Do Parameters Reveal More than Loss for Membership Inference?

Anonymous authors
Paper under double-blind review

Abstract

Membership inference attacks aim to infer whether an individual record was used to train a model. They are used as a key tool for disclosure auditing. While such evaluations are useful to demonstrate risk, they are computationally expensive and often make strong assumptions about potential adversaries' access to models and training environments, and thus do not provide very tight bounds on leakage from potential attacks. We show how prior claims around black-box access being sufficient for optimal membership inference do not hold for stochastic gradient descent, and that optimal membership inference indeed requires white-box access. Our theoretical results suggest a new white-box inference attack *IHA* (**I**nverse **H**essian **A**ttack) that explicitly uses model parameters by taking advantage of computing inverse-Hessian vector products. Our results show that both auditors and adversaries may be able to benefit from access to model parameters, and we advocate for further research into white-box methods for membership privacy auditing.

1 Introduction

Models produced by using machine learning on private training data can leak sensitive information about data used to train or tune the model (Salem et al., 2023). Researchers study these privacy risks by either designing and evaluating attacks to simulate what motivated adversaries may be able to infer in particular settings or by developing privacy methods that can provide strong guarantees, often based on some notion of differential privacy (Dwork et al., 2006), that provide a bound on information disclosure from any attack. Although both developing attacks and formal privacy proofs are important, conducting meaningful privacy audits is different from both approaches. Empirical methods, usually in the form of attack simulations, are inherently limited by the attacks considered and the uncertainty about the possibility of better attacks, while theoretical proofs require many assumptions or result in loose bounds, and any claims based on theoretical results depend on careful analysis that the system as implemented is consistent with the theory. If there is a theoretical result that prescribes an optimal attack, then empirical results with that attack (or approximations of the attack) can offer a more meaningful bound on privacy risk than is possible with theory or experiments alone. While the theory needs to cover all data distributions, experiments with the optimal attack focus on the actual distribution and given model, resulting in tighter and more relevant privacy evaluations.

Privacy audits can also be important in adversarial contexts, where a regulator or external advocate conducts them to test a released model. Auditors with elevated model access (via associated training environments, data, etc.) may be able to take advantage of more information to produce better estimates of what an adversary may be able to do without that information. Auditing is also orthogonal to proofs that establish differential privacy bounds or other privacy notions. As outlined by Cummings et al. (2023), theoretical bounds may be "too conservative or inaccurate in some settings", and it may not always be possible to come up with proofs or theoretical bounds (which usually only apply to membership inference) that ensure models do not "violate disclosure requirements in ways that are not captured by differential privacy". Empirical auditing can provide a more meaningful measure of privacy leakage for these situations.

The most common disclosure auditing approach today is to conduct membership inference attacks (Kumar & Shokri, 2020) and related attacks that attempt to extract specific data (Cummings et al., 2023). While

membership inference assumes the adversary already knows the full candidate record, it may still constitute a direct privacy risk when revealing the inclusion of a known record in the training data itself, which leaks sensitive information. In most scenarios, however, membership disclosure by itself is not a serious privacy risk, but rather used as a proxy for understanding information leakage that may result in more serious privacy violations. Membership inference is simple to define, relatively easy to measure, and aligns well with differential privacy. This has resulted in it being widely used as a method for auditing disclosure risks for machine learning (Kumar & Shokri, 2020; Yeom et al., 2020; Kazmi et al., 2024; Azize & Basu, 2024).

Prior results on membership inference attacks have largely focused on the *black-box* setting, where the attacker only has input–output access to the target model. This focus has been reinforced by folklore and results demonstrating negligible gains from parameter access (known as *white-box* attacks) (Nasr et al., 2018; Carlini et al., 2022). A well-known theoretical result by Sablayrolles et al. (2019) proves that black-box access is sufficient for optimal membership inference (under certain conditions). This result has been the basis of several subsequent works (Ye et al., 2022; Chaudhari et al., 2024). However, the assumptions made in its derivation do not hold for most models, including ones trained with stochastic gradient descent (SGD).

Contributions. Utilizing recent advances in the discrete-time SGD-dynamics literature (Liu et al., 2021; Ziyin et al., 2021), we provide a more accurate formulation of the optimal membership inference attack that that demonstrates the limitations of the results from Sablayrolles et al. (2019) which claimed black-box access is sufficient (Section 3). Our theoretical result also prescribes such an attack, which we call the Inverse Hessian Attack (IHA) (Section 3.3). We empirically demonstrate its effectiveness in simple settings (Section 4).

2 Preliminaries

The section provides background on membership inference (Section 2.1) and SGD dynamics (Section 2.2).

2.1 Membership Inference

Following the framework established by Sablayrolles et al. (2019), let $\mathcal{D}$ be a data distribution from which n records $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ are i.i.d. sampled with $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ being the i-th record. Let $\boldsymbol{w} \in \mathbb{R}^d$ be the model parameters produced by some machine learning algorithm on a training dataset D. Assume $m_1, m_2, \ldots, m_n$ follow a Bernoulli distribution with $\gamma = \mathbb{P}(m_i = 1)$, where m_i is the membership indicator of $\boldsymbol{z}_i$ (i.e., $m_i = 1$ if $\boldsymbol{z}_i \in D$, and $m_i = 0$ otherwise). Given $\boldsymbol{w}$, a *membership inference attack* aims to predict the unknown membership m_i for any given record $\boldsymbol{z}_i$.

Definition 2.1 (Membership Inference)**.** Let $\boldsymbol{w}$ be the parameters of the target model and $\boldsymbol{z}_1$ be a record. Inferring the membership of $\boldsymbol{z}_1$ to the training set of $\boldsymbol{w}$ is equivalent to computing:

$$\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1) = \mathbb{P}(m_1 = 1 \mid \boldsymbol{w}, \boldsymbol{z}_1).$$

Let $\mathbb{P}(\boldsymbol{w} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, m_1, \ldots, m_n)$ be the posterior distribution of model parameters produced by some randomized machine learning algorithm (i.e., stochastic gradient descent). Applying Bayes' theorem, Sablayrolles et al. (2019) derived the following explicit formula for $\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1)$:

Lemma 2.1 (Sablayrolles et al. (2019))**.** Let $\mathcal{T} = \{\boldsymbol{z}_2, \ldots, \boldsymbol{z}_n, m_2, \ldots, m_n\}$. Given model parameters $\boldsymbol{w}$ and a record $\boldsymbol{z}_1$, the optimal membership inference is given by:

$$\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1) = \mathbb{E}_{\mathcal{T}}\left[\sigma\left(\ln\left(\frac{p(\boldsymbol{w} \mid m_1 = 1, \boldsymbol{z}_1, \mathcal{T})}{p(\boldsymbol{w} \mid m_1 = 0, \boldsymbol{z}_1, \mathcal{T})}\right) + \ln\left(\frac{\gamma}{1-\gamma}\right)\right)\right], \tag{1}$$

where $\sigma(u) = (1 + \exp(-u))^{-1}$ is the Sigmoid function, and $\gamma = \mathbb{P}(m_1 = 1)$.

To use Lemma 2.1, one needs to characterize the posterior, $\mathbb{P}(\boldsymbol{w} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, m_1, \ldots, m_n)$, to make explicit the effect of the inferred record $\boldsymbol{z}_1$ on the optimal membership inference $\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1)$. Recent advances in discrete-time SGD dynamics (Liu et al., 2021; Ziyin et al., 2021) literature can help provide a connection between the posterior and model parameters.

2.2 Discrete-time SGD Dynamics

A line of theoretical work (Welling & Teh, 2011; Sato & Nakagawa, 2014; Stephan et al., 2017; Liu et al., 2021; Ziyin et al., 2021) has analyzed the continuous- and discrete-time dynamics of stochastic gradient methods and provided insights for understanding deep learning generalization. Let $L_{\text{tot}}(\boldsymbol{w}) = L(\boldsymbol{w}) + \frac{\alpha}{2}\|\boldsymbol{w}\|_2^2$ be the ℓ_2-regularized total loss that we aim to optimize, where $\alpha \geq 0$ denotes the hyperparameter that controls the regularization strength. Consider an SGD algorithm with the following update rule (for $t = 1, 2, 3, \ldots$):

$$\begin{cases} \boldsymbol{g}_t &= \nabla L_{\text{tot}}(\boldsymbol{w}_{t-1}) + \boldsymbol{\eta}_{t-1}; \\ \boldsymbol{h}_t &= \mu \boldsymbol{h}_{t-1} + \boldsymbol{g}_t; \\ \boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \lambda \boldsymbol{h}_t. \end{cases} \tag{2}$$

Here, $\mu \in [0, 1)$ is the momentum, $\lambda > 0$ is the learning rate, and

$$\boldsymbol{\eta}_t = \frac{1}{S}\sum_{i \in \mathcal{B}_t} (\nabla \ell(\boldsymbol{w}_t, \boldsymbol{z}_i) + \alpha \boldsymbol{w}_t) - \nabla L_{\text{tot}}(\boldsymbol{w}_t) = \frac{1}{S}\sum_{i \in \mathcal{B}_t} \nabla \ell(\boldsymbol{w}_t, \boldsymbol{z}_i) - \nabla L(\boldsymbol{w}_t)$$

represents the unbiased mini-batch noise, where $\mathcal{B}_t$ is a randomly sampled batch of examples with size S from the training dataset D, and $L(\boldsymbol{w}) = \frac{1}{n}\sum_{\boldsymbol{z} \in D} \ell(\boldsymbol{w}, \boldsymbol{z})$.

Assuming a model is trained using SGD according to the update rule defined by Equation 2 on a *quadratic loss* and arrives at a *stationary state*, Liu et al. (2021) established a theoretical connection between the Hessian matrix $\mathbf{H}$, the asymptotic noise covariance $\mathbf{C} = \lim_{t \to \infty} \mathbb{E}_{\boldsymbol{w}_t}[\text{cov}(\boldsymbol{\eta}_t, \boldsymbol{\eta}_t)]$, and the asymptotic model fluctuation $\boldsymbol{\Sigma} = \lim_{t \to \infty} \text{cov}(\boldsymbol{w}_t, \boldsymbol{w}_t)$. To be more specific, we first lay out the two imposed assumptions.

Assumption 1 (Quadratic Loss)**.** The total loss function $L_{\text{tot}}(\boldsymbol{w})$ is either globally quadratic or locally quadratic close to a local minimum $\boldsymbol{w}^*$. Specifically, the loss function can be approximated as:

$$L_{\text{tot}}(\boldsymbol{w}) = L_{\text{tot}}(\boldsymbol{w}^*) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^\top (\mathbf{H}(\boldsymbol{w}^*) + \alpha \mathbf{I}_d)(\boldsymbol{w} - \boldsymbol{w}^*) + o(\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2), \tag{3}$$

where $\boldsymbol{w}^*$ is a local minimum, $\mathbf{H}(\boldsymbol{w}^*)$ denotes the Hessian matrix at $\boldsymbol{w}^*$ with respect to the unregularized loss function $L(\boldsymbol{w})$, and $\mathbf{I}_d$ stands for the $d \times d$ identity matrix.

Assumption 2 (Stationary-State)**.** After a sufficient number of iterations, models trained with SGD defined by Equation 2 arrive at a stationary state, i.e., the asymptotic model fluctuation $\boldsymbol{\Sigma}$ exists and is finite.

Under the above assumptions, Liu et al. (2021) proved the following theorem that describes model fluctuations of discrete SGD in a quadratic potential with connections to the Hessian matrix and the noise covariance:

Theorem 2.2 (SGD Stationary distribution with momentum)**.** Let $\boldsymbol{w}$ be updated with SGD defined by the update rule in Equation 2 with momentum $\mu \in [0, 1)$. Under Assumptions 1 and 2, if we additionally suppose $\mathbf{C}$ commutes with $\mathbf{H}(\boldsymbol{w}^*)$, then the asymptotic model fluctuation satisfies:

$$\boldsymbol{\Sigma} = \left[\frac{\lambda\big(\mathbf{H}(\boldsymbol{w}^*) + \alpha \mathbf{I}_d\big)}{1+\mu} \cdot \left(2\mathbf{I}_d - \frac{\lambda\big(\mathbf{H}(\boldsymbol{w}^*) + \alpha \mathbf{I}_d\big)}{1+\mu}\right)\right]^{-1} \frac{\lambda^2 \mathbf{C}}{1-\mu^2}.$$

Theorem 2.2 requires the existence of a finite stationary noise covariance and the loss function to be quadratic close to a local minimum, which are both mild assumptions (see Liu et al. (2021) for detailed discussions).

In a follow-up work, Ziyin et al. (2021) further derived the explicit dependence of the finite stationary noise covariance $\mathbf{C}$ to the loss and Hessian around a local minimum $\boldsymbol{w}^*$ under certain assumptions:

Theorem 2.3 (SGD Noise Covariance)**.** Let $L_{\text{tot}}(\boldsymbol{w}) = L(\boldsymbol{w}) + \frac{\alpha}{2}\|\boldsymbol{w}\|_2^2$ be the total loss with $\alpha \geq 0$. Assume the model $\boldsymbol{w}$ is optimized with SGD defined by Equation 2 around a local minimum $\boldsymbol{w}^*$. If $L(\boldsymbol{w}^*) \neq 0$, then

$$\mathbf{C} = \frac{2L(\boldsymbol{w}^*)}{S}\mathbf{H}(\boldsymbol{w}^*) - \frac{\alpha^2}{S}\boldsymbol{w}^*\boldsymbol{w}^{*\top} + O(S^{-2}) + O(\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2) + o(L(\boldsymbol{w}^*)),$$

provided that $\boldsymbol{\Sigma}$ is proportional to S^{-1} and $|L(\boldsymbol{w}) - \ell(\boldsymbol{w}, \boldsymbol{z}_i)|$ is small (i.e., of order $o(L(\boldsymbol{w}))$).

The first imposed assumption of $\mathbf{\Sigma} = O(S^{-1})$ has been justified by existing works (Liu et al., 2021; Xie et al., 2021; Mori et al., 2022), while the second assumption assumes that the current total training loss $L(\boldsymbol{w})$ approximates the individual loss for each record $\ell(\boldsymbol{w}, \boldsymbol{z}_i)$ well. Also, note that Theorem 2.3 directly implies that the SGD noise covariance $\mathbf{C}$ commutes with the Hessian matrix $\mathbf{H}(\boldsymbol{w}^*)$.

Based on the above two theorems and only considering the leading term in the noise covariance, we can immediately derive the following formula for the stationary model fluctuation of SGD:

$$\mathbf{\Sigma} = \frac{\lambda}{S(1-\mu)} \left(2L(\boldsymbol{w}^*)\mathbf{H}(\boldsymbol{w}^*) - \alpha^2 \boldsymbol{w}^* \boldsymbol{w}^{*\top} \right) \left(\mathbf{H}(\boldsymbol{w}^*) + \alpha \mathbf{I}_d \right)^{-1} \left(2\mathbf{I}_d - \frac{\lambda}{1+\mu} \big(\mathbf{H}(\boldsymbol{w}^*) + \alpha \mathbf{I}_d \big) \right)^{-1}. \tag{4}$$

We remark that if $L(\boldsymbol{w}^*) = 0$ (i.e., $\boldsymbol{w}^*$ is a global minimum), then $\mathbf{\Sigma} = \mathbf{0}$. In addition, if the Hessian matrix $(\mathbf{H}(\boldsymbol{w}^*) + \alpha \mathbf{I}_d)$ has degenerate rank $r < d$, then $(\mathbf{H}(\boldsymbol{w}^*) + \alpha \mathbf{I}_d)^{-1}$ can be replaced by the corresponding Moore-Penrose pseudo inverse. Accordingly, similar results to Equation 4 can be obtained by considering the projection space spanned by eigenvectors with non-zero eigenvalues. Section 5 of Ziyin et al. (2021) provides more detailed discussions of the imposed assumptions and the implications of the results.

3 Black-Box Access is not Sufficient

In this section, we examine previous assertions concerning optimal membership inference (Section 3.1) and show, for models trained with SGD, that optimal membership inference does require parameter access (Section 3.2). Our theory directly implies an attack (Section 3.3).

3.1 Limitations of Claims of Black-Box Optimality

Sablayrolles et al. (2019) proved the optimality of black-box membership inference under a Bayesian framework. They assume (Equation 1 in Sablayrolles et al. (2019)) that the posterior distribution of model parameters $\boldsymbol{w}$ trained on $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ with membership $m_1, \ldots, m_n$ follows:

$$\mathbb{P}(\boldsymbol{w} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \propto \exp\left(-\frac{1}{T} \sum_{i=1}^{n} m_i \cdot \ell(\boldsymbol{w}, \boldsymbol{z}_i) \right), \tag{5}$$

where T is a temperature parameter that captures the stochasticity of the learning algorithm. This assumption makes subsequent derivations of optimal membership inference much easier, but oversimplifies the training dynamics of typical machine learning algorithms such as SGD. Equation 5 assumes that the posterior of $\boldsymbol{w}$ follows a Boltzmann distribution that only depends on the training loss. This is desirable for Bayesian posterior inference, where the goal is to provide a sampling strategy for an unknown data distribution given a set of observed data samples. This can be achieved using stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011) with shrinking step size λ_t (i.e., $\lim_{t\to\infty} \lambda_t = 0$) and by injecting carefully-designed Gaussian noise $\mathcal{N}(\mathbf{0}, \lambda_t \cdot \mathbf{I}_D)$. However, this special SGLD design differs from the common practice of SGD algorithms used to train neural networks for the following two reasons:

1. All analyses are performed under continuous-time dynamics whereas actual SGD is performed with discrete steps. While related work such as Stephan et al. (2017) cast the continuous-time dynamics of SGD as a multivariate *Ornstein-Uhlenbeck* process (similar to SGLD) whose stationary distribution is proven to be Gaussian (Equations 11 and 12 in Stephan et al. (2017)), they make additional assumptions such as the noise covariance matrix being independent of model parameters.

2. SGLD assumes a vanishing learning rate until convergence, whereas SGD is performed with a non-vanishing step size and for a finite number of iterations in practice. The learning rate of SGD is often large, which can cause model dynamics to drift even further from SGLD (Ziyin et al., 2023), especially in the discrete-time setting (Liu et al., 2021).

We thus characterize the analytical form of the posterior distribution with respect to model parameters trained with SGD:

Theorem 3.1 (Posterior for SGD)**.** Assume the same assumptions as used in Theorems 2.2 and 2.3. Let $\boldsymbol{w}^*$ be the local minimum that SGD (Equation 2) is converging towards. Then, the (conditional) log-probability of observing parameters $\boldsymbol{w}$ is given by (up to constants and negligible terms):

$$-\frac{d}{2}\ln L_* + \sum_{i=1}^{d}\ln\left(\frac{\left(2-\frac{\lambda}{1+\mu}(\sigma_i(\mathbf{H}_*)+\alpha)\right)(\sigma_i(\mathbf{H}_*)+\alpha)}{\sigma_i(\mathbf{H}_*)}\right) - \frac{S(1-\mu)}{2\lambda}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\cdot\frac{\|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2}{L_*}$$
$$-\frac{S(1-\mu)\alpha}{4\lambda}\cdot\left(2-\frac{\lambda\alpha}{1+\mu}\right)\cdot\frac{\nabla L(\boldsymbol{w})^\top\mathbf{H}_*^{-3}\nabla L(\boldsymbol{w})}{L_*} + \frac{S(1-\mu)}{2(1+\mu)}\cdot\frac{L(\boldsymbol{w})}{L_*},$$

where $L_* = L(\boldsymbol{w}^*)$, $\mathbf{H}_* = \mathbf{H}(\boldsymbol{w}^*)$ and $\sigma_i(\mathbf{H}_*)$ denotes the i-th largest eigenvalue of $\mathbf{H}_*$.

A proof for Theorem 3.1 is given in Appendix B. Theorem 3.1 suggests that the posterior distribution of model parameters learned by SGD not only relies on the training loss $L(\boldsymbol{w})$ but is also crucially dependent on terms, such as the Hessian structure $\mathbf{H}^*$, the gradient $\nabla L(\boldsymbol{w})$ and the ℓ_2 distance $\|\boldsymbol{w}-\boldsymbol{w}^*\|_2$, confirming that Equation 5 is insufficient to model the dynamics of a discrete-time SGD algorithm.

3.2 Optimal Membership Inference under Discrete-time SGD

So far, we have explained why the critical assumption imposed by Sablayrolles et al. (2019) about the posterior distribution of $\boldsymbol{w}$ following a Boltzmann distribution (Equation 5) does not hold for typical stochastic gradient methods employed in practice. We now prove a theorem that gives an estimate of the optimal scoring function for membership inference by leveraging the recent theoretical literature on discrete-time SGD dynamics (Liu et al., 2021; Ziyin et al., 2021). Our derivation is based on the assumption that the loss achieved at the local minimum remains unaffected by removing a single training record and that the Hessian structure remains unchanged.

Assumption 3 (Similarity at local minimum)**.** For any $\mathcal{T}$ and $\boldsymbol{z}_1$, let $L_0(\boldsymbol{w}) = \frac{1}{n}\sum_{i=2}^{n} m_i\ell(\boldsymbol{w},\boldsymbol{z}_i)$ and $L_1(\boldsymbol{w}) = \frac{1}{n}(\ell(\boldsymbol{w},\boldsymbol{z}_1) + \sum_{i=2}^{n} m_i\ell(\boldsymbol{w},\boldsymbol{z}_i))$. When the training dataset differs only by a single data point $\boldsymbol{z}_1$, assume that the Hessian matrix structure for models trained with and without the differing point share a similar structure, and the loss function also achieves a similar value at the local minimum:

$$\mathbf{H}_* = \mathbf{H}_1(\boldsymbol{w}_1^*) = \mathbf{H}_0(\boldsymbol{w}_0^*), \quad L_* = L_1(\boldsymbol{w}_1^*) = L_0(\boldsymbol{w}_0^*), \tag{6}$$

where $\boldsymbol{w}_1^*$ (resp. $\boldsymbol{w}_0^*$) is the local minimum that SGD with L_1 (resp. L_0) is converging towards, and $\mathbf{H}_1$ (resp. $\mathbf{H}_0$) denotes the Hessian matrix with respect to L_1 (resp. L_0).

As long as the size of the training dataset is sufficient and the excluded training record $\boldsymbol{z}_1$ is not a low-probability outlier from the data distribution $\mathcal{D}$, we expect Assumption 3 generally holds for SGD algorithms. Under Assumption 3 and a few other assumptions imposed in prior literature on discrete-time SGD dynamics (Liu et al., 2021; Ziyin et al., 2021), we obtain a theorem (proof is in Appendix C) that describes the scoring function for an optimal membership-inference adversary:

Theorem 3.2 (Optimal Membership-Inference Score)**.** Given $\boldsymbol{w}$ produced by an SGD algorithm defined by Equation 2 and a record $\boldsymbol{z}_1$, the optimal membership inference $\mathcal{M}(\boldsymbol{w},\boldsymbol{z}_1)$ is given by:

$$\mathbb{E}_{\mathcal{T}}\left[\sigma\left(\frac{S(1-\mu)}{2nL_*}\cdot\left(\frac{\ell(\boldsymbol{w},\boldsymbol{z}_1)}{1+\mu} - \frac{1}{\lambda}(I_1+I_2+I_3+I_4)\right) + \ln\left(\frac{\gamma}{1-\gamma}\right)\right)\right], \tag{7}$$

where I_1, I_2, I_3, and I_4 are defined as follows:

$$I_1 := \frac{1}{n}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\cdot\|\mathbf{H}_*^{-1}\nabla\ell(\boldsymbol{w},\boldsymbol{z}_1)\|^2,$$
$$I_2 := 2\left(1-\frac{\lambda\alpha}{1+\mu}\right)\cdot\left(\mathbf{H}_*^{-1}\nabla L_0(\boldsymbol{w})\right)^\top\left(\mathbf{H}_*^{-1}\nabla\ell(\boldsymbol{w},\boldsymbol{z}_1)\right),$$
$$I_3 := \frac{\alpha}{2n}\left(2-\frac{\lambda\alpha}{1+\mu}\right)\cdot\left(\mathbf{H}_*^{-1}\nabla\ell(\boldsymbol{w},\boldsymbol{z}_1)\right)^\top\left(\mathbf{H}_*^{-1}\left(\mathbf{H}_*^{-1}\nabla\ell(\boldsymbol{w},\boldsymbol{z}_1)\right)\right),$$
$$I_4 := \alpha\left(2-\frac{\lambda\alpha}{1+\mu}\right)\cdot\left(\mathbf{H}_*^{-1}\nabla L_0(\boldsymbol{w})\right)^\top\left(\mathbf{H}_*^{-1}\left(\mathbf{H}_*^{-1}\nabla\ell(\boldsymbol{w},\boldsymbol{z}_1)\right)\right).$$

Here, L_* and $\mathbf{H}_*$ are defined in Assumption 3, which are dependent on $\mathcal{T}$. Here, $\mathcal{T}$ refers to the set of both member and non-member records along with their corresponding membership indicators, as defined in Lemma 2.1. Note that computing the optimal score requires access to the Hessian and model gradients, both of which require access to model parameters. In fact, knowledge of the learning rate λ, momentum μ, and regularization parameter α are also required, thus requiring complete knowledge of the training setup of the target model. Thus, black-box access is ***not*** sufficient for optimal membership inference.

The first two additional terms I_1 and I_2 can be interpreted as the magnitude and direction, respectively, of a Newtonian step for the given record $\boldsymbol{z}_1$. The first term I_1 characterizes the influence magnitude in ℓ_2-norm of upweighting $\boldsymbol{z_1}$ on the model parameters close to the local minimum (Koh & Liang, 2017), while the second term captures the alignment between the influence of $\boldsymbol{z_1}$ and the averaged influence of the remaining training data. A larger *influence magnitude* of $\boldsymbol{z}_1$ or an increased *influence alignment* suggests a higher risk of membership inference. We remark that the notion of a self-influence function introduced in Cohen & Giryes (2024) naturally relates to I_1, suggesting a similar insight to ours that better membership inference attacks can be designed by leveraging the influence function of the inferred record. The last two additional terms, I_3 and I_4, originate from the extra L_2 regularization term imposed on the training loss of SGD (Section 2.2). When the regularization parameter α is a very small positive constant, the effects of I_3 and I_4 on optimal membership inference will be negligible, particularly compared with those of I_1 and I_2.

3.3 Inverse Hessian Attack

While Theorem 3.2 directly prescribes an optimal membership inference adversary, computing the expectation over all possible models trained using the rest of the training data is infeasible. Our definition of optimal membership inference corresponds to the true leakage of the model (as defined in Section 3.2 of (Ye et al., 2022)). It utilizes worst-case adversary knowledge (membership of all other training records) and white-box access to estimate the influence of the target record, similar to how empirical attacks such as LiRA (Carlini et al., 2022) and RMIA (Zarifzadeh et al., 2023) utilize reference models to account for atypical examples.

Specifically, making use of the insight of Theorem 3.2, we propose a scoring function based on the terms inside the expectation in Equation 7:

$$\mathrm{IHA}(\boldsymbol{z}_1) := \frac{\ell(\boldsymbol{w}, \boldsymbol{z}_1)}{1+\mu} - \frac{1}{\lambda}\big(I_1 + I_2 + I_3 + I_4\big). \tag{8}$$

This score $\mathrm{IHA}(\boldsymbol{z}_1)$, for some given record $\boldsymbol{z}_1$, can be used as the probability of $\boldsymbol{z}_1$ being a member and subsequently serve as a useful attack for privacy auditing, without having to train any reference models. Not having to train reference models offers significant advantages. It helps auditors avoid additional computational costs and, more importantly, eliminates the need for trainers to reserve hold-out data for reference model training. This is particularly beneficial when data availability is a constraint for privacy auditing methods relying on reference models. While the absence of a negative sign with the loss function (like in LOSS) in $\mathrm{IHA}(\boldsymbol{z}_1)$ may seem counter-intuitive at first glance, it can be rewritten such that it is proportional to the negation of the loss function (Appendix C.2).

While membership leakage is typically evaluated on a fixed dataset, the theoretical notion of optimal membership inference is defined for a much larger space. This space encompasses a broad distribution of possible data (including both member and non-member records) and the corresponding models trained on various splits of these datasets. In practice, it's challenging to realize this larger space, but if we could define true positive rates (TPRs) at low false positive rates (FPRs) with respect to this comprehensive space, they would empirically correspond to the optimal membership inference attack. This theoretical framework provides a more robust understanding of membership inference, though its practical implementation remains a significant challenge in privacy auditing.

Empirically, the performance of our audit is also influenced by other factors, such as how efficiently and accurately the inverse-Hessian vector products (iHVPs) can be computed and to what degree our assumptions hold (particularly Assumption 3, which requires the Hessian and loss at local minima being unaffected by the exclusion of a single datapoint).

4 Experiments

To evaluate IHA, we efficiently pre-compute $\nabla L_1(\boldsymbol{w})$ to facilitate the computation of $\nabla L_0(\boldsymbol{w})$ for any given target record $\boldsymbol{z}_1$. For accurate Hessian matrix computations, we address the issue of ill-conditioning due to near-zero and small negative eigenvalues by either damping or using low-rank approximations (damping-based conditioning seems to perform best; see Appendix E for details). To support larger models where direct Hessian computation is infeasible, we extend our method to use approximation methods based on Conjugate Gradients for iHVP computation (Koh & Liang, 2017). Our implementation for reproducing all the experiments is available as open-source code at `https://anonymous.4open.science/r/auditingmi-D16B/`.

Section 4.1 describes the baseline attacks, datasets, and models we use for our experiments. Our results are summarized in Section 4.2, showing that IHA provides a robust privacy auditing baseline, matching or exceeding the performance of current state-of-the-art attacks including attacks that use reference models. This is notable since IHA does ont require training any reference models or the use of hold-out data.

For a given false positive rate (FPR), a threshold is computed using scores for non-members, which is then used to compute the corresponding true positive rate (TPR). This is repeated for multiple FPRs to generate the corresponding ROC curve, which is used to compute the AUC. This experimental design is commonly used for membership-inference evaluations (Yeom et al., 2018; Carlini et al., 2022; Ye et al., 2022).

4.1 Setup

To evaluate IHA, we compare its performance to state-of-the-art baseline attacks with a representative set of datasets and models.

Baseline Attacks. We include LOSS as a baseline that does not use reference models, SIF as it uses iHVP similar to our audit, and LiRA as it is the current state-of-the-art for membership inference. While RMIA (Zarifzadeh et al., 2023) uses fewer reference models, it achieves performance comparable to LiRA and thus for the sake of performance comparison, it suffices to use LiRA with a large number of reference models. We describe the underlying access assumptions for these attacks in Table 1.

Attack	Model Access	Reference Models Used?	Leave-one-out knowledge?
LOSS	Black-box	No	No
LiRA	Black-box	Yes	No
SIF	White-box	No	No
L	Black-box	Yes	Yes
LiRA-L	Black-box	Yes	Yes
IHA (Ours)	White-box	No	Yes

Table 1: Comparison of attacks based on level of model-access, use of reference models, and knowledge of other training members.

LOSS (Yeom et al., 2018). The negative loss is used in this attack as a direct signal for membership inference.

SIF (Cohen & Giryes, 2024). Similar to ours, this attack employs the loss curvature of the target model by computing its Hessian, which is then used to compute a self-influence score. The original attack assigns 0–1 scores to target records. It classifies a given record as a member if its self-influence score is within the specified range and if its predicted class is correct. The latter rule can be ruled out as having many false positives/negatives. Instead of these steps, we choose to use the self-influence as membership scores directly. While the authors used approximation methods for iHVP, we use the exact Hessian for fair comparison.

LiRA (Carlini et al., 2022). There are two variants, *LiRA-Offline*, which uses "offline" models to estimate a Gaussian distribution and then performs one-sided hypothesis testing using loss scores, and *LiRA-Online*, with additionally employs "online" models, i.e., models whose training data includes the target record. The

likelihood ratio for online/offline model score distributions is then used as the score for membership inference. We use LiRA-Online, since it is the strongest of the two variants.

L-Attack (Ye et al., 2022). The L-attack operates in a leave-one-out setting, training reference models on $D \setminus \{z\}$ for any given record z. It uses loss as the target metric and computes attack thresholds for a desired false positive rate (FPR) by leveraging the distribution of losses obtained from reference models.

LiRA-L. We propose combining the LiRA attack for the LOO-setting by utilizing reference models trained under leave-one-out availability, followed by the offline variant of the LiRA attack (Carlini et al., 2022).

Datasets. Since we are limited by the computational constraints of computing iHVPs, we restrict our experiments to datasets where small models can perform adequately.

Purchase-100(S). The task for this dataset (Shokri et al., 2017) is to classify a given purchase into one of 100 categories, given 600 features. We train 2-layer MLPs (32 hidden neurons) with cross-entropy loss, with an average test accuracy of 84%. Experiments by Zarifzadeh et al. (2023) train larger (4-layer MLP) models on 25 K samples from Purchase-100, which is much smaller than the actual dataset, which is why we term it Purchase-100(S) (Small). We also demonstrate results with the 4-layer MLP that achieves similar task accuracy.

Purchase-100. For this version, we train models with 80 K samples. We use the same 2-layer MLP architecture as Purchase-100(S) but achieve a higher test accuracy of 90%. Utilizing more data increases the scope for model performance. We report results for Purchase-100 in Table 2 as the corresponding models are less prone to overfitting, but also report results for Purchase-100(S) for completeness in Appendix D.

MNIST-Odd. We consider the MNIST dataset LeCun et al. (1998), with the modified task of classifying a given digit image as odd or even. This modified task allows us to train models for binary classification using the regression loss, and is thus highly likely to follow the assumptions made in our theory regarding quadratic behavior for the loss function (Assumption 1). We train a logistic regression model with mean-squared error loss, with an average test loss of .078.

FashionMNIST. We use the FashionMNIST (Xiao et al., 2017) dataset, where the task is to classify a given clothing item image into one of ten categories. We train 2-layer MLPs (6 hidden neurons) with cross-entropy loss, with an average test accuracy of 83%.

Models. We train 128 models in the same way as done in Carlini et al. (2022), where data from each model is sampled at random from the actual dataset with a 50% probability. For each target model and target record, there are thus 127 reference models available, half of which are expected to include the target record in the training data. All of our models are trained with momentum ($\mu = 0.9$) and regularization ($\alpha = 5e^{-4}$), with a learning rate $\lambda = 0.01$. For a given false positive rate (FPR), a threshold is computed using scores for non-members, which is then used to compute the corresponding true positive rate (TPR). This is then repeated for multiple FPRs to generate the corresponding ROC curve, which is used to compute the AUC. This experimental design is commonly used for membership-inference evaluations (Yeom et al., 2018; Carlini et al., 2022; Ye et al., 2022).

4.2 Results

As summarized in Table 2, IHA provides a strong privacy auditing baseline that is competitive with current state-of-the-art attacks that require reference models. This is especially useful, considering that IHA does not require training any reference models and thus, does not require any hold-out data to train such reference models. IHA performs much better than the baselines on tabular data (Purchase-100), and is competitive with the baseline for image-based data (MNIST-Odd, Fashion MNIST). While tabular data is a more realistic setting for membership inference and the improved performance on Purchase-100 is promising, we leave further investigation of these factors to future work to better understand the performance discrepancies.

Table 2: Performance of various attacks, reported via attack AUC and true positive rate (TPR) at low false positive rate (FPR). ROC curves for low FPR region are visualized in Figure 1 (Appendix).

Attack	Purchase-100			MNIST-Odd			FashionMNIST		
	AUC	% TPR@FPR		AUC	% TPR@FPR		AUC	% TPR@FPR	
		1%	0.1%		1%	0.1%		1%	0.1%
LOSS (Yeom et al., 2018)	.531 $\pm$.001	0.97	0.00	.500 $\pm$.003	0.97	0.09	.507 $\pm$.002	1.00	0.10
SIF (Cohen & Giryes, 2024)	.531 $\pm$.001	0.97	0.10	.500 $\pm$.002	0.97	0.10	.507 $\pm$.002	0.98	0.10
LiRA (Carlini et al., 2022)	.644 $\pm$.004	4.70	0.98	**.568** $\pm$.005	**2.77**	**0.63**	.578 $\pm$.020	2.98	0.63
IHA (Exact)	**.703** $\pm$.004	**13.69**	**7.52**	.542 $\pm$.004	2.61	0.51	**.594** $\pm$.018	**4.06**	**0.89**

Table 3: Performance of exact and approximation-based variants of IHA, reported via attack AUC and true positive rate (TPR) at low false positive rate (FPR). Statistics are computed on 10000 samples.

Dataset	Exact			CG		
	AUC	%TPR@FPR		AUC	%TPR@FPR	
		1%	0.1%		1%	0.1%
Purchase-100	.701 $\pm$.009	13.74	7.56	.701 $\pm$.009	13.72	7.55
MNIST-Odd	.541 $\pm$.009	2.76	0.43	.541 $\pm$.009	2.76	0.43
FashionMNIST	.593 $\pm$.018	4.10	0.85	.592 $\pm$.018	4.09	0.86

Approximating iHVPs. In order to carry out IHA, an auditor needs to be able to calculate iHVPs and gradients for all training data. While computing gradients is more computationally intensive than simply calculating the loss, the difference is minimal. On the other hand, computing an iHVP involves calculating the Hessian matrix and then inverting it, both of which are computationally expensive processes. Even storing such an inverted Hessian can be problematic ($p \times p$ matrix for a model with p parameters). We thus experiment with evaluating IHA using Conjugate Gradients (Koh & Liang, 2017) to approximate iHVP. While such approximation does not require computing the Hessian directly, the time taken to compute this term for each record is non-trivial. We thus evaluate this approximation-based method on a random sample of 10000 records[1] and find that approximation methods retain most of the attack's performance (Table 3).

We reiterate that the purpose of our comparisons is not to claim a better membership inference attack for adversarial use; the threat models are not comparable, since our attack requires knowledge of all other records $D \setminus \{\boldsymbol{z}_1\}$ for inferring a given target record $\boldsymbol{z}_1$ (relaxing this assumption leads to severe performance degradation, Appendix F). Instead, IHA provides a way to empirically audit models for membership leakage without training reference models, which is desirable both in terms of computing and not having to reserve hold-out data for training reference models. More importantly, **our results suggest untapped potential in exploring parameter access for stronger privacy audits** (and the possibility of new inference attacks from an adversarial lens).

4.3 Ablating over terms inside IHA

As described in Equation 8, calculating IHA requires computing the loss along with the four additional terms I_1, I_2, I_3 and I_4. However, the terms I_3 and I_4 are scaled by α (which is usually very small) and involve an iHVP of an iHVP and thus may be much smaller compared to terms like I_1, I_2 and the loss. We explore variants of IHA, which ignore the terms I_3 and I_4, to see how they impact auditing performance. We also consider variants that use only the terms I_1 and I_2 to understand the importance of their contributions to the performance of IHA, along with the inclusion or not of the loss term to understand the relative importance of parameter-based signals.

[1]For a direct comparison, we recompute metrics for the 10000 samples on which we use approximate-based variants.

Table 4: Performance of IHA on Purchase100-S (MLP-2) when only some of the terms corresponding to Equation 8 are used. I_1 and I_2 seem to be responsible for most of the privacy auditing performance.

Terms Used	AUC	%TPR@FPR	
		1%	0.1%
I_1	.591 $\pm.003$	1.04	0.09
I_2	.704 $\pm.004$	2.63	0.70
I_1, I_2	.779 $\pm.003$	17.61	16.65
I_1, I_2, I_3, I_4	.779 $\pm.003$	17.60	16.64
$\ell(\boldsymbol{w}, \boldsymbol{z}_1)$, I_1	.594 $\pm.003$	1.12	0.12
$\ell(\boldsymbol{w}, \boldsymbol{z}_1)$, I_2	.686 $\pm.004$	1.97	0.48
$\ell(\boldsymbol{w}, \boldsymbol{z}_1)$, I_1, I_2	.791 $\pm.003$	19.96	19.00
All	.791 $\pm.005$	20.09	19.09

The ablation study presented in Table 4 reveals several key insights about the performance of IHA. Excluding I_3 and I_4 has close to no impact on auditing performance, even for low-FPR scenarios. The combination of I_1 and I_2 alone achieves an AUC of .779, which is nearly identical to the performance when all terms are included (AUC .791). Notably, the addition of the loss term $\ell(\boldsymbol{w}, \boldsymbol{z}_1)$ to I_1 and I_2 results in a marginal improvement, increasing the AUC to .791 and slightly boosting the TPR at both 1% and 0.1% FPR. Interestingly, when examined individually, I_2 (AUC .704) performs significantly better than I_1 (AUC .591), suggesting that I_2 captures more relevant information for the auditing task. Including the loss term $\ell(\boldsymbol{w}, \boldsymbol{z}_1)$ has little impact on I_1 but harms performance when used wth I_2. These findings indicate that the majority of the attack's effectiveness stems from the inverse Hessian vector products used in I_1 and I_2, with I_2 being particularly important, while the terms involving weight regularization and nested iHVPs (I_3 and I_4) contribute minimally to the overall performance. While not as impactful as I_2, the loss term still provides valuable information for the auditing process. Based on these results, a simplified version of IHA using only $\ell(\boldsymbol{w}, \boldsymbol{z}_1)$, I_1, and I_2 could potentially offer a favorable trade-off between computational efficiency and auditing effectiveness.

4.4 Comparison with Leave-One-Out Setting

When targeting a record for inference, IHA assumes knowledge of all other $n-1$ records in an n-sized dataset. It is possible that the improved performance of IHA is due to this extra information rather than inherent parameter access. To isolate and analyze these potential sources of increase in leakage, we also assess the performance of a leave-one-out (LOO) membership inference test. Specifically, we evaluate the L-attack (Ye et al., 2022) on 1000 samples, training 100 reference models per record for score calibration. Since targeting each record requires training multiple reference models for the L-attack, evaluating it on a larger sample of data is computationally infeasible.

These results indicate that IHA outperforms the L-attack in terms of AUC, achieving a value of .791 compared to .737 for the L-attack.[2] This suggests that even when controlling for the additional knowledge of all other records in the dataset, the primary source of IHA's superior performance stems from its access to model parameters rather than just the leave-one-out setup. In a way, IHA is utilizing parameter access to circumvent the need for reference models, as it directly aims to measure the influence of the given target record instead of relying on reference models for score calibration.

Interestingly, in comparison, LiRA achieves an AUC of .767, which is lower than IHA but still higher than the L-attack. This suggests that LiRA, even without the extensive reference model training, is more effective than the L-attack, possibly due to its utilization of both “in" and “out" models as opposed to just “out"

[2]Our results for the L-attack are lower than those reported by Ye et al. (2022). For instance, we observe a TPR of .668 at 0.3 FPR, while it was reported to be .968 by (Ye et al., 2022). However, this discrepancy arises from our setting, which uses more data and fewer model parameters. We verified our implementation through direct correspondence with the authors and by replicating their results in the original setting, which used a smaller dataset and more parameters, (resulting in a model prone to overfitting).

models with the L-attack. To try and devise a stronger black-box attack for the LOO setting, we extend LiRA to the LOO setting by utilizing models trained on LOO data as reference models. LiRA-Offline under the LOO setting achieves an AUC of .633, lower than the L-attack (.737). Overall, these results demonstrate the promise of IHA as a privacy auditing tool- yielding results comparable to that of techniques that train hundreds of reference models (for each target record in the worst case, as in L-attack), without using a single reference model.

4.5 Inter-Attack Agreement

Similar to Ye et al. (2022), we compute the agreement rate between ground-truth membership labels and membership predicted by various attacks to understand the ability of our privacy audit to identify vulnerable data, and demonstrate how it differs from existing attacks. Table 5 presents the agreement rate between ground-truth membership labels and membership predicted by various attacks.

Table 5: Agreement rate between ground truth (GT) membership values, and various attacks for 500 training and 500 testing data points. The upper triangle of the table corresponds to the agreement rates of members, whereas the lower triangle corresponds to the agreement rates of non-members. The experimental setup is Purchase-100(S), with effective FPR ≈ 0.05 (a) and effective FPR ≈ 0.3 (b).

	LiRA	**L**	**LiRA-L**	**IHA**	**GT**
LiRA		.794	.826	.718	.220
L	.930		.908	.712	.234
LiRA-L	.916	.938		.768	.154
IHA	.912	.926	.916		.230
GT	.954	.952	.954	.958	

(a) Agreement between methods with FPR 0.05

	LiRA	**L**	**LiRA-L**	**IHA**	**GT**
LiRA		.732	.658	.436	.652
L	.724		.766	.492	.668
LiRA-L	.588	.660		.462	.518
IHA	.612	.660	.584		.640
GT	.702	.702	.706	.706	

(b) Agreement between methods with FPR 0.3

At a low FPR of ≈ 0.05, agreement in predictions for non-members is very high between attacks, with agreement rates above 0.91 for all pairs of attacks. On the other hand, agreement rates for member records are expectedly lower. Interestingly, agreement between LiRA and LiRA-L is higher than IHA and any other attack. This difference is especially pronounced for a higher FPR (Table 5b), where agreement rates are as low as ≈ 0.4 compared to 0.766 for LiRA and LiRA-L. This is very interesting becaue the corresponding TPRs for IHA are higher than LiRA and comparable to that of the L-attack, thus suggesting that the records identified by IHA as being vulnerable are very different from those identified by LiRA or even the L-attack. This also means that a combined (classifying a record as a member only when both attacks classify as a member) attack would have some true positives with a very low FPR.

4.5.1 Runtime Comparison

To compare the computational costs of our proposed audit with existing auditing techniques, we analyze runtime and memory usage statistics across different methods, aiming to evaluate efficiency and practicality in real-world privacy audits (Table 6).

Attack	Time (s)		Memory (MB)		
	Precompute	Time/Sample	Precompute	Runtime	max(Precompute, Runtime)
IHA	43×60	0.16	4228	1324	4228
IHA (Approx)	0	85	0	1638	1638
LiRA	192×60	0.07	276	1228	1228
LOSS	0	0.004	0	906	906

Table 6: Runtime and memory statistics for various auditing techniques. Statistics are computed over 100 randomly-selected members for MLP-2 architecture models trained on Purchase-100 dataset.

While the peak memory consumption of IHA is higher than that of LiRA, the approximate version of IHA is not too far from LiRA in terms of memory consumption. Computing the total runtime is a function of

the number of samples used for the privacy audit, as a "precompute" step is required for LiRA (training reference models) and exact IHA (computing Hessian). For instance, computing the audit for 1K samples takes less overall time for IHA ($\approx$ 1 hour) than it does for LiRA ($\approx$ 3 hours). It should be noted that the Hessian is too large to store on our GPU for IHA and is thus stored on the CPU, which is also why it is slower. We restate that the main benefit of such a privacy audit comes from not having to reserve hold-out (or auxiliary) data, not from any computational cost reduction. Our privacy audit, like the most trivial LOSS attack, only requires member data and some non-member data, whereas other attacks in the literature require shadow/reference models trained on comparable-sized datasets; the limiting factor here is reserving data to train reference models, which a real-world model trainer may not want to do since it reduces the amount of data available for training.

5 Conclusion

Our theoretical result proves that model parameter access is indeed necessary for optimal membership inference, contrary to previous results and the common belief that optimal membership inference can be achieved with only black-box model access. We propose the Inverse Hessian Attack inspired by this theory that provides stronger privacy auditing than existing black-box techniques.

Limitations. IHA is not yet practically realizable for most settings due to the computational expense of calculating the Hessian, or even approximating iHVPs. This restriction poses challenges for real-world cases where fixed compute budgets may be more crucial than the availability of auxiliary data. An auditor might use a subset of parameters to reduce computational costs while performing Hessian-based computations. This aligns with model pruning (Liu et al., 2019), but understanding its impact on membership knowledge within parameters is non-trivial (Yuan & Zhang, 2022). We also note that IHA's performance can be sensitive to the choice of the damping factor, which requires further investigation (Appendix D.1).

Our conclusion aligns well with recent calls in the literature to consider white-box access for rigorous auditing (Casper et al., 2024). While our theory shows that parameter access is required for optimal membership inference, it remains unclear how much better this is compared to the optimal membership inference attack restricted to black-box access. Our empirical studies suggest the gap is non-trivial, but further study is required to understand the theoretical limit of black-box attacks, which is a non-trivial but interesting direction to explore. Exploring the accuracy of iHVP approximation methods to extend IHA to larger models, along with multi-record inference, are both promising directions for future research.

Broader Impact Statement

The increasing integration of AI in sensitive domains like healthcare, finance, and personal data management highlights the critical importance of privacy. Information leakage from AI models can have severe consequences, making effective privacy auditing a necessary safeguard. Our work contributes to this field by theoretically demonstrating that optimal membership inference attacks require white-box access to model parameters, challenging the adequacy of black-box approaches. We also demonstrate with Inverse Hessian Attack how this theory can be used to design empirical privacy audits that do not rely on reference models.

We advocate for the development of more sophisticated privacy auditing tools that fully leverage the elevated access typically available to auditors, such as model parameters, to assess privacy leakage efficiently without extensive data and compute resources. We hope our theoretical and empirical results will reinvigorate interest in the privacy research community to explore white-box attacks, for both adversarial and auditing purposes.

References

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *Journal of Machine Learning Research*, 2017.

Achraf Azize and Debabrota Basu. How much does each datapoint leak your privacy? quantifying the per-datum membership leakage. *arXiv:2402.10065*, 2024.

Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 2024.

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 2024.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*, 2022.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. Black-box access is insufficient for rigorous ai audits. In *ACM Conference on Fairness, Accountability, and Transparency*, 2024.

Harsh Chaudhari, Giorgio Severi, Alina Oprea, and Jonathan Ullman. Chameleon: Increasing label-only membership leakage with adaptive poisoning. In *International Conference on Learning Representations*, 2024.

Gilad Cohen and Raja Giryes. Membership inference attack using self influence functions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Matthew Jagielski, Yangsibo Huang, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, and others. Challenges towards the Next Frontier in Privacy. *arXiv:2304.06929*, 2023.

Daniel DeAlcala, Aythami Morales, Gonzalo Mancera, Julian Fierrez, Ruben Tolosana, and Javier Ortega-Garcia. Is my Data in your AI Model? Membership Inference Test with Application to Face Images. *arXiv:2402.09225*, 2024.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 2006.

Mishaal Kazmi, Hadrien Lautraite, Alireza Akbari, Mauricio Soroco, Qiaoyue Tang, Tao Wang, Sébastien Gambs, and Mathias Lécuyer. PANORAMIA: Privacy Auditing of Machine Learning Models without Retraining. *arXiv:2402.09477*, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

Sasi Kumar and Reza Shokri. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. In *Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs)*, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

Marvin Li, Jason Wang, Jeffrey George Wang, and Seth Neel. Mope: Model perturbation based privacy attacks on language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *International Conference on Machine Learning*, 2021.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019.

Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of sgd. In *International Conference on Machine Learning*, pp. 15959–15975. PMLR, 2022.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning. In *IEEE Symposium on Security and Privacy*, 2018.

Elre Talea Oldewage, Ross M Clarke, and José Miguel Hernández-Lobato. Series of hessian-vector products for tractable saddle-free newton optimisation of neural networks. *Transactions on Machine Learning Research*, 2024.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, 2019.

Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. SoK: Let the privacy games begin! A unified treatment of data inference privacy in machine learning. In *IEEE Symposium on Security and Privacy*, 2023.

Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, 2014.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017.

Mandt Stephan, Matthew D Hoffman, David M Blei, and others. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 2017.

Jasper Tan, Blake Mason, Hamid Javadi, and Richard Baraniuk. Parameters or privacy: A provable tradeoff between overparameterization and membership inference. *Advances in Neural Information Processing Systems*, 2022.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, August 2017. arXiv: cs.LG/1708.07747.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022.

Jiayuan Ye, Anastasia Borovykh, Soufiane Hayou, and Reza Shokri. Leave-one-out distinguishability in machine learning. In *International Conference on Learning Representations*, 2024.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Computer Security Foundations symposium (CSF)*, 2018.

Samuel Yeom, Irene Giacomelli, Alan Menaged, Matt Fredrikson, and Somesh Jha. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *Journal of Computer Security*, 2020.

Soma Yokoi and Issei Sato. Bayesian interpretation of sgd as ito process. *arXiv:1911.09011*, 2019.

Xiaoyong Yuan and Lan Zhang. Membership inference attacks and defenses in neural network pruning. In *USENIX Security Symposium*, 2022.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference by boosting relativity. *arXiv:2312.03262*, 2023.

Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2021.

Liu Ziyin, Hongchao Li, and Masahito Ueda. Law of balance and stationary distribution of stochastic gradient descent. *arXiv:2308.06671*, 2023.

A Related Work

This section reviews methods in membership inference (black-box and white-box), techniques for privacy auditing to predict and mitigate data leakage, and the dynamics of stochastic gradient descent (SGD) along with inverse Hessian vector products (iHVPs).

A.1 Membership Inference

Black-box Membership Inference. Early works on membership inference worked under black-box access, utilizing the model's loss (Shokri et al., 2017) on a given datapoint as a signal for membership. Since then there have been several works focusing on different forms of difficulty calibration—accounting for the inherent "difficulty" of predicting on a record, irrespective of its presence in train data. This calibration has taken several forms; direct score normalization with reference models (Sablayrolles et al., 2019), likelihood tests based on score distributions (Carlini et al., 2022; Zarifzadeh et al., 2023; Ye et al., 2022), and additional models for predicting difficulty (Bertran et al., 2024).

White-box Membership Inference. Nasr et al. (2018) explored white-box access to devise a meta-classifier-based attack that additionally extracts intermediate model activations and gradients to increase leakage but concluded that layers closer to the model's output are more informative for membership inference and report performance not significantly better than a black-box loss-based attack. Recent work by DeAlcala et al. (2024), however, makes the opposite observation, with layers closer to the model's input providing noticeably better performance. Apart from these meta-classifier driven approaches, some works attempt to utilize parameter access much more directly, often utilizing Hessian in one form or another. Cohen & Giryes (2024) defined the self-influence of a datapoint $\boldsymbol{z}_i$ as $(\boldsymbol{g}_i{}^\top \mathbf{H}^{-1} \boldsymbol{g}_i)$ as a signal for membership, using LiSSA (Agarwal et al., 2017) to approximate the iHVP. This has similarities to our result since our optimal membership inference score also involves computing iHVPs. Li et al. (2023) attempted to measure the sharpness for a given model by evaluating fluctuations in model predictions after adding zero-mean noise to the parameters, a step that is supposed to approximate the trace of the Hessian at the given point.

A.2 Privacy Auditing

Ye et al. (2024) proposed using efficient methods to "predict" memorization by not having to run computationally expensive membership inference attacks, with reported speedups of up to $140x$. They showed how their proposed score (LOOD) correlates well with AUC, corresponding to an extremely strong MIA with all-but-one access to records (L-attack (Ye et al., 2022)). However, it is unclear if this computed LOOD is directly comparable across models, making it hard to calibrate these scores to compare the leakage from a model relative to another (an important aspect of internal privacy auditing). Their derivations also involve a connection with the Hessian. Biderman et al. (2024) studied the problem of forecasting memorization in a model for specific training data and proposed using partially trained versions of the model (or smaller models) as a proxy for their computation. While their results support the need for inexpensive auditing methods, their focus is on predicting memorization early in the training process, while ours relates to auditing fully trained models. More recently, Tan et al. (2022) studied the theory behind worst-case membership leakage for the case of linear regression on Gaussian data and derived insights. While this is useful to make an intuitive connection with overfitting, it does not provide a realizable attack or insights for the standard case of models trained with SGD.

A.3 SGD Dynamics and iHVPs

SGD Dynamics. Stephan et al. (2017) approximated the SGD dynamics as an Ornstein-Uhlenbeck process, while Yokoi & Sato (2019) provided a discrete-time weak-order approximation for SGD based on Itô process and finite moment assumption. However, both works rely on strong assumptions about the gradient noises and require a vanishingly small learning rate, largely deviating from the common practice of SGD. To address the limitations of the aforementioned works, Liu et al. (2021) directly analyzed the discrete-time dynamics of SGD and derived the analytic form of the asymptotic model fluctuation with respect to the asymptotic gradient noise covariance and the Hessian matrix. Ziyin et al. (2021) further generalized the results of (Liu et al., 2021) by deriving the exact minibatch noise covariance for discrete-time SGD, which is shown to vary across different kinds of local minima. Our work builds on these advanced theoretical results of discrete-time SGD dynamics but aims to enhance the understanding of optimal membership inference, particularly for models trained with SGD.

iHVPs. Currently literature on approximating inverse-Hessian vector products relies on one of two methods: conjugate gradients (Koh & Liang, 2017) or LiSSA (Agarwal et al., 2017). Both approximation methods rely on efficient computation of exact Hessian-vector products, and use forward and backward propagation as sub-routines. While these methods have utility in certain areas, such as influence functions (Koh & Liang, 2017) and optimization (Oldewage et al., 2024), approximation errors can be non-trivial. For instance, I_1 in the formulation of our attack requires a low approximation error in the norm of an iHVP, while I_2 simultaneously requires a low approximation error in the direction of the iHVP. Recent work on curvature-aware minimization by Oldewage et al. (2024) proposes another method for efficient iHVP approximation as a subroutine, but the authors observed high approximation errors based on both norm and direction.

B Proof for Theorem 3.1

Proof. Recall that $\mathbf{H}_* = \mathbf{H}(\boldsymbol{w}^*)$ and $L_* = L(\boldsymbol{w}^*)$. According to Theorem 2.2 and Theorem 2.3, we obtain

$$\boldsymbol{\Sigma} = \frac{\lambda}{S(1-\mu)}\left(2L_*\mathbf{H}_* - \alpha^2 \boldsymbol{w}^* \boldsymbol{w}^{*\top}\right)\left(\mathbf{H}_* + \alpha\mathbf{I}_d\right)^{-1}\left(2\mathbf{I}_d - \frac{\lambda}{1+\mu}(\mathbf{H}_* + \alpha\mathbf{I}_d)\right)^{-1}, \tag{9}$$

where we only consider the leading terms in Theorems 2.2 and 2.3. Note that Equation 9 holds when the Hessian matrix $\mathbf{H}_*$ has full rank and $L_* \neq 0$. When the Hessian has degenerated rank such that $\operatorname{rank}(\mathbf{H}_* + \alpha\mathbf{I}_d) = r < d$, the following more generalized result can be derived:

$$\mathbf{P}_r\boldsymbol{\Sigma} = \frac{\lambda}{S(1-\mu)}\left(2L_*\mathbf{H}_* - \alpha^2 \boldsymbol{w}^* \boldsymbol{w}^{*\top}\right)\left(\mathbf{H}_* + \alpha\mathbf{I}_d\right)^{+}\left(2\mathbf{I}_d - \frac{\lambda}{1+\mu}\left(\mathbf{H}_* + \alpha\mathbf{I}_d\right)\right)^{-1},$$

where $\mathbf{P}_r = \mathrm{diag}(1, ..., 1, 0, ...0)$ denotes the projection matrix onto non-zero eigenvalues, and $+$ is the Moore-Penrose inverse operator. If $L_* = 0$, meaning $\boldsymbol{w}^*$ is a global minimum, then the asymptotic model fluctuation $\boldsymbol{\Sigma} = \mathbf{0}$. For ease of presentation, we assume the Hessian matrix has full rank in the following proof.

Then, we get:

$$\boldsymbol{\Sigma}^{-1} = \frac{S(1-\mu)}{\lambda}\left(2\mathbf{I}_d - \frac{\lambda}{1+\mu}(\mathbf{H}_* + \alpha\mathbf{I}_d)\right)\left(\mathbf{H}_* + \alpha\mathbf{I}_d\right)\left(2L_*\mathbf{H}_* - \alpha^2\boldsymbol{w}^*\boldsymbol{w}^{*\top}\right)^{-1}. \tag{10}$$

Using the Sherman–Morrison formula, we obtain:

$$\begin{aligned}\left(2L_*\mathbf{H}_* - \alpha^2\boldsymbol{w}^*\boldsymbol{w}^{*\top}\right)^{-1} &= \frac{1}{2L_*}\mathbf{H}_*^{-1} + \frac{\alpha^2}{2L_*(2L_* - \boldsymbol{w}^{*\top}\mathbf{H}_*^{-1}\boldsymbol{w}^*)}\left(\mathbf{H}_*^{-1}\boldsymbol{w}^*\boldsymbol{w}^{*\top}\mathbf{H}_*^{-1}\right) \\ &= \frac{1}{2L_*}\mathbf{H}_*^{-1}\left(\mathbf{I}_d + O(\alpha^2)\right).\end{aligned} \tag{11}$$

Leaving out the second-order term $O(\alpha^2)$ in Equation 11 (since the regularization parameter α is a typically small constant in $[0, 1)$) and plugging it back in Equation 10, we get:

$$\begin{aligned}\boldsymbol{\Sigma}^{-1} &= \frac{S(1-\mu)}{2\lambda L_*}\left(2\mathbf{I}_d - \frac{\lambda}{1+\mu}(\mathbf{H}_* + \alpha\mathbf{I}_d)\right)\left(\mathbf{H}_* + \alpha\mathbf{I}_d\right)\mathbf{H}_*^{-1} \\ &= \frac{S(1-\mu)}{2\lambda L_*}\left(2\mathbf{I}_d - \frac{\lambda}{1+\mu}(\mathbf{H}_* + \alpha\mathbf{I}_d) + 2\alpha\mathbf{H}_*^{-1} - \frac{\lambda\alpha}{1+\mu}(\mathbf{I}_d + \alpha\mathbf{H}_*^{-1})\right) \\ &= \frac{S(1-\mu)}{2\lambda L_*}\left(2\left(1 - \frac{\lambda\alpha}{1+\mu}\right)\mathbf{I}_d - \left(\frac{\lambda}{1+\mu}\right)\mathbf{H}_* + \alpha\left(2 - \frac{\lambda\alpha}{1+\mu}\right)\mathbf{H}_*^{-1}\right).\end{aligned} \tag{12}$$

According to Laplace approximation, we can approximate the posterior distribution of $\boldsymbol{w}$ given $\boldsymbol{w}^*$ as $\mathcal{N}(\boldsymbol{w}^*, \boldsymbol{\Sigma})$. Therefore, making use of Equation 9, we can derive the explicit formula of the log-posterior

distribution as:

$$
\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{w}^*) &= -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{\Sigma}) - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}^*)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{w}^*) \\
&= \frac{1}{2}\sum_{i=1}^{d}\ln\left(\frac{\left(2-\frac{\lambda}{1+\mu}(\sigma_i(\mathbf{H}_*)+\alpha)\right)\left(\sigma_i(\mathbf{H}_*)+\alpha\right)}{2L_*\sigma_i(\mathbf{H}_*)}\right) + \text{const.} \\
&\quad - \frac{S(1-\mu)}{4\lambda L_*}\left[2\left(1-\frac{\lambda\alpha}{1+\mu}\right)\|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2 + \alpha\left(2-\frac{\lambda\alpha}{1+\mu}\right)(\boldsymbol{w}-\boldsymbol{w}^*)^\top\mathbf{H}_*^{-1}(\boldsymbol{w}-\boldsymbol{w}^*)\right. \\
&\qquad \left. - \frac{\lambda}{1+\mu}(\boldsymbol{w}-\boldsymbol{w}^*)^\top\mathbf{H}_*(\boldsymbol{w}-\boldsymbol{w}^*)\right] \\
&= \frac{1}{2}\sum_{i=1}^{d}\ln\left(\frac{\left(2-\frac{\lambda}{1+\mu}(\sigma_i(\mathbf{H}_*)+\alpha)\right)\left(\sigma_i(\mathbf{H}_*)+\alpha\right)}{2L_*\sigma_i(\mathbf{H}_*)}\right) + \text{const.} \\
&\quad - \frac{S(1-\mu)}{2\lambda L_*}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2 \\
&\qquad - \frac{S(1-\mu)\alpha}{4\lambda L_*}\left(2-\frac{\lambda\alpha}{1+\mu}\right)(\boldsymbol{w}-\boldsymbol{w}^*)^\top\mathbf{H}_*^{-1}(\boldsymbol{w}-\boldsymbol{w}^*) \\
&\qquad\quad + \frac{S(1-\mu)}{4(1+\mu)L_*}(\boldsymbol{w}-\boldsymbol{w}^*)^\top\mathbf{H}_*(\boldsymbol{w}-\boldsymbol{w}^*) \\
&= -\frac{d}{2}\ln L_* + \frac{1}{2}\sum_{i=1}^{d}\ln\left(\frac{\left(2-\frac{\lambda}{1+\mu}(\sigma_i(\mathbf{H}_*)+\alpha)\right)\left(\sigma_i(\mathbf{H}_*)+\alpha\right)}{\sigma_i(\mathbf{H}_*)}\right) + \text{const.} \\
&\quad - \frac{S(1-\mu)}{2\lambda}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\cdot\frac{\|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2}{L_*} \\
&\qquad - \frac{S(1-\mu)\alpha}{4\lambda}\cdot\left(2-\frac{\lambda\alpha}{1+\mu}\right)\cdot\frac{\nabla L(\boldsymbol{w})^\top\mathbf{H}_*^{-3}\nabla L(\boldsymbol{w})}{L_*} \\
&\qquad\quad + \frac{S(1-\mu)}{2(1+\mu)}\cdot\frac{L(\boldsymbol{w}^*)}{L_*} + o(\|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2).
\end{aligned} \tag{13}
$$

Here, the last equality holds because of the second-order Taylor expansion:

$$
L(\boldsymbol{w}) = L(\boldsymbol{w}^*) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}^*)^\top\mathbf{H}_*(\boldsymbol{w}-\boldsymbol{w}^*) + o(\|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2).
$$

Omitting the constant and negligible terms, we thus complete the proof of Theorem 3.1. Note that we keep the $O(\alpha^2)$ term in Equation 13 and Theorem 3.1 for the sake of completeness but expect it to be negligible compared with other terms, due to the fact that α is typically set as a very small constant within $[0, 1)$. □

C Optimal Membership-Inference Score

C.1 Proof for Theorem 3.2

Proof. To derive the scoring function for an optimal membership inference, we need to compute the ratio between $p(\boldsymbol{w}|\boldsymbol{w}_1^*)$ and $p(\boldsymbol{w}|\boldsymbol{w}_0^*)$, where $\boldsymbol{w}_0^*$ (resp. $\boldsymbol{w}_1^*$) denotes a local minimum (close to $\boldsymbol{w}$) of the training loss function with respect to $\{\boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$ (resp. $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$). Note that we've obtained the posterior

distribution of $\boldsymbol{w}$ in Theorem 3.1. Therefore, the remaining task is to analyze the following terms:

$$
\begin{aligned}
&\ln p(\boldsymbol{w}|\boldsymbol{w}_1^*) - \ln p(\boldsymbol{w}|\boldsymbol{w}_0^*) \\
&= -\frac{d}{2}\big[\ln L_1(\boldsymbol{w}_1^*) - \ln L_0(\boldsymbol{w}_0^*)\big] + \frac{1}{2}\sum_{i=1}^{d} \ln\left(\frac{\big(2-\frac{\lambda}{1+\mu}(\sigma_i(\mathbf{H}_1(\boldsymbol{w}_1^*))+\alpha)\big)\big(\sigma_i(\mathbf{H}_1(\boldsymbol{w}_1^*))+\alpha\big)}{\big(2-\frac{\lambda}{1+\mu}(\sigma_i(\mathbf{H}_0(\boldsymbol{w}_0^*))+\alpha)\big)\big(\sigma_i(\mathbf{H}_0(\boldsymbol{w}_0^*))+\alpha\big)} \cdot \frac{\sigma_i(\mathbf{H}_0(\boldsymbol{w}_0^*))}{\sigma_i(\mathbf{H}_1(\boldsymbol{w}_1^*))}\right) \\
&\quad - \frac{S(1-\mu)}{2\lambda}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\left(\frac{\|\boldsymbol{w}-\boldsymbol{w}_1^*\|_2^2}{L_1(\boldsymbol{w}_1^*)} - \frac{\|\boldsymbol{w}-\boldsymbol{w}_0^*\|_2^2}{L_0(\boldsymbol{w}_0^*)}\right) + \frac{S(1-\mu)}{2(1+\mu)}\left(\frac{L_1(\boldsymbol{w})}{L_1(\boldsymbol{w}_1^*)} - \frac{L_0(\boldsymbol{w})}{L_0(\boldsymbol{w}_0^*)}\right) \\
&\quad - \frac{S(1-\mu)\alpha}{4\lambda}\left(2-\frac{\lambda\alpha}{1+\mu}\right)\left(\frac{\nabla L_1(\boldsymbol{w})^\top \mathbf{H}_1(\boldsymbol{w}_1^*)^{-3}\nabla L_1(\boldsymbol{w})}{L_1(\boldsymbol{w}_1^*)} - \frac{\nabla L_0(\boldsymbol{w})^\top \mathbf{H}_0(\boldsymbol{w}_0^*)^{-3}\nabla L_0(\boldsymbol{w})}{L_0(\boldsymbol{w}_0^*)}\right), \qquad (14)
\end{aligned}
$$

where $\mathbf{H}_0(\boldsymbol{w}_0^*)$ (resp. $\mathbf{H}_1(\boldsymbol{w}_1^*)$) denotes the Hessian of L_0 (resp. L_1) at $\boldsymbol{w}_0^*$ (resp. $\boldsymbol{w}_1^*$). Since both $\boldsymbol{w}_0^*$ and $\boldsymbol{w}_1^*$ are close to parameters of the observed victim model $\boldsymbol{w}$, so we can approximate the corresponding loss using second-order Taylor expansion. Also, according to Assumption 3, we know $\mathbf{H}_0(\boldsymbol{w}_0^*) = \mathbf{H}_1(\boldsymbol{w}_1^*) = \mathbf{H}_*$ and $L_0(\boldsymbol{w}_0^*) = L_1(\boldsymbol{w}_1^*) = L_*$. Thus, we can simplify Equation 14 and obtain the following form:

$$
\begin{aligned}
&\ln p(\boldsymbol{w}|\boldsymbol{w}_1^*) - \ln p(\boldsymbol{w}|\boldsymbol{w}_0^*) \\
&= -\frac{S(1-\mu)}{2\lambda L_*}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\big(\|\boldsymbol{w}-\boldsymbol{w}_1^*\|_2^2 - \|\boldsymbol{w}-\boldsymbol{w}_0^*\|_2^2\big) + \frac{S(1-\mu)}{2(1+\mu)L_*}\big(L_1(\boldsymbol{w}) - L_0(\boldsymbol{w})\big) \\
&\quad - \frac{S(1-\mu)\alpha}{4\lambda L_*}\left(2-\frac{\lambda\alpha}{1+\mu}\right)\big(\nabla L_1(\boldsymbol{w})^\top \mathbf{H}_*^{-3}\nabla L_1(\boldsymbol{w}) - \nabla L_0(\boldsymbol{w})^\top \mathbf{H}_*^{-3}\nabla L_0(\boldsymbol{w})\big) \\
&= -\frac{S(1-\mu)}{2\lambda L_*}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\big(\nabla L_1(\boldsymbol{w})^\top \mathbf{H}_*^{-1}\mathbf{H}_*^{-1}\nabla L_1(\boldsymbol{w}) - \nabla L_0(\boldsymbol{w})^\top \mathbf{H}_*^{-1}\mathbf{H}_*^{-1}\nabla L_0(\boldsymbol{w})\big) + \frac{S(1-\mu)\ell(\boldsymbol{w},\boldsymbol{z}_1)}{2n(1+\mu)L_*} \\
&\quad - \frac{S(1-\mu)\alpha}{4n\lambda L_*}\left(2-\frac{\lambda\alpha}{1+\mu}\right)\left(2\nabla L_0(\boldsymbol{w})^\top \mathbf{H}_*^{-3}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1) + \frac{1}{n}\ell(\boldsymbol{w},\boldsymbol{z}_1)^\top \mathbf{H}_*^{-3}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1)\right) \\
&= -\frac{S(1-\mu)}{2n\lambda L_*}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\left(2\nabla L_0(\boldsymbol{w})^\top \mathbf{H}_*^{-1}\mathbf{H}_*^{-1}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1) + \frac{1}{n}\|\mathbf{H}_*^{-1}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1)\|_2^2\right) + \frac{S(1-\mu)\ell(\boldsymbol{w},\boldsymbol{z}_1)}{2n(1+\mu)L_*} \\
&\quad - \frac{S(1-\mu)\alpha}{4n\lambda L_*}\left(2-\frac{\lambda\alpha}{1+\mu}\right)\left(2\nabla L_0(\boldsymbol{w})^\top \mathbf{H}_*^{-3}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1) + \frac{1}{n}\ell(\boldsymbol{w},\boldsymbol{z}_1)^\top \mathbf{H}_*^{-3}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1)\right), \qquad (15)
\end{aligned}
$$

where the second equality holds because of the Taylor approximation $\nabla L_i(\boldsymbol{w}) - \nabla L_i(\boldsymbol{w}_i^*) = \mathbf{H}_*(\boldsymbol{w} - \boldsymbol{w}_i^*)$ for $i \in \{0, 1\}$. Moreover, according to Lemma 2.1, we know the optimal membership inference is given by:

$$
\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1) = \mathbb{E}_{\mathcal{T}}\left[\sigma\left(\ln\left(\frac{p(\boldsymbol{w}|m_1=1,\boldsymbol{z}_1,\mathcal{T})}{p(\boldsymbol{w}|m_1=0,\boldsymbol{z}_1,\mathcal{T})}\right) + \ln\left(\frac{\gamma}{1-\gamma}\right)\right)\right], \qquad (16)
$$

where $\sigma(u) = (1+\exp(-u))^{-1}$ is the Sigmoid function, $\mathcal{T} = \{\boldsymbol{z}_2, \ldots, \boldsymbol{z}_n, m_2, \ldots, m_n\}$, and $\gamma = \mathbb{P}(m_i = 1)$. Plugging Equation 15 into Equation 16, we obtain

$$
\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1) = \mathbb{E}_{\mathcal{T}}\left[\sigma\left(\frac{S(1-\mu)}{2nL_*}\left(\frac{\ell(\boldsymbol{w},\boldsymbol{z}_1)}{(1+\mu)} - \frac{1}{\lambda}\big(I_1 + I_2 + I_3 + I_4\big)\right) + t_\gamma\right)\right],
$$

where I_1, I_2, I_3, I_4 and t_γ are defined as:

$$
\begin{aligned}
I_1 &:= \frac{1}{n}\left(1-\frac{\lambda\alpha}{1+\mu}\right)\cdot\|\mathbf{H}_*^{-1}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1)\|_2^2, \\
I_2 &:= 2\left(1-\frac{\lambda\alpha}{1+\mu}\right)\cdot\big(\mathbf{H}_*^{-1}\nabla L_0(\boldsymbol{w})\big)^\top\big(\mathbf{H}_*^{-1}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1)\big), \\
I_3 &:= \frac{\alpha}{2n}\left(2-\frac{\lambda\alpha}{1+\mu}\right)\cdot\big(\mathbf{H}_*^{-1}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1)\big)^\top\big(\mathbf{H}_*^{-1}\big(\mathbf{H}_*^{-1}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1)\big)\big), \\
I_4 &:= \alpha\left(2-\frac{\lambda\alpha}{1+\mu}\right)\cdot\big(\mathbf{H}_*^{-1}\nabla L_0(\boldsymbol{w})\big)^\top\big(\mathbf{H}_*^{-1}\big(\mathbf{H}_*^{-1}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1)\big)\big), \\
t_\gamma &:= \ln\left(\frac{\gamma}{1-\gamma}\right).
\end{aligned}
$$

Therefore, we complete the proof of Theorem 3.2. □

C.2 Connection with LOSS attack

Note that while there are additional terms in our optimal membership-inference score, there is another critical difference: the loss function has its sign flipped when compared to existing results (Yeom et al., 2018; Sablayrolles et al., 2019). While this may seem counter-intuitive at first glance, we show below the addition $-(I_1+I_2+I_3+I_4)$ terms in Equation 7 are expected to be negatively correlated to the loss function, leading to the proposed scoring function, in fact, aligns with the intuition of existing results. For simplicity, we consider the setting without regularization (i.e., $\alpha = 0$) and the Hessian matrix has full rank.

According to the assumption of quadratic loss around $\boldsymbol{w}^*$, we have the following observations:

$$L(\boldsymbol{w}) - L_* = \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}^*)^\top \mathbf{H}_*(\boldsymbol{w}-\boldsymbol{w}^*) = \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}^*)^\top \mathbf{U}^\top \mathrm{diag}\{\sigma_1,\dots,\sigma_d\}\mathbf{U}(\boldsymbol{w}-\boldsymbol{w}^*),$$

where $\mathbf{U}^\top \mathrm{diag}\{\sigma_1,\dots,\sigma_d\}\mathbf{U}$ is the eigenvalue decomposition of $\mathbf{H}$. Let $\boldsymbol{v} = \mathbf{U}(\boldsymbol{w}-\boldsymbol{w}^*)$. Since $\mathbf{U}$ is an orthonormal matrix, we know $\|\boldsymbol{v}\|_2 = \|\boldsymbol{w}-\boldsymbol{w}^*\|_2$. Thus, we obtain

$$\sigma_d \cdot \|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2 \le \sigma_j \cdot \|\boldsymbol{v}\|_2^2 = 2(L(\boldsymbol{w}) - L_*) = \sum_{j=1}^{d} \sigma_j \cdot v_j^2 \le \sigma_1 \cdot \|\boldsymbol{v}\|_2^2 = \sigma_1 \cdot \|\boldsymbol{w}-\boldsymbol{w}^*\|_2^2,$$

which further suggests (provided the Hessian has full rank)

$$\frac{1}{\sigma_1} \le \frac{\|\boldsymbol{w}-\boldsymbol{w}_i^*\|_2^2}{2(L_i(\boldsymbol{w}) - L_*)} \le \frac{1}{\sigma_d} \quad \text{for any } i \in \{0,1\}. \tag{17}$$

Based on Assumption 3, we assume that the Hessian structure and the loss function value remain unchanged with and without a single record $\boldsymbol{z}_1$. We hypothesize that the ratio $\frac{1}{k} = \frac{\|\boldsymbol{w}-\boldsymbol{w}_i^*\|_2^2}{2(L_i(\boldsymbol{w})-L_*)}$ also remains similar for $i=0$ and $i=1$, where $k \in [\sigma_d, \sigma_1]$. Therefore, we have

$$\frac{2\ell(\boldsymbol{w},\boldsymbol{z}_1)}{n\sigma_1} \le \|\boldsymbol{w}-\boldsymbol{w}_1^*\|_2^2 - \|\boldsymbol{w}-\boldsymbol{w}_0^*\|_2^2 \le \frac{2\ell(\boldsymbol{w},\boldsymbol{z}_1)}{n\sigma_d}. \tag{18}$$

Note that the derivation from Equation 17 to Equation 18 is not mathematically rigorous, but as long as the record $\boldsymbol{z}_1$ is not too deviated from the data distribution, we expect the above inequality holds. Plugging Equation 18 into the log-likelihood term inside $\mathcal{M}(\boldsymbol{w},\boldsymbol{z}_1)$ (first equality in Equation 15 with $\alpha = 0$), we get

$$\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{w}_1^*) - \ln p(\boldsymbol{w}|\boldsymbol{w}_0^*) &= \frac{S(1-\mu)}{2L_*}\left(-\frac{1}{\lambda}(\|\boldsymbol{w}-\boldsymbol{w}_1^*\|_2^2 - \|\boldsymbol{w}-\boldsymbol{w}_0^*\|_2^2) + \frac{\ell(\boldsymbol{w},\boldsymbol{z}_1)}{(1+\mu)n}\right) \\
&= \frac{S(1-\mu)}{2L_*}\left(\frac{1}{1+\mu} - \frac{2}{\lambda k}\right)\frac{\ell(\boldsymbol{w},\boldsymbol{z}_1)}{n},
\end{aligned} \tag{19}$$

where k is some real number that falls into $[\sigma_d, \sigma_1]$. In addition, for the case of full-rank Hessian, it is easy to see that the i-th eigenvalue of $\boldsymbol{\Sigma}$ (Equation (4)) can be written as:

$$\underbrace{\frac{S(1-\mu)}{2\lambda L_*}}_{\text{positive}}\left(2 - \frac{\lambda\sigma_i}{1+\mu}\right)^{-1}.$$

Since the covariance matrix $\boldsymbol{\Sigma}$ is positive semi-definite and invertible, it must follow that all of its eigenvalues are positive:

$$2 - \frac{\lambda\sigma_i}{1+\mu} > 0 \quad \Rightarrow \quad \frac{1}{1+\mu} - \frac{2}{\lambda\sigma_i} < 0, \quad \text{for any } i = 1,2,\dots,d.$$

With the above inequalities in mind, by looking at Equation (19), we get:

$$\ln p(\boldsymbol{w}|\boldsymbol{w}_1^*) - \ln p(\boldsymbol{w}|\boldsymbol{w}_0^*) = \underbrace{\frac{S(1-\mu)}{2nL_*}}_{>0}\underbrace{\left(\frac{1}{1+\mu} - \frac{2}{\lambda k}\right)}_{<0}\frac{\ell(\boldsymbol{w},\boldsymbol{z}_1)}{n}. \tag{20}$$

The upper limit on the score (hence the score itself) corresponding to IHA is thus proportional to the negative of the loss function, aligning with intuition (lower loss indicative of overfitting, and thus the record being a member). The score inside IHA can thus be interpreted (up to some approximation error) as $-f(\boldsymbol{w}, \boldsymbol{z}_1)\ell(\boldsymbol{w}, \boldsymbol{z}_1)$ for some $f(\boldsymbol{w}, \boldsymbol{z}_1) > 0$ that essentially accounts for SGD training dynamics, and is a function dependent on parameter access to the target model.

D Purchase-100(S) v/s Purchase-100

Model trainers, under practical settings, would not want to produce sub-optimal models. Under the given experimental settings (access to Purchase100 dataset), it is thus crucial to simulate model training setups that would maximize performance. The switch from Purchase-100(S) to Purchase-100 not only improves model performance but also reduces the performance of MIA attacks (Table 7).

Table 7: Performance of various attacks on Purchase-100(S) and Purchase-100. There is a clear drop in performance when shifting from Purchase-100(S) to Purchase-100. Statistics for IHA (CG) are computed on 10000 samples, not the entire dataset.

Attack	Purchase-100			Purchase-100(S)					
				MLP-2			MLP-4		
	AUC	% TPR@FPR		AUC	% TPR@FPR		AUC	% TPR@FPR	
		1%	0.1%		1%	0.1%		1%	0.1%
LOSS	.529 $\pm$.001	0.97	0.00	.589 $\pm$.003	1.04	0.00	.666 $\pm$.005	0.00	0.00
LiRA	.634 $\pm$.003	4.70	0.98	.743 $\pm$.006	9.56	2.79	.843 $\pm$.004	25.17	9.09
IHA (Ours)	.703 $\pm$.004	13.69	7.52	.791 $\pm$.003	19.95	18.99	—	—	—
IHA (CG)	.701 $\pm$.009	13.72	7.55	.791 $\pm$.005	20.09	19.09	.691 $\pm$.007	1.14	0.16

For instance, AUC values for LOSS drop from ~ 0.59 to ~ 0.53, and LiRA-Online from ~ 0.74 to ~ 0.63. While the smaller version of the dataset has recently been argued not to be very relevant Carlini et al. (2022), we believe this larger version is still interesting to study since such large datasets are practically relevant. We hope that researchers will aim to use the larger version of the dataset and, in general, train target models to maximize performance (as any model trainer would) within the constraints of their experimental design.

D.1 IHA performance on MLP-4

For the MLP-4 architecture on the Purchase-100(S) dataset, we observe a significant drop in performance when using IHA compared to LiRA. We hypothesize that this performance degradation may be due to a mismatch between the damping factor ϵ used in our experiments ($2e^{-1}$) and the optimal damping factor for the eigenvalue distribution of this larger model. To test this hypothesis, we increased ϵ to $5e^{-1}$ and repeated the evaluation with IHA.

This adjustment led to a substantial improvement in performance: the AUC increased to 0.768, with TPRs of 13.11% and 12.12% at 1% and 0.1% FPR, respectively. This result supports our hypothesis and highlights a current limitation of IHA. Specifically, the damping factor ϵ should be scaled according to the model architecture—ideally determined by some percentile of the eigenvalues. However, identifying the optimal scaling method before conducting the audit remains an open question for future work.

It is important to note that further increasing ϵ does not result in a linear performance improvement; in fact, performance declines across architectures when ϵ becomes too large, which is expected. The remaining performance gap is likely due to the absence of reference models or a violation of our assumptions about Hessian behavior. However, the superior TPR at very low (0.1%) FPRs, compared to LiRA, suggests that the former is more likely the cause.

E Implementing the Inverse Hessian Attack

For some given record $\boldsymbol{z}_1$, $\nabla L_0(\boldsymbol{w})$ can be computed by considering all data (except the target record) for which membership is known. To make this step computationally efficient for an audit, we pre-compute $\nabla L_1(\boldsymbol{w})$. Then, if the test record is indeed a member, we can compute $\nabla L_0(\boldsymbol{w})$ as $\nabla L_1(\boldsymbol{w}) - \frac{\nabla \ell(\boldsymbol{w}, \boldsymbol{z}_1)}{n}$. Note that this is equivalent to computing $\nabla L_0(\boldsymbol{w})$ separately for each target record. The Hessian $\mathbf{H}_*$ is also similarly pre-computed using the model's training data.

Conditioning $\mathbf{H}_*$. While computing Hessian matrices for our experiments, we notice the presence of near-zero and small, negative eigenvalues (most of which are likely to arise from precision errors). Such eigenvalues make the Hessian ill-conditioned and thus cannot be inverted directly. We explore two different techniques to mitigate this: damping by adding a small constant ϵ to all the eigenvalues or a low-rank approximation where only eigenvalues (and corresponding eigenvectors) above a certain threshold ϵ are used as a low-rank approximation. We ablate over these two techniques for some candidate values of ϵ. Our results (Table 8) suggest that damping with $\epsilon = 2e^{-1}$ works best across all the datasets we test, which is the setting for which we report our main results.

Table 8: Attack AUCs for various techniques to mitigate ill-conditioned Hessian matrix, with corresponding ϵ values.

Dataset	Low-Rank			Damping		
	$\epsilon = 1e^{-2}$	$\epsilon = 1e^{-1}$	$\epsilon = 2e^{-1}$	$\epsilon = 1e^{-2}$	$\epsilon = 1e^{-1}$	$\epsilon = 2e^{-1}$
MNIST-Odd	.521	.530	.500	.513	.535	.542
FashionMNIST	.551	.557	.541	.533	.582	.594

F Approximating L_0

For auditing purposes, experiments where all but one member is known are useful, but an adversary is unlikely to have this much knowledge of the training data. We experiment with the potential use of IHA where only partial knowledge of the remaining $n-1$ members may be available to approximate ∇L_0, Approximating L_0 with a fraction of the actual dataset could be useful in not only reducing the computational cost of the audit, but also potentially enabling adversarial use of the attack in threat models where the attacker has partial knowledge of the training data. We evaluate IHA for versions where L_0 is approximated using a randomly-sampled fraction of the training data and report results in Table 9.

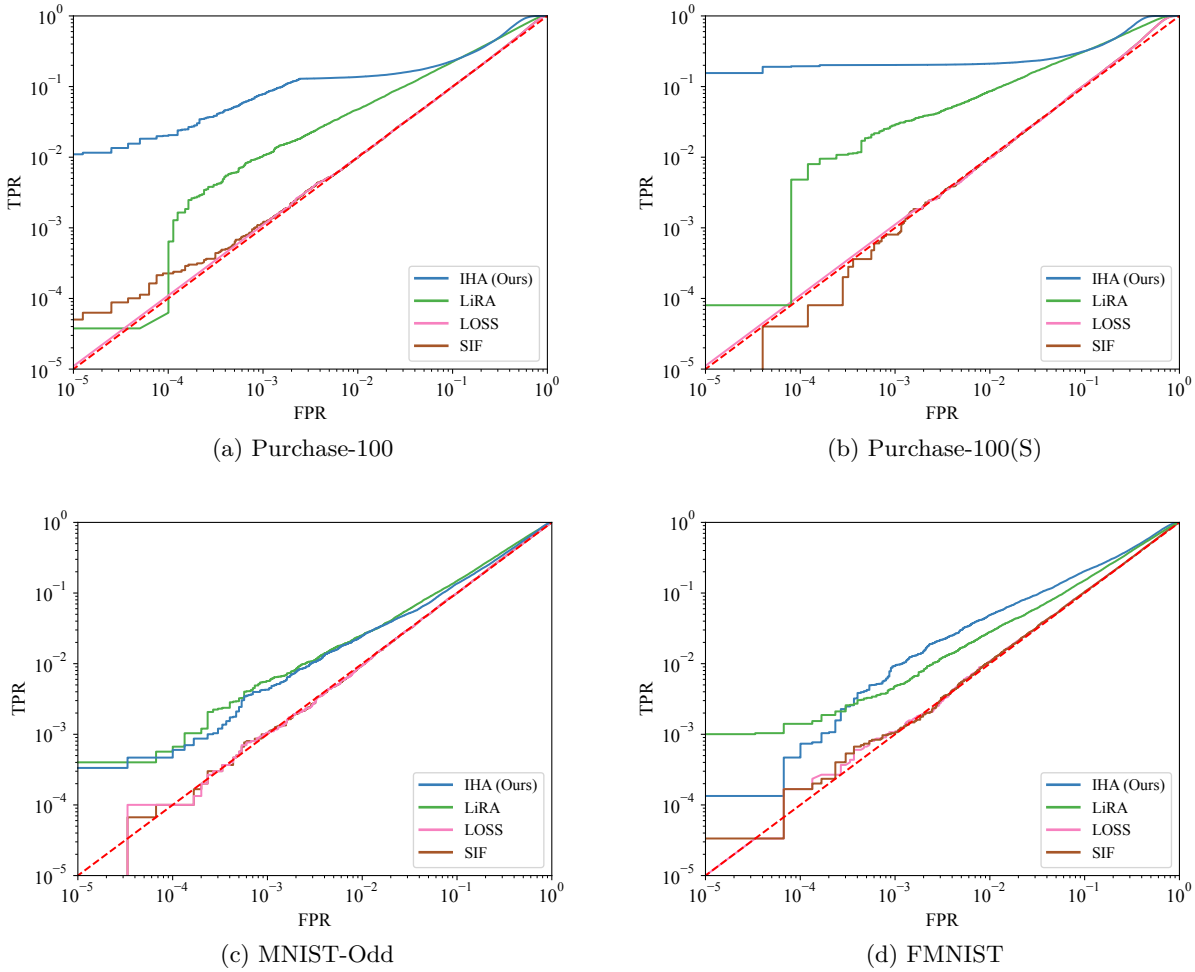
Table 9: Performance of approximation-based variant of IHA on Purchase100-S (MLP-2), when a fraction of data from $D \setminus \{z_1\}$ is used to approximate L_0. Statistics are computed on 10000 samples.

Fraction	AUC	%TPR@FPR	
		1%	0.1%
0.2	.577 $\pm$.005	0.85	0.11
0.4	.607 $\pm$.007	1.79	0.28
0.6	.638 $\pm$.010	3.38	0.86
0.8	.692 $\pm$.012	6.47	1.59
0.9	.733 $\pm$.010	13.53	4.17
1.0	.791 $\pm$.005	20.09	19.09

We see a clear degradation in performance when only a subset of data is used–this is especially true for lower fractions, where AUC can drop by as much as ≈ 0.2. More importantly, even when using 90% of the training data, there is a significant gap in performance. The statistics we compute for IHA thus do completely utilize knowledge of all other training records. While this result suggests that adversarial use of IHA would require

a very strong adversary (that posseses knowledge of nearly all training records), it also hints at how data poisoning attacks could have a large impact on downstream membership inference. A poisoning adversary could hypothetically craft data a way that interferes with L_0 (when relating to the optimal membership adversary) and thus increase/decrease inference risk for other records.

G Additional Results



(a) Purchase-100

(b) Purchase-100(S)

(c) MNIST-Odd

(d) FMNIST

Figure 1: ROC curves for low-FPR region for various attacks and datasets.