



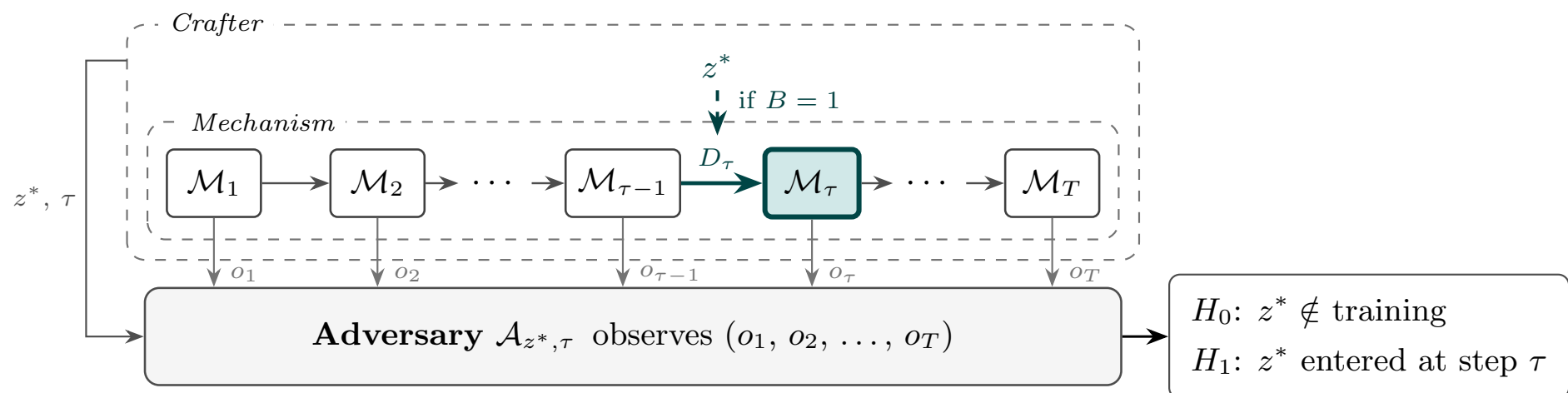
Contribution. Modern AI models are not static: they go through multiple updates in their lifecycle. We propose a *sequential auditing framework* that exploits the released sequence of snapshots $o_1, \dots, o_T$ together with a controlled canary insertion time τ . We derive SeMI*, the Neyman–Pearson optimal Sequential MI attack for the empirical mean mechanism with an *isolation property*, and obtain practical white-box (SeMI^{SGD}) and black-box variants for models trained with (DP-)SGD. **Auditing the sequence yields tighter privacy lower bounds than auditing only the final model.**

The Sequential MI Game

ML models memorize their training data, creating a privacy risk for the individuals in the dataset. (ϵ, δ) -Differential privacy bounds this leakage by design; privacy auditing measures it in practice. The standard tool is a *membership inference* (MI) attack that tries to decide whether a target record z^* was in training.

Modern ML models are increasingly released as a sequence of snapshots $o_1, \dots, o_T$: versioned foundation models, fine-tunes, continual updates, federated rounds. The auditor observes the whole sequence rather than only the final model.

Sequential MI game. A canary z^* is inserted into batch D_τ at step τ with probability 1/2; the adversary observes $(o_1, \dots, o_T)$ and tests H_0 (z^* absent) vs H_1 (z^* at τ).



The crafter controls the mechanism; the adversary sees all releases $o_1, \dots, o_T$.

From MI Attacks to Privacy Audits

(ϵ, δ) -Differential Privacy. A mechanism $\mathcal{M}$ is (ϵ, δ) -DP if replacing any single record changes its output distribution by at most a factor e^ϵ (with slack δ).

Formally, for all neighboring datasets D and D' and any event S we have

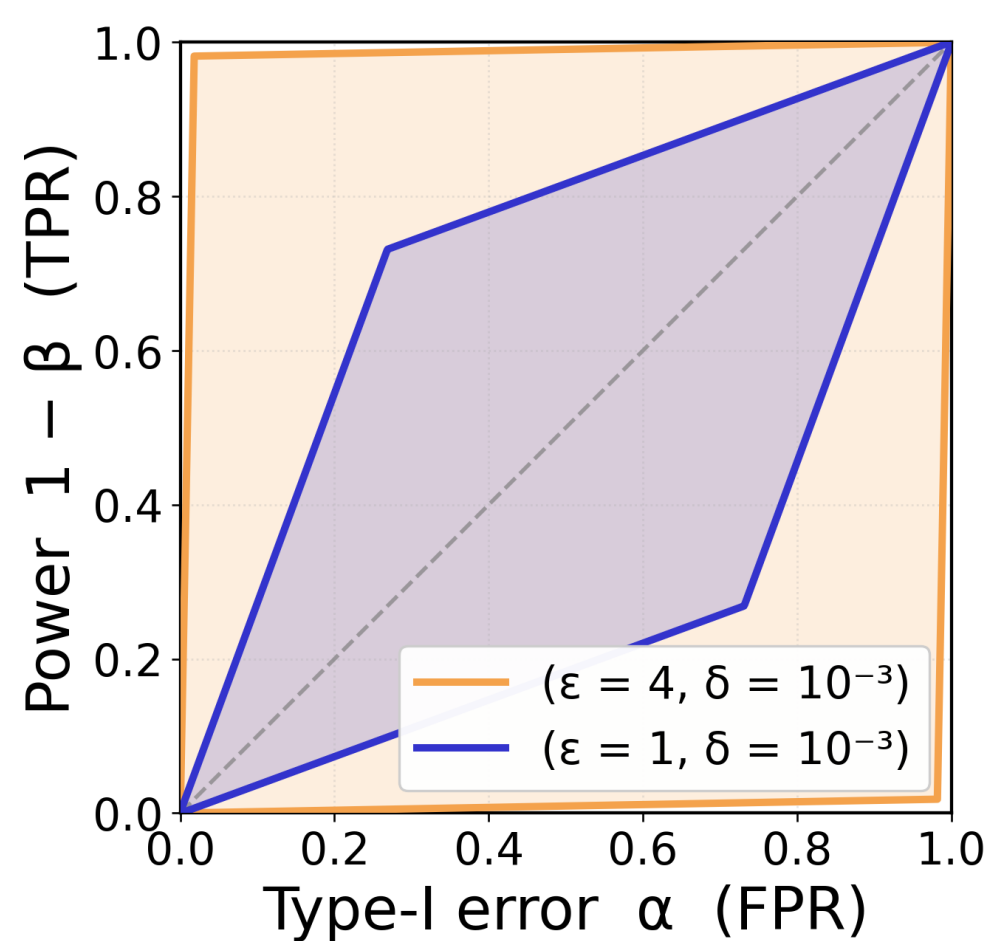
$$\mathbb{P}[\mathcal{M}(D) \in S] \leq e^\epsilon \mathbb{P}[\mathcal{M}(D') \in S] + \delta.$$

Audit principle. An MI test with Type-I/II errors α, β at insertion τ certifies a privacy lower bound on the mechanism:

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

Running R rounds and plugging Clopper–Pearson bounds yields an empirical lower bound $\hat{\epsilon}$.

Tighter attacks $\Rightarrow$ tighter audits. A test sitting outside the (α, β) trade-off region of an (ϵ, δ) -DP mechanism rules out that privacy level.



Any MI test's $(\text{FPR}, 1 - \beta)$ lies in the shaded region; an attack outside the region for a given ϵ certifies a tighter privacy lower bound.

Optimal Attack: SeMI*

The cumulative-mean mechanism releases $\hat{\mu}_t = \frac{1}{nt} \sum_{i \leq t} \sum_k X_{i,k}$ on Gaussian data $\mathcal{D} = \mathcal{N}(\mu, \Sigma)$. **Isolation property.** The sequence likelihood ratio depends only on the transition at τ :

$$\frac{p(\hat{\mu}_{1:T} | \mathbf{H}_1^*)}{p(\hat{\mu}_{1:T} | \mathbf{H}_0)} = \frac{p(\hat{\mu}_\tau | \hat{\mu}_{\tau-1}, \mathbf{H}_1^*)}{p(\hat{\mu}_\tau | \hat{\mu}_{\tau-1}, \mathbf{H}_0)}.$$

Consecutive snapshots reveal the batch mean $\bar{X}_\tau = \tau \hat{\mu}_\tau - (\tau - 1) \hat{\mu}_{\tau-1}$; with $\mathbf{N}_\tau = \bar{X}_\tau - \mu$, $c(n) = n/(n - 1)$, and $m^* = \|z^* - \mu\|_{\Sigma^{-1}}^2$, the log-LR is Mahalanobis-weighted:

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

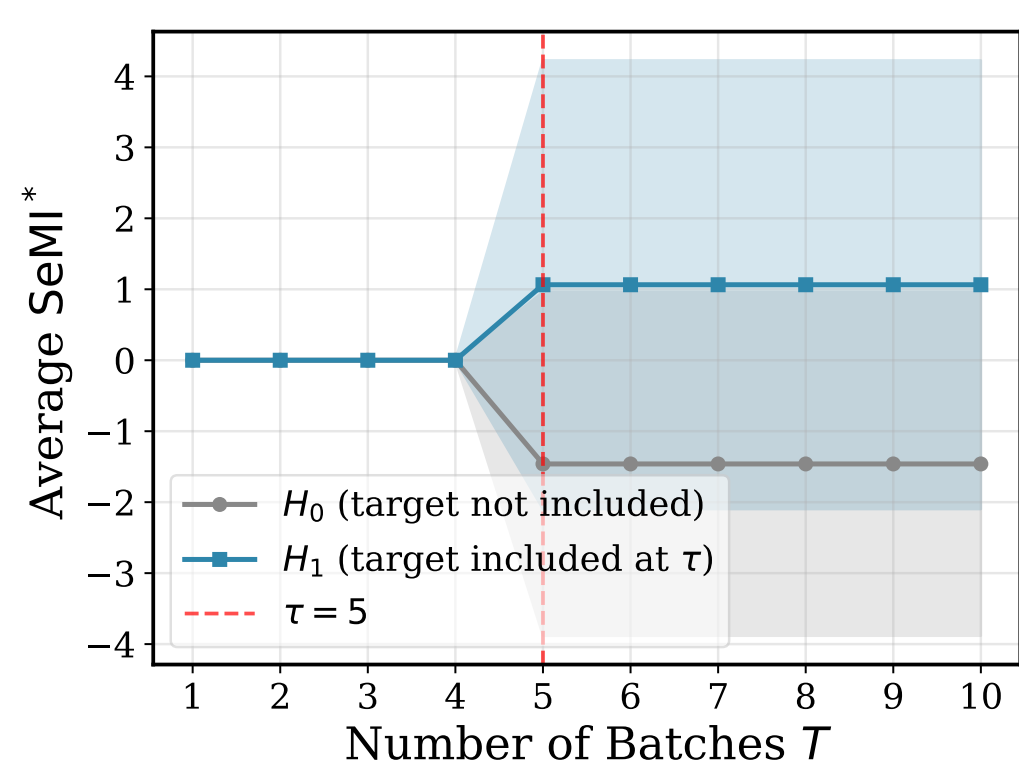
SeMI* is the Neyman–Pearson test that rejects $\mathbf{H}_0$ when $\log \text{LR}_\tau$ exceeds a threshold.

Power of SeMI* depends on (n, m^*) only, independent of τ and T .

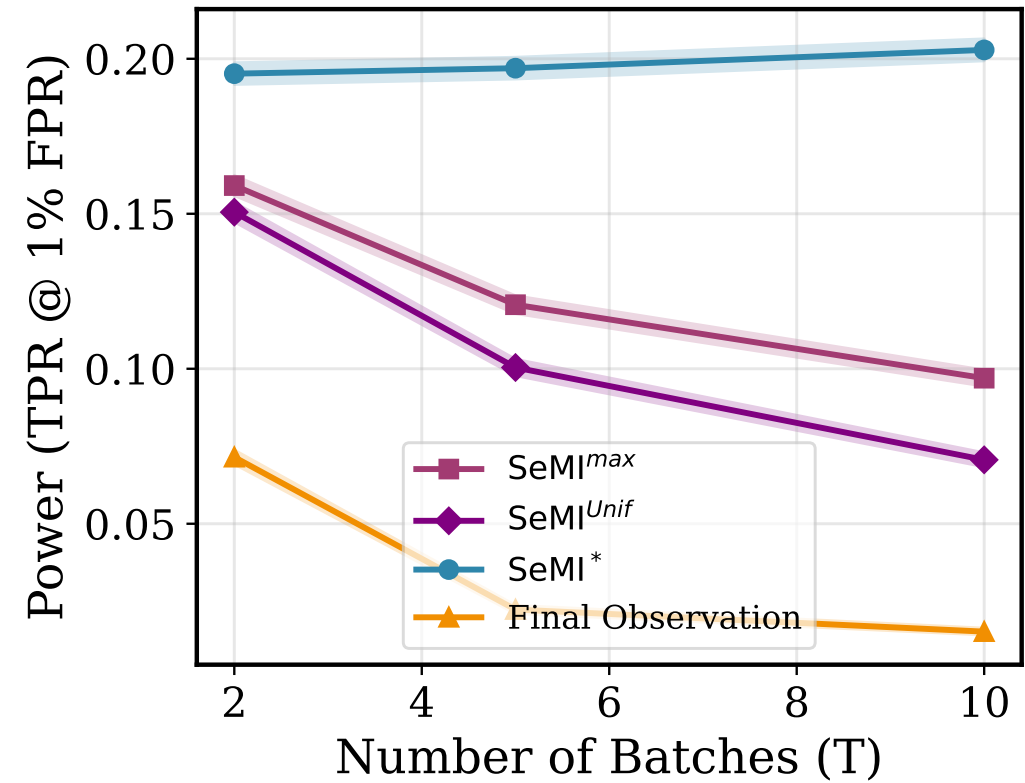
The **final-observation baseline** sees only $\hat{\mu}_T$ and pools nT samples; its per-sample signal $(z^* - \mu)/(nT)$ vanishes as T grows.

Sequential Advantage (Empirical Mean)

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



(a) The SeMI* statistic separates at τ and stays constant in T .



(b) SeMI* keeps near-constant power; the final-observation baseline collapses.

From Mean to DP-SGD: SeMI^{SGD}

SGD **with white-box access** releases $\theta_t = \theta_{t-1} - \eta_t g_t$. Under a Gaussian batch-gradient assumption the isolation property transfers to the parameter trajectory: with the parameter update $\mathbf{N} = \theta_\tau - \theta_{\tau-1} + \eta_\tau \mu_g$ playing the role of the batch mean and $\delta_g = \nabla \ell(\theta_{\tau-1}; z^*) - \mu_g$, the log-LR is again Mahalanobis-weighted:

$$\log \text{LR}_\tau = -\frac{d}{2} \log \frac{n_\tau - 1}{n_\tau} - \frac{n_\tau \mathbf{N}^\top \Sigma_g^{-1} \mathbf{N}}{2(n_\tau - 1) \eta_\tau^2} - \frac{n_\tau \mathbf{N}^\top \Sigma_g^{-1} \delta_g}{(n_\tau - 1) \eta_\tau} - \frac{m^*}{2(n_\tau - 1)}$$

DP-SGD clipping and Gaussian noise preserve this structure.

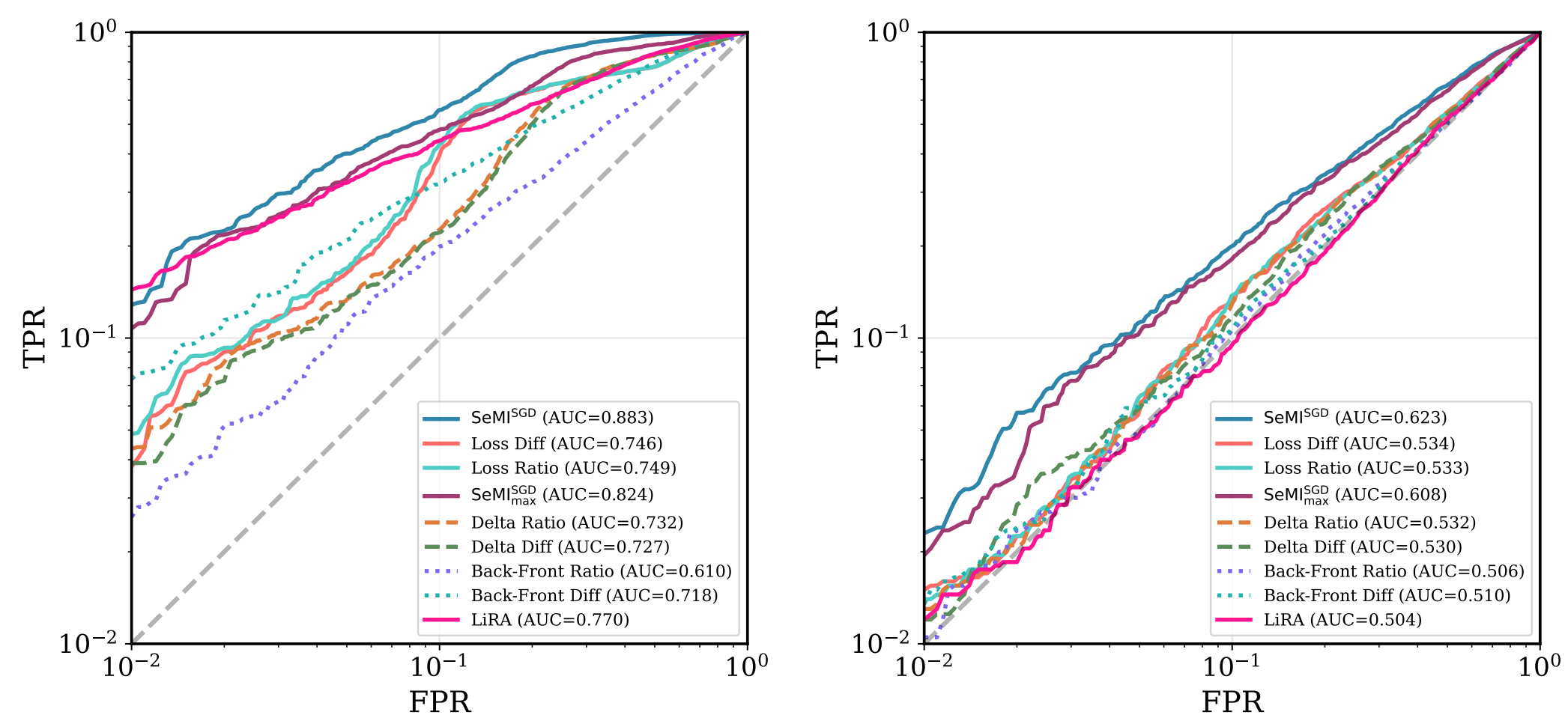
Black-box surrogates. With only the target loss, two per-transition scores contrast $(\theta_{\tau-1}, \theta_\tau)$:

$$\Delta_\tau = \ell(\theta_{\tau-1}; z^*) - \ell(\theta_\tau; z^*) \quad R_\tau = \frac{\ell(\theta_{\tau-1}; z^*)}{\ell(\theta_\tau; z^*)}$$

Maximizing Δ_τ over τ recovers the Delta heuristic; comparing θ_0 to θ_T recovers Back-Front [1].

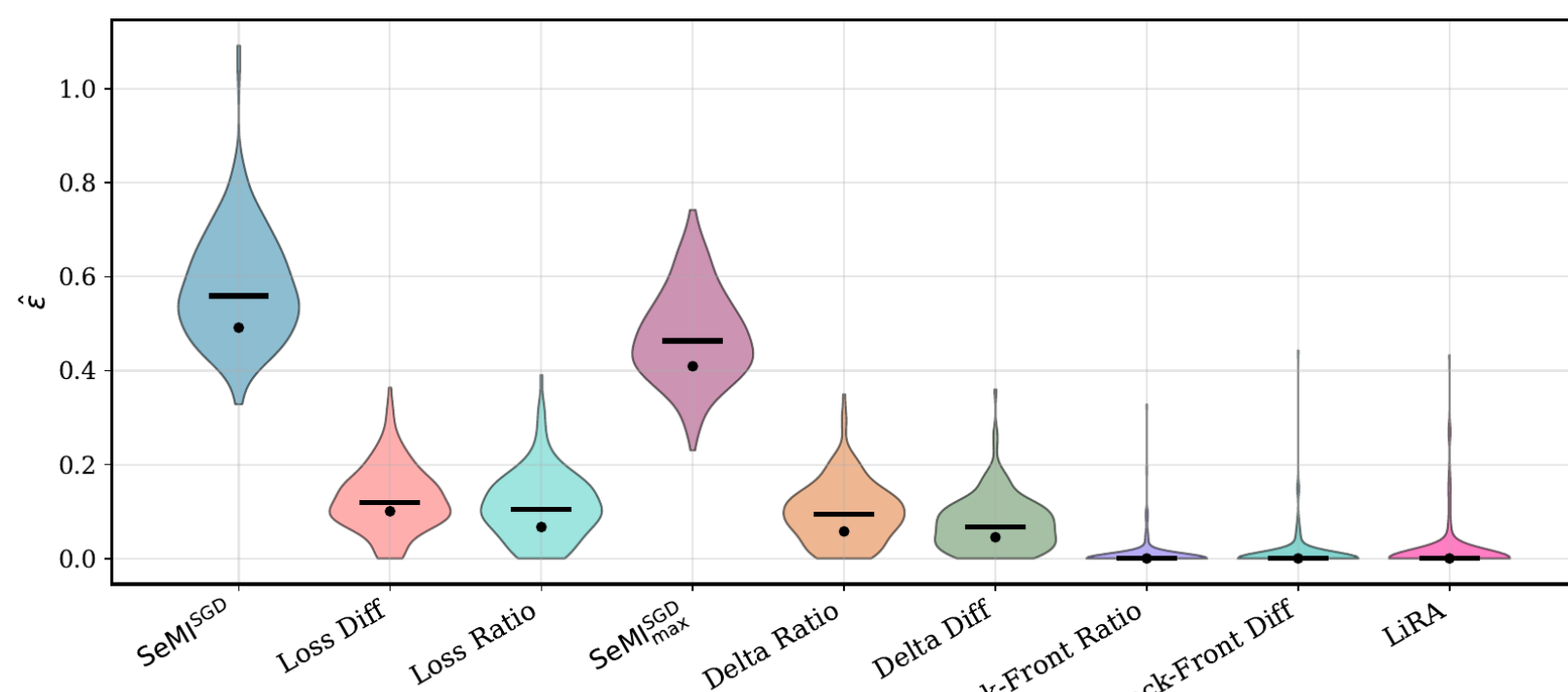
Auditing DP-SGD on Fashion-MNIST

Softmax regression, $L = 2500$ steps over T phases, total $\epsilon = 8$, $\delta = 10^{-4}$, $R = 1900$ rounds. Baselines: Loss Diff/Ratio, Delta, Back-Front [1], and LiRA [2] (snapshot-independent).



(a) White-box, $\tau = 4$, $\epsilon = \infty$.

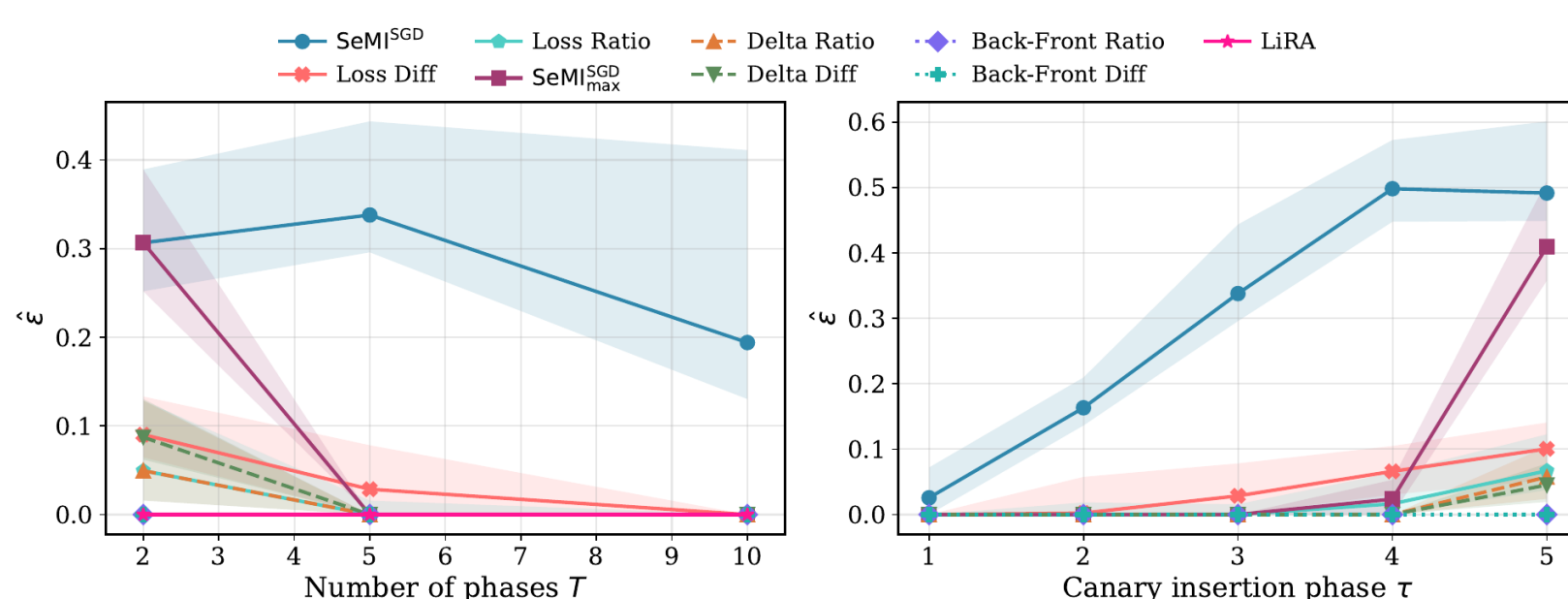
(b) White-box, $\tau = 4$, $\epsilon = 8$.



(c) Certified $\hat{\epsilon}$ across methods ($T = 5$, $\epsilon = 8$).

SeMI^{SGD} dominates the ROC, **keeps the lead at low FPR even at $\epsilon = 8$** , and yields the tightest empirical $\hat{\epsilon}$; knowing τ beats not knowing it.

Tighter Audits from the Sequence



Auditing only the final model under-reports the leakage. At fixed total budget, $\hat{\epsilon}$ grows with the number of snapshots T for SeMI^{SGD} while the final-model LiRA stays flat (left). **Insertion time is a lever:** later τ gives a larger signal (right).

Variants of SeMI

The isolation argument extends beyond the known- τ setting. Each variant is the Neyman–Pearson optimal test for its prior on τ .

Variant	Knowledge of τ	Test statistic	Optimality
SeMI*	known τ	$\log \text{LR}_\tau$	NP against $\mathbf{H}_1^*$
SeMI ^{Unif}	$\tau \sim \mathcal{U}([T])$	$\log \frac{1}{T} \sum_t \text{LR}_t$	NP against $\mathbf{H}_1$
SeMI ^{max}	unknown	$\max_t \log \text{LR}_t$	GLR over $\bigcup_t \mathbf{H}_1^t$
SeMI ^{SGD}	known τ , (DP-)SGD	gradient log-LR	NP under Gaussian-gradient approx.