Evaluation Indicator for Model Inversion Attack

Hiroaki Tanaka
X-Tech Development Department, NTT DOCOMO, INC.
hiroaki.tanaka.sg@nttdocomo.com

Wataru Yamada
X-Tech Development Department, NTT DOCOMO, INC.
wataru.yamada.rz@nttdocomo.com

Keiichi Ochiai
X-Tech Development Department, NTT DOCOMO, INC.
ochiaike@nttdocomo.com

Rina Okada
Secure Platform Laboratories, NTT Corporation.

Satoshi Hasegawa
Secure Platform Laboratories, NTT Corporation.

Daizo Ikeda
Searvice Innovation Department, NTT DOCOMO, INC.

ABSTRACT

The application of machine learning to various fields such as computer vision, speech recognition, and healthcare has been tremendously successful and has gained great popularity. Although very successful, recent studies have found that machine learning models have particular vulnerabilities in terms of security and privacy. In particular, an attack called model inversion restores the training data of a machine learning model using the model output without any prior knowledge of the model. Many researchers have developed defensive methods against this attack; however, no indicator that numerically evaluates the success level of the model inversion has been developed yet. This lack of an evaluation indicator prevents us from determining numerically the most applicable method. To address this, we propose a new approach that evaluates the success of model inversion by considering the difference between an ideal situation for an attacker and the actual situation. In addition, the presented evaluation of the proposed approach shows that it is suitable for evaluating the Model Inversion Attack based on numerical simulation and real data experiments.

ACM Reference Format:
Hiroaki Tanaka, Wataru Yamada, Keiichi Ochiai, Rina Okada, Satoshi Hasegawa, and Daizo Ikeda. 2020. Evaluation Indicator for Model Inversion Attack. In *AdvML '20: Workshop on Adversarial Learning Methods for Machine Learning and Data Mining, August 24, 2020, San Diego, CA.* ACM, New York, NY, USA, 5 pages. https://doi.org/xxxxxxxx

1 INTRODUCTION

Machine learning is used in various fields, including computer vision [9], natural language processing [2], healthcare [3], and some applications of machine learning models that often process sensitive and private datasets such as facial images and gene information. New attacks that are unique to machine learning models have been proposed [6, 12, 14]. An attack called the Model Inversion Attack (MIA) [4, 7] has been proposed that retrieves secret information of a dataset through the output. This attack can divulge private training data, and hence in this paper, we focus on MIA.

AdvML '20, August 24, 2020, San Diego, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/xxxxxxxx

In a similar fashion to previous MIA research [8, 16], we suppose that a malicious attacker reconstructs secret training data using a conditional generative adversarial network (cGAN) [11]. This problem formulation causes a serious problem in that performance degradation cannot be calculated using commonly used performance metrics such as the Peak Signal-to-Noise Ratio (SNR) because the restored data do not correspond to the original data on a one-to-one basis. For instance, considering a facial recognition model, for one input image, there is a corresponding output label. However, since we focus on MIA utilizing cGAN, for a class label, there are many corresponding images. This means that we cannot calculate the value of loss functions whose argument is the pair comprising the true value and predicted value, and so we have no indicator to evaluate the success of MIA. This represents a priority issue because if some security measures are proposed, this lack of an evaluation indicator prevents a safety assessment of these measures. For example, if some security measures are proposed, we would not be able to determine which measure is the best. Furthermore, if some improvements to a security measure are proposed, we cannot determine amelioration.

In this paper, we propose a new approach that evaluates the success of MIA, which does not require a pair comprising the true and predicted values as arguments. The key idea in the proposed approach is to measure the success of MIA based on the difference in probability distributions. This idea contributes to addressing the problem that we cannot evaluate the success of MIA; this means that we will be able to evaluate the success of MIA without calculating the evaluation score one by one. In this paper, we implement the approach based on Kullback-Leibler divergence (KL-divergence) [5]. Through numerical simulation, we show that the proposed approach correctly evaluates the success of MIA as expected. Moreover, through experimentation using real data, we demonstrate that the proposed approach is suitable from a quantitative perspective.

2 RELATED WORK

In previous studies of MIA, the problem formulation is divided into two main groups based on whether or not reconstructed data correspond to correct data on a one-to-one basis. In a scenario where the reconstructed data correspond to the correct data, we can employ a commonly used metric, *e.g.*, Accuracy Score, which requires pairs of predicted and correct values. For example, Zhan *et al.* [16] proposed Generative MIA, which generates sensitive features from non-sensitive features. In their problem formulation, training data comprise sensitive features and non-sensitive features and the malicious attacker can utilize non-sensitive features to generate the

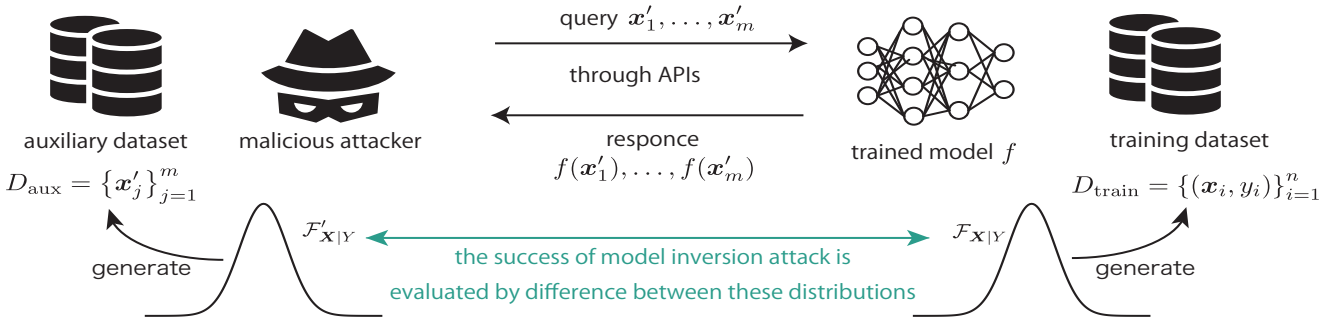

Figure 1: Problem Formulation: A malicious attacker aims to approximate a probability distribution, which is the same as that in the training dataset, and to restore secret training data by resampling the approximated distribution. The attacker is allowed to access the model only through query x' and response $f(x')$. The proposed approach evaluates the success of MIA based on the difference between the conditional distribution of an auxiliary dataset and that in the training dataset.

sensitive features. For example, considering a facial recognition model, Generative MIA reconstructs a complete facial image from an incomplete facial image where both eyes, the nose, and mouth are hidden. They evaluated their proposal using a popular metric, the Peak SNR.

In a scenario where the correct data does not correspond to the predicted data, we must evaluate the attack manually. For instance, Kusano *et al.* [8] proposed PreImageGAN, which approximates the distribution of a training dataset and reconstructs training data from the distribution. In their problem setting, a malicious attacker utilizes an auxiliary dataset that contains similar data to the secret training dataset. For instance, if the training dataset contains facial images of Alice and Bob, the auxiliary dataset contains facial images of other people but does not contain the images of Alice and Bob. Utilizing the auxiliary dataset, PreImageGAN approximates the probability distribution of the secret training dataset by cGAN. They evaluate how correctly PreImageGAN reconstructs the training data based on the following: 1) they generate samples actually from the approximated probability distribution and 2) the authors evaluate how correctly the method reconstructs the training data qualitatively.

In the setting described by Zhan *et al.* the reconstructed data, *i.e.*, sensitive features, from the attacker correspond to the training data one-to-one because the attacker utilizes non-sensitive features of the training data. Specifically, considering a facial recognition model example, when the attacker tries to restore the facial image of Alice from one image where both eyes, the nose, and mouth are hidden, the hidden parts are the corresponding corrects of reconstructed data. Due to this one-to-one correspondence between the reconstructed data and training data, we evaluate how correctly the attacker can regenerate training data based on, for example, the Peak SNR or Attack Accuracy.

Conversely, in the problem formulation by Kusano *et al.*, the attacker approximates the probability distribution of the training data and generates training data from it; therefore, the generated data and training data do not have a one-to-one correspondence. Since there is no one-to-one correspondence, we cannot use the previously noted evaluation indicator. The problem setting herein is similar to that by Kusano *et al.*; therefore, we cannot use an evaluation indicator that requires a one-to-one correspondence between the correct training data and a prediction of the data.

3 PROBLEM FORMULATION

We show the considered problem formulation in Figure 1. A malicious attacker aims to reconstruct data for which the probability distribution is the same as that for secret training dataset D_{train}. The attacker is allowed access to the model only through query x' and response $f(x')$ and does not know information pertaining to the model, *e.g.*, the architecture and algorithm.

As in the case in other studies [8, 14], we suppose that the attacker can access auxiliary dataset $D_{\text{aux}} = \{x'_j\}_{j=1}^m$ to attack target model f. The auxiliary dataset is not the same as the secret training dataset but similar to it. Considering a facial recognition model example, the attacker can use facial images gathered on the Internet to reconstruct the secret dataset that is used to train the target model. This problem setting is not fantasy because the attacker is often able to speculate the type of secret training dataset from the model response [4].

The attacker regenerates training data using cGAN. Let G and D be the generator and discriminator, respectively. The attacker trains G and D using $\{(x'_j, f(x'_j))\}_{i=1}^m$. Then, the generator approximates the conditional distribution of the secret training dataset. Therefore, by generating samples from the approximated probability distribution, the attacker can reconstruct the training data.

4 PROPOSAL

We propose a new approach to evaluate the successfulness of MIA. Specifically, we evaluate the degree of success of the attack by comparing an actual MIA case to the worst case. The worst-case means that the attacker has the training dataset and this is the easiest situation for the attacker to approximate the distribution of D_{train}. Formally, in the worst-case, the attacker approximates the conditional probability distribution by solving

$$\widehat{\theta}_G^{(\mathrm{I})} = \arg\min_{\theta_G} \max_{\theta_D} \int \log D(x, y) \mathcal{F}_{X|Y}(\mathrm{d}x \mid Y = y) + \int \log\{1 - D(G(z, y), y)\} \mathcal{F}_Z(\mathrm{d}z), \quad (1)$$

where $\mathcal{F}_{\boldsymbol{X}|Y}$ represents the conditional distribution of D_{train} given the class label and $\mathcal{F}_{\boldsymbol{Z}}$ represents the distribution of input noise of the generator. However, in an actual situation, the attacker solves

$$\widehat{\boldsymbol{\theta}}_G^{(\mathrm{A})} = \arg\min_{\boldsymbol{\theta}_G} \max_{\boldsymbol{\theta}_D} \int \log D(\boldsymbol{x}, y) \mathcal{F}'_{\boldsymbol{X}|Y}(\mathrm{d}\boldsymbol{x} \mid Y = f(\boldsymbol{x})) + \int \log\{1 - D(G(\boldsymbol{z}, y), y)\} \, \mathcal{F}_{\boldsymbol{Z}}(\mathrm{d}\boldsymbol{z}), \quad (2)$$

where $\mathcal{F}'_{\boldsymbol{X}|Y}$ represents the conditional distribution of D_{aux} given class label Y. In what follows, $\mathcal{F}_{\boldsymbol{X}|Y}(\mathrm{d}\boldsymbol{x} \mid Y = y)$ and $\mathcal{F}'_{\boldsymbol{X}|Y}(\mathrm{d}\boldsymbol{x} \mid Y = f(\boldsymbol{x}))$ are symbolized by $\mathcal{F}_{\boldsymbol{X}|Y}$ and $\mathcal{F}'_{\boldsymbol{X}|Y}$, respectively. Comparing Equation (1) to Equation (2), the difference between $\widehat{\boldsymbol{\theta}}_G^{(\mathrm{I})}$ and $\widehat{\boldsymbol{\theta}}_G^{(\mathrm{A})}$ is a result of the separation between $\mathcal{F}_{\boldsymbol{X}|Y}$ and $\mathcal{F}'_{\boldsymbol{X}|Y}$.

We rewrite how well MIA reconstructs the training data to how well the generator approximates the probability distribution of the training data. In addition, how well the generator approximates the probability is evaluated by the degree of similarity between $\widehat{\boldsymbol{\theta}}_G^{(\mathrm{I})}$ and $\widehat{\boldsymbol{\theta}}_G^{(\mathrm{A})}$. Since the difference between $\widehat{\boldsymbol{\theta}}_G^{(\mathrm{I})}$ and $\widehat{\boldsymbol{\theta}}_G^{(\mathrm{A})}$ is a result of the separation between $\mathcal{F}_{\boldsymbol{X}|Y}$ and $\mathcal{F}'_{\boldsymbol{X}|Y}$, we propose evaluating the success of MIA based on the difference between $\mathcal{F}_{\boldsymbol{X}|Y}$ and $\mathcal{F}'_{\boldsymbol{X}|Y}$. To measure this, we utilize the KL-divergence between $\mathcal{F}_{\boldsymbol{X}|Y}$ and $\mathcal{F}'_{\boldsymbol{X}|Y}$,

$$\int \log\left(\frac{\mathcal{F}_{\boldsymbol{X}|Y}(\mathrm{d}\boldsymbol{x} \mid Y = y)}{\mathcal{F}'_{\boldsymbol{X}|Y}(\mathrm{d}\boldsymbol{x} \mid Y = f(\boldsymbol{x}))}\right) \mathcal{F}_{\boldsymbol{X}|Y}(\mathrm{d}\boldsymbol{x} \mid Y = y),$$

which is the most popular divergence for measuring the difference in probability distributions.

5 EVALUATION

In order to confirm that the proposed approach can measure the degree of success of MIA, we performed a numerical simulation and real data experiment. There are two steps to implement the proposed approach using real data: estimate the KL-divergence from data and then evaluate the success of MIA using the estimated KL-divergence. This means that if we cannot estimate correctly the KL-divergence in the first step, the proposed approach does not function appropriately; however, improving the quality of estimating KL-divergence is not the main topic of this paper. Therefore, we performed not only a real data experiment, but also numerical simulation. In the numerical simulation, the KL-divergence can be theoretically calculated, *i.e.*, we do not need to estimate it from data. Hence, we can evaluate the proposed approach without discussing whether or not a selected method for estimating the KL-divergence is suitable. We confirm that the proposed approach is applicable when the influence from estimating the KL-divergence is removed by numerical simulation, and verify that the proposed approach functions appropriately for real data.

5.1 Numerical Simulation

The purpose of the numerical simulation is to evaluate the proposed approach under ideal conditions where we can estimate the true KL-divergence between $\mathcal{F}_{\boldsymbol{X}|Y}$ and $\mathcal{F}'_{\boldsymbol{X}|Y}$ from real data. Specifically, we suppose that the probability distributions of the training data and auxiliary data are Gaussian. Then, the KL-divergence can be analytically derived [15].

Table 1: Results of numerical simulation. As the KL-divergence becomes large, the null hypothesis tends to be rejected.

d	KL-Divergence	Null Hypothesis (rejected or not)
1	0.5	Not rejected
5	2.5	Not rejected
10	5	Not rejected
50	25	Rejected
100	50	Rejected
500	250	Rejected

Table 2: Details of EMNIST dataset

Dataset	Details
mnist	Contains handwritten numbers from 0 to 9.
letters	Contains a handwritten alphabet For all alphanumerics, upper-case letters and lower-case letters are merged into the same class.
balanced	Contains numbers from 0 to 9 and alphanumerics. For C, I, J, K, L, M, O, P, S, U, V, W, X, Y and Z, upper-case letters and lower-case letters are merged into the same class.

5.1.1 Setting of Numerical Simulation. We generate the training dataset and auxiliary dataset and perform model inversion. The training dataset of class 0 is $\boldsymbol{x}_1^{(0)}, \ldots, \boldsymbol{x}_{1000}^{(0)} \overset{\text{iid}}{\sim} \mathcal{N}(-\mathbf{1}_d, I_d)$ and that of class 1 is $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{1000}^{(1)} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{1}_d, I_d)$ where $\mathbf{1}$ represents a d-dimensional vector whose elements are all 1, I_d represents d-dimensional identity matrix, and $\mathcal{N}$ represents a Gaussian distribution. The auxiliary dataset is $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{500} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_d, I_d)$.

At any fixed d, we apply the Kolmogorov-Smirnov test [10] with a null hypothesis that the probability distribution of a reconstructed dataset is the same as $\mathcal{F}_{X|Y}$. If the null hypothesis is rejected when the KL-divergence is large, this means that when the KL-divergence is large model inversion is not successful.

5.1.2 Results of Numerical Simulation. The results of the numerical simulation are given in Table 1. As the KL-divergence becomes large, the null hypothesis tends to be rejected. Therefore, the proposed approach is suitable as an evaluation indicator for model inversion in an ideal situation.

5.2 Real Data Experiment

Except for the influence of estimating the quality of the KL-divergence, the numerical simulation has revealed that the proposed approach is applicable as an estimation indicator of MIA statistically. On the other hand, in the real data experiment, we aim to verify that the proposed approach is practically suitable as an evaluation indicator, even if the KL-divergence is estimated from data. Precisely, by contrasting regenerated images with the estimated value of the

(a) Auxiliary dataset is mnist (b) Auxiliary dataset is letters (c) Auxiliary dataset is balanced

Figure 2: Reconstructed images for each auxiliary dataset. In all cases, victim training dataset is the training dataset of mnist and the malicious attacker uses test dataset of mnist, letters, or balanced as the auxiliary dataset in (a), (b), and (c) respectively. In each case, the top-left image corresponds to the image of 0, the top-second-left image corresponds to the image of 1, and the bottom-right image corresponds to the image of 9.

Table 3: Estimated KL-divergence between $\mathcal{F}_{X|Y}$ and $\mathcal{F}'_{X|Y}$ For instance, the element whose row is letters and column is 0 represents KL-divergence between $\mathcal{F}_{X|Y}$ and $\mathcal{F}'_{X|Y}$.

Auxiliary dataset	Condition of class label, *i.e.*, y and $f(x)$									
	0	1	2	3	4	5	6	7	8	9
mnist (test dataset)	0.22	0.28	0.16	0.20	0.23	0.20	0.26	0.22	0.19	0.15
letters	3.79	0.60	1.37	4.73	1.63	0.78	9.59	1.83	6.40	1.56
balanced	3.03	0.34	0.68	0.90	0.91	1.01	4.65	0.73	4.92	0.66

estimated KL-divergence, we confirm that the MIA in the class of large KL-divergence cannot be successful.

For the experiment, the EMNIST dataset [1], which includes three datasets, mnist, letters, and balanced are used. These datasets contain 28x28 pixel images of handwritten characters and details of each dataset are given in Table 2. The first reason why we apply the EMNIST dataset is that EMNIST is the standard dataset for machine learning. Second, the purpose of this paper is not to derive an excellent MIA, but to propose an approach that evaluates the success of MIA. Thus, even if the EMNIST dataset is easy to classify and attack, the dataset is sufficient to evaluate the proposal.

The proposed approach enables us to weigh the relative metrics of model safety against the MIA, but not to determine whether or not the model is secure. Furthermore, the quality of estimating the KL-divergence affects whether or not the approach is reliable. Although for an easy dataset such as EMNIST, the proposed approach is functional; however, for complex data that estimate the KL-divergence the proposed approach may not be applicable.

5.2.1 Experimental Configuration. The victim model is trained with the training dataset of mnist, thus the target model is a classifier from grey-scale images to a label of a number from 0 to 9. A malicious attacker utilizes test dataset of mnist, letters, and balance. Here the test dataset of mnist does not contain the same data as mnist. To estimate the KL-divergence, we apply a method proposed by Pérez-Cruz [13].

5.2.2 Results of Real Data Experiment. The reconstructed images are shown in Figure 2 and the estimated KL-divergence corresponding to each case is given in Table 3. According to Table 3, when the auxiliary dataset is mnist, the KL-divergence tends to be small compared to the other two cases. Simultaneously, the corresponding regenerated images in Figure 2(a) seem to be better than the others. As for when the auxiliary dataset is letters, the KL-divergence corresponding to $f(\boldsymbol{x}) = 3, 6, 8$ are especially high and the values of $f(\boldsymbol{x}) = 1, 5$ are especially small. Focusing on the restored images, the images of 3, 6, and 8 in Figure 2(b) cannot be recognized as 3, 6, and 8. In contrast, the image quality of 1 and 5 is better than that of 3, 6, and 8. When the auxiliary dataset is balanced, the KL-divergence of $f(\boldsymbol{x}) = 6, 8$ is particularly larger than the others and the values of $f(\boldsymbol{x}) = 1, 2$ are less than the others. By the same token, in Figure 2(c), the image quality of 6 and 8 is worse than that of the others. Conversely, the images of 1 and 2 are regenerated well.

Considering the results of the experiment, if the KL-divergence is large, the quality of the model inversion becomes worse. Conversely, if the KL-divergence is small, the quality of the model inversion becomes better. This means that the KL-divergence represents the degree of success of MIA, and therefore the proposed approach, by evaluating the success of the model inversion based on the difference between the probability distributions, is functional.

6 CONCLUSION

We proposed a new approach that evaluates the success of MIA by utilizing the difference between probability distributions of the training dataset and auxiliary dataset. To calculate the difference, we employed the KL-divergence and demonstrated that the KL-divergence represents the attack success level based on numerical simulation and a real data experiment. The simulation and experimental results indicate that the proposed approach is suitable as an evaluation indicator of the attack.

In this paper, we supposed that the attacker utilizes cGAN with MIA; however, the attacker may employ other types of GAN for the attack. Thus, as future work, we will verify whether the proposed approach is functional against other types of MIAs. Furthermore, although KL-divergence is applied to calculate the difference in

probability distributions, other types of divergence can be utilized. In the future, we will also inspect what type of divergence is applicable to what type of GAN.

REFERENCES

[1] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. 2017. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2921–2926.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature medicine* 25, 1 (2019), 24–29.

[4] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 1322–1333.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[7] Seira Hidano, Takao Murakami, Shuichi Katsumata, Shinsaku Kiyomoto, and Goichiro Hanaoka. 2018. Model inversion attacks for online prediction systems: Without knowledge of non-sensitive attributes. *IEICE Transactions on Information and Systems* 101, 11 (2018), 2665–2676.

[8] Kosuke Kusano and Jun Sakuma. 2018. Classifier-to-Generator Attack: Estimation of Training Data Distribution from Classifier. (2018).

[9] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. 2020. SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rkl03ySYDH

[10] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[11] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[12] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2019. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4954–4963.

[13] Fernando Pérez-Cruz. 2008. Kullback-Leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*. IEEE, 1666–1670.

[14] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[15] Martin J Wainwright. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press.

[16] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2019. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. *arXiv preprint arXiv:1911.07135* (2019).