Connecting Membership Inference Privacy and Generalization through Instance-Wise Measurements

Leah Woldemariam LSW85@CORNELL.EDU
Anna Scaglione AS337@CORNELL.EDU
Cornell University, USA

Abstract

Membership Inference Attacks (MIAs) seek to assess the privacy risk of a model by extracting membership information, which represents a fundamental unit of information that a model contains. A prevailing intuition in the MIA literature is that decreasing the amount of information in a neural network should improve both privacy risk and generalization ability. Despite the intuitive connection, both theoretical and empirical work has suggested that regularization, whether implicit or explicit, has widely different effects on privacy risk across the individual points in the training data. In this work, we take a first step towards understanding the relationship between privacy and generalization by deriving an instance-wise measurement of Membership Inference Privacy (MIP). We then connect this definition to generalization bounds using a data-dependent prior on the weight distribution.

Keywords: generalization, membership inference privacy, hypothesis testing

1. Introduction

Membership information represents a basic unit of information that a deep learning model can leak about its training data. The goal of a Membership Inference Attack (MIA) is to guess whether or not a given data point was in a training set, serving as an empirical method of evaluating information leakage. Additionally, MIAs have been used to strengthen other attacks by identifying data can be more easily extracted or reconstructed in attacks against LLMs [1]. Although differential privacy provides provable privacy guarantees and defend against MIAs, a general belief in the MIA literature is that explicit regularization techniques should increase the error of MIAs [4, 16, 17, 22, 33]. However, a common conclusion is that while broad statements about a model's susceptibility to MIAs can be made, susceptibility clearly has a dependence on the specific data point being studied: an exceedingly rare instance is more susceptible to MIAs than more ordinary data points [2, 4]. Moreover, the precise influence of regularization on MIA accuracy is contingent on the specific type of regularization used, such as ℓ_2, dropout, or label smoothing [4, 18].

Analogously, generalization is often controlled through these same explicit regularization techniques, and again, discrepancies across the training set appears: recent work theorizes that discrepancies in privacy levels between instances in the training set are inherent to maintaining near-optimal performance [9]. Although these two properties are guided by the same principles, the most frequently reported theoretical connection in the literature is that differential privacy guarantees imply a bound on generalization. Otherwise, the connection between generalization and privacy at the level of individual data points remains mostly unexplored. This motivates our study of instance-wise privacy, allowing for a more clear understanding of the relationship between privacy and generalization.

1.1. Related Work

In addition to discrepencies that arise due to data difficulty, there are also inconsistencies that arise when varying the regularization technique itself: although several techniques such as sharpness-aware minimization, early stopping, and model stacking have been empirically shown to provide a defense against certain MIAs [21, 26], some techniques, namely label smoothing and weight decay, are observed to *increase* susceptibility to MIAs. Even when fixing the regularization technique, tuning the regularization hyper-parameter can pose issues: [4] finds that techniques usually used to improve generalization ability can have a nonlinear relationship with membership inference, revealing the gap in understanding the extent to which these tools act as tunable controllers. These results complicate the exact relationship between generalization and privacy under the context of explicit regularization and the data distribution. A first step to unraveling this mystery would be to understand how membership inference privacy can be defined for each individual instance and the implications of this definition for the generalization ability of a model.

Theoretical works connecting generalization and membership inference privacy vary widely in the settings they consider. In [33], a relationship between generalization and membership inference is analyzed after defining adversarial advantage probabilistically over the data distribution. [29] studies the relationship between the number of non-zero values in the weights and privacy, specifically in the case of logistic regression, using notions of adversarial advantage from [33]. After assuming the distribution of the weights is proportional to a function of the loss of each training point, which [28] notes is a more suitable assumption for SLGD rather than SGD, [25] shows that black-box attacks are comparable to white-box attacks for membership inference.

In this work, we attempt to further uncover the relationship between generalization and privacy by understanding how this relationship simplifies at the example-level. We first highlight the need for an instance-level definition of privacy, which naturally arises when one views membership inference privacy through a hypothesis testing framework. We then decompose this term into a per-iteration leakage term, which we can derive in closed-form under certain assumptions, and relate this term to generalization through data-dependent generalization bounds.

1.2. Background

Notation: We consider the task of supervised classification with a fixed training set Z and $z_i \sim \mathcal{D}$ for all $i \in \{1, \dots, N\}$, where $\mathcal{D}$ is the data distribution over the sample space $\mathcal{Z}$. Let z be a labeled sample, i.e. a pair $z = (x, y)$, where x is the input feature and y is the label. The learned function parameterized by $w \in \mathbb{R}^d$ takes as input x and returns a prediction $\hat{y} = f_w(x)$. Let W denote the random variable representing the weights, W_t the variable at iteration t, and W^t the weights until time t, i.e. $W^t = (W_0, \dots, W_t)$. An instance of the final variable W_T is returned by the learning algorithm $\mathcal{A}(\cdot)$. To define the MIA, we let $Z_i = \{z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_N\}$ be the set that includes z_i and $Z_i' = \{z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_N\}$ be the set where z_i is swapped for $z_i' \sim \mathcal{D}$. The training set Z is chosen to be either Z_i or Z_i' with equal probability. We use $\mathbf{Z_i}$ to denote the random dataset with only z_i' random. We consider SGD as our learning algorithm, where updates at any time t can be written as

$$w_{t+1} = w_t - \eta_t \nabla_w L(b_t, w_t) \tag{1}$$

where the initial condition is random $w_0 \sim W_0$. Because of the random choice of batch samples at every iteration represented by an N-dimensional indicator variable $u_t \sim U$ and the random initial condition, we consider SGD applied over any given Z to define a stochastic process. Let $\ell : \mathcal{Z} \times \mathbb{R}^d \to \mathbb{R}$ be a loss function and define the batched loss as

$$L(b_t, w_t) = \frac{1}{B}\sum_{i=1}^{N} \ell(z_i, w_t)u_i. \tag{2}$$

The loss distribution for a given sample z and the gradient distribution can be written as

$$Q_{Z_i} \triangleq P(\ell(z_i; W_T) \mid Z_i) \qquad G^t_{Z_i} \triangleq P(\nabla L(b_t; W_t) \mid Z_i, W^{t-1}). \tag{3}$$

Additionally, define $P_{Z_i} \triangleq P(W \mid Z_i)$ as the posterior distribution of the weights. $Q_{Z'_i}$, $G^t_{Z'_i}$, and $P_{Z'_i}$ are similarly defined for the dataset Z'_i.

2. An MIA Hypothesis Testing Bound

In order to gain practical insight into the privacy of each point, we focus on the specific issue of membership information leakage. In this section, we will present the key divergence term that bounds the performance of MIAs and later derive an approximation of this term. We will follow the line of MIAs that use loss and output-related information as in [2, 3, 25, 33]. Formally, after a neural network is trained on a dataset Z, an adversary takes as input z and queries the model to get the model prediction $\hat{y}$. Attacks may also consider an adversary with access to the loss $\ell(z, w)$ or predictive distribution. The adversary guesses whether or not z was in the training set that generated w through $\mathcal{A}(Z)$, i.e. whether $z \in Z$ or $z \notin Z$. We formulate this as guessing whether the weights were trained on dataset Z_i or Z'_i. We can now describe an MIA as a hypothesis test between the hypotheses

$$\begin{aligned} H_0 &: \ell(z_i; w) \sim Q_{Z_i} \\ H_1 &: \ell(z_i; w) \sim Q_{\mathbf{Z_i}} \end{aligned} \tag{4}$$

where Q_{Z_i} and $Q_{\mathbf{Z_i}}$ are as defined above, using notation as in [2]. We specifically seek to study the membership inference privacy of fixed instances in order to understand their contribution on the performance of the model. In the alternative hypothesis, we consider the swapped data point to be chosen randomly from the data distribution and later take the expectation over the choice of swapped values.

The lower bound on the sum of false positives $\alpha \triangleq P(T(\ell(z, w)) = 1|H_0)$ and false negatives $\beta \triangleq P(T(\ell(z, w)) = 0|H_1)$ for any test $T : \mathbb{R} \to \{0, 1\}$ is given by

$$\alpha + \beta \geq 1 - ||Q_{Z_i} - Q_{\mathbf{Z_i}}||_{TV}, \tag{5}$$

providing a lower bound on the error of any loss-based MIA. Given the relevance of the term in equation (5), we introduce the following definition.

Definition 1 *The Membership Inference Privacy (MIP) of an example z_i in the training set Z is*

$$\Psi(z_i) \triangleq 1 - \sqrt{KL[Q_{Z_i}||Q_{\mathbf{Z_i}}]}. \tag{6}$$

Following directly from Pinsker's inequality, the MIP $\Psi(z)$ of a data point z is less than or equal to its TV-MIP $\Psi_{TV}(z)$: $\Psi(z) \leq \Psi_{TV}(z) = 1 - ||Q_{Z_i} - Q_{\mathbf{Z_i}}||_{TV}$. Note that MIP is directly tied to differential privacy, as hypothesis testing errors can provide a condition for differential privacy guarantees [15].

2.1. Bounding Membership Inference Privacy

In this section, we seek to bound the membership inference privacy $\Psi(z)$ of a given data point z. We first outline the assumptions made about batch size and sampling. Each batch is sampled without replacement and sampling at a given iteration t can be modelled as drawing a binary vector u from the set of all binary vectors of length N that sum to B, i.e. $\{u : u \in \{0,1\}^N, \sum_{i=1}^N u_i = B\}$. This represents the hypergeometric distribution with a finite population of N items that contain B successes, which asymptotically goes to the Binomial distribution. Thus, each indicator variable included in the batch b_t at time t becomes a Bernoulli random variable. The probability that any given sample is in our batch is akin to the probability of a success given a single draw, so $P(i \in b_t) = \frac{B}{N}$. Formally, we make the following two assumptions.

Assumption 1 *We have N is sufficiently large so that the distribution of u_i is close to in distribution a Bernoulli distribution $Bern(p)$ for all entries i, each independent of one another.*

For a batched updated at time t under dataset Z, we will denote the expected value of the batched update at time t as

$$\mu_Z = \mathbb{E}_u[\nabla L(b_t, w_t)] = \nabla L(Z, w_t) \tag{7}$$

Similarly, when batches are sampled independently and each sample within the batch of size B is chosen without replacement, the covariance matrix can be written as

$$\Sigma_Z = \frac{1}{B}\left[\frac{1}{N}\sum_{i=1}^{N} \nabla \ell(z_i; w_t) \nabla \ell^T(z_i; w_t) - \nabla L(Z, w_t) \nabla L^T(Z, w_t)\right], \tag{8}$$

whose derivation is included in the Appendix.

Assumption 2 *The batched gradient $\nabla L(b_t, w_t)$ has an expected value and covariance that exists and are finite, i.e. $\mu_Z < \infty$ and $\Sigma_Z < \infty$.*

This assumption holds for the many common choices of loss functions and activation function, such as the cross entropy loss and sigmoid activations. We also note that many applications have increasingly large datasets, so a proper choice of batch size B validates Assumption 1. Since each $z_i \in Z$ is drawn from the data distribution $\mathcal{D}$ with an indicator variable u that is asymptotically a vector of i.i.d. Bernoulli random variables at each iteration, then the batched sum can be approximated by a Gaussian by the Central Limit Theorem under Assumptions 1 and 2.

Next, we define $g_i \triangleq \nabla \ell(z_i, w_t)$ as the gradient of the loss for training point i, and let

$$\mu \triangleq \mathbb{E}_z[\nabla \ell(z, w_t)] \qquad \Sigma \triangleq \mathbb{E}_z[(\nabla \ell(z, w_t) - \mu)(\nabla \ell(z, w_t) - \mu)^T] \tag{9}$$

be the expected value and covariance of the gradient over the data distribution, respectively, where we ignore time-indexing for simplicity. We finally arrive at the following derivation.

Theorem 2 (MIP Bound) *As $N \to \infty$, for sufficiently large batch size B,*

$$\mathbb{E}_{z_i'} KL[Q_{Z_i} || Q_{\mathbf{Z_i'}}] \leq \sum_{t=1}^{T} \frac{1}{2} \mathbb{E}_{W_{t-1}} \left[\frac{1}{N^2} ||g_i - \mu||_{\Sigma^-_{Z_i}} + tr(\Sigma^-_{Z_i} \Sigma) \right]. \tag{10}$$

The proof can be found in the Appendix. Intuitively, this result states that the leakage of a fixed instance z_i is related to its expected deviation from the expected gradient μ weighted more heavily when the gradient points in uncommon directions. This directly implies that

$$\alpha + \beta \geq \Psi(z_i) \geq 1 - \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_{W_{t-1}} \left[\frac{1}{N^2} ||g_i - \mu||_{\Sigma^-_{Z_i}} + tr(\Sigma^-_{Z_i} \Sigma) \right], \tag{11}$$

highlighting the importance of the gradient at each iteration in bounding the error of an MIA. We plot some of the MIP scores taken over MNIST in Appendix A.1.

3. Connection to Generalization

Given a loss function as above, the population risk of weights w is defined as $\mathbb{E}_z[\ell(z, w)]$, with empirical risk $\mathbb{E}_w \left[\frac{1}{N} \sum_{i=1}^{N} \ell(z_i, w) \right]$. The generalization error, taken as an expectation over the weights, can be defined as

$$gen(w, Z) \triangleq \mathbb{E}_w[\mathbb{E}_z[\ell(z, w)] - L(Z, w)]. \tag{12}$$

PAC-Bayes generalization bounds use the divergence between a prior and posterior on the weights in order to bound equation 12 for stochastic classifiers, and information-theoretic bounds similarly use the expectation over the divergence for noisy algorithms [7, 23, 24]. Instead of modeling SGD as a discretization of SGLD, as is common in information-theoretic bounds, we can use the following SDE

$$dW_t = -\nabla L(W_t)dt + \sqrt{\eta_t \Sigma_t} dB_t$$

that is often used to model the anisotropic stochasticity in SGD due to batch noise; Σ_t controls the batch noise and dB_t denotes a Brownian process [11, 13, 20, 27, 31]. As we earlier defined P_{Z_i} as the posterior distribution of the weights, we now define P as the prior distribution of the weights. Taking from data-dependent estimates and priors from [23], we can choose the prior to be data-dependent so long as the subset of data it depends on is independent of the training set [23]. In order to connect this to membership inference privacy, we seek to first bound the divergence term below following steps similar to the ones above using an set $\bar{Z}$ independent from Z.

Lemma 3 *(Per-Iteration Divergence) As $N \to \infty$, the expected divergence between gradients at iteration t can be written as*

$$\mathbb{E}_{W_{t-1}} KL[G^t_Z || G^t_{\bar{Z}}] = \frac{1}{2} \mathbb{E}_{W_{t-1}} \left[tr(\Sigma^-_Z \Sigma_{\bar{Z}}) - d + \ln \frac{|\Sigma_Z|}{|\Sigma_{\bar{Z}}|} + ||\mu_Z - \mu_{\bar{Z}}||_{\Sigma^-_Z} \right].$$

This follows from assumptions similar to ones as above, and the full proof is provided in the Appendix. Before providing the bound on the divergence term appearing in the generalization bounds, we state the following remark.

Lemma 4 *As $N \to \infty$, letting $P = \mathbb{E}_{\bar{Z}} P_{\bar{Z}}$ be a data-dependent prior, we can bound the divergence term $KL[P_Z||P]$ with*

$$KL[P_Z||P] \leq \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_{W_{t-1}} \left[tr(\Sigma_Z^{-}\Sigma) + ||\mu_Z - \mu||_{\Sigma_Z^{-}} \right]. \tag{13}$$

This choice of prior is common in PAC-Bayes and information-theoretic bounds that utilize a subset of the training set [8, 19, 23], and our results are in particular similar to those in [19, 30]. The final insight comes from the derivation of the MIPs above, which the generalization bound can be decomposed into.

Lemma 5 *Generalization error can be bounded for some constant C by*

$$gen(w, Z) \leq \sqrt{C \left[\sum_{i=1}^{N} \Psi(z_i) - (N-1)\, \mathbb{E}_{W_{t-1}} \left[tr(\Sigma_Z^{-}\Sigma) \right] \right]}.$$

Proof This comes immediately from replacing the divergence term with our bound in Lemma 4, setting $C = \frac{\sigma^2}{N-N'}$ as a constant, and plugging into the generalization bound from [23]. ■

Again, we are able to connect this definition back to the familiar results on differential privacy [5], although the bound here is a more indirect connection. When combined with the connection between MIP and DP in [15], MIP also implies a bound on generalization error.

3.1. Discussion

Generalization bounds are in part derived in order to gain further understanding in the key terms that control generalization behavior and perhaps motivate optimization techniques, but can be vacuous when used practically [6, 10, 14]. By decomposing our complexity measure into one that depends on each training point, we gain understanding into how the variance of the distribution and deviation of individual points might add to the uncertainty in the bound. In Appendix A.1, we see that high MIP scorers correspond to outliers visually. These bounds pave the way for future work on studying the role of the data distribution and loss function in generalization and privacy.

References

[1] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In 30th USENIX security symposium (USENIX Security 21), pages 2633–2650, 2021.

[2] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In 2022 IEEE Symposium on Security and Privacy (SP), pages 1897–1914. IEEE, 2022.

[3] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In International conference on machine learning, pages 1964–1974. PMLR, 2021.

[4] Antreas Dionysiou and Elias Athanasopoulos. Sok: Membership inference is harder than previously thought. Proceedings on Privacy Enhancing Technologies, 2023.

[5] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In Proceedings of the forty-seventh annual ACM symposium on Theory of computing, pages 117–126, 2015.

[6] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. arXiv preprint arXiv:1703.11008, 2017.

[7] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. Advances in Neural Information Processing Systems, 33:11723–11733, 2020.

[8] Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in pac-bayes bounds. In International Conference on Artificial Intelligence and Statistics, pages 604–612. PMLR, 2021.

[9] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pages 954–959, 2020.

[10] Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Fantastic generalization measures are nowhere to be found. arXiv preprint arXiv:2309.13658, 2023.

[11] Jonas Geiping, Micah Goldblum, Phillip E Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. arXiv preprint arXiv:2109.14119, 2021.

[12] Hrayr Harutyunyan, Alessandro Achille, Giovanni Paolini, Orchid Majumder, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Estimating informativeness of samples with smooth unique information. arXiv preprint arXiv:2101.06640, 2021.

[13] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio. In Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27, pages 392–402. Springer, 2018.

[14] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. arXiv preprint arXiv:1912.02178, 2019.

[15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In International conference on machine learning, pages 1376–1385. PMLR, 2015.

[16] Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In International conference on machine learning, pages 5345–5355. PMLR, 2021.

[17] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. On the effectiveness of regularization against membership inference attacks. arXiv preprint arXiv:2006.05336, 2020.

[18] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In 29th USENIX security symposium (USENIX Security 20), pages 1605–1622, 2020.

[19] Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. arXiv preprint arXiv:1902.00621, 2019.

[20] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). Advances in Neural Information Processing Systems, 34:12712–12725, 2021.

[21] Jiaxiang Liu, Simon Oya, and Florian Kerschbaum. Generalization techniques empirically outperform differential privacy against membership inference. arXiv preprint arXiv:2110.05524, 2021.

[22] Sasi Kumar Murakonda and Reza Shokri. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. arXiv preprint arXiv:2007.09339, 2020.

[23] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. Advances in Neural Information Processing Systems, 32, 2019.

[24] Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In 2018 IEEE International Symposium on Information Theory (ISIT), pages 546–550. IEEE, 2018.

[25] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In International Conference on Machine Learning, pages 5558–5567. PMLR, 2019.

[26] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv preprint arXiv:1806.01246, 2018.

[27] Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In International Conference on Machine Learning, pages 9058–9067. PMLR, 2020.

[28] Anshuman Suri, Xiao Zhang, and David Evans. Do parameters reveal more than loss for membership inference? arXiv preprint arXiv:2406.11544, 2024.

[29] Jasper Tan, Daniel LeJeune, Blake Mason, Hamid Javadi, and Richard G Baraniuk. A blessing of dimensionality in membership inference through regularization. In International Conference on Artificial Intelligence and Statistics, pages 10968–10993. PMLR, 2023.

[30] Ziqiao Wang and Yongyi Mao. On the generalization of models trained with sgd: Information-theoretic bounds and implications. arXiv preprint arXiv:2110.03128, 2021.

[31] Ziqiao Wang and Yongyi Mao. Two facets of sde under an information-theoretic lens: Generalization of sgd via training trajectories and via terminal states. arXiv preprint arXiv:2211.10691, 2022.

[32] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In International Conference on Machine Learning, pages 10367–10376. PMLR, 2020.

[33] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st computer security foundations symposium (CSF), pages 268–282. IEEE, 2018.

[34] Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in sgd. arXiv preprint arXiv:2102.05375, 2021.

Appendix A.

Remark 6 (Mean & Sample Covariance) *When the batches are sampled independently and each sample within the batch of size B is chosen without replacement, the expected batched update at time t is*

$$\mu_Z = \mathbb{E}_u[\nabla L(b_t, w_t)] = \nabla L(Z, w_t) \tag{14}$$

and the covariance matrix can be written as

$$\Sigma_Z = \frac{1}{B}\left[\frac{1}{N}\sum_{i=1}^{N}\nabla\ell(z_i; w_t)\nabla\ell^T(z_i; w_t) - \nabla L(Z, w_t)\nabla L^T(Z, w_t)\right]. \tag{15}$$

Proof First, the expected value:

$$\begin{aligned}
\mathbb{E}_u[L(b_t, w_t)] &= \frac{1}{B}\sum_{i=1}^{N}\mathbb{E}_u\left[\ell(z_i; w_t)u_i\right] \\
&= \frac{1}{B}\sum_{i=1}^{N}\ell(z_i, w_t)\sum_{u\in\Omega}P(u)u_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\ell(z_i, w_t) \triangleq L(Z, w_{t-1}).
\end{aligned}$$

The covariance matrix can be seen from the derivation:

$$\begin{aligned}
\Sigma_Z &= \mathbb{E}_u[(\nabla L(b_t, w_t) - \mu_Z)(\nabla L(b_t, w_t) - \mu_Z)^T] \\
&= \mathbb{E}_u\left[\left(\frac{1}{B}\sum_{i=1}^{N}\nabla\ell(z_i, w_t)u_i - \mu_Z\right)\left(\frac{1}{B}\sum_{i=1}^{N}\nabla\ell(z_i, w_t)u_i - \mu_Z\right)^T\right] \\
&= \frac{1}{B}\left[\frac{1}{N}\sum_{i=1}^{N}\nabla\ell(z_i; w_t)\nabla\ell(z_i; w_t)^T - \nabla L(Z, w_t)\nabla L(Z, w_t)^T\right].
\end{aligned}$$

and is observed in [12, 32, 34]. ■

Remark 7 (LOO Covariance) *For any index i, $\Sigma_{Z_i'}$ can be written in terms of Σ_{Z_i} as*

$$\Sigma_{Z_i'} = \Sigma_{Z_i} + \frac{1}{N}(g_i' g_i'^\top - g_i g_i^\top) - \frac{1}{N}\left(G\delta_i^\top + \delta_i G^\top\right) - \frac{1}{N^2}\delta_i\delta_i^\top. \tag{16}$$

where $\delta_i = g_i' - g_i$.

Proof Let

$$g_i = \nabla\ell(z_i, w_{t-1}), \qquad G = \frac{1}{N}\sum_{i=1}^{N}g_i, \qquad \text{and} \qquad G' = G - \frac{1}{N}g_N + \frac{1}{N}g_N' = G + \frac{1}{N}\delta_N$$

where $\delta = g'_N - g_N$; we swap g_N for g'_N without loss of generality, and Z'_N has a single data point swapped out. Then

$$\begin{aligned}
\Sigma_{Z'_N} &= \left(\frac{1}{N}\sum_{i=1}^{N-1} g_i g_i^\top + \frac{1}{N} g'_N g'^\top_N\right) - G'G'^\top \\
&= \left(\frac{1}{N}\sum_{i=1}^{N-1} g_i g_i^\top + \frac{1}{N} g'_N g'^\top_N\right) - \left(GG^\top + \frac{1}{N}G\delta + \frac{1}{N}\delta G^\top + \frac{1}{N^2}\delta\delta^\top\right) \\
&= \Sigma_{Z_N} + \frac{1}{N}(g'_N g'^\top_N - g_N g_N^\top) - \frac{1}{N}\left(G\delta^\top + \delta G^\top\right) - \frac{1}{N^2}\delta\delta^\top.
\end{aligned}$$

Taking the expectation over the swapped point g'_N, we get that

$$\begin{aligned}
\mathbb{E}_{g'_i}[\Sigma_{Z'}] = \Sigma_Z &+ \frac{1}{N}\left(\Sigma + \mu\mu^\top - g_i g_i^\top\right) \\
&- \frac{1}{N}\left(G(\mu - g_i)^\top + (\mu - g_i)G^\top\right) \\
&- \frac{1}{N^2}\left(\Sigma + (\mu - g_i)(\mu - g_i)^\top\right)
\end{aligned}$$

since $\mathbb{E}[g'_i g'^\top_i] = \Sigma + \mu\mu^\top$. Finally, we will define Δ' as

$$\Delta' \triangleq +\frac{1}{N}(g'_i g'^\top_i - g_i g_i^\top) - \frac{1}{N}\left(G\delta^\top + \delta G^\top\right) - \frac{1}{N^2}\delta\delta^\top$$

so that $\Sigma_{Z'} = \Sigma_Z + \Delta'$ and define $\Delta \triangleq \mathbb{E}_{z'_i}[\Delta']$. ■

Lemma 8 (Instance-Wise Leakage) *The per-iteration MIA leakage for z, i.e. the divergence between two batched updates at time t, can be written as*

$$\mathbb{E}_{W_{t-1}} KL[G^t_{Z_i}||G^t_{\mathbf{Z_i}'}] = \frac{1}{2}\left[\mathbb{E}[tr(\Sigma^-_{Z_i}\Sigma_{Z'_i}) - d + \ln\frac{|\Sigma_{Z_i}|}{|\Sigma_{Z'_i}|}] + \mathbb{E}||\mu_{Z_i} - \mu_{Z'_i}||_{\Sigma^-_{Z_i}}\right].$$

and the expected leakage of the swapped value gives

$$\mathbb{E}_{z'_i, W_{t-1}} KL[G^t_{Z_i}||G^t_{\mathbf{Z_i}'}] = \frac{1}{2}\mathbb{E}_{W_{t-1}}\left[\frac{1}{N^2}||g_i - \mu||_{\Sigma^-_{Z_i}} + tr(\Sigma^-_{Z_i}\Sigma)\right].$$

Proof The first equation comes from the expression of the divergence between two multivariate Gaussians using the statistics above. Next, recall that we are considering a random swap z'_i and that we have defined Δ' such that $\Sigma_{Z'} = \Sigma_Z + \Delta'$ and $\Delta \triangleq \mathbb{E}_{z'_i}[\Delta']$. Taking the expectation over this swap, we get that

$$
\begin{aligned}
\mathbb{E}_{z_i',W_{t-1}} KL[G^t_{Z_i} || G^t_{\mathbf{Z_i}'}] &= \frac{1}{2}\mathbb{E}_{z_i'}\left[\mathbb{E}[tr(\Sigma^{-}_{Z_i}\Sigma_{Z_i'}) - d + \ln\frac{|\Sigma_{Z_i}|}{|\Sigma_{Z_i'}|}] + \mathbb{E}||\mu_{Z_i} - \mu_{Z_i'}||_{\Sigma^{-}_{Z_i}}\right] \\
&= \frac{1}{2}\mathbb{E}_{z_i',W_{t-1}}\left[tr(\Sigma^{-}_{Z_i}\Sigma_{Z_i'}) - d + \ln\frac{|\Sigma_{Z_i}|}{|\Sigma_{Z_i'}|}\right] \\
&+ \frac{1}{2}\mathbb{E}_{W_{t-1}}\left[||\frac{1}{N}(g_i-\mu)||_{\Sigma_{Z_i}^{-}} + \frac{1}{N}tr(\Sigma_{Z_i}^{-}\Sigma)\right] \\
&\approx \frac{1}{2}\left[d + \mathbb{E}_{W_{t-1}} tr\left(\Sigma_{Z_i}^{-}\Delta\right) - d - \mathbb{E}_{W_{t-1}} tr(\Sigma_{Z_i}^{-}\Delta)\right] \\
&+ \frac{1}{2}\mathbb{E}_{W_{t-1}}\left[\frac{1}{N^2}||g_i - \mathbb{E}_{z_i'}g_i'||_{\Sigma_{Z_i}^{-}} + tr(\Sigma^{-}_{Z_i}\Sigma)\right] \\
&= \frac{1}{2}\mathbb{E}_{W_{t-1}}\left[\frac{1}{N^2}||g_i - \mu||_{\Sigma_{Z_i}^{-}} + tr(\Sigma^{-}_{Z_i}\Sigma)\right].
\end{aligned}
$$

where Σ is now the covariance of the gradient for any sample and μ is the expected value across the data distribution:

$$
\Sigma = \mathbb{E}_z[(\nabla\ell(z,w_t)-\mu)(\nabla\ell(z,w_t)-\mu)^T], \qquad \mu = \mathbb{E}_z[\nabla\ell(z,w_t)].
$$

This comes from the following approximations:

$$
\mathbb{E}_{z_i',W_{t-1}}[tr(\Sigma_{Z_i}^{-}\Sigma_{Z_i'})] = \mathbb{E}_{W_{t-1}} tr\left(\Sigma_{Z_i}^{-}\mathbb{E}_{z_i'}\left[\Sigma_{Zi} + \Delta'\right]\right) = d + \mathbb{E}_{W_{t-1}} tr\left(\Sigma_{Z_i}^{-}\Delta\right).
$$

and

$$
\mathbb{E}_{z_i'}\log|\Sigma_{Z_i'}| = \mathbb{E}_{z_i'}\log|\Sigma_{Zi} + \Delta'| \approx \log|\Sigma_{Zi}| + \mathbb{E}_{z_i'} tr(\Sigma_{Z_i}^{-}\Delta') = \log|\Sigma_{Zi}| + tr(\Sigma_{Z_i}^{-}\Delta)
$$

■

Lemma 9 (Expected Weights over History)

$$
\mathbb{E}_{W^{t-1}} KL[P(W_t|W^{t-1})||Q(W_t|W^{t-1})] = \mathbb{E}_{W_{t-1}} KL[P(W_t|W_{t-1})||Q(W_t|W_{t-1})] \tag{17}
$$

Proof

$$
\begin{aligned}
\underset{W^{t-1}}{\mathbb{E}} KL[P(W_t|W^{t-1})||Q(W_t|W^{t-1})] &= \mathbb{E}_{W^{t-1}} KL[P(W_t|W_{t-1})||Q(W_t|W_{t-1})] \\
&= \mathbb{E}_{W_{t-1}} KL[P(W_t|W_{t-1})||Q(W_t|W_{t-1})]
\end{aligned}
$$

This follows from the Markovian nature of the weights in vanilla SGD. ■

Lemma 10 (Conditional Weight & Gradient Distribution)

$$
P(W_t|W_{t-1}) = P(\eta_t \nabla L(b_t, W_t)|W_{t-1}). \tag{18}
$$

Proof Given the previous weights, the stochasticity in the next weights depend only on the stochasticity in the gradients (which depends on the sampling).

$$\begin{aligned} P(W_t = w_t | W_{t-1} = w_{t-1}) &= P(W_{t-1} - \eta_t \nabla L(b_t, W_{t-1}) = w_t | W_{t-1} = w_{t-1}) \\ &= P(\eta_t \nabla L(b_t, W_{t-1}) = w_{t-1} - w_t | W_{t-1} = w_{t-1}) \end{aligned} \tag{19}$$

■

Theorem 11 (MIP Bound) *As $N \to \infty$, for sufficiently large batch size B,*

$$\mathbb{E}_{z_i'} KL[Q_{Z_i} || Q_{\mathbf{Z_i}'}] \leq \sum_{t=1}^{T} \frac{1}{2} \mathbb{E}_{W_{t-1}} \left[\frac{1}{N^2} ||g_i - \mu||_{\Sigma^-_{Z_i}} + tr(\Sigma^-_{Z_i} \Sigma) \right]. \tag{20}$$

Proof We have that

$$\begin{aligned} \mathbb{E}_{z_i'} KL[Q_{Z_i} || Q_{\mathbf{Z_i}'}] &\leq \mathbb{E}_{z_i'} KL[P_{Z_i}(W^T) || P_{\mathbf{Z_i}'}(W^T)] \\ &= \mathbb{E}_{z_i'} \sum_{t=1}^{T} \mathbb{E}_{W^{t-1}} KL[P_{Z_i}(W_t | W^{t-1}) || P_{\mathbf{Z_i}'}(W_t | W^{t-1})] \\ &= \mathbb{E}_{z_i', W_{t-1}} KL[P_{Z_i}(\eta_t \nabla L(b_t, W_t) | W_{t-1}) || P_{\mathbf{Z_i}'}(\eta_t \nabla L(b_t, W_t) | W_{t-1})] \\ &= \mathbb{E}_{z_i', W_{t-1}} KL[P_{Z_i}(\nabla L(b_t, W_t) | W_{t-1}) || P_{\mathbf{Z_i}'}(\nabla L(b_t, W_t) | W_{t-1})] \\ &= \mathbb{E}_{z_i', W_{t-1}} KL[G^t_{Z_i} || G^t_{\mathbf{Z_i}'}] \end{aligned}$$

The first, second, and third steps come from the data processing inequality, lemma 9, and lemma 10 above. The fourth step is true because scaling does not effect the divergence. The final step comes from plugging in the divergence term from lemma 8 above. ■

Next, we provide the proofs from Section 3.

Lemma 12 (Per-Iteration Divergence) *The expected divergence between gradients at iteration t can be written as*

$$\mathbb{E}_{W_{t-1}} KL[G^t_Z || G^t_{\bar{Z}}] = \frac{1}{2} \mathbb{E}_{W_{t-1}} \left[tr(\Sigma^-_Z \Sigma_{\bar{Z}}) - d + \ln \frac{|\Sigma_Z|}{|\Sigma_{\bar{Z}}|} \right] + \mathbb{E}_{W_{t-1}} \left[(\mu_Z - \mu_{\bar{Z}})^T \Sigma^-_Z (\mu_Z - \mu_{\bar{Z}}) \right]$$

Proof This follows the same derivation as lemma 8 above. ■

Theorem 13 (Data-Dependent Prior Divergence)

$$\begin{aligned} KL[P(W|Z) || P^*(W)] \leq \sum_{t=1}^{T} \frac{1}{2} \mathbb{E}_{\bar{Z}, W_{t-1}} &\left[tr(\Sigma^-_Z \Sigma_{\bar{Z}}) - d + \ln \frac{|\Sigma_Z|}{|\Sigma_{\bar{Z}}|} \right] \\ &+ \mathbb{E}_{W_{t-1}} ||\mu_Z - \mathbb{E}_{\bar{Z}} \mu_{\bar{Z}}||_{\Sigma^-_Z} \end{aligned}$$

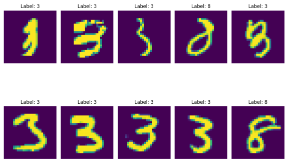

Figure 1: The five highest-scoring (top row) and lowest-scoring (bottom row) images in the MNIST dataset according to MIP scores $\phi(z)$. A two-layer neural network was trained for binary classification on the classes 3 and 8.

Proof First, using convexity, the data-processing inequality, and the chain rule, we get that

$$\begin{aligned}
KL[P(W|Z)||P^*(W)] &\leq \mathbb{E}_{\bar{Z}} KL[P(W|Z)||P(W|\bar{Z})] \\
&\leq \mathbb{E}_{\bar{Z}} KL[P((W_1, \dots, W_T)|Z)||P((W_0, \dots, W_T)|\bar{Z})] \\
&= \mathbb{E}_{\bar{Z}} \left[\sum_{t=1}^{T} \mathbb{E}_{W_{t-1}} KL[P(W_t|W_{t-1}, Z)||P(W_t|W_{t-1}, \bar{Z})] \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\bar{Z}.W_{t-1}} KL[G^t_Z||G^t_{\bar{Z}}]
\end{aligned}$$

We then plug in the lemma above and absorb the expectation over $\bar{Z}$:

$$\begin{aligned}
\mathbb{E}_{\bar{Z}, W_{t-1}} KL[G^t_Z||G^t_{\bar{Z}}] &= \frac{1}{2}\mathbb{E}_{\bar{Z}, W_{t-1}} \left[tr(\Sigma^-_Z \Sigma_{\bar{Z}}) - d + \ln \frac{|\Sigma_Z|}{|\Sigma_{\bar{Z}}|} \right] + \frac{1}{2}\mathbb{E}_{\bar{Z}, W_{t-1}} \left[(\mu_Z - \mu_{\bar{Z}})^T \Sigma^-_Z (\mu_Z - \mu_{\bar{Z}}) \right] \\
&= \frac{1}{2}\mathbb{E}_{W_{t-1}} \left[2tr(\Sigma^-_Z \Sigma) - d + \ln \frac{|\Sigma_Z|}{|\Sigma|} \right] + \frac{1}{2}\mathbb{E}_{W_{t-1}} [(\mu_Z - \mu)^T \Sigma^-_Z (\mu_Z - \mu)] \\
&= \frac{1}{2}\mathbb{E}_{W_{t-1}} \left[tr(\Sigma^-_Z \Sigma) \right] + \frac{1}{2}\mathbb{E}_{W_{t-1}} || \sum_{i=1}^{N} \frac{1}{N} g_i - \sum_{i=1}^{N} \frac{1}{N} \mu ||^2_{\Sigma^-_Z} \\
&\leq \frac{1}{2} \left[\mathbb{E}_{W_{t-1}} tr(\Sigma^-_Z \Sigma) + \sum_{i=1}^{N} \frac{1}{N^2} \mathbb{E}_{W_{t-1}} ||g_i - \mu||^2_{\Sigma^-_Z} \right] \\
&= \sum_{j=1, j \neq i}^{N} \frac{1}{N^2} \mathbb{E}_{W_{t-1}} ||g_i - \mu|| + \Psi(z_i).
\end{aligned}$$

A.1. Preliminary Results

Because the covariance matrix of the gradients is large and low rank, we approximate the distance $||\Sigma^-_{Z_i}(g_i - \mu)||$ using the inner product matrix of the gradients (with dimension $B << d$) and perform this computation layer-wise and sum the scores across each layer. We use this to compute the MIP scores when training a two-layer network on a binary classification task using MNIST.