

Nonlinear equilibrium transitions in a potential game model for federated learning

Kang Liu^a, Ziqi Wang^{b,*}, Enrique Zuazua^{b,c,d}

^a Institut de Mathématiques de Bourgogne, Université Bourgogne Europe, CNRS, Dijon, 21000, France

^b Chair for Dynamics, Control, Machine Learning and Numerics - Alexander von Humboldt Professorship, Department of Mathematics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstrasse 11, Erlangen, 91058, Germany

^c Chair of Computational Mathematics, Fundación Deusto, Avenida de las Universidades, 24, Bilbao, Basque Country, 48007, Spain

^d Departamento de Matemáticas, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid, 28049, Spain

ARTICLE INFO

Communicated by: Dmitry Pelinovsky

Keywords:

Federated learning
Potential game
Nash equilibrium
Best-response algorithm

ABSTRACT

In federated learning (FL), a central server typically allocates training efforts to clients. However, from a market-oriented perspective, clients may independently choose their training efforts based on rational self-interest. To study this setting, we propose a potential game framework in which each client's payoff is determined by its individual effort and the rewards provided by the server. The rewards are influenced by the collective efforts of all clients and can be modulated by a reward factor. We first establish the existence of Nash equilibria (NEs) and then investigate their uniqueness in a stationary setting. We show that the NEs depend nonlinearly on the reward factor and exhibit a nonsmooth transition at a critical value, where the stationary potential loses strict curvature, leading to nonunique NEs and a jump between low-effort and high-effort branches. Furthermore, we prove the convergence of the best-response algorithm for computing NEs in our FL game. Finally, we apply the clients' rational efforts derived from the NEs to FL training with various datasets and models, thereby validating the effectiveness of the identified critical reward factor.

1. Introduction

Federated learning (FL) [1] has shown considerable promise for training large-scale machine learning models across distributed data sources. In FL, clients train local models using local datasets and transmit only the updated model parameters, rather than the raw data, to a central server for aggregation. From the viewpoint of dynamical systems, the iterative client-server updates can be viewed as a discrete approximation of a coupled gradient flow, where each client's local descent contributes to the evolution of the global model parameters along a shared loss landscape.

1.1. Motivations

Incentives as a game-theoretic interaction. Conventional FL often assumes voluntary participation from heterogeneous clients while the server prescribes local workloads in a centrally planned manner. When private costs differ across clients, this can induce strategic under-training or silent dropout [2]. We therefore model incentives as a game-theoretic interaction. The server offers rewards, each client chooses its effort by balancing the reward against its local cost, and the coupled FL aggre-

gation makes each payoff depend on the efforts of others. This yields a Nash equilibrium (NE) characterization of a self-consistent effort profile, and Fig. 1 illustrates how this equilibrium is integrated into the FL training loop. Guided by this viewpoint, we study two sets of questions:

- How should the FL game be modeled (reward rule, local costs, and feasible sets), and under what assumptions do NE existence, uniqueness, and efficient computation hold?
- How does the equilibrium effort profile respond to the reward parameter? In particular, do threshold-type or nonlinear transitions arise, and how should the server tune this reward parameter to improve participation quality and final model performance?

Rationale for potential games. Analyzing NEs in general N -player games is notoriously challenging [3], and tracing how equilibria vary with game parameters is even harder [4]. In FL, clients' updates are typically aggregated through weighted sums, which motivates designing the server's reward as a compatible weighted sum of client efforts. Under this structure, the interaction becomes a *potential game* [5], allowing equilibria to be studied through a single potential function and analyzed with standard optimization tools.

* Corresponding author.

E-mail addresses: kang.liu@u-bourgogne.fr (K. Liu), ziqi.wang@fau.de (Z. Wang), enrique.zuazua@fau.de (E. Zuazua).

<https://doi.org/10.1016/j.physd.2026.135288>

Received 13 November 2025; Received in revised form 15 April 2026; Accepted 30 May 2026

Available online 6 June 2026

0167-2789/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

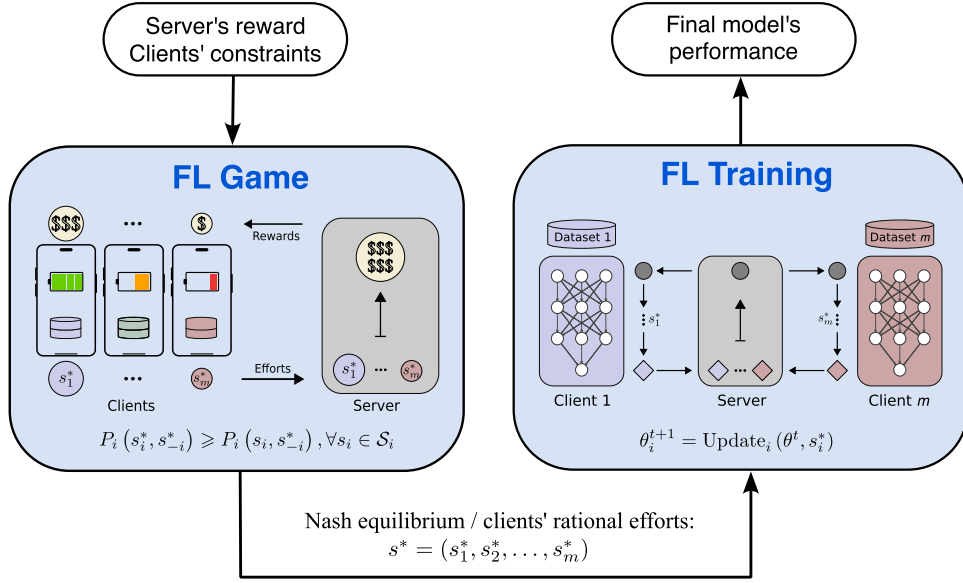


Fig. 1. In the FL game (left), each client i maximizes the payoff P_i , and the NE s^* satisfies $P_i(s_i^*, s_{-i}^*) \geq P_i(s_i, s_{-i}^*)$ for all $s_i \in S_i$. The equilibrium efforts s^* (e.g., local epochs) then drive the FL training stage (right), where the model parameters are updated according to $\theta_i^{t+1} = \text{Update}_i(\theta^t, s_i^*)$.

1.2. Main results

We state the core findings first and defer the full model setup to [Section 2](#).

- Existence via potential structure.** We formulate client participation in FL as a weighted potential game and prove the existence of NEs in [Theorem 2.1](#).
- Uniqueness in the stationary quadratic regime.** In [Theorem 3.1](#), for stationary efforts and quadratic costs, we show uniqueness for all values of the server's reward factor $\lambda \neq \lambda^*$, while at $\lambda = \lambda^*$ the game may admit infinitely many equilibria.
- Nonlinear structure and landscape transition.** [Corollary 3.1](#) shows that the average training effort $\bar{s}^*$ grows nonlinearly with respect to λ , with three thresholds $\lambda_1 < \lambda^* < \lambda_2$ for activation, jump, and saturation (see [Fig. 2](#) and [Table 1](#)). In particular, at $\lambda = \lambda^*$, the curve undergoes a nonsmooth transition. [Remark 3.3](#) explains this by the loss of strict curvature and the emergence of a flat direction in the stationary potential, leading to nonunique NEs and a jump in the equilibrium effort.
- Best-response computation.** In [Theorem 3.2](#) (and the general [Theorem 5.1](#)), we establish convergence guarantees for the best-response algorithm used to compute NEs.
- Theory-experiment consistency.** Across multiple datasets and FL methods, the strongest performance gain occurs near the jump regime around λ^* , consistent with the nonlinear sensitivity predicted by the theory.

1.3. Related work

Game-theoretic analysis has long been used for multi-agent decision-making [\[6,7\]](#). In FL, it is widely used to study fairness, robustness, and strategic participation [\[8\]](#). Existing FL game formulations emphasize different objects (coalitions, payments, or contribution scores), which leads to different analytical guarantees. Below we position our formulation against representative lines of work.

1.3.1. Coalitional game theory for fairness in FL

Coalitional (hedonic) games [\[9,10\]](#) model how clients form stable groups. In FL, model-sharing formulations [\[11,12\]](#) analyze stability and

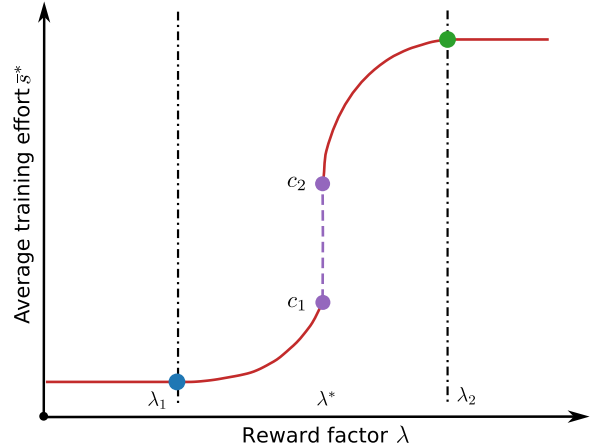


Fig. 2. Illustration of the nonlinear threshold structure and jump phenomenon.

Table 1

Summary of the critical thresholds, see [Theorem 3.1](#) and [Corollary 3.1](#) for more details.

Symbol	Name	Eq.	Definition and role
λ^*	Jump point	(3.3)	The critical reward factor where non-uniqueness occurs.
λ_1	Activation point	(3.8)	If $\lambda \in (0, \lambda_1)$, the unique NE is at minimum efforts.
λ_2	Saturation point	(3.8)	If $\lambda \in (\lambda_2, +\infty)$, the unique NE is at maximum efforts.
c_1	Lower jump bound	(3.4)	If $\lambda \in (\lambda_1, \lambda^*)$, the unique NE satisfies $\bar{s}^* < c_1$.
c_2	Upper jump bound	(3.4)	If $\lambda \in (\lambda^*, \lambda_2)$, the unique NE satisfies $\bar{s}^* > c_2$.

welfare gaps between stable and socially optimal coalitions. These models are well suited to fairness questions driven by grouping and coalition stability, where the key variable is the coalition partition itself. These works focus on a discrete coalition structure rather than a continuous equilibrium effort profile controlled by a server-side reward parameter. As a result, they do not provide our type of NE uniqueness and best-response convergence results.

1.3.2. Stackelberg and auction-based incentive mechanisms

Stackelberg and auction-based mechanisms design payments to induce participation from strategic clients. Representative works study bi-level reward allocation [13], closed-form market equilibria and multi-tier extensions [14,15], and practical marketplace systems with security, reputation, or privacy guarantees [16–18]. Their main strength is explicit economic mechanism design at the server level, often with implementable pricing or allocation rules. These studies mainly characterize payment/allocation rules, whereas our focus is a full threshold analysis of equilibrium effort profiles under a server-tunable reward parameter.

1.3.3. Shapley value-based contribution evaluation

Shapley-based methods evaluate participant contributions after training [19]. Recent work improves scalability via gradient replay and sampling approximations [20–22]. This line is particularly useful for retrospective attribution and reward sharing once model updates are already observed. Because they are post-training valuation methods, they do not analyze strategic effort selection before training, equilibrium thresholds, or best-response computation for NE profiles.

1.3.4. Potential games and their applications

Potential games [5] offer a useful structure because NEs correspond to stationary points of one potential function, which often enables tractable best-response analysis [23]. This framework has been effective in areas such as multi-agent control and resource allocation [24–27]. For incentive problems, this structure is especially attractive because equilibrium existence and dynamics can be studied through one global scalar function. Compared with the FL incentive formulations reviewed above, our contribution is formulated within a single server-tunable game. Within this framework, we establish equilibrium uniqueness, characterize a nonlinear reward threshold structure, and show convergence of best-response dynamics. To the best of our knowledge, these three properties have not been established jointly in the FL incentive formulations reviewed above.

1.4. Notation

In this subsection, we introduce the key notation used in the remainder of the paper. We denote by Γ a game with three components. The first component is a finite set of players $[m] := \{1, \dots, m\}$. The second is a pure strategy set $S_i \subseteq \mathbb{R}^d$ for each player $i \in [m]$, together with $S := \prod_{i \in [m]} S_i$. The third is a payoff function $P_i : S \rightarrow \mathbb{R}$ for each player $i \in [m]$. We say that a game Γ is finite if its strategy set S is finite. For any $i \in [m]$, we define $S_{-i} := \prod_{j \neq i} S_j$. For convenience, for any $s \in S$, we do not distinguish between $P_i(s)$ and $P_i(s_i, s_{-i})$, where $s_i \in S_i$ is the i -th coordinate of s and $s_{-i} \in S_{-i}$ is the tuple of other coordinates.

In a non-cooperative game, each player $i \in [m]$ is self-interested and seeks to maximize their payoff $P_i(s_i, s_{-i})$, leading to the following concept of an NE.

Definition 1.1 (Nash equilibrium). In the game Γ , we call a point $s^* \in S$ an NE if the following inequality holds:

$$P_i(s_i^*, s_{-i}^*) \geq P_i(s_i, s_{-i}^*), \quad \forall s_i \in S_i, \forall i \in [m].$$

A special type of non-cooperative games is known as potential games [5], which can be defined as follows:

Definition 1.2 (Potential game). Let $w = (w_i)_{i \in [m]}$ be a strictly positive vector. The game Γ is said to be a (weighted) potential game if there exists a function $P : S \rightarrow \mathbb{R}$ such that

$$\begin{aligned} P_i(s_i, s_{-i}) - P_i(x_i, s_{-i}) \\ = w_i(P(s_i, s_{-i}) - P(x_i, s_{-i})), \quad \forall s_i, x_i \in S_i, \forall i \in [m]. \end{aligned}$$

1.5. Outline of the paper

The remainder of the paper is organized as follows. In Section 2, we present the design of our FL game and establish the existence of an NE. Next, in Section 3, we focus on the uniqueness of the NE in a stationary scenario and study the convergence of the best-response algorithm. Simulation results are presented and analyzed in Section 4, followed by detailed technical proofs and theorems in Section 5. Finally, we summarize our conclusions and perspectives in Section 6.

2. A potential game framework for FL

This section models client effort selection in FL as a potential game. We first recall the FL updates, then define the FL game, and finally establish the existence of NEs through the potential structure.

We start from the standard FL training objective:

$$\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^m \rho_i \ell_i(\theta), \quad \text{where } \rho_i = \frac{D_i}{D}, \text{ for } i \in [m], \quad (2.1)$$

where m is the number of clients, D_i is client i 's local data size, $D = \sum_{i=1}^m D_i$, and ℓ_i is the local empirical loss. FL training proceeds over communication rounds with local updates and server aggregation. At round $t \in [T]$, each client $i \in [m]$ initializes $\theta_i^{t,0} = \theta^t$ and performs

$$\theta_i^{t,e} = \theta_i^{t,e-1} - \eta_i \nabla \ell_i(\theta_i^{t,e-1}), \quad e = 1, \dots, s_i^t, \quad (2.2)$$

then sends $\theta_i^{t+1} := \theta_i^{t,s_i^t}$ to the server. The server aggregates $\theta^{t+1} = \sum_{i=1}^m \rho_i \theta_i^{t+1}$ and broadcasts θ^{t+1} for the next round.

Rather than assuming that the server prescribes client efforts s_i^t , we model clients as self-interested players who choose effort by balancing rewards against costs. Accordingly, we adopt a finite-player game framework to characterize clients' rational effort choices. In this game, each client maximizes a payoff determined by local training cost and server reward. We denote this game as Γ_{FL} and provide its definition below.

Definition 2.1 (FL game Γ_{FL}). The FL game Γ_{FL} is specified by the player set $[m]$, the strategy profile $S = \prod_{i \in [m]} S_i$ with $S_i \subseteq \mathbb{R}_+^T$, and payoff functions $P_i : S \rightarrow \mathbb{R}$ of the form

$$P_i(s_i, s_{-i}) = \sum_{t=1}^T r_{\lambda}(s_i^t, s_{-i}^t) - c_i(s_i^t), \quad i \in [m], \quad (2.3)$$

where $r_{\lambda}(s_i^t, s_{-i}^t)$ is the server-side reward, and $c_i(s_i^t)$ denotes client i 's local cost.

As shown in (2.3), by adjusting λ , the server can steer the equilibrium profile $s^*(\lambda)$. Next, we introduce the three core components of the FL game, namely strategy sets in Section 2.1, reward design in Section 2.2, and cost design in Section 2.3.

2.1. Clients' Strategy sets

We measure each client's effort by the number of local update steps (or, equivalently, epochs under a fixed conversion). In practical FL systems, client efforts are bounded by resource constraints, and we consider the following two feasible-set models.

Example 2.1 (Total budget case). For client i , let

$$S_i = \left\{ s_i \in \mathbb{R}_+^T \mid b_i \leq \sum_{t \in [T]} s_i^t \leq B_i \right\}, \quad (2.4)$$

where b_i and B_i are the minimum and maximum total efforts over T rounds.

Example 2.2 (Stationary case). In many FL scenarios, clients select one effort level and keep it fixed over rounds. Then

$$S_i = \left\{ s_i \in \mathbb{R}_+^T \mid q_i \leq s_i^t = s_i^\tau \leq Q_i, \forall t, \tau \in [T] \right\}, \quad (2.5)$$

where q_i and Q_i are the minimum and maximum per-round efforts.

2.2. Design of the reward $r_\lambda(s_i^t, s_{-i}^t)$

We consider the incentive rule

$$r_\lambda(s_i^t, s_{-i}^t) = p_\lambda^t s_i^t, \quad \text{where } p_\lambda^t = \lambda \sum_{j=1}^m \rho_j s_j^t. \quad (2.6)$$

Here, $\lambda > 0$ is the server-controlled reward factor. Since the unit price p_λ^t increases with clients' aggregate effort, each client's marginal incentive is strengthened by the efforts of others, creating strategic complementarity among clients. This is consistent with FL evidence that greater effective local computation can improve performance under fixed communication budgets [1,28,29].

2.3. Design of the cost $c_i(s_i^t)$

The function $c_i(s_i^t)$ quantifies client i 's participation burden. A linear model $c_i(s_i^t) = \alpha_i s_i^t$ represents constant marginal cost. However, if we interpret the cost as a measure of the client's reluctance to exert effort, a quadratic function, $c_i(s_i^t) = \alpha_i (s_i^t)^2$, is more appropriate. This captures the tendency of clients to become increasingly unwilling to exert higher levels of local training effort. For instance, on a smartphone, performing the first few local updates when the battery is full may be perceived as inexpensive, whereas executing the same number of updates under low battery appears significantly more costly. Such nonlinear perceptions of cost are well captured by the quadratic model.

2.4. Existence of an NE of Γ_{FL}

Based on the formulation of the strategy set and payoff functions, we now present the results regarding the potential structure and the existence of an NE in the game Γ_{FL} .

Theorem 2.1 (Existence). *Consider the game Γ_{FL} (see Definition 2.1) with the reward function (2.6). Then it admits a w -potential given by*

$$P_{FL}(s) = \sum_{i=1}^m \sum_{t=1}^T \left(\frac{\lambda}{2} \rho_i^2 (s_i^t)^2 - \rho_i c_i(s_i^t) \right) + \sum_{i=1}^T \frac{\lambda}{2} \left(\sum_{i=1}^m \rho_i s_i^t \right)^2, \quad s \in S, \quad (2.7)$$

where $c_i(\cdot)$ is client i 's local cost, and $w_i = 1/\rho_i$ for $i \in [m]$. If S is compact and $c_i(\cdot)$ is lower semi-continuous, then Γ_{FL} possesses at least one NE.

Proof. It is straightforward to verify that P_{FL} given in (2.7) is a w -potential of Γ_{FL} in the sense of Definition 1.2, with $w_i = 1/\rho_i$ for all $i \in [m]$. Assume now that c_i is lower semi-continuous. Then the potential function P_{FL} is upper semi-continuous over the strategy set S . Moreover, if S is compact, P_{FL} attains its maximum on S . This maximizer corresponds to an NE of the game Γ_{FL} , see [5, Lem. 2.1]. $\square$

Beyond the existence result, establishing the uniqueness of the NE is generally more challenging, see Remark 2.1. A possible condition ensuring uniqueness is the strict concavity of the potential function $P_{FL}(s)$. However, this typically requires each cost function c_i to be strongly convex with a sufficiently large strong convexity modulus. For instance, if $c_i(x) = \alpha_i x^2$, then uniqueness holds when $\lambda < 2\alpha_i/(1 + \rho_i^{-1})$ for all i , which is often impractical in real-world scenarios. Furthermore, in time-dependent settings (Example 2.1), uniqueness is easily lost. See Remark 3.8 for further discussion.

Remark 2.1 (Comparison to the classic Cournot competition). Note that due to the positivity of λ , the unit price in (2.6) increases with the aggregate of clients' efforts. This contrasts with the classic Cournot competition [30], where the price decreases as aggregate effort or production increases. The uniqueness of the NE in this increasing case requires careful analysis (see assumptions of [31, Thm. 1]).

In the following section, we investigate the uniqueness of the NE in the stationary and quadratic setting. We show that the NE is unique for all values of λ , except at a critical threshold. This result is established via a detailed analysis of the fixed-point system that defines the NE, as presented in Section 5.

3. Uniqueness results and the best-response algorithm

In this section, we focus on the stationary case of Γ_{FL} with quadratic costs $c_i(x) = \alpha_i x^2$. We first reformulate the stationary game, then present uniqueness and threshold results, establish convergence of the best-response algorithm, and finally discuss extensions to non-uniform data sizes and heterogeneous (time-dependent) settings.

3.1. Stationary game formulation

To simplify the constants in the statements and proofs, we first present the uniform-data-size case, and we discuss the non-uniform counterpart in Remark 3.7. Under stationary strategies (2.5), reward rule (2.6), and quadratic costs, the game Γ_{FL} is equivalently written as the stationary game $\hat{\Gamma}_{FL}$ below.

Definition 3.1 (Stationary FL game $\hat{\Gamma}_{FL}$). The stationary FL game $\hat{\Gamma}_{FL}$ consists of the player set (clients) $[m]$, the strategy sets $\hat{S}_i = [q_i, Q_i]$ for each client, and payoff functions $\hat{P}_i: \hat{S} \rightarrow \mathbb{R}$ with $\hat{S} = \prod_{i \in [m]} \hat{S}_i$, given by

$$\hat{P}_i(s_i, s_{-i}) = \frac{\lambda}{m} s_i \sum_{j=1}^m s_j - \alpha_i (s_i)^2, \quad \text{with } \alpha_i > 0. \quad (3.1)$$

As a special case of Γ_{FL} , Theorem 2.1 implies that $\hat{\Gamma}_{FL}$ is still a potential game and admits NEs. This stationary formulation is the regime where uniqueness can be characterized sharply, as shown in the next subsection.

3.2. Uniqueness of the NE in the stationary game

Let us introduce the crucial constants (summarized in Table 1) for the uniqueness result.

- Concavity threshold:

$$\bar{\lambda} = m\alpha_{\min}, \quad \text{where } \alpha_{\min} := \min_{i \in [m]} \{\alpha_i\}. \quad (3.2)$$

- Jump point $\lambda^* \in (0, \bar{\lambda})$, which is the unique solution to

$$\sum_{i=1}^m \frac{\lambda^*}{2m\alpha_i - \lambda^*} = 1. \quad (3.3)$$

- Two constants c_1 and c_2 that determine whether a jump occurs at λ^* :

$$c_1 := \max_{i \in [m]} \left\{ \frac{2\alpha_i q_i}{\lambda^*} - \frac{q_i}{m} \right\}, \quad c_2 := \min_{i \in [m]} \left\{ \frac{2\alpha_i Q_i}{\lambda^*} - \frac{Q_i}{m} \right\}. \quad (3.4)$$

Theorem 3.1 (Uniqueness). *The following statements hold:*

1. If $\lambda > 0$ and $\lambda \neq \lambda^*$, then the game $\hat{\Gamma}_{FL}$ has a unique NE.

2. If $\lambda = \lambda^*$, then the following holds:

(a) If $c_1 \geq c_2$, then the game $\hat{\Gamma}_{FL}$ has a unique NE;

(b) If $c_1 < c_2$, then the game $\hat{\Gamma}_{FL}$ has infinitely many NEs. A point $s^* \in \hat{S}$ is an NE if and only if there exists $c \in [c_1, c_2]$ such that

$$s_i^* = \frac{\lambda^* c}{2\alpha_i - \lambda^*/m}, \quad \forall i \in [m]. \quad (3.5)$$

Proof. The proof is presented in Section 5. $\square$

Remark 3.1 (Role of the concavity threshold $\bar{\lambda}$). When $\lambda \in (0, \bar{\lambda})$, each payoff $\hat{P}_i(\cdot, s_{-i})$ is concave in s_i , so $\hat{\Gamma}_{FL}$ lies in the classical concave-game framework [32]. Theorem 3.1(1) shows that uniqueness can still hold beyond this range. In our analysis, $\bar{\lambda}$ is mainly used to locate the critical value λ^* through (3.3).

Remark 3.2 (A symmetric benchmark). Assume that $\alpha_i = \alpha_j = \alpha$ for all $i, j \in [m]$. Then

$$\bar{\lambda} = m\alpha, \quad \lambda^* = \frac{2m}{m+1}\alpha, \quad c_1 = \max_{i \in [m]} \{q_i\}, \quad c_2 = \min_{i \in [m]} \{Q_i\}. \quad (3.6)$$

In many FL settings, q_i is set close to one local epoch and Q_i is much larger, so $c_1 < c_2$ is common. This benchmark also clarifies how the model connects to symmetric mean-field limits [33, Section 2].

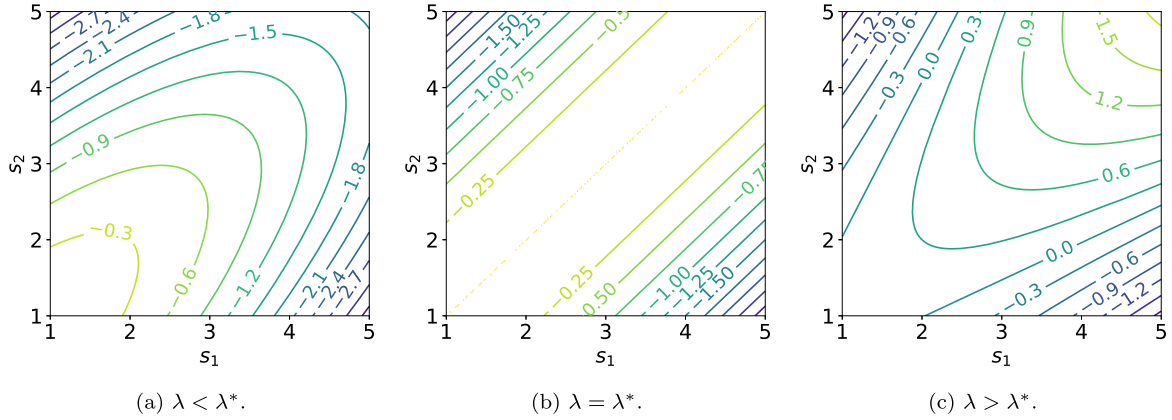


Fig. 3. Two-player stationary potential landscape of Example 3.1. At $\lambda = \lambda^*$, a flat ridge appears along $s_1 = s_2$, yielding infinitely many NEs on the feasible diagonal segment. In contrast, for $\lambda < \lambda^*$ and $\lambda > \lambda^*$, the potential has a unique maximizer, corresponding to a unique NE.

Remark 3.3 (Geometric interpretation of λ^*). In game $\hat{\Gamma}_{\text{FL}}$, consider the potential

$$\hat{P}_{\text{FL}}(s) := \frac{\lambda}{2m} \left(\sum_{i=1}^m s_i \right)^2 + \sum_{i=1}^m \left(\frac{\lambda}{2m} - \alpha_i \right) s_i^2, \quad s \in \hat{S}. \quad (3.7)$$

Its Hessian is

$$\nabla^2 \hat{P}_{\text{FL}} = \frac{\lambda}{m} \mathbf{1}\mathbf{1}^\top + M, \quad \text{with } M := \text{diag} \left(\frac{\lambda}{m} - 2\alpha_1, \dots, \frac{\lambda}{m} - 2\alpha_m \right).$$

For $\lambda < \bar{\lambda} = m\alpha_{\min}$, the matrix M is nonsingular. By the matrix determinant lemma,

$$\det(\nabla^2 \hat{P}_{\text{FL}}) = \det(M) \left(1 + \frac{\lambda}{m} \mathbf{1}^\top M^{-1} \mathbf{1} \right).$$

Therefore, the Hessian is singular if and only if $1 + \lambda/m \sum_{i=1}^m (\lambda/m - 2\alpha_i)^{-1} = 0$, that is,

$$\sum_{i=1}^m \frac{\lambda}{2m\alpha_i - \lambda} = 1.$$

This condition coincides exactly with (3.3). In this case, there exists a nonzero vector $v \in \mathbb{R}^m$ such that

$$\nabla^2 \hat{P}_{\text{FL}} v = 0.$$

Moreover, by the homogeneous quadratic structure of $\hat{P}_{\text{FL}}$, we have

$$\hat{P}_{\text{FL}}(s + tv) = \hat{P}_{\text{FL}}(s), \quad \forall s \in \mathbb{R}^m, \forall t \in \mathbb{R}.$$

Thus, v is a flat direction of the potential. If s^* is a stationary point, then

$$\nabla \hat{P}_{\text{FL}}(s^* + tv) = \nabla \hat{P}_{\text{FL}}(s^*) + t \nabla^2 \hat{P}_{\text{FL}} v = 0,$$

so every feasible point on the affine line $s^* + tv$ is again stationary. Since the stationary potential is concave at $\lambda = \lambda^*$, these feasible stationary points are global maximizers of $\hat{P}_{\text{FL}}$. Hence, whenever the feasible set contains a nontrivial segment in this flat direction, the potential admits multiple global maximizers. Since $\hat{\Gamma}_{\text{FL}}$ is a potential game, these maximizers correspond to NEs, which explains the non-uniqueness of NEs at $\lambda = \lambda^*$. In the case $c_1 < c_2$, this flat segment connects the low-effort and high-effort branches, which gives the geometric mechanism behind the jump of equilibria across λ^* . See Example 3.1 and Fig. 3 for a visual illustration.

Example 3.1 (Two-player toy landscape). Consider $m = 2$ with symmetric parameters $\alpha_1 = \alpha_2 = 1$, $q_1 = q_2 = 1$, and $Q_1 = Q_2 = 5$. Then (3.6) gives $\lambda^* = 4/3$, $c_1 = 1$, and $c_2 = 5$. Fig. 3 shows the potential landscape of (3.7). This example illustrates how the critical threshold λ^* governs the transition between uniqueness and non-uniqueness of NEs.

3.3. Nonlinear threshold structure of the NE and jump phenomenon

In the following corollary, we present a more precise description of the unique NE of the game $\hat{\Gamma}_{\text{FL}}$ when $\lambda \neq \lambda^*$. Before that, we define two additional critical thresholds, namely the activation point λ_1 and the saturation point λ_2 (see Fig. 2), as follows:

$$\lambda_1 := 2 \min_{i \in [m]} \left\{ \frac{\alpha_i q_i}{\bar{q} + \frac{q_i}{m}} \right\}, \quad \lambda_2 := 2 \max_{i \in [m]} \left\{ \frac{\alpha_i Q_i}{\bar{Q} + \frac{Q_i}{m}} \right\}, \quad (3.8)$$

where q_i and Q_i are the minimum and maximum efforts of client $i \in [m]$, while $\bar{q} = \sum_{i=1}^m q_i/m$ and $\bar{Q} = \sum_{i=1}^m Q_i/m$. It follows from the definitions that $0 < \lambda_1 \leq \lambda^* \leq \lambda_2$.

Corollary 3.1 (Precise description of the unique NE). *The following statements hold:*

1. If $\lambda \in (0, \lambda_1)$, then the unique NE of the game $\hat{\Gamma}_{\text{FL}}$ is $s^* = (q_i)_{i \in [m]}$.
2. If $\lambda \in (\lambda_1, \lambda^*)$, then the unique NE of the game $\hat{\Gamma}_{\text{FL}}$, denoted by s^* , satisfies

$$\bar{s}^* := \frac{1}{m} \sum_{i \in [m]} s_i^* < c_1.$$

3. If $\lambda \in (\lambda^*, \lambda_2)$, the unique NE of the game $\hat{\Gamma}_{\text{FL}}$, denoted by s^* , satisfies

$$\bar{s}^* := \frac{1}{m} \sum_{i \in [m]} s_i^* > c_2.$$

4. If $\lambda \in (\lambda_2, +\infty)$, the unique NE of the game $\hat{\Gamma}_{\text{FL}}$ is $s^* = (Q_i)_{i \in [m]}$.

Proof. The proof is presented in Section 5. $\square$

Remark 3.4 (Regime interpretation). Corollary 3.1 yields four regimes, namely inactive incentives $(0, \lambda_1)$, a low-effort branch (λ_1, λ^*) , a high-effort branch (λ^*, λ_2) , and saturation $(\lambda_2, +\infty)$. When $c_1 < c_2$, λ^* is a genuine jump point. In particular, the unique-NE branches on the two sides of λ^* are separated by the gap between c_1 and c_2 , so crossing λ^* moves the equilibrium mean effort from at most c_1 to at least c_2 rather than along a gradual path. This yields a nontrivial jump magnitude of at least $(c_2 - c_1)$ in the branch values. The mechanism combines positive reward coupling in (2.6), bounded strategy sets (2.5), and the loss of strict curvature at λ^* from Remark 3.3. Consequently, reward factors slightly above λ^* form the most sensitive region for incentive tuning.

Remark 3.5 (Robustness beyond the quadratic model). The closed-form characterization of λ^* in (3.3) is specific to the stationary quadratic setting, where best responses are affine and the equilibrium reduces

to a scalar fixed-point equation. Beyond this setting, one may still expect qualitatively similar threshold behavior under positively coupled rewards and bounded efforts, but such extensions are not established in the present paper. In particular, linear costs may lead to bang-bang responses and multiplicity, while sufficiently strong convexity may restore uniqueness and smoother dependence on λ . However, in non-quadratic models, any analogous transition point or jump size would in general have to be characterized implicitly and investigated numerically.

3.4. Computation of NEs via the best-response algorithm

The best-response algorithm [5,23,34] is a widely employed and easily implemented numerical method for finding an NE in potential games. Here, we present the best-response algorithm adapted to our framework, as outlined in Algorithm 1.

Algorithm 1 Best-response algorithm.

```

1: Initialization:  $s^0 \in \hat{S}$ ;
2: for  $k = 1, \dots$  do
3:   for  $i = 1, \dots, m$  do
4:     if  $s_i^{k-1} \notin \arg \max_{s_i \in \hat{S}_i} \hat{P}_i(s_1^k, \dots, s_{i-1}^k, s_i, s_{i+1}^{k-1}, \dots, s_m^{k-1})$  then
5:        $s_i^k \in \arg \max_{s_i \in \hat{S}_i} \hat{P}_i(s_1^k, \dots, s_{i-1}^k, s_i, s_{i+1}^{k-1}, \dots, s_m^{k-1})$ ;
6:     else
7:        $s_i^k = s_i^{k-1}$ ;
8:     end if
9:   end for
10:  if  $s^k = s^{k-1}$  then
11:    return  $s^k$ .
12:  end if
13: end for

```

A classical convergence result for Algorithm 1 pertains to finite potential games [5, Cor. 2.2, Lem. 2.3]. We extend this result to continuous potential games under certain assumptions, as detailed in Assumption 5.1. These results offer convergence guarantees for Algorithm 1 when applied to our FL games.

Recall the definitions of $\bar{\lambda}$ and λ^* from (3.2) and (3.3). We have the following theorem on the convergence of Algorithm 1 applied to the game $\hat{\Gamma}_{\text{FL}}$.

Theorem 3.2. (Convergence of Algorithm 1) Assume that $\lambda \in (0, \bar{\lambda})$ and $\lambda \neq \lambda^*$. For $k \geq 1$, let s^k be the result of Algorithm 1 at the k -th iteration. Then, the sequence $\{s^k\}_{k \geq 1}$ converges to the unique NE of $\hat{\Gamma}_{\text{FL}}$.

Proof. The game $\hat{\Gamma}_{\text{FL}}$ satisfies Assumption 5.1. By applying the general convergence result of the best-response algorithm from Theorem 5.1(2), we obtain that any limit point of the sequence $\{s^k\}_{k \geq 1}$ is an NE of $\hat{\Gamma}_{\text{FL}}$. By Theorem 3.1, if $\lambda \neq \lambda^*$, this NE is unique, which implies that the limit point of $\{s^k\}_{k \geq 1}$ is unique. Combining this with the pre-compactness of $\{s^k\}_{k \geq 1}$, the conclusion follows. $\square$

Building on Theorem 5.1, the following remark establishes the convergence of Algorithm 1 for the specific case $\lambda = \lambda^*$, serving as a complement to Theorem 3.2.

Remark 3.6. Recall the definitions of c_1 and c_2 from (3.4). If $c_1 \geq c_2$, the convergence result in Theorem 3.2 holds for $\lambda = \lambda^*$ (see Theorem 3.1(2a)). Additionally, if $c_1 < c_2$, at the point $\lambda = \lambda^*$, any cluster point of the sequence $\{s^k\}_{k \geq 1}$ is an NE of $\hat{\Gamma}_{\text{FL}}$. In all these cases, for any $K \geq 1$, there exists $k \in \{1, \dots, K\}$ such that s^k is an $\mathcal{O}(1/K)$ -NE of $\hat{\Gamma}_{\text{FL}}$. Here, we refer to Definition 5.1 for the definition of an ϵ -NE.

3.5. Discussion of game models

In this subsection, we discuss both theoretical extensions and the practical interpretation of the FL game. On the theoretical side,

we consider non-uniform data sizes and time-dependent (round-heterogeneous) strategies. On the practical side, we clarify how α_i , q_i , and Q_i can be calibrated and how the model should be interpreted as a reduced-form incentive mechanism, with data-dependent effects incorporated implicitly through parameters.

Remark 3.7 (Non-uniform data sizes). In Definition 3.1, we assume that all clients have equal data sizes to simplify the constants and the analysis in Theorem 3.1. In the non-uniform case, the existence of an NE is guaranteed by Theorem 2.1. Regarding uniqueness, by following the proof in Section 5, we can establish the same results as in Theorem 3.1 and Corollary 3.1, with the constants introduced in (3.2)-(3.4) and (3.8) replaced by

$$\bar{\lambda} = \min_{i \in [m]} \{\alpha_i / \rho_i\}; \quad \lambda^* \in (0, \bar{\lambda}), \text{ s.t. } \sum_{i=1}^m \frac{\lambda^*}{2\alpha_i / \rho_i - \lambda^*} = 1;$$

$$c_1 = \max_{i \in [m]} \left\{ \frac{2\alpha_i q_i}{\lambda^*} - \rho_i q_i \right\}, \quad c_2 = \min_{i \in [m]} \left\{ \frac{2\alpha_i Q_i}{\lambda^*} - \rho_i Q_i \right\};$$

$$\lambda_1 = 2 \min_{i \in [m]} \left\{ \frac{\alpha_i q_i}{\bar{q} + \rho_i q_i} \right\}, \quad \lambda_2 = 2 \max_{i \in [m]} \left\{ \frac{\alpha_i Q_i}{\bar{Q} + \rho_i Q_i} \right\},$$

where $\bar{q} = \sum_{i \in [m]} \rho_i q_i$ and $\bar{Q} = \sum_{i \in [m]} \rho_i Q_i$. Note that if $\rho_i = 1/m$ for all $i \in [m]$ (i.e., the uniform case), then we recover the constants defined in Section 3.2. In addition, we obtain the same results as in Theorem 3.2 and Remark 3.6 for the convergence of Algorithm 1 in this non-uniform case.

Remark 3.8 (Time-dependent games). A key assumption for the uniqueness results presented in this section is the stationarity of training efforts with respect to the communication rounds, i.e., s_i^t is independent of t . However, in the time-dependent case (the original game Γ_{FL} with strategy sets (2.4)), uniqueness can easily be lost. For instance, suppose that there exists a time-dependent NE s^* with rounds $t_1, t_2 \in [T]$ such that $(s^*)^{t_1} \neq (s^*)^{t_2}$. Then, from the potential function (2.7), we can verify that $\tilde{s}^*$ (which differs from s^*) is also an NE of the game Γ_{FL} , where $\tilde{s}^*$ is obtained by swapping the training efforts of s^* at rounds t_1 and t_2 .

Although the uniqueness result does not hold, Algorithm 1 still converges in the time-dependent case, as in Remark 3.6. Let s^k denote the outcome of iteration k in Algorithm 1 adapted to Γ_{FL} . Then, any cluster point of s^k is an NE of Γ_{FL} .

Remark 3.9 (Practical calibration and interpretation). In practice, the server can calibrate α_i , q_i , and Q_i via device profiling, historical participation logs, or lightweight surveys. At the modeling level, however, we emphasize that the proposed game is a reduced-form incentive model. Equilibrium efforts are determined by cost-reward considerations and do not explicitly encode the full local data distribution or the detailed state of the global model. This decoupling is deliberate, because it preserves a transparent potential-game structure while avoiding direct access to private data statistics.

Data-dependent effects can still be incorporated implicitly through model parameters. In the baseline design, ρ_i is tied to relative data size, but it can also be interpreted more broadly as a contribution weight that reflects data quality or usefulness. Likewise, α_i can represent an effective participation cost, including computation, energy, and opportunity costs associated with local data value. One can also introduce model-state dependence by multiplying (2.6) by a round-dependent factor, for example $\gamma_t = 1/\sqrt{t}$, so that incentives are stronger in early rounds and weaker in later rounds.

Therefore, the resulting NE should be interpreted as a tractable proxy for local computation rather than as a closed-form predictor of final model accuracy. Our theory characterizes how incentives shift equilibrium effort. The induced reward-performance relation is then assessed empirically in Section 4. Developing fully data-aware reward rules remains a natural extension (see Section 6).

4. Numerical experiments

In this section, we present numerical experiments for the FL game and the induced FL training process. We first describe the experimental setup. We then investigate the NEs (i.e., the clients' rational training efforts) of the stationary game $\hat{\Gamma}_{\text{FL}}$ under different reward factors λ . Finally, we incorporate the resulting equilibrium efforts into FL training on different datasets and models, thereby validating the practical relevance of the proposed game-theoretic framework. The code for reproducing all experiments is available at <https://github.com/DCN-FAU-AvH/FL-Potential-Game>.

4.1. Experimental setup

FL game scenarios. We consider a stationary FL game with $m \in \{1,000, 3,000\}$ clients over $T = 200$ rounds. The server selects a reward factor $\lambda \in [0, 5]$, and each client i has a local cost coefficient $\alpha_i \in [1, 2]$. We set the training effort s_i to be the number of local epochs, with a fixed minimum effort $q_i = 1$ and a maximum feasible effort $Q_i \in [20, 30]$. This choice is representative of FL practice, where at least one local epoch is required once a client participates. The corresponding thresholds λ_1 , λ^* , and λ_2 are summarized in Table 2.

Datasets and algorithms. We consider representative FL tasks, including image classification on *MNIST* [35], *CIFAR-10*, and *CIFAR-100* [36], sentiment analysis on *IMDB* [37], and character recognition on *FEMNIST* [38]. Training is conducted using two widely adopted FL algorithms, namely FedAvg [1] and FedProx [29].

Data partitioning and training. For *CIFAR-10*, *CIFAR-100*, and *IMDB*, the data are uniformly distributed across 1000 clients, and 10% of the clients are sampled in each round. To induce strong label skew, each client is restricted to data from at most two labels, which yields a pathological non-IID distribution. We use a batch size of 20 and a learning rate of 0.05. For *FEMNIST*, we consider 3000 clients and sample 1% of the clients in each round. This dataset is inherently user-partitioned, naturally non-IID, and highly imbalanced. We use a batch size of 50 and a learning rate of 0.001 for this dataset.

Neural network architectures. For image classification, we use a convolutional neural network with two convolutional layers of 32 and 64 channels, followed by a fully connected layer of size 512, following McMahan et al. [1]. For sentiment analysis, we use an encoder-only Transformer with a 128-dimensional embedding layer and two encoder layers, each equipped with 2-head self-attention and a feedforward sublayer of size 128.

Hardware. All experiments are conducted on a server equipped with an AMD EPYC 7662 processor, 64 GB RAM, and an NVIDIA A100 GPU.

4.2. Evolution of NEs and critical reward factors

In this subsection, we validate the theoretical results in Section 3.2 by examining how the average effort $\bar{s}^*$ varies with the server's reward factor λ . The NEs are computed using Algorithm 1 and are compared with the theoretical values derived from the fixed-point system (5.2). The results shown in Fig. 4 indicate that the equilibria obtained from the theoretical characterization and the best-response method are indistinguishable up to numerical precision, thereby validating both the fixed-point characterization and the best-response solver.

Furthermore, as shown in Fig. 4, the average effort $\bar{s}^*$ increases monotonically with λ for both $m = 1,000$ and $m = 3,000$ clients. In the early regime with $\lambda < \lambda_1$, $\bar{s}^*$ remains at its minimum value $\bar{q}$, as shown in Fig. 4c. In the late regime with $\lambda > \lambda_2$, it saturates at the maximum value $\bar{Q}$, as shown in Fig. 4e. These observations are consistent with

Corollary 3.1 and suggest that the practically relevant range of λ lies in (λ_1, λ_2) , where the model avoids both under-training and saturation.

For the intermediate regime with $\lambda \in (\lambda_1, \lambda_2)$, Fig. 4d shows that $\bar{s}^*$ exhibits a pronounced jump from c_1 to c_2 at λ^* . According to Theorem 3.1(2b), every point between c_1 and c_2 corresponds to an NE. This observation indicates that reward factors slightly above λ^* are especially effective at stimulating collective effort. We examine this implication further in the FL training experiments of the next subsection. Moreover, the jump at λ^* is observed consistently in both the 1,000-client setting in Fig. 4a and the 3,000-client setting in Fig. 4b, which suggests that the theoretical prediction remains robust as the number of clients increases.

Overall, these results highlight the strong sensitivity of the NEs, namely the clients' rational training efforts, to the server's reward factor. To connect the above equilibrium analysis with FL training, we next select four representative equilibrium cases around the thresholds λ_1 , λ^* , and λ_2 , and apply the corresponding effort profiles to practical FL tasks to evaluate their effectiveness.

4.3. Training performance of FL under various NEs

In this subsection, we evaluate how different NEs affect FL training performance across benchmark datasets, models, and algorithms. We emphasize that the proposed equilibrium mechanism is not a new FL algorithm that prescribes clients' training efforts. Rather, it is a modeling framework for the incentive layer. Specifically, when rewards are absent or weak, clients may have limited willingness to participate actively, and the game is used to characterize their rational effort responses under server-side incentives. Accordingly, the goal of this subsection is to assess the practical relevance of this modeling framework by examining how the resulting equilibrium effort profiles affect FL training performance.

We begin with four representative NE cases located near the three critical thresholds λ_1 , λ^* , and λ_2 . Specifically, Case 1 is chosen before the activation point λ_1 . Case 2 is chosen immediately before the jump point λ^* . Case 3 is chosen immediately after the jump point. Case 4 is chosen beyond the saturation point λ_2 . We then apply the equilibrium efforts $(s_i^*)_{i \in [m]}$ under these four cases to various FL settings. In all experiments, each client performs $\lceil s_i^* \rceil$ epochs of local training, where $\lceil \cdot \rceil$ denotes the ceiling function.

The specific values of λ , the average efforts $\bar{s}^*$, and the test results for the four cases are summarized in Table 2 and Fig. 5. The four cases exhibit a broadly consistent pattern across the experiments. The most substantial improvement occurs when moving from Case 2 to Case 3, which underscores the critical role of the jump point λ^* .

The main observations are summarized below.

1. **Case 1, selected before the activation point λ_1 .** This case yields the weakest performance. Both the accuracy and loss curves indicate insufficient training, which is consistent with the low effort level of the clients.
2. **Case 2, selected before the jump point λ^* .** Performance remains close to that of Case 1. The gain is limited overall, which suggests that the modest increase in equilibrium effort is still insufficient to produce a clear improvement in training quality.
3. **Case 3, selected after the jump point λ^* .** This case shows a marked improvement in training performance, with a clear increase in accuracy and a simultaneous reduction in loss. This pattern appears consistently across the tasks and supports the practical relevance of the theoretical jump at λ^* .
4. **Case 4, selected after the saturation point λ_2 .** Further increases in training effort yield only marginal additional gains in accuracy and only slight reductions in loss. This behavior indicates diminishing returns once the reward factor has passed the jump region.

Overall, these results suggest that increasing the reward factor λ can improve training performance, but the most substantial gains occur around the jump point λ^* . Beyond this region, additional incentives

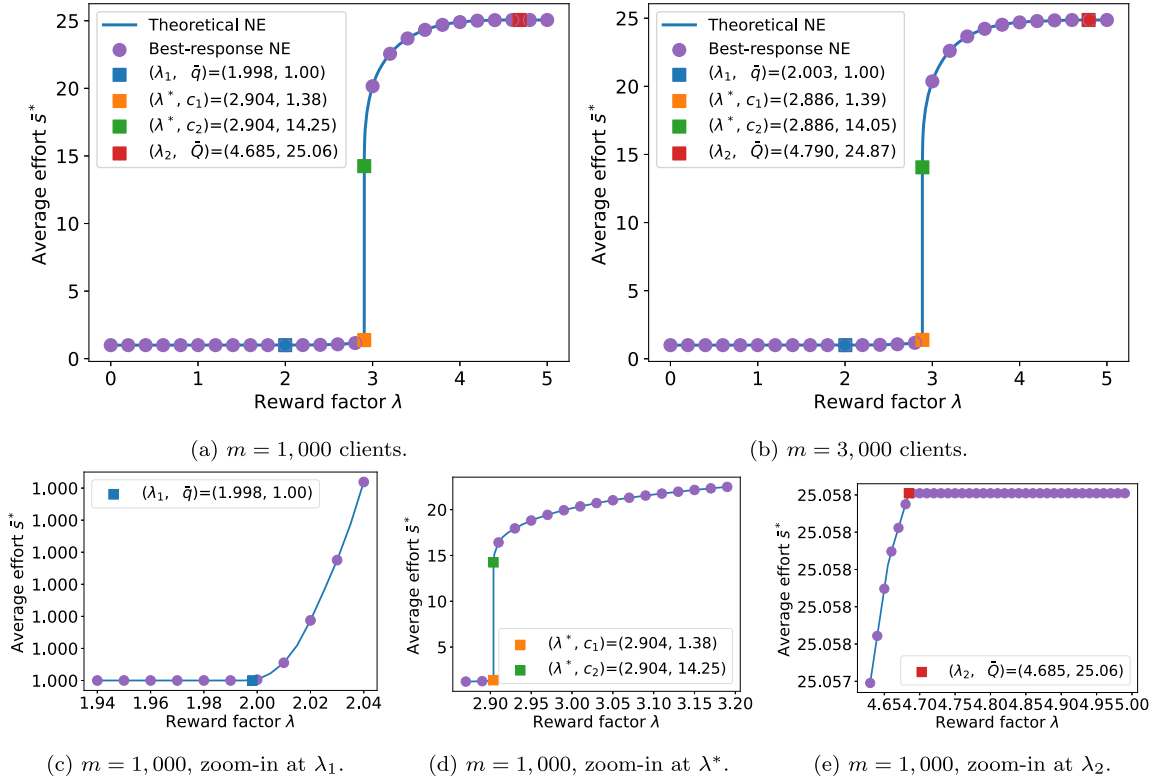


Fig. 4. Average training effort $\bar{s}^*$ as a function of the server's reward factor λ . The activation point λ_1 , jump point λ^* , and saturation point λ_2 are indicated in the legends.

Table 2

Summary of experimental cases. Cases 1–4 correspond, respectively, to the pre-activation, pre-jump, post-jump, and post-saturation regimes, that is, before λ_1 , immediately before λ^* , immediately after λ^* , and after λ_2 . Test accuracies are averaged across different FL algorithms.

Case	Regime	Balanced, non-IID, $m = 1,000$					Imbalanced, non-IID, $m = 3,000$		
		$\lambda_1 = 1.998, \lambda^* = 2.904, \lambda_2 = 4.685$					$\lambda_1 = 2.003, \lambda^* = 2.886, \lambda_2 = 4.790$		
		NE	Test accuracy				NE	Test accuracy	
		λ	$\bar{s}^*$	CIFAR-10	CIFAR-100	IMDB	λ	$\bar{s}^*$	FEMNIST
Case 1	pre-activation	1.99	1.00	0.434	0.120	0.535	2.00	1.00	0.124
Case 2	pre-jump	2.90	1.32	0.477	0.142	0.523	2.88	1.32	0.232
Case 3	post-jump	2.91	16.41	0.558	0.208	0.796	2.89	15.94	0.665
Case 4	post-saturation	4.69	25.05	0.561	0.212	0.819	4.80	24.87	0.706

provided by the server lead only to limited further improvement. If the goal is to obtain a large performance gain without moving deep into the saturation regime, then reward factors slightly above λ^* are natural candidates.

5. Technical proofs and theorems

This section presents the detailed proofs of all theoretical results stated in the paper. We first rewrite the NE conditions as an explicit best-response system and reduce the m -dimensional equilibrium problem to a scalar fixed-point equation for the mean effort. We then study how this scalar equation depends on the reward factor λ . We finally establish convergence guarantees for the best-response algorithm in potential games.

5.1. Proofs of Theorem 3.1 and Corollary 3.1

Recall the parameters $\lambda, m, \alpha_i, q_i$, and Q_i of the game $\hat{\Gamma}_{\text{FL}}$ from Definition 3.1 and Example 2.2. Recall also the payoff function $\hat{P}_i$ in (3.1).

The next lemma derives the fixed-point conditions satisfied by the NEs of $\hat{\Gamma}_{\text{FL}}$ from the quadratic form of $\hat{P}_i$.

Lemma 5.1. *A point $s^* \in \prod_{i \in [m]} [q_i, Q_i]$ is an NE of the game $\hat{\Gamma}_{\text{FL}}$ if and only if the following holds:*

$$s_i^* = \begin{cases} q_i & , \quad \text{if } q_i \geq \frac{\lambda(\bar{s}^* - s_i^*/m)}{2(\alpha_i - \lambda/m)} \text{ and } \alpha_i > \frac{\lambda}{m}, \\ \frac{\lambda \bar{s}^*}{2\alpha_i - \lambda/m}, & \text{if } q_i < \frac{\lambda(\bar{s}^* - s_i^*/m)}{2(\alpha_i - \lambda/m)} < Q_i, \\ Q_i & , \quad \text{if } Q_i \leq \frac{\lambda(\bar{s}^* - s_i^*/m)}{2(\alpha_i - \lambda/m)} \text{ or } \alpha_i \leq \frac{\lambda}{m}, \end{cases} \quad (5.1a)$$

$$s_i^* = \frac{\lambda \bar{s}^*}{2\alpha_i - \lambda/m}, \quad \text{if } q_i < \frac{\lambda(\bar{s}^* - s_i^*/m)}{2(\alpha_i - \lambda/m)} < Q_i, \quad (5.1b)$$

$$s_i^* = Q_i, \quad \text{if } Q_i \leq \frac{\lambda(\bar{s}^* - s_i^*/m)}{2(\alpha_i - \lambda/m)} \text{ or } \alpha_i \leq \frac{\lambda}{m}, \quad (5.1c)$$

where $\bar{s}^* = \sum_{i \in [m]} s_i^*/m$.

The next lemma provides an equivalent formulation of (5.1a)–(5.1c), which will be used in the proof of Theorem 3.1. To this end, for each $i \in [m]$, define the auxiliary mapping $\beta_i: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ by

$$\beta_i(x) := \begin{cases} q_i & , \quad \text{if } q_i \geq \frac{\lambda x}{2\alpha_i - \lambda/m} \text{ and } \alpha_i > \frac{\lambda}{m}, \\ \frac{\lambda x}{2\alpha_i - \lambda/m}, & \text{if } q_i < \frac{\lambda x}{2\alpha_i - \lambda/m} < Q_i, \\ Q_i & , \quad \text{if } Q_i \leq \frac{\lambda x}{2\alpha_i - \lambda/m} \text{ or } \alpha_i \leq \frac{\lambda}{m} \end{cases} \quad (5.2a)$$

$$\beta_i(x) := \frac{\lambda x}{2\alpha_i - \lambda/m}, \quad \text{if } q_i < \frac{\lambda x}{2\alpha_i - \lambda/m} < Q_i, \quad (5.2b)$$

$$\beta_i(x) := Q_i, \quad \text{if } Q_i \leq \frac{\lambda x}{2\alpha_i - \lambda/m} \text{ or } \alpha_i \leq \frac{\lambda}{m} \quad (5.2c).$$

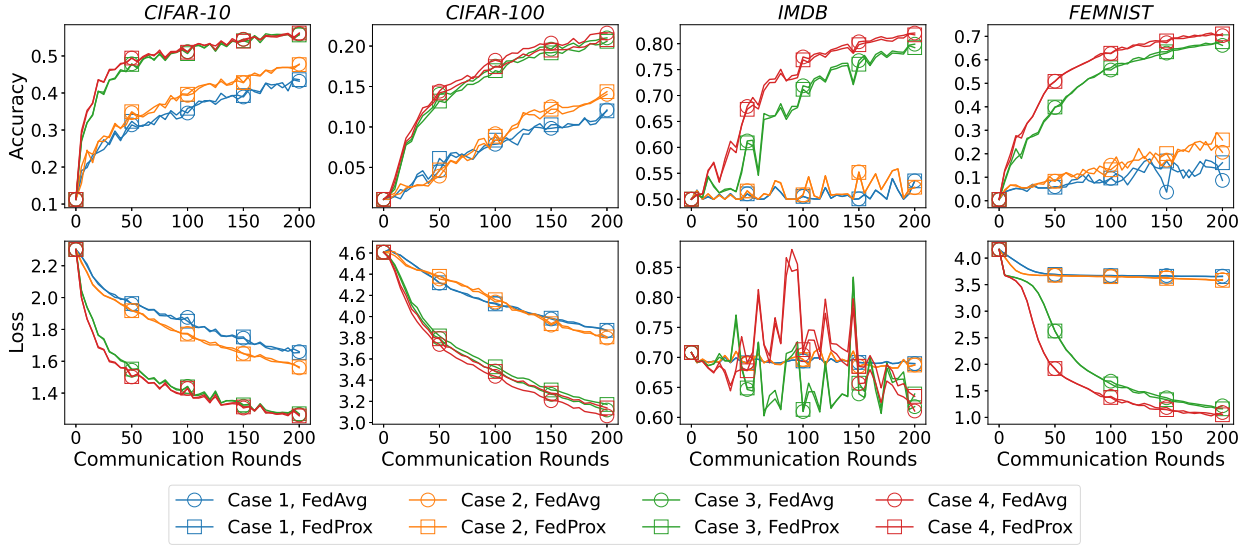


Fig. 5. Results of FL training under four NE cases across different datasets and algorithms. The largest gain is observed when moving from Case 2 to Case 3 in all experiments.

Lemma 5.2. Let s^* be an NE of the game $\hat{\Gamma}_{\text{FL}}$ and let $\bar{s}^* = \sum_{i \in [m]} s_i^*/m$. Then, it holds:

$$s_i^* = \beta_i(\bar{s}^*), \quad \text{for } i \in [m]. \quad (5.1)$$

As a consequence, $\bar{s}^*$ satisfies the following fixed-point equation:

$$\bar{s}^* = \frac{1}{m} \sum_{i \in [m]} \beta_i(\bar{s}^*). \quad (5.2)$$

Proof. By Lemma 5.1, s^* and $\bar{s}^*$ satisfy (5.1a)-(5.1c). For $i \in [m]$ such that $s_i^* = q_i$, it follows from (5.1a) that

$$q_i \geq \frac{\lambda(\bar{s}^* - q_i/m)}{2(\alpha_i - \lambda/m)} \quad \text{and} \quad \alpha_i > \frac{\lambda}{m},$$

which is equivalent to (5.2a) by a direct calculation. This implies that $s_i^* = \beta_i(\bar{s}^*)$ for these values of i . By a similar argument, we obtain that $s_i^* = \beta_i(\bar{s}^*)$ for $i \in [m]$ such that $s_i^* = Q_i$. For $i \in [m]$ such that $s_i^* = \lambda \bar{s}^* / (2\alpha_i - \lambda/m)$, the equality

$$\frac{\lambda(\bar{s}^* - s_i^*/m)}{2(\alpha_i - \lambda/m)} = \frac{\lambda \bar{s}^*}{2\alpha_i - \lambda/m},$$

implies the equivalence between (5.1b) and (5.2b), which gives $s_i^* = \beta_i(\bar{s}^*)$ in this case. $\square$

From (5.2a)-(5.2c), we define the following four set-valued mappings on $\mathbb{R}_+$:

$$I_1(x) := \left\{ i \in [m] \mid q_i \geq \frac{\lambda x}{2\alpha_i - \lambda/m} \text{ and } \alpha_i > \frac{\lambda}{m} \right\};$$

$$I_2(x) := \left\{ i \in [m] \mid q_i < \frac{\lambda x}{2\alpha_i - \lambda/m} < Q_i \right\};$$

$$I_3(x) := \left\{ i \in [m] \mid Q_i \leq \frac{\lambda x}{2\alpha_i - \lambda/m} \text{ and } \alpha_i > \frac{\lambda}{m} \right\};$$

$$I_4(x) := \left\{ i \in [m] \mid \alpha_i \leq \frac{\lambda}{m} \right\}.$$

By construction, the conditions in (5.2a)-(5.2c) are equivalent to $i \in I_1(x)$, $i \in I_2(x)$, and $i \in I_3(x) \cup I_4(x)$, respectively.

Lemma 5.3. Fix any $x \in \mathbb{R}_+$. The following statements hold:

1. The sets $I_1(x)$, $I_2(x)$, $I_3(x)$, and $I_4(x)$ are mutually disjoint, and

$$I_1(x) \cup I_2(x) \cup I_3(x) \cup I_4(x) = [m]. \quad (5.3)$$

2. For any $y \in \mathbb{R}_+$ such that $x \leq y$, we have

$$I_1(y) \subseteq I_1(x), \quad \text{and} \quad I_3(x) \subseteq I_3(y). \quad (5.4)$$

Proof. The proof is straightforward by the definitions of I_1 , I_2 , I_3 , and I_4 . $\square$

Proof. (Proof of Theorem 3.1.) By Lemma 5.2, every equilibrium is determined by the scalar mean effort $\bar{s}$ together with the partition I_1 , I_2 , I_3 , and I_4 . Assume that two distinct equilibria exist and let their means satisfy $\bar{s}^1 < \bar{s}^2$. If $\lambda \neq \lambda^*$, the monotonicity of the index sets and the two fixed-point identities yield incompatible bounds on $\sum_i \lambda / (2\alpha_i - \lambda)$, which gives a contradiction. If $\lambda = \lambda^*$, the same reduction shows that non-uniqueness can only occur on the interior affine branch $s_i = \lambda^* c / (2\alpha_i - \lambda^*/m)$, and the comparison between c_1 and c_2 distinguishes the unique case from the continuum case.

Step 1 (The case $\lambda \in (0, \bar{\lambda})$ and $\lambda \neq \lambda^*$). Assume that there exist two distinct NEs, denoted by $\bar{s}^1$ and $\bar{s}^2$. Let $\bar{s}^1 = \sum_{i \in [m]} s_i^1/m$ and $\bar{s}^2 = \sum_{i \in [m]} s_i^2/m$. By (5.1), we have $s_i^1 = \beta_i(\bar{s}^1)$ and $s_i^2 = \beta_i(\bar{s}^2)$ for all $i \in [m]$. Since $\bar{s}^1 \neq \bar{s}^2$, we have $\bar{s}^1 \neq \bar{s}^2$. Without loss of generality, assume that $\bar{s}^1 < \bar{s}^2$. By the definition of $\bar{\lambda}$ in (3.2) and the inequality $\lambda < \bar{\lambda}$, we have $I_4(\bar{s}^1) = I_4(\bar{s}^2) = \emptyset$. Then (5.2), (5.3), and (5.2a)-(5.2c) imply

$$m\bar{s}^1 = \sum_{i \in I_1(\bar{s}^1)} q_i + \sum_{i \in I_2(\bar{s}^1)} \frac{\lambda \bar{s}^1}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^1)} Q_i, \quad (5.5)$$

$$m\bar{s}^2 = \sum_{i \in I_1(\bar{s}^2)} q_i + \sum_{i \in I_2(\bar{s}^2)} \frac{\lambda \bar{s}^2}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^2)} Q_i. \quad (5.6)$$

Since $\bar{s}^1 < \bar{s}^2$, (5.4) yields $I_1(\bar{s}^2) \subseteq I_1(\bar{s}^1)$ and $I_3(\bar{s}^1) \subseteq I_3(\bar{s}^2)$. Combining this with (5.3), we obtain

$$I_2(\bar{s}^1) \cup I_1(\bar{s}^1) \setminus I_1(\bar{s}^2) = I_2(\bar{s}^2) \cup I_3(\bar{s}^2) \setminus I_3(\bar{s}^1) =: I. \quad (5.7)$$

By (5.2a) and (5.2c), it holds:

$$q_i \geq \frac{\lambda \bar{s}^1}{2\alpha_i - \lambda/m}, \quad \text{for } i \in I_1(\bar{s}^1) \setminus I_1(\bar{s}^2), \quad (5.8)$$

$$Q_i \leq \frac{\lambda \bar{s}^2}{2\alpha_i - \lambda/m}, \quad \text{for } i \in I_3(\bar{s}^2) \setminus I_3(\bar{s}^1). \quad (5.9)$$

Substituting (5.8) into (5.5) and (5.9) into (5.6), and using (5.7), we obtain

$$m\bar{s}^1 \geq \sum_{i \in I_1(\bar{s}^2)} q_i + \sum_{i \in I} \frac{\lambda \bar{s}^1}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^1)} Q_i, \quad (5.10)$$

$$m\bar{s}^2 \leq \sum_{i \in I_1(\bar{s}^2)} q_i + \sum_{i \in I} \frac{\lambda \bar{s}^2}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^1)} Q_i. \quad (5.11)$$

Taking the difference between (5.11) and (5.10), and noting that $\bar{s}^1 < \bar{s}^2$, we have

$$1 \leq \sum_{i \in I} \frac{\lambda}{2m\alpha_i - \lambda}. \quad (5.12)$$

On the other hand, (5.10) implies that if $I_1(\bar{s}^2) \cup I_3(\bar{s}^1) \neq \emptyset$, then

$$1 > \sum_{i \in I} \frac{\lambda}{2m\alpha_i - \lambda}.$$

This contradicts (5.12). Therefore, $I_1(\bar{s}^2) \cup I_3(\bar{s}^1) = \emptyset$, so $I = [m]$. Consequently, (5.10) and (5.12) give

$$1 = \sum_{i \in [m]} \frac{\lambda}{2m\alpha_i - \lambda},$$

which contradicts the assumption $\lambda \neq \lambda^*$, by the definition of λ^* in (3.3). Hence the NE is unique in this case.

Step 2 (The case $\lambda \geq \bar{\lambda}$). Assume that s^1 and s^2 are two distinct NEs and that $\bar{s}^1 < \bar{s}^2$. Arguing as in (5.5)-(5.6) and using $\lambda \geq \bar{\lambda}$, we obtain

$$m\bar{s}^1 = \sum_{i \in I_1(\bar{s}^1)} q_i + \sum_{i \in I_2(\bar{s}^1)} \frac{\lambda \bar{s}^1}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^1)} Q_i + \sum_{i \in I_4(\bar{s}^1)} Q_i,$$

$$m\bar{s}^2 = \sum_{i \in I_1(\bar{s}^2)} q_i + \sum_{i \in I_2(\bar{s}^2)} \frac{\lambda \bar{s}^2}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^2)} Q_i + \sum_{i \in I_4(\bar{s}^2)} Q_i.$$

Since $I_4(x)$ is independent of x , we have $I_4(\bar{s}^1) = I_4(\bar{s}^2) \neq \emptyset$. By the same argument as in Step 1, we obtain (5.12), namely,

$$1 \leq \sum_{i \in I} \frac{\lambda}{2m\alpha_i - \lambda}.$$

In addition, an argument analogous to (5.10) gives

$$\begin{aligned} m\bar{s}^1 &\geq \sum_{i \in I_1(\bar{s}^2)} q_i + \sum_{i \in I} \frac{\lambda \bar{s}^1}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^1)} Q_i + \sum_{i \in I_4(\bar{s}^1)} Q_i \\ &\geq \sum_{i \in I} \frac{\lambda \bar{s}^1}{2\alpha_i - \lambda/m} + \sum_{i \in I_4(\bar{s}^1)} Q_i \\ &> \sum_{i \in I} \frac{\lambda \bar{s}^1}{2\alpha_i - \lambda/m}, \end{aligned}$$

where the last inequality is strict because $I_4(\bar{s}^1)$ is nonempty. This contradicts (5.12), and therefore the NE is unique in this case.

Step 3 (The case $\lambda = \lambda^*$ and $c_1 \geq c_2$). Assume that s^1 and s^2 are two distinct NEs. By the argument in Step 1, we obtain $I_1(\bar{s}^2) \cup I_3(\bar{s}^1) = \emptyset$. Using $I_3(\bar{s}^1) = \emptyset$ in (5.5), and recalling that λ^* solves (3.3), we obtain

$$\sum_{i \in I_1(\bar{s}^1)} \frac{\lambda^* \bar{s}^1}{2\alpha_i - \lambda^*/m} = \sum_{i \in I_1(\bar{s}^1)} q_i.$$

Combining with the definition of $I_1(\bar{s}^1)$, it follows that

$$\frac{\lambda^* \bar{s}^1}{2\alpha_i - \lambda^*/m} = q_i, \quad \text{for } i \in I_1(\bar{s}^1).$$

Given the fact that $I_3(\bar{s}^1) = \emptyset$, we deduce that

$$q_i \leq \frac{\lambda^* \bar{s}^1}{2\alpha_i - \lambda^*/m} < Q_i, \quad \text{for } i \in [m].$$

This is equivalent to

$$\max_{i \in [m]} \left\{ \frac{2\alpha_i q_i}{\lambda^*} - \frac{q_i}{m} \right\} \leq \bar{s}^1, \quad \text{and} \quad \bar{s}^1 < \min_{i \in [m]} \left\{ \frac{2\alpha_i Q_i}{\lambda^*} - \frac{Q_i}{m} \right\}.$$

Recalling the definitions of c_1 and c_2 in (3.4), the previous inequality yields $c_1 \leq \bar{s}^1 < c_2$, which contradicts the assumption $c_1 \geq c_2$. Therefore, the NE is unique.

Step 4 (The case $\lambda = \lambda^*$ and $c_1 < c_2$). Take any $s^* \in \hat{S}$ as in (3.5). Thus, for some $c \in (c_1, c_2)$,

$$s_i^* = \frac{\lambda^* c}{2\alpha_i - \lambda^*/m}, \quad \forall i \in [m].$$

It follows that

$$\bar{s}^* = \frac{1}{m} \sum_{i \in [m]} s_i^* = c \in (c_1, c_2).$$

By direct calculation,

$$\frac{\lambda^* (\bar{s}^* - s_i^*/m)}{2(\alpha_i - \lambda^*/m)} = \frac{c \lambda^*}{2\alpha_i - \lambda^*/m}, \quad \text{for } i \in [m].$$

Since $c \in (c_1, c_2)$, it follows that

$$q_i < \frac{\lambda^* (\bar{s}^* - s_i^*/m)}{2(\alpha_i - \lambda^*/m)} < Q_i, \quad \text{for } i \in [m].$$

Combining this with (3.3), we deduce that s^* satisfies the fixed-point system (5.1a)-(5.1c), and hence s^* is an NE. This proves the sufficiency of (3.5).

Since c can be chosen arbitrarily in (c_1, c_2) , the game $\hat{\Gamma}_{\text{FL}}$ has infinitely many NEs. In particular, the game $\hat{\Gamma}_{\text{FL}}$ has at least two distinct NEs. Then, by an argument similar to that in Step 3, we deduce that for any NE of $\hat{\Gamma}_{\text{FL}}$, denoted by s' , it holds that

$$q_i \leq \frac{\lambda^* s'_i}{2\alpha_i - \lambda^*/m} \leq Q_i, \quad \text{for } i \in [m], \quad (5.13)$$

where $\bar{s}' = \sum_{i \in [m]} s'_i/m$. It also follows from (5.13) and (3.4) that $\bar{s}' \in (c_1, c_2)$. Moreover, by (5.1), $s'_i = \beta_i(\bar{s}')$ for all i . Combining with (5.13), we obtain that

$$s'_i = \frac{\lambda^* \bar{s}'}{2\alpha_i - \lambda^*/m} \quad \text{for } i \in [m].$$

The necessity of condition (3.5) follows. $\square$

Proof of Corollary 3.1. We combine Theorem 3.1 with the fixed-point representation established above. For small and large values of λ , the boundary profiles $(q_i)_{i=1}^m$ and $(Q_i)_{i=1}^m$ satisfy the equilibrium conditions directly. For the intermediate regimes, the same partition argument yields the bounds $\bar{s}^* \leq c_1$ for $\lambda < \lambda^*$ and $\bar{s}^* \geq c_2$ for $\lambda > \lambda^*$.

Step 1 (Proof of points (1) and (4)). Assume that $\lambda \in (0, \lambda_1)$. Since $\lambda_1 \leq \lambda^*$, Theorem 3.1 implies that the NE is unique. It therefore suffices to prove that $(q_i)_{i=1}^m$ is an NE. Let $\bar{q} = \sum_{i \in [m]} q_i/m$. By the definition of λ_1 in (3.8), we obtain

$$q_i \geq \frac{\lambda(\bar{q} - q_i/m)}{2(\alpha_i - \lambda/m)}, \quad \text{for } i \in [m]. \quad (5.14)$$

Therefore, $(q_i)_{i=1}^m$ is an NE because it satisfies the fixed-point system (5.1a)-(5.1c). The case $\lambda \in (\lambda_2, +\infty)$ is proved in the same way, using the definition of λ_2 in (3.8).

Step 2 (Proof of point (2)). Assume that $\lambda < \lambda^*$. Let s^* be the unique NE, and $\bar{s}^* = \sum_{i \in [m]} s_i^*/m$. Since $\lambda^* < \bar{\lambda}$, we have $I_4(\bar{s}^*) = \emptyset$. By (5.2), we obtain that

$$m\bar{s}^* = \sum_{i \in I_1(\bar{s}^*)} q_i + \sum_{i \in I_2(\bar{s}^*)} \frac{\lambda \bar{s}^*}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^*)} Q_i.$$

If $I_1(\bar{s}^*) = \emptyset$, then

$$m\bar{s}^* = \sum_{i \in I_2(\bar{s}^*)} \frac{\lambda \bar{s}^*}{2\alpha_i - \lambda/m} + \sum_{i \in I_3(\bar{s}^*)} Q_i \leq \sum_{i \in [m]} \frac{\lambda \bar{s}^*}{2\alpha_i - \lambda/m}.$$

This implies

$$1 \leq \sum_{i \in [m]} \frac{\lambda}{2\alpha_i - \lambda/m},$$

which contradicts the assumption $\lambda < \lambda^*$. Therefore, $I_1(\bar{s}^*) \neq \emptyset$. Hence

$$m\bar{s}^* \leq \sum_{i \in I_1(\bar{s}^*)} q_i + \sum_{i \in [m] \setminus I_1(\bar{s}^*)} \frac{\lambda \bar{s}^*}{2\alpha_i - \lambda/m} \leq \sum_{i \in I_1(\bar{s}^*)} q_i + \sum_{i \in [m] \setminus I_1(\bar{s}^*)} \frac{\lambda^* \bar{s}^*}{2\alpha_i - \lambda^*/m}.$$

Suppose that

$$\bar{s}^* > c_1 = \max_{i \in [m]} \left\{ \frac{2\alpha_i q_i}{\lambda^*} - \frac{q_i}{m} \right\}. \quad (5.15)$$

Therefore, we prove that s^k is an $(L^2\epsilon^2/(2\alpha_{\min}))$ -NE. From step 1, fixing any $K \geq 1$, there exists $k_0 \leq K$ such that

$$\|s^{k_0} - s^{k_0-1}\|^2 \leq \frac{2}{\delta K}(P^* - P(s^0)),$$

where P^* is the maximum value of P over S and is finite. Taking $k = k_0$, it follows that s^{k_0} is an ϵ' -NE, where

$$\epsilon' = \frac{L^2}{2\alpha_{\min}} \frac{2}{\delta K}(P^* - P(s^0)) = \frac{L^2(P^* - P(s^0))}{\delta \alpha_{\min}} \frac{1}{K}.$$

Step 3 (Proof of point (2)). By the compactness of S , assume that a subsequence of $\{s^k\}_{k \geq 1}$, denoted by $\{s^{q(k)}\}_{k \geq 1}$, converges to some point $s^* \in S$. By step 1, we have that $\|s^{q(k)-1} - s^{q(k)}\| \rightarrow 0$. Similar to step 2, we have

$$P_i(s_i, s_{-i}^*) - P_i(s_i^*, s_{-i}^*) \leq \langle \nabla_i P_i(s_i^*, s_{-i}^*), s_i - s_i^* \rangle - \frac{\alpha_{\min}}{2} \|s_i - s_i^*\|^2.$$

The previous first-order term is decomposed by

$$\begin{aligned} \langle \nabla_i P_i(s_i^*, s_{-i}^*), s_i - s_i^* \rangle &= \langle \nabla_i P_i(s_i^{q(k)}, s_{-i}^{q(k)}), s_i - s_i^{q(k)} \rangle + \langle \nabla_i P_i(s_i^*, s_{-i}^*), s_i^{q(k)} - s_i^* \rangle \\ &\quad - \langle \nabla_i P_i(s_i^*, s_{-i}^*), s_i - s_i^{q(k)} \rangle \\ &\leq L(\|s^{q(k)} - s^{q(k)-1}\| + \|s^* - s^{q(k)}\|) \|s_i - s_i^{q(k)}\| \\ &\quad + \|\nabla_i P_i(s_i^*, s_{-i}^*)\| \|s_i^{q(k)} - s_i^*\|, \end{aligned}$$

where the last inequality follows from a similar argument to step 2 and the Lipschitz continuity of $\nabla_i P_i$. By passing k to infinity, we deduce that $P_i(s_i, s_{-i}^*) - P_i(s_i^*, s_{-i}^*) \leq 0$ for any i and s_i . Therefore, s^* is an NE. $\square$

6. Conclusions and future work

In this paper, we introduce a potential game framework to model the interplay among training costs, incentives, and effort selection in federated learning (FL). We prove the existence of Nash equilibria (NEs) in the proposed game and further investigate the uniqueness of the NE in a stationary scenario where clients maintain constant efforts throughout training. Additionally, we characterize how the NEs evolve with the server's reward factor, identifying three critical thresholds: the activation point, the jump point, and the saturation point. Notably, our findings highlight the significance of the jump point, at which clients' training efforts increase sharply, suggesting that it provides a natural target for server-side reward tuning. Extensive numerical experiments further support our theoretical results. Overall, our work provides a robust foundation for understanding the interplay between clients' rational training efforts and the server's incentives.

For future work, we propose the following perspectives.

1. A first extension of our game model could focus on incorporating a discount factor into the server's reward scheme based on the number of rounds. This aligns with the convergence properties of widely used stochastic gradient algorithms in machine learning, where significant gains in model performance are typically observed during the early training stages [41, Section 4]. This phenomenon is also reflected in our simulation results shown in Fig. 5, where major improvements occur in the first several rounds. Furthermore, adding a discount factor may also help address the issue of non-uniqueness of the NE in heterogeneous cases, as the counterexample mentioned in Remark 3.8 no longer applies in this scenario.
2. A second direction focuses on the asymptotic behavior of our FL game as the number of clients m approaches infinity. In this scenario, a nonatomic version of Γ_{FL} or $\hat{\Gamma}_{\text{FL}}$ could be analyzed. The NEs in this setting are characterized by an optimality condition over the distribution of training efforts among a continuum of clients, see [42, Thm. 1]. It is worth noting that proving the uniqueness of the NE in the nonatomic case is generally challenging when the price increases with respect to the aggregate, as discussed in [42, Prop. 3]. The techniques used to prove the uniqueness results in Section 5 can be naturally adapted to these nonatomic games.

3. Extending our game model to other aspects of FL, such as fairness [12] and privacy [43], is a promising avenue for future research. For example, in a differential privacy scenario, each client's strategy can be the standard deviation of their noise. Therefore, we can model the trade-off between data privacy (which increases with the noise level) and model accuracy (which decreases with the noise level) using our FL game, with the necessary adaptations.
4. Recent studies identify open challenges in federated foundation models [44] and federated reasoning-oriented large language models (LLMs) [45]. Our server-tunable potential game framework naturally extends to these settings by modeling clients' adaptation and reasoning budgets as multidimensional efforts and shaping rewards through global utility proxies such as validation gains, calibration or consistency scores, and safety metrics under bandwidth and risk constraints. Central research questions include the existence and uniqueness of equilibria with coupled efforts and the design of optimal incentives under active alignment or safety constraints.
5. Finally, beyond FL, extending our game model to other applications is also of interest. The current game formulation should be well suited to modeling systems with positive consumption externalities [46]. For example, in technology and social media platforms, the value of the platform to each user increases as more users participate. As a result, such platforms often design incentives that scale with total engagement, thereby reinforcing collective participation.

CRedit authorship contribution statement

Kang Liu: Writing – review & editing, Methodology, Formal analysis, Conceptualization; **Ziqi Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Enrique Zuazua:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Data availability

The code for reproducing all experiments is available at: <https://github.com/DCN-FAU-AvH/FL-Potential-Game>.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors' names are listed in alphabetical order by family name to signify equal contributions. The work was partially supported by the **European Research Council** (ERC) under the European Union's Horizon 2030 research and innovation programme (grant agreement NO: 101096251-CoDeFeL); by the Alexander von Humboldt Professorship program; the European Union's Horizon Europe MSCA project ModConFlex (HORIZON-MSCA-2021-DN-01 project 101073558); the Transregio 154 Project "Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks" of the DFG; the AFOSR 24IOE027 project; the SURE-AI Norwegian Centre for Sustainable, Risk-Averse, and Ethical AI grant 357482, Research Council of Norway; by the Grant PID2023-146872OB-I00-DyCMAMod of MICIU (Spain) and by the COST Actions CA24122 - Multiscale Stochastics, Patterns, and Analysis of Combinatorial Environments and CA24136 - Interactions between Control Theory and Machine Learning.

References

- [1] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas, Communication-efficient learning of deep networks from decentralized data, in:

- Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.
- [2] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, D. Ramage, Federated Learning for Mobile Keyboard Prediction, arXiv preprint arXiv:1811.03604 (2018).
 - [3] C. Daskalakis, P.W. Goldberg, C.H. Papadimitriou, The complexity of computing a nash equilibrium, *Commun. ACM* 52 (2) (2009) 89–97.
 - [4] J. Hofbauer, W.H. Sandholm, Stable games and their dynamics, *J. Econ. Theory* 144 (4) (2009) 1665–1693.
 - [5] D. Monderer, L.S. Shapley, Potential games, *Games Econ. Behav.* 14 (1) (1996) 124–143.
 - [6] E.N. Barron, *Game Theory: An Introduction*, Wiley, 2 edition, 2013.
 - [7] M.J. Osborne, A. Rubinstein, *A Course in Game Theory*, MIT Press, Cambridge, Mass, Cambridge, Mass, 1994.
 - [8] R. Gupta, J. Gupta, Federated learning using game strategies: state-of-the-Art and future trends, *Comput. Netw.* 225 (2023) 109650.
 - [9] S. Banerjee, H. Konishi, T. Sönmez, Core in a simple coalition formation game, *Soc. Choice Welf.* 18 (1) (2001) 135–153.
 - [10] A. Bogomolnaia, M.O. Jackson, The stability of hedonic coalition structures, *Games Econ. Behav.* 38 (2) (2002) 201–230.
 - [11] K. Donahue, J. Kleinberg, Model-sharing games: analyzing federated learning under voluntary participation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, 2021, pp. 5303–5311.
 - [12] K. Donahue, J. Kleinberg, Optimality and stability in federated learning: a game-Theoretic approach, in: *Advances in Neural Information Processing Systems*, 34, 2021, pp. 1287–1298.
 - [13] L.U. Khan, S.R. Pandey, N.H. Tran, W. Saad, Z. Han, M.N.H. Nguyen, C.S. Hong, Federated learning for edge networks: resource optimization and incentive mechanism, *IEEE Commun. Mag.* 58 (10) (2020) 88–93.
 - [14] J. Lee, D. Kim, D. Niyato, Market analysis of distributed learning resource management for internet of things: a game-Theoretic approach, *IEEE Internet Things J.* 7 (9) (2020) 8430–8439.
 - [15] L. Dong, Y. Zhang, Federated learning service market: a game theoretic analysis, in: *2020 International Conference on Wireless Communications and Signal Processing (WCSP)*, 2020, pp. 227–232.
 - [16] Q. Li, Z. Liu, Q. Li, K. Xu, martFL: enabling utility-driven data marketplace with a robust and verifiable federated learning architecture, in: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, ACM*, 2023, pp. 1496–1510.
 - [17] Z. Li, B. Ding, L. Yao, Y. Li, X. Xiao, J. Zhou, Performance-Based pricing for federated learning via auction, *Proc. VLDB Endow.* 17 (6) (2024) 1269–1282.
 - [18] T. Lu, B. Zhang, K. Ren, Privdata network: a privacy-preserving on-chain data factory and trading market, *IEEE Trans. Dependable Secure Comput.* 21 (3) (2024) 1424–1436.
 - [19] L.S. Shapley, A value for n-Person games, in: H.W. Kuhn, A.W. Tucker (Eds.), *Contributions to the Theory of Games, Volume II*, 28 of *Annals of Mathematics Studies*, Princeton University Press, Princeton, NJ, 1953, pp. 307–317.
 - [20] T. Song, Y. Tong, S. Wei, Profit allocation for federated learning, in: *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 2019, pp. 2577–2586.
 - [21] Z. Liu, Y. Chen, H. Yu, Y. Liu, L. Cui, GTG-Shapley: efficient and accurate participant contribution evaluation in federated learning, *ACM Trans. Intell. Syst. Technol.* 13 (4) (2022) 1–21.
 - [22] H. Yu, Z. Liu, Y. Liu, T. Chen, M. Cong, X. Weng, D. Niyato, Q. Yang, A fairness-Aware incentive scheme for federated learning, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 393–399.
 - [23] B. Swenson, R. Murray, S. Kar, On best-response dynamics in potential games, *SIAM J. Control Optim.* 56 (4) (2018) 2734–2767.
 - [24] T. Liu, J. Wang, X. Zhang, D. Cheng, Game theoretic control of multiagent systems, *SIAM J. Control Optim.* 57 (3) (2019) 1691–1709.
 - [25] A. Britzelmeier, A. Dreves, M. Gerdt, Numerical solution of potential games arising in the control of cooperative automatic vehicles, in: *Proceedings of the 2019 Conference on Control and Its Applications*, SIAM, 2019, pp. 38–45.
 - [26] H. Wu, H. Shang, Potential game for dynamic task allocation in multi-Agent system, *ISA Trans.* 102 (2020) 208–220.
 - [27] A. Raschella, F. Bouhafs, M. Mackay, Q. Shi, J. Ortín, J.R. Gállego, M. Canales, A dynamic access point allocation algorithm for dense wireless LANs using potential game, *Comput. Netw.* 167 (2020) 106991.
 - [28] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, *Found. Trends/textregistered Mach. Learn.* 14 (1–2) (2021) 1–210.
 - [29] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: *Proceedings of Machine Learning and Systems*, 2, 2020, pp. 429–450.
 - [30] A.A. Cournot, *The Mathematical Principles of the Theory of Wealth*, Macmillan, New York, NY, New York, NY, 1927. Translated by Nathaniel T. Bacon.
 - [31] F. Szidrovsky, S. Yakowitz, A new proof of the existence and uniqueness of the cournot equilibrium, *Int. Econ. Rev.* 18 (3) (1977) 787–789.
 - [32] J.B. Rosen, Existence and uniqueness of equilibrium points for concave N-Person games, *Econometrica* 33 (3) (1965) 520–534.
 - [33] P. Cardaliaguet, Notes on Mean Field Games, Technical Report, Université Paris-Dauphine, 2010. Lecture notes.
 - [34] S. Durand, B. Gaujal, Complexity and optimality of the best response algorithm in random potential games, in: *International Symposium on Algorithmic Game Theory*, Springer, 2016, pp. 40–51.
 - [35] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2324.
 - [36] A. Krizhevsky, G. Hinton, Learning Multiple Layers of Features from Tiny Images, Technical Report TR-2009-44, University of Toronto, Toronto, ON, Canada, 2009.
 - [37] A. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 142–150.
 - [38] S. Caldas, S.M.K. Duddu, P. Wu, T. Li, J. Konečný, H.B. McMahan, V. Smith, A. Talwalkar, LEAF: A benchmark for federated settings, in: *Proceedings of the 2Nd Workshop on Federated Learning for Data Privacy and Confidentiality*, 2018.
 - [39] Y. Xu, W. Yin, A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion, *SIAM J. Imaging Sci.* 6 (3) (2013) 1758–1789.
 - [40] K. Liu, N. Oudjane, C. Wan, Approximate nash equilibria in large nonconvex aggregative games, *Math. Oper. Res.* 48 (3) (2023) 1791–1809.
 - [41] L. Bottou, F.E. Curtis, J. Nocedal, Optimization methods for large-Scale machine learning, *SIAM Rev.* 60 (2) (2018) 223–311.
 - [42] M.-W. Cheung, R. Lahkar, Nonatomic potential games: the continuous strategy case, *Games Econ. Behav.* 108 (2018) 341–362.
 - [43] X. Zhang, L. Fan, S. Wang, W. Li, K. Chen, Q. Yang, A game-Theoretic framework for privacy-Preserving federated learning, *ACM Trans. Intell. Syst. Technol.* 15 (3) (2024) 1–35.
 - [44] T. Fan, H. Gu, X. Cao, C.S. Chan, Q. Chen, Y. Chen, Y. Feng, Y. Gu, J. Geng, B. Luo, et al., Ten challenging problems in federated foundation models, *IEEE Trans. Knowl. Data Eng.* (2025).
 - [45] S. Wei, Y. Tong, Z. Zhou, Y. Xu, J. Gao, T. Wei, T. He, W. Lv, Federated reasoning LLMs: a survey, *Front. Comput. Sci.* 19 (12) (2025) 1912613.
 - [46] M.L. Katz, C. Shapiro, Network externalities, competition, and compatibility, *Am. Econ. Rev.* 75 (3) (1985) 424–440.