big data and cognitive computing

MDPI

Article

The Geometry of Privacy: A Two-Stage Analysis of Generative Membership Inference in Federated Learning

Borja Arroyo Galende [1,*], Patricia A. Apellániz [1], Alejandro Almodóvar [1], Silvia Uribe [2], Federico Álvarez [1] and Juan Parras [1]

[1] Information Processing and Telecommunications Center, Escuela Técnica Superior de Ingeniería de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain; patricia.alonsod@upm.es (P.A.A.); alejandro.almodovar@upm.es (A.A.); federico.alvarez@upm.es (F.Á.); j.parras@upm.es (J.P.)

[2] Departamento de Sistemas Informáticos, Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain; silviaalba.uribe@upm.es

* Correspondence: borja.arroyog@upm.es

Abstract

We study Membership Inference Attack (MIA) risk in Federated Learning through a two-stage lens that separates (i) whether a target client's contribution is detectable after aggregation and system noise (Stage I: Signal Survival) from (ii) whether a surviving contribution induces a generative membership score change attributable to the target's private data (Stage II: Signal Attribution). Stage I models aggregation as a target–background decomposition and shows that detectability hinges on target–background alignment, which can induce cancellation. Stage II connects the surviving target component to a generative MIA score via a local path representation and Lipschitz/smoothness bounds, avoiding architecture-specific assumptions. Our analysis reveals that the leading attribution term is governed by the alignment between the target update and the score geometry of the target data at an appropriate baseline. We validate the theoretical bounds and illustrate risk trajectories across several scenarios.

Keywords: federated learning; Membership Inference Attack; generative models; privacy; synthetic data

1. Introduction

Strict data privacy regulations like the General Data Protection Regulation (GDPR) [1] and the Health Insurance Portability and Accountability Act (HIPAA) [2] increasingly restrict direct data sharing. Motivated by these constraints and the concurrent need to leverage data across sensitive domains like healthcare and finance, federated learning (FL) has emerged as an established paradigm for collaborative model training [3,4]. FL fundamentally alters the privacy landscape by inserting an aggregation layer between a client's local data and the released global model. Instead of centralizing raw datasets, clients compute model updates locally and transmit only these ephemeral parameter changes, such as gradients or weight differentials, to a coordinating server [5,6]. This decoupled architecture minimizes direct data exposure, inherently positioning FL as a privacy-enhancing alternative to traditional centralized machine learning.

FL deployments generally fall into two categories, each presenting distinct privacy challenges: cross-device and cross-silo [4]. Cross-device FL involves a massive number of unreliable edge devices (e.g., smartphones) with small local datasets, where the primary

Received: 23 March 2026
Revised: 28 April 2026
Accepted: 13 May 2026
Published: 19 May 2026

security concern often involves malicious clients attempting data poisoning or isolating specific updates [5]. Conversely, cross-silo FL typically involves a smaller number of highly reliable institutional participants (e.g., hospitals or banks) with large datasets. In this setting, the threat model often assumes honest-but-curious participants who actively attempt to infer characteristics of other silos' datasets from the shared model updates, making them particularly vulnerable to Membership Inference Attacks (MIAs) [6].

This shared update mechanism introduces a diverse range of privacy and security risks. As a primary example, adversaries may attempt to breach the confidentiality of client datasets directly from exchanged gradients through data reconstruction or gradient inversion methods, seeking to recover original training samples with high fidelity [7,8]. Similarly, malicious or honest-but-curious participants can execute property inference attacks to continuously monitor and extract aggregate statistical characteristics or attributes unintended for sharing [9,10]. Furthermore, active adversaries may exploit the federated aggregation protocol to embed persistent malicious behaviors via federated backdoor and model poisoning attacks [11]. Among these vulnerabilities, MIAs, which specifically aim to determine whether a particular data record was used in the training set [12,13], remain a dominant and challenging concern. Over time, state-of-the-art (SOTA) attacks have achieved significant success against federated deep learning models by exploiting their inherent capacity to memorize individual training examples [14,15].

Despite these vulnerabilities, evaluating membership inference in FL presents a challenge that SOTA approaches often do not fully deconstruct. Current SOTA attacks in FL, including recent trajectory-based methods [16], cross-round comprehensive attacks [17], and resolution-sensitive black-box techniques [18], typically evaluate MIA risk *end-to-end* by directly computing a success metric or training a shadow model on the aggregated global updates [6,9,14,16–18]. This monolithic methodology inherently conflates two distinct privacy barriers: the system-level masking provided by the aggregation process and the fundamental data-level vulnerability of the target example. For instance, an end-to-end attack might fail either because the target's update was thoroughly canceled by other clients' contributions or because the target-data point is intrinsically indistinguishable under the chosen generative score even without aggregation [15,19].

To formally isolate these confounding effects, we decompose MIA risk into two distinct questions that operate at different levels of granularity: Stage I (Signal Survival) operates at the *client (dataset) level*, exploring whether the overall contribution of the target client remains detectable after (robust) aggregation [7,8], and Stage II (Signal Attribution) operates at the *sample level*, assessing whether a surviving contribution changes a generative membership score in a way attributable to specific data points within the dataset of the target [15,19]. This separation geometrically distinguishes what the system aggregation reveals at a systemic level from what the generative score fundamentally attributes at an individual data level, offering a more analytical and granular privacy assessment than monolithic SOTA evaluations.

Contributions.

- We define a two-stage MIA risk analysis to align accountability with the federated learning protocol process. This allows us to define a client and a sample-level metric, Stage I and Stage II, respectively.
- We formalize Stage I as a target–background decomposition and identify alignment-driven cancellation as the key geometric factor governing survival.
- We derive a robust Stage II attribution bound based on a path representation and smoothness, yielding a linear proxy with an explicit quadratic correction.
- We define a two-stage risk metric (TS-MIA) that combines survival and attribution into a single round-level success probability.

- We empirically verify the bounds and visualize the attribution geometry and risk trajectories under IID and non-IID data partitions.

Moreover, Table 1 provides a detailed comparison between existing privacy attack literature and the proposed TS-MIA framework in federated learning, highlighting the unique contributions and insights offered by our two-stage analysis approach.

Table 1. Comparison between existing privacy attack literature and the proposed TS-MIA framework in federated learning.

Attack Literature	What Existing Methods Measure	What TS-MIA Adds	Correlation with Actual Attack Outcomes
End-to-End MIA (e.g., [14,15])	Monolithic evaluation computing a success metric directly from the aggregated global updates or shadow models.	Decomposes risk analytically into Stage I (system-level aggregation masking) and Stage II (intrinsic data-level vulnerability).	Explains baseline failure cases: isolates if an attack failed due to geometric aggregation cancellation ($\lambda \to -1$) or indistinguishable data ($\gamma \to 0$).
Trajectory-based MIA (e.g., LTMIA [16], FedMIA [17])	Empirical privacy leakage by analyzing loss trajectories and cross-round information across multiple FL stages.	Formalizes the score trajectory via continuous local path representations, providing rigorous Lipschitz and smoothness bounds.	TS-MIA's theoretical metrics (SUR and $J(u)$) mathematically underpin the empirical multi-round vulnerabilities exploited by these baselines.
Black-box MIA (e.g., Res-MIA [18])	Attack success relying on output sensitivity (e.g., resolution variations) without direct access to client gradients.	Connects the generative score advantage directly to target-data alignment (γ), avoiding architecture-specific assumptions.	Models the maximum unobservable true score advantage, establishing a theoretical ceiling for the empirical outcomes of black-box methods.
Centralized Generative MIA (e.g., [19])	Log-likelihood ratios to evaluate membership directly on isolated models without aggregation interference.	Introduces the FL target-background decomposition and establishes the W_{base} reference to isolate target attribution post-aggregation.	Generalizes centralized success conditions to FL, showing how the empirical advantage deviates from the true advantage under background noise.
Gradient Inversion (e.g., DLG [7,8])	Exact data reconstruction from shared model updates, measuring pixel/feature fidelity of recovered samples.	Shifts focus from visual reconstruction to statistical score attribution, standardizing the role of aggregation cancellation.	Demonstrates that both reconstruction and MIA depend heavily on Stage I survival (SUR $> \tau_{\text{noise}}$) before any sample-level data extraction can occur.

The rest of the paper is organized as follows. Section 2 provides preliminaries, Section 3 presents Stage I (signal survival), Section 4 presents Stage II (signal attribution), Section 5 defines the two-stage risk metric, and Section 6 reports empirical verification results. Finally, conclusions are presented.

2. Preliminaries: Aggregation, Threat Model and Regularity

This section establishes the formal notation and foundational assumptions used throughout the analysis, covering the standard mechanism of federated aggregation, the threat model outlining attacker capabilities, and the mathematical regularity properties necessary for the subsequent attribution proofs.

2.1. Federated Aggregation and Decomposition

Formalizing how individual client contributions are combined into a global model update is achieved by mathematically describing the aggregation protocol. Subsequently, a target-background decomposition is introduced to isolate the update of a specific target client from the remaining participants' aggregate.

Many prominent federated aggregation protocols, including Federated Averaging (FedAvg) [3] and various personalized or clustered variants [4], compute the global update ΔW as a weighted linear combination of client contributions:

$$\Delta W = \sum_{k=1}^{K} p_k \Delta w_k = \sum_{k=1}^{K} u_k. \tag{1}$$

Fix a target client $k^\star$ and decompose the global update as

$$\Delta W = u + B, \tag{2}$$

where $u := u_{k^\star}$ is the target weighted update and $B := \sum_{j \neq k^\star} u_j$ is the background aggregate. This decomposition is the basis for Stage I, signal survival, and Stage II, signal attribution.

2.2. Attacker Observation Model

We consider a *passive global attacker* (or *honest-but-curious* server [4,5]) that observes the released update up to a noise floor:

$$Y = \Delta W + \xi, \tag{3}$$

where ξ models aggregate uncertainty and perturbations (e.g., stochasticity, quantization [20], secure aggregation artifacts [21], or injected noise for differential privacy [22]). The attacker does not observe individual updates Δw_k nor the target dataset $D_{k^\star}$.

Rather than modelling the attack itself, we model the worst-case scenario for the target client to assess the maximum possible risk. Therefore, we assume that the adversary has access to both B and u.

2.3. Regularity Definitions

To bound the complex trajectory of the generative score during the Stage II analysis, specific mathematical properties from the generative loss landscape are required. To this end, standard analytic regularity criteria such as Lipschitz continuity and smoothness are formally defined.

Definition 1 (Analytic Regularity). *Let $\Omega \subseteq \mathbb{R}^d$ and let $f : \Omega \to \mathbb{R}$.*

- *f is G-Lipschitz on Ω if for all $a, b \in \Omega$,*

$$|f(a) - f(b)| \leq G\|a - b\|.$$

- *f is L-smooth on Ω if ∇f exists on Ω and is L-Lipschitz on Ω, i.e., for all $a, b \in \Omega$,*

$$\|\nabla f(a) - \nabla f(b)\| \leq L\|a - b\|.$$

3. Stage I: Signal Survival

Determining whether the target contribution u can survive aggregation and remain detectable under the observation model defined in (3) depends on analyzing the geometric alignment between target and background updates, followed by the introduction of a scale-free survival proxy.

3.1. Target–Background Alignment and Aggregation Energy

To quantify how interference between the target update and other clients' updates affects the final aggregated model, a geometric target-background alignment parameter is introduced and subsequently used to decompose the energy of the released update.

Definition 2 (Target–background alignment)**.** *In the Euclidean space $\mathbb{R}^d$, the inner product between two nonzero vectors u and B geometrically encodes their magnitudes and the angle $\theta \in [0, \pi]$ between them:*

$$\langle u, B \rangle = \|u\| \, \|B\| \cos(\theta). \tag{4}$$

This standard relation motivates the definition of their alignment λ as this cosine similarity:

$$\lambda := \cos(\theta) = \frac{\langle u, B \rangle}{\|u\| \, \|B\|}. \tag{5}$$

Lemma 1 (Alignment bounds)**.** *The geometric definition $\cos(\theta) \in [-1, 1]$ directly imposes that the alignment is bounded as $-1 \leq \lambda \leq 1$. We can explicitly verify that the algebraic inner product in $\mathbb{R}^d$ naturally respects this geometric constraint via the Cauchy–Schwarz inequality [23]:*

$$\left|\langle u, B \rangle\right| \leq \|u\| \|B\| \quad \Longrightarrow \quad -1 \leq \frac{\langle u, B \rangle}{\|u\| \|B\|} \leq 1.$$

Finally, in FedAvg, one often has $u = p_{k^\star} \Delta w_{k^\star}$ with $p_{k^\star} > 0$. Whenever $\Delta w_{k^\star} \neq 0$, scaling by a positive constant does not change cosine similarity, so (5) is equivalent to $\cos(\Delta w_{k^\star}, B)$, i.e.,

$$\lambda = \frac{\langle \Delta w_{k^\star}, B \rangle}{\|\Delta w_{k^\star}\| \, \|B\|}. \tag{6}$$

Lemma 2 (Energy decomposition with an interference term)**.** *The squared norm of the released update satisfies:*

$$\|\Delta W\|^2 = \|u\|^2 + \|B\|^2 + 2 \, \|u\| \, \|B\| \, \lambda. \tag{7}$$

Proof. To establish this, we expand the squared Euclidean norm of the sum $\Delta W = u + B$ using the linearity and symmetry properties of the inner product space:

1. Definition of Norm: The squared norm is the inner product of the vector with itself:

$$\|\Delta W\|^2 = \langle u + B, \, u + B \rangle.$$

2. Bilinearity Expansion: Distributing the inner product across the sum gives four terms:

$$\langle u + B, \, u + B \rangle = \langle u, \, u \rangle + \langle u, \, B \rangle + \langle B, \, u \rangle + \langle B, \, B \rangle.$$

3. Symmetry and Substitution: By symmetry, $\langle B, \, u \rangle = \langle u, \, B \rangle$. Substituting $\|u\|^2 = \langle u, \, u \rangle$ and $\|B\|^2 = \langle B, \, B \rangle$ yields

$$\|\Delta W\|^2 = \|u\|^2 + \|B\|^2 + 2 \langle u, \, B \rangle.$$

4. Geometric Substitution: To obtain (7), we use the geometric property of the inner product introduced in Definition 2. Assuming $u \neq 0$ and $B \neq 0$, we substitute $\langle u, \, B \rangle = \|u\| \|B\| \cos(\theta) = \|u\| \|B\| \lambda$:

$$\|\Delta W\|^2 = \|u\|^2 + \|B\|^2 + 2 \, \|u\| \, \|B\| \, \lambda. \quad \square$$

Aggregation cancellation.

If $\lambda < 0$, the cross-term in (7) is negative. In the extreme case $B \to -u$, the released update $\Delta W \to 0$. Consequently, the observation Y can be dominated by the noise ξ if $\|\xi\|$ is large enough.

3.2. A Scale-Free Survival Proxy

To obtain a robust scalar measure of the target's relative dominance within the global aggregation, the Signal-to-Update Ratio (SUR) is introduced as a proxy for the target's relative contribution to the released update energy.

Definition 3 (Signal-to-Update Ratio (SUR)). *Define*

$$\mathrm{SUR} := \frac{\|u\|}{\|\Delta W\|}, \tag{8}$$

whenever $\Delta W \neq 0$. *Equivalently,* $\mathrm{SUR}^2 = \|u\|^2 / \|\Delta W\|^2$.

We compute survival using (8) throughout Stage I.

Lemma 3 (Closed form in terms of magnitude ratio and alignment). *Assume* $\Delta W \neq 0$ *and* $u \neq 0$, *and define* $r := \|B\| / \|u\|$. *Then*

$$\mathrm{SUR}^2 = \frac{1}{1 + r^2 + 2r\lambda}. \tag{9}$$

Proof. Divide (7) by $\|u\|^2$ to get $\|\Delta W\|^2 / \|u\|^2 = 1 + r^2 + 2r\lambda$ and invert. □

3.3. Noise-Limited Detectability

Linking the geometric survival proxy to the practical limitations of an attacker's observation requires introducing a necessary noise-floor condition for non-trivial detection. While the SUR structures the geometry of survival itself, absolute detectability under (3) ultimately relies on this designated threshold.

Assumption 1 (Necessary noise-floor condition for detectability). *Assume there exists a detection threshold* $\tau_{\mathrm{noise}} > 0$ *induced by the observation model* (3) *and the attacker test. A necessary condition for non-trivial detection of the target component is*

$$\mathrm{SUR} > \tau_{\mathrm{noise}}. \tag{10}$$

When cancellation is severe ($B \to -u$), $\|\Delta W\| \to 0$ *and* SUR *diverges; however, the aggregate itself carries negligible information, so detectability is limited by noise.*

Remark 1 (Context-dependent survival thresholds). *Assumption 1 is intentionally general, as the precise formulation of the threshold* τ_{noise} *depends heavily on the specific threat model and the defensive mechanisms implemented in the FL system. For instance:*

- *Differential Privacy (DP): If the server injects carefully calibrated noise to ensure DP guarantees [22],* τ_{noise} *scales directly with the variance of this injected noise, requiring a stronger target signal to avoid being drowned out.*
- *Quantization: When client updates are compressed or quantized to reduce communication bandwidth, the discretization artifacts inherently introduce bounded noise into the aggregate observation [20], establishing a baseline threshold below which tiny signals are lost to rounding.*
- *Secure Aggregation (SecAgg): While SecAgg prevents the server from viewing individual updates, practical implementations can still suffer from artifacts, dropouts, or protocol-specific*

noise boundaries [21] that an attacker would need to overcome to reconstruct a reliable signal from the sum.

By abstracting these effects into τ_{noise}, Stage I elegantly isolates the fundamental geometry of aggregation, via the magnitude ratio r and alignment λ, from these highly system-dependent noise instantiations.

4. Stage II: Signal Attribution

Stage II asks the following question: If the target component survives (Stage I, i.e., (10) is fulfilled), does it lead to an attributable change in a generative membership score on samples from D_{k^*}?

4.1. Generative Score and Centered Advantage

Galende et al. [19] cast membership inference as a Bayesian hypothesis test and show, via the Neyman–Pearson lemma [24], that the optimal test statistic for deciding whether a query sample x belongs to the training set is the *log-likelihood ratio* $\ln p(x \mid \theta_v) - \ln p(x \mid \theta_r)$, where θ_v denotes the victim model and θ_r a reference model trained on non-overlapping data. The ratio isolates the *effect* of having seen x during training from the sample's intrinsic complexity, and the sigmoid σ maps it to a membership probability in $[0, 1]$.

We adopt this likelihood-ratio viewpoint in the federated setting. Let $\ell(x; W)$ be a generative score assigned to example x by parameter W. In many generative settings ℓ is a log-likelihood $\log P(x \mid W)$ [15]; more generally, it can be any differentiable surrogate that correlates with sample fit, without assuming a specific architecture. Because raw scores $\ell(x; W)$ are inherently uncalibrated, directly comparing them across samples is unreliable. Instead, mirroring the role of the reference model in [19], we measure the *score advantage*: the change in ℓ caused by a specific parameter update. Just as the likelihood ratio compares a victim model against a reference to cancel out the intrinsic difficulty of generating x, the score advantage compares the model before and after an update to isolate the exact effect of that specific contribution.

In the federated setting, however, the attacker cannot observe the isolated target update u because the server aggregates it with the background updates B. We therefore distinguish two advantage metrics:

$$\Delta\ell(x) := \ell(x; W + u) - \ell(x; W) \tag{11}$$

is the *true (albeit unobservable) score advantage* evaluated on the target's isolated contribution u, while

$$\Delta\tilde{\ell}(x) := \ell(x; W + \Delta W) - \ell(x; W) \tag{12}$$

is the *empirical score advantage* computed from the aggregated update $\Delta W = u + B$. Stage II bounds the deviation between these two quantities.

Following [19], where the sigmoid maps the log-likelihood ratio to a membership probability, we project the unbounded score advantage onto $[0, 1]$ via $\sigma(z) = \frac{1}{1+e^{-z}}$. Because a neutral advantage evaluates to $\sigma(0) = 0.5$, we re-center the mapping so that any positive increase in sample fit yields a strictly positive metric:

$$\phi(z) := \sigma(z) - \frac{1}{2}. \tag{13}$$

Lemma 4 (Properties of σ and ϕ)**.** *The sigmoid σ is smooth, strictly increasing, and satisfies $\sigma(0) = \frac{1}{2}$. Moreover,*

- ***(Bounded)*** *$0 < \sigma(z) < 1$ for all $z \in \mathbb{R}$, hence $-\frac{1}{2} < \phi(z) < \frac{1}{2}$.*

- *(**Lipschitz**) σ is $\frac{1}{4}$-Lipschitz: for all $a, b \in \mathbb{R}$,*

$$|\sigma(a) - \sigma(b)| \leq \frac{1}{4}|a - b|.$$

Equivalently, ϕ is $\frac{1}{4}$-Lipschitz.

Proof. By definition, $\sigma(z) = \frac{1}{1+e^{-z}}$, so $\sigma(0) = \frac{1}{1+1} = \frac{1}{2}$. The remaining properties are derived as follows:

(i) **Boundedness:** Because $e^{-z} > 0$ for all $z \in \mathbb{R}$, the denominator $1 + e^{-z}$ is always strictly greater than 1. Thus, $0 < \frac{1}{1+e^{-z}} < 1$, which directly implies $-\frac{1}{2} < \phi(z) < \frac{1}{2}$.

(ii) **Lipschitz continuity:** The derivative of the sigmoid is given by $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Because $\sigma(z) \in (0, 1)$, this product is strictly positive, meaning σ is strictly increasing. Furthermore, the quadratic expression $y(1 - y)$ is maximized at $y = \frac{1}{2}$, yielding a maximum derivative of $\frac{1}{4}$ at $z = 0$.
By the Mean Value Theorem, for any $a, b \in \mathbb{R}$ with $a < b$, there exists some $c \in (a, b)$ such that

$$\frac{\sigma(b) - \sigma(a)}{b - a} = \sigma'(c).$$

Taking the absolute value and substituting the global upper bound $\sup_z |\sigma'(z)| = \frac{1}{4}$, we obtain

$$|\sigma(a) - \sigma(b)| \leq \frac{1}{4}|a - b|.$$

Therefore, the sigmoid σ is $\frac{1}{4}$-Lipschitz. Since $\phi(z)$ is simply a vertically translated $\sigma(z)$, it shares the same derivative $\phi'(z) = \sigma'(z)$ and thus the same $\frac{1}{4}$-Lipschitz bound. □

4.2. Baseline for Attribution

Recall that the empirical advantage $\Delta\tilde{\ell}$ measures the total score change caused by the aggregated update $\Delta W = u + B$, whereas the unobservable advantage $\Delta\ell$ measures the isolated effect of u starting from the initial model W. To evaluate what portion of the total empirical change is strictly attributable to the target client, we must evaluate the target's impact after accounting for the background shift.

Definition 4 (Stage II baseline)**.** *Following a leave-one-*client*-out auditing approach [15], we define the Stage II baseline at each round as the model that includes all contributions except the target client's. Let W_t denote the global model at the start of round t and $B = \sum_{j \neq k^\star} u_j$ the background aggregate. The baseline is*

$$W_{\text{base}} := W_t + B. \tag{14}$$

W_{base} represents the "null hypothesis" state. Evaluating the score difference exactly from this baseline isolates the final step of the aggregated update: $\ell(x; W_t + \Delta W) - \ell(x; W_{\text{base}}) = \ell(x; W_{\text{base}} + u) - \ell(x; W_{\text{base}})$. This measures the specific contribution of u, answering how much of the empirical advantage is attributable to the target.

Definition 5 (Attributable centered advantage)**.** *Define the attributable centered advantage of the target perturbation u at baseline W_{base} by*

$$J(u) := \mathbb{E}_{x \sim D_{k^\star}}[\phi(\ell(x; W_{\text{base}} + u) - \ell(x; W_{\text{base}}))]. \tag{15}$$

Interpretation.

$J(u) > 0$ means that adding u on top of the background increases the centered sigmoid transformation of the generative score on target data. This models an effect attributable to the target update in the sense used by generative MIA [19].

4.3. Local Path Analysis and Smoothness

To mathematically decompose the macroscopic score change, the local directional increment must first be defined.

Definition 6 (Local directional increment)**.** *For fixed x and direction $u \in \mathbb{R}^d$, define the local increment*

$$\delta(x;u) := \ell(x; W_{\text{base}} + u) - \ell(x; W_{\text{base}}). \tag{16}$$

To analyze this macroscopic score change caused by the parameter update, we mathematically decompose the transition by integrating microscopic directional gradients along a continuous path. We first bound the smoothness of this trajectory.

Assumption 2 (Smoothness along the target segment)**.** *There exists a constant $L_{k^\star} \geq 0$ such that for all $x \sim D_{k^\star}$ and any $t_1, t_2 \in [0,1]$,*

$$\|\nabla_W \ell(x; W_{\text{base}} + t_1 u) - \nabla_W \ell(x; W_{\text{base}} + t_2 u)\| \leq L_{k^\star} \|u\| |t_1 - t_2|.$$

Lemma 5 (Path integral representation of score increments)**.** *Assume that for fixed x, the function $t \mapsto \ell(x; W_{\text{base}} + tu)$ is differentiable for $t \in [0,1]$. Then*

$$\delta(x;u) = \int_0^1 \langle \nabla_W \ell(x; W_{\text{base}} + tu),\, u \rangle \, dt. \tag{17}$$

Proof. Define the path function $F(t) := \ell(x; W_{\text{base}} + tu)$. By the fundamental theorem of calculus, $F(1) - F(0) = \int_0^1 F'(t)\, dt$. By the chain rule, $F'(t) = \langle \nabla_W \ell(x; W_{\text{base}} + tu),\, u \rangle$. Substitute into the integral. □

To streamline the subsequent bounds, we define the directional gradient path shorthand:

$$H(t) := \nabla_W \ell\big(x; W_{\text{base}} + tu\big). \tag{18}$$

Lemma 6 (First-order approximation with explicit quadratic remainder)**.** *Under Assumption 2, for all $x \sim D_{k^\star}$,*

$$\big|\delta(x;u) - \langle H(0),\, u \rangle\big| \leq \frac{L_{k^\star}}{2} \|u\|^2. \tag{19}$$

Proof. Fix $x \sim D_{k^\star}$. By Lemma 5, we represent the local increment as $\delta(x;u) = \int_0^1 \langle H(t),\, u \rangle \, dt$. Subtracting the initial directional gradient $\langle H(0),\, u \rangle$ and using the linearity of the integral yields:

$$\delta(x;u) - \langle H(0),\, u \rangle = \int_0^1 \langle H(t) - H(0),\, u \rangle \, dt.$$

We proceed to bound the absolute value of this difference using three sequential operations:

1. Integral triangle inequality:

$$\Big|\delta(x;u) - \langle H(0),\, u \rangle\Big| \leq \int_0^1 \Big|\langle H(t) - H(0),\, u \rangle\Big| \, dt. \tag{20}$$

2. Cauchy–Schwarz inequality [23]: For the integrand, we have:

$$\left|\langle H(t) - H(0),\, u\rangle\right| \le \|H(t) - H(0)\|\,\|u\|. \tag{21}$$

3. Gradient smoothness: By Assumption 2, the gradients are Lipschitz-continuous along the segment:

$$\|H(t) - H(0)\| \le L_{k^\star}\,\|tu\| = L_{k^\star}\,t\,\|u\|. \tag{22}$$

Substituting (21) and (22) into the integrand of (20), we can explicitly evaluate the integral:

$$\left|\delta(x;u) - \langle H(0),\, u\rangle\right| \le \int_0^1 L_{k^\star}\,t\,\|u\|^2\,dt = \frac{L_{k^\star}}{2}\,\|u\|^2.$$

Finally, substituting $H(0) = \nabla_W \ell(x; W_{\text{base}})$ recovers the stated remainder bound. □

4.4. A Robust Linear Proxy for Attribution

Lemma 6 bounds the difference between the true increment $\delta(x;u)$ and the linear directional score $\langle \nabla_W \ell(x; W_{\text{base}}),\, u\rangle$. We now transfer this control to the attributable advantage $J(u)$ using Lemma 4.

Theorem 1 (Robust Stage II bound: linear proxy plus quadratic correction)**.** *Assume ϕ is $\frac{1}{4}$-Lipschitz (Lemma 4) and Assumption 2 holds. Then,*

$$J(u) = \mathbb{E}_{x\sim D_{k^\star}}[\phi(\langle H(0),\, u\rangle)] \pm \frac{L_{k^\star}}{8}\|u\|^2. \tag{23}$$

Proof. Fix $x \sim D_{k^\star}$. By Lemma 4, ϕ is $\frac{1}{4}$-Lipschitz, so

$$|\phi(\delta(x;u)) - \phi(\langle H(0),\, u\rangle)| \le \frac{1}{4}|\delta(x;u) - \langle H(0),\, u\rangle|.$$

Apply Lemma 6 to bound the right-hand side by $\frac{1}{4}\cdot\frac{L_{k^\star}}{2}\|u\|^2 = \frac{L_{k^\star}}{8}\|u\|^2$:

$$|\phi(\delta(x;u)) - \phi(\langle H(0),\, u\rangle)| \le \frac{L_{k^\star}}{8}\|u\|^2.$$

Take expectations over $x \sim D_{k^\star}$ and use Jensen's inequality ($|\mathbb{E}[\cdot]| \le \mathbb{E}[|\cdot|]$) to obtain (23). □

4.5. Mean Score Geometry and an Alignment Parameter

Theorem 1 reduces attribution to the distribution of directional scores $\langle H(0),\, u\rangle$ plus a controlled quadratic correction. A scalar summary is obtained by the mean score vector and its cosine alignment with u.

Definition 7 (Mean score vector at the Stage II baseline)**.** *Define*

$$s := \mathbb{E}_{x\sim D_{k^\star}}[H(0)] \in \mathbb{R}^d. \tag{24}$$

Definition 8 (Target-data alignment)**.** *Assume $u \neq 0$ and $s \neq 0$. Define*

$$\gamma := \cos(u,\, s) = \frac{\langle u,\, s\rangle}{\|u\|\|s\|} \in [-1, 1]. \tag{25}$$

Equivalently, $\langle s,\, u\rangle = \|s\|\|u\|\gamma$. Since $u = p_{k^\star}\Delta w_{k^\star}$ with $p_{k^\star} > 0$ in FedAvg, $\cos(u,s) = \cos(\Delta w_{k^\star}, s)$.

Interpretation.

The alignment γ measures whether the target update direction increases the generative score on target-data at the baseline W_{base}. This is a target-data geometry quantity, distinct from the federation alignment λ in Stage I.

4.6. A Mean-Score Proxy and a γ-Dominant Regime

To formally connect positive attribution risk to the target-data alignment, it is shown that under sufficient data concentration, a positive alignment gracefully overpowers the quadratic approximation error.

Lemma 7 (Mean-score proxy under concentration)**.** *If the directional scores $\langle H(0), u\rangle$ are concentrated around their mean $\langle s, u\rangle = \|s\|\|u\|\gamma$, then the sign and magnitude of $J(u)$ in (23) are governed by γ up to the quadratic correction scale $\frac{L_{k^\star}}{8}\|u\|^2$.*

Proof. Let $a(x) := \langle H(0), u\rangle$ denote the directional scores, with mean $\mathbb{E}[a] = \langle s, u\rangle = \|s\|\|u\|\gamma$. Theorem 1 establishes that

$$J(u) = \mathbb{E}_{x\sim D_{k^\star}}[\phi(a(x))] \pm \frac{L_{k^\star}}{8}\|u\|^2.$$

We analyze the expectation term by comparing it to the evaluation of ϕ precisely at the mean $\mathbb{E}[a]$. By adding and subtracting $\phi(\mathbb{E}[a])$, we can bound the deviation using Jensen's inequality ($|\mathbb{E}[\cdot]| \le \mathbb{E}[|\cdot|]$):

$$\left|\mathbb{E}_{x\sim D_{k^\star}}[\phi(a(x))] - \phi(\mathbb{E}[a])\right| \le \mathbb{E}_{x\sim D_{k^\star}}\left[\left|\phi(a(x)) - \phi(\mathbb{E}[a])\right|\right].$$

By Lemma 4, the centered sigmoid transformation ϕ is L_ϕ-Lipschitz continuous with $L_\phi = \frac{1}{4}$, implying

$$\left|\phi(a(x)) - \phi(\mathbb{E}[a])\right| \le \frac{1}{4}\left|a(x) - \mathbb{E}[a]\right|.$$

Substituting this bound into the expectation yields

$$\left|\mathbb{E}_{x\sim D_{k^\star}}[\phi(a(x))] - \phi(\mathbb{E}[a])\right| \le \frac{1}{4}\mathbb{E}_{x\sim D_{k^\star}}\left[\left|a(x) - \mathbb{E}[a]\right|\right].$$

The right-hand side is the expected absolute deviation. Applying the Cauchy–Schwarz inequality [23], we introduce the variance $\mathrm{Var}(a)$ associated with the data distribution:

$$\mathbb{E}_{x\sim D_{k^\star}}\left[\left|a(x) - \mathbb{E}[a]\right|\right] \le \sqrt{\mathrm{Var}(a)}.$$

Thus, the leading expectation term evaluates to exactly

$$\mathbb{E}_{x\sim D_{k^\star}}[\phi(a(x))] = \phi(\|s\|\|u\|\gamma) \pm \frac{1}{4}\sqrt{\mathrm{Var}(\langle H(0), u\rangle)}.$$

We substitute this concentrated evaluation back into the overarching bound from Theorem 1, resulting in a complete algebraic decomposition mapping attribution to alignment:

$$J(u) = \phi(\|s\|\|u\|\gamma) \pm \left(\frac{1}{4}\sqrt{\mathrm{Var}(\langle H(0), u\rangle)} + \frac{L_{k^\star}}{8}\|u\|^2\right).$$

Therefore, if the variance of the directional scores is sufficiently small (high concentration), the proxy converges to $\phi(\|s\|\|u\|\gamma)$. In this concentrated regime, the alignment parameter γ dictates the sign and governs the magnitude of the attribution, up to the fixed quadratic correction $\frac{L_{k^\star}}{8}\|u\|^2$. □

Lemma 8 (Sufficient condition for positive attribution risk)**.** *Assume $s \neq 0$ and $u \neq 0$. If $\gamma > 0$ and*

$$\|u\| < \frac{2\,\|s\|\,\gamma}{L_{k^\star}}, \tag{26}$$

then the attribution proxy $\widehat{J}_m$ is strictly guaranteed to be positive, provided the directional scores $a(x)$ are sufficiently concentrated around their mean (as per Lemma 7). This provides an absolute condition under which the positive alignment ($\gamma > 0$) mathematically overpowers the worst-case quadratic error resulting from trajectory curvature.

Proof. By Theorem 1, the worst-case error bound on the un-transformed local increment $\delta(x;u)$ is $\frac{L_{k^\star}}{2}\|u\|^2$. In the concentrated regime (Lemma 7), the expected linear pre-sigmoid signal is exactly $\mathbb{E}[a] = \langle s,\, u\rangle = \|s\|\|u\|\gamma > 0$. The total pre-sigmoid signal, including the worst-case negative approximation error, is strictly bounded from below by the difference:

$$z_{\min} = \|s\|\|u\|\gamma - \frac{L_{k^\star}}{2}\|u\|^2.$$

Because the centered sigmoid $\phi(z)$ is a strictly monotonically increasing function with $\phi(0) = 0$ (Lemma 4), any strictly positive input $z_{\min} > 0$ guarantees a strictly positive output $\phi(z_{\min}) > 0$. Therefore, to guarantee positive attribution, it is necessary and sufficient that this worst-case pre-sigmoid sum remains strictly positive:

$$\|s\|\|u\|\gamma - \frac{L_{k^\star}}{2}\|u\|^2 > 0.$$

Because the update vector is non-zero ($\|u\| > 0$), we can rearrange the inequality by dividing by $\|u\|$:

$$\|s\|\gamma > \frac{L_{k^\star}}{2}\|u\|.$$

Isolating $\|u\|$ yields the exact bound:

$$\|u\| < \frac{2\|s\|\gamma}{L_{k^\star}}.$$

This mathematically guarantees a net positive pre-sigmoid signal, which the strictly increasing property of ϕ translates into a guaranteed positive attribution risk, requiring no linear approximations. □

5. A Joint Two-Stage Metric for MIA Risk

Combining the survival and attribution conditions into a unified probability of success is done by constructing a joint metric that properly accounts for both round-to-round variation and finite-sample estimation.

Stage I describes whether the target component survives aggregation strongly enough to be detectable under the observation model (3). This metric operates at the client level as we are measuring the target-background alignment regarding client updates. Stage II describes whether this surviving component produces an attributable change in a generative membership score on the target dataset. This metric operates at the sample level, as we are estimating densities in the input space. We summarize both stages with a single probability-of-success metric that accounts for (i) round-to-round variation in the background aggregate B and (ii) finite-sample estimation of attribution on $D_{k^\star}$.

5.1. Round-Level Survival Statistic

The survival event is first formalized by introducing a round-dependent signal-to-update ratio and evaluating it against the noise floor threshold.

Recall the decomposition $\Delta W = u + B$ from (2). Define the round-dependent signal-to-update ratio

$$\mathrm{SUR}(B) := \frac{\|u\|}{\|u + B\|}, \tag{27}$$

and the survival event

$$\mathcal{S} := \{\mathrm{SUR}(B) > \tau_{\mathrm{noise}}\}, \tag{28}$$

where $\tau_{\mathrm{noise}} > 0$ is the noise-floor threshold from Assumption 1. Here, B is treated as random across rounds due to client sampling, heterogeneity, and optimization noise.

5.2. Finite-Sample Attribution Estimator

The empirical advantage corresponding to the target contribution is then formalized by defining a finite-sample estimator of the centered advantage across multiple independent samples.

For a given round (hence fixed B) and a target example x, define the score increment

$$\delta(x; u, B) := \ell\big(x; W_{\mathrm{base}} + u\big) - \ell\big(x; W_{\mathrm{base}}\big), \qquad W_{\mathrm{base}} = W_t + B \text{ as in (14)}. \tag{29}$$

Given $m \geq 1$ samples $x_1, \ldots, x_m \sim D_{k^\star}$ drawn i.i.d., define the empirical Stage II advantage

$$\widehat{J}_m(u; B) := \frac{1}{m} \sum_{i=1}^{m} \phi\big(\delta(x_i; u, B)\big). \tag{30}$$

This is the finite-sample analogue of $J(u)$ in Definition 5.

5.3. Two-Stage Success Probability

Bringing these elements together, the final joint risk metric is constructed by considering the combined probability that the target contribution is simultaneously detectable and attributable.

Define the attribution event at sample size m by

$$\mathcal{A}_m := \{\widehat{J}_m(u; B) > 0\}. \tag{31}$$

We define the joint two-stage metric as the probability (over both round-to-round variability in B and the draw of target samples) that the target contribution is simultaneously detectable and attributable:

$$\text{TS-MIA}_m \;:=\; \Pr\left(\mathcal{S} \cap \mathcal{A}_m\right) \;=\; \Pr\big(\mathrm{SUR}(B) > \tau_{\mathrm{noise}} \;\wedge\; \widehat{J}_m(u; B) > 0\big). \tag{32}$$

By the definition of conditional probability, $\text{TS-MIA}_m = \Pr(\mathcal{S})\,\Pr(\mathcal{A}_m \mid \mathcal{S})$, reflecting the two-stage interpretation: survival is necessary for detectability, and attribution is only evaluated and only meaningful conditional on survival.

Remark 2 (Interpretation and limiting cases)**.** *The metric* TS-MIA_m *summarizes MIA risk at the round level. It increases when (i) aggregation cancellation is weak (so* $\mathrm{SUR}(B)$ *is often above* τ_{noise}*) and (ii) the target update tends to increase the centered generative advantage on* $D_{k^\star}$ *(so* $\widehat{J}_m(u; B) > 0$ *often holds). As* $m \to \infty$*, under standard laws of large numbers,* $\widehat{J}_m(u; B)$ *concentrates around* $J(u)$ *from (15), so* TS-MIA_m *approaches the round-level probability* $\Pr\left(\mathrm{SUR}(B) > \tau_{\mathrm{noise}} \;\wedge\; J(u) > 0\right)$ *whenever* $J(u)$ *is well-defined.*

https://doi.org/10.3390/bdcc10050163

6. Empirical Verification

Validating the theoretical bounds and visualizing privacy risks computationally relies on carrying out simulated experiments under a number of federated scenarios and analyzing the resulting risk trajectories. The subsequent empirical verification figures illustrate these outcomes; each figure is accompanied by a concise description of the experimental setup and the qualitative outcome visible in the plot.

6.1. Stage I: SUR Geometry and Cancellation

6.1.1. Setup

We construct synthetic target/background pairs in $\mathbb{R}^{100}$ over a grid of $r \in [0.01, 16]$ (80 values) and $\lambda \in [-1, 1]$ (80 values). With a unit target direction $\hat{u}$ and an orthogonal vector $v_\perp$, we set $B = r(\lambda\hat{u} + \sqrt{1-\lambda^2}v_\perp)$, compute $\Delta W = \hat{u} + B$, and compare $\mathrm{SUR} = \|\hat{u}\|/\|\Delta W\|$ to the closed form in Lemma 3. The figure uses a log color scale, restricts to $r \leq 6$ for visibility, and marks the cancellation region $\|\Delta W\| < \varepsilon$ with $\varepsilon = 0.3$.

6.1.2. Result

The structural breakdown of SUR, as shown in Figure 1, is consistent with the geometric bounds formalized in Lemma 2 and the closed-form equation $\mathrm{SUR}^2 = 1/(1 + r^2 + 2r\lambda)$ in (9) (Lemma 3). The heatmap exhibits the expected ridge of high SUR at small r, where the target dominates. Conversely, it reveals a distinct cancellation band. From the detectability standpoint explicitly formulated in Assumption 1, the region encompassing $\mathrm{SUR} > \tau_{\text{noise}}$ systematically shrinks as the background relative magnitude r grows. Crucially, when $r \gtrsim 1$ and the federation alignment is adversarial ($\lambda \to -1$), the aggregate approaches cancellation; this geometric interference neutralizes the target signal, driving it below the system's noise floor and rendering the contribution undetectable.

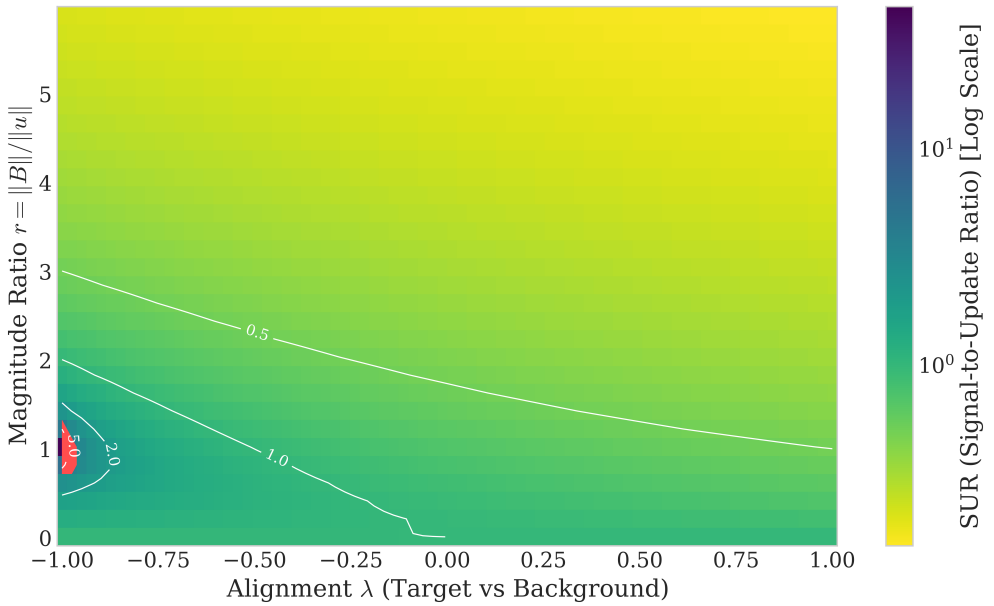

Figure 1. Stage I SUR geometry. Log-scaled heatmap of $\mathrm{SUR} = \|u\|/\|\Delta W\|$ over (λ, r) with $r = \|B\|/\|u\|$, and a red overlay indicating the cancellation region $\|\Delta W\| < \varepsilon$. High SUR occurs when the background is small ($r \to 0$) or cooperative ($\lambda > 0$), while the cancellation band near $\lambda \to -1$ and $r \to 1$ highlights where ΔW is near-zero and detectability becomes noise-limited.

6.2. Stage II: Smoothness Bound Verification

6.2.1. Setup

This experiment verifies Lemma 6 and Theorem 1 in Section 4. We use PathMNIST (train split) and a variational autoencoder (VAE) [25] with three input channels, latent dimension 32, and 10 classes. Inputs are scaled by $1/255$, and double precision is used during perturbation checks. We use 50 images and 30 perturbation scales. In this setup, let $g_{\text{loss}} = \nabla(\text{loss})$ and note that the score gradient is $\nabla \ell = -g_{\text{loss}}$. We define the unit score gradient $\hat{g} = \nabla \ell / \|\nabla \ell\|$ (equivalently, $-g_{\text{loss}}/\|g_{\text{loss}}\|$) and use perturbations $v = (\alpha / \|\nabla \ell\|)\hat{g}$ with $\alpha \sim \text{Uniform}(0.01, 2.0)$. We sort the 30 scales and index them in ascending order; scales with odd indices $(1, 3, 5, \ldots, 29)$ form a calibration set to estimate L via $L_{\text{cal}} = \max(2\,\text{err}/\|v\|^2)$, and even-indexed scales form a test set to verify the quadratic bound $\frac{L_{\text{cal}}}{2}\|v\|^2$. We report a 95% Wilson CI [26] for the pooled test-point violation rate. In this setup, the theoretical score ℓ is instantiated as the negative VAE loss (i.e., $\ell = -\text{loss}$).

6.2.2. Result

Most test points adhere to the explicitly calibrated quadratic bound, as shown in Figure 2, directly validating the core mechanism of Lemma 6. This tight bound confirms that the true unobservable local score increment δ is strongly governed by the linear directional gradient, offset by a predictable quadratic curvature penalty $L/2\|u\|^2$. By validating this error structure empirically, the result underpins the robust attribution derivation in Theorem 1, confirming that Stage II may rely on a tractable linear proxy. The overall test violation rate sits at 6.7% with a 95% Wilson CI of $[5.1\%, 8.7\%]$, computed over all test points pooled across trials. The median empirical smoothness constant is 6.64×10^3.

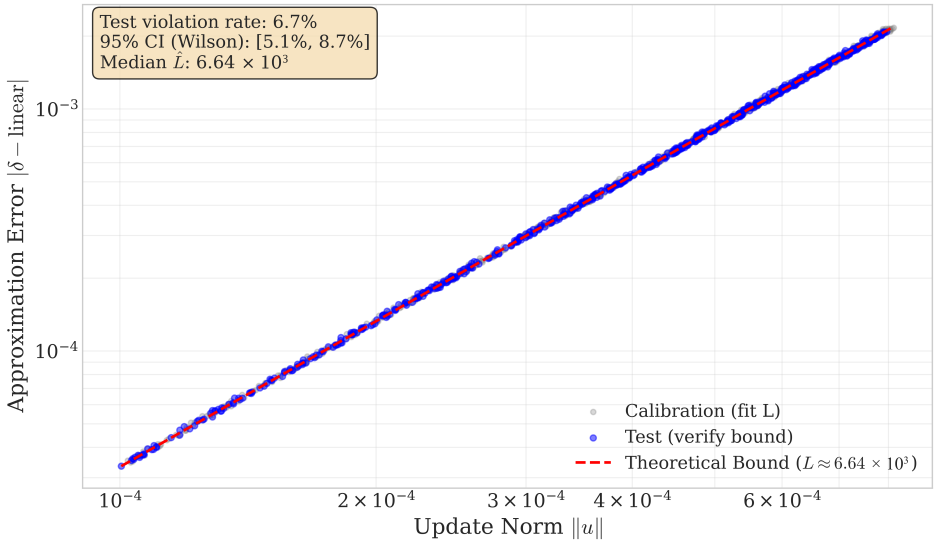

Figure 2. Stage II smoothness bound. Test errors vs. update norm $\|u\|$ with the quadratic bound estimated from calibration points. The log–log scale highlights the expected quadratic scaling; the text box in the figure reports the empirical violation rate and its 95% CI.

6.3. Stage II Geometry: Attribution vs. Alignment

6.3.1. Setup

We use a single PathMNIST sample and compute the VAE gradient g at the baseline parameters. We construct a 2D subspace spanned by $g/\|g\|$ and an orthogonal vector. For alignments $\cos\theta$ over $\theta \in [0, \pi]$ (50 points), we evaluate attribution $J(v)$ for magnitudes chosen so that the linear predictor $z = \|g\|\|v\|\cos\theta$ spans the sigmoid transition region. Concretely, we set $\|v\| = \tau/\|g\|$ with $\tau \in \{0.1, 0.5, 1, 2, 4\}$ so that $|z|$ reaches

$\{0.1, 0.5, 1, 2, 4\}$ at $|\cos\theta| = 1$. VAE noise is controlled with a fixed seed (42). The linear proxy is $-\|g\|\|v\|\cos\theta$ passed through the centered sigmoid.

6.3.2. Result

For small-magnitude perturbations, the true attribution tracks the linear proxy, as shown in Figure 3, corresponding directly to the concentrated regime theorized in Lemma 7, where the response geometry is explicitly dictated by the target-data alignment γ. As the perturbation magnitude scales upward, the attribution exhibits the anticipated controlled saturation governed by the strictly bounded and Lipschitz nature of the centered sigmoid ϕ (Lemma 4). This visual confirms that within the γ-dominant region defined in Lemma 8, positive target-data alignment overcomes the quadratic error limits without escaping the fundamental $[-0.5, 0.5]$ scope of the metric.

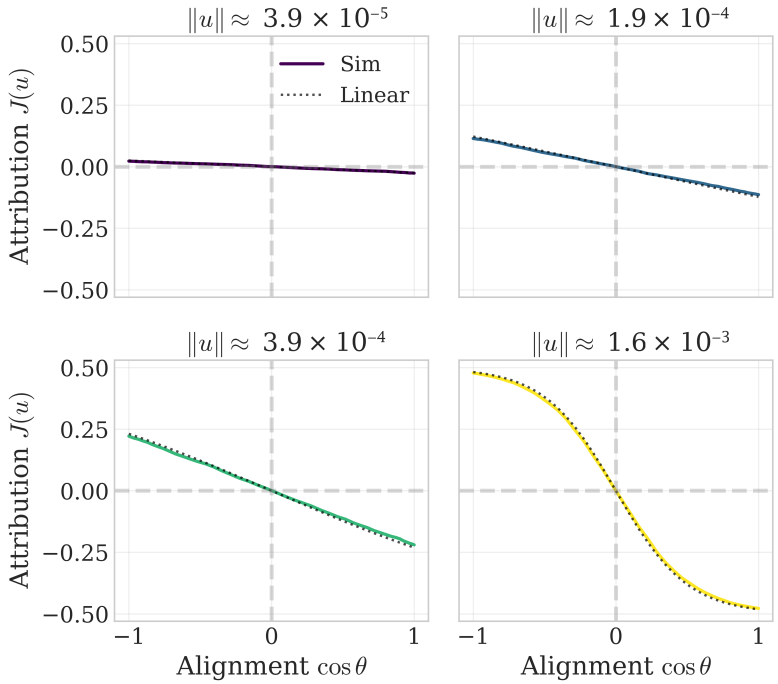

Figure 3. Stage II geometry. $J(v)$ vs. alignment for multiple update magnitudes, with the linear proxy overlaid.

6.4. Risk Trajectories Across FL Scenarios

6.4.1. Setup

We simulate FedAvg with $K = 5$ clients, 20 rounds, 3 local epochs, and Adam at a learning rate of 10^{-3}. Clients train a VAE on PathMNIST (train split) with a data limit of 2000 and inputs scaled by $1/255$. A run seed (1407664341 in this run) generates five per-scenario seeds, each reinitializing the random number generation state before partitioning and training. Each round starts from the current global model; each client performs local training and produces Δw_k; the server averages updates (FedAvg with $p_k = 1/K$) to form ΔW and updates the global model. We compute per-client SUR $= \|(1/K)\Delta w_k\|/\|\Delta W\|$ and J via score differences between the background baseline and the full update. Scenarios are **IID** (random partition), **Non-IID** (label-sorted partition), and **Inverted** (all non-target clients use inverted images). We report mean trajectories with bootstrap CIs across seeds for the indicator $\mathbf{1}\{\text{SUR} > \tau_{\text{noise}} \wedge J > 0\}$ with $\tau_{\text{noise}} = 0.1$.

6.4.2. Result

In this run, depicted in Figure 4, the per-round mean detection rate for IID starts at 1 and quickly decays to 0 (by mid-rounds), Non-IID stays safely anchored at 1 throughout,

and Inverted shows an intermediate, highly variant trajectory (roughly 0.32–1.0 across rounds). Because $K = 5$ keeps the relative background size r low, all scenarios satisfy the Stage I requirement (SUR $> \tau_{\text{noise}}$) throughout. Consequently, the distinct trajectories are determined by the underlying target-data geometry defined in Stage II (the functional sign of J). For the Non-IID paradigm, the target's distinct data distribution ensures a strong, positive target-data alignment ($\gamma > 0$) that fully satisfies the γ-dominant criterion from Lemma 8, overpowering optimization curvature bounds round-by-round. Conversely, since IID clients share homogenized environments, the global model rapidly assimilates the common target morphology. This fundamentally suppresses the target-data alignment ($\gamma \to 0$), effectively protecting the target's privacy as the attribution proxy safely succumbs to trailing quadratic errors and negative J fluctuations that extinguish the overall TS-MIA risk.

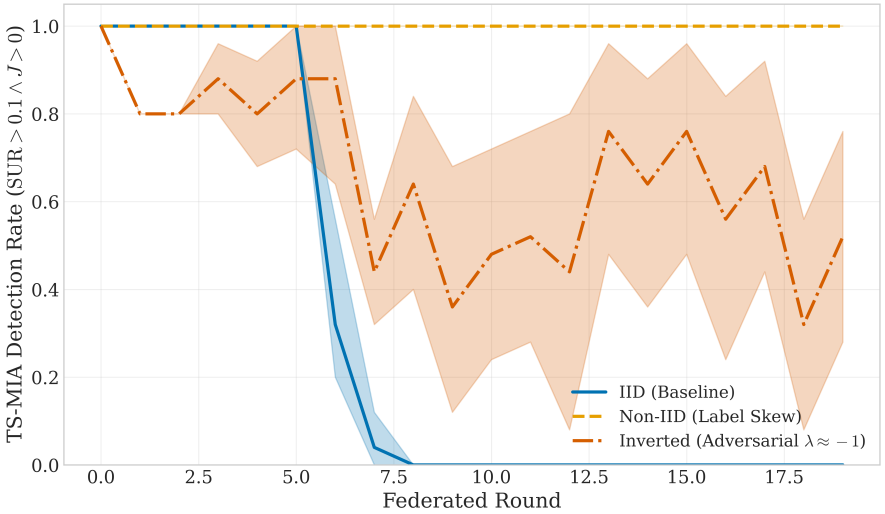

Figure 4. Risk trajectories. Mean TS-MIA detection rate across rounds for IID, Non-IID, and Inverted scenarios. Shaded regions denote bootstrap confidence bands across seeds.

6.5. Non-IID Example Image and Per-Round Metrics

6.5.1. Setup

We display a representative PathMNIST image drawn from the Non-IID partition (client 0, sample 0, with the label shown). The right panel plots three per-round metrics for the Non-IID scenario: mean Risk (TS-MIA), mean SUR, and mean J, each averaged across clients and seeds with shared y-axis bounds.

6.5.2. Result

Figure 5 summarises the results. In the non-IID case, mean risk holds at 1.0 analytically across all observed rounds. The mean SUR sits around 0.23 (range 0.20–0.27), and the mean J operates in the positive region (range 0.010–0.50). Because both formal conditions of the two-stage metric formalized in Equation (32) remain satisfied in our simulation results, specifically SUR $> \tau_{\text{noise}}$ for deterministic survival and $J > 0$ showing the attributable score trace, the combined TS-MIA joint likelihood saturates at 1. This confirms that the non-IID structural label skew embeds a vulnerable regime: the target client's contribution maintains geometrically distinct separation from the aggregation landscape to remain detectable and attributable across all progressive federated steps.

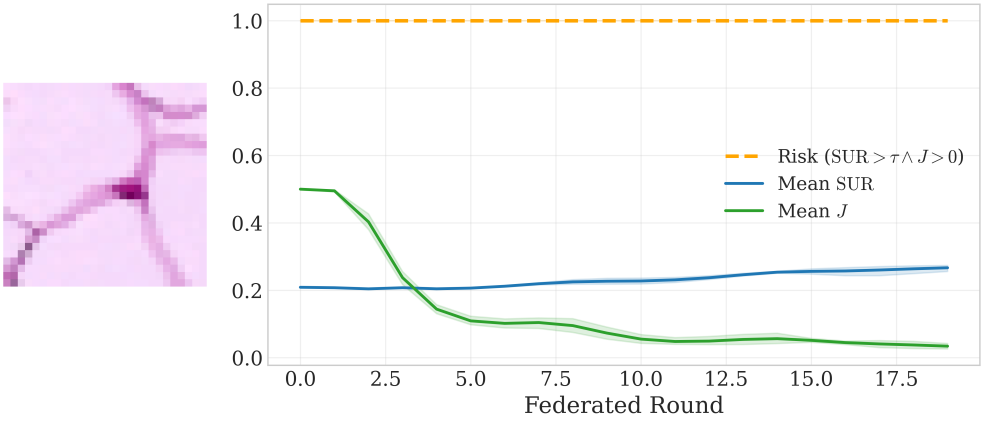

Figure 5. Non-IID example and metrics. (**Left**) A representative pathMNIST sample (Client 0, Label 6) from a label-skewed partition. (**Right**) The non-IID mean risk, mean SUR, and mean J across rounds.

7. Conclusions

This paper introduced a foundational two-stage geometric framework for Membership Inference Attacks (MIAs) on generative models in federated learning. By explicitly separating the confounding factors of aggregation system noise from fundamental data vulnerability, we provided an auditor-evaluable metric, TS-MIA, that accurately captures how data distribution and optimization trajectory shape privacy risks.

Our theoretical analysis of Stage I formalized the role of target-background alignment (λ) via an energy decomposition (Lemma 2). We analytically established the necessary condition for target survival, and closed-form simulations of the Signal-to-Update Ratio (SUR) recovered the exact geometry: adversarial alignments drive the target's surviving signal below the system's noise floor, neutralizing it before Stage II evaluation.

For the surviving updates, our derivation in Stage II transitioned the unobservable generative membership score into a robust linear proxy combined with quadratic error correction (Theorem 1 and Lemma 6). This decoupling confirmed that the target's empirical attribution is impacted by the target-data alignment (γ). The smoothness calibrations across the derived bounds formally demonstrated that small-scale perturbations track the concentrated linear regime, predictably transitioning into the expected controlled saturation layer.

By unifying these dual stages, the consolidated TS-MIA metric dynamically models how structural scenarios dictate distinct patterns of risk. Specifically, as verified mathematically and simulated via FL trajectories, the presence of isolated non-IID label skew anchors the prediction mechanics in a γ-dominant orientation (Lemma 8), locking the MIA likelihood across all communication rounds. Conversely, globally homogenized IID environments swiftly assimilate the target morphology; this structural absorption suppresses analytical alignment, enabling mathematical optimization curvature laws to mitigate downstream privacy leakage. Moving forward, this geometric framework provides both precisely bounded theoretical definitions and practical computational mechanisms to thoroughly pinpoint vulnerabilities and scale privacy policies.

In terms of future work, this article opens a broad research line. First, the application of this framework could be studied under different data modalities and models, including implicit probabilistic models, such as generative adversarial networks, which require surrogate models to estimate densities. Second, the framework provides a set of reference equations that have been selected to properly capture the main principles. However, our equations could be revisited to either relax some of the assumptions or align with other MIA approaches that are not dependent on density ratios. Third, we select the worst-case

scenario to measure the notion of risk. Nevertheless, the chances of this scenario happening are low. Therefore, future studies could explore the threat models that are more likely to happen.

Author Contributions: Conceptualization, B.A.G. and J.P.; methodology, B.A.G. and J.P.; software, B.A.G.; formal analysis, B.A.G. and J.P.; writing—original draft preparation, B.A.G.; writing—review and editing, B.A.G., J.P., A.A., P.A.A., S.U. and F.Á.; visualization, B.A.G.; funding, F.Á. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Synthetic Generation of Hematological Data over Federated Computing Frameworks (SYNTHEMA) project from Horizon Europe under Grant 101095530. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The simulation code is available at https://github.com/BorjaArroyo/mia-fl (accessed on 12 May 2026).

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

Symbol	Meaning
$\langle u, v \rangle$	Inner product
$\|u\|$	Euclidean norm ($\|u\| := \sqrt{\langle u, u \rangle}$)
K	Number of clients in a round
$k^\star$	Target client index
$\Delta w_k \in \mathbb{R}^d$	Client k local update vector (one FL round)
p_k	Aggregation weight (FedAvg), $\sum_{k=1}^{K} p_k = 1$, $p_k \geq 0$
$\Delta W \in \mathbb{R}^d$	Released global update, $\Delta W = \sum_{k=1}^{K} p_k \Delta w_k$
u_k	Weighted client update, $u_k := p_k \Delta w_k$
u	Target weighted update, $u := u_{k^\star} = p_{k^\star} \Delta w_{k^\star}$
B	Background aggregate, $B := \sum_{j \neq k^\star} u_j$
λ	Cosine alignment between u and B
$D_{k^\star}$	Private dataset of the target client
W	Model parameters (vector in $\mathbb{R}^d$)
$\ell(x; W)$	Generative score (log-likelihood or differentiable surrogate)
$\sigma(z)$	Sigmoid, $\sigma(z) = \frac{1}{1+e^{-z}}$
$\phi(z)$	Centered sigmoid, $\phi(z) = \sigma(z) - \frac{1}{2}$
s	Mean score vector, $s := \mathbb{E}_{x \sim D_{k^\star}} [\nabla_W \ell(x; W_{\text{base}})]$
γ	Cosine alignment between u and s
$a(x)$	Directional score, $a(x) := \langle \nabla_W \ell(x; W_{\text{base}}), u \rangle$
$\wedge$	Logical AND (conjunction)

References

1. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). 2016. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 12 May 2026).
2. United States Congress. Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub. L. No. 104-191, 110 Stat. 1936. 1996. Available online: https://www.govinfo.gov/app/details/PLAW-104publ191 (accessed on 12 May 2026).
3. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the Artificial Intelligence and Statistics*; PMLR: New York, NY, USA, 2017; pp. 1273–1282.

4. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
5. Mothukuri, V.; Parizi, R.M.; Pouriyeh, S.; Huang, Y.; Dehghantanha, A.; Choo, K.K.R. A survey on security and privacy of federated learning. *Future Gener. Comput. Syst.* **2021**, *115*, 619–640. [CrossRef]
6. Yin, X.; Zhu, Y.; Hu, J. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–36. [CrossRef]
7. Zhu, L.; Liu, Z.; Han, S. Deep leakage from gradients. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14774–14784. Available online: https://dl.acm.org/doi/10.5555/3454287.3455610 (accessed on 22 March 2026).
8. Geiping, J.; Bauermeister, H.; Dröge, H.; Moeller, M. Inverting gradients-how easy is it to break privacy in federated learning? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 16937–16947.
9. Melis, L.; Song, C.; De Cristofaro, E.; Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*; IEEE: New York, NY, USA, 2019; pp. 691–706. [CrossRef]
10. Wang, Z.; Song, M.; Zhang, Z.; Song, Y.; Wang, Q.; Qi, H. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications*; IEEE: New York, NY, USA, 2019; pp. 2512–2520.
11. Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; Shmatikov, V. How to backdoor federated learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*; PMLR: New York, NY, USA, 2020; pp. 2938–2948.
12. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership Inference Attacks against machine learning models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*; IEEE: New York, NY, USA, 2017; pp. 3–18.
13. Yeom, S.; Giacomelli, I.; Fredrikson, M.; Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of the 2018 IEEE 31st Computer Security Foundations Symposium (CSF)*; IEEE: New York, NY, USA, 2018; pp. 268–282.
14. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*; IEEE: New York, NY, USA, 2019; pp. 739–753. [CrossRef]
15. Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; Tramer, F. Membership Inference Attacks from first principles. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP)*; IEEE: New York, NY, USA, 2022; pp. 1897–1914. [CrossRef]
16. Song, J.; Yuan, J.; Chen, G.; Liu, Y.; Yang, N. Ltmia: A loss trajectory-based Membership Inference Attack method in federated learning. *J. Inf. Secur. Appl.* **2026**, *97*, 104364. [CrossRef]
17. Zhu, G.; Li, D.; Gu, H.; Yao, Y.; Fan, L.; Han, Y. Fedmia: An effective Membership Inference Attack exploiting all for one principle in federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 20643–20653.
18. Zare, M.; Shamsinejadbabaki, P. Res-MIA: A training-free resolution-based membership inference attack on federated learning models. *arXiv* **2026**, arXiv:2601.17378. [CrossRef]
19. Galende, B.A.; Apellániz, P.A.; Parras, J.; Zazo, S.; Uribe, S. Membership Inference Attacks and Differential Privacy: A Study Within the Context of Generative Models. *IEEE Open J. Comput. Soc.* **2025**, *6*, 801–811. [CrossRef]
20. Kang, T.; Li, M.; Li, Y.; Li, J.; Lu, W. The effect of quantization in federated learning: A Rényi differential privacy perspective. *arXiv* **2024**, arXiv:2405.10096. [CrossRef]
21. Gilbert, J.R. Secure aggregation is not all you need: Mitigating privacy attacks with noise tolerance in federated learning. *arXiv* **2022**, arXiv:2211.06324. [CrossRef]
22. Kim, M.; Zong, H. Effects of quantization on federated learning with local differential privacy. *IEEE Access* **2022**, *10*, 11153–11166. [CrossRef]
23. Steele, J.M. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*; Cambridge University Press: Cambridge, UK, 2004. [CrossRef]
24. Neyman, J.; Pearson, E.S. IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. London. Ser. A Contain. Pap. Math. Phys. Character* **1933**, *231*, 289–337. [CrossRef]
25. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. In Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014), Banff, AB, Canada, 14–16 April 2014.
26. Wilson, E.B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **1927**, *22*, 209–212. [CrossRef]