A Membership Inference Attack Framework for Diffusion Models by Fixed-point Iteration

Anonymous Authors[1]

Abstract

The data-intensive nature of Diffusion models amplifies the risks of privacy infringements and copyright disputes, particularly when training on extensive unauthorized data scraped from the internet. Membership Inference Attacks (MIA) aim to determine whether a data sample has been utilized by the target model during training, thereby serving as a pivotal tool for privacy preservation. Traditional MIA primarily employ the prediction loss to distinguish between training member samples and non-members. These methods assume that, compared to non-member samples, member samples, having been encountered by the model during training result in a smaller prediction loss. However, this assumption proves ineffective in diffusion models due to the frequent introduction of random noise during the training process. Rather than estimating the loss, our approach explores this random noise and formulate the MIA as a noise search problem. We propose a fixed-point iteration-based framework to optimize the noise search process. We demonstrate that recent MIA methods for diffusion models can be categorized within this framework, representing a form of single-step fixed-point iteration. Building upon this foundation, we propose the One More iteration Step (OMS) method, further expanding the distinctions between members and non-members. Experimental results indicate that our approach can significantly enhance the accuracy of existing MIA methods for diffusion models.

1. Introduction

Recently, Diffusion models (DMs) (Ho et al., 2020; Song et al., 2020) have been widely recognized for their unrivaled proficiency in generating images of exceptional quality, which are increasingly becoming indistinguishable from their real-world counterparts. Due to the high quality of images generated by diffusion models, more and more AI companies are developing diffusion model based generative tools for commercial art design.

Nonetheless, these models present a double-edged sword. The data-intensive nature of DMs has amplified the risk of privacy infringements and copyright disputes. Training on extensive unauthorized data scraped from the internet, these methods overlook the proprietary rights and privacy of the original owners. A case in point is the recent lawsuit filed by Getty Images against Stability AI, alleging unauthorized use of 12 million of Getty's images for model training (Brittain, 2023). Thus, it is imperative to develop effective tools to detect diffusion models' privacy infringements.

To address these privacy challenges, Membership Inference Attacks (MIA) (Hu et al., 2022) have emerged as a potential solution. The objective of MIA is to ascertain whether a data sample has been utilized in the training process of a machine learning model. Existing MIA methods (Sablayrolles et al., 2019; Salem et al., 2019; Song & Mittal, 2021) typically operate under the assumption that member records are more likely to exhibit a smaller prediction loss compared to nonmember records. Consequently, these methodologies compute the prediction loss and utilize it as a distinguishing factor between member and nonmember records.

Although the use of prediction loss to differentiate between member and nonmember records has been demonstrated to be effective for numerous deterministic models, such as classification models and Generative Adversarial Networks (GANs) (Chen et al., 2020; Hayes et al., 2019; Hilprecht et al., 2019; Choquette-Choo et al., 2021; Hanzlik et al., 2021), its efficacy is diminished in the context of diffusion models due to the intractability of the training loss. More precisely, during the training process of the diffusion model, a random noise is sampled, serving not only as a component of the model's input but also as the training target. However, during the execution of the membership inference attack, it is virtually impossible to replicate the exact noise sampled during the training phase. The discrepancy between the noise used during training and membership inference contributes to the inaccuracy of the loss estimation.

In light of the aforementioned issues, instead of the loss assumption, we assume that **for member samples, there exists a paired noise whereas for nonmember samples, the paired noise does not exist**. This assumption, we argue, is more congruent with the stochastic nature of the diffusion model's training process than the conventional loss assumption. Building on this premise we propose a membership inference framework for the diffusion model by a noise searching mechanism. The core idea is that for member samples, we can find a paired noise that leads to a smaller training loss. We formalize the noise searching process as an optimization problem with the training loss as the optimization objective.

Furthermore, we propose the fixed-point iteration as a method to solve the optimization problem. Our initial analysis focuses on the convergence properties of the fixed-point iteration. Theoretically, we discern a distinct attribute where member samples exhibit a slower convergence rate compared to non-member samples. This implies that the convergence speed can serve as a discriminative feature to differentiate member and non-member samples.

We ascertain that this characteristic can further elucidate the efficacy of existing Membership Inference Attack (MIA) methods for diffusion models. Specifically, from the standpoint of fixed-point iteration, current MIA methods can be interpreted as gauging the speed of convergence through a single iteration. To enhance the measurement of convergence speed, we implement an additional iteration step, termed as One More Step (OMS), beyond the single iteration of existing methods. We discover that OMS can substantially augment the performance of existing methods. We conduct experiments on various datasets and MIA methods. The results demonstrate the effectiveness of the proposed **OMS** and the noise searching MIA framework.

In summary, our paper makes the following contributions:

- We formulate a novel MIA framework for diffusion models in the perspective of noise searching. Moreover, we propose the fixed-point iteration to solve the noise searching optimization problem.
- We ~~analysis~~ analyze existing MIA methods through the proposed framework and reveal that the effectiveness of existing methods lies in the convergence speed of the fixed-point iteration process.
- Inspired by the fixed-point iteration process, we make adjustments to existing MIA methods by iterating **O**ne **M**ore **S**tep (**OMS**).
- We conduct experiments on various datasets and the experimental results demonstrate the correctness of the proposed MIA framework and the effectiveness of **OMS** adjustments.

2. Related Work

Membership Inference Attack (MIA). The Membership Inference Attack (MIA), initially introduced by Shokri et al. (Shokri et al., 2017), is a technique designed to extract privacy information from machine learning models. Its primary objective is to predict the presence of a specific data record in the training set of a given model.

The effectiveness of Membership Inference Attacks (MIAs) fundamentally relies on the hypothesis that machine learning models exhibit differential responses to member records versus unfamiliar nonmember records. Given the manner to exploit model's reactions, existing methods can be divided into two categories, model-based methods and metric-based methods. In the realm of model-based methods (Shokri et al., 2017; Salem et al., 2019; Long et al., 2020; Chen et al., 2020; Truex et al., 2019), a shadow model is trained to mimic the responses of the target machine learning model. Subsequently, attack algorithms are formulated, predicated on the reactions of the shadow model, with the ultimate objective of achieving generalization to the target model. For instance, Shokri et al. (Shokri et al., 2017) employ multiple shadow models to augment the attack success rate of the MIA task. Salem (Salem et al., 2019) discerns that the success of the shadow model is attributable to the transferability of the target machine learning model's output distribution. Long et al. (Long et al., 2020) initially select vulnerable (outlier) samples and find that these outlier samples are subject to a heightened privacy risk by shadow model-based attacks. Despite the significant advancements in the field, model-based methods are characterized by their computational intensity and exhibit susceptibility to alterations in the model's architecture.

Methods grounded in metrics (Sablayrolles et al., 2019; Yeom et al., 2018; Salem et al., 2020; Bentley et al., 2020) primarily employ a metric (typically the loss value) as a representative measure of the model's response to each sample. The membership of a specific sample is subsequently determined based on the numerical values of the selected metric. For example, Sablayrolles et al. (Sablayrolles et al., 2019) delve into the investigation of the metric within a white-box context, concluding that the most effective metric is the training loss function. Yeom et al. (Yeom et al., 2018) leverage the average training loss as their chosen metric and present a discussion on the interplay between model overfitting and the performance of membership inference. Bentley et al. (Bentley et al., 2020) examine the relationship between the generalization gap and membership inference, positing that a deficiency in generalization escalates the risk of model privacy leakage.

MIA for diffusion models. Given the computational intensity of training a shadow model with comparable parameters, model-based methods are deemed unsuitable for diffusion models. As a result, current MIA tailored for diffusion models (Duan et al., 2023; Matsumoto et al., 2023; Kong et al., 2023) are metric-based methods. These methods share a core assumption with MIAs applied to other models, which assumes that the loss value for members is smaller than that for nonmembers. A significant contribution to this field is made by Matsumoto et al. (Matsumoto et al., 2023), who introduce a pioneering MIA method for diffusion models. This approach utilizes the training loss of the model and identifies a timestep at which the divergence of loss between members and non-members is maximized, thereby optimizing the performance of membership inference. SecMI (Duan et al., 2023) exploits the approximated posterior error as a proxy to estimate the training loss, which subsequently serves as the membership inference metric. PIA (Kong et al., 2023) endeavors to reconstruct the training loss through the complete sampling path, thereby leveraging the training loss throughout the entire diffusion process.

Despite achieving substantial performance, these methods still suffer from the inaccuracy of loss estimation. During the training phase of the diffusion model, the loss value is dictated by the training target (a random Gaussian noise). However, when executing the MIA, replicating the exact noise sampled during the training process is virtually unattainable. Instead, we propose a novel MIA framework predicated on noise searching. This innovative approach promises to enhance the overall performance of MIA and provide insight into the principles of MIA methods for diffusion models.

3. Preliminaries ~~of Diffusion Models~~

Diffusion models. Denoising Models (DDPM) (Ho et al., 2020; Song et al., 2020) have emerged as a novel branch of generative models. With the impressive ability to approximate the distribution of image data, diffusion models have made breakthroughs in many visual content generation tasks (Rombach et al., 2022; Ruiz et al., 2023; Liu et al., 2023; Wu et al., 2023; Du et al., 2023; Kawar et al., 2023; Bhunia et al., 2023) and have broken the long-term domination of GANs (Dhariwal & Nichol, 2021; Mazé & Ahmed, 2023; Müller-Franzes et al., 2022). DDPMs consist of a forward and a reverse process. The forward process, also named as the diffusion process, gradually adds Gaussian noise to the input image x_0 in T time steps according to a predefined variance schedule $\beta_1, ..., \beta_T$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, this process can be simplified to:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

When t is large enough, the $\bar{\alpha}$ is approaching 0, making x_t an isotropic Gaussian noise.

The reverse process aims to recover the data distribution from the Gaussian noise. The reverse process in one step can be represented as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_t) \tag{3}$$

where Σ_t is a constant depending on the variance schedule β_t and $\mu_\theta(x_t, t)$ is determined by a neural network:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) \tag{4}$$

By recursively leveraging the reverse step, Gaussian noise can be recovered to the original image. To train the DDPM, we first sample an image x_0, a timestep t and a random noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. We can obtain a noisy image x_t using the forward process (Equation 2). We then input both the noisy image x_t and the timestep t into a U-Net (Ronneberger et al., 2015) ϵ_θ to predict the noise within x_t. The optimization objective for the denoising U-Net can be written as:

$$\mathcal{L} = \mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_0, t, \epsilon)||_2^2] \tag{5}$$

Fixed-point Iteration. The fixed-point iteration (Smart, 1980) is a widely used approach for solving equations. It is commonly employed to establish existence and uniqueness theorems for various classes of operator equations, including differential equations and integral equations. Given the implicit function $x = f(x)$. The process of fixed-point iteration can be represented as:

$$x^n = f(x^{n-1}) \quad n = 1, 2, ... \tag{6}$$

Formally, the fixed-point iteration is assured to converge provided that the function f is contractive on its domain if there exists the constant $0 \leq L < 1$ such that:

$$\forall x, y \quad ||f(x) - f(y)|| \leq L||x - y|| \tag{7}$$

The constant L also impacts the convergence rate of the sequence x_n, as:

$$\begin{aligned} ||x_{n+1} - x_n|| &= ||f(x_n) - f(x_{n-1})|| \\ &\leq L||x_n - x_{n-1}|| \\ &\leq L^n||x_1 - x_0|| \end{aligned} \tag{8}$$

4. Method

Given a data record x_0, the goal of membership inference attack (MIA) is to identify whether x_0 is in the training set of the target diffusion model ϵ_θ. Existing MIAs (Duan et al., 2023; Matsumoto et al., 2023; Kong et al., 2023) tailored for diffusion models predominantly assume that the loss values for members are lower than those for non-members. However, these loss-based approaches are susceptible to inaccuracies in loss estimation, which are caused by noise inconsistency between the training and inference stages (Section 4.1). In contrast, we propose a novel MIA framework that employs noise searching, an approach we believe aligns more closely with the stochastic nature of the model's training process. We formulate the noise searching process as an optimization problem and propose to use the fixed-point iteration to solve this problem (Section 4.2). Subsequently, we analyze the convergence properties of the fixed-point iteration and find that members converge more slowly than non-members (Section 4.3). From the standpoint of convergence speed, we reinterpret the reasons for the effectiveness of existing MIA methods and propose an enhanced method by incorporating **O**ne **M**ore iteration **S**tep (**OMS**) (Section 4.5).

4.1. Noise Inconsistency between Training and Inference

The diffusion model's training procedure can be described by Equation 5. To elaborate, given the input image x_0 and a specific timestep t, a random noise ϵ_{train} is sampled from the standard normal distribution. This noise is utilized to perturb x_0 into a corrupted version x_t according to the schedule predefined in Equation 2. Subsequently, the diffusion model, parameterized by θ generates a prediction of the noise within x_t (denoted as $\epsilon_\theta(x_0, t, \epsilon_{train})$). The training loss for the diffusion model is computed as the distance between the predicted noise and the sampled noise ϵ_{train}.

During the inference phase, due to the infeasibility of the training noise ϵ_{train}, an alternate noise ϵ_{inf} is sampled to calculate the loss value. However, it is important to note that we can not guarantee that the sampled ϵ_{inf} is identical or approximately similar to the noise ϵ_{train}. This inconsistency in noise significantly impacts the accuracy of the loss values, consequently affecting the efficacy of existing Membership Inference Attacks (MIA) for diffusion models (Duan et al., 2023; Matsumoto et al., 2023; Kong et al., 2023).

4.2. MIA by Noise Searching

We introduce a novel MIA framework for diffusion models by noise searching. Specifically, given the record image x_0, we intend to find its corresponding noise ϵ_{train} that minimizes the training loss as defined in Equation 5. We formulate the process of noise searching as an optimization problem:

$$\begin{aligned} &\min_{\epsilon} ||\epsilon - \epsilon_\theta(x_0, t, \epsilon)||_p \\ &s.t. \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \tag{9}$$

Compared with loss-based methods, our proposed approach emphasizes the identification of ϵ_{train}. We posit that this approach aligns more coherently with the inherent stochasticity of the diffusion model's training process.

Fixed-point iteration. Given a record x_0 and a timestep t, the predicted noise $\epsilon_\theta(x_0, t, \epsilon)$ is solely dependent on ϵ. This dependency can be represented as an implicit function $\epsilon = f(\epsilon)$. The ideal noise, which pairs with the record x_0 during the training process, also serves as the solution for the implicit function f. We address this problem utilizing the fixed-point iteration (Smart, 1980). The process can be represented as follows:

$$\epsilon^n = f(\epsilon^{n-1}), \quad n = 1, 2, ... \tag{10}$$

We posit that the fixed-point iteration process essentially satisfies the constraints inherent in the optimization problem. This is predicated on the fact that the model ϵ_θ is trained to match a noise that adheres to a standard normal distribution. Consequently, we also hypothesize that the output of the model conforms to a standard normal distribution.

Note that we do not use more advanced methods for solving implicit functions such as Newton-Raphson or Conjugate Gradient (Nocedal & Wright, 1999), this is because Newton-Raphson method needs to compute gradient and Conjugate Gradients need to find high dimension gradients which are computationally ~~computationally~~ overload.

4.3. Convergence of Fixed-point Iteration

~~Fixed-point iterations are assured to converge provided that the function f is an a-contraction (Berinde & Takens, 2007), that is, there exists a constant $0 \leq a < 1$ such that: $d(f(x)), f(y)) \leq ad(x, y), \forall x, y$(11) where $d(\cdot, \cdot)$ denotes the distance function.~~ In our specific case where the function f is represented by a highly complex and nonlinear neural network, it is not theoretically guaranteed that the constant L exists, as suggested in Equation 8. ~~function f is a-contractive across all the regions~~. However, empirical observations indicate that the iterative process defined by Equation 10 results in a diminishing difference across successive iterations $(\epsilon^i - f(\epsilon^i))$.

As illustrated in Figure 1, we show the Relative Loss Value varying with the number of iteration (solid lines for members and dashed lines for nonmembers). The Relative Loss Value is defined as the ratio of ith difference $(f(\epsilon^i) - \epsilon^i)$ to the initial difference $(f(\epsilon^0) - \epsilon^0)$. It can be observed that the fixed-point iteration converges in several iterations. Interestingly, it is observed that the convergence speed for members is slower compared to nonmembers.

Convergence Speed. Initially, we establish a relationship between the speed of convergence and the slope of the function f. Subsequently, we elucidate the influence of the training process on the slope and provide an explanation for the slower convergence speed observed for members in comparison to nonmembers.

Figure 2 provides a graphical illustration of low-dimensional fixed-point iteration process. It is observed that a flatter

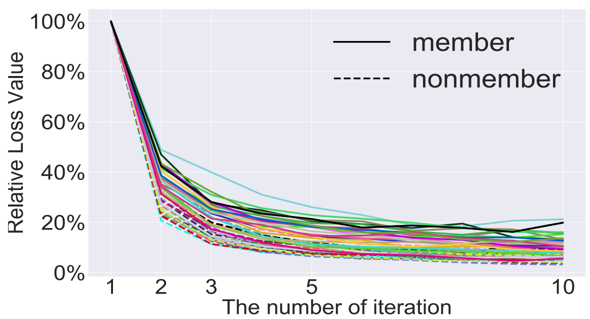


Figure 1. The Relative Loss Value for 50 records when $t = 100$ (half for members and half for nonmembers). We empirically show that the fixed-point iteration converges for all samples we tested.

slope of the function f results in a more rapid convergence ~~$(\epsilon^0 - \epsilon^i)$~~ of the fixed-point iteration. Besides, from the perspective of Equation 8, we observe that the slope $\left|\frac{\partial f(\epsilon)}{\partial \epsilon}\right|$ is a good approximation for the constant L:

$$\begin{aligned}||\epsilon^{n+1} - \epsilon^n|| &= ||f(\epsilon^n) - f(\epsilon^{n-1})|| \\ &= ||\frac{\partial f(\epsilon)}{\partial \epsilon}|_{\epsilon=\epsilon^{n-1}} \cdot \delta^n + \mathcal{O}(||\delta^n||^2) \\ &\leq ||\frac{\partial f(\epsilon)}{\partial \epsilon}|_{\epsilon=\epsilon^{n-1}}|| \cdot ||\delta^n|| + \mathcal{O}(||\delta^n||^2)\end{aligned} \tag{12}$$

where $\delta^n = \epsilon^n - \epsilon^{n-1}$. When δ is sufficiently small, $||\mathcal{O}(\delta^2)||$ can be ignored. The convergence speed is mainly dependent on the slope $\frac{\partial f(\epsilon)}{\partial \epsilon}$.

We further demonstrate that the slope of f escalates as the model undergoes the training process. We assume that the learned distribution ~~$p(x)$~~ $p_\theta(x)$ adheres to a normal distribution with a mean of ~~μ~~ μ_θ and a variance of ~~σ^2~~ σ_θ^2. As the training advances, the learned distribution ~~$p(x)$~~ $p_\theta(x)$ becomes increasingly concentrated in data intensive areas, which means the variance ~~σ^2~~ σ_θ^2 diminishes as the training progresses. According to Theorem 1, the slope of f is inversely proportional to the variance ~~σ^2~~ σ_θ^2. This indicates that the slope escalates as the training progresses and further results in slower convergence.

In conclusion, as the training process progresses, the variance ~~σ^2~~ σ_θ^2 of the learned distribution ~~σ^2~~ σ_θ^2 diminishes, resulting in a steeper slope of f and consequently, a slower convergence of the fixed-point iteration. Subsequently, we will leverage the differential convergence speed observed between members and nonmembers to execute the MIA task.

Theorem 1. *Assume that the learned distribution $p(x)$ adheres to a normal distribution with a mean of ~~μ~~ μ_θ and a variance of ~~σ^2~~ σ_θ^2. The slope of the function f is inversely proportional to ~~σ^2~~ σ_θ^2:*

$$\left|\frac{\partial f(\epsilon)}{\partial \epsilon}\right| \propto \frac{1}{\sigma_\theta^2} \tag{13}$$

Proof. By Newton-Leibniz law of calculus, we have:

$$\frac{\partial f(\epsilon)}{\partial \epsilon} = \frac{\partial f(\epsilon)}{\partial x}\frac{\partial x}{\partial \epsilon} = \sqrt{1-\bar{\alpha}}\frac{\partial f(\epsilon)}{\partial x} \tag{14}$$

For simplicity, we omit the subscript t. From the score matching perspective (Song et al., 2020), the neural network aims to estimate:

$$f(\epsilon) = \nabla_x \log p_\theta(x) = \frac{\nabla_x p_\theta(x)}{p_\theta(x)} \tag{15}$$

Considering that $p_\theta(x) \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, we have:

$$\left|\frac{\partial f(\epsilon)}{\partial \epsilon}\right| = \sqrt{1-\bar{\alpha}}\left|\nabla_x \frac{\mu - x}{\sigma_\theta^2}\right| = \frac{\sqrt{1-\bar{\alpha}}}{\sigma_\theta^2} \tag{16}$$

4.4. Convergence speed as MIA metric

In previous discussion, we elucidate the differential convergence speeds of members and nonmembers. Subsequently, we will leverage this finding to perform the membership inference attacks. As indicated by Equation 8, the constant L serves as an effective surrogate for convergence speed. However, it is impractical to obtain an accurate L. Existing methods (Virmaux & Scaman, 2018; Anil et al., 2019; Fazlyab et al., 2019) merely offer a loose upper bound and requires a number of queries, which are not suitable for MIA. Consequently, we propose an alternative: approximating the convergence speed using the residual (δ^n). The main strength to choose δ^n as the metric is its ability to capitalize on both the loss value and the convergence speed. From Equation 8, we derive:

$$\delta^n \leq L^{n-1}\delta^1 \tag{17}$$

The $\delta^1 = f(\epsilon^0) - \epsilon^0$ also recognized as the loss function, which is advantageous for distinguishing between members and nonmembers, given that members typically exhibit a smaller loss value. To further leverage both the loss and the convergence speed, we use $\epsilon^n - \epsilon^0$:

$$\begin{aligned}||\epsilon^n - \epsilon^0|| \leq ||\delta^n|| + ||\delta^{n-1}|| + ... + ||\delta^1|| \\ (L^{n-1} + L^{n-2} + ... + 1) \cdot ||\delta^1||\end{aligned} \tag{18}$$

We also provide ablations using various metric to differentiate members and nonmembers (Section 5.3).

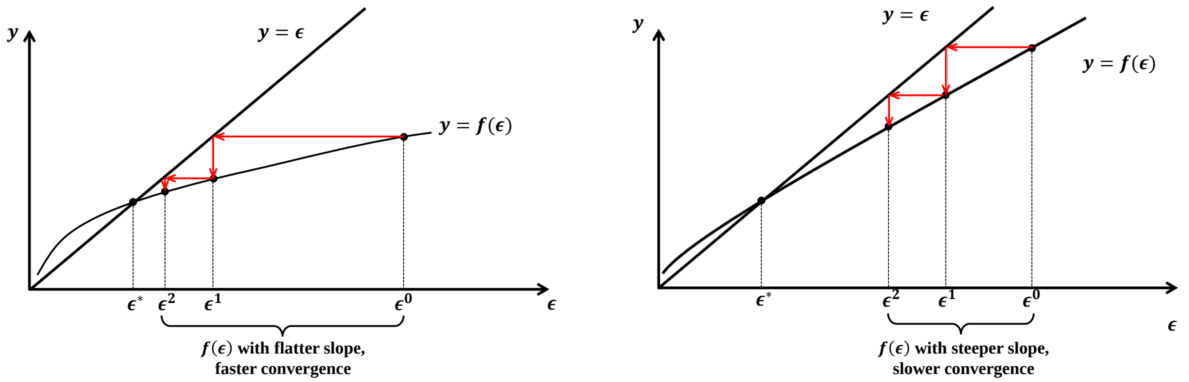


Figure 2. Illustration of the low-dimension fixed-point iteration process. It is observed that an increase in the slope of the function f results in a slower convergence speed.

Table 1. The ASR and AUC metrics of current baselines with/without One More Step (OMS). The symbol Δ is employed to denote the performance gain attributable to the implementation of our OMS method.

	Cifar10		Cifar100		LFW		Ave	
Method	ASR	AUC	ASR	AUC	ASR	AUC	ASR	AUC
NaiveLoss	76.08	82.18	73.17	79.74	80.39	88.50	76.55	83.47
+ OMS	82.08	88.84	80.13	87.55	87.92	94.79	83.38	90.39
$\Delta\uparrow$	**6.01**	**6.65**	**6.96**	**7.81**	**7.53**	**6.29**	**6.83**	**6.92**
SecMI	83.59	90.31	78.84	85.92	80.50	88.88	80.98	88.37
+ OMS	86.20	92.44	84.72	91.63	85.31	93.83	85.41	92.63
$\Delta\uparrow$	**2.62**	**2.12**	**5.88**	**5.71**	**6.60**	**4.95**	**5.03**	**4.26**
PIA	88.50	94.72	85.53	92.45	74.80	79.67	82.94	88.95
+ OMS	92.24	97.39	90.18	96.23	81.11	85.40	87.84	93.01
$\Delta\uparrow$	**3.74**	**2.67**	**4.65**	**3.78**	**6.31**	**5.73**	**4.71**	**4.06**

4.5. Rethinking exiting MIA methods

We reevaluate the efficacy of existing (MIA) methods from the standpoint of convergence speed. As depicted in Figure 2, the speed of convergence can be quantified as the distance between the initial point ϵ^0 and the ith iteration ϵ^i. Present MIA methods estimate the training loss of diffusion models (as per Equation 5), which can be interpreted as the l_2 distance between the initial point ϵ^0 and the first iteration ϵ^1. Although some methodologies, such as the one proposed by Duan et al. (Duan et al., 2023), estimate the loss by reconstructing the image x rather than the noise, we assert that the image reconstruction process can be reformulated to align with Equation 5. To elaborate, without loss of generality, we consider the reconstruction process that initiates from x_{ta}, diffuses to x_{tb}, and subsequently reverses to the reconstructed $\tilde{x}_{ta}$. The diffusion process from x_{ta} to x_{tb} can be represented as follows:

$$\frac{x_{ta}-\sqrt{1-\bar{\alpha}_{ta}}\epsilon}{\sqrt{\alpha_{ta}}} = \frac{x_{tb}-\sqrt{1-\bar{\alpha}_{tb}}\epsilon}{\sqrt{\alpha_{tb}}}$$

$$x_{tb} = \frac{\sqrt{\bar{\alpha}_{tb}}}{\sqrt{\bar{\alpha}_{ta}}}x_{ta} + \frac{\sqrt{\bar{\alpha}_{ta}\bar{\beta}_{tb}} - \sqrt{\bar{\alpha}_{tb}\bar{\beta}_{ta}}}{\sqrt{\bar{\alpha}_{ta}}}\epsilon \tag{19}$$

where $\bar{\beta}_{tb} = \prod_{s=1}^{t}\beta_s$. The diffusion process from x_{tb} to x_{ta} is similar to Equation 19, but the noise is predicted by the neural network $\epsilon_\theta(x_0, t_b, \epsilon)$:

$$\tilde{x}_{ta} = \frac{\sqrt{\bar{\alpha}_{ta}}}{\sqrt{\bar{\alpha}_{tb}}}x_{tb} + \frac{\sqrt{\bar{\alpha}_{tb}\bar{\beta}_{ta}} - \sqrt{\bar{\alpha}_{ta}\bar{\beta}_{tb}}}{\sqrt{\bar{\alpha}_{tb}}}\epsilon_\theta(x_0, t_b, \epsilon) \tag{20}$$

Finally, the distance between image can be converted to the distance between noise:

$$||x_{ta} - \tilde{x}_{ta}||_p = \frac{\sqrt{\bar{\alpha}_{ta}\bar{\beta}_{tb}} - \sqrt{\bar{\alpha}_{tb}\bar{\beta}_{ta}}}{\sqrt{\bar{\alpha}_{tb}}}||\epsilon - \epsilon_\theta(x_0, t_b, \epsilon)||_p \tag{21}$$

Drawing from Equation 21, we posit that the image reconstruction process can be equivalently viewed as the fixed-point iteration delineated in Equation 10.

Iterate One More Step. From the perspective of convergence speed, we observe that current MIA methods solely consider the distance between the initial point ϵ^0 and the

Table 2. The True Positive Rate (TPR) at very low False Positive Rate (FPR). TPR@1% and TPR@0.1% stands for TPR when the FPR is set at 1% and 0.1% separately. The symbol Δ is employed to denote the performance gain attributable to the implementation of our OMS.

	Cifar10		Cifar100		LFW	
Method	TPR@1%	TPR@0.1%	TPR@1%	TPR@0.1%	TPR@1%	TPR@0.1%
NaiveLoss	3.62	0.26	6.21	0.80	11.89	1.73
+ OMS	9.43	0.74	12.83	1.55	38.03	8.81
$\Delta\uparrow$	**5.80**	**0.48**	**6.61**	**0.75**	**26.14**	**7.08**
SecMI	10.81	0.56	5.00	0.02	27.20	0.20
+ OMS	14.39	0.90	20.14	0.84	57.82	0.20
$\Delta\uparrow$	**3.58**	**0.34**	**15.14**	**0.82**	**30.62**	0.00
PIA	26.82	1.18	22.33	1.85	3.60	0.40
+ OMS	57.31	8.32	50.05	10.26	6.80	0.62
$\Delta\uparrow$	**30.48**	**7.14**	**27.73**	**8.41**	**3.20**	**0.22**

first iteration ϵ^1. To enhance the measurement of convergence speed, we implement an additional iteration step, termed as **O**ne **M**ore **S**tep (**OMS**), beyond the single iteration employed by existing methods. In essence, we employ the distance between ϵ^0 and ϵ^2 as a discriminative measure to ascertain between members and non-members. (A detailed description of each MIA method is provided in the Appendix for further reference.)

5. Experiment

5.1. Experimental Setup

Datasets, Diffusion Models and Member Sets We use Cifar10, Cifar100 and LFW datasets in our experiments. We train the denoising diffusion model (Ho et al., 2020) at 32 × 32 resolution for Cifar10 and Cifar100 datasets. We also train a 128 × 128 resolution diffusion model for the LFW dataset. For Cifar10 and Cifar100 datasets, we select half of the image for each class as the training set to guarantee the diversity (25000 images in total). For LFW dataset, we choose half of the identity as the training set (2875 identities), the other half as the test set (3200 images in total). When executing the membership inference, the training set is considered as members while test set as the non-members. We train the diffusion model using 1 NVIDIA V100 GPU with the batchsize set to 16 for LFW and 50 for Cifar10 and Cifar100 separately. We use the AdamW optimizer (Loshchilov & Hutter, 2018) with a fixed learning rate of 0.0001. For Cifar10 and Cifar100 dataset, the training iterations are set to 250K (500 epoch). For LFW dataset, the training iterations are set to 80K (400 epoch).

Evaluation Metrics. To evaluate the performance of our proposed method, we employ widely used metrics including Attack Success Rate (ASR) and Area-Under-the-ROC-curve (AUC). The ASR is defined as the propotion of successful attacks relative to the total number of attacks (Duan et al., 2023). Additionally, we utilize the True Positive Rate (TPR) at extremely low False Positive Rate (FPR) to gauge the confidence level of the prediction, as suggested by Carlini et al. (Carlini et al., 2023). In this context, TPR@1%FPR and TPR@0.1%FPR denote the True Positive Rate when the False Positive Rate is set at 1% and 0.1% respectively.

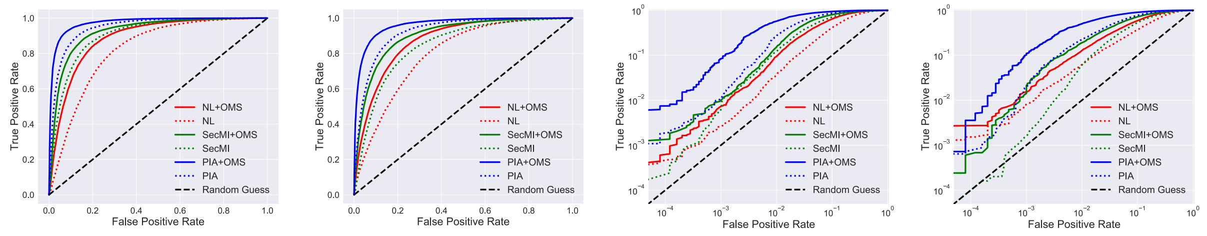
Table 3. Ablation studies of distance metrics. We show the performance (AUC and ASR) with/without OMS under various metrics on Cifar10 dataset.

		L1	MSE	Sim
ASR	w/o OMS	76.08	78.86	78.62
	with OMS	82.08	82.20	82.67
	$\Delta\uparrow$	**6.01**	**3.34**	**4.05**
AUC	w/o OMS	82.18	85.45	85.25
	with OMS	88.84	88.94	89.45
	$\Delta\uparrow$	**6.65**	**3.49**	**4.20**

5.2. Comparasion to Baselines

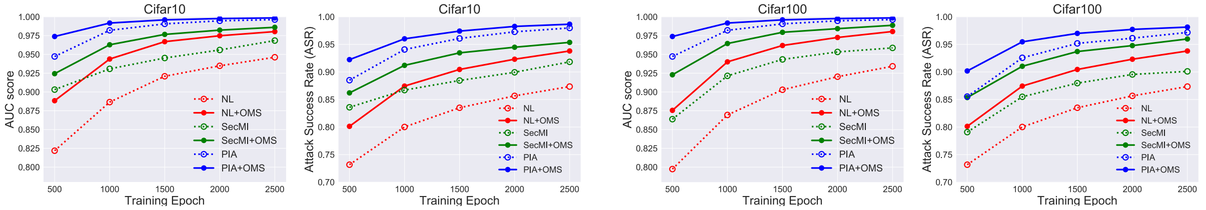
We perform one more fixed-point iteration to current MIA methods, including PIA (Kong et al., 2023), SecMI (Duan et al., 2023) and NaiveLoss (Matsumoto et al., 2023). (We give a more detailed description of these baselines in the Appendix.) We show the performance with/without One More Step (OMS) in Table 1. Compared to the baselines, OMS achieves significant performance boosts of 6.92/4.26/4.06 on the Average AUC metric, indicating the superiority of performing multiple fixed-point iterations than only one iteration. We also consider the TPR at very low FPR, e.g. 1% and 0.1% FPR in Table 2 (TPR@1% and TPR@0.1% for short). The results show that OMS can significantly enhance the prediction's confidence, which means that our proposed OMS can identify more members than the baselines.

We also present the overall ROC curves and the log-scaled ROC curves with and without the incorporation of OMS in Figure 3. For the same membership inference baselines, we use dotted lines to represent the curve without OMS, while solid lines illustrate the curve after incorporating OMS. As can be seen from Figure 3, an additional iteration prompts a shift of all baseline ROC curves towards the upper left cor-



(a) ROC on Cifar10. **(b) ROC on Cifar100.** **(c) Log-scaled ROC on Cifar10.** **(d) Log-scaled ROC on Cifar100.**

Figure 3. The ROC and log scaled ROC curves on Cifar10 and Cifar100 datasets with/without One More Step (OMS). The ROC and log scaled ROC provide compelling evidence that OMS can substantially enhance the performance of existing MIA methods. The NL is short for the method NaiveLoss (Matsumoto et al., 2023).



(a) AUC and ASR with different Training Epochs on Cifar10 **(b) AUC and ASR with different Training Epochs on Cifar100**

Figure 4. Ablation studies of the training epoch. We plot the performance (AUC and ASR) with/without OMS while varying the training epoch. The NL is short for the method NaiveLoss (Matsumoto et al., 2023).

ner. This shift signifies an improvement in the performance of our proposed method. Furthermore, the log-ROC curve reveals that the implementation of fixed-point iteration can amplify the prediction confidence of existing membership inference baselines, This enhancement is particularly noticeable in scenarios demanding high prediction confidence, even when the FPR is low.

5.3. Ablation Study

Proxy for convergence speed. We execute attacks utilizing various metrics to approximate the convergence speed, with the results illustrated in Figure 6. It is observed that the performance of the residual diminishes as n increases. This can be attributed to the fact that the residual leverages L^{n-1} to differentiate between members and nonmembers (Equation 8). Furthermore, we discern that $\epsilon^n - \epsilon^0$ demonstrates a substantial performance enhancement over $\epsilon^1 - \epsilon^0$. This is primarily due to the fact that $\epsilon^1 - \epsilon^0$ depends on the loss value while $\epsilon^n - \epsilon^0$ also incorporates the convergence speed (Equation 18).

Training Epoch. Numerous studies (Yeom et al., 2018; Leino & Fredrikson, 2020; Salem et al., 2019) have underscored the propensity of models to memorize training instances as the number of training epochs increases. In light of these findings, we have conducted an evaluation of our method throughout the training process. The results are depicted in Figure 4. A key observation is that our method is robust to the number of training epochs. Notably, our method enhances the performance of existing MIA methods consistently throughout the entire training process.

Timestep. In order to assess the influence of the timestep, we conduct attacks on the target model, varying the timestep from 50 to 190 in increments of 20. The boost in AUC is illustrated in Figure 5 (The boost represents the AUC difference with/without OMS). It is noteworthy that an additional iteration can significantly enhance the performance of current baselines when the timestep is small. However, this enhancement diminishes as the timestep increases. This may attributed to the added noise. When the timestep gets larger, the noise becomes the dominant factor in model's input, thereby disrupting the process of fixed-point iteration.

Distance Metric. In order to demonstrate the robustness of our proposed method, we conduct an evaluation under various distance metrics. In addition to the most commonly employed metrics such as L1-norm and Mean Squared Error (MSE), we incorporate the use of cosine similarity (Sim). This metric is particularly relevant as the predicted output aligns with the gradient direction of the dataset manifold. The Sim metric serves as an effective measure for consistency indirection. The results are shown in Table 3. We

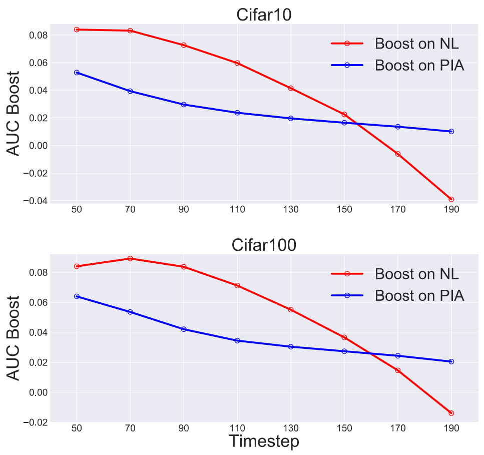


Figure 5. Ablation studies of the timestep. We plot the performance boost (quantified by the AUC) that results from OMS while varying the timestep.

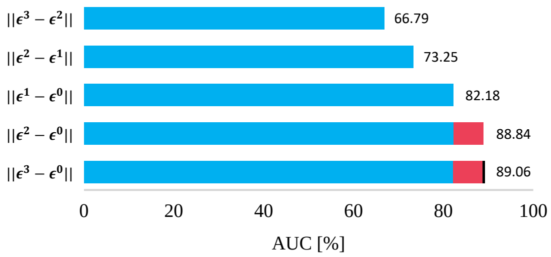


Figure 6. Ablation studies of the convergence rate proxy. We explore the residual ($\epsilon^n - \epsilon^{n-1}$) as the proxy, along with its ensembles ($\epsilon^n - \epsilon^0$) on Cifar10.

can observe that OMS demonstrates utility across all these distance metrics.

The number of iteration. We perform ablation studies to investigate the impact of the number of fixed-point iterations. The outcomes of these studies are depicted in Figure 7. It can be observed that the performance initially exhibits a significant increase, followed by a slight decrease across all training epochs. The initial increase underscores the effectiveness to perform one more iteration, while the subsequent decrease could potentially be attributed to the oscillations that are frequently observed during the optimization process. More generally, the fixed-point iteration can be viewed as:

$$x^{n+1} = \alpha^n x^n + (1 - \alpha^n) \tag{22}$$

We employ the naive fixed-point iteration which sets all $\alpha^n = 0$, taking a fully updated step. It is plausible to anticipate oscillations after several iterations. ~~Given that we employ the initial fixed-point iteration with a fixed step size, it is plausible to anticipate oscillations after several iterations.~~

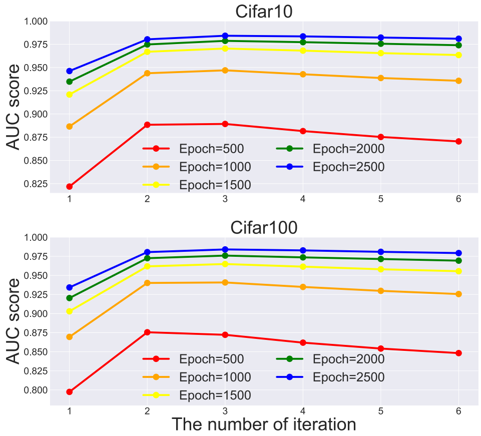


Figure 7. Ablation studies of the number of iteration. We plot the AUC score varing with the number of iteration numbers.

6. Conclusion and Future Works

In this study, we approach the Membership Inference Attack (MIA) task from a novel perspective, namely, noise searching. We formulate the noise searching process as an optimization problem and utilize fixed-point iteration to solve this problem. We analyse the convergence properties of the fixed-point iteration and deduce that members and non-members exhibit different convergence speed. Subsequently, we elucidate the efficacy of current MIA methods within the unified perspective of convergence speed. Ultimately, we enhance current MIA methods by incorporating **O**ne **M**ore iteration **S**tep (**OMS**), which significantly boost the performance of existing MIA methods. Our proposed MIA framework offers a unified perspective for comprehending the principles of MIA tasks for diffusion models. We hope that our work will contribute to the ongoing research on the privacy risks associated with diffusion models.

Future Works. From a theoretical standpoint, the exploration of the existence and uniqueness theorem, particularly in the context of the diffusion model being the function, presents an intriguing avenue for research. In terms of acceleration, our investigation is currently confined to the initial fixed-point iteration, specifically, the Picard iteration (Smart, 1980). However, the scope for further research is vast. It would be compelling to extend our inverstigation to other iterations, such as Krasnoselkij, Mann and Ishikawa iterations (Babu & Vara Prasad, 2006; Berinde, 2004a;b).

Ethical Statement

The primary objective of our research is to devise a method capable of discerning whether a particular sample was included in the training dataset. The proposed method offers a multitude of beneficial applications, encompassing the detection of privacy violations and the assessment of model privacy. While acknowledging the potential for malevolent entities to misuse our method for privacy attacks, we underscore the capacity of privacy protection techniques, such as differential privacy, to counteract such threats. It is crucial to note that the development of these techniques is not intended to facilitate malicious activities, but rather to advance the field of privacy protection. In the spirit of open science and to further the progress of privacy protection technology, we plan to make our code publicly accessible. We trust that our contributions will be used responsibly to enhance privacy protection measures and promote ethical practices in machine learning research.

References

Anil, C., Lucas, J., and Grosse, R. Sorting out lipschitz function approximation. In International Conference on Machine Learning, pp. 291–301. PMLR, 2019.

Babu, G. and Vara Prasad, K. Mann iteration converges faster than ishikawa iteration for the class of zamfirescu operators. Fixed Point Theory and Applications, 2006: 1–6, 2006.

Bentley, J. W., Gibney, D., Hoppenworth, G., and Jha, S. K. Quantifying membership inference vulnerability via generalization gap and other model metrics. arXiv preprint arXiv:2009.05669, 2020.

Berinde, V. Comparing krasnoselskij and mann iterations for lipschitzian generalized pseudocontractive operators. In Proc. Int. Conf. Fixed Point Theory and Applications (Valencia, 2003), Yokohama Publishers, Yokohama, 2004a.

Berinde, V. Picard iteration converges faster than mann iteration for a class of quasi-contractive operators. Fixed Point Theory and Applications, 2004:1–9, 2004b.

Berinde, V. and Takens, F. Iterative approximation of fixed points, volume 1912. Springer, 2007.

Bhunia, A. K., Khan, S., Cholakkal, H., Anwer, R. M., Laaksonen, J., Shah, M., and Khan, F. S. Person image synthesis via denoising diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5968–5976, 2023.

Brittain, B. Getty images lawsuit says stability ai misused photos to train ai. Reuters, Feb 6th, 2023.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In 32nd USENIX Security Symposium (USENIX Security 23), pp. 5253–5270, 2023.

Chen, D., Yu, N., Zhang, Y., and Fritz, M. Gan-leaks: A taxonomy of membership inference attacks against generative models. In Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, pp. 343–362, 2020.

Choquette-Choo, C. A., Tramer, F., Carlini, N., and Papernot, N. Label-only membership inference attacks. In International conference on machine learning, pp. 1964–1974. PMLR, 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34:8780–8794, 2021.

Du, C., Chen, Q., He, T., Tan, X., Chen, X., Yu, K., Zhao, S., and Bian, J. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. In Proceedings of the 31st ACM International Conference on Multimedia, pp. 4281–4289, 2023.

Duan, J., Kong, F., Wang, S., Shi, X., and Xu, K. Are diffusion models vulnerable to membership inference attacks? International Conference on Machine Learning, 2023.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of lipschitz constants for deep neural networks. Advances in neural information processing systems, 32, 2019.

Hanzlik, L., Zhang, Y., Grosse, K., Salem, A., Augustin, M., Backes, M., and Fritz, M. Mlcapsule: Guarded offline deployment of machine learning as a service. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3300–3309, 2021.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: Membership inference attacks against generative models. In Proceedings on Privacy Enhancing Technologies (PoPETs), volume 2019, pp. 133–152. De Gruyter, 2019.

Hilprecht, B., Härterich, M., and Bernau, D. Monte carlo and reconstruction membership inference attacks against generative models. Proc. Priv. Enhancing Technol., 2019 (4):232–249, 2019.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR), 54(11s): 1–37, 2022.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6007–6017, 2023.

Kong, F., Duan, J., Ma, R., Shen, H., Zhu, X., Shi, X., and Xu, K. An efficient membership inference attack for the diffusion model by proximal initialization. arXiv preprint arXiv:2305.18355, 2023.

Leino, K. and Fredrikson, M. Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In 29th USENIX security symposium (USENIX Security 20), pp. 1605–1622, 2020.

Liu, J., Wang, X., Fu, X., Chai, Y., Yu, C., Dai, J., and Han, J. Mfr-net: Multi-faceted responsive listening head generation via denoising diffusion model. In Proceedings of the 31st ACM International Conference on Multimedia, pp. 6734–6743, 2023.

Long, Y., Wang, L., Bu, D., Bindschaedler, V., Wang, X., Tang, H., Gunter, C. A., and Chen, K. A pragmatic approach to membership inferences on machine learning models. In 2020 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 521–534. IEEE, 2020.

Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. 2018.

Matsumoto, T., Miura, T., and Yanai, N. Membership inference attacks against diffusion models. arXiv preprint arXiv:2302.03262, 2023.

Mazé, F. and Ahmed, F. Diffusion models beat gans on topology optimization. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Washington, DC, 2023.

Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nebelung, S., Kather, J. N., et al. Diffusion probabilistic models beat gans on medical images. arXiv preprint arXiv:2212.07501, 2022.

Nocedal, J. and Wright, S. J. Numerical optimization. Springer, 1999.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241. Springer, 2015.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22500–22510, 2023.

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

Salem, A., Zhang, Y., Humbert, M., Fritz, M., and Backes, M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In Network and Distributed Systems Security Symposium 2019. Internet Society, 2019.

Salem, A., Bhattacharya, A., Backes, M., Fritz, M., and Zhang, Y. {Updates-Leak}: Data set inference and reconstruction attacks in online learning. In 29th USENIX security symposium (USENIX Security 20), pp. 1291–1308, 2020.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP), pp. 3–18. IEEE, 2017.

Smart, D. R. Fixed point theorems, volume 66. Cup Archive, 1980.

Song, L. and Mittal, P. Systematic evaluation of privacy risks of machine learning models. In 30th USENIX Security Symposium (USENIX Security 21), pp. 2615–2632, 2021.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2020.

Truex, S., Liu, L., Gursoy, M. E., Yu, L., and Wei, W. Demystifying membership inference attacks in machine learning as a service. IEEE Transactions on Services Computing, 14(6):2073–2089, 2019.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. Advances in Neural Information Processing Systems, 31, 2018.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7623–7633, 2023.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st computer security foundations symposium (CSF), pp. 268–282. IEEE, 2018.

A. More details on Baselines.

In this section, we elucidate the specifics of each baseline. Furthermore, we provide an interpretation of each baseline from the standpoint of fixed-point iteration.

NaiveLoss (Matsumoto et al., 2023) employs the training objective (Equation 5) of the diffusion model as a discriminative criterion. This is computed using the training loss at timestep 350 in the original paper. However, in our experiments, we empirically find that NaiveLoss achieves optimal performance around timestep 100. Consequently, in our experiments, we employ the loss at timestep 100:

$$\mathcal{L} = |\epsilon - \epsilon_\theta(x_0, t = 100, \epsilon)| \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{23}$$

Interpreted through the lens of fixed-point iteration, this can be construed as the distance between the initial point ϵ^0 and the first iteration ϵ^1, with the initial point being a random noise.

PIA (Kong et al., 2023) employs the distance between trajectories to evaluate the training loss. Our empirical observations indicate that the timestep does not exert a significant influence on the MIA performance. Consequently, we adhere to all the configurations specified in the original paper for our experiments.

$$\mathcal{L} = ||\epsilon' - \epsilon_\theta(x_0, t = 200, \epsilon')||_4 \tag{24}$$

$$\epsilon' = \epsilon_\theta(x_0, t = 0, \epsilon = \emptyset) \tag{25}$$

Viewed from the perspective of fixed-point iteration, this can be interpreted as the l_4 distance between the initial point ϵ^0 and the first iteration ϵ^1. Notably, the initial point is a noise predicted from the clean image x_0.

SecMI (Duan et al., 2023) utilizes the posterior estimation error as a proxy for the training loss:

$$\mathcal{L} = ||x_{100} - \bar{x}_{100}|| \tag{26}$$

where $x_{\bar{100}}$ is generated by initially diffusing x_{100} to x_{110}, followed by reversing x_{110} to $\bar{x}_{100}$. This is equivalent to $||\epsilon' - \epsilon_\theta(x_0, t = 110, \epsilon')||$ (refer to Equation 21 in the main text). We empirically find that SecMI achieves optimal performance around timestep 100, which is consistent with the original paper. Consequently, we adhere to all the configurations specified in the original paper for our experiments.

Viewed from the perspective of fixed-point iteration, this can be interpreted as the l_2 distance between the initial point ϵ^0 and the first iteration ϵ^1. Notably, the initial point is a noise predicted from a diffusing sequence from x_0 to x_{100} with an interval of 10.