DROPOUT METHODS FOR BILEVEL TRAINING TASK

Peiran Yu *
Department of Computer Science
University of Maryland
College Park, MD 20740, USA
{pyu123}@umd.edu

Junyi Li
Department of Electrical and Computer Engineering University of Pittsburgh
Pittsburgh, PA 15260
{junyili.ai}@gmail.com

Heng Huang †
Department of Computer Science
University of Maryland
College Park, MD 20740, USA
{henghuanghh}@gmail.com

ABSTRACT

Bilevel optimization problems appear in many widely used machine learning tasks. Bilevel optimization models are sensitive to small changes, and bilevel training tasks typically involve limited datasets. Therefore, overfitting is a common challenge in bilevel training tasks. This paper considers the use of dropout to address this problem. We propose a bilevel optimization model that depends on the distribution of dropout masks. We investigate how the dropout rate affects the hypergradient of this model. We propose a dropout bilevel method to solve the dropout bilevel optimization model. Subsequently, we analyze the resulting dropout bilevel method from an optimization perspective. Analyzing the optimization properties of methods with dropout is essential because it provides convergence guarantees for methods using dropout. However, there has been limited investigation in this research direction. We provide the complexity of the resulting dropout bilevel method in terms of reaching an ϵ stationary point of the proposed stochastic bilevel model. Empirically, we demonstrate that overfitting occurs in data cleaning and meta-learning, and the method proposed in this work mitigates this issue.

1 INTRODUCTION

Bilevel optimization appear in many problems widely used in machine learning tasks such as data cleaning (Shaban et al., 2019), hyperparameter optimization, meta-learning (Franceschi et al., 2018) and reinforcement learning (Yang et al., 2019). The bilevel optimization problem involves two minimization problems that are stacked on top of each other, where the solution to one optimization problem depends on the solution of the other. Take data cleaning as an example. Suppose we have a corrupted training data set $\mathcal{D}_{tr}$ with N_{tr} data points and a clean data set $\mathcal{D}_{val}$ with N_{val} data points. Let $f(w;\xi)$ be a network parametrized by w. The aim of data cleaning is to train the model with corrupted data by solving

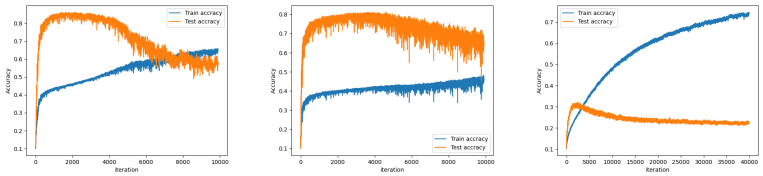
$$\min_{\lambda\in\mathbb{R}^{N_{tr}}} \frac{1}{|N_{val}|}\sum_{i=1}^{N_{val}} l(f(w(\lambda);\eta_i);\eta_i) \text{ s.t.} w(\lambda) \in \operatorname*{Arg\,min}_{w} \frac{1}{|N_{tr}|}\sum_{i=1}^{N_{tr}} \sigma(\lambda_i) l(f(w;\xi_i);\xi_i) \tag{1}$$

where $\eta_i \in \mathcal{D}_{val}$, $\xi_i \in \mathcal{D}_{tr}$, l is the loss function and $\sigma(\lambda_i)$ is the weight of the loss w.r.t. the data ξ_i. If ξ_i is corrupted, we hope $\sigma(\lambda_i)$ can neutralize the influence of ξ_i.

Another important machine learning task using bilevel optimization is the meta-learning (Franceschi et al., 2018; Lorraine et al., 2020b). Suppose we have N different training tasks. The jth task

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper.

†



(a) MNIST (b) FMNIST (c) CIFAR10

Figure 1: Overfitting in data cleaning tasks

has a pair of training and validation data sets $\{(\mathcal{D}_{j,v}, \mathcal{D}_{j,tr})\}$. The j-th task is to train a network $f(w_j, \lambda; \xi)$, where ξ is an input, w_j denotes the parameters for task j and λ denotes the parameters shared for all N tasks. The goal of meta-learning is to find good shared parameters λ. To this end, the following bilevel model is considered:

$$\min_{\lambda} \frac{1}{N} \sum_{j=1}^{N} l_j(w_j, \lambda; \mathcal{D}_{j,v}) \text{ s.t. } w(\lambda) \in \underset{w:=(w_1,\ldots,w_N)}{\operatorname{Arg\,min}} \frac{1}{N} \sum_{j=1}^{N} l_j(w_j, \lambda; \mathcal{D}_{j,tr}), \tag{2}$$

where $l_j(w_j, \lambda; \mathcal{D})$ is the loss of the network f for the jth task on a data set $\mathcal{D}$.

As shown in (1) and (2), bilevel training tasks are more complex than single level training tasks. Thus, bilevel optimization models can be sensitive to changes in the input data and parameters. In addition, bilevel training tasks usually suffer from limited data sets. As evidenced by Figure 1, in the data cleaning problem, overfitting happens since the increase in the training accuracy leads to a decrease in the testing accuracy of the classifier. Regarding meta-learning, overfitting in the model learned through bilevel optimization can result in a diminished ability to learn new tasks. As shown in our experiments section, in classification problems, the training accuracy is high while the testing accuracy is not. Therefore, overfitting is a common challenge in bilevel training tasks and we need to addressing it.

In single level machine learning optimization problems such as empirical risk minimization, there are many ways to redeem the overfitting problem, including dropout, early stopping, data augmentation, etc. Among these methods, dropout is highly effective, simple to apply and computationally efficient, (Srivastava et al., 2014; Labach et al., 2019). Proposed by Srivastava et al. (2014), dropout randomly drops neurons of the neural network during training, leading to only a portion of parameters being updated in each epoch.Dropout has been successfully applied to fully connected networks (Ba & Frey, 2013; Srivastava et al., 2014), convolutional neural networks (CNNs) (Wu & Gu, 2015; Srivastava et al., 2014; Park & Kwak, 2017), and recurrent layers in recurrent neural networks (RNNs) (Pham et al., 2014; Zaremba et al., 2014). In this work, we investigate how dropout behaves when adopted to solving the bilevel optimization problem.

Since the dropout only randomly zero out part of the neurons in a network, we propose a bilevel optimization model that characterizes this randomness. Since current existing bilevel methods are not directly applicable to this new model, we use a representative bilevel method from Li et al. (2022) as an example to see how existing bilevel methods can be adapted to this new model. We then investigate theoretical convergence guarantees of the resulting dropout bilevel method. Existing analysis of the dropout method from an optimization perspective is very limited. As far as we search, only (Li et al., 2016; Mianjy & Arora, 2020; Senen-Cerda & Sanders, 2020; Senen-Cerda & Sanders, 2022) investigated this direction previously. However, these works only focused on single level optimization problems. For bilevel problems, there are no theoretical convergence guarantees for dropout methods. In this work, we fill in this gap and study the convergence properties for a dropout bilevel method. The challenges in analyzing the dropout bilevel method, which the single-level model considered in (Senen-Cerda & Sanders, 2020) does not face, include how to analyze the hypergradient. Analyzing the hypergradient is hard to analyze in two ways: 1. it is related to the solution of the lower level problem; 2. both upper and lower level objectives are composed with random dropout masks.

Our contributions are summarized as follows:

- We form a statistical bilevel optimization problem that includes the distribution of the dropout masks. To solve this problem, we propose a dropour variant of an existing method

from (Li et al., 2022) as an example of a bilevel method. We investigate the inductive relations between the variables that are used in the dropout bilevel method to approximate the hypergradients of the outer-level objective. These variables are affected by the dropout masks, and the variance term in the inductive relations is affected by the dropout rates.

- We show that the complexity of the dropout bilevel method depends on the dropout rate. Unlike the complexity of the bilevel methods for the bilevel model without dropout, the complexity of the dropout bilevel method for the dropout bilevel model has additional errors that are only brought by the dropout. The challenge in this analysis is how to handle the distributions of dropout masks that appear when estimating the error of the hypergradient of the upper-level objective.
- We apply the proposed method to data cleaning problems. In our experiments, we observe the overfitting problem. In addition, we observe that the proposed dropout bilevel method redeems the overfitting problem. We also observe that the dropout bilevel method converges. Furthermore, we observed that accuracy changes across iterations are more stable when using dropout.

1.1 Related work

Theoretical properties of dropout methods for single-level optimization problems have been investigated in (Baldi & Sadowski, 2013; Gal & Ghahramani, 2016; Zhai & Wang, 2018; Li et al., 2016). In particular, Baldi & Sadowski (2013) introduced a general formalism for studying dropout and use it to analyze the averaging and regularization properties of dropout. Gal & Ghahramani (2016) and (Gal & Ghahramani, 2016) formed dropout training as approximate Bayesian inference in deep Gaussian processes and studied the model uncertainty. (Zhai & Wang, 2018; Gao & Zhou, 2016; Wang et al., 2019) studied Rademacher complexity bound of the dropout method.

Li et al. (2016); Mianjy & Arora (2020); Senen-Cerda & Sanders (2020); Senen-Cerda & Sanders (2022) investigated the convergence of the dropout method for single level risk minimization tasks. Li et al. (2016) view dropout as a data-dependent regularizer added to the training loss. In contrast, our approach involves forming a stochastic minimization model that treats dropout masks as a random linear transformation within the training loss. Mianjy & Arora (2020) considered a dropout method for a 2-layers network. When the loss function is logistic loss (convex) and the activation function is Relu, they provided the convergence rate of testing error. In this work, we consider general multi-layer neural networks and investigate training error. In addition, they assume data distribution is separable by a margin in a particular Reproducing Kernel Hilbert space, while we do not have assumptions on the data distribution. (Senen-Cerda & Sanders, 2020) is closely related to our work. They studies the convergence properties of dropout stochastic gradient methods for minimizing the empirical risks of multiple fully connected networks. Different from (Li et al., 2016; Mianjy & Arora, 2020; Senen-Cerda & Sanders, 2020; Senen-Cerda & Sanders, 2022), we focus on bilevel training tasks.

Popular methods to solve bilevel optimization problems in machine learning have been proposed in (Franceschi et al., 2017; 2018; Finn et al., 2017a; Li et al., 2022; Gould et al., 2016; Lorraine et al., 2020a; Bae & Grosse, 2020). The major of them are gradient-based methods. They can be further divided into two types based on the way they approximate the hypergradient. The first type is iterative differentiation (ITD) (Franceschi et al., 2017; 2018; Finn et al., 2017a; Liu et al., 2020; Ghadimi & Wang, 2018; Ji et al., 2021; Rajeswaran et al., 2019), and the second type is approximate implicit differentiation (AID) (Chen et al., 2022; Ji et al., 2021; Li et al., 2022; Gould et al., 2016; Lorraine et al., 2020a). As far as we know, there is no existing work in bilevel optimization that considers the overfitting problem, let alone analyzing the dropout bilevel method for this problem. In this work, we select an ITD method with a relatively simple structure as an example to investigate how dropout affects bilevel training tasks.

2 Preliminaries

In this paper, we denote $\mathbb{R}^n$ the n-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and Euclidean norm $\| \cdot \|$. We denote the spectrum norm of a matrix $A \in \mathbb{R}^{n \times m}$ as $\|A\|$ and the Frobenius norm of A as $\|A\|_F$. For any matrices A and B, we denote $\mathrm{trace}(A^T B) := \langle A, B \rangle$. For

a ramdom variable ξ defined on a probability space (Ξ, Σ, P), we denote its expectation as $\mathbb{E}\xi$. Given an event A and a function f of ξ, the conditional expectation of ξ is denoted as $\mathbb{E}_{\xi|A} f(\xi)$. For a function $F : \mathbb{R}^n \to \mathbb{R}^m$, we denote the function $F(x, y)$ with respect to y for a fixed x as $F(x, \cdot)$ and denote the function $F(x, y)$ with respect to x for a fixed y as $F(\cdot, y)$. For a differentiable function f, we say x is an ϵ stationary point of f if $\|\nabla f(x)\| \leq \epsilon$. For a twice differential function $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, we denote $\nabla_x F(x, y)$ and $\nabla_y F(x, y)$ as the partial gradients $\frac{\partial F(x,y)}{\partial y}$ and $\frac{\partial F(x,y)}{\partial y}$ correspondingly. We denote $\nabla^2 F(x, y)$ as the Hessian matrix of F, $\nabla_{xy} F(x, y) := \frac{\partial^2 F(x,y)}{\partial x \partial y}$ and $\nabla_{yy} F(x, y) := \frac{\partial^2 F(x,y)}{\partial y \partial y}$. For a multiple valued differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$, we denote its Jacobian at x as $J(f(x))$. With a little abuse of notation, given a distribution P, let $g(x; \xi)$ be a function that depends on $x \in \mathbb{R}^n$ and a data point $\xi \sim P$.

In general, a bilevel optimization training task is formed as follows:

$$\min_{\lambda \in \mathbb{R}^{n_\lambda}} F(\lambda, w(\lambda)) := F(\lambda, w(\lambda); \mathcal{D}_{val}), \text{ s.t. } w(\lambda) \in \arg\min_{w \in \mathbb{R}^{n_w}} G(\lambda, w) := G(\lambda, w; \mathcal{D}_{tr}), \tag{3}$$

where $\mathcal{D}_{val}$ is a validation data set, $\mathcal{D}_{tr}$ is a training data set, $F : \mathbb{R}^{n_\lambda} \to \mathbb{R}$ and $G : \mathbb{R}^{n_\lambda + n_w} \to \mathbb{R}$ are differentiable functions[1].

As mentioned in Section 2, there are various approaches to solve (3). Due to the nested structure of bilevel problems, the methods for (3) are inherently complex. To better understand how dropout is implemented in methods for (3), we use the fully single loop algorithm (FSLA) proposed in (Li et al., 2022) as an example to investigate dropout. This method has relatively simple formulas at each iteration and low computational cost per iteration. Step 6 in FSLA is one step of SGD for the lower level problem and Step 9 generates an approximation of the hypergradient of the objective in the upper level.

2.1 Dropout

Let $f(w; \xi)$ be an l-layer network, where w is the weight and ξ is an input data. In particular, we assume

$$f(w; \xi) := f(w; \xi) = a_l(w_l(a_{l-1}(w_{l-1} \cdots a_1(w_1 \xi)))), \tag{4}$$

where $w_i \in \mathbb{R}^{n_i} \times \mathbb{R}^{o_i}$ is the parameters in the ith layer[2], a_i is the activation function in the ith layer. We let $w = ((w_1)_1, \ldots, (w_1)_{n_1}, \ldots, (w_l)_1, \ldots, (w_l)_{n_l})$ be the collection of all rows in all weight matrices, where $(w_i)_s$ is the transpose of the jth row of w_i. Thus w is an no-dimensional vector, where $n = \sum n_i$, $o = \sum o_i$. Let l be a loss function. When training f with loss function l, the forward pass is

$$F(w; \xi) = l(f(w; \xi); \xi). \tag{5}$$

In the ith layer, the forward pass of an input $\xi_i \in \mathbb{R}^{n_i}$ using dropout can be formed as

$$r_i = m_i \circ a_i(w_i \xi_i), \tag{6}$$

where $\xi_0 = \xi$, $\circ$ represent the Hadamard product of two vectors, and $m_i \in \mathbb{R}^{n_i}$ with $(m_i)_j \sim Bernoulli(p_i)$ with $p_i \in [0, 1]$. m_j is called a dropout mask at the ith layer. Note that this formation of dropout masks is more general than the original dropout masks proposed in Srivastava et al. (2014). In Srivastava et al. (2014), the dropout rate is the same for all rows of weights in the same layer. Here, we allow the drop rate of each row of the weight matrix to be different.

Note that the function values of many well known activation functions at 0 is 0. For example, tanh, centered sigmoid and Relu has this property. Through out this work, we make the following assumption.

Assumption 1. *For any a_i used in* (4)*, $a_i(0) = 0$.*

Under assumption 1, r_i in (6) can be alternatively formed as

$$r_i = a_i(m_i \circ (w_i \xi_i)) = a_i \left(\begin{bmatrix} (m_i)_1 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & (m_i)_{n_i} \end{bmatrix} w_i \xi_i \right). \tag{7}$$

[1]With a little abuse of notation, without misunderstanding, we denote the $F(\lambda, w(\lambda); \mathcal{D}_{val})$ as a function $F(\lambda)$ and denote $G(\lambda, w; \mathcal{D}_{tr})$ as a function $G(\lambda, w)$.

[2]For simplicity, we include the parameters of bias in w_i with a corresponding fixed input of 1.

Denote $diag((m_i)_1, \ldots, (m_i)_{n_i}) := \begin{bmatrix} (m_i)_1 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & (m_i)_{n_i} \end{bmatrix}$. Recalling that for any layer i, $(m_i)_j \sim Bernoulli(p_i)$. The right hand formula in (7) means the dropout method randomly pick the jth row of the weight matrix w_i with probability p_i and sets the rest to 0. Thus, given a loss function l, under Assumption 1, the forward pass with dropout in fact calculates

$$l(f(mw;\xi);\xi) = F(mw;\xi),$$

where F is defined in (5) and

$$m := diag\left((m_1)_1, \ldots, (m_1)_{n_1}, \ldots, (m_l)_1, \ldots, (m_l)_{n_l}\right). \tag{8}$$

As we can see, the jth diagonal element $m_{jj} \sim Bernoulli(p_i)$ when $j \in \{\sum_k^{i-1} n_k + 1, \ldots, \sum_k^i n_k\}$. Note that in above discussion, each row of the weight matrix in each layer is always viewed as one element. For notation simplicity, without loss of generality, in the rest of this paper, when we mention the weight of a network, we view the weight matrix in the ith layer as a vector of $\mathbb{R}^{n_i}$, where n_i is the number of outputs in the ith layer. Then, we view w, the collection of all rows of all layers, as a vector in $\mathbb{R}^n$.

3 Dropout in Bilevel Problems

As introduced in Section 2.1, the objectives are the composition of them with the dropout masks. Thus, we propose the following variant of (3) that considers the distribution of the dropout masks:

$$\begin{aligned} &\min_{\lambda \in \mathbb{R}^{n_\lambda}} \tilde{F}(\lambda) := \mathbb{E}_{m_\lambda, m_w} F(m_\lambda \lambda, m_w w(\lambda)), \\ &\text{s.t. } w(\lambda) \in \arg\min_{w \in \mathbb{R}^{n_w}} \tilde{G}(\lambda, w) := \mathbb{E}_{m_\lambda, m_w} G(m_\lambda \lambda, m_w w), \end{aligned} \tag{9}$$

where F and G are the same as in (3), m_λ and m_w are random diagonal matrices with $(m_\lambda)_{ii} \sim Bernoulli(p_{i,\lambda})$, $(m_w)_{ii} \sim Bernoulli(p_{i,w})$ respectively. This idea of modeling with the dropout masks is also considered in analyzing dropout SGD for the single level training tasks in (Senen-Cerda & Sanders, 2020). For bilevel learning tasks, this model is new. We add dropout masks on both upper and lower level objective functions to include the application where λ and w are both weights of a network, Franceschi et al. (2018); Zügner & Günnemann (2019); Finn et al. (2017a); Snell et al. (2017).

Now we adopt the dropout bilevel method. based on FSLA. Before presenting a general form of FSLA with dropout, let's consider applying it to the example mentioned in (1). The data cleaning problem (1) is a case of (3) with $F(\lambda, w(\lambda); \mathcal{D}_{val}) = \frac{1}{|N_{val}|} \sum_{i=1}^{N_{val}} l(f(w(\lambda); \eta_i); \eta_i)$ and $G(\lambda, w; \mathcal{D}_{tr}) = \frac{1}{|N_{tr}|} \sum_{i=1}^{N_{tr}} \sigma(\lambda_i) l(f(w; \xi_i); \xi)$. As we illustrate in Section 2.1, suppose m is a dropout masks added in the forward pass of the network used in (1) in each iteration. Then the forward pass of the upper level objective used in FSLA becomes $F(\lambda, mw(\lambda); \mathcal{D}_{val})$, where m is defined in (8). The forward pass of the lower level objective used in FSLA becomes $G(\lambda, mw; \mathcal{D}_{tr})$. By the chain rule, the backward pass needed in FSLA calculate the follows:

$$\begin{aligned} &\nabla F((\cdot), mw(\cdot); \mathcal{D}_{val})(\lambda) = m \nabla F((\cdot), mw(\cdot); \mathcal{D}_{val})(\lambda); \\ &\nabla G(\cdot, m(\cdot); \mathcal{D}_{tr})(\lambda, w) = m \nabla G(\lambda, mw; \mathcal{D}_{tr}); \\ &\nabla^2 G(\cdot, m(\cdot); \mathcal{D}_{tr})(\lambda, w) = m \nabla^2 G(\lambda, mw; \mathcal{D}_{tr}) m. \end{aligned} \tag{10}$$

Based on this, we present Algorithm 1. We add a projection in step 8 to avoid v blow up. This is also important in the theoretical analysis. In the next section, we analyze the convergence of Algorithm 1.

4 Analysis of Algorithm 1

A challenge of analyzing (9), which the single level model considered in (Senen-Cerda & Sanders, 2020) does not have, is how to analyze the hypergradient $\nabla \tilde{F}(\lambda)$. $\nabla \tilde{F}(\lambda)$ is hard to analyze due to

Algorithm 1 FSLA with dropout

Input: $\beta, \alpha, \gamma > 0$, λ^0, w^0, dropout rates $\{p_{i,\lambda}\}_{i=1}^{n_\lambda}$, $\{p_{i,w}\}_{i=1}^{n_w}$, $r > 0$, $\nabla \bar{F}(\lambda^0, w^0, v^{-1})))$.
for $k = 0, \ldots, T$. **do**
  $\alpha_k = \frac{\delta}{\sqrt{k}}$; $\gamma_k = \gamma\alpha_k$; $\beta_k = \beta\alpha_k$; $\eta_k = \eta\alpha_k$.
  Generate random diagonal matrices $\overline{m}_\lambda^k$, $\overline{m}_w^k$, $\underline{m}_\lambda^k$ and $\underline{m}_w^k$ with $(\overline{m}_\lambda^k)_{ii}, (\underline{m}_\lambda^k)_{ii} \sim Bernoulli(p_{i,\lambda})$, $(\overline{m}_w^k)_{ii}, (\underline{m}_w^k)_{ii} \sim Bernoulli(p_{i,w})$.
  Sample $\xi^k \in \mathcal{D}_{tr}$, $\eta^k \in \mathcal{D}_{val}$.
  $w^{k+1} = w^k - \gamma_k \underline{m}_w^k \nabla_w G(\underline{m}_\lambda^k \lambda^k, \underline{m}_w^k w^k; \xi^k)$.
  $\tilde{v}^{k+1} = \beta \overline{m}_w^k \nabla_w F(\overline{m}_\lambda^k \lambda^k, \overline{m}_w^k w^{k+1}; \eta^k) + \left(I - \beta_k \underline{m}_w^k \nabla_{ww}^2 G(\underline{m}_\lambda^k \lambda^k, \underline{m}_w^k w^k; \xi^k) \underline{m}_w^k\right) v^k$.
  $v^{k+1} = \mathrm{Proj}_{B(0,r)} \tilde{v}^{k+1}$.
  $\nabla \bar{F}(\lambda^k, w^{k+1}, v^{k+1}) = \overline{m}_\lambda^k \nabla_\lambda F(\overline{m}_\lambda^k \lambda^k, \overline{m}_w w^{k+1}; \eta^k) - \underline{m}_{\lambda^k} \nabla_{\lambda w}^2 G(\underline{m}_\lambda^k \lambda^k, \underline{m}_w^k w^{k+1}; \xi^k) \underline{m}_w^k v^{k+1}$.
  $d^{k+1} = \nabla \bar{F}(\lambda^k, w^{k+1}, v^{k+1}) + (1 - \eta_k)(d^k - \nabla \bar{F}(\lambda^{k-1}, w^k, v^k)))$.
  $\lambda^{k+1} = \lambda^k - \alpha_k d^{k+1}$.
end for

its relation with the solution of the lower level problem and the fact that both upper and lower level objectives are composed with random dropout masks. To analyze this and Algorithm 1 for (9), we first make the following assumptions that are standard in bilevel optimization literature.

Assumption 2. *Let $F(\lambda, w)$ be defined as in* (3)*. Suppose the following assumptions hold:*

(i) Suppose F is Lipschitz continuous with modulus L_F.

(ii) For any fixed λ, $\nabla_\lambda F(\lambda, \cdot)$ and $\nabla_w F(w, \cdot)$ are Lipschitz continuous with modulus L_{12}^F and L_{22}^F.

(iii) There exists $C^F > 0$ such that $\max\{\|\nabla_w F(\lambda, w)\|, \|\nabla_\lambda F(\lambda, w)\|\} \leq C^F$ for any λ and w.

(iv) For any fixed w, $\nabla_w F(\cdot, w)$ is Lipschitz continuous with modulus $L_{w\lambda}^F > 0$.

Assumption 3. *Let $G(\lambda, w)$ be defined as in* (3)*. Suppose the following assumptions hold:*

(i) G is twice continuously differentiable. $\nabla G(\lambda, w)$ is Lipschitz continuous with modulus L_G.

(ii) For any λ, $G(\lambda, \cdot)$ is strongly convex with modulus μ.

(iii) For any λ, $\nabla_{ij}^2 G(\lambda, \cdot)$ is Lipschitz continuous with modulus L_{G^w}.

(iv) For any w, $\nabla_{i,j}^2 G(\cdot, w)$ is Lipschitz continuous with modulus L_{G^λ}.

(v) There exists C_G such that $\max_{i,j}\{\|\nabla_{ij}^2 G(\lambda, w)\|, \|\nabla_{ij}^2 G(\lambda, w)\|\} \leq C_G$ for any λ and w.

Denote $M = \begin{bmatrix} m_\lambda & 0 \\ 0 & m_w \end{bmatrix}$ and $F_M(\lambda) := F(m_\lambda(\cdot), m_w w(\cdot))(\lambda)$. Then

$$\nabla \tilde{F}(\lambda) = \mathbb{E}_{m_\lambda, m_w} \nabla F_M(\lambda). \tag{11}$$

Under Assumptions 2 and 3, it is easy to see that using the chain rule,

$$\nabla F_M(\lambda) = m_\lambda \nabla_\lambda F(m_\lambda \lambda, m_w w(\lambda)) + J(w(\lambda))^T m_w \nabla_w F(m_\lambda \lambda, m_w w(\lambda)), \tag{12}$$

where $J(w(\lambda))$ is the Jacobian of $w(\lambda)$. The next proposition gives a closed form of $J(w(\lambda))$ and its property.

Proposition 1. *Consider* (9) *and suppose Assumption 3 holds. Denote $W = \begin{bmatrix} \lambda \\ w \end{bmatrix}$ and $M = \begin{bmatrix} m_\lambda & 0 \\ 0 & m_w \end{bmatrix}$. Let $\underline{p}_w = \min_i p_{i,w}$. Then $\tilde{G}$ is strongly convex with modulus $\underline{p}_w \mu$ and the following equalities hold:*

$$\nabla \tilde{G}(\lambda, w) = \mathbb{E} M \nabla G\left(MW\right); \ \nabla^2 \tilde{G}(\lambda, w) = \mathbb{E} M \nabla^2 G\left(MW\right) M;$$

$$J(w(\lambda)) = -\left(\nabla^2_{ww}\tilde{G}(\lambda, w(\lambda))\right)^{-1} \nabla^2_{\lambda w}\tilde{G}(\lambda, w(\lambda)). \tag{13}$$

In addition, for any λ, it holds that $w(\lambda)$ is Lipschitz continuous with some modulus L_w. Also, for any w, it holds that $\|J(\lambda(w))\|_F \leq \frac{C_G}{\mu}$.

Based on Proposition 1, we can estimate the variance of $\nabla\tilde{G}(\lambda, w)$.

Lemma 1. *Let G be defined as in* (3)*. Let m_λ and m_w are diagonal matrices with $(m_\lambda)_{ii} \sim Bernoulli(p_{\lambda,i})$ and $(m_w)_{jj} \sim Bernoulli(p_{w,j})$. Let $p_i = p_{i,\lambda}$ if $i \in \{1, \ldots, n_\lambda\}$ and $p_i = p_{i,w}$ if $i \in \{n_\lambda + 1, \ldots, n_\lambda + n_w\}$. Then*

$$\mathbb{E}\|\nabla^2 G\left(M(\cdot)\right)(W) - \mathbb{E}_M \nabla^2 G\left(M(\cdot)\right)(W)\|^2 \leq \sum_{i,j} p_i p_j \left(1 - p_i p_j\right) C_G^2. \tag{14}$$

Next, we show how the update of the lower level parameters behaves. Since Algorithm 1 uses the stochastic gradient and Hessian. We add the following assumptions.

Assumption 4. *Consider* (9)*. Let ξ be randomly picked from $\mathcal{D}_{tr}$ and η be randomly picked from $\mathcal{D}_{val}$. Suppose $\mathbb{E}_\eta \nabla_\lambda F(\lambda, w; \eta) = \nabla_\lambda F(\lambda, w)$ and $\mathbb{E}_\xi \nabla_w G(\lambda, w; \xi) = \nabla_w G(\lambda, w)$. Suppose $\mathbb{E}_\eta\|\nabla_\lambda F(\lambda, w; \eta) - \nabla_\lambda F(\lambda, w)\|^2 \leq \sigma^2$, $\mathbb{E}_\xi\|\nabla_w G(\lambda, w; \xi) - \nabla_w G(\lambda, w)\|^2 \leq \sigma^2$, and for any i, j, $\mathbb{E}_\xi\|\nabla^2_{i,j} G(\lambda, w; \xi) - \nabla^2_{i,j} G(\lambda, w)\| \leq \sigma_h^2$.*

Now we show the inductive relations of $\{w^k\}$ generated by Algorithm 1.

Lemma 2. *Consider* (9)*. Suppose Assumptions 2, 3 and 4 hold. Let $\{(\lambda^k, w^k)\}$ be generated by Algorithm 1. Let γ in Algorithm 1 be small enough such that $\gamma\mu < 1$. Let $\overline{p}_w := \max_i p_{i,w}$. Then there exists $\iota \in (0, 1)$ such that*

$$\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 \leq \iota\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2 + O(\alpha_k^2)\mathbb{E}\|d^{k-1}\|^2 + O(\gamma_k^2)\overline{p}_w\sigma^2. \tag{15}$$

Remark 1. *Note that $w(\lambda^k)$ is the true solution of the lower level problem given λ^k. This induction has additional errors related to the updates of the upper level variable λ^k and the variance. Note that the variance term in* (26) *decreases linearly with the maximum dropout rate.*

Next, we present the inductive relations of $\{v^k\}$ generated by Algorithm 1. In fact, combining (13) with (12), it is easy to see that

$$\nabla F_M(\lambda) = m_\lambda \nabla_\lambda F(m_\lambda\lambda, m_w w(\lambda)) - \nabla^2_{\lambda w}\tilde{G}(\lambda, w(\lambda)) v_{m_w,\lambda}. \tag{16}$$

where $v_{m_w,\lambda} := \left(\nabla^2_{ww}\tilde{G}(\lambda, w(\lambda))\right)^{-1} m_w \nabla_w F(m_\lambda\lambda, m_w w(\lambda))$. This implies that

$$v_{m_w,\lambda} = \beta m_w \nabla_w F(m_\lambda\lambda, m_w w(\lambda)) + \left(I - \beta\nabla^2_{ww}\tilde{G}(\lambda, w(\lambda))\right) v_{m_w,\lambda}.$$

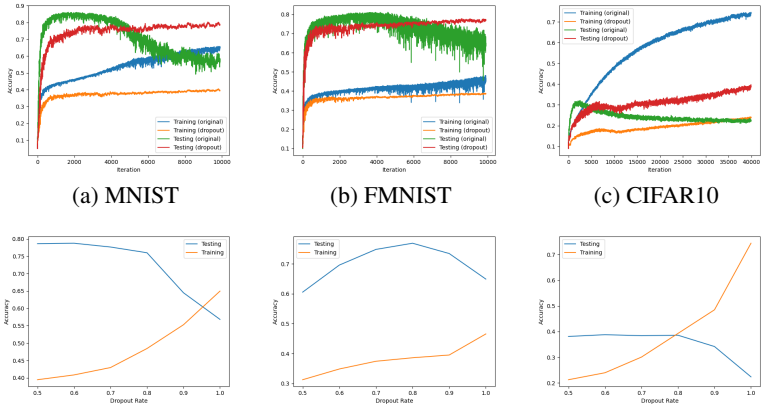
Comparing the above equation with the update of v^{k+1} in Algorithm 1, we see that v^{k+1} is an approximation of $v_{\overline{m}_w^k,\lambda^k}$. The next theorem shows how the difference v^k and $v_{\overline{m}_w^k,\lambda^k}$ varies with iteration k and the dropout rate.

Theorem 1. *Consider* (9)*. Suppose Assumptions 2, 3 and 4 hold. Let $\{(\lambda^k, w^k)\}$ be generated by Algorithm 1. Denote $V^k := v^k - v_{\overline{m}_w^{k-1},\lambda^{k-1}}$. Denote $\overline{p}_\lambda := \max_i p_{i,\lambda}$, $\underline{p}_\lambda := \min_i p_{i,\lambda}$, $\overline{p}_w := \max_i p_{i,w}$ and $\underline{p}_w := \min_i p_{i,w}$. Suppose $r \geq C_v := \frac{L_{22}^F}{\mu}$. Then there exists $\varrho \in (0, 1)$ such that*

$$\begin{aligned}\mathbb{E}\|V^{k+1}\|^2 &\leq \varrho\mathbb{E}\|V^k\|^2 + O((\overline{p}_w)^2\beta_k^2)\sigma^2 + O(\overline{p}_w\beta_k^2)\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 \\ &+ O(\alpha_k^2)\mathbb{E}\|d^{k-1}\|^2 + O((\overline{p}_w)^2\beta_k^2)\sigma_h^2 + O(\bar{p}_\lambda\bar{p}_w(1 - \underline{p}_\lambda\underline{p}_w)\beta_k^2).\end{aligned}$$

Remark 2. *The last term in the above theorem are introduced by the dropout of the network and disappear when no dropout is applied to the network parametrized by w.*

Now, based on Theorem 1 and Lemma 2, we analyze the convergence of Algorithm 1. Thanks to (11), to find an ϵ stationary point to (9) such that $\|\nabla\tilde{F}(\lambda)\|^2 \leq \epsilon$, it suffices to find λ that satisfies $\mathbb{E}\|\nabla F_M(\lambda)\|^2 \leq \epsilon$.



(a) MNIST (b) FMNIST (c) CIFAR10

(d) MNIST (e) FMNIST (f) CIFAR10

Figure 2: Results from data cleaning. The first line reports how training and testing accuracy change as the number of iterations increases. The second line details the training and testing accuracy in the final iteration when different dropout rates are applied.

Theorem 2. *Consider* (9). *Suppose Assumptions 2, 3 and 4 hold. Let* $\{(w^k, \lambda^k)\}$ *be generated by Algorithm 1. Suppose* F *is bounded below by* $\bar{F}$. *Then there exist small* δ *in Algorithm 1 such that*

$$\frac{1}{T-1}\sum_{k=1}^{T-1}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \leq \frac{1}{\sqrt{T}}\frac{1}{\underline{p}_w\underline{p}_\lambda}I_0 + \frac{\ln(T+1)}{\sqrt{T}}O\left(\frac{\bar{p}_\lambda^2}{\underline{p}_\lambda\underline{p}_w} + \underline{p}_w\bar{p}_w\right)(\sigma^2+\sigma_h^2)$$
$$+\frac{\ln(T+1)}{\sqrt{T}}O\left(\underline{p}_w\left(\bar{p}_w^2(1-\underline{p}_w^2)+\bar{p}_w(1-\underline{p}_w)\right)\right).$$

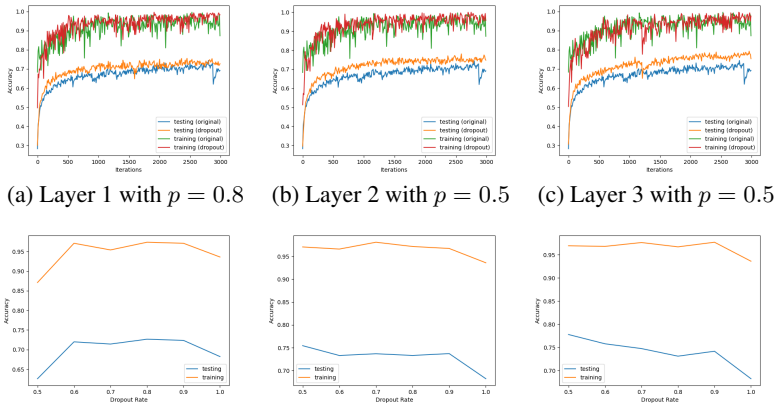
Remark 3. *The first term shows that the convergence rate is slower when* $\underline{p}_\lambda\underline{p}_w$ *is small, which can be confirmed in our experiments Figure 2. The last term in the above results is introduced by dropout, and it disappears when* $\underline{p}_\lambda = \underline{p}_w = 1$, *i.e., no dropout is applied.*

On the other hand, we have to point that the convergence of $\mathbb{E}\nabla F_{M^k}(\lambda^k)$ *does not imply that of* $F(\lambda)$. *Consider a simple 2-dimensional case. Let* $f(x_1,x_2) := \frac{1}{2}x_1^2x_2 + 1000x_1$. *Let* $x^k = (x_1^k, x_2^k)$ *be with* $x_1^k = -\frac{1000p^2+2000(1-p)p}{k}$ *and* $x_2^k = k$. *Let* $m = \mathrm{diag}\{m_1, m_2\}$ *with* $m_1 \approx Bernoulli(p)$ *and* $m_2 \approx Bernoulli(p)$. *Then* $\lim_{k\to\infty}\mathbb{E}_m\nabla f(mx^k) = 0$. *However,* $\lim_{k\to\infty}\nabla f(x^k) = (-1000p^2 - 2000(1-p)p + 1000, 0)$. *Therefore, the convergence of* $\nabla\mathbb{E}_m f(mx^k)$ *does not necessarily imply the convergence of* $\nabla f(x^k)$, *the error between* $\nabla f(x^k)$ *and* $\nabla\mathbb{E}_m f(mx^k)$ *can not be closed unless* p *approaches* 1. *Therefore, there are additional errors brought only by the dropout. This example implies that the dropout method may not be optimizing the original bilevel problem. This is expected because when the original training loss is optimized, the model fits the training data too well and this will increase the chance of overfitting.*

5 Experiments

In this section, we test Algorithm 1 on data cleaning tasks (1) and meta learning problem (2). The experiments were conducted on a machine equipped with an Intel Xeon E5-2683 CPU and 4 Nvidia Tesla P40 GPUs.

Data cleaning We followed the approach outlined in (Srivastava et al., 2014) and utilized a fully connected network. The experiments were performed using the MNIST dataset (LeCun et al., 2010). The network architecture employed was 784-1024-1024-2048-10, with ReLU activation functions used for all hidden units. To introduce unclean data, we randomly mislabel 60% of the training data. When training on MNIST and FMNIST, we use a fully connected network with the architecture $784-1024-1024-2048-10$ and employ ReLU activation functions for all hidden units. In all experiments conducted on MNIST and FMNIST, we set the dropout rates of all layers to the same value, denoted as p. Dropout is only applied when updating the lower level parameters. We set $\gamma = 0.01$ and train $10000/10000$ iterations for $5000/10000$ MNIST/FMNIST data points. When



(a) Layer 1 with $p = 0.8$ (b) Layer 2 with $p = 0.5$ (c) Layer 3 with $p = 0.5$

(d) Layer 1 (e) Layer 2 (f) Layer 3

Figure 3: Results from few-shot learning on the Omniglot dataset. The first line reports how training and testing accuracy change as the number of iterations increases. The second line details the training and testing accuracy in the final iteration when different dropout rates are applied to only one layer of the network.

training on CIFAR10, we use 4-layer convolutional neural networks + 1 fully connected layer. In all CIFAR10 experiments, we apply dropout after each convolutional layer, using a consistent dropout rate of p. We set $\gamma = 0.1$. We train 40000 iterations for 40000 data points.

Meta-learning We conduct experiments with the few-shot learning task, following the experimental protocols of (Vinyals et al., 2016), we performed learning tasks over the Omniglot dataset. We set train/validation/test with 102/172/423, respectively. We perform 5-way-1-shot classification. More specifically, we perform 5 training tasks $(N = 5)$. For each task, we randomly sample 5 characters from the alphabet over that client and for each character, and select 1 data points for training and 15 samples for validation. We use a 4-layer convolutional neural network + 1 fully connected layer. Each convolutional layer has 64 filters of 3×3 and is followed by batch-normalization layers (Finn et al., 2017b). The parameters of convolutional layers are shared between different tasks (λ) and the last linear layer is the task-specific parameter w_j. In each experiment, we only add dropout to one CNN layer. In all experiments, we let β, α and γ in FSLA and Algorithm 1 be 0.05, 0.1 and 0.8 respectively.

Results We report the results in Figures 2 and 3. In the first line of both figures, we plot the accuracy against the iteration in each training progress. In the second line of both figures, we plot how different dropout rates affect the training and testing accuracy. Figure 2 shows when the training accuracy increases, the testing accuracy decreases. This observation indicates the occurrence of overfitting in data cleaning tasks. The proper addition of dropout during training can enhance testing accuracy in response. On the other hand, Figure 2 demonstrates that training the network with dropout leads to convergence at a higher testing accuracy and greater stability. Figure 3 shows the accuracy of the method when adding dropout with different rates on the different layers. As we can see, adding a proper dropout to any layer improves the testing accuracy.

6 Conclusion

In this paper, we explore the application of dropout in bilevel training tasks. We propose a stochastic bilevel model that is dependent on the distribution of dropout masks. We adapt an existing bilevel method with dropout. We analyze the convergence properties of the resulting method for the proposed model. We investigate the inductive relations of the variables attributes to an approximation of the hypergradients. In addition, we show how the dropout rates affect the complexity. We believe that other state-of-art bilevel methods can also be adapted to address the stochastic bilevel model with random dropout masks, and our convergence analysis serve as the first example for such adaptations.

REFERENCES

Jimmy Ba and Brendan Frey. Adaptive dropout for training deep neural networks. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/7b5b23f4aadf9513306bcd59afb6e4c9-Paper.pdf.

Juhan Bae and Roger B. Grosse. Delta-stn: Efficient bilevel optimization for neural networks using structured response jacobians. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f754186469a933256d7d64095e963594-Abstract.html.

Pierre Baldi and Peter J. Sadowski. Understanding dropout. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2814–2822, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/71f6278d140af599e06ad9bf1ba03cb0-Abstract.html.

Ziyi Chen, Bhavya Kailkhura, and Yi Zhou. A fast and convergent proximal algorithm for regularized nonconvex and nonsmooth bi-level optimization. 2022. URL https://doi.org/10.48550/arXiv.2203.16615.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August*, 2017a.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August*, 2017b.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August*, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1563–1572. PMLR, 2018. URL http://proceedings.mlr.press/v80/franceschi18a.html.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 1019–1027, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/076a0c97d09cf1a0ec3e19c7f2529f2b-Abstract.html.

Wei Gao and Zhi-Hua Zhou. Dropout rademacher complexity of deep neural networks. *Sci. China Inf. Sci.*, 59(7):072104:1–072104:12, 2016. doi: 10.1007/s11432-015-5470-z. URL https://doi.org/10.1007/s11432-015-5470-z.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. 2018. URL https://arxiv.org/abs/1802.02246.

Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. 2016. URL http://arxiv.org/abs/1607.05447.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July*, 2021.

Alex Labach, Hojjat Salehinejad, and Shahrokh Valaee. Survey of dropout methods for deep neural networks. *CoRR*, abs/1904.13310, 2019. URL `http://arxiv.org/abs/1904.13310`.

Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist,2.`

Junyi Li, Bin Gu, and Heng Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1*, 2022.

Zhe Li, Boqing Gong, and Tianbao Yang. Improved dropout for shallow and deep learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2531–2539, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July*, 2020.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August, Online [Palermo, Sicily, Italy]*, 2020a.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1540–1552. PMLR, 2020b. URL `http://proceedings.mlr.press/v108/lorraine20a.html`.

Poorya Mianjy and Raman Arora. On convergence and generalization of dropout training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21151–21161. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/f1de5100906f31712aaa5166689bfdf4-Paper.pdf`.

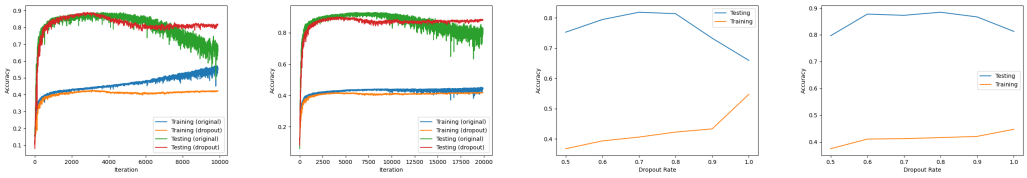
Sungheon Park and Nojun Kwak. Analysis on the dropout effect in convolutional neural networks. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (eds.), *Computer Vision – ACCV 2016*, pp. 189–204, Cham, 2017. Springer International Publishing. ISBN 978-3-319-54184-6.

Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. Dropout improves recurrent neural networks for handwriting recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pp. 285–290, 2014. doi: 10.1109/ICFHR.2014.55.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, Vancouver, BC, Canada*, 2019.

Albert Senen-Cerda and Jaron Sanders. Almost sure convergence of dropout algorithms for neural networks. *CoRR*, abs/2002.02247, 2020. URL `https://arxiv.org/abs/2002.02247`.

Albert Senen-Cerda and Jaron Sanders. Asymptotic convergence rate of dropout on shallow linear neural networks. *SIGMETRICS Perform. Eval. Rev.*, 50(1):105–106, jul 2022. ISSN 0163-5999. doi: 10.1145/3547353.3530965. URL `https://doi.org/10.1145/3547353.3530965`.



(a) 10000 training data (b) 40000 training data (c) 10000 training data (d) 40000 training data

Figure 4: MNIST

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April, Naha, Okinawa, Japan*, 2019.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, Long Beach, CA, USA*, 2017.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958, 2014.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, Barcelona, Spain*, 2016.

Haotian Wang, Wenjing Yang, Zhenyu Zhao, Tingjin Luo, Ji Wang, and Yuhua Tang. Rademacher dropout: An adaptive dropout for deep neural network via optimizing generalization gap. *Neurocomputing*, 357:177–187, 2019. doi: 10.1016/j.neucom.2019.05.008. URL https://doi.org/10.1016/j.neucom.2019.05.008.

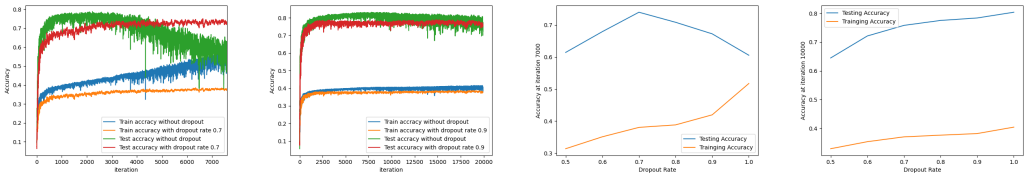
Haibing Wu and Xiaodong Gu. Towards dropout training for convolutional neural networks. *Neural Networks*, 71:1–10, 2015. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2015.07.007. URL https://www.sciencedirect.com/science/article/pii/S0893608015001446.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/9713faa264b94e2bf346a1bb52587fd8-Paper.pdf.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *CoRR*, abs/1409.2329, 2014. URL http://arxiv.org/abs/1409.2329.

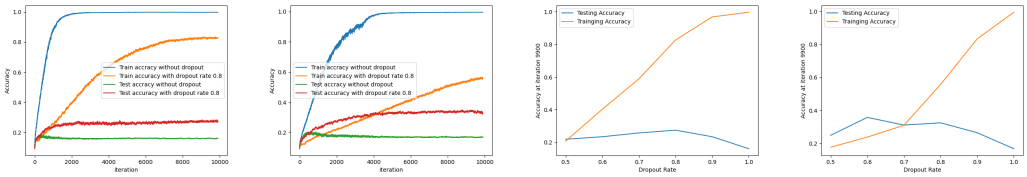
Ke Zhai and Huan Wang. Adaptive dropout with rademacher complexity regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1uxsye0Z.

Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9*, 2019.



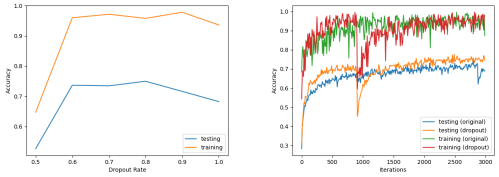
(a) 5000 training data (b) 40000 training data (c) 5000 training data (d) 40000 training data

Figure 5: FMNIST



(a) 5000 training data (b) 10000 training data (c) 5000 training data (d) 10000 training data

Figure 6: CIFAR10



(a) Layer 4 (b) Layer 4 with $p = 0.2$

Figure 7: Mata-learning

A Additional Experiments

A.1 Data Cleaning

We use the same network and hyperparameters of the FLSA with dropout as in the full paper. We train $10000/20000$ iterations for 10000/40000 MNIST. We train $10000/20000$ iterations for 5000/40000 for FMNIST. The results are in Figure 4 and 5.

Figure 2d 4c 4d indicates that as the dataset size increases, the lines representing testing accuracy and training accuracy become more parallel. This implies that in larger datasets, overfitting is less likely to happen.

In data learning rasks on CIFAR10, we let each convolutional layer be defined by its input channel, output channel, kernel size, stride, and padding, set as [64, 3, 3, 3, 2, 0], [64, 64, 3, 3, 2, 0], [64, 64, 3, 3, 2, 0], and [4, 64, 2, 2, 1, 0], respectively. The linear layer connects an input of size 64*4 to an output size of 10. Throughout the network, ReLU activation functions are employed for all hidden units. Additionally, we train $10000/10000$ iterations for 5000/10000 data points. The results are in 6.

A.2 Meta-learning

We additionally test how dropout affects the last convolution layer, see Figure 7.

B PROOFS OF PROPOSITION 1

We restate a detailed version of the property here.

Proposition 2. *Consider* (9) *and suppose Assumption 3 holds. Denote* $W = \begin{bmatrix}\lambda \\ w\end{bmatrix}$ *and* $M = \begin{bmatrix} m_\lambda & 0 \\ 0 & m_w \end{bmatrix}$. *Let* $\underline{p}_w = \min_i p_{i,w}$. *Then* $\tilde{G}$ *is strongly convex with modulus* $\underline{p}_w \mu$ *and*

$$\begin{aligned} \nabla \tilde{G}(\lambda, w) &= \sum_{m_w \in M_w, m_\lambda \in M_\lambda} p_{m_w, m_\lambda} M \nabla G\left(MW\right) = \mathbb{E} M \nabla G\left(MW\right); \\ \nabla^2 \tilde{G}(\lambda, w) &= \mathbb{E} M \nabla^2 G\left(MW\right) M. \end{aligned} \tag{17}$$

Also $w(\lambda)$ *is Lipschitz continuous with constant* $L_w := \frac{C_g}{\underline{p}_w \mu}$. *In addition, for any* λ, *it holds that* $J(w(\lambda)) = -\left(\nabla^2_{ww} \tilde{G}(\lambda, w(\lambda))\right)^{-1} \nabla^2_{\lambda w} \tilde{G}(\lambda, w(\lambda))$ *is Lipschitz continuous with modulus* $L_J := \frac{L_{G^\lambda_\lambda} + L_{G^w_\lambda} L_w}{\underline{p}_w \mu} + \frac{(L_{G^\lambda_w} + L_{G^w_w} L_w) C_G}{\underline{p}_w^2 \mu^2}$. *Also, for any* w, *it holds that* $\|J(\lambda(w))\|_F \leq \frac{C_G}{\underline{p}_w \mu}$.

Proof. Note that

$$\tilde{G}(\lambda, w) = \sum_{m_w \in M_w, m_\lambda \in M_\lambda} p_{m_w, m_\lambda} G(m_\lambda \lambda, m_w w)$$

where P_{m_w, m_λ} is the probability of the case (m_w, m_λ) happends and (M_w, M_λ) is the set of all cases of m_w, m_λ. Thus $\tilde{G}(\lambda, w)$ is also twice continuously differentiable.

Note that $G(m_\lambda(\cdot), m_w(\cdot))(\lambda, m_w)$ can be equally written as $G\left(\begin{bmatrix} m_\lambda & 0 \\ 0 & m_w \end{bmatrix}(\cdot)\right)\left(\begin{bmatrix}\lambda \\ w\end{bmatrix}\right)$. Then,

$$\nabla \tilde{G}(\lambda, w) = \sum_{m_w \in M_w, m_\lambda \in M_\lambda} p_{m_w, m_\lambda} M \nabla G\left(MW\right) = \mathbb{E} M \nabla G\left(MW\right). \tag{18}$$

and

$$\nabla^2 \tilde{G}(\lambda, w) = \sum_{m_w \in M_w, m_\lambda \in M_\lambda} p_{m_w, m_\lambda} M \nabla^2 G\left(MW\right) M = \mathbb{E} M \nabla^2 G\left(MW\right) M. \tag{19}$$

These together with Assumption 3, the fact that $p_{m_w, m_\lambda} \in [0, 1]$ and the fact that all entries of the diagonal matrix M belongs to $\{0, 1\}$ shows that

- $\tilde{G}$ is twice continuously differentiable. In addition, suppose $\nabla_w \tilde{G}(\lambda, \cdot)$ is Lipschitz continuous with modulus $L_G > 0$ for any λ.
- For any λ, $\nabla^2_{\lambda w} \tilde{G}(\lambda, \cdot)$ and $\nabla^2_{ww} \tilde{G}(\lambda, \cdot)$ are Lipschitz continuous with modulus $L_{G^w_\lambda}$ and $L_{G^w_w}$.

Now we show that for any λ, $\tilde{G}(\lambda, \cdot)$ is strongly convex with modulus μ. Let $v \in \mathbb{R}^{n_w}$. Using (19), we have that

$$\begin{aligned} & v \nabla^2_{ww} \tilde{G}(\lambda, w) v = v \left(\mathbb{E} m_w \nabla^2 G\left(m_\lambda \lambda, m_w w\right) m_w\right) v \\ & = \mathbb{E} m_w v \nabla^2 G\left(m_\lambda \lambda, m_w w\right) m_w v \\ & \geq \mathbb{E} \mu \|m_w v\|^2 = \sum_{i=1}^{n_w} \mathbb{E}_{(m_w)_{ii}} \mu \|(m_w)_{ii} v_i\|^2 \\ & = \sum_{i=1}^{n_w} \mu p_{i,w} \|v_i\|^2 \geq \mu \underline{p}_w \|v\|^2, \end{aligned} \tag{20}$$

where the first inequality uses Assumption 3 and $\underline{p}_w = \min_i p_{i,w}$. Thus, using Lemma 2.1 and Lemma 2.2 in (Ghadimi & Wang, 2018), for any λ, it holds that

- $w(\lambda)$ is Lipschitz continuous with constant $\frac{C_g}{\underline{p}_w \mu}$.
- $$J(w(\lambda)) = -\left(\nabla^2_{ww}\tilde{G}(\lambda, w(\lambda))\right)^{-1} \nabla^2_{\lambda w}\tilde{G}(\lambda, w(\lambda)) \tag{21}$$

  is Lipschitz continuous with modulus $L_J := \frac{L_{G^\lambda_\lambda} + L_{G^w_\lambda} L_w}{\underline{p}_w \mu} + \frac{(L_{G^\lambda_w} + L_{G^w_w} L_w) C_G}{\underline{p}_w^2 \mu^2}$. Also, for any w, it holds that $\|J(w(\lambda)\|_F \leq \frac{C_G}{\underline{p}_w \mu}$.

□

C PROOFS OF LEMMA 1

We restate the lemma here.

Lemma 3. *Let G be defined as in* (3)*. Let m_λ and m_w are diagonal matrices with $(m_\lambda)_{ii} \sim Bernoulli(p_{\lambda,i})$ and $(m_w)_{jj} \sim Bernoulli(p_{w,j})$. Let $p_i = p_{i,\lambda}$ if $i \in \{1, \ldots, n_\lambda\}$ and $p_i = p_{i,w}$ if $i \in \{n_\lambda + 1, \ldots, n_\lambda + n_w\}$. Then*

$$\mathbb{E}\|\nabla^2 G\left(M(\cdot)\right)(W) - \mathbb{E}_M \nabla^2 G\left(M(\cdot)\right)(W)\|^2 \leq \sum_{i,j} p_i p_j \left(1 - p_i p_j\right) C_G^2. \tag{22}$$

Similarly, we have

$$\mathbb{E}\|\nabla_w F\left(M(\cdot)\right)(W) - \mathbb{E}_M \nabla_w F\left(M(\cdot)\right)(W)\|^2 \leq \sum_{i,j} p_i \left(1 - p_i\right) (C^F)^2. \tag{23}$$

Proof. Note that $G(m_\lambda(\cdot), m_w(\cdot))(\lambda, m_w)$ can be equally written as $G\left(\begin{bmatrix} m_\lambda & 0 \\ 0 & \underline{m}_w \end{bmatrix}(\cdot)\right)\left(\begin{bmatrix} \lambda \\ w \end{bmatrix}\right)$. Denote $W = \begin{bmatrix} \lambda \\ w \end{bmatrix}$ and $M = \begin{bmatrix} m_\lambda & 0 \\ 0 & \underline{m}_w \end{bmatrix}$. Then,

$$\nabla G\left(M(\cdot)\right)(W) = M \nabla G\left(MW\right)$$

and

$$\begin{aligned} &\nabla^2 G\left(M(\cdot)\right)(W) = M \nabla^2 G\left(MW\right) M \\ &= \sum_{i,j} diag(0, \ldots, 0, M_{ii}, 0, \ldots, 0) H_G\left(MW\right) diag(0, \ldots, 0, M_{jj}, 0, \ldots, 0) \\ &= \sum_{i,j} \begin{bmatrix} 0 & 0 & \ldots & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & M_{ii} \frac{\partial^2 G}{\partial W_j \partial W_i}\left(MW\right) M_{jj} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \ldots & 0 \end{bmatrix} \end{aligned} \tag{24}$$

Thus,

$$\mathbb{E}_M \nabla^2 G\left(M(\cdot)\right)(W) = \mathbb{E}_M \sum_{i,j} \begin{bmatrix} 0 & 0 & \ldots & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & p_i p_j \frac{\partial^2 G}{\partial W_j \partial W_i}\left(MW\right) & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \ldots & 0 \end{bmatrix}$$

Fix any $i_0, j_0 \in \{1, \ldots, n_\lambda + n_w\}$. It holds that

$$\begin{aligned}
&\mathbb{E}_M \left\| M_{i_0 i_0} H_G\left(MW\right) M_{j_0 j_0} - \mathbb{E}_M \frac{\partial^2 G\left(MW\right)}{\partial W_j \partial W_i}\left(W\right) \right\|^2 \\
&= \mathbb{E}_M \left\| M_{i_0 i_0} \frac{\partial^2 G(MW)}{\partial W_j \partial W_i} M_{j_0 j_0} - p_{i_0} p_{j_0} \frac{\partial^2 G\left(MW\right)}{\partial W_j \partial W_i} \right\|^2 \\
&= p_{i_0} p_{j_0} \left\| \frac{\partial^2 G\left(MW\right)}{\partial W_j \partial W_i} - p_{i_0} p_{j_0} \frac{\partial^2 G\left(MW\right)}{\partial W_j \partial W_i} \right\|^2 + \left(1 - p_{i_0} p_{j_0}\right) \left\| p_{i_0} p_{j_0} \frac{\partial^2 G}{\partial W_j \partial W_i}\left(MW\right) \right\|^2 \\
&\leq p_{i_0} p_{j_0} \left(1 - p_{i_0} p_{j_0}\right)^2 \left\| \frac{\partial^2 G\left(MW\right)}{\partial W_j \partial W_i} \right\|^2 + \left(1 - p_{i_0} p_{j_0}\right) \left(p_{i_0} p_{j_0}\right)^2 \left\| \frac{\partial^2 G}{\partial W_j \partial W_i}\left(MW\right) \right\|^2 \\
&\leq \left(p_{i_0} p_{j_0} \left(1 - p_{i_0} p_{j_0}\right)^2 + \left(1 - p_{i_0} p_{j_0}\right) \left(p_{i_0} p_{j_0}\right)^2 \right) C_G^2 = \left(p_{i_0} p_{j_0} \left(1 - p_{i_0} p_{j_0}\right)\right) C_G^2.
\end{aligned}$$

Since i_0 and j_0 are arbitrarily chosen, we have that

$$\begin{aligned}
&\mathbb{E}\|\nabla^2 G\left(M(\cdot)\right)(W) - \mathbb{E}_M \nabla^2 G\left(M(\cdot)\right)(W)\|^2 \\
&\leq \mathbb{E}\|\nabla^2 G\left(M(\cdot)\right)(W) - \mathbb{E}_M \nabla^2 G\left(M(\cdot)\right)(W)\|_F^2 \leq \sum_{i_0, j_0} p_{i_0} p_{j_0} \left(1 - p_{i_0} p_{j_0}\right) C_G^2.
\end{aligned}$$

□

D PROOFS OF LEMMA 2

We first present a lemma that will be used repeatedly.

Lemma 4. *Let $m_1 \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_1}$ and $m_2 \in \mathbb{R}^{n_2} \times \mathbb{R}^{n_2}$ be diagonal matrices. If $\|v(m_1)\|^2 \leq c_1$ for a constant c_1, then*

$$\mathbb{E}_{m_1} \|m_1 v(m_1)\|^2 \leq \max_j p_j c_1.$$

If $\|A_{ij}\| \leq c$,

$$\mathbb{E}_{m_1, m_2} \|m_1 A m_2\|^2 \leq \max p_j \max p_i n_1 n_2 c_2^2.$$

Proof. Noting $\|m_1 v(m_1)\|^2 = \sum_j (m_1)_{jj} \|v_j(m)\|^2$, it holds that

$$\begin{aligned}
&\mathbb{E}_{m_1} \|m_1 v\|^2 \leq \mathbb{E}_{m_1} \sum_j (m_1)_{jj} \|v(m_1)_j\|^2 = \sum_j p_j \mathbb{E}_{m_1 | (m_1)_{jj} = 1} \|v(m_1)_j\|^2 \\
&\leq \sum_j (\max_{j'} p_{j'}) \|v(m_1)_j\|^2 \leq \max_j p_j c_1.
\end{aligned}$$

On the other hand,

$$\mathbb{E}_{m_1, m_2} \|m_1 A m_2\|^2 \leq \mathbb{E}_{m_1, m_2} \sum_{i,j} (m_1)_{jj} A_{ji}^2 (m_2)_{ii} \leq \max p_j \max p_i n_1 n_2 c_2^2.$$

□

We present a detailed version of Lemma 2 here.

Lemma 5. *Consider* (9). *Suppose Assumptions 2, 3 and 4 hold. Let $\{(\lambda^k, w^k)\}$ be generated by Algorithm 1. Let γ in Algorithm 1 be small enough such that $\gamma\mu < 1$. Then*

$$\begin{aligned}
&\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 \leq (1 - \frac{1}{2}\gamma_k \underline{p}_w \mu) \mathbb{E}\|w^k - w(\lambda^{k-1})\|^2 \\
&+ (\frac{2}{\gamma_k \underline{p}_w \mu} - 1)(1 - \gamma_k \underline{p}_w \mu) L_w^2 \alpha_k^2 \mathbb{E}\|d^{k-1}\|^2 + \gamma_k^2 \overline{p}_w \sigma^2
\end{aligned} \tag{25}$$

and

$$\mathbb{E}\|w^{T+2}-w(\lambda^{T+1})\|^2-\mathbb{E}\|w^1-w(\lambda^0)\|^2\leq-\frac{1}{2}\sum_{k=1}^{T+1}\gamma_k\underline{p}_w\mu\mathbb{E}\|w^k-w(\lambda^{k-1})\|^2 + \sum_{k=1}^{T+1}(\frac{2}{\gamma_k\underline{p}_w\mu}-1)(1-\gamma_k\underline{p}_w\mu)L_w^2\alpha_k^2\mathbb{E}\|d^{k-1}\|^2+\sum_{k=1}^{T+1}\gamma_k^2\overline{p}_w\sigma^2. \tag{26}$$

Proof. Using the Lipschitz continuity of $\tilde{G}$, it holds that Using the definition of w^{k+1} in Algorithm 1,

$$\begin{aligned}&\mathbb{E}_{\underline{m}_\lambda^k,\underline{m}_w^k,\xi^k|\mathcal{M}^{k-1}}\|w^{k+1}-w(\lambda^k)\|^2=\|w^k-\gamma_k\underline{m}_w^k\nabla_w G(\underline{m}_\lambda^k\lambda^k,\underline{m}_w^k w^k;\xi^k)-w(\lambda^k)\|^2\\&=\|w^k-w(\lambda^k)\|^2+\gamma_k\left\langle\nabla\tilde{G}(\lambda^k,w^k),w^k-w(\lambda^k)\right\rangle\\&+\mathbb{E}_{\underline{m}_\lambda^k,\underline{m}_w^k,\xi^k|\mathcal{M}^{k-1}}\|\gamma_k\underline{m}_w^k\nabla_w G(\underline{m}_\lambda^k\lambda^k,\underline{m}_w^k w^k;\xi^k))\|^2\\&\leq\|w^k-w(\lambda^k)\|^2+\gamma_k\left\langle\nabla\tilde{G}(\lambda^k,w^k),w^k-w(\lambda^k)\right\rangle+\gamma_k^2\overline{p}_w\sigma^2,\end{aligned} \tag{27}$$

where the first equality uses Assumption 4 and (17), the second inequality follows from Assumption 4 and Lemma 4. Note from Proposition 2 that $\tilde{G}$ is strongly convex and $w(\lambda^k)$ is the minimizer of $\min_w\tilde{G}(\lambda^k,w)$. Thus,

$$\left\langle\nabla\tilde{G}(\lambda^k,w^k),w^k-w(\lambda^k)\right\rangle\geq\underline{p}_w\mu\|w^k-w(\lambda^k)\|^2.$$

This together with (27) gives

$$\begin{aligned}&\mathbb{E}_{\underline{m}_\lambda^k,\underline{m}_w^k,\xi^k|\mathcal{M}^{k-1}}\|w^{k+1}-w(\lambda^k)\|^2\leq(1-\gamma_k\underline{p}_w\mu)\|w^k-w(\lambda^k)\|^2+\gamma_k^2\overline{p}_w\sigma^2\\&\leq(1+\kappa_3^k)(1-\gamma_k\underline{p}_w\mu)\|w^k-w(\lambda^{k-1})\|^2+(1+\frac{1}{\kappa_3^k})(1-\gamma_k\underline{p}_w\mu)\|w(\lambda^k)-w(\lambda^{k-1})\|^2+\gamma_k^2\overline{p}_w\sigma^2\\&\overset{(a)}{\leq}(1+\kappa_3^k)(1-\gamma_k\underline{p}_w\mu)\|w^k-w(\lambda^{k-1})\|^2+(1+\frac{1}{\kappa_3^k})(1-\gamma_k\underline{p}_w\mu)L_w^2\alpha_k^2\|d^{k-1}\|^2+\gamma_k^2\overline{p}_w\sigma^2\\&=(1-\frac{1}{2}\gamma_k\underline{p}_w\mu)\|w^k-w(\lambda^{k-1})\|^2+(\frac{2}{\gamma_k\underline{p}_w\mu}-1)(1-\gamma_k\underline{p}_w\mu)L_w^2\alpha_k^2\|d^{k-1}\|^2+\gamma_k^2\overline{p}_w\sigma^2\end{aligned}$$

where $\kappa_3^k:=\frac{1}{2}\frac{\gamma_k\underline{p}_w\mu}{1-\gamma_k\underline{p}_w\mu}$ such that $(1+\kappa_3^k)(1-\gamma_k\underline{p}_w\mu)<1$, (a) makes use the definition of λ^k and the fact that $w(\lambda)$ is L_w-Lipschitz continuous by Proposition 2, and the last inequality makes use of the definition of κ_3^k. Taking expectation w.r.t. $\mathcal{M}^{k-1}$ we obtain

$$\begin{aligned}&\mathbb{E}\|w^{k+1}-w(\lambda^k)\|^2\\&\leq(1-\frac{1}{2}\gamma_k\underline{p}_w\mu)\mathbb{E}\|w^k-w(\lambda^{k-1})\|^2+(\frac{2}{\gamma_k\underline{p}_w\mu}-1)(1-\gamma_k\underline{p}_w\mu)L_w^2\alpha_k^2\mathbb{E}\|d^{k-1}\|^2+\gamma_k^2\overline{p}_w\sigma^2,\end{aligned}$$

Rearranging the above inequality, we have

$$\mathbb{E}\|w^{k+1}-w(\lambda^k)\|^2-\mathbb{E}\|w^k-w(\lambda^{k-1})\|^2\leq-\frac{1}{2}\gamma_k\underline{p}_w\mu\mathbb{E}\|w^k-w(\lambda^{k-1})\|^2 + (\frac{2}{\gamma_k\underline{p}_w\mu}-1)(1-\gamma_k\underline{p}_w\mu)L_w^2\alpha_k^2\mathbb{E}\|d^{k-1}\|^2+\gamma_k^2\overline{p}_w\sigma^2.$$

Summing the above inequality from $k=1$ to $k=T$ and divide both sides with T, we have

$$\mathbb{E}\|w^{T+1}-w(\lambda^T)\|^2-\mathbb{E}\|w^1-w(\lambda^0)\|^2\leq-\frac{1}{2}\sum_{k=1}^{T}\gamma_k\underline{p}_w\mu\mathbb{E}\|w^k-w(\lambda^{k-1})\|^2 + \sum_{k=1}^{T}(\frac{2}{\gamma_k\underline{p}_w\mu}-1)(1-\gamma_k\underline{p}_w\mu)L_w^2\alpha_k^2\mathbb{E}\|d^{k-1}\|^2+\sum_{k=1}^{T}\gamma_k^2\overline{p}_w\sigma^2.$$

□

E PROOFS OF THEOREM 1

We first show the property of $v_{m,\lambda}$ w.r.t. λ.

Lemma 6. *Consider* (9). *Suppose Assumptions 2, 3 and 4 hold. Let* $\{(\lambda^k, w^k)\}$ *be generated by Algorithm 1. Let* $v_{\overline{m}_w^k,\lambda^k}$ *be defined in* (16). *Denote* $\overline{p}_\lambda := \max_i p_{i,\lambda}$, $\underline{p}_\lambda := \min_i p_{i,\lambda}$, $\overline{p}_w := \max_i p_{i,w}$ *and* $\underline{p}_w := \min_i p_{i,w}$. *Then the following conclusions hold:*

(i) There exists C_v *such that* $\|v_{\overline{m}_w^k,\lambda^k}\| \le C_v$.

(ii) There exists C_w *such that*

$$\|\mathbb{E}_{M^k} v_{\overline{m}_w^k,\lambda^k} - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}}\|^2 \le C_w \alpha_k^2 \|d^k\|^2.$$

Proof. (i) follows from Assumptions 2 and 3. Now we prove (ii). Denote $\mathcal{M}^k := \{M^1, \dots, M^k, \xi_1, \dots, \xi_k, \dots, \eta_1, \dots, \eta_k\}$. Denote $F_M(\lambda) := F(m_\lambda(\cdot), m_w w(\cdot))(\lambda)$. It holds that

$$\begin{aligned}
&\mathbb{E}_{M^k} v_{\overline{m}_w^k,\lambda^k} - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}} \\
&= \left(\nabla^2_{ww}\tilde{G}(\lambda^k, w(\lambda^k))\right)^{-1} \nabla_w F_{M^k}(\lambda^k, w(\lambda^k)) \\
&- \left(\nabla^2_{ww}\tilde{G}(\lambda^{k-1}, w(\lambda^{k-1}))\right)^{-1} \nabla_w F_{M^{k-1}}(\lambda^{k-1}, w(\lambda^{k-1})) \\
&= \left(\left(\nabla^2_{ww}\tilde{G}(\lambda^k, w(\lambda^k))\right)^{-1} - \left(\nabla^2_{ww}\tilde{G}(\lambda^{k-1}, w(\lambda^{k-1}))\right)^{-1}\right) \nabla_w F_{M^k}(\lambda^k, w(\lambda^k)) \\
&+ \left(\nabla^2_{ww}\tilde{G}(\lambda^{k-1}, w(\lambda^{k-1}))\right)^{-1} \left(\nabla_w F_{M^k}(\lambda, w(\lambda^k)) - \nabla_w F_{M^{k-1}}(\lambda^{k-1}, w(\lambda^{k-1}))\right).
\end{aligned} \tag{28}$$

Thanks to Assumption 3, there exists $L_{G^{-1}}$ such that

$$\left\|\left(\nabla^2_{ww}\tilde{G}(\lambda^k, w(\lambda^k))\right)^{-1} - \left(\nabla^2_{ww}\tilde{G}(\lambda^{k-1}, w(\lambda^{k-1}))\right)^{-1}\right\|^2 \le L^2_{G^{-1}} \|w(\lambda^k) - w(\lambda^{k-1})\|^2. \tag{29}$$

Combining this with (28), (29) with Assumptions 2 and 3, we have that

$$\begin{aligned}
&\|\mathbb{E}_{M^k} v_{\overline{m}_w^k,\lambda^k} - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}}\|^2 \\
&\le L^2_{G^{-1}} C^F \|w(\lambda^k) - w(\lambda^{k-1})\|^2 + \frac{1}{\underline{p}_w^2 \mu^2} \bar{p}_w^2 (L^F_{ww})^2 \left(\alpha_k^2 \|d^k\|^2 + \|w(\lambda^k) - w(\lambda^{k-1})\|^2\right) \\
&\le \left(L^2_{G^{-1}} C^F \left(\frac{C_g}{\underline{p}_w \mu}\right)^2 + \frac{1}{\underline{p}_w^2 \mu^2} \bar{p}_w^2 (L^F_{ww})^2 (1 + \left(\frac{C_g}{\underline{p}_w \mu}\right)^2)\right) \alpha_k^2 \|d^k\|^2.
\end{aligned}$$

where the second inequality uses (4) and the third inquelity uses Proposition 2. □

Now we present the detailed version of Theorem 1.

Theorem 3. *Consider* (9). *Suppose Assumptions 2, 3 and 4 hold. Let* $\{(\lambda^k, w^k)\}$ *be generated by Algorithm 1. Denote* $V^k := \mathbb{E}\|v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}})\|^2$, *where* $v_{\overline{m}_w^{k-1},\lambda^{k-1}}$ *is defined in* (16). *Denote* $\overline{p}_\lambda := \max_i p_{i,\lambda}$, $\underline{p}_\lambda := \min_i p_{i,\lambda}$, $\overline{p}_w := \max_i p_{i,w}$ *and* $\underline{p}_w := \min_i p_{i,w}$. *Suppose* $r \ge C_v$, *where* C_v *is defined as in Lemma 6 (i). Then*

$$\begin{aligned}
V^{k+1} &\le \frac{2}{\underline{p}_w \mu} \beta_k L^F_{ww} \overline{p}_w \mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 + (1 - \frac{1}{2}\beta_k \underline{p}_w \mu)^2 V^k + \frac{2}{\beta \underline{p}_w \mu} C_w \alpha_k \mathbb{E}\|d^k\|^2 \\
&+ \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1 - \underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1 - \underline{p}_w) n (C^F)^2\right).
\end{aligned}$$

and

$$
\begin{aligned}
V^{T+2} - V^1 \leq & \sum_{k=1}^{T+1} \left(-\frac{1}{2}\underline{p}_w \mu + \frac{1}{4}\beta_k \underline{p}_w^2 \mu^2 \right) \beta_k V^k \\
& + \frac{2}{\underline{p}_w \mu} \sum_{k=1}^{T+1} \beta_k L_{ww}^F \bar{p}_w \mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 + \frac{2}{\beta \underline{p}_w \mu} C_w \sum_{k=1}^{T+1} \alpha_k \mathbb{E}\|d^k\|^2 \\
& + \sum_{k=1}^{T+1} \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1 - \underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1 - \underline{p}_w) n (C^F)^2 \right).
\end{aligned} \tag{30}
$$

Proof. Denote $\mathcal{M}^k := \{M^1, \ldots, M^k, \xi_1, \ldots, \xi_k, \ldots, \eta_1, \ldots, \eta_k\}$. Denote $\mathcal{B}^k := \{\xi^k, \eta^k\}$. Denote $F_M(\lambda) := F(m_\lambda(\cdot), m_w w(\cdot))(\lambda)$.

Let $\tilde{v}_d^{k+1} := \beta_k \nabla_w \mathbb{E}_{M^k} F_{M^k}(\lambda^k, w^{k+1}) - \left(I - \beta_k \nabla_{ww}^2 \tilde{G}(\lambda^k, w^k) \right) v^k$. Therefore, $\mathbb{E}_{M^k, \xi^k, \eta^k} v_d^{k+1} = \tilde{v}_d^{k+1}$. Using Assumption 2, Assumption 3, Lemma 3 and Lemma 4, we have

$$
\begin{aligned}
& v_d^{k+1} \mathbb{E}_{M^k, \xi^k, \eta^k} \|\tilde{v}_d^{k+1} - \mathbb{E}_{M^k} v_{\overline{m}_w^k, \lambda^k}\|^2 \\
& \leq \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1 - \underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1 - \underline{p}_w) n (C^F)^2 \right).
\end{aligned} \tag{31}
$$

On the other hand, the definition of $v_{\overline{m}_w^k, \lambda^k}$ shows that

$$
\mathbb{E}_{M^k} v_{\overline{m}_w^k, \lambda^k} = \beta_k \nabla_w \mathbb{E}_{M^k} F_{M^k}(\lambda^k, w(\lambda^k)) + \left(I - \beta_k \nabla_{ww}^2 \tilde{G}(\lambda^k, w(\lambda^k) \right) \mathbb{E}_M^k v_{\overline{m}_w^k, \lambda^k}.
$$

Combining this with (31) and the definition of v^{k+1}, we have that

$$
\begin{aligned}
& \mathbb{E}\|v^{k+1} - \mathbb{E}_{M^k} v_{\overline{m}_w^k, \lambda^k}\|^2 \leq \|\tilde{v}^{k+1} - v_{\overline{m}_w^k, \lambda^k}\|^2 \\
& = \mathbb{E}\|\tilde{v}^{k+1} - v_{\overline{m}_w^k, \lambda^k} - \tilde{v}_d^{k+1} - v_{\overline{m}_w^k, \lambda^k} + \tilde{v}_d^{k+1} - v_{\overline{m}_w^k, \lambda^k}\|^2 \\
& = \mathbb{E}\|\tilde{v}^{k+1} - \tilde{v}_d^{k+1}\|^2 + \|\tilde{v}_d^{k+1} - v_{\overline{m}_w^k, \lambda^k}\|^2 \\
& \leq (1 + \frac{1}{\kappa_1^k}) \beta_k^2 \mathbb{E}\|\mathbb{E}_{M^k} F_{M^k}(\lambda^k, w^k) - \mathbb{E}_{M^k} F_{M^k}(\lambda^k, w(\lambda^k))\|^2 \\
& + (1 + \kappa_1^k) \mathbb{E}\| \left(I - \beta_k \nabla_{ww}^2 \tilde{G}(\lambda^k, w(\lambda^k)) \right) (v^k - \mathbb{E}_{M^k} v_{\overline{m}_w^k, \lambda^k})\|^2 \\
& + \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1 - \underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1 - \underline{p}_w) n (C^F)^2 \right),
\end{aligned} \tag{32}
$$

where the first inequality is because $r \geq C_v$ and the projection on $B(0, r)$ is a contraction.

For the first term in the above inequality, using Lemma 4 and Assumption 2, it holds that

$$
\mathbb{E}\|\mathbb{E}_{M^k} F_{M^k}(\lambda^k, w^k) - \mathbb{E}_{M^k} F_{M^k}(\lambda^k, w(\lambda^k))\|^2 \leq L_{ww}^F \bar{p}_w \|w^{k+1} - w(\lambda^k)\|^2. \tag{33}
$$

For the second term in (32), using Proposition 2, it holds that

$$
\begin{aligned}
& \| \left(I - \beta_k \nabla_{ww}^2 \tilde{G}(\lambda^k, w(\lambda^k)) \right) \mathbb{E}_{M^k} (v^k - v_{\overline{m}_w^k, \lambda^k})\|^2 \\
& \leq (1 - \beta_k \underline{p}_w \mu)^2 \|(v^k - \mathbb{E}_{M^k} v_{\overline{m}_w^k, \lambda^k})\|^2 \\
& \leq (1 + \kappa_2^k)(1 - \beta_k \underline{p}_w \mu)^2 \|(v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1}, \lambda^{k-1}})\|^2 \\
& + (1 + \frac{1}{k_2^k})(1 - \beta_k \underline{p}_w \mu)^2 \|\mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1}, \lambda^{k-1}} - \mathbb{E}_{M^k} v_{\overline{m}_w^k, \lambda^k}\|^2.
\end{aligned} \tag{34}
$$

where $\kappa_2^k > 0$. This together with Lemma 6 implies that

$$
\begin{aligned}
& \| \left(I - \beta_k \nabla_{ww}^2 \tilde{G}(\lambda^k, w(\lambda^k)) \right) \mathbb{E}_{M^k} (v^k - v_{\overline{m}_w^k, \lambda^k})\|^2 \\
& \leq (1 - \beta_k \underline{p}_w \mu)^2 \|(v^k - \mathbb{E}_{M^k} v_{\overline{m}_w^k, \lambda^k})\|^2 \\
& \leq (1 + \kappa_2^k)(1 - \beta_k \underline{p}_w \mu)^2 \|(v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1}, \lambda^{k-1}})\|^2 + (1 + \frac{1}{k_2^k})(1 - \beta_k \underline{p}_w \mu)^2 C_w \alpha_k^2 \|d^k\|^2
\end{aligned} \tag{35}
$$

Combining (35), (33) and (32), we have that

$$
\begin{aligned}
&\mathbb{E}\|v^{k+1} - \mathbb{E}_{M^k} v_{\overline{m}_w^k,\lambda^k}\|^2 \leq \|\tilde{v}^{k+1} - v_{\overline{m}_w^k,\lambda^k}\|^2 \\
&= \mathbb{E}\|\tilde{v}^{k+1} - v_{\overline{m}_w^k,\lambda^k} - \tilde{v}_d^{k+1} - v_{\overline{m}_w^k,\lambda^k} + \tilde{v}_d^{k+1} - v_{\overline{m}_w^k,\lambda^k}\|^2 \\
&= \mathbb{E}\|\tilde{v}^{k+1} - \tilde{v}_d^{k+1}\|^2 + \|\tilde{v}_d^{k+1} - v_{\overline{m}_w^k,\lambda^k}\|^2 \\
&\leq (1 + \frac{1}{\kappa_1^k})\beta_k^2 L_{ww}^F \overline{p}_w \|w^{k+1} - w(\lambda^k)\|^2 \\
&+ (1+\kappa_1^k)(1+\kappa_2^k)(1-\beta_k \underline{p}_w \mu)^2 \|(v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}})\|^2 \\
&+ (1+\kappa_1^k)(1+\frac{1}{k_2^k})(1-\beta_k \underline{p}_w \mu)^2 C_w \alpha_k^2 \|d^k\|^2 \\
&+ \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1-\underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1-\underline{p}_w) n (C^F)^2\right).
\end{aligned} \tag{36}
$$

Let $\kappa_1 = \kappa_2 = \frac{\beta_k \underline{p}_w \mu}{2(1-\beta_k \underline{p}_w \mu)}$. Then $1+\kappa_1 = \frac{1}{2} + \frac{1}{2-\beta_k \underline{p}_w \mu}$, $1 + \frac{1}{\kappa_1} = 1 + \frac{1}{\kappa_2} = \frac{2-\beta_k \underline{p}_w \mu}{\beta_k \underline{p}_w \mu} \leq \frac{2}{\beta_k \underline{p}_w \mu}$. Thus, (36)

$$
\begin{aligned}
&\mathbb{E}\|v^{k+1} - \mathbb{E}_{M^k} v_{\overline{m}_w^k,\lambda^k}\|^2 \leq \|\tilde{v}^{k+1} - v_{\overline{m}_w^k,\lambda^k}\|^2 \\
&\leq \frac{2}{\beta_k \underline{p}_w \mu} \beta_k^2 L_{ww}^F \overline{p}_w \|w^{k+1} - w(\lambda^k)\|^2 \\
&+ (\frac{1}{2} + \frac{1}{2-\beta_k \underline{p}_w \mu})(\frac{1}{2} + \frac{1}{2-\beta_k \underline{p}_w \mu})(1-\beta_k \underline{p}_w \mu)^2 \|(v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}})\|^2 \\
&+ (\frac{1}{2} + \frac{1}{2-\beta_k \underline{p}_w \mu}) \frac{2}{\beta_k \underline{p}_w \mu} (1-\beta_k \underline{p}_w \mu)^2 C_w \alpha_k^2 \|d^k\|^2 \\
&+ \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1-\underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1-\underline{p}_w) n (C^F)^2\right) \\
&\leq \frac{2}{\underline{p}_w \mu} \beta_k L_{ww}^F \overline{p}_w \mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 \\
&+ (1 - \frac{1}{2}\beta_k \underline{p}_w \mu)^2 \mathbb{E}\|v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}}\|^2 + \frac{2}{\beta \underline{p}_w \mu} C_w \alpha_k \mathbb{E}\|d^k\|^2 \\
&+ \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1-\underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1-\underline{p}_w) n (C^F)^2\right).
\end{aligned}
$$

Rearranging the above inequality, we have

$$
\begin{aligned}
&\mathbb{E}\|v^{k+1} - \mathbb{E}_{M^k} v_{\overline{m}_w^k,\lambda^k}\|^2 - \mathbb{E}\|v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}}\|^2 \\
&\leq \left(-\frac{1}{2}\underline{p}_w \mu + \frac{1}{4}\beta_k \underline{p}_w^2 \mu^2\right) \beta_k \mathbb{E}\|v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}}\|^2 \\
&+ \frac{2}{\underline{p}_w \mu} \beta_k L_{ww}^F \overline{p}_w \mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 + \frac{2}{\beta \underline{p}_w \mu} C_w \alpha_k \mathbb{E}\|d^k\|^2 \\
&+ \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1-\underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1-\underline{p}_w) n (C^F)^2\right).
\end{aligned}
$$

Summing the above inequality from $k=1$ to $k=T+1$, we have

$$
\begin{aligned}
&\mathbb{E}\|v^{T+2} - \mathbb{E}_{M^{T+1}} v_{\overline{m}_w^{T+1},\lambda^{T+1}}\|^2 - \mathbb{E}\|v^1 - \mathbb{E}_{M^0} v_{\overline{m}_w^0,\lambda^0}\|^2 \\
&\leq \sum_{k=1}^{T+1} \left(-\frac{1}{2}\underline{p}_w \mu + \frac{1}{4}\beta_k \underline{p}_w^2 \mu^2\right) \beta_k \mathbb{E}\|v^k - \mathbb{E}_{M^{k-1}} v_{\overline{m}_w^{k-1},\lambda^{k-1}}\|^2 \\
&+ \frac{2}{\underline{p}_w \mu} \sum_{k=1}^{T+1} \beta_k L_{ww}^F \overline{p}_w \mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 + \frac{2}{\beta \underline{p}_w \mu} C_w \sum_{k=1}^{T+1} \alpha_k \mathbb{E}\|d^k\|^2 \\
&+ \sum_{k=1}^{T+1} \beta_k^2 \left(4\bar{p}_w \sigma^2 + 4\bar{p}_w^2 \sigma_h^2 + 4\bar{p}_w^2 (1-\underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1-\underline{p}_w) n (C^F)^2\right).
\end{aligned}
$$

□

F PROOF OF THEOREM 2

We first present the following lemma that is needed in the proofs of Theorem 2.

Lemma 7. *Consider* (9)*. Let assumptions in Theorem 2 hold. Then, it holds that*

$$\begin{aligned}
&\mathbb{E}F_{M^{k+1}}(\lambda^{k+1}) - \mathbb{E}F_{M^k}(\lambda^k) + \Gamma_3\left(\mathbb{E}\|d^{k+1} - \bar{\nabla}\bar{F}(\lambda^k, w^{k+1}, v^{k+1})\|^2 - \mathbb{E}\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2\right) \\
&\leq -\frac{\alpha_k}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \\
&- \Gamma_3\alpha_k\eta_k\mathbb{E}\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2 + \Gamma_3\alpha_k\eta_k^2\left((\bar{p}_\lambda)^2\sigma^2 + \underline{p}_\lambda\underline{p}_w\sigma_h^2\right) \\
&+ \Gamma_3\alpha_k\overline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2(1+3L_w^2) + 2C_G^2 3C_w\right)\alpha_{k-1}^2\mathbb{E}\|d^{k-1}\|^2 \\
&+ \Gamma_3\alpha_k\underline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2\right)\left(3\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2 + 3\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2\right) \\
&+ 2\Gamma_3\alpha_k\underline{p}_\lambda\underline{p}_w C_G^2\left(3V^k + 3V^{k+1}\right) \\
&+ \alpha_k\overline{p}_w\left((L_{12}^F)^2 + nL_{G^w}^2 r^2\right)\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 + \alpha_k\overline{p}_\lambda\overline{p}_w nC_G^2 V^{k+1} \\
&+ \left(\frac{\alpha_k^2 L_{\bar{F}}}{2} - \frac{\alpha_k}{2}\right)\mathbb{E}\|d^{k+1}\|^2,
\end{aligned} \tag{37}$$

Proof. Denote $\mathcal{M}^k := \{M^1, \ldots, M^k, \xi_1, \ldots, \xi_k, \ldots, \eta_1, \ldots, \eta_k\}$. Denote $\mathcal{B}^k := \{\xi^k, \eta^k\}$. Denote $F_M(\lambda) := F(m_\lambda(\cdot), m_w w(\cdot))(\lambda)$.

Thanks to Proposition 2, there exists $L_{\bar{F}}$ such that $\nabla F_M(\lambda)$ is $L_{\bar{F}}$-Lipschitz continuous. Thus,

$$\begin{aligned}
&\mathbb{E}F_{M^{k+1}}(\lambda^{k+1}) \leq \mathbb{E}F_{M^{k+1}}(\lambda^k) + \mathbb{E}\left\langle\nabla F_{M^{k+1}}(\lambda^k), \lambda^{k+1} - \lambda^k\right\rangle + \mathbb{E}\frac{L_{\bar{F}}}{2}\|\lambda^{k+1} - \lambda^k\|^2 \\
&= \mathbb{E}F_{M^k}(\lambda^k) - \alpha_k\mathbb{E}\left\langle\nabla F_{M^k}(\lambda^k), d^{k+1}\right\rangle + \alpha_k^2\mathbb{E}\frac{L_{\bar{F}}}{2}\|d^{k+1}\|^2 \\
&= \mathbb{E}F_{M^k}(\lambda^k) - \frac{\alpha_k}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 + \frac{\alpha_k}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k) - d^{k+1}\|^2 \\
&+ \left(\frac{\alpha_k^2 L_{\bar{F}}}{2} - \frac{\alpha_k}{2}\right)\mathbb{E}\|d^{k+1}\|^2,
\end{aligned} \tag{38}$$

where the first equality is because M^{k+1} and M^k have the same distributions and therefore

$$\begin{aligned}
\mathbb{E}F_{M^{k+1}}(\lambda^k) &= \mathbb{E}_{\mathcal{M}^{k-1}}\mathbb{E}_{M^{k+1}, M^k|\mathcal{M}^{k-1}}F_{M^{k+1}}(\lambda^k) \\
&= \mathbb{E}_{\mathcal{M}^{k-1}}\mathbb{E}_{M^k|\mathcal{M}^{k-1}}F_{M^k}(\lambda^k) = \mathbb{E}F_{M^k}(\lambda^k).
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\nabla F_{M^{k+1}}(\lambda^k) &= \mathbb{E}_{\mathcal{M}^{k-1}}\mathbb{E}_{M^{k+1}, M^k|\mathcal{M}^{k-1}}\nabla F_{M^{k+1}}(\lambda^k) \\
&= \mathbb{E}_{\mathcal{M}^{k-1}}\mathbb{E}_{M^k|\mathcal{M}^{k-1}}\nabla F_{M^k}(\lambda^k) = \mathbb{E}\nabla F_{M^k}(\lambda^k).
\end{aligned}$$

Next, we bound the third term on the right hand side of (38). Using the chain rule, it holds that

$$\nabla F_M(\lambda) = m_\lambda\nabla_\lambda F(m_\lambda\lambda, m_w w(\lambda)) + J(w(\lambda))^T m_w\nabla_w F(m_\lambda\lambda, m_w w(\lambda)),$$

where $J(w(\lambda))$ is the Jacobian of $w(\lambda)$. Using Proposition 2, the above equality ca be further passed to

$$\begin{aligned}
&\nabla F_M(\lambda) \\
&= m_\lambda\nabla_\lambda F(m_\lambda\lambda, m_w w(\lambda)) - \nabla_{\lambda w}^2\tilde{G}(\lambda, w(\lambda))\underbrace{\left(\nabla_{ww}^2\tilde{G}(\lambda, w(\lambda))\right)^{-1} m_w\nabla_w F(m_\lambda\lambda, m_w w(\lambda))}_{v_{m_w,\lambda}}.
\end{aligned}$$

Denote

$$\bar{\nabla}\bar{F}(\lambda^k, w^{k+1}, v^{k+1}) := \overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w w^{k+1}) - \underline{m}_{\lambda^k}\nabla_{\lambda w}^2 G(\underline{m}_\lambda^k\lambda^k, \underline{m}_w^k w^{k+1})\underline{m}_w^k v^{k+1}. \tag{39}$$

This together with the definition of d^{k+1} in Algorithm 1 gives

$$\begin{aligned}&\|d^{k+1}-\nabla F_{M^k}(\lambda^k)\|^2\\&\leq 2\|d^{k+1}-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})\|^2+2\|\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})-\nabla F_{M^k}(\lambda^k)\|^2.\end{aligned}\tag{40}$$

Now we bound the first term on the right hand side of the above inequality.

$$\begin{aligned}&\mathbb{E}\|d^{k+1}-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})\|^2\\&=\mathbb{E}\|\nabla\bar{F}(\lambda^k,w^{k+1})+(1-\eta_k)(d^k-\nabla\bar{F}(\lambda^{k-1},w^k)))-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})\|^2\\&=(1-\eta_k)^2\mathbb{E}\|d^k-\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)\|^2\\&+\mathbb{E}\|(1-\eta_k)(\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)-\nabla\bar{F}(\lambda^{k-1},w^k,v^k)))\\&+\nabla\bar{F}(\lambda^k,w^{k+1},v^{k+1})-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})\|^2\\&\leq(1-\eta_k)^2\mathbb{E}\|d^k-\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)\|^2\\&+2\eta^2\mathbb{E}\|\nabla\bar{F}(\lambda^k,w^{k+1},v^{k+1})-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})\|\\&+2(1-\eta_k)^2\mathbb{E}\|\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)-\nabla\bar{F}(\lambda^{k-1},w^k,v^k)\\&+\nabla\bar{F}(\lambda^k,w^{k+1},v^{k+1})-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})\|^2\\&\leq(1-\eta_k)^2\mathbb{E}\|d^k-\bar{\nabla}\bar{F}(\lambda^{k-1},w^k)\|^2+\eta_k^2\mathbb{E}\|\nabla\bar{F}(\lambda^k,w^{k+1})-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1},v^{k+1})\|^2\\&+(1-\eta_k)^2\mathbb{E}_{\xi^k,\eta^k|\mathcal{M}^k/\{\xi^k,\eta^k\}}\|\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1})\|^2\\&\leq(1-\eta_k)^2\|(d^k-\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)\|^2+\eta_k^2\left((\bar{p}_\lambda)^2\sigma^2+\underline{p}_\lambda\underline{p}_w\sigma_h^2\right)\\&+(1-\eta_k)^2\mathbb{E}\|\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})\|^2\end{aligned}\tag{41}$$

where the second inequality uses Lemma 4 and Assumption 4, the last inequality uses the fact that $\mathbb{E}\|X\|^2\geq\mathbb{E}\|X-\mathbb{E}X\|^2$. For the last term in the above inequality, using the definition $\bar{\nabla}\bar{F}(\lambda^{k-1},w^k)$, it holds that

$$\begin{aligned}&\mathbb{E}\|\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)-\bar{\nabla}\bar{F}(\lambda^k,w^{k+1},v^{k+1})\|^2\\&=2\mathbb{E}\|\overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^{k-1}\lambda^{k-1},\overline{m}_w w^k)-\overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k,\overline{m}_w w^{k+1})\|^2\\&+2\mathbb{E}\|\underline{m}_{\lambda^k}\nabla^2_{\lambda w}G(\underline{m}_\lambda^{k-1}\lambda^{k-1},\underline{m}_w^k w^k)\underline{m}_w^k v^k-\underline{m}_{\lambda^k}\nabla^2_{\lambda w}G(\underline{m}_\lambda^k\lambda^k,\underline{m}_w^k w^{k+1})\underline{m}_w^k v^{k+1}\|^2\\&\leq\left(2\overline{p}_\lambda(L_{12}^F)^2+2\underline{p}_\lambda\underline{p}_w r^2\right)\left(\mathbb{E}\|\lambda^{k-1}-\lambda^k\|+\mathbb{E}\|w^{k+1}-w^k\|^2\right)+2\underline{p}_\lambda\underline{p}_w C_G^2\mathbb{E}\|v^k-v^{k+1}\|^2,\end{aligned}\tag{42}$$

Note that

$$\begin{aligned}&\mathbb{E}\|w^{k+1}-w^k\|^2\\&\leq 3\mathbb{E}\|w^k-w(\lambda^{k-1})\|^2+3\mathbb{E}\|w^{k+1}-w(\lambda^k)\|^2+3L_w^2\mathbb{E}\|\lambda^k-\lambda^{k-1}\|^2\\&=3\mathbb{E}\|w^k-w(\lambda^{k-1})\|^2+3\mathbb{E}\|w^{k+1}-w(\lambda^k)\|^2+3L_w^2\alpha_{k-1}^2\mathbb{E}\|d^{k-1}\|^2.\end{aligned}\tag{43}$$

On the other hand, using the definition of $\tilde{v}^k$ and v^k, Lemma 6 and Proposition 2 we have that

$$\mathbb{E}\|v^k-v^{k+1}\|^2\leq 3V^k+3V^{k+1}+3C_w\alpha_k^2\mathbb{E}\|d^{k-1}\|^2.\tag{44}$$

Combining (41), (42), (43) and (44), we have

$$
\begin{aligned}
&\mathbb{E}\|d^{k+1} - \bar{\nabla}\bar{F}(\lambda^k, w^{k+1}, v^{k+1})\|^2 \\
&\leq (1-\eta_k)^2\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2 + \eta_k^2\left((\bar{p}_\lambda)^2\sigma^2 + \underline{p}_\lambda\underline{p}_w\sigma_h^2\right) \\
&+ (1-\eta_k)^2\overline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2(1+3L_w^2) + 2C_G^2 3C_w\right)\alpha_{k-1}^2\mathbb{E}\|d^{k-1}\|^2 \\
&+ (1-\eta_k)^2\underline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2\right)\left(3\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2 + 3\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2\right) \\
&+ 2(1-\eta_k)^2\underline{p}_\lambda\underline{p}_w C_G^2\left(3V^k + 3V^{k+1}\right) \\
&\leq (1-\eta_k)\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2 + \eta_k^2\left((\bar{p}_\lambda)^2\sigma^2 + \underline{p}_\lambda\underline{p}_w\sigma_h^2\right) \\
&+ \overline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2(1+3L_w^2) + 2C_G^2 3C_w\right)\alpha_{k-1}^2\mathbb{E}\|d^{k-1}\|^2 \\
&+ \underline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2\right)\left(3\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2 + 3\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2\right) \\
&+ 2\underline{p}_\lambda\underline{p}_w C_G^2\left(3V^k + 3V^{k+1}\right).
\end{aligned} \tag{45}
$$

Rearranging terms in the above inequality, we have

$$
\begin{aligned}
&\mathbb{E}\|d^{k+1} - \bar{\nabla}\bar{F}(\lambda^k, w^{k+1}, v^{k+1})\|^2 - \mathbb{E}\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2 \\
&\leq -\eta_k\mathbb{E}\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2 + \eta_k^2\left((\bar{p}_\lambda)^2\sigma^2 + \underline{p}_\lambda\underline{p}_w\sigma_h^2\right) \\
&+ \overline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2(1+3L_w^2) + 2C_G^2 3C_w\right)\alpha_{k-1}^2\mathbb{E}\|d^{k-1}\|^2 \\
&+ \underline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2\right)\left(3\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2 + 3\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2\right) \\
&+ 2\underline{p}_\lambda\underline{p}_w C_G^2\left(3V^k + 3V^{k+1}\right).
\end{aligned} \tag{46}
$$

Now we bound the second term in (40). Using the definition of $\bar{\nabla}\bar{F}(\lambda^k, w^{k+1})$ and $\nabla F_{M^k}(\lambda^k)$, we have

$$
\begin{aligned}
&\mathbb{E}\|\bar{\nabla}\bar{F}(\lambda^k, w^{k+1}) - \nabla F_{M^k}(\lambda^k)\|^2 \\
&\leq \mathbb{E}\|\overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w w^{k+1}) - \overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w^k w(\lambda^k))\|^2 \\
&+ \mathbb{E}\|\underline{m}_{\lambda^k}\nabla_{\lambda w}^2 G(\underline{m}_\lambda^k\lambda^k, \underline{m}_w^k w^{k+1})\underline{m}_w^k v^{k+1} - \nabla_{\lambda w}^2\tilde{G}(\lambda^k, w(\lambda^k))v_{\overline{m}_w^k,\lambda^k}\|^2 \\
&\leq \mathbb{E}\|\overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w w^{k+1}) - \overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w^k w(\lambda^k))\|^2 \\
&+ \mathbb{E}\|\underline{m}_{\lambda^k}\nabla_{\lambda w}^2 G(\underline{m}_\lambda^k\lambda^k, \underline{m}_w^k w^{k+1})\underline{m}_w^k\left(v^{k+1} - v_{\overline{m}_w^k,\lambda^k}\right)\|^2 \\
&+ \mathbb{E}\|\left(\underline{m}_{\lambda^k}\nabla_{\lambda w}^2 G(\underline{m}_\lambda^k\lambda^k, \underline{m}_w^k w^{k+1})\underline{m}_w^k - \nabla_{\lambda w}^2\tilde{G}(\lambda^k, w(\lambda^k))\right)v_{\overline{m}_w^k,\lambda^k}\|^2 \\
&\leq \mathbb{E}\|\overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w w^{k+1}) - \overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w^k w(\lambda^k))\|^2 \\
&+ \mathbb{E}\|\underline{m}_{\lambda^k}\nabla_{\lambda w}^2 G(\underline{m}_\lambda^k\lambda^k, \underline{m}_w^k w^{k+1})\underline{m}_w^k\left(v^{k+1} - v_{\overline{m}_w^k,\lambda^k}\right)\|^2 \\
&+ \mathbb{E}\|\left(\underline{m}_{\lambda^k}\nabla_{\lambda w}^2 G(\underline{m}_\lambda^k\lambda^k, \underline{m}_w^k w^{k+1})\underline{m}_w^k - \underline{m}_{\lambda^k}\nabla_{\lambda w}^2 G(\underline{m}_\lambda^k\lambda^k, \underline{m}_w^k w(\lambda^k))\underline{m}_w^k\right)v_{\overline{m}_w^k,\lambda^k}\|^2 \\
&+ \mathbb{E}\|\left(\underline{m}_{\lambda^k}\nabla_{\lambda w}^2 G(\underline{m}_\lambda^k\lambda^k, \underline{m}_w^k w(\lambda^k))\underline{m}_w^k - \nabla_{\lambda w}^2\tilde{G}(\lambda^k, w(\lambda^k))\right)v_{\overline{m}_w^k,\lambda^k}\|^2.
\end{aligned}
$$

Using Lemma 4, (23) together with Assumption 2, 3 and 4, the above inequality can be further passed to

$$
\begin{aligned}
&\mathbb{E}\|\bar{\nabla}\bar{F}(\lambda^k) - \nabla F_{\overline{m}_\lambda^k}(\lambda^k)\|^2 \\
&\leq \mathbb{E}\|\overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w w^{k+1}) - \overline{m}_\lambda^k\nabla_\lambda F(\overline{m}_\lambda^k\lambda^k, \overline{m}_w^k w(\lambda^k))\|^2 \\
&+ \overline{p}_\lambda\overline{p}_w n C_G^2\mathbb{E}_{\underline{m}_w^k|\mathcal{M}^{k-1}}\mathbb{E}\|\underline{m}_w^k\left(v^{k+1} - E_{m^k}v_{\overline{m}_w^k,\lambda^k}\right)\|^2 \\
&+ \overline{p}_\lambda\overline{p}_w n L_{G^w}\mathbb{E}\|w^k - w(\lambda^k)\|\|E_{m^k}v_{\overline{m}_w^k,\lambda^k}\|^2.
\end{aligned}
$$

Using Lemma 6, Lemma 4 and the fact that $\underline{m}_w^k$ is a diagonal matrix with all entries belongs to $\{0,1\}$, the above inequality can be further passed to

$$\begin{aligned}
&\mathbb{E}\|\bar{\nabla}\bar{F}(\lambda^k) - \nabla F_{\overline{m}_\lambda^k}(\lambda^k)\|^2 \\
&\leq \left((L_{12}^F)^2\overline{p}_w + \bar{p}_\lambda\overline{p}_w n L_{G^w}^2 r^2\right)\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 + \overline{p}_\lambda\overline{p}_w n C_G^2 \mathbb{E}\|v^{k+1} - v_{\overline{m}_w^k,\lambda^k}\|^2 \\
&\leq \overline{p}_w\left((L_{12}^F)^2 + n L_{G^w}^2 r^2\right)\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 + \overline{p}_\lambda\overline{p}_w n C_G^2 V^{k+1}
\end{aligned} \tag{47}$$

Summing (38), $\frac{\alpha_k}{2}$ times of (40), α_k times of (47) and Γ_3 times of (46), we have that

$$\begin{aligned}
&\mathbb{E}F_{M^{k+1}}(\lambda^{k+1}) - \mathbb{E}F_{M^k}(\lambda^k) + \Gamma_3\left(\mathbb{E}\|d^{k+1} - \bar{\nabla}\bar{F}(\lambda^k, w^{k+1}, v^{k+1})\|^2 - \mathbb{E}\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2\right) \\
&\leq -\frac{\alpha_k}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \\
&- \Gamma_3\alpha_k\eta_k\mathbb{E}\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2 + \Gamma_3\alpha_k\eta_k^2\left((\bar{p}_\lambda)^2\sigma^2 + \underline{p}_\lambda\underline{p}_w\sigma_h^2\right) \\
&+ \Gamma_3\alpha_k\overline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2(1+3L_w^2) + 2C_G^2 3C_w\right)\alpha_{k-1}^2\mathbb{E}\|d^{k-1}\|^2 \\
&+ \Gamma_3\alpha_k\underline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2\right)\left(3\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2 + 3\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2\right) \\
&+ 2\Gamma_3\alpha_k\underline{p}_\lambda\underline{p}_w C_G^2\left(3V^k + 3V^{k+1}\right) \\
&+ \alpha_k\overline{p}_w\left((L_{12}^F)^2 + n L_{G^w}^2 r^2\right)\mathbb{E}\|w^{k+1} - w(\lambda^k)\|^2 + \alpha_k\overline{p}_\lambda\overline{p}_w n C_G^2 V^{k+1} \\
&+ \left(\frac{\alpha_k^2 L_{\bar{F}}}{2} - \frac{\alpha_k}{2}\right)\mathbb{E}\|d^{k+1}\|^2.
\end{aligned}$$

This completes the proof. □

Now we present the detailed version of Theorem 2.

Theorem 4. *Consider* (9). *Suppose Assumptions 2, 3 and 4 hold. Let* $\{(w^k, \lambda^k)\}$ *be generated by Algorithm 1. Suppose* F *is bounded below by* $\bar{F}$. *Then there exist small* δ *in Algorithm 1 such that*

$$\begin{aligned}
&\frac{1}{T-1}\sum_{k=1}^{T-1}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \leq \frac{1}{\sqrt{T}}\frac{1}{\underline{p}_w\underline{p}_\lambda}I_0 \\
&+ \frac{\ln(T+1)}{\sqrt{T}}O\left(\frac{\bar{p}_\lambda^2}{\underline{p}_\lambda\underline{p}_w} + \underline{p}_w\bar{p}_w\right)\left(\sigma^2 + \sigma_h^2\right) + \frac{\ln(T+1)}{\sqrt{T}}O\left(\underline{p}_w\left(\bar{p}_w^2(1-\underline{p}_w^2) + \bar{p}_w(1-\underline{p}_w)\right)\right)
\end{aligned} \tag{48}$$

Proof. Rearrange (37) and summing it from 1 to T, we have

$$\begin{aligned}
&\sum_1^T\frac{\alpha_k}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \leq -\mathbb{E}F_{M^{T+1}}(\lambda^{T+1}) + \mathbb{E}F_{M^0}(\lambda^0) \\
&+ \Gamma_3\left(\mathbb{E}\|(d^1 - \bar{\nabla}\bar{F}(\lambda^0, w^1, v^1)\|^2 - \mathbb{E}\|d^{T+1} - \bar{\nabla}\bar{F}(\lambda^k, w^{T+1}, v^{T+1})\|^2\right) \\
&- \Gamma_3\sum_1^T\alpha_k\eta_k\mathbb{E}\|(d^k - \bar{\nabla}\bar{F}(\lambda^{k-1}, w^k, v^k)\|^2 + \Gamma_3\alpha_k\eta_k^2\left((\bar{p}_\lambda)^2\sigma^2 + \underline{p}_\lambda\underline{p}_w\sigma_h^2\right) \\
&+ \sum_1^{T+1}\left(\Gamma_3\Gamma_d\alpha_k\alpha_{k-1}^2 + \frac{\alpha_k^2 L_{\bar{F}}}{2} - \frac{\alpha_k}{2}\right)\mathbb{E}\|d^{k-1}\|^2 \\
&+ \Gamma_w\sum_1^{T+1}\alpha_k\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2 + 4\Gamma_3\sum_1^{T+1}\alpha_k\underline{p}_\lambda\underline{p}_w C_G^2 3V^k,
\end{aligned}$$

where $\Gamma_v := 4\Gamma_3\underline{p}_\lambda\underline{p}_w C_G^2 3 + \overline{p}_\lambda\overline{p}_w n C_G^2$, $\Gamma_w := 2\Gamma_3\underline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2\right)3 + \overline{p}_w\left((L_{12}^F)^2 + n L_{G^w}^2 r^2\right)$, and $\Gamma_d := \Gamma_3\overline{p}_\lambda\left(2(L_{12}^F)^2 + 2r^2(1+3L_w^2) + 2C_G^2 3C_w\right)$. Summing

the above inequality with Γ_2 times of (30), we have

$$
\begin{aligned}
&\sum_1^T \frac{\alpha_k}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 + \Gamma_2\left(V^{T+2}-V^1\right)\\
&\le -\mathbb{E}F_{M^{T+1}}(\lambda^{T+1}) + \mathbb{E}F_{M^0}(\lambda^0)\\
&+\Gamma_3\left(\mathbb{E}\|(d^1-\bar{\nabla}\bar{F}(\lambda^0,w^1,v^1)\|^2 - \mathbb{E}\|d^{T+1}-\bar{\nabla}\bar{F}(\lambda^k,w^{T+1},v^{T+1})\|^2\right)\\
&-\Gamma_3\sum_1^T \alpha_k\eta_k\mathbb{E}\|(d^k-\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)\|^2 + \Gamma_3\alpha_k\eta_k^2\left((\bar{p}_\lambda)^2\sigma^2 + \underline{p}_\lambda\underline{p}_w\sigma_h^2\right)\\
&+\sum_1^{T+1}\left(\Gamma_3\Gamma_d\alpha_k\alpha_{k-1}^2 + \frac{\alpha_k^2 L_{\bar{F}}}{2} - \frac{\alpha_k}{2} + \Gamma_2\frac{2}{\beta\underline{p}_w\mu}C_w\alpha_k\right)\mathbb{E}\|d^{k-1}\|^2\\
&+\sum_1^{T+1}\left(\Gamma_w\alpha_k + \Gamma_2\frac{2}{\underline{p}_w\mu}\beta_k L_{ww}^F\bar{p}_w\right)\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2\\
&+\sum_1^{T+1}\left(\Gamma_2\left(-\frac{1}{2}+\frac{1}{4}\beta_k\underline{p}_w\mu\right)\underline{p}_w\mu\beta_k + 4\Gamma_3\underline{p}_\lambda\underline{p}_w C_G^2 3\alpha_k\right)V^k\\
&+\Gamma_2\sum_{k=1}^{T+1}\beta_k^2\left(4\bar{p}_w\sigma^2 + 4\bar{p}_w^2\sigma_h^2 + 4\bar{p}_w^2(1-\underline{p}_w^2)nC_G^2 + 4\bar{p}_w(1-\underline{p}_w)n(C^F)^2\right)
\end{aligned}
$$

Summing this inequality with Γ_1 times of (26), we have

$$
\begin{aligned}
&\sum_1^T \frac{\alpha_k}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 + \Gamma_2\left(V^{T+2}-V^1\right) + \Gamma_1\left(\mathbb{E}\|w^{T+2}-w(\lambda^{T+1})\|^2 - \mathbb{E}\|w^1-w(\lambda^0)\|^2\right)\\
&\le -\mathbb{E}F_{M^{T+1}}(\lambda^{T+1}) + \mathbb{E}F_{M^0}(\lambda^0)\\
&+\Gamma_3\left(\mathbb{E}\|(d^1-\bar{\nabla}\bar{F}(\lambda^0,w^1,v^1)\|^2 - \mathbb{E}\|d^{T+1}-\bar{\nabla}\bar{F}(\lambda^k,w^{T+1},v^{T+1})\|^2\right)\\
&-\Gamma_3\sum_1^T \alpha_k\eta_k\mathbb{E}\|(d^k-\bar{\nabla}\bar{F}(\lambda^{k-1},w^k,v^k)\|^2\\
&+\sum_1^{T+1}\left(\Gamma_3\Gamma_d\alpha_k\alpha_{k-1}^2 + \frac{\alpha_k^2 L_{\bar{F}}}{2} - \frac{\alpha_k}{2} + \Gamma_2\frac{2}{\beta\underline{p}_w\mu}C_w\alpha_k + \Gamma_1\frac{2(1-\gamma_k\underline{p}_w\mu)}{\gamma\underline{p}_w\mu}L_w^2\alpha_k\right)\mathbb{E}\|d^{k-1}\|^2\\
&+\sum_1^{T+1}\left(\Gamma_w\alpha_k + \Gamma_2\frac{2}{\underline{p}_w\mu}\beta_k L_{ww}^F\bar{p}_w - \Gamma_1\frac{1}{2}\gamma_k\underline{p}_w\mu\right)\mathbb{E}\|w^k - w(\lambda^{k-1})\|^2\\
&+\sum_{k=1}^{T+1}\left(\Gamma_3\alpha_k\eta_k^2\bar{p}_\lambda^2 + \Gamma_1\gamma_k^2\bar{p}_w + \Gamma_2\beta_k^2 4\bar{p}_w\right)\sigma^2\\
&+\sum_1^{T+1}\left(\Gamma_2\left(-\frac{1}{2}\underline{p}_w\mu + \frac{1}{4}\beta_k\underline{p}_w^2\mu^2\right)\beta_k + 4\Gamma_3\underline{p}_\lambda\underline{p}_w C_G^2 3\alpha_k\right)V^k\\
&+\sum_{k=1}^{T+1}\left(\Gamma_2\beta_k^2 4\bar{p}_w^2 + \Gamma_3\alpha_k\eta_k^2\underline{p}_\lambda\underline{p}_w\right)\sigma_h^2\\
&+\sum_{k=1}^{T+1}\Gamma_2\beta_k^2\left(4\bar{p}_w^2(1-\underline{p}_w^2)nC_G^2 + 4\bar{p}_w(1-\underline{p}_w)n(C^F)^2\right)
\end{aligned}
\tag{49}
$$

Let $\Gamma_2 := \frac{\beta\underline{p}_w\mu}{16C_w}, \Gamma_1 := \frac{\gamma\underline{p}_w\mu}{16L_w^2}$ be such that $-\frac{\alpha_k}{2} + \Gamma_2\frac{2}{\beta\underline{p}_w\mu}C_w\alpha_k + \Gamma_1\frac{2(1-\gamma_k\underline{p}_w\mu)}{\gamma\underline{p}_w\mu}L_w^2\alpha_k < 0$. Then there exists small δ with $\alpha_k = \frac{\delta}{\sqrt{T}}$ such that $\Gamma_3\Gamma_d\alpha_k\alpha_{k-1}^2 + \frac{\alpha_k^2 L_{\bar{F}}}{2} - \frac{\alpha_k}{2} +$

$\Gamma_2 \frac{2}{\beta \underline{p}_w \mu} C_w \alpha_k + \Gamma_1 \frac{2(1-\gamma_k \underline{p}_w \mu)}{\gamma \underline{p}_w \mu} L_w^2 \alpha_k \leq 0$. Let $\Gamma_3 := \frac{1}{24 \underline{p}_\lambda \underline{p}_w C_G^2}$ and β be small enough such that $\Gamma_2 \left(-\frac{1}{2}\underline{p}_w \mu + \frac{1}{4}\beta_k \underline{p}_w^2 \mu^2\right) \beta_k + 4\Gamma_3 \underline{p}_\lambda \underline{p}_w C_G^2 3\alpha_k \leq 0$. Let γ be big enough such that $\Gamma_w \alpha_k + \Gamma_2 \frac{2}{\underline{p}_w \mu} \beta_k L_{ww}^F \bar{p}_w - \Gamma_1 \frac{1}{2}\gamma_k \underline{p}_w \mu \leq 0$. Then (49) can be further passed to

$$
\begin{aligned}
&\sum_1^T \frac{\alpha_k}{2} \mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 + \Gamma_2 \left(V^{T+2} - V^1\right) + \Gamma_1 \left(\mathbb{E}\|w^{T+2} - w(\lambda^{T+1})\|^2 - \mathbb{E}\|w^1 - w(\lambda^0)\|^2\right) \\
&\leq -\mathbb{E}F_{M^{T+1}}(\lambda^{T+1}) + \mathbb{E}F_{M^0}(\lambda^0) + \Gamma_3 \mathbb{E}\|(d^1 - \bar{\nabla}\bar{F}(\lambda^0, w^1, v^1)\|^2 \\
&+ \sum_{k=1}^{T+1} \left(\Gamma_3 \alpha_k \eta_k^2 \bar{p}_\lambda^2 + \Gamma_1 \gamma_k^2 \bar{p}_w + \Gamma_2 \beta_k^2 4\bar{p}_w\right) \sigma^2 + \sum_{k=1}^{T+1} \left(\Gamma_2 \beta_k^2 4\bar{p}_w^2 + \Gamma_3 \alpha_k \eta_k^2 \underline{p}_\lambda \underline{p}_w\right) \sigma_h^2 \\
&+ \sum_{k=1}^{T+1} \Gamma_2 \beta_k^2 \left(4\bar{p}_w^2 (1 - \underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1 - \underline{p}_w) n (C^F)^2\right) \\
&\leq -\bar{F} + \mathbb{E}F_{M^0}(\lambda^0) + \Gamma_3 \mathbb{E}\|(d^1 - \bar{\nabla}\bar{F}(\lambda^0, w^1, v^1)\|^2 \\
&+ \sum_{k=1}^{T+1} \left(\Gamma_3 \alpha_k \eta_k^2 \bar{p}_\lambda^2 + \Gamma_1 \gamma_k^2 \bar{p}_w + \Gamma_2 \beta_k^2 4\bar{p}_w\right) \sigma^2 + \sum_{k=1}^{T+1} \left(\Gamma_2 \beta_k^2 4\bar{p}_w^2 + \Gamma_3 \alpha_k \eta_k^2 \underline{p}_\lambda \underline{p}_w\right) \sigma_h^2 \\
&+ \sum_{k=1}^{T+1} \Gamma_2 \beta_k^2 \left(4\bar{p}_w^2 (1 - \underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1 - \underline{p}_w) n (C^F)^2\right)
\end{aligned}
\tag{50}
$$

where the second inequality uses the assumption that F is lower bounded by $\bar{F}$. Rearranging the above inequality we have

$$
\begin{aligned}
&\sum_1^T \frac{\alpha_T}{2} \mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \leq \sum_1^T \frac{\alpha_k}{2} \mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \\
&\leq \Gamma_2 V^1 + \Gamma_1 \mathbb{E}\|w^1 - w(\lambda^0)\|^2 - \bar{F} + \mathbb{E}F_{M^0}(\lambda^0) + \Gamma_3 \mathbb{E}\|(d^1 - \bar{\nabla}\bar{F}(\lambda^0, w^1, v^1)\|^2 \\
&+ \sum_{k=1}^{T+1} \left(\Gamma_3 \alpha_k \eta_k^2 \bar{p}_\lambda^2 + \Gamma_1 \gamma_k^2 \bar{p}_w + \Gamma_2 \beta_k^2 4\bar{p}_w\right) \sigma^2 + \sum_{k=1}^{T+1} \left(\Gamma_2 \beta_k^2 4\bar{p}_w^2 + \Gamma_3 \alpha_k \eta_k^2 \underline{p}_\lambda \underline{p}_w\right) \sigma_h^2 \\
&+ \sum_{k=1}^{T+1} \Gamma_2 \beta_k^2 \left(4\bar{p}_w^2 (1 - \underline{p}_w^2) n C_G^2 + 4\bar{p}_w (1 - \underline{p}_w) n (C^F)^2\right)
\end{aligned}
\tag{51}
$$

Using the definition of w_1 in Algorithm 1, we have that

$$
\begin{aligned}
&\mathbb{E}\|w^1\|^2 = \mathbb{E}\|w^0 - \gamma_k \underline{m}_w^0 \nabla_w G(\underline{m}_\lambda^0 \lambda^0, \underline{m}_w^0 w^0; \xi^0)\| \\
&\leq 4\|w^0\|^2 + 4\mathbb{E}\|\gamma_k \underline{m}_w^0 \nabla_w G(\lambda^0, w^0)\|^2 \\
&+ 4\mathbb{E}\|\gamma_k \underline{m}_w^0 \left(\nabla_w G(\underline{m}_\lambda^0 \lambda^0, \underline{m}_w^0 w^0) - \nabla_w G(\lambda^0, w^0)\right)\|^2 + \gamma_k^2 \sigma^2 \\
&\leq 4\|w^0\|^2 + 4\mathbb{E}\|\gamma_k \underline{m}_w^0 \nabla_w G(\lambda^0, w^0)\|^2 \\
&+ 4\mathbb{E}\|\gamma_k \left(\nabla_w G(\underline{m}_\lambda^0 \lambda^0, \underline{m}_w^0 w^0) - \nabla_w G(\lambda^0, w^0)\right)\|^2 + \gamma_k^2 \sigma^2 \\
&\leq 4\|w^0\|^2 + 4\gamma_k^2 \bar{p}_w \|\nabla_w G(\lambda^0, w^0)\|^2 + 4\mathbb{E}\gamma_k^2 L_G^2 \|(\underline{m}_\lambda^0 \lambda^0, \underline{m}_w^0 w^0) - (\lambda^0, w^0)\|^2 + \gamma_k^2 \sigma^2 \\
&\leq 4\|w^0\|^2 + 4\gamma_k^2 \bar{p}_w \|\nabla_w G(\lambda^0, w^0)\|^2 + 4\gamma_k^2 L_G^2 \|(\lambda^0, w^0)\|^2 + \gamma_k^2 \sigma^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbb{E}\|w^1 - w(\lambda^0)\|^2 &\leq 8\|w^0\|^2 + 8\gamma_0^2 \bar{p}_w \|\nabla_w G(\lambda^0, w^0)\|^2 \\
&+ 8\gamma_0^2 L_G^2 \|(\lambda^0, w^0)\|^2 + 2\gamma_0^2 \sigma^2 + 2\|w(\lambda^0)\|^2
\end{aligned}
\tag{52}
$$

For the term V^1, we first bound $\|v^1\|$. By definition,

$$
\begin{aligned}
&\|v^1\| \\
&\leq 3\mathbb{E}\|\beta_k \overline{m}_w^0 \nabla_w F(\overline{m}_\lambda^0 \lambda^0, \overline{m}_w^0 w^1; \eta^0)\|^2 + 3\mathbb{E}\|\underline{m}_w^0 v^0\|^2 + \mathbb{E}\|\beta_k \underline{m}_w^0 \nabla_{ww}^2 G(\lambda^0, w^0; \xi^0) \underline{m}_w^0 v^0\|^2 \\
&\leq 3\mathbb{E}\|\nabla_w F(\overline{m}_\lambda^0 \lambda^0, \overline{m}_w^0 w^1; \eta^0)\|^2 + 3\|v^0\|^2 + 3\|\beta_k \nabla_{ww}^2 G(\lambda^0, w^0; \xi^0) v^0\|^2 \\
&\leq 6\mathbb{E}\|\nabla_w F(\overline{m}_\lambda^0 \lambda^0, \overline{m}_w^0 w^1)\|^2 + 6\sigma^2 + 3\|v^0\|^2 + 6\|\beta_k \nabla_{ww}^2 G(\lambda^0, w^0) v^0\|^2 + 6n\sigma_h^2 \\
&+ 2\frac{1}{\underline{p}_w \mu}\mathbb{E}\|\nabla_w F(\overline{m}_\lambda^0 \lambda^0, \overline{m}_w^0 w(\lambda^0))\|^2 \\
&\leq 6(L_{12}^F)^2\|\lambda^0\|^2 + 6(L_{22}^F)^2\|w^1\|^2 + 6\sigma^2 + 3\|v^0\|^2 + 6\|\beta_k \nabla_{ww}^2 G(\lambda^0, w^0) v^0\|^2 + 12n\sigma_h^2
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\mathbb{E}\|v_{\overline{m}_w^0, \lambda^0}\|^2 \leq \mathbb{E}\| \left(\nabla_{ww}^2 \tilde{G}(\lambda^0, w(\lambda^0))\right)^{-1} \overline{m}_w^0 \nabla_w F(\overline{m}_\lambda^0 \lambda^0, \overline{m}_w^0 w(\lambda^0))\|^2 \\
&\leq \frac{1}{\underline{p}_w \mu}\mathbb{E}\|\nabla_w F(\overline{m}_\lambda^0 \lambda^0, \overline{m}_w^0 w(\lambda^0))\|^2 \leq \frac{1}{\underline{p}_w \mu}\left(12(L_{12}^F)^2\|\lambda^0\|^2 + 12(L_{22}^F)^2\|w(\lambda^0)\|^2\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
V^1 &\leq 12(L_{12}^F)^2\|\lambda^0\|^2 + 12(L_{22}^F)^2\|w^1\|^2 + 12\sigma^2 + 6\|v^0\|^2 + 12\|\beta_k \nabla_{ww}^2 G(\lambda^0, w^0) v^0\|^2 \\
&+ 12n\sigma_h^2 + 2\frac{1}{\underline{p}_w \mu}\left(12(L_{12}^F)^2\|\lambda^0\|^2 + 12(L_{22}^F)^2\|w(\lambda^0)\|^2\right).
\end{aligned} \tag{53}
$$

Using (52), the above inequality can be further passed to

$$
\begin{aligned}
&\mathbb{E}\|V_1\|^2 \leq 12(L_{12}^F)^2\|\lambda^0\|^2 + 12\sigma^2 + 6\|v^0\|^2 + 12\|\beta_k \nabla_{ww}^2 G(\lambda^0, w^0) v^0\|^2 + 12n\sigma_h^2 \\
&+ 2\frac{1}{\underline{p}_w \mu}\left(12(L_{12}^F)^2\|\lambda^0\|^2 + 12(L_{22}^F)^2\|w(\lambda^0)\|^2\right) \\
&+ 12(L_{22}^F)^2\left(4\|w^0 - w(\lambda^0)\|^2 + 4\gamma_k^2 \overline{p}_w \|\nabla_w G(\lambda^0, w^0)\|^2 + 4\gamma_k^2 L_G^2\|(\lambda^0, w^0)\|^2 + \gamma_k^2\sigma^2\right).
\end{aligned}
$$

As for $\mathbb{E}F_{M^0}(\lambda^0)$, using the definition of F_M and Assumption 2, it holds that

$$
\|F_{M^0}(\lambda^0)\| \leq \|f(\lambda^0)\| + C^F\|\lambda^0\| = \|F(\lambda^0, w(\lambda^0))| \tag{54}
$$

Using the definition in (39)

$$
\begin{aligned}
&\|\bar{\nabla}\bar{F}(\lambda^0, w^1, v^1)\|^2 = \|\overline{m}_\lambda^1 \nabla_\lambda F(\overline{m}_\lambda^0 \lambda^0, \overline{m}_w w^1) - \underline{m}_{\lambda^0} \nabla_{\lambda w}^2 G(\underline{m}_\lambda^0 \lambda^0, \underline{m}_w^0 w^1) \underline{m}_w^0 v^1\|^2 \\
&\leq L_F^2 + C_G^2 r^2
\end{aligned} \tag{55}
$$

and

$$
\begin{aligned}
\|d^1\|^2 &= \|\nabla\bar{F}(\lambda^0, w^1, v^1) + (1-\eta_0)(d^0 - \nabla\bar{F}(\lambda^0, w^0, v^{-1})))\|^2 \\
&\leq \|\overline{m}_\lambda^k \nabla_\lambda F(\overline{m}_\lambda^0 \lambda^0, \overline{m}_w w^1; \eta^0) - \underline{m}_{\lambda^0} \nabla_{\lambda w}^2 G(\underline{m}_\lambda^0 \lambda^0, \underline{m}_w^0 w^1; \xi^0) \underline{m}_w^k v^1\| \\
&+ \|(1-\eta_0)(d^0 - \nabla\bar{F}(\lambda^0, w^0, v^{-1})))\|^2 \\
&\leq (C^F)^2 + 2C_G^2 r^2 + \|(1-\eta_0)(d^0 - \nabla\bar{F}(\lambda^0, w^0, v^{-1})))\|^2
\end{aligned} \tag{56}
$$

Combining (53), (54), (55), (56) , we know that there exists I_0 such that (58) ca be further passed to

$$
\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{1}^{T}\frac{1}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \leq \frac{1}{\underline{p}_w\underline{p}_\lambda}I_0 \\
&+\sum_{k=1}^{T+1}\left(\Gamma_3\alpha_k\eta_k^2\bar{p}_\lambda^2+\Gamma_1\gamma_k^2\bar{p}_w+\Gamma_2\beta_k^2 4\bar{p}_w\right)\sigma^2+\sum_{k=1}^{T+1}\left(\Gamma_2\beta_k^2 4\bar{p}_w^2+\Gamma_3\alpha_k\eta_k^2\underline{p}_\lambda\underline{p}_w\right)\sigma_h^2 \\
&+\sum_{k=1}^{T+1}\Gamma_2\beta_k^2\left(4\bar{p}_w^2(1-\underline{p}_w^2)nC_G^2+4\bar{p}_w(1-\underline{p}_w)n(C^F)^2\right) \\
&\leq \frac{1}{\underline{p}_w\underline{p}_\lambda}I_0+\ln(T+1)\left(\Gamma_3\delta\eta\bar{p}_\lambda^2+\Gamma_1\gamma^2\delta^2\bar{p}_w+\Gamma_2\beta^2\delta^2 4\bar{p}_w\right)\sigma^2 \\
&+\ln(T+1)\left(\Gamma_2\beta^2\delta^2 4\bar{p}_w^2+\Gamma_3\delta\eta^2\underline{p}_\lambda\underline{p}_w\right)\sigma_h^2 \\
&+\ln(T+1)\Gamma_2\beta^2\delta^2\left(4\bar{p}_w^2(1-\underline{p}_w^2)nC_G^2+4\bar{p}_w(1-\underline{p}_w)n(C^F)^2\right).
\end{aligned}
\tag{57}
$$

Deviding both sides by $\sqrt{T}$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{1}^{T}\frac{1}{2}\mathbb{E}\|\nabla F_{M^k}(\lambda^k)\|^2 \\
&\leq \frac{1}{\sqrt{T}}\frac{1}{\underline{p}_w\underline{p}_\lambda}I_0+\frac{\ln(T+1)}{\sqrt{T}}\left(\Gamma_3\delta\eta\bar{p}_\lambda^2+\Gamma_1\gamma^2\delta^2\bar{p}_w+\Gamma_2\beta^2\delta^2 4\bar{p}_w\right)\sigma^2 \\
&+\ln(T+1)\left(\Gamma_2\beta^2\delta^2 4\bar{p}_w^2+\Gamma_3\delta\eta^2\underline{p}_\lambda\underline{p}_w\right)\sigma_h^2 \\
&+\frac{\ln(T+1)}{\sqrt{T}}\Gamma_2\beta^2\delta^2\left(4\bar{p}_w^2(1-\underline{p}_w^2)nC_G^2+4\bar{p}_w(1-\underline{p}_w)n(C^F)^2\right) \\
&=\frac{1}{\sqrt{T}}\frac{1}{\underline{p}_w\underline{p}_\lambda}I_0 \\
&+\frac{\ln(T+1)}{\sqrt{T}}O\left(\frac{\bar{p}_\lambda^2}{\underline{p}_\lambda\underline{p}_w}+\underline{p}_w\bar{p}_w\right)\left(\sigma^2+\sigma_h^2\right) \\
&+\frac{\ln(T+1)}{\sqrt{T}}O\left(\underline{p}_w\left(\bar{p}_w^2(1-\underline{p}_w^2)+\bar{p}_w(1-\underline{p}_w)\right)\right).
\end{aligned}
\tag{58}
$$

This completes the proofs. □