Parameters or Privacy: A Provable Tradeoff Between Overparameterization and Membership Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

A surprising phenomenon in modern machine learning is the ability of a highly *overparameterized* model to generalize well (small error on the test data) even when it is trained to memorize the training data (zero error on the training data). This has led to an arms race towards increasingly overparameterized models (c.f., deep learning). In this paper, we study an underexplored hidden cost of overparameterization: the fact that overparameterized models are more vulnerable to *privacy attacks*, in particular the *membership inference* attack that predicts the (potentially sensitive) examples used to train a model. We significantly extend the relatively few empirical results on this problem by theoretically proving for an overparameterized linear regression model that the membership inference vulnerability increases with the number of parameters. Moreover, a range of empirical studies indicates that more complex, nonlinear models exhibit the same behavior. Finally, we study different methods for mitigating such attacks in the overparameterized regime, such as noise addition and regularization, and conclude that simply reducing the parameters of an overparameterized model is an effective strategy to protect it from membership inference without greatly decreasing its generalization error.

1 Introduction

As more and more machine learning models are trained on sensitive user data (e.g., customer behavior, medical records, personal media), there is a growing concern that these models may serve as a gateway for malicious adversaries to access their private training data [1, 2]. For example, the possibility of performing *membership inference* (MI), which is the task of identifying whether a specific data point was included in a model's training set or not, is greatly detrimental to user privacy. In settings like healthcare, knowledge of mere inclusion in a training dataset (e.g., hospital visitation records) already reveals sensitive information. Moreover, MI can enable the extraction of verbatim training data from a released model [1].

The general motivation for this paper is the recent realization that privacy issues around machine learning might be exacerbated by today's trend towards increasingly *overparameterized models* that have more parameters than training data points and so can be trained to memorize (attain zero error on) the training data. Surprisingly, some overparameterized models (e.g., large regression models [3], massive deep networks [4, 5]) generalize extremely well [6, 7]. Limited empirical evidence suggests that overparameterization and memorization may lead to greater privacy vulnerabilities [8–10]. However, there has been little to no analytical work on this important problem.

In this paper, we take the first steps towards an analytical understanding of how the number of parameters in a machine learning model relates to MI vulnerability. In a theoretical direction, we prove for linear regression on Gaussian data in the overparameterized regime that increasing the number of parameters of the model increases its vulnerability to MI (Theorem 3.2). In a supporting

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

empirical direction, we demonstrate that the same behavior holds for a range of more complex models: a latent space model, a time-series model, and a nonlinear random ReLU features model (Section 5). Furthermore, we study ridge-regularized linear regression and similarly prove that that MI vulnerability increases with the number of parameters. Surprisingly, however, we also demonstrate analytically that additional regularization *increases* this vulnerability in overparameterized linear models. Hence, ridge-regularization is not an effective defense against MI and can even be harmful.

Our analyses indicate that reducing the number of parameters of a highly overparameterized model is a simple yet effective strategy to protect a model against MI. We analytically prove that this strategy provides a better generalization vs. privacy tradeoff than the popular technique of noise addition (Section 4). As the trend towards increasingly overparameterized machine learning models accelerates, *our results make the case for less overparameterization when privacy is a concern.*

Related work. The problem of membership inference has received considerable attention, and we refer the reader to [11] for a survey of the prior art. Many works have demonstrated how deep neural networks can be highly vulnerable to MI attacks (cf., [12–15] for example). Various types of attacks, such as shadow models [12], confidence-based attacks [16, 17], and label-only attacks [18] have risen to further expose privacy vulnerability of machine learning models. While we study the effect parameterization has on MI vulnerability, several studies examine how other aspects of the machine learning pipeline, such as data augmentation [19], dropout [14], sparsity [20], and ensembles [21] affect MI vulnerability. [15] show that pruning a neural network can defend against MI, lending further experimental support to the link between overparameterization and membership inference. [9] analyzes the relationship between overfitting (as measured by generalization error) and membership inference but does not connect this to the number of parameters. However, much recent work has shown that overparameterization does not always lead to increased generalization error and indeed sometimes can even decrease it [22, 6, 3, 7], suggesting that more work is needed linking overparameterization and membership inference beyond mere generalization error. Despite overwhelming empirical evidence that overparameterized models are vulnerable to MI attacks, to the best of our knowledge, there has been no theoretical study of the effect of this connection.

MI can be seen as a special case of differential privacy (DP) [9]. We refer the reader to [23] for a survey of DP results. [24] bound the minimum dataset size as a function of dimension to ensure privacy, however their results do not extend to data of continuous values (e.g., regression data) as is done in this work and does not account for the randomness in the data generating process, which as argued in [13], is a necessary condition to bound the optimal MI risk.

Our analysis is based on optimal membership inference attacks, a framework also employed by other studies of membership inference [9, 13, 10]. While our theoretical analysis focuses on linear regression, many works have studied overparameterized linear models as an interpretable and tractable test case of more complex overparameterized models such as deep networks [25, 26, 3, 6, 22], suggesting links of our results to more complex nonlinear models.

Contributions. In this paper, we establish the first theoretical connection between overparameterization and vulnerability to membership inference attacks. Our contributions are as follows:

1. We derive the optimal MI accuracy for linear regression on Gaussian data and show that increasing the number of parameters increases the vulnerability to MI.
2. We empirically show for more complex models that increasing the number of parameters also increases MI vulnerability.
3. We theoretically show for overparameterized linear models that increased ridge regularization increases MI vulnerability.
4. We compare reducing the number of parameters vs. noise addition for mitigating MI and show that parameter reduction generalizes better given the same level of MI vulnerability.

2 The Membership Inference Problem

We now introduce our notation and formally state the membership inference (MI) problem. Let $D \in \mathbb{Z}^+$ denote the data dimension. Let $z = (\boldsymbol{x}, y) \in \mathbb{R}^D \times \mathbb{R}$ denote a data point, and let $\mathcal{D}$ be a distribution over the data points. Consider a set $\mathcal{S}$ sampled $\mathcal{S} \sim \mathcal{D}^n$ of $n \in \mathbb{Z}^+$ data points and denote it by $\mathcal{S} = \{z_1, z_2, ..., z_n\} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)\}$. Let f_S represent a machine learning model obtained by applying a training algorithm T on the dataset S. In particular, f_S is a deterministic

function in $\mathbb{R}^D \to \mathbb{R}$. Typically, it is a function that minimizes $\sum_{i=1}^{n} \ell(f_S(x_i), y_i)$ for some loss function $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$. We denote $\widehat{y}_i = f_S(x_i)$.

Given a data point z_0 and a trained model f_S, *membership inference* is the task of identifying if z_0 was included in the set S used to train f_S. Different MI attacks are characterized by the types of information accorded to the adversary. In this work, we focus on a blind black-box setting where the adversary has access to $\boldsymbol{x}_0$, $f_S(\boldsymbol{x}_0)$, n, the training algorithm T, and the data distribution $\mathcal{D}$, but not to the ground truth y_0, the loss value $\ell(f_S(\boldsymbol{x}_0), y_0)$, or any learned parameters of f_S. Thus, the adversary is a function $A : \mathbb{R}^d \times \mathbb{R} \to \{0, 1\}$ such that given $(\boldsymbol{x}_0, f_S(\boldsymbol{x}_0))$, it outputs 0 or 1, representing its prediction as to whether $\boldsymbol{x}_0$ is a member of S based on $f_S(\boldsymbol{x}_0)$.

There are two main interests in studying blind MI wherein the adversary does not have access to the loss value. First, there are many realistic scenarios where the ground truth variable of interest is unknown for the general population, and hence the need for a model f_S to predict it. Second, since we study the interpolation regime where the training loss can be driven down to exactly 0, MI becomes trivial given the loss value: an adversary that predicts any data point that achieves 0 loss as a member will have perfect accuracy. Multiple works in the literature also study this blind setting [17, 27]. We emphasize that our results on the blind adversary lower bound the performance of adversaries that additionally have access to the true labels.

MI is often defined as an experiment to facilitate precise analysis, and we follow the setup of [9] summarized below.

Experiment 1. **Membership inference experiment.** Given a data distribution $\mathcal{D}$, an integer n, a machine learning training algorithm f_S, and an adversary A, the MI experiment is as follows:

1. Sample $\mathcal{S} \sim \mathcal{D}^n$.
2. Apply the training algorithm on $\mathcal{S}$ to obtain $f_{\mathcal{S}}$.
3. Sample $m \in \{0, 1\}$ uniformly at random.
4. If $m = 0$, sample a data point $(\boldsymbol{x}_0, y_0) \sim \mathcal{D}$. Else, sample a data point $(\boldsymbol{x}_0, y_0) \in \mathcal{S}$ uniformly at random.
5. Observe the adversary's prediction $A(x_0, f_S(x_0)) \in \{0, 1\}$.

We wish to quantify the optimal performance of the adversary A and turn to the popular *membership advantage* metric defined in [9]. The membership advantage of A is the difference between the true positive rate and false positive rate of its predictions.

Definition 2.1 ([9])**.** The *membership advantage* of an adversary A is:

$$\text{Adv}(A) := \mathbb{P}(A((\boldsymbol{x}_0, f_S(\boldsymbol{x}_0))) = 1 \mid m = 1) - \mathbb{P}(A((\boldsymbol{x}_0, f_S(\boldsymbol{x}_0))) = 1 \mid m = 0),$$

where $\mathbb{P}(\cdot)$ is taken jointly over all randomness in Exp. 1.

Membership advantage may equivalently be written as $\text{Adv}(A) = 2\mathbb{P}(A((\boldsymbol{x}_0, f_S(\boldsymbol{x}_0)) = m) - 1$, where $\mathbb{P}(A((\boldsymbol{x}_0, f_S(\boldsymbol{x}_0)) = m)$ is the probability that the adversary is correct.

3 Theoretical Results

3.1 Optimal Membership Inference Via Hypothesis Testing

In this paper, rather than studying current MI attacks, we study the optimal MI adversary: the adversary that maximizes membership advantage. As such, our analysis and results are not restricted to the current known attack strategies, which are constantly evolving, and instead serve as upper bounds for the performance of any MI attack, now or in the future. The following proposition supplies the theoretically optimal MI adversary.

Proposition 3.1. *The adversary that maximizes membership advantage is:*

$$A^*(\boldsymbol{x}_0, f_S(\boldsymbol{x}_0)) = \begin{cases} 1 & \text{if } P(\widehat{y}_0 \mid m = 1, \boldsymbol{x}_0) > P(\widehat{y}_0 \mid m = 0, \boldsymbol{x}_0), \\ 0 & \text{otherwise}, \end{cases}$$

where $\widehat{y}_0 = f_S(\boldsymbol{x}_0)$ and P denotes the distribution function for $\widehat{y}_0$ over the randomness in the membership inference experiment conditioned on $\boldsymbol{x}_0$.

As observed by [28, 10], the optimal adversary performs a hypothesis test with respect to the posterior probabilities, with the two hypotheses being:

$$H_0 : \mathcal{S} \sim \mathcal{D}^n, (\boldsymbol{x}_0, y_0) \sim \mathcal{D} \quad \text{and} \quad H_1 : \mathcal{S} \sim \mathcal{D}^n, (\boldsymbol{x}_0, y_0) \sim \mathcal{S},$$

so that maximizing membership advantage is achieved by performing the a likelihood ratio test.

3.2 Linear Regression with Gaussian Data

We begin our study of overparameterization's effect on MI with linear regression with Gaussian data. We find that in the sufficiently overparameterized regime, increasing the number of parameters increases vulnerability to MI. We denote by n, p, D the number of data points, number of model parameters, and data dimensionality, respectively. Let $n, p, D \in \mathbb{Z}^+$ be given such that $p \leq D$, and let $\sigma > 0$ be given. Consider a D-dimensional random vector $\beta \sim \mathcal{N}\left(0, \frac{1}{D}\boldsymbol{I}_D\right)$, representing the true coefficients. We consider data points $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $y_i = \boldsymbol{x}_i^\top \beta + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Sampling n data points, we denote by $\boldsymbol{X}$ the $n \times D$ matrix whose i^{th} row is $x_i^\top$ and by $\boldsymbol{y}$ the n-dimensional vector of elements y_i. Further, let $\boldsymbol{X}_p$ be the $n \times p$ matrix containing the first p columns of $\boldsymbol{X}$. Least squares linear regression finds the minimum-norm p-dimensional vector $\hat{\beta}$ that minimizes $\|\boldsymbol{y} - \boldsymbol{X}_p\beta\|_2^2$, which is $\hat{\beta} = \boldsymbol{X}_p^\dagger \boldsymbol{y}$ where $\boldsymbol{X}_p^\dagger$ denotes the Moore-Penrose inverse of $\boldsymbol{X}_p$. Then, for a given feature vector $\boldsymbol{x}_0$, the model predicts $\hat{y}_0 = \boldsymbol{x}_{0,p}^\top \hat{\beta}$, where $\boldsymbol{x}_{0,p}$ is the vector containing the first p elements of $\boldsymbol{x}_0$. In this setting, we can derive the membership advantage of the optimal adversary as a function of the number of parameters p used by the model.

Theorem 3.2. *Let $n, p, D \in \mathbb{Z}^+$ be given such that $p \leq D$. Let $\boldsymbol{x}_0$ be a given D-dimensional vector. Let $\beta \sim \mathcal{N}(0, \frac{1}{D}\boldsymbol{I}_D), \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_D), \epsilon_i \sim \mathcal{N}(0, \sigma^2)$. If $m = 1$, let $\boldsymbol{x}_k = \boldsymbol{x}_0$ for k chosen uniformly at random from $1, 2, ..., n$. Let $y_i = \boldsymbol{x}_i^\top \beta + \epsilon_i$. Let $\boldsymbol{X}$ be the $n \times D$ matrix whose rows are $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$. Then, as $n, p, D \to \infty$ such that $\frac{p}{n} \to \gamma \in (1, \infty)$, we have:*

$$\widehat{y_0} \mid m = 0, \boldsymbol{x}_0 \sim \mathcal{N}(0, \sigma_0^2) \quad \textit{and} \quad \widehat{y_0} \mid m = 1, \boldsymbol{x}_0 \sim \mathcal{N}(0, \sigma_1^2) \tag{1}$$

where

$$\sigma_0^2 := \left(\frac{n}{p}\right)\left(\frac{1}{D} + \frac{1 + \sigma^2 - \frac{p}{D}}{p - n - 1}\right)\|\boldsymbol{x}_{0,p}\|^2 \quad \textit{and} \quad \sigma_1^2 := \sigma^2 + \frac{\|\boldsymbol{x}_0\|^2}{D}. \tag{2}$$

Consider the case when $\sigma_1 > \sigma_0$. The resulting optimal membership inference algorithm A^* is

$$A^*(\boldsymbol{x}_0, \hat{y}_0) := \mathbb{1}\left[\hat{y_0}^2 > \alpha^2\right] \quad \textit{where} \quad \alpha := \sqrt{\frac{\sigma_0^2 \sigma_1^2 \log\left(\frac{\sigma_1^2}{\sigma_0^2}\right)}{\sigma_1^2 - \sigma_0^2}}. \tag{3}$$

with membership inference advantage Adv*:*

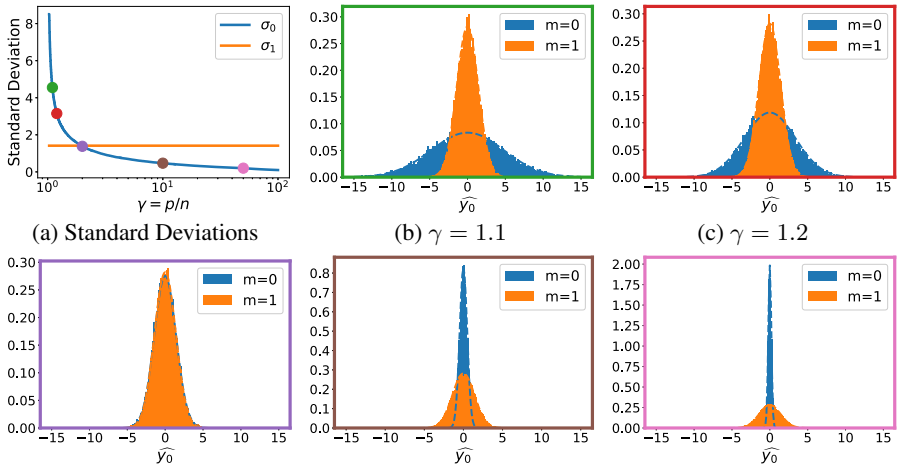
$$\text{Adv}(A^*) = \mathbb{E}_{\boldsymbol{x}_0}\left[2\left\{\Phi\left(\frac{\alpha}{\sigma_0}\right) - \Phi\left(\frac{\alpha}{\sigma_1}\right)\right\}\right], \tag{4}$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal, and $\alpha, \sigma_0, \sigma_1$ are conditioned on $\boldsymbol{x}_0$.

Remark 3.3. The case where $\sigma_1 < \sigma_0$ occurs when γ is small ($p \approx n$). The same membership advantage result holds in this, reversing the roles of σ_0 and σ_1 in $\text{Adv}(A^*)$ in Eq. (4).

Remark 3.4. The above result holds using the asymptotic distributions as $n, p, D \to \infty$. In Lemma B.1 in the Appendix, we derive the non-asymptotic distributions for the predictions of the minimum-norm least squares interpolator, though they cannot be written in closed form.

In the Theorem 3.2, Eq. (1) shows that the posterior distributions of the outputs for test points (when $m = 0$) and training points (when $m = 1$) are both 0-mean Gaussians but with variances σ_0^2 and σ_1^2 respectively given in Eq. (2). Recall from Prop. 3.1 that the optimal MI adversary is a likelihood ratio test (LRT) between the distribution of the model's output for new test points (when $m = 0$) and that for reused training points (when $m = 1$). This is reflected in Eq. (5) where we note that α is the standard sufficient statistic for an LRT between two 0-mean Gaussian distributions with deviations σ_0 and σ_1. Finally, we compute the membership advantage of this adversary in Eq. (4) by taking an expectation over the random draw of $\boldsymbol{x}_0$, as defined in Experiment 1, and noting that σ_0, σ_1, and α are all functions of $\boldsymbol{x}_0$. Membership advantage is defined to be a difference between the true and



(a) Standard Deviations (b) $\gamma = 1.1$ (c) $\gamma = 1.2$

(d) $\gamma = 2$ (e) $\gamma = 10$ (f) $\gamma = 50$

Figure 1: In the highly overparameterized regime, increasing the number of parameters yields more distinguishable posterior distributions. (a) Standard deviations of $\hat{y}_0$ for minimum-norm least squares as a function of parameterization for $\gamma > 1$. (b-f) The Gaussian distributions in Eq. (1) (broken lines), as well as empirical histograms of 20,000 $\hat{y}_0$ samples for different γ values. For these plots, $n = 400$, $D = 20,000$, $\sigma = 1$. The prediction variance for $m = 0$ decreases as parameterization increases, while the variance for $m = 1$ is constant.

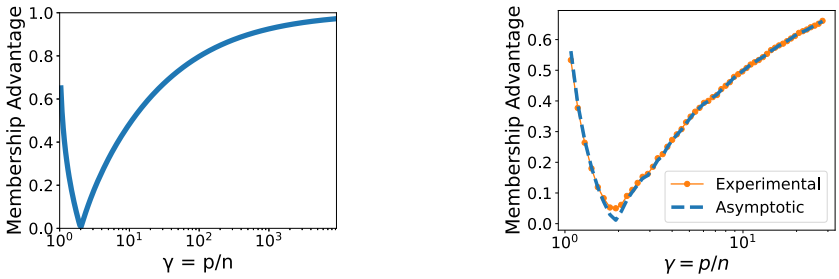
false positive probabilities in Defn. 2.1. This difference is given by $\Phi(\alpha/\sigma_0) - \Phi(\alpha/\sigma_1)$ leading to the expression in Eq. (4).

To understand the implication of the result for MI, first observe that when $m = 1$ and $\boldsymbol{x}_0$ is a training point that is memorized by the model, the variance of the model's output is equal to the variance of the measurement y_0 itself independent of p. On the contrary, when $m = 0$ and $\boldsymbol{x}_0$ is a test point, the variance of the model's output is *decreasing* with p (Figure 1a). Hence, though the output distribution means stay the same, an LRT can easily distinguish these two distributions. In the extreme case, suppose $p, D \gg n$, then $\sigma_0^2 \to 0$, while σ_1^2 remains a nonzero value. Hence $\mathrm{Adv}(A^*) \to 1$.

We confirm our derivations of the output distributions numerically in Figure 1 where we plot the predicted distributions from Eq. (1) to empirical histograms generated by repeatedly computing the minimum norm least squares solution over 20,000 independent trials[1]. We also plot the variances from Eq. (2) in Fig. 1a. Visually, given $f(\boldsymbol{x}_0)$ membership inference reduces to identifying whether it is more likely a sample from the blue distribution and or the orange, and the adversary simply predicts the more likely outcome. These distributions cross at $\pm\alpha$ so it is sufficient to compare $\hat{y}_0^2$ to α^2 to perform the LRT rather than compute the likelihoods exactly.

In Figure 2a, we plot the membership advantage for $\gamma > 1$ using Eq. (4). Since the difference between the variances of the model's output when $m = 0$ and $m = 1$ increases with γ for $\gamma > 2$ (cf., Fig. 1a), it becomes easier to distinguish the two distributions. This results in an increase in the membership advantage. The initial decrease in membership advantage when $\gamma \leq 2$ is a consequence of $\sigma_1 \leq \sigma_0$ in this regime as is shown in Fig. 1a. Since σ_0 is decreasing in p, initially, this decrease makes the train and test output distributions harder to distinguish leading to lower membership advantage and 0 advantage for $\gamma = 2$. However, for $\gamma > 2$, σ_0 decreases past σ_1 and membership advantage approaches 1 as $\gamma \to \infty$. In practice, the $\gamma < 2$ regime (only slightly overparameterized) is less interesting since models suffer larger generalization error in this setting (cf., Theorem 1 of [22]) and are rarely (if ever) used in practice. A key takeaway is that extreme overparameterization increases the vulnerability of a machine learning model to MI. While Theorem 3.2 operates in the asymptotic regime, we empirically approximate membership advantage for $n = 100$ and $D = 10,000$ by approximating the posterior distributions with histograms of samples. As we see in Figure 2b, the asymptotic formula agrees very closely with the non-asymptotic experiment.

[1] We detail additional experimental details and the computational hardware in Appendix E.



(a) Asymptotic membership advantage

(b) Empirical verification for non-asymptotic case

Figure 2: Increasing the number of parameters increases membership advantage. (a) Theoretical membership advantage for linear regression with Gaussian data (Eq. 4) as a function of the number of parameters for 100 sampled $\boldsymbol{x}_0$'s, $n = 10^3$, $D = 10^7$. (b) We empirically approximate membership advantage by estimating the two posterior distributions as empirical histograms with 200,000 samples each for $n = 100$, $D = 10,000$, and $\sigma = 1$. We plot it alongside the theoretical asymptotic membership advantage, showing a close agreement between the two.

4 Mitigating Membership Inference Attacks

In the previous section, we established that as models become increasingly overparameterized, it is easier for the attacker to perform membership inference. It may seem reasonable to regularize to guard against this phenomenon. We begin with a negative result showing that Ridge regularization actually *increases* membership advantage. We then present successful techniques in Section 4.2.

4.1 Ridge-Regularized Linear Regression

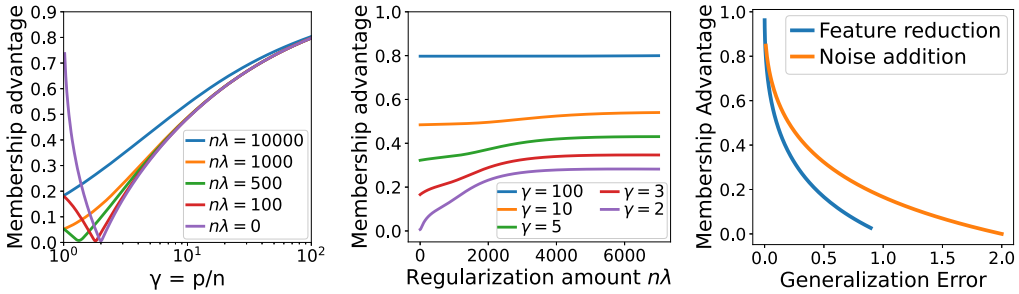
In this section, we analyze membership inference in *ridge-regularized* linear regression in the same setting as Section 3.2 except that now $\hat{\beta}_\lambda = (\boldsymbol{X}_p^\top \boldsymbol{X}_p + n\lambda \boldsymbol{I}_p)^\dagger \boldsymbol{X}_p^\top y$, where λ is a regularization parameter. Larger values of λ yield greater regularization, and $\lambda \to 0$ reduces to the case of Section 3.2. Regularization is a common method to reduce overfitting, and one might think that it would reduce model's vulnerability to membership inference attacks in this setting.

A surprising observation resulting from this analysis is that, in the highly overparameterized region ($\gamma \gg 1$), ridge regularization actually *increases* the model's vulnerability to MI attacks for linear regression with Gaussian data. In Theorem D.1, presented in the appendix, we derive an analogous result for the distribution of the ridge regularized estimate as Theorem 3.2 demonstrated for the unregularized case. In particular, if A_λ^* is the adversary for the λ-regularized problem, its membership advantage $\mathrm{Adv}(A_\lambda^*)$ is an increasing function in λ when $\gamma \gg 1$. To visualize this, we plot the theoretical membership advantages for 100 sampled $\boldsymbol{x}_0$'s for differing regularization strengths in Figure 3. In Figure 3a, we plot membership advantage as a function of the overparameterization ratio γ for a few $n\lambda$ values. Conversely, In Figure 3b we plot membership advantage versus regularization $n\lambda$ for a few different values of γ. In particular, note how in both subplots, increasing λ never decreases the membership advantage. This observation has been made empirically (but not analytically) for techniques that have similar regularizing effects, such as dropout and ensembling [14, 21].

Intuitively, the reason ridge regularization increases MI vulnerability is because while it decreases both the variances of the model's outputs on training and test points, it decreases that of training points at a faster rate. Thus, it actually makes the output distributions more distinguishable from each other. In Figure 5 in the appendix, we plot the gap between the standard deviations for the $m = 0$ and $m = 1$ case to visualize this effect.

4.2 Improving Privacy: Noise Addition vs. Feature Reduction

Feature reduction: A consequence of Thm. 3.2 is that one can reduce vulnerability to MI attacks by simply *decreasing* the number of parameters. This is similar in spirit to [15] who show that pruning neural networks improves robustness to MI attacks. However, as has been shown in a variety of works (e.g., [6, 22, 7]) *increasing* the number of parameters leads to a "double descent" behavior where



(a) Ridge (vs. parameters) (b) Ridge (vs. regularization) (c) MA vs. accuracy

Figure 3: Different methods to try to reduce membership advantage (MA). (a) MA vs. overparamterization γ for a few ridge-regularization strengths λ. Stronger ridge regularization harms privacy. (b) MA vs. λ for a few γ's, again showing how ridge regularization increases MA. (c) Trade-off between MA and mean squared error for noise addition and feature/dimension reduction. Feature reduction produces a better trade-off.

generalization performance improves. This is captured in the following proposition where we define generalization error as the mean squared error for random $\boldsymbol{x}_0$ sampled from the data distribution.

Proposition 4.1. *In the same setting as Theorem 3.2, the generalization error is given by*

$$\mathbb{E}[(y - \hat{\beta}^\top \boldsymbol{x}_0)^2] = \left(\frac{n}{p}\right)\left(\frac{1}{D} + \frac{1 + \sigma^2 - \frac{p}{D}}{p - n - 1}\right) \mathbb{E}\left[\|\boldsymbol{x}_{0,p}\|^2\right].$$

The proposition follows immediately as a consequence of Theorem 3.2. Since the error decreases as p grows (since $p/D \leq 1$) the choice of p induces *a direct trade-off in the overparameterized regime between a linear model's generalization error and the membership advantage it suffers.*

Noise addition: An alternative and popular method to increase a model's privacy is noise addition [23, 29], performed by perturbing the model output with independent noise before releasing it: $\widehat{y_0} = \boldsymbol{x}_{0,p}^\top \hat{\beta} + \bar{\epsilon}$ where $\bar{\epsilon} \sim \mathcal{N}(0, \bar{\sigma}^2)$. Recall that MI is possible for overparameterized linear regression models because the variance of the output of test points $(m = 0)$ is lower than the variance of the output for training points $(m = 1)$ (Fig. 2b). Thus, intuitively, if the model adds some random noise to its output when a training point $\boldsymbol{x}_0$ is queried, then it can make these output distributions harder to distinguish. However, adding noise decreases the model's accuracy, inducing a different generalization error vs. membership advantage trade-off. In the following lemma, we compute the membership advantage and generalization error of noise addition using the full set of features $p = D$.

Lemma 4.2. *Consider the setup of Theorem 3.2 restricted to $p = D$, $\sigma_1 > \sigma_0$, and $\gamma > 1$, suppose independent 0-mean, variance $\bar{\sigma}^2$ Gaussian noise is added to the outputs when $m = 0$ and $\boldsymbol{x}_0$ is not a member of the training set. Then $y \mid m = 0, \boldsymbol{x}_0 \sim \mathcal{N}(0, \sigma_0^2 + \bar{\sigma}^2)$. The optimal adversary is*

$$A^*_{\bar{\sigma}}(\boldsymbol{x}_0, \hat{y}_0) := \mathbb{1}\left[\hat{y_0}^2 > \alpha^2_{\bar{\sigma}}\right] \quad \textit{where} \quad \alpha_{\bar{\sigma}} := \sqrt{\frac{(\sigma_0^2 + \bar{\sigma}^2)\sigma_1^2 \log\left(\frac{\sigma_1^2}{(\sigma_0^2 + \bar{\sigma}^2)}\right)}{\sigma_1^2 - (\sigma_0^2 + \bar{\sigma}^2)}}.$$

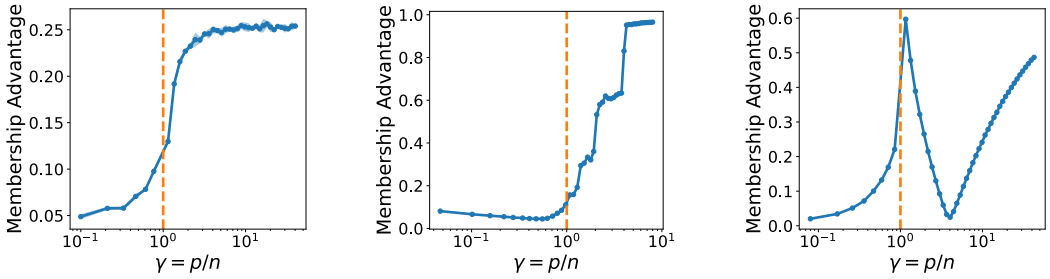
Its membership inference advantage is

$$Adv(A^*_{\bar{\sigma}}) = \mathbb{E}_{\boldsymbol{x}_0}\left[2\left\{\Phi\left(\frac{\alpha_{\bar{\sigma}}}{\sqrt{\sigma_0^2 + \bar{\sigma}^2}}\right) - \Phi\left(\frac{\alpha_{\bar{\sigma}}}{\sigma_1}\right)\right\}\right].$$

Additionally, the generalization error incurred is

$$\mathbb{E}[(y - \hat{\beta}^\top \boldsymbol{x}_0)^2] = \frac{n}{D}\left(\frac{1}{D} + \frac{\sigma^2}{D - n - 1}\right) \mathbb{E}\left[\|\boldsymbol{x}_0\|^2\right] + \bar{\sigma}^2.$$

The terms in Lemma 4.2 are similar to those in Theorem 3.2 except for the fact that they also depend on the variance of the additive noise $\bar{\sigma}^2$ in the case that $m = 0$ ($\boldsymbol{x}_0$ is not a member) and are denoted



(a) Latent space model (b) Time-series data (c) Random ReLU features

Figure 4: Empirically measured membership advantage vs. parameterization for various data models. The dashed line at $\gamma = 1$ divides the underparameterized and overparameterized regions. Shaded regions (typically too small to see) depict standard deviations over multiple experiments.

with a subscripted $\bar{\sigma}^2$ to reflect this. As expected, the generalization error increases with the variance of the added noise due to the $+\bar{\sigma}^2$ term. Furthermore, it is easily verified numerically that the membership inference advantage decreases with the variance of the added noise. When $\bar{\sigma}^2 = 0$, the variance on test points σ_0^2 is much less than the variance for training points σ_1^2. Indeed, in the limit, it is possible to add sufficient noise such that $\sigma_1^2 = \sigma_0^2 + \bar{\sigma}^2$, rendering the membership advantage 0, though possibly at the cost of impermissible generalization performance.

While both feature reduction and noise addition yield a privacy–accuracy trade-off, we observe that feature reduction provides a more desirable trade-off. In Figure 3c, we plot curves depicting the trade-off between achieving low generalization error (high accuracy) shown on the x-axis and low membership advantage shown on the y-axis for both parameter reduction and noise addition. The bottom left in the plot represents the nexus of perfect generalization and no membership advantage. Hence, lower and further to the left in the figure represents better performance. To generate both curves, we repeat the setting of Fig. 2a except that γ is no smaller than 30 to ensure that $\sigma_1 > \sigma_0$ and that the model is highly overparameterized. For the blue curve, we directly change how overparameterized the model is (γ) and plot the resulting membership advantage vs. generalization error using Thm 3.2 and Prop 4.1. The orange curve depicts the performance of noise addition to control membership inference. In this setting, the model uses all available features and instead perturbs the output with noise. In particular, lower values of generalization error on the left side of the plot are achieved by adding very small amounts of noise. By contrast, low levels of membership advantage on the bottom right side of the figure are achieved by adding larger amounts of noise. *Notably, the blue curve is below the orange at all points. Hence, noise addition is less effective for controlling membership inference advantage than simply using a smaller model.*

5 More Complex Models

In this section, we empirically show that the trend of more overparameterization leading to increased MI vulnerability exists even for more complex data models beyond linear regression on Gaussian data. In these experiments, we measure membership advantage conditioned on a single sampled $\boldsymbol{x}_0$ as a function of the number of parameters, and so we begin each experiment by sampling a $\boldsymbol{x}_0$ vector. Then we sample a training dataset $\boldsymbol{X}$, measurements y, as well as other random elements according to the data model being investigated. To obtain an $m = 0$ prediction, we learn a model on $\boldsymbol{X}$ that we then apply on $\boldsymbol{x}_0$. To obtain an $m = 1$ prediction, we first replace a row of $\boldsymbol{X}$ with $\boldsymbol{x}_0$ before learning the model and applying it to $\boldsymbol{x}_0$. We keep $\boldsymbol{x}_0$ fixed throughout the experiment, but resample all other random data (such as $\boldsymbol{X}$) many times to collect a large set of $m = 0$ and $m = 1$ prediction samples. We build a histogram of these samples by assigning them into fine discrete bins to obtain approximations $\hat{p}(\widehat{y_0} \mid m = 1, \boldsymbol{x}_0)$ and $\hat{p}(\widehat{y_0} \mid m = 0, \boldsymbol{x}_0)$ of the conditional densities needed for the optimal adversary (cf., Prop 3.1). To approximate membership advantage, we sum up the differences $\hat{p}(\widehat{y_0} \mid m_1, \boldsymbol{x}_0) - \hat{p}(\widehat{y_0} \mid m_0, \boldsymbol{x}_0)$ over all bins where $\hat{p}(\widehat{y_0} \mid m_1, \boldsymbol{x}_0) > \hat{p}(\widehat{y_0} \mid m_0, \boldsymbol{x}_0)$. Finally, we report both the means and standard deviations (as shaded error regions) of membership advantage across 5 repetitions of this experiment.

Linear Regression on Latent Space Model: We first consider the latent space model from [3], where the output variable y_i is a noisy linear function of a data point's d latent features in $\boldsymbol{z}_i$, but one

only observes a vector $\boldsymbol{x}_i$ containing $p \geq d$ covariates rather than the direct features $\boldsymbol{z}_i$. Let $\boldsymbol{Z}$ be an $n \times d$ matrix where each row is the vector of latent features for an observation. We then have:

$$\boldsymbol{y} = \boldsymbol{Z}\beta + \epsilon, \qquad \boldsymbol{X}_{i,j} = \boldsymbol{w}_j^\top \boldsymbol{z}_i + u_{i,j}, \qquad \widehat{y}_0 = \boldsymbol{x}_0^\top \boldsymbol{X}_P^\dagger y,$$

where $\boldsymbol{w}_j$ is a d-dimensional vector, and $u_{i,j}$ is a noise term. In this experiment, we set $n = 200$, $d = 20$, and vary p. We sample a single $\boldsymbol{x}_0 \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and a single set of $\boldsymbol{w}_j$ vectors, each from $\mathcal{N}(0, \boldsymbol{I}_d)$, for the whole experiment. We also let $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\beta \sim \mathcal{N}\left(0, \frac{1}{d}\boldsymbol{I}_d\right)$ and $u_{i,j} \sim \mathcal{N}(0, 1)$. We plot the empirical membership advantage in Fig. 4a. Similar to our theoretical results in Section 3.2, membership advantage here increases with the degree of overparameterization.

Noise-Free Time-Series Regression: In this experiment, we consider a model that aims to interpolate a time-series signal using frequency components as the features. For example, consider a patient who visits a hospital at irregular times t_i to get their blood glucose level measured. After obtaining a number of measurements taken over time, the hospital fits a time-series signal representing the patient's blood glucose level at any time. Using this learned model, an adversary wishes to identify whether the patient visited the hospital at a particular time t_0, that is, if t_0 is one such time point included in the hospital's training dataset for learning the patient's blood glucose level.

To formalize this, we fix $D = 1024$ as the number of frequency components each signal contains and $M = 2D - 1$ be the number of time samples for each signal. Let $\boldsymbol{W}$ be the matrix whose elements $\boldsymbol{W}_{kl} = \frac{1}{\sqrt{M+2}} \cos\left(\frac{2\pi kl}{M}\right)$. That is, each column of the matrix is half a period of an M-dimensional discrete cosine with frequency $\frac{2\pi k}{M}$. Sample a $\beta \sim \mathcal{N}\left(0, \frac{1}{D}\boldsymbol{I}_D\right)$, and let $\boldsymbol{z} = \boldsymbol{W}\beta$ denote the true length-D signal. Thus, the signal is a random linear combination of cosines. We randomly select $n = 128$ indices from $1, 2, \ldots, D$ uniformly without replacement, and let $\boldsymbol{X}$ be the $n \times D$ matrix whose rows are the rows of $\boldsymbol{W}$ at these selected indices. Then, $y = \boldsymbol{X}\beta$ is the signal observed at the randomly selected n indices. The regressor learns $\hat{\beta} = \boldsymbol{X}_p^\dagger \boldsymbol{y}$ and then predicts the signal at any other time point t_0 as $\widehat{y_0} = \boldsymbol{x}_0^\top \hat{\beta}$, where $\boldsymbol{x}_0$ is row t_0 of $\boldsymbol{W}$. Thus, identifying whether t_0 was a time point in the training dataset is equivalent to identifying if $\boldsymbol{x}_0$ was in $\boldsymbol{X}$.

We plot in Figure 4b the membership advantage of this task for a single random point t_0 for different values of p, the number of frequency components the regressor decides to use in learning the signal. As shown in the plot, we again observe the same phenomenon that, as the number of parameters p increases, the model becomes more vulnerable to membership inference.

Random ReLU Features: We next consider a nonlinear data model based on Random ReLU feature networks [30, 31]. Let $\boldsymbol{Z}$ be a random $n \times D$ matrix whose elements are iid standard normal. Let $\boldsymbol{V}$ be a random $D \times p$ matrix whose rows are sampled iid from the surface of the unit sphere in $\mathbb{R}^p$. Let $\boldsymbol{X} = \max(\boldsymbol{Z}\boldsymbol{V})$, where the $\max$ is taken elementwise. The target variables are given by $\boldsymbol{y} = \boldsymbol{Z}\beta + \epsilon$, where $\beta \sim \mathcal{N}\left(0, \frac{1}{D}\boldsymbol{I}_D\right)$ and $\epsilon \sim \mathcal{N}(0, \boldsymbol{I}_n)$. Finally, for the data point $\boldsymbol{x}_0$, let its prediction be $\widehat{y_0} = \boldsymbol{x}_0^\top \boldsymbol{X}^\dagger y$. We plot the membership advantage in Figure 4c with $n = 100$, $D = 5000$, and $\sigma = 1$. We again observe that in the highly overparametrized regime, membership advantage increases with parameters.

Key takeaways from this work: We have studied theoretically optimal membership inference attacks given by a likelihood ratio test (LRT). These attacks are very similar to the popular family of loss threshold attacks [13] or can be estimated directly via shadow models [12]. Since the LRT adapts to the scale of the outputs, we observe that ridge regularization does not reduce membership advantage (MA). While it is possible to add noise to reduce MA, we see that for the same MA, simply reducing the number of parameters achieves lower generalization error (higher accuracy) and is therefore preferable. We show experimentally that the same phenomena holds for a variety of data models. Though we study regression, an interesting avenue of future work may be to analyze MI in classification by employing techniques from [32, 33] to compute the output distribution of high dimensional linear classifiers. Additionally, while many works demonstrate that neural networks are vulnerable to MI attacks [12, 16–19, 14, 20, 21], it may be interesting to study this more directly as we have done for other models in Fig. 4. As a first step towards this, we direct the interested reader to Section 5 of [15] who show that by varying the amount of pruning, they can alter the membership advantage for networks trained on ImageNet, Cifar-100, and others. Finally, explicitly studying trade-offs between differential privacy more generally versus generalization error is an interesting avenue of future work.

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*, pages 2633–2650, 2021.

[2] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys*, 54(2), Mar. 2021. ISSN 0360-0300. doi: 10.1145/3436755. URL https://doi.org/10.1145/3436755.

[3] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[4] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.

[5] Junyang Lin, An Yang, Jinze Bai, Chang Zhou, Le Jiang, Xianyan Jia, Ang Wang, Jie Zhang, Yong Li, Wei Lin, et al. M6-10t: A sharing-delinking paradigm for efficient multi-trillion parameter pretraining. *arXiv preprint arXiv:2110.03888*, 2021.

[6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[7] Yehuda Dar, Vidya Muthukumar, and Richard G Baraniuk. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*, 2021.

[8] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated White-Box membership inference. In *29th USENIX Security Symposium*, pages 1605–1622, August 2020.

[9] 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*, pages 268–282, 2018.

[10] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2021.

[11] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *arXiv preprint arXiv:2103.07853*, 2021.

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

[13] 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, 2019.

[14] Erick Galinkin. The influence of dropout on membership inference in differentially private models. *arXiv preprint arXiv:2103.09008*, 2021.

[15] Yijue Wang, Chenghong Wang, Zigeng Wang, Shanglin Zhou, Hang Liu, Jinbo Bi, Caiwen Ding, and Sanguthevar Rajasekaran. Against membership inference attack: Pruning is all you need. *arXiv preprint arXiv:2008.13578*, 2020.

[16] Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136*, 2017.

[17] 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.

[18] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1964–1974, Jul. 2021.

[19] Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In *International Conference on Machine Learning*, pages 5345–5355, 2021.

[20] Aadesh Bagmar, Shishira R Maiya, Shruti Bidwalka, and Amol Deshpande. Membership inference attacks on lottery ticket networks. *arXiv preprint arXiv:2108.03506*, 2021.

[21] Shahbaz Rezaei, Zubair Shafiq, and Xin Liu. Accuracy-privacy trade-off in deep ensemble. *arXiv preprint arXiv:2105.05381*, 2021.

[22] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[23] Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19, 2008.

[24] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938, 2018.

[25] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[26] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[27] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. Practical blind membership inference attack via differential comparisons. *arXiv preprint arXiv:2101.01341*, 2021.

[28] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, and Reza Shokri. Enhanced membership inference attacks against machine learning models. *arXiv preprint arXiv:2111.09679*, 2021.

[29] Rathindra Sarathy and Krishnamurty Muralidhar. Evaluating Laplace noise addition to satisfy differential privacy for numeric data. *Transactions on Data Privacy*, 4(1):1–17, 2011.

[30] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1177–1184, 2007.

[31] Youngmin Cho. *Kernel Methods for Deep Learning*. PhD thesis, University of California, San Diego, 2012.

[32] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

[33] Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4030–4034. IEEE, 2021.

[34] Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011.

[35] Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21:66–1, 2020.

[36] Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918–943, 2021.

[37] R Dennis Cook and Liliana Forzani. On the mean and variance of the generalized inverse of a singular Wishart matrix. *Electronic Journal of Statistics*, 5:146–158, 2011.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] , please see Section A.1
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] , please see Section A.2
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]