
Fault-Tolerant Federated Reinforcement Learning with Theoretical Guarantee

Flint Xiaofeng Fan^{1,3}, Yining Ma², Zhongxiang Dai¹, Wei Jing⁴,
Cheston Tan³, Bryan Kian Hsiang Low¹

¹Dept. of Computer Science, National University of Singapore, Republic of Singapore

²Dept. of ISEM, National University of Singapore, Republic of Singapore

³Institute for Infocomm Research, A*STAR, Republic of Singapore

⁴Alibaba DAMO Academy, Hangzhou, China

¹{xiaofeng, daizhongxiang, lowkh}@comp.nus.edu.sg, ²yiningma@u.nus.edu

³{stufanxf, cheston-tan}@i2r.a-star.edu.sg, ⁴jw334405@alibaba-inc.com

Abstract

The growing literature of *Federated Learning* (FL) has recently inspired *Federated Reinforcement Learning* (FRL) to encourage multiple agents to federatively build a *better* decision-making policy without sharing raw trajectories. Despite its promising applications, existing works on FRL fail to I) provide theoretical analysis on its convergence, and II) account for random system failures and adversarial attacks. Towards this end, we propose the first FRL framework the convergence of which is guaranteed and tolerant to less than half of the participating agents being random system failures or adversarial attackers. We prove that the sample efficiency of the proposed framework is guaranteed to improve with the number of agents and is able to account for such potential failures or attacks. All theoretical results are empirically verified on various RL benchmark tasks.

1 Introduction

Reinforcement learning (RL) has recently been applied to many real-world decision-making problems such as gaming, robotics, healthcare, etc. [1–3]. However, despite its impressive performances in simulation, RL often suffers from poor sample efficiency, which hinders its success in real-world applications [4, 5]. For example, when RL is applied to provide clinical decision support [3, 6, 7], its performance is limited by the number (i.e., sample size) of admission records possessed by a hospital, which cannot be synthetically generated [3]. As this challenge is usually faced by many agents (e.g., different hospitals), a natural solution is to encourage multiple RL agents to share their trajectories, to collectively build a better decision-making policy that one single agent can not obtain by itself. However, in many applications, raw RL trajectories contain sensitive information (e.g., the medical records contain sensitive information about patients) and thus sharing them is prohibited. To this end, the recent success of *Federated Learning* (FL) [8–11] has inspired the setting of *Federated Reinforcement Learning* (FRL) [12], which aims to *federatively* build a *better* policy from multiple RL agents without requiring them to share their raw trajectories. FRL is practically appealing for addressing the sample inefficiency of RL in real systems, such as autonomous driving [13], fast personalization [14], optimal control of IoT devices [15], robots navigation [16], and resource management in networking [17]. Despite its promising applications, FRL is faced by a number of major challenges, which existing works are unable to tackle.

Firstly, existing FRL frameworks are not equipped with theoretical convergence guarantee, and thus lack an assurance for the sample efficiency of practical FRL applications, which is a critical drawback

due to the high sampling cost of RL trajectories in real systems [4]. Unlike FL where training data can be collected offline, FRL requires every agent to sample trajectories by interacting with the environment during learning. However, interacting with real systems can be slow, expensive, or fragile. This makes it critical for FRL to be sample-efficient and hence highlights the requirement for convergence guarantee of FRL, without which no assurance on its sample efficiency is provided for practical applications. To fill this gap, we establish on recent endeavors in stochastic variance-reduced optimization techniques to develop a variance-reduced federated policy gradient framework, the convergence of which is guaranteed. We prove that the proposed framework enjoys a sample complexity of $O(1/\epsilon^{5/3})$ to converge to an ϵ -stationary point in the single-agent setting, which matches recent results of variance-reduced policy gradient [18, 19]. More importantly, the aforementioned sample complexity is guaranteed to *improve* at a rate of $O(1/K^{2/3})$ upon the federation of K agents. This guarantees that an agent achieves a better sample efficiency by joining the federation and benefits from more participating agents, which are highly desirable in FRL.

Another challenge inherited from FL is that FRL is vulnerable to random failures or adversarial attacks, which poses threats to many real-world RL systems. For example, robots may behave arbitrarily due to random hardware issues; clinical data may provide inaccurate records and hence create misleading trajectories [3]; autonomous vehicles, on which RL is commonly deployed, are subject to adversarial attacks [20]. As we will show in experiments, including such random failures or adversary agents in FRL can significantly deteriorate its convergence or even result in unlearnability. Of note, random failures and adversarial attacks in FL systems are being encompassed by the Byzantine failure model [21], which is considered as the most stringent fault formalism in distributed computing [22, 23] – a small fraction of agents may behave arbitrarily and possibly adversarially, with the goal of breaking or at least slowing down the convergence of the system. As algorithms proven to be correct in this setting are guaranteed to converge under arbitrary system behavior (e.g., exercising failures or being attacked) [9, 24], we study the fault tolerance of our proposed FRL framework using the Byzantine failure model. We design a gradient-based Byzantine filter on top of the variance-reduced federated policy gradient framework. We show that, when a certain percentage (denoted by $\alpha < 0.5$) of agents are Byzantine agents, the sample complexity of the FRL system is worsened by *only an additive term* of $O(\alpha^{4/3}/\epsilon^{5/3})$ (Section 4). Therefore, when $\alpha \rightarrow 0$, (i.e., an ideal system with zero chance of failure), the filter induces no impact on the convergence.

Contributions. In this paper, we study the *federated* reinforcement learning problem with theoretical guarantee in the potential presence of faulty agents. We introduce *Federated Policy Gradient with Byzantine Resilience* (FedPG-BR), the first FRL framework that is theoretically principled and practically effective for the FRL setting, accounting for random systematic failures and adversarial attacks. In particular, FedPG-BR (a) enjoys a guaranteed sample complexity which improves with more participating agents, and (b) is tolerant to the Byzantine fault in both theory and practice. We discuss the details of problem setting and the technical challenges (Section 3) and provide theoretical analysis of FedPG-BR (Section 4). We also demonstrate its empirical efficacy on various RL benchmark tasks (Section 5).

2 Background

Stochastic Variance-Reduced Gradient aims to solve $\min_{\theta \in \mathbb{R}^d} [J(\theta) \triangleq \frac{1}{B} \sum_{i=1}^B J_i(\theta)]$. Under the common assumption of all function components J_i being smooth and convex in θ , *gradient descent* (GD) achieves linear convergence in the number of iterations of parameter updates [25, 26]. However, every iteration of GD requires B gradient computations, which can be expensive for large B . To overcome this problem, *stochastic GD* (SGD) [27, 28] samples a single data point per iteration, which incurs lower per-iteration cost yet results in a sub-linear convergence rate [29]. For a better trade-off between convergence rate and per-iteration computational cost, the *stochastic variance-reduced gradient* (SVRG) method has been proposed, which reuses past gradient computations to reduce the variance of the current gradient estimate [30–33]. More recently, *stochastically controlled stochastic gradient* (SCSG) has been proposed for convex [34] or smooth non-convex objective function [35], to further reduce the computational cost of SVRG especially when required ϵ is small in finding ϵ -approximate solution. Refer to Appendix A.1 for more details on SVRG and SCSG.

Reinforcement Learning (RL) can be modelled as a discrete-time Markov Decision Process (MDP) [36]: $M \triangleq \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho\}$. \mathcal{S} represents the state space, \mathcal{A} is the action space, $\mathcal{P}(s'|s, a)$

defines the transition probability from state s to s' after taking action a , $\mathcal{R}(s, a) : \mathcal{S} \times \mathcal{A} \mapsto [0, R]$ is the reward function for state-action pair (s, a) and some constant $R > 0$, $\gamma \in (0, 1)$ is the discount factor, and ρ is the initial state distribution. An agent’s behavior is controlled by a policy π , where $\pi(a|s)$ defines the probability that the agent chooses action a at state s . We consider episodic MDPs with trajectory horizon H . A trajectory $\tau \triangleq \{s_0, a_0, s_1, a_1, \dots, s_{H-1}, a_{H-1}\}$ is a sequence of state-action pairs traversed by an agent following any stationary policy, where $s_0 \sim \rho$. $\mathcal{R}(\tau) \triangleq \sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t)$ gives the cumulative discounted reward for a trajectory τ .

Policy Gradient (PG) methods have achieved impressive successes in model-free RL [37, 38, etc.]. Compared with deterministic value-function based methods such as Q-learning, PG methods are generally more effective in high-dimensional problems and enjoy the flexibility of stochasticity. In PG, we use π_θ to denote the policy parameterized by $\theta \in \mathbb{R}^d$ (e.g., a neural network), and $p(\tau|\pi_\theta)$ to represent the trajectory distribution induced by policy π_θ . For brevity, we use θ to denote the corresponding policy π_θ . The performance of a policy θ can be measured by $J(\theta) \triangleq \mathbb{E}_{\tau \sim p(\cdot|\theta)}[\mathcal{R}(\tau)|M]$. Taking the gradient of $J(\theta)$ with respect to θ gives

$$\nabla_\theta J(\theta) = \int_{\tau} \mathcal{R}(\tau) \nabla_\theta p(\tau | \theta) d\tau = \mathbb{E}_{\tau \sim p(\cdot|\theta)} [\nabla_\theta \log p(\tau | \theta) \mathcal{R}(\tau) | M] \quad (1)$$

Then, the policy θ can be optimized by gradient ascent. Since computing (1) is usually prohibitive, stochastic gradient ascent is typically used. In each iteration, we sample a batch of trajectories $\{\tau_i\}_{i=1}^B$ using the current policy θ , and update the policy by $\theta \leftarrow \theta + \eta \widehat{\nabla}_B J(\theta)$, where η is the step size and $\widehat{\nabla}_B J(\theta)$ is an estimate of (1) using the sampled trajectories $\{\tau_i\}_{i=1}^B$: $\widehat{\nabla}_B J(\theta) = \frac{1}{B} \sum_{i=1}^B \nabla_\theta \log p(\tau_i | \theta) \mathcal{R}(\tau_i)$. The most common policy gradient estimators, such as REINFORCE [39] and GPOMDP [40], can be expressed as

$$\widehat{\nabla}_B J(\theta) = \frac{1}{B} \sum_{i=1}^B g(\tau_i | \theta) \quad (2)$$

where $\tau_i = \{s_0^i, a_0^i, s_1^i, a_1^i, \dots, s_{H-1}^i, a_{H-1}^i\}$ and $g(\tau_i | \theta)$ is an *unbiased* estimate of $\nabla_\theta \log p(\tau_i | \theta) \mathcal{R}(\tau_i)$. We provide formal definition of $g(\tau_i | \theta)$ in Appendix A.2.

SVRPG. A key issue for PG is the high variance of the estimator based on stochastic gradients (2) which results in slow convergence. Similar to SGD for finite-sum optimization, PG requires $O(1/\epsilon^2)$ trajectories to find an ϵ -stationary point such that $\mathbb{E}[\|\nabla J(\theta)\|^2] \leq \epsilon$ [19]. That is, PG typically requires a large number of trajectories to find a well-performing policy. To reduce the variance of the gradient estimator in PG (2), SVRG has been applied to policy evaluation [41, 42] and policy optimization [43]. The work of Papini et al. [18] has adapted the theoretical analysis of SVRG to PG to introduce the *stochastic variance-reduced PG* (SVRPG) algorithm. More recently, Xu et al. [19] has refined the analysis of SVRPG [18] and shown that SVRPG enjoys a sample complexity of $O(1/\epsilon^{5/3})$. These works have demonstrated both theoretically and empirically that SVRG is a promising approach to reduce the variance and thus improve the sample efficiency of PG methods.

Fault tolerance refers to the property that enables a computing system to continue operating properly without interruption when one or more of its workers fail. Among the many fault formalisms, the Byzantine failure model has a rich history in distributed computing [22, 23] and is considered as the most stringent fault formalism in fault-tolerant FL system design [9, 24]. Originated from the *Byzantine generals problem* [21], the Byzantine failure model allows an α -fraction (typical $\alpha < 0.5$) of workers to behave arbitrarily and possibly adversarially, with the goal of breaking or at least slowing down the convergence of the algorithm. As algorithms proven to be resilient to the Byzantine failures are guaranteed to converge under arbitrary system behavior (hence fault-tolerant) [22, 23], it has motivated a significant interest in providing distributed *supervised learning* with Byzantine resilience guarantees [e.g., 44–49]. However, there is yet no existing work studying the correctness of Byzantine resilience in the context of FRL.

3 Fault-tolerant federated reinforcement learning

3.1 Problem statement

Our problem setting is similar to that of FL [8] where a central server is assumed to be trustworthy and governs the federation of K distributed agents $k \in \{1, \dots, K\}$. In each round $t \in \{1, \dots, T\}$,

the central server broadcasts its parameter θ_0^t to all agents. Each agent then independently samples a batch of trajectories $\{\tau_{t,i}^{(k)}\}_{i=1}^{B_t}$ by interacting with the environment using the obtained policy, e.g., $\{\tau_{t,i}^{(k)}\}_{i=1}^{B_t} \sim p(\cdot|\theta_0^t)$. However, different from FL where each agent computes the parameter updates and sends the updated parameter to the server for aggregation [8], agents in our setup do not compute the updates locally, but instead send the gradient computed w.r.t. their local trajectories $\mu_t^{(k)} \triangleq \widehat{\nabla}_{B_t} J(\theta_0^t)$ directly to the server. The server then aggregates the gradients, performs a policy update step, and starts a new round of federation.

Of note, every agent including the server is operating in a separate copy of the MDP. No exchange of raw trajectories is required, and no communication between any two agents is allowed. To account for potential failures and attacks, we allow an α -fraction of agents to be Byzantine agents with $\alpha \in [0, 0.5)$. That is, in each round t , a good agent always sends its computed $\mu_t^{(k)}$ back to the server, while a Byzantine agent may return any arbitrary vector.¹ The server has no information regarding whether Byzantine agents exist and cannot track the communication history with any agent. In every round, the server can only access the K gradients received from agents, and thereby uses them to detect Byzantine agents so that it only aggregates the gradients from those agents that are believed to be non-Byzantine agents.

Notations. Following the notations of SCSG [35], we use θ_0^t to denote the server’s initial parameter in round t and θ_n^t to represent the updated parameter at the n -th step in round t . $\tau_{t,i}^{(k)}$ represents agent k ’s i -th trajectory sampled using θ_0^t . $\|\cdot\|$ denotes Euclidean norm and Spectral norms for vectors and matrices, respectively. $O(\cdot)$ hides all constant terms.

3.2 Technical challenges

There is an emerging interest in Byzantine-resilient distributed *supervised learning* [e.g., 44–50]. However, a direct application of those works to FRL is not possible due to that the objective function $J(\theta)$ of RL, which is conditioned on $\tau \sim p(\cdot|\theta)$, is different from the supervised classification loss seen in the aforementioned works, resulting in the following issues:

Non-stationarity: unlike in supervised learning, the distribution of RL trajectories is affected by the value of the policy parameter which changes over time (e.g., $\tau \sim p(\cdot|\theta)$). We deal with the non-stationarity using importance sampling [51] (Section 3.3).

Non-concavity: the objective function $J(\theta)$ is typically non-concave. To derive the theoretical results accounting for the non-concavity, we need the L -smoothness assumption on $J(\theta)$, which is a reasonable assumption and commonly made in the literature [52] (Section 4). Hence we aim to find an ϵ -approximate solution (i.e., a commonly used objective in non-convex optimization):

Definition 1 (ϵ -approximate solution). *A point θ is called ϵ -stationary if $\|\nabla J(\theta)\|^2 \leq \epsilon$. Moreover, the algorithm is said to achieve an ϵ -approximate solution in t rounds if $\mathbb{E}[\|\nabla J(\theta)\|^2] \leq \epsilon$, where the expectation is with respect to all randomness of the algorithm until round t .*

High variance in gradient estimation: the high variance in estimating (2) renders the FRL system vulnerable to variance-based attacks which conventional Byzantine-resilient optimization works fail to defend [47]. To combat this issue, we adapt the SCSG optimization [35] to federated policy gradient for a refined control over the estimation variance, hence enabling the following assumption which we exploit to design our Byzantine filtering step:

Assumption 2 (On bounded variance of the gradient estimator). *There is a constant σ such that $\|g(\tau|\theta) - \nabla J(\theta)\| \leq \sigma$ for any $\tau \sim p(\tau|\theta)$ for all policy π_θ .*

Remark. *Assumption 2 is also seen in Byzantine-resilient optimization [46, 48] and may be relaxed to $\mathbb{E}\|g(\tau|\theta) - \nabla J(\theta)\| \leq \theta$ which is a standard assumption commonly used in stochastic non-convex optimization [e.g. 35, 53]. In this work, the value of σ is the maximum difference between optimal gradient $\nabla J(\theta)$ and the gradient estimate $g(\tau|\theta)$ w.r.t. any trajectories induced by policy π_θ . For complex real-world problems with continuous, high-dimensional controls, σ may be upper-bounded, provided that the MDP is Lipschitz continuous Pirota et al. [52]. The deviation can be obtained by referring to Proposition 2 of Pirota et al. [52].²*

¹A Byzantine agent may not be Byzantine in every round.

²The value of σ can be estimated at the server.

3.3 Algorithm description

The pseudocode for the proposed *Federated Policy Gradient with Byzantine Resilience* (FedPG-BR) is shown in Algorithm 1. FedPG-BR starts with a randomly initialized parameter $\tilde{\theta}_0$ at the server. At the beginning of the t -th round, the server keeps a snapshot of its parameter from the previous round (i.e., $\theta_0^t \leftarrow \tilde{\theta}_{t-1}$) and broadcasts this parameter to all agents (line 3). Every (good) agent k samples B_t trajectories $\{\tau_{t,i}^{(k)}\}_{i=1}^{B_t}$ using the policy θ_0^t (line 5), computes a gradient estimate $\mu_t^{(k)} \triangleq 1/B_t \sum_{i=1}^{B_t} g(\tau_{t,i}^{(k)} | \theta_0^t)$ where g is either the REINFORCE or the GPOMDP estimator (line 6), and sends $\mu_t^{(k)}$ back to the server. For a Byzantine agent, it can send an arbitrary vector instead of the correct gradient estimate. After all gradients are received, the server performs the *Byzantine filtering* step, and then computes the batch gradient μ_t by averaging those gradients that the server believes are from non-Byzantine agents (line 7). For better clarity, we present the subroutine **FedPG-Aggregate** for Byzantine filtering and gradient aggregation in Algorithm 1.1, which we discuss in detail separately.

The aggregation is then followed by the SCSG inner loop [35] with N_t steps, where N_t is sampled from a geometric distribution with parameter $\frac{B_t}{B_t+b_t}$ (line 8). At step n , the server *independently* samples b_t ($b_t \ll B_t$) trajectories $\{\tau_{n,j}^t\}_{j=1}^{b_t}$ using its current policy θ_n^t (line 10), and then updates the policy parameter θ_n^t based on the following semi-stochastic gradient (lines 11 and 12):

$$v_n^t \triangleq \frac{1}{b_t} \sum_{j=1}^{b_t} [g(\tau_{n,j}^t | \theta_n^t) - \omega(\tau_{n,j}^t | \theta_n^t, \theta_0^t) g(\tau_{n,j}^t | \theta_0^t)] + \mu_t. \quad (3)$$

The last two terms serve as a correction to the gradient estimate to reduce variance and improve the convergence rate of Algorithm 1. Of note, the semi-stochastic gradient above (3) differs from that used in SCSG due to the additional term of $\omega(\tau | \theta_n^t, \theta_0^t) \triangleq p(\tau | \theta_0^t) / p(\tau | \theta_n^t)$. This term is known as the *importance weight* from $p(\tau | \theta_n^t)$ to $p(\tau | \theta_0^t)$ to account for the aforementioned non-stationarity of the distribution in RL [18, 19]. In particular, directly computing $g(\tau_{n,j}^t | \theta_0^t)$ results in a biased estimation because the trajectories $\{\tau_{n,j}^t\}_{j=1}^{b_t}$ are sampled from the policy θ_n^t instead of θ_0^t . We prove in Lemma 8 (Appendix E) that this importance weight results in an unbiased estimate of the gradient, i.e., $\mathbb{E}_{\tau \sim p(\cdot | \theta_n)}[\omega(\tau | \theta_n, \theta_0) g(\tau | \theta_0)] = \nabla J(\theta_0)$.

Here we describe the details of our Byzantine filtering step (i.e., the subroutine **FedPG-Aggregate** in Algorithm 1.1), which is inspired by the works of Alistarh et al. [46] and Khanduri et al. [48] in distributed supervised learning. In any round t , we use \mathcal{G} to denote the set of true good agents and use \mathcal{G}_t to denote the set of agents that are believed to be good by the server. Our Byzantine filtering

Algorithm 1 FedPG-BR

- 1: **Input:** $\tilde{\theta}_0 \in \mathbb{R}^d$, batch size B_t , mini batch size b_t , step size η_t
 - 2: **for** $t = 1$ **to** T **do**
 - 3: $\theta_0^t \leftarrow \tilde{\theta}_{t-1}$; broadcast to all agents
 - 4: **for** $k = 1$ **to** K **do**
 - 5: Sample B_t trajectories $\{\tau_{t,i}^{(k)}\}_{i=1}^{B_t}$ from $p(\cdot | \theta_0^t)$
 - 6: $\mu_t^{(k)} \triangleq \begin{cases} \frac{1}{B_t} \sum_{i=1}^{B_t} g(\tau_{t,i}^{(k)} | \theta_0^t) & \text{for } k \in \mathcal{G} \\ * & \text{for } k \notin \mathcal{G} \end{cases}$; push $\mu_t^{(k)}$ to server
 - 7: $\mu_t \leftarrow \mathbf{FedPG-Aggregate}(\{\mu_t^{(k)}\}_{k=1}^K)$
 - 8: Sample $N_t \sim \text{Geom}(\frac{B_t}{B_t+b_t})$
 - 9: **for** $n = 0$ **to** $N_t - 1$ **do**
 - 10: Sample b_t trajectories $\{\tau_{n,j}^t\}_{j=1}^{b_t}$ from $p(\cdot | \theta_n^t)$
 - 11: $v_n^t \triangleq \frac{1}{b_t} \sum_{j=1}^{b_t} [g(\tau_{n,j}^t | \theta_n^t) - \omega(\tau_{n,j}^t | \theta_n^t, \theta_0^t) g(\tau_{n,j}^t | \theta_0^t)] + \mu_t$
 - 12: $\theta_{n+1}^t = \theta_n^t + \eta_t v_n^t$
 - 13: $\tilde{\theta}_t \leftarrow \theta_{N_t}^t$
 - 14: **Output:** $\tilde{\theta}_a$ uniformly randomly picked from $\{\tilde{\theta}_t\}_{t=1}^T$
-

Algorithm 1.1 FedPG-Aggregate

- 1: **Input:** Gradient estimates from K agents in round t : $\{\mu_t^{(k)}\}_{k=1}^K$, variance bound σ , filtering threshold $\mathfrak{T}_\mu \triangleq 2\sigma\sqrt{\frac{V}{B_t}}$, where $V \triangleq 2\log(\frac{2K}{\delta})$ and $\delta \in (0, 1)$
 - 2: $S_1 \triangleq \{\mu_t^{(k)}\}$ where $k \in [K]$ s.t. $\left| \left\{ k' \in [K] : \left\| \mu_t^{(k')} - \mu_t^{(k)} \right\| \leq \mathfrak{T}_\mu \right\} \right| > \frac{K}{2}$
 - 3: $\mu_t^{\text{mom}} \leftarrow \operatorname{argmin}_{\mu_t^{(\tilde{k})}} \left\| \mu_t^{(\tilde{k})} - \operatorname{mean}(S_1) \right\|$ where $\tilde{k} \in S_1$
 - 4: **R1:** $\mathcal{G}_t \triangleq \left\{ k \in [K] : \left\| \mu_t^{(k)} - \mu_t^{\text{mom}} \right\| \leq \mathfrak{T}_\mu \right\}$
 - 5: **if** $|\mathcal{G}_t| < (1 - \alpha)K$ **then**
 - 6: $S_2 \triangleq \{\mu_t^{(k)}\}$ where $k \in [K]$ s.t. $\left| \left\{ k' \in [K] : \left\| \mu_t^{(k')} - \mu_t^{(k)} \right\| \leq 2\sigma \right\} \right| > \frac{K}{2}$
 - 7: $\mu_t^{\text{mom}} \leftarrow \operatorname{argmin}_{\mu_t^{(\tilde{k})}} \left\| \mu_t^{(\tilde{k})} - \operatorname{mean}(S_2) \right\|$ where $\tilde{k} \in S_2$
 - 8: **R2:** $\mathcal{G}_t \triangleq \left\{ k \in [K] : \left\| \mu_t^{(k)} - \mu_t^{\text{mom}} \right\| \leq 2\sigma \right\}$
 - 9: **Return:** $\mu_t \triangleq \frac{1}{|\mathcal{G}_t|} \sum_{k \in \mathcal{G}_t} \mu_t^{(k)}$
-

consists of two filtering rules denoted by R1 (lines 2-4) and R2 (lines 6-8). R2 is more intuitive to understand, so we start by introducing R2. Firstly, in line 6, the server constructs a set S_2 of *vector medians* [46] where each element of S_2 is chosen from $\{\mu_t^{(k)}\}_{k=1}^K$ if it is close (within 2σ in Euclidean distance) to more than $K/2$ elements. Next, the server finds a *Mean of Median* vector μ_t^{mom} from S_2 , which is defined as any $\mu_t^{(\tilde{k})} \in S_2$ that is the closest to the mean of the vectors in S_2 . After μ_t^{mom} is selected, the server can construct the set \mathcal{G}_t by filtering out any $\mu_t^{(k)}$ whose distance to μ_t^{mom} is larger than 2σ (line 8). This filtering rule is designed based on Assumption 2 which implies that the maximum distance between any two good agents is 2σ , and our assumption that at least half of the agents are good (i.e., $\alpha < 0.5$). We show in Appendix D that under these two assumptions, R2 guarantees that all good agents are included in \mathcal{G}_t (i.e., $|\mathcal{G}_t| \geq (1 - \alpha)K$). We provide a graphical illustration (Fig. 4 in Appendix D) on that if any Byzantine agent is included in \mathcal{G}_t , its distance to the true gradient $\nabla J(\theta_0^t)$ is at most 3σ , which ensures that its impact on the algorithm is limited. Note that the pairwise computation among the weights of all the agents can be implemented using the Euclidean Distance Matrix Trick [54].

R1 (lines 2-4) is designed in a similar way: R1 ensures that all good agents are *highly likely* to be included in \mathcal{G}_t by exploiting Lemma 14 (Appendix E) to guarantee that *with high probability*, all good agents are *concentrated in a smaller region*. That is, define $V \triangleq 2\log(2K/\delta)$ and $\delta \in (0, 1)$, then with probability of $\geq 1 - \delta$, the maximum distance between any two good agents is $\mathfrak{T}_\mu \triangleq 2\sigma\sqrt{V/B_t}$. Having all good agents in a smaller region improves the filtering strategy, because it makes the Byzantine agents less likely to be selected and reduces their impact even if they are selected. Therefore, R1 is applied first such that if R1 fails to include all good agents in \mathcal{G}_t (line 4) which happens with probability $< \delta$, R2 is then employed as a backup to ensure that \mathcal{G}_t always include all good agents. Therefore, these two filtering rules ensure in any round t that (a) gradients from good agents are never filtered out, and that (b) if gradients from Byzantine agents are not filtered out, their impact is limited since their maximum distance to $\nabla J(\theta_0^t)$ is bounded by 3σ .

4 Theoretical results

Here, we firstly put in place a few assumptions required for our theoretical analysis, all of which are common in the literature.

Assumption 3 (On policy derivatives). *Let $\pi_\theta(a|s)$ be the policy of an agent at state s . There exist constants $G, M > 0$ s.t. the log-density of the policy function satisfies, for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$*

$$|\nabla_\theta \log \pi_\theta(a|s)| \leq G, \quad \|\nabla_\theta^2 \log \pi_\theta(a|s)\| \leq M, \quad \forall a \in \mathcal{A}, \forall s \in \mathcal{S}$$

Assumption 3 provides the basis for the smoothness assumption on the objective function $J(\theta)$ commonly used in non-convex optimization [32, 33] and also appears in Papini et al. [18], Xu et al. [19]. Specifically, Assumption 3 implies:

Proposition 4 (On function smoothness). *Under Assumption 3, $J(\boldsymbol{\theta})$ is L -smooth with $L \triangleq HR(M + HG^2)/(1 - \gamma)$. Let $g(\tau|\boldsymbol{\theta})$ be the REINFORCE or GPOMDP gradient estimators. Then for all $\boldsymbol{\theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, it holds that*

$$\|g(\tau|\boldsymbol{\theta})\| \leq C_g, \quad \|g(\tau|\boldsymbol{\theta}_1) - g(\tau|\boldsymbol{\theta}_2)\| \leq L_g \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

where $L_g \triangleq HM(R + |C_b|)/(1 - \gamma)$, $C_g \triangleq HG(R + |C_b|)/(1 - \gamma)$ and C_b is the baseline reward.

Proposition 4 is important for deriving a fast convergence rate and its proof can be found in Xu et al. [19]. Next, we need an assumption on the variance of the importance weights (Section 3).

Assumption 5 (On variance of the importance weights). *There exists a constant $W < \infty$ such that for each policy pairs in Algorithm 1, it holds*

$$\text{Var}(\omega(\tau|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) \leq W, \quad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d, \tau \sim p(\cdot|\boldsymbol{\theta}_1)$$

Assumption 5 has also been made by Papini et al. [18], Xu et al. [19]. Now we present the convergence guarantees for our FedPG-BR algorithm:

Theorem 6 (Convergence of FedPG-BR). *Assume uniform initial state distribution across agents, and the gradient estimator is set to be the REINFORCE or GPOMDP estimator. Under Assumptions 2, 3, and 5, if we choose $\eta_t \leq \frac{1}{2\Psi B_t^{2/3}}$, $b_t = 1$, and $B_t = B \geq 4\Phi L^{-2}$ where $\Phi \triangleq L_g + C_g^2 C_w$, $\Psi \triangleq (L(L_g + C_g^2 C_w))^{1/3}$, L, L_g, C_g are defined in Proposition 4 and C_w is defined in Lemma 9, $\delta \in (0, 1)$ such that $e^{\frac{\delta B_t}{2(1-2\delta)}} \leq \frac{2K}{\delta} \leq e^{\frac{B_t}{2}}$ and $\delta \leq \frac{1}{5KB_t}$, then the output $\tilde{\boldsymbol{\theta}}_\alpha$ of Algorithm 1 satisfies*

$$\mathbb{E}[\|\nabla J(\tilde{\boldsymbol{\theta}}_\alpha)\|^2] \leq \frac{2\Psi [J(\tilde{\boldsymbol{\theta}}^*) - J(\tilde{\boldsymbol{\theta}}_0)]}{TB^{1/3}} + \frac{8\sigma^2}{(1-\alpha)^2 KB} + \frac{96\alpha^2 \sigma^2 V}{(1-\alpha)^2 B}$$

where $0 \leq \alpha < 0.5$ and $\tilde{\boldsymbol{\theta}}^*$ is a global maximizer of J .

This theorem leads to many interesting insights. When $K = 1, \alpha = 0$, Theorem 6 reduces to $\mathbb{E}\|\nabla J(\tilde{\boldsymbol{\theta}}_\alpha)\|^2 \leq 2\Psi[J(\tilde{\boldsymbol{\theta}}^*) - J(\tilde{\boldsymbol{\theta}}_0)]/TB^{1/3} + 8\sigma^2/B$. The second term here $O(1/B)$, which also shows up in SVRPG [18, 19], results from the full gradient approximation in Equation (2) in each round. In this case, our theorem implies that $\mathbb{E}\|\nabla J(\tilde{\boldsymbol{\theta}}_\alpha)\|^2 = O(\Psi[J(\tilde{\boldsymbol{\theta}}^*) - J(\tilde{\boldsymbol{\theta}}_0)]/TB^{1/3})$ which is consistent with SCSG for L -smooth non-convex objective functions [35]. Moreover, using $\mathbb{E}[\text{Traj}(\epsilon)]$ to denote the expected number of trajectories required by each agent to achieve $\mathbb{E}\|\nabla J(\tilde{\boldsymbol{\theta}}_\alpha)\|^2 \leq \epsilon$, Theorem 6 leads to:

Corollary 7 (Sample complexity of FedPG-BR). *Under the same assumptions as Theorem 6, let $\epsilon > 0$, we have: (i) $\mathbb{E}[\text{Traj}(\epsilon)] = O(\frac{1}{\epsilon^{5/3} K^{2/3}} + \frac{\alpha^{4/3}}{\epsilon^{5/3}})$; (ii) When $\alpha = 0$, we have $\mathbb{E}[\text{Traj}(\epsilon)] = O(\frac{1}{\epsilon^{5/3} K^{2/3}})$; (iii) When $K = 1$, we have $\mathbb{E}[\text{Traj}(\epsilon)] = O(\frac{1}{\epsilon^{5/3}})$*

We present a straightforward comparison of the sample complexity of related works in Table 1. Both REINFORCE and GPOMDP have a sample complexity of $O(1/\epsilon^2)$ since they use stochastic gradient-based optimization. Xu et al. [19] has made a refined analysis of SVRPG to improve its sample

Table 1: Sample complexities of relevant works to achieve $\mathbb{E}\|\nabla J(\boldsymbol{\theta})\|^2 \leq \epsilon$.

SETTINGS	METHODS	COMPLEXITY
$K = 1$	REINFORCE [39]	$O(1/\epsilon^2)$
	GPOMDP [40]	$O(1/\epsilon^2)$
	SVRPG [18]	$O(1/\epsilon^2)$
	SVRPG [19]	$O(1/\epsilon^{5/3})$
	FedPG-BR	$O(1/\epsilon^{5/3})$
$K > 1, \alpha = 0$	FedPG-BR	$O(\frac{1}{\epsilon^{5/3} K^{2/3}})$
$K > 1, \alpha > 0$	FedPG-BR	$O(\frac{1}{\epsilon^{5/3} K^{2/3}} + \frac{\alpha^{4/3}}{\epsilon^{5/3}})$

complexity from $O(1/\epsilon^2)$ [18] to $O(1/\epsilon^{5/3})$. Corollary 7 (iii) reveals that the sample complexity of FedPG-BR in the single-agent setup agrees with that of SVRPG derived by Xu et al. [19].

When $K > 1$, $\alpha = 0$, Corollary 7 (ii) implies that the total number of trajectories required by each agent is upper-bounded by $O(1/(\epsilon^{5/3}K^{2/3}))$. This result gives us the theoretical grounds to encourage more agents to participate in the federation, since the number of trajectories each agent needs to sample decays at a rate of $O(1/K^{2/3})$. This guaranteed improvement in sample efficiency is highly desirable in practical systems with a large number of agents.

Next, for a more realistic system where an α -fraction ($\alpha > 0$) of the agents are Byzantine agents, Corollary 7 (i) assures us that the total number of trajectories required by each agent will be increased by only an additive term of $O(\alpha^{4/3}/\epsilon^{5/3})$. This term is unavoidable due to the presence of Byzantine agents in FRL systems. However, the bound implies that the impact of Byzantine agents on the overall convergence is limited, which aligns with the discussions on our Byzantine filtering strategy (Section 3.3), and will be empirically verified in our experiments. Moreover, the impact of Byzantine agents on the convergence vanishes when $\alpha \rightarrow 0$. That is, when the system is ideal ($\alpha = 0$), our Byzantine filtering step induces no effect on the convergence.

5 Experiments

We evaluate the empirical performances of FedPG-BR with and without Byzantine agents on different RL benchmarks, including CartPole balancing [55], LunarLander, and the 3D continuous locomotion control task of Half-Cheetah [56]. In all experiments, we measure the performance online such that in each iteration, we evaluate the current policy of the server by using it to independently interact with the test MDP for 10 trajectories and reporting the mean returns. Each experiment is independently repeated 10 times with different random seeds and policy initializations, both of which are shared among all algorithms for fair comparisons. The results are averaged over the 10 independent runs with 90% bootstrap confidence intervals. Due to space constraints, some experimental details are deferred to Appendix F.

Performances in ideal systems with $\alpha = 0$. We firstly evaluate the performances of FedPG-BR in ideal systems with $\alpha = 0$, i.e., no Byzantine agents. We compare FedPG-BR ($K = 1, 3, 10$) with vanilla policy gradient using GPOMDP³ and SVRPG. The results in all three tasks are plotted in Fig. 1. The figures show that FedPG-BR ($K = 1$) and SVRPG perform comparably, both outperforming GPOMDP. This aligns with the results in Table 1 showing that FedPG-BR ($K = 1$) and SVRPG share the same sample complexity, and both are provably more sample-efficient than GPOMDP. Moreover, the performance of FedPG-BR is improved significantly with the federated of only $K = 3$ agents, and improved even further when $K = 10$. This corroborates our theoretical insights implying that the federation of more agents (i.e., larger K) improves the sample efficiency of FedPG-BR (Section 4), and verifies the practical performance benefit offered by the participation of more agents.

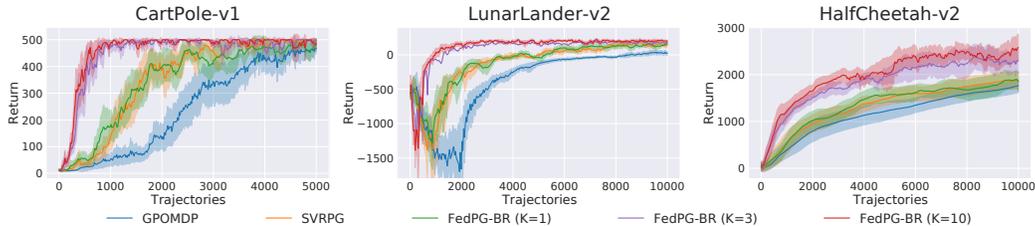


Figure 1: Performance of FedPG-BR in ideal systems with $\alpha = 0$ for the three tasks.

Performances in practical systems with $\alpha > 0$. Next, we investigate the impact of Byzantine agents (i.e., random failures or adversarial attacks) on the sample efficiency, which is critical for the practical deployment of FRL algorithms. In this experiment, we use $K = 10$ agents among which 3 are Byzantine agents, and we simulate different types of Byzantine failures: (a) *Random Noise (RN)*: each Byzantine agent sends a random vector to the server; (b) *Random Action (RA)*: every Byzantine agent ignores the policy from the server and takes actions randomly, which is used to

³Since GPOMDP has been repeatedly found to be comparable to or better than REINFORCE [18, 19].

simulate random system failures (e.g., hardware failures) and results in false gradient computations since the trajectories are no longer sampled according to the policy; (c) *Sign Flipping (SF)*: each Byzantine agent computes the correct gradient but sends the scaled negative gradient (multiplied by -2.5), which is used to simulate adversarial attacks aiming to manipulate the direction of policy update at the server.

For comparison, we have adapted both GPOMDP and SVRPG to the FRL setting (pseudocode is provided in Appendix A.3). Fig. 2 shows the results using the HalfCheetah task. We have also included the performances of GPOMDP, SVRPG and FedPG-BR in the single-agent setting ($K = 1$) as dotted-line (mean value of 10 independent runs) for reference. The figures show that for both GPOMDP and SVRPG, the 3 Byzantine agents cause the performance of their federated versions to be worse than that in the single-agent setting. Particularly, RA agents (middle figure) render GPOMDP and SVRPG unlearnable, i.e., unable to converge at all. In contrast, our FedPG-BR is robust against all three types of Byzantine failures. That is, FedPG-BR ($K = 10$ $B = 3$) with 3 Byzantine agents still significantly outperforms the single-agent setting, and more importantly, *performs comparably to FedPG-BR ($K = 10$) with 10 good agents*. This is because our Byzantine filtering strategy can effectively filter out those Byzantine agents. These results demonstrate that even in practical systems which are subject to random failures or adversarial attacks, FedPG-BR is still able to deliver superior performances. This provides an assurance on the reliability of our FedPG-BR algorithm to promote its practical deployment, and significantly improves the practicality of FRL. The results for the CartPole and LunarLander tasks, which yield the same insights as discussed here, can be found in Appendix G.

Performance of FedPG-BR against FedPG attack. We have discussed (Section 3.3) and shown through theoretical analysis (Section 4) that even when our Byzantine filtering strategy fails, the impact of the Byzantine agents on the performance of our algorithm is still limited. Here we verify this empirically. To this end, we design a new type of Byzantine agents who have perfect knowledge about our Byzantine filtering strategy, and call it *FedPG attacker*. The goal of FedPG attackers are to collude with each other to attack our algorithm without being filtered out. To achieve this, FedPG attackers firstly estimate $\nabla J(\theta_0^t)$ using the mean of their gradients $\bar{\mu}_t$, and estimate σ by calculating the maximum Euclidean distance between the gradients of any two FedPG attackers as $2\bar{\sigma}$. Next, all FedPG attackers send the vector $\bar{\mu}_t + 3\bar{\sigma}$ to the server. Recall we have discussed in Section 3.3 that if a Byzantine agent is not filtered out, its distance to the true gradient $\nabla J(\theta_0^t)$ is at most $3\bar{\sigma}$. Therefore, if $\bar{\mu}_t$ and $\bar{\sigma}$ are estimated accurately, the vectors from the FedPG attackers can exert negative impact on the convergence while still evading our Byzantine filtering.

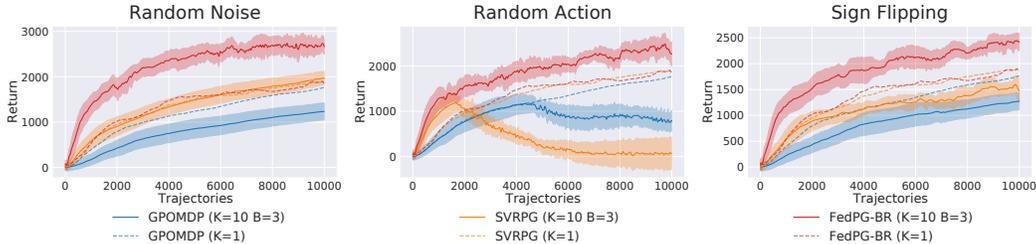


Figure 2: Performance of FedPG-BR in practical systems with $\alpha > 0$ for HalfCheetah. Each subplot corresponds to a different type of Byzantine failure exercised by the 3 Byzantine agents.

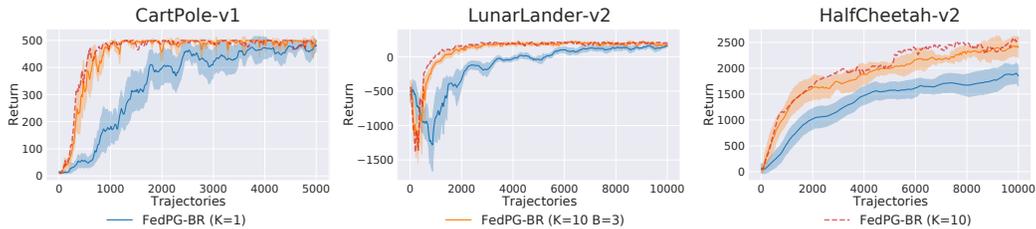


Figure 3: Performance of FedPG-BR in practical systems against FedPG attack.

We again use $K = 10$ agents, among which 3 are FedPG attackers. The results (Fig. 3) show that in all three tasks, even against such strong attackers with perfect knowledge of our Byzantine filtering strategy, FedPG-BR ($K = 10$ $B = 3$) still manages to significantly outperform FedPG-BR ($K = 1$) in the single-agent setting. Moreover, the performance of FedPG-BR ($K = 10$ $B = 3$) is only marginally worsened compared with FedPG-BR ($K = 10$) with 10 good agents. This corroborates our theoretical analysis showing that although we place no assumption on the gradients sent by the Byzantine agents, they only contribute an additive term of $O(\alpha^{4/3}/\epsilon^{5/3})$ to the sample complexity (Section 4). These results demonstrate the empirical robustness of FedPG-BR even against strong attackers, hence further highlighting its practical reliability.

6 Conclusion and future work

Federation is promising in boosting the sample efficiency of RL agents, without sharing their trajectories. Due to the high sampling cost of RL, the design of FRL systems appeals for theoretical guarantee on its convergence which is, however, vulnerable to failures and attacks in practical setup, as demonstrated. This paper provides the theoretical ground to study the sample efficiency of FRL with respect to the number of participating agents, while accounting for faulty agents. We verify the empirical efficacy of the proposed FRL framework in systems with and without different types of faulty agents on various RL benchmarks.

Variance control is the key to exploiting Assumption 2 on the bounded variance of PG estimators in our filter design. As a result, our framework is restricted to the variance-reduced policy gradient methods. Intuitively, it is worth studying the fault-tolerant federation of other policy optimization methods. Another limitation of this work is that agents are assumed to be homogeneous, while in many real-world scenarios, RL agents are heterogeneous. Therefore, it would be interesting to explore the possibility of heterogeneity of agents in fault-tolerant FRL in future works. Moreover, another interesting future work is to apply our Byzantine filtering strategy to other federated sequential decision-making problems such as federated bandit [57–59] and federated/collaborative Bayesian optimization [60–62], as well as other settings of collaborative multi-party ML [63–75], to equip them with theoretically guaranteed fault tolerance.

Acknowledgments and Disclosure of Funding

This research/project is supported by A*STAR under its RIE2020 Advanced Manufacturing and Engineering (AME) Industry Alignment Fund – Pre Positioning (IAF-PP) (Award A19E4a0101) and its A*STAR Computing and Information Science Scholarship (ACIS) awarded to Flint Xiaofeng Fan. Wei Jing is supported by Alibaba Innovative Research (AIR) Program.

References

- [1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. arXiv:1312.5602, 2013.
- [2] Levine Sergey, Nolan Wagener, and Pieter Abbeel. Learning contact-rich manipulation skills with guided policy search. In *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA*, pages 26–30, 2015.
- [3] Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, and Mengling Feng. Reinforcement learning for clinical decision support in critical care: comprehensive review. *Journal of Medical Internet Research*, 22(7):e18477, 2020.
- [4] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. arXiv:1904.12901, 2019.
- [5] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv:2005.01643, 2020.

- [6] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.
- [7] Rongmei Lin, Matthew D Stanley, Mohammad M Ghassemi, and Shamim Nemati. A deep deterministic policy gradient approach to medication dosing and surveillance in the icu. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4927–4931. IEEE, 2018.
- [8] Jakub Konečný, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. arXiv:1610.02527, 2016.
- [9] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. arXiv:1912.04977, 2019.
- [10] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: vision, hype and reality for data privacy and protection. arXiv:1907.09693, 2019.
- [11] Qinbin Li, Zeyi Wen, and Bingsheng He. Practical federated gradient boosting decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4642–4649, 2020.
- [12] Hankz Hankui Zhuo, Wenfeng Feng, Qian Xu, Qiang Yang, and Yufeng Lin. Federated reinforcement learning. arXiv:1901.08277, 2019.
- [13] Xinle Liang, Yang Liu, Tianjian Chen, Ming Liu, and Qiang Yang. Federated transfer reinforcement learning for autonomous driving. arXiv:1910.06001, 2019.
- [14] Chetan Nadiger, Anil Kumar, and Sherine Abdelhak. Federated reinforcement learning for fast personalization. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 123–127. IEEE, 2019.
- [15] Hyun-Kyo Lim, Ju-Bong Kim, Joo-Seong Heo, and Youn-Hee Han. Federated reinforcement learning for training control policies on multiple iot devices. *Sensors*, 20(5):1359, 2020.
- [16] Boyi Liu, Lujia Wang, and Ming Liu. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters*, 4(4):4555–4562, 2019.
- [17] Shuai Yu, Xu Chen, Zhi Zhou, Xiaowen Gong, and Di Wu. When deep reinforcement learning meets federated learning: Intelligent multi-timescale resource management for multi-access edge computing in 5G ultra dense network. *IEEE Internet of Things Journal*, 2020.
- [18] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirota, and Marcello Restelli. Stochastic variance-reduced policy gradient. arXiv:1806.05618, 2018.
- [19] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551. PMLR, 2020.
- [20] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2267–2281, 2019.
- [21] LESLIE Lamport, ROBERT SHOSTAK, and MARSHALL PEASE. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.
- [22] Nancy A. Lynch. *Distributed Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996. ISBN 9780080504704.
- [23] Miguel Castro and Barbara Liskov. Practical byzantine fault tolerance. In *OSDI*, volume 99, pages 173–186, 1999.

- [24] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.
- [25] Augustin Cauchy. Méthode générale pour la résolution des systemes d’équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.
- [26] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- [27] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [28] Léon Bottou and Yann Cun. Large scale online learning. *Advances in neural information processing systems*, 16:217–224, 2003.
- [29] Arkadi Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.
- [30] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.
- [31] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.
- [32] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707, 2016.
- [33] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016.
- [34] Lihua Lei and Michael I Jordan. Less than a single pass: Stochastically controlled stochastic gradient method. arXiv:1609.03261, 2016.
- [35] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2348–2358, 2017.
- [36] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [37] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.
- [38] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.
- [39] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [40] Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.
- [41] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. arXiv:1702.07944, 2017.
- [42] Zilun Peng, Ahmed Touati, Pascal Vincent, and Doina Precup. Svrg for policy evaluation with fewer gradient evaluations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2697–2703. International Joint Conferences on Artificial Intelligence Organization, 2020.
- [43] Tianbing Xu, Qiang Liu, and Jian Peng. Stochastic variance reduction for policy gradient estimation. arXiv:1710.06034, 2017.

- [44] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 118–128, 2017.
- [45] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. arXiv:1803.01498, 2018.
- [46] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4613–4623, 2018.
- [47] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/ec1c59141046cd1866bbcbdfb6ae31d4-Paper.pdf>.
- [48] Prashant Khanduri, Saikiran Bulusu, Pranay Sharma, and Pramod K Varshney. Byzantine resilient non-convex svrg with distributed batch gradient computations. arXiv:1912.04531, 2019.
- [49] Zeyuan Allen-Zhu, Faeze Ebrahimian, Jerry Li, and Dan Alistarh. Byzantine-resilient non-convex stochastic gradient descent. arXiv:2012.14368, 2020.
- [50] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International Conference on Machine Learning*, pages 6893–6901. PMLR, 2019.
- [51] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, pages 5442–5454, 2018.
- [52] Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2-3):255–283, 2015.
- [53] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR, 2016.
- [54] Samuel Albanie. Euclidean distance matrix trick. Technical report, 2019. URL https://www.robots.ox.ac.uk/~albanie/notes/Euclidean_distance_trick.pdf.
- [55] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.
- [56] Yan Duan, Xi Chen, Rein Houthoofd, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.
- [57] Abhimanyu Dubey and Alex Pentland. Differentially-private federated linear bandits. In *Proc. NeurIPS*, 2020.
- [58] Chengshuai Shi, Cong Shen, and Jing Yang. Federated multi-armed bandits with personalization. In *Proc. AISTATS*, pages 2917–2925. PMLR, 2021.
- [59] Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. In *Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, pages 3–4, 2021.
- [60] Zhongxiang Dai, Bryan Kian Hsiang Low, and Patrick Jaillet. Federated Bayesian optimization via Thompson sampling. In *Proc. NeurIPS*, 2020.
- [61] Zhongxiang Dai, Bryan Kian Hsiang Low, and Patrick Jaillet. Differentially private federated Bayesian optimization with distributed exploration. In *Proc. NeurIPS*, 2021.

- [62] Rachael Hwee Ling Sim, Yehong Zhang, Bryan Kian Hsiang Low, and Patrick Jaillet. Collaborative Bayesian optimization with fair regret. In *Proc. ICML*, pages 9691–9701. PMLR, 2021.
- [63] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Proc. NeurIPS*, 2021.
- [64] Xinyi Xu, Zhaoxuan Wu, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Validation free and replication robust volume-based data valuation. In *Proc. NeurIPS*, 2021.
- [65] Trong Nghia Hoang, Shenda Hong, Cao Xiao, Bryan Kian Hsiang Low, and Jimeng Sun. Aid: Active distillation machine to leverage pre-trained black-box models in private data settings. In *Proc. TheWebConf*, pages 3569–3581, 2021.
- [66] Thanh Chi Lam, Nghia Hoang, Bryan Kian Hsiang Low, and Patrick Jaillet. Model fusion for personalized learning. In *Proc. ICML*, pages 5948–5958. PMLR, 2021.
- [67] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *Proc. ICML*, pages 8927–8936. PMLR, 2020.
- [68] Nghia Hoang, Thanh Lam, Bryan Kian Hsiang Low, and Patrick Jaillet. Learning task-agnostic embedding of multiple black-box experts for multi-task model fusion. In *Proc. ICML*, pages 4282–4292. PMLR, 2020.
- [69] Ruofei Ouyang and Bryan Kian Hsiang Low. Gaussian process decentralized data fusion meets transfer learning in large-scale distributed cooperative perception. *Autonomous Robots*, 44(3): 359–376, 2020.
- [70] Minh Hoang, Nghia Hoang, Bryan Kian Hsiang Low, and Carleton Kingsford. Collective model fusion for multiple black-box experts. In *Proc. ICML*, pages 2742–2750. PMLR, 2019.
- [71] Trong Nghia Hoang, Quang Minh Hoang, Bryan Kian Hsiang Low, and Jonathan How. Collective online learning of Gaussian processes in massive multi-agent systems. In *Proc. AAAI*, volume 33, pages 7850–7857, 2019.
- [72] Ruofei Ouyang and Bryan Kian Hsiang Low. Gaussian process decentralized data fusion meets transfer learning in large-scale distributed cooperative perception. In *Proc. AAAI*, 2018.
- [73] Jie Chen, Bryan Kian Hsiang Low, Yujian Yao, and Patrick Jaillet. Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems. *IEEE Transactions on Automation Science and Engineering*, 12(3):901–921, 2015.
- [74] Jie Chen, Bryan Kian Hsiang Low, and Colin Keng-Yan Tan. Gaussian process-based decentralized data fusion and active sensing for mobility-on-demand system. In *Proc. RSS*, 2013.
- [75] Jie Chen, Bryan Kian Hsiang Low, Colin Keng-Yan Tan, Ali Oran, Patrick Jaillet, John M Dolan, and Gaurav S Sukhatme. Decentralized data fusion and active sensing with mobile sensors for modeling and predicting spatiotemporal traffic phenomena. In *Proc. UAI*, 2012.
- [76] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.
- [77] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

A More on the background

A.1 SVRG and SCSG

Here we provide the pseudocode for SVRG (Algorithm 2) and SCSG (Algorithm 3) seen in Lei et al. [35]. The idea of SVRG (Algorithm 2) is to reuse past *full* gradient computations (line 3) to reduce the variance of the current *stochastic* gradient estimate (line 7) before the parameter update (line 8). Note that $N = 1$ corresponds to a GD step (i.e., $v_{k-1}^{(j)} \leftarrow g_j$ in line 7). For $N > 1$, $v_{k-1}^{(j)}$ is the corrected gradient in SVRG and is an unbiased estimate of the true gradient $\nabla J(\theta)$. SVRG achieves linear convergence $O(1/T)$ using the semi-stochastic gradient.

Algorithm 2 SVRG

- 1: **Input:** Number of stages T , initial iteratre $\tilde{\theta}_0$, number of gradient steps N , step size η
 - 2: **for** $t = 1$ **to** T **do**
 - 3: $g_t \leftarrow \nabla J(\tilde{\theta}_{t-1}) = \frac{1}{n} \sum_{i=1}^n \nabla J_i(\tilde{\theta}_{t-1})$
 - 4: $\theta_0^{(t)} \leftarrow \tilde{\theta}_{t-1}$
 - 5: **for** $k = 1$ **to** N **do**
 - 6: Randomly pick $i_k \in [n]$
 - 7: $v_{k-1}^{(t)} \leftarrow \nabla J_{i_k}(\theta_{k-1}^{(t)}) - \nabla J_{i_k}(\theta_0^{(t)}) + g_t$
 - 8: $\theta_k^{(t)} \leftarrow \theta_{k-1}^{(t)} - \eta_t v_{k-1}^{(t)}$
 - 9: $\tilde{\theta}_t \leftarrow \theta_{N_t}^{(t)}$
 - 10: **Output:** $\tilde{\theta}_T$ (Convex case) or $\tilde{\theta}_t$ uniformly picked from $\{\tilde{\theta}_t\}_{t=1}^T$ (Non-Convex case)
-

More recently, *Stochastically Controlled Stochastic Gradient* (SCSG) has been proposed [34], to further reduce the computational cost of SVRG. The key difference is that SCSG (Algorithm 3) considers a sequence of time-varying batch sizes (B_t and b_t) and employs geometric sampling to generate the number of parameter update steps N_t in each iteration (line 6), instead of fixing the batch sizes and the number of updates as done in SVRG. Particularly when finding an ϵ -approximate solution (Definition 1) for optimizing smooth non-convex objectives, Lei et al. [35] proves that SCSG is never worse than SVRG in convergence rate and significantly outperforms SVRG when the required ϵ is small.

Algorithm 3 SCSG for smooth non-convex objectives

- 1: **Input:** Number of stages T , initial iteratre $\tilde{\theta}_0$, batch size B_t , mini-batch size b_t , step size η_t
 - 2: **for** $t = 1$ **to** T **do**
 - 3: Uniformly sample a batch $\mathcal{I}_t \subset \{1, \dots, n\}$ with $|\mathcal{I}_t| = B_t$
 - 4: $g_t \leftarrow \nabla J_{\mathcal{I}_t}(\tilde{\theta}_{t-1})$
 - 5: $\theta_0^{(t)} \leftarrow \tilde{\theta}_{t-1}$
 - 6: Generate $N_t \sim \text{Geom}(B_t/(B_t + b_t))$
 - 7: **for** $k = 1$ **to** N_t **do**
 - 8: Randomly pick $\tilde{\mathcal{I}}_{k-1} \subset [n]$ with $|\tilde{\mathcal{I}}_{k-1}| = b_t$
 - 9: $v_{k-1}^{(t)} \leftarrow \nabla J_{\tilde{\mathcal{I}}_{k-1}}(\theta_{k-1}^{(t)}) - \nabla J_{\tilde{\mathcal{I}}_{k-1}}(\theta_0^{(t)}) + g_t$
 - 10: $\theta_k^{(t)} \leftarrow \theta_{k-1}^{(t)} + \eta_t v_{k-1}^{(t)}$
 - 11: $\tilde{\theta}_t \leftarrow \theta_{N_t}^{(t)}$
 - 12: **Output:** $\tilde{\theta}_T$ (P-L case) or sample $\tilde{\theta}_T^*$ from $\{\tilde{\theta}_t\}_{t=1}^T$ with $P(\tilde{\theta}_T^* = \tilde{\theta}_t) \propto \eta_t B_t / b_t$ (Smooth case)
-

As a member of the SVRG-like algorithms, SCSG enjoys the same convergence rate of SVRG while being computationally cheaper than SVRG for tasks with small ϵ requirements [34], which is highly desired in RL, hence the motivation of FedPG-BR to adapt SCSG.

A.2 Gradient estimator

Use $g(\tau|\theta)$ to denote the *unbiased* estimator to the true gradient $J(\theta)$. The common gradient estimators are the REINFORCE and the GPOMDP estimators, which are considered as baseline estimators in [18] and [19]. The REINFORCE [39]:

$$g(\tau|\theta) = \left[\sum_{h=0}^{H-1} \nabla_{\theta} \log \pi_{\theta}(a_h | s_h) \right] \left[\sum_{h=0}^{H-1} \gamma^h \mathcal{R}(s_h, a_h) - C_b \right]$$

And the GPOMDP [40]

$$g(\tau|\theta) = \sum_{h=0}^{H-1} \left[\sum_{t=0}^h \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right] (\gamma^h r(s_h, a_h) - C_{b_h})$$

where C_b and C_{b_h} are the corresponding baselines. Under Assumption 3, whether we use the REINFORCE or the GPOMDP estimator, Proposition 4 holds [18, 19].

Algorithm 4 GPOMDP (for federation of K agents)

Input: number of iterations T , batch size B , step size η , initial parameter $\tilde{\theta}_0 \in \mathbb{R}^d$
for $t = 1$ **to** T **do**
 $\theta^t \leftarrow \tilde{\theta}_{t-1}$; broadcast to agents
 for $k = 1$ **to** K **do**
 Sample B trajectories $\{\tau_{t,i}^{(k)}\}$ from $p(\cdot|\theta^t)$
 $\mu_t^{(k)} = \frac{1}{B} \sum_{i=1}^B g(\tau_{t,i}^{(k)}|\theta^t)$; push $\mu_t^{(k)}$ to server
 $\mu_t = \frac{1}{K} \sum_{k=1}^K \mu_t^{(k)}$
 $\tilde{\theta}_t \leftarrow \theta^t + \eta \mu_t$
Output θ_{out} : uniformly randomly picked from $\{\tilde{\theta}_t\}_{t=1}^T$

A.3 Federated GPOMDP and SVRPG

Closely following the problem setting of FedPG-BR, we adapt both GPOMDP and SVRPG to the FRL setting. The pseudocode is shown in Algorithm 4 and Algorithm 5.

Algorithm 5 SVRPG (for federation of K agents)

Input: number of epochs T , epoch size N , batch size B , mini-batch size b , step size η , initial parameter $\tilde{\theta}_0 \in \mathbb{R}^d$
for $t = 1$ **to** T **do**
 $\theta_0^t \leftarrow \tilde{\theta}_{t-1}$; broadcast to agents
 for $k = 1$ **to** K **do**
 Sample B trajectories $\{\tau_{t,i}^{(k)}\}$ from $p(\cdot|\theta_0^t)$
 $\mu_t^{(k)} = \frac{1}{B} \sum_{i=1}^B g(\tau_{t,i}^{(k)}|\theta_0^t)$; push $\mu_t^{(k)}$ to server
 $\mu_t = \frac{1}{K} \sum_{k=1}^K \mu_t^{(k)}$
 for $n = 0$ **to** $N - 1$ **do**
 Sample b trajectories $\{\tau_{n,j}^t\}$ from $p(\cdot|\theta_n^t)$
 $v_n^t = \frac{1}{b} \sum_{j=1}^b [g(\tau_{n,j}^t|\theta_n^t) - \omega(\tau_{n,j}^t|\theta_n^t, \theta_0^t)g(\tau_{n,j}^t|\theta_0^t)] + \mu_t$
 $\theta_{n+1}^t = \theta_n^t + \eta v_n^t$
 $\tilde{\theta}_t \leftarrow \theta_N^t$
Output θ_{out} : uniformly randomly picked from $\{\tilde{\theta}_t\}_{t=1}^T$

B Proof of Theorem 6

In our proof, we follow the suggestion from Lei et al. [35] to set $b_t = 1$ to derive better theoretical results. Refer to Section F in this appendix for the value of b_t used in our experiments.

Proof. From the L-smoothness of the objective function $J(\boldsymbol{\theta})$, we have

$$\begin{aligned}
\mathbb{E}_{\tau_n^t}[J(\boldsymbol{\theta}_{n+1}^t)] &\geq \mathbb{E}_{\tau_n^t} \left[J(\boldsymbol{\theta}_n^t) + \langle \nabla J(\boldsymbol{\theta}_n^t), \boldsymbol{\theta}_{n+1}^t - \boldsymbol{\theta}_n^t \rangle - \frac{L}{2} \|\boldsymbol{\theta}_{n+1}^t - \boldsymbol{\theta}_n^t\|^2 \right] \\
&= J(\boldsymbol{\theta}_n^t) + \eta_t \langle \mathbb{E}_{\tau_n^t}[v_n^t], \nabla J(\boldsymbol{\theta}_n^t) \rangle - \frac{L\eta_t^2}{2} \mathbb{E}_{\tau_n^t}[\|v_n^t\|^2] \\
&\geq J(\boldsymbol{\theta}_n^t) + \eta_t \langle \nabla J(\boldsymbol{\theta}_n^t) + e_t, \nabla J(\boldsymbol{\theta}_n^t) \rangle \\
&\quad - \frac{L\eta_t^2}{2} [(2L_g + 2C_g^2 C_w) \|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2 + 2\|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + 2\|e_t\|^2] \quad (4) \\
&= J(\boldsymbol{\theta}_n^t) + \eta_t(1 - L\eta_t) \|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + \eta_t \langle e_t, \nabla J(\boldsymbol{\theta}_n^t) \rangle \\
&\quad - L\eta_t^2 (L_g + C_g^2 C_w) \|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2 - L\eta_t^2 \|e_t\|^2
\end{aligned}$$

where (4) follows from Lemma 11. Use \mathbb{E}_t to denote the expectation with respect to all trajectories $\{\tau_1^t, \tau_2^t, \dots\}$, given N_t . Since $\{\tau_1^t, \tau_2^t, \dots\}$ are independent of N_t , \mathbb{E}_t is equivalently the expectation with respect to $\{\tau_1^t, \tau_2^t, \dots\}$. The above inequality gives

$$\begin{aligned}
\mathbb{E}_t[J(\boldsymbol{\theta}_{n+1}^t)] &\geq \mathbb{E}_t[J(\boldsymbol{\theta}_n^t)] + \eta_t(1 - L\eta_t) \mathbb{E}_t \|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + \eta_t \mathbb{E}_t \langle e_t, \nabla J(\boldsymbol{\theta}_n^t) \rangle \\
&\quad - L\eta_t^2 (L_g + C_g^2 C_w) \mathbb{E}_t \|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2 - L\eta_t^2 \|e_t\|^2
\end{aligned}$$

Taking $n = N_t$ and using \mathbb{E}_{N_t} to denote the expectation w.r.t. N_t , we have from the above

$$\begin{aligned}
\mathbb{E}_{N_t} \mathbb{E}_t[J(\boldsymbol{\theta}_{N_t+1}^t)] &\geq \mathbb{E}_{N_t} \mathbb{E}_t[J(\boldsymbol{\theta}_{N_t}^t)] + \eta_t(1 - L\eta_t) \mathbb{E}_{N_t} \mathbb{E}_t \|\nabla J(\boldsymbol{\theta}_{N_t}^t)\|^2 + \eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle e_t, \nabla J(\boldsymbol{\theta}_{N_t}^t) \rangle \\
&\quad - L\eta_t^2 (L_g + C_g^2 C_w) \mathbb{E}_{N_t} \mathbb{E}_t \|\boldsymbol{\theta}_{N_t}^t - \boldsymbol{\theta}_0^t\|^2 - L\eta_t^2 \|e_t\|^2
\end{aligned}$$

Rearrange,

$$\begin{aligned}
\eta_t(1 - L\eta_t) \mathbb{E}_{N_t} \mathbb{E}_t \|\nabla J(\boldsymbol{\theta}_{N_t}^t)\|^2 &\leq \mathbb{E}_{N_t} \mathbb{E}_t[J(\boldsymbol{\theta}_{N_t+1}^t)] + L\eta_t^2 (L_g + C_g^2 C_w) \mathbb{E}_{N_t} \mathbb{E}_t \|\boldsymbol{\theta}_{N_t}^t - \boldsymbol{\theta}_0^t\|^2 \\
&\quad - \eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle e_t, \nabla J(\boldsymbol{\theta}_{N_t}^t) \rangle + L\eta_t^2 \|e_t\|^2 - \mathbb{E}_{N_t} \mathbb{E}_t[J(\boldsymbol{\theta}_{N_t}^t)] \\
&= \frac{1}{B_t} (\mathbb{E}_t \mathbb{E}_{N_t}[J(\boldsymbol{\theta}_{N_t}^t)] - J(\boldsymbol{\theta}_0^t)) - \eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle e_t, \nabla J(\boldsymbol{\theta}_{N_t}^t) \rangle \\
&\quad + L\eta_t^2 (L_g + C_g^2 C_w) \mathbb{E}_{N_t} \mathbb{E}_t \|\boldsymbol{\theta}_{N_t}^t - \boldsymbol{\theta}_0^t\|^2 + L\eta_t^2 \|e_t\|^2 \quad (5)
\end{aligned}$$

where (5) follows from Lemma 16 with Fubini's theorem. Note that $\tilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_{N_t}^t$ and $\tilde{\boldsymbol{\theta}}_{t-1} = \boldsymbol{\theta}_0^t$. If we take expectation over all the randomness and denote it by \mathbb{E} , we get

$$\begin{aligned}
\eta_t(1 - L\eta_t) \mathbb{E} \|\nabla J(\tilde{\boldsymbol{\theta}}_t)\|^2 &= \frac{1}{B_t} \mathbb{E} [J(\tilde{\boldsymbol{\theta}}_t) - J(\tilde{\boldsymbol{\theta}}_{t-1})] - \eta_t \mathbb{E} \langle e_t, \nabla J(\tilde{\boldsymbol{\theta}}_t) \rangle \\
&\quad + L\eta_t^2 (L_g + C_g^2 C_w) \mathbb{E} \|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|^2 + L\eta_t^2 \mathbb{E} \|e_t\|^2 \\
&= \frac{1}{B_t} \mathbb{E} [J(\tilde{\boldsymbol{\theta}}_t) - J(\tilde{\boldsymbol{\theta}}_{t-1})] - \frac{1}{B_t} \mathbb{E} \langle e_t, \tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1} \rangle \\
&\quad + L\eta_t^2 (L_g + C_g^2 C_w) \mathbb{E} \|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|^2 + \eta_t(1 + L\eta_t) \mathbb{E} \|e_t\|^2 \quad (6) \\
&\leq \frac{1}{B_t} \mathbb{E} [J(\tilde{\boldsymbol{\theta}}_t) - J(\tilde{\boldsymbol{\theta}}_{t-1})] \\
&\quad + \frac{1}{2\eta_t B_t} \left[-\frac{1}{B_t} + \eta_t^2 (2L_g + 2C_g^2 C_w) \right] \mathbb{E} \|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|^2 \\
&\quad + \frac{1}{B_t} \mathbb{E} \langle \nabla J(\tilde{\boldsymbol{\theta}}_t), \tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1} \rangle + \frac{\eta_t}{B_t} \mathbb{E} \|\nabla J(\tilde{\boldsymbol{\theta}}_t)\|^2 + \frac{\eta_t}{B_t} \mathbb{E} \|e_t\|^2 \\
&\quad + L\eta_t^2 (L_g + C_g^2 C_w) \mathbb{E} \|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|^2 + \eta_t(1 + L\eta_t) \mathbb{E} \|e_t\|^2 \quad (7)
\end{aligned}$$

where (6) follows from Lemma 12 and (7) follows from Lemma 13. Rearrange,

$$\begin{aligned}
\eta_t(1 - L\eta_t - \frac{1}{B_t}) \mathbb{E} \|\nabla J(\tilde{\boldsymbol{\theta}}_t)\|^2 &+ \frac{1 - 2\eta_t^2 (L_g + C_g^2 C_w) B_t - 2L\eta_t^3 (L_g + C_g^2 C_w) B_t^2}{2\eta_t B_t^2} \mathbb{E} \|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|^2 \\
&\leq \frac{1}{B_t} \mathbb{E} [J(\tilde{\boldsymbol{\theta}}_t) - J(\tilde{\boldsymbol{\theta}}_{t-1})] + \frac{1}{B_t} \mathbb{E} \langle \nabla J(\tilde{\boldsymbol{\theta}}_t), \tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1} \rangle + \eta_t(1 + L\eta_t + \frac{1}{B_t}) \mathbb{E} \|e_t\|^2 \quad (8)
\end{aligned}$$

Now we can apply Lemma 17 on $\mathbb{E} \langle \nabla J(\tilde{\theta}_t), \tilde{\theta}_t - \tilde{\theta}_{t-1} \rangle$ using $a = \tilde{\theta}_t - \tilde{\theta}_{t-1}$, $b = \nabla J(\tilde{\theta}_t)$, and $\beta = \frac{1-2\eta_t^2(L_g+C_g^2C_w)B_t-2L\eta_t^3(L_g+C_g^2C_w)B_t^2}{\eta_t B_t}$ to get

$$\begin{aligned} \frac{1}{B_t} \mathbb{E} \langle \nabla J(\tilde{\theta}_t), \tilde{\theta}_t - \tilde{\theta}_{t-1} \rangle &\leq \frac{1-2\eta_t^2(L_g+C_g^2C_w)B_t-2L\eta_t^3(L_g+C_g^2C_w)B_t^2}{2\eta_t B_t^2} \mathbb{E} \|\tilde{\theta}_t - \tilde{\theta}_{t-1}\|^2 \\ &\quad + \frac{\eta_t}{2[1-2\eta_t^2(L_g+C_g^2C_w)B_t-2L\eta_t^3(L_g+C_g^2C_w)B_t^2]} \mathbb{E} \|\nabla J(\tilde{\theta}_t)\|^2 \end{aligned} \quad (9)$$

Combining (8) and (9) and rearrange, we have

$$\begin{aligned} &\eta_t(1-L\eta_t - \frac{1}{B_t} - \frac{1}{2[1-2\eta_t^2(L_g+C_g^2C_w)B_t-2L\eta_t^3(L_g+C_g^2C_w)B_t^2]}) \mathbb{E} \|\nabla J(\tilde{\theta}_t)\|^2 \\ &\leq \frac{1}{B_t} \mathbb{E} [J(\tilde{\theta}_t) - J(\tilde{\theta}_{t-1})] + \eta_t(1+L\eta_t + \frac{1}{B_t}) \mathbb{E} \|e_t\|^2 \\ &\leq \frac{1}{B_t} \mathbb{E} [J(\tilde{\theta}_t) - J(\tilde{\theta}_{t-1})] + \eta_t(1+L\eta_t + \frac{1}{B_t}) \left[\frac{4\sigma^2}{(1-\alpha)^2 K B_t} + \frac{48\alpha^2\sigma^2 V}{(1-\alpha)^2 B_t} \right] \end{aligned} \quad (10)$$

where (10) follows from Lemma 15. We want to choose η_t such that $1-2\eta_t^2(L_g+C_g^2C_w)B_t-2L\eta_t^3(L_g+C_g^2C_w)B_t^2 > 0$. Denoting $\Phi = L_g+C_g^2C_w$, we have

$$\begin{aligned} 1-2\eta_t^2\Phi B_t-2L\eta_t^3\Phi B_t^2 &\geq 0 \\ 2\eta_t^2\Phi B_t+2L\eta_t^3\Phi B_t^2 &\leq 1 \end{aligned}$$

We can then choose η_t to have $2\eta_t^2\Phi B_t \leq \frac{1}{2}$ and $2L\eta_t^3\Phi B_t^2 \leq \frac{1}{2}$, which implies:

$$(i) \quad \eta_t \leq \frac{1}{(4\Phi B_t)^{1/2}}$$

$$(ii) \quad \eta_t \leq \frac{1}{(4L\Phi B_t^2)^{1/3}}$$

Therefore, we can choose $\eta_t \leq \frac{1}{2(L\Phi B_t^2)^{1/3}} < \frac{1}{(4L\Phi B_t^2)^{1/3}}$, and set $B_t \geq 4\Phi L^{-2}$ to ensure both condition (i) and (ii) are satisfied, together with $\frac{1}{(4L\Phi B_t^2)^{1/3}} \leq \frac{1}{(4\Phi B_t)^{1/2}}$, and $L\eta_t + \frac{1}{B_t} \leq 1$. Thus, by choosing $\eta_t \leq \frac{1}{2\Psi B_t^{2/3}}$, where $\Psi = (L\Phi)^{1/3} = (L(L_g+C_g^2C_w))^{1/3}$, we can obtain the following from (10):

$$\eta_t \mathbb{E} \|\nabla J(\tilde{\theta}_t)\|^2 \leq \frac{1}{B_t} \mathbb{E} [J(\tilde{\theta}_t) - J(\tilde{\theta}_{t-1})] + 2\eta_t \left[\frac{4\sigma^2}{(1-\alpha)^2 K B_t} + \frac{48\alpha^2\sigma^2 V}{(1-\alpha)^2 B_t} \right]$$

Replacing $\eta_t = \frac{1}{2\Psi B_t^{2/3}}$ and rearranging, we have

$$\begin{aligned} \mathbb{E} \|\nabla J(\tilde{\theta}_t)\|^2 &\leq \frac{1}{B_t \eta_t} \mathbb{E} [J(\tilde{\theta}_t) - J(\tilde{\theta}_{t-1})] + 2 \left[\frac{4\sigma^2}{(1-\alpha)^2 K B_t} + \frac{48\alpha^2\sigma^2 V}{(1-\alpha)^2 B_t} \right] \\ &\leq \frac{2\Psi \mathbb{E} [J(\tilde{\theta}_t) - J(\tilde{\theta}_{t-1})]}{B_t^{1/3}} + \frac{8\sigma^2}{(1-\alpha)^2 K B_t} + \frac{96\alpha^2\sigma^2 V}{(1-\alpha)^2 B_t} \end{aligned}$$

Replacing B_t with constant batch size B and telescoping over $t = 1, 2, \dots, T$, we have for $\tilde{\theta}_a$ from our algorithm:

$$\begin{aligned} \mathbb{E} \|\nabla J(\tilde{\theta}_a)\|^2 &\leq \frac{2\Psi \mathbb{E} [J(\tilde{\theta}_T) - J(\tilde{\theta}_0)]}{TB^{1/3}} + \frac{8\sigma^2}{(1-\alpha)^2 K B} + \frac{96\alpha^2\sigma^2 V}{(1-\alpha)^2 B} \\ &\leq \frac{2\Psi [J(\tilde{\theta}^*) - J(\tilde{\theta}_0)]}{TB^{1/3}} + \frac{8\sigma^2}{(1-\alpha)^2 K B} + \frac{96\alpha^2\sigma^2 V}{(1-\alpha)^2 B} \end{aligned}$$

which completes the proof. \square

C Proof of Corollary 7

Proof. Recall $\Psi = (L(L_g + C_g^2 C_w))^{1/3}$. From Theorem 6, we have

$$\mathbb{E}\|\nabla J(\tilde{\theta}_a)\|^2 \leq \underbrace{\frac{2\Psi\mathbb{E}[J(\tilde{\theta}^*) - J(\tilde{\theta}_0)]}{TB^{1/3}}}_{T=O(\frac{1}{\epsilon B^{1/3}})} + \underbrace{\frac{8\sigma^2}{(1-\alpha)^2 KB}}_{B_K=O(\frac{1}{\epsilon K})} + \underbrace{\frac{96\alpha^2\sigma^2 V}{(1-\alpha)^2 B}}_{B_\alpha=O(\frac{\alpha^2}{\epsilon})}$$

To guarantee that the output of Algorithm 1 is ϵ -approximate, i.e., $\mathbb{E}\|\nabla J(\tilde{\theta}_a)\|^2 \leq \epsilon$, we need the number of rounds T and the batch size B to meet the following:

$$(i)T = O(\frac{1}{\epsilon B^{1/3}}), (ii)B_K = O(\frac{1}{\epsilon K}), \text{ and } (iii)B_\alpha = O(\frac{\alpha^2}{\epsilon})$$

By union bound and using $\mathbb{E}[Traj(\epsilon)]$ to denote the total number of trajectories required by each agent to sample, the above implies that

$$\begin{aligned} \mathbb{E}[Traj(\epsilon)] &\leq TB_K + TB_\alpha \\ &\leq O(\frac{1}{\epsilon^{5/3} K^{2/3}} + \frac{\alpha^{4/3}}{\epsilon^{5/3}}) \end{aligned}$$

in order to obtain an ϵ -approximate policy, which completes the proof for Corollary 7 (i). Note that the total number of trajectories generated across the whole FRL system, denoted by $\mathbb{E}[Traj_{total}(\epsilon)]$ is thus bounded by:

$$\mathbb{E}[Traj_{total}(\epsilon)] \leq O(\frac{K^{1/3}}{\epsilon^{5/3}} + \frac{K\alpha^{4/3}}{\epsilon^{5/3}})$$

Now for an ideal system where $\alpha = 0$:

$$\begin{aligned} \mathbb{E}[Traj(\epsilon)] &\leq O(\frac{1}{\epsilon^{5/3} K^{2/3}}) \\ \mathbb{E}[Traj_{total}(\epsilon)] &\leq O(\frac{K^{1/3}}{\epsilon^{5/3}}) \end{aligned}$$

which completes the proof for Corollary 7 (ii). Moreover, when $K = 1$, the number of trajectories required by the agent using FedPG-BR is

$$\mathbb{E}[Traj(\epsilon)] \leq O(\frac{1}{\epsilon^{5/3}})$$

which is Corollary 7 (iii) and is coherent with the recent analysis of SVRPG [19]. \square

D More on the Byzantine Filtering Step

In this section, we continue our discussion on our *Byzantine Filtering Step* in Section 3.3. We include the pseudocode for the subroutine **FedPG-Aggregate** below for ease of reference:

As discussed in Section 3.3, R2 (line 8 in Algorithm 1.1) ensures that \mathcal{G}_t always include all good agents and for any Byzantine agents being included, their impact on the convergence of Algorithm 1 is limited since their maximum distance to $\nabla J(\theta_0^t)$ is bounded by 3σ . Here we give proofs for the claims.

Claim D.1. *Under Assumption 2 and $\forall \alpha < 0.5$, the filtering rule R2 in Algorithm 1.1 ensures that, in any round t , all gradient estimates sent from non-Byzantine agents are included in \mathcal{G}_t , i.e., $|\mathcal{G}_t| \geq (1 - \alpha)K$.*

Proof. First, from Assumption 2:

$$\|\mu_t^{(k)} - \nabla J(\theta_0^t)\| \leq \sigma, \forall k \in \mathcal{G}$$

Algorithm 1.1 FedPG-Aggregate

- 1: **Input:** Gradient estimates from K agents in round t : $\{\mu_t^{(k)}\}_{k=1}^k$, Variance Bound σ , filtering threshold $\mathfrak{T}_\mu \triangleq 2\sigma\sqrt{\frac{V}{B_t}}$, where $V \triangleq 2\log(\frac{2K}{\delta})$ and $\delta \in (0, 1)$
 - 2: $S_1 \triangleq \{\mu_t^{(k)}\}$ where $k \in [K]$ s.t. $\left| \left\{ k' \in [K] : \left\| \mu_t^{(k')} - \mu_t^{(k)} \right\| \leq \mathfrak{T}_\mu \right\} \right| > \frac{K}{2}$
 - 3: $\mu_t^{\text{mom}} \leftarrow \operatorname{argmin}_{\mu_t^{(\tilde{k})}} \left\| \mu_t^{(\tilde{k})} - \operatorname{mean}(S_1) \right\|$ where $\tilde{k} \in S_1$
 - 4: $R1: \mathcal{G}_t \triangleq \left\{ k \in [K] : \left\| \mu_t^{(k)} - \mu_t^{\text{mom}} \right\| \leq \mathfrak{T}_\mu \right\}$
 - 5: **if** $|\mathcal{G}_t| < (1 - \alpha)K$ **then**
 - 6: $S_2 \triangleq \{\mu_t^{(k)}\}$ where $k \in [K]$ s.t. $\left| \left\{ k' \in [K] : \left\| \mu_t^{(k')} - \mu_t^{(k)} \right\| \leq 2\sigma \right\} \right| > \frac{K}{2}$
 - 7: $\mu_t^{\text{mom}} \leftarrow \operatorname{argmin}_{\mu_t^{(\tilde{k})}} \left\| \mu_t^{(\tilde{k})} - \operatorname{mean}(S_2) \right\|$ where $\tilde{k} \in S_2$
 - 8: $R2: \mathcal{G}_t \triangleq \left\{ k \in [K] : \left\| \mu_t^{(k)} - \mu_t^{\text{mom}} \right\| \leq 2\sigma \right\}$
 - 9: **Return:** $\mu_t \triangleq \frac{1}{|\mathcal{G}_t|} \sum_{k \in \mathcal{G}_t} \mu_t^{(k)}$
-

it implies that $\left\| \mu_t^{(k_1)} - \mu_t^{(k_2)} \right\| \leq 2\sigma, \forall k_1, k_2 \in \mathcal{G}$. So, for any value of the *vector median* [46] in $S_2 = \{\mu_t^{(k)}\}$ (defined in line 6):

$$\left\| \mu_t^{(k)} - \nabla J(\theta_0^t) \right\| \leq 3\sigma, \forall \mu_t^{(k)} \in S_2$$

An intuitive illustration is provided in Fig. 4. Next, consider the worst case where all values sent by the K agents are included in S_2 : for all $(1 - \alpha)K$ good agents, they send the same value $\mu_t^{(k)}$, s.t., $\left\| \nabla J(\theta_0^t) - \mu_t^{(k)} \right\| = \sigma, \forall k \in \mathcal{G}$; and for all αK Byzantine agents, they send the same value $\mu_t^{(k')}$ s.t., $\left\| \nabla J(\theta_0^t) - \mu_t^{(k')} \right\| = 3\sigma, \forall k' \in S_2 \setminus \mathcal{G}$. Then the mean of values in S_2 satisfies:

$$\begin{aligned} \left\| \mu_t^{\text{mean}} - \nabla J(\theta_0^t) \right\| &= \frac{(1 - \alpha)K \cdot \sigma + \alpha K \cdot 3\sigma}{K} \\ &= (1 - \alpha)\sigma + 3\alpha\sigma \\ &= \sigma + 2\alpha\sigma \\ &< 2\sigma \end{aligned}$$

where the last inequality holds for $\alpha < 0.5$ which is our assumption. Then the value μ_t^{mom} of Algorithm 1 will be set to any $\mu_t^{(k)}$ from S_2 , of which is the closet to μ_t^{mean} .

The selection of μ_t^{mom} implies $\left\| \mu_t^{\text{mom}} - \nabla J(\theta_0^t) \right\| \leq \sigma$. Therefore, by constructing a region of \mathcal{G}_t that is centred at μ_t^{mom} and 2σ in radius (line 8), \mathcal{G}_t can cover all estimates from non-Byzantine agents and hence ensure $|\mathcal{G}_t| \geq (1 - \alpha)K$. \square

Claim D.2. *Under Assumption 2 and $\alpha < 0.5$, the filtering rule R2 in Algorithm 1.1 ensures that, in any round t , $\left\| \mu_t^{(k)} - \nabla J(\theta_0^t) \right\| \leq 3\sigma, \forall k \in \mathcal{G}_t$.*

Proof. This lemma is a straightforward result following the proof of Claim D.1. \square

Remark. *Claim D.2 implies that, in any round t , if an estimate sent from Byzantine agent is included in \mathcal{G}_t , then its impact on the convergence of Algorithm 1 is limited since its distance to $\nabla J(\theta_0^t)$ is bounded by 3σ . Fig. 4 provides an intuitive illustration for this claim.*

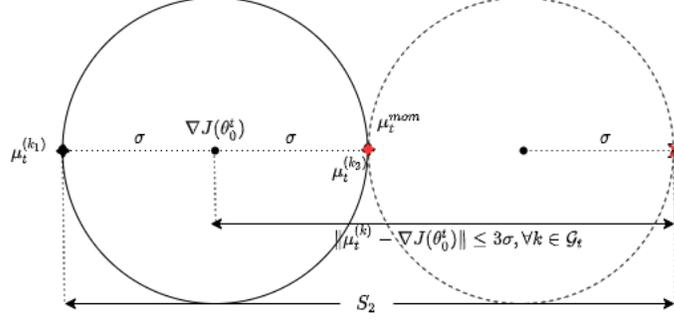


Figure 4: Graphical illustration of the Byzantine filtering strategy where $\mu_t^{(k_1)}, \mu_t^{(k_2)}$ are two good gradients while the red cross represents one Byzantine gradient which falls within S_2 . μ_t^{mom} will be chosen at the red diamond.

As discussed above, R2 ensures that all good agents are included in \mathcal{G}_t , i.e., a region in which all good agents are *concentrated*. R1 (lines 2-4) is designed in a similar way and aims to improve the practical performance of FedPG-BR by exploiting Lemma 14: all good agents are *highly likely* to be *concentrated* in a much *smaller region*.

Claim D.3. Define $V \triangleq 2 \log(2K/\delta)$ and $\delta \in (0, 1)$, the filtering R1 in Algorithm 1 ensure

$$\|\mu_t^{(k)} - \nabla J(\theta_0^t)\| \leq \sigma \sqrt{\frac{V}{B_t}}, \forall k \in \mathcal{G}$$

with probability of at least $1 - \delta$.

Proof. From Assumption 2, $\|\mu_t^{(k)} - \nabla J(\theta_0^t)\| \leq \sigma, \forall k \in \mathcal{G}$. We have

$$\begin{aligned} \|\mu_t^{(k)} - \nabla J(\theta_0^t)\| &= \left\| \frac{1}{B_t} \sum_{i=1}^{B_t} g(\tau_{t,i}^{(k)} | \theta_0^t) - \nabla J(\theta_0^t) \right\| \\ &= \frac{1}{B_t} \sqrt{\left\| \sum_{i=1}^{B_t} g(\tau_{t,i}^{(k)} | \theta_0^t) - \nabla J(\theta_0^t) \right\|^2} \end{aligned} \quad (11)$$

Consider $X_i \triangleq g(\tau_{t,i}^{(k)}) - \nabla J(\theta_0^t)$ and apply Lemma 14 on (11), we have

$$\begin{aligned} \Pr \left[\left\| \sum_{i=1}^{B_t} X_i \right\|^2 \leq 2 \log\left(\frac{2}{\delta}\right) \sigma^2 B_t \right] &\geq 1 - \delta \\ \Pr \left[\frac{1}{B_t} \sqrt{\left\| \sum_{i=1}^{B_t} X_i \right\|^2} \leq \frac{1}{B_t} \sqrt{2 \log\left(\frac{2}{\delta}\right) \sigma^2 B_t} \right] &\geq 1 - \delta \end{aligned}$$

With $V \triangleq 2 \log(2K/\delta)$ and $\delta \in (0, 1)$, the above inequality yields the Claim. \square

Therefore, the first filtering R1 (lines 2-4) of FedPG-BR constructs a region of \mathcal{G}_t centred at μ_t^{mom} with radius of $2\sigma \sqrt{\frac{V}{B_t}}$, which ensures in any round t that, with probability $\geq 1 - \delta$, (a) all good agents are included in \mathcal{G}_t , and (b) if gradients from Byzantine agents are included in \mathcal{G}_t , their impact is limited since their maximum distance to $\nabla J(\theta_0^t)$ is bounded by $3\sigma \sqrt{\frac{V}{B_t}}$ (The proof is similar to that of Claim D.2). Compared to R2, R1 can construct a *smaller region* that the server believes contains all good agents. If any Byzantine agent is included, their impact is also *smaller*, with probability of at least $1 - \delta$. Therefore, R1 is applied first such that if R1 fails (line 5) which happens with probability $< \delta$, R2 is then employed as a backup to ensure that \mathcal{G}_t always includes all good agents.

E Useful technical lemmas

Lemma 8 (Unbiaseness of importance sampling).

$$\begin{aligned}\mathbb{E}_{\tau \sim p(\cdot | \boldsymbol{\theta}_n)}[\omega(\tau | \boldsymbol{\theta}_n, \boldsymbol{\theta}_0)g(\tau | \boldsymbol{\theta}_0)] &= \mathbb{E}_{\tau \sim p(\cdot | \boldsymbol{\theta}_0)}[g(\tau | \boldsymbol{\theta}_0)] \\ &= \nabla J(\boldsymbol{\theta}_0)\end{aligned}$$

Proof. Drop t from notation and use τ_n to denote trajectories sampled from $\boldsymbol{\theta}_n$ at step n . From the definition of gradient estimation, we have

$$\begin{aligned}g(\tau_n | \boldsymbol{\theta}_0) &= \mathbb{E}_{\tau \sim p(\cdot | \boldsymbol{\theta}_n)}[\nabla_{\boldsymbol{\theta}_0} p(\boldsymbol{\theta}_0) r(\tau)] \\ &= \int p(\cdot | \boldsymbol{\theta}_n) \nabla_{\boldsymbol{\theta}_0} p(\boldsymbol{\theta}_0) r(\tau) d\tau \\ &= \int \frac{p(\cdot | \boldsymbol{\theta}_0)}{p(\cdot | \boldsymbol{\theta}_n)} p(\cdot | \boldsymbol{\theta}_n) \nabla_{\boldsymbol{\theta}_0} p(\boldsymbol{\theta}_0) r(\tau) d\tau \\ &= \int p(\cdot | \boldsymbol{\theta}_0) \frac{p(\cdot | \boldsymbol{\theta}_n)}{p(\cdot | \boldsymbol{\theta}_0)} \nabla_{\boldsymbol{\theta}_0} p(\boldsymbol{\theta}_0) r(\tau) d\tau \\ &= \mathbb{E}_{\tau \sim p(\cdot | \boldsymbol{\theta}_0)} \left[\frac{p(\cdot | \boldsymbol{\theta}_n)}{p(\cdot | \boldsymbol{\theta}_0)} \nabla_{\boldsymbol{\theta}_0} p(\boldsymbol{\theta}_0) r(\tau) \right] \\ &= \frac{p(\cdot | \boldsymbol{\theta}_n)}{p(\cdot | \boldsymbol{\theta}_0)} g(\tau_0 | \boldsymbol{\theta}_0)\end{aligned}$$

Then,

$$\begin{aligned}\omega(\tau | \boldsymbol{\theta}_n, \boldsymbol{\theta}_0)g(\tau_n | \boldsymbol{\theta}_0) &= \frac{p(\cdot | \boldsymbol{\theta}_0)}{p(\cdot | \boldsymbol{\theta}_n)} g(\tau_n | \boldsymbol{\theta}_0) \\ &= g(\tau_0 | \boldsymbol{\theta}_0)\end{aligned}$$

which gives the lemma. \square

Lemma 9 (Adapted from [19]). *Let $\omega(\tau | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = p(\tau | \boldsymbol{\theta}_1) / p(\tau | \boldsymbol{\theta}_2)$, under Assumptions 3 and 5, it holds that*

$$\text{Var}(\omega(\tau | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) \leq C_w \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2$$

where $C_w = H(2HG^2 + M)(W + 1)$. Furthermore, we have

$$\begin{aligned}\mathbb{E}_{\tau_n^t} \|1 - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)\|^2 &= \text{Var}_{\boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t}(\omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)) \\ &\leq C_w \|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2\end{aligned}$$

Proof. The proof can be found in Xu et al. [19]. \square

Lemma 10. *For $X_1, X_2 \in \mathbb{R}^d$, we have*

$$\|X_1 + X_2\|^2 \leq 2\|X_1\|^2 + 2\|X_2\|^2$$

Lemma 11.

$$\mathbb{E}_{\tau_n^t} [\|v_n^t\|^2] \leq (2L_g + 2C_g^2 C_w) \|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2 + 2 \|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + 2 \|e_t\|^2$$

Proof. We follow the suggestion of Lei et al. [35] to set $b_t = 1$ to deliver better theoretical results. However in our experiments, we do allow b_t to be sampled from different values. With $b_t = 1$ and $\mu_t = \frac{1}{|\mathcal{G}_t|} \sum_{k \in \mathcal{G}_t} \mu_t^{(k)}$, we have the flowing definition according to Algorithm 1:

$$v_n^t \triangleq g(\tau_n^t | \boldsymbol{\theta}_n^t) - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)g(\tau_n^t | \boldsymbol{\theta}_0^t) + u_t \quad (11-12)$$

which is the SCSG update step. Define $e_t \triangleq u_t - \nabla J(\boldsymbol{\theta}_0^t)$, we then have

$$\begin{aligned}\mathbb{E}_{\tau_n^t} [v_n^t] &= \nabla J(\boldsymbol{\theta}_n^t) - \nabla J(\boldsymbol{\theta}_0^t) + e_t + \nabla J(\boldsymbol{\theta}_0^t) \\ &= \nabla J(\boldsymbol{\theta}_n^t) + e_t\end{aligned} \quad (11-13)$$

Note that $\nabla J(\boldsymbol{\theta}_n^t) - \nabla J(\boldsymbol{\theta}_0^t) = \mathbb{E}_{\tau_n^t}[g(\tau_n^t | \boldsymbol{\theta}_n^t) - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)g(\tau_n^t | \boldsymbol{\theta}_0^t)]$ as we have showed that the importance weighting term results in unbiased estimation of the true gradient in Lemma 8. Then from $\mathbb{E}\|X\|^2 = \mathbb{E}\|X - \mathbb{E}X\|^2 + \|\mathbb{E}X\|^2$,

$$\begin{aligned} \mathbb{E}_{\tau_n^t}[\|v_n^t\|^2] &= \mathbb{E}_{\tau_n^t}\|v_n^t - \mathbb{E}_{\tau_n^t}[v_n^t]\|^2 + \|\mathbb{E}_{\tau_n^t}[v_n^t]\|^2 \\ &= \mathbb{E}_{\tau_n^t}\|g(\tau_n^t | \boldsymbol{\theta}_n^t) - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)g(\tau_n^t | \boldsymbol{\theta}_0^t) + u_t - (\nabla J(\boldsymbol{\theta}_n^t) + e_t)\|^2 + \|\mathbb{E}_{\tau_n^t}[v_n^t]\|^2 \\ &= \mathbb{E}_{\tau_n^t}\|g(\tau_n^t | \boldsymbol{\theta}_n^t) - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)g(\tau_n^t | \boldsymbol{\theta}_0^t) - (\nabla J(\boldsymbol{\theta}_n^t) - \nabla J(\boldsymbol{\theta}_0^t))\|^2 + \|\nabla J(\boldsymbol{\theta}_n^t) + e_t\|^2 \\ &\leq \mathbb{E}_{\tau_n^t}\|g(\tau_n^t | \boldsymbol{\theta}_n^t) - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)g(\tau_n^t | \boldsymbol{\theta}_0^t)\|^2 + 2\|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + 2\|e_t\|^2 \end{aligned} \quad (11-14)$$

where (11-14) follows from $\mathbb{E}\|X - \mathbb{E}X\|^2 \leq \mathbb{E}\|X\|^2$ and Lemma 10. Note that

$$\begin{aligned} &\mathbb{E}_{\tau_n^t}\|g(\tau_n^t | \boldsymbol{\theta}_n^t) - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)g(\tau_n^t | \boldsymbol{\theta}_0^t)\|^2 \\ &= \mathbb{E}_{\tau_n^t}\|g(\tau_n^t | \boldsymbol{\theta}_n^t) + g(\tau_n^t | \boldsymbol{\theta}_0^t) - g(\tau_n^t | \boldsymbol{\theta}_0^t) - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t)g(\tau_n^t | \boldsymbol{\theta}_0^t)\|^2 \\ &= \mathbb{E}_{\tau_n^t}\|g(\tau_n^t | \boldsymbol{\theta}_n^t) - g(\tau_n^t | \boldsymbol{\theta}_0^t) + (1 - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t))g(\tau_n^t | \boldsymbol{\theta}_0^t)\|^2 \\ &\leq 2\mathbb{E}_{\tau_n^t}\|g(\tau_n^t | \boldsymbol{\theta}_n^t) - g(\tau_n^t | \boldsymbol{\theta}_0^t)\|^2 + 2\mathbb{E}_{\tau_n^t}\|(1 - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t))g(\tau_n^t | \boldsymbol{\theta}_0^t)\|^2 \end{aligned} \quad (11-15)$$

where (11-15) follows from Lemma 10. Combining (11-14) and (11-15), we have

$$\begin{aligned} \mathbb{E}_{\tau_n^t}[\|v_n^t\|^2] &\leq 2\mathbb{E}_{\tau_n^t}\|g(\tau_n^t | \boldsymbol{\theta}_n^t) - g(\tau_n^t | \boldsymbol{\theta}_0^t)\|^2 \\ &\quad + 2\mathbb{E}_{\tau_n^t}\|(1 - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t))g(\tau_n^t | \boldsymbol{\theta}_0^t)\|^2 + 2\|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + 2\|e_t\|^2 \\ &\leq 2L_g\|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2 + 2C_g^2\mathbb{E}_{\tau_n^t}\|(1 - \omega(\tau_n^t | \boldsymbol{\theta}_n^t, \boldsymbol{\theta}_0^t))\|^2 + 2\|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + 2\|e_t\|^2 \end{aligned} \quad (11-16)$$

$$\leq 2L_g\|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2 + 2C_g^2C_w\|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2 + 2\|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + 2\|e_t\|^2 \quad (11-17)$$

$$= (2L_g + 2C_g^2C_w)\|\boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t\|^2 + 2\|\nabla J(\boldsymbol{\theta}_n^t)\|^2 + 2\|e_t\|^2 \quad (11-18)$$

where (11-16) is from Lemma 4 and (11-17) follows from Lemma 9 \square

Lemma 12.

$$\eta_t \mathbb{E} \langle e_t, \mathbb{E} \nabla J(\tilde{\boldsymbol{\theta}}_t) \rangle = \frac{1}{B_t} \mathbb{E} \langle e_t, \tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1} \rangle - \eta_t \mathbb{E} \|e_t\|^2$$

Proof. Consider $M_n^t = \langle e_t, \boldsymbol{\theta}_n^t - \boldsymbol{\theta}_0^t \rangle$. We have

$$M_{n+1}^t - M_n^t = \langle e_t, \boldsymbol{\theta}_{n+1}^t - \boldsymbol{\theta}_n^t \rangle = \eta_t \langle e_t, v_n^t \rangle$$

Taking expectation with respect to τ_n^t , we have

$$\begin{aligned} \mathbb{E}_{\tau_n^t} [M_{n+1}^t - M_n^t] &= \eta_t \langle e_t, \mathbb{E}_{\tau_n^t}[v_n^t] \rangle \\ &= \eta_t \langle e_t, \nabla J(\boldsymbol{\theta}_n^t) \rangle + \eta_t \|e_t\|^2 \end{aligned}$$

following from (11-13). Use \mathbb{E}_t to denote the expectation with respect to all trajectories $\{\tau_1^t, \tau_2^t, \dots\}$, given N_t . Since $\{\tau_1^t, \tau_2^t, \dots\}$ are independent of N_t , \mathbb{E}_t is equivalently the expectation with respect to $\{\tau_1^t, \tau_2^t, \dots\}$. We have

$$\mathbb{E}_t [M_{n+1}^t - M_n^t] = \eta_t \langle e_t, \mathbb{E}_t \nabla J(\boldsymbol{\theta}_n^t) \rangle + \eta_t \|e_t\|^2$$

Taking $n = N_t$ and denoting \mathbb{E}_{N_t} the expectation w.r.t. N_t , we have

$$\mathbb{E}_{N_t} \mathbb{E}_t (M_{N_t+1}^t - M_{N_t}^t) = \eta_t \langle e_t, \mathbb{E}_{N_t} \mathbb{E}_t \nabla J(\boldsymbol{\theta}_{N_t}^t) \rangle + \eta_t \|e_t\|^2.$$

Using Fubini's theorem, Lemma 16 and using the fact $\boldsymbol{\theta}_{N_t}^t = \tilde{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}_0^t = \tilde{\boldsymbol{\theta}}_{t-1}$,

$$\begin{aligned} \mathbb{E}_{N_t} \mathbb{E}_t (M_{N_t+1}^t - M_{N_t}^t) &= -\mathbb{E}_t \mathbb{E}_{N_t} (M_{N_t}^t - M_{N_t+1}^t) \\ &= -\left(\frac{1}{B_t/(B_t+1)} - 1\right)(M_0^t - \mathbb{E}_{N_t} \mathbb{E}_t M_{N_t}^t) \\ &= \frac{1}{B_t} \mathbb{E}_{N_t} \mathbb{E}_t \langle e_t, \tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1} \rangle \\ &= \eta_t \langle e_t, \mathbb{E}_{N_t} \mathbb{E}_t \nabla J(\boldsymbol{\theta}_{N_t}^t) \rangle + \eta_t \|e_t\|^2 \end{aligned}$$

Taking expectation with respect to the whole past yields the lemma. \square

Lemma 13.

$$-2\eta_t \mathbb{E} \langle e_t, \tilde{\theta}_t - \tilde{\theta}_{t-1} \rangle \leq \left[-\frac{1}{B_t} + \eta_t^2 (2L_g + 2C_g^2 C_w) \right] \mathbb{E} \|\tilde{\theta}_t - \tilde{\theta}_{t-1}\|^2 + 2\eta_t^2 \mathbb{E} \|e_t\|^2 \\ + 2\eta_t \mathbb{E} \langle \nabla J(\tilde{\theta}_t), \tilde{\theta}_t - \tilde{\theta}_{t-1} \rangle + 2\eta_t^2 \mathbb{E} \|\nabla J(\tilde{\theta}_t)\|^2$$

Proof. We have from the update equation $\theta_{n+1}^t = \theta_n^t + \eta_t v_n^t$, then,

$$\mathbb{E}_{\tau_n^t} \|\theta_{n+1}^t - \theta_0^t\|^2 = \mathbb{E}_{\tau_n^t} \|\theta_n^t + \eta_t v_n^t - \theta_0^t\|^2 \\ = \|\theta_n^t - \theta_0^t\|^2 + \eta_t^2 \mathbb{E}_{\tau_n^t} \|v_n^t\|^2 + 2\eta_t \langle \mathbb{E}_{\tau_n^t} [v_n^t], \theta_n^t - \theta_0^t \rangle \\ \leq \|\theta_n^t - \theta_0^t\|^2 + \eta_t^2 [(2L_g + 2C_g^2 C_w) \|\theta_n^t - \theta_0^t\|^2 + 2\|\nabla J(\theta_n^t)\|^2 + 2\|e_t\|^2] \\ + 2\eta_t \langle e_t, \theta_n^t - \theta_0^t \rangle + 2\eta_t \langle \nabla J(\theta_n^t), \theta_n^t - \theta_0^t \rangle \quad (13-19) \\ = [1 + \eta_t^2 (2L_g + 2C_g^2 C_w)] \|\theta_n^t - \theta_0^t\|^2 + 2\eta_t \langle \nabla J(\theta_n^t), \theta_n^t - \theta_0^t \rangle \\ + 2\eta_t \langle e_t, \theta_n^t - \theta_0^t \rangle + 2\eta_t^2 \|\nabla J(\theta_n^t)\|^2 + 2\eta_t^2 \|e_t\|^2$$

where (13-19) follows the result of (11-18). Use \mathbb{E}_t to denote the expectation with respect to all trajectories $\{\tau_1^t, \tau_2^t, \dots\}$, given N_t . Since $\{\tau_1^t, \tau_2^t, \dots\}$ are independent of N_t , \mathbb{E}_t is equivalently the expectation with respect to $\{\tau_1^t, \tau_2^t, \dots\}$. We have

$$\mathbb{E}_t \|\theta_{n+1}^t - \theta_0^t\|^2 \leq [1 + \eta_t^2 (2L_g + 2C_g^2 C_w)] \mathbb{E}_t \|\theta_n^t - \theta_0^t\|^2 + 2\eta_t \mathbb{E}_t \langle \nabla J(\theta_n^t), \theta_n^t - \theta_0^t \rangle \\ + 2\eta_t \mathbb{E}_t \langle e_t, \theta_n^t - \theta_0^t \rangle + 2\eta_t^2 \mathbb{E}_t \|\nabla J(\theta_n^t)\|^2 + 2\eta_t^2 \|e_t\|^2$$

Now taking $n = N_t$ and denoting E_{N_t} the expectation w.r.t. N_t we have

$$-2\eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle e_t, \theta_{N_t}^t - \theta_0^t \rangle \\ \leq [1 + \eta_t^2 (2L_g + 2C_g^2 C_w)] \mathbb{E}_{N_t} \mathbb{E}_t \|\theta_{N_t}^t - \theta_0^t\|^2 - \mathbb{E}_{N_t} \mathbb{E}_t \|\theta_{N_t+1}^t - \theta_0^t\|^2 \\ + 2\eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle \nabla J(\theta_{N_t}^t), \theta_{N_t}^t - \theta_0^t \rangle + 2\eta_t^2 \mathbb{E}_{N_t} \mathbb{E}_t \|\nabla J(\theta_{N_t}^t)\|^2 + 2\eta_t^2 \|e_t\|^2 \\ = \left[-\frac{1}{B_t} + \eta_t^2 (2L_g + 2C_g^2 C_w) \right] \mathbb{E}_{N_t} \mathbb{E}_t \|\theta_{N_t}^t - \theta_0^t\|^2 \\ + 2\eta_t \mathbb{E}_{N_t} \mathbb{E}_t \langle \nabla J(\theta_{N_t}^t), \theta_{N_t}^t - \theta_0^t \rangle + 2\eta_t^2 \mathbb{E}_{N_t} \mathbb{E}_t \|\nabla J(\theta_{N_t}^t)\|^2 + 2\eta_t^2 \|e_t\|^2 \quad (13-20)$$

where (13-20) follows Lemma 16 using Fubini's theorem. Rearranging, replacing $\theta_{N_t}^t = \tilde{\theta}_t$ and $\theta_0^t = \tilde{\theta}_{t-1}$ and taking expectation w.r.t the whole past yields the lemma. \square

Lemma 14 (Pinelis' inequality [76]; Lemma 2.4 [46]). *Let the sequence of random variables $X_1, X_2, \dots, X_N \in \mathbb{R}^d$ represent a random process such that we have $\mathbb{E}[X_n | X_1, \dots, X_{n-1}]$ and $\|X_n\| \leq M$. Then,*

$$\mathbb{P} \left[\|X_1 + \dots + X_N\|^2 \leq 2 \log(2/\delta) M^2 N \right] \geq 1 - \delta$$

Lemma 15 (Adapted from [48]). *If we choose δ and B_t in Algorithm 1 such that:*

$$(i) e^{\frac{\delta B_t}{2(1-2\delta)}} \leq \frac{2K}{\delta} \leq e^{\frac{B_t}{2}}$$

$$(ii) \delta \leq \frac{1}{5KB_t}$$

then we have the following bound for $\mathbb{E}\|e_t\|^2$:

$$\mathbb{E} \|e_t\|^2 \leq \frac{4\sigma^2}{(1-\alpha)^2 K B_t} + \frac{48\alpha^2 \sigma^2 V}{(1-\alpha)^2 B_t}$$

Proof. The proof of this lemma is similar to that of Lemma 7 of Khanduri et al. [48]. The key difference lays on the base conditions used to define the probabilistic events.

In FedPG-BR, the following refined conditions (results of Claims D.1 and D.2) are used,

$$\|\mu_t^{\text{mom}} - \nabla J(\theta)\| \leq \sigma, \|\mu_t^{(k)} - \mu_t^{\text{mom}}\| \leq 2\sigma, \|\mu_t^{(k)} - \nabla J(\theta)\| \leq 3\sigma, \forall k \in \mathcal{G}_t$$

whereas Khanduri et al. [48] needs the following:

$$\|\mu_t^{\text{med}} - \nabla J(\boldsymbol{\theta})\| \leq 3\sigma, \|\mu_t^{(k)} - \mu_t^{\text{med}}\| \leq 4\sigma, \|\mu_t^{(k)} - \nabla J(\boldsymbol{\theta})\| \leq 7\sigma, \forall k \in \mathcal{G}_t$$

The detailed proof of Lemma 15 can be obtained following the derivation of Lemma 7 of Khanduri et al. [48] by modifying the base conditions. \square

Lemma 16. *If $N \sim \text{Geom}(\Gamma)$ for $\Gamma > 0$. Then for any sequence D_0, D_1, \dots with $\mathbb{E}\|D_N\| \leq \infty$, we have*

$$\mathbb{E}[D_N - D_{N+1}] = \left(\frac{1}{\Gamma} - 1\right)(D_0 - \mathbb{E}D_N)$$

Proof. The proof can be found in Lei et al. [35]. \square

Lemma 17 (Young’s inequality (Peter-Paul inequality)). *For all real numbers a and b and all $\beta > 0$, we have*

$$ab \leq \frac{a^2}{2\beta} + \frac{\beta b^2}{2}$$

F Experimental details

F.1 Hyperparameters

We follow the setups of SVRPG [18] to parameterize the policies using neural networks. For all the algorithms under comparison in the experiments (Section 5), Adam[77] is used as the gradient optimizer. The 10 random seeds are [0–9]. All other hyperparameters used in all the experiments are reported in Table 2.

Table 2: Hyperparameters used in the experiments.

Hyperparameters	Algorithms	CartPole-v1	LunarLander-v2	HalfCheetach-v2
NN policy	-	Categorical MLP	Categorical MLP	Gaussian MLP
NN hidden weights	-	16,16	64,64	64,64
NN activation	-	ReLU	Tanh	Tanh
NN output activation	-	Tanh	Tanh	Tanh
Step size (Adam) η	-	1e-3	1e-3	8e-5
Discount factor γ	-	0.999	0.990	0.995
Maximum trajectories	-	5000	10000	10000
Task horizon H (for training)	-	500	1000	500
Task horizon H (for test)	-	500	1000	1000
α (for practical setup)	-	0.3	0.3	0.3
Number of runs	-	10	10	10
Batch size B_t	GPOMDP	16	32	48
	SVRPG	16	32	48
	FedPG-BR	sampled from [12, 20]	sampled from [26, 38]	sampled from [46, 50]
Mini-Batch size b_t	GPOMDP	-	-	-
	SVRPG	4	8	16
	FedPG-BR	4	8	16
Number of steps N_t	GPOMDP	1	1	1
	SVRPG	3	3	3
	FedPG-BR	$N_t \sim \text{Geom}(\frac{B_t}{B_t+b_t})$	$N_t \sim \text{Geom}(\frac{B_t}{B_t+b_t})$	$N_t \sim \text{Geom}(\frac{B_t}{B_t+b_t})$
Variance bound σ (Estimated by server)	GPOMDP	-	-	-
	SVRPG	-	-	-
	FedPG-BR	0.06	0.07	0.9
Confidence parameter δ	GPOMDP	-	-	-
	SVRPG	-	-	-
	FedPG-BR	0.6	0.6	0.6

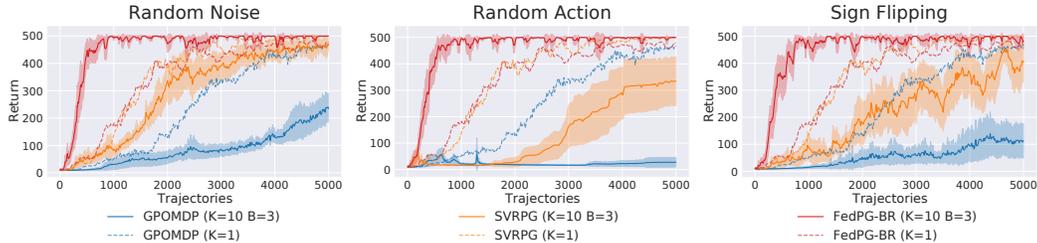


Figure 5: Performance of FedPG-BR in practical systems with $\alpha > 0$ for CartPole. Each subplot corresponds to a different type of Byzantine failure exercised by the 3 Byzantine agents.

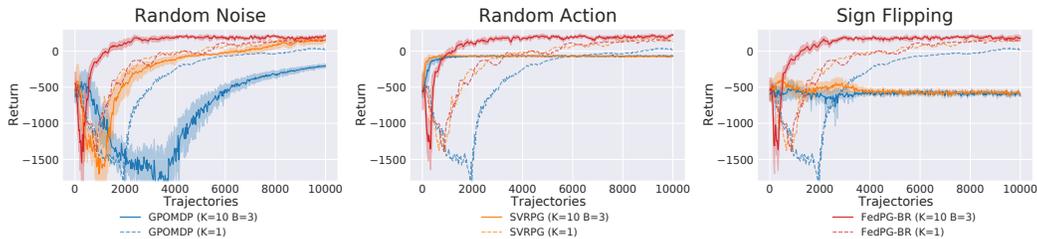


Figure 6: Performance of FedPG-BR in practical systems with $\alpha > 0$ for LunarLander. Each subplot corresponds to a different type of Byzantine failure exercised by the 3 Byzantine agents.

F.2 Computing Infrastructure

All experiments are conducted on a computing server without GPUs. The server is equipped with 14 cores (28 threads) *Intel(R) Core(TM) i9-10940X CPU @ 3.30GHz* and 64G memory. The average runtime for each run of FedPG-BR ($K=10$ $B=3$) is 2.5 hours for the CartPole task, 4 hours for the HalfCheetah task, and 12 hours for the LunarLander task.

G Additional experiments

G.1 Performance of FedPG-BR in practical systems with $\alpha > 0$ for the CartPole and the LunarLander tasks

The results for the CartPole and the LunarLander tasks which yield the same insights as discussed in experiments (Section 5) are plotted in Figure 5 and Figure 6. As discussed earlier, for both GPOMDP and SVRPG, the federation of more agents in practical systems which are subject to the presence of Byzantine agents, i.e., random failures or adversarial attacks, causes the performance of their federation to be worse than that in the single-agent setting. In particular, RA agents (middle figure) and SF agents (right figure) render GPOMDP and SVRPG unlearnable, i.e., unable to converge at all. This is in contrast to the performance of FedPG-BR. That is, FedPG-BR ($K = 10B = 3$) is able to deliver superior performances even in the presence of Byzantine agents for all three tasks: CartPole (Figure 5), LunarLander (Figure 6), and HalfCheetah (Figure 2 in Section 5). This provides an assurance on the reliability of our FedPG-BR algorithm to promote its practical deployment, and significantly improves the practicality of FRL.

G.2 Performance of FedPG-BR against the Variance Attack

We have discussed in Section 3.2 where the high variance in PG estimation renders the FRL system vulnerable to variance-based attacks such as the Variance Attack (VA) proposed by Baruch et al. [47]. The VA attackers collude together to estimate the population mean and the standard-deviation of gradients at each round, and move the mean by the largest value such that their values are still within the population variance. Intuitively, this non-omniscient attack works by exploiting the high variance in gradient estimation of the population and crafting values that contribute most to the population variance, hence gradually shifting the population mean. According to Cao et al. [20],

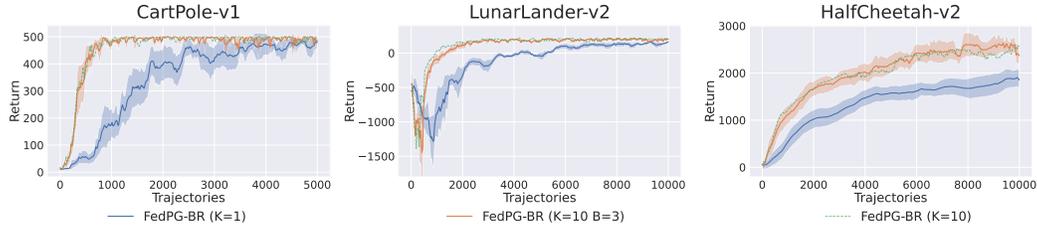


Figure 7: Performance of FedPG-BR in practical systems with $\alpha > 0$ for CartPole. Among the $K = 10$ participating agents, 3 Byzantine agents are colluding together to launch the VA attack.

existing defenses will fail to remove those non-omniscient attackers and the convergence will be significantly worsened if the population variance is large enough.

We are thus motivated to look for solutions that theoretically reduce the variance in policy gradient estimation. Inspired by the variance-reduced policy gradient works [e.g., 18, 19], we adapt the SCSG optimization [35] to our federated policy gradient framework for a refined control over the estimation variance. Through our adaptation, we are able to control the variance by the semi-stochastic gradient (line 11 in Algorithm 1), hence resulting in the fault-tolerant FRL system that can defend the VA attackers. Each plot in Figure 7 shows the experiment for each of the three tasks correspondingly, where 3 Byzantine agents are implemented as the VA attackers [20] (z^{max} is 0.18 in our setup). We again include the corresponding single-agent performance ($K = 1$) and the federation of 10 good agents ($K = 10$) in the plots for reference. The results show that in all three tasks, FedPG-BR ($K = 10B = 3$) still manages to significantly outperform FedPG-BR ($K = 1$) in the single-agent setting. Furthermore, the performance of FedPG-BR ($K = 10B = 3$) is barely worsened compared with FedPG-BR ($K = 10$) with 10 good agents. This shows that, with the adaptation of SCSG, our fault-tolerant FRL system can perfectly defend the VA attack from the literature, which further corroborates our analysis on our Byzantine filtering step (Section 3.3) showing that if gradients from Byzantine agents are not filtered out, their impact is limited since their maximum distance to $\nabla J(\theta_0^t)$ is bounded by 3σ (Claim D.2).

G.3 Environment Setup

On a Linux system, navigate into the root directory of this project and execute the following commands:

```
$ conda create -n FT-FRL pytorch=1.5.0
$ conda activate FT-FRL
$ pip install -r requirements.txt
$ cd codes
```

To run experiments in HalfCheetah, a mujoco license⁴ is required. After obtaining the license, install the mujoco-py library by following the instructions from OpenAI.⁵

G.4 Examples

To reproduce the results of FedPG-BR ($K = 10$) in Figure 1 for the HalfCheetah task, run the following command:

```
$ python run.py --env_name HalfCheetah-v2 --FT_FedPG
--num_worker 10 --num_Byzantine 0
--log_dir ./logs_HalfCheetah --multiple_run 10
--run_name HalfCheetah_FT-FRL_W10B0
```

To reproduce the results of FedPG-BR ($K = 10 B = 3$) in Figure 2 where 3 Byzantine agents are Random Noise in the HalfCheetah task environment, run the following command:

⁴<http://www.mujoco.org>

⁵<https://github.com/openai/mujoco-py>

```
$ python run.py --env_name CartPole-v1 --FT_FedPG
--num_worker 10 --num_Byzantine 3
--attack_type random-noise
--log_dir ./logs_Cartpole --multiple_run 10
--run_name Cartpole_FT-FRL_W10B3
```

Replace ‘--FT_FedPG’ with ‘--SVRPG’ for the results of SVRPG in the same experiment. All results including all statistics will be logged into the directory indicated by ‘--log_dir’, which can be visualized in tensorboard.

To visualize the behavior of the learnt policy, run the experiment in evaluation mode with rendering option on. For example:

```
$ python run.py --env_name CartPole-v1 --FT_FedPG
--eval_only --render
--load_path PATH_TO_THE_SAVED_POLICY_MODEL
```