

Sequential Membership Inference Attacks

Thomas Michel, Debabrota Basu, Emilie Kaufmann

Inria, Univ. Lille, CNRS, Centrale Lille, CRISStAL

CAp 2026



Paper



Models are released as a sequence



Spam filter

retrained as users flag messages



A hospital fine-tunes a public chest X-ray model

two releases: the public base model, then the fine-tuned model



Smartphone keyboard next-word language model

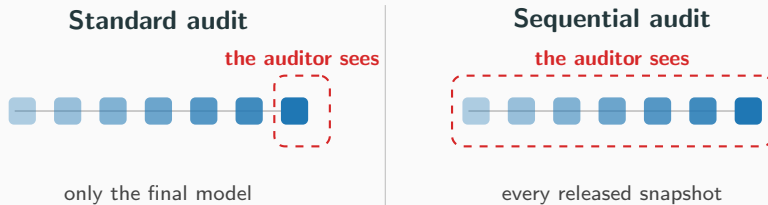
federated learning across phones, with DP



What the auditor sees

A **privacy audit** measures how much a released model reveals about any single training record.

Auditing only the final model under-reports leakage: we give the optimal attack that uses the whole sequence.



- 1 Optimal sequential MI attack
- 2 Extension to (DP-)SGD
- 3 Tighter empirical DP-SGD audits

Differential privacy and auditing

Differential privacy.

$$\mathbb{P}[\mathcal{M}(D) \in S] \leq e^\epsilon \mathbb{P}[\mathcal{M}(D') \in S] + \delta$$

for neighbouring training datasets D, D' (differ in one record), mechanism $\mathcal{M}$, and every event S .

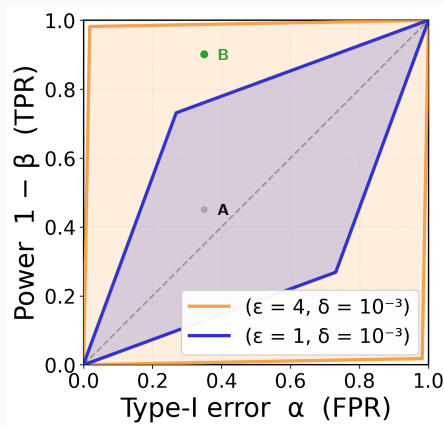
Membership inference.

A test deciding whether $z^* \in D$; errors α (false alarm), β (miss).

Audit (Kairouz et al., 2015):

$$\epsilon \geq \log \max\left(\frac{1-\delta-\alpha}{\beta}, \frac{1-\delta-\beta}{\alpha}\right)$$

**A stronger attack certifies a larger ϵ :
better attacks mean tighter audits.**

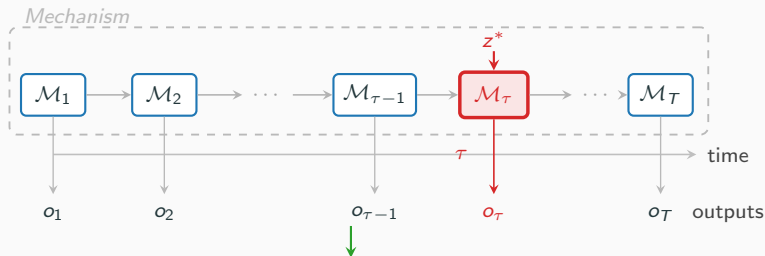


The (α, β) trade-off region of an (ϵ, δ) -DP mechanism.

A: allowed by both ϵ . B: only by $\epsilon = 4 \Rightarrow$ certifies $\epsilon > 1$. 4

The Sequential MI game

The **crafter** picks a target z^* and a time τ ; with probability $1/2$ it inserts z^* into the mechanism at step τ .



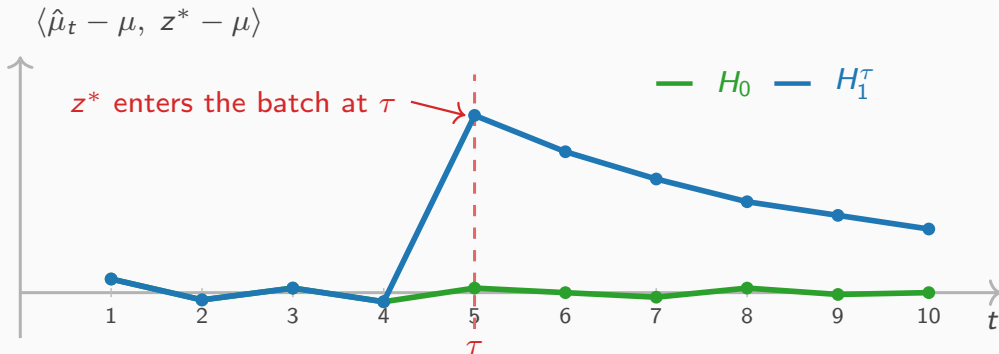
Adversary observes $(o_1, \dots, o_T)$ and decides:

H_0 : z^* absent. H_1^T : z^* inserted at step τ .

The empirical-mean mechanism

Setup. $X_{i,k} \sim \mathcal{N}(\mu, \Sigma) \Rightarrow \hat{\mu}_t = \frac{1}{nt} \sum_{i \leq t} \sum_k X_{i,k}$

Released sequence: $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_T$.



The isolation property

$$\begin{aligned}\text{LR} &= \frac{p(\hat{\mu}_{1:T} \mid H_1^\tau)}{p(\hat{\mu}_{1:T} \mid H_0)} = \prod_{t=1}^T \frac{p(\hat{\mu}_t \mid \hat{\mu}_{t-1}, H_1^\tau)}{p(\hat{\mu}_t \mid \hat{\mu}_{t-1}, H_0)} \\ &= \underbrace{\frac{p(\hat{\mu}_1 \mid H_1^\tau)}{p(\hat{\mu}_1 \mid H_0)}}_{=1} \cdots \frac{p(\hat{\mu}_\tau \mid \hat{\mu}_{\tau-1}, H_1^\tau)}{p(\hat{\mu}_\tau \mid \hat{\mu}_{\tau-1}, H_0)} \cdots \underbrace{\frac{p(\hat{\mu}_T \mid \hat{\mu}_{T-1}, H_1^\tau)}{p(\hat{\mu}_T \mid \hat{\mu}_{T-1}, H_0)}}_{=1}\end{aligned}$$

For $t \neq \tau$, given $\hat{\mu}_{t-1}$, the law of $\hat{\mu}_t$ is the same under both hypotheses, so the factor is 1.

$$\text{LR} = \frac{p(\hat{\mu}_\tau \mid \hat{\mu}_{\tau-1}, H_1^\tau)}{p(\hat{\mu}_\tau \mid \hat{\mu}_{\tau-1}, H_0)}$$

Only the τ -transition carries the canary's signal. Snapshots before and after τ are uninformative conditional on it.

SeMI*: the optimal test

$$\log \text{LR}_\tau = \frac{d}{2} \log c(n) - \frac{N_\tau^\top \Sigma^{-1} N_\tau}{2(n-1)} + c(n) N_\tau^\top \Sigma^{-1} (z^* - \mu) - \frac{m^*}{2(n-1)}$$

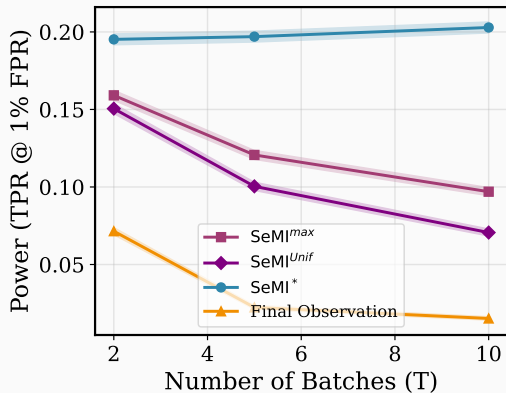
$$N_\tau = \bar{X}_\tau - \mu, \quad c(n) = \frac{n}{n-1}, \quad m^* = \|z^* - \mu\|_{\Sigma^{-1}}^2$$

The power of SeMI* depends only on the batch size n and the Mahalanobis distance m^* , not on τ or T .

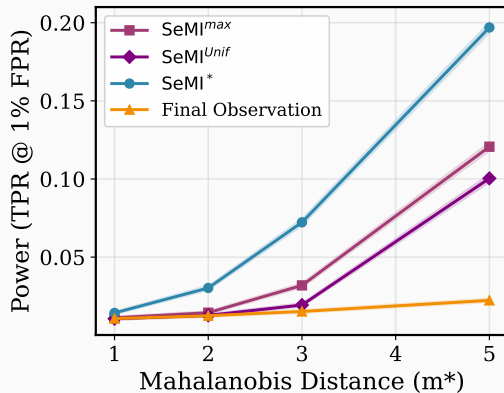
Contrast. A test that sees only the *final* model $\hat{\mu}_T$ pools nT samples; its per-sample signal $(z^* - \mu)/(nT)$ vanishes as T grows.

Power of SeMI* on the empirical mean

$T = 10$ batches of size $n = 10$ from $\mathcal{N}(0, 1)$, target at Mahalanobis distance 5, $\tau = 5$, $R = 50\,000$ runs.



(a) Power vs T : SeMI* stays near-constant; the final-observation baseline collapses.



(b) Power vs Mahalanobis distance m^* : SeMI* gains most when less common data is chosen as the target.

SeMI^{SGD}: lift to (DP-)SGD

Released sequence. The (DP-)SGD trajectory.

$$\theta_t = \theta_{t-1} - \eta_t g_t$$

Isolation transfers (Gaussian batch-gradient assumption).

The LR of the parameter trajectory depends on the data only through the gradient at step τ .

White-box SeMI^{SGD}: the same statistic, read in gradient space.

$$\bar{X}_\tau \longmapsto g_\tau = (\theta_{\tau-1} - \theta_\tau)/\eta_\tau$$

$$z^* \longmapsto g^* = \nabla_\theta \ell(\theta_{\tau-1}; z^*)$$

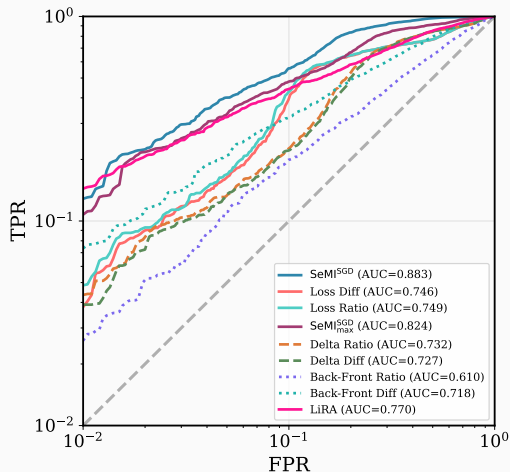
$$m^* = \|z^* - \mu\|_{\Sigma^{-1}}^2 \longmapsto m^* = \|g^* - \mu_g\|_{\Sigma_g^{-1}}^2$$

Black-box surrogates. $\Delta_\tau = \ell(\theta_{\tau-1}; z^*) - \ell(\theta_\tau; z^*)$, $R_\tau = \ell(\theta_{\tau-1}; z^*)/\ell(\theta_\tau; z^*)$

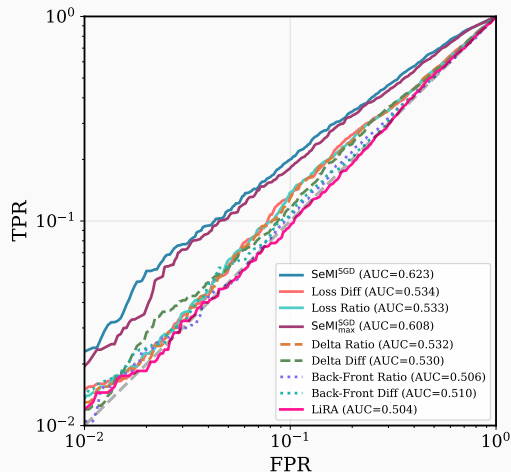
Maximised over the insertion time τ , they recover the Back-Front and Delta heuristics of Jagielski et al. (2023).

Auditing DP-SGD on Fashion-MNIST

Softmax regression, $L = 2500$ steps over T phases, total $\varepsilon = 8$, $\delta = 10^{-4}$, $R = 1900$ rounds. Baselines: Loss Diff, Loss Ratio, Delta, Back-Front (Jagielski 2023), LiRA (Carlini 2022).

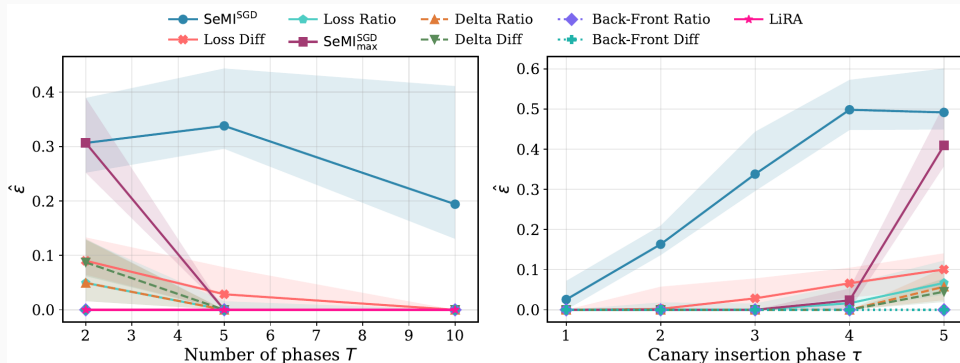


(a) $\varepsilon = \infty$: SeMI^{SGD} dominates.



(b) $\varepsilon = 8$: SeMI^{SGD} keeps the lead at low FPR.

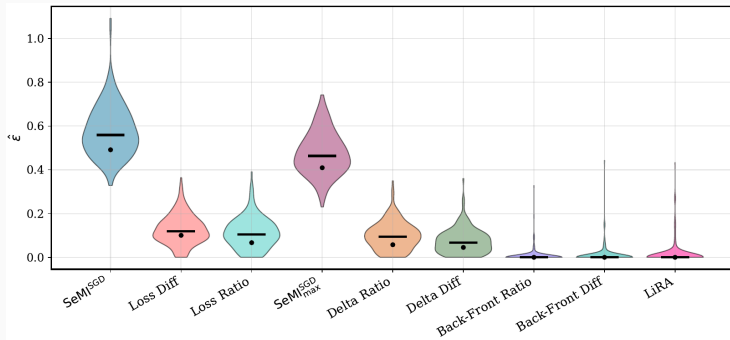
Tighter DP-SGD audits, across releases



For every number of releases T , SeMI^{SGD} certifies the tightest bound; the snapshot-independent LiRA stays flat near zero. The bound also depends strongly on the insertion time τ , a lever the auditor controls.

The tightest DP-SGD audit

Empirical privacy lower bound $\hat{\varepsilon}$ per attack (Fashion-MNIST, DP-SGD, $T = 5$, total $\varepsilon = 8$).



Among informed attacks (those that use τ), SeMI^{SGD} certifies the largest median $\hat{\varepsilon}$; SeMI^{max} follows even without knowing τ . The loss surrogates trail, and the snapshot-independent Back-Front and LiRA stay near zero.

Takeaways

Optimal sequential MI.

Neyman–Pearson optimal attack on the empirical mean, lifted to (DP-)SGD.

Two levers for the auditor.

Target selection (Azize et al., 2025; Carlini et al., 2022) and insertion time (this work).

Tighter empirical audits.

On DP-SGD with Fashion-MNIST, SeMI^{SGD} dominates every baseline.

Open.

Single-run audits (Steinke et al., 2023), extended sequentially; crafter-chosen insertion-time distributions.



Paper

Thank you!