GenAI Copyright Evidence with Operational Meaning

Eli Chien [*1] **Amit Saha** [*1] **Yinan Huang** [*1] **Pan Li** [1]

Abstract

The remarkable success of generative AI models, enabled by large-scale training on massive and diverse datasets, has raised growing concerns about whether their outputs constitute copyright infringement. Under U.S. copyright law, two key elements must be established for infringement: the model is trained on the copyrighted content (`Access`) and its outputs are substantially similar to the copyrighted content (`Similarity`). However, determining infringement is inherently complex, and legal practices often rely on subjective assessments. In this paper, we focus on designing criteria that provide quantitative evidence to help determine AI copyright infringement. We introduce a game-theoretic framework that formalizes `Access` and `Similarity` as a membership inference game and a data reconstruction game, respectively, between a plaintiff and a defendant. The plaintiff's performance in these games serves as a quantifiable criterion with a clear operational meaning, aligned with the real-world legal context. We also prove that the widely adopted Near-Access-Free (NAF) copyright framework fails to provide meaningful guarantees for either game. Our theoretical findings are supported by empirical evaluations on image diffusion models, highlighting the potential of our framework for informing legal thresholds and guiding AI copyright regulation.

1. Introduction

Generative AI models have achieved remarkable success, driven by training on extensive and diverse datasets spanning various domains, including images, text, code, music, and more (Ramesh et al., 2021; 2022; Rombach et al., 2022; Brown et al., 2020; Achiam et al., 2023; Li et al., 2022; Dhariwal et al., 2020). However, this success raises growing concerns about whether the outputs of these models constitute copyright infringement (Lee et al., 2023; Franceschelli & Musolesi, 2022; Samuelson, 2023; Zirpoli, 2023). Given the vast volume of training data, excluding all copyrighted content is often impractical. Moreover, using copyrighted material does not necessarily result in infringement, as long as the model's output is not a direct *copy* of protected content or can be justified as fair use (Elkin-Koren et al., 2024). A notable example is the recent lawsuit filed by The New York Times against OpenAI for *copying* millions of the Times's copyrighted news articles, in-depth investigations, opinion pieces, reviews, how-to guides, and more[1]. This case highlights the need for formal AI copyright regulations and a pressing question: "How can we determine whether an AI model infringes on copyright?"

This is a complex and ambitious question, one that even experts in AI copyright law struggle to answer with a single objective criterion. This challenge also limits the progress of establishing proper AI copyright regulations. Copyright laws and legal precedents are inherently subjective; for example, *the ordinary observer test* is one of the most widely applied copyright tests, where an "ordinary observer" will determine if a result is substantially similar (Con, 1988; Scheffler et al., 2022). This is subjective by nature since it depends on how the chosen "ordinary observer" feels. Rather than defining a legal standard, we propose to focus on designing criteria that provide quantitative evidence to assist lawyers and judges in determining whether the outputs of an AI model constitute copyright infringement, and can be further utilized as an AI copyright regulation.

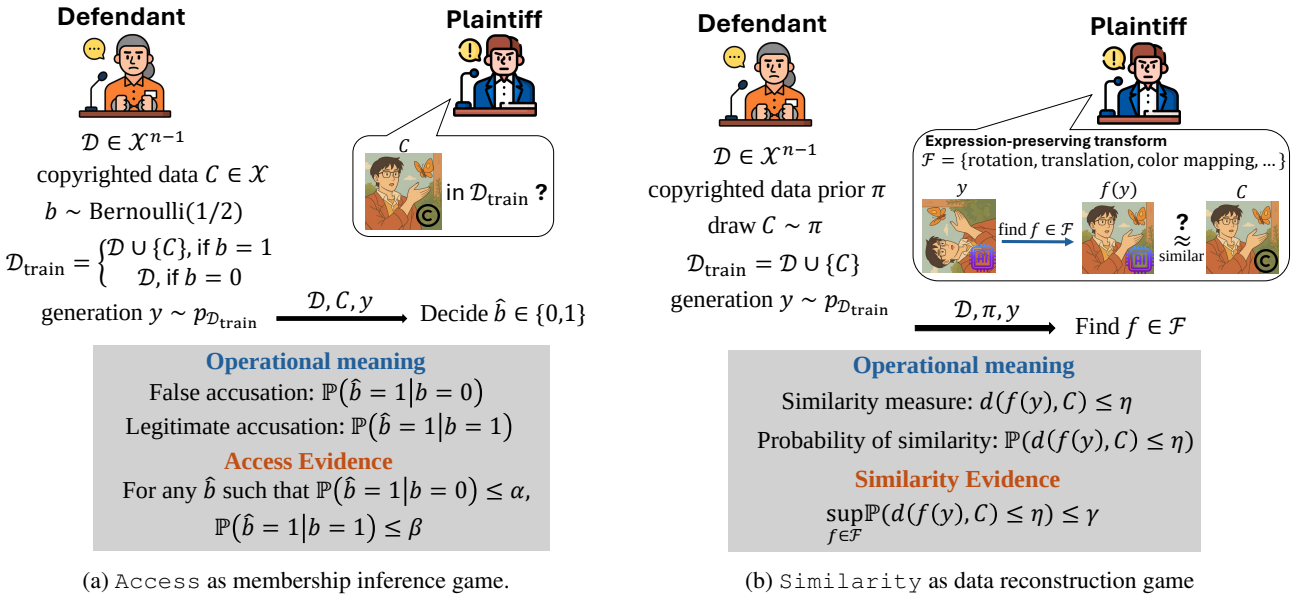
We begin by revisiting U.S. copyright law, specifically as interpreted by the Ninth Circuit: "*On the plaintiff's copyright infringement claim, the plaintiff has the burden of proving by a preponderance of the evidence that: . . ., and the defendant copied original expression from the copyrighted work.*"[2] To establish that an AI model's output constitutes a copy of copyrighted content, the plaintiff must demonstrate at least two necessary (but not necessarily sufficient) key

[*]Equal contribution [1]School of Electrical and Computer Engineering, Georgia Institute of Technology, U.S.A.. Correspondence to: Eli Chien <elichien@icloud.com>, Pan Li <panli@gatech.edu>.

Published at ICML 2025 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Vancouver, Canada. PMLR 267, 2025.

[1]https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

[2]https://www.ce9.uscourts.gov/jury-instructions/node/261



(a) `Access` as membership inference game.

(b) `Similarity` as data reconstruction game

Figure 1. Copyright evidence framework of `Access` and `Similarity`.

elements: (1) `Access`—that the AI model was trained on the copyrighted content, and (2) `Similarity`—that the model's output is substantially similar to the copyrighted work. If the defendant exhibits strong evidence of disproving either `Access` or `Similarity` by satisfying some criteria, they are unlikely to infringe copyright. Such criteria can thus serve as the foundation of AI copyright regulations.

Our contributions. We propose formal definitions that provide shreds of evidence pertaining to the two key elements of copyright infringement—`Access` and `Similarity`—with clear operational meanings. Having a clear operational meaning is critical, as the criterion must be understandable to lawyers and judges for serving as evidence in the actual copyright lawsuit, as well as the AI copyright regulations. Our key idea is to formulate it as two game-theoretic problems among two players: the plaintiff and the defendant. The goal of the plaintiff is to prove `Access` (and `Similarity` respectively) while the defendant aims to disprove it, similar to the actual lawsuit scenario. We show that the membership inference game (Shokri et al., 2017) directly corresponds to `Access`, while the data reconstruction game (Guo et al., 2022; Balle et al., 2022) is related to `Similarity`. Our criterion provides a formal quantitative measure regarding these two aspects by characterizing the "performance of the plaintiff" in each game and thus clear operational meanings. As a result, our criterion serves as a strong candidate for AI copyright regulations with qualitative measures. Furthermore, we prove that the popular Near-Access-Free (NAF) copyright definition (Vyas et al., 2023) does not provide meaningful guarantees in these two games, which the authors argued to be related to `Access` and `Similarity`. By constructing counterexamples, we demonstrate that even a 0-NAF (perfectly compliant) model can still violate both `Access` and `Similarity`. We provide experiments on image generative (diffusion) models against practical attackers in the aforementioned games, which illustrate that NAF and the achieving algorithms are inadequate for providing evidence of both `Access` and `Similarity`. Our results underscore the advantage of our criterion, which is aligned with real-world copyright law.

Due to the space limit, we defer the discussion on related works in Appendix A.

2. The Theory of Copyright Evidence Framework

An ideal copyright criterion should align with real-world copyright law and regulations, and must exhibit clear operational meaning that allows general audiences and law experts outside the computer science field to understand and utilize it. In this section, we propose a copyright evidence framework that focuses on providing theoretical evidence with clear operational meaning pertaining to `Access` (AI model was trained on the copyrighted content) (Figure 1a) and `Similarity` (AI model's output is substantially similar to the copyrighted work) (Figure 1b). We focus on providing rigorous evidence about aspects that are necessary conditions for copyright infringement, bypassing the challenge of defining subjective aspects in copyright law. While our main focus is to establish the copyright evidence framework instead of algorithms tailored to it, we briefly discuss a naive approach for achieving it and how it can be

potentially improved at the end of this section. We relegate a detailed comparison of copyright evidence to the privacy literature in Appendix B.

Notation. Let $\mathcal{D} = \{x_1, ..., x_n\} \in \mathcal{X}^n$ be a dataset. A training algorithm $\mathcal{A}$ applied to $\mathcal{D}$ produces a generative model $\mathcal{A}(D) = p_{\mathcal{D}}$. We simply write p whenever the training dataset is clear in the context. We denote $p(\cdot|z)$ as the generating probability distribution over $\mathcal{X}$ given a prompt z. Under the promptless setting, we simply use $p(\cdot)$. A generation y from p is denoted by $y \sim p$. We denote $C \in \mathcal{X}$ as the copyrighted data of interest.

While our discussion in this section primarily focuses on the promptless setting $p(\cdot)$, the formulation and the proposed criterion of copyright evidence naturally generalizes to the prompt-conditioned case $p(\cdot|z)$ for any given prompt z, as a promptless model can be viewed as a special instance of a prompt-conditioned model with a fixed prompt.

2.1. Criterion of Defendant's Evidence on `Access`

To establish copyright infringement through `Access`, the plaintiff must demonstrate that the defendant (i.e., the model developer) used a piece of copyrighted data during the training process. This naturally leads to a game-theoretic setting between the plaintiff and the defendant: the defendant trains a generative model, while the plaintiff seeks to determine whether the model has utilized the copyrighted data based on its generated samples. This is well-known as the membership inference game (Shokri et al., 2017).

Definition 2.1 (Membership inference game)**.** Let $\mathcal{D}$ be a dataset and C be a piece of copyrighted data. The defendant tosses a fair coin b: if the outcome is heads ($b = 1$), set $\mathcal{D}_{\text{train}} = \mathcal{D} \cup \{C\}$ and $\mathcal{D}_{\text{train}} = \mathcal{D}$ otherwise ($b = 0$). She then trains a generative model $p_{\mathcal{D}_{\text{train}}}$. The plaintiff aims to make a decision $\hat{b}$ to decide if $b = 1$ ($\mathcal{D}_{\text{train}} = \mathcal{D} \cup \{C\}$) or $b = 0$ ($\mathcal{D}_{\text{train}} = \mathcal{D}$) using the following information: (1) dataset $\mathcal{D}$ and the copyrighted data C; (2) generated samples $y \sim p_{\mathcal{D}_{\text{train}}}$.

The performance of the plaintiff in a membership inference game can serve as quantitative evidence for the legal notion of `Access`. If the plaintiff performs poorly in the game—they fail to prove that the copyrighted data is contained in the training dataset—it constitutes strong evidence against `Access`. Accordingly, a low plaintiff performance serves as a strong AI copyright regulation as well.

Criterion of `Access` Evidence. The performance of a membership inference attack is characterized by the Receiver Operating Characteristic curve (ROC curve), which shows the the tradeoff curve between true positive rate (TPR) $\mathbb{P}(\hat{b} = 1|b = 1)$ (when the plaintiff correctly identifies access) and false positive rate (FPR) $\mathbb{P}(\hat{b} = 1|b = 0)$ (when the plaintiff falsely accuses the defendant) at different decision thresholds. AUROC (Area under the ROC curve) is a widely-used summary metric that equally treats positive and negative cases. However, we argue that AUROC is not a proper metric for copyright access criterion, due to the asymmetric role of membership prediction and non-membership prediction. The accurate prediction of non-membership $\mathbb{P}(\hat{b} = 0|b = 0)$ is not a primary concern, since AI models without access to the copyrighted data cannot violate copyright. Consequently, the prediction of membership, given by the TPR probability $\mathbb{P}(\hat{b} = 1|b = 1)$, is more relevant. To deal with this evaluation asymmetry, a better evaluation metric than AUROC is the TPR at a low FPR (Carlini et al., 2022) to ensure strong power of the plaintiff. Based on this analysis, we propose a formal definition based on the performance of a defendant's model in the membership inference game.

Definition 2.2 ((α, β)-Access-Evidence)**.** A generative model g satisfies (α, β)-Access-Evidence w.r.t. copyrighted data C and dataset $\mathcal{D}$, if for all the membership predictors $\hat{b}$ defined in Definition 2.1 (w.r.t. $\mathcal{D}, C$) such that $\mathbb{P}(\hat{b} = 1|b = 0) \leq \alpha$, the predictor $\hat{b}$ satisfies $\mathbb{P}(\hat{b} = 1|b = 1) \leq \beta$.

Operational meaning of α, β. The parameter β is a threshold for the probability that the plaintiff correctly identifies the data access. A lower β restricts the "successful rate" of membership prediction, thereby enforcing stronger regulation on the AI model's use of copyrighted data. The parameter α is a threshold for the probability of the plaintiff making a false accusation against the defendant. A plaintiff with a high false accusation rate is unreliable, and thus their claims about data access should not be considered meaningful. For example, a plaintiff could trivially accuse every model of using copyrighted data regardless of evidence. Although this strategy would achieve perfect detection when copyrighted data is actually used, the corresponding error probability, $\mathbb{P}(\hat{b} = 1|b = 0) = 1$, renders such claims meaningless. By setting α (typically a small number, e.g., 0.05), we filter out such invalid predictors, ensuring that only meaningful predictors are considered.

2.2. Criterion of Defendant's Evidence on `Similarity`

Expression-preserving Transforms. In practice, U.S. courts also test for "substantial similarity" to decide whether infringement has occurred (pet; alt; Samuelson, 2012). The overall goal of any such test is to determine whether a transformed, derived work retains protectable elements of the original work's expression, or whether it has diverged sufficiently to constitute a new, non-infringing work. However, defining and testing "substantial similarity" is often subjective (Scheffler et al., 2022). While there are some natural similarity measures for common domains, such as

the ℓ_2 distance $d(x, y) = \|x - y\|_2$ for images, edit distance $d(x, y) = \text{Edit}(x, y)$ for natural languages, one key pitfall of these similarity measures is that they only consider the "absolute similarity" and ignore the possibility of underlying shared expression. One naive example of this possibility, in the context of image generation, is when a model generates a rotated or color-hashed version of a copyrighted image in the training dataset. This transformed image retains "substantial similarity" from the perspective of lawmakers and can be argued to infringe on the original work's copyright, but it may also have large ℓ_2 distance from the original work. In other words, certain transformations of the work that do not add new expressions may not be considered transformative enough to avoid copyright infringement.

Let us denote the set of such abstract transformations by $\mathcal{F}$, and assume $\mathcal{F}$ is given in the following discussion. It is worth noticing that determinations about what constitutes an expression-preserving transformation are domain-specific and often made on a case-by-case basis. Of course, $\mathcal{F}$ should at least include the identity map, i.e., direct comparison. In the end, we will also discuss how to bypass the difficulty of determining $\mathcal{F}$ by relaxing it to all possible reconstruction functions in the context of a data reconstruction game, providing a stronger regulating criterion.

`Similarity` as Data Reconstruction Game. Suppose the plaintiff is able to identify an expression-preserving transformation $f \in \mathcal{F}$ such that the transformed generated sample closely resembles the copyrighted content. Then it is strong evidence of potential copyright infringement. Note that this process can be viewed as the plaintiff selecting a *reconstruction function* from $\mathcal{F}$, with the goal of accurately recovering the original content C from the generated sample y. This perspective naturally leads to a data reconstruction game between the plaintiff and the defendant.

Definition 2.3 (Data reconstruction game)**.** Let $\mathcal{D}$ be a dataset with copyrighted data C. Let $p_{\mathcal{D}\cup\{C\}}$ be a generative model trained on dataset $\mathcal{D} \cup \{C\}$ by the defendant. The plaintiff aims to find a transformation $f \in \mathcal{F}$ to recover C based on the following information: (1) the dataset $\mathcal{D}$; (2) generated samples $y \sim p_{\mathcal{D}\cup\{C\}}$; (3) prior (knowledge) π about C.

Criterion of `Similarity` Evidence. The performance of the plaintiff in the data reconstruction game quantifies the evidence for `Similarity`. To reduce the likelihood of generating copyright-infringing content, we seek a provable guarantee such that none of the transformations in $\mathcal{F}$ leads to a low reconstruction error of C. A possible candidate to measure the reconstruction error is the (normalized) Mean Square Error (MSE) (Guo et al., 2022). In our case, we can write the expected reconstruction error as $\inf_{f\in\mathcal{F}} \mathbb{E}_{C\sim\pi, y\sim p_{\mathcal{D}\cup\{C\}}} d(f(y), C)$ with $d(f(y), C) = \ell_2(f(y), C) = \|f(y) - C\|_2/\sqrt{d}$, where d is the dimension of the generated data. However, we argue that this average-case error is not a suitable metric in the context of copyright similarity, since a model generating copyrighted content with high probability could yield a large MSE. For instance, assume $y = C$ with probability 0.99 and $y = C + \Delta$ otherwise. This deviation Δ can be arbitrarily large, e.g, $\Delta = 10^6$, which "hacks" the regulation based on MSE. The model can still generate the copyrighted content C with probability 0.99, which is a clear copyright infringement. Alternatively, we propose to directly restrict the probability that similar samples are generated.

Definition 2.4 ((η, γ)-Similarity-Evidence)**.** A generative model p satisfies (η, γ)-Similarity-Evidence w.r.t. dataset $\mathcal{D}$, prior π and function class $\mathcal{F}$, if

$$\sup_{f\in\mathcal{F}} \mathbb{P}_{C\sim\pi, y\sim p_{\mathcal{D}\cup\{C\}}}(d(f(y), C) \le \eta) \le \gamma. \quad (1)$$

In practice, the set of expression-preserving transformations $\mathcal{F}$ is domain-specific and may be difficult to determine or characterize. In this case, from the perspective of reducing the risk of copyright infringement and regulating generated content, one conservative choice is to consider all possible reconstruction functions f applied to samples $y \sim p_{\mathcal{D}\cup\{C\}}$. This relaxation provides a stronger criterion and bypasses the difficulty of specifying $\mathcal{F}$ when it is hard to define.

Operational Meaning of η, γ. The error parameter η measures the reconstruction error, which quantifies the threshold under which the generated and copyrighted contents are considered "similar" after applying all kinds of expression-preserving transforms. The probability parameter γ measures how likely "similar" contents will be generated by the AI models. The specific choice of η, γ should be determined by the lawmaker, and the operational meaning helps to make decisions more clearly.

We leave the additional discussion on the choice of prior π, as well as the naive achieving algorithm using noisy gradient training in Appendix C.

Other results. In Appendix D we prove that the popular NAF copyright criterion is unable to provide meaningful evidence regarding both `Access` and `Similarity` in our copyright games. It shows the importance of any copyright criterion, even merely as evidence for a partial aspect of the copyright, should have a clear operational meaning. Otherwise, it may disconnect with the real-world copyright law and is inappropriate to be set as part of the copyright regulations. In Appendix E, we support our results by running practical algorithms that are available for the plaintiff in both `Access` and `Similarity` copyright games, where the defendant will take either NAF-based strategies or procedures that can naively achieve our copyright evidence.

Acknowledgements

We thank Wei-Ning Chen for the helpful discussion. E. Chien, A. Saha, Y. Huang, and P. Li are supported by NSF awards CIF-2402816, PHY-2117997, and JPMC faculty award.

References

Computer associates international, inc. v. altai, inc., 982 f.2d 693 (2d cir. 1992).

Peter pan fabrics, inc. v. martin weiner corp., 274 f.2d 487 (2d cir. 1960).

Concrete machinery v. classic lawn ornaments, 1988.

Abad, J., Donhauser, K., Pinto, F., and Yang, F. Copyright-protected language generation via adaptive model fusion. *arXiv preprint arXiv:2412.06619*, 2024.

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 2496–2506. PMLR, 2020.

Balle, B., Cherubin, G., and Hayes, J. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1138–1156. IEEE, 2022.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Borjesson, P. and Sundberg, C.-E. Simple approximations of the error function q (x) for communications applications. *IEEE Transactions on Communications*, 27(3):639–643, 1979.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

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

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models, 2023. URL https://arxiv.org/abs/2301.13188.

Chiba-Okabe, H. and Su, W. J. Tackling copyright issues in ai image generation through originality estimation and genericization. *Scientific Reports*, 15(1):10621, 2025.

Chu, T., Song, Z., and Yang, C. How to protect copyright data in optimization of large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17871–17879, 2024.

Cohen, A. Differential privacy in the clean room: Copyright protections for generative ai.

Deng, J. and Ma, J. Computational copyright: Towards a royalty model for AI music generation platforms. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024. URL https://openreview.net/forum?id=CIQxqQvkEE.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

Dwork, C. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Elkin-Koren, N., Hacohen, U., Livni, R., and Moran, S. Can copyright be reduced to privacy? In *5th Symposium on Foundations of Responsible Computing (FORC 2024)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2024.

Franceschelli, G. and Musolesi, M. Copyright in generative deep learning. *Data & Policy*, 4:e17, 2022.

Golatkar, A., Achille, A., Zancato, L., Wang, Y.-X., Swaminathan, A., and Soatto, S. Cpr: Retrieval augmented generation for copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12374–12384, 2024.

Guo, C., Karrer, B., Chaudhuri, K., and van der Maaten, L. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning*, pp. 8056–8071. PMLR, 2022.

Hayes, J., Balle, B., and Mahloujifar, S. Bounding training data reconstruction in dp-sgd. *Advances in neural information processing systems*, 36:78696–78722, 2023.

He, L., Huang, Y., Shi, W., Xie, T., Liu, H., Wang, Y., Zettlemoyer, L., Zhang, C., Chen, D., and Henderson, P. Fantastic copyrighted beasts and how (not) to generate them. *arXiv preprint arXiv:2406.14526*, 2024.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 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, 2023. URL https://arxiv.org/abs/2305.18355.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lee, K., Cooper, A. F., and Grimmelmann, J. Talkin"bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*, 2023.

Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Min, S., Gururangan, S., Wallace, E., Shi, W., Hajishirzi, H., Smith, N. A., and Zettlemoyer, L. SILO language models: Isolating legal risk in a nonparametric datastore. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ruk0nyQPec.

Neyman, J. and Pearson, E. S. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

Panaitescu-Liess, M.-A., Che, Z., An, B., Xu, Y., Pathmanathan, P., Chakraborty, S., Zhu, S., Goldstein, T., and Huang, F. Can watermarking large language models prevent copyrighted text generation and hide training data? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25002–25009, 2025.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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.

Samuelson, P. A fresh look at tests for nonliteral copyright infringement. *Nw. UL Rev.*, 107:1821, 2012.

Samuelson, P. Generative ai meets copyright. *Science*, 381 (6654):158–161, 2023.

Scheffler, S., Tromer, E., and Varia, M. Formalizing human ingenuity: A quantitative framework for copyright law's substantial similarity. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pp. 37–49, 2022.

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.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

Vyas, N., Kakade, S. M., and Barak, B. On provable copyright protection for generative models. In *International Conference on Machine Learning*, pp. 35277–35299. PMLR, 2023.

Wang, J. T., Deng, Z., Chiba-Okabe, H., Barak, B., and Su, W. J. An economic solution to copyright challenges of generative ai. *arXiv preprint arXiv:2404.13964*, 2024.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

Zhang, R., Hussain, S. S., Neekhara, P., and Koushanfar, F. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1813–1830, 2024.

Zhao, X., Ananth, P., Li, L., and Wang, Y.-X. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023a.

Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.-M., and Lin, M. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023b.

Zirpoli, C. T. Generative artificial intelligence and copyright law. 2023.

A. Related Works

Although there are many works on copyright questions for generative AI (Somepalli et al., 2023; Min et al., 2024; He et al., 2024; Panaitescu-Liess et al., 2025), remarkably few have attempted to tackle the problem from a theoretical perspective. (Vyas et al., 2023) is the first to propose a mathematical framework, NAF, that attempts to quantify the degree of copyright infringement for a generative AI system. They also propose a black-box reduction, the Copy-Protection-Δ algorithm (CP-Δ) that converts any generative model to a model satisfying the NAF criterion. This claim is powerful and appealing, where there has already been some follow-up work developing new algorithms that satisfy the NAF criterion (Golatkar et al., 2024; Abad et al., 2024). However, we show that NAF fails to even provide meaningful evidence for both `Access` and `Similarity`, let alone whether it is an appropriate measure for AI copyright infringement.

Prior works have also critiqued the NAF framework from different aspects. (Elkin-Koren et al., 2024) argues that copyright cannot be reduced to algorithmic stability, such as differential privacy (DP) (Dwork et al., 2014) and NAF, due to its inherent complexity, such as fair use. (Henderson et al., 2023; Lee et al., 2023) argues against NAF in cases when the original expression of one work appears in other works, which leads to a similar conclusion above. We agree with these statements, motivating our focus on providing evidence for `Access` and `Similarity` instead of directly determining copyright infringement, leaving the judgment to actual copyright law experts. In fact, we believe it is impossible to have a single mathematical criterion for this purpose, as we state in the introduction. (Cohen) is the closest related work to ours, where they show that when users use many prompts, the NAF-guaranteed model can generate training data samples. In contrast, we prove that even in the promptless setting, the NAF-guaranteed model's output can be close to the copyrighted training data in ℓ_2 distance with high probability.

Unlike NAF, (Scheffler et al., 2022) addresses a distinct problem: determining "substantial similarity" in legal contexts. The authors propose a complexity-theoretic similarity test based on the description length required to derive one specific work from another. Their framework is designed to measure the similarity between two specific samples, regardless of whether they are AI-generated. In contrast, we directly regulate the ability of the AI model to generate similar output before its actual generation process, and thus, orthogonal to their approach. (Chiba-Okabe & Su, 2025) proposes a distance-based originality measure and the corresponding genericization for reducing the risk of copyright infringement. Their work can be viewed as another attempt that tries to provide evidence regarding `Similarity`, but it is still different from our proposal of leveraging the data reconstruction game and it does not tackle the problem regarding `Access`. (Chu et al., 2024) proposes measuring the degree of copyright infringement by comparing the average loss on copyrighted versus non-copyrighted data in the training set, and aims to mitigate the risk of generating copyrighted content by increasing this loss gap during training. However, the loss gap is a heuristic and does not offer a rigorous guarantee of copyright protection. Besides trying to provide theoretical measures related to AI copyright infringement, there are works that instead focus on designing platforms for distributing revenues to copyrighted content holders based on Shapely values (Wang et al., 2024) or other data attribution techniques (Deng & Ma, 2024). This is a very interesting direction but orthogonal to our work.

Another line of research focuses on watermarking generative models, that is, injecting detectable signals into generated samples to enable identification of whether a sample originates from a specific model (Kirchenbauer et al., 2023; Zhao et al., 2023b;a; Zhang et al., 2024). Although watermarking was not originally designed to address copyright concerns, recent empirical studies have shown that watermarking language models can reduce the generation of copyrighted content and mitigate membership inference attacks on copyrighted training data (Panaitescu-Liess et al., 2025). Nevertheless, watermarking alone does not provide a formal framework for measuring copyright infringement, nor does it offer a rigorous guarantee of copyright protection.

B. Comparison of Copyright Evidence to Privacy Definitions

Comparing (α, β)-Access-Evidence with Differential Privacy. Differential privacy is a popular notion of data privacy that provides provable protection against membership inference attack (Dwork, 2006; Dwork et al., 2014; Balle et al., 2020). The idea is to ensure the probabilistic distribution of model weights/outputs trained on a dataset is robust to the removal of any individual data point. A generative model satisfying differential privacy thus offers guarantees for the membership inference game and thus for (α, β)-Access-Evidence. However, there are some key differences between Definition 2.2 and differential privacy, which suggests differential privacy may not be an ideal notion in the context of `Access`. For example, a (ε, δ)-DP generative model satisfies $(\alpha, e^{\varepsilon}\alpha + \delta)$-Access-Evidence. Nevertheless, the original definition of (ε, δ)-DP is not designed to directly characterize the trade-off curve between α and β, and it does not distinguish membership prediction and non-membership prediction, which may make the derived bounds (e.g., $e^{\varepsilon}\alpha + \delta$) too loose. f-DP (Dong et al., 2022) instead

directly defines the differential privacy in terms of the trade-off between α and β. If A generative model is f-DP, then it satisfies $(\alpha, 1 - f(\alpha))$-Access-Evidence for any α by definition. Still, f-DP (and (ε, δ)-DP as well) requires the privacy guarantee to hold for any dataset and any data point removal, which can be overly restrictive and may compromise model utility. In contrast, for copyright `Access`, we may only be concerned with a particular dataset $\mathcal{D}$ held by model developers and some specified copyrighted data C. It allows us to retain a better model performance under our Access-Evidence constraint compared to the differential privacy framework.

Comparing (η, γ)-Similarity-Evidence with Other Data Reconstruction Guarantees. (Guo et al., 2022) proposes to measure the degree of data reconstruction guarantees by MSE $\mathbb{E}_{x \sim g_{\mathcal{D} \cup \{C\}}} \|f(x) - C\|$. However, as we discussed earlier, the notion of average error is not a suitable metric for the copyright context, since a large MSE could still yield a high probability of generating verbatim copy of the protected data. On the other hand, (Balle et al., 2022; Hayes et al., 2023) proposes to directly bound the probability of approximately correct reconstruction $\mathbb{P}_{C \sim \pi, x \sim g_{\mathcal{D} \cup \{C\}}} (d(f(x), C) \leq \eta) \leq \gamma$. This is a stronger requirement compared to our Definition 2.4, in the sense that the guarantee of (Balle et al., 2022; Hayes et al., 2023) holds for arbitrary reconstruction functions and arbitrary dataset $\mathcal{D}$. As a result, one would sacrifice more model utility in satisfying such a guarantee. In contrast, for copyright `Similarity`, only a particular class of transformation functions $\mathcal{F}$ and a specific dataset $\mathcal{D}$ held by model developers are considered.

C. Additional Explanation of Copyright Evidence

The Choice of Prior π. In Definition 2.3, one important component is the prior π that the data reconstructor can leverage. Indeed, the more π concentrated around the copyrighted content, the "harder" to ensure low (η, γ) in Definition 2.4. Consider the case that $\mathcal{F}$ is the set of all possible reconstruction functions. Set $\pi = \mathbf{1}_C$, the Dirac delta measure on C. Then there is no generative model p that can satisfy Definition 2.4 with low (η, γ), since there is always a trivial reconstruction function $f(y) = \text{argmax}_y \pi(y) = C$ that perfectly recovers C without leveraging the generated output $y \sim p$. In this case, the "success" of the plaintiff in the data reconstruction game is not because the defendant's output is substantially similar to the copyrighted content, but instead because the prior π reveals too much information about the copyrighted data C. It is hence required that the lawmaker choose a reasonable π so that the data reconstruction game reflects the main purpose pertaining to `Similarity`. The non-informative prior $\pi = \text{Uniform}(\mathcal{X})$ indicates that any information leakage is purely from samples $y \sim p$. It is also possible to choose a more informative prior (even $\pi = \mathbf{1}_C$) when the choice of $\mathcal{F}$ avoids the trivial reconstruction issue mentioned above. We leave the choice of π to be decided by the lawmaker as well.

Remark C.1 (Naive achieving algorithm). A straightforward approach to providing guarantees for `Access` and `Similarity`, as defined in Definitions 2.2 and 2.4, is to train generative models using noisy gradient methods such as DP-SGD (Abadi et al., 2016) based on the analysis of (Dong et al., 2022; Balle et al., 2022; Hayes et al., 2023). However, we stress that **this does not imply that our copyright evidence reduces to privacy guarantees.** As discussed, privacy guarantees impose stricter requirements on the model, necessitating greater noise injection during training and thus leading to degraded model utility. We believe there exist algorithms tailored for copyright that are better suited than the naive application of DP-SGD for achieving our copyright evidence goals, highlighting a promising direction for future research.

D. NAF Cannot Provide Evidence for `Access` and `Similarity`

We demonstrate that the NAF criterion fails to provide appropriate copyright evidence with respect to `Access` and `Similarity`. This highlights the importance of our proposed criteria and, more broadly, underscores that any copyright criterion should have a clear and rigorous operational meaning aligned with the principles of copyright law. We begin by introducing the NAF criterion.

Definition D.1 (Near Access Freeness (Vyas et al., 2023))**.** Let $\mathcal{C}$ be a set of copyrighted datapoints, let $\text{safe} : \mathcal{C} \to \mathcal{M}$ be a safe reference model, and let Δ be a divergence between distributions. We say that a generative model p is k_z-*near access-free* on some prompt z with respect to $\mathcal{C}$, safe, and Δ if for all $C \in \mathcal{C}$, $\Delta\left(p(\cdot|z) \| \text{safe}_C(\cdot|z)\right) \leq k_z$.

Algorithm 1 CP-Δ algorithm (Vyas et al., 2023)

Input: Divergence $\Delta \in \{\Delta_{\text{max}}, \Delta_{\text{KL}}\}$, dataset $\mathcal{D}$, learning algorithm $\mathcal{A}$.

Learning: Partition $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$ and set $q_i = \mathcal{A}(\mathcal{D}_i),\ i \in [2]$.

return *Model p, where $N(z)$ is the normalizing constant and*

$$p(y|z) = \begin{cases} \frac{\min(q_1(y|z), q_2(y|z))}{N(z)} & \text{if } \Delta = \Delta_{max} \\ \frac{\sqrt{q_1(y|z) \cdot q_2(y|z)}}{N(z)} & \text{if } \Delta = \Delta_{KL} \end{cases}$$

(Vyas et al., 2023) focuses on the sharded-safe model defined as follows. For any learning algorithm $\mathcal{A}$, dataset $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$ with partition $\mathcal{D}_1, \mathcal{D}_2$, and copyrighted data $C \in \mathcal{D}$, we have $\text{safe}_C = \mathcal{A}(\mathcal{D}_i)$, where $C \notin \mathcal{D}_i,\ i \in [2]$. In the rest of the manuscript, we will focus on this choice of safe model as well. They propose the Copy-Protection-Δ algorithm (CP-Δ Algorithm 1), which is an ad-hoc approach allowing generative models to satisfy a k_z-NAF guarantee where k_z depends on the divergence between distributions q_1, q_2 conditioned on the prompt z. See Appendix F for more details. Note that the "best" model under the NAF criterion is the one satisfying the 0-NAF guarantee due to the non-negativity of divergence Δ. The authors of (Vyas et al., 2023) interpret the k-NAF guarantee as leaking k bits of information about the copyrighted content.

We challenge this claim by theoretically proving that a 0-NAF guarantee does **not** provide meaningful evidence for both `Access` and `Similarity`. We will prove this by constructing counterexamples, where even if the defendant obtains a 0-NAF generative model via the CP-Δ algorithm, the model is still vulnerable against membership inference attacks (MIA) and data reconstruction attacks (DRA). Numerical results are presented in Figure 2 against provably successful attackers. Our results indicate that even if the defendant provides a 0-NAF guarantee of their model, it is possible that the plaintiff can still accurately infer whether a copyrighted content C is used (`Access`) or the model output will be close to C with high probability (`Similarity`). We now state our main results regarding this argument.

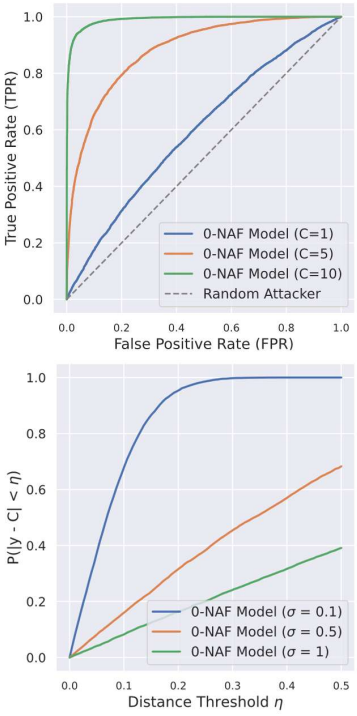


Figure 2. 0-NAF model against MIA (top) and output distance to C (bottom) based on counterexamples in Theorem D.2 and D.3.

Theorem D.2 (0-NAF Does Not Provide Evidence for `Access`)**.** *There exists a dataset $\mathcal{D}$, generative model p and a membership inference attacker* `Attack` *such that p is 0-NAF with divergence choice Δ_{KL} or Δ_{max} but* `Attack` *achieves FPR α and TPR β with any $\alpha, \beta \in (0, 1)$.*

Theorem D.2 states that even if a generative model p has 0-NAF guarantee, there is still an attack (i.e. membership inference) that can discern if p is trained with specific copyrighted content that achieves arbitrarily high accuracy. As a result, **Theorem D.2 implies that NAF does not provide meaningful evidence for `Access`**, since even a 0-NAF model can be proven to have access to specific copyright content in the training dataset under the membership inference attack game.

Proof Sketch. Our proof relies on providing a dataset $\mathcal{D}$ and the learning algorithm $\mathcal{A}$ that will output an generative model q. We show that there exists an instance $(\mathcal{D}, \mathcal{A})$ such that running CP-Δ (Algorithm 1) provides 0-NAF guarantees, but there exists a membership inference attacker `Attack` with arbitrarily well performance. Specifically, consider the dataset $\mathcal{D} = \{0, 0, -C, C\}$ for some copyrighted data $C > 0$ (a scalar here for simplicity) and the learned generative model $q = \mathcal{A}(\mathcal{D}) = N(\mu(\mathcal{D}), 1)$. This is the simplified Gaussian mixture model that learns the mean with a deterministic rule of averaging the training data. Firstly, note that the partitioned dataset $\mathcal{D}_1 = \{0, 0\}, \mathcal{D}_2 = \{-C, C\}$ both gives identical generative model $q_1 = q_2 = N(0, 1)$. It is obvious that the final generative model p of Algorithm 1 satisfies the 0-NAF guarantee. In the meanwhile, Algorithm 1 will clearly give a different generative model p' when the input dataset is $\mathcal{D}' = \{0, 0, -C\}$ that does not contain the copyright content C. By setting `Attack` to be the likelihood ratio test, we can characterize the TPR and FPR rate by the KL divergence between p, p'. Finally, we conclude our proof by showing both $\Delta_{\text{KL}}(p||p'), \Delta_{\text{KL}}(p'||p) \to \infty$ as $C \to \infty$ for both options of $\Delta_{\text{KL}}, \Delta_{\text{max}}$ in the CP-Δ algorithm. □

We prove that **a 0-NAF guarantee does not provide meaningful evidence for `Similarity`** as well.

Theorem D.3 (0-NAF Does Not Provide Evidence for `Similarity`)**.** *There exists a dataset $\mathcal{D}$ and a generative model p, such that p is 0-NAF with divergence choice Δ_{KL} or Δ_{max}, but still generates substantially similar output to the copyrighted content $C \in \mathcal{D}$ with arbitrarily high probability. Specifically, for any $C > 0$ and $\gamma \in (0, 1]$, there exists an instance $(\mathcal{D}, C, p)$ with 0-NAF guarantee such that $\mathbb{P}_{Z \sim p}(\|Z - C\| \leq \eta) \geq 1 - \gamma$.*

Proof Sketch. We again consider the simplified Gaussian mixture model as the generative model $q = \mathcal{A}(\mathcal{D}) = N(\mu(\mathcal{D}), \sigma^2)$. Consider the dataset $\mathcal{D} = \{C_1, C_2, C\}$ for some copyrighted data $C \in \mathbb{R}$ and $(C_1 + C_2)/2 = C$. Then both the partitioned

dataset $\mathcal{D}_1 = \{C_1, C_2\}$ and $\mathcal{D}_2 = \{C\}$ give identical generative model $q_1 = q_2 = N(C, \sigma^2)$. It is obvious that the final generative model p of Algorithm 1 satisfies the 0-NAF guarantee and $p = N(C, \sigma^2)$. This construction guarantees that for any $C > 0$ and $\gamma \in (0, 1]$, there is always a sufficiently small σ such that Theorem D.3 holds. □

The counterexamples presented above, though simple and specific, already demonstrate that the NAF criterion is insufficient as evidence for either `Access` or `Similarity`. In contrast, our proposed notions of evidence, Definitions 2.2 and 2.4, have precise operational meanings and are grounded in a formal game-theoretic formulation between plaintiff and defendant. These definitions directly quantify the strength of evidence regarding `Access` and `Similarity`, providing a more principled and interpretable foundation for assessing copyright-related claims.

E. Experiments: Against the Real-World Validator

In theory, we have demonstrated that NAF is inadequate as a criterion for disproving `Access` and `Similarity`. We show that there exist strong MIA or DRA approaches for the plaintiff to confidently predict if the defendant's 0-NAF model was trained using the copyrighted data C, or is highly likely to output a sample y that is substantially similar to C. However, such strong approaches may not be available in practice, where practical MIA and DRA approaches are often weaker. In this section, we empirically evaluate the strength of copyright evidence provided by different copyright protection mechanisms for image generation tasks. This simulates how the defendant and plaintiff may realistically play the copyright game with practical approaches during a copyright trial.

E.1. Experiment Setup

We train conditional diffusion models using the DDPM framework (Ho et al., 2020) on the CIFAR-10 dataset (Krizhevsky et al., 2009). We test three approaches of the defendant (i.e. model provider) in the MIA and DRA game. First, we consider the case where the defendant trains a baseline diffusion model. The second case considers modified diffusion models with k-NAF guarantee. The third case assumes the defendant has trained a diffusion model with DP-Adam (Abadi et al., 2016) ($\varepsilon \in \{100, 500, 1000\}$, $\delta = 10^{-5}$) via Opacus (Yousefpour et al., 2021), which can naively provide (α, β)-Access-Evidence (Definition 2.2) and (η, γ)-Similarity-Evidence. See Remark C.1 for further details. Notably, (Carlini et al., 2023) reports that training a diffusion model on CIFAR-10 with $\varepsilon \geq 50$ can already diverge. This aligns with our findings, where diffusion models trained with $(\varepsilon = 100, \delta = 10^{-5})$ have relatively poor generation quality. We emphasize that it is possible to develop more sophisticated algorithms that are tailored for our criterion, and we employ off-the-shelf DP algorithms as a naive solution to merely show that our criterion can be satisfied.

Additionally, it is worth noting that while CP-Δ (Algorithm 1) provides a k-NAF guarantee, the sampling mechanism does not have explicit control of the resulting value k. (Vyas et al., 2023) also proposes a rejection-sampling-based approach, CP-k, which allows a generative model to satisfy k-NAF for any k. See Appendix I and Algorithm 2 for further details. At a high level, the CP-k algorithm takes three models: a draft model p, trained on the full dataset, and q_1, q_2 trained on the datasets $\mathcal{D}_1, \mathcal{D}_2$ respectively, as in CP-Δ. Then, for each generated output $y \sim p$ with prompt z, the CP-k algorithm will release sample y only if the maximum log-likelihood ratio $\max_{i=\{1,2\}} \log(p(y|z)/q_i(y|z)) \leq k$. When α_k denotes the corresponding one-shot acceptance probability, the CP-k algorithm achieves a $(k + \log(1/\alpha_k))$-NAF guarantee (Vyas et al., 2023). In our experiments, we study different acceptance probabilities $\alpha_k \in [0, 1]$ and examine how the CP-k algorithm affects the performance of MIA and DRA.

For the plaintiff, we describe the corresponding MIA and DRA approaches for evaluation of `Access` and `Similarity` evidence below.

Access Evaluation. We employ proximal initialization attacks (PIA) (Kong et al., 2023), the state-of-the-art MIA approach for diffusion models, to distinguish whether the defendant's model was trained with a particular copyrighted training sample. Performance is characterized by the TPR-FPR tradeoff curve, as well as TPR at low FPR, according to our criterion (α, β)-Access-Evidence (Definition 2.2). Following the literature, we also report the FID score (Heusel et al., 2017) for image generation quality (utility).

Similarity Evaluation. We adapt the data extraction attack of (Carlini et al., 2023) to our setting: the plaintiff conditionally (using prompts containing class information) samples images from the defendant's model, then selects the top 10% of the samples that are most likely in the training dataset based on MIA scores computed using PIA on the reconstructed images. Note that this attack only considers the simplest expression-preserving operations $\mathcal{F} = \{\text{identity}\}$, yet it already serves as a

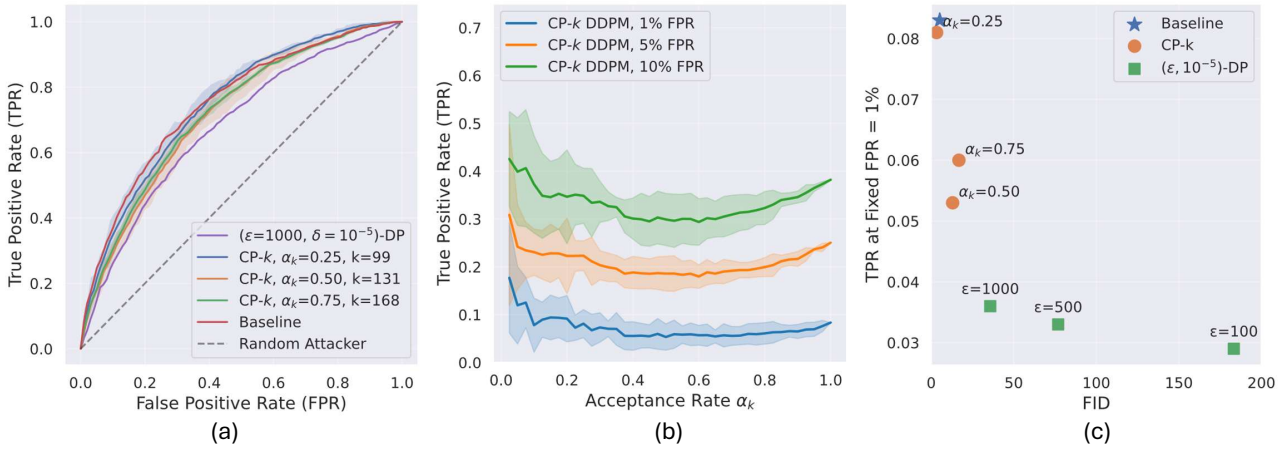


Figure 3. The performance against MIA. (a) TPR-FPR tradeoff curve. The shaded interval indicates 2 standard deviations computed over 5 independent rejection sampling trials for CP-k. (b) TPR at low FPR performance in varying α_k. (c) TPR@FPR=1% versus image quality measured in FID.

lower-bound of the actual similarity probability γ in (η, γ)-Access-Evidence for a general $\mathcal{F}$. We also emphasize that there are stronger DRAs for the plaintiff, e.g., training reconstruction networks (Hayes et al., 2023).

We characterize performance by computing the fraction of reconstructions that have $d(y, C) \leq \eta$ for a fixed reconstruction threshold η, which is an empirical estimate of $\mathbb{P}(d(y, C) \leq \eta)$, thereby aligning with our proposed notion of (η, γ)-Similarity-Evidence (Definition 2.4). Throughout the DRA experiment, we use the normalized ℓ_2 distance $d(y, C) = \ell_2(y, C) = \|y - C\|_2/\sqrt{d}$, where d denotes the dimension of the generated data.

We provide full experimental details in Appendix I.

E.2. (α, β)-Access-Evidence: The Membership Inference Game

We first study the TPR-FPR tradeoff for all tested methods (Figure 3 (a)). Compared to baselines, the CP-k mechanism slightly worsens the MIA (plaintiff) performance in general. However, we find that having a smaller acceptance rate (or equivalently, smaller k) in the CP-k algorithm does not monotonically lead to worse MIA performance (Figure 3 (b,c)). This trend indicates that the CP-k algorithm (and k-NAF criterion) is inappropriate for meaningful `Access` evidence. On the other hand, DP-Adam provides more consistent behavior in controlling the performance of MIA via the injected noise scale. Nevertheless, it significantly degrades the image generation quality as well (Figure 3(c)). As discussed in Remark C.1, using DP-Adam for (α, β)-Access-Evidence may be suboptimal. We envision that an access-evidence-tailored approach can achieve a much better evidence-utility tradeoff.

E.3. (η, γ)-Similarity-Evidence: The Data Reconstruction Game

We now examine how obtaining a stronger k-NAF guarantee via CP-k affects the DRA performance in Figure 4. Surprisingly, we found that rejecting more generated images via CP-k (smaller α_k) algorithm actually makes the overall system *more vulnerable* to DRA. This implies that the plaintiff has a higher probability of extracting images from the defendant's model that are similar to the copyrighted image. We further investigate this by checking the distance distribution of all generated images and the rejected images (Figure 4 (b)). We find that the CP-k algorithm actually rejects images with large ℓ_2 distance from copyrighted samples, which makes the final output images more likely to be close to the copyrighted images. Indeed, since the NAF criterion is agnostic to the underlying metric space of data, the CP-k model is not guaranteed to reject samples close to copyrighted data. This again indicates that the NAF criterion is inadequate to provide evidence against `Similarity`, and supports our claim that (η, γ)-Similarity-Evidence, which incorporates the metric of sample space,

Figure 5. Perceptual comparisons of reconstruction (left) and copyrighted image (right) with different distances within the CIFAR-10 "car" class.

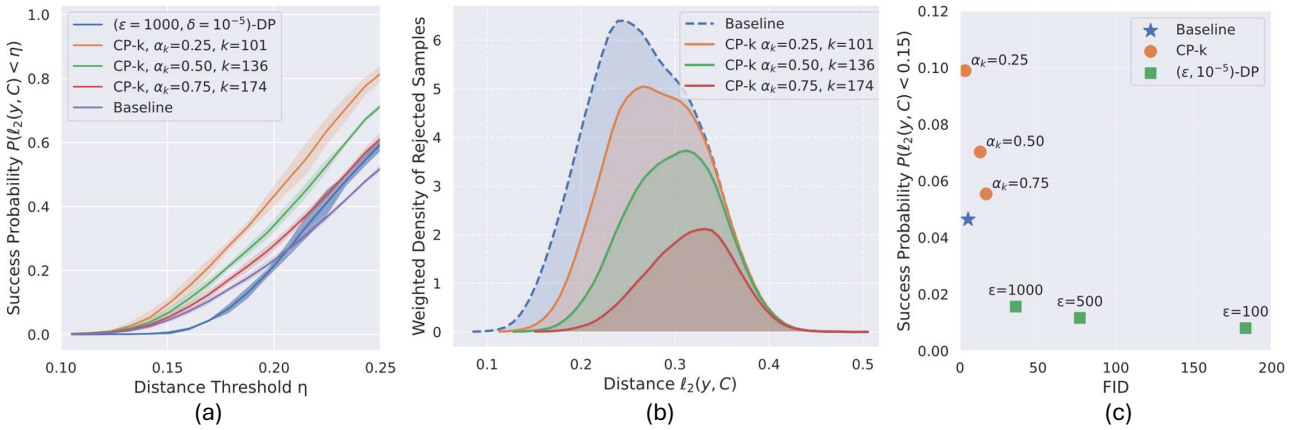


Figure 4. The performance against DRA. (a) Probability of success reconstruction based on different distance threshold η. (b) Distribution of rejected samples in CP-k with respect to their ℓ_2 distance to copyright data C. The density is weighted by the acceptance rate α_k. The baseline distribution indicates the density with respect to all generated images (i.e., $\alpha_k = 1$, no rejection). (c) Success probability versus image quality measured in FID.

is a more reasonable criterion.

On the other hand, DP-Adam can degrade the DRA success probability when the distance threshold $\eta \leq 0.2$, corresponding to the plaintiff being less favored in the data reconstruction game. From Figure 5, having a distance greater than 0.25 corresponds to a noticeable difference between the reconstruction and a copyrighted image. This is also why we focus our study on the threshold $\eta \leq 0.25$. Finally, similar to the MIA experiment, DP training deteriorates the image generation quality. Developing a similarity-evidence-tailored procedure with a better evidence-utility tradeoff is an important future direction.

F. Proof of Theorem D.2

Theorem (0-NAF Does Not Provide Evidence for `Access`)**.** *There exists a dataset $\mathcal{D}$, generative model p and a membership inference attacker* `Attack` *such that p is 0-NAF with divergence choice Δ_{KL} or Δ_{max} but* `Attack` *achieves FPR α and TPR β with any $\alpha, \beta \in (0, 1)$.*

Let us first introduce a helpful technical lemma and the NAF guarantee of the CP-Δ algorithm.

Lemma F.1. *Let p, q be any probability distribution over the domain $\mathcal{X}$. Consider the Bhattacharya distance with $S = Supp(p) \cup Supp(q)$, defined by*

$$d_B(p, q) = -\log\left(\int_S \sqrt{p(t)q(t)}\,\mathrm{d}t\right).$$

Then

$$\Delta_{KL}(p\|q) \geq 2d_B(p, q).$$

Our proof of the following theorem makes use of F.1.

Theorem F.2 (NAF guarantee of CP-Δ)**.** *Let p be the model returned by CP-Δ, and q_1, q_2 be model trained on dataset partitions $\mathcal{D}_1, \mathcal{D}_2$ with learning algorithm $\mathcal{A}$ respectively. Then p is k_z-NAF with respect to $\mathcal{C}$,* SHARDED-SAFE*, and Δ, where*

$$k_x \leq \begin{cases} -\log(1 - TV(q_1(\cdot|z), q_2(\cdot|z))) \text{ if } \Delta = \Delta_{max} \\ -2\log\big(1 - H^2(q_1(\cdot|z), q_2(\cdot|z))\big) \text{ if } \Delta = \Delta_{KL}. \end{cases} \tag{2}$$

Here, z is any fixed prompt, and $TV(\cdot,\cdot)$ and $H(\cdot,\cdot)$ are total variation and Hellinger distance, respectively.

With this result, we are ready to state our proof. We will start by proving the case of 0-NAF with the divergence Δ_{KL}, and then finish the proof with the divergence $\Delta_{\max}$.

F.1. The case of $\Delta_{\mathbf{KL}}$

Proof. We will prove this by construction. Our key idea is to show that there exists a dataset $\mathcal{D}$ with partition $\mathcal{D}_1, \mathcal{D}_2$ and learning algorithm $\mathcal{A}$, such that the CP-Δ algorithm (Algorithm 1) returns a generative model p with a 0-NAF guarantee, yet remains arbitrarily vulnerable to membership inference attacks. We claim that the main source of this vulnerability is the notion that a model's satisfaction of a 0-NAF guarantee is entirely independent of the adjacent dataset $\mathcal{D}' = \mathcal{D} \setminus \{C\}$ for copyrighted data $C \in \mathcal{D}$). Consequently, it is impossible for the 0-NAF guarantee to provide meaningful protection on the indistinguishability of cases between accessing $\mathcal{D}$ or $\mathcal{D}'$.

Consider a dataset $\mathcal{D} = \{x_1, \dots x_{n-1}, x_n\}$ with copyrighted data x_n. Set $\mathcal{D}' = \mathcal{D} \setminus \{x_n\}$. Next, we partition both datasets evenly.

$$\mathcal{D}_1 = \{x_1 \dots, x_{\lceil n/2 \rceil}\}, \quad \mathcal{D}_2 = \{x_{\lceil n/2 \rceil + 1}, \dots, x_n\}, \quad \mathcal{D}_1' = \mathcal{D}_1, \quad \mathcal{D}_2' = \mathcal{D}_2 \setminus \{x_n\}. \tag{3}$$

Using the CP-Δ algorithm, we train the generative models on each of these respective partitions. Let us denote $q_i = \mathcal{A}(\mathcal{D}_i)$ and $q_i' = \mathcal{A}(\mathcal{D}_i')$ for $i = 1, 2$. It is not hard to see that $q_1 = q_1'$ in this scenario. In the following, we denote p, p' as the output of CP-Δ algorithm with dataset $\mathcal{D}$ and $\mathcal{D}'$ respectively.

We first prove that in this case, for any prompt z, we have

$$\Delta_{\mathrm{KL}}(p(\cdot|z) \| p'(\cdot|z)) \geq -2 \log \max_y q_1(y|z) + 2 d_B(q_2(\cdot|z), q_2'(\cdot|z)) \tag{4}$$

$$- d_B(q_1(\cdot|z), q_2(\cdot|z)) - d_B(q_1(\cdot|z), q_2'(\cdot|z)). \tag{5}$$

Our goal here is to establish a lower bound on $\Delta_{\mathrm{KL}}(p(\cdot|z) \| p'(\cdot|z))$, which can later be translated to TPR and FPR in the membership inference attack game. This is based on the close relation of KL divergence and hypothesis testing, which we will further elaborate on later.

By Lemma F.1, we establish

$$\Delta_{\mathrm{KL}}(p(\cdot|z) \| p'(\cdot|z)) \geq 2 d_B((p(\cdot|z), p'(\cdot|z)). \tag{6}$$

Expanding the right side gives

$$\begin{aligned}
2 d_B((p(\cdot|z), p'(\cdot|z)) &= -2 \log \left(\int \sqrt{p(y|z) p'(y|z)} \mathrm{d}y \right) \\
&= -2 \log \left(\int q_1(y|z) \sqrt{\frac{q_2(y|z) q_2'(y|z)}{N(z) N'(z)}} \mathrm{d}y \right) \\
&\geq -2 \log \left(\frac{\max_y q_1(y|z)}{\sqrt{N(z) N'(z)}} \int \sqrt{q_2(y|z) q_2'(y|z)} \mathrm{d}y \right) \\
&= -2 \log \max_y q_1(y|z) + \log N(z) + \log N'(z) - 2 \log \int \sqrt{q_2(y|z) q_2'(y|z)} dy \\
&= -2 \log \max_y q_1(y|z) + \log N(z) + \log N'(z) + 2 d_B(q_2(\cdot|z), q_2'(\cdot|z)).
\end{aligned}$$

For $N(z)$, we have

$$\begin{aligned}
N(z) &= 1 - \mathtt{H}^2(q_1(\cdot|z), q_2(\cdot|z)) \\
&= 1 - \left(1 - \int \sqrt{q_1(y|z) q_2(y|z)} \mathrm{d}y \right) \\
&= \int \sqrt{q_1(y|z) q_2(y|z)} \mathrm{d}y.
\end{aligned}$$

Using the same logic for $N'(z)$, we have

$$\log N(z) = -d_B(q_1(\cdot|z), q_2(\cdot|z)), \quad \log N'(z) = -d_B(q_1(\cdot|z), q_2'(\cdot|z)).$$

Substituting back into our inequality, we find that

$$\Delta_{\mathrm{KL}}(p(\cdot|z)\|p'(\cdot|z)) \geq -2\log\max_y q_1(y|z) + 2d_B(q_2(\cdot|z), q_2'(\cdot|z)) \\ - d_B(q_1(\cdot|z), q_2(\cdot|z)) - d_B(q_1(\cdot|z), q_2'(\cdot|z)),$$

which is exactly equation (4).

Now consider the case $n = 4$, where $\mathcal{D} = \{0, 0, -C, C\}$ and C is our copyrighted content. Let us choose the density estimation algorithm $\mathcal{A}(\mathcal{D}) = N(\mu(\mathcal{D}), \frac{1}{2\pi})$ as our generative model and learning algorithm, where $\mu(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} x$. We can view this as learning a Gaussian mixture model from data with one center, which will converge to the maximum-likelihood estimate of the mean from data (Bishop, 2006). In this scenario, we have

$$q_1 = \mathcal{N}\left(0, \frac{1}{2\pi}\right), \quad q_2 = \mathcal{N}\left(0, \frac{1}{2\pi}\right), \quad q_2' = \mathcal{N}\left(-C, \frac{1}{2\pi}\right). \tag{7}$$

This implies that

$$d_B(q_1(\cdot|z), q_2(\cdot|z)) = 0, \quad d_B(q_1(\cdot|z), q_2'(\cdot|z)) = \frac{C^2}{8}, \quad d_B(q_2(\cdot|z), q_2'(\cdot|z)) = \frac{C^2}{8}. \tag{8}$$

By substitute these quantities back to equation (4), we have

$$\Delta_{\mathrm{KL}}(p(\cdot|z)\|p'(\cdot|z)) \geq -2\log\max_y q_1(y|z) + \frac{C^2}{8} \geq \frac{C^2}{8}, \tag{9}$$

where the last inequality is due to the fact that $q_1(\cdot|z) = \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \frac{1}{2\pi}$ where the maximum happens at $q_1(0|z) = \frac{1}{\sqrt{2\pi\sigma^2}} = 1$. Clearly, choosing C arbitrarily large leads to an arbitrarily large KL divergence. It is also important to notice that the same bound holds for $\Delta_{\mathrm{KL}}(p'(\cdot|z)\|p(\cdot|z))$ by symmetry. As a result we have

$$\min\left(\Delta_{\mathrm{KL}}(p(\cdot|z)\|p'(\cdot|z)), \Delta_{\mathrm{KL}}(p'(\cdot|z)\|p(\cdot|z))\right) \geq \frac{C^2}{8}. \tag{10}$$

Finally, note that by the Neyman-Pearson Lemma (Neyman & Pearson, 1933), the most powerful test for hypothesis testing, or equivalently, the strongest attacker in the MIA game, is the likelihood ratio test. Let us consider the attacker `Attack`: declaring the generative model p has access to the copyrighted content C if $\log\left(\frac{p(y|z)}{p'(y|z)}\right) \geq 0$ and no access otherwise. In this case, we show that `Attack` can achieve arbitrarily high TPR with arbitrarily low FPR given C is sufficiently large. Namely, for any $\alpha, \beta \in (0, 1)$, we can always find a sufficiently large C such that

$$\mathbb{P}_{y\sim p(\cdot|z)}\left(\log\left(\frac{p(y|z)}{p'(y|z)}\right) \geq 0\right) \geq \beta, \quad \mathbb{P}_{y\sim p'(\cdot|z)}\left(\log\left(\frac{p(y|z)}{p'(y|z)}\right) \geq 0\right) \leq \alpha. \tag{11}$$

The proof is quite standard in the hypothesis testing literature; see, for example, the lecture note from Robert Nowak[3]. We include all details here for the self-contained purpose. Firstly, consider the random variable $\Lambda = \log\left(\frac{p(Y|z)}{p'(Y|z)}\right)$. We first show that $Y \sim p(\cdot|z)$ or $Y \sim p'(\cdot|z)$. That is,

$$\mathbb{P}\left(\Lambda - \mathbb{E}\Lambda \geq \epsilon\right) \leq a\exp\left(\frac{-bt^2}{2}\right), \tag{12}$$

for some constant $a, b \geq 0$.

Recall that under our setting, we have

$$p(y|z) \propto \exp\left(-\frac{y^2}{2\sigma^2}\right), \; p'(y|z) \propto \exp\left(-\frac{y^2 + (y-C)^2}{4\sigma^2}\right), \tag{13}$$

[3] https://nowak.ece.wisc.edu/ece830/ece830_spring15_lecture7.pdf

where $\sigma^2 = 2\pi$ by our choice. Then, by some manipulation, we have

$$\Lambda(y) = \frac{-y^2 + (y-C)^2}{4\sigma^2} + c_1 = \frac{-2Cy + C^2}{4\sigma^2} + c_1, \tag{14}$$

for some constant c_1 that is independent of y. Clearly, $\Lambda(Y)$ is nothing but a linear transformation of Y. As a result, if Y is subgaussian, then $\Lambda(Y)$ is also subgaussian. The case $Y \sim p(\cdot|x) = N(0, \sigma^2)$ is straightforward. It is also not hard to show that $Y \sim p'(\cdot|z)$ is subgaussian as well since

$$p'(y|z) \propto \exp\left(-\frac{y^2 + (y-C)^2}{4\sigma^2}\right) = \exp\left(-\frac{(y - C/2)^2}{2\sigma^2}\right)\exp\left(-\frac{C^2}{8\sigma^2}\right). \tag{15}$$

As a result, we have the following tail bound due to the lecture notes of Robert Nowak.

$$\mathbb{P}_{Y\sim p}\left(\mathbb{E}_{Y\sim p}\Lambda(Y) - \Lambda(Y) \geq \epsilon\right) \leq \exp\left(-c_2\epsilon^2\right) \tag{16}$$

$$\mathbb{P}_{Y\sim p'}\left(\Lambda(Y) - \mathbb{E}_{Y\sim p'}\Lambda(Y) \geq \epsilon\right) \leq \exp\left(-c_3\epsilon^2\right), \tag{17}$$

where $c_2, c_3 \geq 0$ are some constant depending on the subgaussian property of p, p' respectively.

Now let us first analyze the FPR.

$$\text{FPR} = \mathbb{P}_{Y\sim p'}\left(\Lambda(Y) \geq 0\right) = \mathbb{P}_{Y\sim p'}\left(\Lambda(Y) - \mathbb{E}_{Y\sim p'}\Lambda(Y) \geq -\mathbb{E}_{Y\sim p'}\Lambda(Y)\right). \tag{18}$$

Note that

$$\mathbb{E}_{Y\sim p'}\Lambda(Y) = \int p'(y|z)\log\left(\frac{p(y|z)}{p'(y|z)}\right)\mathrm{d}y = -\Delta_{\text{KL}}(p'(\cdot|z)||p(\cdot|z)). \tag{19}$$

Let us choose $\epsilon = \Delta_{\text{KL}}(p'(\cdot|z)||p(\cdot|z)) \geq 0$ (non-negativity of KL divergence) and apply the tail bound in (16), which leads to

$$\begin{aligned}
\text{FPR} = \mathbb{P}_{Y\sim p'}\left(\Lambda(Y) \geq 0\right) &= \mathbb{P}_{Y\sim p'}\left(\Lambda(Y) - \mathbb{E}_{Y\sim p'}\Lambda(Y) \geq -\mathbb{E}_{Y\sim p'}\Lambda(Y)\right) && (20)\\
&\leq \exp\left(-c_3\Delta_{\text{KL}}(p'(\cdot|z)||p(\cdot|z))^2\right) && (21)\\
&\leq \exp\left(-\frac{c_3}{8}C^2\right), && (22)
\end{aligned}$$

where the last inequality is due to equation (10). Finally, we are left to prove the bound for TPR. Following a similar argument we have

$$\begin{aligned}
\text{TPR} = \mathbb{P}_{Y\sim p}\left(\Lambda(Y) \geq 0\right) &= \mathbb{P}_{Y\sim p}\left(\Lambda(Y) - \mathbb{E}_{Y\sim p}\Lambda(Y) \geq -\mathbb{E}_{Y\sim p}\Lambda(Y)\right) && (23)\\
&= 1 - \mathbb{P}_{Y\sim p}\left(\Lambda(Y) - \mathbb{E}_{Y\sim p}\Lambda(Y) \leq -\mathbb{E}_{Y\sim p}\Lambda(Y)\right) && (24)\\
&= 1 - \mathbb{P}_{Y\sim p}\left(\mathbb{E}_{Y\sim p}\Lambda(Y) - \Lambda(Y) \geq \mathbb{E}_{Y\sim p}\Lambda(Y)\right). && (25)
\end{aligned}$$

Note that

$$\mathbb{E}_{Y\sim p}\Lambda(Y) = \int p(y|z)\log\left(\frac{p(y|z)}{p'(y|z)}\right)\mathrm{d}y = \Delta_{\text{KL}}(p(\cdot|z)||p'(\cdot|z)). \tag{26}$$

Let us choose $\epsilon = \Delta_{\text{KL}}(p(\cdot|x)||p'(\cdot|x))$ and apply the tail bound in (16), which leads to

$$\begin{aligned}
&\mathbb{P}_{Y\sim p}\left(\mathbb{E}_{Y\sim p}\Lambda(Y) - \Lambda(Y) \geq \mathbb{E}_{Y\sim p}\Lambda(Y)\right) && (27)\\
&\leq \exp\left(-c_2\Delta_{\text{KL}}(p(\cdot|z)||p'(\cdot|z))^2\right) \leq \exp\left(-\frac{c_2}{8}C^2\right), && (28)
\end{aligned}$$

where the last inequality is due to equation (10). Apparently, we can always choose C large enough so that both types of errors are arbitrarily close to zero. Together we complete the proof for the case of Δ_{KL}. □

F.2. The case of $\Delta_{\max}$

Proof. We again work with the same example $\mathcal{D} = \{0, 0, -C, C\}$ and C is our copyrighted content. We also choose the density estimation algorithm $\mathcal{A}(\mathcal{D}) = N(\mu(\mathcal{D}), 1)$ as our generative model and learning algorithm similar to the case of Δ_{KL}. Note that we choose the variance $\sigma^2 = 1$ for simplicity and our proof holds for the choice $\sigma^2 = \frac{1}{2\pi}$ as well. Then we have

$$q_1 = \mathcal{N}(0, 1), \quad q_2 = \mathcal{N}(0, 1), \quad q_2' = \mathcal{N}(-C, 1). \tag{29}$$

The major difference between the case of Δ_{KL} is that the CP-Δ algorithm led to different output distributions. In the case of $\Delta_{\max}$, according to Algorithm 1 we have

$$p(y|z) = \frac{\min(q_1(y|z), q_2(y|z))}{N(z)}, \text{ where } N(z) = \int_{-\infty}^{\infty} \min(q_1(y|z), q_2(y|z))\mathrm{d}y. \tag{30}$$

Similarly, we have $p'(y|z) = \frac{\min(q_1(y|z), q_2'(y|z))}{N'(z)}$ on the adjacent dataset $\mathcal{D}' = \{0, 0, -C\}$. Our goal is again to establish a divergent lower bound of $\Delta_{\mathrm{KL}}(p(\cdot|z)||p'(\cdot|z))$ (and the other direction), which lead to the TPR and FPR bound for the underlying hypothesis problem as before.

First, let us denote the Q-function, which is the tail CDF of the standard normal

$$Q(t) := \int_t^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \mathrm{d}y. \tag{31}$$

In the meantime, note that

$$q_1(y|z) \geq q_2'(y|z) \; \forall \, y \in \left(-\infty, -\frac{C}{2}\right], \; q_1(y|z) \leq q_2'(y|z) \; \forall \, y \in \left(-\frac{C}{2}, \infty\right). \tag{32}$$

As a result, we first derive the normalizing constant $N'(z)$ via standard manipulation and the definition of Q-function.

$$N'(z) = \int_{-\infty}^{\infty} \min(q_1(y|z), q_2'(y|z))\mathrm{d}y = \int_{-\infty}^{-\frac{C}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \mathrm{d}y + \int_{-\frac{C}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y+C)^2}{2}} \mathrm{d}y \tag{33}$$

$$\overset{(a)}{=} Q\left(\frac{C}{2}\right) + Q\left(\frac{C}{2}\right) = 2Q\left(\frac{C}{2}\right). \tag{34}$$

The equality (a) is by the fact that standard normal is symmetric around its mean. Namely, for the first term we have $\int_{-\infty}^{-\frac{C}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \mathrm{d}y = \int_{\frac{C}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \mathrm{d}y = Q(\frac{C}{2})$. The second term is directly by the definition of Q-function $\int_{-\frac{C}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y+C)^2}{2}} dy = Q(\frac{-C}{2} + C) = Q(\frac{C}{2})$. In the meanwhile, note that $p = N(0, 1)$ due to the fact that $q_1 = q_2 = N(0, 1)$ and thus $N(z) = 1$.

Now we can directly lower bound the KL divergence between p, p'. We proceed with the direction $\Delta_{\mathrm{KL}}(p(\cdot|z)||p'(\cdot|z))$ first.

$$\Delta_{\mathrm{KL}}(p(\cdot|z)||p'(\cdot|z)) = \int p(y|z) \log\left(\frac{p(y|z)}{p'(y|z)}\right)\mathrm{d}y \tag{35}$$

$$= \int_{-\infty}^{\frac{-C}{2}} q_1(y|x) \log\left(\frac{q_1(y|z)}{q_1(y|z)}\right)\mathrm{d}y + \int_{\frac{-C}{2}}^{\infty} q_1(y|z) \log\left(\frac{q_1(y|z)}{q_2'(y|z)}\right)\mathrm{d}y + \log(N') \tag{36}$$

$$= \int_{\frac{-C}{2}}^{\infty} q_1(y|z) \log\left(\frac{q_1(y|z)}{q_2'(y|z)}\right)\mathrm{d}y + \log(N') \tag{37}$$

$$= \int_{\frac{-C}{2}}^{\infty} q_1(y|z)(Cy + \frac{C^2}{2})dy + \log(N'). \tag{38}$$

$$= \int_{\frac{-C}{2}}^{\infty} q_1(y|z)(Cy)\mathrm{d}y + \frac{C^2}{2}(1 - Q(\frac{C}{2})) + \log(N'). \tag{39}$$

For the first term, note that

$$\int_{\frac{-C}{2}}^{\infty} q_1(y|z)Cydy = \int_{\frac{-C}{2}}^{\frac{C}{2}} q_1(y|z)Cydy + \int_{\frac{C}{2}}^{\infty} q_1(y|z)Cydy \geq \frac{C^2}{2}Q(\frac{C}{2}), \tag{40}$$

where the inequality holds for all $C \geq 0$.

Together we have for any $C \geq 0$,

$$\Delta_{\text{KL}}(p(\cdot|x)||p'(\cdot|z)) \geq \frac{C^2}{2} + \log(N'(z)) = \frac{C^2}{2} + \log\left(2Q(\frac{C}{2})\right). \tag{41}$$

Finally, we may further lower bound Q function via standard results in the literature (Borjesson & Sundberg, 1979) to obtain an explicit formula with respect to C.

$$Q(t) \geq \frac{t}{(1+t^2)\sqrt{2\pi}}e^{-\frac{t^2}{2}}, \ \forall z > 0. \tag{42}$$

As a result, denoting $t = \frac{C}{2}$, we have

$$\log(2Q(t)) \geq \log\left(\frac{2t}{(1+t^2)\sqrt{2\pi}}e^{-\frac{t^2}{2}}\right) = \log\left(\frac{2t}{(1+t^2)\sqrt{2\pi}}\right) - \frac{t^2}{2}. \tag{43}$$

Substituting $t = \frac{C}{2}$ once again gives

$$\log\left(2Q(\frac{C}{2})\right) \geq \log\left(\frac{4C}{(4+C^2)\sqrt{2\pi}}\right) - \frac{C^2}{8}. \tag{44}$$

Altogether, we have

$$\Delta_{\text{KL}}(p(\cdot|z)||p'(\cdot|z)) \geq \frac{C^2}{2} + \log\left(2Q(\frac{C}{2})\right) \geq \frac{3C^2}{8} + \log\left(\frac{4C}{(4+C^2)\sqrt{2\pi}}\right). \tag{45}$$

Clearly, this lower bound goes to ∞ as $C \to \infty$, which leads to an arbitrarily high TPR rate in the hypothesis testing. Finally, we prove a similar result for the other direction in a slightly more complicated manner.

$$\Delta_{\text{KL}}(p'(\cdot|z)||p(\cdot|z)) = \int p'(y|z)\log\left(\frac{p'(y|z)}{p(y|z)}\right)dy \tag{46}$$

$$= \frac{1}{N'}\int_{-\infty}^{\frac{-C}{2}} q_1(y|z)\log\left(\frac{q_1(y|z)}{q_1(y|z)}\right)dy + \frac{1}{N'}\int_{\frac{-C}{2}}^{\infty} q_2'(y|z)\log\left(\frac{q_2'(y|z)}{q_1(y|z)}\right)dy - \frac{\log(N')}{N'} \tag{47}$$

$$= \frac{1}{N'}\int_{\frac{-C}{2}}^{\infty} q_2'(y|z)\log\left(\frac{q_2'(y|z)}{q_1(y|z)}\right)dy - \frac{\log(N')}{N'} \tag{48}$$

$$= \frac{1}{N'}\int_{\frac{-C}{2}}^{\infty} q_2'(y|z)(-Cy - \frac{C^2}{2})dy - \frac{\log(N')}{N'}. \tag{49}$$

For the first term, we have

$$\int_{\frac{-C}{2}}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{(y+C)^2}{2}}(-Cy - \frac{C^2}{2})dy \overset{(a)}{=} \int_{\frac{C}{2}}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}(-Ct + \frac{t^2}{2})dz \tag{50}$$

$$= \int_{\frac{C}{2}}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}(-Ct)dz + \frac{t^2}{2}(1 - Q(\frac{C}{2})). \tag{51}$$

where in (a) we apply the change of variable $t = y + C$. For the first term above, we unfortunately cannot lower bound it as before due to the negative factor. Nevertheless, observe that

$$\int_{\frac{C}{2}}^{\infty} \frac{t}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \leq \int_{0}^{\infty} \frac{t}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dz = \frac{1}{2}\int_{-\infty}^{\infty} \frac{|t|}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \overset{(b)}{=} \frac{1}{\sqrt{2\pi}}, \tag{52}$$

where (b) is nothing but the expectation of half-normal. As a result, we have

$$\Delta_{\mathrm{KL}}(p'(\cdot|x)||p(\cdot|z)) \geq \frac{1}{N'}\left(\frac{-C}{\sqrt{2\pi}} + \frac{t^2}{2}(1-Q(\frac{C}{2})) - \log(N')\right) \geq \frac{1}{N'}\left(\frac{-C}{\sqrt{2\pi}} - \log(N')\right). \tag{53}$$

In the meanwhile, recall that $N' = 2Q(\frac{C}{2})$ and that we always have $N' \leq 1$. Next, we turn to analyze $\log(N')$. By leveraging the upper bound of the Q function in the literature (Borjesson & Sundberg, 1979):

$$Q(t) \leq \frac{1}{t\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \ \forall t > 0. \tag{54}$$

we have for $C > 2$,

$$-\log(N') = -\log\left(2Q(\frac{C}{2})\right) \geq -\log\left(\frac{2}{\sqrt{2\pi}} e^{-\frac{C^2}{8}}\right) = -\log\left(\frac{2}{\sqrt{2\pi}}\right) + \frac{C^2}{8}. \tag{55}$$

As a result, for $C > 2$ we the following lower bound always hold

$$\Delta_{\mathrm{KL}}(p'(\cdot|z)||p(\cdot|z)) \geq \frac{1}{N'}\left(\frac{-C}{\sqrt{2\pi}} + \frac{t^2}{2}(1-Q(\frac{C}{2})) - \log(N')\right) \tag{56}$$

$$\geq \frac{1}{N'}\left(\frac{-C}{\sqrt{2\pi}} - \log\left(\frac{2}{\sqrt{2\pi}}\right) + \frac{C^2}{8}\right) \tag{57}$$

$$\overset{(a)}{\geq} \frac{-C}{\sqrt{2\pi}} - \log\left(\frac{2}{\sqrt{2\pi}}\right) + \frac{C^2}{8}, \tag{58}$$

where (a) is due to the fact that we always have $N' \leq 1$. Evidently, this lower bound goes to ∞ as $C \to \infty$, which leads to arbitrarily low FPR rate in the hypothesis testing. Together, we complete the proof. □

G. Proof of Theorem D.3

Theorem (0-NAF Does Not Provide Evidence for `Similarity`)**.** *There exists a dataset $\mathcal{D}$, a generative model p such that p is 0-NAF with divergence choice Δ_{KL} or Δ_{max} but still generate substantially similar output to the copyrighted content $C \in \mathcal{D}$ with arbitrarily high probability. Specifically, for any $C > 0$ and $\gamma \in (0, 1]$, there exists an instance $(\mathcal{D}, C, p)$ satisfying a 0-NAF guarantee such that*

$$\mathbb{P}_{y\sim p}(\|y - C\| \leq C) \geq 1 - \gamma. \tag{59}$$

Proof. We again analyze a similar setting as in the proof of D.2. Consider the dataset $\mathcal{D} = \{C_1, C_2, C\}$ where $\frac{C_1+C_2}{2} = C$. Then we obtain the partitioned dataset as before: $\mathcal{D}_1 = \{C_1, C_2\}$ and $\mathcal{D}_2 = \{C\}$. Then we run the CP-Δ algorithm with the choice of $\mathcal{A}(\mathcal{D}) = N(\mu(\mathcal{D}), \sigma^2)$ similar as the case of `Access`, where we have

$$q_1 = N(\mu(\mathcal{D}_1), \sigma^2) = N(C, \sigma^2), \ q_2 = N(\mu(\mathcal{D}_2), \sigma^2) = N(C, \sigma^2). \tag{60}$$

Since $q_1 = q_2$, for both divergence choices $\Delta_{\mathrm{KL}}, \Delta_{\max}$ the final generative mode p from the CP-Δ algorithm is 0-NAF and $p = N(C, \sigma^2)$. Apparently, the probability of the generative model p generating the copyrighted content C is arbitrarily high as $\sigma \to 0$. More specifically,

$$\mathbb{P}_{y\sim p}(\|y - C\| \leq C) = 1 - 2Q\left(\frac{C}{\sigma}\right), \tag{61}$$

where Q is the Q-function, the tail probability of the standard Gaussian distribution. By setting $\gamma = 2Q(\frac{C}{\sigma})$, it is apparent that for any $C > 0$ and $\gamma \in (0, 1]$, we can always find a sufficiently small σ to satisfy the inequality stated in the theorem. Together, we complete the proof. □

H. Proof of Lemma F.1

The proof of Lemma F.1 is relatively straightforward. We can write $d_B(p,q)$ as

$$d_B(p,q) = -\log\left(\mathbb{E}_{y\sim p}\left[\left(\frac{q(y)}{p(y)}\right)^{1/2}\right]\right).$$

Consider the convex function $\phi(t) = -\log t$. Applying Jensen's inequality, we obtain:

$$-\log\left(\mathbb{E}_{y\sim p}\left[\left(\frac{q(y)}{p(y)}\right)^{1/2}\right]\right) \le \mathbb{E}_{y\sim p}\left[-\log\left(\left(\frac{q(y)}{p(y)}\right)^{1/2}\right)\right] \tag{62}$$

$$= \mathbb{E}_{y\sim p}\left[\frac{1}{2}\log\frac{p(y)}{q(y)}\right] \tag{63}$$

$$= \frac{1}{2}\,\mathbb{E}_{y\sim p}\left[\log\frac{p(y)}{q(y)}\right] \tag{64}$$

$$= \frac{1}{2}\,d_{\mathrm{KL}}(p,q). \tag{65}$$

Equivalently, we have

$$d_{KL}(p,q) \ge 2\,d_B(p,q).$$

I. Experiment Settings

We provide supplementary details regarding the models, datasets, implementation specifics, and evaluation protocols used in the empirical evaluation presented in Section E.

I.1. Models and Samplers

In this section, we review the denoising diffusion probabilistic model (Ho et al., 2020), detail the CP-k rejection-sampling procedure for samplers satisfying the NAF criterion (Vyas et al., 2023), and briefly outline differentially private training schemes (DP-Adam) used as comparisons in Section E.

I.1.1. Conditional Denoising Diffusion Probabilistic Models

Conditional DDPMs extend the unconditional diffusion framework (Ho et al., 2020) by incorporating class label information y into the reverse process for image generation. Given a data distribution p, a data sample $y_0 \sim p$ and condition (prompt) z, the *forward* (noising) process defines a Markov chain

$$q(y_t \mid y_{t-1}) = \mathcal{N}\big(y_t; \sqrt{1-\beta_t}\,y_{t-1},\ \beta_t\,I\big), \quad t = 1,\dots,T, \tag{66}$$

where $\{\beta_t\}_{t=1}^T$ is a fixed variance schedule. One can show in closed form that

$$q(y_t \mid y_0) \ = \ \mathcal{N}\big(y_t;\ \sqrt{\bar\alpha_t}\,y_0,\ (1-\bar\alpha_t)\,I\big), \quad \bar\alpha_t = \prod_{s=1}^{t}(1-\beta_s).$$

The *reverse* (denoising) process is parameterized by a neural network $\epsilon_\theta(y_t, t, z)$ which predicts the noise added at step t conditioned on z. The learned kernel is then

$$p_\theta(y_{t-1} \mid y_t, z) = \mathcal{N}\Big(y_{t-1};\ \mu_\theta(y_t,t,z),\ \beta_t\,I\Big),$$

with

$$\mu_\theta(y_t,t,z) = \frac{1}{\sqrt{1-\beta_t}}\left(y_t - \frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\,\epsilon_\theta(y_t,t,z)\right)$$

The network is trained by minimizing the simplified noise prediction objective

$$\mathcal{L}(\theta) = \mathbb{E}_{y_0,z,\epsilon,t}\Big\|\epsilon - \epsilon_\theta\big(\sqrt{\bar\alpha_t}\,y_0 + \sqrt{1-\bar\alpha_t}\,\epsilon,\ t,\ z\big)\Big\|_2^2, \tag{67}$$

Algorithm 2 CP-k sampling procedure (Vyas et al., 2023)

Input: Dataset $\mathcal{D}$, learning algorithm $\mathcal{A}$, draft model $p(\cdot|z) = \mathcal{A}(\mathcal{D})$, threshold k, sharded models q_1, q_2.
repeat
  if $\max_{i\in\{1,2\}} \log \frac{p(y|z)}{q_i(y|z)} \leq k$ **then**
    break
  end
until *Sample* $y \sim p(\cdot|z)$;
return *Sample* y

with $x_0 \sim p$, $\epsilon \sim \mathcal{N}(0, I)$, and $t \sim \text{Uniform}\{1, \ldots, T\}$. At generation time, one samples $y_T \sim \mathcal{N}(0, I)$ and iteratively applies $p_\theta(y_{t-1} \mid y_t, z)$ for $t = T, \ldots, 1$, yielding a sample y_0 consistent with the prompt z.

I.1.2. CP-k ALGORITHM

As described in Section E, we use the CP-k algorithm to provide k-NAF guarantees on our trained models. We provide a precise algorithmic description in Algorithm 2, as described in (Vyas et al., 2023), for the sake of completeness.

While CP-Δ provides a k-NAF guarantee, the sampling mechanism does not have explicit control of the resulting value k. As a result, (Vyas et al., 2023) also proposes a rejection sampler CP-k, summarized in Algorithm 2, which only requires sampling from a "draft" model p and checking a single log-ratio bound against each sharded model. With a fixed threshold k, at each iteration, one draws $y \sim p(\cdot|z)$, where z is a prompt, and accepts it if

$$\max_{i\in\{1,2\}} \log \frac{p(y|z)}{q_i(y|z)} \leq k.$$

By a result of (Vyas et al., 2023), the resulting sampler satisfies a k_z-NAF guarantee, with $k_z = k + \log(1/\alpha_k(z))$, while incurring at most $1 - \alpha_k(z)$ total-variation distance from p, where

$$\alpha_k(z) \;=\; \mathbb{P}_{y\sim p}\Big[\max_{i\in\{1,2\}} \log \tfrac{p(y|z)}{q_i(y|z)} \leq k\Big]$$

is the single-shot acceptance probability, which monotonically increases in k. This trade-off provides direct user control over the k_z-NAF guarantee. In practice, we pick a desired acceptance rate $\alpha_k(z)$ and compute a corresponding threshold k from the lower $\alpha_k(z)-$quantile of the empirical distribution of $\max_{i\in\{1,2\}} \log(p(y|z), q_i(y|z))$.

I.1.3. DP-ADAM

To train diffusion models with (ϵ, δ)-DP, we employ DP-Adam, a variant of the original DP-SGD method introduced in (Abadi et al., 2016). We note that the mechanism for maintaining privacy (i.e. adding noise and clipping gradients) remains the same between both samplers, as post-processing guarantees that the privacy loss is the same.

I.2. Attack Methods

In this section, we discuss the implementation details of the MIA and DRA attacks used in our empirical evaluation of `Access` and `Similarity` evidence.

I.2.1. MEMBERSHIP INFERENCE

As discussed in Section E, we use proximal initialization attacks (PIA) (Kong et al., 2023) to evaluate the performance of models in the `Access` game. We briefly describe this attack here for completeness.

Fix a real sample y_0. One first obtains the model's own noise estimate at $t = 0$, given by $\varepsilon_0 = \varepsilon_\theta(y_0, 0)$. Then, we estimate the noised input at any later timestep t via the deterministic forward map

$$y_t = \sqrt{\bar{\alpha}_t}\, y_0 + \sqrt{1 - \bar{\alpha}_t}\, \varepsilon_0.$$

A second query yields $\varepsilon_\theta(y_t, t)$, and the attack score is measure by the ℓ_p norm difference.

$$R_{t,p}(y_0) = \big\|\varepsilon_0 - \varepsilon_\theta(y_t, t)\big\|_p.$$

Since training samples tend to reproduce the model's proximal initialization more faithfully, smaller values of $R_{t,p}$ indicate higher likelihood of membership in the training set (Kong et al., 2023). Hence, the formal attack may be written as

$$f(y_0) = \mathbf{1}[R_{t,p} < \tau],$$

where τ is some threshold adjusted based on the desired FPR. In our experiments, we set $p = 4$ and choose t to maximize the AUROC of the attack curve for each relevant attack, since the plaintiff aims to show the most vulnerability in the defendant's model and need not fix the parameters of their attacks across defendants. In general, regardless of the choice of model, this maximizer was given by $t \approx 200$.

For baselines and DP models, we employ a direct implementation of this attack. For the CP-k sampler, we estimate the log probability ratio, $\max_{i\in\{1,2\}} \log(p(y|z), q_i(y|z))$ with p, q_i as given in Appendix I.1.2, of member and nonmember samples by running the forward DDPM noising process on y_t. When a certain sample exceeds the threshold, we simply exclude it from the attack. This mimics the realistic scenario where a defendant's model, which implements CP-k, will not release information about samples that exceed the fixed threshold k.

I.2.2. Data Reconstruction Attacks

Additionally, as described in Section E, we employ a modified DRA (Carlini et al., 2023) to evaluate the `Similarity` game.

Consider a fixed a class prompt z. We first draw n independent generations $y^i \sim p(\cdot|z)$. Next, for each y^i, we compute its MIA score $R_{t,p}(y^i)$ exactly as in the proximal initialization attack above, and sort the $\{y^i\}$ in ascending order of $R_{t,p}$. We then select the top ρn samples as candidate reconstructions, where $\rho \in (0, 1]$.

To decide whether a candidate y is a successful reconstruction, we set

$$d(y, C) = \min_{y' \in \mathcal{D}_z} d(y, y')$$

be the distance from y to its nearest neighbor in the true class-z training set $\mathcal{D}_z$. We consider y to be a successful reconstruction if $d(y, C) \le \eta$. Hence the empirical reconstruction success rate is, with $k = \lfloor \rho n \rfloor$,

$$\mathbb{P}(d(y, C) \le \eta) \approx \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}\big(d(y^i, C) \le \eta\big).$$

In our experiments we fix $\rho = 0.10$ and use the normalized distance $\ell_2(\cdot, \cdot) = d(\cdot, \cdot)$.