A First Order Meta Stackelberg Method for Robust Federated Learning

Yunian Pan [*1] **Tao Li** [*1] **Henger Li** [2] **Tianyi Xu** [2] **Zizhan Zheng** [2] **Quanyan Zhu** [1]

Abstract

Previous research has shown that federated learning (FL) systems are exposed to an array of security risks. Despite the proposal of several defensive strategies, they tend to be non-adaptive and specific to certain types of attacks, rendering them ineffective against unpredictable or adaptive threats. This work models adversarial federated learning as a Bayesian Stackelberg Markov game (BSMG) to capture the defender's incomplete information of various attack types. We propose meta-Stackelberg learning (meta-SL), a provably efficient meta-learning algorithm, to solve the equilibrium strategy in BSMG, leading to an adaptable FL defense. We demonstrate that meta-SL converges to the first-order ε-equilibrium point in $O(\varepsilon^{-2})$ gradient iterations, with $O(\varepsilon^{-4})$ samples needed per iteration, matching the state of the art. Empirical evidence indicates that our meta-Stackelberg framework performs exceptionally well against potent model poisoning and backdoor attacks of an uncertain nature.

1. Introduction

Federated learning (FL) provides a way for several devices possessing private data to collaboratively train a learning model without the need to share their local data (McMahan et al., 2017). Nonetheless, FL systems remain susceptible to antagonistic attacks, including untargeted model poisoning and specific backdoor attacks. To counter these vulnerabilities, a range of robust aggregation techniques like Krum (Blanchard et al., 2017), coordinate-wise median (Yin et al., 2018), trimmed mean (Yin et al., 2018), and FLTrust (Cao et al., 2021) have been suggested for defense against non-specific attacks. Furthermore, different post-training protective measures like Neuron Clipping (Wang et al., 2022) and Pruning (Wu et al., 2020) have been recently introduced to reduce the impact of backdoor attacks.

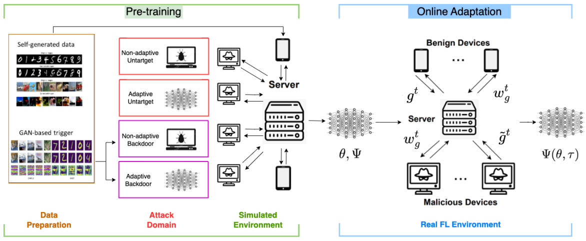


Figure 1. A schematic illustration of the meta-Stagberg game framework. In the pertaining stage, a simulated environment is constructed using generated data and a set of attacks sampled from the attack domain consisting of various attack strategies. The defender utilizes meta-Stackelberg learning (Algorithm 2) to obtain the meta policy θ and the adaptation Ψ in (2). Then, in the online execution, the defender can adapt its defense to $\Psi(\theta, \tau)$ using received feedback τ in the presence of unknown attacks.

Existing defenses are typically built to resist specific attack types and attacks that do not evolve in response to defensive measures. In this work, we introduce a meta-Stackelberg game (meta-SG) framework that delivers robust defensive performance, even against adaptive attacks such as the reinforcement learning (RL)-based attack (Li et al., 2022a), which current state-of-the-art defenses struggle to address, or an amalgamation of different attack types like the simultaneous occurrence of model poisoning and backdoor attacks (see Section 5).

Our meta-SG defense framework is established on several key observations. Firstly, the issue of robust FL in the face of a non-adaptive attack can be perceived as a Markov decision process (MDP), where the state represents model updates from selected devices, and the action refers to the gradient for updating the global model. Moreover, when the attack is known beforehand, the defender can employ the limited amount of local data at the server and publicly accessible data to construct an (approximate) MDP model and determine a robust defense policy by pre-training prior to the commencement of FL training. Secondly, for situations where the attack is adaptive but with specific parameters, we consider a Markov game between the attacker and the defender and establish a robust defense by solving the Stackelberg equilibrium of the game, wherein the defender is the leader and the attacker the follower. This approach applies

[*] Equal contribution [1] Department of Electrical and Computer Engineering, New York, NY, USA [2] Department of Computer Science, Tulane University, New Orleans, LA, USA. Correspondence to: Tao Li <taoli@nyu.edu>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

to both single and multiple concurrent attacks and may yield an (almost) optimal defense. Thirdly, in more realistic scenarios where attacks are unknown or uncertain, this situation can be treated as a Bayesian Stackelberg Markov game (BSMG), offering a comprehensive model for adversarial FL. Uncertain attacks refer to those involved in the pre-training stage but undisclosed in the online FL process, leaving the defender unsure about their existence. On the other hand, unknown attacks point to those excluded in the pre-training, of which the defender is unaware. Nonetheless, the standard solution concept for BSMG, the Bayesian Stackelberg equilibrium, aims at the expected case and does not adapt to the actual attack.

In this study, we introduce a novel solution concept, the meta-Stackelberg equilibrium (meta-SE), for BSMG as a systematic approach to creating resilient and adaptive defenses for federated learning. By merging meta-learning with Stackelberg reasoning, meta-SE provides a computationally efficient method to handle information asymmetry in adversarial FL and facilitates strategic adaptation during online execution amidst multiple (adaptive) attackers. Prior to training an FL model, a meta-policy is trained by solving the BSMG using experiences sampled from a set of potential attacks. During FL training, when confronted with an actual attacker, the meta-policy rapidly adapts using a relatively small batch of samples gathered in real-time. Importantly, our proposed Meta-SG framework only requires a rough estimate of potential attacks during meta-training due to the generalization capability offered by meta-learning.

To solve the BSMG in the pre-training stage, we develop a meta-Stackelberg learning (meta-SL) algorithm, based on the concept of debiased meta-reinforcement learning (Fallah et al., 2021). Meta-SL is proven to converge to the first-order ε-approximate meta-SE in $O(\varepsilon^{-2})$ iterations, and the corresponding sample complexity per iteration is $O(\varepsilon^{-4})$. Such algorithmic complexity aligns with the latest complexity results in nonconvex bi-level stochastic optimization (Ji et al., 2021). Due to the conflicting interests between the defender and the attacker in FL, the ensuing BSMG is strictly competitive, which can be seen as a generalization of zero-sum. Therefore, meta-SL does not require second-order derivatives of the attacker's value function (the low-level problem), even though the Hessian of the defender's value function remains due to the meta adaptation. Inspired by Reptile (Nichol et al., 2018), a first-order meta-learning algorithm, we propose a fully first-order pre-training algorithm, referred to as Reptile Meta-SL, as a substitute for meta SL in our experiments. Reptile Meta-SL uses only the first-order stochastic gradients of the attacker's and defender's objective functions to solve for the approximate equilibrium. As evidenced by numerical results in Section 5 and Appendix, it is effective in managing adaptive and/or uncertain (or unknown) attacks.

Our contributions can be summarized as follows, with the discussion of related work relocated to the Appendix due to space constraints:

- We tackle vital security issues in federated learning in the face of multiple adaptive (non-adaptive) attackers of uncertain or unknown types.
- We formulate a Bayesian Stackelberg game model (Section 2.2) to encapsulate the information asymmetry in adversarial FL under uncertain or unknown adaptive attacks.
- To provide the defender with strategic adaptability, we introduce a new equilibrium concept, the meta-Stackelberg equilibrium (Definition 3.1). Here, the defender (the leader) commits to a meta policy and an adaptation strategy, leading to a data-driven method to handle information asymmetry.
- To learn the meta equilibrium defense during the pre-training phase, we develop meta-Stackelberg learning (Algorithm 2), an efficient first-order meta RL algorithm. This algorithm provably converges to ε-approximate equilibrium in $O(\varepsilon^{-2})$ gradient steps with $O(\varepsilon^{-4})$ samples per iteration, matching the state-of-the-art efficiency in stochastic bilevel optimization.
- We carry out comprehensive experiments in real-world scenarios to demonstrate the outstanding performance of our proposed method.

2. Model Formulation

2.1. Federated Learning and Threat Model

FL objective. Consider a learning system that includes one server and n clients, each client possesses its own private dataset $D_i = (x_i^j, y_i^j)_{j=1}^{|D_i|}$ and $|D_i|$ signifies the size of the dataset for the i-th client. Let $U = \{D_1, D_2, \dots, D_n\}$ represent the compilation of all client datasets. The objective of federated learning is defined as identifying a model w that minimizes the average loss across all the devices: $\min_w F(w, U) := \frac{1}{n}\sum_{i=1}^{n} f(w, D_i)$, where $f(w, D_i) := \frac{1}{|D_i|}\sum_{j=1}^{|D_i|} \ell(w, (x_i^j, y_i^j))$ is the local empirical loss with $\ell(\cdot,\cdot)$ being the loss function.

Attack objective. We consider two major categories of attacks, namely, backdoor attacks and untargeted model poisoning attacks. Our framework can be extended to other attack scenarios. For simplicity, assume that the first M_1 malicious clients carry out the backdoor attack and the following M_2 malicious clients undertake the poisoning attack. The model poisoning attack aims to maximize the average model loss, i.e., $\max_w F(w)$; the backdoor attack aims to preserve decent performance on clean test inputs ("main task") while causing misclassification of poisoned test inputs to one or more target labels ("backdoor task"). Each malicious client in the backdoor attack produces a poisoned data set $D'_{i\leq M_1}$, obtained by altering a subset of data samples $(x_i^j, y_i^j) \in D_i$ to $(\hat{x}_i^j, c^*)$, where $\hat{x}_i^j$ is the tainted sample with a backdoor trigger inserted, and $c^* \neq y_i^j, c^* \in C$ is the targeted label. Let $U' = \{D'_1, D'_2, \dots, D'_{M_1}\}$ denote the compilation of poisoned datasets. The objective function in the backdoor attack is defined as: $\min_w F'(w) = \lambda F(w, U) + (1-\lambda) F(w, U')$, where $\lambda \in [0, 1]$ serves to balance between the main task and the backdoor task.

FL process. The federated learning process works in an adversarial setting as follows. At each round t out of H FL rounds, the server randomly selects a subset of clients $\mathcal{S}^t$ and sends them the most recent global model w_g^t. Every benign client in $\mathcal{S}^t$ updates the model using their local data via one or more iterations of stochastic gradient descent and returns the model update g^t to the server. Conversely, an adversary in $\mathcal{S}^t$ creates a malicious model update $\widetilde{g}^t$ clandestinely and sends it back. The server then collects the set of model updates $\{\widetilde{g}_i^t \cup \widetilde{g}_j^t \cup g_k^t\}_{i,j,k \in \mathcal{S}^t, i \in [M_1], j \in [M_2], k \notin [M_1] \cup [M_2]}$, utilizing an aggregation rule $Aggr$ to combine them and updates the global model $w_g^{t+1} = w_g^t - Aggr(\widetilde{g}_i^t \cup \widetilde{g}_j^t \cup g_k^t)$, which is then sent to clients in round $t+1$. At the final round T, the server applies a post-training defense $h(\cdot)$ on the global model to generate the final global model $\widehat{w}_g^T = h(w_g^T)$.

Attacker type and behavior. We anticipate multiple types of attacks occurring simultaneously, emanating from various categories. For clarity, we consider a single mastermind attacker present within the FL system who controls a group of malicious clients employing diverse attack strategies, which may be either non-adaptive or adaptive. Non-adaptive attacks involve a fixed attack strategy that solves a short-sighted optimization problem, disregarding the defense mechanism implemented by the server (i.e., the robust aggregation rule and the post-training defense). Such attacks include explicit boosting (EB) (Bhagoji et al., 2019), inner product manipulation (IPM) (Xie et al., 2020), and local model poisoning attack (LMP) (Fang et al., 2020)), federated backdoor attack (BFL) (Bagdasaryan et al., 2020), distributed backdoor attack (DBA) (Xie et al., 2019), projected gradient decent (PGD) backdoor attack (Wang et al., 2020). On the other hand, an adaptive attack, such as the RL-based model poisoning attack (Li et al., 2022a) and RL-based backdoor attack (Li et al., 2023a), designs model updates by simulating the server's reactions to optimize a long-term objective. These adaptive and non-adaptive untargeted/backdoor attack methods mentioned above constitute the attack domain in our adversarial FL setting. One significant hurdle in addressing adversarial attacks is the *information asymmetry* (Li et al., 2022d). This is when the server (i.e., the defender) lacks knowledge of the behavior and identities of malicious clients in a realistic black-box scenario. We denote the collective attack configuration of malicious clients as the type of the mastermind attacker, detailing M_1, M_2, attack behaviors (adaptive or not), and other required parameters of the attack method.

2.2. Bayesian Stackelberg Markov Game Model

In this study, we propose a comprehensive framework for robust defense against potent unknown or uncertain attacks. The central principle is to construct RL-based defenses by simulating unknown attack behavior using RL-based attacks. As demonstrated in prior research (Li et al., 2022a; 2023a), RL-based attacks serve as a robust baseline for both model poisoning and backdoor attacks. Therefore, a defense that is resilient to RL-based attacks could potentially safeguard the system against other (less potent) attacks. To manage the high-dimensional state and action spaces, we integrate a set of lightweight defenses in the training stage and post-training stage. The groundbreaking element of our approach is the use of RL to optimize these defenses, moving away from the conventional fixed and manually-tuned hyperparameters. This approach requires a Bayesian Stackelberg Markov game formulation, encapsulated in the tuple $G = (\mathcal{P}, Q, S, O, A, \mathcal{T}, r, \gamma)$, where $\gamma \in (0, 1)$ is the reward discounting factor:

- The player set $\mathcal{P} = \{\mathcal{D}, \mathcal{A}\}$ contains $\mathcal{D}$ as the leader (defender), and $\mathcal{A}$ as the follower (attacker) who controls multiple malicious clients.
- $Q(\cdot) : \Xi \to [0, 1]$ denotes the probability distribution over the attacker's private types. $\Xi := \{\xi_i\}_{i=1}^{|\Xi|}$ where ξ_i denotes i-th type attacks.
- O is the observation space; the observation for the server (i.e., defender) at round t is w_g^t (the server does not have access to the client's identities); the observation for the attacker at round t is $s^t := (w_g^t, \mathbf{I}^t)$ since the attacker controls these malicious clients. $\mathbf{I}^t \in \{0, 1\}^{|\mathcal{S}^t|}$ is the identity vector for the random client subset $\mathcal{S}^t \subseteq \{1, \dots, n\}$, where the identities of malicious and benign devices are 1 and 0 respectively. Notice that the clients' identities are independent of players' actions.
- $A = \{A_{\mathcal{D}}, A_\xi\}$ is the joint action set, where $A_{\mathcal{D}}$ and A_ξ denote the set of defense actions and type-ξ attack actions, respectively; in the FL setting, $a_{\mathcal{D}}^t = \widehat{w}_g^{t+1} := h(w_g^{t+1})$. The attacker's action is characterized by the joint actions of malicious clients $a_{A_\xi}^t := \{\widetilde{g}_i^t\}_{i=1}^{M_1} \cup \{\widetilde{g}_i^t\}_{i=M_1+1}^{M_2}$. Note that a malicious device not sampled at round t does not send any information to the server; hence its action has no effect on the model update. The subscript ξ is suppressed if it is clear from the context.
- $\mathcal{T} : S \times A \to \Delta(S)$ is the state transition function, which represents the probability of reaching a state $s' \in S$ from current state $s \in S$, where the defender and the attacker chose actions $a_{\mathcal{D}}^t$ and $a_{A_\xi}^t$ respectively.
- $r = \{r_{\mathcal{D}}, r_{\mathcal{A}_\xi}\}$, where $r_{\mathcal{D}} : S \times A \to \mathbb{R}_{\leq 0}$ and $r_{\mathcal{A}_\xi} : S \times A \to \mathbb{R}$ are the reward functions for the defender and the attacker, respectively. Define the expected reward at round t as $r_{\mathcal{D}}^t := -\mathbb{E}[F(\widehat{w}_g^{t+1})]$ and $r_{\mathcal{A}_\xi}^t := \rho \mathbb{E}[F'(\widehat{w}_g^{t+1})] - (1-\rho)\mathbb{E}[F(\widehat{w}_g^{t+1})]$, $\rho = M_1/(M_1 + M_2)$, if $\mathbf{1} \cdot \mathbf{I}^t > 0$, and $r_{\mathcal{A}_\xi}^t := 0$ otherwise.

3. Meta-Stackelberg Equilibrium

Let the defender's and the attacker's policies be parameterized by neural networks $\pi_{\mathcal{D}}(a^t_{\mathcal{D}}|s^t;\theta)$, $\pi_{\mathcal{A}}(a^t_{\mathcal{A}}|s^t;\phi)$ with model weights $\theta \in \Theta$ and $\phi \in \Phi$, respectively. Given the two players' policies θ, ϕ and the private attack type ξ, the defender's expected utility is defined as $J_{\mathcal{D}}(\theta,\phi,\xi) := \mathbb{E}_{a^t_{\mathcal{A}}\sim\pi_{\mathcal{A}}(\cdot;\phi,\xi),a^t_{\mathcal{D}}\sim\pi_{\mathcal{D}}(\cdot;\theta)}[\sum_{t=1}^{H}\gamma^t r_{\mathcal{D}}(s^t,a^t_{\mathcal{D}},a^t_{\mathcal{A}})]$. Similarly, the attacker's expected utility is $J_{\mathcal{A}}(\theta,\phi,\xi) := \mathbb{E}_{a^t_{\mathcal{A}}\sim\pi_{\mathcal{A}}(\cdot;\phi,\xi),a^t_{\mathcal{D}}\sim\pi_{\mathcal{D}}(\cdot;\theta)}[\sum_{t=1}^{H}\gamma^t r_{\mathcal{A}}(s^t,a^t_{\mathcal{D}},a^t_{\mathcal{A}})]$. Denote by $\tau_\xi := (s^k,a^k_{\mathcal{D}},a^k_{\mathcal{A}})_{k=1}^{H}$ the trajectory of the BSMG under type-ξ attacker, which is subject to the distribution $q(\theta,\phi,\xi) := \prod_{t=1}^{H}\pi_{\mathcal{D}}(a^t_{\mathcal{D}}|s^t;\theta)\pi_{\mathcal{A}}(a^t_{\mathcal{A}}|s^t;\phi,\xi)\mathcal{T}(s^{t+1}|s^t,a^t_{\mathcal{D}},a^t_{\mathcal{A}})$. In the later development of meta-SG, we consider the gradient $\nabla_\theta J_{\mathcal{D}}(\theta,\phi,\xi)$ and its sample estimate $\hat{\nabla}_\theta J_{\mathcal{D}}(\tau_\xi)$ based on the trajectory τ_ξ. The estimation is due to the policy gradient theorem (Sutton et al., 2000) reviewed in Appendix D, and we note that such an estimate takes a batch of τ_ξ (the batch size is N_b) for variance reduction. For simplicity, we use the one-trajectory estimate denoted by $\hat{\nabla}_\theta J_{\mathcal{D}}(\tau_\xi)$.

A natural defense strategy to tackle the information asymmetry is to find a Bayesian Stackelberg equilibrium (BSE):

$$\begin{aligned}&\max_{\theta\in\Theta}\mathbb{E}_{\xi\sim Q(\cdot)}[J_{\mathcal{D}}(\theta,\phi^*_\xi,\xi)]\\ &\text{s.t. } \phi^*_\xi \in \arg\max J_{\mathcal{A}}(\theta,\phi,\xi), \forall\xi\in\Xi.\end{aligned} \tag{1}$$

(1) admits a simple characterization for optimal defense, yet its limitation is evident. The attacker's actions (equivalently, the aggregated models) reveal partial information about its hidden type (its attack objective), which the defender does not properly handle, as the strategy is fixed throughout the BSMG. Consequently, the defender does not adapt to the specific attacker in the online execution.

To equip the defender with responsive intelligence in the face of unknown multi-type attacks, we propose a new equilibrium concept, meta-Stackelberg equilibrium in Definition 3.1. The intuition of this meta-equilibrium is that $\Psi(\theta,\tau_\xi)$ is tailored to each realized ξ when the defender observes the attacker's moves included in τ_ξ.

Definition 3.1 (Meta Stackelberg Equilibrium)**.** A triple of the defender's meta policy θ, the adaptation mapping Ψ, and the attacker's type-dependent policy ϕ is a meta Stackelberg equilibrium if it satisfies

$$\begin{aligned}&\max_{\theta\in\Theta,\Psi} V(\theta) := \mathbb{E}_{\xi\sim Q}\mathbb{E}_{\tau\sim q}[J_{\mathcal{D}}(\Psi(\theta,\tau),\phi^*_\xi,\xi)],\\ &\text{s.t. } \phi^*_\xi \in \arg\max \mathbb{E}_{\tau\sim q}J_{\mathcal{A}}(\Psi(\theta,\tau),\phi,\xi), \forall\xi\in\Xi,\end{aligned} \tag{2}$$

where $q = q(\theta,\phi,\xi)$ is the trajectory distribution.

In practice, $\Psi(\theta,\tau)$ is simply fixed as a one-step (or multi-step, see Appendix C) SGD operation, i.e., $\Psi(\theta,\tau) = \theta + \eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau)$ to leave θ as the only variable to be optimized. In comparison with meta-defense and BSE-defense, the proposed meta-SE defense highlights **strategic adaptation** in adversarial FL modeled by the BSMG. A detailed discussion on this meta-equilibrium is deferred to Appendix B.

4. Meta-Stackelberg Learning

Based on the aforementioned meta-Stackelberg equilibrium, we introduce the meta-learning-based defense approach (Li & Zheng, 2023) (referred to as the meta-defense in the sequel) by considering non-adaptive attack methods. The goal of meta-defense is to find a meta-policy and an adaptation rule such that the adapted policy gives satisfying defense performance. The mathematical characterization is presented in Appendix C.

The meta-defense framework includes three stages: **pre-training**, **online adaptation**, and **post-training**. The **pre-training** stage is implemented in a simulated environment (as discussed in our technical report), which allows sufficient alternative training with trajectories generated from random potential attacks, which includes both adaptive (e.g., RL-based attacks as discussed in our technical report) and non-adaptive (e.g., IPM and LMP) attacks. After obtaining a meta-policy, the defender will interact with the real FL environment in the **online adaptation** stage to tune its defense policy using feedback (i.e., rewards) received in the face of real attacks. In real-world FL training, the server typically waits for $1 \sim 10$ minutes before receiving responses from the clients, allowing the defender to update the defense policy using SGD in the interim. In the **post-training** stage, the defender finally performs a post-training defense on the global model.

4.1. Robust FL via Meta-Stackelberg Learning

Even though meta-learning enables an adaptable defense, it fails to address adaptative attacks in the online adaptation phase, e.g., the RL attacker learns to evade the adapted defense. To create **strategic** online adaptation, the defender needs to learn the Stackelberg equilibrium of the BSMG using meta-learning, preparing itself for the worst possible case: adaptive attacks in the online phase.

Inspired by a first-order meta-learning algorithm called Reptile (Nichol et al., 2018), we propose Reptile meta-Stackelberg learning (meta-SL) defense (Algorithm 2 solving for the BSMG) in adversarial FL. The algorithms start from an initial meta-defense policy model θ^0 and a set of initial attack policies $\{\phi_i\}_{i\leq m}$ from the attack domain. Meta-SL alternatively updates the defender's and attacker's policies, and each iteration comprises three major steps. **Step 1 attack sampling**: in each iteration, the algorithm first sample a batch of K attacks from $Q(\xi)$. **Step 2 policy adaptation**: for each attack, meta-SL adapts the defender's policy θ^t_k to the k-th attack by performing l-step gradient descent to the defender's value function (see Appendix C) with step size η. **Step 3 meta policy update**: meta-SL first derives the attacker's best response policy for each sampled type by applying gradient descent to the attacker's value function until convergence. After this alternative learning, the algorithm then performs l-step gradient descent to the defender's value function under the convergent attack policy (of each type) for adaptation. Finally, the average of all adapted policies of the sampled attacks becomes the new meta policy.

4.2. First-order Approximate Meta Equilibrium

Now we unfold the theoretical analysis for the **pre-training** stage, which we refer to as meta-Stackelberg learning (meta-SL). The main task of meta-SL is solving (2), a bi-level optimization problem. We employ a bi-level approach, applying gradient ascent to the upper-level problem (the defender's) where the gradient estimation involves the optimizer of the lower-level problem (the attacker's). The details are deferred to Appendix C.

In general, the meta-SE (Definition 3.1) may not be feasible (Nouiehed et al., 2019), we hereby propose a weaker characterization that only involves the first-order necessary conditions. To simplify our exposition, we let $\mathcal{L}_{\mathcal{D}}(\theta,\phi,\xi) := \mathbb{E}_{\tau\sim q}J_{\mathcal{D}}(\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau),\phi,\xi)$ and $\mathcal{L}_{\mathcal{A}}(\theta,\phi,\xi) := \mathbb{E}_{\tau\sim q}J_{\mathcal{A}}(\theta+\hat{\nabla}_\theta J_{\mathcal{D}}(\tau),\phi,\xi)$, for a fixed type $\xi\in\Xi$.. In the sequel, we will assume $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{A}}$ to be continuously twice differentiable and Lipschitz-smooth with respect to both θ and ϕ as in (Li et al., 2022b), and the Lipschitz assumptions are deferred to Appendix D.

Definition 4.1 (ε-meta First-Order Stackelberg Equilibrium)**.** For a small $\varepsilon\in(0,1)$, a set of parameters $(\theta^*,\{\phi^*_\xi\}_{\xi\in\Xi})\in\Theta\times\Phi^{|\Xi|}$ is a ε-*meta First-Order Stackelbeg Equilibrium* (ε-meta-FOSE) of the meta-SG if it satisfies the following conditions for $\xi\in\Xi$,

$$\begin{aligned}\max_{\theta\in\Theta\bigcap B(\theta^*)}\langle\nabla_\theta\mathcal{L}_{\mathcal{D}}(\theta^*,\phi^*_\xi,\xi),\theta-\theta^*\rangle\leq\varepsilon,\\ \max_{\phi\in\Phi\bigcap B(\phi^*_\xi)}\langle\nabla_\phi\mathcal{L}_{\mathcal{A}}(\theta^*,\phi^*_\xi,\xi),\phi-\phi^*_\xi\rangle\leq\varepsilon,\end{aligned}\tag{3}$$

where $B(\theta^*):=\{\theta\in\Theta:\|\theta-\theta^*\|\leq1\}$, and $B(\phi^*_\xi):=\{\theta\in\Theta:\|\phi-\phi^*_\xi\|\leq1\}$. When $\varepsilon=0$, the parameter set $(\theta^*,\{\phi^*_\xi\}_{\xi\in\Xi})$ is said to be the meta-FOSE.

The necessary equilibrium condition for Definition 3.1 can be reduced to $\|\nabla_\theta\mathcal{L}_{\mathcal{D}}(\theta^*,\phi_\xi,\xi)\|\leq\varepsilon$ and $\|\nabla_\phi\mathcal{L}_{\mathcal{A}}(\theta^*,\phi_\xi,\xi)\|\leq\varepsilon$ in the unconstraint settings. Since we utilize stochastic gradient in practice, all inequalities mentioned above shall be considered in expectation. These conditions, along with the positive-semi-definiteness of the Hessians, construct the optimality conditions for a local solution for the meta-SE, which may not exist even in the zero-sum cases (Jin et al., 2019). The advantage of considering meta-FOSE is that its existence is guaranteed.

Theorem 4.2. *Under the condition that Θ and Φ are compact and convex, the meta-SG admits at least one meta-FOSE.*

For the rest of this section, we assume the attacker is unconstrained, i.e., Φ is a finite-dimensional Euclidean space.

4.3. Sufficiency for First-Order Estimation in Strictly Competitive Games

Finding a meta-FOSE for (2) is challenging due to the non-convex equilibrium constraint at the lower level. To see this more clearly, consider differentiating the defender's value function: $\nabla_\theta V=\mathbb{E}_{\xi\sim Q}[\nabla_\theta\mathcal{L}_{\mathcal{D}}(\theta,\phi_\xi,\xi)+(\nabla_\theta\phi_\xi(\theta))^\top\nabla_\phi\mathcal{L}_{\mathcal{D}}(\theta,\phi_\xi,\xi)]$, where $\nabla_\theta\phi_\xi(\cdot)$ is locally characterized by the implicit function theorem, i.e., $\nabla_\theta\phi_\xi(\theta)=(-\nabla^2_\phi\mathcal{L}_{\mathcal{A}}(\theta,\phi,\xi))^{-1}\nabla^2_{\phi\theta}\mathcal{L}_{\mathcal{A}}(\theta,\phi,\xi)$. Therefore, the gradient estimation requires iteratively estimating the second-order information for the attacker (lower level) objective, which can be costly and prohibitive in many scenarios (Song et al., 2019). Hence, we introduce the following assumption to bypass the technicality involved in calculating $\nabla_\theta\phi_\xi$.

Assumption 4.3 (Strict-Competitiveness)**.** The BSMG is strictly competitive, i.e., there exist constants $c<0$, d such that $\forall\xi\in\Xi$, $s\in S$, $a_{\mathcal{D}},a_{\mathcal{A}}\in A_{\mathcal{D}}\times A_\xi$, $r_{\mathcal{D}}(s,a_{\mathcal{D}},a_{\mathcal{A}})=cr_{\mathcal{A}}(s,a_{\mathcal{D}},a_{\mathcal{A}})+d$.

The above assumption is a direct extension of the strict-competitiveness (SC) notion in matrix games (Adler et al., 2009). One can treat the SC notion as a generalization of zero-sum games: if one joint action $(a_{\mathcal{D}},a_{\mathcal{A}})$ leads to payoff increases for one player, it must decrease the other's payoff. In adversarial FL, the untargeted attack naturally makes the game zero-sum (hence, SC), and the backdoor attack also leads to the SC (see Appendix D). The purpose of introducing Assumption 4.3 is to establish the Danskin-type result (Bernhard & Rapaport, 1995) for the Stackelberg game with nonconvex value functions (see Lemma 4.5), which spares us from the Hessian inversion.

Another key regularity assumption we impose on the nonconvex value functions is adapted from the Polyak-Łojasiewicz (PL) condition (Karimi et al., 2016), which is customary in nonconvex analysis.

Assumption 4.4 (Stackelberg Polyak-Łojasiewicz condition)**.** There exists a positive constant μ such that for any $(\theta,\phi)\in\Theta\times\Phi$ and $\xi\in\Xi$, the following inequalities hold: $\frac{1}{2\mu}\|\nabla_\phi\mathcal{L}_{\mathcal{D}}(\theta,\phi,\xi)\|^2\geq\max_\phi\mathcal{L}_{\mathcal{D}}(\theta,\phi,\xi)-\mathcal{L}_{\mathcal{D}}(\theta,\phi,\xi)$, $\frac{1}{2\mu}\|\nabla_\phi\mathcal{L}_{\mathcal{A}}(\theta,\phi,\xi)\|^2\geq\max_\phi\mathcal{L}_{\mathcal{A}}(\theta,\phi,\xi)-\mathcal{L}_{\mathcal{A}}(\theta,\phi,\xi)$.

Under Assumption 4.4, the first-order estimation is sufficient by Lemma 4.5.

Lemma 4.5. *Under Assumptions 4.4 and regularity conditions, there exists $\{\phi_\xi : \phi_\xi \in \arg\max_\phi \mathcal{L}_\mathcal{A}(\theta, \phi, \xi)\}_{\xi \in \Xi}$, such that $\nabla_\theta V(\theta) = \nabla_\theta \mathbb{E}_{\xi \sim Q, \tau \sim q} J_\mathcal{D}(\theta + \eta \hat{\nabla}_\theta J_\mathcal{D}(\tau), \phi_\xi, \xi)$. Moreover, there exists a constant $L > 0$ such that the defender value function $V(\theta)$ is L-Lipschitz-smooth.*

4.4. Non-Asymptotic Iteration Complexity

We now present the main iteration complexity results. Lemma 4.6 states that one can stabilize the lower-level simulated RL attacks under mild conditions, with proper choices of the batch size ($\mathcal{O}(\varepsilon^{-4})$) and the attacker learning iteration ($\mathcal{O}(\log \varepsilon^{-1})$). Moreover, the defender's gradient feedback can be approximated using the last iterate of the inner loop.

Lemma 4.6. *Under Assumption 4.4 and regularity assumptions. For any given $\varepsilon \in (0, 1)$, at any iteration $t \in 1, \dots, N_\mathcal{D}$, if the attacker learning iteration $N_\mathcal{A}$ and the batch size N_b are large enough such that $N_\mathcal{A} \sim \mathcal{O}(\log \varepsilon^{-1})$ and $N_b \sim \mathcal{O}(\varepsilon^{-4})$, then, for any $\xi \in \Xi$, the attack policy is stabilized, i.e.,*

$$\mathbb{E}\left[\max_\phi \langle \nabla_\phi \mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N_\mathcal{A}), \xi), \phi - \phi_\xi^t(N_\mathcal{A})\rangle\right] \leq \varepsilon.$$

Further, the defender's gradient feedback can be ε-approximated, i.e.,

$$\mathbb{E}\left[\|\nabla_\theta V(\theta^t) - \nabla_\theta \mathbb{E}_{\xi \sim Q} \mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N_\mathcal{A}), \xi)\|\right] \leq \varepsilon,$$

where the expectation $\mathbb{E}[\cdot]$ is taken over all the randomness from the algorithm.

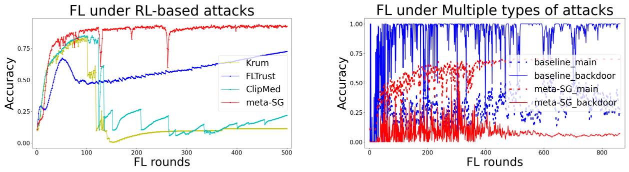
Equipped with Lemma 4.6, we apply the standard analysis for first-order methods in a non-convex setting to the outer loop, leading to the main complexity result in Theorem 4.7.

Theorem 4.7. *Under assumption 4.4 and regularity assumptions, for any given $\varepsilon \in (0, 1)$, let the learning rates $\kappa_\mathcal{A}$ and $\kappa_\mathcal{D}$ be properly chosen (see Appendix D); let $N_\mathcal{A}$ and N_b be chosen as required by Lemma 4.6, then, meta-SL finds a ε-meta-FOSE within $N_\mathcal{D} \sim \mathcal{O}(\varepsilon^{-2})$ iterations.*

Theorem 4.7 implies that meta-SL requires $\mathcal{O}(\varepsilon^{-2} \log \varepsilon^{-1})$ iterations of first-order evaluation, which matches the $\Omega(\varepsilon^{-2})$ lower bound for the general non-convex smooth setting (Carmon et al., 2020), (up to a logarithmic factor) indicating the efficiency of meta-SL.

5. Experiments

For the detailed setup of the experiment and corresponding results, please refer to our technical report (Li et al., 2023b).



(a)

(b)

Figure 2. Advantages of the Meta-SG framework against (a) the RL-based model poisoning attack (Li et al., 2022a) on MNIST with 20% malicious devices and (b) a mix of the backdoor attack against FL (BFL) (Bagdasaryan et al., 2020) (5% malicious devices) and the inner product manipulation (IPM) based model poisoning attack (Xie et al., 2020) (10% malicious devices) on CIFAR-10. The baseline defense combines the training-stage FLTrust and the post-training Neuron Clipping.

In our experiments, we evaluate our meta-SG defense using the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) datasets. The evaluation is performed under a range of advanced attacks, including non-adaptive and adaptive untargeted model poisoning attacks (specifically, IPM (Xie et al., 2020), LMP (Fang et al., 2020), RL (Li et al., 2022a)), backdoor attacks (BFL (Bagdasaryan et al., 2020), BRL (Li et al., 2023a)), and a combination thereof. Various robust defenses are taken into account as baselines, including training-stage defenses such as Krum (Blanchard et al., 2017), Clipping Median (Yin et al., 2018; Sun et al., 2019; Li et al., 2022a), FLTrust (Cao et al., 2021), and post-training defenses like Neuron Clipping (Wang et al., 2022), Pruning (Wu et al., 2020).

As shown in Figure 2(a), meta-SG demonstrates excellent accuracy in federated learning models when facing the RL-based model poisoning attack. Furthermore, as seen in Figure 2(b), meta-SG maintains high model accuracy for unpoisoned data and significantly lowers backdoor accuracy in scenarios involving both backdoor and model poisoning attacks. In contrast, the baseline defense that merely combines a training-stage defense with a post-training defense leads to low model accuracy and fails to shield the FL system from a backdoor attack. Due to the page limit, more experiment results are deferred to our technical report (Li et al., 2023b).

6. Conclusion and Future Work

In this work, we have proposed a data-driven approach to tackle information asymmetry in adversarial federated learning, which can also be applied to a variety of scenarios in adversarial machine learning, where information asymmetry is prevalent. We have offered a meta-equilibrium solution concept that is computationally tractable and strategically adaptable. In addition, theoretical guarantees on sample complexity and convergence rate have been established under mild assumptions, indicating the computational efficiency of training the meta-equilibrium-based FL defense.

References

Adler, I., Daskalakis, C., and Papadimitriou, C. H. A Note on Strictly Competitive Games. In *Internet and Network Economics*, pp. 471–474, 2009. ISBN 9783642108402. doi: 10.1007/978-3-642-10841-9_44.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948. PMLR, 2020.

Baruch, G., Baruch, M., and Goldberg, Y. A little is enough: Circumventing defenses for distributed learning. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2019.

Bernhard, P. and Rapaport, A. On a theorem of Danskin with an application to a theorem of Von Neumann-Sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24 (8):1163–1181, 1995. ISSN 0362-546X. doi: 10.1016/0362-546x(94)00186-l.

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. signsgd with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations(ICLR)*, 2018.

Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning(ICML)*, 2019.

Bhaskar, U., Cheng, Y., Ko, Y. K., and Swamy, C. Hardness results for signaling in bayesian zero-sum and network routing games. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 479–496, 2016.

Blanchard, P., Guerraoui, R., Stainer, J., et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2017.

Cao, X., Fang, M., Liu, J., and Gong, N. Z. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *NDSS*, 2021.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.

Chen, Z., Kailkhura, B., and Zhou, Y. An accelerated proximal algorithm for regularized nonconvex and nonsmooth bi-level optimization. *Machine Learning*, 112 (5):1433–1463, 2023. ISSN 0885-6125. doi: 10.1007/s10994-023-06329-6.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Fallah, A., Georgiev, K., Mokhtari, A., and Ozdaglar, A. On the convergence theory of debiased model-agnostic meta-reinforcement learning, 2021.

Fang, M., Cao, X., Jia, J., and Gong, N. Local model poisoning attacks to byzantine-robust federated learning. In *29th USENIX Security Symposium*, 2020.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Fudenberg, D. and Tirole, J. *Game Theory*. MIT Press, Cambridge, MA, 1991.

Ge, Y., Li, T., and Zhu, Q. Scenario-Agnostic Zero-Trust Defense with Explainable Threshold Policy: A Meta-Learning Approach. *arXiv preprint arXiv:2303.03349*, 2023.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Gupta, A., Lanctot, M., and Lazaridou, A. Dynamic population-based meta-learning for multi-agent communication with natural language. *Advances in Neural Information Processing Systems*, 34:16899–16912, 2021.

Harris, K., Anagnostides, I., Farina, G., Khodak, M., Wu, Z. S., and Sandholm, T. Meta-learning in games. *arXiv preprint arXiv:2209.14110*, 2022.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Jin, C., Netrapalli, P., and Jordan, M. I. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *ArXiv*, abs/1902.00618, 2019.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016.

Kayaalp, M., Vlaski, S., and Sayed, A. H. Dif-maml: Decentralized multi-agent meta-learning. *IEEE Open Journal of Signal Processing*, 3:71–93, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kwon, J., Kwon, D., Wright, S., and Nowak, R. A Fully First-Order Method for Stochastic Bilevel Optimization. *arXiv*, 2023. doi: 10.48550/arxiv.2301.10945.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, H. and Zheng, Z. Robust moving target defense against unknown attacks: A meta-reinforcement learning approach. In *Decision and Game Theory for Security: 13th International Conference, GameSec 2022, Pittsburgh, PA, USA, October 26–28, 2022, Proceedings*, pp. 107–126. Springer, 2023.

Li, H., Sun, X., and Zheng, Z. Learning to attack federated learning: A model-based reinforcement learning attack framework. In *Advances in Neural Information Processing Systems*, 2022a.

Li, H., Wu, C., Zhu, S., and Zheng, Z. Learning to backdoor federated learning. *arXiv preprint arXiv:2303.03320*, 2023a.

Li, H., Xu, T., Li, T., Pan, Y., Zhu, Q., and Zheng, Z. A first order meta stackelberg method for robust federated learning (technical report). *arXiv preprint arXiv:2306.13273*, 2023b.

Li, T. and Zhu, Q. On the Price of Transparency: A Comparison between Overt Persuasion and Covert Signaling. *arXiv preprint arXiv:2304.00096*, 2023.

Li, T., Lei, H., and Zhu, Q. Sampling attacks on meta reinforcement learning: A minimax formulation and complexity analysis. *arXiv preprint arXiv:2208.00081*, 2022b.

Li, T., Peng, G., Zhu, Q., and Baar, T. The Confluence of Networks, Games, and Learning a Game-Theoretic Framework for Multiagent Decision Making Over Networks. *IEEE Control Systems*, 42(4):35–67, 2022c. ISSN 1066-033X. doi: 10.1109/mcs.2022.3171478.

Li, T., Zhao, Y., and Zhu, Q. The role of information structures in game-theoretic multi-agent learning. *Annual Reviews in Control*, 53:296–314, 2022d. ISSN 1367-5788. doi: 10.1016/j.arcontrol.2022.03.003.

Manshaei, M. H., Zhu, Q., Alpcan, T., Bacşar, T., and Hubaux, J.-P. Game theory meets network security and privacy. *ACM Comput. Surv.*, 45(3), jul 2013. ISSN 0360-0300. doi: 10.1145/2480741.2480742. URL https://doi.org/10.1145/2480741.2480742.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Nguyen, T. D., Rieger, P., Chen, H., Yalame, H., Möllering, H., Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Zeitouni, S., et al. Flame: Taming backdoors in federated learning. *Cryptology ePrint Archive*, 2021.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods, 2019.

Pillutla, K., Kakade, S. M., and Harchaoui, Z. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

Rieger, P., Nguyen, T. D., Miettinen, M., and Sadeghi, A.-R. Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection. *arXiv preprint arXiv:2201.00763*, 2022.

Shejwalkar, V. and Houmansadr, A. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021.

Shen, W., Li, H., and Zheng, Z. Coordinated attacks against federated learning: A multi-agent reinforcement learning approach. In *ICLR 2021 Workshop on Security and Safety in Machine Learning Systems (SecML)*, 2021.

Song, X., Gao, W., Yang, Y., Choromanski, K., Pacchiano, A., and Tang, Y. Es-maml: Simple hessian-free meta learning. *arXiv preprint arXiv:1910.01215*, 2019.

Sun, Z., Kairouz, P., Suresh, A. T., and McMahan, H. B. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pp. 1057—1063. MIT press, 2000.

Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., Sohn, J.-y., Lee, K., and Papailiopoulos, D. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems*, 33:16070–16084, 2020.

Wang, H., Xiang, Z., Miller, D. J., and Kesidis, G. Universal post-training backdoor detection. *arXiv preprint arXiv:2205.06900*, 2022.

Wu, C., Yang, X., Zhu, S., and Mitra, P. Mitigating backdoor attacks in federated learning. *arXiv preprint arXiv:2011.01767*, 2020.

Xie, C., Huang, K., Chen, P.-Y., and Li, B. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Xie, C., Koyejo, O., and Gupta, I. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 261–270. PMLR, 2020.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659. PMLR, 2018.

Zhang, X., Hu, C., He, B., and Han, Z. Distributed reptile algorithm for meta-learning over multi-agent systems. *IEEE Transactions on Signal Processing*, 70:5443–5456, 2022.

Zhao, Y. and Zhu, Q. Stackelberg meta-learning for strategic guidance in multi-robot trajectory planning. *arXiv preprint arXiv:2211.13336*, 2022.

A. Broader Impacts and Limitations

Meta Equilibrium and Information Asymmetry. Information asymmetry is a prevailing phenomenon arising in a variety of contexts, including adversarial machine learning (e.g. FL discussed in this work), cyber security (Manshaei et al., 2013), and large-scale network systems (Li et al., 2022c). Our proposed meta-equilibrium (Definition 3.1) offers a data-driven approach tackling asymmetric information structure in dynamic games without Bayesian-posterior beliefs (see Appendix B). Achieving the strategic adaptation through stochastic gradient descent, the meta-equilibrium is computationally superior to perfect Bayesian equilibrium and better suited for real-world engineering systems involving high-dimensional continuous parameter spaces. It is expected that the meta-equilibrium can also be relevant to other adversarial learning contexts, cyber defense, and decentralized network systems.

First-order Method with Strict Competitiveness. Due to the hardness of the stochastic bilevel optimization problem, we have expanded our search scope with an alternative solution concept that merely involves the first-order necessary conditions for meta-SE. Our analytical result relies on the special game structure induced by the strict competitiveness assumption, which essentially "aligns" the defender/attacker objectives leveraging the nature of policy gradient, despite them being general-sum. Relaxing this assumption allows our framework to deal with a more general class of problems, yet may potentially disrupt the Danskin-type structure of gradient estimation. For simplicity of exposition, we neglected the stochastic analysis for the defender policy gradient estimation in the outer loop of the algorithm, the concentration of which depends on the trajectory batch size and attacker-type sample size. We leave the outer loop sample-complexity analysis to future work.

Incomplete Universal Defense. Our aim is to establish a comprehensive framework for universal federated learning defense. This framework ensures that the server remains oblivious to any details pertaining to the environment or potential attackers. Still, it possesses the ability to swiftly adapt and respond to uncertain or unknown attackers during the actual federated learning process. Nevertheless, achieving this universal defense necessitates an extensive attack set through pre-training, which often results in a protracted convergence time toward a meta-policy. Moreover, the effectiveness and efficiency of generalizing from a wide range of diverse attack distributions pose additional challenges. Considering these, we confine our experiments in this paper to specifically address a subset of uncertainties and unknowns. This includes parameters that determine attack types, the number of malicious devices, the heterogeneity of local data distributions, backdoor triggers, backdoor targets, and other relevant aspects. However, we acknowledge that our focus is not all-encompassing, and there may be other factors that remain unexplored in our research.

B. Further Justification on Meta Equilibrium

This section offers further justification for the meta-equilibrium, and we argue that meta-equilibrium provides a data-driven approach to address incomplete information in dynamic games. Note that information asymmetry is prevalent in the adversarial machine learning context, where the attacker enjoys an information advantage (e.g., the attacker's type). The proposed meta-equilibrium notion can shed light on these related problems beyond the adversarial FL context.

We begin with the insufficiency of Bayesian Stackelberg equilibrium (1) in handling information asymmetry, a customary solution concept in security studies (Li et al., 2022d). One can see from (1) that such an equilibrium is of ex-ante type: the defender's strategy is determined before the game starts. It targets an "representative" attacker (an average of all types). As the game unfolds, new information regarding the attacker's private type is revealed (e.g., through the global model updates). However, this ex-ante strategy does not enable the defender to handle this emerging information as the game proceeds. Using game theory language, the defender fails to adapt its strategy in the interim stage.

To create interim adaptability in this dynamic game of incomplete information, one can consider introducing the belief system to capture the defender's learning process on the hidden type. Let I^t be the defender's observations up to time t, i.e., $I^t := (s^k, a_{\mathcal{D}}^k)_{k=1}^t s^{t+1}$. Denote by $\mathcal{B}$ the belief generation operator $b^{t+1}(\xi) = \mathcal{B}[I^t]$. With the Bayesian equilibrium framework, the belief generation can be defined recursively as below

$$b^{t+1}(\xi) = \mathcal{B}[s^t, a_{\mathcal{D}}^t, b^t] := \frac{b^t(\xi)\pi_{\mathcal{A}}(a_{\mathcal{A}}^t|s^t;\xi)\mathcal{T}(s^{t+1}|s^t, a_{\mathcal{A}}^t, a_{\mathcal{D}}^t)}{\sum_{\xi'} b^t(\xi')\pi_{\mathcal{A}}(a_{\mathcal{A}}^t|s^t;\xi')\mathcal{T}(s^{t+1}|s^t, a_{\mathcal{A}}^t, a_{\mathcal{D}}^t)}. \tag{B1}$$

Since b^t is the defender's belief on the hidden type at time t, its belief-dependent Markovian strategy is defined as $\pi_{\mathcal{D}}(s^t, b^t)$. Therefore, the interim equilibrium, also called Perfect Bayesian Equilibrium (PBE) (Fudenberg & Tirole, 1991) is given by a tuple $(\pi_{\mathcal{D}}^*, \pi_{\mathcal{A}}^*, \{b^t\}_{t=1}^H)$ satisfying

$$\begin{aligned}
&\pi_{\mathcal{D}}^* = \arg\max \mathbb{E}_{\xi\sim Q}\mathbb{E}_{\pi_{\mathcal{D}},\pi_{\mathcal{A}}^*}[\sum_{t=1}^H r_{\mathcal{D}}(s^t, a_{\mathcal{D}}^t, a_{\mathcal{A}}^t) b^t(\xi)] \\
&\pi_{\mathcal{A}}^* = \arg\max \mathbb{E}_{\pi_{\mathcal{D}},\pi_{\mathcal{A}}}[\sum_{t=1}^H r_{\mathcal{A}}(s^t, a_{\mathcal{D}}^t, a_{\mathcal{A}}^t)], \forall \xi, \\
&\{b^k\}_{k=1}^H \text{ satisfies (B1) for realized actions and states.}
\end{aligned} \tag{PBE}$$

In contrast with (1), this perfect Bayesian equilibrium notion (PBE) enables the defender to make good use of the information revealed by the attacker, and subsequently adjust its actions according to the revealed information through the belief generation. From a game-theoretic viewpoint, both (PBE) and (2) create strategic online adaptation: the defender can infer and adapt to the attacker's private type through the revealed information since different types aim at different objectives, hence, leading to different actions. Compared with PBE, the proposed meta-equilibrium notion is better suited for large-scale complex systems where players' decision variables can be high-dimensional and continuous, as argued in the ensuing paragraph.

To achieve the strategic adaptation, PBE relies on the Bayesian-posterior belief updates, which soon become intractable as the denominator in (B1) involves integration over high-dimensional space and discretization inevitably leads to the curse of dimensionality. Despite the limited practicality, PBE is inherently difficult to solve even in finite-dimensional cases. It is shown in (Bhaskar et al., 2016) that the equilibrium computation in games with incomplete information is NP-hard, and how to solve for PBE in dynamic games remains an open problem. Even though there have been encouraging attempts at solving PBE in two-stage games (Li & Zhu, 2023), it is still challenging to address PBE computation in generic Markov games.

C. Algorithms

This section elaborates on meta-learning defense and meta-Stackelberg learning. To begin with, we first review the policy gradient method (Sutton et al., 2000) in RL and its Monte-Carlo estimation. To simplify our exposition, we fix the attacker's policy ϕ, and then BSMG reduces to a single-agent MDP, where the optimal policy to be learned is the defender's θ.

Policy Gradient The idea of the policy gradient method is to apply gradient ascent to the value function $J_{\mathcal{D}}$. Following (Sutton et al., 2000), we obtain $\nabla_\theta J_{\mathcal{D}} := \mathbb{E}_{\tau\sim q(\theta)}[g(\tau;\theta)]$, where $g(\tau;\theta) = \sum_{t=1}^H \nabla_\theta \log \pi(a_{\mathcal{D}}^t | s^t; \theta) R(\tau)$ and $R(\tau) = \sum_{t=1}^H \gamma^t r(s^t, a_{\mathcal{D}}^t)$. Note that for simplicity, we suppress the parameter ϕ, ξ in the trajectory distribution q, and instead view it as a function of θ. In numerical implementations, the policy gradient $\nabla_\theta J_{\mathcal{D}}$ is replaced by its Monte-Carlo (MC) estimation using sample trajectory. Suppose a batch of trajectories $\{\tau_i\}_{i=1}^{N_b}$, and N_b denotes the batch size, then the MC estimation is

$$\hat{\nabla}_\theta J_{\mathcal{D}}(\theta, \tau) := 1/N_b \sum_{\tau_i} g(\tau_i; \theta) \tag{C1}$$

The same deduction also holds for the attacker's problem when fixing the defense θ.

Meta-Learning FL Defense Meta-learning-based defense (meta defense) mainly targets non-adaptive attack methods, where $\pi_{\mathcal{A}}(\cdot; \phi, \xi)$ is a pre-fixed attack strategy following some rulebook, such as IPM (Xie et al., 2020) and LMP (Fang et al., 2020). In this case, the BSMG reduces to single-agent MDP for the defender, where the transition kernel is determined by the attack method. Mathematically, the meta-defense problem is given by

$$\max_{\theta, \Psi} \mathbb{E}_{\xi\sim Q(\cdot)}[J_{\mathcal{D}}(\Psi(\theta, \tau), \phi, \xi)] \tag{C2}$$

Since the attack type is hidden from the defender, the adaptation mapping Ψ is usually defined in a data-driven manner. For example, $\Psi(\theta, \tau)$ can be defined as a one-step stochastic gradient update with learning rate η: $\Psi(\theta, \tau) = \theta + \eta \hat{\nabla} J_{\mathcal{D}}(\tau_\xi)$ (Finn et al., 2017) or a recurrent neural network in (Duan et al., 2016). This work mainly focuses on gradient adaptation for the purpose of deriving theoretical guarantees in Appendix D.

With the one-step gradient adaptation, the meta-defense problem in (C2) can be simplified as

$$\max_{\theta} \mathbb{E}_{\xi\sim Q(\cdot)}\mathbb{E}_{\tau\sim q(\theta)}[J_{\mathcal{D}}(\theta + \eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau), \phi, \xi)] \tag{C3}$$

Recall that the attacker's strategy is pre-determined, ϕ, ξ can be viewed as fixed parameters, and hence, the distribution q is a function of θ. To apply the policy gradient method to (C3), one needs an unbiased estimation of the gradient of the objective function in (C3). Consider the gradient computation using the chain rule:

$$\begin{aligned}
&\nabla_\theta \mathbb{E}_{\tau\sim q(\theta)}[J_{\mathcal{D}}(\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau),\phi,\xi)] \\
&= \mathbb{E}_{\tau\sim q(\theta)}\{\underbrace{\nabla_\theta J_{\mathcal{D}}(\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau),\phi,\xi)(I+\eta\hat{\nabla}^2_\theta J_D(\tau))}_{\text{①}} \\
&\quad \underbrace{+\, J_{\mathcal{D}}(\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau))\nabla_\theta \sum_{t=1}^{H}\pi(a^t|s^t;\theta)\}.}_{\text{②}}
\end{aligned} \tag{C4}$$

The first term results from differentiating the integrand $J_{\mathcal{D}}(\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau),\phi,\xi)$ (the expectation is taken as integration), while the second term is due to the differentiation of $q(\theta)$. One can see from the first term that the above gradient involves a Hessian $\hat{\nabla}^2 J_{\mathcal{D}}$, and its sample estimate is given by the following. For more details on this Hessian estimation, we refer the reader to (Fallah et al., 2021).

$$\hat{\nabla}^2 J_{\mathcal{D}}(\tau) = \frac{1}{N_b}\sum_{i=1}^{N_b}[g(\tau_i;\theta)\nabla_\theta \log q(\tau_i;\theta)^\mathsf{T} + \nabla_\theta g(\tau_i;\theta)] \tag{C5}$$

Finally, to complete the sample estimate of $\nabla_\theta \mathbb{E}_{\tau\sim q(\theta)}[J_{\mathcal{D}}(\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau),\phi,\xi)]$, one still needs to estimate $\nabla_\theta J_{\mathcal{D}}(\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau),\phi,\xi)$ in the first term. To this end, we need to first collect a batch of sample trajectories τ' using the adapted policy $\theta' = \theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau)$. Then, the policy gradient estimate of $\hat{\nabla}_\theta J_{\mathcal{D}}(\theta')$ proceeds as in (C1). To sum up, constructing an unbiased estimate of (C4) takes two rounds of sampling. The first round is under the meta policy θ, which is used to estimate the Hessian (C5) and to adapt the policy to θ'. The second round aims to estimate the policy gradient $\nabla_\theta J_{\mathcal{D}}(\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau),\phi,\xi)$ in the first term in (C4).

In the experiment, we employ a first-order meta-learning algorithm called Reptile (Nichol et al., 2018) to avoid the Hessian computation. The gist is to simply ignore the chain rule and update the policy using the gradient $\nabla_\theta J_{\mathcal{D}}(\theta',\phi,\xi)|_{\theta'=\theta+\eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau)}$. Naturally, without the Hessian term, the gradient in this update is biased, yet it still points to the ascent direction as argued in (Nichol et al., 2018), leading to effective meta policy. The advantage of Reptile is more evident in multi-step gradient adaptation. Consider a l-step gradient adaptation, the chain rule computation inevitably involves multiple Hessian terms (each gradient step brings a Hessian term) as shown in (Fallah et al., 2021). In contrast, Reptile only requires first-order information, and the meta-learning algorithm (l-step adaptation) is given by Algorithm 1.

Algorithm 1 Reptile Meta-Reinforcement Learning with l-step adaptation

```
Input: the type distribution $Q(\xi)$, step size parameters $\kappa, \eta$
Output: $\theta^T$
randomly initialize $\theta^0$
for iteration $t = 1$ to $T$ do
    Sample a batch $\hat{\Xi}$ of $K$ attack types from $Q(\xi)$;
    for each $\xi \in \hat{\Xi}$ do
        $\theta^t_\xi(0) \leftarrow \theta^t$
        for $k = 0$ to $l - 1$ do
            Sample a batch trajectories $\tau$ of the horizon length $H$ under $\theta^t_\xi(k)$;
            Evaluate $\hat{\nabla}_\theta J_{\mathcal{D}}(\theta^t_\xi(k), \tau)$ using MC in (C1);
            $\theta^t_\xi(k+1) \leftarrow \theta^t_\xi(k) + \kappa\hat{\nabla}_\theta J_{\mathcal{D}}(\theta^t, \tau)$
        end for
    end for
    Update $\theta^{t+1} \leftarrow \theta^t + 1/K \sum_{\xi\in\hat{\Xi}}(\theta^t_\xi(l) - \theta^t)$;
end for
```

Meta-Stackelberg Learning Recall that in meta-SE, the attacker's policy ϕ^*_ξ is not pre-fixed, instead, it is the best response to the defender's adapted policy. To obtain this best response, one needs alternative training: fixing the defense policy, and applying gradient ascent to the attacker's problem until convergence. It should be noted that the proposed meta-SL utilizes the unbiased gradient estimation in (C5), which paves the way for theoretical analysis in Appendix D. Yet, we turn to the Reptile to speed up pre-straining in the experiments. We present both algorithms in Algorithm 2, and only consider one-step adaptation for simplicity. The multi-step version is a straightforward extension of Algorithm 2.

Algorithm 2 (Reptile) Meta-Stackelberg Learning with one-step adaptation

```
**Input:** the type distribution $Q(\xi)$, initial defense meta policy $\theta^0$, pre-trained attack policies $\{\phi^0_\xi\}_{\xi\in\Xi}$, step size parameters $\kappa_{\mathcal{D}}$, $\kappa_{\mathcal{A}}$, $\eta$, and iterations numbers $N_{\mathcal{A}}, N_{\mathcal{D}}$;
**Output:** $\theta^{N_{\mathcal{D}}}$
**for** iteration $t = 0$ to $N_{\mathcal{D}} - 1$ **do**
    Sample a batch $\hat{\Xi}$ of $K$ attack types from $Q(\xi)$;
    **for** each $\xi \in \hat{\Xi}$ **do**
        Sample a batch of trajectories using $\phi^t$ and $\phi^t_\xi$;
        Evaluate $\hat{\nabla}_\theta J_D(\theta^t, \phi^t_\xi, \xi)$ using (C1);
        Perform one-step adaptation $\theta^t_\xi \leftarrow \theta^t + \eta\hat{\nabla}_\theta J_D(\theta^t_\xi(k), \phi^t_\xi, \xi)$;
        $\phi^t_\xi(0) \leftarrow \phi^t_\xi$;
        **for** $k = 0, \dots, N_{\mathcal{A}} - 1$ **do**
            Sample a batch of trajectories using $\theta^t_\xi$ and $\phi^t_\xi(k)$;
            $\phi^t_\xi(k+1) \leftarrow \phi^t_\xi(k) + \kappa_{\mathcal{A}}\hat{\nabla}_\phi J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(k), \xi)$;
        **end for**
        **if** Reptile **then**
            Sample a batch of trajectories using $\theta^t_\xi$ and $\phi^t_\xi(N_{\mathcal{A}})$;
            Evaluate $\hat{\nabla} J_D(\xi) := \hat{\nabla}_\theta J_{\mathcal{D}}(\theta, \phi^t_\xi(N_{\mathcal{A}}), \xi)|_{\theta=\theta^t_\xi}$ using (C1);
        **else**
            Sample a batch of trajectories using $\theta^t$ and $\phi^t_\xi(N_{\mathcal{A}})$;
            Evaluate the Hessian using (C5);
            Sample a batch of trajectories using $\theta^t_\xi$ and $\phi^t_\xi(N_{\mathcal{A}})$;
            Evaluate $\hat{\nabla} J_D(\xi) := \hat{\nabla}_\theta J_{\mathcal{D}}(\theta^t_\xi, \phi^t_\xi(N_{\mathcal{A}}), \xi)$ using (C4);
        **end if**
        $\bar{\theta}^t_\xi \leftarrow \theta^t + \kappa_{\mathcal{D}}\hat{\nabla} J_D(\xi)$;
    **end for**
    $\theta^{t+1} \leftarrow \theta^t + 1/K \sum_{\xi\sim\hat{\Xi}}(\bar{\theta}^t_\xi - \theta_t)$, $\phi^{t+1}_\xi \leftarrow \phi^t_\xi(N_{\mathcal{A}})$;
**end for**
```

D. Theoretical Results

D.1. Existence of Meta-SG

Theorem D.1 (Theorem 4.2). *Under the conditions that Θ and Φ are compact and convex, the meta-SG admits at least one meta-FOSE.*

Proof. Clearly, $\Theta \times \Phi^{|\Xi|}$ is compact and convex, let $\phi \in \Phi^{|\Xi|}$, $\phi_\xi \in \Phi$ be the (type-aggregated) attacker's strategy, since the consider twice continuously differentiable utility functions $\ell_{\mathcal{D}}(\theta, \phi) := \mathbb{E}_{\xi\sim Q}\mathcal{L}_{\mathcal{D}}(\theta, \phi_\xi, \xi)$ and $\ell_\xi(\theta, \phi) := \mathcal{L}_{\mathcal{A}}(\theta, \phi_\xi, \xi)$ for all $\xi \in \Xi$. Then, there exists a constant $\gamma_c > 0$, such that the auxiliary utility functions:

$$\begin{aligned} \tilde{\ell}_{\mathcal{D}}(\theta; (\theta', \phi')) &\equiv \ell_{\mathcal{D}}(\theta, \phi) - \frac{\gamma_c}{2}\|\theta - \theta'\|^2 \\ \tilde{\ell}_\xi(\phi_\xi; (\theta', \phi')) &\equiv \ell_\xi(\theta', (\phi_\xi, \phi'_{-\xi})) - \frac{\gamma_c}{2}\|\phi_\xi - \phi'_\xi\|^2 \quad \forall \xi \in \Xi \end{aligned} \tag{D6}$$

are γ_c-strongly concave in spaces $\theta \in \Theta$, $\phi_\xi \in \Phi$ for all $\xi \in \Xi$, respectively for any fixed $(\theta', \phi') \in \Theta \times \Phi^{|\Xi|}$.

Define the self-map $h : \Theta \times \Phi^{|\Xi|} \to \Theta \times \Phi^{|\Xi|}$ with $h(\theta', \phi') \equiv (\bar{\theta}(\theta', \phi'), \bar{\phi}(\theta', \phi'))$, where

$$\bar{\theta}(\theta', \phi') = \arg\max_{\theta \in \Theta} \tilde{\ell}_{\mathcal{D}}(\theta, \phi'), \qquad \bar{\phi}_\xi(\theta', \phi') = \arg\max_{\phi_\xi \in \Phi} \tilde{\ell}_\xi(\theta', (\phi_\xi, \phi'_{-\xi})).$$

Due to compactness, h is well-defined. By strong concavity of $\tilde{\ell}_{\mathcal{D}}(\cdot; (\theta', \phi'))$ and $\tilde{\ell}_\xi(\cdot; (\theta', \phi'))$, it follows that $\bar{\theta}, \bar{\phi}$ are continuous self-mapping from $\Theta \times \Phi^{|\Xi|}$ to itself. By Brouwer's fixed point theorem, there exists at least one $(\theta^*, \phi^*) \in \Theta \times \Phi^{|\Xi|}$ such that $h(\theta^*, \phi^*) = (\theta^*, \phi^*)$. Then, one can verify that (θ^*, ϕ^*) is a meta-FOSE of the meta-SG with utility function $\ell_{\mathcal{D}}$ and ℓ_ξ, $\xi \in \Xi$, in view of the following inequality

$$\begin{aligned} \langle \nabla_\theta \tilde{\ell}_{\mathcal{D}}(\theta^*; (\theta^*, \phi^*)), \theta - \theta^* \rangle &= \langle \nabla_\theta \ell_{\mathcal{D}}(\theta^*, \phi^*), \theta - \theta^* \rangle \\ \langle \nabla_{\phi_\xi} \tilde{\ell}_\xi(\theta^*; (\theta^*, \phi^*)), \phi_\xi - \phi_\xi^* \rangle &= \langle \nabla_{\phi_\xi} \ell_\xi(\theta^*, \phi^*), \phi_\xi - \phi_\xi^* \rangle, \end{aligned}$$

therefore, the equilibrium conditions for meta-SG with utility functions $\tilde{\ell}_{\mathcal{D}}$ and $\{\tilde{\ell}_\xi\}_{\xi \in \Xi}$ are the same as with utility functions $\ell_{\mathcal{D}}$ and $\{\ell_\xi\}_{\xi \in \Xi}$, hence the claim follows. □

D.2. Proofs: Non-Asymptotic Analysis

In the sequel, we make the following smoothness assumptions for every attack type $\xi \in \Xi$. In addition, we assume, for analytical simplicity, that all types of attackers are unconstrained, i.e., Φ is the Euclidean space with proper finite dimension.

Assumption D.2 ((ξ-wise) Lipschitz smoothness)**.** The functions $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{A}}$ are continuously diffrentiable in both θ and ϕ. Furthermore, there exists constants L_{11}, L_{12}, L_{21}, and L_{22} such that for all $\theta, \theta_1, \theta_2 \in \Theta$ and $\phi, \phi_1, \phi_2 \in \Phi$, we have, for any $\xi \in \Xi$,

$$\|\nabla_\theta \mathcal{L}_{\mathcal{D}}(\theta_1, \phi, \xi) - \nabla_\theta \mathcal{L}_{\mathcal{D}}(\theta_2, \phi, \xi)\| \leq L_{11} \|\theta_1 - \theta_2\| \tag{D7}$$
$$\|\nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta, \phi_1, \xi) - \nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta, \phi_2, \xi)\| \leq L_{22} \|\phi_1 - \phi_2\| \tag{D8}$$
$$\|\nabla_\theta \mathcal{L}_{\mathcal{D}}(\theta, \phi_1, \xi) - \nabla_\theta \mathcal{L}_{\mathcal{D}}(\theta, \phi_2, \xi)\| \leq L_{12} \|\phi_1 - \phi_2\| \tag{D9}$$
$$\|\nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta_1, \phi, \xi) - \nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta_2, \phi, \xi)\| \leq L_{12} \|\theta_1 - \theta_2\| \tag{D10}$$
$$\|\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta, \phi_1, \xi) - \nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta, \phi_2, \xi)\| \leq L_{21} \|\phi_1 - \phi_2\|. \tag{D11}$$

Lemma D.3 (Implicit Function Theorem (IFT) for Meta-SG)**.** *Suppose for $(\bar{\theta}, \bar{\phi}) \in \Theta \times \Phi^{|\Xi|}$, $\xi \in \Xi$ we have $\nabla_\phi \mathcal{L}_{\mathcal{A}}(\bar{\theta}, \bar{\phi}, \xi) = 0$ the Hessian $\nabla_\phi^2 \mathcal{L}_{\mathcal{A}}(\bar{\theta}, \bar{\phi}, \xi)$ is non-singular. Then, there exists a neighborhood $B_\varepsilon(\bar{\theta}), \varepsilon > 0$ centered around $\bar{\theta}$ and a C^1-function $\phi(\cdot) : B_\varepsilon(\bar{\theta}) \to \Phi^{|\Xi|}$ such that near $(\bar{\theta}, \bar{\phi})$ the solution set $\{(\theta, \phi) \in \Theta \times \Phi^{|\Xi|} : \nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi) = 0\}$ is a C^1-manifold locally near $(\bar{\theta}, \bar{\phi})$. The gradient $\nabla_\theta \phi(\theta)$ is given by $-(\nabla_\phi^2 \mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi))^{-1} \nabla_{\phi\theta}^2 \mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi)$.*

Lemma D.4. *Under assumptions D.2, 4.4, there exists $\{\phi_\xi : \phi_\xi \in \arg\max_\phi \mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi)\}_{\xi \in \Xi}$, such that*

$$\nabla_\theta V(\theta) = \nabla_\theta \mathbb{E}_{\xi \sim Q, \tau \sim q} J_{\mathcal{D}}(\theta + \eta \hat{\nabla}_\theta J_{\mathcal{D}}(\tau), \phi_\xi, \xi).$$

Moreover, the function $V(\theta)$ is L-Lipschitz-smooth, where $L = L_{11} + \frac{L_{12} L_{21}}{\mu}$

$$\|\nabla_\theta V(\theta_1) - \nabla_\theta V(\theta_2)\| \leq L \|\theta_1 - \theta_2\|.$$

Proof of Lemma D.4. First, we show that for any $\theta_1, \theta_2 \in \Theta, \xi \in \Xi$, and $\phi_1 \in \arg\max_\phi \mathcal{L}_{\mathcal{A}}(\theta_1, \phi, \xi)$, there exists $\phi_2 \in \arg\max_\phi \mathcal{L}_{\mathcal{A}}(\theta_2, \phi, \xi)$ such that $\|\phi_1 - \phi_2\| \leq \frac{L_{12}}{\mu} \|\theta_1 - \theta_2\|$. Indeed, based on smoothness assumption (D11) and (D10),

$$\begin{aligned} \|\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta_1, \phi_1, \xi) - \nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta_2, \phi_1, \xi)\| &\leq L_{21} \|\theta_1 - \theta_2\|, \\ \|\nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta_1, \phi_1, \xi) - \nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta_2, \phi_1, \xi)\| &\leq L_{12} \|\theta_1 - \theta_2\|. \end{aligned}$$

Since $\phi_2 \in \arg\max_\phi \mathcal{L}_{\mathcal{A}}(\theta_2, \phi, \xi)$, $\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta_2, \phi_2, \xi) = 0$. Apply PL condition to $\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta, \phi_2, \xi)$,

$$\begin{aligned} \max_\phi \mathcal{L}_{\mathcal{A}}(\theta_1, \phi, \xi) - \mathcal{L}_{\mathcal{A}}(\theta_1, \phi_2, \xi) &\leq \frac{1}{2\mu} \|\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta_1, \phi_2, \xi)\|^2 \\ &= \frac{1}{2\mu} \|\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta_1, \phi_2, \xi) - \nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta_2, \phi_2, \xi)\|^2 \\ &\leq \frac{L_{21}^2}{2\mu} \|\theta_1 - \theta_2\|^2 \qquad \text{by (D11).} \end{aligned}$$

Since PL condition implies quadratic growth, we also have

$$\mathcal{L}_{\mathcal{A}}(\theta_1, \phi_1, \xi) - \mathcal{L}_{\mathcal{A}}(\theta_1, \phi_2, \xi) \geq \frac{\mu}{2}\|\phi_1 - \phi_2\|^2.$$

Combining the two inequalities above we obtain the Lipschitz stability for $\phi^*_\xi(\cdot)$, i.e.,

$$\|\phi_1 - \phi_2\| \leq \frac{L_{21}}{\mu}\|\theta_1 - \theta_2\|.$$

Second, show that $\nabla_\theta V(\theta)$ can be directly evaluated at $\{\phi^*_\xi\}_{\xi\in\Xi}$. Inspired by Danskin's theorem, we first made the following argument, consider the definition of directional derivative. Let $\ell(\theta, \phi) := \nabla_\theta \mathbb{E}_{\xi,\tau} J_{\mathcal{D}}(\theta + \eta\hat{\nabla} J_{\mathcal{D}}(\tau), \xi)$. For a constant τ and an arbitrary direction d,

$$\begin{aligned}
&\ell(\theta + \tau d, \phi^*(\theta + \tau d)) - \ell(\theta, \phi^*(\theta)))\\
&= \ell(\theta + \tau d, \phi^*(\theta + \tau d)) - \ell(\theta + \tau d, \phi^*(\theta)) + \ell(\theta + \tau d, \phi^*(\theta)) - \ell(\theta, \phi^*(\theta))\\
&= \nabla_\phi \ell(\theta + \tau d, \phi^*(\theta))^\top \underbrace{[\phi^*(\theta + \tau d) - \phi^*(\theta))]}_{\Delta\phi} + o(\Delta\phi^2)\\
&+ \tau\nabla_\theta \ell(\theta, \phi^*(\theta))^T d + o(d^2).
\end{aligned}$$

Hence, a sufficient condition for the first equation is $\nabla_\phi \ell(\theta + \tau d, \phi^*(\theta)) = 0$, meaning that $\ell_D(\theta, \phi)$ and $\mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi)$ share the first-order stationarity at every ϕ when fixing θ. Indeed, by Lemma D.3, we have, the gradient is locally determined by

$$\begin{aligned}
\nabla_\theta V &= \mathbb{E}_{\xi\sim Q}[\nabla_\theta \mathcal{L}_{\mathcal{D}}(\theta, \phi_\xi, \xi) + (\nabla_\theta \phi_\xi(\theta))^\top \nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta, \phi_\xi, \xi)]\\
&= \mathbb{E}_{\xi\sim Q}\left[\nabla_\theta \mathcal{L}_{\mathcal{D}}(\theta, \phi_\xi, \xi) - [(\nabla^2_\phi \mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi))^{-1}\nabla^2_{\phi\theta}\mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi)]^\top \nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta, \phi_\xi, \xi)\right].
\end{aligned}$$

Given a trajectory $\tau := (s^1, a^t_{\mathcal{D}}, a^t_{\mathcal{A}}, \ldots, a^H_{\mathcal{D}}, a^H_{\mathcal{A}}, s^{H+1})$, let $R_{\mathcal{D}}(\tau, \xi) := \sum_{t=1}^H \gamma^{t-1} r_{\mathcal{D}}(s_t, a_t, \xi)$ and $R_{\mathcal{D}}(\tau, \xi) := \sum_{t=1}^H \gamma^{t-1} r_{\mathcal{D}}(s_t, a_t, \xi)$. Denote by $\mu(\tau; \theta, \phi)$ the trajectory distribution, that the log probability of μ is given by

$$\log\mu(\tau; \theta, \phi) = \sum_{t=1}^H (\log \pi_{\mathcal{D}}(a^t_{\mathcal{D}}|s^t; \theta + \eta\hat{\nabla}_\theta J_{\mathcal{D}}(\tau)) + \log \pi_{\mathcal{A}}(a^t_{\mathcal{A}}|s^t; \phi) + \log P(s^{t+1}|a^t_{\mathcal{D}}, a^t_{\mathcal{A}}, s^t)$$

According to the policy gradient theorem, we have

$$\begin{aligned}
\nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta, \phi, \xi) &= \mathbb{E}_\mu[R_{\mathcal{D}}(\tau, \xi)\sum_{t=1}^H \nabla_\phi \log(\pi_{\mathcal{A}}(a^t_{\mathcal{A}}|s^t; \phi))],\\
\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi) &= \mathbb{E}_\mu[R_{\mathcal{A}}(\tau, \xi)\sum_{t=1}^H \nabla_\phi \log(\pi_{\mathcal{A}}(a^t_{\mathcal{A}}|s^t; \phi))].
\end{aligned}$$

By SC Assumption 4.3, when $\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta, \phi, \xi) = 0$, there exists $c < 0$, d, such that $\nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta, \phi, \xi) = \mathbb{E}_\mu[cR_{\mathcal{A}}(\tau, \xi)\sum_{t=1}^H \nabla_\phi \log(\pi_{\mathcal{A}}(a^t_{\mathcal{A}}|s^t; \phi))] + \mathbb{E}_\mu[\sum_{t=1}^H \gamma^{t-1} d \sum_{t=1}^H \nabla_\phi \log(\pi_{\mathcal{A}}(a^t_{\mathcal{A}}|s^t; \phi))] = 0$. Hence $\nabla_\theta V = \mathbb{E}_{\xi\sim Q}[\nabla_\theta \mathcal{L}_{\mathcal{D}}(\theta, \phi_\xi, \xi)]$.

Third, $V(\theta)$ is also Lipschitz smooth. As we notice that, $\ell_{\mathcal{D}}$ is Lipschitz smooth since $\mathbb{E}_{\xi\sim Q}$ is a linear operator, we have,

$$\begin{aligned}
&\|\nabla_\theta V(\theta_1) - \nabla_\theta V(\theta_2)\|\\
&\leq \|\nabla_\theta \mathbb{E}_{\xi\sim Q}\mathcal{L}_{\mathcal{D}}(\theta_1, \phi_1, \xi) - \nabla_\theta \mathbb{E}_{\xi\sim Q}\mathcal{L}_{\mathcal{D}}(\theta_2, \phi_2, \xi)\|\\
&= \|\nabla_\theta \ell_{\mathcal{D}}(\theta_1, \phi_1) - \nabla_\theta \ell_{\mathcal{D}}(\theta_2, \phi_1) + \nabla_\theta \ell_{\mathcal{D}}(\theta_2, \phi_1) - \nabla_\theta \ell_{\mathcal{D}}(\theta_2, \phi_2)\|\\
&\leq \|\nabla_\theta \ell_{\mathcal{D}}(\theta_1, \phi_1) - \nabla_\theta \ell_{\mathcal{D}}(\theta_2, \phi_1)\| + \|\nabla_\theta \ell_{\mathcal{D}}(\theta_2, \phi_1) - \nabla_\theta \ell_{\mathcal{D}}(\theta_2, \phi_2)\|\\
&\leq L_{11}\|\theta_1 - \theta_2\| + L_{12}\|\phi_1 - \phi_2\|\\
&\leq (L_{11} + \frac{L_{12}L_{21}}{\mu})\|\theta_1 - \theta_2\|,
\end{aligned}$$

which implies the Lipschitz constant $L = L_{11} + \frac{L_{12}L_{21}}{\mu}$. □

It is impossible to present the convergence theory without the assistance of some standard assumptions in batch reinforcement learning, of which the justification can be found in (Fallah et al., 2021). We also require some additional information about the parameter space and function structure. These assumptions are all stated in Assumption D.5.

Assumption D.5.

(a) The following policy gradients are bounded, $\|\nabla_\phi \mathcal{L}_\mathcal{D}(\theta, \phi, \xi)\| \leq G^2$, $\|\mathcal{L}_\mathcal{A}(\theta, \phi, \xi)\| \leq G^2$ for all $\theta, \phi \in \Theta \times \Phi$ and $\xi \in \Xi$.

(b) The policy gradient estimations are unbiased.

(c) The variances for the stochastic gradients are bounded, i.e., for all $theta_\xi^t, \phi_\xi^t, \xi$,

$$\mathbb{E}[\|\hat{\nabla}_\phi J(\theta_\xi^t, \phi_\xi^t, \xi) - \nabla_\phi J(\theta_\xi^t, \phi_\xi^t, \xi)\|^2] \leq \frac{\sigma^2}{N_b}.$$

(d) The parameter space Θ has diameter $D_\Theta := \sup_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\|$; the initialization θ^0 admits at most D_V function gap, i.e., $D_V := \max_{\theta \in \Theta} V(\theta) - V(\theta^0)$.

(e) It holds that the parameters satisfy $0 < \mu < -cL_{22}$.

Equipped with Assumption D.5 we are able to unfold our main result Theorem 4.7, before which we show in Lemma D.6 that ϕ_ξ^* can be efficiently approximated by the inner loop in the sense that $\nabla_\theta \mathbb{E}_{\xi \sim Q} \mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N_\mathcal{A}), \xi) \approx \nabla_\theta V(\theta^t)$, where $\phi_\xi^t(N_\mathcal{A})$ is the last iterate output of the attacker policy.

Lemma D.6. *Under Assumption D.5, 4.4, 4.3, and D.2, let* $\rho := 1 + \frac{\mu}{cL_{22}} \in (0, 1)$, $\bar{L} = \max\{L_{11}, L_{12}, L_{22}, L_{21}, V_\infty\}$ *where* $V_\infty := \max\{\max \|\nabla V(\theta)\|, 1\}$*. For all* $\varepsilon > 0$*, if the attacker learning iteration* $N_\mathcal{A}$ *and batch size* N_b *are large enough such that*

$$N_\mathcal{A} \geq \frac{1}{\log \rho^{-1}} \log \frac{32 D_V^2 (2V_\infty + LD_\Theta)^4 \bar{L} |c| G^2}{L^2 \mu^2 \varepsilon^4}$$

$$N_b \geq \frac{32 \mu L_{21}^2 D_V^2 (2V_\infty + LD_\Theta)^4}{|c| L_{22}^2 \sigma^2 \bar{L} L \varepsilon^4},$$

then, for $z_t := \nabla_\theta \mathbb{E}_{\xi \sim Q} \mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N_\mathcal{A}), \xi) - \nabla_\theta V(\theta^t)$*,*

$$\mathbb{E}[\|z_t\|] \leq \frac{L\varepsilon^2}{4D_V(2V_\infty + LD_\Theta)^2},$$

and

$$\mathbb{E}[\|\nabla_\phi \mathcal{L}_\mathcal{A}(\theta^t, \phi_\xi^t(N), \xi)\|] \leq \varepsilon.$$

Proof of Lemma D.6. Fixing a $\xi \in \Xi$, due to Lipschitz smoothness,

$$\begin{aligned} &\mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N), \xi) - \mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N-1), \xi) \\ &\leq \langle \nabla_\phi \mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N-1), \xi), \phi_\xi^t(N) - \phi_\xi^t(N-1) \rangle + \frac{L_{22}}{2} \|\phi_\xi^t(N) - \phi_\xi^t(N-1)\|^2. \end{aligned}$$

The inner loop updating rule ensures that when $\kappa_\mathcal{A} = \frac{1}{L_{21}}$, $\phi_\xi^t(N) - \phi_\xi^t(N-1) = \frac{1}{L_{21}} \hat{\nabla}_\phi J_\mathcal{A}(\theta_\xi^t, \phi_\xi^t(N-1), \xi)$. Plugging it into the inequality, we arrive at

$$\begin{aligned} &\mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N), \xi) - \mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N-1), \xi) \\ &\leq \frac{1}{L_{21}} \langle \nabla_\phi \mathcal{L}_\mathcal{D}(\theta^t, \phi_\xi^t(N-1), \xi), \hat{\nabla}_\phi J_\mathcal{A}(\theta_\xi^t, \phi_\xi^t(N-1), \xi) \rangle + \frac{L_{22}}{2L_{21}^2} \|\hat{\nabla}_\phi J_\mathcal{A}(\theta_\xi^t, \phi_\xi^t(N-1), \xi)\|^2. \end{aligned}$$

Therefore, we let $(\mathcal{F}_n^t)_{0 \leq n \leq N}$ be the filtration generated by $\sigma(\{\phi_\xi^t(\tau)\}_{\xi \in \Xi} | \tau \leq n)$ and take conditional expectations on $\mathcal{F}_n^t$:

$$\begin{aligned} &\mathbb{E}[V(\theta^t) - \ell_\mathcal{D}(\theta^t, \phi^t(N)) | \mathcal{F}_{N-1}^t] \leq V(\theta^t) - \ell_\mathcal{D}(\theta^t, \phi^t(N-1)) \\ &\mathbb{E}_\xi \left[\frac{1}{L_{21}} \langle \nabla_\phi \mathcal{L}_\mathcal{D}, \nabla_\phi J_\mathcal{A}(\theta_\xi^t, \phi_\xi^t(N-1), \xi) \rangle + \frac{L_{22}}{2L_{21}^2} \|\hat{\nabla}_\phi J_\mathcal{A}(\theta_\xi^t, \phi_\xi^t(N-1), \xi)\|^2 \right]. \end{aligned}$$

By variance-bias decomposition, and Assumption D.5 (b) and (c),

$$\begin{aligned}
&\mathbb{E}[\|\hat{\nabla}_\phi J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi)\|^2 | \mathcal{F}^t_{N-1}] \\
&= \mathbb{E}[\|\hat{\nabla}_\phi J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi) - \nabla_\phi J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi) + \nabla_\phi J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi)\|^2 | \mathcal{F}^t_{N-1}] \\
&= \mathbb{E}[\|(\hat{\nabla}_\phi - \nabla_\phi) J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi)\|^2 | \mathcal{F}^t_{N-1}] + \mathbb{E}[\|\nabla_\phi J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi)\|^2 | \mathcal{F}^t_{N-1}] \\
&\quad + \mathbb{E}[2\langle(\hat{\nabla}_\phi - \nabla_\phi) J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi), \nabla_\phi J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi)\rangle | \mathcal{F}^t_{N-1}] \\
&\le \frac{\sigma^2}{N_b} + \|\nabla_\phi J_{\mathcal{A}}(\theta^t_\xi, \phi^t_\xi(N-1), \xi)\|^2.
\end{aligned}$$

Applying the PL condition (Assumption 4.4), and Assumption D.5 (a) we obtain

$$\begin{aligned}
&\mathbb{E}[V(\theta^t) - \ell_{\mathcal{D}}(\theta, \phi^t(N)) | \phi^{N-1}] - V(\theta^t) - \ell_{\mathcal{D}}(\theta, \phi^t(N-1)) \\
&\le \mathbb{E}_\xi \left[\frac{1}{L_{21}} \langle \nabla_\phi \mathcal{L}_{\mathcal{D}}, \nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta^t, \phi^t_\xi(N-1), \xi)\rangle + \frac{L_{22}}{2L_{21}^2}\left(\frac{\sigma^2}{N_b} + \|\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta^t, \phi^t_\xi(N-1), \xi)\|^2\right)\right] \\
&= \mathbb{E}_\xi \left[-\frac{1}{2L_{22}} \|\nabla_\phi \mathcal{L}_{\mathcal{D}}\|^2 + \frac{1}{2L_{22}} \|\nabla_\phi (\mathcal{L}_{\mathcal{D}} + \frac{L_{22}}{L_{21}} \mathcal{L}_{\mathcal{A}})(\theta^t, \phi^t_\xi(N-1), \xi)\|^2 + \frac{L_{22}\sigma^2}{2L_{21}^2 N_b}\right] \\
&\le \frac{\mu}{cL_{21}} (\max_\phi \ell_{\mathcal{D}}(\theta^t, \phi) - \ell_{\mathcal{D}}(\theta^t, \phi^t(N-1))) + \frac{L_{22}\sigma^2}{2L_{21}^2 N_b},
\end{aligned}$$

rearranging the terms yields

$$\mathbb{E}[V(\theta^t) - \ell_{\mathcal{D}}(\theta^t, \phi^t(N)) | \mathcal{F}^t_n] \le \rho(V(\theta^t) - \ell_{\mathcal{D}}(\theta^t, \phi^t(N-1))) + \frac{L_{22}\sigma^2}{2L_{21}^2 N_b},$$

where we use the fact that $-\max_\phi \ell_{\mathcal{D}}(\theta^t, \phi) \le -V(\theta^t)$. Telescoping the inequalities from $\tau = 0$ to $\tau = N$, we arrive at

$$\mathbb{E}[V(\theta^t) - \ell_{\mathcal{D}}(\theta^t, \phi^t(N))] \le \rho^N (V(\theta^t) - \ell_{\mathcal{D}}(\theta^t, \phi^t(0))) + \frac{1-\rho^N}{1-\rho}\left(\frac{L_{22}\sigma^2}{2L_{21}^2 N_b}\right).$$

PL-condition implies quadratic growth, we also know that $V(\theta^t) - \ell_{\mathcal{D}}(\theta^t, \phi^t(N)) \le \mathbb{E}_\xi \frac{1}{2\mu} \|\nabla_\phi \mathcal{L}_{\mathcal{D}}(\theta^t, \phi^t_\xi(N), \xi)\|^2 \le \frac{1}{2\mu} G^2$, by Assumption 4.3,

$$\begin{aligned}
\|\phi^*_\xi(\theta^t) - \phi^t_\xi(N)\|^2 &\le \frac{2}{\mu}(\mathcal{L}_{\mathcal{A}}(\theta^t, \phi^*_\xi, \xi) - \mathcal{L}_{\mathcal{A}}(\theta^t, \phi^t_\xi(N), \xi)) \\
&\le \frac{2|c|}{\mu} \left| \mathcal{L}_{\mathcal{D}}(\theta^t, \phi^*_\xi, \xi) - \mathcal{L}_{\mathcal{D}}(\theta^t, \phi^t_\xi(N), \xi)\right|
\end{aligned}$$

Hence, with Jensen inequality and choice of $N_{\mathcal{A}}$ and N_b,

$$\begin{aligned}
\mathbb{E}[\|z_t\|] &= \mathbb{E}[\|\nabla_\theta V(\theta^t) - \mathbb{E}_\xi \nabla_\theta \mathcal{L}_{\mathcal{D}}(\theta^t, \phi^t_\xi(N_{\mathcal{A}}), \xi)\|] \\
&\le L_{12} \mathbb{E}[\|\phi^t_\xi(N_{\mathcal{A}}) - \phi^*_\xi\|] \\
&\le L_{12} \sqrt{\frac{2|c|}{\mu} \mathbb{E}[V(\theta^t) - \ell_{\mathcal{D}}(\theta^t, \phi^t(N_{\mathcal{A}}))]} \\
&\le L_{12} \sqrt{\frac{|c|}{\mu^2} \rho^{N_{\mathcal{A}}} G^2 + (1 - \rho^{N_{\mathcal{A}}}) \frac{|c| L_{22}^2 \sigma^2}{\mu L_{21}^2 N_b}}.
\end{aligned}$$

Now we adjust the size of $N_{\mathcal{A}}$ and N_b to make $\mathbb{E}[\|z_t\|]$ small enough, to this end, we set

$$\begin{aligned}
\rho^{N_{\mathcal{A}}} \frac{|c| G^2}{\mu^2} &\le \frac{\varepsilon^4 L^2}{32 D_V^2 (2V_\infty + LD_\Theta)^4 \bar{L}} \\
\frac{|c| L_{22}^2 \sigma^2}{L_{21}^2 N_b} &\le \frac{\varepsilon^4 L^2 \mu^2}{32 D_V^2 (2V_\infty + LD_\Theta)^4 \bar{L}},
\end{aligned}$$

which further indicates that

$$\begin{aligned} N_{\mathcal{A}} &\geq \frac{1}{\log \rho^{-1}} \log \frac{32 D_V^2 (2V_\infty + LD_\Theta)^4 \bar{L} |c| G^2}{L^2 \mu^2 \varepsilon^4} \\ N_b &\geq \frac{32 \mu L_{21}^2 D_V^2 (2V_\infty + LD_\Theta)^4}{|c| L_{22}^2 \sigma^2 \bar{L} L \varepsilon^4}. \end{aligned}$$

In the setting above, it is not hard to verify that

$$\mathbb{E}[\|z_t\|] \leq \frac{L\varepsilon^2}{4D_V(2V_\infty + LD_\Theta)^2} \leq \varepsilon.$$

Also note that $\|\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta^t, \phi_\xi^t(N_{\mathcal{A}}), \xi)\| = \|\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta^t, \phi_\xi^t(N_{\mathcal{A}}), \xi) - \nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta^t, \phi_\xi^*, \xi)\|$, given the proper choice of $N_{\mathcal{A}}$ and N_b, one has

$$\begin{aligned} &\mathbb{E}\|\nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta^t, \phi_\xi^t(N_{\mathcal{A}}), \xi) - \nabla_\phi \mathcal{L}_{\mathcal{A}}(\theta^t, \phi_\xi^*, \xi)\| \\ &\leq L_{21} \mathbb{E}[\|\phi_\xi^t(N_{\mathcal{A}}) - \phi_\xi^*\|] \leq \frac{L\varepsilon^2}{4D_V(2V_\infty + LD_\Theta)^2} \leq \varepsilon, \end{aligned}$$

which indicates the ξ-wise inner loop stability. □

Now we are ready to provide the convergence guarantee of the first-order outer loop.

Theorem D.7. *Under Assumption D.5, Assumption 4.3, and Assumption D.2, let the stepsizes be, $\kappa_{\mathcal{A}} = \frac{1}{L_{22}}$, $\kappa_{\mathcal{D}} = \frac{1}{L}$, if $N_{\mathcal{D}}, N_{\mathcal{A}}$, and N_b are large enough,*

$$N_{\mathcal{D}} \geq N_{\mathcal{D}}(\varepsilon) \sim \mathcal{O}(\varepsilon^{-2}) \quad N_{\mathcal{A}} \geq N_{\mathcal{A}}(\varepsilon) \sim \mathcal{O}(\log \varepsilon^{-1}), \quad N_b \geq N_b(\varepsilon) \sim \mathcal{O}(\varepsilon^{-4})$$

then there exists $t \in \mathbb{N}$ such that $(\theta^t, \{\phi_\xi^t(N_{\mathcal{A}})\}_{\xi \in \Xi})$ is ε-meta-FOSE.

Proof. According to the update rule of the outer loop, (here we omit the projection analysis for simplicity)

$$\theta^{t+1} - \theta^t = \frac{1}{L} \hat{\nabla}_\theta \ell_{\mathcal{D}}(\theta^t, \phi^t(N_{\mathcal{A}})),$$

one has, due to unbiasedness assumption, let $(\mathcal{F}_t)_{0 \leq t \leq N_{\mathcal{D}}}$ be the filtration generated by $\sigma(\theta^t | k \leq t)$

$$\begin{aligned} \mathbb{E}[\langle \nabla_\theta \ell_{\mathcal{D}}(\theta^t, \phi^t(N_{\mathcal{A}})), \theta^{t+1} - \theta^t \rangle | \mathcal{F}_t] &= \frac{1}{L} \mathbb{E}[\|\nabla_\theta \ell_{\mathcal{D}}(\theta^t, \phi^t(N_{\mathcal{A}}))\|^2 | \mathcal{F}_t] \\ &= L\mathbb{E}\|\theta^{t+1} - \theta^t\|^2 | \mathcal{F}_t], \end{aligned}$$

which leads to

$$\mathbb{E}[\langle \nabla_\theta \ell_{\mathcal{D}}(\theta^t, \phi^*), \theta^{t+1} - \theta^t \rangle | \mathcal{F}_t] = \mathbb{E}[\langle z_t, \theta^t - \theta^{t+1} \rangle | \mathcal{F}_t] + L\mathbb{E}[\|\theta^{t+1} - \theta^t\|^2 \|].$$

Since $V(\cdot)$ is L-Lipschitz smooth,

$$\begin{aligned} \mathbb{E}[V(\theta^t) - V(\theta^{t+1})] &\leq \mathbb{E}[\langle \nabla_\theta V(\theta^t), \theta^t - \theta^{t+1} \rangle] + \frac{L}{2} \mathbb{E}[\|\theta^{t+1} - \theta^t\|^2] \\ &\leq \mathbb{E}[\langle z_t, \theta^{t+1} - \theta^t \rangle] - \mathbb{E}[\langle \nabla_\theta \ell_{\mathcal{D}}(\theta^t, \phi^t(N_{\mathcal{A}})), \theta^{t+1} - \theta^t \rangle] + \frac{L}{2} \mathbb{E}[\|\theta^{t+1} - \theta^t\|^2] \\ &\leq \mathbb{E}[\langle z_t, \theta^{t+1} - \theta^t \rangle] - \frac{L}{2} \mathbb{E}[\|\theta^{t+1} - \theta^t\|^2]. \end{aligned} \tag{D12}$$

Fixing a $\theta \in \Theta$, let $e_t := \langle \nabla_\theta \ell_{\mathcal{D}}(\theta^t, \phi^t(N_{\mathcal{A}})), \theta - \theta^t \rangle$, we have

$$\begin{aligned} \mathbb{E}[e_t | \mathcal{F}_t] &= L\mathbb{E}[\langle \theta^{t+1} - \theta^t, \theta - \theta^t \rangle | \mathcal{F}_t] \\ &= \mathbb{E}[\langle \nabla_\theta \ell_{\mathcal{D}}(\theta^t, \phi^t(N_{\mathcal{A}})) - \nabla_\theta V(\theta^t), \theta^{t+1} - \theta^t \rangle + \langle \nabla_\theta V(\theta^t), \theta^{t+1} - \theta^t \rangle] \\ &\quad + L\mathbb{E}[\langle \theta^{t+1} - \theta^t, \theta - \theta^{t+1} \rangle] \\ &\leq \mathbb{E}[(\|z_t\| + V_\infty + LD_\Theta)\|\theta^{t+1} - \theta^t\|] \end{aligned} \tag{D13}$$

By the choice of N_b, we have, since $V_\infty = \max\{\max_\theta \|\nabla V(\theta)\|, 1\}$,

$$\mathbb{E}[\|z_t\|] \leq L_{12}\mathbb{E}[\|\phi^N - \phi^*\|] \leq \frac{L\varepsilon^2}{4D_V(2V_\infty + LD_\Theta)} \leq V_\infty.$$

Thus, the relation (D13) can be reduced to

$$\mathbb{E}[e_t] \leq (2V_\infty + LD_\Theta)\mathbb{E}[\|\theta^{t+1} - \theta^t\|].$$

Telescoping (D12) yields

$$-D_V \leq \mathbb{E}[V(\theta^0) - V(\theta^{N_\mathcal{D}})] \leq D_\Theta \sum_{t=0}^{T-1} \mathbb{E}[\|z_t\|] - \frac{L}{2(2V_\infty + LD_\Theta)^2}\mathbb{E}[\sum_{t=0}^{T-1} \mathbb{E}[e_t^2|\mathcal{F}_t].$$

Thus, setting $N_\mathcal{D} \geq \frac{4D_V(2V_\infty + LD_\Theta)^2}{L\varepsilon^2}$, and then by Lemma 4.6, we obtain that,

$$\frac{1}{N_\mathcal{D}} \sum_{t=0}^{N_\mathcal{D}-1} \mathbb{E}[e_t^2] \leq \frac{\varepsilon^2}{2} + \frac{2D_V(2V_\infty + LD_\Theta)^2}{LN_\mathcal{D}} \leq \varepsilon^2$$

which implies there exists $t \in \{0, \ldots, N_\mathcal{D} - 1\}$ such that $\mathbb{E}[e_t^2] \leq \varepsilon^2$.

□

E. Related Works

Poisoning/backdoor attacks and defenses in FL. Various methods for compromising the integrity of a federated learning target model have been introduced, including targeted poisoning attacks which strive to misclassify a particular group of inputs, as explored in the studies by (Bhagoji et al., 2019; Baruch et al., 2019). Other techniques, such as those studied by (Fang et al., 2020; Xie et al., 2020; Shejwalkar & Houmansadr, 2021), focus on untargeted attacks with the aim of diminishing the overall model accuracy. The majority of existing strategies often utilize heuristics-based methods (e.g., (Xie et al., 2020)), or they focus on achieving a short-sighted goal ((Fang et al., 2020; Shejwalkar & Houmansadr, 2021)). On the other hand, malicious participants can easily embed backdoors into the aggregated model while maintaining the model's performance on the main task with model replacement (Bagdasaryan et al., 2020). To enhance the surreptitious nature of these poisoned updates, triggers can be distributed across multiple cooperative malicious devices, as discussed by Xie et al. (2019)(Xie et al., 2019), and edge-case backdoors can be employed, as demonstrated by Wang et al. (2020) (Wang et al., 2020). However, these methods can be sub-optimal, especially when there's a need to adopt a robust aggregation rule. Additionally, these traditional methods typically demand access to the local updates of benign agents or precise parameters of the global model for the upcoming round (Xie et al., 2020; Fang et al., 2020) in order to enact a significant attack. In contrast to these methods, RL-based approach (Li et al., 2022a; Shen et al., 2021; Li et al., 2023a) employs reinforcement learning for the attack, reducing the need for extensive global knowledge while focusing on a long-term attack goal.

Several defensive strategies have been suggested to counter model poisoning attacks, which broadly fall into two categories: those based on robust aggregation and those centered around detection. Robust-aggregation-based defenses encompass techniques such as dimension-wise filtering. These methods treat each dimension of local updates individually, as explored in studies by (Bernstein et al., 2018; Yin et al., 2018). Another strategy is client-wise filtering, the goal of which is to limit or entirely eliminate the influence of clients who might harbor malicious intent. This approach has been examined in the works of (Blanchard et al., 2017; Pillutla et al., 2022; Sun et al., 2019). Some defensive methods necessitate the server having access to a minimal amount of root data, as detailed in the study by (Cao et al., 2021). Naive backdoor attacks are limited by even simple defenses like norm-bounding (Sun et al., 2019) and weak differential private (Geyer et al., 2017) defenses. Despite to the sophisticated design of state-of-the-art non-addaptive backdoor attacks against federated learning, post-training stage defenses (Wu et al., 2020; Nguyen et al., 2021; Rieger et al., 2022) can still effectively erase suspicious neurons/parameters in the backdoored model.

Multi-agent meta learning. Meta-learning, and in particular meta-reinforcement-learning aim to create a generalizable policy that can fast adapt to new tasks by exploiting knowledge obtained from past tasks (Duan et al., 2016; Finn et al., 2017).

The early use cases of meta-learning have been primarily single-agent tasks, such as few-shot classification and single-agent RL (Finn et al., 2017). A recent research thrust is to extend the meta-learning idea to multi-agent systems (MAS), which can be further categorized into two main directions: 1) distributed meta-learning in MAS (Kayaalp et al., 2022; Zhang et al., 2022); 2) meta-learning for generalizable equilibrium-seeking (Gupta et al., 2021; Harris et al., 2022; Zhao & Zhu, 2022; Ge et al., 2023). The former focuses on a decentralized operation of meta-learning over networked computation units to reduce computation/storage expenses. The latter is better aligned with the original motivation of meta-learning, which considers how to solve a new game (or multi-agent decision-making) efficiently by reusing past experiences from similar occasions.

In stark contrast to the existing research efforts, our work leverages the adaptability of meta-learning to address information asymmetry in dynamic games of incomplete information, leading to a new equilibrium concept: meta-equilibrium (see Definition 3.1). What distinguishes our work from the aforementioned ones is that 1) every entity in our meta-SG is a self-interest player acting rationally without any coordination protocol; 2) meta-learning in our work is beyond a mere solver for computing long-established equilibria (e.g., Stackelberg equilibrium); it brings up a non-Bayesian approach to processing information in dynamic games (see Appendix B), which is computationally more tractable. This meta-equilibrium notion has been proven effective in combating information asymmetry in adversarial FL. Since asymmetric information is prevalent in security studies, our work can shed light on other related problems.

First-order methods in bilevel optimization. The meta-SG problem in (2) amounts to a stochastic bilevel optimization. The meta-SL in Algorithm 2 admits a much simpler gradient estimation than what one would often observe in the bilevel optimization literature (Chen et al., 2023; Kwon et al., 2023), where the gradient estimate for the upper-level problem involves a Hessian inverse (Chen et al., 2023) or some first-order correction terms (Kwon et al., 2023). The key intuition behind this simplicity lies in the strict competitiveness (see Assumption 4.3). Informally speaking, (2) is more akin to minimax programming (Nouiehed et al., 2019; Li et al., 2022b), even though it is a general-sum game. However, the data-driven meta-adaptation within the value function in Equation (2) leads to a more involved gradient estimation. since the data induces extra randomness in addition to policy gradient estimates (Fallah et al., 2021). Perhaps, the closest to our work is (Li et al., 2022b) where the authors investigate adversarial meta-RL and arrive at a similar Stackelberg formulation. However, (Li et al., 2022b) considers a minimax relaxation to the original Stackelberg formulation, leading to simpler nonconvex programming. Our work is among the first endeavors to investigate fully first-order algorithms for solving general-sum Stackelberg games.