BYZANTINE-ROBUST FEDERATED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study federated learning (FL) with adversarial clients, where the goal is to minimize the average loss of the honest (non-adversarial) clients without knowing their identity. Under heterogeneity, a single shared model parameter is statistically inappropriate: it cannot capture the distinct data-generating processes across clients, incurring an irreducible model-heterogeneity bias and severely limiting robustness to adversarial clients (a.k.a. Byzantine-robustness). We address this problem through representation learning, where each client learns a personalized linear head, while collaboratively estimating a shared nonlinear representation through Byzantine-robust aggregation. We demonstrate that the heterogeneity among honest representation gradients is only controlled by the representation error and statistical errors that decay either with the number of data samples per client (τ) or the number of iterations (T). In particular, our non-asymptotic parameter recovery bound reveals three terms: (i) an initialization-dependent error that goes away with T, (ii) finite-sample noise terms that decreases with τ and the number of honest clients, and (iii) a stochastic gradient variance term that also reduces with T. Importantly, with no irreducible model-heterogeneity bias in our bounds. We extend our regression analysis to multiclass classification, and empirically validate it on CIFAR-10, FEMNIST, and School Exam Score datasets.

1 INTRODUCTION

Learning from decentralized data requires accounting for heterogeneous clients while defending against adversarial participants (also referred to as *Byzantine* clients Lamport et al. (2019); Guerraoui et al. (2024a)). This is crucial for federated learning (FL), where clients keep their data locally and communicate only model updates to a central server (McMahan et al., 2017; Konečnỳ et al., 2016). Typically, the client models are non-identical, and thus a single common model parameter is statistically inappropriate. In addition, adversarial clients make the problem even harder by sharing arbitrary updates to corrupt the underlying learning process. Byzantine-robust FL addresses the latter through *robust aggregators* (Blanchard et al., 2017; Yin et al., 2018; El-Mhamdi et al., 2018; Small, 1990; Guerraoui et al., 2024b; Toso et al., 2026).

In particular, the class of (f, κ)-robust aggregators (Guerraoui et al., 2024b), where f is the number of adversarial clients and κ[1] the robustness coefficient, controls its deviation from the honest average through the heterogeneity of the honest clients' updates. Consequently, under a single common model parameter, the robust aggregation error is nonzero even under an infinite number of local data samples (Ghosh et al., 2019; Karimireddy et al., 2020a; Allouah et al., 2023a; Mishchenko et al., 2023; Toso et al., 2026). Indeed, this is the fundamental limitation of Byzantine-robust FL under heterogeneity: the server cannot tell the difference between honest heterogeneity and adversarial behavior when the honest clients are heterogeneous (Karimireddy et al., 2020a). Hence, the fundamental question: *Should we instead separate the common and client-specific components of each local model, collaborate robustly to learn what is shared and personalize what is specific?*

We address this question by leveraging a common nonlinear representation and client-specific linear heads. More precisely, clients can have different models while relying on a common representation of their covariates (features). Learning common feature representations have been proven to improve statistical efficiency in multitask and personalized federated learning (Tripuraneni et al., 2020; Du et al., 2020; Collins et al., 2021; Yang et al., 2023; Mishchenko et al., 2023; Zhang et al., 2024a;b).

[1] κ typically scales with f/n, with n being the total number of participating clients (Allouah et al., 2023a)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Common representations have also proved useful for learning across similar dynamical systems (Zhang et al., 2023; 2024a; Lee et al., 2025; Fallah et al., 2025). The intuition is that, although the representation gradients depend on the client-specific heads, under a realizability assumption, each composition of the shared representation and a personalized head can fit the underlying client data distribution. Therefore, after fitting the local heads, the remaining representation-gradient heterogeneity is governed only by representation-recovery and statistical errors, rather than by a persistent model-parameter heterogeneity bias arising from fitting a single common model across clients.

Motivated by this, we consider alternating optimization to (i) fit each personalized head with clients' specific data samples and to (ii) robustly aggregate the representation updates of all participating clients (honest and adversarial). We then bound the honest representation gradient heterogeneity and use it to establish the convergence and parameter recovery guarantees. In the end, our bounds comprise an initialization-dependent term that decreases with number of training rounds T, finite-sample noise term that scales inversely with the number of data samples per client τ, and stochastic gradient term that scales inversely with $\sqrt{T}$. We first provide the results for nonlinear regression and then instantiate it to multiclass classification with squared loss.

Informal main result. Let $\mathcal{H}$ denote the set of honest clients and let T be the number of communication rounds. Suppose that the number of samples per client satisfies $\tau \geq \tau_{\mathsf{burn\text{-}in}} > 0$,with $\tau_{\mathsf{burn\text{-}in}}$ specified in Section 2.3 (equation 2), and let the robust aggregation coefficient $\kappa = \mathcal{O}(f/n)$ satisfy $\kappa \leq \underline{\kappa}$ with $\underline{\kappa}$ defined in Section 2.3 (equation 4). Then, up to logarithmic factors, with high probability, our main result (Theorem 1) proves

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbb{E}\nabla_{\mathbf{B}} L_{\mathcal{H}}\left(\{\mathbf{h}_i^{(t+1)}\}_{i\in\mathcal{H}}, \mathbf{B}^{(t)}\right)\right\|^2 \lesssim \underbrace{\frac{\Delta_0}{\sqrt{T}}}_{\text{initialization}} + \underbrace{\kappa\left(\frac{\mathsf{d}_{\mathsf{rep}}}{\tau|\mathcal{H}|}+\frac{\mathsf{d}_{\mathsf{head}}}{\tau}\right)\sigma^2}_{\text{finite samples}} + \underbrace{\left(\kappa+\kappa^2+\frac{1+\kappa}{|\mathcal{H}|}\right)\frac{\sigma_{\mathsf{g}}^2}{\sqrt{T}}}_{\text{stochastic gradients}},$$

where $L_{\mathcal{H}}(\cdot)$ is the empirical honest averaged loss and σ^2, σ_{g}^2 are the label noise and stochastic gradient variances, respectively. In addition, Δ_0 denotes the initialization-dependent error and $\mathsf{d}_{\mathsf{rep}}$ and $\mathsf{d}_{\mathsf{head}}$ are the underlying representation ($\mathbf{B}^{(t)}$) and client-specific head ($\mathbf{h}_i^{(t+1)}, \forall i \in \mathcal{H}$) dimensions, respectively. We note that the first and last terms decay with the number of rounds, while the finite-sample terms decreases with the local sample size. Importantly, there is no persistent model-heterogeneity term, demonstrating the benefit of collaboratively learning what is shared and personalizing what is specific under adversarial attacks. This rigorously complements the benefit of partial personalization pointed out in Mishchenko et al. (2023)[2] by carefully providing a high-probability, finite-sample guarantee for a common nonlinear representation under arbitrary adversarial updates.

Contributions. Our main contributions are summarized as follows:

• We propose adversarially robust nonlinear representation learning for heterogeneous FL. Each honest client learns a personalized head, while collaborating robustly without knowing honest ones' identity to learn a common nonlinear feature representation.

• We prove that the representation-gradient heterogeneity is controlled only by the representation-recovery error, finite-sample estimation error, and the stochastic-gradient variance (Lemma 1). Therefore, our guarantee is not affected by any model and data heterogeneity bias that hurts traditional Byzantine-robust FL (Karimireddy et al., 2020a; Allouah et al., 2023a).

• We establish a high-probability ergodic convergence guarantee for the common representation under adversarial attacks (Theorem 1). We then derive the non-asymptotic guarantees for recovering the underlying representation parameter (Theorem 3) and (Corollary 1). Our bounds characterize the dependence on the number of local samples per client, the number of honest clients, the stochastic-gradient noise, and the robustness coefficient of the aggregation rule.

• We go beyond regression and extend our parameter recovery guarantee to multiclass classification. We demonstrate that under a class-probability margin condition, the classification error inherits the nonlinear prediction-error rate from our regression analysis (Corollary 2).

Notation. For a positive integer n, we write $[n] := \{1, \ldots, n\}$. The Euclidean norm for vectors and spectral norm for matrices are denoted by $\|\cdot\|$, while $\|\cdot\|_F$ denotes the Frobenius norm. For a matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalues, and $a \vee b := \max\{a, b\}$. We

[2]We position our paper within the literature throughout and include a related work section in Appendix B.

write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ for conditional expectation with respect to the filtration $\mathcal{F}_t$. We also write $h \lesssim g$ when $h \leq Cg$ for a constant $C > 0$. We write $a \asymp b$ if there exist constants $C_1, C_2 > 0$ such that $C_1 b \leq a \leq C_2 b$. For a real-valued random variable X, its sub-Gaussian norm (Vershynin, 2019, Definition 2.6.4) is defined as follows:

$$\|X\|_{\psi_2} := \inf \left\{ t > 0 : \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2 \right\}.$$

The random variable X is said to be sub-Gaussian if $\|X\|_{\psi_2} < \infty$. For a random vector $Z \in \mathbb{R}^d$ and a random matrix $M \in \mathbb{R}^{p \times q}$, we define

$$\|Z\|_{\psi_2} := \sup_{u \in \mathbb{S}^{d-1}} \|\langle u, Z \rangle\|_{\psi_2} \text{ and } \|M\|_{\psi_2} := \sup_{U \in \mathbb{S}_F^{p \times q}} \|\langle U, M \rangle_F\|_{\psi_2}.$$

Z and M are sub-Gaussian when its corresponding ψ_2-norm is finite.

2 Problem Formulation

We first describe the adversarial FL setup, including the (f, κ)-robust aggregation definition, and then introduce our nonlinear representation learning problem with client-specific heads and a shared representation. Our adversarial setup follows the standard Byzantine-robust FL framework considered in Karimireddy et al. (2020a); Farhadkhani et al. (2022); Guerraoui et al. (2024b).

2.1 Adversarially Robust FL

We consider a federated system with n clients and a central server. An unknown subset $\mathcal{B} \subset [n]$ of cardinality $|\mathcal{B}| = f < n/2$ is adversarial and may share arbitrary updates with the server with the intention of corrupting the underlying learning process (Guerraoui et al., 2024b). The other clients form the honest set $\mathcal{H} = [n] \setminus \mathcal{B}$, with $|\mathcal{H}| = n - f$.

Data. At every training round t, each honest client $i \in \mathcal{H}$ draws two independent batches of τ i.i.d. samples from a client-specific distribution that we denote by D_i on $\mathbb{R}^d \times \mathbb{R}^q$. The first batch is used to estimate the personalized head and the second to evaluate the representation loss and gradient (this is also referred to as "debiasing" in (Zhang et al., 2024a)). Both batches of data are independent of the history before the current round and independent across honest clients $i \in \mathcal{H}$. To avoid cumbersome notation, we suppress the round and batch superscripts whenever the role of the samples is clear.

The server aggregates client updates using a robust aggregation rule. Given updates $u_1, \ldots, u_n \in \mathbb{R}^p$, define the honest average $\bar{u}_{\mathcal{H}} := \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} u_i$. We next define the class of (f, κ)-robust aggregators (Allouah et al., 2023a; Guerraoui et al., 2024b).

Definition 1 ((f, κ)-robust aggregator)**.** *An aggregation rule $\mathsf{F} : (\mathbb{R}^p)^n \to \mathbb{R}^p$ is (f, κ)-robust if, for every collection of inputs and every honest set $\mathcal{H}$ with $|\mathcal{H}| = n - f$,*

$$\|\mathsf{F}(u_1, \ldots, u_n) - \bar{u}_{\mathcal{H}}\|^2 \leq \frac{\kappa}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|u_i - \bar{u}_{\mathcal{H}}\|^2.$$

We also emphasize that the (f, κ)-robustness condition is satisfied by many aggregation rules, including Krum (Blanchard et al., 2017), geometric median (Small, 1990; Acharya et al., 2022), coordinate-wise median and trimmed mean (Yin et al., 2018), and minimum-diameter averaging (El-Mhamdi et al., 2021). We note that κ typically scales as f/n, since standard robust aggregation rules can be combined with the nearest-neighbor mixing (NNM) preprocessing step of Allouah et al. (2023a) to achieve this information-theoretically optimal scaling. Therefore, the effect of adversarial clients is determined by κ and the heterogeneity of the honest updates. Our analysis then controls this heterogeneity using a shared representation across clients as discussed below.

2.2 Nonlinear Representation Learning and Personalization

We consider that the observations of every honest client $i \in \mathcal{H}$ are generated according to

$$Y_{i,k} = \mathbf{h}_i^\star \phi_{\mathbf{B}^\star}(X_{i,k}) + V_{i,k}, \text{ for all } k = 1, \ldots, \tau,$$

where $\mathbf{B}^\star \in \mathbb{R}^p$ parameterizes a shared nonlinear representation $\phi_{\mathbf{B}^\star} : \mathbb{R}^d \to \mathbb{R}^r$, $\mathbf{h}_i^\star \in \mathbb{R}^{q \times r}$ is a client-specific linear head, and $V_{i,k} \in \mathbb{R}^q$ is label noise. The client-specific heads capture heterogeneity, while $\mathbf{B}^\star$ represents what is shared across clients. We assume that the covariates are uniformly bounded, i.e., $\|X_{i,k}\| \leq R$, for every honest clients $i \in \mathcal{H}$ and sample $k \in [\tau]$.

For any candidate head $\mathbf{h}$ and representation parameter $\mathbf{B}$, we define the client-specific and honest-average empirical losses on the current data batch as follows:

$$L_i(\mathbf{h}, \mathbf{B}) := \frac{1}{\tau} \sum_{k=1}^{\tau} \|Y_{i,k} - \mathbf{h}\phi_{\mathbf{B}}(X_{i,k})\|^2 \text{ and } L_{\mathcal{H}}(\{\mathbf{h}_i\}_{i\in\mathcal{H}}, \mathbf{B}) := \frac{1}{|\mathcal{H}|} \sum_{i\in\mathcal{H}} L_i(\mathbf{h}_i, \mathbf{B}).$$

Let $r_i(X_{i,k}; \mathbf{h}, \mathbf{B}) := \mathbf{h}\phi_{\mathbf{B}}(X_{i,k}) - Y_{i,k}$ and $J_{\mathbf{B}}(X) := \nabla_{\mathbf{B}}\phi_{\mathbf{B}}(X) \in \mathbb{R}^{r\times p}$, and thus

$$\nabla_{\mathbf{B}} L_i(\mathbf{h}, \mathbf{B}) = \frac{2}{\tau} \sum_{k=1}^{\tau} J_{\mathbf{B}}^{\top}(X_{i,k}) \mathbf{h}^{\top} r_i(X_{i,k}; \mathbf{h}, \mathbf{B}).$$

Goal. We aim to learn the shared representation $\mathbf{B}$ and the personalized heads $\{\mathbf{h}_i\}_{i\in\mathcal{H}}$ by minimizing $L_{\mathcal{H}}(\cdot)$ despite not knowing $\mathcal{H}$ operating under adversarial attacks.

To do so, we leverage alternating minimization Collins et al. (2021); Zhang et al. (2024a). At round t, every client estimates its local head on the head-fitting batch and sends a stochastic representation-gradient update evaluated on the independent representation batch. The server then robustly aggregates the representation-gradient momenta. More precisely, we have

$$\textbf{Head estimation.} \quad \mathbf{h}_i^{(t+1)} \in \arg\min_{\mathbf{h}} L_i(\mathbf{h}, \mathbf{B}^{(t)}),$$

$$\textbf{Representation update.} \quad \mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} - \eta \mathsf{F}\left(m_{\mathbf{B},1}^{(t)}, \ldots, m_{\mathbf{B},n}^{(t)}\right), \tag{1}$$

where $\{m_{\mathbf{B},i}\}_{i\in[n]}$ denotes the representation-gradient momentum given by $m_{\mathbf{B},i}^{(t)} := \beta m_{\mathbf{B},i}^{(t-1)} + (1-\beta)\widehat{\nabla}_{\mathbf{B}} L_i^{(t)}(\mathbf{h}^{(t+1)}, \mathbf{B}^{(t)})$, for some momentum coefficient $\beta \in (0,1)$. We also assume that $m_{\mathbf{B},i}^{(0)} = \mathbf{0}$. In addition, here $\widehat{\nabla}_{\mathbf{B}} L_i^{(t)}(\cdot)$ denotes the stochastic representation gradient computed by client i at round t, and $\widehat{\nabla}_{\mathbf{B}} L_{\mathcal{H}}^{(t)}(\cdot) := \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \widehat{\nabla}_{\mathbf{B}} L_i^{(t)}(\cdot)$ is its average over the honest clients.

The head update adapts the predictor to each client's local model. We assume that this update is solved exactly, corresponding to a sufficiently large number of local optimization steps. This allows us to isolate the effect of learning a common representation on the convergence of adversarially robust FL. Extending our analysis to account for local head estimation error is left for future work.

On the other hand, the representation update keeps only what is common across clients. We reemphasize here that the independent batches (for local heads and representation) here are an artifact that separates the error incurred when fitting the head from the noise in the representation update. In practice, the same local dataset can be partitioned or resampled (Zhang et al., 2024a).

As discussed above, for an (f,κ)-robust aggregator, the effect of adversarial clients is controlled by the heterogeneity among the honest updates. As the server aggregates representation-gradient momenta, our analysis proceeds in two steps. We first bound the heterogeneity among the honest representation gradients. We then invoke Lemmas 12 and 13, deferred to the appendix, to control the additional error introduced by stochastic gradients and momentum. Therefore, for later use, we define the representation gradient heterogeneity per iteration and averaged over rounds as follows:

$$\bar{G}^{(t)} := \frac{1}{|\mathcal{H}|} \sum_{i\in\mathcal{H}} \left\| \nabla_{\mathbf{B}} L_i(\mathbf{h}_i^{(t+1)}, \mathbf{B}^{(t)}) - \nabla_{\mathbf{B}} L_{\mathcal{H}}(\{\mathbf{h}_j^{(t+1)}\}_{j\in\mathcal{H}}, \mathbf{B}^{(t)}) \right\|^2 \text{ and } \bar{G}_T := \frac{1}{T}\sum_{t=0}^{T-1} \bar{G}^{(t)}.$$

2.3 Assumptions

Let us here collect the assumptions used throughout the analysis.

Assumption 1 (Sub-Gaussian noise)**.** *The label noise $V_{i,k}$ is mean-zero and σ^2-sub-Gaussian. In particular, for every honest client $i \in \mathcal{H}$, sample $k \in [\tau]$, unit vector $u \in \mathbb{S}^{q-1}$, and $\lambda \in \mathbb{R}$, we have*

$$\mathbb{E}[V_{i,k} \mid X_{i,k}] = 0 \text{ and } \mathbb{E}\left[\exp(\lambda u^{\top} V_{i,k}) \mid X_{i,k}\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Definition 2. *For $A \in \mathbb{R}^{m\times n}$, define $\|A\|_{\infty\to 2} := \sup_{\|z\|_\infty \le 1} \|Az\|_2$.*

Assumption 2. *There exist constants $\bar{\phi}, \bar{J}, H > 0$ such that $\|\phi_{\mathbf{B}}(X)\| \le \bar{\phi}$, $\|J_{\mathbf{B}}(X)\|_{\infty\to 2} \le \bar{J}$, $\|\mathbf{h}_i^{(t+1)}\| \le H$, and $\|\mathbf{h}_i^{\star}\| \le H$ for all honest clients, covariates, and iterates. Moreover, the*

honest-average loss is L_1-smooth with respect to $\mathbf{B}$ in a neighborhood of $\mathbf{B}^\star$, and every client loss is L_2-smooth with respect to its client-specific head. In addition, the population representation gradient is L_3-Lipschitz in the heads, i.e., for any $\mathbf{B}$ and head collections $\{\mathbf{h}_i\}_i, \{\mathbf{h}'_i\}_i$,

$$\|\nabla_{\mathbf{B}} \mathbb{E}_X L_{\mathcal{H}}(\{\mathbf{h}_i\}_i, \mathbf{B}) - \nabla_{\mathbf{B}} \mathbb{E}_X L_{\mathcal{H}}(\{\mathbf{h}'_i\}_i, \mathbf{B})\| \le \frac{L_3}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{h}_i - \mathbf{h}'_i\|.$$

On the other hand, for the Jacobian, there exists $L_4 > 0$ such that $\|J_{\mathbf{B}}(X) - J_{\mathbf{B}'}(X)\|_{2\to 2} \le L_4 \|\mathbf{B} - \mathbf{B}'\|$ for all X and all $\mathbf{B}, \mathbf{B}'$ satisfying $\|\mathbf{B} - \mathbf{B}^\star\|, \|\mathbf{B}' - \mathbf{B}^\star\| \le \rho_0$, for some $\rho_0 > 0$.

For each honest client, we define the empirical and population feature covariances as follows:

$$\widehat{\Sigma}_{i,\phi}(\mathbf{B}) := \frac{1}{\tau} \sum_{k=1}^{\tau} \phi_{\mathbf{B}}(X_{i,k}) \phi_{\mathbf{B}}(X_{i,k})^\top \text{ and } \Sigma_{i,\phi}(\mathbf{B}) := \mathbb{E}_{X \sim \mathsf{D}_i}[\phi_{\mathbf{B}}(X) \phi_{\mathbf{B}}(X)^\top],$$

and let $\mathcal{N}(\mathbf{B}^\star, \rho_0) := \{\mathbf{B} \in \mathbb{R}^p : \|\mathbf{B} - \mathbf{B}^\star\| \le \rho_0\}$ be the ball of radius ρ_0 centered at $\mathbf{B}^\star$.

Assumption 3. *There exists $\mu_1 > 0$ such that $\lambda_{\min}(\Sigma_{i,\phi}(\mathbf{B})) \ge \mu_1$ for every honest client i and every $\mathbf{B} \in \mathcal{N}(\mathbf{B}^\star, \rho_0)$. In addition, the honest-client population provides positive curvature for identifying the shared representation, i.e., we assume that*

$$\mu_3 := \lambda_{\min}\left(\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbb{E}_X \left[J_{\mathbf{B}^\star}^\top(X) \mathbf{h}_i^{\star\top} \mathbf{h}_i^\star J_{\mathbf{B}^\star}(X) \right] \right) > 0.$$

Assumption 4. *For $\nu_i^{(t)} := \widehat{\nabla}_{\mathbf{B}} L_i^{(t)}(\mathbf{h}_i^{(t+1)}, \mathbf{B}^{(t)}) - \nabla_{\mathbf{B}} L_i(\mathbf{h}_i^{(t+1)}, \mathbf{B}^{(t)})$, conditioned on $\mathcal{F}_t$, the vectors $\{\nu_i^{(t)}\}_{i \in \mathcal{H}}$ are independent, mean-zero, and satisfy $\|\nu_i^{(t)}\|_{\psi_2 | \mathcal{F}_t} \le \sigma_{\mathsf{g}}$ almost surely. Equivalently, for every honest client and iteration, we have*

$$\mathbb{E}[\widehat{\nabla}_{\mathbf{B}} L_i^{(t)}(\mathbf{h}_i^{(t+1)}, \mathbf{B}^{(t)})] = \nabla_{\mathbf{B}} L_i(\mathbf{h}_i^{(t+1)}, \mathbf{B}^{(t)}).$$

We emphasize that these assumptions are standard conditions used in non-asymptotic analyses of representation learning and Byzantine-robust stochastic optimization. In particular, bounded features and Jacobians, local smoothness, and nondegenerate feature covariance provide the regularity and local identifiability required in our nonlinear setting. Similar covariance, noise, and regularity conditions appear in analyses of linear and nonlinear representation learning (Collins et al., 2021; Zhang et al., 2024a;b). Our client-diversity condition in Assumption 3 is the nonlinear analogue of the task-diversity conditions used to identify a shared representation across heterogeneous tasks (Tripuraneni et al., 2020; Du et al., 2020; Collins et al., 2021). Finally, unbiased stochastic gradients with bounded conditional variance are standard in Byzantine-robust stochastic optimization (Karimireddy et al., 2020a; Farhadkhani et al., 2022; Allouah et al., 2023a).

For the reminder of the paper it is taken for granted that Assumptions 1–4 hold.

3 Theoretical Guarantees

We are now ready to state our main results. The analysis proceeds in three steps. Lemma 1 controls the heterogeneity that drives the robust-aggregation error, Theorem 1 establishes convergence of the population representation gradient, and Corollary 1 converts this guarantee into the parameter recovery error bound. We fix a small probability of failure $\delta \in (0, 1)$ and present our results in a probabilistic manner. We first collect the quantities used throughout.

$$\Delta_{\mathbf{B},T+1} := \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{B}^{(t)} - \mathbf{B}^\star\|^2, \quad \bar{Q}_T := \frac{1}{T} \sum_{t=0}^{T-1} \left\| \mathbb{E}_X \nabla_{\mathbf{B}} L_{\mathcal{H}} \left(\{\mathbf{h}_i^{(t+1)}\}_{i \in \mathcal{H}}, \mathbf{B}^{(t)} \right) \right\|^2,$$

$$\bar{E}_T := \frac{1}{T|\mathcal{H}|\tau} \sum_{t=0}^{T-1} \sum_{i \in \mathcal{H}} \sum_{k=1}^{\tau} \left\| \mathbf{h}_i^{(t+1)} \phi_{\mathbf{B}^{(t)}}(X_{i,k}) - \mathbf{h}_i^\star \phi_{\mathbf{B}^\star}(X_{i,k}) \right\|^2.$$

Here, $\Delta_{\mathbf{B},T+1}$ and $\bar{E}_T$ denote the average representation-recovery and prediction-recovery errors, respectively, whereas $\bar{Q}_T$ denotes the squared norm of the population honest-average representation gradient, averaged over rounds $t \in \{0, 1, \dots, T-1\}$. We also define

Non-asymptotic learning rates. $\mathfrak{s}_{\mathcal{H},\tau} := \dfrac{\mathsf{d}_{\mathsf{rep}} + \log(|\mathcal{H}|T/\delta)}{\tau|\mathcal{H}|}$ and $\mathfrak{s}_\tau := \dfrac{\mathsf{d}_{\mathsf{head}} + \log(|\mathcal{H}|T/\delta)}{\tau}$.

Initialization error. $\Delta_0 := L_{\mathcal{H}}(\{\mathbf{h}_i^{(1)}\}_i, \mathbf{B}^{(0)}) - L_{\mathcal{H}}(\{\mathbf{h}_i^{\star}\}_i, \mathbf{B}^{\star}) + \frac{1}{L_1}\left\|\nabla_{\mathbf{B}} L_{\mathcal{H}}(\{\mathbf{h}_i^{(1)}\}_i, \mathbf{B}^{(0)})\right\|^2 + (1+\kappa)\Delta_{\mathbf{B}}^{(0)}$,

where $\mathsf{d}_{\mathsf{rep}} = p$ and $\mathsf{d}_{\mathsf{head}} = p \vee rq$. We also define $p_\sigma := (H\bar{\phi}+\sigma)^2+\sigma^2$ and $\Delta_{\mathbf{B}}^{(0)} := \|\mathbf{B}^{(0)}-\mathbf{B}^{\star}\|^2$. The rate $\mathfrak{s}_{\mathcal{H},\tau}$ will capture the complexity of learned the common representation with finite number of data samples, while $\mathfrak{s}_\tau$ will capture the local head error when fitting the learned representation. Throughout this section, Assumptions 1-4 hold, the server uses an (f,κ)-robust aggregator with $f < n/2$, and the local head update is the empirical risk minimizer described (i.e., in equation 1).

Our analysis is local, we assume that the initial representation is sufficiently close to $\mathbf{B}^{\star}$ and that all iterates remain in $\mathcal{N}(\mathbf{B}^{\star}, \rho)$, where $\rho := \min\{\rho_0, \mu_3/\bar{p}\}$ and $\bar{p} := \frac{3}{2}\bar{J}H^2L_4 + \frac{1}{2}H^2L_4^2\rho_0$, and that $\mu_3 \geq 4L_3H\bar{J}\bar{\phi}/\mu_1$. Finally, for a constant $C > 0$, define the burn-in sample size

$$\tau_{\mathsf{burn\text{-}in}} := \max\left\{\frac{16C^2\bar{\phi}^4}{\mu_1^2}, \frac{4C\bar{\phi}^2}{\mu_1}\right\}\log(2r|\mathcal{H}|T/\delta), \tag{2}$$

and assume $\tau \geq \tau_{\mathsf{burn\text{-}in}}$. This guarantees that the empirical feature covariances from the T rounds are well-conditioned. Below, we use $\lesssim$ to omit constants depending only on the fixed problem parameters $(\bar{J}, H, \bar{\phi}, \mu_1, \mu_3, L_{1:4})$. In addition, for conciseness, we define

$$\bar{\eta} := \min\left\{\frac{1}{24L_1}, \frac{\mu_3}{8\bar{J}^2H^2q\left(2H^2\bar{J}^2+16H^2\bar{J}^2\bar{\phi}^4/\mu_1^2\right)}, \frac{1}{32\bar{J}^2H^2q\bar{\phi}^2\mu_3}, \frac{1}{2\mu_3 q}, \frac{8}{\mu_3}\right\}, \tag{3}$$

and for any $\eta \leq \bar{\eta}$, we set $\beta^2 = 1 - 24\eta L_1$ and define $\bar{a} := \beta^2(1+\eta L_1)(1+4\eta L_1)$ with $\bar{b} := 4\eta L_1(1+\eta L_1)\beta^2$. We also note that $1-\bar{a} \asymp \eta L_1$ and $\bar{b}/(1-\bar{a}) \asymp 1$. In addition, the robust coefficient $\kappa = \mathcal{O}(f/n)$, with $f/n < 1/2$, is assumed to satisfy

$$\kappa \leq \underline{\kappa} := C_\kappa\left[\bar{J}^4qH^4\left(1+\frac{\bar{\phi}^4}{\mu_1^2}\right)\left(\frac{1}{\mu_3}+\frac{1}{\mu_3^2}\right)\right]^{-1}, \text{ for some } C_\kappa > 0. \tag{4}$$

We note that for any (f,κ)-robust aggregator when combined with nearest-neighbor mixing (NNM) (Allouah et al., 2023a), its robustness coefficient satisfies $\kappa = \mathcal{O}(f/n)$, which is information-theoretically optimal (Allouah et al., 2023a). Therefore, the condition in equation 4 is satisfied when f/n is sufficiently small. We also emphasize that conditions on κ are also required in Karimireddy et al. (2020a); Allouah et al. (2023b).

Lemma 1 (Gradient Heterogeneity Bound)**.** *Suppose that $\eta \leq \bar{\eta}$, $\tau \geq \tau_{\mathsf{burn\text{-}in}}$, and $\kappa \leq \underline{\kappa}$. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$\bar{G}_T \lesssim \frac{\Delta_{\mathbf{B}}^{(0)}}{\eta T} + p_\sigma\mathfrak{s}_{\mathcal{H},\tau} + \sigma^2\mathfrak{s}_\tau + \eta\left(1+\frac{1}{|\mathcal{H}|}\right)\sigma_{\mathsf{g}}^2\log(T/\delta) + \bar{Q}_T. \tag{5}$$

Discussion: The bound separates initialization, finite-sample, and stochastic-gradient errors. In particular, the heterogeneity of the honest representation gradients decreases with representation learning through $\Delta_{\mathbf{B}}^{(0)}/(\eta T)$ and $\bar{Q}_T$. That is, it is not controlled by a model parameter heterogeneity bias as in typical adversarially robust FL (Allouah et al., 2023a). Thus, after fitting the personalized heads, honest clients agree on the direction used to learn the common representation up to errors that go away with more communication rounds T or data samples τ. The proof is in Appendix E.

Importantly, our analysis does not require a uniform gradient-dissimilarity condition, such as the G- or (G,B)-dissimilarity assumptions commonly used in heterogeneous FL (Li et al., 2020; Karimireddy et al., 2020b; Allouah et al., 2023b; Gorbunov et al., 2022). Instead, building on the analysis of the gradient heterogeneity dynamics in adversarially robust FL from Toso et al. (2026), we directly control the trajectory of $\bar{G}_T$ in this more intricate setting of nonlinear representation learning.

Theorem 1 (Convergence Bound)**.** *Suppose that the conditions of Lemma 1 hold. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$\bar{Q}_T \lesssim \frac{\Delta_0}{\eta T} + \kappa p_\sigma\,\mathfrak{s}_{\mathcal{H},\tau} + \kappa\sigma^2\,\mathfrak{s}_\tau + \eta\left(\kappa+\kappa^2+\frac{1+\kappa}{|\mathcal{H}|}\right)\sigma_{\mathsf{g}}^2\log(T/\delta). \tag{6}$$

In addition, suppose that $\eta = \min\{\bar{\eta}, 1/\sqrt{T}\}$, then it holds that

$$\bar{Q}_T \lesssim \frac{\Delta_0}{\sqrt{T}} + \kappa p_\sigma\,\mathfrak{s}_{\mathcal{H},\tau} + \kappa\sigma^2\,\mathfrak{s}_\tau + \left(\kappa+\kappa^2+\frac{1+\kappa}{|\mathcal{H}|}\right)\frac{\sigma_{\mathsf{g}}^2\log(T/\delta)}{\sqrt{T}}. \tag{7}$$

Discussion: In our ergodic convergence bound, the optimization error and stochastic-gradient term both decrease as $1/\sqrt{T}$. The statistical terms vanish with τ and are inflated by κ. Therefore, under equation 4, adversarial clients affect the rate through aggregation robustness coefficient κ that typically scales as f/n, but most importantly, it does not introduce an irreducible model parameter heterogeneity bias, therefore demonstrating the benefit of learning what is common and personalizing what is specific. We provide the proof in Appendix F.

Key takeaway: Partial personalization is known to mitigate heterogeneity in federated optimization (Mishchenko et al., 2023). We rigorously characterize this benefit in adversarially robust FL with nonlinear representation learning by establishing high-probability, non-asymptotic guarantees under arbitrary adversarial updates. By separating the client-specific heads from the common representation, we demonstrate that all remaining errors decrease with the number of rounds T, the number of local samples τ, or the number of honest clients $|\mathcal{H}|$. We emphasize that to the best of our knowledge this is the first time such analysis is provided for adversarially robust FL. In contrast to single-model Byzantine-robust FL, our guarantees have no irreducible model parameter heterogeneity bias.

Corollary 1 (Parameter Recovery Error Bound)**.** *Suppose the conditions of Theorem 1. Then, for every $\delta \in (0,1)$, with probability at least $1-2\delta$, it holds that*

$$\bar{E}_T \lesssim \frac{\Delta_0}{\sqrt{T}} + (1+\kappa+\kappa^2)\left(p_\sigma\, \mathfrak{s}_{\mathcal{H},\tau} + \sigma^2\, \mathfrak{s}_\tau\right) + \left(\kappa + \kappa^2 + \frac{1+\kappa}{|\mathcal{H}|}\right) \frac{\sigma_{\mathrm{g}}^2 \log(T/\delta)}{\sqrt{T}}. \tag{8}$$

Discussion: With η of order $1/\sqrt{T}$, the initialization and stochastic-gradient terms vanish at the standard ergodic rate (Allouah et al., 2023a). When $\kappa = 0$, corresponding to the absence of adversarial clients, the second term in the bound above reduces to $p_\sigma \mathfrak{s}_{\mathcal{H},\tau} + \sigma^2 \mathfrak{s}_\tau$. This is consistent with the statistical recovery guarantees for nonlinear representation learning in multitask settings established in Zhang et al. (2024b): $\mathfrak{s}_{\mathcal{H},\tau}$ captures the complexity of collaboratively learning the common representation, whereas $\mathfrak{s}_\tau$ captures the client-specific complexity of fitting the personalized heads on that representation. Our setting is more intricate as it additionally accounts for arbitrary adversarial clients and stochastic updates aggregated. We defer the proof to Appendix G.

4 Multiclass classification

We now demonstrate that the nonlinear regression guarantee can be transferred to classification. For this, we consider a q-class problem. Let $C_{i,k} \in [q]$ denote the categorical label associated with $X_{i,k}$, and let $e_c \in \mathbb{R}^q$ denote the c-th canonical basis vector. We set $Y_{i,k} = e_{C_{i,k}}$ and train the raw class scores using the multiclass squared loss given by

$$M_i(\mathbf{h}, \mathbf{B}) := \frac{1}{\tau} \sum_{k=1}^{\tau} \left\| e_{C_{i,k}} - \mathbf{h}\phi_{\mathbf{B}}(X_{i,k}) \right\|^2.$$

This square loss with vector-valued class encodings is a standard classification surrogate and has been studied both empirically and theoretically in Hui and Belkin (2020); Hu et al. (2022); Vigogna et al. (2022). In particular, we define the conditional class-probability vector

$$\pi_i(X) := \mathbb{E}[Y_i \mid X] = \big(\mathbb{P}(C_i = 1 \mid X), \ldots, \mathbb{P}(C_i = q \mid X)\big)^\top,$$

and suppose that it is realizable by the shared representation, namely $\pi_i(X) = \mathbf{h}_i^\star \phi_{\mathbf{B}^\star}(X)$. For any score vector $z \in \mathbb{R}^q$, the conditional squared risk decomposes as follows:

$$\mathbb{E}[\|Y_i - z\|^2 \mid X] = \mathbb{E}[\|Y_i - \pi_i(X)\|^2 \mid X] + \|z - \pi_i(X)\|^2,$$

such that its population minimizer is precisely the class-probability vector. In addition, we note that $V_i := Y_i - \pi_i(X)$ is conditionally mean-zero and bounded, and thus it satisfies Assumption 1. Thus, the multiclass squared-loss problem is a direct instance of our regression model analyzed above.

For a vector $z \in \mathbb{R}^q$, $[z]_c$ denotes its c-th coordinate. To make each classifier single-valued, we use the same deterministic tie-breaking convention in both definitions, i.e., if several classes attain the maximum score, we select the class with the smallest index. The learned and Bayes classifiers are

$$\widehat{C}_i^{(t+1)}(X) := \min\left(\arg\max_{c \in [q]} \left[\mathbf{h}_i^{(t+1)} \phi_{\mathbf{B}^{(t)}}(X)\right]_c\right) \text{ and } C_i^\star(X) := \min\left(\arg\max_{c \in [q]} \pi_{i,c}(X)\right),$$

where $\pi_{i,c}(X) := \mathbb{P}(C_i = c \mid X)$.

Let $\bar{\mathcal{C}}_T := \frac{1}{T|\mathcal{H}|\tau} \sum_{t=0}^{T-1} \sum_{i\in\mathcal{H}} \sum_{k=1}^{\tau} \mathbf{1}\left\{\widehat{C}_i^{(t+1)}(X_{i,k}) \neq C_i^\star(X_{i,k})\right\}$ be the error rate between the learned and optimal Bayes classifiers.

Corollary 2. *Suppose the conditions of Theorem 4 hold and suppose that there exists a margin parameter $\gamma > 0$ such that, for every honest client $i \in \mathcal{H}$,*

$$\pi_{i,C_i^\star(X)}(X) - \max_{c\neq C_i^\star(X)} \pi_{i,c}(X) \geq \gamma \tag{9}$$

almost surely. Then, with probability at least $1 - 2\delta$, it holds that

$$\bar{\mathcal{C}}_T \leq \frac{2\bar{E}_T}{\gamma^2} \lesssim \frac{\Delta_0}{\gamma^2\sqrt{T}} + \frac{(1+\kappa+\kappa^2)}{\gamma^2}\left(p_\sigma \mathfrak{s}_{\mathcal{H},T} + \sigma^2 \mathfrak{s}_T\right) + \left(\kappa + \kappa^2 + \frac{1+\kappa}{|\mathcal{H}|}\right) \frac{\sigma_{\mathrm{g}}^2 \log(T/\delta)}{\gamma^2\sqrt{T}}. \tag{10}$$

Discussion. The margin condition (i.e., inequality (9)) ensures that the predicted class remains unchanged whenever the score-estimation error is sufficiently small relative to the separation between the most likely class and its closest competitor. Therefore, $\widehat{C}_i^{(t+1)}(X_{i,k}) \neq C_i^\star(X_{i,k})$ can occur only if the learned class-probability scores incur a squared prediction error of order at least γ^2. Therefore, Corollary 2 converts the parameter recovery error guarantee (equation 8) into a classification guarantee: the excess error rate is bounded by the parameter recovery error, with the same dependence on the number of rounds T, local sample size τ, number of honest clients $|\mathcal{H}|$, stochastic-gradient variance σ_{g}^2 and label noise variance σ^2, up to γ^{-2}. The proof is in Appendix H.

5 Experiments

We now validate our theoretical results on three heterogeneous federated datasets: CIFAR-10 image classification (Krizhevsky et al., 2009), FEMNIST character classification from the LEAF benchmark (Caldas et al., 2018), and School Exam Score regression (Zhou et al., 2011; Li et al., 2015). CIFAR-10 is partitioned into 100 clients, each with examples from two classes. FEMNIST keeps its natural writer-based partition. For the School Exam Score dataset each of the 139 schools defines one regression client. We evaluate NNM-preprocessed Krum and coordinate-wise trimmed mean (Allouah et al., 2023a) under ALIE and Mimic attacks. Full details are provided in Appendix I.

Baseline. We compare against the standard common-model formulation of Byzantine-robust FL (Karimireddy et al., 2020a; Allouah et al., 2023a; Toso et al., 2026). The baseline robustly aggregates updates of one model shared by every client, whereas our approach aggregates only the common-representation updates and keeps the linear heads personalized.

Table 1: Average local test accuracy (%) on CIFAR-10 and FEMNIST with 50 honest and 5 Byzantine clients per round. Entries report mean ± standard deviation over three seeds.

			NNM+Krum		NNM+TrMean	
Dataset	Loss	Method	ALIE	Mimic	ALIE	Mimic
CIFAR-10	Cross-entropy	Baseline	50.46 ± 0.57	46.39 ± 1.76	50.46 ± 0.71	47.17 ± 1.58
	Cross-entropy	Rep. learning	$\mathbf{88.79 \pm 0.42}$	$\mathbf{88.50 \pm 0.59}$	$\mathbf{88.76 \pm 0.44}$	$\mathbf{88.54 \pm 0.61}$
	Multiclass	Rep. learning	88.33 ± 0.53	87.95 ± 0.51	88.42 ± 0.52	87.89 ± 0.60
FEMNIST	Cross-entropy	Baseline	72.95 ± 1.95	70.34 ± 1.23	73.09 ± 2.03	70.66 ± 1.47
	Cross-entropy	Rep. learning	91.89 ± 0.31	92.43 ± 0.15	92.02 ± 0.28	92.51 ± 0.07
	Multiclass	Rep. learning	$\mathbf{93.56 \pm 0.24}$	$\mathbf{93.60 \pm 0.07}$	$\mathbf{93.57 \pm 0.22}$	$\mathbf{93.67 \pm 0.13}$

Classification. Table 1 reveals the fundamental limitation of adversarially robust FL when learning a common model parameter under heterogeneous honest clients. With cross-entropy, representation learning improves CIFAR-10 accuracy by 38.3–42.1 percentage points and FEMNIST accuracy by 18.9–22.1 points across all aggregator-attack pairs. The benefit persist for two distinct robust aggregators and attacks. This is precisely the result in Lemma 1: after personalization, the honest-update heterogeneity is tied to representation recovery and statistical errors instead of a persistent model-heterogeneity bias.

The multiclass squared loss also attains 87.89–88.42% on CIFAR-10 and 93.56–93.67% on FEMNIST. On FEMNIST, it improves over cross-entropy by 1.16–1.65 points. On CIFAR-10, the difference remains below 0.7 points. These results directly complement Corollary 2: the regression guarantees can be translated to classification.

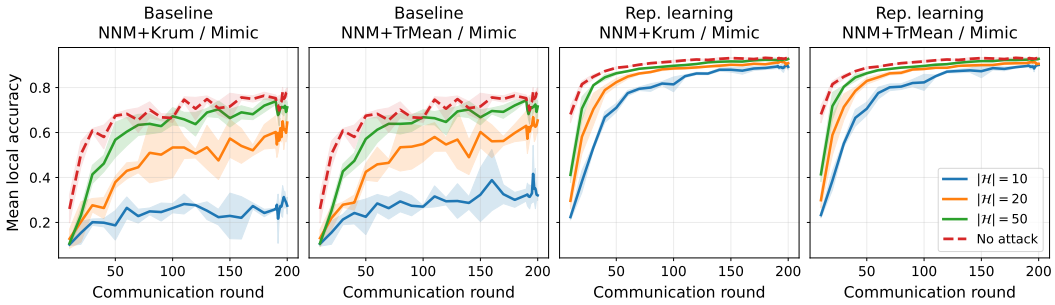


Figure 1: FEMNIST accuracy under the Mimic attack using cross-entropy loss and a fixed number of five Byzantine clients. Each curve reports the mean accuracy and standard deviation across three random seeds as the number of honest clients per round increases.

Table 2: School Exam test MSE using the squared loss with 20 honest and 5 Byzantine clients per round. Results report mean $\pm$ standard deviation over three seeds. Lower is better.

	NNM+Krum		NNM+TrMean	
Method	ALIE	Mimic	ALIE	Mimic
Baseline	0.7611 ± 0.0452	0.6532 ± 0.0201	0.7672 ± 0.0465	0.6531 ± 0.0226
Rep. learning	$\mathbf{0.6324 \pm 0.0272}$	$\mathbf{0.6300 \pm 0.0276}$	$\mathbf{0.6316 \pm 0.0275}$	$\mathbf{0.6301 \pm 0.0276}$

Pooled mean predictor	Per-school mean predictor
0.9913 ± 0.0306	0.8826 ± 0.0263

Figure 1 shows the effect of increasing the number of honest clients $|\mathcal{H}|$ under the Mimic attack. For both aggregators, representation learning approaches the no-attack curve as $|\mathcal{H}|$ increases and remains near 90% accuracy even with only ten honest updates. In contrast, the baseline is substantially below the no-attack curve and deteriorates sharply for small $|\mathcal{H}|$. This supports our results in Theorem 1 and Corollary 1, whose client-averaged statistical and stochastic-gradient terms decrease with $|\mathcal{H}|$, while no irreducible model-heterogeneity term remains.

Regression. On the School Exam Score dataset, representation learning achieves an MSE of approximately 0.63 in every adversarial configuration, reducing the common-model baseline error by 3.5%–17.7% and outperforming both constant predictors (i.e., pooled mean and per-school mean). The pooled-mean predictor assigns every example the average training label across all schools, whereas the per-school-mean predictor assigns each example the average training label of its corresponding school. These results support Corollary 1, that is, robust aggregation can operate on a genuinely shared representation, while the personalized heads absorb school-specific heterogeneity, improving regression without incurring an irreducible model-parameter heterogeneity bias.

6 Conclusion and Future Work

We established rigorous high-probability, non-asymptotic guarantees for nonlinear representation learning in heterogeneous federated systems with adversarial clients. By separating each client model into a personalized head and a common nonlinear representation, we demonstrated that the heterogeneity among honest representation gradients is governed by representation-recovery, initialization error, finite-sample, and stochastic gradient terms, without an irreducible model-heterogeneity bias. Our bounds explicitly reveal how recovery improves with the number of communication rounds T, local samples τ, and honest clients $|\mathcal{H}|$. We further demonstrated that these guarantees extend to multiclass classification and validated our theoretical guarantees on CIFAR-10, FEMNIST, and School Exam Score datasets.

We leave for future work to determine whether the restriction on the robustness coefficient κ (i.e., equation 4) is fundamental. In particular, Toso et al. (2026), for single-model Byzantine-robust linear and nonlinear regression demonstrated that a restriction on κ can instead be replaced by a sufficiently large sample burn-in condition. Establishing whether a similar guarantee also holds for our setting would clarify the statistical limits of personalization under adversarial clients.

AI USE STATEMENT

We used generative AI tools only for the purpose of editing and polishing the manuscript for clarity, grammar, and readability. We emphasize that we did not use generative AI tools to conceive the project idea or develop any of the theoretical results, formulate mathematical claims, construct proofs, process data, or interpret the results. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. In *Concurrency: the works of leslie lamport*, pages 203–226. 2019.

Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. Byzantine machine learning: A primer. *ACM Computing Surveys*, 56(7):1–39, 2024a.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. Pmlr, 2017.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pages 5650–5659. Pmlr, 2018.

El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *International conference on machine learning*, pages 3521–3530. PMLR, 2018.

Christopher G Small. A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277, 1990.

Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. Robust machine learning. *Cham, Switzerland: Springer*, 2024b.

Leonardo F. Toso, James Anderson, Nirupam Gupta, and Rafael Pinot. On the Gradient Heterogeneity Dynamics of Adversarially Robust Federated Regression. *arXiv preprint arXiv:2609.25705*, 2026.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 1232–1300. PMLR, 2023a.

Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. *arXiv preprint arXiv:2006.09365*, 2020a.

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

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pages 2089–2099. PMLR, 2021.

Yi-Rui Yang, Kun Wang, and Wu-Jun Li. Fedrep: A byzantine-robust, communication-efficient and privacy-preserving framework for federated learning. *arXiv preprint arXiv:2303.05206*, 2023.

Thomas T Zhang, Leonardo Felipe Toso, James Anderson, and Nikolai Matni. Sample-efficient linear representation learning from non-iid non-isotropic data. In *International Conference on Learning Representations*, volume 2024, pages 16551–16562, 2024a.

Thomas T Zhang, Bruce D Lee, Ingvar Ziemann, George J Pappas, and Nikolai Matni. Guarantees for nonlinear representation learning: Non-identical covariates, dependent data, fewer samples. *arXiv preprint arXiv:2410.11227*, 2024b.

Thomas T Zhang, Katie Kang, Bruce D Lee, Claire Tomlin, Sergey Levine, Stephen Tu, and Nikolai Matni. Multi-task imitation learning for linear dynamical systems. In *Learning for Dynamics and Control Conference*, pages 586–599. PMLR, 2023.

Bruce D Lee, Leonardo F Toso, Thomas T Zhang, James Anderson, and Nikolai Matni. Regret analysis of multi-task representation learning for linear-quadratic adaptive control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18062–18070, 2025.

Kasra Fallah, Leonardo F Toso, and James Anderson. Adversarially robust multitask adaptive control. *arXiv preprint arXiv:2511.05444*, 2025.

Roman Vershynin. High-dimensional probability. *Cambridge Series in Statistical and Probabilistic Mathematics*, 47, 2019.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pages 6246–6283. PMLR, 2022.

Anish Acharya, Abolfazl Hashemi, Prateek Jain, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Robust training in high dimensions via block coordinate geometric median descent. In *International Conference on Artificial Intelligence and Statistics*, pages 11145–11168. PMLR, 2022.

El Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyên Hoang, and Sébastien Rouault. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). *Advances in neural information processing systems*, 34:25044–25057, 2021.

Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and Geovani Rizk. Robust distributed learning: Tight error bounds and breakdown point under data heterogeneity. *Advances in neural information processing systems*, 36:45744–45776, 2023b.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

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

Eduard Gorbunov, Samuel Horváth, Peter Richtárik, and Gauthier Gidel. Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top. *arXiv preprint arXiv:2206.00529*, 2022.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Tianyang Hu, Jun Wang, Wenjia Wang, and Zhenguo Li. Understanding square loss in training overparametrized neural network classifiers. *Advances in Neural Information Processing Systems*, 35:16495–16508, 2022.

Stefano Vigogna, Giacomo Meanti, Ernesto De Vito, and Lorenzo Rosasco. Multiclass learning with margin: exponential rates with no bias-variance trade-off. In *International Conference on Machine Learning*, pages 22260–22269. PMLR, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Jiayu Zhou, Jianhui Chen, Jieping Ye, et al. Malsar: Multi-task learning via structural regularization. *Arizona State University*, 21:1–50, 2011.

Ya Li, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Multi-task model and feature joint learning. In *IJCAI*, pages 3643–3649, 2015.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning. *Advances in Neural Information Processing Systems*, 35:10572–10586, 2022.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Marc González, Rachid Guerraoui, Rafael Pinot, Geovani Rizk, John Stephan, and François Taïani. ByzFL: Research Framework for Robust Federated Learning, 2025.