Do Parameters Reveal More than Loss for Membership Inference?

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2024

Abstract

Membership inference attacks aim to infer whether an individual record was used to train a model, serving 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 most useful settings such as stochastic gradient descent, and that optimal membership inference indeed requires white-box access. We validate our findings with 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 audits 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 [20]. 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 (DP) [10], that bound the potential effectiveness of any attack of a certain type. 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 around better attacks, while theoretical proofs require many assumptions or result in loose bounds. On the other hand, empirical audits provide a more meaningful bound than is possible with theory or experiments alone. 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. Since auditors have elevated model access (via associated training environments, data, etc.), they can take advantage of more information to produce better estimates of what an adversary may be able to do without that information.

The most common disclosure auditing approach today is to conduct membership inference attacks [13] as well as related attacks that attempt to extract specific data [8]. Membership inference privacy poses the question of being able to determine the inclusion of a given record in a model's training data. Membership inference is popular in the research and industrial privacy communities

because it is simple to define, relatively easy to measure, and aligns well with DP. This has resulted in it being widely used as a proxy for auditing privacy leakage for machine learning [2, 11, 13, 30]. 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 largely been reinforced by folklore and results demonstrating negligible gains from parameter access (white-box attacks) [5, 17]. A well-known result by Sablayrolles et al. [19], in fact, theoretically proves that black-box access is sufficient for optimal membership inference. However, the assumptions made in its derivation do not hold for most trained models, including ones trained with stochastic gradient descent (SGD).

Contributions. Utilizing recent advances in the discrete-time SGD-dynamics literature [16, 33], we provide a more accurate formulation of the optimal membership inference attack that invalidates previous claims [19] about black-box access being optimal (Section 3). Our theoretical result also prescribes such an attack, which we call IHA (Inverse Hessian Attack) and empirically demonstrate its effectiveness in simple settings (Section 3.4). Motivated by our findings, we advocate for further research in white-box inference attacks (Section 4).

2. Preliminaries

Membership Inference. Following the framework established in [19], let $\mathcal{D}$ be a 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 variable of $\boldsymbol{z}_i$ (i.e., $m_i = 1$ if $\boldsymbol{z}_i \in D$, and $m_i = 0$ otherwise). Given $\boldsymbol{w}$, an *membership inference attack* aims to predict the unknown membership m_i for any given record $\boldsymbol{z}_i$.

Definition 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. [19] derived the following explicit formula for $\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1)$.

Lemma 2 (Sablayrolles et al. [19]) *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, one need to characterize the posterior $\mathbb{P}(\boldsymbol{w} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, m_1, \ldots, m_n)$ to explicit out the effect of the inferred record $\boldsymbol{z}_1$ on the optimal membership inference $\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1)$. This is where recent advances in discrete-time SGD dynamics [16, 33] literature can help provide a connection with model parameters. Due to the limited space, we defer the introduction of additional preliminary theoretical results about discrete-time SGD dynamics to Appendix B.

3. Black-Box Access is not Sufficient

3.1. Limitations of Claims of Black-Box Optimality

Sablayrolles et al. [19] proved the optimality of black-box membership inference under a Bayesian framework. In particular, they assume (Equation 1 in [19]) 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:

$$\mathbb{P}(\boldsymbol{w} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \propto \exp\left(-\frac{1}{T} \sum_{i=1}^{n} m_i \cdot \ell(\boldsymbol{w}, \boldsymbol{z}_i) \right), \tag{2}$$

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 2 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, and is shown achievable using SGLD [25] 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. [23] 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 11-12 in [23]), 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 [34], especially under the discrete-time setting [16].

3.2. Optimal Membership Inference under Discrete-time SGD

So far, we have explained why the critical assumption imposed in [19] about the posterior distribution of $\boldsymbol{w}$ following a Boltzmann distribution (Equation 2) does not hold for typical stochastic gradient methods employed in practice. We now derive the optimal membership inference for models produced by SGD by leveraging the recent theoretical literature on discrete-time SGD dynamics [16, 33]. Specifically, we consider an SGD algorithm with the following update rule: for $t = 1, 2, 3, \ldots$

$$\boldsymbol{g}_t = \nabla L(\boldsymbol{w}_{t-1}) + \boldsymbol{\eta}_{t-1}, \quad \boldsymbol{h}_t = \mu \cdot \boldsymbol{h}_{t-1} + \boldsymbol{g}_t, \quad \boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \lambda \cdot \boldsymbol{h}_t. \tag{3}$$

Here, $\mu \in [0, 1)$ is the momentum, $\lambda > 0$ is the learning rate, and $\boldsymbol{\eta}_t = \frac{1}{S} \sum_{i \in \mathcal{B}_t} \nabla \ell(\boldsymbol{w}_{t-1}, \boldsymbol{z}_i) - \nabla L(\boldsymbol{w}_{t-1})$ stands for the mini-batch noise, where $\mathcal{B}_t$ is a randomly sampled batch of examples with size S from D, $\ell(\boldsymbol{w}, \boldsymbol{z})$ is the individual loss, and $L(\boldsymbol{w}) = \frac{1}{n} \sum_{\boldsymbol{z} \in D} \ell(\boldsymbol{w}, \boldsymbol{z})$ denotes the total loss.

In addition, we assume that the loss achieved at the local minimum remains unaffected by the removal of a single training record and that the Hessian structure remains unchanged.

Assumption 1 (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 $L_1(\boldsymbol{w}) = \frac{1}{n}(\ell(\boldsymbol{w}, \boldsymbol{z}_1) + \sum_{i=2}^{n} m_i \ell(\boldsymbol{w}, \boldsymbol{z}_i))$. Assume the Hessian matrix shares a similar structure when the model's training data differs by a single point, and the loss function also achieves a similar value at the local minimum, i.e.,*

$$\mathbf{H}_* = \mathbf{H}_1(\boldsymbol{w}_1^*) = \mathbf{H}_0(\boldsymbol{w}_0^*), \quad L_* = L_1(\boldsymbol{w}_1^*) = L_0(\boldsymbol{w}_0^*), \tag{4}$$

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 too deviated from the data distribution $\mathcal{D}$, we expect Assumption 1 generally holds for SGD algorithms. Under Assumption 1 and a few other assumptions imposed in prior literature on discrete-time SGD dynamics [16, 33], we obtain the following form for the optimal membership-inference adversary.

Theorem 3 (Optimal Membership-Inference Score) *Given $\boldsymbol{w}$ produced by an SGD algorithm defined by Equation 3 and a record $\boldsymbol{z}_1$, the optimal membership inference $\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1)$ is given by:*

$$\mathbb{E}_{\mathcal{T}}\left[\sigma\left(\frac{S(1-\mu)}{2nL_*} \cdot \left(\frac{\ell(\boldsymbol{w}, \boldsymbol{z}_1)}{1+\mu} - I_1 - I_2\right) + \ln\left(\frac{\gamma}{1-\gamma}\right)\right)\right], \tag{5}$$

where I_1 and I_2 are defined as follows:

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

Here, L_* and $\mathbf{H}_*$ are defined in Assumption 1, which are dependent on $\mathcal{T}$. The proof of Theorem 3 and the complete set of required assumptions are presented in Appendix B-D. 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 and momentum μ are also required, thus requiring complete knowledge of the training setup of the target model. Black-box access is thus **not** optimal for membership inference. The 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 $\boldsymbol{z}_1$.

3.3. Inverse Hessian Attack

While Theorem 3 directly describes the optimal membership inference adversary, computing the expectation over $\mathcal{T}$ is infeasible. We thus make use of the insight of Theorem 3 to propose a scoring function based on the terms inside the expectation:

$$\text{IHA}(\boldsymbol{z}_1) := \frac{\ell(\boldsymbol{w}, \boldsymbol{z}_1)}{1+\mu} - \frac{1}{\lambda}\left(\frac{1}{n}\|\mathbf{H}_*^{-1}\nabla\ell(\boldsymbol{w}, \boldsymbol{z}_1)\|^2 + 2\big(\mathbf{H}_*^{-1}\nabla L_0(\boldsymbol{w})\big)^\top\big(\mathbf{H}_*^{-1}\nabla\ell(\boldsymbol{w}, \boldsymbol{z}_1)\big)\right). \tag{6}$$

This score $\text{IHA}(\boldsymbol{z}_1)$, for some given record $\boldsymbol{z}$, can be used as the probability of $\boldsymbol{z}_1$ being a member and subsequently serve as a useful attack for privacy auditing. While the absence of a negative sign with the loss function (like in LOSS) in $\text{IHA}(\boldsymbol{z}_1)$ may seem counter-intuitive at first glance, it can be rewritten such that it is proportional to the negation of the loss function (Appendix D.2). While the optimal attack prescribed by our theory requires expectation over all possible $\mathcal{T}$, we directly use

Table 1: Performance of various attacks, reported via attack AUC and true positive rate (TPR) at low false positive rate (FPR). For corresponding ROC curves, see Appendix F.1.

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 [29]	$.531_{\pm.001}$	.100	.010	$.500_{\pm.002}$	.100	.010	$.507_{\pm.002}$	.099	.010
SIF [7]	$.530_{\pm.001}$	.100	.010	$.500_{\pm.002}$	.100	.010	$.507_{\pm.002}$	.099	.010
LiRA [5]	$.645_{\pm.003}$	.221	.048	$\mathbf{.569}_{\pm.005}$	**.156**	**.028**	$.581_{\pm.021}$	.166	.108
Reference [19]	$.615_{\pm.003}$	.198	.039	$.525_{\pm.001}$	.126	.018	$.532_{\pm.009}$	.126	.016
IHA (Ours)	$\mathbf{.709}_{\pm.008}$	**.254**	**.154**	$.538_{\pm.009}$	.132	.025	$\mathbf{.588}_{\pm.012}$	**.180**	**.036**

$\text{IHA}(z_1)$ without any reference models[1]. Apart from the absence of reference models to compute this expectation, the performance of our attack 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 1).

3.4. Experiments

We evaluate IHA over multiple datasets with models where the Hessian can be computed directly, with some damping applied to deal with near-zero eigenvalues (Appendix E.2). We evaluate our attack on three different datasets across linear regression and 2-layer neural networks. See Appendix E.1 for details. As visible in Table 1, IHA provides a strong privacy auditing baseline at par with current state-of-the-art attacks that use reference models and outperforms baselines in most cases.

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 $D \setminus \{z_1\}$ for inferring a given target record z_1. 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. Conclusion

Our theoretical result proves that model parameter access is indeed helpful for membership inference, contrary to previous results and the common belief that parameter access is not beneficial. We propose a corresponding attack inspired by this theory (Inverse Hessian Attack) that provides stronger privacy auditing than existing black-box techniques. However, it should be noted that IHA is not practically realizable for most settings due to the computational expense of calculating the Hessian. Attempts to approximate iHVPs [1, 12, 18] can introduce errors that negatively impact performance. Our conclusion aligns well with recent calls in the literature to consider white-box access for rigorous auditing [6]. 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.

1. Adapting our attack to work with reference models in the general setting remains a future interesting direction.

References

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

[2] Achraf Azize and Debabrota Basu. How Much Does Each Datapoint Leak Your Privacy? Quantifying the Per-datum Membership Leakage. *arXiv:2402.10065*, 2024.

[3] 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.

[4] 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.

[5] 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.

[6] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. *arXiv:2401.14446*, 2024.

[7] 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.

[8] 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.

[9] 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.

[10] 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*, pages 265–284. Springer, 2006.

[11] 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.

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

[13] 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.

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

[15] 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.

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

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

[18] Elre T Oldewage, Ross M Clarke, and José Miguel Hernández-Lobato. Series of hessian-vector products for tractable saddle-free newton optimisation of neural networks. *arXiv:2310.14901*, 2023.

[19] 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.

[20] 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.

[21] Issei Sato and Hiroshi Nakagawa. Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Ito Process. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 982–990, Bejing, China, June 2014. PMLR. URL https://proceedings.mlr.press/v32/satoa14.html. Issue: 2.

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

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

[24] 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.

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

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

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

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

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

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

Appendix A. Related Works

A.1. Membership Inference

Black-box Membership Inference. Early works on membership inference worked under black-box access, utilizing the model's loss [22] 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 it being present in train data. This calibration has taken several forms; direct score normalization with reference models [19], likelihood tests based on score distributions [5, 27, 32], and additional models for predicting difficulty [3].

White-box Membership Inference. Nasr et al. [17] 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 [9], 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 and Giryes [7] 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 [1] to approximate the iHVP. This has similarities to our result since our optimal membership inference score also involves computing iHVPs. Li et al. [15] 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. [28] 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 [27]). 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. [4] studied the problem of forecasting memorization in a model for specific training data. The authors propose 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. [24] 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. [23] approximated the SGD dynamics as an Ornstein-Uhlenbeck process, while Yokoi and Sato [31] 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. [16] 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. [33] further generalized the results of [16] 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 [12] or LiSSA [1]. 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 [12] and optimization [18], 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. [18] proposes another method for efficient iHVP approximation as a subroutine, but the authors observed high approximation errors based on both norm and direction.

Appendix B. Additional Preliminaries

Discrete-time SGD Dynamics. A line of theoretical work [16, 21, 23, 25, 33] has analyzed the continuous- and discrete-time dynamics of stochastic gradient methods and provided insights for understanding deep learning generalization. In particular, assuming the model is trained using SGD according to the update rule defined by Equation 3 on a *quadratic loss* with the Hessian matrix $\mathbf{H}$ and arrives at a *stationary state*, Liu et al. [16] established a connection, shown in the following theorem, between the Hessian matrix $\mathbf{H}$, the asymptotic noise covariance $\mathbf{C} = \lim_{t\to\infty} \mathbb{E}_{\boldsymbol{w}_t}[\mathrm{cov}(\boldsymbol{\eta}_t, \boldsymbol{\eta}_t)]$, and the asymptotic model fluctuation $\boldsymbol{\Sigma} = \lim_{t\to\infty} \mathrm{cov}(\boldsymbol{w}_t, \boldsymbol{w}_t)$. To be more specific, we begin by reinstating their assumptions.

Assumption 2 (Stationary-State) *After a sufficient number of iterations, models trained with SGD defined by the update rule in Equation 3 arrive at a stationary state, i.e., the stationary model fluctuation $\boldsymbol{\Sigma}$ exists and is finite.*

Assumption 3 (Quadratic Loss) *The loss function $L(\boldsymbol{w})$ is either globally quadratic or locally quadratic close to a local minimum $\boldsymbol{w}^*$. Specifically, the loss function can be approximated as:*

$$L(\boldsymbol{w}) = L(\boldsymbol{w}^*) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^\top \mathbf{H}(\boldsymbol{w}^*)(\boldsymbol{w} - \boldsymbol{w}^*) + o(\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2), \tag{7}$$

where $\boldsymbol{w}^*$ is a local minimum and $\mathbf{H}(\boldsymbol{w}^*)$ denotes the Hessian matrix at $\boldsymbol{w}^*$.

Theorem 4 (SGD Stationary distribution with momentum) *Let $\boldsymbol{w}$ be updated with SGD defined by the update rule in Equation 3 with momentum $\mu \in [0, 1)$. Assumptions 2 and 3, if we additionally suppose $\mathbf{C}$ commutes with $\mathbf{H}(\boldsymbol{w}^*)$, then the stationary model fluctuation satisfies:*

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

Theorem 4 requires the existence of a finite stationary noise covariance and the loss function to be quadratic close to a local minimum, which are mild assumptions (see [16] for detailed discussions). In follow-up work, Ziyin et al. [33] further derived the explicit dependence of the asymptotic noise covariance $\mathbf{C}$ to the loss and Hessian around a local minimum $\boldsymbol{w}^*$ under mild assumptions.

Theorem 5 (SGD Noise Covariance) *Let $L(\boldsymbol{w})$ be the training loss and the model $\boldsymbol{w}$ is optimized with SGD defined by Equation 3 in the neighborhood of a local minimum $\boldsymbol{w}^*$. If $L(\boldsymbol{w}^*) \neq 0$, then*

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

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

Theorem 5 implies that the SGD noise covariance $\mathbf{C}$ commutes with the Hessian matrix $\mathbf{H}(\boldsymbol{w}^*)$. If only considering the leading term in the noise covariance, then one can immediately derive the following formula for the stationary model fluctuation of SGD based on the above two theorems:

$$\boldsymbol{\Sigma} = \frac{2\lambda L(\boldsymbol{w}^*)}{S(1-\mu)} \cdot \left[2\mathbf{I}_d - \frac{\lambda \mathbf{H}(\boldsymbol{w}^*)}{1+\mu}\right]^{-1}. \tag{8}$$

We remark that if $L(\boldsymbol{w}^*) = 0$ (i.e., $\boldsymbol{w}^*$ is a global minimum), then $\boldsymbol{\Sigma} = 0$. In addition, if the Hessian matrix $\mathbf{H}(\boldsymbol{w}^*)$ has degenerated rank $r < d$, then $\mathbf{H}^{-1}$ can be replaced by the Moore-Penrose pseudo inverse $\mathbf{H}$. Accordingly, similar results to Equation 8 can be obtained by considering the projection space spanned by eigenvectors with non-zero eigenvalues. We refer to Section 5 of [33] to readers for more detailed discussions of the imposed assumptions and the implications of the results.

Appendix C. Posterior for SGD

To prove our main theoretical result of optimal membership inference, we first need to characterize the analytical form of the posterior distribution with respect to model parameters trained with SGD.

Theorem 6 (Posterior for SGD) *Assume the same assumptions as used in Theorems 4 and 5. Let $\boldsymbol{w}^*$ be the local minimum that SGD (Equation 3) is converging towards. Then, the (conditional) log-probability of observing parameters $\boldsymbol{w}$ is given by (up to constants and negligible terms):*

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

where $\sigma_i(\mathbf{H}(\boldsymbol{w}^*))$ denotes the i-th largest eigenvalue of $\mathbf{H}(\boldsymbol{w}^*)$.

Proof According to Theorem 4 and Theorem 5, we obtain

$$\boldsymbol{\Sigma} = \frac{2\lambda L(\boldsymbol{w}^*)}{S(1-\mu)} \cdot \left[2\mathbf{I}_d - \frac{\lambda \mathbf{H}(\boldsymbol{w}^*)}{1+\mu}\right]^{-1}. \tag{9}$$

Note that the above equation holds when the Hessian matrix $\mathbf{H}(\boldsymbol{w}^*)$ has full rank and $L(\boldsymbol{w}^*) \neq 0$. When the Hessian has degenerated rank $r < d$, the following more generalized result can be derived:

$$\mathbf{P}_r\boldsymbol{\Sigma} = \frac{2\lambda L(\boldsymbol{w}^*)}{S(1-\mu)} \cdot \mathbf{P}_r\left[2\mathrm{I}_d - \frac{\lambda \mathbf{H}(\boldsymbol{w}^*)}{1+\mu}\right]^{-1},$$

where $\mathbf{P}_r = \mathrm{diag}(1, ..., 1, 0, ...0)$ denotes the projection matrix onto non-zero eigenvalues. If $L(\boldsymbol{w}^*) = 0$, meaning $\boldsymbol{w}^*$ is a global minimum, then the asymptotic model fluctuation $\boldsymbol{\Sigma} = \mathbf{0}$. For the ease of presentation, let $\mathbf{H}_* = \mathbf{H}(\boldsymbol{w}^*)$, $L_* = L(\boldsymbol{w}^*)$ and the Hessian matrix have full rank in the following proof. 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:

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

Here, the last equality holds because of the second-order Taylor expansion of $L(\boldsymbol{w})$ at $\boldsymbol{w}^*$. Omitting the constant and negligible terms in the above equation, we thus complete the proof of Theorem 6. ■

Appendix D. Optimal Membership-Inference Score

D.1. Proof for Theorem 3

Proof To derive the 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 6. Therefore, the remaining task is to analyze the following terms:

$$\begin{aligned}
&\ln p(\boldsymbol{w}|\boldsymbol{w}_1^*) - \ln p(\boldsymbol{w}|\boldsymbol{w}_0^*) \\
&= -\frac{D}{2}\left[\ln L_1(\boldsymbol{w}_1^*) - \ln L_0(\boldsymbol{w}_0^*)\right] + \frac{1}{2}\sum_{i=1}^{D}\ln\left(\frac{2 - \frac{\lambda}{1+\mu}\sigma_i(\mathbf{H}_1(\boldsymbol{w}_1^*))}{2 - \frac{\lambda}{1+\mu}\sigma_i(\mathbf{H}_0(\boldsymbol{w}_0^*))}\right) \\
&\quad - \frac{S(1-\mu)}{2\lambda}\cdot\left(\frac{\|\boldsymbol{w}-\boldsymbol{w}_1^*\|_2^2}{L_1(\boldsymbol{w}_1^*)} - \frac{\|\boldsymbol{w}-\boldsymbol{w}_0^*\|_2^2}{L_0(\boldsymbol{w}_0^*)}\right) + \frac{S(1-\mu)}{2(1+\mu)}\cdot\left(\frac{L_1(\boldsymbol{w})}{L_1(\boldsymbol{w}_1^*)} - \frac{L_0(\boldsymbol{w})}{L_0(\boldsymbol{w}_0^*)}\right), \quad (10)
\end{aligned}$$

where the constant and small $o(\cdot)$ terms are neglected, and $\mathbf{H}_0(\boldsymbol{w}_0^*)$ (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 1, 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 10 and obtain the following form:

$$
\begin{aligned}
&-\frac{S(1-\mu)}{2\lambda L_*}\big(\|\boldsymbol{w}-\boldsymbol{w}_1^*\|_2^2 - \|\boldsymbol{w}-\boldsymbol{w}_0^*\|_2^2\big) + \frac{S(1-\mu)}{2(1+\mu)L_*}\big(L_1(\boldsymbol{w}) - L_0(\boldsymbol{w})\big) \\
&= -\frac{S(1-\mu)}{2\lambda L_*}\big(\nabla L_1(\boldsymbol{w})^\top \mathbf{H}_*^{-1}\mathbf{H}_*^{-1}\nabla L_1(\boldsymbol{w}) - \nabla L_0(\boldsymbol{w})^\top \mathbf{H}_*^{-1}\mathbf{H}_*^{-1}\nabla L_0(\boldsymbol{w})\big) + \frac{S(1-\mu)\ell(\boldsymbol{w},\boldsymbol{z}_1)}{2n(1+\mu)L_*} \\
&= -\frac{S(1-\mu)}{2\lambda L_* n}\left(2\nabla L_0(\boldsymbol{w})^\top \mathbf{H}_*^{-1}\mathbf{H}_*^{-1}\nabla \ell(\boldsymbol{w},\boldsymbol{z}_1) + \frac{1}{n}\|\mathbf{H}_*^{-1}\nabla\ell(\boldsymbol{w},\boldsymbol{z}_1)\|^2\right) + \frac{S(1-\mu)\ell(\boldsymbol{w},\boldsymbol{z}_1)}{2n(1+\mu)L_*},
\end{aligned}
\tag{11}
$$

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, we know the optimal membership inference is given by:

$$
\mathcal{M}(\boldsymbol{w},\boldsymbol{z}_1) = \mathbb{E}_{\mathcal{T}}\left[\sigma\left(\ln\left(\frac{p(\boldsymbol{w}|m_1=1,\boldsymbol{z}_1,\mathcal{T})}{p(\boldsymbol{w}|m_1=0,\boldsymbol{z}_1,\mathcal{T})}\right) + \ln\left(\frac{\gamma}{1-\gamma}\right)\right)\right], \tag{12}
$$

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 11 into Equation 12, we obtain

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

where I_1, I_2 and t_γ are defined as:

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

Therefore, we complete the proof of Theorem 3. ■

D.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 [19, 29]. While this may seem counter-intuitive at first glance, we show below the addition $-(I_1+I_2)$ terms in Equation 5 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.

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

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

where $\mathbf{U}^\top \text{diag}\{\sigma_1, \ldots, \sigma_d\}\mathbf{U}$ 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

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

which further suggests (provided the Hessian has full rank)

$$\frac{1}{\sigma_1} \leq \frac{\|\boldsymbol{w} - \boldsymbol{w}_i^*\|_2^2}{2(L_i(\boldsymbol{w}) - L_*)} \leq \frac{1}{\sigma_d} \quad \text{for any } i \in \{0, 1\}. \tag{13}$$

Based on Assumption 1, 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

$$\frac{2\ell(\boldsymbol{w}, \boldsymbol{z}_1)}{n\sigma_1} \leq \|\boldsymbol{w} - \boldsymbol{w}_1^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}_0^*\|_2^2 \leq \frac{2\ell(\boldsymbol{w}, \boldsymbol{z}_1)}{n\sigma_d}. \tag{14}$$

Note that the derivation from Equation 13 to Equation 14 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 14 into the log-likelihood term inside $\mathcal{M}(\boldsymbol{w}, \boldsymbol{z}_1)$ (Equation 11), we get

$$\begin{aligned} \ln p(\boldsymbol{w}|\boldsymbol{w}_1^*) - \ln p(\boldsymbol{w}|\boldsymbol{w}_0^*) &= \frac{S(1-\mu)}{2L_*}\left(-\frac{1}{\lambda}(\|\boldsymbol{w} - \boldsymbol{w}_1^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}_0^*\|_2^2) + \frac{\ell(\boldsymbol{w}, \boldsymbol{z}_1)}{(1+\mu)n}\right) \\ &= \frac{S(1-\mu)}{2L_*} \cdot \left(\frac{1}{1+\mu} - \frac{2}{\lambda k}\right) \cdot \frac{\ell(\boldsymbol{w}, \boldsymbol{z}_1)}{n}, \end{aligned} \tag{15}$$

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 8) 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:

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

With the above inequalities in mind, by looking at Equation (15), we get:

$$\ln p(\boldsymbol{w}|\boldsymbol{w}_1^*) - \ln p(\boldsymbol{w}|\boldsymbol{w}_0^*) = \underbrace{\frac{S(1-\mu)}{2nL_*}}_{>0} \cdot \underbrace{\left(\frac{1}{1+\mu} - \frac{2}{\lambda k}\right)}_{<0} \cdot \frac{\ell(\boldsymbol{w}, \boldsymbol{z}_1)}{n}. \tag{16}$$

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.

Appendix E. Experimental Details

We train 128 models in the same way as prior literature [5], 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 (in expectation) include the target record in the training data, and the other half do not. All of our models are trained without momentum ($\mu = 0$). For a given FPR, a threshold is computed using scores for non-members, which is then used to compute the corresponding 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 common and standard for membership-inference evaluations [5, 27, 29].

E.1. Datasets and Models

MNIST-Odd. We consider the MNIST dataset [14], with the modified task of classifying a given digit image as odd or even. We train a logistic regression model with mean-squared error loss, with an average test loss of .078.

FashionMNIST. We use the FashionMNIST [26] 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%.

Purchase-100(S). The task for this dataset [22] 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 in the prior literature [32] train models on 25K samples from Purchase-100, which is much smaller than the actual dataset, which is why we term it Purchase-100(S) (Small).

Purchase-100. For this version, we train models with 80K 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.

E.1.1. 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 2).

For instance, AUC values for LOSS drop from ~ 0.59 to ~ 0.53, and LiRA-Online from ~ 0.73 to ~ 0.65. While the smaller version of the dataset has recently been argued not to be very relevant [5], 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.

E.2. IHA: Implementation Details

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,

Table 2: 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.

Attack	Purchase-100			Purchase-100(S)		
	AUC	TPR@FPR		AUC	TPR@FPR	
		1%	0.1%		1%	0.1%
LOSS [29]	$.531_{\pm.001}$	.100	.010	$.590_{\pm.004}$	.104	.010
SIF [7]	$.530_{\pm.001}$	.100	.010	$.590_{\pm.005}$	.104	.010
LiRA [5]	$.645_{\pm.003}$	.221	.048	$.742_{\pm.007}$	.330	.093
Reference [19]	$.615_{\pm.003}$	.198	.039	$.686_{\pm.005}$	.269	.067
IHA (Ours)	$.709_{\pm.008}$	.254	.154	$.801_{\pm.003}$	.329	.223

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 3) 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 3: 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	.514	.537	.538
FashionMNIST	.551	.557	.541	.520	.579	.588

E.3. Baseline Attacks

LOSS [29]. For this attack, the negative loss is used directly as a signal for membership inference.

SIF. [7] This attack, similar to ours, also utilizes the loss curvature of the target model by calculating its Hessian, which is then used to compute self-influence as a 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 immediately 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 utilize approximation methods for iHVP, we utilize the exact Hessian for fair comparison.

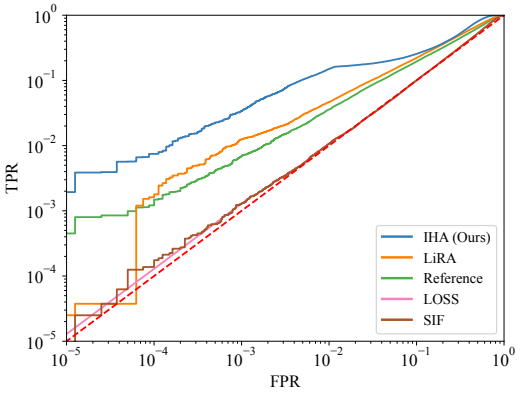
Reference [19]. The reference-based attack performs difficulty calibration based on the expected loss on the target record for models trained without the target record. This is achieved by averaging the loss using “offline" models.

LiRA [5]. There are two variants, LiRA-Offline and LiRA-Online. The former utilizes “offline" models to estimate a Gaussian distribution and then performs one-sided hypothesis testing using loss scores. The LiRA-Online variant additionally “online" models, i.e., models whose training data included 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 stronger of the two variants.

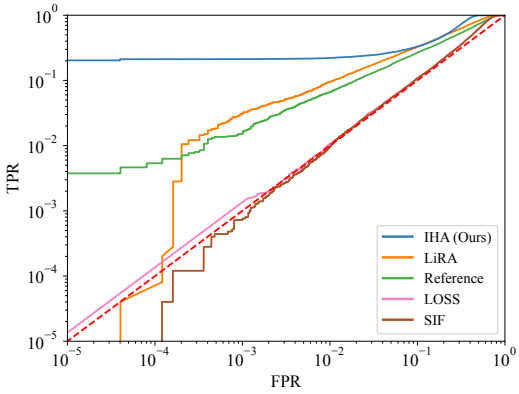
Appendix F. Additional Results

F.1. RoC Curves

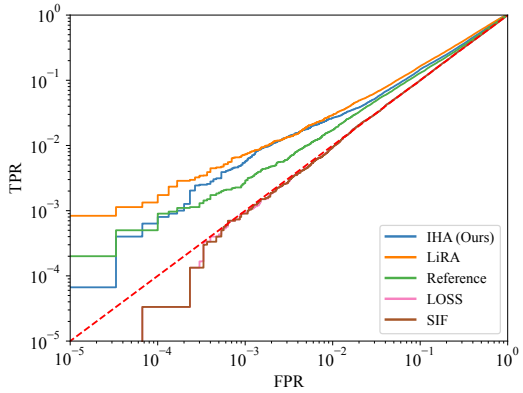
Here we report RoC curves for various attacks and datasets, focusing on the low FPR region.



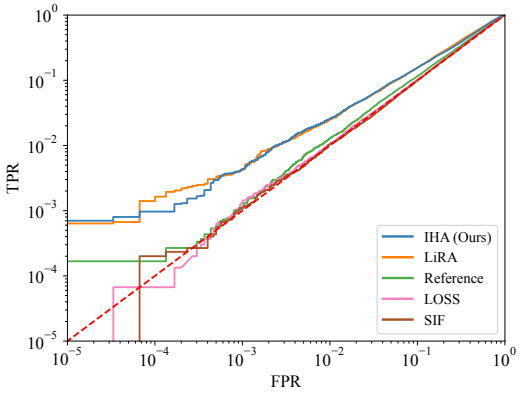
(a) Purchase-100



(b) Purchase-100(S)



(c) MNIST-Odd



(d) FMNIST

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