14 Membership Inference Attacks

Membership inference attacks [Hom+08], also known as *tracing* or *fingerprinting* attacks, play a major role in the theory and practice of privacy in statistics and machine learning. In the simplest setting, a curator holds a dataset S consisting of records for individuals from some underlying population. The curator publishes some summary $\hat{\theta} = \mathcal{A}(S)$, such as a predictive or generative model trained on S, marginal tables for S, or synthetic data that mimics statistical properties of S. A membership inference attacker, given a data record Z and the published output $\hat{\theta}$, aims to determine if Z is a member of the dataset S or, alternatively, a random member of the population.

From the attacker's point of view, membership inference can be modeled as a hypothesis testing problem (Figure 8). We model the population as a probability distribution $\mathcal{D}$, and model the sample $S = \{X_1, \ldots, X_n\}$ as consisting of n records drawn independently and identically from $\mathcal{D}$. The target record Z is chosen in one of two ways: In the IN scenario, Z is a uniformly random record from S. In the OUT scenario, $Z = X_0$ is an independent sample from $\mathcal{D}$. The attacker (or tester) must distinguish these two possibilities. The attack's *false positive rate* (FPR) is the probability that it outputs "IN" in the OUT scenario, and the *true positive rate* (TPR) is the probability of saying "IN" in the IN scenario.

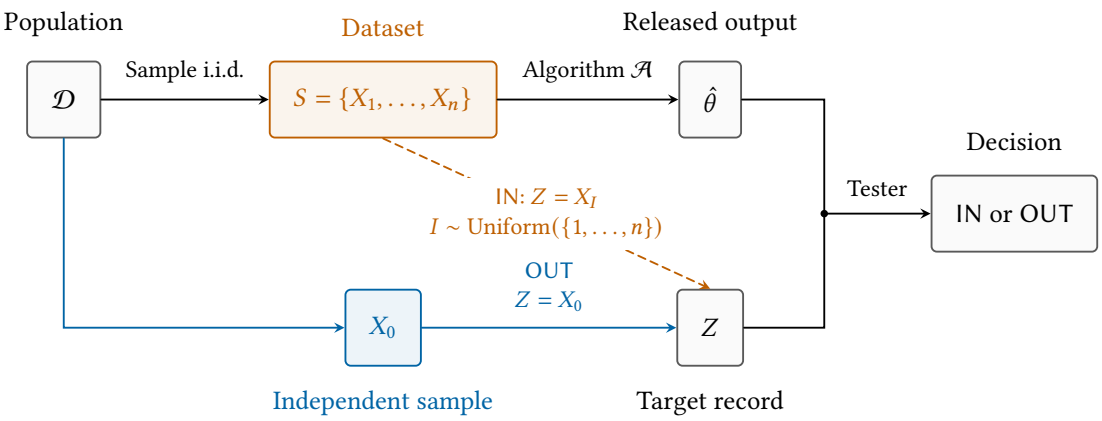


Figure 8: The basic membership inference setting. The tester observes the released output $\hat{\theta}$ and a target record Z. In the IN scenario, $Z = X_I$ for a uniformly random $I \in \{1, \ldots, n\}$; in the OUT scenario, $Z = X_0$ is drawn independently from $\mathcal{D}$. The two paths into Z represent alternative scenarios.

One might ask why membership inference is a privacy problem at all, especially given that the attacker already has the data record Z. First, the data may be split into groups based on some sensitive attribute, and learning membership in one of the groups is a clear privacy violation. For example, in the original work on membership inference [Hom+08, SOJH09], the data was split into a case group and a control group for a medical study, and membership in the case group indicates having a rare medical condition. Second, membership inference is the simplest meaningful privacy attack, and thus the easiest and most broadly applicable way to identify potential privacy issues that indicate more serious problems such as data reconstruction or training data extraction. As a result, membership inference is often used as a diagnostic tool for training data leakage in machine learning, either as an end goal [SSSS17] or to provide empirical lower bounds on the privacy parameters for differential privacy.

14.1 The Simplest Membership Inference Attack

We will start with a very simple membership inference attack against the empirical mean of Gaussian data. The attack is particularly effective in the overparameterized setting where the dimension is much larger than the sample size.

Let $\mathcal{D} = \mathcal{N}(\mu, I_d)$ for some $\mu \in \mathbb{R}^d$, and draw $S = (X_1, \ldots, X_n)$ independently from $\mathcal{D}$. The algorithm releases the empirical mean

$$\bar{X} = \mathcal{A}(S) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We assume that the attacker knows the population mean μ. Given a target record Z and the released mean $\bar{X}$, the attacker computes the score

$$T_{\mathcal{D}}(Z; \bar{X}) = \left\langle \bar{X} - \mu \,,\; Z - \mu \right\rangle.$$

The idea is that a member of the sample pulls $\bar{X}$ toward itself, making this inner product positive. An independent target has no such influence. The attacker outputs IN when the score exceeds a threshold τ, and OUT otherwise.

Theorem 14.1 (Gaussian membership inference from the empirical mean)**.** *Let $\bar{X}$ be the empirical mean of n independent samples from $\mathcal{D} = \mathcal{N}(\mu, I_d)$. If, for a sufficiently large universal constant C,*

$$d \geq Cn,$$

then there is a membership inference attack based on thresholding the score

$$T_{\mathcal{D}}(Z; \bar{X}) = \left\langle \bar{X} - \mu \,,\; Z - \mu \right\rangle$$

at $\tau = d/2n$. The attack has false positive rate at most $1/10$ and true positive rate at least $9/10$, for every population mean μ.

To visualize this effect, fix a target value $Z = z$. In the OUT case, knowing Z does not change the distribution of $\bar{X}$. In the IN case, we may take $Z = X_1$ by symmetry, so fixing $Z = z$ fixes one term in the empirical mean. Thus,

$$\text{OUT:} \quad \bar{X} \mid Z = z \sim \mathcal{N}\left(\mu, \frac{1}{n} I_d\right),$$
$$\text{IN:} \quad \bar{X} \mid X_1 = z \sim \mathcal{N}\left(\mu + \frac{z-\mu}{n}, \frac{n-1}{n^2} I_d\right).$$

Conditioning on membership shifts the mean by $(z-\mu)/n$ and barely changes the covariance when n is large (Figure 9). Without conditioning on the target, $\bar{X} \sim \mathcal{N}(\mu, I_d/n)$ in both cases; the membership signal comes from the dependence between Z and $\bar{X}$.

The OUT case. Let $Z \sim \mathcal{D}$ be independent of S. Conditional on $\bar{X}$, the vector $Z - \mu \sim \mathcal{N}(0, I_d)$ is independent of $\bar{X} - \mu$, so

$$T_{\mathcal{D}}(Z; \bar{X}) \mid \bar{X} \sim \mathcal{N}\left(0, \|\bar{X} - \mu\|_2^2\right).$$

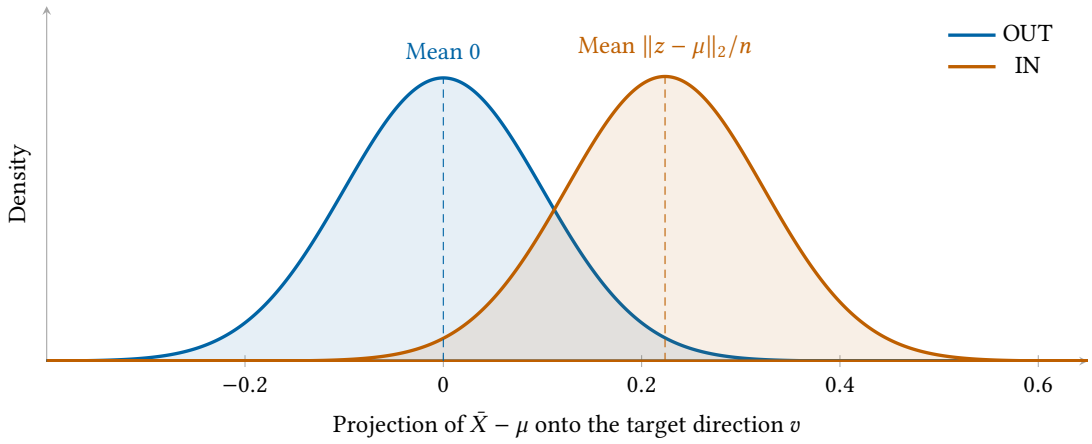


Figure 9: The distribution of the empirical mean projected onto the direction $v = (z-\mu)/\|z-\mu\|_2$, for a fixed target z. Here $n = 100$, $d = 500$, and $\|z-\mu\|_2 = \sqrt{d}$. In the OUT case, $\langle \bar{X}-\mu\,,\ v\rangle$ is centered at 0; conditioning on $X_1 = z$ in the IN case shifts its mean to $\sqrt{d}/n$, with almost the same variance.

In particular, $\mathbb{E}[T_{\mathcal{D}}(Z;\bar{X}) \mid \bar{X}] = 0$, and hence

$$\mathop{\mathbb{E}}_{\text{OUT}}[T_{\mathcal{D}}(Z;\bar{X})] = 0.$$

Using the conditional second moment and the fact that $\bar{X}-\mu \sim \mathcal{N}(0, I_d/n)$, we also obtain

$$\begin{aligned}\mathop{\text{Var}}_{\text{OUT}}(T_{\mathcal{D}}(Z;\bar{X})) &= \mathbb{E}\big[T_{\mathcal{D}}(Z;\bar{X})^2\big] \\ &= \mathbb{E}\big[\|\bar{X}-\mu\|_2^2\big] = \sum_{j=1}^{d}\mathbb{E}\big[(\bar{X}_j-\mu_j)^2\big] = \frac{d}{n}.\end{aligned}$$

Thus the OUT score is centered at 0 and has standard deviation $\sqrt{d/n}$.

The IN case. By symmetry, every choice of member has the same score distribution, so take $Z = X_1$. Expanding the empirical mean gives

$$T_{\mathcal{D}}(X_1;\bar{X}) = \frac{1}{n}\|X_1-\mu\|_2^2 + \frac{1}{n}\sum_{i=2}^{n}\langle X_i-\mu\,,\ X_1-\mu\rangle.$$

The first term is the target's contribution to its own score. Each term in the sum has mean zero because $X_i-\mu$ and $X_1-\mu$ are independent and centered. Therefore,

$$\mathop{\mathbb{E}}_{\text{IN}}[T_{\mathcal{D}}(X_1;\bar{X})] = \frac{1}{n}\,\mathbb{E}\big[\|X_1-\mu\|_2^2\big] = \frac{d}{n}.$$

The variance is

$$\mathop{\text{Var}}_{\text{IN}}(T_{\mathcal{D}}(X_1;\bar{X})) = \frac{d(n+1)}{n^2} = \left(1+\frac{1}{n}\right)\frac{d}{n},$$

which is almost identical to the OUT variance when n is large. To see the two contributions, the squared-norm term has variance $2d/n^2$, the sum of cross terms has variance $(n-1)d/n^2$, and these two terms have covariance zero.

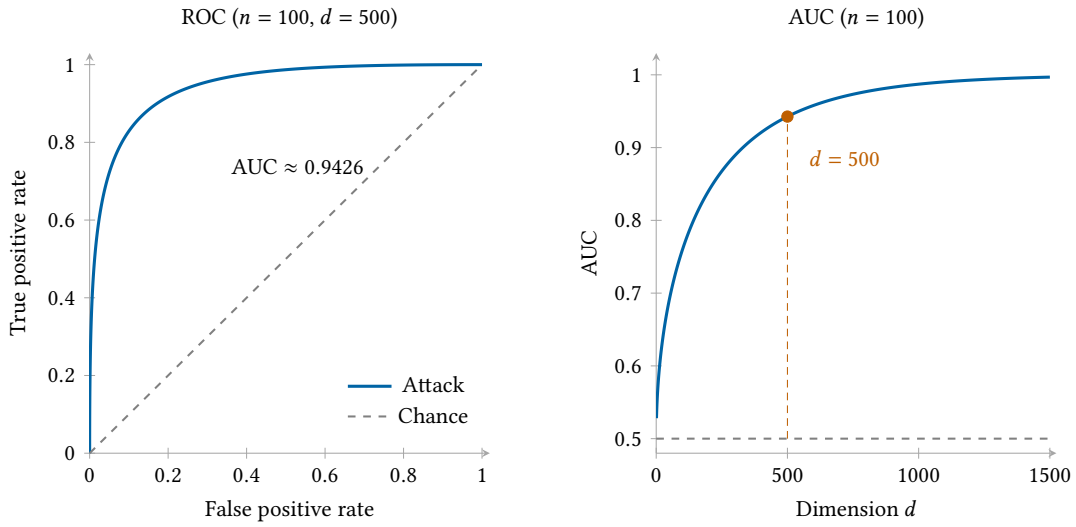


Figure 10: Attack performance using the Gaussian approximations $T_{\mathsf{OUT}} \sim \mathcal{N}(0, d/n)$ and $T_{\mathsf{IN}} \sim \mathcal{N}(d/n, d(n+1)/n^2)$. Left: the ROC curve for $n = 100$ and $d = 500$ plots the true positive rate against the false positive rate, with AUC approximately 0.9426. Right: AUC increases with d for fixed $n = 100$; the orange point marks $d = 500$. Under these approximations, $\text{AUC} = \Phi\left(\sqrt{d/(2n+1)}\right)$, where Φ is the standard Gaussian cumulative distribution function. Dashed gray lines indicate chance performance.

Distinguishing IN from OUT. Figure 10 shows the ROC curve and area under the curve (AUC) using Gaussian approximations.

14.2 Robust Membership Inference Attacks

In practical use cases, membership inference attacks are used to evaluate the privacy of a specific algorithm $\mathcal{A}$. However, in theory membership inference attacks can have a remarkable feature called *robustness*. An attack is robust if it succeeds for any mechanism $\mathcal{A}$ that accomplishes some task with sufficient level of accuracy. The notion of accuracy varies by task—for example, for training a predictive model it might mean classification error at most α, or for estimating the mean of the population it might mean error at most α in Euclidean norm. But robust membership inference attacks have a compelling practical motivation, because they are attacks that cannot be easily thwarted by simple tweaks.

Gaucher et al. [GLSUZ26] show that the Gaussian example above extends to arbitrary accurate mean estimators. Let $S = (X_1, \ldots, X_n)$ consist of independent samples from $\mathcal{D} = \mathcal{N}(\mu, I_d)$, and let $\hat{\mu} = \mathcal{A}(S)$ be the released estimate. The algorithm may add noise or modify the empirical mean in any way. The population mean μ is unknown to the algorithm but known to the attacker.

Theorem 14.2 (Robust Gaussian membership inference, informal)**.** *Suppose $\mathcal{A}$ estimates every population mean $\mu \in \mathbb{R}^d$ accurately, with $\|\hat{\mu} - \mu\|_2 \leq \alpha\sqrt{d}$, so the root mean squared error per coordinate is at most α. If, for a sufficiently large universal constant C,*

$$d \gtrsim C\big(n + \alpha^2 n^2\big),$$

then there is a membership inference attack based on thresholding the same score,

$$T_{\mathcal{D}}(Z; \hat{\mu}) = \langle \hat{\mu} - \mu\,,\; Z - \mu \rangle.$$

The attack has false positive rate at most $1/10n$ and, for some population mean μ, correctly identifies a member of S with constant probability.

Ignoring logarithmic factors, the key regime is $d \gg n$ and $\alpha \ll \sqrt{d}/n$. The conclusion holds for every sufficiently accurate algorithm $\mathcal{A}$, regardless of how it introduces noise. Thus, even a noisy estimate can reveal membership if its error is small enough.