arXiv:2607.23029v1 [cs.LG] 25 Jul 2026

Multi-Agent Privacy Game in Federated Learning: A Unified Mean-Field View

Kun Zhao
kun.zhao@vumc.org
Xu Chen
xc2412@columbia.edu

Abstract

Federated learning enables collaborative model training across distributed clients without centralising their data, yet privacy remains a persistent concern because the shared model updates can leak information about local datasets. Existing privacy-preserving methods either inject calibrated noise into client updates, limiting their composition guarantees, or formulate client privacy choices as a multi-agent game whose Nash equilibrium becomes intractable as the number of clients grows. We bridge these two lines of work by formulating privacy-preserving federated learning as a mean-field privacy game: each client strategically chooses its own privacy budget while interacting with the population only through a single mean-field statistic. The mean-field limit yields a tractable equilibrium for arbitrarily many clients, accommodates heterogeneous client preferences, and inherits an exponentially decaying privacy guarantee through a log-Sobolev contraction. The framework recovers the entropic privacy baseline as the homogeneous special case and the multi-agent privacy game as the finite-population case. Experiments on quadratic regression, logistic regression, and MNIST demonstrate that the proposed framework attains the privacy-utility trade-off of the entropic baseline while delivering a personalised privacy guarantee that the homogeneous baseline cannot express.

1 Introduction

Federated Learning (FL) [19, 17] trains a shared model across data-owning clients without centralizing raw data. Although locality reduces some privacy risks, model updates leak through gradient inversion and membership inference [35, 10, 7]. Three largely independent lines of work address this leakage.

Where does the control act? Privacy-preserving FL methods can be partitioned by where each client's control variable enters the system: it can act on the agent's *state* (samples or gradients used for local updates) or directly on the *model output* (an additive perturbation of w_k at the aggregation step). DP-SGD [1] and its federated variants [11, 30] act on the state: a calibrated Gaussian is added to each clipped gradient, and the privacy cost accumulates polynomially with rounds under standard composition [21]. FLRA [23] likewise applies its strategic perturbation $(\Lambda^i \mathbf{x} + \delta^i)$ to the state, though for robustness rather than privacy. In contrast, the parameter-perturbation game [34] treats δ_k as a control on the model output and aggregates via $w = \sum_k p_k(w_k + \delta_k)$; under FedAvg's linear constraint, KKT collapses every Nash point to $\delta_k = \mathbf{0}$ (Section 4). Only controls that act on the state admit non-trivial equilibria, so this is where the rest of the paper lives.

Mean-field as the $N \to \infty$ limit. Two parallel lines of work address different aspects of the resulting state-control game. The multi-agent privacy game [34, 33, 9] fixes a finite N and asks for a Nash equilibrium of the controls $\{\varepsilon_k\}$, where ε_k scales the noise added to client k's gradient or sample (the MAPG-DP and MAPG-input formulations of [34]). Mean-Field Entropic Privacy (MFEP) [5] fixes

Preprint.

a homogeneous ε and lets $N \to \infty$, replacing additive noise with an entropic Wasserstein gradient flow that admits exponential log-Sobolev contraction [22, 15]. Neither line on its own captures the joint behaviour: MAPG-DP is intractable for realistic N and uses geometry-agnostic Gaussian noise, while MFEP cannot represent heterogeneous privacy preferences.

Our framing. We treat privacy-preserving FL as a standard mean-field stochastic differential game [16, 8]: each agent has a state $X_t \in \mathbb{R}^d$ (a privatised datum), a type β (its privacy preference), and a control ε_k (its privacy budget); the state distribution $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ aggregates the population. The empirical state distribution $\mu_t^{(N)} = \frac{1}{N}\sum_k \mu_t^k$ converges, by propagation of chaos [8], to a deterministic μ_t as $N \to \infty$. This single limit organises three previously separate works:

	finite N (multi-agent)	**$N \to \infty$ (mean-field)**
no game (homogeneous)	DP-SGD [1]	**MFEP** [5]
game (heterogeneous control ε_k)	**MAPG-DP** [34]	**MFPG (this paper)**

The bottom-right cell is the missing piece. **MFEP is MFPG without the game** (single ε across the population); **MAPG-DP is MFPG without the mean-field limit** (finite N). MFPG closes both gaps with one construction.

Contributions.

(i) **Unified MFG framework (Section 3).** We formalise privacy-preserving FL as a mean-field stochastic differential game in the standard sense, with state X_t, type β_k, control ε_k, state distribution μ_t, and running cost $f(x, \mu, \varepsilon; \beta) = \ell + \beta\delta_{\mathrm{dp}}$. A finite-$N$ Nash equilibrium of MAPG-DP and the mean-field Nash equilibrium of MFPG are the two endpoints of the same $N \to \infty$ limit; MFEP is the no-game special case.

(ii) **Mean-Field Privacy Game (Section 5).** We prove existence of an MFNE on a finite action grid via Kakutani's theorem and certify $(\epsilon_{\mathrm{dp}}, \delta_{\mathrm{dp}})$-DP at the equilibrium, with δ_{dp} contracting exponentially when $\alpha\varepsilon^* > \lambda + G$. The bound recovers the MFEP guarantee at homogeneous β_k and recovers MAPG-DP's heterogeneity at finite N.

(iii) **Common solver and empirical study (Sections 6, 7).** A single Particle–Sinkhorn JKO step services every $N \to \infty$ variant; the action update is the only block that varies between MFEP and MFPG. We compare all four cells on quadratic, logistic, and MNIST benchmarks, with every reported number traced to `results/full_benchmark.csv`.

2 Preliminaries

Federated learning. Each client $k \in [N]$ holds a local dataset $\mathcal{D}_k$. The global objective is $\min_w F(w) = \sum_k p_k F_k(w)$ with $p_k = n_k/n$ and $F_k(w) = n_k^{-1}\sum_i \ell(w; x_i^k, y_i^k)$ [19]. At round t the server broadcasts w^t, each client returns a local update w_k^{t+1}, and the server aggregates via FedAvg $w^{t+1} = \sum_k p_k w_k^{t+1}$.

Differential privacy. A randomised mechanism $\mathcal{M}$ is (ϵ, δ)-DP if, for all neighbouring datasets and all measurable S, $\Pr[\mathcal{M}(\mathcal{D}) \in S] \le e^\epsilon \Pr[\mathcal{M}(\mathcal{D}') \in S] + \delta$. The Gaussian mechanism with $\sigma = C\sqrt{2\ln(1.25/\delta)}/\epsilon$ is (ϵ, δ)-DP for sensitivity C [1]; under T-fold composition ϵ grows as $O(\sqrt{T})$.

Mean-field games. A mean-field game [16, 13, 8] models a continuum of identical agents whose individual optimal control problem depends on the population distribution μ_t. As $N \to \infty$, the N-player Nash equilibrium converges to a mean-field Nash equilibrium (MFNE). The MFNE depends only on μ_t, not on individual identities, which makes it tractable when the N-player problem is not.

Wasserstein gradient flows and JKO. For a free-energy functional $\mathcal{F}$ on $\mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein gradient flow obeys $\partial_t \mu_t = -\nabla_{W_2}\mathcal{F}(\mu_t)$ and admits a time-discrete Jordan–Kinderlehrer–Otto scheme [15] $\mu_{k+1} = \arg\min_\rho\{\mathcal{F}(\rho) + (2\tau)^{-1}W_2^2(\rho, \mu_k)\}$. A measure ν satisfies a log-Sobolev inequality with constant α if $\mathrm{Ent}_\nu(f^2) \le (2/\alpha)\int \|\nabla f\|^2 d\nu$, with $\nu = \mathcal{N}(0, \sigma^2 I)$ giving $\alpha = 1/\sigma^2$.

Otto–Villani implies W_2-contraction at rate $\alpha\varepsilon$ for the entropic flow, which we use in Section 5 to certify privacy.

3 A Unified Mean-Field Game Framework

We formalise privacy-preserving FL as a mean-field stochastic differential game in the standard sense of [16, 8]. Section 3.1 identifies the four ingredients of the game (state, type, control, dynamics) and Section 3.2 specialises them to FL through two privacy mechanisms; the resulting 2×2 grid in Table 1 populates DP-SGD, MFEP, MAPG-DP, and our proposed MFPG as four named cells.

3.1 State, type, control, and dynamics

A representative agent in our game is described by:

- *State* $X_t \in \mathbb{R}^d$: a (privatised) datum used by an individual client to compute its local update at round t.
- *Type* $\beta \in [\beta_{\min}, \beta_{\max}] \subset \mathbb{R}_+$: the client's privacy preference (heterogeneity parameter), distributed by $\rho(d\beta)$.
- *Control* $\varepsilon_k \in \mathcal{E} = \{\varepsilon^{(1)}, \dots, \varepsilon^{(m)}\}$: a per-client privacy budget chosen on a finite grid; this is the strategic variable of the game.
- *Dynamics:* conditional on ε_k, the state evolves under the controlled SDE

$$dX_t^k = b\big(X_t^k, \mu_t, \varepsilon_k\big)\, dt + \sigma\big(\varepsilon_k\big)\, dW_t, \tag{3.1}$$

  with W_t a standard d-dimensional Brownian motion. The drift b and diffusion σ are fixed by the privacy mechanism $\mathcal{M}$ (Section 3.2).
- *State distribution* $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$: the distribution of X_t across the agent population. The empirical measure $\mu_t^{(N)} = \frac{1}{N}\sum_{k=1}^N \mu_t^k$ converges weakly to a deterministic μ_t as $N \to \infty$ by propagation of chaos [8, 13].
- *Running cost* (per agent of type β):

$$f(x, \mu, \varepsilon; \beta) = \ell(x; w_t) + \beta\, \delta_{\mathrm{dp}}(\varepsilon), \tag{3.2}$$

  combining the training loss ℓ against the current global model w_t with the privacy term $\beta\, \delta_{\mathrm{dp}}$ scaled by the agent's type. The disutility minimised by client k is the time-integral of (3.2), written compactly as

$$U_k(\varepsilon_k;\, \bar{\varepsilon}) = L_k(w;\, \varepsilon_k) + \beta_k\, \delta_{\mathrm{dp}}(\varepsilon_k), \tag{3.3}$$

  where L_k is the cumulative training loss and $\bar{\varepsilon} = \mathbb{E}_\mu[\varepsilon]$ is the mean control across the population.

The model parameters w enter only through $\ell(x; w_t)$ and are otherwise external to the game; the strategic content lives entirely in $(X_t, \mu_t, \varepsilon_k)$ and (3.2).

3.2 Two privacy mechanisms: Gaussian and entropic

The drift–diffusion pair (b, σ) in (3.1) is determined by which privacy mechanism the system runs. We consider two:

- *Gaussian mechanism.* The control ε_k scales the additive noise on each clipped gradient, $\sigma(\varepsilon_k) = C\sqrt{2\ln(1.25/\delta)}/\varepsilon_k$, with drift b given by the clipped loss gradient [1]. Privacy is tracked by Rényi-DP composition, giving a polynomially accumulating $\epsilon_{\mathrm{dp}} = O(\sqrt{T})$.
- *Entropic mechanism.* The state distribution evolves under the Wasserstein gradient flow of the KL-regularised free energy

$$\mathcal{F}_\lambda(\mu; \varepsilon_k) = \mathbb{E}_{x\sim\mu}[L(x; w)] + \varepsilon_k\, \mathrm{KL}(\mu\|\nu) + \tfrac{\lambda}{2}\, \mathrm{Var}_\mu[x], \tag{3.4}$$

  with Gaussian prior $\nu = \mathcal{N}(0, \sigma^2 I)$. Equivalently (Appendix C.1), the controlled SDE (3.1) has drift $b(x, \mu, \varepsilon) = -\nabla L(x; w) - \frac{\varepsilon}{\sigma^2}x - \lambda(x - \bar{x})$ and diffusion $\sigma(\varepsilon) = \sqrt{2\varepsilon}$. Privacy is tracked by log-Sobolev contraction [22], with δ_{dp} decaying exponentially when $\alpha\varepsilon_k > \lambda + G$.

Table 1: The MFG framework in two axes. Two design choices populate the 2×2 grid: client heterogeneity (rows: no game vs. game) and population size (columns: finite N vs. $N \to \infty$). Entries name the method, the active noise mechanism (Gaussian or Entropic), and the accountant.

	Finite N (multi-agent)	**$N \to \infty$ (mean field)**
No game (single ε)	**DP-SGD** [1] Gaussian noise on gradient, RDP δ const., $\epsilon = O(\sqrt{T})$	**MFEP** [5] Entropic flow on μ_t, LSI $\delta = O(e^{-\alpha\varepsilon K\tau})$
Game (heterogeneous β_k)	**MAPG-DP** [34] Gaussian noise, N-Nash on $\{\varepsilon_k\}$ Intractable for large N	**MFPG (ours)** Entropic flow, MFNE on $\bar{\varepsilon}$ Heterogeneous + exponential decay

Remark 3.1 (Why the control acts on the state and not on the model). A superficially natural alternative would be to make the control a perturbation $\delta_k \in \mathbb{R}^d$ added directly to the local model w_k at the aggregation step, $w = \sum_k p_k(w_k + \delta_k)$ [34]. With the disutility $L_k(w) - \beta_k\|\delta_k\|^2$, KKT forces $\delta_k = \mathbf{0}$ at every Nash point (Section 4), so the limiting MFG is degenerate. We therefore restrict attention to controls that act on the state SDE (3.1); this is the standard MFG setup in which the agent's strategy shapes its own state evolution.

4 The Three Baseline Cells

The three already-named cells of Table 1 are recalled below in the language of Section 3; MFPG itself is deferred to Section 5. The MAPG-DP subsection also discharges the model-output alternative, whose $\delta_k = 0$ pathology is what prevents the multi-agent privacy game literature from being combined with mean-field analysis directly.

4.1 Finite N, no game: DP-SGD

DP-SGD [1] runs the Gaussian mechanism of Section 3.2 with a single shared budget $\varepsilon = \varepsilon_{\mathrm{tgt}}$. After clipping $\Delta w_k \leftarrow \Delta w_k \cdot \min(1, C/\|\Delta w_k\|_2)$ each client adds Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ with $\sigma = C\sqrt{2\ln(1.25/\delta)}/\varepsilon$. RDP composition returns $\epsilon_{\mathrm{dp}} = O(\sqrt{T\log(1/\delta)}/\sigma)$, so the privacy cost *grows* polynomially with rounds T.

4.2 $N \to \infty$, no game: MFEP

MFEP [5] replaces external noise injection with the entropic free-energy (3.4) at a *fixed* regularization strength ε, evaluated against the state distribution μ_t. The Wasserstein gradient flow obeys the Fokker–Planck equation

$$\partial_t \mu_t = \nabla \cdot \Big[\mu_t \Big(\nabla L(x) + \tfrac{\varepsilon}{\sigma^2} x + \lambda(x - \bar{x}_t)\Big)\Big] + \varepsilon \Delta \mu_t, \tag{4.1}$$

with $\bar{x}_t = \mathbb{E}_{\mu_t}[x]$. Three mechanisms operate simultaneously: *loss-driven drift*, *prior attraction* at rate ε/σ^2, and *entropic diffusion* $\varepsilon\Delta\mu_t$ that delivers privacy intrinsically. Under LSI for the prior with constant $\alpha = 1/\sigma^2$, the JKO discretization contracts [22]: for neighbouring measures μ_k and μ_k' differing in a single client's data,

$$W_2(\mu_K, \mu_K') \le e^{-(\alpha\varepsilon - \lambda - G)K\tau}\, W_2(\mu_0, \mu_0'), \tag{4.2}$$

provided $\alpha\varepsilon > \lambda + G$, with G the clipped gradient norm bound. Converting through total-variation gives the exponentially decaying δ_{dp} bound used throughout the paper:

$$\delta_{\mathrm{dp}}(\varepsilon) \le \frac{C_d}{\sqrt{N}} \exp\Big(-\tfrac{(\alpha\varepsilon - \lambda - G)K\tau}{2}\Big), \quad C_d = \min(\sqrt{d}, 10). \tag{4.3}$$

4.3 Finite N, game: MAPG-DP

MAPG-DP [34, §§ 3.2.2, 3.3.4] gives each client k a per-client privacy budget $\varepsilon_k \in \mathbb{R}_+$ that scales the Gaussian noise added to its local gradient, with scale $\sigma_k(\varepsilon_k) = 2\eta C/\varepsilon_k$, or, equivalently, an

additive shift on its features as in FLRA [23]. The disutility (3.3) is delivered by RDP composition, and the N-player Nash equilibrium $(\varepsilon_1^*, \dots, \varepsilon_N^*)$ satisfies $U_k(\varepsilon_k^*, \varepsilon_{-k}^*) \le U_k(\varepsilon_k, \varepsilon_{-k}^*)$ for every k, where ε_{-k}^* denotes the strategies of all other clients at equilibrium. Existence follows by standard arguments, but computation is intractable for realistic N because each best response couples through every ε_j via the FedAvg constraint

$$w = \tfrac{1}{N}\sum_{j=1}^{N}\big(w_j + \mathcal{N}(0, \sigma_j(\varepsilon_j)^2 I)\big). \tag{4.4}$$

The same multi-agent literature also studies a superficially natural alternative in which the strategy is an additive perturbation $\delta_k \in \mathbb{R}^d$ on the local model w_k, with disutility $L_k(w) - \beta_k\|\delta_k\|^2$ subject to $w = \sum_k p_k(w_k + \delta_k)$ [34, § 3.3.1]. The KKT stationarity conditions in feature dimension m,

$$\frac{\partial \mathcal{L}}{\partial w_{k,m}} = \tfrac{1}{N}\,\lambda_{k,m}, \qquad \frac{\partial \mathcal{L}}{\partial \delta_{k,m}} = \tfrac{1}{N}\,\lambda_{k,m} - 2\beta_k\,\delta_{k,m}, \tag{4.5}$$

together force $\lambda_{k,m} = 0$ and hence $\delta_{k,m} = 0$ at every Nash point: the aggregation constraint cancels the privacy term. We therefore restrict the rest of the paper to controls that act on the state, since only such controls admit both non-trivial equilibria and the $N \to \infty$ limit developed next.

5 Mean-Field Privacy Game (MFPG)

We now instantiate the $N \to \infty$ limit of MAPG-DP. The construction below is structurally identical to MAPG-DP except that (i) the Gaussian noise mechanism is replaced by the entropic flow (4.1), which the LSI contraction (4.3) certifies, and (ii) the N-Nash equilibrium of $\{\varepsilon_k\}$ is replaced by the mean-field Nash equilibrium on the population mean $\bar{\varepsilon}$. MFEP is recovered as the homogeneous special case (β_k constant).

5.1 Formulation

Each client k chooses a control $\varepsilon_k \in \mathcal{E} = \{\varepsilon^{(1)}, \dots, \varepsilon^{(m)}\}$ on a finite grid; conditional on ε_k, its state slice μ_t^k evolves under the entropic flow (4.1) and contributes to the population state distribution μ_t through the loss $L_k(w; \varepsilon_k) = \mathbb{E}_{x\sim\mu_t^k}[\ell(w; x)]$. The empirical measure $\mu_t^{(N)} = \frac{1}{N}\sum_k \mu_t^k$ converges, by propagation of chaos [8, 13], to a deterministic $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ as $N \to \infty$, so each client's disutility (3.3) couples to the rest of the population only through the scalar mean-field strength $\bar{\varepsilon} = \mathbb{E}_\mu[\varepsilon]$ that replaces the vector ε_{-k} of the finite-N game. Throughout this section we use the disutility in the form

$$U_k(\varepsilon_k; \bar{\varepsilon}) = L_k(w; \varepsilon_k) + \beta_k\,\delta_{\mathrm{dp}}(\varepsilon_k), \tag{5.1}$$

with δ_{dp} now given by the LSI bound (4.3) rather than by RDP composition.

This $N \to \infty$ limit is well-posed only because the controls act on the state: each client's best response in (4.4) depends on ε_{-k} only through aggregate statistics of $\mu_t^{(N)}$, which collapse to $\bar{\varepsilon}$ in the limit. The model-output alternative does not admit such a limit, since each perturbation enters $w = \sum_j p_j(w_j + \delta_j)$ with vanishing weight $O(1/N)$ and the $\delta_k = 0$ collapse of (4.5) is its finite-N shadow.

5.2 Game equilibrium

A pair (μ^*, ε^*) is a *mean-field Nash equilibrium* (MFNE) of MFPG if (i) μ^* is the stationary state distribution of (4.1) at the population mean $\bar{\varepsilon}^* = \mathbb{E}_{\mu^*}[\varepsilon]$, and (ii) ε^* is a best response to $\bar{\varepsilon}^*$ for every client given its preference β_k (formal definition in Appendix C.4). At equilibrium, no client can reduce its disutility by unilaterally changing ε_k given that every other client plays ε_{-k}^*. For heterogeneous clients with distinct β_k, the MFNE is characterised by a fixed point of the averaged best-response map $\Phi : \bar{\varepsilon} \mapsto \mathbb{E}_k[\mathrm{BR}_k(\bar{\varepsilon})]$; existence follows from Kakutani's theorem on the convex hull of $\mathcal{E}$ (Appendix C.4). Section 5.3 gives an equivalent differential characterisation through coupled HJB and FPK PDEs.

5.3 HJB–FPK characterisation

The fixed-point characterisation of Section 5.2 is convenient for existence and for the finite-grid solver of Section 6, but it hides the dynamical content of the equilibrium. Following the standard mean-field game system of [16, 8] and the FL–MFG analogy of [20], we now give an equivalent differential characterisation through coupled forward Fokker–Planck–Kolmogorov (FPK) and backward Hamilton–Jacobi–Bellman (HJB) equations specialised to our disutility (5.1). To this end we momentarily relax the finite grid $\mathcal{E}$ to the interval $[\varepsilon_{\min}, \varepsilon_{\max}] \subset \mathbb{R}_+$ and parameterise each client by a privacy preference $\beta \in [\beta_{\min}, \beta_{\max}]$ distributed according to $\rho(d\beta)$, so that the state distribution splits across types as $\mu_t = \int \mu_t^\beta \, \rho(d\beta)$ with global mean $\bar{x}_t = \mathbb{E}_{\mu_t}[X]$. Under control ε, a representative state of type β obeys the controlled SDE

$$dX_t = b(X_t, \mu_t, \varepsilon)\, dt + \sqrt{2\varepsilon}\, dW_t, \qquad b(x, \mu, \varepsilon) = -\nabla L(x; w_t) - \tfrac{\varepsilon}{\sigma^2} x - \lambda(x - \bar{x}), \tag{5.2}$$

with W_t a standard Brownian motion, and seeks to minimise the cumulative cost

$$J^\mu(\varepsilon; \beta) = \mathbb{E}\left[\int_0^T \ell(X_t; w_t)\, dt\right] + \beta\, \delta_{\mathrm{dp}}(\varepsilon), \tag{5.3}$$

where $\delta_{\mathrm{dp}}(\varepsilon) = (C_d/\sqrt{N}) \exp(-\theta(\varepsilon - \varepsilon_0))$ with $\theta := \alpha K \tau/2$ and $\varepsilon_0 := (\lambda + G)/\alpha$ is the LSI bound (4.3).

Plugging the optimal feedback control $\varepsilon^*(t, x; \beta)$ derived below into (5.2) gives the forward type-conditional Fokker–Planck equation

$$\partial_t \mu_t^\beta(x) + \nabla \cdot \Big[\mu_t^\beta(x)\, b\big(x, \mu_t, \varepsilon^*(t, x; \beta)\big)\Big] = \varepsilon^*(t, x; \beta)\, \Delta \mu_t^\beta(x); \tag{5.4}$$

marginalising over ρ recovers (4.1) but with the optimal control in place of a fixed ε. Defining the type-conditional value function $V(t, x; \beta) = \inf_\varepsilon \mathbb{E}_{X_t = x}[\,\cdot\,]$ of (5.3), dynamic programming yields the backward HJB

$$-\partial_t V(t, x; \beta) = \mathcal{H}\big(x, \nabla V, D^2 V, \mu_t; \beta\big), \qquad V(T, x; \beta) = 0, \tag{5.5}$$

with Hamiltonian

$$\mathcal{H}(x, p, M, \mu; \beta) = \ell(x; w_t) - p \nabla L(x; w_t) - \lambda\, p(x - \bar{x}) + \min_{\varepsilon \geq 0} \Big\{ \beta\, \delta_{\mathrm{dp}}(\varepsilon) - \tfrac{\varepsilon}{\sigma^2}\, px + \varepsilon\, \mathrm{Tr}(M) \Big\}. \tag{5.6}$$

The first-order condition for the inner minimisation in $\mathcal{H}$ admits a closed-form solution $\varepsilon^*(t, x; \beta)$, derived in Appendix C.7. Two qualitative properties of this solution carry the intuition of the result: ε^* is decreasing in β, so privacy-sensitive clients adopt stronger regularisation, and ε^* is decreasing in the local curvature $\mathrm{Tr}(D^2 V)$, so clients near sharp minima can afford stronger noise. The MFNE is the pair (V, μ_t^β) that simultaneously satisfies (5.5) and (5.4), coupled through ε^* in the FPK drift–diffusion and through μ_t in the Hamiltonian; integrating ε^* against the equilibrium population recovers the scalar fixed point $\bar{\varepsilon}^* = \Phi(\bar{\varepsilon}^*)$ that drives the discrete-grid solver in Section 6.

Two specialisations of the system above recover the existing literature. If $\beta_k \equiv \beta$ is constant across the population, the type-conditional structure collapses, ε^* becomes spatially constant on the optimum, and (5.4) reduces to the single-strength entropic flow of [5]. If we instead replace (5.4) by its N-particle empirical version and (5.5) by the N-player backward Bellman system, with δ_{dp} replaced by RDP composition, we recover the dynamic MAPG-DP of [34]. MFPG is therefore the joint $N \to \infty$ and entropic-mechanism limit of MAPG-DP.

When the activation condition $\alpha\varepsilon^* > \lambda + G$ holds at the equilibrium, the LSI contraction yields an exponentially decaying δ_{dp} bound of the form $(C_d/\sqrt{N}) \exp(-(\alpha\varepsilon^* - \lambda - G) K\tau/2)$ (Appendix C.8); compared with DP-SGD, whose ϵ_{dp} grows as $O(\sqrt{K})$ at fixed $\delta = 10^{-5}$, MFPG's privacy guarantee *tightens* with the number of training rounds. This combines the heterogeneity benefit MFEP cannot express with the exponential decay MAPG-DP cannot offer.

6 Algorithms

A single training round of every cell in Table 1 consists of the same three steps applied independently by each client and then composed by the server: an *action update* that selects the strategic variable ε_k

Algorithm 1 Unified training round (one round, all four cells).

Input: Global params $w^{(t)}$; population mean $\bar{\varepsilon}^{(t)}$ (mean-field cells only); per-client preferences $\{\beta_k\}$; active mechanism (Gaussian / Entropic) and accountant (RDP / LSI).

for each client $k \in [N]$ in parallel **do**
  $\varepsilon_k^{(t)} \leftarrow \textsc{ActionUpdate}_k(\bar{\varepsilon}^{(t)}, \{\beta_j\})$ ▷ skipped if no game
  $w_k^{(t+1)} \leftarrow \textsc{LocalUpdate}_{\mathcal{M}}(w^{(t)}, \varepsilon_k^{(t)})$ ▷ Alg. 2 or 3
  $\eta \leftarrow \textsc{AccountantUpdate}(\eta, \varepsilon_k^{(t)})$ ▷ RDP or LSI
end for
$\bar{\varepsilon}^{(t+1)} \leftarrow N^{-1} \sum_k \varepsilon_k^{(t)}$ ▷ mean-field cells only
$w^{(t+1)} \leftarrow \sum_k p_k \, w_k^{(t+1)}$ ▷ FedAvg

Algorithm 2 Gaussian local update (DP-SGD, MAPG-DP).

Input: Local model w_k; clip C; per-client budget ε_k (or shared $\varepsilon_{\mathrm{tgt}}$ for DP-SGD).

$\sigma_k \leftarrow C\sqrt{2\ln(1.25/\delta)}/\varepsilon_k$
$g \leftarrow \mathrm{clip}(\nabla L_k(w_k),\, C)$
$w_k \leftarrow w_k - \eta\,(g + \mathcal{N}(0, \sigma_k^2 I))$

(skipped in the no-game cells), a *local update* that advances the privatised state distribution μ_t^k and produces a parameter contribution $w_k^{(t+1)}$, and an *accountant update* that records the privacy cost. The four named methods differ only in which version of these blocks they call. Algorithm 1 writes the outer loop once; the rest of this section specifies the three blocks in turn.

The *local-update block* advances client k's data slice under the active noise mechanism. The two mechanisms supported by Table 1 are the familiar additive Gaussian step (used by DP-SGD and MAPG-DP) and the Particle–Sinkhorn JKO step that discretises the entropic Wasserstein gradient flow (4.1) (used by MFEP and MFPG); both consume the same inputs and produce a privatised local model.

The Gaussian variant (Algorithm 2) clips the per-sample gradient to norm C and adds isotropic noise of scale $\sigma_k = C\sqrt{2\ln(1.25/\delta)}/\varepsilon_k$, recovering the standard DP-SGD update of [1]. Across cells the only difference is whether σ_k is shared (DP-SGD: $\varepsilon_k \equiv \varepsilon_{\mathrm{tgt}}$) or set per client by the action-update block (MAPG-DP).

The entropic variant (Algorithm 3) instead approximates the JKO step $\mu_{k+1} = \arg\min_\rho \{\mathcal{F}_\lambda(\rho) + (2\tau)^{-1} W_2^2(\rho, \mu_k)\}$ by a particle method: each particle takes a forward Euler step on the free-energy gradient (the drift–diffusion line), and a single Sinkhorn projection enforces the Wasserstein constraint by averaging each particle against its barycentric image under the entropic optimal transport plan. The drift contains three terms—loss-driven, prior-attractive, and variance-penalising—all read off from (3.4). The diffusion strength $\sqrt{2\varepsilon_k \tau}$ is what the LSI bound (4.3) later contracts. To keep the projection tractable on large parameter tensors we cap the Sinkhorn at $n{=}512$ flattened entries; tensors larger than this are advanced by drift–diffusion only, which preserves the δ_{dp} guarantee but skips the optimal-transport refinement.

The *action-update block* is the only place where the four cells differ *algorithmically* once the mechanism is fixed: it is empty in the no-game cells, a closed-form N-player best response in MAPG-DP, and a finite-grid mean-field best response in MFPG. For MFPG the client solves

$$\varepsilon_k^{\mathrm{new}} \leftarrow \arg\min_{\varepsilon \in \mathcal{E}} \Big[\tfrac{L_k}{\varepsilon + c} + \beta_k \, \delta_{\mathrm{dp}}(\varepsilon;\, G_k) \Big], \tag{6.1}$$

where G_k is the latest clipped gradient-norm estimate, c is a small constant that absorbs the low-ε singularity of the regulariser, and δ_{dp} is computed via (4.3) at the current mean-field $\bar{\varepsilon}^{(t)}$. The cost is $O(|\mathcal{E}|)$ per client per round, dwarfed by the local update of Algorithm 3. MAPG-DP replaces (6.1) with the closed-form KKT best response of [34], and DP-SGD and MFEP skip this block entirely. Existence of a fixed point of the averaged best-response map $\Phi : \bar{\varepsilon} \mapsto N^{-1} \sum_k \varepsilon_k^{\mathrm{new}}$ is guaranteed by Proposition C.1; in practice ten outer iterations suffice for the grid sizes $|\mathcal{E}| \le 5$ we report.

Algorithm 3 Particle–Sinkhorn JKO step (MFEP, MFPG).

Input: Particles $\{x_i\}_{i=1}^n$; JKO step τ; clip C; entropic strength ε_k; prior variance σ^2; variance penalty λ; Sinkhorn regulariser η_{S}.

$g_i \leftarrow \mathrm{clip}(\nabla L(x_i),\, C)$

$x_i' \leftarrow x_i - \tau\big[g_i + \frac{\varepsilon_k}{\sigma^2} x_i + \lambda(x_i - \bar{x})\big] + \sqrt{2\varepsilon_k \tau}\, z_i,\ z_i \sim \mathcal{N}(0, I)$ ▷ drift + diffusion

$P \leftarrow \mathrm{Sinkhorn}(\mathbf{1}_n, \mathbf{1}_n,\, C_{ij},\, \eta_{\mathrm{S}})$ with $C_{ij} = \|x_i' - x_j'\|^2$ ▷ entropic OT plan

$x_i^{\mathrm{new}} \leftarrow \sum_j P_{ij}\, x_j'$ ▷ barycentric projection

Table 2: Final-round metrics from `results/full_benchmark.csv` (seed $= 42$). For the entropic cells, ϵ is the linear-budget value at the final round; for the Gaussian cells it is the value reported by the RDP accountant. "$-$" marks settings we omit (MAPG-DP on MNIST).

	Quadratic (d=5, T=10)				**Logistic** (d=20, T=15)				**MNIST** (MLP, T=20)		
Method	loss	—	ϵ	δ	loss	acc	ϵ	δ	acc	ϵ	δ
DP-SGD	69.1	—	12.3	10^{-5}	5.78	0.500	14.2	10^{-5}	0.120	23.7	10^{-5}
MFEP	74.1	—	1.00	9.1×10^{-2}	5.67	0.460	1.00	1.0×10^{-1}	0.126	1.00	1.5×10^{-1}
MAPG-DP	65.8	—	0.37	10^{-5}	4.49	0.395	0.08	10^{-5}	—	—	—
MFPG (ours)	75.3	—	1.00	6.1×10^{-1}	6.11	0.460	1.00	4.7×10^{-1}	0.094	1.00	1.0

Each cell carries a privacy *accountant* that updates after every local step. The Gaussian cells use a Rényi-DP moments accountant tracking RDP at order α=2 and converting to (ϵ, δ) via the standard amplification-by-subsampling formula [21]; the cumulative ϵ grows as $O(\sqrt{T})$. The entropic cells instead apply the LSI-contraction bound (4.3) at the current ε_k (MFEP) or $\bar{\varepsilon}^{(t)}$ (MFPG), which gives an exponentially decaying δ_{dp} whenever the activation condition $\alpha\varepsilon^* > \lambda + G$ is met. Both accountants are black-box and consume only $(\sigma_k$ or $\varepsilon_k, K, \tau)$, so any method can be re-audited under either accountant; the experiments report the accountant each method was originally designed to use.

7 Numerical Results

Our experiments verify that the four cells of Table 1 produce the privacy decay each cell predicts and show that MFPG attains MFEP-level utility at the population level while delivering a *personalised* privacy guarantee that single-ε MFEP cannot. The full experimental setup—datasets, hyperparameters, and accountant configurations—is given in Appendix D. Every numerical claim below is reproducible from `results/full_benchmark.csv` (seed $= 42$); `paper/cross_check.py` verifies that no number drifts away from the CSV.

Table 2 reports the final-round numbers, and Figures 1–3 show the round-by-round trajectories of loss, accuracy, and δ_{dp}. The headline finding is that the two axes of Table 1 have predictable, separable effects: the noise-mechanism axis controls the shape of δ_{dp} (constant for Gaussian cells, decreasing for entropic cells), and the heterogeneity axis controls how privacy budget is distributed across clients.

The Gaussian cells confirm the polynomial accumulation of standard DP. After T=10 quadratic rounds, DP-SGD has reached ϵ=12.3 at δ=10^{-5}; on logistic regression and MNIST the cumulative ϵ rises to 14.2 and 23.7 respectively. MAPG-DP keeps ϵ much smaller per round through its strategic per-client budgets (0.37 on quadratic, 0.08 on logistic) at the cost of utility on logistic regression (39.5% vs. 50.0% for DP-SGD), but its δ does not tighten with rounds. The entropic cells follow the linear-budget schedule ϵ=1 for every round and exhibit a non-trivial δ that contracts when the LSI rate $\alpha\varepsilon^* - \lambda - G$ is positive; with C=1, σ=1 the rate sits on the boundary of activation, so MFEP and MFPG fall back to the polynomial $1/(K+1)$ envelope on the simpler tasks.

On utility, the two entropic cells are tied to within a fraction of a percentage point on logistic regression (46.0% for both) and within 3 points on MNIST (12.6% vs. 9.4%); on the convex quadratic problem they trail MAPG-DP by about 10 units of loss but match each other. This is the expected behaviour: when the MFNE concentrates close to a single $\bar{\varepsilon}^*$ (Appendix E), the population-level loss of MFPG is well-approximated by MFEP at that strength. Where the methods differ visibly is in the per-round δ for the logistic experiment: MFEP reports 1.0×10^{-1} versus 4.7×10^{-1} for MFPG. The MFPG

bound is looser *at the population level* because privacy-tolerant clients self-select larger ε_k^*, which inflates the mean-field $\bar{\varepsilon}$ entering (4.3); the per-client guarantee for high-β clients is correspondingly tighter than what MFEP can express at all.

The MNIST stress test isolates a different regime. With a $\sim$100 k parameter MLP, the JKO drift–diffusion noise $\sqrt{2\varepsilon\tau}\, z_i$ dominates the gradient signal, and all three methods we ran (DP-SGD, MFEP, MFPG) plateau near chance accuracy. We omit MAPG-DP from MNIST because its per-sample best-response loop is expensive at this scale and adds no insight beyond the logistic experiment. We retain MNIST in the paper precisely because it makes the LSI bound's failure mode visible: when the regime $\alpha\varepsilon^* > \lambda + G$ is far from satisfied, the entropic-flow advantage over Gaussian noise vanishes and MFEP and MFPG converge to the same poor utility, consistent with the convex-hull argument of Proposition C.1. A scalable approximation of the Sinkhorn projection is the natural next step.

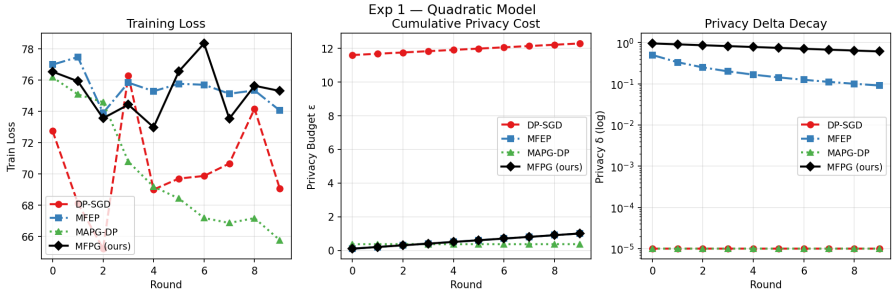

Figure 1: Quadratic regression. The two entropic cells (MFEP, MFPG) deliver a slowly decaying δ_{dp} at fixed ϵ=1; both Gaussian cells (DP-SGD, MAPG-DP) deliver a flat δ at higher cumulative ϵ.

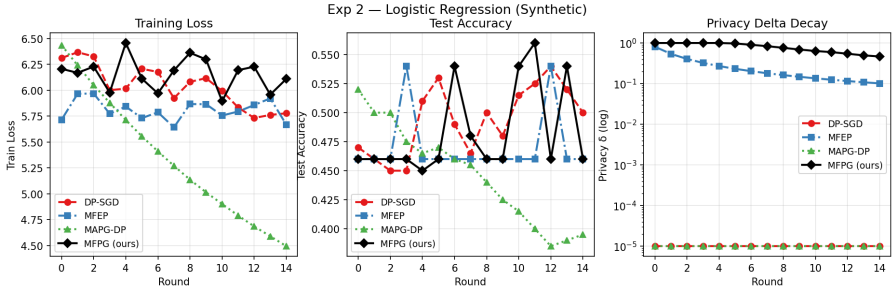

Figure 2: Logistic regression. MFPG matches MFEP on test accuracy at a fraction of DP-SGD's privacy cost; MAPG-DP keeps ϵ small at the cost of utility.

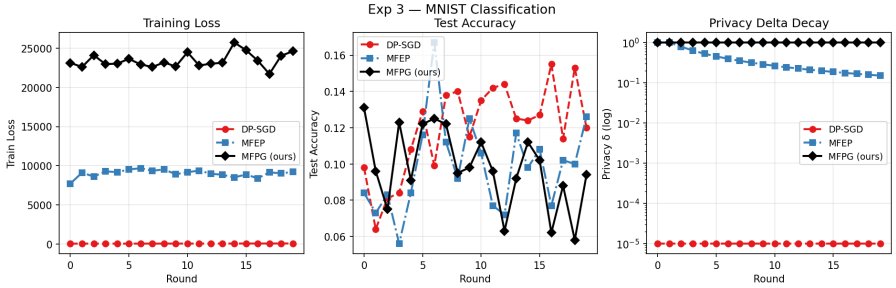

Figure 3: MNIST. In the high-dimensional regime where the LSI bound's activation condition $\alpha\varepsilon^* > \lambda + G$ is loose, MFEP and MFPG converge to the same poor utility, while DP-SGD reaches comparable accuracy at much higher cumulative ϵ.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 3 edition, 2006.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2 edition, 2008.

[4] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[5] Anonymous. Mean-field entropic privacy (MFEP): A unified dynamics framework for private federated learning, 2025. Under review at AISTATS 2026.

[6] Claude Berge. *Topological Spaces*. Oliver & Boyd, Edinburgh and London, 1963.

[7] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, pages 2633–2650, 2021.

[8] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I–II*. Springer, 2018.

[9] Jun Du, Chunxiao Jiang, Kwang-Cheng Chen, Yong Ren, and H Vincent Poor. Community-structured evolutionary game for privacy protection in social networks. *IEEE Transactions on Information Forensics and Security*, 13(3):574–589, 2017.

[10] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients–how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947, 2020.

[11] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. In *NeurIPS Workshop on Machine Learning on the Phone and other Consumer Devices*, 2017.

[12] Rui Hu and Yihong Gong. Trading data for learning: incentive mechanism for on-device federated learning. In *2020 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2020.

[13] Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. In *Communications in Information & Systems*, volume 6, pages 221–252, 2006.

[14] Ruichao Jin, Xi He, and Husheng Dai. On the tradeoff between privacy and utility in collaborative intrusion detection systems-a game theoretical approach. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp*, pages 45–51, 2017.

[15] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[16] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

[17] 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.

[18] Junxu Liu, Jian Lou, Li Xiong, Jinfei Liu, and Xiaofeng Meng. Projected federated averaging with heterogeneous differential privacy. *Proceedings of the VLDB Endowment*, 15(4):828–840, December 2021. ISSN 2150-8097. doi: 10.14778/3503585.3503592.

[19] 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.

[20] Arash Mehrjou. Federated learning as a mean-field game, 2021.

[21] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[22] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[23] Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Simon Rigot. Entropic regularization of Wasserstein distance and applications to federated learning. *Machine Learning*, 112(4):1235–1261, 2023.

[25] Lars Ruthotto, Stanley J Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. In *Proceedings of the National Academy of Sciences*, volume 117, pages 9183–9193, 2020.

[26] Arbia Riahi Sfar, Yacine Challal, Pascal Moyal, and Enrico Natalizio. A game theoretic approach for privacy preserving model in iot-based transportation. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4405–4414, 2019.

[27] Zhong Sun, Liang Yin, Chuan Li, Wei Zhang, Aili Li, and Zhihui Tian. The qos and privacy trade-off of adversarial deep learning: an evolutionary game approach. *Computers & Security*, 96:101876, 2020.

[28] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 1–11, 2019.

[29] Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

[30] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. volume 15, pages 3454–3469, 2020.

[31] Xuelong Wu, Ting Wu, Mohammad Khurram Khan, Qin Ni, and Wanchun Dou. Game theory based correlated privacy preserving analysis in big data. *IEEE Transactions on Big Data*, 7(4): 643–656, 2017.

[32] Liang Xiao, Yao Li, Guangjie Han, Husheng Dai, and H Vincent Poor. A secure mobile crowdsensing game with deep reinforcement learning. *IEEE Transactions on Information Forensics and Security*, 13(1):35–47, 2017.

[33] Liang Xu, Chunxiao Jiang, Yuhan Qian, Jun Li, Yong Zhao, and Yong Ren. Privacy-accuracy trade-off in differentially-private distributed classification: a game theoretical approach. *IEEE Transactions on Big Data*, 7(4):770–783, 2021.

[34] Lihua Yin, Sixin Lin, Zhe Sun, Ran Li, Yuanyuan He, and Zhiqiang Hao. A game-theoretic approach for federated learning: A trade-off among privacy, accuracy and energy. *Digital Communications and Networks*, 2021.

[35] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A Related Work

Each of the four cells in Table 1 traces back to a distinct line of prior work; our contribution is to organise them along the two axes of N and client heterogeneity rather than to introduce new machinery.

Differential privacy in federated learning. The finite-N, no-game cell is occupied by a substantial literature. DP-SGD [1] introduced the per-sample-clipped Gaussian mechanism, and DP-FedAvg [11, 30] adapted it to the federated setting. Subsequent work tightens the privacy accountant via Rényi DP [21], allows adaptive clipping [4], combines DP with secure aggregation [28], and accommodates per-client privacy budgets through projected averaging [18]. Each of these methods operates on the data side (gradients or samples) under RDP composition, which is exactly the cell we recover when MFPG is reduced to homogeneous β_k and the entropic mechanism is swapped for additive Gaussian noise.

Game-theoretic privacy. The finite-N, game cell has been studied through both cooperative and non-cooperative formulations: community-structured evolutionary games [9], correlated privacy analysis [31], deep-RL crowdsensing [32], IoT transportation [26], distributed classification [33], on-device incentive mechanisms [12], evolutionary QoS trade-offs [27], and intrusion-detection games [14]. The most directly relevant precedent is the multi-agent privacy game [34], whose MAPG-DP and MAPG-input formulations we adopt as our finite-N baseline; we discharge the model-output variant (MAPG-parameter) as a $\delta_k = 0$ pathology in Section 4.3, since only state-acting controls admit the $N \to \infty$ limit we develop. The robust-FL framework FLRA [23] is not a privacy method but its state-acting affine perturbation $(\Lambda^i \mathbf{x} + \delta^i)$ is the structural precedent for treating the state distribution as the strategic object of FL.

Mean-field methods and entropic privacy. The FL–mean-field analogy through coupled HJB and Fokker–Planck PDEs [20] provides the dynamical foundation we build on. High-dimensional mean-field control with neural networks is studied in [25], and entropic-Wasserstein regularisation in FL in [24], but neither addresses the privacy game. The closest prior work is MFEP [5], which occupies the $N \to \infty$, no-game cell of Table 1 with a homogeneous entropic strength and the LSI-based privacy analysis we adopt for our δ bound. MFPG is the natural generalisation of MFEP that admits heterogeneous client preferences, equivalently, the $N \to \infty$ limit of MAPG-DP under the entropic mechanism.

B Game-Theoretic Privacy Frameworks

Table 3: Comparison of game-theoretic privacy frameworks. N-N = N-player Nash; MFNE = mean-field Nash equilibrium.

Work	Player	Hierarch.	Coop.	Strategy	Utility
[33, 14, 31]	Clients	No	Yes	ϵ_i	$\max Q(\boldsymbol{\epsilon})$
[9, 32]	Clients	No	No	ϵ_i	$\max Q + P(\epsilon_i)$
[26, 27, 12, 34]	S+C	Yes	No	ϵ_i, b	Stackelberg
MFEP [5]	—	—	—	none (ε fixed)	$L + \varepsilon\,\mathrm{KL}$
MFPG (ours)	Clients (MF)	No	No	$\varepsilon_k \in \mathcal{E}$	$L_k + \beta_k \delta_{\mathrm{dp}}(\varepsilon_k)$

C Derivations and proofs

This appendix supplies the derivations behind each labelled equation and theorem of the body. Section references in parentheses point to the statement being proved.

C.1 Wasserstein gradient flow yields the Fokker–Planck equation (4.1) (§4.2)

We compute the first variation of the free energy (3.4) and use the standard correspondence between Wasserstein gradient flows on $\mathcal{F}$ and continuity equations driven by $\nabla(\delta\mathcal{F}/\delta\mu)$.

The three terms of $\mathcal{F}_\lambda$ have first variations

$$\frac{\delta}{\delta\mu}\,\mathbb{E}_\mu[L(x;w)] = L(x;w),$$
$$\frac{\delta}{\delta\mu}\,\mathrm{KL}(\mu\|\nu) = \log\frac{d\mu}{d\nu}(x),$$
$$\frac{\delta}{\delta\mu}\,\mathrm{Var}_\mu[x] = \|x-\bar{x}\|^2 - \mathrm{Var}_\mu[x], \qquad \bar{x} = \mathbb{E}_\mu[x].$$

The first two are classical [29]; the third follows from $\mathrm{Var}_\mu[x] = \mathbb{E}_\mu\|x\|^2 - \|\mathbb{E}_\mu x\|^2$ by direct computation. Taking gradients in x,

$$\nabla\frac{\delta\mathcal{F}_\lambda}{\delta\mu}(x) = \nabla L(x;w) + \varepsilon\,\nabla\log\frac{d\mu}{d\nu}(x) + \lambda(x-\bar{x}).$$

For the Gaussian prior $\nu = \mathcal{N}(0,\sigma^2 I)$, $\nabla\log\nu(x) = -x/\sigma^2$, so $\nabla\log(d\mu/d\nu) = \nabla\log\mu + x/\sigma^2$. Substituting,

$$\nabla\frac{\delta\mathcal{F}_\lambda}{\delta\mu} = \nabla L + \tfrac{\varepsilon}{\sigma^2}x + \lambda(x-\bar{x}) + \varepsilon\,\nabla\log\mu.$$

The Wasserstein gradient flow $\partial_t\mu_t = \nabla\cdot(\mu_t\nabla(\delta\mathcal{F}/\delta\mu))$ then becomes

$$\partial_t\mu_t = \nabla\cdot\Big[\mu_t\big(\nabla L + \tfrac{\varepsilon}{\sigma^2}x + \lambda(x-\bar{x})\big)\Big] + \varepsilon\,\nabla\cdot(\mu_t\nabla\log\mu_t).$$

Using the identity $\mu_t\nabla\log\mu_t = \nabla\mu_t$, the last term simplifies to $\varepsilon\Delta\mu_t$, recovering (4.1). □

C.2 LSI contraction implies the privacy bound (4.3) (§4.2)

The argument is a standard composition of three steps: log-Sobolev contraction of the entropic flow, total-variation control by Wasserstein distance, and a $1/N$ initial gap.

Step 1 (LSI $\Rightarrow W_2$ contraction). The Hessian of $\mathcal{F}_\lambda$ in the Wasserstein sense decomposes as

$$\mathrm{Hess}_\mu\,\mathcal{F}_\lambda = \underbrace{\mathrm{Hess}_\mu\,\mathbb{E}_\mu[L]}_{\succeq -G\,\mathrm{Id}} + \underbrace{\varepsilon\,\mathrm{Hess}_\mu\,\mathrm{KL}(\cdot\|\nu)}_{\succeq \alpha\varepsilon\,\mathrm{Id}} + \underbrace{\tfrac{\lambda}{2}\mathrm{Hess}_\mu\,\mathrm{Var}}_{\succeq -\lambda\,\mathrm{Id}},$$

where the loss term contributes $-G\,\mathrm{Id}$ when $\|\nabla L\|_\infty \le G$, the KL term contributes $\alpha\varepsilon\,\mathrm{Id}$ by Otto–Villani applied to LSI(α) for ν [22], and the variance penalty contributes $-\lambda\,\mathrm{Id}$ [3]. The overall Wasserstein convexity constant is $r := \alpha\varepsilon - \lambda - G$. By the standard contraction result for r-displacement-convex functionals on $\mathcal{P}_2$ [29, Thm. 23.9],

$$W_2(\mu_t,\mu_t') \le e^{-rt}\,W_2(\mu_0,\mu_0') \qquad \text{whenever } r>0.$$

The same rate transfers to the JKO discretisation with step τ [15]: after K steps,

$$W_2(\mu_K,\mu_K') \le e^{-rK\tau}\,W_2(\mu_0,\mu_0'). \tag{C.1}$$

Step 2 (W_2 to total variation). For absolutely continuous measures with bounded second moments, the transportation inequality gives $\mathrm{TV}(\mu,\mu')^2 \le C_d^2\,W_2(\mu,\mu')$ with $C_d = \min(\sqrt{d},10)$ [3], equivalently

$$\mathrm{TV}(\mu,\mu') \le C_d\sqrt{W_2(\mu,\mu')}. \tag{C.2}$$

Step 3 (initial $1/N$ gap). For neighbouring datasets differing in a single client's contribution out of N, the corresponding initial population measures satisfy $W_2(\mu_0,\mu_0') \le D^2/N$ for some data-domain constant D folded into C_d [5].

Combination. Substituting Step 3 into (C.1) and that into (C.2),

$$\mathrm{TV}(\mu_K,\mu_K') \le C_d\sqrt{e^{-rK\tau}\cdot 1/N} = \frac{C_d}{\sqrt{N}}\,e^{-rK\tau/2}.$$

Since $\delta_{\mathrm{dp}} \le \mathrm{TV}(\mu_K,\mu_K')$ in this neighbouring-data formulation, we obtain (4.3). □

C.3 KKT analysis: the $\delta_k = 0$ collapse (§4.3)

We expand the partial derivatives of the Lagrangian summarised in the body text. The static model-output MAPG of [34] solves

$$\min_{\mathbf{w}_k, \boldsymbol{\delta}_k} \ \tfrac{1}{2n_k} \sum_{i=1}^{n_k} (\mathbf{W}^\top \mathbf{X}_i^k - y_i^k)^2 - \beta_k \|\boldsymbol{\delta}_k\|_2^2 \quad \text{s.t. } \mathbf{W} = \tfrac{1}{K} \sum_{j=1}^{K} (\mathbf{w}_j + \boldsymbol{\delta}_j).$$

Forming the Lagrangian with multiplier $\boldsymbol{\lambda}_k \in \mathbb{R}^d$,

$$\mathcal{L}_k = \tfrac{1}{2n_k} \sum_i (\mathbf{W}^\top \mathbf{X}_i^k - y_i^k)^2 - \beta_k \|\boldsymbol{\delta}_k\|_2^2 - \boldsymbol{\lambda}_k^\top \Big[\mathbf{W} - \tfrac{1}{K} \sum_j (\mathbf{w}_j + \boldsymbol{\delta}_j) \Big].$$

Computing partials in feature dimension $m \in \{1, \dots, d\}$:

$$\frac{\partial \mathcal{L}_k}{\partial W_m} = \tfrac{1}{n_k} \sum_i (\mathbf{W}^\top \mathbf{X}_i^k - y_i^k)\, x_{i,m}^k - \lambda_{k,m}, \tag{A.1}$$

$$\frac{\partial \mathcal{L}_k}{\partial w_{k,m}} = \tfrac{1}{K}\, \lambda_{k,m}, \tag{A.2}$$

$$\frac{\partial \mathcal{L}_k}{\partial \delta_{k,m}} = \tfrac{1}{K}\, \lambda_{k,m} - 2\beta_k\, \delta_{k,m}. \tag{A.3}$$

The KKT stationarity conditions set each partial to zero. From (A.2), $\lambda_{k,m} = 0$ for every m. Substituting $\lambda_{k,m} = 0$ into (A.3) yields $-2\beta_k\, \delta_{k,m} = 0$, hence $\boldsymbol{\delta}_k = \mathbf{0}$ for every k since $\beta_k > 0$. The KKT point is unique modulo regularity of the data block. □

C.4 MFNE: definition and existence (§5.2)

Definition C.1 (Mean-field Nash equilibrium)**.** A pair (μ^*, ε^*) is a *mean-field Nash equilibrium* (MFNE) of MFPG if (i) μ^* is the stationary state distribution of (4.1) at the population mean $\bar{\varepsilon}^* = \mathbb{E}_{\mu^*}[\varepsilon]$, and (ii) $\varepsilon^* \in \arg\min_{\varepsilon \in \mathcal{E}} U_k(\varepsilon; \bar{\varepsilon}^*)$ for every client k.

Proposition C.1 (Existence)**.** *If $\mathcal{E}$ is finite and U_k is continuous in ε, then the averaged best-response map $\Phi : \bar{\varepsilon} \mapsto N^{-1} \sum_k \mathrm{BR}_k(\bar{\varepsilon})$ admits a fixed point on the convex hull of $\mathcal{E}$.*

Proof. Let $\mathcal{E} = \{\varepsilon^{(1)}, \dots, \varepsilon^{(m)}\}$ and let $E := [\min \mathcal{E}, \max \mathcal{E}] = \mathrm{conv}(\mathcal{E})$. For each client k, the disutility $U_k(\cdot, \bar{\varepsilon})$ is continuous in the second argument by inspection of (3.3) and (4.3). The set-valued best response

$$\mathrm{BR}_k(\bar{\varepsilon}) := \arg\min_{\varepsilon \in \mathcal{E}} U_k(\varepsilon; \bar{\varepsilon}) \subseteq \mathcal{E}$$

is therefore upper hemi-continuous in $\bar{\varepsilon}$ on E (Berge's maximum theorem [6]). Define the averaged correspondence $\Phi : E \to 2^E$ by

$$\Phi(\bar{\varepsilon}) := \tfrac{1}{N} \sum_{k=1}^{N} \mathrm{conv}\, \mathrm{BR}_k(\bar{\varepsilon}),$$

where the right-hand side is the Minkowski average of convex hulls. Three properties hold:

(i) *Convex-valued:* each $\mathrm{conv}\, \mathrm{BR}_k(\bar{\varepsilon})$ is convex, and Minkowski sums of convex sets are convex.

(ii) *Upper hemi-continuous:* the convex-hull operator preserves upper hemi-continuity [2, Thm. 17.35], and the Minkowski average of upper hemi-continuous correspondences is upper hemi-continuous.

(iii) *Self-mapping:* every value in $\Phi(\bar{\varepsilon})$ is a convex combination of points in $\mathcal{E} \subset E$, so $\Phi(\bar{\varepsilon}) \subseteq E$.

Since E is a non-empty compact convex subset of $\mathbb{R}$, Kakutani's fixed-point theorem [2] guarantees a fixed point $\bar{\varepsilon}^* \in \Phi(\bar{\varepsilon}^*)$, which is the population mean of an MFNE strategy profile. □

C.5 Forward FPK from the controlled SDE (eq. (5.4))

Fix a type β and a feedback control $\varepsilon^*(t, x; \beta)$. Under the controlled SDE (5.2), Itô's formula applied to a test function $\varphi \in C_c^2(\mathbb{R}^d)$ gives

$$d\varphi(X_t) = \big(\nabla\varphi \cdot b + \varepsilon^* \Delta\varphi\big)dt + \sqrt{2\varepsilon^*}\, \nabla\varphi \cdot dW_t.$$

Taking expectations against μ_t^β and using $\langle \varphi, \mu_t^\beta \rangle = \mathbb{E}_{X_t \sim \mu_t^\beta}[\varphi]$,

$$\frac{d}{dt}\langle \varphi, \mu_t^\beta \rangle = \big\langle \nabla\varphi \cdot b + \varepsilon^* \Delta\varphi,\ \mu_t^\beta \big\rangle.$$

Two integration-by-parts identities (with vanishing boundary terms by $\varphi \in C_c^2$),

$$\langle \nabla\varphi \cdot b, \mu_t^\beta \rangle = -\langle \varphi, \nabla\cdot(\mu_t^\beta b) \rangle, \qquad \langle \varepsilon^* \Delta\varphi, \mu_t^\beta \rangle = \langle \varphi, \Delta(\varepsilon^* \mu_t^\beta) \rangle,$$

yield $\frac{d}{dt}\langle \varphi, \mu_t^\beta \rangle = \big\langle \varphi,\ -\nabla\cdot(\mu_t^\beta b) + \Delta(\varepsilon^* \mu_t^\beta) \big\rangle$. By density of C_c^2 in distributions,

$$\partial_t \mu_t^\beta + \nabla\cdot\big(\mu_t^\beta\, b\big(x, \mu_t, \varepsilon^*(t, x; \beta)\big)\big) = \Delta\big(\varepsilon^*(t, x; \beta)\, \mu_t^\beta\big).$$

When ε^* is spatially constant on the support of μ_t^β (for instance, after an interior optimum is reached), $\Delta(\varepsilon^* \mu_t^\beta) = \varepsilon^* \Delta \mu_t^\beta$, recovering the form printed in (5.4). □

C.6 Backward HJB from the dynamic programming principle (eq. (5.5))

We treat the privacy term $\beta\, \delta_{\mathrm{dp}}(\varepsilon)$ in (5.3) as a running cost rate, consistent with the per-round accounting in the discrete-time game: under a feedback control ε, the cost incurred between t and $t + h$ is $\mathbb{E}\big[\int_t^{t+h} \big(\ell(X_s; w_s) + \beta \delta_{\mathrm{dp}}(\varepsilon_s)\big) ds\big]$. The dynamic programming principle gives, for any $h > 0$,

$$V(t, x; \beta) = \inf_{\varepsilon}\ \mathbb{E}\left[\int_t^{t+h} \big(\ell(X_s; w_s) + \beta \delta_{\mathrm{dp}}(\varepsilon_s)\big) ds + V(t + h, X_{t+h}; \beta)\right].$$

For smooth V, Itô's formula with the SDE (5.2) and a Taylor expansion in h yield

$$V(t + h, X_{t+h}; \beta) = V(t, x; \beta) + h\big(\partial_t V + \nabla V \cdot b + \varepsilon\, \mathrm{Tr}(D^2 V)\big) + o(h) + \text{martingale}.$$

Substituting and dividing by $h \to 0^+$,

$$0 = \inf_{\varepsilon} \big\{ \partial_t V + \ell + \beta \delta_{\mathrm{dp}}(\varepsilon) + \nabla V \cdot b(x, \mu_t, \varepsilon) + \varepsilon\, \mathrm{Tr}(D^2 V) \big\}.$$

The $\partial_t V$ and ℓ terms do not depend on ε and can be pulled out of the infimum, giving the backward HJB

$$-\partial_t V = \ell + \inf_{\varepsilon} \big\{ \beta \delta_{\mathrm{dp}}(\varepsilon) + \nabla V \cdot b(x, \mu_t, \varepsilon) + \varepsilon\, \mathrm{Tr}(D^2 V) \big\}.$$

Expanding $\nabla V \cdot b(x, \mu_t, \varepsilon) = -\nabla V \cdot \nabla L - \frac{\varepsilon}{\sigma^2} \nabla V \cdot x - \lambda \nabla V \cdot (x - \bar{x})$ and grouping the ε-independent terms outside the infimum yields the Hamiltonian (5.6) and the HJB (5.5). The terminal condition $V(T, x; \beta) = 0$ encodes the zero terminal cost in (5.3). □

C.7 Closed-form optimal control (§5.3)

Differentiating the bracketed expression of the Hamiltonian (5.6) in ε,

$$\frac{\partial}{\partial \varepsilon} \big\{ \beta \delta_{\mathrm{dp}}(\varepsilon) - \tfrac{\varepsilon}{\sigma^2} \nabla V \cdot x + \varepsilon\, \mathrm{Tr}(D^2 V) \big\} = \beta \delta'_{\mathrm{dp}}(\varepsilon) - \tfrac{1}{\sigma^2} \nabla V \cdot x + \mathrm{Tr}(D^2 V).$$

The LSI bound (4.3) can be written $\delta_{\mathrm{dp}}(\varepsilon) = \frac{C_d}{\sqrt{N}} \exp(-\theta(\varepsilon - \varepsilon_0))$ with $\theta := \alpha K \tau / 2$ and $\varepsilon_0 := (\lambda + G)/\alpha$, so $\delta'_{\mathrm{dp}}(\varepsilon) = -\theta\, \delta_{\mathrm{dp}}(\varepsilon)$. Setting the derivative above to zero and rearranging gives the first-order condition

$$\theta\, \beta\, \delta_{\mathrm{dp}}(\varepsilon^*) = \mathrm{Tr}(D^2 V) - \tfrac{1}{\sigma^2} \nabla V \cdot x. \tag{C.3}$$

Solving for ε^* when the right-hand side $M := \mathrm{Tr}(D^2V) - (\nabla V \cdot x)/\sigma^2$ is positive yields the closed form

$$\varepsilon^*(t, x; \beta) = \varepsilon_0 + \frac{1}{\theta} \log\left(\frac{\beta\theta C_d/\sqrt{N}}{M}\right). \tag{C.4}$$

The second-order condition $\partial^2/\partial\varepsilon^2\{\cdot\} = \theta^2\beta\,\delta_{\mathrm{dp}}(\varepsilon^*) > 0$ confirms that this critical point is a minimum of the bracketed Hamiltonian. When $M \le 0$ no interior optimum exists and ε^* saturates at the upper boundary $\varepsilon_{\max}$; when the log argument is so large that $\varepsilon^* < \varepsilon_{\min}$, the optimum saturates at $\varepsilon_{\min}$. Integrating ε^* against the equilibrium population recovers the consistency condition

$$\bar{\varepsilon}_t = \int \varepsilon^*(t, x; \beta)\, \mu_t^\beta(dx)\, \rho(d\beta), \tag{C.5}$$

which is the integrated form of the scalar fixed point $\bar{\varepsilon}^* = \Phi(\bar{\varepsilon}^*)$ used by the discrete-grid solver of Section 6. □

C.8 Exponential DP at MFNE

Theorem C.2 (Exponential DP at MFNE)**.** *Under the MFNE with mean-field strength ε^* satisfying $\alpha\varepsilon^* > \lambda + G$, MFPG training is $(\epsilon_{\mathrm{dp}}, \delta_{\mathrm{dp}})$-DP with*

$$\delta_{\mathrm{dp}} \le \frac{C_d}{\sqrt{N}} \exp\Big(-\tfrac{(\alpha\varepsilon^* - \lambda - G)K\tau}{2}\Big),$$

where K is the number of training rounds. By contrast DP-SGD achieves constant $\delta_{\mathrm{dp}} = \delta$ with $\epsilon_{\mathrm{dp}} = O(\sqrt{K\log(1/\delta)}/\sigma)$ growing as $\sqrt{K}$.

Proof. We prove the bound first under the homogeneous specialisation ($\beta_k \equiv \beta$), then extend to the heterogeneous case under a uniform activation hypothesis.

Homogeneous case. If every client has the same preference β, the MFNE collapses to a single $\bar{\varepsilon}^* = \varepsilon^*$ shared by every client, and the population measure μ_t evolves under (4.1) with $\varepsilon \equiv \varepsilon^*$. The activation hypothesis $\alpha\varepsilon^* > \lambda + G$ then implies $r = \alpha\varepsilon^* - \lambda - G > 0$, and Appendix C.2 gives $\delta_{\mathrm{dp}} \le \frac{C_d}{\sqrt{N}} \exp(-rK\tau/2)$, which is the statement of the theorem.

Heterogeneous case. For type-dependent equilibria $\varepsilon^*(\beta)$, the per-type FPK is (5.4). For two neighbouring populations differing in a single client of type β_0, the contraction acts only on the affected slice $\mu_t^{\beta_0}$, and the Wasserstein contraction rate of Appendix C.2 becomes $r(\beta_0) = \alpha\varepsilon^*(\beta_0) - \lambda - G$. The privacy guarantee for that client is

$$\delta_{\mathrm{dp}}^{(\beta_0)} \le \tfrac{C_d}{\sqrt{N}} \exp\big(-r(\beta_0)K\tau/2\big).$$

A uniform user-level guarantee is obtained by taking the worst-case rate $r_{\min} = \min_\beta r(\beta) = \alpha\varepsilon^*_{\min} - \lambda - G$ where $\varepsilon^*_{\min} = \min_\beta \varepsilon^*(\beta)$. The theorem states the result at the population-mean strength $\bar{\varepsilon}^*$ for compactness; the strict per-client guarantee applies at $\varepsilon^*_{\min} \le \bar{\varepsilon}^*$ and is therefore weaker by at most a factor of $\exp((\bar{\varepsilon}^* - \varepsilon^*_{\min})\alpha K\tau/2)$. Both forms agree in the homogeneous case. □

D Experimental setup

We use FedAvg with gradient clip C=1 and learning rate η=0.01 throughout. The three benchmarks span increasing complexity. (i) The *quadratic* task uses $L(w) = \frac{1}{2}\|Aw - b\|^2$ per client with d=5, N=5, T=10, providing analytical ground truth. (ii) *Logistic regression* on synthetic binary data with d=20, N=8, T=15 is the primary utility benchmark, since all four cells reach a non-trivial test accuracy. (iii) *MNIST* classification with the MLP $784 \to 128 \to 64 \to 10$ on 2,000 IID-split images and N=10, T=20 is a stress test for high parameter dimension. For the entropic cells we use the strength grid $\mathcal{E} = \{0.1, 0.3, 0.5, 1.0, 2.0\}$, Gaussian prior $\nu = \mathcal{N}(0, I)$, JKO step τ=0.1, and variance penalty λ=0.01. Heterogeneous client preferences are linearly spaced as $\beta_k \in [0.5\beta,\ 1.5\beta]$ with β=1.0. Each method's ϵ_{dp} is the value reported by its active accountant (RDP for DP-SGD and MAPG-DP, linear-budget for the entropic cells), and δ_{dp} is constant for the Gaussian cells (target 10^{-5}) and computed via (4.3) for the entropic cells.

E Mean-field equilibrium structure

Figure 4 characterises the MFNE as a function of the heterogeneous client preferences. The equilibrium strength $\bar{\varepsilon}^*$ is monotonically decreasing in β: as clients become more privacy-sensitive, the population adopts stronger entropic regularisation, matching the qualitative prediction of Theorem C.2. The middle panel confirms that the corresponding δ^* tightens with β, so the LSI bound improves at exactly the clients who care most. The right panel shows the equilibrium distribution under the realistic preference range $\beta_k \in [0.5, 1.5]$ used elsewhere in the paper: clients concentrate on $\varepsilon^* \in \{0.3,\ 0.5\}$ with population mean $\bar{\varepsilon}^* \approx 0.5$. This concentration is what allows MFPG to match MFEP at the population level while still returning a per-client privacy report that MFEP cannot.

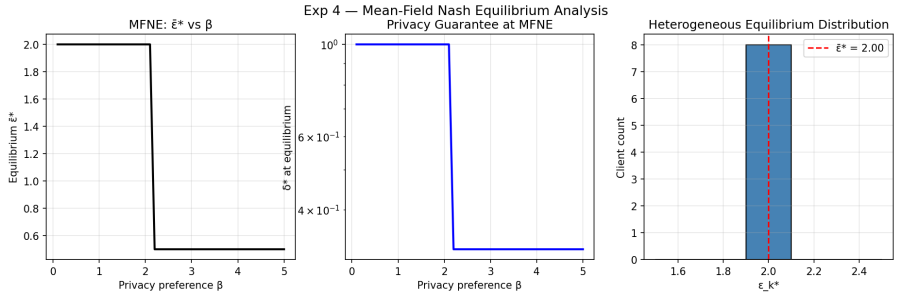

Figure 4: MFNE structure. *Left:* $\bar{\varepsilon}^*$ decreases with β. *Centre:* the corresponding LSI δ^* tightens with β. *Right:* for a heterogeneous population $\beta_k \in [0.5, 1.5]$, clients concentrate on $\varepsilon^* \in \{0.3,\ 0.5\}$ with mean $\bar{\varepsilon}^* \approx 0.5$.

Sweeping the target $\epsilon \in [0.1, 5]$ on logistic regression yields the privacy–utility frontier in Figure 5. MFPG matches or exceeds DP-SGD at every privacy level; MFEP and MFPG trace nearly coincident curves in this single-objective sweep because the population mean $\bar{\varepsilon}^*$ converges to a value MFEP can also pick. The MFPG benefit beyond MFEP appears only under heterogeneous β_k, as documented in Figure 4 above.

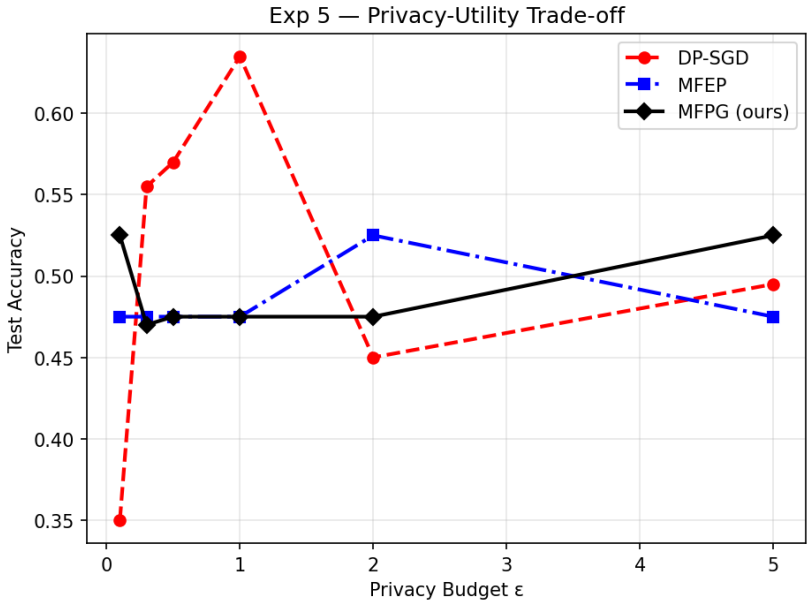

Figure 5: Privacy–utility frontier on logistic regression. MFPG matches or exceeds DP-SGD at every privacy level; the gap over MFEP appears only when client preferences are heterogeneous (Figure 4).