AN EQUIVALENCE BETWEEN DATA POISONING AND BYZANTINE GRADIENT ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

To address the resilience of distributed learning, the "Byzantine" literature considers a strong threat model where workers can report arbitrary gradients to the parameter server. While this model helped generate several fundamental results, it has however sometimes been considered unrealistic, when the workers are mostly trustworthy machines. In this paper, we show a surprising equivalence between this model and data poisoning, a threat considered much more realistic. More specifically, we prove that any gradient attack can be reduced to data poisoning in a personalized federated learning system that provides PAC guarantees (which we show are both desirable and realistic in various personalized federated learning contexts such as linear regression and classification). Maybe most importantly, we derive a simple and practical attack that may be constructed against classical personalized federated learning models, and we show both theoretically and empirically the effectiveness of this attack.

1 INTRODUCTION

Learning algorithms nowadays typically leverage data generated by a large number of users (Smith et al., 2013; Wang et al., 2019a;b) to learn a common model that fits a large population (Konecný et al., 2015), or to construct a *personalized* model for each individual (Ricci et al., 2011). Autocompletion (Lehmann & Buschek, 2021), conversational (Shum et al., 2018) and recommendation (Ie et al., 2019) algorithms are examples of such algorithms deployed at scale. To be effective, besides huge amounts of (distributed) data (Brown et al., 2020; Fedus et al., 2021), these algorithms require a high level of customization. This has motivated a lot of research on *personalized federated learning* (Fallah et al., 2020; Hanzely et al., 2020; Dinh et al., 2020).

However, in applications such as content recommendation, activists, companies, and politicians have strong incentives to promote certain views, products or ideologies (Hoang, 2020; Hoang et al., 2021). Remember for instance that, on YouTube, two views out of three result from algorithmic recommendations (Solsman, 2018). Perhaps unsurprisingly, this has led to vast amounts of fabricated activities to bias algorithms (Bradshaw & Howard, 2019; Neudert et al., 2019), like "fake reviews" (Wu et al., 2020). The scale of this phenomenon is well illustrated by the case of Facebook which, in 2019 alone, reported the removal of around 6 billion fake accounts from its platform (Fung & Garcia, 2019). This is particularly concerning in the era of "stochastic parrots" (Bender et al., 2021): climate denialists are incentivized to pollute textual datasets with claims like "climate change is a hoax", as autocompletion, conversational and recommendation algorithms trained on such data will more likely spread these views (McGuffie & Newhouse, 2020). This raises serious concerns about the vulnerability of personalized federated learning to such misleading data. Data poisoning attacks clearly constitute now a major machine learning security issue (Kumar et al., 2020).

Overall, in highly adversarial environments like social media, given the advent of *deep fakes* (Johnson & Diakopoulos, 2021), we should expect that *most data are strategically crafted and labeled*. In this context, the authentication of the data provider seems critical. In particular, the safety of learning algorithms arguably demands that they be trained solely on *cryptographically signed* data, that is, data that provably come from a known source. But even signed data cannot be wholeheartedly trusted since users usually have preferences over what ought to be recommended to others. They thus have incentives to behave strategically in order to promote certain views or products.

To address data poisoning, the Byzantine learning literature usually considers that each federated learning worker may behave arbitrarily (Blanchard et al., 2017; Mhamdi et al., 2018; El-Mhamdi et al., 2021). Recall that at each iteration of the federated learning stochastic gradient descent, each worker is given the updated model, and is asked to compute the gradient of the loss function with respect to (a batch of) its local data. Byzantine learning usually assumes that a malicious (Byzantine) worker may report *any* gradient; without having to justify whether such a gradient could have been generated through data poisoning. In fact, the gradient attack threat model has sometimes been claimed to be unrealistic in practical federated learning (Shejwalkar et al., 2021), especially when the workers are machines owned by trusted entities (cross-silo FL (Kairouz et al., 2021)).

We prove in this paper an equivalence between gradients attacks and fabricated data injection, in a general and desirable collaborative learning framework. Thereby, our paper provides the first practically compelling argument for the necessity to protect federated learning against gradient attacks.

Contributions. As a preamble of our main result, we formalize PAC* learning[1] (Valiant, 1984) for personalized learning, and prove that a simple and general solution to personalized federated linear regression and classification is indeed PAC* learning. Our proof leverages a new concept called *gradient-PAC* learning*, which is of independent interest. We prove that it is sufficient to guarantee personalized federated PAC* learning, and that it is verified by basic learning algorithms, like linear and logistic regression. This is an important and highly nontrivial contribution of this paper.

Our main contribution is to then prove that PAC* learning in personalized federated learning essentially implies an equivalence between data poisoning and gradient attacks. More precisely, we show how any (converging) gradient attack can be turned into a data poisoning attack, with the same resulting harm. Given how easy it generally is to create fake accounts on web platforms and to inject data poisoning by generating fake activities, this result should arguably greatly increase the concerns over the vulnerabilities of federated learning with user-generated data.

Finally, we propose a simple but very general *strategic* gradient attack, called the *counter-gradient attack* (CGA), which any participant to federated learning can deploy to bias the global model towards any target model that better suits their interest. We prove the effectiveness of this attack under fairly general assumptions, which apply to many proposed personalized learning frameworks including (Hanzely et al., 2020; Dinh et al., 2020). We then show empirically how this gradient attack can be turned into a devastating data poisoning attack, with remarkably few data.

Related work. Byzantine learning has provided both negative and positive results in Byzantine resilience (Blanchard et al., 2017; Mhamdi et al., 2018; Baruch et al., 2019), some of which apply almost straightforwardly to personalized federated learning (El-Mhamdi et al., 2020; 2021). Such results study the resilience against a minority of *adversarial* users. Note that our paper focuses on a slightly different kind of malicious users. Namely, like Suya et al. (2021), we study the resilience against *strategic* users, who aim to bias the learned models towards a specific target model.

The study of the resilience against strategic users is part of the research on *strategyproof* learning. Many special cases of strategyproofness have been tackled, including regression (Chen et al., 2018; Dekel et al., 2010; Perote & Perote-Peña, 2004; Ben-Porat & Tennenholtz, 2017), classification (Meir et al., 2012; Chen et al., 2020; Meir et al., 2011; Hardt et al., 2016), statistical estimation (Cai et al., 2015), and clustering (Perote & Sevilla, 2003). However, none of these papers tackles a general *personalized federated learning* scheme. Typically, for linear regression, Chen et al. (2018) and Perote & Perote-Peña (2004) assume that each user can only provide a single data point. This greatly restricts the users' ability to contribute to the learning model. And while Dekel et al. (2010) allows multiple contributions, they either require payments, which might not be possible (e.g., due to ethical reasons), or they restrict the model to one dimension or a constant function in $\mathbb{R}^d$. Conversely, Suya et al. (2021) show how to arbitrarily manipulate convex learning models through multiple data injections, in the case where a single model is learned from all data at once.

A large literature has focused on *data poisoning*, with either a focus on *backdoor* (Dai et al., 2019; Zhao et al., 2020; Severi et al., 2021; Truong et al., 2020; Schwarzschild et al., 2021) or *triggerless* attacks (Biggio et al., 2012; Muñoz-González et al., 2017; Shafahi et al., 2018; Zhu et al., 2019;

[1] We omit complexity considerations for the sake of generality. We define PAC* to be PAC without such considerations.

Huang et al., 2020; Barreno et al., 2006; Aghakhani et al., 2021; Geiping et al., 2021). However, most of this research analyzed data poisoning without *signed* data. One noteworthy exception is Mahloujifar et al. (2019), whose universal attack amplifies the probability of a (bad) property.

Structure of the paper. The rest of the paper is organized as follows. Section 2 presents a general model of personalized learning. Section 3 proves the PAC* learning properties for federated linear regression and classification. Section 4 then proves the equivalence between data poisoning and gradient attacks, for PAC* learning federated learning. Section 5 describes a simple and general gradient attack, which is shown to be converging and to arbitrarily manipulate the ℓ_2^2 regularization. Section 6 shows how to turn it into a data poisoning attack. Section 7 concludes.

2 A GENERAL PERSONALIZED FEDERATED LEARNING FRAMEWORK

We consider a set $[N] = \{1, \ldots, N\}$ of users. Each user $n \in [N]$ has a local *signed* dataset $\mathcal{D}_n$, and aims to learn a local model $\theta_n \in \mathbb{R}^d$. Users may collaborate to do so. Personalized learning must then input a tuple of users' local datasets $\vec{\mathcal{D}} \triangleq (\mathcal{D}_1, \ldots, \mathcal{D}_N)$, and output a tuple of local models $\vec{\theta}^* \triangleq (\theta_1^*, \ldots, \theta_N^*)$. In this paper, like many others, we assume that the users perform federated learning to do so, by leveraging the computation of a common global model $\rho \in \mathbb{R}^d$. Intuitively, the global model is an aggregate of all users' local models, which local users can leverage to improve their model. The common global model will typically allow users with too few data to obtain an effective local model, while it may be mostly discarded by users whose local datasets are large.

More formally, we consider a very general personalized learning framework which generalizes the models proposed by Dinh et al. (2020) and Hanzely et al. (2020). Namely, we consider that the personalized learning algorithm outputs a global minimum $(\rho^*, \vec{\theta}^*)$ of a global loss given by

$$\textsc{Loss}(\rho, \vec{\theta}, \vec{\mathcal{D}}) \triangleq \sum_{n \in [N]} \mathcal{L}_n(\theta_n, \mathcal{D}_n) + \sum_{n \in [N]} \mathcal{R}(\rho, \theta_n), \tag{1}$$

where $\mathcal{R}$ is a regularization, typically with a minimum at $\theta_n = \rho$. For instance, Hanzely et al. (2020) and Dinh et al. (2020) define $\mathcal{R}(\rho, \theta_n) \triangleq \lambda \left\|\rho - \theta_n\right\|_2^2$, which we shall call the ℓ_2^2 regularization. But other regularizations may be considered, like the ℓ_2 regularization $\mathcal{R}(\rho, \theta_n) \triangleq \lambda \left\|\rho - \theta_n\right\|_2$, or the smooth-$\ell_2$ regularization $\mathcal{R}(\rho, \theta_n) \triangleq \lambda \sqrt{1 + \left\|\rho - \theta_n\right\|_2^2}$. Note that, for all such regularizations, the limit $\lambda \to \infty$, essentially yields the classical non-personalized federated learning framework.

In this paper, we will focus on personalized learning algorithms that successfully recover a user's true model, if the user provides a large enough dataset, as defined below.

Definition 1. *A personalized learning algorithm is PAC*-learning if, for any subset $H \subset [N]$, any true models $\vec{\theta}^\dagger$, any $\varepsilon, \delta > 0$, and any datasets $\vec{\mathcal{D}}_{-H}$ from other users $n \notin H$, there exists $\mathcal{I}$ such that, if all users $h \in H$ provide honest datasets $\mathcal{D}_h$ with at least $|\mathcal{D}_h| \geq \mathcal{I}$ data points, then, with probability at least $1 - \delta$, we have $\left\|\theta_h^*\left(\vec{\mathcal{D}}\right) - \theta_h^\dagger\right\|_2 \leq \varepsilon$ for all users $h \in H$.*

PAC* learning is arguably a very desirable property. Indeed, it guarantees that any honest active user will not be discouraged to participate in federated learning. In the following section, we show how PAC* learning can be achieved in practice, by considering specific local loss functions $\mathcal{L}_n$.

3 EXAMPLES OF PAC* LEARNING SYSTEMS

To the best of our knowledge, PAC* learnability is a new concept in the context of personalized federated learning. It is thus important to show that it is not meaningless. To achieve this, in this section, we provide *sufficient* conditions for a personalized learning model to be PAC* learnable.

3.1 SUMS OVER EXPECTATIONS

First, let us argue that local losses $\mathcal{L}_n$ should be sums of losses per input, i.e.

$$\mathcal{L}_n(\theta_n, \mathcal{D}_n) = \nu \left\|\theta_n\right\|_2^2 + \sum_{x \in \mathcal{D}_n} \ell(\theta_n, x), \tag{2}$$

for some "loss per input" function ℓ and some weight $\nu > 0$. Indeed, intuitively, if one considers an expectation $\mathbb{E}_{x\sim\mathcal{D}_n}\left[\ell(\theta_n, x)\right]$ rather than a sum, as is done by Hanzely et al. (2020), Dinh et al. (2020) and El-Mhamdi et al. (2021), then the weight of an honest active user's local loss will not increase as a node provides more and more data, which will hinder the ability of θ_n to fit the user's local data. Interestingly, in the sequel, we will prove that, at least for linear and logistic regression, using a sum is essentially a sufficient condition of PAC* learnability.

In fact, we can provide three other intuitive, theoretical and empirical arguments for using a sum of losses per input rather than an expectation. First, intuitively, using an expectation yields the same influence to any two nodes, even when one (honest) node provides a much larger dataset $\mathcal{D}_n$ than the other, and should thus intuitively be regarded as "more reliable". Second, if the loss is regarded as a Bayesian negative log-posterior, given a prior $\exp\left(-\sum_{n\in[N]}\nu\left\|\theta_n\right\|_2 - \sum_{n\in[N]}\mathcal{R}(\rho,\theta_n)\right)$ on the local and global models, then the term that fits local data should equal the negative log-likelihood of the data, given the models $(\rho, \vec{\theta})$. Assuming that the distribution of each data point $x \in \mathcal{D}_n$ is independent from all other data points, and depends only on the local model θ_n, this negative log-likelihood yields a sum over data points; not an expectation.

	$\mathbb{E}L$	ΣL	$\mathbb{E}NN$	ΣNN
idle user's model	0.52	0.80	0.55	0.79
active user's model	0.58	0.80	0.56	0.79
global model	0.55	0.80	0.58	0.79

Table 1: Accuracy of trained models, depending on the use of expectation (denoted $\mathbb{E}$) or sum (Σ), and on the use of linear classifier (L) or a 2-layer neural net (NN). Here, all users are honest and an ℓ_2^2 regularization is used, but there is a large heterogeneity in the amount of data per user.

Third and finally, we constructed a setting where 10 "idle" users draw randomly 10 data points from the FashionMNIST dataset, while one "active" user has all of the FashionMNIST dataset. We then learned local and global models, with $\mathcal{R}(\rho,\theta) \triangleq \lambda\left\|\rho-\theta\right\|_2^2$, $\lambda = 1$. We considered both a linear classifier and a two-layer neural network. To see a strong difference between sum and average, we made the FashionMNIST dataset harder to learn, by randomly labeling 60% of the training set. Table 1 reports the accuracy of local and global models in the different settings. Appendix A contains more details, like convergence graphs for different values of λ, as well as plots of the case where no noise is added to the training set. Our results clearly and robustly indicate that the use of sums outperforms the use of expectations.

3.2 GRADIENT-PAC*

In order to prove PAC* learning guarantees, we need to guarantee that each user will robustly learn its true models, assuming it learns them from a large amount of honest data. The following definition formalizes what is meant by "robustly learn".

Definition 2 (Gradient-PAC*). *Denote $\mathcal{E}(\mathcal{D}, \theta^\dagger, \mathcal{I}, A, B, \alpha)$ the event*

$$\forall\theta\in\mathbb{R}^d,\ \left(\theta-\theta^\dagger\right)^T\nabla\mathcal{L}\left(\theta,\mathcal{D}\right)\geq A\mathcal{I}\min\left\{\left\|\theta-\theta^\dagger\right\|_2,\left\|\theta-\theta^\dagger\right\|_2^2\right\}-B\mathcal{I}^\alpha\left\|\theta-\theta^\dagger\right\|_2. \quad (3)$$

The loss ℓ is gradient-PAC if, for any $\mathcal{K} > 0$, there exist constants $A_\mathcal{K}, B_\mathcal{K} > 0$ and $\alpha_\mathcal{K} < 1$, such that for any true model $\theta^\dagger \in \mathbb{R}^d$ with $\left\|\theta^\dagger\right\|_2 \leq \mathcal{K}$, assuming that the dataset $\mathcal{D}$ is obtained by honestly collecting and labeling $\mathcal{I}$ data points according to the true model $\theta^\dagger$, the probability of the event $\mathcal{E}(\mathcal{D}, \theta^\dagger, \mathcal{I}, A_\mathcal{K}, B_\mathcal{K}, \alpha_\mathcal{K})$ goes to 1 as $\mathcal{I} \to \infty$.*

Intuitively, this definition asserts that, as we collect more data from an honest user, then, with high probability, the gradient of the loss at any point θ too far from $\theta^\dagger$ will point away from $\theta^\dagger$. In particular, gradient descent is then essentially guaranteed to draw θ closer to $\theta^\dagger$. The right-hand side of equation 3 is subtly chosen to be strong enough to yield PAC* learning guarantees, and weak enough to be verified by linear and logistic regression, as proved by the following sections.

3.3 LEAST SQUARE ERROR IS GRADIENT-PAC*

Assume that a data point x consists of a vector *query* (or *feature*) $\mathcal{Q}_i \in \mathbb{R}^d$, and of a real-valued *answer* (or *label*) given by $\mathcal{A}_i = \mathcal{Q}_i^T \theta^\dagger + \xi_i$, where ξ_i is a zero-mean sub-Gaussian random noise with parameter σ_ξ, independent from $\mathcal{Q}_i$.

Theorem 1. *If queries $\mathcal{Q}_i$ are drawn i.i.d from a sub-Gaussian distribution with parameter $\sigma_\mathcal{Q}$ and positive definite matrix $\Sigma = \mathbb{E}\left[\mathcal{Q}_i \mathcal{Q}_i^T\right]$, then $\ell(\theta, \mathcal{Q}_i, \mathcal{A}_i) = \frac{1}{2}(\theta^T \mathcal{Q}_i - \mathcal{A}_i)^2$ is gradient-PAC*.*

Sketch of proof. Remarkably, the discrepancy between the empirical and the expected loss functions depends only on a few key random variables, such as $\min \text{SP}\left(\frac{1}{\mathcal{I}} \sum \mathcal{Q}_i \mathcal{Q}_i^T\right)$ and $\sum \xi_i \mathcal{Q}_i$, which can be controlled by appropriate concentration bounds. The full proof is given in Appendix B.2. □

3.4 LOGISTIC REGRESSION IS GRADIENT-PAC*

Again, consider $\mathcal{Q}_i \in \mathbb{R}^d$. Now, denote $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid logistic function. Then with probability $\sigma(\mathcal{Q}_i^T \theta^\dagger)$, a truthful answer with model $\theta^\dagger$ yields $\mathcal{A}_i = 1$. Otherwise, $\mathcal{A}_i = -1$.

Theorem 2. *If the queries $\mathcal{Q}_i$ are drawn i.i.d. from $\tilde{\mathcal{Q}}$ whose support $\text{SUPP}(\tilde{\mathcal{Q}})$ is bounded and spans the full vector space $\mathbb{R}^d$, then $\ell(\theta, \mathcal{Q}_i, \mathcal{A}_i) = -\ln(\sigma(\mathcal{A}_i \theta^T \mathcal{Q}_i))$ is gradient-PAC*.*

Sketch of proof. Crucially, for $|b| \leq \mathcal{K}$, $(a-b)(\sigma(a) - \sigma(b)) \geq c_\mathcal{K} \min(|a-b|, |a-b|^2)$. Essentially, this proves that gradient-PAC* would hold if the empirical loss was replaced by the expected loss (times $\mathcal{I}$). The actual theorem, however, is nontrivial. It leverages topological considerations to derive a critical uniform concentration bound. The full proof is given in Appendix B.3. □

3.5 A SUFFICIENT CONDITION FOR PAC* LEARNING

Now, under very mild assumptions on $\mathcal{R}$ (not even convexity!), we prove that the gradient-PAC* learnability through ℓ suffices to guarantee that personalized learning will be PAC* learning.

Theorem 3. *Suppose that the regularization function $\mathcal{R}$ is (sub)differentiable and nonnegative everywhere, and that $\mathcal{R}(\rho, \theta) \to \infty$ as $\|\rho - \theta\|_2 \to \infty$. If ℓ is gradient-PAC* and nonnegative, then the personalized learning algorithm is PAC*-learning.*

Sketch of proof. Once other users' datasets are fixed, the learning of an honest user's model is biased. But as the user provides more data, by virtue of gradient-PAC*, the local loss becomes predominant, which guarantees weak PAC*-learning. Appendix C provides a full proof. □

4 THE EQUIVALENCE BETWEEN DATA POISONING AND GRADIENT ATTACKS

In this section, we present our main result, which is the equivalence between data poisoning and gradient attacks for strategic users whose goal is to bias the global model toward a specific desired model. This threat model is closely related to "model-targeted attacks", studied in Suya et al. (2021). Our proof goes through an intermediary step, which we call a *model attack*.

4.1 EQUIVALENCE BETWEEN DATA POISONING AND MODEL ATTACKS

In this section, we prove the equivalence between data poisoning and what we call *model attacks*. But first, let us define precisely what we mean by data poisoning and by model attacks. A data poisoning consists of a strategic (mis)reporting of a dataset $\mathcal{D}_s$ by a strategic user $s \in [N]$, which leads to learning $\rho^*(\vec{\mathcal{D}})$ and $\vec{\theta}^*(\vec{\mathcal{D}})$. By opposition, we say that the strategic user s performs a model attack if they instead report a strategically crafted model $\theta_s^\spadesuit$, and refuse to be given any model $\theta_s^* \neq \theta_s^\spadesuit$. More precisely, we define the following modified loss function:

$$\text{LOSS}_s(\rho, \vec{\theta}_{-s}, \theta_s^\spadesuit, \vec{\mathcal{D}}_{-s}) \triangleq \text{LOSS}(\rho, (\theta_s^\spadesuit, \vec{\theta}_{-s}), (\emptyset, \vec{\mathcal{D}}_{-s})) \tag{4}$$

where $\vec{\theta}_{-s}$ and $\vec{\mathcal{D}}_{-s}$ are variables and datasets for users $n \neq s$. We then define $\rho^*(\theta_s^\spadesuit, \vec{\mathcal{D}}_{-s})$ and $\vec{\theta}_{-s}^*(\theta_s^\spadesuit, \vec{\mathcal{D}}_{-s})$ as a minimum of the modified loss function, and $\theta_s^*(\theta_s^\spadesuit, \vec{\mathcal{D}}_{-s}) \triangleq \theta_s^\spadesuit$. The following theorem shows that any data poisoning can be obtained through a model attack.

Theorem 4 (Reduction from model attack to data poisoning)**.** *Consider any data $\vec{\mathcal{D}}$ and user $s \in [N]$. Assume the global loss with datasets $\vec{\mathcal{D}}$ has a global minimum $(\rho^*, \vec{\theta}^*)$. Then $(\rho^*, \vec{\theta}^*_{-s})$ is also a global minimum of the modified loss with datasets $\vec{\mathcal{D}}_{-s}$ and strategic reporting $\theta^{\spadesuit}_s = \theta^*_s(\vec{\mathcal{D}})$.*

Proof. The proof is almost straightforward. It is given in Appendix E.1. □

Now, intuitively, by answering a large number of random queries with model $\theta^{\spadesuit}_s \in \mathbb{R}^d$, by virtue of PAC* learning, strategic user s can essentially guarantee that the personalized learning framework will be learning $\theta^*_s \approx \theta^{\spadesuit}_s$. But, a priori, it may seem unclear if this will imply a biasing of the global model and of other local models essentially identitical to the one obtained through the model attack that imposes $\theta^*_s = \theta^{\spadesuit}_s$. In the sequel, we show that the answer is yes.

Theorem 5 (Reduction from data poisoning to model attack)**.** *Assume PAC* learning. Suppose also that $\mathcal{R}$ is continuous and that $\mathcal{R}(\rho, \theta) \to \infty$ when $\|\rho - \theta\|_2 \to \infty$. Consider any datasets $\mathcal{D}_{-s}$ and any attack model $\theta^{\spadesuit}_s$ such that the modified loss LOSS_s has a unique minimum $\rho^*(\theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s}), \vec{\theta}^*_{-s}(\theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s})$. Then, for any $\varepsilon, \delta > 0$, there exists $\mathcal{I}$ such that if user s's dataset $\mathcal{D}_s$ contains at least $\mathcal{I}$ inputs drawn from model $\theta^{\spadesuit}_s$, then, with probability at least $1 - \delta$, we have*

$$\left\|\rho^*(\vec{\mathcal{D}}) - \rho^*(\theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s})\right\|_2 \leq \varepsilon \ \text{ and } \ \forall n, \ \left\|\theta^*_n(\vec{\mathcal{D}}) - \theta^*_n(\theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s})\right\|_2 \leq \varepsilon. \tag{5}$$

Sketch of proof. Given PAC* learning, given a large enough honest dataset $\mathcal{D}_s$, strategic user s can guarantee $\theta^*_s(\vec{\mathcal{D}}) \approx \theta^{\spadesuit}_s$. By carefully bounding the effect of the approximation on the loss using the Heine-Cantor theorem, we show that this implies $\rho^*(\vec{\mathcal{D}}) \approx \rho^*(\theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s})$ and $\theta^*_n(\vec{\mathcal{D}}) \approx \theta^*_n(\theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s})$ for all $n \neq s$ too. The precise analysis is nontrivial. □

4.2 Reduction from model attacks to gradient attacks

Suppose now that learning is based on (stochastic) gradient descent, where the variables ρ and $\vec{\theta}$ are alternatively optimized. At each iteration, the global model ρ^t is thus updated based on $\nabla_\rho \text{LOSS} = \sum_{n \in [N]} g^{\dagger,t}_n$, which is a sum of user-dependent gradients $g^{\dagger,t}_n \triangleq \nabla_\rho \mathcal{R}(\rho, \theta^t_n)$. We denote $\text{GRAD}(\rho) \triangleq \overline{\{\nabla_\rho \mathcal{R}(\rho, \theta) \mid \theta \in \mathbb{R}^d\}}$ the closure set of plausible (sub)gradients at ρ. But now, let us assume that instead of reporting $g^{\dagger,t}_s \triangleq \nabla_\rho \mathcal{R}(\rho, \theta^t_n)$, the strategic user s can report any gradient g^t_s within the closure set $\text{GRAD}(\rho)$ of gradients that they could have legitimately obtained. In the case of an ℓ^2_2 regularization, where $\mathcal{R}(\rho, \theta) = \|\rho - \theta\|^2_2$, we clearly have $\text{GRAD}(\rho) = \mathbb{R}^d$ for all $\rho \in \mathbb{R}^d$. It can be easily shown that, for smooth-ℓ_2, $\text{GRAD}(\rho)$ is the closed ball $\mathcal{B}(0, \lambda)$.

Now, to encompass both ℓ^2_2 and smooth-ℓ_2 regularization, a well as linear regression and logistic regression, we make the following assumptions on the components of the global loss.

Assumption 1. *Assume that ℓ is convex and L_ℓ-smooth, and that $\mathcal{R}(\rho, \theta) = \mathcal{R}_0(\rho - \theta)$, where $\mathcal{R}_0 : \mathbb{R}^d \to \mathbb{R}$ is locally strongly convex (i.e. strongly convex on any convex compact set), $L_{\mathcal{R}_0}$-smooth and satisfy $\mathcal{R}_0(z) = \Omega(\|z\|_2)$ as $\|z\|_2 \to \infty$.*

For the last few years, gradient attacks have been widely studied by the Byzantine learning literature. Recently, Shejwalkar et al. (2021) argued that they are not a realistic threat model. Below, we prove that, in our setting, they are actually as concerning as model attacks (and, thus, as data poisoning). Indeed, instead of reporting g^t_s, a strategic user can leverage their gradient attack to learn a model $\theta^{\spadesuit}_s$, which will have the same effect as the gradient attack, as guaranteed by the following theorem.

Theorem 6 (Reduction from model attack to gradient attack)**.** *Under Assumption 1, if g^t_s converges and if $\eta_t = \eta$ is a constant small enough, then ρ^t will converge too. Denote ρ^∞ its limit. Then for any $\varepsilon > 0$, there is $\theta^{\spadesuit}_s \in \mathbb{R}^d$ such that $\left\|\rho^\infty - \rho^*(\theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s})\right\|_2 \leq \varepsilon$.*

Sketch of proof. Denote g^∞_s the limit of g^t_s. Gradient descent then behaves as though it was minimizing the loss plus $\rho^T g^\infty_s$ (and ignoring $\mathcal{R}(\rho, \theta_s)$). Essentially, classical gradient descent theory then guarantees the convergence of ρ^t to ρ^∞, though the precise proof is nontrivial. Then, since GRAD is closed and $g^\infty_s \in \text{GRAD}$, we can construct $\theta^{\spadesuit}_s$ which approximately yields the gradient g^∞_s. The full proof (which also yields the necessary upper bound on η) is given in Appendix E. □

Since it is clear that any model attack can be achieved by the corresponding honest gradient attack, we conclude that model attacks and gradient attacks are essentially equivalent. In light of our previous results, this implies that gradient attacks are essentially equivalent to data poisoning.

5 THE COUNTER-GRADIENT ATTACK

We now define a simple, general and practical gradient attack, which we call the counter-gradient attack (CGA). Intuitively, at iteration t, this attack estimates the sum $g_{-s}^{\dagger,t}$ of the gradients of other users based on its at the previous iteration, which can be inferred from the way the global model ρ^{t-1} was updated into ρ^t. More precisely, apart from initialization $\hat{g}_{-s}^1 \triangleq 0$, CGA makes the estimation

$$\hat{g}_{-s}^t \triangleq \frac{\rho^{t-1} - \rho^t}{\eta_{t-1}} - g_s^{t-1} = g_{-s}^{\dagger,t-1}. \tag{6}$$

Strategic user s then reports the plausible gradient that moves the global model closest to the user's target model $\theta_s^\dagger$, assuming others report $\hat{g}_{-s}^t$. In other words, at every iteration, CGA reports

$$g_s^t \in \underset{g \in \textsc{Grad}(\rho)}{\arg\min} \left\| \rho^t - \eta_t(\hat{g}_{-s}^t + g) - \theta_s^\dagger \right\|_2. \tag{7}$$

Note that, to perform this attack, user s only needs to know the previous and current learning rates η_{t-1} and η_t, the previous and current global models ρ^{t-1} and ρ^t, and their target model $\theta_s^\dagger$.

5.1 COMPUTATION OF CGA

Note that if $\textsc{Grad}(\rho) = \mathbb{R}^d$, then CGA can be easily computed.

Proposition 1. *Assume* $\textsc{Grad}(\rho) = \mathbb{R}^d$. *Then* CGA *reports* $g_s^t = g_s^{t-1} + \frac{\rho^t - \theta_s^\dagger}{\eta_t} - \frac{\rho^{t-1} - \rho^t}{\eta_{t-1}}$.

Proof. This follows from the fact that the minimum value of the problem in equation 7 is zero. □

Moreover, if $\textsc{Grad}(\rho)$ is convex, then the counter-gradient attack boils down to the computation of an orthogonal projection. In particular, it is easy to verify that for the smooth-ℓ_2 regularization, the counter-gradient attack is again very simple to compute.

Proposition 2. *Assume* $\mathcal{R}(\rho, \theta) = \lambda\sqrt{1 + \|\rho - \theta\|_2^2}$. *Denote* $h_s^t \triangleq g_s^{t-1} + \frac{\rho^t - \theta_s^\dagger}{\eta_t} - \frac{\rho^{t-1} - \rho^t}{\eta_{t-1}}$. *If* $\|h_s^t\|_2 \leq \lambda$, *then* CGA *reports* $g_s^t = h_s^t$. *Otherwise, it reports* $g_s^t = \lambda h_s^t / \|h_s^t\|_2$.

Proof. Equation (7) boils down to minimizing the distance between $\frac{\rho^t - \theta_s^\dagger}{\eta_t} - \hat{g}_{-s}^t$ and $\textsc{Grad}(\rho)$, which is the (convex) ball $\mathcal{B}(0, \lambda)$. This minimum is the orthogonal projection on $\mathcal{B}(0, \lambda)$. □

5.2 THEORETICAL ANALYSIS

We now prove that CGA is perfectly successful against ℓ_2^2 regularization. To do so, we suppose that, at each iteration t and for each user $n \neq s$, the local models θ_n are fully optimized with respect to ρ^t, and the true gradients of $g_n^{\dagger,t}$ are used for the gradient descent of ρ.

Theorem 7. *Consider* ℓ_2^2 *regularization. Assume that* ℓ *is convex and* L_ℓ*-smooth, and that* $\eta_t = \eta$ *is small enough. Then* CGA *is converging and optimal, as* $\rho^t \to \theta_s^\dagger$.

Sketch of proof. The main challenge is to guarantee that the other users' gradients $g_n^{\dagger,t}$ for $n \neq s$ remain sufficiently stable over time to guarantee convergence, which can be done by leveraging L-smoothness. The full proof, with the necessary upper-bound on η, is given in Appendix F. □

The analysis of the convergence against smooth-ℓ_2 is unfortunately significantly more challenging. Here, we simply make a remark about what CGA yields in this case, if it converges.

Proposition 3. *If* CGA *against smooth-ℓ_2 regularization converges for $\eta_t = \eta$, then it either achieves perfect manipulation, or it is eventually partially honest, in the sense that the gradient by* CGA *correctly points towards $\theta_s^\dagger$.*

Proof. Denote P the projection onto the closed ball $\mathcal{B}(0, \lambda)$. If CGA converges, then, by Proposition 2, $P\left(g_s^\infty + \frac{\rho^\infty - \theta_s^\dagger}{\eta}\right) = g_s^\infty$. This implies in particular that $\rho^\infty - \theta_s^\dagger$ and g_s^∞ must be colinear. If perfect manipulation is not achieved (i.e. $\rho^\infty \neq \theta_s^\dagger$), then we must have $g_s^\infty = \lambda \frac{\rho^\infty - \theta_s^\dagger}{\left\|\rho^\infty - \theta_s^\dagger\right\|_2}$. □

5.3 Empirical evaluation of CGA

We deployed CGA to bias the federated learning of MNIST. We consider a strategic user whose target model is one that labels 0's as 1's, 1's as 2's, and so on, until 9's that are labeled as 0's. In particular, this target model has a nil accuracy. Figure 16 shows that such a user effectively hacks the ℓ_2^2 regularization against 10 honest users who each have 6,000 data points of MNIST, in the case where local models only undergo a single gradient step at each iteration, but fails to hack the ℓ_2 regularization. Further details on the experiment are given in Appendix H.

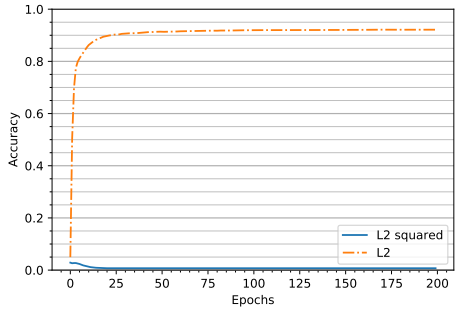


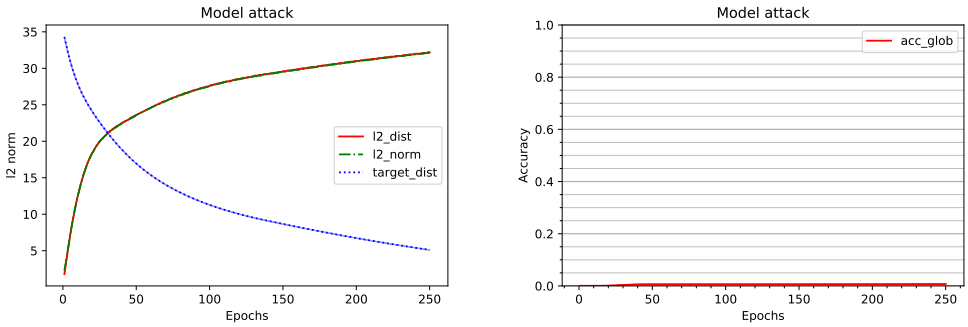
Figure 1: Accuracy of global models under attack by CGA for both ℓ_2 and ℓ_2^2 regularizations.

6 Practical data poisoning against ℓ_2^2

We now show how to turn a gradient attack into data poisoning, against ℓ_2^2 regularization.

6.1 From gradient attack to model attack

Assume that we found a gradient g_s^∞ such that $\rho^\infty = \theta_s^\dagger$. It is trivial to transform it into a model attack by setting $\theta_s^\spadesuit \triangleq \theta_s^\dagger - \frac{1}{2} g_s^\infty$, as guaranteed by the following result, and as depicted by Figure 2.



(a) Distance between ρ^t and $\theta_s^\dagger$ (target_dist).

(b) Accuracy of the global model

Figure 2: Successful model attack against ℓ_2^2 by combining CGA and Proposition 4.

Proposition 4. *Consider the ℓ_2^2 regularization. Suppose that $g_s^t \to g_s^\infty$ and $\rho^t \to \theta_s^\dagger$, with a constant learning rate $\eta_t = \eta$. Then, under the model attack $\theta_s^\spadesuit \triangleq \theta_s^\dagger - \frac{1}{2\lambda} g_s^\infty$, the gradient at $\rho = \theta_s^\dagger$ vanishes, i.e. $\nabla_\rho \text{LOSS}(\theta_s^\dagger, \vec{\theta}^*_{-s}(\theta_s^\dagger, \vec{\mathcal{D}}_{-s}), \theta_s^\spadesuit, \mathcal{D}_{-s}) = 0$.*

Proof. Given that the learning rate is constant, the convergence $\rho^t \to \theta_s^\dagger$ implies that the sum of honest users' gradients at $\rho = \theta_s^\dagger$ must equal $-g_s^\infty$. Therefore, to achieve $\rho^* = \theta_s^\dagger$ with a model attack, it suffices to send a model $\theta_s^\spadesuit$ such that the gradient of $\lambda \left\|\rho - \theta_s^\spadesuit\right\|_2^2$ with respect to ρ at $\rho = \theta_s^\dagger$ equals g_s^∞. Since the gradient is $\lambda(\theta_s^\dagger - \theta_s^\spadesuit)$, $\theta_s^\spadesuit \triangleq \theta_s^\dagger - \frac{1}{2\lambda} g_s^\infty$ does the trick. □

6.2 From model attack to data poisoning for linear regression

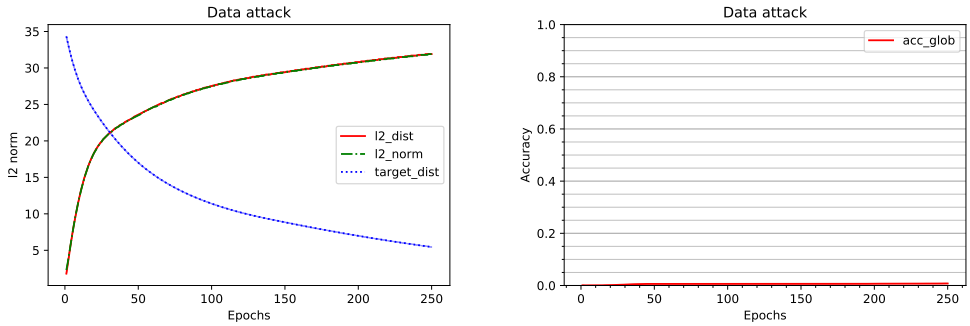
In least square linear regression, any model attack can be turned into a single data poisoning attack, as proved by the following theorem whose proof is given in Appendix G.

Theorem 8. *Consider the ℓ_2^2 regularization and linear regression. For any data $\mathcal{D}_{-s}$ and any target value $\theta_s^\dagger$, there is a datapoint $(\mathcal{Q}, \mathcal{A})$ to be injected by user s such that $\rho^*(\{(\mathcal{Q}, \mathcal{A})\}, \mathcal{D}_{-s}) = \theta_s^\dagger$.*

6.3 From model attack to data poisoning for linear classification

We now consider linear classification, with the case of MNIST. By Theorem 5, any model attack can be turned into data poisoning, provided sufficiently many (random) data points are labeled by the strategic user. However, this may require creating too many data labelings, especially if the norm of $\theta_s^\spadesuit$ is large (which is the case if s is alone against many active users), as suggested by Theorem 2.

For efficient data poisoning, define the indifference affine subspace $V \subset \mathbb{R}^d$ as the set of images with equiprobable labels. Intuitively, labeling images close to V is very informative, as it informs us directly about the separating hyperplanes. To generate images, we first draw random images, which we then project orthogonally on V. We then add a small noise, before probabilistically labeling the image with model $\theta_s^\spadesuit$. Note that this leads us to consider images not in $[0, 1]^d$. Figure 3 shows the effectiveness of the resulting data poisoning attack, with only 2,000 data points, as opposed to the other nodes' 6,000 data points. More details and explanations are provided in Appendix I.



(a) Distance between ρ^t and $\theta_s^\dagger$ (target_dist).

(b) Accuracy of the global model

Figure 3: Successful model attack against ℓ_2^2 by combining CGA and Proposition 4.

7 Concluding Remarks

We showed in this paper that, unlike what has been argued by, e.g., Shejwalkar et al. (2021), the gradient attack threat is not unrealistic. For personalized federated learning with PAC* guarantees, gradient attacks can be constructed from *strategic* data reporting. More generally, our work stresses how critical Byzantine learning research is. For instance, El-Mhamdi et al. (2021) proved lower bounds on what *any* collaborative learning algorithm can guarantee in heterogeneous environments, under Byzantine gradient attacks. Our work implies that, at least for certain personalized federated learning problems, these lower bounds also hold for data poisoning attacks, which are known to be common for many high-risk applications. Arguably, a lot more security measures are urgently needed to make large-scale learning algorithms safe.

ETHICS STATEMENT

The safety of algorithms is arguably a prerequisite to their ethics. After all, an arbitrarily manipulable large-scale algorithm will unavoidably endanger the targets of the entities that successfully design such algorithms. Typically, unsafe large-scale recommendation algorithms may be hacked by health disinformation campaigns that aim to promote non-certified products, e.g., by falsely pretending that they cure COVID-19. Such algorithms must not be regarded as ethical, even if they were designed with the best intentions. We believe that our work helps understand the vulnerabilities of such algorithms, and will motivate further research in the ethics and security of machine learning.

REPRODUCIBILITY STATEMENT

All our experiments are run on the classical datasets MNIST and FashionMNIST. We provide all of the source codes to reproduce the experiments:

- The sum versus expectation experiments can be run by executing this file: `https://www.dropbox.com/sh/qdgmz9air24nhyr/AAAtycEkxc_1hGbvU5YG18z4a?dl=0`
- The counter-gradient attack experiments can be run by executing this file: `https://www.dropbox.com/sh/bycqkccgmk4muzn/AACRD1yeTglLSHEd1OOAzmVqa?dl=0`
- The data poisoning attack experiments can be run by executing this file: `https://www.dropbox.com/sh/qodnl6ivzti8hch/AADgX4EYuSOotiMCAHyTIiGMa?dl=0`

The experiments are seeded and the CuDNN backend is configured in deterministic mode in order to reduce the sources of non-determinism. We also turn of the benchmark mode. Executing the codes will generate the figures and statistics of our main paper, and most of the figures of our Appendix. Our other figures can be obtained by adjusting the hyperparameters of our codes.

The full description of the architecture and optimisation algorithm used is described in Appendix A. The experimental setup details of each experiment are provided in the Appendix, along with additional results. The Appendix also contains the full proofs of our theorems.

REFERENCES

Hojjat Aghakhani, Dongyu Meng, Yu-Xiang Wang, Christopher Kruegel, and Giovanni Vigna. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability, 2021.

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, ASIACCS '06, pp. 16–25, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932720. doi: 10.1145/1128817.1128824. URL `https://doi.org/10.1145/1128817.1128824`.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/ec1c59141046cd1866bbbcdfb6ae31d4-Paper.pdf`.

Omer Ben-Porat and Moshe Tennenholtz. Best response regression. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/1ce927f875864094e3906a4a0b5ece68-Paper.pdf`.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Madeleine Clare Elish,

William Isaac, and Richard S. Zemel (eds.), *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pp. 610–623. ACM, 2021. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/2012/papers/880.pdf.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 119–129, 2017. URL http://papers.nips.cc/paper/6617-machine-learning-with-adversaries-byzantine-tolerant-gradient-descent.

Samantha Bradshaw and Philip N Howard. *The global disinformation order: 2019 global inventory of organised social media manipulation*. Project on Computational Propaganda, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. Optimum statistical estimation with strategic data sources. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pp. 280–296. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Cai15.html.

Yiling Chen, Chara Podimata, Ariel D. Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, pp. 9–26, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358293. doi: 10.1145/3219166.3219175. URL https://doi.org/10.1145/3219166.3219175.

Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15265–15276. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ae87a54e183c075c494c4d397d126a66-Paper.pdf.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019. doi: 10.1109/ACCESS.2019.2941376. URL https://doi.org/10.1109/ACCESS.2019.2941376.

Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2010.03.003. URL https://www.sciencedirect.com/science/article/pii/S0022000010000309.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f4f1f13c8289ac1b1ee0ff176b56fc60-Abstract.html.

El-Mahdi El-Mhamdi, Rachid Guerraoui, Arsany Guirguis, Lê Nguyên Hoang, and Sébastien Rouault. Genuinely distributed byzantine machine learning. In Yuval Emek and Christian Cachin (eds.), *PODC '20: ACM Symposium on Principles of Distributed Computing, Virtual Event, Italy, August 3-7, 2020*, pp. 355–364. ACM, 2020. doi: 10.1145/3382734.3405695. URL https://doi.org/10.1145/3382734.3405695.

El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê Nguyên Hoang, and Sébastien Rouault. Collaborative learning in the jungle. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, December 6-14, 2021*, 2021.

El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. Distributed momentum for byzantine-resilient stochastic gradient descent. In *9th International Conference on Learning Representations, ICLR 2021, Vienna, Austria, May 4–8, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=H8UHdhWG6A3.

Alireza Fallah, Aryan Mokhtari, and Asuman E. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/24389bfe4fe2eba8bf9aa9203a44cdad-Abstract.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961, 2021. URL https://arxiv.org/abs/2101.03961.

Brian Fung and Ahiza Garcia. Facebook has shut down 5.4 billion fake accounts this year. *CNN Business*, 2019.

Jonas Geiping, Liam H Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=01olnfLIbD.

Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/187acf7982f3c169b3075132380986e4-Abstract.html.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pp. 111–122, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

Lê Nguyên Hoang. Science communication desperately needs more aligned recommendation algorithms. *Frontiers in Communication*, 5:115, 2020.

Lê Nguyên Hoang, Louis Faucon, and El-Mahdi El-Mhamdi. Recommendation algorithms, a neglected opportunity for public health. *Revue Médecine et Philosophie*, 4(2):16–24, 2021.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. doi: 10.1017/9781139020411.

W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8ce6fc704072e351679ac97d4a985574-Abstract.html.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Tushar Chandra, and Craig Boutilier. Slateq: A tractable decomposition for reinforcement learning with recommendation sets. In Sarit Kraus (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 2592–2599. ijcai.org, 2019. doi: 10.24963/ijcai.2019/360. URL https://doi.org/10.24963/ijcai.2019/360.

Deborah G. Johnson and Nicholas Diakopoulos. What to do about deepfakes. *Commun. ACM*, 64 (3):33–35, 2021. doi: 10.1145/3447255. URL https://doi.org/10.1145/3447255.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2021.

Jakub Konecný, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *CoRR*, abs/1511.03575, 2015. URL http://arxiv.org/abs/1511.03575.

Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops, SP Workshops, San Francisco, CA, USA, May 21, 2020*, pp. 69–75. IEEE, 2020. doi: 10.1109/SPW50608.2020.00028. URL https://doi.org/10.1109/SPW50608.2020.00028.

Florian Lehmann and Daniel Buschek. Examining autocompletion as a basic concept for interaction with generative AI. *i-com*, 19(3):251–264, 2021. doi: 10.1515/icom-2020-0025. URL https://doi.org/10.1515/icom-2020-0025.

Saeed Mahloujifar, Mohammad Mahmoody, and Ameer Mohammed. Data poisoning attacks in multi-party learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4274–4283. PMLR, 2019. URL http://proceedings.mlr.press/v97/mahloujifar19a.html.

Kris McGuffie and Alex Newhouse. The radicalization risks of GPT-3 and advanced neural language models. *CoRR*, abs/2009.06807, 2020. URL https://arxiv.org/abs/2009.06807.

Reshef Meir, Shaull Almagor, Assaf Michaely, and Jeffrey S. Rosenschein. Tight bounds for strategyproof classification. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '11, pp. 319–326, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0982657153.

Reshef Meir, Ariel D. Procaccia, and Jeffrey S. Rosenschein. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.03.008. URL https://www.sciencedirect.com/science/article/pii/S000437021200029X.

El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3518–3527. PMLR, 2018. URL http://proceedings.mlr.press/v80/mhamdi18a.html.

Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In Bhavani M. Thuraisingham, Battista Biggio, David Mandell Freeman, Brad Miller, and Arunesh Sinha (eds.), *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pp. 27–38. ACM, 2017. doi: 10.1145/3128572.3140451. URL https://doi.org/10.1145/3128572.3140451.

Lisa-Maria Neudert, Philip Howard, and Bence Kollanyi. Sourcing and automation of political news and information during three european elections. *Social Media+ Society*, 5(3): 2056305119863147, 2019.

Javier Perote and Juan Perote-Peña. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47(2):153–176, 2004. ISSN 0165-4896. doi: https://doi.org/10.1016/S0165-4896(03)00085-4. URL https://www.sciencedirect.com/science/article/pii/S0165489603000854.

Javier Perote and Olavide Sevilla. The impossibility of strategy-proof clustering. *Economics Bulletin*, 2003.

Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (eds.), *Recommender Systems Handbook*, pp. 1–35. Springer, 2011. doi: 10.1007/978-0-387-85820-3_1. URL https://doi.org/10.1007/978-0-387-85820-3_1.

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9389–9398. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/schwarzschild21a.html.

Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. Explanation-guided backdoor poisoning attacks against malware classifiers. In Michael Bailey and Rachel Greenstadt (eds.), *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pp. 1487–1504. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/severi.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6106–6116, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/22722a343513ed45f14905eb07621686-Abstract.html.

Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on federated learning. *CoRR*, abs/2108.10241, 2021. URL https://arxiv.org/abs/2108.10241.

Heung-Yeung Shum, Xiaodong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers Inf. Technol. Electron. Eng.*, 19(1):10–26, 2018. doi: 10.1631/FITEE.1700826. URL https://doi.org/10.1631/FITEE.1700826.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pp. 1374–1383. The Association for Computer Linguistics, 2013. URL https://www.aclweb.org/anthology/P13-1135/.

Joan E. Solsman. Youtube's ai is the puppet master over most of what you watch. *CNET*, 2018.

Fnu Suya, Saeed Mahloujifar, Anshuman Suri, David Evans, and Yuan Tian. Model-targeted poisoning attacks with provable convergence. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10000–10010. PMLR, 2021. URL http://proceedings.mlr.press/v139/suya21a.html.

Loc Truong, Chace Jones, Brian Hutchinson, Andrew August, Brenda Praggastis, Robert Jasper, Nicole Nichols, and Aaron Tuor. Systematic evaluation of backdoor data poisoning attacks on image classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pp. 3422–3431. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPRW50498.2020.00402. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w47/Truong_Systematic_Evaluation_of_Backdoor_Data_Poisoning_Attacks_on_Image_Classifiers_CVPRW_2020_paper.html.

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3261–3275, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Yuanyuan Wu, Eric W. T. Ngai, Pengkun Wu, and Chong Wu. Fake online reviews: Literature review, synthesis, and directions for future research. *Decis. Support Syst.*, 132:113280, 2020. doi: 10.1016/j.dss.2020.113280. URL https://doi.org/10.1016/j.dss.2020.113280.

Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 14431–14440. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01445. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Zhao_Clean-Label_Backdoor_Attacks_on_Video_Recognition_Models_CVPR_2020_paper.html.

Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7614–7623. PMLR, 2019. URL http://proceedings.mlr.press/v97/zhu19a.html.

Appendices

A COMPARISON SUM OVER EXPECTATIONS

In this section, we report the results of experiments where we compare a sum-based local loss to an expectation-based local loss.

We considered 1 node (the active user) with the whole FashionMNIST training dataset (60,000 data points) and 10 nodes (the idle users) with 10 data points each (randomly drawn from FashionMNIST training set).

We compared two different classifiers to which we refer as a "linear model" and "2-layers neural network", both using *CrossEntropy* loss. The linear model has $(784 + 1) \times 10$ parameters. The neural network has 2 layers of 784 parameters with bias, with *ReLU* activation in between, adding up to $((784 + 1) \times 784 + (784 + 1) \times 10$.

Note also that, in all our experiments, we did not consider any local regularization, i.e. we set $\nu \triangleq 0$. All our experiments are seeded with seed 999.

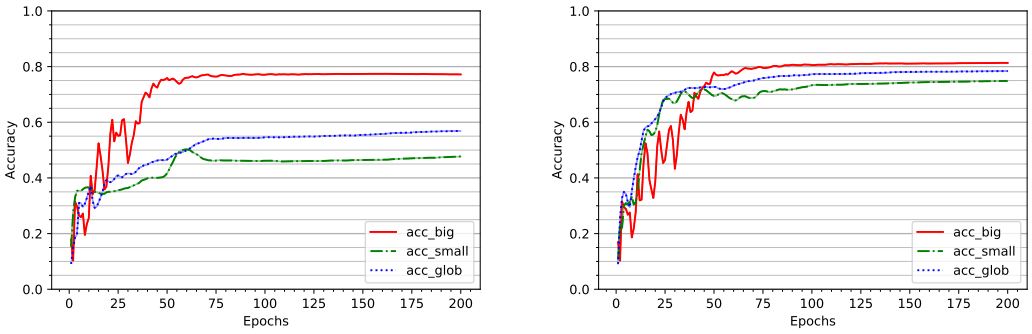
A.1 NOISY FASHIONMNIST

In this section, 60% of the training dataset is labeled randomly, before being distributed, to make the problem harder.

On each plot, we display the top-1 accuracy on the MNIST test dataset (10 000 images) for the active user, for the global model and for one of the idle users (in Table 1, the mean accuracy for idle users is reported), as we vary the value of λ. Intuitively, λ models how much we want the local models to be similar.

In the case of learning FashionMNIST, given that the data is i.i.d., larger values of λ are more meaningful (though our experiments show that they may hinder convergence speed). However, in settings where users have different data distributions, e.g. because the labels depend on users' preferences, then smaller values of λ may be more relevant.

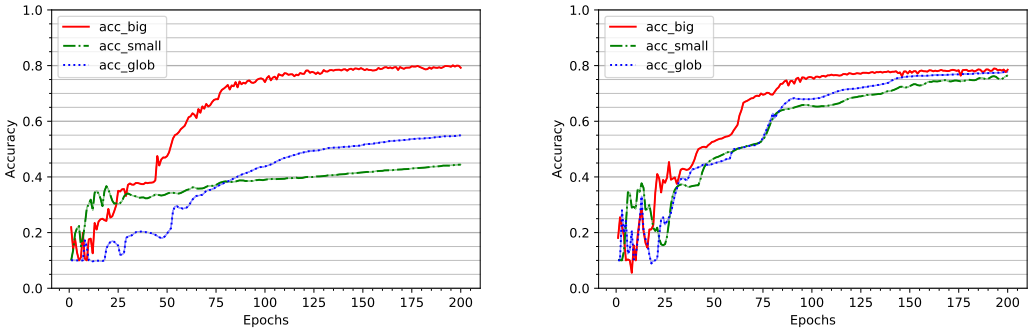
Note that the use of a common value of λ in both cases is slightly misleading, as using the sum intuitively decreases the comparative weight of the regularization term. To reduce this effect, for this experiment only, we divide the local losses by the average of the number of data points per node for the sum version. This way, if the number of points is equal for all nodes, the two losses will be exactly the same. What's more, our experiments seem to robustly show that using the sum consistently outperforms the expectation, for both a linear classifier and a 2-layer neural network, for the problem of noisy FashionMNIST classification.



(a) Using the average

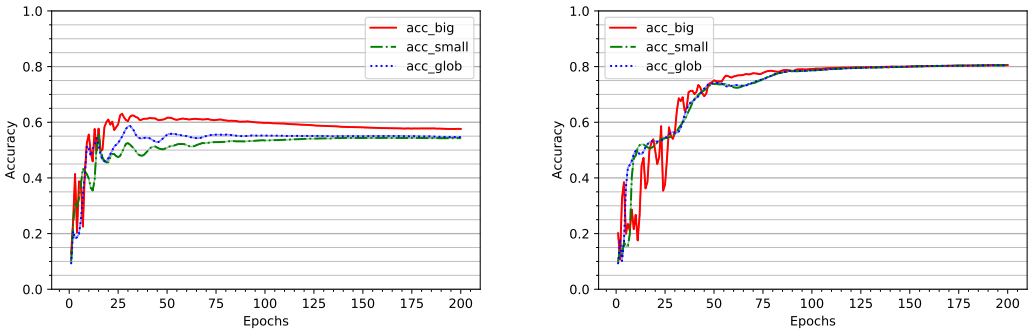
(b) Using the sum

Figure 4: Linear model on noisy FashionMNIST, for $\lambda = 0.01$.



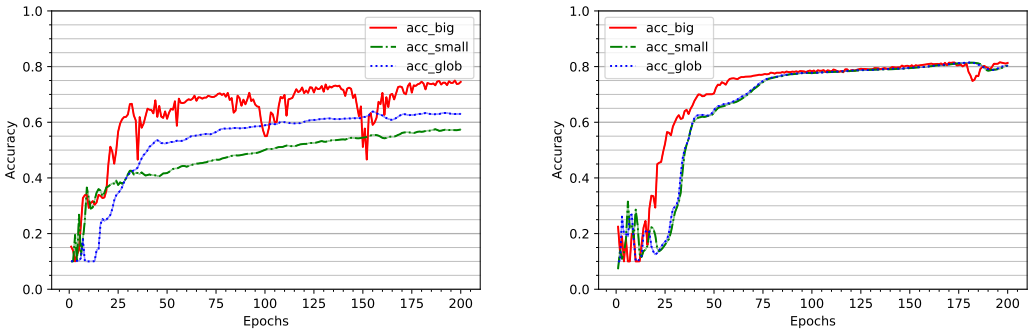
(a) Using the average (b) Using the sum

Figure 5: 2-layer neural network on noisy FashionMNIST, for $\lambda = 0.01$.



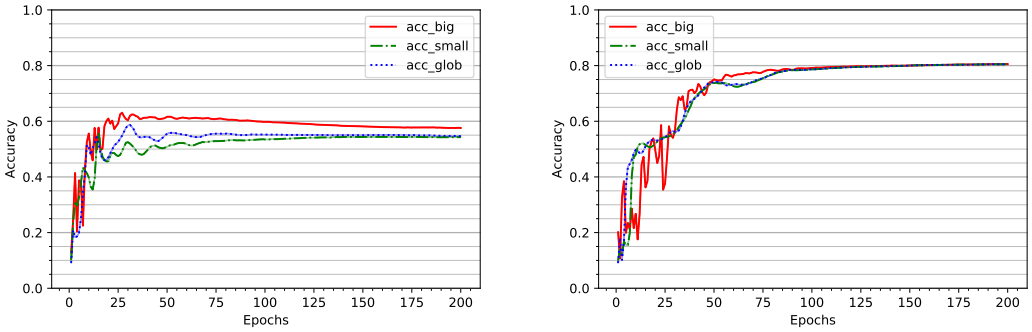
(a) Using the average (b) Using the sum

Figure 6: Linear model on noisy FashionMNIST, for $\lambda = 0.1$.



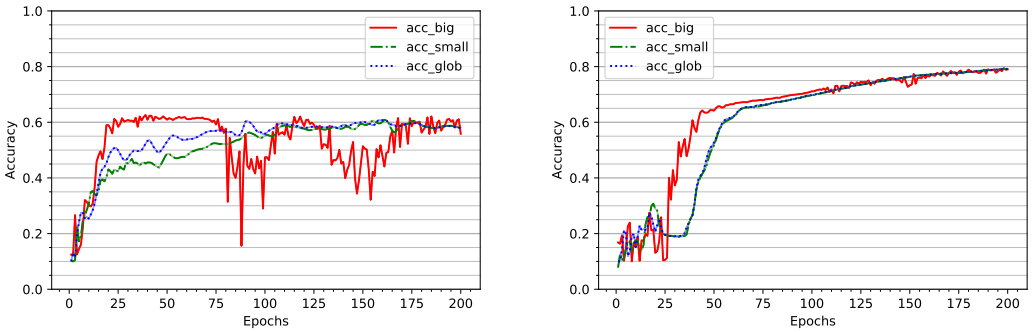
(a) Using the average (b) Using the sum

Figure 7: 2-layer neural network on noisy FashionMNIST, for $\lambda = 0.1$.



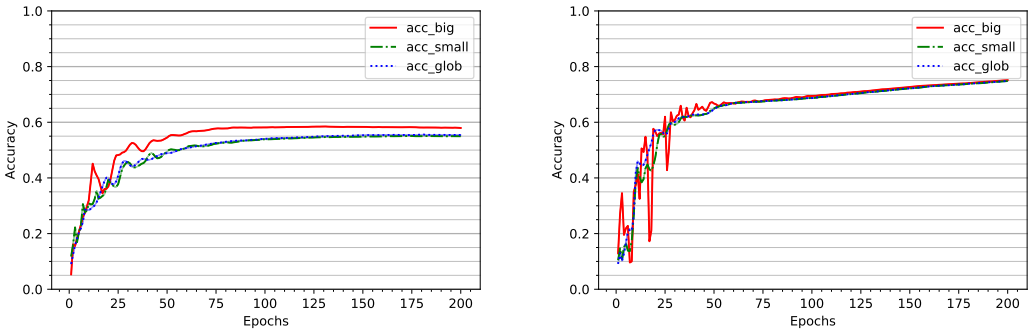
(a) Using the average (b) Using the sum

Figure 8: Linear model on noisy FashionMNIST, for $\lambda = 1$.



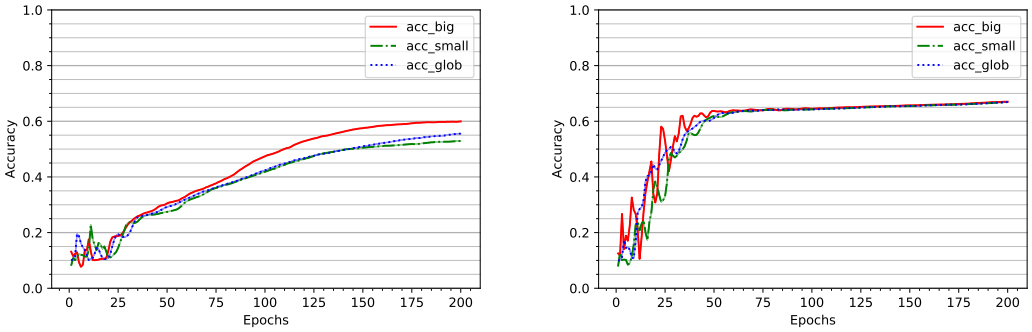
(a) Using the average (b) Using the sum

Figure 9: 2-layer neural network on noisy FashionMNIST, for $\lambda = 1$.



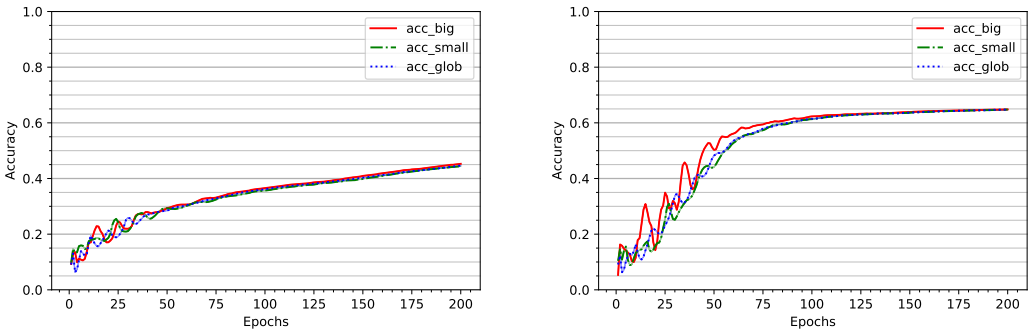
(a) Using the average (b) Using the sum

Figure 10: Linear model on noisy FashionMNIST, for $\lambda = 10$.



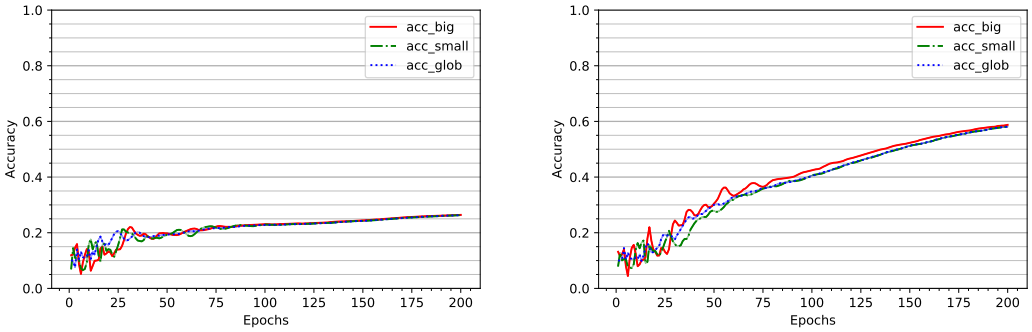
(a) Using the average (b) Using the sum

Figure 11: 2-layer neural network on noisy FashionMNIST, for $\lambda = 10$.



(a) Using the average (b) Using the sum

Figure 12: Linear model on noisy FashionMNIST, for $\lambda = 100$.

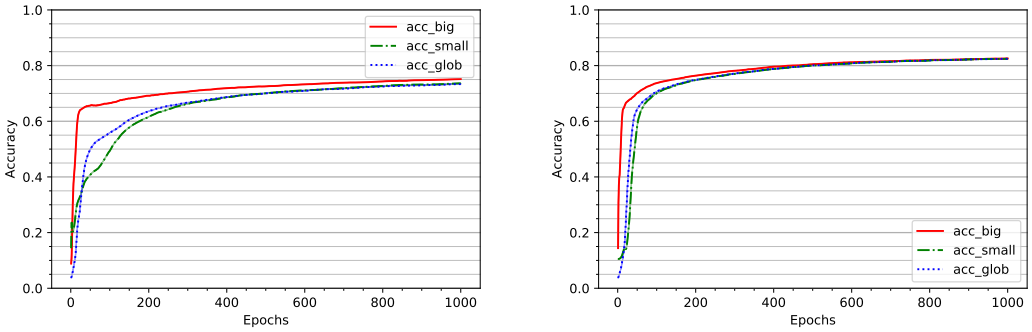


(a) Using the average (b) Using the sum

Figure 13: 2-layer neural network on noisy FashionMNIST, for $\lambda = 100$.

A.2 FASHIONMNIST WITHOUT NOISE

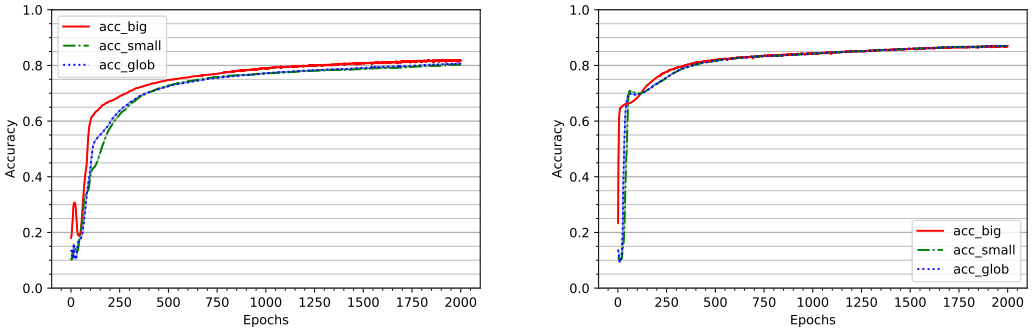
Recall that we introduced noise into FashionMNIST to make the problem harder to learn and observe a clear difference between the average and the sum. In this section, we present results of our experiments when the noise is removed.



(a) Using the average

(b) Using the sum

Figure 14: Linear model on FashionMNIST (without noise), for $\lambda = 1$.



(a) Using the average

(b) Using the sum

Figure 15: 2-layer neural network on FashionMNIST (without noise), for $\lambda = 1$.

Even without noise, the difference between using the sum and using the expectation still seems important. We acknowledge, however, that the plots suggest that even though we ran this experiment for 10 times more (and 5 times more for the linear model) than other experiments, we might not have reached convergence yet, and that the use of the expectation might still eventually gets closer to the case of sum. We believe that the fact that the difference between sum and expectation in the absence of noise is weak is due to the fact that the FashionMNIST dataset is sufficiently linearly separable. Thus, we achieve a near-zero loss in both cases, which make the sum and the expectation close at optimum.

Even in this case, however, we observed that the sum clearly outperforms the expectation especially, in the first epochs. We argue that the reason for this is the following. By taking the average in local losses, the weights of the data of idle nodes are essentially blown out of proportion. As a result, the optimizer will very quickly fit these data. However, the signal from the data of the active node will then be too weak, so that the optimizer has to first almost perfectly fit the idle nodes' data before it can catch the signal of the active node's data and hence the average achieves weaker convergence performances than the sum.

B Linear regression and classification are gradient PAC*

Throughout this section, we use the following terminology.

Definition 3. *Consider a parameterized event $\mathcal{E}(\mathcal{I})$. We say that the event $\mathcal{E}$ occurs with high probability if $\mathbb{P}\left[\mathcal{E}(\mathcal{I})\right] \to 1$ as $\mathcal{I} \to \infty$.*

B.1 PRELIMINARIES

Define $\|\Sigma\|_2 \triangleq \max_{\|x\|_2 \neq 0}(\|\Sigma x\|_2 / \|x\|_2)$ the ℓ_2 operator norm of the matrix Σ. For symmetric matrices Σ, this is also the largest eigenvalue in absolute value.

Theorem 9 (Covariance concentration, Theorem 6.5 in Wainwright (2019))**.** *Denote $\Sigma = \mathbb{E}\left[\mathcal{Q}_i \mathcal{Q}_i^T\right]$, where $\mathcal{Q}_i \in \mathbb{R}^d$ is from a $\sigma_{\mathcal{Q}}$-sub-Gaussian random distribution $\tilde{\mathcal{Q}}$. Then, there are universal constants c_1, c_2 and c_3 such that, for any set $\{\mathcal{Q}_i\}_{i\in[\mathcal{I}]}$ of i.i.d. samples from $\tilde{\mathcal{Q}}$, and any $\delta > 0$, the sample covariance $\widehat{\Sigma} = \frac{1}{\mathcal{I}} \sum \mathcal{Q}_i \mathcal{Q}_i^T$ satisfies the bound*

$$\mathbb{P}\left[\frac{1}{\sigma_{\mathcal{Q}}^2} \left\|\hat{\Sigma} - \Sigma\right\|_2 \geq c_1 \left(\sqrt{\frac{d}{\mathcal{I}}} + \frac{d}{\mathcal{I}}\right) + \delta\right] \leq c_2 \exp\left(-c_3 \mathcal{I} \min(\delta, \delta^2)\right). \tag{8}$$

Theorem 10 (Weyl's Theorem, Theorem 4.3.1 in Horn & Johnson (2012))**.** *Let A and B be Hermitian*[2] *and let the respective eigenvalues of A and B and $A + B$ be $\{\lambda_i(A)\}_{i=1}^d$, $\{\lambda_i(B)\}_{i=1}^d$, and $\{\lambda_i(A + B)\}_{i=1}^d$, each increasingly ordered. Then*

$$\lambda_i(A + B) \leq \lambda_{i+j}(A) + \lambda_{d-j}(B), \quad j = 0, 1, ..., d - i, \tag{9}$$

and

$$\lambda_{i+j}(A) + \lambda_{j+1}(B) \leq \lambda_i(A + B), \quad j = 0, ..., i - 1, \tag{10}$$

for each $i = 1, ..., d$.

Lemma 1. *Consider two symmetric definite positive matrices S and Σ. Denote ρ_{min} and λ_{min} their minimal eigenvalues. Then $|\rho_{min} - \lambda_{min}| \leq \|S - \Sigma\|_2$.*

Proof. This is a direct consequence of Theorem 10, for $A = S$, $B = \Sigma - S$, $i = 1$, and $j = 0$. □

Corollary 1. *There are universal constants c_1, c_2 and c_3 such that, for any $\sigma_{\mathcal{Q}}$-sub-Gaussian vector distribution $\tilde{\mathcal{Q}} \in \mathbb{R}^d$ and any $\delta > 0$, the sample covariance $\widehat{\Sigma} = \frac{1}{\mathcal{I}} \sum \mathcal{Q}_i \mathcal{Q}_i^T$ satisfies the bound*

$$\mathbb{P}\left[\frac{1}{\sigma_{\mathcal{Q}}^2} \left|\min \text{SP}(\hat{\Sigma}) - \min \text{SP}(\Sigma)\right| \geq c_1 \left(\sqrt{\frac{d}{\mathcal{I}}} + \frac{d}{\mathcal{I}}\right) + \delta\right] \leq c_2 \exp\left(-c_3 \mathcal{I} \min(\delta, \delta^2)\right), \tag{11}$$

where $\min \text{SP}(\hat{\Sigma})$ and $\min \text{SP}(\Sigma)$ are the minimal eigenvalues of $\hat{\Sigma}$ and Σ.

Proof. This follows from Theorem 9 and Lemma 1. □

Lemma 2. *With high probability, $\min \text{SP}(\hat{\Sigma}) \geq \min \text{SP}(\Sigma)/2$.*

Proof. Denote $\lambda_{min} \triangleq \min \text{SP}(\Sigma)$ and $\widehat{\lambda}_{min} \triangleq \min \text{SP}(\hat{\Sigma})$. Since each $\mathcal{Q}_i$ is drawn i.i.d. from a $\sigma_{\mathcal{Q}}$-sub-Gaussian, we can apply Corollary 1. Namely, there are constants c_1, c_2 and c_3, such that for any $\delta > 0$, we have

$$\mathbb{P}\left[\left|\widehat{\lambda}_{min} - \lambda_{min}\right| \geq c_1 \sigma_{\mathcal{Q}}^2 \left(\sqrt{\frac{d}{\mathcal{I}}} + \frac{d}{\mathcal{I}}\right) + \delta \sigma_{\mathcal{Q}}^2\right] \leq c_2 \exp\left(-c_3 \mathcal{I} \min\left\{\delta, \delta^2\right\}\right). \tag{12}$$

We now set $\delta \triangleq \lambda_{min}/(4\sigma_{\mathcal{Q}}^2)$ and we consider $\mathcal{I}$ large enough so that $c_1 \left(\sqrt{\frac{d}{\mathcal{I}}} + \frac{d}{\mathcal{I}}\right) \leq \lambda_{min}/(4\sigma_{\mathcal{Q}}^2)$. With high probability, we then have $\widehat{\lambda}_{min} \geq \lambda_{min}/2$. □

[2] For real matrices, Hermitian is the same as symmetric.

B.2 LINEAR REGRESSION IS GRADIENT-PAC

B.2.1 LEMMAS FOR LINEAR REGRESSION

Before moving to the main proof that linear regression is gradient-PAC, we first prove a few useful lemmas. These lemmas will rest on the following well-known theorems.

Theorem 11 (Lemma 2.7.7 in Vershynin (2018))**.** *If X and Y are sub-Gaussian, then XY is sub-exponential.*

Theorem 12 (Equation 2.18 in Wainwright (2019))**.** *If $X_1, \ldots, X_{\mathcal{I}}$ are iid sub-exponential variables, then there exist constants c_4, c_5 such that, for all $\mathcal{I}$, we have*

$$\forall t \in [0, c_4],\ \mathbb{P}\left[|X - \mathbb{E}[X]| \geq t\mathcal{I}\right] \leq 2\exp\left(-c_5 \mathcal{I} t^2\right). \tag{13}$$

Lemma 3. *For all $j \in [d]$, the random variables $X_i \triangleq \xi_i \mathcal{Q}_i[j]$ are iid, sub-exponential and have zero mean.*

Proof. The fact that these variables are iid follows straightforwardly from the fact that the noises ξ_i are iid, and the queries $\mathcal{Q}_i$ are also iid. Moreover, both are sub-Gaussian, and by Theorem 11, the product of sub-Gaussian variables is sub-exponential. Finally, we have $\mathbb{E}[X] = \mathbb{E}[\xi \mathcal{Q}[j]] = \mathbb{E}[\xi]\,\mathbb{E}[\mathcal{Q}[j]] = 0$, using the independence of the noise and the query, and the fact that noises have zero mean ($\mathbb{E}[\xi] = 0$). □

Lemma 4. *There exists B such that $\left\|\sum_{i\in\mathcal{I}} \xi_i \mathcal{Q}_i\right\|_2 \leq B\mathcal{I}^{3/4}$ with high probability.*

Proof. By Lemma 3, the terms $\xi_i \mathcal{Q}_i[j]$ are iid, sub-exponential and have zero mean. Therefore, by Theorem 12, there exist constants c_4 and c_5 such that for any coordinate $j \in [d]$ of $\xi_i \mathcal{Q}_i$ and for all $0 \leq u \leq c_4$, we have

$$\mathbb{P}\left[\left|\sum_{i\in\mathcal{I}} \xi_i \mathcal{Q}_i[j]\right| \geq \mathcal{I}u\right] \leq 2\exp\left(-c_5 \mathcal{I} u^2\right). \tag{14}$$

Plugging $u = v\mathcal{I}^{(-1/4)}$ into the inequality for some small enough constant v, and using union bound then yields

$$\mathbb{P}\left[\left\|\sum_{i\in\mathcal{I}} \xi_i \mathcal{Q}_i\right\|_2 \geq \mathcal{I}^{(3/4)} v\sqrt{d}\right] \leq \mathbb{P}\left[\left\|\sum_{i\in\mathcal{I}} \xi_i \mathcal{Q}_i\right\|_\infty \geq \mathcal{I}^{(3/4)} v\right] \leq 2d\exp\left(-c_5\sqrt{\mathcal{I}} v^2\right). \tag{15}$$

Defining $B \triangleq v\sqrt{d}$ yields the lemma. □

B.2.2 PROOF THAT LINEAR REGRESSION IS GRADIENT-PAC

We now move on to proving that least square linear regression is gradient-PAC.

Proof of Theorem 1. Note that $\nabla_\theta \ell(\theta, \mathcal{Q}, \mathcal{A}) = (\theta^T \mathcal{Q} - \mathcal{A})\mathcal{Q}$. Thus, on input $i \in [\mathcal{I}]$, we have

$$\nabla_\theta \ell(\theta, \mathcal{Q}_i, \mathcal{A}(\mathcal{Q}_i, \theta^\dagger)) = \left((\theta - \theta^\dagger)^T \mathcal{Q}_i\right)\mathcal{Q}_i - \xi_i \mathcal{Q}_i. \tag{16}$$

Moreover, we have

$$(\theta - \theta^\dagger)^T \nabla_\theta \left(\nu \|\theta\|_2^2\right) = 2\nu(\theta - \theta^\dagger)^T \theta = 2\nu \left\|\theta - \theta^\dagger\right\|_2^2 + 2\nu(\theta - \theta^\dagger)^T \theta^\dagger. \tag{17}$$

As a result, we have

$$(\theta - \theta^\dagger)^T \nabla_\theta \mathcal{L}(\theta, \mathcal{D}) = \tag{18}$$

$$\mathcal{I}(\theta - \theta^\dagger)^T \widehat{\Sigma}(\theta - \theta^\dagger) - (\theta - \theta^\dagger)^T \left(\sum_{i\in\mathcal{I}} \xi_i \mathcal{Q}_i\right) + 2\nu \left\|\theta - \theta^\dagger\right\|_2^2 + 2\nu(\theta - \theta^\dagger)^T \theta^\dagger. \tag{19}$$

But now, with high probability, we have $(\theta-\theta^\dagger)^T\widehat{\Sigma}(\theta-\theta^\dagger) \geq (\lambda_{min}/2)\left\|\theta-\theta^\dagger\right\|_2^2$ (Lemma 2) and $\left\|\sum_{i\in\mathcal{I}}\xi_i\mathcal{Q}_i\right\|_2 \leq B\mathcal{I}^{(3/4)}$ (Lemma 4). Using the fact that $\left\|\theta^\dagger\right\|_2 \leq \mathcal{K}$ and the Cauchy-Schwarz inequality, we have

$$(\theta-\theta^\dagger)^T\nabla_\theta\mathcal{L}(\theta,\mathcal{D}) \geq (\frac{\lambda_{min}}{2}\mathcal{I}+\nu)\left\|\theta-\theta^\dagger\right\|_2^2 - (B\mathcal{I}^{(3/4)}+2\nu\mathcal{K})\left\|\theta-\theta^\dagger\right\|_2. \tag{20}$$

Denoting $A_\mathcal{K} \triangleq \frac{\lambda_{min}}{2}$ and $B_\mathcal{K} \triangleq B+2\nu\mathcal{K}$ and using the fact that $\mathcal{I}\geq 1$, we then have

$$(\theta-\theta^\dagger)^T\nabla_\theta\mathcal{L}(\theta,\mathcal{D}) \geq A_\mathcal{K}\mathcal{I}\left\|\theta-\theta^\dagger\right\|_2^2 - B_\mathcal{K}\mathcal{I}^{(3/4)}\left\|\theta-\theta^\dagger\right\|_2 \tag{21}$$

$$\geq A_\mathcal{K}\mathcal{I}\min\left\{\left\|\theta-\theta^\dagger\right\|_2, \left\|\theta-\theta^\dagger\right\|_2^2\right\} - B_\mathcal{K}\mathcal{I}^{(3/4)}\left\|\theta-\theta^\dagger\right\|_2, \tag{22}$$

with high probability. This corresponds to saying Assumption 2 is satisfied for $\alpha = 3/4$. □

B.3 Logistic regression

B.3.1 Lemmas about the sigmoid function

We first prove two useful lemmas about the following logistic distance function.

Definition 4. *We define the logistic distance function by* $\Delta(a,b) \triangleq (a-b)(\sigma(a)-\sigma(b))$.

Lemma 5. *If $a,b\in\mathbb{R}$ such that for some $k>0$, $|a|\leq k$ and $|b|\leq k$, then there exists some constant $c_k>0$ such that*

$$\Delta(a,b) \geq c_k\left|a-b\right|^2. \tag{23}$$

Proof. Note that the derivative of $\sigma(z)$ is strictly positive, symmetric ($\sigma'(z)=\sigma'(-z)$) and monotonically decreasing for $z\geq 0$. Therefore, for any $z\in[-k,k]$, we know $\sigma'(z)\geq c_k \triangleq \sigma'(k)$. Thus, by the mean value theorem, we have

$$\frac{\sigma(a)-\sigma(b)}{a-b} \geq c_k. \tag{24}$$

Multiplying both sides by $(a-b)^2$ then yields the lemma. □

Lemma 6. *If $b\in\mathbb{R}$, and $|b|\leq k$, for some $k>0$, then there exists a constant d_k, such that for any $a\in\mathbb{R}$, we have*

$$\Delta(a,b) \geq d_k\left|a-b\right| - d_k \tag{25}$$

Proof. Assume $|a-b|\geq 1$ and define $d_k \triangleq \sigma(k+1)-\sigma(k)$. If $b\geq 0$, since $\sigma'(z)$ is decreasing for $z\geq 0$, we have $\sigma(b)-\sigma(b-1)\geq\sigma(b+1)-\sigma(b)\geq d_k$, and by symmetry, a similar argument holds for $b\leq 0$. Thus, we have

$$\left|\sigma(a)-\sigma(b)\right| \geq \min\left\{\sigma(b)-\sigma(b-1), \sigma(b+1)-\sigma(b)\right\} \geq d_k. \tag{26}$$

Therefore,

$$(a-b)(\sigma(a)-\sigma(b)) \geq d_k\left|a-b\right| \geq d_k\left|a-b\right| - d_k. \tag{27}$$

For the case of $|a-b|\leq 1$, we also have $(a-b)(\sigma(a)-\sigma(b))\geq 0 \geq d_k\left|a-b\right|-d_k$. □

B.3.2 A uniform lower bound

Definition 5. *Denote* $\mathbb{S}^{d-1} \triangleq \left\{\mathbf{u}\in\mathbb{R}^d \,\middle|\, \|\mathbf{u}\|_2 = 1\right\}$ *the hypersphere in* $\mathbb{R}^d$.

Lemma 7. *Assume* $\textsc{Supp}(\tilde{\mathcal{Q}})$ *spans* $\mathbb{R}^d$. *Then, for all* $\mathbf{u}\in\mathbb{S}^{d-1}$, $\mathbb{E}\left[\left|\mathcal{Q}^T\mathbf{u}\right|\right]>0$.

Proof. Let $\mathbf{u}\in\mathbb{S}^{d-1}$. We know that there exists $\mathcal{Q}_1,\ldots,\mathcal{Q}_d\in\textsc{Supp}(\tilde{\mathcal{Q}})$ and $\alpha_1,\ldots,\alpha_d\in\mathbb{R}$ such that $\mathbf{u}$ is colinear with $\sum\alpha_j\mathcal{Q}_j$. In particular, we then have $\mathbf{u}^T\sum\alpha_j\mathcal{Q}_j = \sum\alpha_j(\mathcal{Q}_j^T\mathbf{u})\neq 0$. Therefore, there must be a query $\mathcal{Q}_*\in\textsc{Supp}(\tilde{\mathcal{Q}})$ such that $\mathcal{Q}_*^T\mathbf{u}\neq 0$, which implies $a\triangleq\left|\mathcal{Q}_*^T\mathbf{u}\right| > 0$ By continuity of the scalar product, there must then also exist $\varepsilon>0$ such that, for any $\mathcal{Q}\in\mathcal{B}(\mathcal{Q}_*,\varepsilon)$, we have $\left|\mathcal{Q}^T\mathbf{u}\right|\geq a/2$, where $\mathcal{B}(\mathcal{Q}_*,\varepsilon)$ is an Euclidean ball centered on $\mathcal{Q}_*$ and of radius ε.

But now, by definition of the support, we know that $p \triangleq \mathbb{P}\left[\mathcal{Q} \in \mathcal{B}(\mathcal{Q}_*, \varepsilon)\right] > 0$. By the law of total expectation, we then have

$$\begin{aligned}\mathbb{E}\left[\left|\mathcal{Q}^T\mathbf{u}\right|\right] &= \mathbb{E}\left[\left|\mathcal{Q}^T\mathbf{u}\right| \,\middle|\, \mathcal{Q} \in \mathcal{B}(\mathcal{Q}_*, \varepsilon)\right]\mathbb{P}\left[\mathcal{Q} \in \mathcal{B}(\mathcal{Q}_*, \varepsilon)\right] \\ &\quad + \mathbb{E}\left[\left|\mathcal{Q}^T\mathbf{u}\right| \,\middle|\, \mathcal{Q} \notin \mathcal{B}(\mathcal{Q}_*, \varepsilon)\right]\mathbb{P}\left[\mathcal{Q} \notin \mathcal{B}(\mathcal{Q}_*, \varepsilon)\right] \quad (28)\\ &\geq ap/2 + 0 > 0, \quad (29)\end{aligned}$$

which is the lemma. □

Lemma 8. *Assume that, for all unit vectors $\mathbf{u} \in \mathbb{S}^{d-1}$, we have $\mathbb{E}\left[\left|\mathcal{Q}^T\mathbf{u}\right|\right] > 0$, and that $\textsc{Supp}(\tilde{\mathcal{Q}})$ is bounded by $M_{\mathcal{Q}}$. Then there exists $C > 0$ such that, with high probability,*

$$\forall \mathbf{u} \in \mathbb{S}^{d-1},\ \sum_{i\in\mathcal{I}} \left|\mathcal{Q}_i^T\mathbf{u}\right| \geq C\mathcal{I}. \quad (30)$$

Proof. By continuity of the scalar product and the expectation operator, and by compactness of $\mathbb{S}^{d-1}$, we know that

$$C_0 \triangleq \inf_{\mathbf{u}\in\mathbb{R}^d} \mathbb{E}\left[\left|\mathcal{Q}^T\mathbf{u}\right|\right] > 0. \quad (31)$$

Now define $\varepsilon \triangleq C_0/4M_{\mathcal{Q}}$. Note that $\mathbb{S}^{d-1} \subset \bigcup_{\mathbf{u}\in\mathbb{S}^{d-1}} \mathcal{B}(\mathbf{u}, \varepsilon)$. Thus we have a covering of the hypersphere by open sets. But since $\mathbb{S}^{d-1}$ is compact, we know that we can extract a finite covering. In other words, there exists a finite subset $S \subset \mathbb{S}^{d-1}$ such that $\mathbb{S}^{d-1} \subset \bigcup_{\mathbf{u}\in S} \mathcal{B}(\mathbf{u}, \varepsilon)$. Put differently, for any $\mathbf{v} \in \mathbb{S}^{d-1}$, there exists $\mathbf{u} \in S$ such that $\|\mathbf{u} - \mathbf{v}\|_2 \leq \varepsilon$.

Now consider $\mathbf{u} \in S$. Given that $\textsc{Supp}(\tilde{\mathcal{Q}})$ is bounded, we know that $\left|\mathcal{Q}_i^T\mathbf{u}\right| \in [0, M_{\mathcal{Q}}]$. Moreover, such variables $\left|\mathcal{Q}_i^T\mathbf{u}\right|$ are iid. By Hoeffding's inequality, for any $t > 0$, we have

$$\mathbb{P}\left[\left|\sum_{i\in\mathcal{I}} \left|\mathcal{Q}_i^T\mathbf{u}\right| - \mathcal{I}\mathbb{E}\left[\left|\mathcal{Q}^T\ \mathbf{u}\right|\right]\right| \geq \mathcal{I}t\right] \leq 2\exp\left(\frac{-2\mathcal{I}t^2}{M_{\mathcal{Q}}}\right). \quad (32)$$

Choosing $t = C_0/2$ then yields

$$\begin{aligned}\mathbb{P}\left[\sum_{i\in\mathcal{I}} \left|\mathcal{Q}_i^T\mathbf{u}\right| \leq \frac{C_0\mathcal{I}}{2}\right] &\leq \mathbb{P}\left[\left|\sum_{i\in\mathcal{I}} \left|\mathcal{Q}_i^T\mathbf{u}_{\theta-\theta^\dagger}\right| - \mathcal{I}\mathbb{E}\left[\left|\mathcal{Q}^T\ \mathbf{u}_{\theta-\theta^\dagger}\right|\right]\right| \geq \frac{\mathcal{I}C_0}{2}\right] \quad (33)\\ &\leq 2\exp\left(\frac{-\mathcal{I}{C_0}^2}{2M_{\mathcal{Q}}}\right). \quad (34)\end{aligned}$$

Taking a union bound for $\mathbf{u} \in S$ then guarantees

$$\mathbb{P}\left[\forall \mathbf{u} \in S,\ \sum_{i\in\mathcal{I}} \left|\mathcal{Q}_i^T\mathbf{u}\right| \geq \frac{C_0\mathcal{I}}{2}\right] \geq 1 - 2\left|S\right|\exp\left(\frac{-\mathcal{I}{C_0}^2}{2M_{\mathcal{Q}}}\right), \quad (35)$$

which clearly goes to 1 as $\mathcal{I} \to \infty$. Thus $\forall \mathbf{u} \in S,\ \sum_{i\in\mathcal{I}} \left|\mathcal{Q}_i^T\mathbf{u}\right| \geq \frac{C_0\mathcal{I}}{2}$ holds with high probability.

Now consider $\mathbf{v} \in \mathbb{S}^{d-1}$. We know that there exists $\mathbf{u} \in S$ such that $\|\mathbf{u} - \mathbf{v}\|_2 \leq \varepsilon$. Then, we have

$$\begin{aligned}\sum_{i\in[\mathcal{I}]} \left|\mathcal{Q}_i^T\mathbf{v}\right| &= \sum_{i\in[\mathcal{I}]} \left|\mathcal{Q}_i^T\mathbf{u} + \mathcal{Q}_i^T(\mathbf{v} - \mathbf{u})\right| \quad (36)\\ &\geq \sum_{i\in[\mathcal{I}]} \left|\mathcal{Q}_i^T\mathbf{u}\right| - \mathcal{I}M_{\mathcal{Q}}\left\|\mathbf{v} - \mathbf{u}\right\|_2 \quad (37)\\ &\geq \frac{C_0\mathcal{I}}{2} - \mathcal{I}M_{\mathcal{Q}}\frac{C_0}{4M_{\mathcal{Q}}} = \frac{C_0\mathcal{I}}{4}, \quad (38)\end{aligned}$$

which proves the lemma. □

B.3.3 LOWER BOUND ON THE DISCREPANCY BETWEEN TRUE AND REPORTED ANSWERS

Lemma 9. *Assume that $\tilde{\mathcal{Q}}$ has a bounded support, whose interior contains the origin. Suppose also that $\left\|\theta^\dagger\right\|_2 \leq \mathcal{K}$. Then there exists $A_\mathcal{K}$ such that, with high probability, we have*

$$\sum_{i\in[\mathcal{I}]} \Delta(\mathcal{Q}_i^T\theta, \mathcal{Q}_i^T\theta^\dagger) \geq A_\mathcal{K}\mathcal{I}\min\left\{\left\|\theta-\theta^\dagger\right\|_2, \left\|\theta-\theta^\dagger\right\|_2^2\right\}. \tag{39}$$

Proof. Note that by Cauchy-Schwarz inequality we have

$$\left|\mathcal{Q}_i^T\theta^\dagger\right| \leq \|\mathcal{Q}_i\|_2 \left\|\theta^\dagger\right\|_2 \leq M_\mathcal{Q}\mathcal{K}. \tag{40}$$

Thus, Lemma 6 implies the existence of a positive constant $d_\mathcal{K}$, such that for all $\theta \in \mathbb{R}^d$, we have

$$\sum_{i\in\mathcal{I}} \Delta\left(\mathcal{Q}_i^T\theta, \mathcal{Q}_i^T\theta^\dagger\right) \geq \sum_{i\in\mathcal{I}} \left(d_\mathcal{K}\left|\mathcal{Q}_i^T\theta - \mathcal{Q}_i^T\theta^\dagger\right| - d_\mathcal{K}\right) \tag{41}$$

$$= -d_\mathcal{K}\mathcal{I} + d_\mathcal{K}\left\|\theta-\theta^\dagger\right\|_2 \sum_{i\in\mathcal{I}} \left|\mathcal{Q}_i^T\mathbf{u}_{\theta-\theta^\dagger}\right|, \tag{42}$$

where $\mathbf{u}_{\theta-\theta^\dagger} \triangleq (\theta-\theta^\dagger)/\left\|\theta-\theta^\dagger\right\|_2$ is the unit vector in the direction of $\theta-\theta^\dagger$.

Now, by Lemma 8, we know that, with high probability, for all unit vectors $\mathbf{u} \in \mathbb{S}^{d-1}$, we have $\sum\left|\mathcal{Q}_i^T\mathbf{u}\right| \geq C\mathcal{I}$. Thus, for $\mathcal{I}$ sufficiently large, for any $\theta \in \mathbb{R}^d$, with high probability, we have

$$\sum_{i\in\mathcal{I}} \Delta(\mathcal{Q}_i^T\theta, \mathcal{Q}_i^T\theta^\dagger) \geq \frac{d_\mathcal{K}C_{min}}{2}\mathcal{I}\left\|\theta-\theta^\dagger\right\|_2 - d_\mathcal{K}\mathcal{I}. \tag{43}$$

Defining $e_\mathcal{K} \triangleq \frac{d_\mathcal{K}C_{min}}{4}$, and $f_\mathcal{K} \triangleq \frac{4}{C_{min}}$, for $\left\|\theta-\theta^\dagger\right\|_2 > f_\mathcal{K}$, we then have

$$\sum_{i\in\mathcal{I}} \Delta(\mathcal{Q}_i^T\theta, \mathcal{Q}_i^T\theta^\dagger) \geq e_\mathcal{K}\mathcal{I}\left\|\theta-\theta^\dagger\right\|_2. \tag{44}$$

We now focus on the case of $\left\|\theta-\theta^\dagger\right\|_2 \leq f_\mathcal{K}$. The triangle inequality yields $\|\theta\|_2 \leq \left\|\theta-\theta^\dagger\right\|_2 + \left\|\theta^\dagger\right\|_2 \leq f_\mathcal{K} + \mathcal{K}$. By Cauchy-Schwarz inequality, we then have $\left|\mathcal{Q}_i^T\theta\right| \leq (f_\mathcal{K}+\mathcal{K})M_\mathcal{Q} \triangleq g_\mathcal{K}$ and $\left|\mathcal{Q}_i^T\theta^\dagger\right| \leq \mathcal{K}M_\mathcal{Q} \leq g_\mathcal{K}$. Thus, by Lemma 5, we know there exists some constant $c_\mathcal{K}$ such that

$$\sum_{i\in\mathcal{I}} \left(\sigma(\mathcal{Q}_i^T\theta) - \sigma(\mathcal{Q}_i^T\theta^\dagger)\right)\left(\mathcal{Q}_i^T\theta - \mathcal{Q}_i^T\theta^\dagger\right) \geq \sum_{i\in\mathcal{I}} c_\mathcal{K}\left|\mathcal{Q}_i^T\theta - \mathcal{Q}_i^T\theta^\dagger\right|^2 \tag{45}$$

$$= \sum_{i\in\mathcal{I}} c_\mathcal{K}(\theta-\theta^\dagger)^T\mathcal{Q}_i\mathcal{Q}_i^T(\theta-\theta^\dagger) \tag{46}$$

$$= c_\mathcal{K}(\theta-\theta^\dagger)^T\left(\sum_{i\in\mathcal{I}}\mathcal{Q}_i\mathcal{Q}_i^T\right)(\theta-\theta^\dagger). \tag{47}$$

Since distribution $\tilde{\mathcal{Q}}$ is bounded (and thus sub-Gaussian), by Theorem 9, with high probability, we have

$$(\theta-\theta^\dagger)^T\left(\sum_{i\in\mathcal{I}}\mathcal{Q}_i\mathcal{Q}_i^T\right)(\theta-\theta^\dagger) \geq \frac{\lambda_{min}}{2}\mathcal{I}\left\|\theta-\theta^\dagger\right\|_2^2, \tag{48}$$

where λ_{min} is the smallest eigenvalue of $\mathbb{E}\left[\mathcal{Q}_i\mathcal{Q}_i^T\right]$. Thus, for $\left\|\theta-\theta^\dagger\right\|_2 \leq f_\mathcal{K}$, we have

$$\sum_{i\in\mathcal{I}} \left(\sigma(\mathcal{Q}_i^T\theta) - \sigma(\mathcal{Q}_i^T\theta^\dagger)\right)\left(\mathcal{Q}_i^T\theta - \mathcal{Q}_i^T\theta^\dagger\right) \geq \frac{\lambda_{min}c_\mathcal{K}}{2}\mathcal{I}\left\|\theta-\theta^\dagger\right\|_2^2. \tag{49}$$

Combining this with (44), and defining $A_\mathcal{K} \triangleq \min\left\{\frac{\lambda_{min}c_\mathcal{K}}{2}, e_\mathcal{K}\right\}$, we then obtain the lemma. □

B.3.4 PROOF THAT LOGISTIC REGRESSION IS GRADIENT-PAC

Now we proceed with the proof that logistic regression is gradient-PAC.

Proof of Theorem 2. Note that $\sigma(-z) = e^{-z}\sigma(z) = 1-\sigma(z)$ and $\sigma'(z) = e^{-z}\sigma^2(z)$. We then have

$$\nabla_\theta \ell(\theta, \mathcal{Q}, \mathcal{A}) = -\frac{\sigma'(\mathcal{A}\mathcal{Q}^T\theta)\mathcal{A}\mathcal{Q}}{\sigma(\mathcal{A}\mathcal{Q}^T\theta)} = -e^{-\mathcal{A}\mathcal{Q}^T\theta}\sigma(\mathcal{A}\mathcal{Q}^T\theta)\mathcal{A}\mathcal{Q} \tag{50}$$

$$= -\sigma(-\mathcal{A}\mathcal{Q}^T\theta)\mathcal{A}\mathcal{Q} = \left(\sigma(\mathcal{Q}^T\theta) - \mathbb{1}\left[\mathcal{A}=1\right]\right)\mathcal{Q}, \tag{51}$$

where $\mathbb{1}\left[\mathcal{A}=1\right]$ is the indicator function that outputs 1 if $\mathcal{A}=1$, and 0 otherwise. As a result,

$$(\theta-\theta^\dagger)^T\nabla_\theta\mathcal{L}(\theta,\mathcal{D}) = \tag{52}$$

$$(\theta-\theta^\dagger)^T\left(\sum_{i\in\mathcal{I}}\left(\sigma(\mathcal{Q}_i^T\theta) - \mathbb{1}\left[\mathcal{A}_i=1\right]\right)\mathcal{Q}_i\right) + 2\nu(\theta-\theta^\dagger)^T\theta \tag{53}$$

$$= (\theta-\theta^\dagger)^T\left(\sum_{i\in\mathcal{I}}\left(\sigma(\mathcal{Q}_i^T\theta) - \sigma(\mathcal{Q}_i^T\theta^\dagger) + \sigma(\mathcal{Q}_i^T\theta^\dagger) - \mathbb{1}\left[\mathcal{A}_i=1\right]\right)\mathcal{Q}_i\right) \tag{54}$$

$$+ 2\nu\left\|\theta-\theta^\dagger\right\|_2^2 + 2\nu(\theta-\theta^\dagger)^T\theta^\dagger \tag{55}$$

$$= \sum_{i\in[\mathcal{I}]}\Delta\left(\mathcal{Q}_i^T\theta, \mathcal{Q}_i^T\theta^\dagger\right) + (\theta-\theta^\dagger)^T\left(\sum_{i\in\mathcal{I}}\left(\sigma(\mathcal{Q}_i^T\theta^\dagger) - \mathbb{1}\left[\mathcal{A}_i=1\right]\right)\mathcal{Q}_i\right) \tag{56}$$

$$+ 2\nu\left\|\theta-\theta^\dagger\right\|_2^2 + 2\nu(\theta-\theta^\dagger)^T\theta^\dagger. \tag{57}$$

By Lemma 9, with high probability, we have

$$\sum_{i\in[\mathcal{I}]}\Delta\left(\mathcal{Q}_i^T\theta, \mathcal{Q}_i^T\theta^\dagger\right) \geq A_\mathcal{K}\mathcal{I}\min\left\{\left\|\theta-\theta^\dagger\right\|_2, \left\|\theta-\theta^\dagger\right\|_2^2\right\}. \tag{58}$$

To control the second term of (56), note that the random vectors $Z_i \triangleq \left(\sigma(\mathcal{Q}_i^T\theta^\dagger) - \mathbb{1}\left[\mathcal{A}_i=1\right]\right)\mathcal{Q}_i$ are iid with norm at most $M_\mathcal{Q}$. Moreover, since $\mathbb{E}\left[\mathbb{1}\left[\mathcal{A}_i=1\right]|\mathcal{Q}_i\right] = \sigma(\mathcal{Q}_i^T\theta^\dagger)$, by the tower rule, we have $\mathbb{E}\left[Z_i\right] = \mathbb{E}\left[\mathbb{E}\left[Z_i|\mathcal{Q}_i\right]\right] = 0$. Therefore, by applying Hoeffding's bound to every coordinate of Z_i, and then taking a union bound, for any $B>0$, we have

$$\mathbb{P}\left[\left\|\sum_{i\in\mathcal{I}}Z_i\right\|_2 \geq B\mathcal{I}^{3/4}\right] \leq 2d\exp\left(-\frac{B^2\sqrt{\mathcal{I}}}{2dM_\mathcal{Q}^2}\right). \tag{59}$$

Applying now Cauchy-Schwarz inequality, with high probability, we have

$$\left|(\theta-\theta^\dagger)^T\left(\sum_{i\in\mathcal{I}}\left(\sigma(\mathcal{Q}_i^T\theta^\dagger) - \mathbb{1}\left[\mathcal{A}_i=1\right]\right)\mathcal{Q}_i\right)\right| \leq B\mathcal{I}^{3/4}\left\|\theta-\theta^\dagger\right\|_2.$$

Combining this with (49) and using $\left\|\theta^\dagger\right\|_2^2 \leq \mathcal{K}$, we then have

$$(\theta-\theta^\dagger)^T\nabla_\theta\mathcal{L}(\theta,\mathcal{D}) \tag{60}$$

$$\geq (A_\mathcal{K}\mathcal{I}+\nu)\left\{\left\|\theta-\theta^\dagger\right\|_2, \left\|\theta-\theta^\dagger\right\|_2^2\right\} - (B\mathcal{I}^{(3/4)} + 2\nu\mathcal{K})\left\|\theta-\theta^\dagger\right\|_2 \tag{61}$$

$$\geq A_\mathcal{K}\mathcal{I}\left\{\left\|\theta-\theta^\dagger\right\|_2, \left\|\theta-\theta^\dagger\right\|_2^2\right\} - B_\mathcal{K}\mathcal{I}^{(3/4)}\left\|\theta-\theta^\dagger\right\|_2, \tag{62}$$

where $B_\mathcal{K} = B + 2\nu\mathcal{K}$. This shows that Assumption 2 is satisfied for logistic loss for $\alpha = 3/4$, and $A_\mathcal{K}$ and $B_\mathcal{K}$ as previously defined.

□

C PROOFS OF PAC-LEARNABILITY

Consider true parameters $\vec{\theta}^\dagger$ and a subset $H \subset [N]$ of honest users. Denote $\vec{\mathcal{D}}_{-H}$ the datasets provided by users $n \in [N] - H$. Each honest user $h \in H$ provides an honest dataset $\mathcal{D}_h$ of cardinality at least $\mathcal{I} \geq 1$. Consider the bound $K_H \triangleq \max_{h \in H} \left\|\theta_h^\dagger\right\|_2$ on the parameter norm of honest active users $h \in H$.

C.1 BOUNDS ON THE OPTIMA

Before proving the theorem, we prove a useful lemma that bounds the set of possible values for the global model and honest local models.

Lemma 10. *Assume that $\mathcal{R}$ and ℓ are nonnegative. For $\mathcal{I}$ large enough, if all honest active nodes $h \in H$ provide at least $\mathcal{I}$ data, then, with high probability, $\vec{\theta}_H^*$ must lie in a compact subset of $\mathbb{R}^{d \times H}$ that does not depend on $\mathcal{I}$.*

Proof. Denote $L^0 \triangleq \textsc{Loss}(0, (\vec{\theta}_H^\dagger, 0_{-H}), (\emptyset, \vec{\mathcal{D}}_{-H}))$. Essentially, we will show that, if $\vec{\theta}_H^*$ is too far from $\vec{\theta}_H^\dagger$, then the loss will take values strictly larger than L^0.

Assumption 2 implies the existence of an event $\mathcal{E}$ that occurs with probability at least $P_0 \triangleq P(K_H, \mathcal{I})^{|H|}$, under which, for any $\theta_h \in \mathbb{R}^d$, we have

$$\left(\theta_h - \theta_h^\dagger\right)^T \nabla \mathcal{L}_h\left(\theta_h\right) \geq A_{K_H} \mathcal{I} \min\left\{\left\|\theta_h - \theta_h^\dagger\right\|_2, \left\|\theta_h - \theta_h^\dagger\right\|_2^2\right\} - B_{K_H} \mathcal{I}^\alpha \left\|\theta_h - \theta_h^\dagger\right\|_2, \quad (63)$$

which implies

$$\mathbf{u}_{\left(\theta_h - \theta_h^\dagger\right)}^T \nabla \mathcal{L}_h\left(\theta_h\right) \geq A_{K_H} \mathcal{I} \min\left\{1, \left\|\theta_h - \theta_h^\dagger\right\|_2\right\} - B_{K_H} \mathcal{I}^\alpha. \quad (64)$$

Note also that $P_0 \to 1$ as $\mathcal{I} \to \infty$. We now integrate both sides over the line segment from $\theta_h^\dagger$ to θ_h. The fundamental theorem of calculus for line integrals then yields

$$\mathcal{L}_h\left(\theta_h\right) - \mathcal{L}_h\left(\theta_h^\dagger\right) = \left\|\theta_h - \theta_h^\dagger\right\|_2 \int_{t=0}^1 \mathbf{u}_{\left(\theta_h - \theta_h^\dagger\right)}^T \nabla \mathcal{L}\left(\theta_h^\dagger + t(\theta_h - \theta_h^\dagger)\right) dt \quad (65)$$

$$\geq \left\|\theta_h - \theta_h^\dagger\right\|_2 \int_{t=0}^1 \left(A_{K_H} \mathcal{I} \min\left\{1, t\left\|\theta_h - \theta_h^\dagger\right\|_2\right\} - B_{K_H} \mathcal{I}^\alpha\right) dt \quad (66)$$

$$= \left\|\theta_h - \theta_h^\dagger\right\|_2 \int_{t=0}^1 \left(A_{K_H} \mathcal{I} \min\left\{1, t\left\|\theta_h - \theta_h^\dagger\right\|_2\right\}\right) dt - B_{K_H} \mathcal{I}^\alpha \left\|\theta_h - \theta_h^\dagger\right\|_2. \quad (67)$$

Now, if $\left\|\theta_h - \theta_h^\dagger\right\|_2 > 2$, we then have

$$\mathcal{L}_h\left(\theta_h\right) - \mathcal{L}_h\left(\theta_h^\dagger\right) \geq \left(\frac{A_{K_H} \mathcal{I}}{2} - B_{K_H} \mathcal{I}^\alpha\right) \left\|\theta_h - \theta_h^\dagger\right\|_2 \quad (68)$$

$$\geq A_{K_H} \mathcal{I} - 2 B_{K_H} \mathcal{I}^\alpha. \quad (69)$$

Now for $\mathcal{I} > \mathcal{I}_1 \triangleq \max\left\{2L^0 / A_{K_H}, (4B_{K_H}/A_{K_H})^{\frac{1}{1-\alpha}}\right\}$, we have

$$\mathcal{L}_h\left(\theta_h\right) - \mathcal{L}_h\left(\theta_h^\dagger\right) > L^0. \quad (70)$$

This implies that if $\left\|\theta_h - \theta_h^\dagger\right\|_2 > 2$ for any $h \in H$, then we have

$$\textsc{Loss}(0, (\vec{\theta}_H^\dagger, 0_{-H}), \vec{\mathcal{D}}) < \textsc{Loss}(\rho, (\vec{\theta}_H, \vec{\theta}_{-H}), \vec{\mathcal{D}}), \quad (71)$$

regardless of ρ and θ_{-H}. Therefore, we must have $\left\|\theta_h^\dagger - \theta_h^*\right\|_2 \leq 2$. Such inequalities describe a bounded closed subset of $\mathbb{R}^{d \times H}$, which is thus compact. □

Lemma 11. *Assume that $\mathcal{R}(\rho,\theta) \to \infty$ as $\|\rho-\theta\|_2 \to \infty$, and that $\left\|\theta_h^\dagger - \theta_h^*\right\|_2 \leq 2$ for all honest users $h \in H$. Then ρ^* must lie in a compact subset of $\mathbb{R}^d$ that does not depend on $\mathcal{I}$.*

Proof. Consider an honest user h'. Given our assumption on $\mathcal{R} \to \infty$, we know that there exists D_{K_H} such that if $\|\rho - \theta_{h'}^*\|_2 \geq D_{K_H}$, then $\mathcal{R}(\rho,\theta_{h'}^*) \geq L^0 + 1$. Thus any global optimum ρ^* must satisfy $\left\|\rho^* - \theta_{h'}^\dagger\right\|_2 \leq \|\rho^* - \theta_{h'}^*\|_2 + \left\|\theta_{h'}^* - \theta_{h'}^\dagger\right\|_2 \leq D_{K_H} + 2$. □

C.2 Proof of Theorem 3

Proof of Theorem 3. Fix $\varepsilon, \delta > 0$. We want to show the existence of some value of $\mathcal{I}(\varepsilon,\delta,\vec{\mathcal{D}}_{-H},\vec{\theta}^\dagger)$ that will guarantee (ε,δ)-PAC learning for honest users.

By lemmas 10 and 11, we know that the set C of possible values for $(\rho^*,\vec{\theta}_H^*)$ is compact. Now, we define

$$E_{K_H} \triangleq \max_{(\rho,\theta)\in C} \|\nabla_\theta \mathcal{R}(\rho,\theta)\|_2 \tag{72}$$

the maximum of the norm of achievable gradients at the optimum. We know this maximum exists since C is compact.

Using the optimality of $(\rho^*,\vec{\theta}^*)$, for all $h \in H$, we have

$$\begin{aligned}
0 &\in (\theta_h^* - \theta_h^\dagger)^T \nabla_{\theta_h} \textsc{Loss}(\rho^*,\vec{\theta}^*) && (73)\\
&= (\theta_h^* - \theta_h^\dagger)^T \nabla \mathcal{L}_h(\theta_h^*) + (\theta_h^* - \theta_h^\dagger)^T \nabla_{\theta_h} \mathcal{R}(\rho^*,\theta_h^*) && (74)\\
&\geq (\theta_h^* - \theta_h^\dagger)^T \nabla \mathcal{L}_h(\theta_h^*) - \left\|\theta_h^* - \theta_h^\dagger\right\|_2 \|\nabla_{\theta_h} \mathcal{R}(\rho^*,\theta_h^*)\|_2 && (75)\\
&\geq (\theta_h^* - \theta_h^\dagger)^T \nabla \mathcal{L}_h(\theta_h^*) - E_{K_H} \left\|\theta_h^* - \theta_h^\dagger\right\|_2 . && (76)
\end{aligned}$$

We now apply assumption 2 for $\theta = \theta_h^*$ (for $h \in H$). Thus, there exists some other event $\mathcal{E}'$ with probability at least P_0, under which, for all $h \in H$, we have

$$0 \geq A_{K_H}\mathcal{I} \min\left\{\left\|\theta_h^* - \theta_h^\dagger\right\|_2, \left\|\theta_h^* - \theta_h^\dagger\right\|_2^2\right\} - B_{K_H}\mathcal{I}^\alpha \left\|\theta_h^* - \theta_h^\dagger\right\|_2 - E_{K_H}\left\|\theta_h^* - \theta_h^\dagger\right\|_2 . \tag{77}$$

Now if $\mathcal{I} > \mathcal{I}_2 \triangleq \max\left\{2E_{K_H}/A_{K_H}, (2B_{K_H}/A_{K_H})^{\frac{1}{1-\alpha}}\right\}$ this inequality cannot hold for $\left\|\theta_h^* - \theta_h^\dagger\right\|_2 \geq 1$. Therefore, for $\mathcal{I} > \mathcal{I}_2$, we have $\left\|\theta_h^* - \theta_h^\dagger\right\|_2 < 1$, and thus,

$$0 \geq A_{K_H}\mathcal{I}\left\|\theta_h^* - \theta_h^\dagger\right\|_2^2 - B_{K_H}\mathcal{I}^\alpha \left\|\theta_h^* - \theta_h^\dagger\right\|_2 - E_{K_H}\left\|\theta_h^* - \theta_h^\dagger\right\|_2 \tag{78}$$

and thus,

$$\left\|\theta_h^* - \theta_h^\dagger\right\|_2 \leq \frac{B_{K_H}\mathcal{I}^\alpha + E_{K_H}}{A_{K_H}\mathcal{I}} . \tag{79}$$

Now note that $\mathbb{P}[\mathcal{E} \wedge \mathcal{E}'] = 1 - \mathbb{P}[\neg\mathcal{E} \vee \neg\mathcal{E}'] \geq 1 - \mathbb{P}[\neg\mathcal{E}] - \mathbb{P}[\neg\mathcal{E}'] = 2P_0 - 1$. It now suffices to consider $\mathcal{I}$ larger than $\mathcal{I}_2$ and large enough so that $P(K_H,\mathcal{I})^{|H|} \geq 1 - \delta/2$ (whose existence is guaranteed by Assumption 2, and which guarantees $2P_0 - 1 \geq 1 - \delta$) and so that $\frac{B_{K_H}\mathcal{I}^\alpha + E_{K_H}}{A_{K_H}\mathcal{I}} \leq \varepsilon$ to obtain the theorem. □

D Convexity lemmas

D.1 General lemmas

Definition 6. *We say that $f : \mathbb{R}^d \to \mathbb{R}$ is locally strongly convex if, for any convex compact set $C \subset \mathbb{R}^d$, there exists $\mu > 0$ such that f is μ-strongly convex on C, i.e. for any $x, y \in C$ and any $\lambda \in [0,1]$, we have*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|_2^2 . \tag{80}$$

It is well-known that if f is differentiable, this condition amounts to saying that $\|\nabla f(x) - \nabla f(y)\|_2 \geq \mu \|x - y\|_2$ for all $x, y \in C$. And if f is twice differentiable, then it amounts to saying $\nabla^2 f(x) \succeq \mu I$ for all $x \in C$.

Lemma 12. *If f is locally strongly convex and g is convex, then $f + g$ is locally strongly convex.*

Proof. Indeed, $(f+g)(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|_2^2 + \lambda g(x) + (1-\lambda)g(y) = \lambda(f+g)(x) + (1-\lambda)(f+g)(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|_2^2$. □

Definition 7. *We say that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if it is differentiable and if its gradient is L-Lipschitz continuous, i.e. for any $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 . \tag{81}$$

Lemma 13. *If f is L_f-smooth and g is L_g-smooth, then $f + g$ is $(L_f + L_g)$-smooth.*

Proof. Indeed, $\|\nabla(f+g)(x) - \nabla(f+g)(y)\|_2 \leq \|\nabla f(x) - \nabla f(y)\|_2 + \|\nabla g(x) - \nabla g(y)\|_2 \leq L_f \|x-y\|_2 + L_g \|x-y\|_2 = (L_f + L_g)\|x-y\|_2$. □

Lemma 14. *Suppose that $f : \mathbb{R}^d \times \mathbb{R}^{d'} \mapsto \mathbb{R}$ is locally strongly convex and L-smooth, and that, for any $x \in X$, where $X \subset \mathbb{R}^d$ is a convex compact subset, the map $y \mapsto f(x, y)$ has a minimum $y^*(x)$. Note that local strong convexity guarantees the uniqueness of this minimum. Then, there exists K such that the function y^* is K-Lipschitz continuous on X.*

Proof. The existence and uniqueness of $y^*(x)$ hold by strong convexity. Fix x, x'. By optimality of y^*, we know that $\nabla_y f(x, y^*(x)) = \nabla_y f(x', y^*(x')) = 0$. We then have the following bounds

$$\begin{aligned}
\mu \|y^*(x) - y^*(x')\|_2 &\leq \|\nabla_y f(x, y^*(x)) - \nabla_y f(x, y^*(x'))\|_2 = \|\nabla_y f(x, y^*(x'))\|_2 && (82)\\
&= \|\nabla_y f(x, y^*(x')) - \nabla_y f(x', y^*(x'))\|_2 && (83)\\
&\leq \|\nabla f(x, y^*(x')) - \nabla f(x', y^*(x'))\|_2 && (84)\\
&\leq L \|(x - x', y^*(x') - y^*(x'))\|_2 = L \|x - x'\|_2 , && (85)
\end{aligned}$$

where we first used the local strong convexity assumption, then the fact that $\nabla_y f(x, y^*(x)) = 0$, then the fact that $\nabla_y f(x', y^*(x')) = 0$, and then the L-smooth assumption. □

Lemma 15. *Suppose that $f : \mathbb{R}^d \times \mathbb{R}^{d'} \mapsto \mathbb{R}$ is locally strongly convex and L-smooth, and that, for any $x \in X$, where $X \subset \mathbb{R}^d$ is a convex compact subset, the map $y \mapsto f(x, y)$ has a minimum $y^*(x)$. Define $g(x) \triangleq \min_{y \in Y} f(x, y)$. Then g is convex and differentiable on X and $\nabla g(x) = \nabla_x f(x, y^*(x))$.*

Proof. First we prove that g is convex. Let $x_1, x_2 \in \mathbb{R}^d$, and $\lambda_1, \lambda_2 \in [0, 1]$ with $\lambda_1 + \lambda_2 = 1$. For any $y_1, y_2 \in \mathbb{R}^{d'}$, we have

$$\begin{aligned}
g(\lambda_1 x_1 + \lambda_2 x_2) &= \min_{y \in \mathbb{R}^{d'}} f(\lambda_1 x_1 + \lambda_2 x_2, y) && (86)\\
&\leq f(\lambda_1 x_1 + \lambda_2 x_2, \lambda_1 y_1 + \lambda_2 y_2) && (87)\\
&\leq \lambda_1 f(x_1, y_1) + \lambda_2 f(x_2, y_2). && (88)
\end{aligned}$$

Taking the infimum of the right-hand side over y_1 and y_2 yields $g(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 g(x_1) + \lambda_2 g(x_2)$, which proves the convexity of g.

Now denote $h(x) = \nabla_x f(x, y^*(x))$. We aim to show that $\nabla g(x) = h(x)$. Let $\varepsilon \in \mathbb{R}^d$ small enough so that $x + \varepsilon \in X$. Now note that we have

$$\begin{aligned}
g(x + \varepsilon) &= \min_{y \in \mathbb{R}^{d'}} f(x + \varepsilon, y) \leq f(x + \varepsilon, y^*(x)) && (89)\\
&= f(x, y^*(x)) + \varepsilon^T \nabla_x f(x, y^*(x)) + o(\|\varepsilon\|_2) && (90)\\
&= g(x) + \varepsilon^T h(x) + o(\|\varepsilon\|_2), && (91)
\end{aligned}$$

which shows that $h(x)$ is a superderivative of g at x. We now show that it is also a subderivative. To do so, first note that its value at $x+\varepsilon$ is approximately the same, i.e.

$$\begin{aligned}\|h(x+\varepsilon)-h(x)\|_2 &\leq \|\nabla_x f(x+\varepsilon, y^*(x+\varepsilon)) - \nabla_x f(x, y^*(x+\varepsilon))\|_2 \\ &\quad + \|\nabla_x f(x, y^*(x+\varepsilon)) - \nabla_x f(x, y^*(x))\|_2 \qquad (92)\\ &\leq L\|\varepsilon\|_2 + L\|y^*(x+\varepsilon)-y^*(x)\|_2 \leq \left(L+\frac{L^2}{\mu}\right)\|\varepsilon\|_2, \qquad (93)\end{aligned}$$

where we used the L-smoothness of f and Lemma 14. Now notice that

$$\begin{aligned}g(x) &= \min_{y\in\mathbb{R}^{d'}} f(x,y) \leq f(x, y^*(x+\varepsilon)) = f((x+\varepsilon)-\varepsilon, y^*(x+\varepsilon)) \qquad (94)\\ &= f(x+\varepsilon, y^*(x+\varepsilon)) - \varepsilon^T\nabla_x f(x+\varepsilon, y^*(x+\varepsilon)) + o(\|\varepsilon\|_2) \qquad (95)\\ &= g(x+\varepsilon) - \varepsilon^T h(x) - \varepsilon^T\left(h(x+\varepsilon)-h(x)\right) + o(\|\varepsilon\|_2), \qquad (96)\end{aligned}$$

But we know that $\|h(x+\varepsilon)-h(x)\|_2 = \mathcal{O}(\|\varepsilon\|_2)$. Rearranging the terms then yields

$$g(x+\varepsilon) \geq g(x) + \varepsilon^T h(x) - o(\|\varepsilon\|_2), \qquad (97)$$

which shows that $h(x)$ is also a subderivative. Therefore, we know that $g(x+\varepsilon) = g(x)+\varepsilon^T h(x) + o(\|\varepsilon\|_2)$, which boils down to saying that g is differentiable in $x\in X$, and that $\nabla g(x) = h(x)$. □

Lemma 16. *Suppose that $f : X\times\mathbb{R}^{d'} \to \mathbb{R}$ is μ-strongly convex, where $X\subset\mathbb{R}^d$ is closed and convex. Then $g : X\to\mathbb{R}$, defined by $g(x) = \inf_{y\in Y} f(x,y)$, is well-defined and μ-strongly convex too.*

Proof. The function $y\mapsto f(x,y)$ is still strongly convex, which means that it is at least equal to a quadratic approximation around 0, which is a function that goes to infinity in all directions as $\|y\|_2\to\infty$. This proves that the infimum must be reached within a compact set, which implies the existence of a minimum. Thus g is well-defined. Moreover, for any $x_1, x_2\in X, y_1, y_2\in\mathbb{R}^{d'}$, and $\lambda_1,\lambda_2\geq 0$ with $\lambda_1+\lambda_2=1$, we have

$$\begin{aligned}g(\lambda_1 x_1+\lambda_2 x_2) &= \inf_y f(\lambda_1 x_1+\lambda_2 x_2, y) \qquad (98)\\ &\leq f(\lambda_1 x_1+\lambda_2 x_2, \lambda_1 y_1+\lambda_2 y_2) \qquad (99)\\ &\leq \lambda_1 f(x_1,y_1)+\lambda_2 f(x_2,y_2) - \frac{\mu}{2}\lambda_1\lambda_2\|(x_1-x_2, y_1-y_2)\|_2^2 \qquad (100)\\ &\leq \lambda_1 f(x_1,y_1)+\lambda_2 f(x_2,y_2) - \frac{\mu}{2}\lambda_1\lambda_2\|x_1-x_2\|_2^2, \qquad (101)\end{aligned}$$

where we used the μ-strong convexity of f. Taking the infimum over y_1, y_2 implies the μ-strong convexity of g. □

D.2 Applications to Loss

Lemma 17. *Under Assumption 1,* Loss *is locally strongly convex and L-smooth.*

Proof. All terms of Loss are L_0-smooth, for an appropriate value of L_0. By Lemma 13, their sum is thus also L-smooth, for an appropriate value of L. Now, given Lemma 12, to prove that Loss is locally strongly convex, it suffices to prove that $\nu\sum\|\theta_n\|_2^2 + \mathcal{R}_0(\rho-\theta_1)$ is locally strongly convex. Consider any convex compact set $C\subset\mathbb{R}^{d\times(1+N)}$. Since $\mathcal{R}_0$ is locally strongly convex, we know that there exists $\mu>0$ such that $\nabla^2\mathcal{R}_0\succeq\mu I$. As a result,

$$\begin{aligned}(\rho,\vec{\theta})^T\left(\nabla^2\text{Loss}\right)(\rho,\vec{\theta}) &\geq \nu\sum_{n\in[N]}\|\theta_n\|_2^2 + \mu\|\rho-\theta_1\|_2^2 \qquad (102)\\ &= \nu\|\theta_1\|_2^2 + \mu\|\rho\|_2^2 + \mu\|\theta_1\|_2^2 - 2\mu\rho^T\theta_1 + \nu\sum_{n\neq 1}\|\theta_n\|_2^2. \qquad (103)\end{aligned}$$

Now define $\alpha \triangleq \sqrt{\frac{2\mu}{\nu+2\mu}}$. Clearly, $0 < \alpha < 1$. Moreover, $0 \leq \left\|\frac{1}{\alpha}\theta_1 - \alpha\rho\right\|_2^2 = \frac{1}{\alpha^2}\|\theta_1\|_2^2 + \alpha^2\|\rho\|_2^2 - 2\rho^T\theta_1$. Therefore $2\rho^T\theta_1 \leq \alpha^2\|\rho\|_2^2 + \frac{1}{\alpha^2}\|\theta_1\|_2^2$, which thus implies

$$(\rho,\vec{\theta})^T\left(\nabla^2\text{Loss}\right)(\rho,\vec{\theta}) \geq \left(\nu + \mu\left(1-\alpha^{-2}\right)\right)\|\theta_1\|_2^2 + \mu\left(1-\alpha^2\right)\|\rho\|_2^2 + \nu\sum_{n\neq 1}\|\theta_n\|_2^2 \quad (104)$$

$$\geq \frac{\nu}{2}\|\theta_1\|_2^2 + \frac{2\nu\mu}{\nu+2\mu}\|\rho\|_2^2 + \nu\sum_{n\neq 1}\|\theta_n\|_2^2 \geq \min\left\{\frac{\nu}{2}, \frac{2\nu\mu}{\nu+2\mu}\right\}\left\|(\rho,\vec{\theta})\right\|_2^2, \quad (105)$$

which proves that $\nabla^2\text{Loss} \succeq \kappa I$, with $\kappa > 0$. This shows that Loss is locally strongly convex. □

Lemma 18. *Under Assumption 1, $\rho \mapsto \vec{\theta}^*(\rho,\vec{\mathcal{D}})$ is Lipchitz continuous on any compact set.*

Proof. Define $f_n(\rho,\theta_n) \triangleq \nu\|\theta_n\|_2^2 + \sum_{x\in\mathcal{D}_n}\ell(\theta_n,x) + \lambda\|\rho-\theta_n\|_2^2$. If ℓ is L-smooth, then f_n is clearly $(|\mathcal{D}_n|L + \nu + \lambda)$-smooth. Moreover, if ℓ is convex, then for any ρ, the function $\theta_n \mapsto f_n(\rho,\theta_n)$ is at least ν-strongly convex. Thus Lemma 14 applies, which guarantees that $\rho \mapsto \vec{\theta}^*(\rho,\vec{\mathcal{D}})$ is Lipchitz. □

Lemma 19. *Under Assumption 1, $\rho \mapsto \text{Loss}(\rho,\vec{\theta}^*(\rho,\vec{\mathcal{D}}),\vec{\mathcal{D}})$ is L-smooth and locally strongly convex.*

Proof. By Lemma 17, the global loss is known to be L-smooth, for some value of L and locally strongly convex. Denoting $f : \rho \mapsto \text{Loss}(\rho,\vec{\theta}^*(\rho,\vec{\mathcal{D}}),\vec{\mathcal{D}})$, we then have

$$\|\nabla f(\rho) - \nabla f(\rho')\|_2 \leq \left\|\nabla_\rho\text{Loss}(\rho,\vec{\theta}^*(\rho,\vec{\mathcal{D}}),\vec{\mathcal{D}}) - \nabla_\rho\text{Loss}(\rho',\vec{\theta}^*(\rho',\vec{\mathcal{D}}),\vec{\mathcal{D}})\right\|_2 \quad (106)$$

$$\leq L\left\|(\rho,\vec{\theta}^*(\rho,\vec{\mathcal{D}})) - (\rho',\vec{\theta}^*(\rho',\vec{\mathcal{D}}))\right\|_2 \quad (107)$$

$$\leq L\|\rho-\rho'\|_2, \quad (108)$$

which proves that f is L-smooth.

For strong convexity, note that since the global loss function is locally strongly convex, for any compact convex set C, there exists μ such that $\text{Loss}(\rho,\vec{\theta},\vec{\mathcal{D}})$ is μ-strongly convex on $C = (C_1, C_2) \subset (\mathbb{R}^d, \mathbb{R}^{N\times d})$, therefore, by Lemma 16, $f(\rho)$ will also be μ-strongly convex on C_1 which means that $f(\rho)$ is locally strongly convex. □

E Proof of the reductions

E.1 Proof of the reduction from model attack to data poisoning

Proof of Theorem 4. We omit making the dependence of the optima on $\vec{\mathcal{D}}$ explicit, and we consider any other models ρ and $\vec{\theta}_{-s}$. We have the following inequalities:

$$\text{Loss}_s(\rho^*,\vec{\theta}^*_{-s},\theta_s^\spadesuit,\vec{\mathcal{D}}) = \text{Loss}(\rho^*,\vec{\theta}^*,\vec{\mathcal{D}}) - \mathcal{L}(\theta_s^*,\mathcal{D}_s) \quad (109)$$

$$\leq \text{Loss}(\rho,(\vec{\theta}_{-s},\theta_s^*),\vec{\mathcal{D}}) - \mathcal{L}(\theta_s^*,\mathcal{D}_s) = \text{Loss}_s(\rho,\vec{\theta}_{-s},\theta_s^\spadesuit,\vec{\mathcal{D}}), \quad (110)$$

where we used the optimality of $(\rho^*,\vec{\theta}^*)$ in the second line, and where we repeatedly used the fact that $\theta_s^* = \theta_s^\spadesuit$. This proves that $(\rho^*,\vec{\theta}^*_{-s})$ is a global minimum of the modified loss. □

E.2 Proof of the reduction from data poisoning to model attack

Proof of Theorem 5. Let $\varepsilon,\delta > 0$ and $\theta_s^\spadesuit \in \mathbb{R}^d$. Denote $\rho^\spadesuit \triangleq \rho^*(\theta_s^\spadesuit,\vec{\mathcal{D}}_{-s})$ and $\vec{\theta}^\spadesuit \triangleq \vec{\theta}^*(\theta_s^\spadesuit,\vec{\mathcal{D}}_{-s})$ the result of strategic user s's model attack. We define the compact set C by

$$C \triangleq \left\{\rho,\vec{\theta}_{-s}\,\middle|\,\left\|\rho-\rho^\spadesuit\right\|_2 \leq \varepsilon \wedge \forall n\neq s,\ \left\|\theta_n - \theta_n^\spadesuit\right\|_2 \leq \varepsilon\right\} \quad (111)$$

We define $D \triangleq \overline{\mathbb{R}^{d\times N} - C}$ the closure of the complement of C. Clearly, $\rho^{\spadesuit}, \vec{\theta}^{\spadesuit}_{-s} \notin D$. We aim to show that, when strategic user s reveals a large dataset $\mathcal{D}_s$ whose answers are provided using the attack model $\theta^{\spadesuit}_s$, then the same holds for any global minimum of the global loss $\rho^*(\vec{\mathcal{D}}), \vec{\theta}^*_{-s}(\vec{\mathcal{D}}) \in C$. Note that, to prove this, it suffices to prove that the modified loss takes too large values, even when $\theta^{\spadesuit}_s$ is replaced by $\theta^*_s(\vec{\mathcal{D}})$.

Let us now formalize this. Denote $L^{\spadesuit} \triangleq \textsc{Loss}_s(\rho^{\spadesuit}, \vec{\theta}^{\spadesuit}_{-s}, \theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s})$. We define

$$\eta \triangleq \inf_{\rho, \vec{\theta}_{-s} \in D} \textsc{Loss}_s(\rho, \vec{\theta}_{-s}, \theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s}) - L^{\spadesuit}. \tag{112}$$

By a similar argument as that of Theorem 3, using the assumption $\mathcal{R} \to \infty$ at infinity, we know that the infimum is actually a minimum. Moreover, given that the minimum of the modified loss $\textsc{Loss}_s$ is unique, we know that the value of the loss function at this minimum is different from its value at $\rho^{\spadesuit}, \vec{\theta}^{\spadesuit}_{-s}$. As a result, we must have $\eta > 0$.

Now, since the function $\mathcal{R}$ is differentiable, it must be continuous. By the Heine–Cantor theorem, it is thus uniformly continuous on all compact sets. Thus, there must exist $\kappa > 0$ such that, for all models θ_s satisfying $\left\|\theta_s - \theta^{\spadesuit}_s\right\|_2 \leq \kappa$, we have

$$\left|\mathcal{R}(\theta_s, \rho^{\spadesuit}) - \mathcal{R}(\theta^{\spadesuit}_s, \rho^{\spadesuit})\right| \leq \eta/3. \tag{113}$$

Now, Theorem 3 guarantees the existence of $\mathcal{I}$ such that, if user s provides a dataset $\mathcal{D}_s$ of least $\mathcal{I}$ answers with the model $\theta^{\spadesuit}_s$, then with probability at least $1-\delta$, we will have $\left\|\theta^*_s(\vec{\mathcal{D}}) - \theta^{\spadesuit}_s\right\|_2 \leq \min(\kappa, \varepsilon)$. Under this event, we then have

$$\textsc{Loss}_s\left(\rho^{\spadesuit}, \vec{\theta}^{\spadesuit}_{-s}, \theta^*_s(\vec{\mathcal{D}}), \vec{\mathcal{D}}_{-s}\right) \leq L^{\spadesuit} + \eta/3. \tag{114}$$

Then

$$\inf_{\rho, \vec{\theta}_{-s} \in D} \textsc{Loss}_s(\rho, \vec{\theta}_{-s}, \theta^*_s(\vec{\mathcal{D}}), \vec{\mathcal{D}}_{-s}) \geq \inf_{\rho, \vec{\theta}_{-s} \in D} \textsc{Loss}_s(\rho, \vec{\theta}_{-s}, \theta^{\spadesuit}_s, \vec{\mathcal{D}}_{-s}) - \eta/3 \tag{115}$$

$$\geq L^{\spadesuit} + \eta - \eta/3 \geq L^{\spadesuit} + 2\eta/3 \tag{116}$$

$$> \textsc{Loss}_s\left(\rho^{\spadesuit}, \vec{\theta}^{\spadesuit}_{-s}, \theta^*_s(\vec{\mathcal{D}}), \vec{\mathcal{D}}_{-s}\right). \tag{117}$$

This shows that there is a high probability event under which the minimum of $\rho, \vec{\theta}_{-s} \mapsto \textsc{Loss}_s\left(\rho, \vec{\theta}_{-s}, \theta^*_s(\vec{\mathcal{D}}), \vec{\mathcal{D}}_{-s}\right)$ cannot be reached in D. This is equivalent to what the theorem we needed to prove states. □

E.3 Proof of reduction from model attack to gradient attack

Denote g^{∞}_s the limit of the attack gradients g^t_s. We now define

$$\textsc{Loss}^1_s(\rho) \triangleq \inf_{\vec{\theta}_{-s}} \left\{\textsc{Loss}(\rho, \vec{\theta}, \vec{\mathcal{D}}) - \mathcal{L}_s(\theta_s, \mathcal{D}_s) - \mathcal{R}(\rho, \theta_s)\right\} + \rho^T g^{\infty}_s \tag{118}$$

$$= \inf_{\vec{\theta}_{-s}} \left\{\sum_{n\neq s} \mathcal{L}_n(\theta_n, \mathcal{D}_n) + \sum_{n\neq s} \mathcal{R}(\rho, \theta_n)\right\} + \rho^T g^{\infty}_s, \tag{119}$$

and prove that ρ^t will converge to the minimizer of $\textsc{Loss}^1_s(\rho)$. By Lemma 19, we show that $\textsc{Loss}^1_s(\rho)$ is both locally strongly convex and L-smooth.

Now define $\zeta^t_s \triangleq g^t_s - g^{\infty}_s$. We then know $\zeta^t_s \to 0$ and at iteration t of the optimization algorithm, we will take one step in the direction $G^t \triangleq \nabla \textsc{Loss}^1_s(\rho^t) + \zeta^t_s$, i.e.,

$$\rho^{t+1} = \rho^t - \eta_t G^t. \tag{120}$$

We now prove the following lemma that bounds the difference between the function value in two successive iterations.

Lemma 20. *If* $\text{Loss}_s^1(\rho)$ *is* L*-smooth and* $\eta_t \leq 1/L$*, we have*

$$\text{Loss}_s^1(\rho^{t+1}) - \text{Loss}_s^1(\rho^t) \leq -\frac{\eta_t}{2}\left\|G^t\right\|_2^2 + \eta_t {\zeta_s^t}^T G^t. \tag{121}$$

Proof. Since Loss_s^1 is L-smooth, we have

$$\text{Loss}_s^1(\rho^{t+1}) \leq \text{Loss}_s^1(\rho^t) + (\rho^{t+1} - \rho^t)^T \nabla \text{Loss}_s^1(\rho^t) + \frac{L}{2}\left\|\rho^{t+1} - \rho^t\right\|_2^2. \tag{122}$$

Now plugging $\rho^{t+1} - \rho^t = -\eta_t G^t$ and $\nabla \text{Loss}_s^1(\rho^t) = G^t - \zeta_s^t$ into the inequality implies

$$\text{Loss}_s^1(\rho^{t+1}) - \text{Loss}_s^1(\rho^t) \leq \left(-\eta_t G^t\right)^T \left(G^t - \zeta_s^t\right) + \frac{L}{2}\left\|-\eta_t G^t\right\|_2^2 \tag{123}$$

$$\leq -\frac{\eta_t}{2}\left\|G^t\right\|_2^2 + \eta_t {\zeta_s^t}^T G^t, \tag{124}$$

where we used the fact $\eta_t \leq 1/L$. □

E.3.1 The global model remains bounded

Lemma 21. *There is* M *such that, for all* t*,* $\text{Loss}_s^1(\rho^t) \leq M$*.*

Proof. Consider the closed ball $\mathcal{B}(\rho^*, 1)$ centered on ρ^* and of radius 1. By Lemma 19, we know that Loss_s^1 is locally strongly convex and thus there exists a $\mu_1 > 0$ such that Loss_s^1 is μ_1-strongly convex on $\mathcal{B}(\rho^*, 1)$. Now consider a point ρ_1 on the boundary of $\mathcal{B}(\rho^*, 1)$. By strong convexity we have

$$\left\|\nabla \text{Loss}_s^1(\rho_1)\right\|_2^2 \geq (\rho_1 - \rho^*)^T \nabla \text{Loss}_s^1(\rho_1) \geq \mu_1 \left\|\rho_1 - \rho^*\right\|_2^2 = \mu_1. \tag{125}$$

Now similarly, by the convexity of Loss_s^1 on $\mathbb{R}^d$, for any $\rho \in \mathbb{R}^d - \mathcal{B}(\rho^*, 1)$, we have $\left\|\nabla \text{Loss}_s^1(\rho_1)\right\|_2 \geq \sqrt{\mu_1}$. Now since $\zeta_s^t \to 0$, there exists an iteration T_1 after which ($t \geq T_1$), we have $\left\|\zeta_s^t\right\|_2 \leq \frac{1}{4}\sqrt{\mu_1}$, and thus $\left\|G^t\right\|_2 \geq \left\|\nabla \text{Loss}_s^1(\rho^t)\right\|_2 - \left\|\zeta_s^t\right\|_2 \geq \frac{3}{4}\sqrt{\mu_1}$. Thus, Lemma 20 implies that for $t \geq T_1$, if $\left\|\rho^t - \rho^*\right\|_2 \geq 1$, then

$$\text{Loss}_s^1(\rho^{t+1}) - \text{Loss}_s^1(\rho^t) \leq -\frac{\eta}{2}\left\|G^t\right\|_2^2 + \eta {\zeta_s^t}^T G^t \tag{126}$$

$$\leq -\frac{\eta}{2}\left\|G^t\right\|_2^2 + \eta \left\|\zeta_s^t\right\|_2 \left\|G^t\right\|_2 \tag{127}$$

$$\leq -\frac{\eta}{2}\left\|G^t\right\|_2 \left(\left\|G^t\right\|_2 - 2\left\|\zeta_s^t\right\|_2\right) \tag{128}$$

$$\leq -\frac{\eta}{2}\frac{3}{4}\sqrt{\mu_1}\left(\frac{3}{4}\sqrt{\mu_1} - \frac{2}{4}\sqrt{\mu_1}\right) \leq -\frac{3\eta}{32}\mu_1 < 0. \tag{129}$$

Thus, for $\left\|\rho^t - \rho^*\right\|_2 \geq 1$, the loss cannot increase at the next iteration.

Now consider the case $\left\|\rho^t - \rho^*\right\|_2 < 1$ for $t \geq T_1$. The smoothness of Loss_s^1 implies $\left\|\nabla \text{Loss}_s^1(\rho^t)\right\|_2 < L$. Therefore,

$$\left\|\rho^{t+1} - \rho^*\right\|_2 = \left\|\rho^t - \eta(\nabla \text{Loss}_s^1(\rho^t) + \zeta_s^t) - \rho^*\right\|_2 \tag{130}$$

$$\leq \left\|\rho^{t+1} - \rho^*\right\|_2 + \eta(L + \frac{1}{4}\sqrt{\mu_1}) \leq 1 + \eta(L + \frac{1}{4}\sqrt{\mu_1}). \tag{131}$$

Now we define $M_1 \triangleq \max_{\rho \in \mathcal{B}\left(\rho^*, 1+\eta(L+\frac{1}{4}\sqrt{\mu_1})\right)} \text{Loss}_s^1(\rho)$, the maximum function value in the closed ball $\mathcal{B}\left(\rho^*, 1 + \eta(L + \frac{1}{4}\sqrt{\mu_1})\right)$. Therefore, we have $\text{Loss}_s^1(\rho^{t+1}) \leq M_1$. So far we proved that for $t \geq T_1$, in each iteration of gradient descent either the function value will not increase or it will be upper-bounded by M_1. This implies that for all t, the function value $\text{Loss}_s^1(\rho^t)$ is upper-bounded by

$$M \triangleq \max\left\{\max_{t \leq T_1}\left\{\text{Loss}_s^1(\rho^t)\right\}, M_1\right\}. \tag{132}$$

This concludes the proof. □

Lemma 22. *There is a compact set X such that, for all t, $\rho^t \in X$.*

Proof. Now since Loss_s^1 is μ_1-strongly convex in $\mathcal{B}(\rho^*, 1)$, for any point $\rho \in \mathbb{R}^d$ such that $\|\rho - \rho^t\|_2 = 1$, we have

$$\text{Loss}_s^1(\rho) \geq \text{Loss}_s^1(\rho^*) + \frac{\mu_1}{2}\|\rho - \rho^*\|_2^2 = \text{Loss}_s^1(\rho^*) + \frac{\mu_1}{2}. \tag{133}$$

But now by the convexity of Loss_s^1 in $\mathbb{R}^d$, for any ρ such that $\|\rho - \rho^*\|_2 \geq 1$, we have

$$\text{Loss}_s^1(\rho) \geq \text{Loss}_s^1(\rho^*) + \|\rho - \rho^*\|_2 \frac{\mu_1}{2}. \tag{134}$$

This implies that if $\|\rho^t - \rho^*\|_2 > \frac{2}{\mu_1}\left(M_2 - \text{Loss}_s^1(\rho^*)\right)$, then $\text{Loss}_s^1(\rho^t) > M_2$. Therefore, we must have $\|\rho^t - \rho^*\|_2 \leq \frac{2}{\mu_1}\left(M_2 - \text{Loss}_s^1(\rho^*)\right)$, for all $t \geq 0$. This describes a closed ball, which is a compact set. □

E.3.2 Convergence of the global model under converging gradient attack

Lemma 23. *Suppose $u_t \geq 0$ verifies $u_{t+1} \leq \alpha u_t + \delta_t$, with $\delta_t \to 0$. Then $u_t \to 0$.*

Proof. We now show that for any $\varepsilon > 0$, there exists an iteration $T(\varepsilon)$, such that for $t \geq T(\varepsilon)$, we have $u_t \leq \varepsilon$. For this, note that by induction, we observe that, for all $t \geq 0$,

$$u_{t+1} \leq u_0 \alpha^{t+1} + \sum_{\tau=0}^{t} \alpha^\tau \delta_{t-\tau}. \tag{135}$$

Since $\delta_t \to 0$, there exists an iteration $T_2(\varepsilon)$ such that for all $t \geq T_2(\varepsilon)$, we have $\delta_t \leq \frac{\varepsilon(1-\alpha)}{2}$. Therefore, for $t \geq T_2(\varepsilon)$, we have

$$u_{t+1} \leq u_0 \alpha^{t+1} + \sum_{\tau=0}^{t-T_2(\varepsilon)} \alpha^\tau \delta_{t-\tau} + \sum_{\tau=t-T_2(\varepsilon)+1}^{t} \alpha^\tau \delta_{t-\tau} \tag{136}$$

$$\leq u_0 \alpha^{t+1} + \frac{\varepsilon(1-\alpha)}{2} \sum_{\tau=0}^{t-T_2(\varepsilon)} \alpha^\tau + \sum_{s=0}^{T_2(\varepsilon)-1} \alpha^{t-s} \delta_s \tag{137}$$

$$\leq \left(u_0 + \sum_{s=0}^{T_2(\varepsilon)-1} \alpha^{-s-1} \delta_s\right) \alpha^{t+1} + \frac{\varepsilon(1-\alpha)}{2} \sum_{\tau=0}^{\infty} \alpha^\tau. \tag{138}$$

Denoting $M_0(\varepsilon) \triangleq \sum_{s=0}^{T_2(\varepsilon)-1} \alpha^{-s-1}\delta_s$, we then have

$$u_{t+1} \leq \left(u_0 + M_0(\varepsilon)\right) \alpha^{t+1} + \frac{\varepsilon}{2}. \tag{139}$$

Therefore, for $t \geq \frac{\ln \frac{\varepsilon}{2(u_0 + M_0(\varepsilon))}}{\ln \alpha}$, we have

$$u_{t+1} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \tag{140}$$

This proves that $u_t \to 0$. □

We first prove the first part of Theorem 6.

Lemma 24. *Under Assumption 1 and $\eta_t = \eta \leq 1/L$, if g_s^t converges, then so does ρ^t.*

Proof. Define X based on Lemma 22. Since Loss_s^1 is locally strongly convex, there exists $\mu_2 > 0$ such that Loss_s^1 is μ_2-strongly convex in a convex compact set X containing ρ^t for all $t \geq 0$. By the strong convexity of $\text{Loss}_s^1(\rho)$, we have

$$\text{Loss}_s^1(\rho^t) - \text{Loss}_s^1(\rho^*) \leq (\rho^t - \rho^*)^T \nabla \text{Loss}_s^1(\rho^t) - \frac{\mu_2}{2}\left\|\rho^t - \rho^*\right\|_2^2 \tag{141}$$

$$= (\rho^t - \rho^*)^T \left(G^t - \zeta_s^t\right) - \frac{\mu_2}{2}\left\|\rho^t - \rho^*\right\|_2^2. \tag{142}$$

Now, using the fact

$$(\rho^t - \rho^*)^T G^t = \frac{1}{\eta}(\rho^t - \rho^*)^T(\rho^t - \rho^{t+1}) \tag{143}$$

$$= \frac{1}{2\eta}\left(\left\|\rho^t - \rho^*\right\|_2^2 + \left\|\rho^t - \rho^{t+1}\right\|_2^2 - \left\|\rho^{t+1} - \rho^*\right\|_2^2\right) \tag{144}$$

$$= \frac{1}{2\eta}\left(\eta^2\left\|G^t\right\|_2^2 + \left\|\rho^t - \rho^*\right\|_2^2 - \left\|\rho^{t+1} - \rho^*\right\|_2^2\right) \tag{145}$$

$$= \frac{\eta}{2}\left\|G^t\right\|_2^2 + \frac{1}{2\eta}\left(\left\|\rho^t - \rho^*\right\|_2^2 - \left\|\rho^{t+1} - \rho^*\right\|_2^2\right), \tag{146}$$

we have

$$\text{Loss}_s^1(\rho^t) - \text{Loss}_s^1(\rho^*) \leq \tag{147}$$

$$\frac{\eta}{2}\left\|G^t\right\|_2^2 + \frac{1}{2\eta}\left(\left\|\rho^t - \rho^*\right\|_2^2 - \left\|\rho^{t+1} - \rho^*\right\|_2^2\right) - (\rho^t - \rho^*)^T\zeta_s^t - \frac{\mu_2}{2}\left\|\rho^t - \rho^*\right\|_2^2. \tag{148}$$

But now note that $\text{Loss}_s^1(\rho^t) - \text{Loss}_s^1(\rho^*) \geq \text{Loss}_s^1(\rho^t) - \text{Loss}_s^1(\rho^{t+1})$. Thus, combining Equation (148) and Lemma 20 yields

$$-\eta{\zeta_s^t}^T G^t \leq \frac{1}{2\eta}\left(\left\|\rho^t - \rho^*\right\|_2^2 - \left\|\rho^{t+1} - \rho^*\right\|_2^2\right) - (\rho^t - \rho^*)^T\zeta_s^t - \frac{\mu_2}{2}\left\|\rho^t - \rho^*\right\|_2^2. \tag{149}$$

By rearranging the terms, we then have

$$\left\|\rho^{t+1} - \rho^*\right\|_2^2 \leq (1 - \mu_2\eta)\left\|\rho^t - \rho^*\right\|_2^2 - \eta\left(\rho^{t+1} - \rho^*\right)^T\zeta_s^t \tag{150}$$

$$\leq (1 - \mu_2\eta)\left\|\rho^t - \rho^*\right\|_2^2 + \eta\left\|\rho^{t+1} - \rho^*\right\|_2\left\|\zeta_s^t\right\|_2. \tag{151}$$

Now note that $\eta \leq 1/L < 1/\mu_2$ and thus $0 < 1 - \mu_2\eta < 1$. We now define two sequences $u_t \triangleq \left\|\rho^t - \rho^*\right\|_2$ and $\delta_t = \eta\left\|\zeta_s^t\right\|_2$. We already know that $\delta_t \to 0$, and we want to show u_t also converges to 0. By Equation (151), we have

$$u_{t+1}^2 \leq (1 - \eta\mu_2)u_t^2 + \delta_t u_{t+1}, \tag{152}$$

which implies

$$\left(u_{t+1} - \frac{\delta_t}{2}\right)^2 = u_{t+1}^2 - u_{t+1}\delta_t + \frac{\delta_t^2}{4} \leq (1 - \eta\mu_2)u_t^2 + \frac{\delta_t^2}{4}, \tag{153}$$

and thus

$$u_{t+1} \leq \sqrt{(1 - \eta\mu_2)u_t^2 + \frac{\delta_t^2}{4}} + \frac{\delta_t}{2} \leq \sqrt{(1 - \eta\mu_2)u_t^2} + \frac{\delta_t}{2} + \frac{\delta_t}{2} \leq \left(1 - \frac{\eta\mu_2}{2}\right)u_t + \delta_t. \tag{154}$$

Lemma 23 allows to conclude. □

E.3.3 Reduction from model attack to converging gradient attack

Proof of Theorem 6. Lemma 24 already provided the convergence part of Theorem 6. We now move forward to proving the second part of the theorem, showing that for any $\varepsilon > 0$, there exists $\theta_s^\spadesuit \in \mathbb{R}^d$ such that $\left\|\rho^\infty - \rho^*(\theta_s^\spadesuit, \vec{\mathcal{D}}_{-s})\right\|_2 \leq \varepsilon$. We define

$$\text{Loss}_s^2(\rho, \theta_s) \triangleq \inf_{\vec{\theta}_{-s}}\left\{\text{Loss}(\rho, \vec{\theta}, \vec{\mathcal{D}}) - \mathcal{L}_s(\theta_s, \mathcal{D}_s)\right\} \tag{155}$$

$$= \text{Loss}_s^1(\rho) + \mathcal{R}(\rho, \theta_s) - \rho^T g_s^\infty, \tag{156}$$

and $\rho^*(\theta_s)$, its minimizer. Therefore, we have

$$\nabla_\rho \text{Loss}_s^2(\rho, \theta_s) = \nabla_\rho \text{Loss}_s^1(\rho) + \nabla_\rho \mathcal{R}(\rho, \theta_s) - g_s^\infty. \tag{157}$$

By Lemma 19, we know that LOSS_s^2 is locally strongly convex. Therefore, there exists $\mu_3 > 0$ such that $\text{LOSS}_s^2(\rho)$ is μ_3-strongly convex in $\{\rho | \left\|\rho - \rho^*(\theta_s)\right\|_2 \leq 1\}$. Therefore, for any $0 < \varepsilon < 1$, if $\left\|\rho^\infty - \rho^*(\theta_s)\right\|_2 > \varepsilon$, we then have

$$\varepsilon \left\|\nabla_\rho \text{LOSS}_s^2(\rho^\infty, \theta_s^\spadesuit)\right\|_2 \geq (\rho^\infty - \rho^*(\theta_s))^T \nabla_\rho \text{LOSS}_s^2(\rho^\infty, \theta_s^\spadesuit) \tag{158}$$

$$\geq \mu_3 \left\|\rho^\infty - \rho^*(\theta_s)\right\|_2^2 = \mu_3 \geq \mu_3 \varepsilon^2, \tag{159}$$

and thus $\left\|\nabla_\rho \text{LOSS}_s^2(\rho^\infty, \theta_s^\spadesuit)\right\|_2 \geq \mu_3 \varepsilon$.

Now since $g_s^\infty \in \text{GRAD}(\rho^\infty)$ there exists $\theta_s^\spadesuit \in \mathbb{R}^d$ such that[3] $\left\|\nabla_\rho \mathcal{R}(\rho^\infty, \theta_s^\spadesuit) - g_s^\infty\right\|_2 \leq \frac{\mu_3 \varepsilon}{2}$ which yields

$$\left\|\nabla_\rho \text{LOSS}_s^2(\rho^\infty, \theta_s^\spadesuit)\right\|_2 = \left\|\nabla_\rho \text{LOSS}_s^1(\rho^\infty) + \nabla_\rho \mathcal{R}(\rho^\infty, \theta_s^\spadesuit) - g_s^\infty\right\|_2 \tag{160}$$

$$= \left\|\nabla_\rho \mathcal{R}(\rho^\infty, \theta_s^\spadesuit) - g_s^\infty\right\|_2 \leq \frac{\mu_3 \varepsilon}{2}, \tag{161}$$

which is a contradiction. Therefore, we must have $\left\|\rho^\infty - \rho^*(\theta_s^\spadesuit, \vec{\mathcal{D}}_{-s})\right\|_2 \leq \varepsilon$. □

F CONVERGENCE OF CGA AGAINST ℓ_2^2

To write our proof, we define $\text{LOSS}_{-s}^\rho : \mathbb{R}^d \to \mathbb{R}$ by

$$\text{LOSS}_{-s}^\rho(\rho) \triangleq \inf_{\vec{\theta}} \left\{\text{LOSS}(\rho, \vec{\theta}, \vec{\mathcal{D}}) - \mathcal{L}_s(\theta_s, \mathcal{D}_s) - \mathcal{R}(\rho, \theta_s)\right\} \tag{162}$$

$$= \inf_{\vec{\theta}} \sum_{n \neq s} \mathcal{L}(\theta_n, \mathcal{D}_n) + \lambda \sum_{n \neq s} \left\|\rho - \theta_n\right\|_2^2. \tag{163}$$

In other words, it is the loss when local models are optimized, and when the data of strategic user s are removed.

Lemma 25. *Assuming ℓ_2^2 regularization and convex loss-per-input functions ℓ, for any datasets $\vec{\mathcal{D}}$,* LOSS *is strongly convex. As a result, so is* LOSS_{-s}^ρ.

Proof. Note that the global loss can be written as a sum of convex function, and of $\nu \sum \left\|\theta_n\right\|_2^2 + \left\|\rho - \theta_1\right\|_2^2$. Using tricks similar to the proof of Lemma 17, we see that the loss is strongly convex. The latter part of the lemma is then a straightforward application of Lemma 16. □

We now move on to the proof of Theorem 7. Note that our statement of the proof was not fully explicit, especially about the upper bound on the constant learning rate η. Here, we prove that it holds for $\eta_t = \eta \leq 1/3L$, where L is a constant such that LOSS_{-s}^ρ is L-smooth. The existence of L is guaranteed by Lemma 19.

Proof of Theorem 7. Note that by Lemma 15, LOSS_{-s}^ρ is convex, differentiable and L-smooth, and $\nabla \text{LOSS}_{-s}^\rho(\rho^t) = g_{-s}^{\dagger,t}$. For ℓ_2^2 regularization, we have $\text{GRAD}(\rho) = \mathbb{R}^d$ for all $\rho \in \mathbb{R}^d$. Then the minimum of equation 7 is zero, which is obtained when $g_s^t \triangleq \frac{\rho^t - \theta_s^\dagger}{\eta} - \hat{g}_{-s}^t = g_s^{t-1} + \frac{\rho^t - \theta_s^\dagger}{\eta} + \frac{\rho^t - \rho^{t-1}}{\eta}$. Note that

$$\rho^{t+1} = \rho^t - \eta g_{-s}^{\dagger,t} - \eta g_s^t \tag{164}$$

$$= \rho^t - \eta g_{-s}^{\dagger,t} - (\rho^t - \theta_s^\dagger) + (\rho^{t-1} - \rho^t) - \eta g_s^{t-1} \tag{165}$$

$$= \theta_s^\dagger - \eta_t (g_{-s}^{\dagger,t} + g_s^{t-1}) + \eta (g_{-s}^{\dagger,t-1} + g_s^{t-1}) \tag{166}$$

$$= \theta_s^\dagger - \eta (g_{-s}^{\dagger,t} - g_{-s}^{\dagger,t-1}). \tag{167}$$

Therefore, $\rho^{t+1} - \rho^t = \eta (g_{-s}^{\dagger,t} - g_{-s}^{\dagger,t-1}) - \eta (g_{-s}^{\dagger,t-1} - g_{-s}^{\dagger,t-2})$.

[3] In fact, if g_s^∞ belongs to the interior of $\text{GRAD}(\rho^\infty)$, we can guarantee $\nabla_\rho \mathcal{R}(\rho^\infty, \theta_s^\spadesuit) = g_s^\infty$.

Then, using the L-smoothness of Loss^{ρ}_{-s}, and denoting $u_t \triangleq \left\|\rho^{t+1} - \rho^t\right\|_2$, we have $u_{t+1} \leq L\eta_t u_t + L\eta_{t-1}u_{t-1}$. Now assume that $\eta \leq 1/3L$. Then $u_{t+1} \leq \frac{1}{3}(u_t + u_{t-1})$. We then know that $u_{t+2} \leq \frac{1}{3}(u_{t+1} + u_t) \leq \frac{1}{3}(\frac{1}{3}(u_t + u_{t-1}) + u_t) = \frac{4}{9}u_t + \frac{1}{9}u_{t-1}$.

Now define $v_t \triangleq u_t + u_{t-1}$. We then have $v_{t+2} \leq u_{t+2} + u_{t+1} \leq \frac{7}{9}u_t + \frac{4}{9}u_{t-1} \leq \frac{7}{9}(u_t + u_{t-1}) \leq \frac{7}{9}v_t$. By induction, we know that $v_t \leq (7/9)^{(t-1)/2}\max\{v_0, v_1\} \leq (\sqrt{7}/3)^t\left((\sqrt{7}/3)\max\{v_0, v_1\}\right)$. Thus, defining $\alpha \triangleq \sqrt{7}/3 < 1$, there exists $C > 0$ such that $u_t \leq v_t \leq C\alpha^t$. This implies that $\sum\left\|\rho^{t+1} - \rho^t\right\|_2 \leq \sum C\alpha^t < \infty$. Thus $\sum(\rho^{t+1} - \rho^t)$ converges, which implies the convergence of ρ^t to a limit ρ^∞. By L-smoothness, we know that $g^{\dagger,t}_{-s}$ must converge too. Taking equation 167 to the limit then implies $\rho^\infty = \theta^\dagger_s$. This shows that the strategic user achieves precisely what they want with CGA. It is thus optimal. □

G SINGLE DATA POISONING FOR LEAST SQUARE LINEAR REGRESSION

Proof of Theorem 8. We define the minimized loss with respect to ρ and without strategic user s by

$$\text{Loss}^*_{-s}(\rho, \vec{\mathcal{D}}_{-s}) \triangleq \min_{\vec{\theta}_{-s}\in\mathbb{R}^{d\times(N-1)}} \left\{\sum_{n\neq s}\mathcal{L}_n(\theta_n, \mathcal{D}_n) + \sum_{n\neq s}\lambda\left\|\theta_n - \rho\right\|_2^2\right\}. \tag{168}$$

Now consider a subgradient $g \in \nabla_\rho \text{Loss}^*_{-s}(\theta^\dagger_s, \vec{\mathcal{D}}_{-s})$ of the minimized loss at $\theta^\dagger_s$. For $x \triangleq \frac{-g}{2\lambda}$, then have $-g \in \nabla\left(\lambda\left\|x\right\|_2^2\right)$. We then define $\theta^\spadesuit_s \triangleq \theta^\dagger_s - x$.

$$0 = g - g \in \nabla_\rho \text{Loss}^*_{-s}(\theta^\dagger_s, \vec{\mathcal{D}}_{-s}) + \nabla_\rho\left(\lambda\left\|\theta^\spadesuit_s - \theta^\dagger_s\right\|_2^2\right) \tag{169}$$

$$= \nabla_\rho \text{Loss}_s(\theta^\dagger_s, \vec{\theta}^*_{-s}(\theta^\spadesuit_s, \vec{\mathcal{D}}_{-s}), \theta^\spadesuit_s, \vec{\mathcal{D}}_{-s}), \tag{170}$$

where Loss_s is defined by (4). Now consider the data point $(\mathcal{Q}, \mathcal{A}) = (g, g^T\theta^\spadesuit_s - 1)$. For $\mathcal{D}_s = \{(\mathcal{Q}, \mathcal{A})\}$, we then have $\nabla\mathcal{L}_s(\theta^\spadesuit_s, \mathcal{D}_s) = g$, which implies

$$\nabla_{\theta_s}\text{Loss}(\theta^\dagger_s, (\theta^\spadesuit_s, \vec{\theta}^*_{-s}(\theta^\spadesuit_s, \vec{\mathcal{D}}_{-s}), \vec{\mathcal{D}}) = 0. \tag{171}$$

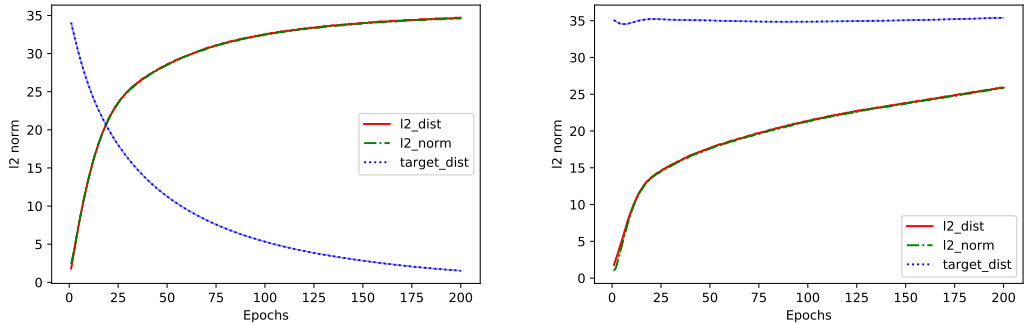
Combining it all together with the uniqueness of the solution then yields

$$\underset{(\rho,\vec{\theta})}{\arg\min}\left\{\text{Loss}(\rho, \vec{\theta}, \vec{\mathcal{D}})\right\} = \left(\theta^\dagger_s, \left(\theta^\spadesuit_s, \vec{\theta}^*_{-s}(\theta^\spadesuit_s, \vec{\mathcal{D}}_{-s})\right)\right), \tag{172}$$

which is what we wanted. □

H CGA ON MNIST

In this section, CGA is executed against 10 honest users, each one having 6,000 randomly and data points of MNIST, drawn randomly and independently. CGA is run by a strategic user whose target model $\theta^\dagger_s$ labels 0's as 1's, 1's as 2's, and so on, until 9's as 0's. We learn $\theta^\dagger_s$ by relabeling the MNIST training dataset and learning from the relabeled data. We use $\lambda = 1$, Adam optimizer and a decreasing learning rate.



(a) Using ℓ_2^2

(b) Using ℓ_2

Figure 16: Norm of global model, distance to initialisation and distance to target, under attack by CGA. In particular, we see that the attack against ℓ_2^2 is successful, as the distance between the global model and the target model goes to zero.

I DATA POISONING AGAINST LINEAR CLASSIFICATION

I.1 GENERATING EFFICIENT POISONING DATA

For every label $a \in \{1, \dots, 9\}$, we define $y_a \triangleq \theta_a^{\spadesuit} - \theta_0^{\spadesuit}$, and $c_a \triangleq -(\theta_{a0}^{\spadesuit} - \theta_{00}^{\spadesuit})$ (where $\theta_{a0}^{\spadesuit}$ is the bias of the linear classifier). The indifference subspace V is then the set of images $\mathcal{Q} \in \mathbb{R}^d$ such that $\mathcal{Q}^T y_a = c_a$ for all $a \in \{1, \dots, 9\}$.

To project any image $X \in \mathbb{R}^d$ on V, let us first construct an orthogonal basis of the vector space orthogonal to V, using the Gram-Schmidt algorithm. Namely, we first define $z_1 \triangleq y_1$. Then, for any answer $a \in \{1, \dots, 9\}$, we define

$$z_a \triangleq y_a - \sum_{b<a} y_a^T z_b \frac{z_b}{\|z_b\|_2^2}. \tag{173}$$

It is easy to check that for $b < a$, we have $z_a^T z_b = 0$. Moreover, if $\mathcal{Q} \in V$, then

$$z_a^T \mathcal{Q} = y_a^T \mathcal{Q} - \sum_{b<a} \frac{(y_a^T z_b)(z_b^T \mathcal{Q})}{\|z_b\|_2^2} = c_a - \sum_{b<a} \frac{(y_a^T z_b)(z_b^T \mathcal{Q})}{\|z_b\|_2^2}. \tag{174}$$

By induction, we see that $z_a^T \mathcal{Q}$ is a constant independent from $\mathcal{Q}$. Indeed, for $a = 1$, this is clear as $z_1^T \mathcal{Q} = y_1^T \mathcal{Q} = c_1$. Moreover, for $a > 1$, then, in the computation of $z_a^T \mathcal{Q}$, $\mathcal{Q}$ always appear as $z_b^T \mathcal{Q}$ for $b < a$. Moreover, denoting c'_a the constant such that $z_a^T \mathcal{Q} = c'_a$ for all $a \in \{1, \dots 9\}$, we see that these constants can be computed by

$$c'_a = c_a - \sum_{b<a} \frac{y_a^T z_b}{\|z_b\|_2^2} c'_b. \tag{175}$$

Finally, we can simply perform repeated projection onto the hyperplanes where a is equally probable as the answer 0. To do this, we first define the orthogonal projection $P(X, y, c)$ of $X \in \mathbb{R}^d$ on the hyperplane $x^T y = c$, which is given by

$$P(X, y, c) = X - (X^T y - c) \frac{y}{\|y\|_2^2}. \tag{176}$$

It is straightforward to verify that $P(X, y, c)^T y = c$ and that $P(P(X, y, c), y, c) = P(X, y, c)$. We then canonically define repeated projection by induction, as

$$P(X, (y_1, \dots, y_{k+1}), (c_1, \dots, c_{k+1})) \triangleq P(P(X, (y_1, \dots, y_k), (c_1, \dots, c_k)), y_{k+1}, c_{k+1}). \tag{177}$$

Now consider any image $X \in \mathbb{R}^d$. Its projection can be obtained by setting

$$\mathcal{Q} \triangleq P(X, (z_1, \dots, z_9), (c'_1, \dots c'_9)) + \xi. \tag{178}$$

Note that to avoid being exactly on the boundary, and thus retrieve information about the scales of $\theta^{\spadesuit}$ and on which side of the boundary favors which label, we add a small noise ξ, to make sure $\mathcal{Q}$ does not lie exactly on V (which would lead to multiple solutions for the learning), but small enough so that the probabilities of the different label remain close to 0.1 (the equiprobable probability).

We acknowledge that images obtained this way may not be in $[0,1]^d$, like the images of the MNIST dataset. In general, one could search for points $\mathcal{Q} \in V \cap [0,1]^d$. Note that in theory, by Theorem 2 (or a generalization of it), labeling random images in $[0,1]^d$ should suffice. However, in the case where $V \cap [0,1]^d$ is empty (typically if no image in $[0,1]$ is argued by model $\theta_s^{\spadesuit}$ as realistically a 9), this procedure may require the labeling of significantly more images to be successful.

I.2 A BRIEF THEORY OF DATA POISONING FOR LINEAR CLASSIFICATION

Using the efficient poisoning data fabrication, we thus have a set of images $(\mathcal{Q}, p(\mathcal{Q}))$, where $p_a(\mathcal{Q})$ is the probability assigned to image $\mathcal{Q}$ and label a. This defines the following local loss for the strategic node:

$$\mathcal{L}_s(\theta_s, \mathcal{D}_s) = \sum_{(\mathcal{Q}, p(\mathcal{Q})) \in \mathcal{D}_s} \sum_{a \in \{0,1,\ldots,9\}} p_a(\mathcal{Q}) \ln \sigma_a(\theta_s, \mathcal{Q}), \tag{179}$$

where $\sigma_a(\theta_s, \mathcal{Q}) = \frac{\exp(\theta_{sa}^T \mathcal{Q} + \theta_{sa0})}{\sum \exp(\theta_{sb}^T \mathcal{Q} + \theta_{sb0})}$ is the probability that image $\mathcal{Q}$ has label a, according to the model θ_s. We acknowledge that such labelings of queries is unusual. Evidently, in practice, an image may be labeled N times, and the number of labels N_a it received can be set to be approximately $N_a \approx N p_a(\mathcal{Q})$.

It is noteworthy that the gradient of the loss function is then given by

$$\left(\theta_s - \theta_s^{\spadesuit}\right)^T \nabla_{\theta_s} \mathcal{L}_s(\theta_s, \mathcal{D}_s) = \sum_{\mathcal{Q} \in \mathcal{D}_s} \sum_{a \in \{0,1,\ldots,9\}} \left(\sigma_a(\theta_s, \mathcal{Q}) - \sigma_a(\theta_s^{\spadesuit}, \mathcal{Q})\right) \left(\theta_{sa} - \theta_{sa}^{\spadesuit}\right)^T \mathcal{Q}^+, \tag{180}$$

where we defined $\mathcal{Q}^+ \triangleq (1, \mathcal{Q})$ (which allows to factor in the bias of the model. This shows that $\nabla_{\theta_s} \mathcal{L}_s(\theta_s, \mathcal{D}_s)$ points systematically away from $\theta_s^{\spadesuit}$, and thus that gradient descent will move towards $\theta_s^{\spadesuit}$.

In fact, if the set of images $\mathcal{Q}$ cover all dimensions (which occurs if there are $\Omega(d)$ images, which is the case for 2,000 images, since $d = 784$), then gradient descent will always move the model in the direction of $\theta_s^{\spadesuit}$, which will be the minimum. Moreover, by overweighting each data $(\mathcal{Q}, p(\mathcal{Q}))$ by a factor α (as though the image $\mathcal{Q}$ was labeled α times), we can guarantee PAC* learning, which means that we will have $\theta_s^* \approx \theta_s^{\spadesuit}$, even in the personalized federated learning framework. This shows why data poisoning should work in theory, with relatively few data injections.

Note that the number of other users does make learning harder. Indeed, the gradient of the regularization $\mathcal{R}(\rho, \theta_s)$ at $\rho = \theta_s^{\dagger}$ and $\theta_s = \theta_s^{\spadesuit}$ is equal to $2\lambda \left\|\theta_s^{\dagger} - \theta_s^{\spadesuit}\right\|_2$. As the number $N-1$ of other users grows, we should expect this distance to grow roughly proportionally to N. In order to make strategic user s robustly learn $\theta_s^{\spadesuit}$, the norm of the gradient of the local loss $\mathcal{L}_s$ at $\theta_s^{\dagger}$ must be vastly larger than $2\lambda \left\|\theta_s^{\dagger} - \theta_s^{\spadesuit}\right\|_2$. This means that the value of α (or, equivalently, the number of data injected in $\mathcal{D}_s$) must also grow proportionally to N.

I.3 INITIALIZATION OF THE LEARNING ALGORITHM

The convergence to the optimum is slow. But given that the problem is convex, we focus here mostly on showing that the minimum is indeed a poisoned model. To boost the convergence, we initialize our learning algorithm at a point close to what we expect to be the minimum, by taking this minimum and adding a Gaussian noise, and then we observe the convergence to this minimum.