Incentivizing Truthful Collaboration in Heterogeneous Federated Learning

Dimitar Chakarov* CHAKAROV@TTIC.EDU
Toyota Technological Institute at Chicago

Nikita Tsoy NIKITA.TSOY@INSAIT.AI
Kristian Minchev KRISTIAN.MINCHEV@INSAIT.AI
Nikola Konstantinov NIKOLA.KONSTANTINOV@INSAIT.AI
INSAIT, Sofia University "St. Kliment Ohridski"

Abstract

It is well-known that Federated Learning (FL) is vulnerable to manipulated updates from clients. In this work we study the impact of data heterogeneity on clients' incentives to manipulate their updates. We formulate a game in which clients may upscale their gradient updates in order to "steer" the server model to their advantage. We develop a payment rule that disincentivizes sending large gradient updates, and steers the clients towards truthfully reporting their gradients. We also derive explicit bounds on the clients' payments and the convergence rate of the global model, which allows us to study the trade-off between heterogeneity, payments and convergence.

Keywords: Federated Learning, Heterogeneity, Game Theory, Incentive Design

1. Introduction

Federated learning (FL) [22] enables the efficient training of machine learning models on large datasets, distributed among multiple stakeholders, via gradient updates shared with a central server. FL has the potential to provide state-of-the-art models in multiple domains where high-quality training data is scarce and distributed, for example healthcare, finance and agriculture [15].

Unfortunately, the distributed nature of standard FL protocols makes them susceptible to clients misreporting their gradient updates. Indeed, prior work has shown that a small fraction of malicious participants can damage the learned model with seemingly benign updates [2, 3]. Furthermore, it is known that the presence of market competition [10], privacy concerns [26] and high data gathering costs [17] may incentivize clients to share updates that are harmful for the global model. These issues bring the practical merit of federated learning in the presence of misaligned incentives into question.

In this paper, we argue that incentives for update manipulation may appear even between clients who are solely interested in their own accuracy, as long as they have different data distributions. Since data heterogeneity is ubiquitous in common federated learning scenarios [15, 22], this implies that clients could be incentivized to manipulate their updates in realistic scenarios and even without the presence of explicitly conflicting goals, such as those arising from competition and privacy concerns. To incentivize clients to share truthful gradient updates we propose a budget-balanced payment scheme, such that (1) truthful reporting results in utility that is ε-close to optimal for a client i given everyone else is sending truthful updates, and (2) the best response of client i when everyone else is

* Part of this work was done while DC was visiting INSAIT, Sofia University "St. Kliment Ohridski."

sending truthful updates is ε-close to truthful. We provide an explicit convergence rate for the global objective function and bound the total payments for each client. Finally, we look at the impact of different types of data heterogeneity on clients' payments and the convergence rate.

2. Related work

Robustness in federated learning. Several works consider Byzantine-robust distributed learning and show that malicious workers can prevent any protocol from converging at optimal rates [2, 3, 29]. Prior work has also explored the robustness of FL to noise and bias towards subgroups, e.g. [1, 11]. We refer to [25] for a recent survey. Our work takes a different approach towards securing FL from harmful updates, as we model the clients as rational and seek to ensure that honest reporting is maximally beneficial for them.

Incentives in collaborative learning. A major research direction is that of studying whether clients have an incentive to join the FL protocol, relative to participation costs (e.g. compute resources and data collection costs). We refer to [27, 30] for recent surveys. Among works that consider incentives for manipulating updates, [17] focus on free-riding, [13] study defection, [4] look at agents who wish to keep their data collection low, while [14] try to incentivize diverse data contributions. Moreover, [10] and [26] consider manipulating updates due to incentives stemming from competition and privacy, respectively. Another behavior driven by the heterogeneity of client data is studied by [8, 9], who model FL as a coalitional game, where players need to decide how to cluster in groups to improve their performance. In contrast, we study non-cooperative games and consider actions that alter the client updates and hence may be harmful to the central model. In our setting the clients are solely interested in their own accuracy and utility, but may still have conflicting incentives due to data heterogeneity.

Heterogeneity in FL. The impact of data heterogeneity on the quality of the learned model is of central interest in the FL literature. Some works study how to train a central model and also provide personalization for each individual client, in order to maximize the accuracy for each client [20–23]. Others focus on the impact of heterogeneity on model convergence and on providing algorithms that provide more accurate centralized models in a heterogeneous environment [16, 18, 19, 24, 28]. These works tackle issues from data heterogeneity from the server's perspective, while we take a mechanism-design point of view and focus on the clients' behavior. Prior work [6] has also explored how to adapt to heterogeneity from a client's perspective—how to optimally weight local and server updates, however, we study how clients might manipulate their messages to the server, thus damaging the global training process, and try to mitigate this potential manipulation.

3. Setting and Preliminaries

In this section we introduce our heterogeneous FL setup. Then we motivate and formally define a game that describes the interaction between self-interested clients.

3.1. FL setup and protocol

Learning setup. We consider a setting with N clients that seek to obtain an accurate model by exchanging messages (gradient updates) with a central server, which orchestrates the FL protocol. All clients work with a shared loss function $f(\theta; z)$ that is differentiable in θ for every z. Each client

i has her own distribution D_i over the data $z \in \mathcal{Z}$.[1] Each client is interested in minimizing the expected loss with respect to their own distribution, so the objective function of client i is $F_i(\theta) = \mathbb{E}_{z\sim D_i}[f(\theta; z)]$. The server's objective is to minimize the average loss of all clients, i.e. $F(\theta) = \frac{1}{N}\sum_{i=1}^{N} F_i(\theta)$. Throughout the paper we assume that for each client $i \in [N]$ their objective $F_i(\theta)$ is L-Lipschitz, H-smooth and m-strongly convex on a compact set $\Theta \subset \mathbb{R}^d$. We assume that $\nabla F_i(\theta) = \mathbb{E}_{z\sim D_i}[\nabla f(\theta; z)]$ and that the gradients have bounded variance $\mathbb{E}_{z\sim D_i}\left[\|\nabla f(\theta; z) - \nabla F_i(\theta)\|_2^2\right] \leq \sigma^2$ for all $\theta \in \Theta$. Let $R = \sup_{\theta,\theta'} \|\theta - \theta'\|$ be the largest distance between any two points in Θ.[2]

FL protocol. We consider the standard FedSGD protocol, where the server asks the clients to send stochastic gradients at the current model, with respect to their own distribution. Client i computes a stochastic gradient $g_i(\theta) := \nabla f(\theta; z)$ by sampling from their distribution $z \sim D_i$, such that $\mathbb{E}_{z\sim D_i}[g_i(\theta)] = \mathbb{E}_{z\sim D_i}[\nabla f(\theta; z)] = \nabla F_i(\theta)$. Let $e_i(\theta) = g_i(\theta) - \nabla F_i(\theta)$ be the gradient noise, so that $\mathbb{E}\left[\|e_i(\theta)\|^2\right] = \mathrm{Var}[g_i(\theta)]$. The server then updates the central model by averaging the updates and taking an SGD step, i.e. $\theta_{t+1} = \Pi_\Theta\left(\theta_t - \gamma_t \frac{1}{N}\sum_{i=1}^{N} g_i(\theta_t)\right)$, where γ_t is the learning rate at step t and Π_Θ is a projection back onto Θ.

3.2. Heterogeneity assumptions

To enable the convergence analysis we evoke an assumption reminiscent to the bounded first-order heterogeneity assumption from the FL literature on the convergence rates of local and mini-batch SGD [16, 18, 19, 24, 28]. Assumption 1 restricts the size of the gradient ∇F_i of client i's objective F_i relative to the gradient ∇F of the aggregate objective F. While standard first-order assumptions previously used in [16, 18, 19, 24, 28] (for formal statements see Assumption 14 and 15 in Appendix B) usually require the gradients of the objectives to be close in some vector norm, we only require them to be close in magnitude.

Assumption 1 (Bounded Gradient Difference) *For every client i and every $\theta \in \Theta$, we have $\left|\|\nabla F_i(\theta)\|^2 - \|\nabla F(\theta)\|^2\right| \leq \zeta^2$. Moreover, as a consequence $\left|\|\nabla F_i(\theta)\|^2 - \|\nabla F_j(\theta)\|^2\right| \leq 2\zeta^2$.*

Next, Assumption 2 below controls the difference between the variance of stochastic gradients. This allows us to study scenarios, where the objectives are sufficiently similar, but the variance of some client i (relative to others) might induce an incentive to misreport.

Assumption 2 (Bounded Variance Difference) *For any pair of clients $i, j \in [N]$ and any $\theta \in \Theta$, we have $|\mathrm{Var}[g_i(\theta)] - \mathrm{Var}[g_j(\theta)]| \leq \rho^2$.*

3.3. Game-theoretic framework

In this subsection we first provide an example scenario illustrating why clients may be able to increase their utility by upscaling their updates. Then, we introduce our game setup which captures these interactions.

1. As is standard in the stochastic optimization analysis of FL, we do not assume any parametric form of the distributions, but instead formulate assumptions on the objective functions and (stochastic) gradients that arise from each distribution.
2. Because Θ is compact, it is bounded, and so the supremum is finite.

Motivating example. At each round of the FL protocol, the server effectively tries to estimate the mean of the gradients of all participants. Since the server just averages the received updates, the players can easily "shift" the global mean towards their local mean by upscaling their message.

To see why, consider the following simple example. Assume that N clients seek to estimate their respective means $\mu_1, \ldots, \mu_N \in \mathbb{R}$, where without loss of generality $\mu_1 > \mu_2 > \ldots > \mu_N$. Each client has an independent sample $x_i \sim \mathcal{N}(\mu_i, \sigma^2)$, and sends a message m_i (supposedly their sample x_i) to the server. Then the server computes an aggregate $\overline{\mu} = \sum_{i=1}^N m_i/N$ and broadcasts it to all clients. Each client i would like to receive an estimate of their local mean with minimal mean squared error $\mathbb{E}[(\overline{\mu} - \mu_i)^2]$.

Proposition 3 *Let $\mu = \frac{1}{N}\sum_{i=1}^N \mu_i$ and assume that $\mu_1(\mu_1 - \mu) > \sigma^2/N$ and that everyone but the first client truthfully reports their sample. If the first client truthfully reports their sample x_1, then $\mathbb{E}\left[\overline{\mu}\right] = \mu$ and $Var(\overline{\mu}) = \sigma^2/N$, and so $\mathbb{E}\left[(\overline{\mu} - \mu_1)^2\right] = (\mu - \mu_1)^2 + \sigma^2/N$.*

However, if the first client misreports cx_i, where $c > 1$ is some appropriate constant, the client can reduce their MSE compared to the truthful message by $\frac{(\mu_1(\mu_1-\mu)-\sigma^2/N)^2}{\mu_1^2+\sigma^2}$.

Hence, in this setting the first player is incentivized to increase their reported sample in order to reduce their error. The assumed inequality $\mu_1(\mu_1 - \mu) > \sigma^2/N$ holds whenever the heterogeneity of the data distributions $(\mu - \mu_1)$ is large with respect to the variance σ^2/N of the aggregated message and the larger this gap, the larger the benefits of misreporting. Intuitively, truthful reporting results a server result close to μ, which is far from the target means of the "extremal" clients whose μ_i is far from μ. Hence, these clients can bias the aggregated mean in their favor by altering their update.

The heterogeneous FL game. Given that FL clients can direct the server model in their interest via upscaling their update, we consider a game in which the clients seek to improve their own loss function by manipulating the messages they send to the server. Specifically, each client sends message $m_t^i = a_t^i g_i(\theta_t)$, where $\left|a_t^i\right| \geq 1$. The rest of the process remains unchanged, with the server computing $\bar{m}_t = \frac{1}{N}\sum_{i=1}^N m_t^i$ and $\theta_{t+1} = \Pi_\Theta\left(\theta_t - \gamma_t \bar{m}_t\right)$, and communicating θ_{t+1} to all clients at each round. As is standard in game theory, we assume that the clients are rational, i.e. they would like to select actions a_t^i that increase their utility. We let the utility U_i of player i to be $U_i = R_i - p_i$, where R_i is the reward the client gets from the end model, and p_i is the total payment the client makes to the server for participating in the FL protocol. In the canonical case, we consider the reward of client i to be $R_i(\theta) = -F_i(\theta)$ (see Theorem 6), that is, the negative of the expected loss of the end model on the client's distribution. Then, we also show how to extend our analysis to more general reward functions in Corollary 7.

Desiderata. We seek to design a payment mechanism, or payment rule, such that the clients are incentivized to send meaningful updates, in order for the final model to achieve small loss on the global objective $F(\theta)$. In particular, we would like our payment mechanism to satisfy two properties: (1) be ε-approximately incentive compatible, meaning that whenever all clients $j \neq i$ are reporting truthfully, then truthful reporting is ε-close to the optimal utility for client i as well (see Definition 4); and (2) the best response strategy of client i when everyone else is reporting truthfully is ε-close to truthful reporting (see Definition 5). Below we formally define these two desirable properties for a federated learning protocol.[3]

3. The curious reader might want to compare the usual definition of Bayesian Incentive Compatibility from mechanism design in Appendix C with the one in this section.

Definition 4 (Bayesian Incentive Compatibility) *A federated learning protocol M is ε-*Bayesian Incentive Compatible *(BIC) if:*

$$\mathbb{E}\left[U_i\left(M,\{\mathbf{1}_1,\ldots,\mathbf{1}_i,\ldots,\mathbf{1}_N\}\right)\right] \geq \mathbb{E}\left[U_i\left(M,\{\mathbf{1}_1,\ldots,\mathbf{a}_i,\ldots,\mathbf{1}_N\}\right)\right] - \varepsilon,$$

where $\mathbf{1}_i = (1,\ldots,1) \in \mathbb{R}^T$ denotes fully truthful participation by client i, $\mathbf{1}_j$ denotes truthful participation by client $j \neq i$, $\mathbf{a}_i = \left\{a_t^i\right\}_{t=1}^T$ is some arbitrary strategy of client i, and $U_i(M,\mathbf{v})$ is the utility of client i from M when clients are using the strategy profile $\mathbf{v}$. Note that expectation is taken over the randomness in the clients' distributions, and any randomness (possibly none) in the protocol.

Definition 5 (Approximately truthful reporting) *A strategy a_t^i of client i is approximately truthful if it satisfies $\mathbb{E}\left[\left\|a_t^i g_i(\theta_t) - g_i(\theta_t)\right\|^2\right] \leq \varepsilon^2$. Moreover, in our analysis we require that the best response of client i to truthful participation by clients $j \neq i$ is approximately truthful.*

4. Theoretical Results

In this section, we study a payment scheme which we prove is both ε-BIC and incentivizes approximately truthfull reporting. We also provide explicit bounds on the penalties a client may pay and on the achieved rates of convergence of the global model.

At each step the server "charges" client i the payment:

$$p_t^i(\boldsymbol{m}_t) = C_t\left[\left\|m_t^i\right\|^2 - \frac{1}{N-1}\sum_{j\neq i}\left\|m_t^j\right\|^2\right], \tag{1}$$

where C_t is some client-independent constant (see the individual results below for definition). The total payment for each client is then $p_i = \sum_{i=1}^T p_t^i(\boldsymbol{m}_t)$. Note that this payment rule is *budget-balanced*, that is at each step the server neither makes nor loses money because $\sum_{i=1}^N p_t^i(\boldsymbol{m}_t) = 0$.

4.1. Approximately truthful reporting

First we show that the payment scheme is both ε-BIC and incentivizes approximately truthful reporting.

Theorem 6 (Properties of the payment scheme) *Suppose that for all clients i the objective function F_i is L-Lipschitz, H-smooth and m-strongly-convex. Also assume that the gradient noise $e_i(\theta) = g_i(\theta) - \nabla F_i(\theta)$ is D-Lipschitz with probability 1. Set $C_t = \frac{\sqrt{2\mathcal{C}_t}\gamma_t L}{N\varepsilon}$, where $\mathcal{C}_t = \prod_{l=t+1}^T c_l$ and $c_l = 2\left(1 - 2\gamma_l m + \gamma_l^2(H^2+D^2)\right)$. Then Payment Rule (1) is $O(\varepsilon)$-BIC (as $\varepsilon \to 0$) and the best response strategy a_t^i of client i to truthful participation from everyone else satisfies $\mathbb{E}\left[\left\|a_t^i g_i(\theta_t) - g_i(\theta_t)\right\|^2\right] \leq \varepsilon^2$ for all t.*

We also note that a similar result holds for any reward function which is Lipschitz in θ. Two relevant examples of such reward functions are (1) when a group of clients S_1 seeks to decrease their average loss, i.e. $R_i = -\sum_{i\in S_1} F_i(\theta_T)$ (reminiscent of [8, 9]), and (2) when a client benefits if a group of other clients S_2 does poorly, i.e. $R_i = \sum_{i\in S_2} F_i(\theta_T)$ (similar to [10]).

Corollary 7 (General Lipschitz reward) *In the setting of Theorem 6 suppose that each client i has an L'-Lipschitz reward function $R_i : \Theta \to \mathbb{R}_{\geq 0}$, so that $|R_i(\theta) - R_i(\theta')| \leq L'|\theta - \theta'|$. Then the payment from Theorem 6 with $C_t = \frac{\sqrt{2\mathcal{C}_t}\gamma_t L'}{N\varepsilon}$ is $O(\varepsilon)$-BIC (as $\varepsilon \to 0$) and the best response strategy a_t^i of client i to truthful participation from everyone else satisfies $\mathbb{E}\left[\left\|a_t^i g_i(\theta_t) - g_i(\theta_t)\right\|^2\right] \leq \varepsilon^2$ for all t. Note that setting $R_i = -F_i$ recovers the original result.*

Intuitively, the two results demonstrate that if all other players $j \neq i$ are truthful, then it is in the interest of player i to be (approximately) truthful as well, and this will yield close to optimal utility.

4.2. Payments and convergence

Finally, we provide explicit upper bounds on the total penalty paid by each player, as well as on the convergence speed for the loss function of each client, whenever the clients are approximately truthful. These bounds allow us to discuss the interplay between learning quality and penalties, as well as their dependence on the parameters of the FL protocol.

Proposition 8 (Bound on individual payments) *Suppose all participants are approximately truthful at each time step, i.e, $\mathbb{E}\left[\left\|a_t^i g_i(\theta_t) - g_i(\theta_t)\right\|^2\right] \leq \varepsilon^2$. Then the total payment is bounded by $\sum_{t=1}^T p_t^i(\boldsymbol{m}_t) \leq \frac{\sqrt{2}LG}{N}\left[\varepsilon^2 + 2\varepsilon\left(HR + \sigma\right) + 2\zeta^2 + \rho^2\right]$, where $G = \sum_{t=1}^T \gamma_t \sqrt{\mathcal{C}_t}$.*

Theorem 9 (Convergence rate) *Suppose all players are approximately truthful at each time step, i.e. $\mathbb{E}\left[\left\|a_t^i g_i(\theta_t) - g_i(\theta_t)\right\|^2\right] \leq \varepsilon^2$. Assume that there exist scalars $M \geq 0$ and $M_V \geq 0$, such that $\mathbb{E}\left[\left\|e_i(\theta_t)\right\|^2\right] \leq M + M_V \left\|\nabla F_i(\theta_t)\right\|^2$ for every t. Next, fix an integer constant $\eta = \frac{4H(M_V+N)}{mN}$ and set the learning rate to be $\gamma_t = \frac{4}{m(\eta+t)}$. Given that $\mathbb{P}\left[\exists t \leq T : \Pi_\Theta\left(\theta_t - \gamma_t \bar{m}_t\right) \neq \theta_t - \gamma_t \bar{m}_t\right] = O\left(\frac{1}{NT}\right)$, then we have $\mathbb{E}\left[F(\theta_T) - F(\theta^*)\right] \leq \frac{16H(\varepsilon^2+M+M_V\zeta^2)}{3Nm^2(\eta+T)} + O\left(\frac{1}{NT}\right) + O\left(\frac{1}{T^2}\right)$.*

Discussion. The following two scenarios explore the effect of higher levels of heterogeneity on the expected payment of each player and the convergence rate of the federal learning algorithm. In particular, we seek to understand what is the trade-off between heterogeneity and payments/convergence rate. Both of these directly rely on Theorems 6 and 9.

Example 1 (Constant heterogeneity bounds) *Suppose ζ and ρ are both constants chosen before the learning process is ever run. Then the total payment made by each player is at most $O\left(\frac{1}{N}\right)$ and the convergence rate becomes $O\left(\frac{1}{NT}\right)$. Hence, both the expected payments and the convergence rate decrease linearly in N, while preserving the ε-BIC property and the approximate truthful reporting property.*

Example 2 (Scaling the heterogeneity bounds for large N) *Notice that even if the heterogeneity bounds increase with the number of clients, our results still give reasonable bounds on the individual payments and the convergence rate. Suppose ζ and ρ are both of order $O\left(\sqrt[4]{N}\right)$. Then the total payment made by each player is at most $O\left(\frac{1}{\sqrt{N}}\right)$ and the convergence rate becomes $O\left(\frac{1}{\sqrt{N}T}\right)$. Hence, as we increase the number of participants N we can simultaneously (1) increase the threshold for heterogeneity; and (2) reduce the maximal individual payment, while (3) preserving convergence.*

Acknowledgements. This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure). The authors would like to thank Florian Dorner and Kumar Kshitij Patel for several helpful discussions at various stages of the project.

References

[1] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

[2] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[3] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[4] Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, and Han Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *International Conference on Machine Learning*, pages 1005–1014. PMLR, 2021.

[5] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[6] El Mahdi Chayti, Sai Praneeth Karimireddy, Sebastian U Stich, Nicolas Flammarion, and Martin Jaggi. Linear speedup in personalized collaborative learning. *arXiv preprint arXiv:2111.05968*, 2021.

[7] Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.

[8] Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. *AAAI Conference on Artificial Intelligence*, 2021.

[9] Kate Donahue and Jon Kleinberg. Optimality and stability in federated learning: A game-theoretic approach. *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[10] Florian E Dorner, Nikola Konstantinov, Georgi Pashaliev, and Martin Vechev. Incentivizing honesty among competitors in collaborative learning and optimization. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[11] Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

[13] Minbiao Han, Kumar Kshitij Patel, Han Shao, and Lingxiao Wang. On the effect of defections in federated learning and how to prevent them. *arXiv preprint arXiv:2311.16459*, 2023.

[14] Baihe Huang, Sai Praneeth Karimireddy, and Michael I Jordan. Evaluating and incentivizing diverse data contributions in collaborative learning. *arXiv preprint arXiv:2306.05592*, 2023.

[15] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 2021.

[16] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[17] Sai Praneeth Karimireddy, Wenshuo Guo, and Michael Jordan. Mechanisms that incentivize data sharing in federated learning. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022.

[18] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

[19] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

[20] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

[21] Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized federated learning through local memorization. In *International Conference on Machine Learing (ICML)*, 2022.

[22] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2017.

[23] Konstantin Mishchenko, Rustem Islamov, Eduard Gorbunov, and Samuel Horváth. Partially personalized federated learning: Breaking the curse of data heterogeneity. *arXiv preprint arXiv:2305.18285*, 2023.

[24] Kumar Kshitij Patel, Margalit Glasgow, Ali Zindari, Lingxiao Wang, Sebastian U Stich, Ziheng Cheng, Nirmit Joshi, and Nathan Srebro. The limits and potentials of local sgd for distributed heterogeneous learning with intermittent communication. *Conference on Learning Theory (COLT)*, 2024.

[25] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *IEEE Symposium on Security and Privacy*, 2022.

[26] Nikita Tsoy, Anna Mihalkova, Teodora N Todorova, and Nikola Konstantinov. Provable mutual benefits from federated learning in privacy-sensitive domains. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2024.

[27] Xuezhen Tu, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Yang Zhang, and Juan Li. Incentive mechanisms for federated learning: From economic and game theoretic perspective. *arXiv preprint arXiv:2111.11850*, 2021.

[28] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33: 6281–6292, 2020.

[29] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, 2018.

[30] Yufeng Zhan, Jie Zhang, Zicong Hong, Leijie Wu, Peng Li, and Song Guo. A survey of incentive mechanism design for federated learning. *IEEE Transactions on Emerging Topics in Computing*, 10(2):1035–1044, 2021.

Appendix A. Theoretical Refresher

Definition 10 (Lipschitzness) *A function* $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}^m$ *is* L-Lipschitz *if*

$$\|f(x) - f(y)\|_m \leq L \|x - y\|_n$$

for all $x, y \in \mathcal{X}$.

Definition 11 (Smoothness) *Let* $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ *be a differentiable function. Then* f *is* H-smooth *if*

$$\|\nabla f(x) - \nabla f(y)\| \leq H \|x - y\|,$$

for all $x, y \in \mathcal{X}$. *In other words,* f *is* H*-smooth if its gradient* ∇f *is* H*-Lipschitz. Moreover, this condition is equivalent to*

$$|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \leq \frac{H}{2} \|x - y\|^2 .$$

Definition 12 (Strong convexity) *Let* $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ *be a differentiable function. Then* f *is* m-strongly-convex *if*

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{m}{2} \|x - y\|^2 .$$

Proposition 13 *Let* $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ *be a differentiable function that is both* H*-smooth and* m*-strongly-convex, then*

$$\frac{m}{2} \|x - y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{H}{2} \|x - y\|^2 ,$$

In other words, the error between f *and its linear approximation is bounded by quadratics from both above and below.*

Appendix B. Similar heterogeneity assumptions from previous works

Assumption 14 (Bounded First-Order Heterogeneity, [19, 28]) *For any client* i *and any* $\theta \in \Theta$ *we have:*

$$\sup_{\theta \in \Theta, i \in [N]} \|\nabla F_i(\theta) - \nabla F(\theta)\|_2^2 \leq \zeta^2.$$

Assumption 15 (Relaxed First-Order Heterogeneity, [16]) *There exist constants* $G \geq 0$ *and* $D \geq 1$*, such that for every* $\theta \in \Theta$ *the following holds:*

$$\frac{1}{N} \sum_{i=1}^{N} \|\nabla F_i(\theta)\|^2 \leq G^2 + D^2 \|\nabla F(\theta)\|^2 .$$

Remark 16 *If Assumption 15 holds with* $D = 0$ *and* $\theta^* \in \arg\min_{\theta \in \Theta} F(\theta)$*, then we recover another assumption used in the literature—*Bounded First-Order Heterogeneity at Optima *[12, 19, 28].*

Appendix C. Note on Incentive Compatibility

The usual definition of Bayesian Incentive Compatibility from Mechanism Design is the following. Suppose that each player i has a true type $t_i \in T_i$, and a utility function $u_i(t, o)$ that takes as input a type t and an outcome o, and outputs a value. Let $o(t_i, t_{-i})$ be the outcome of the mechanism M on input (t_i, t_{-i}). Let $p_i(t_i, t_{-i})$ be the payment of player i according to M.

Definition 17 (Bayesian Incentive Compatibility) *A mechanism M is* Bayesian Incentive Compatible *(BIC) if:*

$$\mathbb{E}\left[u_i(t_i, o(t_i, t_{-i})) - p_i(t_i, t_{-i})\right] \geq \mathbb{E}\left[u_i(t_i, o(t'_i, t_{-i})) - p_i(t'_i, t_{-i})\right],$$

where $t'_i \in T_i$ is any type for player i, and expectation is taken with respect to the randomness in the types of everyone but player i.

Moreover, we can relax the condition to have an approximately BIC mechanism up to an additive constant, denoted ε-BIC:

$$\mathbb{E}\left[u_i(t_i, o(t_i, t_{-i})) - p_i(t_i, t_{-i})\right] \geq \mathbb{E}\left[u_i(t_i, o(t_i, t_{-i})) - p_i(t'_i, t_{-i})\right] - \varepsilon.$$

Appendix D. Proof of Proposition 3

One can easily check that if the first player is truthful, then the MSE is as stated. Now, consider the case where the first client can lie by selecting a constant $c > 1$. In particular, if we select a constant c such that

$$1 < c < \frac{2\mu_1(\mu_1 - \mu)N + \mu_1^2 - \sigma^2}{\mu_1^2 + \sigma^2},$$

then the we have $\mathbb{E}\left[\overline{\mu}\right] = \frac{c-1}{N}\mu_1 + \mu$ and $Var(\overline{\mu}) = \frac{\sigma^2}{N} + \frac{c^2-1}{N^2}\sigma^2$. Then the MSE for the first player is

$$\begin{aligned}
\mathbb{E}\left[(\overline{\mu} - \mu_1)^2\right] &= \left(\mu + \frac{c - N - 1}{N}\mu_1\right)^2 + \frac{\sigma^2}{N} + \frac{c^2 - 1}{N^2}\sigma^2 \\
&= (\mu - \mu_1)^2 + 2\frac{c-1}{N}\mu_1(\mu - \mu_1) + \frac{(c-1)^2}{N^2}\mu_1^2 + \frac{c^2-1}{N^2}\sigma^2 + \frac{\sigma^2}{N} \\
&= (\mu - \mu_1)^2 + \frac{\sigma^2}{N} + \frac{c-1}{N^2}\left(2\mu_1\left(\mu - \mu_1\right)N + (c-1)\,\mu_1^2 + (c+1)\,\sigma^2\right) \\
&= (\mu - \mu_1)^2 + \frac{\sigma^2}{N} + \frac{c-1}{N^2}\left(c\left(\mu_1^2 + \sigma^2\right) - 2\mu_1\left(\mu_1 - \mu\right)N - \mu_1^2 + \sigma^2\right).
\end{aligned}$$

Note that the last term is a quadratic in c and hence is minimized at

$$c = \frac{\mu_1^2 + \mu_1(\mu_1 - \mu)}{\mu_1^2 + \sigma^2} > 1.$$

Note that $c > 1$ is guaranteed by the assumed inequality $\mu_1(\mu_1 - \mu) > \sigma^2/N$. In this case, the MSE of the client is

$$\mathbb{E}\left[(\bar{\mu} - \mu_1)^2\right] = (\mu - \mu_1)^2 + \frac{\sigma^2}{N} - \frac{(\sigma^2/N - \mu_1(\mu_1 - \mu))^2}{\mu_1^2 + \sigma^2},$$

which completes our proof.

Appendix E. Proof of Theorem 6

Claim 18 (Per-turn bound on trajectory difference) *Fix a client i. Let $\theta = \{\theta_t\}_{t=1}^{T+1}$ and $\theta' = \{\theta'_t\}_{t=1}^{T+1}$, be two trajectories obtained from two distinct strategy profiles $\{a_t^i\}_{t=1}^{T}$ and $\{\bar{a}_t^i\}_{t=1}^{T}$ of client i, while everyone else is doing the same in both scenarios. Then at time $t+1$:*

$$\mathbb{E}\left[\left\|\theta_{t+1} - \theta'_{t+1}\right\|^2\right] \leq c_t \mathbb{E}\left[\left\|\theta_t - \theta'_t\right\|^2\right] + \frac{2\gamma_t^2}{N^2}(a_t^i - \bar{a}_t^i)^2 \mathbb{E}\left[\left\|g_i(\theta'_t)\right\|^2\right],$$

where $c_t = 2\left(1 - \frac{2\gamma_t m A_t}{N} + \frac{\gamma_t^2 A_t^2 (H^2 + D^2)}{N^2}\right)$ and $A_t = \sum_{j=1}^{N} a_t^j$.

Proof Observe the following sequence

$$\begin{aligned}
\mathbb{E}\left[\left\|\theta_{t+1} - \theta'_{t+1}\right\|^2\right] &= \mathbb{E}\left[\left\|\Pi_\Theta\left(\theta_t - \gamma_t \bar{m}_t\right) - \Pi_\Theta\left(\theta'_t - \gamma_t \bar{m}'_t\right)\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|(\theta_t - \theta'_t) - \gamma_t(\bar{m}_t - \bar{m}'_t)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|(\theta_t - \theta'_t) - \frac{\gamma_t}{N}\sum_{j=1}^{N}\left(a_t^j g_j(\theta_t) - \bar{a}_t^j g_j(\theta'_t)\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{(\bar{a}_t^i - a_t^i)\gamma_t}{N} g_i(\theta'_t) + (\theta_t - \theta'_t) - \frac{\gamma_t}{N}\sum_{j=1}^{N} a_t^j\left(g_j(\theta_t) - g_j(\theta'_t)\right)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\frac{(\bar{a}_t^i - a_t^i)\gamma_t}{N} g_i(\theta'_t)\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\|(\theta_t - \theta'_t) - \frac{\gamma_t}{N}\sum_{j=1}^{N} a_t^j\left(g_j(\theta_t) - g_j(\theta'_t)\right)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|(\theta_t - \theta'_t) - \frac{\gamma_t}{N}\sum_{j=1}^{N} a_t^j\left(g_j(\theta_t) - g_j(\theta'_t)\right)\right\|^2\right] \\
&\quad + \frac{2\gamma_t^2 (a_t^i - \bar{a}_t^i)^2 \mathbb{E}\left[\|g_i(\theta'_t)\|^2\right]}{N^2} \\
&= 2\mathbb{E}\left[\left\|\theta_t - \theta'_t\right\|^2\right] + 2\mathbb{E}\left[\left\|\frac{\gamma_t}{N}\sum_{j=1}^{N} a_t^j\left(g_j(\theta_t) - g_j(\theta'_t)\right)\right\|^2\right] \\
&\quad - 4\mathbb{E}\left[\frac{\gamma_t}{N}\sum_{j=1}^{N} a_t^j\left\langle \theta_t - \theta'_t, g_j(\theta_t) - g_j(\theta'_t)\right\rangle\right] \\
&\quad + \frac{2\gamma_t^2 (a_t^i - \bar{a}_t^i)^2 \mathbb{E}\left[\|g_i(\theta'_t)\|^2\right]}{N^2}
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\mathbb{E}\left[\left\|\theta_t - \theta_t'\right\|^2\right] + \frac{2\gamma_t^2 A_t^2 (H^2 + D^2)}{N^2}\mathbb{E}\left[\left\|\theta_t - \theta_t'\right\|^2\right] \\
&\quad - \frac{4\gamma_t m A_t}{N}\mathbb{E}\left[\left\|\theta_t - \theta_t'\right\|^2\right] + \frac{2\gamma_t^2 (a_t^i - \bar{a}_t^i)^2 \mathbb{E}\left[\|g_i(\theta_t')\|^2\right]}{N^2} \\
&= 2\left(1 - \frac{2\gamma_t m A_t}{N} + \frac{\gamma_t^2 A_t^2 (H^2 + D^2)}{N^2}\right)\mathbb{E}\left[\left\|\theta_t - \theta_t'\right\|^2\right] \\
&\quad + \frac{2\gamma_t^2}{N^2}(a_t^i - \bar{a}_t^i)^2 \mathbb{E}\left[\left\|g_i(\theta_t')\right\|^2\right]
\end{aligned}
$$

The first line follow because the projection operator is non-expansive. The next three lines are rearrangement. The fifth line uses the inequality $(x+y)^2 \leq 2x^2 + 2y^2$.[4] On line eight (second to last inequality), the second term follows from H-smoothness of F_i and D-Lipschitzness of e_i, while the third term follows from the m-strong-convexity of F_i. ∎

Corollary 19 (Bound on trajectory difference) *Suppose that $\theta = \{\theta_t\}_{t=1}^{T+1}$ is obtained from fully truthful participation from all clients. Then at time $T+1$:*

$$
\mathbb{E}\left[\left\|\theta_{T+1} - \theta_{T+1}'\right\|^2\right] \leq \frac{2}{N^2}\sum_{t=1}^{T} \gamma_t^2 \mathcal{C}_t (\bar{a}_t^i - 1)^2 \mathbb{E}\left[\left\|g_i(\theta_t')\right\|^2\right],
$$

where $\mathcal{C}_t = \prod_{l=t+1}^{T} c_l$ *and* $c_l = 2\left(1 - 2\gamma_l m + \gamma_l^2 (H^2 + D^2)\right)$.

Proof Apply Claim 18 and telescope. Because the reference strategy is all reporting truthfully, then we have $A = N$ in Claim 18. ∎

Claim 20 (Bound on reward) *Suppose that $\theta = \{\theta_t\}_{t=1}^{T+1}$ is obtained from fully truthful participation from all clients. Then at time $T+1$:*

$$
\mathbb{E}\left[\left|F_i(\theta_{T+1}) - F_i(\theta_{T+1}')\right|\right] \leq \frac{\sqrt{2}L}{N}\sum_{t=1}^{T} \gamma_t \sqrt{\mathcal{C}_t}\left(\bar{a}_t^i - 1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta_t')\|^2\right]},
$$

where $\mathcal{C}_t = \prod_{t'=1}^{t} c_{t'}$ *and* $c_l = 2\left(1 - 2\gamma_l m + \gamma_l^2 (H^2 + D^2)\right)$.

Proof Observe:

$$
\begin{aligned}
\left(\mathbb{E}\left[\left|F_i(\theta_{T+1}) - F_i(\theta_{T+1}')\right|\right]\right)^2 &\leq L^2\left(\mathbb{E}\left[\left\|\theta_{T+1} - \theta_{T+1}'\right\|\right]\right)^2 \\
&\leq L^2 \mathbb{E}\left[\left\|\theta_{T+1} - \theta_{T+1}'\right\|^2\right]
\end{aligned}
$$

4. We do this roundabout bounding, so that all terms are squared norms. Also notice that $(x+y)^2 \leq 2x^2 + 2y^2$ is equivalent to $0 \leq (x-y)^2$, which is trivially true.

$$
\begin{aligned}
&\leq L^2\mathcal{C}_T\mathbb{E}\left[\left\|\theta_1-\theta_1'\right\|^2\right]+\frac{2L^2}{N^2}\sum_{t=1}^{T}\gamma_t^2\mathcal{C}_t(\bar{a}_t^i-1)^2\mathbb{E}\left[\left\|g_i(\theta_t')\right\|^2\right]\\
&=\frac{2L^2}{N^2}\sum_{t=1}^{T}\gamma_t^2\mathcal{C}_t(\bar{a}_t^i-1)^2\mathbb{E}\left[\left\|g_i(\theta_t')\right\|^2\right]
\end{aligned}
$$

The first line follows from $F_i(\cdot)$ being L-Lipschitz. The second line follows from Cauchy-Schwartz. The third line applies Corollary 19, and uses the assumption $\theta_1 = \theta_1'$. Finally, to get the desired inequality we observe that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for non-negative x, y. ■

Claim 21 (Bound on payment) *Suppose that θ and θ' differ only at the gradient reported by client i at time t, so $\bar{a}_t^i > 1$, and that θ is obtained from truthfulness. Then*

$$
\mathbb{E}\left[p_t^i(\boldsymbol{m}_t)-p_t^i(\boldsymbol{m}_t')\right]\leq -C_t(\bar{a}_t^i-1)^2\mathbb{E}\left[\|g_i(\theta_t)\|^2\right].
$$

Proof For what's to follow keep in mind that $\bar{a}_t^i \geq 1$ by assumption. Observe:

$$
\begin{aligned}
\mathbb{E}\left[p_t^i(\boldsymbol{m}_t)-p_t^i(\boldsymbol{m}_t')\right]&=C_t\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]-C_t\mathbb{E}\left[\frac{1}{N-1}\sum_{j\neq i}\|g_j(\theta_t)\|^2\right]\\
&\quad -C_t\mathbb{E}\left[\left\|\bar{a}_t^i g_i(\theta_t)\right\|^2\right]+C_t\mathbb{E}\left[\frac{1}{N-1}\sum_{j\neq i}\|g_j(\theta_t)\|^2\right]\\
&=C_t\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]-C_t\mathbb{E}\left[\left\|\bar{a}_t^i g_i(\theta_t)\right\|^2\right]\\
&=C_t(1-(\bar{a}_t^i)^2)\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]\\
&\leq -C_t(\bar{a}_t^i-1)^2\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]
\end{aligned}
$$

The last line follows from the assumption $\bar{a}_t^i \geq 1$. ■

Proposition 22 (Bound on utility difference between trajectories) *Suppose that θ and θ' differ only at the gradient reported by client i at time t, so $\bar{a}_t^i > 1$, and that θ is obtained from truthfulness. Then*

$$
\begin{aligned}
\mathbb{E}\left[-F_i(\theta_{t+1}')-p_t^i(\boldsymbol{m}_t')-\left(-F_i(\theta_{t+1})-p_t^i(\boldsymbol{m}_t)\right)\right]\leq \frac{\sqrt{2\mathcal{C}_t}\gamma_t L}{N}\left(\bar{a}_t^i-1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]}\\
-C_t(\bar{a}_t^i-1)^2\mathbb{E}\left[\|g_i(\theta_t)\|^2\right].
\end{aligned}
$$

Moreover, for $C_t = \dfrac{\sqrt{2\mathcal{C}_t}\gamma_t L}{N\varepsilon}$ we have:

$$
\mathbb{E}\left[-F_i(\theta_{t+1}')-p_t^i(\boldsymbol{m}_t')-\left(-F_i(\theta_{t+1})-p_t^i(\boldsymbol{m}_t)\right)\right]\leq\frac{\sqrt{2\mathcal{C}_t}\gamma_t L\varepsilon}{N},
$$

and the best strategy $\bar{a}_t^i$ of client i is such that $\mathbb{E}\left[\left\|\bar{a}_t^i g_i(\theta_t)-g_i(\theta_t)\right\|^2\right]\leq\varepsilon^2$.

Proof The first inequality is a direct combination of Claim 20 and Claim 21. Now we tackle the second portion.

First, we write:

$$\mathbb{E}\left[-F_i(\theta'_{t+1}) - p_t^i(\boldsymbol{m}'_t) - \left(-F_i(\theta_{t+1}) - p_t^i(\boldsymbol{m}_t)\right)\right] \leq \frac{\sqrt{2\mathcal{C}_t}\gamma_t L}{N}\left(\bar{a}_t^i - 1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]} - C_t(\bar{a}_t^i - 1)^2\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]$$

Next, the right-hand side expression is a downwards-curved quadratic, and has roots

$$\left(\bar{a}_t^i - 1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta'_t)\|^2\right]} = 0$$

and

$$\left(\bar{a}_t^i - 1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]} = \frac{\sqrt{2\mathcal{C}_t}\gamma_t L}{NC_t}.$$

Then if $C_t = \frac{\sqrt{2\mathcal{C}_t}\gamma_t L}{N\varepsilon}$, for both roots we have $\left(\bar{a}_t^i - 1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]} \leq \varepsilon$. Now observe that the quadratic is positive only between the two roots, i.e. when $0 \leq \left(\bar{a}_t^i - 1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]} \leq \varepsilon$. Therefore,

$$\begin{aligned}\mathbb{E}\left[F_i(\theta'_{t+1}) - p_t^i(\boldsymbol{m}'_t) - \left(F_i(\theta_{t+1}) - p_t^i(\boldsymbol{m}_t)\right)\right] &\leq \frac{\sqrt{2\mathcal{C}_t}\gamma_t L}{N}\left(\bar{a}_t^i - 1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]} \\ &\quad - C_t(\bar{a}_t^i - 1)^2\mathbb{E}\left[\|g_i(\theta_t)\|^2\right] \\ &\leq \frac{\sqrt{2\mathcal{C}_t}\gamma_t L}{N}\left(\bar{a}_t^i - 1\right)\sqrt{\mathbb{E}\left[\|g_i(\theta_t)\|^2\right]} \\ &\leq \frac{\sqrt{2\mathcal{C}_t}\gamma_t L\varepsilon}{N}.\end{aligned}$$

■

Claim 23 (Total bound on utility difference between trajectories) *Suppose θ' denotes some arbitrary reporting strategy and θ denotes truthful reporting. That is we don't require them to be the same up to the last step. Then:*

$$\mathbb{E}\left[-F_i(\theta'_{T+1}) - \sum_{t=1}^{T} p_t^i(\boldsymbol{m}'_t) - \left(-F_i(\theta_{T+1}) - \sum_{t=1}^{T} p_t^i(\boldsymbol{m}_t)\right)\right] \leq \frac{\sqrt{2}L}{N}\left(\sum_{t=1}^{T}\gamma_t\sqrt{\mathcal{C}_t}\right)\varepsilon.$$

Proof Combine Claim 20 and Proposition 22. ■

Appendix F. Proof of Proposition 8

We repeat the statement for completeness.

Claim 24 *Suppose all participants are approximately truthful at each time step, i.e* $\mathbb{E}\left[\left\|a_t^i g_i(\theta_t) - g_i(\theta_t)\right\|^2\right] \leq \varepsilon^2$. *Then the total payment is bounded by*

$$\sum_{t=1}^{T} p_t^i(\boldsymbol{m}_t) \leq \frac{\sqrt{2}LG}{N}\left[\varepsilon^2 + 2\varepsilon\left(HR + \sigma\right) + 2\zeta^2 + \rho^2\right],$$

where $G = \sum_{t=1}^{T} \gamma_t \sqrt{\mathcal{C}_t}$.

Proof First, we consider a single time step:[5]

$$\begin{aligned}
\frac{p_t^i(\boldsymbol{m}_t)}{C_t} &= \mathbb{E}\left[\left\|\bar{a}_t^i g_i(\theta_t)\right\|^2 - \frac{1}{N-1}\sum_{j\neq i}\left\|\bar{a}_t^j g_j(\theta_t)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\bar{a}_t^i e_i(\theta_t)\right\|^2\right] + \left\|\bar{a}_t^i \nabla F_i(\theta_t)\right\|^2 - \frac{1}{N-1}\left(\sum_{j\neq i}\mathbb{E}\left[\left\|\bar{a}_t^j e_j(\theta_t)\right\|^2\right] + \left\|\bar{a}_t^j \nabla F_j(\theta_t)\right\|^2\right) \\
&\leq \mathbb{E}\left[\left(\left(\bar{a}_t^i\right)^2 - 1\right)\left\|g_i(\theta_t)\right\|^2\right] + 2\zeta^2 + \rho^2 \\
&= \mathbb{E}\left[\left(\bar{a}_t^i - 1\right)^2\left\|g_i(\theta_t)\right\|^2\right] + \mathbb{E}\left[2\left(\bar{a}_t^i - 1\right)\left\|g_i(\theta_t)\right\|^2\right] + 2\zeta^2 + \rho^2 \\
&\leq \varepsilon^2 + 2\varepsilon\sqrt{\mathbb{E}\left[\left\|g_i(\theta_t)\right\|^2\right]} + 2\zeta^2 + \rho^2 \\
&\leq \varepsilon^2 + 2\varepsilon\left(HR + \frac{\sigma}{\sqrt{n}}\right) + 2\zeta^2 + \rho^2
\end{aligned}$$

The third line applies Assumption 1 and 2. The forth line is rearrangement, and the fifth applies the assumption $\mathbb{E}\left[\left\|a_t^i g_i(\theta_t) - g_i(\theta_t)\right\|^2\right] \leq \varepsilon^2$. Corollary 29 yields the last inequality.

Therefore, over all time steps we have:

$$\sum_{t=1}^{T} p_t^i(\boldsymbol{m}_t) \leq \frac{\sqrt{2}LG}{N}\left[\varepsilon^2 + 2\varepsilon\left(HR + \frac{\sigma}{\sqrt{n}}\right) + 2\zeta^2 + \rho^2\right].$$

■

Appendix G. Proof of Theorem 9

What follows is the proof of Theorem 9, which establishes the convergence rate in the approximately truthful setting. First, we mention two useful results due to [5] and [7], then we bound the variance of the aggregate gradient at each step (Claim 27). The full proof is at the end.

5. We ignore the constant term for now.

G.1. Results from the literature

Lemma 25 (Equation 4.23, Theorem 4.7 from [5]) *Let F be a continuously differentiable function that is H-smooth and m-strongly-convex. Let $g(\theta)$ be a stochastic gradient of F at θ, such that $\mathbb{E}[g(\theta)] = \nabla F(\theta)$. Suppose there exist scalars $M, M_V \geq 0$, such that $\mathrm{Var}[g(\theta_t)] \leq M + M_V \|\nabla F_i(\theta_t)\|^2$ for every t. If we run SGD with $\gamma_t = \frac{\gamma}{\eta+t}$, where $\gamma > \frac{1}{m}$, $\eta > 0$ and $\gamma_1 \leq \frac{1}{H(M_V+1)}$, then*[6]

$$\mathbb{E}[F(\theta_{t+1}) - F(\theta^*)] \leq (1 - \gamma_t m)\mathbb{E}[F(\theta_t) - F(\theta^*)] + \frac{\gamma_t^2 HM}{2}.$$

Lemma 26 (Lemma 1 from [7]) *Suppose that $\{b_n\}_{n\in\mathbb{N}}$ is a sequence of real numbers such that for $n \geq n_0$,*

$$b_{n+1} \leq \left(1 - \frac{c}{n}\right) b_n + \frac{c_1}{n^{p+1}}$$

where $c > p > 0$, $c_1 > 0$. Then

$$b_n \leq \frac{c_1}{c-p}\frac{1}{n^p} + O\left(\frac{1}{n^{p+1}} + \frac{1}{n^c}\right).$$

G.2. Proof

Claim 27 (Bound on the variance of the aggregated gradient) *Suppose that there exist scalars $M, M_V \geq 0$, such that for every t we have:*

$$\mathbb{E}\left[\|e_i(\theta_t)\|^2\right] \leq M + M_V \|\nabla F(\theta_t)\|^2.$$

If all participants are approximately truthful at each time step, i.e $\mathbb{E}\left[\left\|a_t^i g_i(\theta_t) - g_i(\theta_t)\right\|^2\right] \leq \varepsilon^2$ for all clients i and all time steps t, then:

$$\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{n} a_t^i g_i(\theta_t)\right] \leq \frac{2\left(\varepsilon^2 + M + 2M_V\zeta^2\right)}{N} + \frac{2M_V}{N}\|\nabla F(\theta_t)\|^2.$$

Proof

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} a_t^i g_i(\theta_t)\right] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} a_t^i g_i(\theta_t)\right\|^2\right] - \left\|\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} a_t^i g_i(\theta_t)\right]\right\|^2 \\
&= \frac{1}{N^2}\mathbb{E}\left[\left\|\sum_{i=1}^{N} a_t^i e_i(\theta_t) + \sum_{i=1}^{N} a_t^i \nabla F_i(\theta_t)\right\|^2\right] - \frac{1}{N^2}\left\|\sum_{i=1}^{N} a_t^i \nabla F_i(\theta_t)\right\|^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|a_t^i e_i(\theta_t)\right\|^2\right] + \frac{1}{N^2}\left\|\sum_{i=1}^{N} a_t^i \nabla F_i(\theta_t)\right\|^2 - \frac{1}{N^2}\left\|\sum_{i=1}^{N} a_t^i \nabla F_i(\theta_t)\right\|^2
\end{aligned}$$

6. In the original results, there is two additional variables μ and μ_G. Here we can set them to $\mu_G = \mu = 1$, so we choose to simplify the write-up and ignore them.

$$\leq \frac{2\varepsilon^2}{N} + \frac{2}{N^2}\sum_{i=1}^{N} \mathbb{E}\left[\|e_i(\theta_t)\|^2\right]$$
$$\leq \frac{2\left(\varepsilon^2 + M\right)}{N} + \frac{2M_V}{N^2}\sum_{i=1}^{N} \|\nabla F_i(\theta_t)\|^2$$
$$\leq \frac{2\left(\varepsilon^2 + M + M_V\zeta^2\right)}{N} + \frac{2M_V}{N}\|\nabla F(\theta_t)\|^2$$

■

Proof [Proof of Theorem 9] First, we condition on the event that there is no $t \leq T$, such that $\Pi_\Theta\left(\theta_t - \gamma_t \bar{m}_t\right) \neq \theta_t - \gamma_t \bar{m}_t$. Once we have Claim 27 under out belt, we invoke Lemma 25 with $M = \frac{2(\varepsilon^2+M+M_V\zeta^2)}{N}$, $M_V = \frac{2M_V}{N}$, and $\gamma_t = \frac{\gamma}{\eta+t} = \frac{4}{m(\eta+t)}$, where $\eta = \frac{4H(2M_V+N)}{mN}$. This yields:

$$\mathbb{E}\left[F(\theta_{t+1}) - F(\theta^*)\right] \leq \left(1 - \frac{4}{\eta+t}\right)\mathbb{E}\left[F(\theta_t) - F(\theta^*)\right] + \frac{16H(\varepsilon^2 + M + M_V\zeta^2)}{Nm^2(\eta+t)^2}.$$

Finally, we use Lemma 26 with $p = 1$, $c = 4$ and $c_1 = \frac{16H(\varepsilon^2+M+M_V\zeta^2)}{Nm^2}$, to get:

$$\mathbb{E}\left[F(\theta_t) - F(\theta^*)\right] \leq \frac{16H(\varepsilon^2 + M + M_V\zeta^2)}{3Nm^2(\eta+t)} + O\left(\frac{1}{t^2} + \frac{1}{t^4}\right).$$

Now, if there is a t, such that $\Pi_\Theta\left(\theta_t - \gamma_t \bar{m}_t\right) \neq \theta_t - \gamma_t \bar{m}_t$, we can still bound $\mathbb{E}\left[F(\theta_t) - F(\theta^*)\right]$ by a constant because Θ is bounded and F is L-Lipschitz. Recall that by assumption

$$\mathbb{P}\left[\exists t \leq T : \Pi_\Theta\left(\theta_t - \gamma_t \bar{m}_t\right) \neq \theta_t - \gamma_t \bar{m}_t\right] = O\left(\frac{1}{NT}\right).$$

Therefore, combining everything we get the bound:

$$\mathbb{E}\left[F(\theta_T) - F(\theta^*)\right] \leq \frac{16H(\varepsilon^2 + M + M_V\zeta^2)}{3Nm^2(\eta+t)} + O\left(\frac{1}{NT}\right) + O\left(\frac{1}{T^2}\right).$$

■

Appendix H. Miscellaneous claims

Claim 28 *Let F and g satisfy the same conditions as F_i and g_i from Section 3.1. Let $\theta^* \in \Theta$ be a minimizer of F. For any $\theta \in \Theta$ we have*

$$\mathbb{E}\left[\|g_i(\theta)\|^2\right] \leq 2H^2R^2 + 2\|\nabla F_i(\theta^*)\|^2 + \sigma^2.$$

Proof Let $e(\theta) = g(\theta) - \nabla F(\theta)$ be the gradient noise. Observe:

$$\mathbb{E}\left[\|g_i(\theta)\|^2\right] = \mathbb{E}\left[\|\nabla F_i(\theta) + e_i(\theta)\|^2\right]$$
$$= \mathbb{E}\left[\|\nabla F_i(\theta)\|^2\right] + 2\mathbb{E}\left[\langle \nabla F_i(\theta), e_i(\theta)\rangle\right] + \mathbb{E}\left[\|e_i(\theta)\|^2\right]$$

$$
\begin{aligned}
&= \|\nabla F_i(\theta)\|^2 + \mathbb{E}\left[\|e_i(\theta)\|^2\right] \\
&\leq 2H^2 \|\theta - \theta^*\|^2 + 2\|\nabla F_i(\theta^*)\|^2 + \sigma^2 && \because\ F_i \text{ is } H\text{-smooth} \\
&\leq 2H^2R^2 + 2\|\nabla F_i(\theta^*)\|^2 + \sigma^2 && \because\ R = \sup_{\theta,\theta' \in \Theta} \left\|\theta - \theta'\right\|
\end{aligned}
$$

The third line uses the fact that $\mathbb{E}\left[e_i(\theta)\right] = 0$ for any fixed $\theta \in \Theta$. The forth line observes that the variance of $e_i(\theta)$ is $\frac{1}{n}$ of the variance of $f(\theta; z)$, and uses smoothness. ■

Corollary 29 *For any client i and any $\theta \in \Theta$ we have*

$$
\mathbb{E}\left[\|g_i(\theta)\|\right] \leq \sqrt{\mathbb{E}\left[\|g_i(\theta)\|^2\right]} \leq \sqrt{H^2R^2 + \sigma^2} \leq HR + \sigma.
$$

Proof Notice that $\sqrt{\cdot}$ is concave and $\|g_i(\theta)\|^2$ is a non-negative random variable, so Jensen's inequality gives the desired. ■