega})) \cap U} \ell(\mathbf{u}, \boldsymbol{\omega})$ but also admits a uniform convergence towards the fixed-point set $\text{Fix}(\mathcal{T}(\cdot, \boldsymbol{\omega}))$ with respect to $\|\mathbf{u}^k(\boldsymbol{\omega}) - \mathcal{T}(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})\|_{\mathbf{H}_{lb}}^2$ for $\boldsymbol{\omega} \in \Omega$.

Theorem A.1 *Let $\{\mathbf{u}^k(\boldsymbol{\omega})\}$ be the sequence generated by Eq. (17) with $\mu \in (0, 1)$ and $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{lb})}{L_\ell})$, where $\lambda_{\min}(\mathbf{H}_{lb})$ denotes the smallest eigenvalue of matrix \mathbf{H}_{lb} . Then, we have for any $\boldsymbol{\omega} \in \Omega$,*

$$\lim_{k \rightarrow \infty} \text{dist}(\mathbf{u}^k(\boldsymbol{\omega}), \text{Fix}(\mathcal{T}(\cdot, \boldsymbol{\omega}))) = 0,$$

and

$$\lim_{k \rightarrow \infty} \ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega}) = \varphi(\boldsymbol{\omega}).$$

Furthermore, there exists $C > 0$ such that for any $\boldsymbol{\omega} \in \Omega$,

$$\|\mathbf{u}^k(\boldsymbol{\omega}) - \mathcal{T}(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})\|_{\mathbf{H}_{lb}}^2 \leq C \sqrt{\frac{1 + \ln(1+k)}{k^{\frac{1}{4}}}}.$$

Proof: The property for any $\boldsymbol{\omega} \in \Omega$,

$$\begin{aligned} \lim_{k \rightarrow \infty} \text{dist}(\mathbf{u}^k(\boldsymbol{\omega}), \text{Fix}(\mathcal{T}(\cdot, \boldsymbol{\omega}))) &= 0, \\ \text{and } \lim_{k \rightarrow \infty} \ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega}) &= \varphi(\boldsymbol{\omega}). \end{aligned}$$

follows from Proposition A.1 immediately. Since U and Ω are both compact, and $\ell(\mathbf{u}, \boldsymbol{\omega})$ is continuous on $U \times \Omega$, we have that $\ell(\mathbf{u}, \boldsymbol{\omega})$ is uniformly bounded above on $U \times \Omega$ and thus $\varphi(\boldsymbol{\omega}) = \inf_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \boldsymbol{\omega})) \cap U} \ell(\mathbf{u}, \boldsymbol{\omega})$ is bounded on Ω . And combining with the assumption that $\ell(\mathbf{u}, \boldsymbol{\omega})$ is bounded below by M_0 on $U \times \Omega$, we can obtain from the Proposition A.2 that there exists $C > 0$ such that for any $\boldsymbol{\omega} \in \Omega$, we have

$$\|\mathbf{u}^k(\boldsymbol{\omega}) - \mathcal{T}(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})\|_{\mathbf{H}}^2 \leq C \sqrt{\frac{1 + \ln(1+k)}{k^{\frac{1}{4}}}}.$$

□

Thanks to the uniform convergence property of the sequence $\{\mathbf{u}^k(\boldsymbol{\omega})\}$, inspired by the arguments used in (Liu et al., 2022), we can establish the convergence on both \mathbf{u} and $\boldsymbol{\omega}$ of BMO (Algorithm 1) towards the solution of optimization problem in Eq. (4) in the following proposition and theorem.

Proposition A.3 *Suppose U and Ω are compact,*

(a) $\{\mathbf{u}^K(\boldsymbol{\omega})\} \subset U$ for any $\boldsymbol{\omega} \in \Omega$, and for any $\epsilon > 0$, there exists $k(\epsilon) > 0$ such that whenever $k > k(\epsilon)$,

$$\sup_{\boldsymbol{\omega} \in \Omega} \|\mathbf{u}^k(\boldsymbol{\omega}) - \mathcal{T}(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})\| \leq \epsilon.$$

(2) For each $\boldsymbol{\omega} \in \Omega$,

$$\lim_{k \rightarrow \infty} \varphi_k(\boldsymbol{\omega}) \rightarrow \varphi(\boldsymbol{\omega}).$$

Let $\boldsymbol{\omega}^K \in \text{argmin}_{\boldsymbol{\omega} \in \Omega} \varphi_K(\boldsymbol{\omega})$, then we have

(1) any limit point $(\bar{\mathbf{u}}, \bar{\boldsymbol{\omega}})$ of the sequence $\{(\mathbf{u}^K(\boldsymbol{\omega}^K), \boldsymbol{\omega}^K)\}$ satisfies $\bar{\boldsymbol{\omega}} \in \text{argmin}_{\boldsymbol{\omega} \in \Omega} \varphi(\boldsymbol{\omega})$ and $\bar{\mathbf{u}} = \mathcal{T}(\bar{\mathbf{u}}, \bar{\boldsymbol{\omega}})$.

(2) $\inf_{\boldsymbol{\omega} \in \Omega} \varphi_K(\boldsymbol{\omega}) \rightarrow \inf_{\boldsymbol{\omega} \in \Omega} \varphi(\boldsymbol{\omega})$ as $K \rightarrow \infty$.

Proof: For any limit point $(\bar{\mathbf{u}}, \bar{\omega})$ of the sequence $\{(\mathbf{u}^K(\omega^K), \omega^K)\}$, let $\{(\mathbf{u}^i(\omega^i), \omega^i)\}$ be a subsequence of $\{(\mathbf{u}^K(\omega^K), \omega^K)\}$ such that $\mathbf{u}^i(\omega^i) \rightarrow \bar{\mathbf{u}} \in U$ and $\omega^i \rightarrow \bar{\omega} \in \Omega$. It follows from the assumption that for any $\epsilon > 0$, there exists $k(\epsilon) > 0$ such that for any $i > k(\epsilon)$, we have

$$\|\mathbf{u}^i(\omega^i) - \mathcal{T}(\mathbf{u}^i(\omega^i), \omega^i)\| \leq \epsilon.$$

By letting $i \rightarrow \infty$, and since \mathcal{T} is closed on U , we have

$$\|\bar{\mathbf{u}} - \mathcal{T}(\bar{\mathbf{u}}, \bar{\omega})\| \leq \epsilon.$$

As ϵ is arbitrarily chosen, we have $\bar{\mathbf{u}} = \mathcal{T}(\bar{\mathbf{u}}, \bar{\omega})$ and thus $\bar{\mathbf{u}} \in \text{Fix}(\mathcal{T}(\cdot, \bar{\omega}))$.

Next, as ℓ is continuous at $(\bar{\mathbf{u}}, \bar{\omega})$, for any $\epsilon > 0$, there exists $k(\epsilon) > 0$ such that for any $i > k(\epsilon)$, it holds

$$\ell(\bar{\mathbf{u}}, \bar{\omega}) \leq \ell(\mathbf{u}^i(\omega^i), \omega^i) + \epsilon.$$

Then, we have, for any $i > k(\epsilon)$ and $\omega \in \Omega$,

$$\begin{aligned} \varphi(\bar{\omega}) &= \inf_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \bar{\omega})) \cap U} \ell(\mathbf{u}, \bar{\omega}) \\ &\leq \ell(\bar{\mathbf{u}}, \bar{\omega}) \\ &\leq \ell(\mathbf{u}^i(\omega^i), \omega^i) + \epsilon \\ &= \varphi_i(\omega^i) + \epsilon \\ &\leq \varphi_i(\omega) + \epsilon. \end{aligned} \tag{40}$$

Taking $i \rightarrow \infty$ and by the assumption, we have for any $\omega \in \Omega$,

$$\varphi(\bar{\omega}) \leq \lim_{i \rightarrow \infty} \varphi_i(\omega) + \epsilon = \varphi(\omega) + \epsilon.$$

By letting $\epsilon \rightarrow 0$, we have

$$\varphi(\bar{\omega}) \leq \varphi(\omega), \quad \forall \omega \in \Omega,$$

which implies $\bar{\omega} \in \arg \min_{\omega \in \Omega} \varphi(\omega)$. The second conclusion can be obtained by the same arguments used in the proof of (Liu et al., 2022)[Theorem 1]. \square

From Proposition A.3, we can derive the following theorem.

Theorem A.2 *Let $\{\mathbf{u}^k(\omega)\}$ be the sequence generated by Eq. (17) with $\mu \in (0, 1)$ and $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{\ell})}{L_{\ell}})$. Then, let $\omega^K \in \arg \min_{\omega \in \Omega} \varphi_K(\omega)$, and we have*

- (1) *any limit point $(\bar{\mathbf{u}}, \bar{\omega})$ of the sequence $\{(\mathbf{u}^K(\omega^K), \omega^K)\}$ is a solution to the problem in Eq. (4), i.e., $\bar{\omega} \in \arg \min_{\omega \in \Omega} \varphi(\omega)$ and $\bar{\mathbf{u}} = \mathcal{T}(\bar{\mathbf{u}}, \bar{\omega})$.*
- (2) *$\inf_{\omega \in \Omega} \varphi_K(\omega) \rightarrow \inf_{\omega \in \Omega} \varphi(\omega)$ as $K \rightarrow \infty$.*

Proof: As shown in Theorem A.1, there exists $C > 0$ such that for any $\omega \in \Omega$,

$$\|\mathbf{u}^k(\omega) - \mathcal{T}(\mathbf{u}^k(\omega), \omega)\|_{\mathbf{H}}^2 \leq C \sqrt{\frac{1 + \ln(1+k)}{k^{\frac{1}{4}}}}.$$

Since $\sqrt{\frac{1 + \ln(1+k)}{k^{\frac{1}{4}}}} \rightarrow 0$ as $k \rightarrow \infty$ and $\{\mathbf{u}^k(\omega)\} \subset U$ from Eq. (17), condition (a) in Proposition A.3 holds. Next, it follows from Theorem A.1 that $\varphi_k(\omega) = \ell(\mathbf{u}^k(\omega), \omega) \rightarrow \varphi(\omega)$ as $k \rightarrow \infty$ for any $\omega \in \Omega$ and thus condition (b) in Proposition A.3 is satisfied and the conclusion follows from Proposition A.3 immediately. \square

A.2. Stationary Analysis

Here we provide the convergence analysis of our algorithm with respect to stationary points, i.e., for any limit point $\bar{\omega}$ of the sequence $\{\omega^K\}$ satisfies $\nabla\varphi(\bar{\omega}) = 0$, where $\varphi(\omega)$ is defined in Eq. (15).

We consider the special case where $U = \mathbb{R}^n$, and $\text{Fix}(\mathcal{T}(\cdot, \omega))$ has a unique fixed point, i.e. the solution set $\mathcal{S} = \text{Fix}(\mathcal{T}(\cdot, \omega))$ is a singleton, and we denote the unique solution by $\mathbf{u}^*(\omega)$. Our analysis is partly inspired by (Liu et al., 2022) and (Grazzi et al., 2020).

Assumption A.3 Ω is a compact set and $U = \mathbb{R}^n$. $\text{Fix}(\mathcal{T}(\cdot, \omega))$ is nonempty for any $\omega \in \Omega$. $\ell(\mathbf{u}, \omega)$ is twice continuously differentiable on $\mathbb{R}^n \times \Omega$. For any $\omega \in \Omega$, $\ell(\cdot, \omega) : \mathbb{R}^n \rightarrow \mathbb{R}$ is L_ℓ -smooth, convex and bounded below by M_0 .

For $\mathcal{D}(\cdot, \omega)$ we request a stronger assumption than Assumption A.2 that $\mathcal{D}(\cdot, \omega)$ is contractive with respect to $\|\cdot\|_{\mathbf{H}_\omega}$ throughout this part, to guarantee the uniqueness of the fixed point.

Assumption A.4 There exist $\mathbf{H}_{ub} \succeq \mathbf{H}_{lb} \succ 0$, such that for each $\omega \in \Omega$, there exists $\mathbf{H}_{ub} \succeq \mathbf{H}_\omega \succeq \mathbf{H}_{lb}$ such that

(1) $\mathcal{D}(\cdot, \omega)$ is contractive with respect to $\|\cdot\|_{\mathbf{H}_\omega}$, i.e., there exists $\bar{\rho} \in (0, 1)$, such that for all $(\mathbf{u}_1, \mathbf{u}_2) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$\|\mathcal{D}(\mathbf{u}_1, \omega) - \mathcal{D}(\mathbf{u}_2, \omega)\|_{\mathbf{H}_\omega} \leq \bar{\rho} \|\mathbf{u}_1 - \mathbf{u}_2\|_{\mathbf{H}_\omega}. \quad (41)$$

(2) $\mathcal{D}(\cdot, \omega)$ is closed, i.e.,

$$\text{gph } \mathcal{D}(\cdot, \omega) := \{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \mathbf{v} = \mathcal{D}(\mathbf{u}, \omega)\}$$

is closed.

In this part, for succinctness we will write \mathbf{H} instead of \mathbf{H}_ω , and denote $\hat{\mathcal{S}}(\omega) := \text{argmin}_{\mathbf{u} \in \text{Fix}(\mathcal{T}(\cdot, \omega)) \cap U} \ell(\mathbf{u}, \omega)$. We begin with the following lemma.

Lemma A.6 (Liu et al., 2022)[Lemma 5] Let $\{a_k\}$ and $\{b_k\}$ be sequences of non-negative real numbers. Assume that $b_k \rightarrow 0$ and there exist $\rho \in (0, 1)$, and $n_0 \in \mathbb{N}$, such that $a_{k+1} \leq \rho a_k + b_k$, $\forall k \geq n_0$. Then $a_k \rightarrow 0$ as $k \rightarrow \infty$.

Proposition A.4 Suppose Assumption A.3 and Assumption A.4 are satisfied, $\frac{\partial}{\partial \mathbf{u}} \mathcal{T}(\mathbf{u}, \omega)$ and $\frac{\partial}{\partial \omega} \mathcal{T}(\mathbf{u}, \omega)$ are Lipschitz continuous with respect to \mathbf{u} , and $\hat{\mathcal{S}}(\omega)$ is nonempty for all $\omega \in \Omega$. Let $\{\mathbf{u}^k(\omega)\}$ be the sequence generated by Eq. (17) with $\mu \in (0, 1)$ and $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{lb})}{L_\ell})$. Then we have

$$\sup_{\omega \in \Omega} \|\nabla\varphi_k(\omega) - \nabla\varphi(\omega)\|_{\mathbf{H}} \rightarrow 0, \text{ as } k \rightarrow \infty.$$

Proof: According to the update scheme of \mathbf{u}^{k+1} given in Eq. (17), since $U = \mathbb{R}^n$, we have for all $\omega \in \Omega$,

$$\mathbf{u}^{k+1}(\omega) = \mu \mathbf{v}_u^{k+1}(\omega) + (1 - \mu) \mathbf{v}_l^{k+1}(\omega) = \mu \left(\mathbf{u}^k(\omega) - s_{k+1} \mathbf{H}^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right) + (1 - \mu) \mathcal{T}(\mathbf{u}^k(\omega), \omega). \quad (42)$$

As $\mathbf{u}^*(\omega)$ is the fixed point of \mathcal{T} , we have $\mathcal{T}(\mathbf{u}^*(\omega), \omega) = \mathbf{u}^*(\omega)$. Thus

$$\mathbf{u}^{k+1}(\omega) - \mathbf{u}^*(\omega) = \mu (\mathbf{u}^k(\omega) - \mathbf{u}^*(\omega)) + (1 - \mu) (\mathcal{T}(\mathbf{u}^k(\omega), \omega) - \mathcal{T}(\mathbf{u}^*(\omega), \omega)) - \mu s_{k+1} \mathbf{H}^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega).$$

Then

$$\begin{aligned} & \|\mathbf{u}^{k+1}(\omega) - \mathbf{u}^*(\omega)\|_{\mathbf{H}} \\ & \leq \mu \|\mathbf{u}^k(\omega) - \mathbf{u}^*(\omega)\|_{\mathbf{H}} + (1 - \mu) \|\mathcal{T}(\mathbf{u}^k(\omega), \omega) - \mathcal{T}(\mathbf{u}^*(\omega), \omega)\|_{\mathbf{H}} + \mu s_{k+1} \left\| \mathbf{H}^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}}. \end{aligned}$$

From the contraction of $\mathcal{D}(\cdot, \omega)$, we have $\mathcal{T}(\cdot, \omega)$ is contractive with respect to $\|\cdot\|_{\mathbf{H}}$, i.e., for all $(\mathbf{u}_1, \mathbf{u}_2) \in \mathbb{R}^n \times \mathbb{R}^n$, $\|\mathcal{T}(\mathbf{u}_1, \omega) - \mathcal{T}(\mathbf{u}_2, \omega)\|_{\mathbf{H}} \leq \rho \|\mathbf{u}_1 - \mathbf{u}_2\|_{\mathbf{H}}$, where $\rho \in (0, 1)$. The L_ℓ -smoothness of ℓ yields that

$$\left\| \mathbf{H}^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} \leq \frac{1}{\sqrt{\lambda_{\min}(\mathbf{H})}} \left\| \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\| \leq \frac{1}{\sqrt{\lambda_{\min}(\mathbf{H})}} \left(\left\| \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^*(\omega), \omega) \right\| + L_\ell \|\mathbf{u}^k(\omega) - \mathbf{u}^*(\omega)\| \right),$$

and $\left\| \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^*(\omega), \omega) \right\|$ is bounded for all $\omega \in \Omega$. Denote $M_\ell^* := \sup_{\omega \in \Omega} \left\| \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^*(\omega), \omega) \right\|$, and thus

$$\left\| \mathbf{H}^{-1} \frac{\partial}{\partial \mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} \leq \frac{1}{\sqrt{\lambda_{\min}(\mathbf{H})}} (M_\ell^* + L_\ell \|\mathbf{u}^k(\omega) - \mathbf{u}^*(\omega)\|). \quad (43)$$

Therefore,

$$\begin{aligned} & \|\mathbf{u}^{k+1}(\omega) - \mathbf{u}^*(\omega)\|_{\mathbf{H}} \\ & \leq (\mu + (1 - \mu)\rho) \|\mathbf{u}^k(\omega) - \mathbf{u}^*(\omega)\|_{\mathbf{H}} + \frac{\mu s_{k+1}}{\sqrt{\lambda_{\min}(\mathbf{H})}} (M_\ell^* + L_\ell \|\mathbf{u}^k(\omega) - \mathbf{u}^*(\omega)\|) \\ & \leq \left(\mu + (1 - \mu)\rho + \frac{\mu s_{k+1} L_\ell}{\lambda_{\min}(\mathbf{H})} \right) \|\mathbf{u}^k(\omega) - \mathbf{u}^*(\omega)\|_{\mathbf{H}} + \frac{\mu s_{k+1} M_\ell^*}{\sqrt{\lambda_{\min}(\mathbf{H})}}. \end{aligned}$$

As $s_{k+1} \rightarrow 0$ and $\rho \in (0, 1)$, there exists $n_0 \in \mathbb{N}$, such that $\mu + (1 - \mu)\rho + \frac{\mu s_{k+1} L_\ell}{\lambda_{\min}(\mathbf{H})} \in (0, 1)$, $\forall k > n_0$. Then we obtain from Lemma A.6 that

$$\sup_{\omega \in \Omega} \|\mathbf{u}^k(\omega) - \mathbf{u}^*(\omega)\|_{\mathbf{H}} \rightarrow 0, \text{ as } k \rightarrow \infty. \quad (44)$$

By taking derivative with respect to ω on both sides of Eq. (42), we have

$$\frac{\partial \mathbf{u}^{k+1}(\omega)}{\partial \omega} = \mu \left(\frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} - s_{k+1} \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} \right) + (1 - \mu) \left(\frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} + \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \omega} \right). \quad (45)$$

From $\mathcal{T}(\mathbf{u}^*(\omega), \omega) = \mathbf{u}^*(\omega)$, we obtain

$$\frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} = \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \mathbf{u}} \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} + \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \omega}. \quad (46)$$

Then combining Eq. (45) and Eq. (46) derives

$$\begin{aligned} \frac{\partial \mathbf{u}^{k+1}(\omega)}{\partial \omega} - \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} &= \mu \left(\frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} - \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right) + (1 - \mu) \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \left(\frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} - \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right) \\ &+ (1 - \mu) \left(\frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} - \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \mathbf{u}} \right) \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \\ &+ (1 - \mu) \left(\frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \omega} - \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \omega} \right) \\ &- \mu s_{k+1} \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \left(\frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} - \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right) \\ &- \mu s_{k+1} \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega}. \end{aligned}$$

Hence,

$$\begin{aligned}
 \sup_{\omega \in \Omega} \left\| \frac{\partial \mathbf{u}^{k+1}(\omega)}{\partial \omega} - \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right\|_{\mathbf{H}} &\leq \left(\mu + (1 - \mu) \sup_{\omega \in \Omega} \left\| \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \right\|_{\mathbf{H}} + \mu s_{k+1} \sup_{\omega \in \Omega} \left\| \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} \right) \\
 &\quad \times \sup_{\omega \in \Omega} \left\| \frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} - \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right\|_{\mathbf{H}} \\
 &\quad + (1 - \mu) \sup_{\omega \in \Omega} \left\| \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \mathbf{u}} - \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \right\|_{\mathbf{H}} \sup_{\omega \in \Omega} \left\| \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right\|_{\mathbf{H}} \\
 &\quad + (1 - \mu) \sup_{\omega \in \Omega} \left\| \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \omega} - \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \omega} \right\|_{\mathbf{H}} \\
 &\quad + \mu s_{k+1} \sup_{\omega \in \Omega} \left\| \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} \sup_{\omega \in \Omega} \left\| \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right\|_{\mathbf{H}}.
 \end{aligned} \tag{47}$$

Next, before using Lemma A.6, we first show that the last three terms on the right hand side of the above inequality converge to 0 as $k \rightarrow \infty$. From the Lipschitz continuity of $\frac{\partial \mathcal{T}}{\partial \mathbf{u}}$ and $\frac{\partial \mathcal{T}}{\partial \omega}$, since $\mathbf{u}^{k+1}(\omega)$ uniformly converges to $\mathbf{u}^*(\omega)$ with respect to $\|\cdot\|_{\mathbf{H}}$ as $k \rightarrow \infty$ as proved in Eq. (44), we have $\sup_{\omega \in \Omega} \left\| \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \mathbf{u}} - \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \right\|_{\mathbf{H}} \rightarrow 0$ and $\sup_{\omega \in \Omega} \left\| \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \omega} - \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \omega} \right\|_{\mathbf{H}} \rightarrow 0$ as $k \rightarrow \infty$. From Eq. (46), $\left(\mathcal{I} - \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \mathbf{u}} \right) \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} = \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \omega}$, i.e., $\frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} = \left(\mathcal{I} - \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \mathbf{u}} \right)^{-1} \frac{\partial \mathcal{T}(\mathbf{u}^*(\omega), \omega)}{\partial \omega}$. Along with the Lipschitz continuity of $\frac{\partial \mathcal{T}}{\partial \mathbf{u}}$ and $\frac{\partial \mathcal{T}}{\partial \omega}$, we have $\frac{\partial \mathbf{u}^*(\omega)}{\partial \omega}$ is continuous on the compact set Ω , and thus

$$\sup_{\omega \in \Omega} \left\| \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right\|_{\mathbf{H}} < +\infty. \tag{48}$$

From the twice continuous differentiability of ℓ , it holds that $\sup_{\omega \in \Omega} \left\| \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} \leq \sup_{\omega \in \Omega} \frac{1}{\sqrt{\lambda_{\min}(\mathbf{H})}} \left\| \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\| < +\infty$. Then from $s_{k+1} \rightarrow 0$ as $k \rightarrow \infty$, we have $\mu s_{k+1} \sup_{\omega \in \Omega} \left\| \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} \sup_{\omega \in \Omega} \left\| \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right\|_{\mathbf{H}} \rightarrow 0$ as $k \rightarrow \infty$. Thus, the last three terms in Eq.(47) converge to 0 as $k \rightarrow \infty$.

As for the coefficient $\mu + (1 - \mu) \sup_{\omega \in \Omega} \left\| \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \right\|_{\mathbf{H}} + \mu s_{k+1} \sup_{\omega \in \Omega} \left\| \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}}$ in Eq.(47), from the contraction of \mathcal{T} and the Lipschitz continuity of $\frac{\partial \mathcal{T}}{\partial \mathbf{u}}$, we have $\sup_{\omega \in \Omega} \left\| \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \right\|_{\mathbf{H}} \leq \rho < 1$. Since $\sup_{\omega \in \Omega} \left\| \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} < +\infty$ and $s_{k+1} \rightarrow 0$ as $k \rightarrow \infty$, i.e., $\mu s_{k+1} \sup_{\omega \in \Omega} \left\| \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} \rightarrow 0$ as $k \rightarrow \infty$, there exists $n_0 \in \mathbb{N}$, such that

$$\mu + (1 - \mu) \sup_{\omega \in \Omega} \left\| \frac{\partial \mathcal{T}(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \right\|_{\mathbf{H}} + \mu s_{k+1} \sup_{\omega \in \Omega} \left\| \mathbf{H}^{-1} \nabla_{\mathbf{u}\mathbf{u}} \ell(\mathbf{u}^k(\omega), \omega) \right\|_{\mathbf{H}} \in (0, 1), \forall k > n_0.$$

Therefore, by applying Lemma A.6 on Eq. (47), we obtain

$$\sup_{\omega \in \Omega} \left\| \frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} - \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} \right\|_{\mathbf{H}} \rightarrow 0, \text{ as } k \rightarrow \infty. \tag{49}$$

Finally, we will prove

$$\sup_{\omega \in \Omega} \left\| \nabla \varphi_k(\omega) - \nabla \varphi(\omega) \right\|_{\mathbf{H}} \rightarrow 0, \text{ as } k \rightarrow \infty.$$

From the definition of $\varphi_k(\omega) = \ell(\mathbf{u}^k(\omega), \omega)$ and $\varphi(\omega) = \ell(\mathbf{u}^*(\omega), \omega)$, we have

$$\begin{aligned}
 \nabla \varphi_k(\omega) &= \frac{\partial \ell(\mathbf{u}^k(\omega), \omega)}{\partial \mathbf{u}} \frac{\partial \mathbf{u}^k(\omega)}{\partial \omega} + \frac{\partial \ell(\mathbf{u}^k(\omega), \omega)}{\partial \omega}, \\
 \nabla \varphi(\omega) &= \frac{\partial \ell(\mathbf{u}^*(\omega), \omega)}{\partial \mathbf{u}} \frac{\partial \mathbf{u}^*(\omega)}{\partial \omega} + \frac{\partial \ell(\mathbf{u}^*(\omega), \omega)}{\partial \omega}.
 \end{aligned}$$

Thus,

$$\begin{aligned} \nabla\varphi_k(\boldsymbol{\omega}) - \nabla\varphi(\boldsymbol{\omega}) &= \frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} \left(\frac{\partial\mathbf{u}^k(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} - \frac{\partial\mathbf{u}^*(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} \right) \\ &\quad + \left(\frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} - \frac{\partial\ell(\mathbf{u}^*(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} \right) \frac{\partial\mathbf{u}^*(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} + \left(\frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\boldsymbol{\omega}} - \frac{\partial\ell(\mathbf{u}^*(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\boldsymbol{\omega}} \right). \end{aligned}$$

Then we have the following estimation

$$\begin{aligned} \sup_{\boldsymbol{\omega} \in \Omega} \|\nabla\varphi_k(\boldsymbol{\omega}) - \nabla\varphi(\boldsymbol{\omega})\|_{\mathbf{H}} &\leq \sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} \right\|_{\mathbf{H}} \sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\mathbf{u}^k(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} - \frac{\partial\mathbf{u}^*(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} \right\|_{\mathbf{H}} \\ &\quad + \sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} - \frac{\partial\ell(\mathbf{u}^*(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} \right\|_{\mathbf{H}} \sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\mathbf{u}^*(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} \right\|_{\mathbf{H}} \\ &\quad + \sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\boldsymbol{\omega}} - \frac{\partial\ell(\mathbf{u}^*(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\boldsymbol{\omega}} \right\|_{\mathbf{H}}. \end{aligned} \quad (50)$$

We have obtained $\sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\mathbf{u}^k(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} - \frac{\partial\mathbf{u}^*(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} \right\|_{\mathbf{H}} \rightarrow 0$, as $k \rightarrow \infty$ in Eq. (49), and $\sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\mathbf{u}^*(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}} \right\|_{\mathbf{H}} < +\infty$ in Eq. (48). Then from the L_ℓ -smoothness of $\ell(\cdot, \boldsymbol{\omega})$ and the twice continuous differentiability of ℓ on $\mathbb{R}^n \times \Omega$, where Ω is compact, we have $\sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} - \frac{\partial\ell(\mathbf{u}^*(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} \right\|_{\mathbf{H}}$ and $\sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\boldsymbol{\omega}} - \frac{\partial\ell(\mathbf{u}^*(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\boldsymbol{\omega}} \right\|_{\mathbf{H}}$ converge to 0 as $k \rightarrow \infty$. Also, it holds that $\sup_{\boldsymbol{\omega} \in \Omega} \left\| \frac{\partial\ell(\mathbf{u}^k(\boldsymbol{\omega}), \boldsymbol{\omega})}{\partial\mathbf{u}} \right\|_{\mathbf{H}} < +\infty$ from Eq. (43). Therefore, the three terms on the right hand side of Eq. (50) all converge to 0 as $k \rightarrow \infty$, which derives

$$\sup_{\boldsymbol{\omega} \in \Omega} \|\nabla\varphi_k(\boldsymbol{\omega}) - \nabla\varphi(\boldsymbol{\omega})\|_{\mathbf{H}} \rightarrow 0, \text{ as } k \rightarrow \infty.$$

□

Theorem A.3 Suppose Assumption A.3 and Assumption A.4 are satisfied, $\frac{\partial}{\partial\mathbf{u}}\mathcal{T}(\mathbf{u}, \boldsymbol{\omega})$ and $\frac{\partial}{\partial\boldsymbol{\omega}}\mathcal{T}(\mathbf{u}, \boldsymbol{\omega})$ are Lipschitz continuous with respect to \mathbf{u} , and $\hat{\mathcal{S}}(\boldsymbol{\omega})$ is nonempty for all $\boldsymbol{\omega} \in \Omega$. Let $\{\boldsymbol{\omega}^k\}$ be the sequence generated by Eq. (17) with $\mu \in (0, 1)$ and $s_k = \frac{s}{k+1}$, $s \in (0, \frac{\lambda_{\min}(\mathbf{H}_{lb})}{L_\ell})$. Let $\boldsymbol{\omega}^K$ be an ε_K -stationary point of $\varphi_K(\boldsymbol{\omega})$, i.e.,

$$\|\nabla\varphi_K(\boldsymbol{\omega}^K)\| = \varepsilon_K.$$

Then if $\varepsilon_K \rightarrow 0$, we have that any limit point $\bar{\boldsymbol{\omega}}$ of the sequence $\{\boldsymbol{\omega}^K\}$ is a stationary point of φ , i.e.,

$$\nabla\varphi(\bar{\boldsymbol{\omega}}) = 0.$$

Proof: For any limit point $\bar{\boldsymbol{\omega}}$ of the sequence $\{\boldsymbol{\omega}^K\}$, let $\{\boldsymbol{\omega}^l\}$ be a subsequence of $\{\boldsymbol{\omega}^K\}$ such that $\boldsymbol{\omega}^l \rightarrow \bar{\boldsymbol{\omega}} \in \Omega$. For any $\epsilon > 0$, as shown in Proposition A.4, there exists k_1 such that

$$\sup_{\boldsymbol{\omega} \in \Omega} \|\nabla\varphi_k(\boldsymbol{\omega}) - \nabla\varphi(\boldsymbol{\omega})\| \leq \epsilon/2, \quad \forall k \geq k_1.$$

Since $\varepsilon_k \rightarrow 0$, there exists $k_2 > 0$ such that $\varepsilon_k \leq \epsilon/2$ for any $k \geq k_2$. Then, for any $l \geq \max(k_1, k_2)$, we have

$$\|\nabla\varphi(\boldsymbol{\omega}^l)\| \leq \|\nabla\varphi(\boldsymbol{\omega}^l) - \nabla\varphi_l(\boldsymbol{\omega}^l)\| + \|\nabla\varphi_l(\boldsymbol{\omega}^l)\| \leq \epsilon.$$

Taking $l \rightarrow \infty$ in the above inequality, and by the continuity of $\nabla\varphi$, we get

$$\|\nabla\varphi(\bar{\boldsymbol{\omega}})\| \leq \epsilon.$$

Since ϵ is arbitrarily chosen, we obtain $\nabla\varphi(\bar{\boldsymbol{\omega}}) = 0$.

□

B. Detailed descriptions for \mathcal{D} in Section 4

B.1. Proximal Gradient Method (\mathcal{D}_{PG})

Consider the following convex minimization problem

$$\min_{\mathbf{u} \in \mathbb{R}^n} f(\mathbf{u}) + g(\mathbf{u}) \quad (51)$$

where $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ are proper, closed, and convex functions, and f is a continuously differentiable function with a Lipschitz continuous gradient. The proximal gradient method for solving problem Eq. (51) reads as

$$\mathbf{u}^{k+1} = \underset{\mathbf{u}}{\operatorname{argmin}} \left\{ f(\mathbf{u}^k) + \langle \nabla f(\mathbf{u}^k), \mathbf{u} - \mathbf{u}^k \rangle + g(\mathbf{u}) + \frac{1}{2\gamma} \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{G}}^2 \right\}, \quad (52)$$

where $\mathbf{G} \succeq 0$ and $\|\mathbf{u}\|_{\mathbf{G}}^2 := \langle \mathbf{u}, \mathbf{G}\mathbf{u} \rangle$. By parameterizing functions f, g and matrix \mathbf{G} by hyper-parameter $\boldsymbol{\omega}$ in Eq. (52) to make them learnable, we can obtain \mathcal{D}_{PG} in the following form,

$$\mathcal{D}_{\text{PG}}(\mathbf{u}^k, \boldsymbol{\omega}) = \underset{\mathbf{u}}{\operatorname{argmin}} \left\{ f(\mathbf{u}^k, \boldsymbol{\omega}) + \langle \nabla_{\mathbf{u}} f(\mathbf{u}^k, \boldsymbol{\omega}), \mathbf{u} - \mathbf{u}^k \rangle + g(\mathbf{u}, \boldsymbol{\omega}) + \frac{1}{2\gamma} \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{G}(\boldsymbol{\omega})}^2 \right\}. \quad (53)$$

Next we show that \mathcal{D}_{PG} satisfies Assumption A.2 under the following standing assumption.

Assumption B.1 For any $\boldsymbol{\omega} \in \Omega$, $f(\cdot, \boldsymbol{\omega})$ and $g(\cdot, \boldsymbol{\omega})$ are proper closed convex functions and $f(\cdot, \boldsymbol{\omega})$ are L_f -smooth. And there exist $\mathbf{H}_{ub} \succeq \mathbf{H}_{lb} \succ 0$ such that $\mathbf{H}_{ub} \succeq \mathbf{G}(\boldsymbol{\omega}) \succeq \mathbf{H}_{lb}$ for each $\boldsymbol{\omega} \in \Omega$.

Proposition B.1 Suppose Assumption B.1 holds and $\gamma \in (0, 2\lambda_{\min}(\mathbf{H}_{lb})/L_f)$. Then \mathcal{D}_{PG} satisfies Assumption A.2.

Proof: Since $\mathcal{D}_{\text{PG}}(\cdot, \boldsymbol{\omega}) = (\mathcal{I} + \gamma\mathbf{G}(\boldsymbol{\omega})^{-1}\partial_{\mathbf{u}}g(\cdot, \boldsymbol{\omega}))^{-1}(\mathcal{I} - \gamma\mathbf{G}(\boldsymbol{\omega})^{-1}\nabla_{\mathbf{u}}f(\cdot, \boldsymbol{\omega}))$, and $\nabla_{\mathbf{u}}f(\cdot, \boldsymbol{\omega})$ is L_f -Lipschitz continuous, by (Cui et al., 2019)[Lemma 3.2], (Bauschke et al., 2011)[Proposition 4.25] and Assumption B.1, we have $\mathcal{D}_{\text{PG}}(\cdot, \boldsymbol{\omega})$ satisfies Assumption A.2 (1) when setting $\mathbf{G}(\boldsymbol{\omega})$ as $\mathbf{H}_{\boldsymbol{\omega}}$ for any $\boldsymbol{\omega} \in \Omega$, with $\gamma \in (0, 2\lambda_{\min}(\mathbf{G}_{lb})/L_f)$. The closedness of $\mathcal{D}_{\text{PG}}(\cdot, \boldsymbol{\omega})$ follows from the outer-semicontinuity of $\partial_{\mathbf{u}}g(\cdot, \boldsymbol{\omega})$ and continuity of $\nabla_{\mathbf{u}}f(\cdot, \boldsymbol{\omega})$, and then the proof is completed. \square

B.2. Proximal Augmented Lagrangian Method (\mathcal{D}_{ALM})

Consider the following convex minimization problem with linear constraints

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{R}^n} f(\mathbf{u}) \\ \text{s.t. } \mathcal{A}\mathbf{u} = \mathbf{b}, \end{aligned} \quad (54)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a proper closed convex function. Proximal ALM for solving problem Eq. (54) is given by

$$\begin{aligned} \mathbf{u}^{k+1} &= \underset{\mathbf{u}}{\operatorname{argmin}} \left\{ f(\mathbf{u}) + \langle \boldsymbol{\lambda}^k, \mathcal{A}\mathbf{u} - \mathbf{b} \rangle + \frac{\beta}{2} \|\mathcal{A}\mathbf{u} - \mathbf{b}\|^2 + \frac{1}{2} \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{G}}^2 \right\}, \\ \boldsymbol{\lambda}^{k+1} &= \boldsymbol{\lambda}^k + \beta(\mathcal{A}\mathbf{u}^{k+1} - \mathbf{b}), \end{aligned} \quad (55)$$

where $\beta > 0$. We can parameterize functions f , matrices \mathcal{A} and \mathbf{G} , and vector \mathbf{b} by hyper-parameter θ in Eq. (55) to make them learnable as the following

$$\begin{aligned} \mathbf{u}^{k+1} &= \underset{\mathbf{u}}{\operatorname{argmin}} \left\{ f(\mathbf{u}, \theta) + \langle \boldsymbol{\lambda}^k, \mathcal{A}(\theta)\mathbf{u} - \mathbf{b} \rangle + \frac{\beta}{2} \|\mathcal{A}(\theta)\mathbf{u} - \mathbf{b}(\theta)\|^2 + \frac{1}{2} \|\mathbf{u} - \mathbf{u}^k\|_{\mathbf{G}(\theta)}^2 \right\}, \\ \boldsymbol{\lambda}^{k+1} &= \boldsymbol{\lambda}^k + \beta(\mathcal{A}(\theta)\mathbf{u}^{k+1} - \mathbf{b}(\theta)). \end{aligned} \quad (56)$$

By setting β as the hyper-parameter and letting $\boldsymbol{\omega} := (\theta, \beta)$ as hyper-parameters for scheme Eq. (56), we can define \mathcal{D}_{ALM} by scheme Eq. (56) as the following

$$\mathcal{D}_{\text{ALM}}(\mathbf{u}^k, \boldsymbol{\omega}) = \mathbf{u}^{k+1}. \quad (57)$$

We can show that \mathcal{D}_{ALM} satisfies Assumption A.2 under the following standing assumption.

Assumption B.2 For any $\omega \in \Omega$, $f(\cdot, \theta)$ is proper closed convex functions. And there exist $\beta_{ub} \geq \beta_{lb} > 0$ and $\mathbf{G}_{ub} \succeq \mathbf{G}_{lb} \succ 0$ such that $\beta \in [\beta_{lb}, \beta_{ub}]$ and $\mathbf{G}_{ub} \succeq \mathbf{G}(\omega) \succeq \mathbf{G}_{lb}$ for each $\omega \in \Omega$.

Proposition B.2 Suppose Assumption B.2 holds. Then \mathcal{D}_{ALM} satisfies Assumption A.2.

Proof: As shown in (He et al., 2020), $\mathcal{D}_{\text{ALM}}(\cdot, \omega) = (\Phi_\omega + \mathbf{H}_\omega)^{-1} \mathbf{H}_\omega$ with

$$\Phi_\omega(\mathbf{u}, \omega) = \begin{pmatrix} \partial_{\mathbf{u}} f(\mathbf{u}, \theta) + \mathcal{A}(\theta)^\top \boldsymbol{\lambda} \\ -\mathcal{A}(\theta) \mathbf{u} + \mathbf{b}(\theta) \end{pmatrix} \quad \text{and} \quad \mathbf{H}_\omega = \begin{pmatrix} \mathbf{G}(\theta) & 0 \\ 0 & \frac{1}{\beta} \mathbf{I} \end{pmatrix}.$$

Since Φ_ω is maximal monotone mapping (see, e.g. (Rockafellar, 1970)[Corollary 37.5.2]), and $\mathbf{H}_\omega \succ 0$ from Assumption B.2, it can be easily verified that $\mathcal{D}_{\text{ALM}}(\cdot, \omega)$ is firmly non-expansive with respect to $\|\cdot\|_{\mathbf{H}_\omega}$, and thus $\mathcal{D}_{\text{ALM}}(\cdot, \omega)$ satisfies Assumption A.2 (1) for any $\omega \in \Omega$. The closedness of $\mathcal{D}_{\text{ALM}}(\cdot, \omega)$ follows from the outer-semicontinuity of $\partial_{\mathbf{u}} f(\cdot, \omega)$. Next, Assumption B.2 implies the existences of $\mathbf{H}_{ub} \succeq \mathbf{H}_{lb} \succ 0$ such that $\mathbf{H}_{ub} \succeq \mathbf{H}_\omega \succeq \mathbf{H}_{lb}$ for each $\omega \in \Omega$ and then the conclusion follows immediately. \square

B.3. Composition of \mathcal{D}_{num} and \mathcal{D}_{net} ($\mathcal{D}_{\text{num}} \circ \mathcal{D}_{\text{net}}$)

Proposition B.3 Suppose \mathcal{D}_{num} and \mathcal{D}_{net} satisfy Assumption A.2 with the same \mathbf{H}_ω . Then $\mathcal{D}_{\text{num}} \circ \mathcal{D}_{\text{net}}$ satisfies Assumption A.2.

Proof: Since \mathcal{D}_{num} and \mathcal{D}_{net} satisfy Assumption A.2 with the same \mathbf{H}_ω , it can be easily verified from the definition that $\mathcal{D}_{\text{num}} \circ \mathcal{D}_{\text{net}}$ satisfies Assumption A.2 (1) with \mathbf{H}_ω . For the closedness of $\mathcal{D}_{\text{num}}(\cdot, \omega) \circ \mathcal{D}_{\text{net}}(\cdot, \omega)$, for any fixed $\omega \in \Omega$, consider any sequence $\{(\mathbf{u}^k, \mathbf{v}^k)\} \in \text{gph}(\mathcal{D}_{\text{num}}(\cdot, \omega) \circ \mathcal{D}_{\text{net}}(\cdot, \omega))$ satisfying $(\mathbf{u}^k, \mathbf{v}^k) \rightarrow (\bar{\mathbf{u}}, \bar{\mathbf{v}})$. Since $\mathcal{D}_{\text{net}}(\cdot, \omega)$ satisfies Assumption A.2 (1) with $\mathbf{H}_\omega \succ 0$, and $\{\mathbf{u}^k\}$ is bounded, we have $\mathcal{D}_{\text{net}}(\mathbf{u}^k, \omega)$ is bounded, and thus there exists a subsequence $\{(\mathbf{u}^i, \mathbf{v}^i)\} \subseteq \{(\mathbf{u}^k, \mathbf{v}^k)\}$ such that $\mathcal{D}_{\text{net}}(\mathbf{u}^i, \omega) \rightarrow \bar{\mathbf{v}}$. Then it follows from the closedness of $\mathcal{D}_{\text{net}}(\cdot, \omega)$ and $\mathcal{D}_{\text{num}}(\cdot, \omega)$, that $(\bar{\mathbf{u}}, \bar{\omega}) \in \text{gph} \mathcal{D}_{\text{net}}(\cdot, \omega)$ and $(\bar{\omega}, \bar{\mathbf{v}}) \in \text{gph} \mathcal{D}_{\text{num}}(\cdot, \omega)$, and thus $(\bar{\mathbf{u}}, \bar{\mathbf{v}}) \in \text{gph}(\mathcal{D}_{\text{num}}(\cdot, \omega) \circ \mathcal{D}_{\text{net}}(\cdot, \omega))$. Therefore, the closedness of $\mathcal{D}_{\text{num}}(\cdot, \omega) \circ \mathcal{D}_{\text{net}}(\cdot, \omega)$ follows and the proof is completed. \square

Remark 1 Given any non-expansive \mathcal{D}_{net} (which can be achieved by spectral normalization) and any positive-definite matrix \mathbf{H}_ω , by setting $\mathcal{D}_{\text{net}^*} = \mathbf{H}_\omega^{-1/2} \mathcal{D}_{\text{net}} \mathbf{H}_\omega^{1/2}$, we can obtain that $\mathcal{D}_{\text{net}^*}$ satisfies Assumption A.2 with \mathbf{H}_ω .

B.4. Summary of Operator \mathcal{D}

Table 4. Summary of operator \mathcal{D} and \mathbf{H}_ω .

\mathcal{D}	Operator	\mathbf{H}_ω
\mathcal{D}_{num}	PG : $\mathbf{u}^{k+1} = \underset{\mathbf{u}}{\text{argmin}} \left\{ f(\mathbf{u}^k, \omega) + \langle \nabla_{\mathbf{u}} f(\mathbf{u}^k, \omega), \mathbf{u} - \mathbf{u}^k \rangle + g(\mathbf{u}, \omega) + \frac{1}{2\gamma} \ \mathbf{u} - \mathbf{u}^k\ _{\mathbf{G}(\omega)}^2 \right\}$ ALM : $\begin{cases} \mathbf{u}^{k+1} = \underset{\mathbf{u}}{\text{argmin}} \left\{ f(\mathbf{u}, \theta) + \langle \boldsymbol{\lambda}^k, \mathcal{A}(\theta) \mathbf{u} - \mathbf{b} \rangle + \frac{\beta}{2} \ \mathcal{A}(\theta) \mathbf{u} - \mathbf{b}(\theta)\ ^2 + \frac{1}{2} \ \mathbf{u} - \mathbf{u}^k\ _{\mathbf{G}(\theta)}^2 \right\} \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \beta(\mathcal{A}(\theta) \mathbf{u}^{k+1} - \mathbf{b}(\theta)) \end{cases}$	$\mathbf{G}(\omega)$ $\begin{pmatrix} \mathbf{G}(\theta) & 0 \\ 0 & \frac{1}{\beta} \mathbf{I} \end{pmatrix}$
\mathcal{D}_{net}	Various networks with non-expansive property(1-Lipschitz continuous)	\mathbf{I}
$\mathcal{D}_{\text{num}} \circ \mathcal{D}_{\text{net}}$	$\mathcal{D}_{\text{num}} \circ \left(\mathbf{H}_{\text{num}, \omega}^{-1/2} \mathcal{D}_{\text{net}} \mathbf{H}_{\text{num}, \omega}^{1/2} \right)$ where \mathcal{D}_{net} is non-expansive	$\mathbf{H}_{\text{num}, \omega}$

C. Experimental Details

Our experiments were mainly conducted on a PC with Intel Core i9-10900KF CPU (3.70GHz), 128GB RAM and two NVIDIA GeForce RTX 3090 24GB GPUs. In all experiments, we use synthetic datasets, and adopt the Adam optimizer for updating variable ω .

C.1. Sparse Coding

For sparse coding, we set batch size=128, random seed=1126, training set size=10000. The testing set size depends on the size of each image. Because we conduct unsupervised single image training, we do not use the MSE loss between the clear picture and the generated picture as the upper loss, but instead use the same unsupervised loss as in (Xie et al., 2019). Here \mathcal{D}_{net} is set to be $\mathcal{D}_{\text{DLADMM}}$ and is given as follows

$$\begin{aligned} \mathbf{u}_1^{k+1} &= \underset{\mathbf{u}_1}{\operatorname{argmin}} \left\{ \|\mathbf{u}_1\|_1 + \langle \beta \mathbf{Q}^T (\mathbf{Q}\mathbf{u}_1^k + \mathbf{u}_2^k - \mathbf{b} + \boldsymbol{\lambda}^k / \beta), \mathbf{u}_1 \rangle + \frac{\rho_1}{2} \|\mathbf{u}_1 - \mathbf{u}_1^k\|^2 \right\}, \\ \mathbf{u}_2^{k+1} &= \underset{\mathbf{u}_2}{\operatorname{argmin}} \left\{ \|\mathbf{u}_2\|_1 + \langle \beta (\mathbf{Q}\mathbf{u}_1^{k+1} + \mathbf{u}_2^k - \mathbf{b} + \boldsymbol{\lambda}^k / \beta), \mathbf{u}_2 \rangle + \frac{\rho_2}{2} \|\mathbf{u}_2 - \mathbf{u}_2^k\|^2 \right\}, \\ \boldsymbol{\lambda}^{k+1} &= \boldsymbol{\lambda}^k + \gamma \beta (\mathbf{Q}\mathbf{u}_1^{k+1} + \mathbf{u}_2^{k+1} - \mathbf{b}). \end{aligned} \quad (58)$$

For comparing with DLADMM more directly, we set \mathcal{D}_{num} to be \mathcal{I} (the identity operator) here. We choose $(\rho_1, \rho_2) \in \{\rho_1 \geq \beta L_{\mathbf{Q}}^2, \rho_2 \geq \beta L_{\mathbf{I}}^2\}$ to ensure the non-expansive hypothesis. All methods follow the general setting of hyper-parameters given in Table 5.

Table 5. Values for hyper-parameters of sparse coding.

Hyper-parameters	Value	Hyper-parameters	Value
Epochs	100	Learning rate	$0.0002 * 0.5^{\text{epoch}/30}$
Stage	15	Ratio β	0.1
Batch size	128	Ratio γ	1

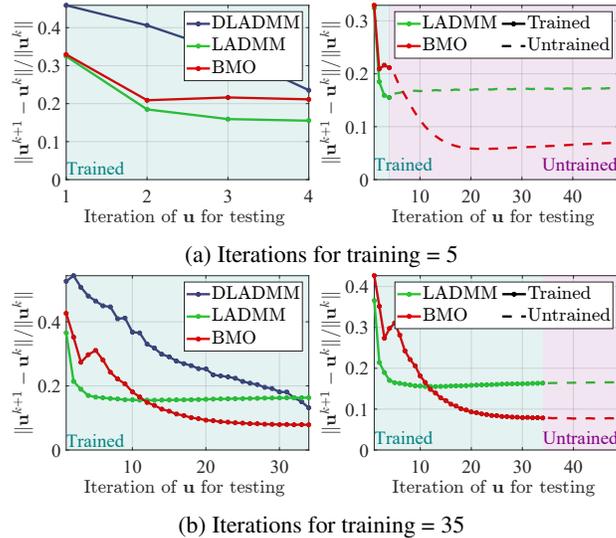


Figure 6. Convergence curves of $\|\mathbf{u}^{k+1} - \mathbf{u}^k\| / \|\mathbf{u}^k\|$ with respect to k , the number of iterations of \mathbf{u} for testing, after (a) 5 and (b) 35 iterations for training. Solid lines on the right column represent the iterations for testing are less than those for training (trained iterations), while dotted lines represent the iterations for testing are more than those for training (untrained iterations). It can be seen that our method can successfully learn a non-expansive mapping with different number of iterations for training.

In addition to Figure 3, we show more results to demonstrate the impact on the number of iterations for training in Figure 6, and we can see that our method remain stable when the number of iterations for training changes. Note that for DLADMM,

the number of iterations for training have to be more than those for testing, so in the right column we only show the curves of LADMM and BMO.

C.2. Image Deconvolution

As for network architectures \mathcal{D}_{net} , we use DRUNet which consists of four scales. Each scale has an identity skip connection between 2×2 strided convolution (SConv) downscaling and 2×2 transposed convolution (TConv) upscaling operators. The number of channels in each layer from the first scale to the fourth scale are 64, 128, 256 and 512, respectively. Four successive residual blocks are adopted in the downscaling and upscaling of each scale. For numerical update \mathcal{D}_{num} , by using the auxiliary variable $\mathbf{z} = \mathbf{W}\mathbf{u}$, we transform the problem as $\|\mathbf{Q}\mathbf{W}^{-1}\mathbf{z} - \mathbf{b}\|_2^2$ with a regularization term $\|\mathbf{z}\|_1$. For this application, the numerical operator \mathcal{D}_{PG} is given by

$$\mathbf{z}^{k+1} = \underset{\mathbf{u}}{\operatorname{argmin}} \left\{ \|\mathbf{z}^k\|_1 + \langle \mathbf{W}^{-\top} \mathbf{Q}^\top (\mathbf{Q}\mathbf{W}^{-1}\mathbf{z}^k - \mathbf{b}), \mathbf{z} - \mathbf{z}^k \rangle + \frac{1}{2} \|\mathbf{z} - \mathbf{z}^k\|_{\mathbf{G}(\omega)}^2 \right\}, \quad (59)$$

where $\mathbf{G}(\omega)$ is a parameterized diagonal matrix defined in Appendix B.2. Here we use MSE as the upper loss function. We follow the general setting of hyper-parameters given in Table 6.

Table 6. Values for hyper-parameters of image deconvolution.

Hyper-parameters	Value	Hyper-parameters	Value
Epochs	10000	Learning rate	0.0001
Stage	8	Ratio μ	0.3
Batch size	1	Ratio α	0.9

C.3. Rain Streak Removal

In the rain streak removal task, for dataset, we use Rain100L and Rain100H (Yang et al., 2019). For network architecture \mathcal{D}_{net} , we use a 3-layer convolutional network with \mathbf{u}_b , \mathbf{u}_r and \mathbf{b} as the network input to estimate \mathbf{u}_r , and a 2-layer convolutional network with \mathbf{u}_r and \mathbf{b} as the input to estimate \mathbf{u}_b . In the network for estimating \mathbf{u}_r , we use some prior information of \mathbf{u}_r as input just like (Wang et al., 2020). Such a problem falls into the form of problem in Eq. (54) as the following,

$$f(\mathbf{u}) = \frac{1}{2} \|\mathbf{u}_b + \mathbf{u}_r - \mathbf{b}\|_2^2 + \kappa_b \|\mathbf{v}_b\|_1 + \kappa_r \|\mathbf{v}_r\|_1 \text{ s.t. } \mathcal{A}\mathbf{u} = 0,$$

where

$$\mathbf{u} = [\mathbf{u}_b; \mathbf{u}_r; \mathbf{v}_b; \mathbf{v}_r], \mathcal{A} = \begin{pmatrix} \mathbf{I} & 0 & -\mathbf{I} & 0 \\ 0 & \nabla & 0 & -\mathbf{I} \end{pmatrix}.$$

Then the numerical operator \mathcal{D}_{ALM} with hyper-parameters β , $\rho_{\mathbf{u}_b}$, $\rho_{\mathbf{u}_r}$, $\rho_{\mathbf{v}_b}$ and $\rho_{\mathbf{v}_r}$ reads as

$$\begin{aligned} \mathbf{u}_b^{k+1} &= \arg \min_{\mathbf{u}_b} \left\{ \langle (\mathbf{u}_b^k + \mathbf{u}_r^k - \mathbf{b}) + \beta \nabla^\top (\nabla \mathbf{u}_b^k - \mathbf{v}_b^k) + \boldsymbol{\lambda}_b^k, \mathbf{u}_b \rangle + \frac{\rho_{\mathbf{u}_b}}{2} \|(\mathbf{u}_b - \mathbf{u}_b^k)\|_2^2 \right\}, \\ \mathbf{u}_r^{k+1} &= \arg \min_{\mathbf{u}_r} \left\{ \langle (\mathbf{u}_b^k + \mathbf{u}_r^k - \mathbf{b}) + \beta (\mathbf{u}_r^k - \mathbf{v}_r^k) + \boldsymbol{\lambda}_r^k, \mathbf{u}_r \rangle + \frac{\rho_{\mathbf{u}_r}}{2} \|(\mathbf{u}_r - \mathbf{u}_r^k)\|_2^2 \right\}, \\ \mathbf{v}_b^{k+1} &= \arg \min_{\mathbf{v}_b} \left\{ \kappa_b \|\mathbf{v}_b\|_1 + \langle -\beta (\nabla \mathbf{u}_b^{k+1} - \mathbf{v}_b^k) + \boldsymbol{\lambda}_b^k, \mathbf{v}_b \rangle + \frac{\rho_{\mathbf{v}_b}}{2} \|(\mathbf{v}_b - \mathbf{v}_b^k)\|_2^2 \right\}, \\ \mathbf{v}_r^{k+1} &= \arg \min_{\mathbf{v}_r} \left\{ \kappa_r \|\mathbf{v}_r\|_1 + \langle -\beta (\mathbf{u}_r^{k+1} - \mathbf{v}_r^k) + \boldsymbol{\lambda}_r^k, \mathbf{v}_r \rangle + \frac{\rho_{\mathbf{v}_r}}{2} \|(\mathbf{v}_r - \mathbf{v}_r^k)\|_2^2 \right\}, \\ \boldsymbol{\lambda}_b^{k+1} &= \boldsymbol{\lambda}_b^k + \beta (\nabla \mathbf{u}_b^{k+1} - \mathbf{v}_b^{k+1}), \\ \boldsymbol{\lambda}_r^{k+1} &= \boldsymbol{\lambda}_r^k + \beta (\mathbf{u}_r^{k+1} - \mathbf{v}_r^{k+1}). \end{aligned} \quad (60)$$

By setting $\omega = (\beta, \rho_{\mathbf{u}_b}, \rho_{\mathbf{u}_r}, \rho_{\mathbf{v}_b})$, we have

$$\mathbf{H}_\omega = \begin{pmatrix} \rho_{\mathbf{u}_b} - \beta \mathbf{I} & 0 & 0 & 0 & 0 \\ 0 & \rho_{\mathbf{u}_r} - \beta \nabla^\top \nabla & 0 & 0 & 0 \\ 0 & 0 & \rho_{\mathbf{v}_b} - \beta \mathbf{I} & 0 & 0 \\ 0 & 0 & 0 & \rho_{\mathbf{u}_b} - \beta \mathbf{I} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\beta} \mathbf{I} \end{pmatrix}. \quad (61)$$

In practice, we choose Ω such that $\mathbf{H}_\omega \succ 0$ for each $\omega \in \Omega$, and \mathbf{H}_ω can be inverted quickly by Fourier transform. Here we use MSE as the upper loss function. We follow the general setting of hyper-parameters given in Table 7.

Table 7. Values for hyper-parameters of rain streak removal.

Hyper-parameter	Value	Hyper-parameter	Value
Epochs	100	Learning rate	0.001
Stage	17	Ratio μ	0.1
Batch size	16	Ratio α	0.9