Local FDR Membership Inference Attacks: Multiple Testing and the Role of Ridge Regularization

Jinyoung Hong **Bonwoo Lee** **Jeongyoun Ahn***
Korea Advanced Institute of Science & Technology
Daejeon, 34141 South Korea
urared@kaist.ac.kr, righthim@kaist.ac.kr, jyahn@kaist.ac.kr

Abstract

Membership inference attacks (MIAs) are commonly formulated as single-sample hypothesis tests with false-positive rate control. However, realistic adversaries often test many candidate samples and aggregate the declared members, making membership inference a multiple-testing problem. In this regime, per-sample calibration can produce unreliable sets of inferred members. We develop a false discovery rate (FDR)-controlled framework for membership inference based on the classical local FDR formulation. We instantiate the framework for high-dimensional ridge regression. Under a conditional sufficiency condition, we establish an FDR guarantee of the proposed method and derive an asymptotic characterization of its detection power in high-dimensional regimes under an isotropic Gaussian design. Our analysis shows that, under FDR control, stronger ridge regularization reduces membership inference risk. Experiments on synthetic and real-world datasets validate the theoretical findings and demonstrate that the proposed methods provide more reliable membership discoveries than existing attacks.

1 Introduction

Membership inference attacks (MIAs) seek to determine whether a given data point was used to train a machine learning model. The stakes are high in data-sensitive settings: a model trained on patients with a particular disease can, under a successful MIA, inadvertently expose an individual's health status [24]. As large-scale models are deployed in increasingly high-stakes environments, understanding and mitigating these vulnerabilities has become central to trustworthy machine learning. Early approaches relied on shadow-model heuristics, but the field has since moved toward more principled, high-power methods [7, 32]. A particularly important strand of this work formalizes membership inference as a statistical hypothesis testing problem, enabling explicit control of the false positive rate (FPR) [28, 32].

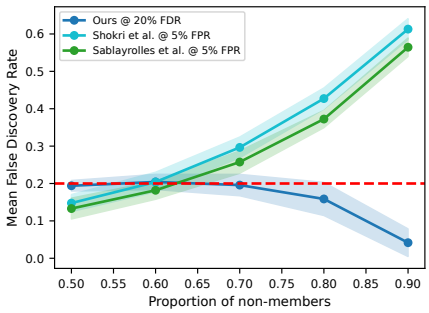


Figure 1: Mean false discovery rate as a function of the proportion of non-members in the candidate pool. FPR-calibrated baselines deteriorate rapidly as non-members dominate. See Appendix E for details.

Most existing formulations, however, treat MIA as a decision problem for a *single* target sample. In practice, an adversary is more likely to test many candidates at once, aggregating positive findings to reconstruct as much of the training set as possible. For example, a privacy auditor or adversary may have access to a pool of candidate records drawn from a public

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

dataset or a data provider's database. They can query the released model for each record but do not know which records were used for training. Their goal may be to identify a reliable *subset* of records likely to have been used in training, for data provenance investigations or sample-specific privacy-risk assessments, subject to a prespecified error-rate constraint. This is inherently a *multiple-testing* problem, and per-sample false positive control is not sufficient. When true members are rare relative to the candidate pool, even a small per-sample FPR can produce many false discoveries: the false discovery rate (FDR)—the expected proportion of false positives among all flagged samples—grows as the proportion of non-members increases. The adversary is then left with a sizable but unreliable set of discoveries. Figure 1 makes this concrete: at 50% non-members, FPR-calibrated baselines [24, 22] and our method all achieve comparable FDR, but as the proportion of non-members grows, the baselines' realized FDR rises sharply, exceeding 60% at 90% non-members, while our method remains controlled at the nominal 20% level.

The inadequacy of per-sample FPR control in large-scale attacks motivates a multiple-testing approach to membership inference. The multiple testing literature offers principled tools for controlling errors at the level of an entire collection of hypotheses [12, 2], and FDR control provides exactly what per-sample FPR control lacks: a guarantee on the reliability of the full set of declared members, not merely of each individual decision. FDR control thus offers a principled way to balance reliability and power in large-scale membership inference attacks, limiting false discoveries while preserving the ability to surface true members.

We propose an FDR-controlled MIA framework grounded in the classical local FDR formulation [13], which assigns each candidate a posterior probability of non-membership and naturally accommodates heterogeneity across samples. We instantiate this framework for high-dimensional ridge regression—a setting that is both theoretically tractable and directly relevant to the regularization–MIA tradeoff. In high dimensions, overparameterized models are known to memorize training data, increasing vulnerability to MIAs, and ridge regularization is a natural defense candidate. The leave-one-out structure of ridge residuals yields closed-form conditional distributions under membership and non-membership, enabling both a rigorous FDR guarantee and an explicit asymptotic characterization of detection power as $p/n \to \gamma \in (0, \infty)$ under an isotropic Gaussian design. Our analysis reveals that, under FDR control, stronger ridge regularization reduces membership inference risk—a finding validated on both synthetic and real-world data, where the proposed methods achieve superior FDR calibration and FDR–TPR trade-offs relative to existing attacks.

Contributions.

- We develop an FDR-controlled MIA framework based on the local FDR formulation, instantiated via two algorithms—a dataset-level and a sample-level procedure—that differ in their conservativeness and are each equipped with an FDR guarantee under a conditional sufficiency condition.
- For high-dimensional ridge regression, we derive the minimal attainable FDR level as a closed-form function of the ridge leverage scores, providing a quantitative measure of a model's exposure to MIAs.
- We establish an explicit formula for the asymptotic sample-level true positive rate as $p/n \to \gamma \in (0, \infty)$ under an isotropic Gaussian design, expressed in terms of the Stieltjes transform of the Marchenko–Pastur distribution. This characterization reveals a sharp threshold behavior: the TPR collapses to zero once regularization exceeds a critical level.
- We validate the proposed methods against FPR-calibrated and FDR-calibrated baselines on synthetic and real-world data, demonstrating superior FDR calibration and FDR–TPR trade-offs across varying target FDR levels and non-member proportions.

Additional experiments and the proofs omitted from the main text are provided in the appendix.

1.1 Related work

Membership inference attacks. Since the seminal work of Shokri et al. [24], a large body of work has developed attacks based on shadow models, model outputs, and losses [23, 22, 28]. A more recent line of work formalizes MIA as a hypothesis testing problem: Carlini et al. [7] propose the likelihood-ratio attack, designed and evaluated for the low-FPR regime; Ye et al. [28] propose

a unified hypothesis testing framework that recovers prior attacks and gains power from reference models; and Zarifzadeh et al. [32], Bertran et al. [4] develop low-cost, high-power likelihood-ratio tests. All of these methods operate in the single-sample regime and provide FPR guarantees, leaving the multiple-testing setting unaddressed. Zhao et al. [33] is the closest prior work to ours, proposing an FDR-controlled MIA via conformity scores. However, their method lacks a rigorous FDR guarantee and an asymptotic TPR characterization, and tends to be highly conservative in practice, as our experiments confirm.

Overfitting, regularization, and membership leakage. Yeom et al. [30] establish a formal connection between privacy leakage and overfitting, while Dionysiou [11] cautions that MIA success depends heavily on the threat model and evaluation protocol. On the defense side, Kaya et al. [16] show that common regularization techniques do not uniformly reduce MIA vulnerability and can even strengthen certain attacks. Most relevant to our theoretical setting is Tan et al. [27], who analyze membership advantage in linear and ridge regression under a Gaussian design. Unlike our work, their analysis targets the single-sample regime and does not consider FDR control.

Multiple testing. Multiple testing concerns simultaneous inference across a family of null hypotheses, where naive per-test calibration fails to control group-level error. Classical criteria include the family-wise error rate (FWER), which controls the probability of any false rejection, and the FDR [2], which controls the expected proportion of false rejections and is less conservative than FWER in large-scale settings. The Benjamini–Hochberg (BH) procedure [2] provides FDR control under independence and positive dependence, and was extended to arbitrary dependence by Benjamini and Yekutieli [3]. Storey [25] introduced the q-value and the positive FDR (pFDR), offering a Bayesian perspective on FDR that connects naturally to local FDR methods and requires estimating the null proportion π_0. A parallel line of work develops oracle and adaptive compound decision procedures for local FDR control [26, 5]. Our work builds on the local FDR framework of Efron [13], which defines a posterior null probability for each hypothesis given its test statistic, and in particular on covariate-adaptive extensions that incorporate auxiliary information to improve power [31, 15, 18]. We adapt this framework to the membership inference setting, where the ridge leverage scores and bias terms serve as natural covariates.

2 Setup and performance criteria

2.1 Multiple-target membership inference

Let $\mathcal{D}$ be a data distribution over $\mathbb{R}^{p+1}$, and let $S := \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be N i.i.d. target samples drawn from $\mathcal{D}$. For a given null probability $\pi_0 \in (0, 1)$, membership indicators $M_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1 - \pi_0)$ are drawn independently for $i = 1, \ldots, N$, specifying whether each target sample is included in training. The data owner releases the model f trained on the selected samples; the adversary observes S and the prediction vector $\widehat{Y} := (f(x_1), \ldots, f(x_N))$, but not the membership indicators. Unlike the conventional single-target formulation, the goal here is to infer the full membership vector $M = (M_1, \ldots, M_N)$ simultaneously, via an algorithm $\mathcal{A} : (\mathbb{R}^{p+1})^N \times \mathbb{R}^N \to \{0, 1\}^N$ that outputs $\widehat{M} := \mathcal{A}(S, \widehat{Y})$. This simultaneous inference problem is naturally cast as multiple hypothesis testing: for each target sample i, the adversary tests

$$\mathsf{H}_{i0} : M_i = 0 \quad \text{vs.} \quad \mathsf{H}_{i1} : M_i = 1,$$

where the null hypothesis represents non-membership.

We instantiate this formulation for ridge regression. Suppose the target samples are generated from the linear model

$$y_i = x_i^\top \beta + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $x_i \in \mathbb{R}^p$ is drawn from a covariate distribution $\mathcal{D}_X$. Let $I_{\text{tr}} := \{i \in \{1, \ldots, N\} : M_i = 1\}$ be the training index set, with X_{tr} and Y_{tr} the corresponding design matrix and response vector. For ridge regularization parameter $\lambda > 0$, the released model is

$$\widehat{\beta} = (X_{\text{tr}}^\top X_{\text{tr}} + \lambda I_p)^{-1} X_{\text{tr}}^\top Y_{\text{tr}}.$$

The adversary then observes the predictor matrix $X := [x_1, \ldots, x_N]^\top$ and prediction vector $\widehat{Y} := X\widehat{\beta}$, and outputs the inferred membership vector $\widehat{M} = \mathcal{A}(X, Y, \widehat{Y}) \in \{0, 1\}^N$.

2.2 Performance criteria

We evaluate membership inference attacks using two criteria: the false discovery rate, which measures the reliability of the declared member set, and the true positive rate, which measures its completeness.

Definition 1 (False discovery rate)**.** *Let R be the total number of declared members, and V be the number of non-members falsely declared as members. The FDR is defined as*

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right].$$

The ratio $V/\max\{R, 1\}$ is the proportion of declared members that are in fact non-members; controlling the FDR therefore provides a reliability guarantee for the discovered set as a whole.

Definition 2 (True positive rate)**.** *For a given membership vector M, the sample-level true positive rate of an attack $\mathcal{A}$ for a member sample i with $M_i = 1$ is*

$$\mathrm{TPR}(i) = \mathbb{P}(\widehat{M}_i = 1 \mid M),$$

and the true positive rate of $\mathcal{A}$ is

$$\mathrm{TPR} = \frac{1}{|I_{\mathrm{tr}}|} \sum_{i \in I_{\mathrm{tr}}} \mathrm{TPR}(i).$$

The sample-level TPR measures the probability that a given member is correctly identified, and the overall TPR is the average of these probabilities over all true members.

3 FDR-controlled membership inference via local FDR

We introduce two FDR-controlled MIA algorithms grounded in the local FDR framework and establish their FDR guarantees. We instantiate the framework for ridge regression, which admits an exact leave-one-out characterization of residuals that makes both FDR control and asymptotic analysis tractable. The framework extends naturally beyond linear models, as discussed in Section 3.3.

3.1 Residual-based local FDR attack

The covariate-modulated local FDR [31] uses the posterior probability of the null hypothesis, conditional on the observed statistic and auxiliary covariates, as the decision statistic. This makes it a natural fit for membership inference: the posterior probability of non-membership can be computed directly from the conditional residual distributions under H_{i0} and H_{i1}, and the resulting statistic admits a target FDR level as a tuning parameter.

We derive these conditional distributions via the leave-one-out structure of ridge residuals. For $i = 1, \ldots, N$, denote the augmented index set $I_{\mathrm{tr},i} = I_{\mathrm{tr}} \cup \{i\}$, the corresponding design matrix $X_{\mathrm{tr},i} = [x_j^\top]_{j \in I_{\mathrm{tr},i}}$, and the associated ridge hat matrix

$$H_{\lambda,i} := X_{\mathrm{tr},i}(X_{\mathrm{tr},i}^\top X_{\mathrm{tr},i} + \lambda I_p)^{-1} X_{\mathrm{tr},i}^\top.$$

Let h_i and d_i denote the diagonal entries corresponding to sample i of $H_{\lambda,i}$ and $(I - H_{\lambda,i})^2$, respectively, and let μ_i denote the corresponding component of $(I - H_{\lambda,i})X_{\mathrm{tr},i}\beta$. Here, h_i is the ridge leverage score for sample i and μ_i is the residual bias induced by fitting the ridge estimator on $X_{\mathrm{tr},i}$. We collect these into the nuisance summary $C_i := (h_i, d_i, \mu_i)$, and recall that $X := [x_1, \ldots, x_N]^\top$ denotes the full predictor matrix of the target samples.

Proposition 3 (Distribution of the residual [10])**.** *Conditional on X and C_i, the residual $r_i := y_i - x_i^\top \widehat{\beta}$ satisfies*

$$\begin{aligned}
\mathsf{H}_{i0} : M_i = 0 \quad &\Longrightarrow \quad r_i \mid X, C_i \;\sim\; \mathcal{N}\!\left(\frac{\mu_i}{1-h_i},\, \frac{d_i\sigma^2}{(1-h_i)^2}\right), \\
\mathsf{H}_{i1} : M_i = 1 \quad &\Longrightarrow \quad r_i \mid X, C_i \;\sim\; \mathcal{N}\big(\mu_i,\, d_i\sigma^2\big).
\end{aligned}$$

Proposition 3 characterizes the conditional distribution of r_i under both hypotheses, yielding closed-form expressions for the densities $f_{i0}(r_i \mid X, C_i)$ and $f_{i1}(r_i \mid X, C_i)$ under H_{i0} and H_{i1}, respectively. These densities determine the posterior probability of non-membership for each candidate sample, which we take as the local FDR statistic.

Definition 4 (Covariate-modulated local FDR)**.** *The covariate-modulated local FDR for sample i is*

$$\mathrm{lfdr}(r_i, X, C_i) = \mathbb{P}(M_i = 0 \mid r_i, X, C_i) = \frac{\pi_0 f_{i0}(r_i \mid X, C_i)}{\pi_0 f_{i0}(r_i \mid X, C_i) + (1-\pi_0) f_{i1}(r_i \mid X, C_i)}.$$

Samples with small $\mathrm{lfdr}(r_i, X, C_i)$ provide strong posterior evidence for membership, and are natural candidates for rejection. Algorithm 1 exploits this by selecting the largest rejection set — ordered by increasing local FDR — whose average local FDR does not exceed the target level q, directly translating Definition 4 into an FDR-controlled decision rule.

Algorithm 1 Dataset-level local FDR MIA

Input: Residuals r, covariate matrix X, nuisance summaries $\{C_i\}_{i=1}^N$, noise variance σ^2, null probability π_0, target FDR level $q \in (0,1)$

Output: Predicted membership vector $\widehat{M} \in \{0,1\}^N$

1: Compute $\ell_i \leftarrow \mathrm{lfdr}(r_i, X, C_i)$ for $i = 1, \ldots, N$
2: Sort: $\ell_{(1)} \le \ell_{(2)} \le \cdots \le \ell_{(N)}$
3: Compute $A_j \leftarrow \frac{1}{j}\sum_{s=1}^{j} \ell_{(s)}$ for $j = 1, \ldots, N$
4: Set $k \leftarrow \max\big(\{j : A_j \le q\} \cup \{0\}\big)$
5: **return** $\widehat{M}$ with $\widehat{M}_{(i)} = 1$ for $i = 1, \ldots, k$ and $\widehat{M}_{(i)} = 0$ otherwise

The term *dataset-level* does not imply that Algorithm 1 makes a single decision about whether the candidate dataset and the training dataset have a nonempty intersection. Rather, it indicates that the threshold defining the rejection region depends on the candidate dataset. Algorithm 1 admits an FDR guarantee under the condition stated in Section 3.2, but relies on accurately estimated nuisance quantities C_i and σ^2. When these estimates are noisy, a more conservative rule is preferable. Algorithm 2 imposes the additional requirement that each rejected sample individually satisfies $\ell_i \le q$, tightening the dataset-level criterion to a per-sample one.

Algorithm 2 Sample-level local FDR MIA

Input: Residuals r, covariate matrix X, nuisance summaries $\{C_i\}_{i=1}^N$, noise variance σ^2, null probability π_0, target FDR level $q \in (0,1)$

Output: Predicted membership vector $\widehat{M} \in \{0,1\}^N$

1: Compute $\ell_i \leftarrow \mathrm{lfdr}(r_i, X, C_i)$ for $i = 1, \ldots, N$
2: **return** $\widehat{M}$ with $\widehat{M}_i = 1$ if $\ell_i \le q$ and $\widehat{M}_i = 0$ otherwise

Algorithm 2 rejects sample i only if $\ell_i \le q$, whereas Algorithm 1 selects the largest prefix of the sorted local FDR sequence with average at most q, and may therefore reject samples with $\ell_i > q$. The rejection set of Algorithm 2 is thus always contained in that of Algorithm 1. From the perspective of compound decision theory [26], Algorithm 1 makes rejection decisions jointly by pooling evidence across candidates, and such joint rules dominate marginal ones in terms of power. Algorithm 2 sacrifices some power in exchange for robustness to estimation error in C_i and σ^2.

3.2 False discovery rate guarantee

Controlling FDR requires that the average local FDR over the rejection set faithfully estimates the true FDR. Because the residual vector r is dependent across samples, this is not automatic: the local FDR statistics $\{\ell_i\}_{i=1}^N$ inherit this dependence, and ℓ_i may not equal the exact marginal posterior probability of non-membership for sample i when the joint data $Z_1, \ldots, Z_N$ carry additional information beyond $Z_i := (r_i, C_i)$. We therefore impose the following conditional sufficiency condition, which ensures each ℓ_i is the exact marginal posterior.

Assumption 5. $\mathbb{P}(M_i = 0 \mid X, Z_1, \ldots, Z_N) = \mathbb{P}(M_i = 0 \mid X, Z_i)$.

Assumption 5 states that, conditional on X, the summary Z_i contains all information relevant to inferring the non-membership of the sample i, that is, the joint data $Z_1, \ldots, Z_N$ provide no additional information beyond Z_i itself. This is a conditional independence condition on the membership posteriors, analogous to but distinct from the full independence of test statistics assumed in classical local FDR procedures [13, 26]. It should be interpreted as a sufficient condition for exact FDR control rather than a necessary condition for the empirical validity of the procedure. The assumption holds in idealized orthogonal designs and is expected to be a useful approximation in high-dimensional nearly orthogonal regimes [6, 1], but need not hold exactly in correlated finite-sample settings. For Algorithm 2, the assumption can be relaxed to achieve asymptotic FDR control under a block orthogonality condition and a nonvanishing proportion of discoveries. Details are provided in Appendix B.

We now state the FDR guarantee for both proposed algorithms.

Proposition 6. *Under Assumption 5, Algorithms 1 and 2 control the false discovery rate at level q, that is,* $\mathrm{FDR} \leq q$.

The proof hinges on the observation that, under Assumption 5, the conditional expected number of false discoveries satisfies $\mathbb{E}[V \mid W] = \sum_{i:\, \widehat{M}_i = 1} \ell_i$, where $W = (X, Z_1, \ldots, Z_N)$. Since both algorithms ensure the average local FDR over the rejection set does not exceed q, the FDR bound follows immediately. The full proof is provided in the Appendix A.

3.3 Extensions to other model classes

The local FDR framework of Definition 4 requires only the conditional densities f_{i0} and f_{i1} under non-membership and membership, respectively, and is therefore not specific to ridge regression. It applies to any model class for which these two distributions can be characterized or accurately approximated. Immediate extensions include other linear smoothers, such as smoothing splines, where analogous leave-one-out identities hold.

For nonlinear models, f_{i0} and f_{i1} can be approximated via leave-one-out analysis, jackknife procedures, or influence functions [17]. Exact LOO formulas are available for kernel ridge regression and Gaussian processes, making these immediate extensions of the present framework. The formulation of local FDR MIA for kernel ridge regression and the corresponding experimental results are provided in Appendix D. Approximate LOO characterizations exist for sparse regression and for generalized linear models [8, 19], enabling approximate local FDR attacks in these settings. Neural networks are a particularly relevant direction: Ye et al. [29] show that the predictive distributions of a network trained with and without a target sample can be approximated using neural network Gaussian processes, providing natural plug-in estimates of f_{i0} and f_{i1} in Definition 4. Extending the FDR guarantee of Proposition 6 to these settings would require an analog of Assumption 5, which remains an important direction for future work.

4 Minimal FDR level and asymptotic power

4.1 Minimal attainable FDR level

A natural question from the model builder's perspective is: how small can an adversary's FDR target be while still making any discoveries? This minimal attainable FDR level characterizes the model's inherent vulnerability to membership inference — the lower it is, the more exposed the model. For Algorithm 2, whether any rejection is possible depends on whether the local FDR statistic ℓ_i can fall below the target level q for at least one sample. For fixed C_i, this requires q to exceed a sample-specific threshold, which we now characterize. Define

$$q(h, d, \mu, \sigma) := \left[1 + \left(\frac{1}{\pi_0} - 1\right)\left(\frac{1}{1-h}\right)\exp\left(\frac{\mu^2}{2\sigma^2} \cdot \frac{h}{d(2-h)}\right)\right]^{-1}.$$

Proposition 7. *Under Assumption 5, the minimal target level for which Algorithm 2 has a nonempty rejection region for sample i is $q_i^\star = q(h_i, d_i, \mu_i, \sigma)$.*

From a model builder's perspective, $q_i^\star$ provides a quantitative measure of the model's exposure to MIAs: any adversary targeting a FDR level below $q_i^\star$ will make no discoveries. Interventions that

reduce the leverage scores h_i — such as adding noise to the training data or increasing λ — raise $q_i^\star$, making successful inference harder.

Under pairwise orthogonality of the rows of X, Proposition 7 simplifies to a closed-form expression in terms of the regularization strength $\nu_i = \lambda/\|x_i\|_2^2$ and the signal-to-noise ratio $x_i^\top \beta/\sigma$.

Corollary 8. *If the rows of X are pairwise orthogonal, then*

$$q_i^\star = q\left(\frac{1}{1+\nu_i}, \frac{\nu_i^2}{(1+\nu_i)^2}, \frac{\nu_i x_i^\top \beta}{1+\nu_i}, \sigma\right) = \left[1 + \left(\frac{1}{\pi_0} - 1\right)\frac{1+\nu_i}{\nu_i}\exp\left(\frac{(x_i^\top \beta)^2}{2\sigma^2(1+2\nu_i)}\right)\right]^{-1},$$

where $\nu_i = \lambda/\|x_i\|_2^2$ measures the relative strength of regularization for sample i.

Since $q_i^\star$ is nondecreasing in ν_i and converges to π_0 as $\nu_i \to \infty$, increasing λ progressively shields the model from membership inference. In the limit of strong regularization, no attack can achieve FDR below π_0, the prior non-membership probability.

4.2 Asymptotic detection power

Proposition 7 characterizes the threshold below which no attack can succeed, but does not quantify how many true members can be detected once that threshold is exceeded. To address this, we derive an explicit formula for the asymptotic sample-level TPR of Algorithm 2 in the high-dimensional regime $p/n \to \gamma \in (0, \infty)$. This regime is relevant to modern overparameterized models, where the number of parameters is comparable to or exceeds the training set size. The resulting characterization reveals how the aspect ratio γ, the regularization strength λ, and the signal $s = x_i^\top \beta$ jointly determine detection power, and provides a theoretical basis for the empirical finding that stronger regularization reduces MIA risk.

Set $\alpha := \lambda/n$ and let $m_\alpha = m(-\alpha)$, $m'_\alpha = m'(-\alpha)$, where $m(z)$ denotes the Stieltjes transform of the Marchenko–Pastur distribution with aspect ratio γ. Define the limiting minimal FDR level

$$q_\bullet^\star = q\left(\frac{\gamma m_\alpha}{1+\gamma m_\alpha}, \frac{1+\gamma m_\alpha - \alpha\gamma m'_\alpha}{(1+\gamma m_\alpha)^2}, \frac{\alpha m_\alpha s}{1+\gamma m_\alpha}, \sigma\right),$$

which is the high-dimensional limit of $q_i^\star$ in Proposition 7. Assume the predictors are i.i.d. from $\mathcal{N}(0, I_p)$, and for sample i let $x_{i,p}$ and β_p denote the predictor and true parameter at dimension p. Suppose $\|x_{i,p}\|_2^2/p \to 1$ and $x_{i,p}^\top \beta_p \to s$ as $p/n \to \gamma$.

Proposition 9. *Under the above conditions, the conditional sample-level true positive rate of Algorithm 2 converges almost surely to $(\Phi(z_+) - \Phi(z_-))\,\mathbb{I}(q \geq q_\bullet^\star)$, where $z_- \leq z_+$ are the two roots of the limiting quadratic derived in Appendix A.3.*

In the Gaussian design setting, $m(z)$ admits a closed-form expression, so Proposition 9 yields a fully explicit formula for the limiting TPR. Since the rejection set of Algorithm 2 is contained in that of Algorithm 1, the limiting TPR of Algorithm 2 serves as a lower bound on that of Algorithm 1.

The isotropic Gaussian design is a standard assumption in high-dimensional statistical theory [14], enabling the use of random matrix theory tools such as the Marchenko–Pastur law. While it is an idealization relative to real-world datasets, the experimental results in Section 5 on a highly correlated real-world dataset suggest that the qualitative findings, in particular the effect of ridge regularization on detection power, extend well beyond this setting.

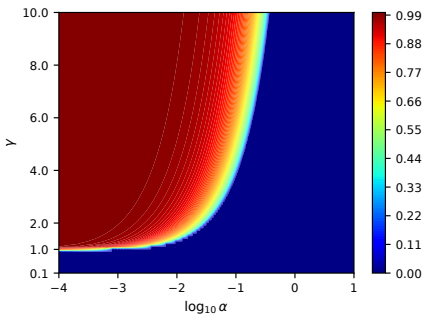
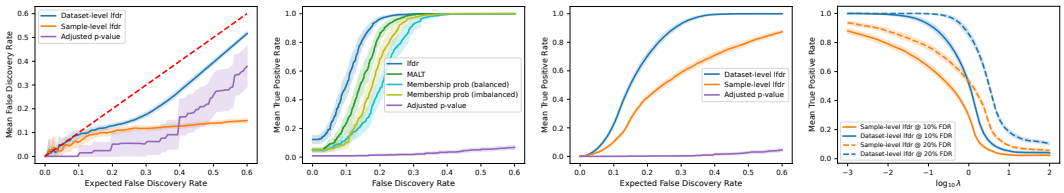


Figure 2: Asymptotic sample-level TPR of Algorithm 2 under the isotropic Gaussian setting.

Figure 2 plots the limiting TPR against $\log_{10}\alpha$ for $q = 0.2$, $\pi_0 = 0.9$, and $s/\sigma = 1$. Two findings stand out. First, for small γ, the TPR collapses to zero regardless of α, indicating that low-dimensional models are inherently resistant to inference. Second, and most importantly, stronger ridge regularization



(a) (b) (c) (d)

Figure 3: MIA results on synthetic data ($\pi_0 = 0.9$). (a) Realized vs. target FDR for FDR-controlling methods. (b) FDR–TPR trade-off curve for each score. (c) TPR vs. predetermined target FDR level q for FDR-controlling methods. (d) TPR vs. $\log_{10} \lambda$ at target FDR levels of 10% and 20%.

monotonically reduces asymptotic power: the TPR remains high under weak regularization but drops sharply once α exceeds a critical level, with this threshold increasing with γ. This sharp transition is consistent with the experimental results in Section 5; additional validation of the effect of γ is provided in Appendix E.

5 Experimental results

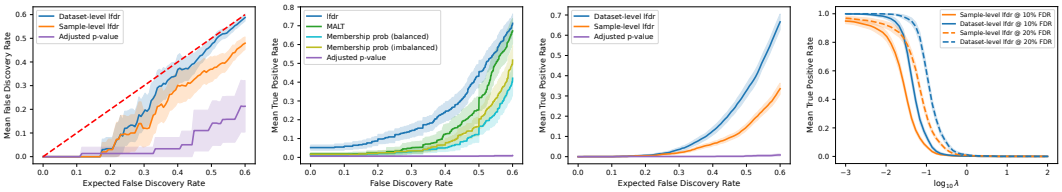
We evaluate the proposed methods against two categories of baselines. The first consists of FPR-calibrated attacks: MALT [22] and the shadow-model attack of Shokri et al. [24], which we refer to as Membership prob. For Membership prob we implement two variants: a balanced version that assumes equal prior membership probability ($\pi_0 = 0.5$), and an imbalanced version that incorporates the true prior π_0 to adjust the posterior membership probability. The second category consists of an existing FDR-controlling MIA method, the Adjusted p-value method of Zhao et al. [33]. This comparison allows us to assess both whether FDR control comes at a cost in power relative to FPR-calibrated attacks and whether our method improves over the existing FDR-controlled approach. Conventional multiple-testing procedures, including the Benjamini–Hochberg (BH) procedure and Storey's q-value method, are excluded due to their near-zero detection power in MIA. We refer to Algorithm 1 as dataset-level lfdr and Algorithm 2 as sample-level lfdr. Since the adversary does not have access to the true C_i and σ^2, we estimate these quantities from a reference dataset. Details of the implementation and additional experiments with different γ, π_0, and λ are provided in Appendices C and E, respectively.

5.1 Simulation study

Setup. Predictors are sampled i.i.d. from $\mathcal{N}(0, I_{200})$ and responses are generated from the linear model (1) with regression coefficients drawn from $\mathrm{Unif}([-1, 1]^{200})$ and noise variance $\sigma^2 = 4$. The training set consists of $N_{\mathrm{in}} = 100$ members, and the number of non-members is set to $N_{\mathrm{out}} = \lfloor \pi_0 N_{\mathrm{in}} / (1 - \pi_0) \rfloor$ to match the target null probability π_0. Unless otherwise stated, $\pi_0 = 0.9$ and $\lambda = 2$. The reference dataset used for nuisance estimation has size $N_{\mathrm{refer}} = 50 N_{\mathrm{in}}$. The adversary is assumed to know N_{in}, π_0, and λ. All results are averaged over 50 independent repetitions.

FDR calibration. Figure 3a plots the realized FDR against the target level q for the three FDR-controlling methods (MALT and Membership prob are excluded as they do not provide FDR control). The dataset-level local FDR MIA tracks the target level most closely, with realized FDR near the diagonal over a broad range of q, consistent with the theoretical guarantee under the nearly orthogonal Gaussian design. The sample-level local FDR MIA and Adjusted p-value are both more conservative, which ensures valid FDR control but reduces power when the adversary must commit to a target level in advance.

FDR–TPR trade-off. Figure 3b shows the FDR–TPR curves as the decision threshold varies. Since the dataset-level and sample-level MIAs share the same ranking of candidates, they trace the same curve, referred to as local FDR. It achieves the highest TPR in the low- and moderate-FDR regimes, which are the practically relevant ranges for high-confidence membership claims. Compared with local FDR, the existing MIA methods exhibit less favorable FDR–TPR trade-offs. Notably, Adjusted p-value yields the worst FDR–TPR trade-off in this experiment, despite being designed for



(a) (b) (c) (d)

Figure 4: MIA results on the real-world data ($\pi_0 = 0.9$). (a) Realized vs. target FDR for FDR-controlling methods. (b) FDR–TPR trade-off curve for each score. (c) TPR vs. predetermined target FDR for FDR-controlling methods. (d) TPR vs. $\log_{10} \lambda$ at target FDR levels of 10% and 20%.

FDR-controlling MIA. In the practical setting where the adversary must predetermine the target FDR level, Figure 3c confirms that the dataset-level procedure achieves strictly higher TPR at any fixed q, highlighting the practical value of accurate FDR calibration.

Effect of ridge regularization. Figure 3d plots the TPR of both local FDR MIAs against $\log_{10} \lambda$ at target FDR levels of 10% and 20%. As λ increases, the TPR of both procedures decreases monotonically and eventually collapses to zero, with the dataset-level procedure consistently more powerful than the sample-level rule at the same target FDR. This sharp decline is in precise agreement with Proposition 9, which predicts a threshold beyond which the TPR collapses to zero. Together, these results confirm that stronger ridge regularization substantially reduces the feasibility of successful membership inference under FDR control.

5.2 Real-world data experiment

Setup. We evaluate the proposed methods on the UCI Communities and Crime regression dataset [20]. After preprocessing and standardizing predictors and response, the dataset comprises 1,994 observations and 96 predictors, of which 64 have a variance inflation factor (VIF) exceeding 10, indicating severe multicollinearity. We fix $N_{\text{in}} = 50$, sample the training set uniformly at random and set $N_{\text{out}} = \lfloor \pi_0 N_{\text{in}} / (1 - \pi_0) \rfloor$; the remaining observations form the reference dataset. Unless otherwise stated, $\pi_0 = 0.9$ and $\lambda = 1$. The adversary is again assumed to know N_{in}, π_0, and λ. All results are averaged over 50 independent repetitions.

Results. Despite the severe multicollinearity of the dataset, the proposed local FDR methods achieve the highest TPR across the full FDR range considered, as shown in Figures 4b and 4c. Compared with the simulation study, both procedures are somewhat more conservative in FDR calibration (Figure 4a), which is expected since the conditional sufficiency assumption of Assumption 5 does not hold exactly under multicollinearity. Nevertheless, this conservativeness does not eliminate the power advantage over the baselines: the proposed methods maintain superior FDR–TPR trade-offs throughout. The dataset-level procedure continues to outperform the sample-level rule at any fixed target FDR, consistent with the pattern observed in the simulation study. Finally, Figure 4d confirms that stronger ridge regularization continues to reduce membership inference risk even in this highly correlated setting, extending the theoretical finding of Proposition 9 beyond the isotropic Gaussian regime.

6 Discussion

We formulated membership inference on multiple target samples as a multiple-testing problem and proposed local FDR-based attacks that provide rigorous FDR control. The leave-one-out characterization of ridge residuals yields closed-form conditional distributions under membership and non-membership, enabling FDR-controlled rejection rules, an explicit asymptotic power characterization, and superior FDR calibration over existing attacks. Experiments on synthetic and real-world data confirm favorable FDR–TPR trade-offs across a range of target FDR levels.

A central finding of our theoretical analysis is that stronger ridge regularization monotonically raises the minimal attainable FDR level and reduces asymptotic detection power, with a sharp threshold

beyond which reliable inference is no longer possible. The sharp-threshold behavior further suggests that, in principle, a model builder can choose λ to ensure that any FDR-controlled attack is rendered ineffective, providing a quantitative guideline for privacy-aware model selection.

Several limitations point to directions for future work. First, the theoretical FDR guarantee relies on the conditional sufficiency assumption (Assumption 5). Developing FDR guarantees for more general dependent residual statistics remains an important open problem. Second, the proposed attack assumes the adversary knows the ridge parameter, which may be unknown or deliberately concealed by the model builder. Extending the framework to settings with unknown λ is a natural next step. Experiments examining the sensitivity of the proposed methods to the choice of λ are provided in Appendix E. Third, our analysis does not systematically investigate how the predictive performance of the target model influences MIA power under FDR control. Existing work has related membership leakage to overfitting and generalization error [30, 16, 11], and characterizing this relationship in the FDR-controlled setting is an important direction for future work.

Acknowledgments and Disclosure of Funding

This work was supported by National IT Industry Promotion Agency (NIPA) grant funded by the Korea government (MSIT) (AI-Based Dark Factory Design-Deployment Solution Development, No. RS-2026-25621604), and National Research Foundation of Korea (NRF) grants (RS-2026-25597323, RS-2026-25517529).

References

[1] J. Ahn, J. S. Marron, K. M. Muller, and Y.-Y. Chi. The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766, aug 5 2007. ISSN 0006-3444. doi: 10.1093/biomet/asm050.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[3] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[4] M. Bertran, S. Tang, A. Roth, M. Kearns, J. H. Morgenstern, and S. Z. Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36:314–330, 2023.

[5] T. T. Cai and W. Sun. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481, 12 2009. ISSN 0162-1459. doi: 10.1198/jasa.2009.tm08415.

[6] T. T. Cai, J. Fan, and T. Jiang. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 14(136):1837–1864, 2013.

[7] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022.

[8] M. Celentano, A. Montanari, and Y. Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5), oct 1 2023. ISSN 0090-5364. doi: 10.1214/23-aos2327.

[9] R. Couillet and Z. Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

[10] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403, 1978.

[11] A. Dionysiou. Sok: Membership inference is harder than previously thought. *Zenodo (CERN European Organization for Nuclear Research)*, 2023.

[12] C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[13] B. Efron. Size, power and false discovery rates. *The Annals of Statistics*, pages 1351–1377, 2007.

[14] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2), apr 1 2022. ISSN 0090-5364. doi: 10.1214/21-aos2133.

[15] N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580, 2016. ISSN 1548-7091.

[16] Y. Kaya, S. Hong, and T. Dumitras. On the effectiveness of regularization against membership inference attacks. *arXiv preprint arXiv:2006.05336*, 2020.

[17] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

[18] L. Lei and W. Fithian. AdaPT: An interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):649–679, jun 4 2018. ISSN 1369-7412.

[19] K. R. Rad and A. Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996, 2020.

[20] M. Redmond. Communities and Crime Unnormalized. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C5PC8X.

[21] E. Rio. Covariance inequalities for strongly mixing processes. In *Annales de l'IHP Probabilités et statistiques*, volume 29, pages 587–597, 1993.

[22] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.

[23] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

[24] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017. doi: 10.1109/SP.2017.41.

[25] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[26] W. Sun and T. T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 9 2007. ISSN 0162-1459.

[27] J. Tan, B. Mason, H. Javadi, and R. Baraniuk. Parameters or privacy: A provable tradeoff between overparameterization and membership inference. *Advances in Neural Information Processing Systems*, 35:17488–17500, 2022.

[28] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pages 3093–3106, 2022.

[29] J. Ye, A. Borovykh, S. Hayou, and R. Shokri. Leave-one-out distinguishability in machine learning. *arXiv preprint arXiv:2309.17310*, 2023.

[30] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[31] R. W. Zablocki, A. J. Schork, R. A. Levine, O. A. Andreassen, A. M. Dale, and W. K. Thompson. Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics*, 30(15):2098–2104, 2014.

[32] S. Zarifzadeh, P. Liu, and R. Shokri. Low-cost high-power membership inference attacks. *arXiv preprint arXiv:2312.03262*, 2023.

[33] C. Zhao, W. Qian, A. Chen, and M. Huai. Membership inference attacks with false discovery rate control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1216–1227, 2025.

A Proofs

A.1 Proof of Proposition 6

Proof. We first prove the FDR guarantee of Algorithm 1. Let $W = (X, Z_1, \ldots, Z_N)$.

$$\begin{aligned}\mathbb{E}[V \mid W] &= \mathbb{E}\left[\sum_{i=1}^{N} \widehat{M}_i \cdot \mathbb{I}(M_i = 0) \mid W\right] = \sum_{i:\widehat{M}_i=1}^{N} \mathbb{P}(M_i = 0 \mid W) \\ &= \sum_{i:\widehat{M}_i=1}^{N} \mathbb{P}(M_i = 0 \mid X, Z_i) = \sum_{i:\widehat{M}_i=1}^{N} \mathrm{lfdr}(r_i, X, C_i).\end{aligned}$$

Algorithm 1 chooses the rejection set where the average local FDR of rejected samples is at most q. That is,

$$\frac{1}{\max\{R, 1\}} \sum_{i:\widehat{M}_i=1}^{N} \mathrm{lfdr}(r_i, X, C_i) \le q. \tag{2}$$

Therefore,

$$\begin{aligned}\mathrm{FDR} &= \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{V}{\max\{R, 1\}} \mid W\right]\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[V \mid W]}{\max\{R, 1\}}\right] \le \mathbb{E}[q] = q.\end{aligned}$$

In addition, Algorithm 2 rejects the sample i if

$$\mathrm{lfdr}(r_i, X, C_i) \le q.$$

It also satisfy the inequality (2), therefore Algorithm 2 controls the false discovery rate at level q. □

A.2 Proof of Proposition 7

Proof. The rejection region of Algorithm 2 is

$$\frac{f_{i1}(r_i \mid X, C_i)}{f_{i0}(r_i \mid X, C_i)} \ge \frac{1/q - 1}{1/\pi_0 - 1} =: c_q.$$

Take the logarithm on both sides yields

$$\log \frac{f_{i1}(r_i \mid X, C_i)}{f_{i0}(r_i \mid X, C_i)} = -\log(1 - h_i) + \frac{-h_i(2 - h_i)r_i^2 + 2\mu_i h_i r_i}{2d_i\sigma^2}.$$

The rejection region becomes

$$h_i(2-h_i)r_i^2 - 2\mu_i h_i r_i + 2d_i\sigma^2 \log\left(c_q(1-h_i)\right) \leq 0. \tag{3}$$

(3) has a solution if

$$(\mu_i h_i)^2 - h_i(2-h_i) \times 2d_i\sigma^2 \log\left(c_q(1-h_i)\right) \geq 0$$

After simplification,

$$\frac{1}{1-h_i} \exp\left(\frac{\mu_i}{2\sigma^2} \cdot \frac{h_i}{d_i(2-h_i)}\right) \geq c_q.$$

Note that

$$c_q = \frac{1/q - 1}{1/\pi_0 - 1},$$

therefore solving for q yields

$$q \geq \left[1 + \left(\frac{1}{\pi_0} - 1\right)\left(\frac{1}{1-h_i}\right)\exp\left(\frac{\mu_i^2}{2\sigma^2} \cdot \frac{h_i}{d_i(2-h_i)}\right)\right]^{-1} =: q_i^\star.$$

□

A.3 Proof of Proposition 9

First, we prove the following lemma.

Lemma 10. *Assume the setting of Proposition 9. Let h_i, d_i, and μ_i be the quantities from Proposition 3 computed with X_{tr}. Then, conditional on $x_{i,p}$,*

$$h_i \xrightarrow{\text{a.s.}} \frac{\gamma m_\alpha}{1+\gamma m_\alpha}, \qquad \mu_i \xrightarrow{\text{a.s.}} = \frac{\alpha m_\alpha s}{1+\gamma m_\alpha}, \qquad d_i \xrightarrow{\text{a.s.}} \frac{1+\gamma m_\alpha - \alpha\gamma m'_\alpha}{(1+\gamma m_\alpha)^2}$$

where $m_\alpha = m(-\alpha)$, $m'_\alpha = m'(-\alpha)$, and $m(z)$ is the Stieltjes transform of the Marchenko–Pastur law with aspect ratio γ,

$$m_\alpha = \frac{-(\alpha+1-\gamma) + \sqrt{(\alpha+1-\gamma)^2 + 4\gamma\alpha}}{2\gamma\alpha},$$

$$m'_\alpha = \frac{m_\alpha(1+\gamma m_\alpha)}{\alpha+1-\gamma+2\gamma\alpha m_\alpha}.$$

Proof. Write

$$S := \frac{1}{n} X_{\mathrm{tr}}^\top X_{\mathrm{tr}}, \qquad Q := (S + \alpha I_p)^{-1}.$$

Then

$$H = X_{\mathrm{tr}}(X_{\mathrm{tr}}^\top X_{\mathrm{tr}} + n\alpha I_p)^{-1} X_{\mathrm{tr}}^\top = \frac{1}{n} X_{\mathrm{tr}} Q X_{\mathrm{tr}}^\top.$$

Moreover,

$$\begin{aligned} H^2 &= X_{\mathrm{tr}}(X_{\mathrm{tr}}^\top X_{\mathrm{tr}} + n\alpha I_p)^{-1} X_{\mathrm{tr}}^\top X_{\mathrm{tr}}(X_{\mathrm{tr}}^\top X_{\mathrm{tr}} + n\alpha I_p)^{-1} X_{\mathrm{tr}}^\top \\ &= H - n\alpha X_{\mathrm{tr}}(X_{\mathrm{tr}}^\top X_{\mathrm{tr}} + n\alpha I_p)^{-2} X_{\mathrm{tr}}^\top \\ &= \frac{1}{n} X_{\mathrm{tr}} Q X_{\mathrm{tr}}^\top - \frac{\alpha}{n} X_{\mathrm{tr}} Q^2 X_{\mathrm{tr}}^\top. \end{aligned}$$

Hence

$$h_i = e_i^\top H e_i = \frac{1}{n} x_{i,p}^\top Q x_{i,p},$$

and

$$d_i = e_i^\top (I_n - H)^2 e_i = 1 - 2e_i^\top H e_i + e_i^\top H^2 e_i = 1 - h_i - \frac{\alpha}{n} x_{i,p}^\top Q^2 x_{i,p}.$$

For μ_i, we also have

$$\begin{aligned}\mu_i &= e_i^\top (I_n - H) X_{\mathrm{tr}} \beta_p \\ &= x_{i,p}^\top \beta_p - x_{i,p}^\top (X_{\mathrm{tr}}^\top X_{\mathrm{tr}} + n\alpha I_p)^{-1} X_{\mathrm{tr}}^\top X_{\mathrm{tr}} \beta_p \\ &= n\alpha\, x_{i,p}^\top (X_{\mathrm{tr}}^\top X_{\mathrm{tr}} + n\alpha I_p)^{-1} \beta_p \\ &= \alpha x_{i,p}^\top Q \beta_p.\end{aligned}$$

Let X_{-i} be the matrix obtained by deleting the row $x_{i,p}$ from X_{tr}, and define

$$Q_{-i} := \left(\frac{1}{n} X_{-i}^\top X_{-i} + \alpha I_p \right)^{-1}.$$

Then

$$Q^{-1} = Q_{-i}^{-1} + \frac{1}{n} x_{i,p} x_{i,p}^\top.$$

By the Sherman–Morrison formula,

$$Q = Q_{-i} - \frac{Q_{-i} x_{i,p} x_{i,p}^\top Q_{-i}}{n + x_{i,p}^\top Q_{-i} x_{i,p}}.$$

Set

$$h_{-i} := \frac{1}{n} x_{i,p}^\top Q_{-i} x_{i,p}, \qquad \mu_{-i} := \alpha x_{i,p}^\top Q_{-i} \beta_p, \qquad a_{-i} := \frac{1}{n} x_{i,p}^\top Q_{-i}^2 x_{i,p}.$$

Note that

$$Q x_{i,p} = Q_{-i} x_{i,p} - \frac{Q_{-i} x_{i,p} x_{i,p}^\top Q_{-i} x_{i,p}}{n + x_{i,p}^\top Q_{-i} x_{i,p}} = Q_{-i} x_{i,p} - \frac{h_{-i} Q_{-i} x_{i,p}}{1 + h_{-i}},$$

$$Q x_{i,p} = \frac{Q_{-i} x_{i,p}}{1 + h_{-i}}.$$

We obtain

$$h_i = \frac{1}{n} x_{i,p}^\top Q x_{i,p} = \frac{h_{-i}}{1 + h_{-i}}, \qquad \mu_i = \alpha x_{i,p}^\top Q \beta_p = \frac{\mu_{-i}}{1 + h_{-i}},$$

$$\frac{1}{n} x_{i,p}^\top Q^2 x_{i,p} = \frac{1}{n} (Q x_{i,p})^\top (Q x_{i,p}) = \frac{a_{-i}}{(1 + h_{-i})^2}.$$

Consequently,

$$d_i = 1 - \frac{h_{-i}}{1 + h_{-i}} - \alpha \frac{a_{-i}}{(1 + h_{-i})^2} = \frac{1 + h_{-i} - \alpha a_{-i}}{(1 + h_{-i})^2}. \tag{4}$$

Since replacing n by $n - 1$ does not change the limiting aspect ratio, Theorem 2.4 of [9] gives the same deterministic equivalents for Q_{-i}. Conditional on $x_{i,p}$,

$$\frac{1}{p} x_{i,p}^\top Q_{-i} x_{i,p} \xrightarrow{\text{a.s.}} m_\alpha, \qquad x_{i,p}^\top Q_{-i} \beta_p \xrightarrow{\text{a.s.}} m_\alpha s.$$

The same deterministic-equivalent argument applied to the derivative of the resolvent, equivalently to Q_{-i}^2, yields

$$\frac{1}{p} x_{i,p}^\top Q_{-i}^2 x_{i,p} \xrightarrow{\text{a.s.}} m'_\alpha$$

since

$$\frac{d}{dz} (S - z I_p)^{-1} = (S - z I_p)^{-2}.$$

Therefore,

$$h_{-i} = \frac{p}{n} \cdot \frac{1}{p} x_{i,p}^\top Q_{-i} x_{i,p} \xrightarrow{\text{a.s.}} \gamma m_\alpha, \qquad a_{-i} = \frac{p}{n} \cdot \frac{1}{p} x_{i,p}^\top Q_{-i}^2 x_{i,p} \xrightarrow{\text{a.s.}} \gamma m'_\alpha,$$

and

$$\mu_{-i} = \alpha x_{i,p}^\top Q_{-i}\beta_p \xrightarrow{\text{a.s.}} \alpha m_\alpha s.$$

By the continuous mapping theorem,

$$h_i = \frac{h_{-i}}{1+h_{-i}} \xrightarrow{\text{a.s.}} \frac{\gamma m_\alpha}{1+\gamma m_\alpha},$$

and

$$\mu_i = \frac{\mu_{-i}}{1+h_{-i}} \xrightarrow{\text{a.s.}} \frac{\alpha m_\alpha s}{1+\gamma m_\alpha}.$$

Finally, using (4),

$$d_i = \frac{1+h_{-i}-\alpha a_{-i}}{(1+h_{-i})^2} \xrightarrow{\text{a.s.}} \frac{1+\gamma m_\alpha - \alpha\gamma m'_\alpha}{(1+\gamma m_\alpha)^2}.$$

□

Now we prove Proposition 9.

Proof. Under H_{i1}, Proposition 3 gives

$$r_i \mid X_{\text{tr}}, x_{i,p} \sim \mathcal{N}\big(\mu_i, d_i\sigma^2\big).$$

Standardize r_i by writing

$$z := \frac{r_i - \mu_i}{\sigma\sqrt{d_i}}.$$

Then $z \sim \mathcal{N}(0,1)$ conditionally on $(X_{\text{tr}}, x_{i,p})$.

By Proposition 7, rejection occurs exactly when

$$h_i(2-h_i)r_i^2 - 2\mu_i h_i r_i + 2d_i\sigma^2 \log\{c_q(1-h_i)\} \le 0.$$

Substituting $r_i = \mu_i + \sigma\sqrt{d_i}\, z$ and simplifying gives the equivalent condition

$$h_i(2-h_i)z^2 + 2h_i(1-h_i)\frac{\mu_i}{\sigma\sqrt{d_i}}z - h_i^2\frac{\mu_i^2}{d_i\sigma^2} + 2\log\{c_q(1-h_i)\} \le 0. \tag{5}$$

Define the finite-sample discriminant

$$\Delta_{n,i} := \frac{\mu_i^2 h_i^2}{\sigma^2 d_i} - 2h_i(2-h_i)\log\{c_q(1-h_i)\}.$$

When $\Delta_{n,i} \ge 0$, the two roots of (5) are

$$z_{i,\pm} = \frac{-h_i(1-h_i)\frac{\mu_i}{\sigma\sqrt{d_i}} \pm \sqrt{\Delta_{n,i}}}{h_i(2-h_i)}, \qquad z_{i,-} \le z_{i,+}.$$

For Algorithm 2, define the sample-level true positive rate, conditional on $X_{\text{tr}}, x_{i,p}$,

$$\text{TPR}_n(X_{\text{tr}}) := \mathbb{P}\big(\text{lfdr}(r_i, X, \mu_i, h_i) \le q \mid X_{\text{tr}}, x_{i,p}\big).$$

Then,

$$\text{TPR}_n(X_{\text{tr}}) = \Phi(z_{i,+}) - \Phi(z_{i,-})$$

where Φ is the cumulative distribution function of the standard normal distribution. If the quadratic has no real root, then the rejection region is empty and $\text{TPR}_n(X_{\text{tr}}) = 0$.

By Lemma 10, conditional on $x_{i,p}$,

$$h_i \xrightarrow{\text{a.s.}} h_\bullet := \frac{\gamma m_\alpha}{1+\gamma m_\alpha}, \qquad \mu_i \xrightarrow{\text{a.s.}} \mu_\bullet := \frac{\alpha m_\alpha s}{1+\gamma m_\alpha},$$
$$d_i \xrightarrow{\text{a.s.}} d_\bullet := \frac{1+\gamma m_\alpha - \alpha\gamma m'_\alpha}{(1+\gamma m_\alpha)^2}.$$

Consequently,

$$\Delta_{n,i} \xrightarrow{\text{a.s.}} \Delta_\bullet,$$

conditionally on $x_{i,p}$, where

$$\Delta_\bullet := \frac{\mu_\bullet^2 h_\bullet^2}{\sigma^2 d_\bullet} - 2h_\bullet(2-h_\bullet)\log\{c_q(1-h_\bullet)\}.$$

The condition $\Delta_\bullet \geq 0$ is equivalent to

$$\log\{c_q(1-h_\bullet)\} \leq \frac{\mu_\bullet^2 h_\bullet}{2d_\bullet(2-h_\bullet)\sigma^2}.$$

Equivalently,

$$q \geq q_\bullet^\star,$$

where

$$\begin{aligned} q_\bullet^\star &= \left[1 + \left(\frac{1}{\pi_0} - 1\right)\left(\frac{1}{1-h_\bullet}\right)\exp\left(\frac{\mu_\bullet^2 h_\bullet}{2d_\bullet(2-h_\bullet)\sigma^2}\right)\right]^{-1} \\ &= \left[1 + \left(\frac{1}{\pi_0} - 1\right)(1+\gamma m_\alpha)\exp\left(\frac{\gamma\alpha^2 m_\alpha^3 (s/\sigma)^2}{2(2+\gamma m_\alpha)(1+\gamma m_\alpha - \alpha\gamma m'_\alpha)}\right)\right]^{-1}. \end{aligned}$$

First suppose $q > q_\bullet^\star$. Then $\Delta_\bullet > 0$, so $\Delta_{n,i} > 0$ for all sufficiently large n, almost surely conditionally on $x_{i,p}$. By continuity,

$$z_{i,\pm} \xrightarrow{\text{a.s.}} z_\pm,$$

where

$$z_\pm = \frac{-h_\bullet(1-h_\bullet)\frac{\mu_\bullet}{\sigma\sqrt{d_\bullet}} \pm \sqrt{\Delta_\bullet}}{h_\bullet(2-h_\bullet)}.$$

Therefore,

$$\mathrm{TPR}_n(X_{\mathrm{tr}}) = \Phi(z_{i,+}) - \Phi(z_{i,-}) \xrightarrow{\text{a.s.}} \Phi(z_+) - \Phi(z_-),$$

conditionally on $x_{i,p}$.

Next suppose $q < q_\bullet^\star$. Then $\Delta_\bullet < 0$, and hence $\Delta_{n,i} < 0$ for all sufficiently large n, almost surely conditionally on $x_{i,p}$. Therefore the rejection region is eventually empty, and

$$\mathrm{TPR}_n(X_{\mathrm{tr}}) \xrightarrow{\text{a.s.}} 0.$$

At the boundary $q = q_\bullet^\star$, the two limiting roots coincide, hence $\Phi(z_+) - \Phi(z_-) \to 0$, so the same conclusion remains valid.

Finally, $0 \leq \mathrm{TPR}_n(X_{\mathrm{tr}}) \leq 1$, so the sample-level true positive rate of Algorithm 2 with given $x_{i,p}$ converges to

$$\mathrm{TPR}(i \mid x_{i,p}) = \mathbb{E}\big[\mathrm{TPR}_n(X_{\mathrm{tr}}) \mid x_{i,p}\big] \to \begin{cases} \Phi(z_+) - \Phi(z_-), & q \geq q_\bullet^\star, \\ 0, & q < q_\bullet^\star, \end{cases}$$

by dominated convergence. □

B Relaxation of Assumption 5

Assumption 5 is sufficient for exact finite-sample FDR control, but it can be relaxed for Algorithm 2 when the dependence among the local posterior errors is sufficiently weak. This provides settings beyond pairwise orthogonal or nearly orthogonal designs in which the procedure remains asymptotically reliable. First, we state the α-mixing relaxation of Assumption 5. For a given predetermined FDR level q, denote $U_i := (\mathbb{I}(M_i = 0) - l_i)\mathbb{I}(l_i \leq q)$. We consider the following conditional α-mixing coefficients for the sequence $\{U_i\}_{i=1}^N$:

$$\alpha(k \mid X) := \max_{1 \leq m \leq N-k} \sup_{\substack{A \in \sigma(U_1,\ldots,U_m), \\ B \in \sigma(U_{m+k},\ldots,U_N)}} \left|\mathbb{P}(A \cap B \mid X) - \mathbb{P}(A \mid X)\mathbb{P}(B \mid X)\right|.$$

Then the relaxed conditions for Assumption 5 are

$$\sum_{k=1}^{N-1}(N-k)\alpha(k \mid X) = o(N^2) \quad \text{and} \quad \sum_{i=1}^{N}\mathbb{I}(l_i \le q) = \Theta_p(N).$$

The first condition is a relaxation of the independence assumption. The second condition requires a nonvanishing proportion of discoveries. This is natural for a meaningfully chosen q: under the Bayesian formulation, the number of true members is concentrated around $(1-\pi_0)N$, and an informative MIA with nonvanishing power should declare a number of members scales as $\Theta(N)$. Then we can control the FDR asymptotically as follows. First, decompose FDP as below.

$$\begin{aligned}
\mathrm{FDP} &= \frac{\sum_{i=1}^{N}\mathbb{I}(M_i=0)\mathbb{I}(l_i \le q)}{\max\left\{\sum_{i=1}^{N}\mathbb{I}(l_i \le q), 1\right\}} \\
&= \frac{\sum_{i=1}^{N} l_i \mathbb{I}(l_i \le q)}{\max\left\{\sum_{i=1}^{N}\mathbb{I}(l_i \le q), 1\right\}} + \frac{\sum_{i=1}^{N}(\mathbb{I}(M_i=0)-l_i)\mathbb{I}(l_i \le q)}{\max\left\{\sum_{i=1}^{N}\mathbb{I}(l_i \le q), 1\right\}} \\
&\le q + \frac{\sum_{i=1}^{N} U_i}{\max\left\{\sum_{i=1}^{N}\mathbb{I}(l_i \le q), 1\right\}}.
\end{aligned}$$

Theorem 1.1 of Rio [21] implies that

$$|\mathrm{Cov}(U_i, U_j \mid X)| \le 4\alpha(|i-j| \mid X)\|U_i\|_\infty \|U_j\|_\infty \le 4\alpha(|i-j| \mid X).$$

Thus,

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{i=1}^{N} U_i\right| \ge t \middle| X\right) &\le \frac{\sum_{i=1}^{N}\mathrm{Var}(U_i \mid X) + \sum_{i\ne j}\mathrm{Cov}(U_i, U_j \mid X)}{t^2} \\
&\le \frac{N + 8\sum_{k=1}^{N-1}(N-k)\alpha(k \mid X)}{t^2} \\
&= o\left(\frac{N^2}{t^2}\right).
\end{aligned}$$

Since $\sum_{i=1}^{N}\mathbb{I}(l_i \le q) = \Theta_p(N)$, it follows that

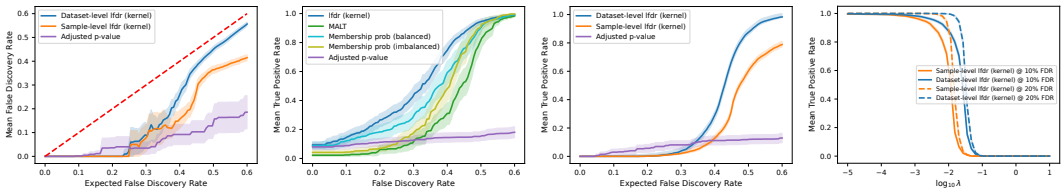
$$\frac{\sum_{i=1}^{N} U_i}{\max\left\{\sum_{i=1}^{N}\mathbb{I}(l_i \le q), 1\right\}} \xrightarrow{p} 0.$$

Thus, $\limsup_{N\to\infty}\mathrm{FDR} \le q$, as the FDR is the expectation of the FDP.

Next, we provide a concrete sufficient setting for $X_1, \ldots, X_N$ to satisfy the mixing condition: block orthogonality condition. Suppose there exists a partition $B_1 \cup \cdots \cup B_L = \{1, \ldots, N\}$ such that $\max_a |B_a| = o(N)$ for any $a \ne b$, if $i \in B_a$ and $j \in B_b$ then $\langle x_i, x_j \rangle = 0$. Within each block, the predictors may be arbitrarily correlated.

Under this condition, the ridge estimator decomposes across the mutually orthogonal row subspaces. Therefore, for $i \in B_a$, the residual r_i and the nuisance summary C_i depend only on the observations and membership indicators in block B_a. Since the responses and membership indicators are independent across samples, the collections $\{U_i : i \in B_a\}$ are conditionally independent across blocks given X. Thus, $l_i \perp l_j$ if $|i-j| \ge \max_a |B_a|$. After reindexing the observations so that the blocks are contiguous, $\alpha(k \mid X) = 0$ for $k \ge \max_a |B_a|$, so the relaxed condition is satisfied: $\sum_{k=1}^{N-1}(N-k)\alpha(k \mid X) = O(N \max_a |B_a|) = o(N^2)$.

This relaxation shows that exact conditional sufficiency is not essential for the asymptotic reliability of the sample-level procedure. In particular, the condition $\max_a |B_a| = o(N)$ allows that the FDR can be asymptotically controlled even if asymptotically almost every data points are correlated.



(a) (b) (c) (d)

Figure 5: MIA results for kernel ridge regression on the real-world data ($\pi_0 = 0.9$).

C Estimation procedure of nuisance quantities

The oracle attacks require the nuisance quantities h_i, d_i and μ_i, which depend on the unknown training membership. A practical adversary must therefore replace them with estimates, and must also estimate σ^2, the noise variance. Because μ_i depends on β, estimating μ_i also requires estimating β. This estimation strategy is analogous to a shadow-model approach based on an auxiliary reference dataset. In Section 5, we estimate σ^2 and β using the full reference dataset. To estimate C_i, we split the reference dataset into J sub-datasets, each with the same size as the target training set, since leverage depends strongly on sample size. For the jth reference sub-dataset, let $\widehat{C}_i^{(j)}$ denote the corresponding estimates, and define

$$\widehat{\mathrm{lfdr}}_i^{(j)} = \mathrm{lfdr}\left(r_i, X, \widehat{C}_i^{(j)}\right)$$

We then average the estimates over the J reference sub-datasets to obtain the final local FDR estimate for the i-th sample.

D Local FDR MIA in kernel ridge regression

For the kernel ridge regression model, local FDR MIA can be applied immediately using same leave-one-out identities in the ridge regression with different nuisance summaries C_i. Consider

$$y_j = f_0(x_j) + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma^2), \qquad f_0 \in \mathcal{H}_k,$$

where $\mathcal{H}_k$ is the reproducing kernel Hilbert space (RKHS) associated with a kernel k. The kernel ridge estimator is

$$\widehat{\alpha} = (K_{\mathrm{tr}} + \lambda I)^{-1} Y_{\mathrm{tr}}, \qquad \widehat{f}(x) = k(x, X_{\mathrm{tr}})^\top \widehat{\alpha}.$$

For a candidate (x_i, y_i), let

$$I_{\mathrm{tr},i} = I_{\mathrm{tr}} \cup \{i\}$$

denote the augmented index set, and let $K_{\mathrm{tr},i}$ be its kernel matrix. Define the associated linear-smoother matrix

$$H_{\lambda,i} = K_{\mathrm{tr},i}(K_{\mathrm{tr},i} + \lambda I)^{-1},$$

and

$$h_i = [H_{\lambda,i}]_{ii}, \qquad d_i = [(I - H_{\lambda,i})^2]_{ii}, \qquad \mu_i = [(I - H_{\lambda,i}) f_{0,i}]_i,$$

where

$$f_{0,i} = \big(f_0(x_j)\big)_{j \in I_{\mathrm{tr},i}}.$$

The leave-one-out identity for kernel ridge regression then gives the same form of conditional residual distributions as in Proposition 3. Thus, after estimating (h_i, d_i, μ_i) using the same reference-data strategy as in Appendix C, these two densities can be plugged into the local FDR procedure.

We repeat the experiments in Section 5.2 using kernel ridge regression as the target model, with a radial basis function (RBF) kernel whose variance parameter is set to 0.5. Although we do not establish that Assumption 5 holds for kernel ridge regression, the realized FDR remains below the prespecified target levels across all target FDR levels (Figure 5a). Moreover, the kernel local FDR score achieves the highest mean TPR among the competing methods at each matched FDR level (Figure 5b). These findings suggest that the proposed local FDR approach can also be effective for kernel ridge regression, beyond the linear setting considered in our theoretical analysis. Figure 5d

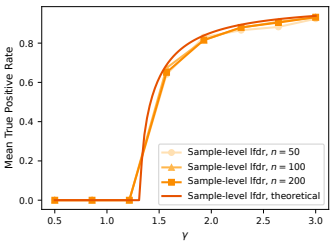


Figure 6: Mean TPR as a function of γ at 10% target FDR level and $\alpha = 10^{-2}$ in the simulation study.

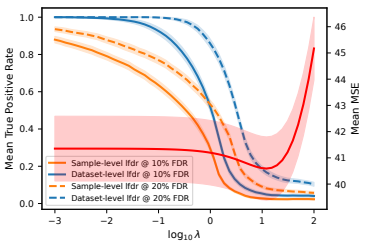


Figure 7: Privacy–utility trade-off shown in Figure 3d, with the mean test MSE curve overlaid.

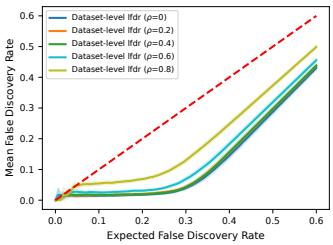


Figure 8: Realized vs. target FDR for the dataset-level local FDR MIA under autoregressive covariance structures with different ρ.

further shows that increasing the kernel ridge regularization parameter substantially reduces attack power, consistent with the qualitative findings for linear ridge regression.

These empirical findings, however, do not establish a general FDR guarantee for nonlinear target models. Deriving such a guarantee would require additional theoretical conditions, potentially analogous to Assumption 5. For more complex models such as neural networks, the required member and non-member distributions would generally need to be approximated using, for example, leave-one-out approximations, influence functions, or suitable reference models. Establishing both the accuracy of such approximations and the corresponding FDR guarantee remains future work.

E Additional experiments

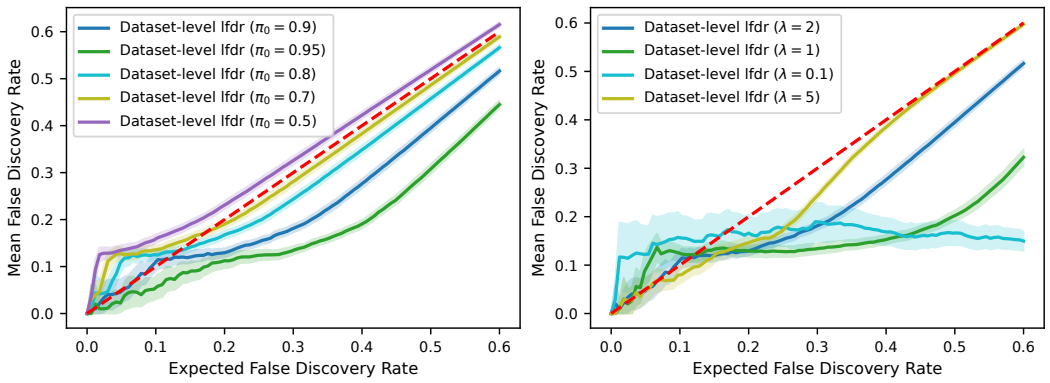
This section presents additional experiments that complements the discussion in Section 5.

Effect of γ. Figure 6 shows the mean TPR of the sample-level local FDR MIA as a function of the aspect ratio $\gamma = p/n$, at a target FDR level 10% and regularization scaling parameter $\alpha = 10^{-2}$, under the simulation setting described in Section 5.1. The empirical curves for finite sample sizes $n = 50, 100, 200$ closely follow the limiting TPR derived in Section 4.2. This provides empirical support for the asymptotic characterization in Proposition 9. The figure also shows a threshold-like behavior: for smaller values of γ, the TPR is close to zero, whereas for larger γ the attack becomes substantially more powerful. This pattern is consistent with the limiting theory, where the rejection region is empty when the target FDR level is below the limiting minimal attainable level.

Privacy–utility trade-off. Figure 7 shows that increasing λ can substantially reduce the TPR of the local FDR attacks while preserving and initially improving predictive performance. However, excessive regularization eventually increases the test MSE. Thus, stronger regularization should not be interpreted as an unconditionally beneficial privacy defense and λ should be selected by explicitly balancing privacy risk and predictive utility. For example, the model builder can select an optimal ridge parameter $\lambda^\star$ as the value that provides the strongest privacy protection while maintaining an acceptable level of predictive performance. Alternatively, $\lambda^\star$ may be chosen at the knee point of the empirical Pareto frontier between the TPR of the local FDR MIA and predictive performance. As a theoretical next step, the asymptotic TPR expression derived in this paper could be combined with an asymptotic prediction-risk formula for ridge regression. This would yield an explicit characterization of $\lambda^\star$, rather than relying only on empirical model selection.

Violation of Assumption 5 under autoregressive dependence. To examine the effect of dependence, we repeated the simulation study using Gaussian covariates with an autoregressive covariance structure, $\Sigma_{jk} = \sigma^2 \rho^{|j-k|}$ where $\sigma = 3$ and $\rho \in \{0, 0.2, 0.4, 0.6, 0.8\}$. Increasing ρ induces stronger correlations among the covariates, making Assumption 5 less plausible. All other simulation parameters were held fixed at the values used in Section 5.1.

The realized FDR generally increases with ρ, and the largest increase observed at $\rho = 0.8$. Although it remains below the target level q in most settings, the realized FDR slightly exceeds the target at very



(a) (b)

Figure 9: Realized vs. target FDR for the dataset-level local FDR MIA under misspecification of (a) π_0 and (b) λ. The true parameter values are $\pi_0 = 0.9$ and $\lambda = 2$.

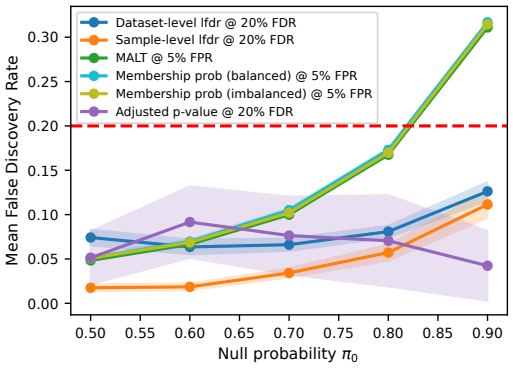


Figure 10: Mean realized FDR as a function of the null probability π_0 in the simulation study. FDR-based attacks target 20% FDR, whereas FPR-based baselines target 5% FPR.

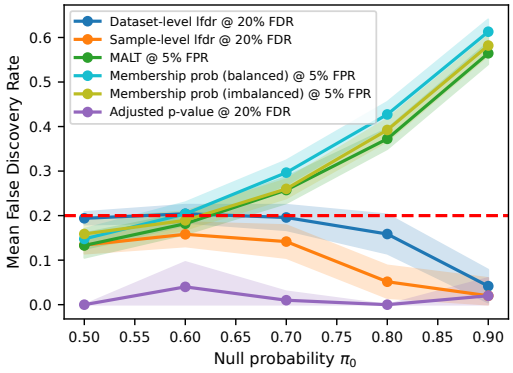


Figure 11: Mean realized FDR as a function of the null probability π_0 in the real-world data experiment. FDR-based attacks target 20% FDR, whereas FPR-based baselines target 5% FPR.

small values of q. These results suggest that the procedure becomes less conservative as covariate dependence strengthens, especially at low target levels. However, because ρ is only an indirect proxy for violations of Assumption 5, this experiment should be interpreted as a sensitivity analysis rather than evidence of FDR control under general dependence.

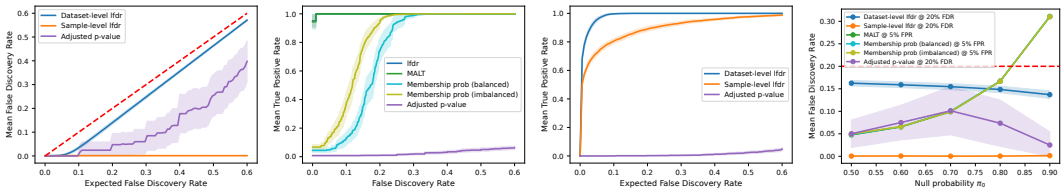
Misspecification of π_0 and λ. The experiments in Section 5 assume that the adversary knows the true values of π_0 and λ. In practice, however, these parameters may be misspecified. We examine the sensitivity of FDR calibration to such misspecification using the simulation setting in Section 5.1.

Figure 9a shows that underestimating π_0 assigns a greater prior probability to membership and leads to a more liberal rejection rule. However, substantial underestimation can cause the realized FDR to exceed the target level, especially at low target FDR levels. By contrast, overestimating π_0 makes the procedure more conservative and generally reduces realized FDR. Therefore, when π_0 is uncertain and reliable FDR control is the primary objective, using a conservative upper estimate of π_0 can reduce the risk of anti-conservative inference.

Unlike π_0, λ affects the nuisance summaries $C_i = (h_i, d_i, \mu_i)$ and therefore changes the conditional residual distributions used to compute the local FDR. Figure 9b shows that misspecification of λ can produce noticeable FDR calibration errors, especially at low target FDR levels. In practice, such misspecification may arise when the adversary uses reference or shadow models whose hyperparameters differ from those of the target model.

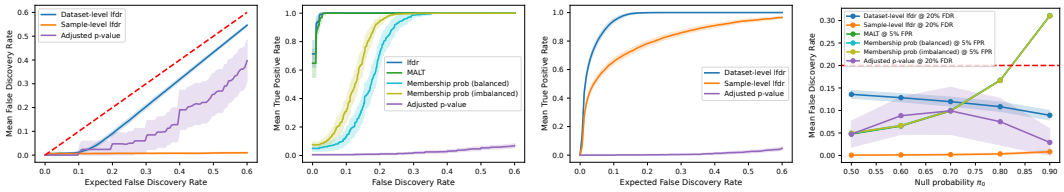
FDR on the null probability. We investigate the FDR calibration as the non-member proportion varies. Figure 10 and 11 fix the target FDR at 20% for FDR-based attacks and the target FPR at 5% for FPR-based baselines, while varying $\pi_0 \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. The FPR-based baselines become progressively less reliable as π_0 increases, even under low FPR control. In contrast, the FDR-based methods exhibit more stable behavior, tracking the target FDR level closely across the full range of non-member proportions. This illustrates why per-sample FPR control is insufficient when MIA is applied to multiple targets.

Experiments with different λ. Figures 12–21 present additional simulation and real-world data experiments for $\lambda \in \{0.01, 0.1, 1, 2, 5, 10\}$. Results already reported in the main text are omitted. These additional experiments show similar trends: the dataset-level local FDR MIA achieves the tightest FDR control and the most favorable FDR–TPR trade-off across most settings. Figure 16a further illustrates a case in which the sample-level local FDR MIA provides a more conservative alternative when the dataset-level local FDR MIA exceeds the specified FDR level.



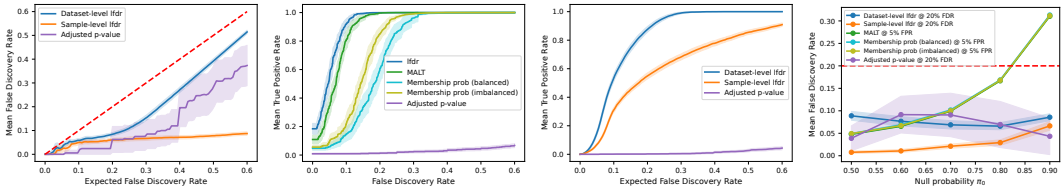
(a) (b) (c) (d)

Figure 12: Simulation results with $\lambda = 0.01$. $\pi_0 = 0.9$ except (d).



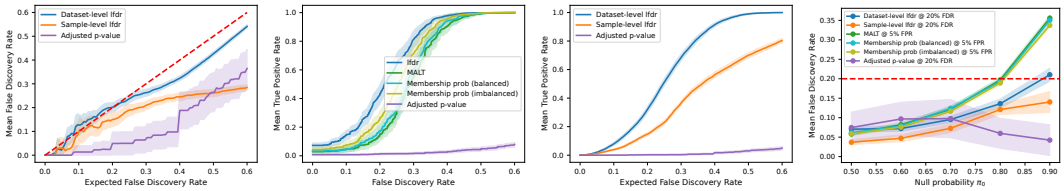
(a) (b) (c) (d)

Figure 13: Simulation results with $\lambda = 0.1$. $\pi_0 = 0.9$ except (d).



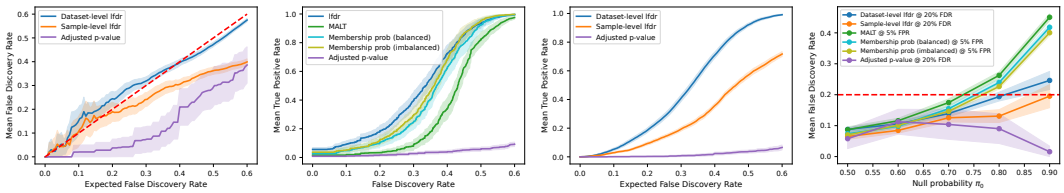
(a) (b) (c) (d)

Figure 14: Simulation results with $\lambda = 1$. $\pi_0 = 0.9$ except (d).



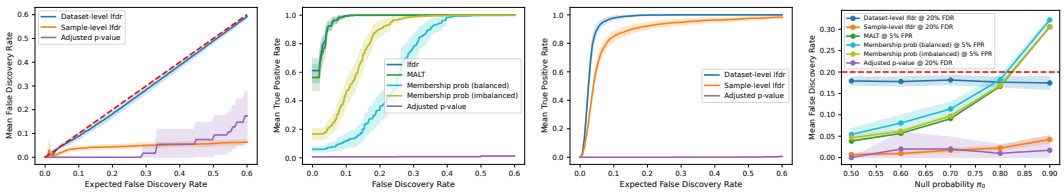
(a) (b) (c) (d)

Figure 15: Simulation results with $\lambda = 5$. $\pi_0 = 0.9$ except (d).



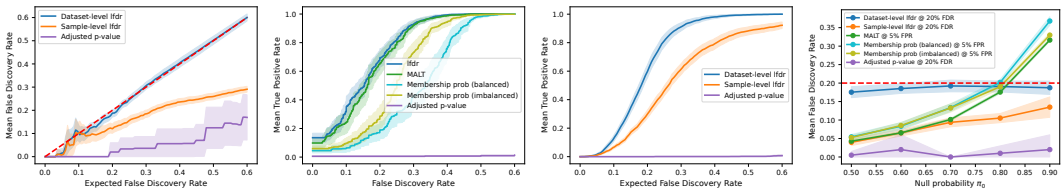
(a) (b) (c) (d)

Figure 16: Simulation results with $\lambda = 10$. $\pi_0 = 0.9$ except (d).



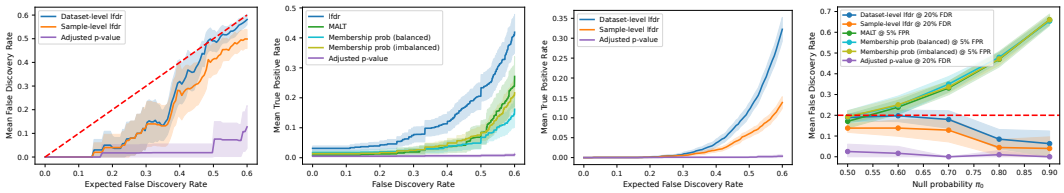
(a) (b) (c) (d)

Figure 17: Real-world data experiment with $\lambda = 0.01$. $\pi_0 = 0.9$ except (d).



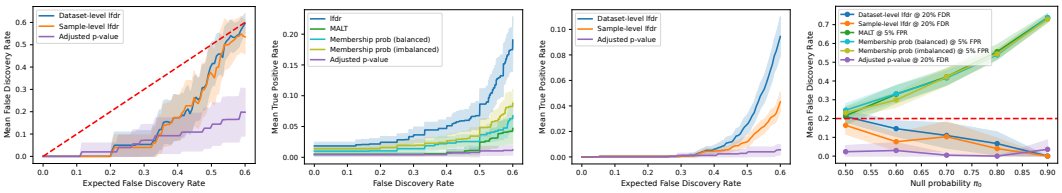
(a) (b) (c) (d)

Figure 18: Real-world data experiment with $\lambda = 0.1$. $\pi_0 = 0.9$ except (d).



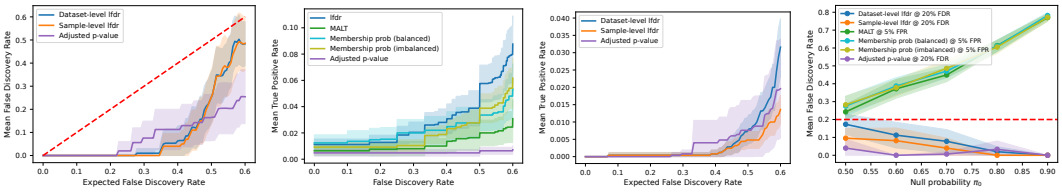
(a) (b) (c) (d)

Figure 19: Real-world data experiment with $\lambda = 2$. $\pi_0 = 0.9$ except (d).



(a) (b) (c) (d)

Figure 20: Real-world data experiment with $\lambda = 5$. $\pi_0 = 0.9$ except (d).



(a) (b) (c) (d)

Figure 21: Real-world data experiment with $\lambda = 10$. $\pi_0 = 0.9$ except (d).

F Compute resources

All experiments were conducted using Google Colaboratory environment, where CPU is AMD EPYC 7B12.

G Societal impacts

This work studies membership inference attacks from the perspective of false discovery rate control. A positive societal impact of this framework is that it can support more reliable privacy auditing: by controlling the expected proportion of false discoveries among declared members, the proposed methods can help practitioners assess membership leakage in a statistically calibrated manner. At the same time, the proposed method can also have negative societal impacts if misused, since it allows more reliable attacks. Our work is intended as an auditing and risk-assessment tool, and highlights the need for appropriate safeguards, including regularization, when deploying models trained on sensitive data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1 provides the contribution of our work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 suggests the limitation and future work of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Appendix A provides the proofs of the propositions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 and codes can reproduce our experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We disclose our codes in Supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and codes in the Supplementary material shows the detail of experimental setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We plot a 2-sigma confidence interval in each figure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information of the used compute resources in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper conforms to the NeurIPS Code of Ethics. The work develops and evaluates a statistical methodology and does not involve human subjects, personally identifiable information, sensitive data, crowdsourced data collection, or high-risk deployment scenarios. The experiments use synthetic or publicly available datasets for methodological evaluation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes a Societal Impact discussion in Section G. It describes the positive impact of FDR-controlled membership inference as a tool for more reliable privacy auditing and risk assessment, while also noting the potential negative impact that better-calibrated attacks could be misused to infer whether specific individuals were included in a training dataset.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release high-risk data or models. Although membership inference attacks are dual-use and may pose privacy risks, our experiments are conducted in controlled settings for privacy-risk assessment and mitigation. Any released code is intended only for reproducing the reported experiments, and we do not release tools targeting private users, confidential datasets, or deployed systems.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use UCI Communities and Crime dataset [20], which has CC BY 4.0 license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce or release any new assets, such as datasets, benchmarks, software packages, or pretrained models. All experiments are conducted using existing datasets or synthetically generated data described in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM was used solely to improve the clarity and readability of the writing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.