Gaussian Membership Inference Privacy

Tobias Leemann[*]
University of Tübingen
Technical University of Munich

Martin Pawelczyk[*]
Harvard University

Gjergji Kasneci
Technical University of Munich

Abstract

We propose a novel and practical privacy notion called f-Membership Inference Privacy (f-MIP), which explicitly considers the capabilities of realistic adversaries under the membership inference attack threat model. Consequently, f-MIP offers interpretable privacy guarantees and improved utility (e.g., better classification accuracy). In particular, we derive a parametric family of f-MIP guarantees that we refer to as μ-Gaussian Membership Inference Privacy (μ-GMIP) by theoretically analyzing likelihood ratio-based membership inference attacks on stochastic gradient descent (SGD). Our analysis highlights that models trained with standard SGD already offer an elementary level of MIP. Additionally, we show how f-MIP can be amplified by adding noise to gradient updates. Our analysis further yields an analytical membership inference attack that offers two distinct advantages over previous approaches. First, unlike existing state-of-the-art attacks that require training hundreds of shadow models, our attack does not require *any* shadow model. Second, our analytical attack enables straightforward auditing of our privacy notion f-MIP. Finally, we quantify how various hyperparameters (e.g., batch size, number of model parameters) and specific data characteristics determine an attacker's ability to accurately infer a point's membership in the training set. We demonstrate the effectiveness of our method on models trained on vision and tabular datasets.

1 Introduction

Machine learning (ML) has seen a surge in popularity and effectiveness, leading to its widespread application across various domains. However, some of these domains, such as finance and healthcare, deal with sensitive data that cannot be publicly shared due to ethical or regulatory concerns. Therefore, ensuring data privacy becomes crucial at every stage of the ML process, including model development and deployment. In particular, the trained model itself [5, 31] or explanations computed to make the model more interpretable [29, 32] may leak information about the training data if appropriate measures are not taken. For example, this is a problem for recent generative Diffusion Models [7] and Large Language models, where the data leakage seems to be amplified by model size [6].

Differential privacy (DP) [14] is widely acknowledged as the benchmark for ensuring provable privacy in academia and industry [10]. DP utilizes randomized algorithms during training and guarantees that the output of the algorithm will not be significantly influenced by the inclusion or exclusion of any individual sample in the dataset. This provides information-theoretic protection against the maximum amount of information that an attacker can extract about any specific sample in the dataset, even when an attacker has full access to and full knowledge about the predictive model.

While DP is an appealing technique for ensuring privacy, DP's broad theoretical guarantees often come at the expense of a significant loss in utility for many ML algorithms. This utility loss cannot be further reduced by applying savvier algorithms: Recent work [26, 27] confirms that an attacker can

---

[*]Equal contribution. Corresponding authors: `tobias.leemann@uni-tuebingen.de` and `martin.pawelczyk.1@gmail.com`.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

be implemented whose empirical capacity to differentiate between neighboring datasets D and D' when having access to privatized models matches the theoretical upper bound. This finding suggests that to improve a model's utility, we need to take a step back and inspect the premises underlying DP. For example, previous work has shown that privacy attacks are much weaker when one imposes additional realistic restrictions on the attacker's capabilities [26].

In light of these findings, we revisit the DP threat model and identify three characteristics of an attacker that might be overly restrictive in practice. First, DP grants the attacker full control over the dataset used in training including the capacity to poison all samples in the dataset. For instance, DP's protection includes pathological cases such as an empty dataset and a dataset with a single, adversarial instance [27]. Second, in many applications, it is more likely that the attacker only has access to an API to obtain model predictions [13, 31] or to model gradients [20]. Finally, one may want to protect typical samples from the data distribution. As argued by Triastcyn & Faltings [38], it may be over-constraining to protect images of dogs in a model that is conceived and trained with images of cars.

Such more realistic attackers have been studied in the extensive literature on Membership Inference (MI) attacks (e.g., [5, 41]), where the attacker attempts to determine whether a sample from the data distribution was part of the training dataset. Under the MI threat model, Carlini et al. [5] observe that ML models with very lax ($\epsilon > 5000$) or no ($\epsilon = \infty$) DP-guarantees still provide some defense against membership inference attacks [5, 41]. Hence, we hypothesize that standard ML models trained with low or no noise injection may already offer some level of protection against realistic threats such as MI, despite resulting in very large provable DP bounds.

To build a solid groundwork for our analysis, we present a hypothesis testing interpretation of MI attacks. We then derive f-Membership Inference Privacy (f-MIP), which bounds the trade-off between an MI attacker's false positive rate (i.e., FPR, type I errors) and false negative rate (i.e., FNR, type II errors) in the hypothesis testing problem by some function f. We then analyze the privacy leakage of a gradient update step in stochastic gradient descent (SGD) and derive the first analytically optimal attack based on a likelihood ratio test. However, for f-MIP to cover practical scenarios, post-processing and composition operations need to be equipped with tractable privacy guarantees as well. Using f-MIP's handy composition properties, we analyze full model training via SGD and derive explicit f-MIP guarantees. We further extend our analysis by adding carefully calibrated noise to the SGD updates to show that f-MIP may be guaranteed without any noise or with less noise than the same parametric level of f-DP [11], leading to a smaller loss of utility.

Our analysis comes with a variety of novel insights: We confirm our hypothesis that, unlike for DP, no noise ($\tau^2 = 0$) needs to be added during SGD to guarantee f-MIP. Specifically, we prove that the trade-off curves of a single SGD step converge to the family of Gaussian trade-offs identified by Dong et al. [11] and result in the more specific μ-Gaussian Membership Inference Privacy (μ-GMIP). The main contributions this research offers to the literature on privacy preserving ML include:

1. **Interpretable and practical privacy notion**: We suggest the novel privacy notion of f-MIP that addresses the realistic threat of MI attacks. f-MIP considers the MI attacker's full trade-off curve between false positives and false negatives. Unlike competing notions, f-MIP offers appealing composition and post-processing properties.

2. **Comprehensive theoretical analysis**: We provide (tight) upper bounds on any attacker's ability to run successful MI attacks, i.e., we bound any MI attacker's ability to identify whether points belong to the training set when ML models are trained via gradient updates.

3. **Verification and auditing through novel attacks**: As a side product of our theoretical analysis, which leverages the Neyman-Pearson lemma, we propose a novel set of attacks for auditing privacy leakages. An important advantage of our analytical Gradient Likelihood-Ratio (GLiR) attack is its computational efficiency. Unlike existing attacks that rely on training hundreds of shadow models to approximate the likelihood ratio, our attack does not require any additional training steps.

4. **Privacy amplification through noise addition**: Finally, our analysis shows how one can use noisy SGD (also known as Differentially Private SGD [1]) to reach f-MIP while maintaining worst-case DP guarantees. Thereby our work establishes a theoretical connection between f-MIP and f-DP [11], which allows to translate an f-DP guarantee into an f-MIP guarantee and vice versa.

2

2 Related Work

Privacy notions. DP and its variants provide robust, information-theoretic privacy guarantees by ensuring that the probability distribution of an algorithm's output remains stable even when one sample of the input dataset is changed [14]. For instance, a DP algorithm is ε-DP if the probability of the algorithm outputting a particular subset E for a dataset S is not much higher than the probability of outputting E for a dataset S_0 that differs from S in only one element. DP has several appealing features, such as the ability to combine DP algorithms without sacrificing guarantees.

A few recent works have proposed to carefully relax the attacker's capabilities in order to achieve higher utility from private predictions [4, 13, 17, 38]. For example, Dwork & Feldman [13] suggest the notion of "privacy-preserving prediction" to make private model predictions through an API interface. Their work focuses on PAC learning guarantees of any class of Boolean functions. Similarly, Triastcyn & Faltings [38] suggest "Bayesian DP", which is primarily based on the definition of DP, but restricts the points in which the datasets S and S_0 may differ to those sampled from the data distribution. In contrast, Izzo et al. [17] introduces a notion based on MI attacks, where their approach guarantees that an adversary $\mathcal{A}$ does not gain a significant advantage in terms of accuracy when distinguishing whether an element x was in the training data set compared to just guessing the most likely option. However, they only constrain the accuracy of the attacker, while we argue that it is essential to bound the entire trade-off curve, particularly in the low FPR regime, to prevent certain re-identification of a few individuals [5]. Our work leverages a hypothesis testing formulation that covers the entire trade-off curve thereby offering protection also to the most vulnerable individuals. Additionally, our privacy notion maintains desirable properties such as composition and privacy amplification through subsampling, which previous notions did not consider.

Privacy Attacks on ML Models. Our work is also related to auditing privacy leakages through a common class of attacks called MI attacks. These attacks determine if a given instance is present in the training data of a particular model [5, 7, 9, 15, 23, 29, 30, 31, 32, 35, 36, 40, 41]. Compared to these works, our work suggests a new much stronger class of MI attacks that is analytically derived and uses information from model gradients. An important advantage of our analytically derived attack is its computational efficiency, as it eliminates the need to train any additional shadow models.

3 Preliminaries

The classical notion of (ε, δ)-differential privacy [14] is the current workhorse of private ML and can be described as follows: An algorithm is DP if for any two neighboring datasets S, S' (that differ by one instance) and any subset of possible outputs, the ratio of the probabilities that the algorithm's output lies in the subset for inputs S, S' is bounded by a constant factor. DP is a rigid guarantee, that covers *every* pair of datasets S and S', including pathologically crafted datasets (for instance, Nasr et al. [27] use an empty dataset) that might be unrealistic in practice. For this reason, we consider a different attack model in this work: The MI game [41]. This attack mechanism on ML models follows the goal of inferring an individual's membership in the training set of a learned ML model. We will formulate this problem using the language of hypothesis testing and trade-off functions, a concept from hypothesis testing theory [11]. We will close this section by giving several useful properties of trade-off functions which we leverage in our main theoretical results presented in Sections 4 and 5.

3.1 Membership Inference Attacks

The overarching goal of privacy-preserving machine learning lies in protecting personal data. To this end, we will show that an alternative notion of privacy can be defined through the success of a MI attack which attempts to infer whether a given instance was present in the training set or not. Following Yeom et al. [41] we define the standard MI experiment as follows:

Definition 3.1 (Membership Inference Experiment [41])**.** *Let $\mathcal{A}$ be an attacker, A be a learning algorithm, N be a positive integer, and $\mathcal{D}$ be a distribution over data points $x \in D$, where the vector x may also be a tuple of data and labels. The MI experiment proceeds as follows: The model and data owner $\mathcal{O}$ samples $S \sim \mathcal{D}_N$ (i.e, sample n points i.i.d. from $\mathcal{D}$) and trains $A_S = A(S)$. They choose $b \in \{0, 1\}$ uniformly at random and draw $x' \sim \mathcal{D}$ if $b = 0$, or $x' \sim S$ if $b = 1$. Finally, the attacker is successful if $\mathcal{A}(x', A_S, N, \mathcal{D}) = b$. $\mathcal{A}$ must output either 0 or 1.*

3

We note that the membership inference threat model features several key differences to the threat model underlying DP, which are listed in Table 1. Most notably, in MI attacks, the datasets are sampled from the distribution $\mathcal{D}$, whereas DP protects all datasets. This corresponds to granting the attacker the capacity of full dataset manipulation. Therefore, the MI attack threat model is sensible in cases where the attacker cannot manipulate the dataset through injection of malicious samples also called "canaries". This may be realistic for financial and healthcare applications, where the data is often collected from actual events (e.g., past trades) or only a handful of people (trusted hospital staff) have access to the records. In such scenarios, it might be overly restrictive to protect against worst-case canary attacks as attackers cannot freely inject arbitrary records into the training datasets. Furthermore, MI attacks are handy as a fundamental ingredient in crafting data extraction attacks [6]. Hence we expect a privacy notion based on the MI threat model to offer protection against a broader class of reconstruction attacks. Finally, being an established threat in the literature [5, 9, 31, 40, 41], MI can be audited through a variety of existing attacks.

	f-DP threat model	f-MIP threat model (this work)
Goal	Distinguish between S and S' for *any* S, S' that differ in at most one instance.	Distinguish whether $\boldsymbol{x}' \in S$ (training data set) or not.
Dataset access	Attacker has full data access. For example, the attacker can poison or adversarially construct datasets on which ML models could be trained; e.g., $S = \{\}$ and $S' = \{10^6\}$.	Attacker has no access to the training data set; i.e., the model owner privately trains their model free of adversarially poisoned samples.
Protected Instances	The instance in which S and S' differ is arbitrary. This includes OOD samples and extreme outliers.	The sample $\boldsymbol{x}'$ for which membership is to be inferred is drawn from the data distribution $\mathcal{D}$. Therefore, MI is concerned with typical samples that can occur in practice.
Best used	When the specific attack model is unknown. Offers a form of general protection.	When dataset access (e.g. canary injection) of an attacker can be ruled out and the main attack goal lies in revealing private training data (e.g., membership inference, data reconstruction).
Model knowledge	The attacker knows the model architecture and has full access to the model in form of its parameters, hyperparameters and its model outputs.	

Table 1: Comparing the threat models underlying f-DP and f-MIP.

3.2 Membership Inference Privacy as a Hypothesis Testing Problem

While DP has been studied through the perspective of a hypothesis testing formulation for a while [3, 11, 19, 39], we adapt this route to formulate membership inference attacks. To this end, consider the following hypothesis test:

$$\text{H}_0 : \boldsymbol{x}' \notin S \text{ vs. } \text{H}_1 : \boldsymbol{x}' \in S. \tag{1}$$

Rejecting the null hypothesis corresponds to detecting the presence of the individual $\boldsymbol{x}'$ in S, whereas failing to reject the null hypothesis means inferring that $\boldsymbol{x}'$ was not part of the dataset S. The formulation in (1) is a natural vehicle to think about any attacker's capabilities in detecting members of a train set in terms of false positive and true positive rates. The motivation behind these measures is that the attacker wants to reliably identify the subset of data points belonging to the training set (i.e., true positives) while incurring as few false positive errors as possible [5]. In other words, the attacker wants to maximize their true positive rate at any chosen and ideally low false positive rate (e.g., 0.001). From this perspective, the formulation in (1) allows to define membership inference privacy via trade-off functions f which exactly characterize the relation of false negative rates (i.e., 1-TPR) and false positive rates that an optimal attacker can achieve.

Definition 3.2. *(Trade-off function [11]) For any two probability distributions P and Q on the same space, denote the trade-off function $Test(P; Q) : [0; 1] \to [0; 1]$*

$$\text{Test}(P; Q)(\alpha) = \inf \{FNR \mid FPR = \alpha\}, \tag{2}$$

where the infimum is taken over all (measurable) rejection rules ("tests") which lead to a FPR of α between distributions P, Q.

4

Not every function makes for a valid trade-off function. Instead, trade-off functions possess certain characteristics that are handy in their analysis.

Definition 3.3 (Characterization of trade-off functions [11]). *A function $f : [0,1] \to [0,1]$ is a trade-off function if f is convex, continuous at zero, non-increasing, and $f(r) \leq 1 - r$ for $r \in [0,1]$.*

We additionally introduce a semi-order on the space of trade-off functions to make statements on the hardness of different trade-offs in relation to each other.

Definition 3.4 (Comparing trade-offs). *A trade-off function f is uniformly at least as hard as another trade-off function g, if $f(r) \geq g(r)$ for all $0 \leq r \leq 1$. We write $f \geq g$.*

If $\text{Test}(P; Q) \geq \text{Test}(P'; Q')$, testing P vs Q is uniformly at least as hard as testing P' vs Q'. Intuitively, this means that for a given FPR α, the best test possible test on $(P; Q)$ will result in an equal or higher FNR than the best test on $(P'; Q')$.

3.3 Noisy Stochastic Gradient Descent (Noisy SGD)

Most recent large-scale ML models are trained via stochastic gradient descent (SGD). Noisy SGD (also known as DP-SGD) is a variant of classical SGD that comes with privacy guarantees. We consider the algorithm as in the work by Abadi et al. [1], which we restate for convenience in Appendix A. While its characteristics with respect to DP have been extensively studied, we take a fundamentally different perspective in this work and study the capabilities of this algorithm to protect against membership inference attacks. In summary, the algorithm consists of three fundamental steps: *gradient clipping* (i.e., $\boldsymbol{\theta}_i \coloneqq \boldsymbol{g}(\boldsymbol{x}_i, y_i) \cdot \max(1, C/\|\boldsymbol{g}(\boldsymbol{x}_i, y_i)\|)$ where $\boldsymbol{g}(\boldsymbol{x}_i, y_i) = \nabla \mathcal{L}(\boldsymbol{x}_i, y_i)$ is the gradient with respect to the loss function $\mathcal{L}$), *aggregation* (i.e., $\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta}_i$) and *adding Gaussian noise* (i.e., $\tilde{\boldsymbol{m}} = \boldsymbol{m} + Y$ where $Y \sim \mathcal{N}(\boldsymbol{0}, \tau^2 \boldsymbol{I})$ with variance parameter τ^2). To obtain privacy bounds for this algorithm, we study MI attacks for means of random variables. This allows us to bound the MI vulnerability of SGD.

4 Navigating Between Membership Inference Privacy and DP

In this section, we formally define our privacy notion f-MIP. To this end, it will be handy to view MI attacks as hypothesis tests.

4.1 Membership Inference Attacks from a Hypothesis Testing Perspective

Initially, we define the following distributions of the algorithm's output

$$A_0 = A(\boldsymbol{X} \cup \{\boldsymbol{x}\}) \text{ with } \boldsymbol{X} \sim \mathcal{D}^{n-1}, \boldsymbol{x} \sim \mathcal{D} \text{ and } A_1(\boldsymbol{x}') = A(\boldsymbol{X} \cup \{\boldsymbol{x}'\}) \text{ with } \boldsymbol{X} \sim \mathcal{D}^{n-1}, \quad (3)$$

where we denote other randomly sampled instances that go into the algorithm by $\boldsymbol{X} = \{\boldsymbol{x}_1, ... \boldsymbol{x}_{n-1}\}$. Here A_0 represents the output distribution under the null hypothesis (H_0) where the sample $\boldsymbol{x}'$ is not part of the training dataset. On the other hand, A_1 is the output distribution under the alternative hypothesis (H_1) where $\boldsymbol{x}'$ was part of the training dataset. The output contains randomness due to the instances drawn from the distribution $\mathcal{D}$ and due to potential inherent randomness in A.

We observe that the distribution A_1 depends on the sample $\boldsymbol{x}'$ which is known to the attacker. The attacker will have access to samples for which A_0 and $A_1(\boldsymbol{x}')$ are simpler to distinguish and others where the distinction is harder. To reason about the characteristics of such a stochastically composed test, we define a composition operator that defines an optimal test in such a setup. To obtain a global FPR of α, an attacker can target different FPRs $\bar{\alpha}(\boldsymbol{x}')$ for each specific test. We need to consider the optimum over all possible ways of choosing $\bar{\alpha}(\boldsymbol{x}')$, which we refer to as *test-specific FPR function*, giving rise to the following definition.

Definition 4.1 (Stochastic composition of trade-off functions). *Let $\mathcal{F}$ be a family of trade-off functions, $h : D \subset \mathbb{R}^d \to \mathcal{F}$ be a function that maps an instance of the data domain to a corresponding trade-off function, and $\mathcal{D}$ be a probability distribution on D. The set of valid test-specific FPR functions $\bar{\alpha} : D \to [0,1]$ that result in a global FPR of $\alpha \in [0,1]$ given the distribution $\mathcal{D}$ is defined through*

$$\mathcal{E}(\alpha, \mathcal{D}) = \{\bar{\alpha} : D \to [0,1] \mid \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}} [\bar{\alpha}(\boldsymbol{x}')] = \alpha\}. \quad (4)$$

5

For a given test-specific FPR function, $\bar{\alpha}$ the global false negative rate (type II error) β is given by

$$\beta_h(\bar{\alpha}) = \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}} \left[h(\boldsymbol{x})(\bar{\alpha}(\boldsymbol{x})) \right], \tag{5}$$

where $\bar{\alpha}(\boldsymbol{x})$ is the argument to the trade-off function $h(\boldsymbol{x}) \in \mathcal{F}$. For a global $\alpha \in [0, 1]$ the stochastic composition of these trade-functions is defined as

$$\left(\bigotimes_{\boldsymbol{x} \sim \mathcal{D}} h(\boldsymbol{x}) \right)(\alpha) = \min_{\bar{\alpha} \in \mathcal{E}(\alpha, \mathcal{D})} \{ \beta_h(\bar{\alpha}) \}, \tag{6}$$

(supposing the minimum exists), the smallest global false negative rate possible at a global FPR of α.

This definition specifies the trade-off function $\bigotimes_{\boldsymbol{x} \sim \mathcal{D}} h(\boldsymbol{x}) : [0, 1] \rightarrow [0, 1]$ of a stochastic composition of several trade-offs. While it is reminiscent of the "most powerful test" (MPT) [28], there are several differences to the MPT that are important in our work. Most prominently, a straightforward construction of the MPT to MI problems does not work since the adversary does not only run one hypothesis test to guess whether one sample belongs to the training data set or not; instead, the adversary draws multiple samples and runs sample-dependent and (potentially) different hypotheses tests for each drawn sample. This is necessary due to the form of the alternative hypotheses in the formulation of the test in (3), which depends on the sample $\boldsymbol{x}'$. We therefore require a tool to compose the results from different hypothesis tests. Finally, we prove that the trade-off of the stochastic composition has the properties of a trade-off function (see App. D.1):

Theorem 4.1 (Stochastic composition of trade-off functions). *The stochastic composition $\bigotimes_{\boldsymbol{x} \sim \mathcal{D}} h(\boldsymbol{x})$ of trade-off functions $h(\boldsymbol{x})$ maintains the characteristics of a trade-off function, i.e., it is convex, non-increasing, $\left(\bigotimes_{\boldsymbol{x} \sim \mathcal{D}} h(\boldsymbol{x}) \right)(r) \leq 1 - r$ for all $r \in [0, 1]$, and it is continuous at zero.*

4.2 f-Membership Inference Privacy (f-MIP)

This rigorous definition of the stochastic composition operator allows us to define membership inference privacy from a hypothesis testing perspective.

Definition 4.2 (f-Membership Inference Privacy). *Let f be a trade-off function. An algorithm[1] $A : D^n \rightarrow \mathbb{R}^d$ is said to be f-membership inference private (f-MIP) with respect to a data distribution $\mathcal{D}$ if*

$$\bigotimes_{\boldsymbol{x}' \sim \mathcal{D}} \text{Test}\left(A_0; A_1(\boldsymbol{x}') \right) \geq f, \tag{7}$$

where $\boldsymbol{x}' \sim \mathcal{D}$ and $\bigotimes$ denotes the stochastic composition built from individual trade-off functions of the MI hypotheses tests for random draws of $\boldsymbol{x}'$.

In this definition, both sides are functions dependent on the false positive rate α. A prominent special case of a trade-off function is the Gaussian trade-off, which stems from testing one-dimensional normal distributions of unit variance that are spaced apart by $\mu \in \mathbb{R}_{\geq 0}$. Therefore, defining the following special case of f-MIP will be useful.

Definition 4.3 (μ-Gaussian Membership Inference Privacy). *Let Φ be the cumulative distribution function (CDF) of a standard normal distribution. Define $g_\mu(\alpha) := \Phi(\Phi^{-1}(1 - \alpha) - \mu)$ to be the trade-off function derived from testing two Gaussians; one with mean 0 and one with mean μ. An algorithm A is μ-Gaussian Membership Inference private (μ-GMIP) with privacy parameter μ if it is g_μ-MIP, i.e., it is MI private with trade-off function g_μ.*

Remark 4.1. *DP can also be defined via the Gaussian trade-off function, which results in μ-Gaussian Differential Privacy (μ-GDP, [11]). While the trade-off curves for both μ-GDP and μ-GMIP have the same parametric form, they have different interpretations: μ-GDP describes the trade-off function that an attacker with complete knowledge (left column in Table 1) could achieve while μ-GMIP describes the trade-off function that an attacker with MI attack capability can achieve (right column in Table 1). In the next section we will quantify their connection further.*

---

[1] When using the term "algorithm", we also include randomized mappings.

6

4.3 Relating f-MIP and f-DP

We close this section by providing first results regarding the relation between f-DP and f-MIP. As expected, f-DP is strictly stronger than f-MIP, which can be condensed in the following result:

Theorem 4.2 (f-DP implies f-MIP). *Let an algorithm $A : D^n \to \mathbb{R}^d$ be f-differentially private [11]. Then, algorithm A will also be f-membership inference private.*

We proof this result in Appendix F.1. This theorem suggests one intuitive, simple and yet actionable approach to guarantee Membership Inference Privacy. This approach involves the use of DP learning algorithms such as DP-SGD [1], which train models using noised gradients. However, as we will see in the next section, using noise levels to guarantee f-DP is usually suboptimal to guarantee f-MIP.

5 Implementing f-MIP through Noisy SGD

We would now like to obtain a practical learning algorithm that comes with f-MIP guarantees. As the dependency between the final model parameters and the input data is usually hard to characterize, we follow the common approach and trace the information flow from the data to the model parameters through the training process of stochastic gradient descent [1, 33]. Since the gradient updates are the only path where information flows from the data into the model, it suffices to privatize this step.

5.1 f-MIP for One Step of Noisy SGD

We start by considering a single SGD step. Following prior work [1, 33], we make the standard assumption that only the mean over the individual gradients $\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta}_i$ is used to update the model (or is published directly) where $\boldsymbol{\theta}_i \in \mathbb{R}^d$ is a sample gradient. Consistent with the definition of the membership inference game, the attacker tries to predict whether a specific gradient $\boldsymbol{\theta}'$ was part of the set $\{\boldsymbol{\theta}_i\}_i$ that was used to compute the model's mean gradient $\boldsymbol{m}$ or not. We are interested in determining the shape of the attacker's trade-off function. For the sake of conciseness, we directly consider one step of noisy SGD (i.e., one averaging operation with additional noising, see Algorithm 2 from the Appendix), which subsumes a result for the case without noise by setting $\tau^2 = 0$. We establish the following theorem using the Central Limit Theorem (CLT) for means of adequately large batches of n sample gradients, which is proven in Appendix E.

Theorem 5.1 (One-step noisy SGD is f-membership inference private). *Denote the cumulative distribution function of the non-central chi-squared distribution with d degrees of freedom and non-centrality parameter γ by $F_{\chi_d^2(\gamma)}$. Let the gradients $\boldsymbol{\theta}' \in \mathbb{R}^d$ of the test points follow a distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, let $K \geq \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}'\|_2^2$ and define $n_{effective} = n + \frac{\tau^2 n^2}{C^2}$. For sufficiently large batch sizes n, one step of noisy SGD is f-membership inference private with trade-off function given by:*

$$\beta(\alpha) \approx 1 - F_{\chi_d^2((n_{effective}-1)K)}\left(\frac{n_{effective}}{n_{effective} - 1} F_{\chi_d^2(n_{effective}K)}^{-1}(\alpha)\right). \tag{8}$$

The larger the number of parameters d and the batch size n grow, the more the trade-off curve approaches the μ-GMIP curve, which we show next (see Figure 1).

Corollary 5.1 (One step noisy SGD is approx. μ-GMIP). *For large d, n, noisy SGD is approximately $g_{\mu_{Step}}$-GMIP. In particular, $\beta(\alpha) \approx \Phi(\Phi^{-1}(1 - \alpha) - \mu_{Step})$ with privacy parameter:*

$$\mu_{step} = \frac{d + (2n_{effective} - 1)K}{n_{effective}\sqrt{2d + 4n_{effective}K}}. \tag{9}$$

This result is striking in its generality as it also covers models trained without additional noise or gradient cropping ($n_{\text{effective}} = n$ in that case). Unlike for DP, even standard models trained with non-noisy SGD offer an elementary level of MIP. Our result further explicitly quantifies four factors

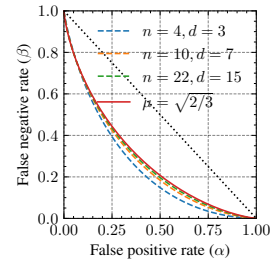



Figure 1: **Trade-off function convergence.** The trade-off function from Theorem 5.1 converges to the one from Corollary 5.1 where $\tau^2{=}0$ and $K{=}d$.

7

that lead to attack success: the batch size n, the number of parameters d, the strength of the noise τ^2 and the worst-case data-dependent *gradient susceptibility* $\|\mathbf{\Sigma}^{-1/2}\boldsymbol{\theta}'\|_2^2$. The closeness of the trade-off function to the diagonal, which is equivalent to the attacker randomly guessing whether a gradient $\boldsymbol{\theta}'$ was part of the training data or not, is majorly determined by the ratio of d to n. The higher the value of d relative to n, the easier it becomes for the attacker to identify training data points. Furthermore, a higher gradient susceptibility K, which measures the atypicality of a gradient with respect to the gradient distribution, increases the likelihood of MI attacks succeeding in identifying training data membership. It is worth noting that if we do not restrict the gradient distribution or its support, then there might always exist gradient samples that significantly distort the mean, revealing their membership in the training dataset. This phenomenon is akin to the δ parameter in DP, which also allows exceptions for highly improbable events.

Remark 5.1 (Magnitude of μ_{Step}). *When the dimensions of the uncorrelated components in $\mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\theta}'$ are also independent, we expect K to follow a χ^2-distribution with d degrees of freedom and thus $K \in \mathcal{O}(d)$. In the standard SGD-regime ($\tau^2 = 0$) with $d, n \gg 1$, we obtain $\mu \in \mathcal{O}\left(\sqrt{d/n}\right)$.*

Remark 5.2 (On Optimality). *The dependency on d when $\tau^2 > 0$ is a consequence of our intentionally broad proving strategy. Our proof approach consists of two key steps: First, we establish the optimal LRT under general gradient distributions, without adding noise or imposing any cropping constraints (See Appendix E.1). This initial step serves as the foundation for our subsequent analysis and is (1) as general as possible covering all distributions with finite variance and is (2) optimal in the sense of the Neyman-Pearson Lemma, i.e., it cannot be improved. This means that our result covers all models trained with standard SGD ($\tau^2 = 0$ and $C = \infty$) and is remarkable in its generality as it is the first to suggest clear conditions when adding noise is not required to reach f-MIP. Second, we specialize our findings to cropped random variables with added noise (See Appendix E.2). This analysis could potentially be improved by considering individual gradient dimensions independently.*

5.2 Composition and Subsampling

In the previous section, we have derived the trade-off function for a single step of SGD. Since SGD is run over multiple rounds, we require an understanding of how the individual trade-off functions can be composed when a sequence of f-MIP operations is conducted, and a random subset of the entire data distribution is used as an input for the privatized algorithm. The next lemma provides such a result for μ-GMIP and follows from a result that holds for hypotheses tests between Gaussian random variables due to Dong et al. [11] (see Appendix D.3 for details and more results).

Lemma 5.1 (Asymptotic convergence of infinite DP-SGD). *Let n be the batch size in SGD, and N be the entire size of the dataset. If a single SGD-Step is at least as hard as μ_{step}-GMIP with respect to the samples that were part of the batch and $\frac{n\sqrt{t}}{N} \to c$ as $\lim_{t\to\infty}$ (the batch size is gradually decreased), then the noisy SGD algorithm will be μ-GMIP with*

$$\mu = \sqrt{2}c\sqrt{\exp(\mu_{step}^2)\Phi\left(1.5\mu_{step}\right) + 3\Phi\left(-0.5\mu_{step}\right) - 2}. \tag{10}$$

Note that this result also provides a (loose) bound for the case where exactly T iterations are run with a batch size of n' with $c = \frac{n'\sqrt{T}}{N}$ (through using $n(t) = n'$ if $t \leq T$, else $n(t) = \frac{n'\sqrt{T}}{\sqrt{t}}$). With this result in place, we can defend against MI attacks using the standard noisy SGD algorithm.

6 Experimental Evaluation

Datasets and Models. We use three datasets that were previously used in works on privacy risks of ML models [32]: The CIFAR-10 dataset which consists of 60k small images [21], the Purchase tabular classification dataset [25] and the Adult income classification dataset from the UCI machine learning repository [12]. Following prior work by Abadi et al. [1], we use a model pretrained on CIFAR-100 and finetune the last layer on CIFAR-10 using a ResNet-56 model for this task [16] where the number of fine-tuned parameters equals $d = 650$. We follow a similar strategy on the Purchase dataset, where we use a three-layer neural network. For finetuning, we use the 20 most common classes and $d = 2580$ parameters while the model is pretrained on 80 classes. On the adult dataset, we use a two-layer network with 512 random features in the first layer trained from scratch on the

8

dataset such that $d = 1026$. We refer to Appendix C.1 for additional training details. We release our code online.[2]

6.1 Gradient Attacks Based on the Analytical LRT

To confirm our theoretical analysis for one step of SGD and its composition, we implement the gradient attack based on the likelihood ratio test derived in the proof of Theorem 5.1. We provide a sketch of the implementation in Algorithm 1 and additional details in Appendix C.3. An essential requirement in the construction of the empirical test is the estimation of the true gradient mean $\boldsymbol{\mu}$ and the true inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ since these quantities are essential parts of both the test statistic $\hat{S}$ and the true gradient susceptibility term $\hat{K}$ needed for the analytical attack. The attacker uses their access to the gradient distribution (which is standard for membership inference attacks [5, 29] and realistic in federated learning scenarios [20]), to estimate the distribution parameters. In practice, however, the empirical estimates of $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}^{-1}$ and thus $\hat{K}$ will be noisy and therefore we do not expect that the empirical trade-off curves match the analytical curves exactly.

Algorithm 1: Gradient Likelihood Ratio (GLiR) Attack

Require: Training data distribution $\mathcal{D}$, batch size n, number of parameters d, query point $\boldsymbol{x} \in D$, averaged gradients of each batch $\boldsymbol{m}_t \in \mathbb{R}^d$ for training steps $t = 1, \dots, T$, parameter gradient computation function $\nabla_{\boldsymbol{w}_t} \mathcal{L} : D \to \mathbb{R}^d$ of training, threshold η

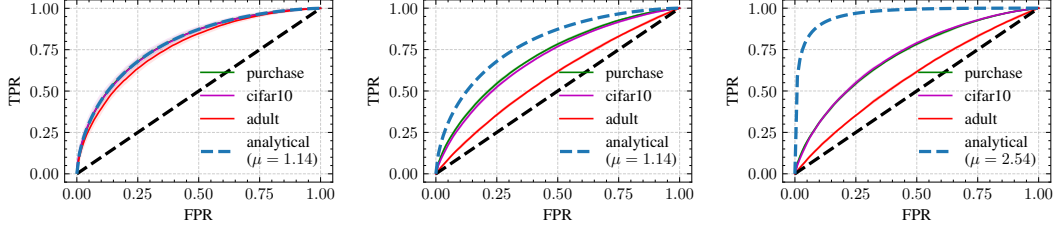
1: $p_{\text{total}} \leftarrow 0$
2: **for** $t = 1, \dots, T$ **do**
3:      $B = \{\boldsymbol{b}_1, \dots, \boldsymbol{b}_m\} \sim D^m$ ▷ Sample background data
4:      $\boldsymbol{g}_i = \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{b}_i), i = 1 \dots m$ ▷ Compute background gradients
5:      $\hat{\boldsymbol{\Sigma}} = \text{Cov} \{\boldsymbol{g}_1, \dots, \boldsymbol{g}_m\} \in \mathbb{R}^{d \times d}$ ▷ Approximate covariance $\boldsymbol{\Sigma}$
6:      $\hat{\boldsymbol{\mu}} = \text{Mean} \{\boldsymbol{g}_1, \dots, \boldsymbol{g}_m\} \in \mathbb{R}^d$ ▷ Approximate mean $\boldsymbol{\mu}$
7:      $\boldsymbol{\theta} = \nabla_{w_t} \mathcal{L}(\boldsymbol{x})$ ▷ Compute gradients for the query point
8:      $\hat{S} = (n-1)(\boldsymbol{m}_t - \boldsymbol{\theta})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{m}_t - \boldsymbol{\theta})$ ▷ Compute test statistic
9:      $\hat{K} = \|\hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\mu}})\|_2^2$ ▷ Estimate gradient susceptibility
10:     $p_{\text{step}} = \log \text{F}^{-1}_{\chi^2_d(n\hat{K})}(\hat{S})$ ▷ Compute $\log p$-value under H_0
11:     $p_{\text{total}} \leftarrow p_{\text{total}} + p_{\text{step}}$
12: **end for**
13: **return** `Train` if $p_{\text{total}} < \eta$, **else** `Test`

Using our novel Gradient Likelihood Ratio (GLiR) attack we can audit our derived guarantees and their utility. First, we audit our one-step guarantees from Theorem 5.1. To compare the models, we adapt the batch size n such that all models reach the same level of μ-GMIP. In Figure 2a, we use a simulated gradient distribution with known parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}$ and d. In this case, we can estimate K accurately and observe that our bounds are tight when the distribution parameters and thus the respective gradient susceptibilities can be computed accurately. We provide additional ablation studies that gauge the approximation quality of with small values for d, n and different simulated distributions in Appendix C.2. When the parameters are unknown and we have to estimate the parameters, our attacks become weaker and do not match the analytical prediction (see Figure 2b).

We also audit our composition guarantees. We do five SGD-steps in Figure 2c. While there is a small gain in attack performance on the CIFAR-10 dataset (e.g., at FPR=0.25), the attack performance on the other datasets remains largely unaffected. This mismatch occurs since the theoretical analysis is based on the premise that the attacker gains access to independently sampled gradient means for each step to separate training and non-training points, but in practice we do not gain much new information as the model updates are not statistically independent and too incremental to change the gradient means significantly between two subsequent steps. Therefore, a practical attacker does not gain much additional information through performing several steps instead of one. Future work is required to model these dependencies and potentially arrive at a tighter composition result under incremental parameter updates. We provide results for additional existing membership inference attacks, for instance the recent loss-based likelihood-ratio attack by Carlini et al. [5] in Appendix C.4, which all show weaker success rates than the gradient-based attack that proved most powerful in our setting.

6.2 Comparing Model Utility under μ-GDP and μ-GMIP

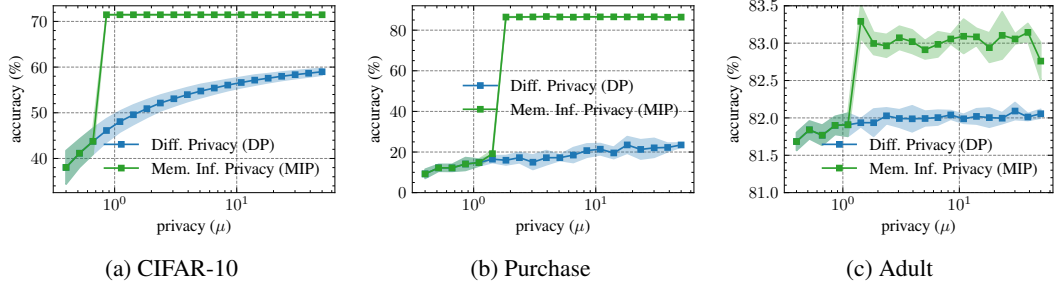
Here we compare the utility under our privacy notion to the utility under differential privacy. We sample 20 different privacy levels ranging from $\mu \in [0.4, \dots, 50]$ and calibrate the noise in the SGD iteration to reach the desired value of μ. We can do so both for μ-GMIP using the result in Equation (10) and using the result by Dong et al. [11, Corollary 4] for μ-GDP, which result in the same

---
[2]https://github.com/tleemann/gaussian_mip

9

(a) Single step of simulated gradient distribution with known parameters.

(b) Single step with real model gradients and estimated parameters.

(c) As in (b), but now composition of 5 steps for real model gradients.

Figure 2: **Auditing f-MIP with our gradient attack (GLiR) when $\tau^2 = 0$.** We show trade-off curves when the gradient distribution is known (a) and when the gradients are obtained from a trained model that was finetuned on various data sets (b, c). The analytical solutions are computed with a value of $K = d$ and using the composition result for k steps in Appendix D.3 for (c).

(a) CIFAR-10

(b) Purchase

(c) Adult

Figure 3: **Utility of DP versus MIP.** Model performance on three datasets across different privacy levels μ (small μ denotes high privacy) using the notions of μ-Gaussian Differential Privacy (parametric form of f-DP, [11]) and μ-Gaussian Membership Inference Privacy (parametric form of f-MIP, ours) on three datasets. GMIP usually allows for substantially increased accuracy over the corresponding GDP guarantee with the same attack success rates controlled by μ. However, the attacker under GMIP runs membership inference (MI) attacks while GDP allows for a wider set of privacy threat models. For more details on differences in the underlying threat models see Table 1.

attack success rates while μ-GDP allows for stronger privacy threat models. Due to Theorem 4.2, we never need to add more noise for μ-GMIP than for μ-DP. Further details are provided in Appendix C.1. Figure 3 shows a comparison of the accuracy that the models obtain. We observe that the model under GMIP results in significantly higher accuracy for most values of μ. As $\mu \to 0$ both privacy notions require excessive amounts of noise such that the utility decreases towards the random guessing accuracy. On the other hand, for higher values of μ, there is no need to add any noise to the gradient to obtain μ-GMIP, allowing to obtain the full utility of the unconstrained model. This indicates that useful GMIP-bounds do not necessarily require noise. For instance, on the CIFAR-10 model, no noise is required for $\mu \geq 0.86$ which is a reasonable privacy level [11]. Overall, these results highlight that useful and interpretable privacy guarantees can often be obtained without sacrificing utility.

7 Conclusion and Future Work

In the present work, we derived the general notion of f-Membership Inference Privacy (f-MIP) by taking a hypothesis testing perspective on membership inference attacks. We then studied the noisy SGD algorithm as a model-agnostic tool to implement f-Membership Inference Privacy, while maintaining Differential Privacy (DP) as a worst-case guarantee. Our analysis revealed that significantly less noise may be required to obtain f-MIP compared to DP resulting in increased utility. Future work is required to better model the dependencies when composing subsequent SGD steps which could lead to improved bounds in practice. Furthermore, our analysis shows that when the capacity of the attacker is further restricted, e.g., to API access of predictions, there remains a gap between our theoretical bounds and loss-based membership inference attacks that can be implemented for real models. More work is required to either produce more sophisticated attacks or derive theoretical bounds for even less powerful attackers to close this gap.

10

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016. 2, 5, 7, 8, 14, 16

[2] Andrew, G., Kairouz, P., Oh, S., Oprea, A., McMahan, H. B., and Suriyakumar, V. One-shot empirical privacy estimation for federated learning. *arXiv preprint arXiv:2302.03098*, 2023. 14

[3] Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 2496–2506. PMLR, 2020. 4

[4] Bassily, R., Thakkar, O., and Guha Thakurta, A. Model-agnostic private learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 3

[5] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2021. 1, 2, 3, 4, 9, 14, 15, 17, 18

[6] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021. 1, 4, 16

[7] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 1, 3, 14

[8] Chen, D., Yu, N., Zhang, Y., and Fritz, M. Gan-leaks: A taxonomy of membership inference attacks against generative models. In Ligatti, J., Ou, X., Katz, J., and Vigna, G. (eds.), *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pp. 343–362. ACM, 2020. doi: 10.1145/3372297.3417238. URL https://doi.org/10.1145/3372297.3417238. 14

[9] Choquette-Choo, C. A., Tramèr, F., Carlini, N., and Papernot, N. Label-only membership inference attacks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume abs/2007.14321, 2020. 3, 4, 14

[10] Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Jagielski, M., Huang, Y., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., et al. Challenges towards the next frontier in privacy. *arXiv preprint arXiv:2304.06929*, 2023. 1

[11] Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022. 2, 3, 4, 5, 6, 7, 8, 9, 10, 14, 16, 18, 20, 21, 33, 34

[12] Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. 8

[13] Dwork, C. and Feldman, V. Privacy-preserving prediction. In *Conference On Learning Theory*, pp. 1693–1702. PMLR, 2018. 2, 3

[14] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006. 1, 3

[15] Haim, N., Vardi, G., Yehudai, G., Shamir, O., et al. Reconstructing training data from trained neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 14, 16

[16] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 8

[17] Izzo, Z., Yoon, J., Arik, S. O., and Zou, J. Provable membership inference privacy. *arXiv preprint arXiv:2211.06582*, 2022. 3

11

[18] Jagielski, M., Ullman, J., and Oprea, A. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 22205–22216, 2020. 14

[19] Kairouz, P., Oh, S., and Viswanath, P. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML)*, pp. 1376–1385, 2015. 4

[20] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. 2, 9, 15

[21] Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009. 8

[22] Lehmann, E. L., Romano, J. P., and Casella, G. *Testing statistical hypotheses*, volume 3. Springer, 2005. 24

[23] Long, Y., Bindschaedler, V., Wang, L., Bu, D., Wang, X., Tang, H., Gunter, C. A., and Chen, K. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018. 3, 14

[24] Maddock, S., Sablayrolles, A., and Stock, P. Canife: Crafting canaries for empirical privacy measurement in federated learning. In *International Conference on Learning Representations (ICLR)*, 2023. 14

[25] Nasr, M., Shokri, R., and Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646, 2018. 8

[26] Nasr, M., Songi, S., Thakurta, A., Papernot, N., and Carlin, N. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pp. 866–882. IEEE, 2021. 1, 2, 14

[27] Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramèr, F., Jagielski, M., Carlini, N., and Terzis, A. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1631–1648. USENIX Association, 2023. 1, 2, 3, 14

[28] Neyman, J. and Pearson, E. S. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. 6

[29] Pawelczyk, M., Lakkaraju, H., and Neel, S. On the Privacy Risks of Algorithmic Recourse. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023. 1, 3, 9, 14, 15, 17

[30] Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jegou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. 3, 14

[31] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017. 1, 2, 3, 4, 14

[32] Shokri, R., Strobel, M., and Zick, Y. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 231–241, 2021. 1, 3, 8, 14, 15

[33] Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013. 7, 16

[34] Steinke, T., Nasr, M., and Jagielski, M. Privacy auditing with one (1) training run. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 14

12

[35] Tan, J., Mason, B., Javadi, H., and Baraniuk, R. Parameters or privacy: A provable tradeoff between overparameterization and membership inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17488–17500, 2022. 3, 14

[36] Tan, J., LeJeune, D., Mason, B., Javadi, H., and Baraniuk, R. G. A blessing of dimensionality in membership inference through regularization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 10968–10993. PMLR, 2023. 3, 14

[37] Thudi, A., Shumailov, I., Boenisch, F., and Papernot, N. Bounding membership inference. *arXiv preprint arXiv:2202.12232*, 2022. 14

[38] Triastcyn, A. and Faltings, B. Bayesian differential privacy for machine learning. In *International Conference on Machine Learning (ICML)*, pp. 9583–9592. PMLR, 2020. 2, 3

[39] Wasserman, L. and Zhou, S. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. 4

[40] Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022. 3, 4, 14

[41] Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018. 2, 3, 4, 14, 15

[42] Zanella-Béguelin, S., Wutschitz, L., Tople, S., Salem, A., Rühle, V., Paverd, A., Naseri, M., Köpf, B., and Jones, D. Bayesian estimation of differential privacy. In *International Conference on Machine Learning (ICML)*. PMLR, 2023. 14

13

A   Algorithms

Reviewing Noisy SGD. Noisy SGD, also known as DP-SGD when appropriately parameterized, is the most prevalent algorithm to train differentiable machine learning models subject to DP privacy constraints. In the main text, we have shown that this algorithm, when appropriately parameterized, can be used to train f-MIP models, too. Since our gradient attack relies on the inner workings of the algorithm, we review it here for the reader's convenience. DP-SGD works by clipping the individual gradients in each batch, taking the mean over the these gradients in a batch, and finally adding noise of magnitude τ to them. This process is then iterated over T epochs. Pseudo code is shown in Algorithm 2.

We note that the parametrization of the noise level is different across recent works. While τ corresponds to the noise added to the entire batch, other works such as [1, 11] use different parameters to characterize the noise level. For instance, Dong et al. [11] add noise of magnitude $\tau^2 = \frac{4\sigma^2 C^2}{n^2}$ and use σ to characterize the noise level.

Algorithm 2 Noisy Stochastic Gradient Descent (Noisy SGD)

Require: Training data $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, loss function $\mathcal{L}$, learning rate η, batch size n, number of iterations T, gradient norm bound $C \in \mathbb{R}_+$, noise scale $\tau \in \mathbb{R}_+$
1: Initialize model parameters $\boldsymbol{\theta}$ randomly
2: **for** $t = 1, \dots, T$ **do**
3:      Sample a batch B_t of size n uniformly at random from D
4:      **Compute gradients**
5:      For each $(\boldsymbol{x}_i, y_i) \in B_t$ compute $\boldsymbol{g}(\boldsymbol{x}_i, y_i) = \nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_i, y_i)$
6:      **Clip gradients** (to have norm at most C)
7:      $\boldsymbol{g}(\boldsymbol{x}_i, y_i) \leftarrow \boldsymbol{g}(\boldsymbol{x}_i, y_i) \cdot \max\left(1, \frac{C}{\|\boldsymbol{g}(\boldsymbol{x}_i, y_i)\|}\right), (\boldsymbol{x}_i, y_i) \in B_t$
8:      **Aggregate and noise gradients**
9:      $\tilde{\boldsymbol{g}} \leftarrow \left(\frac{1}{n} \sum_i \boldsymbol{g}(\boldsymbol{x}_i, y_i)\right) + \mathcal{N}\left(\boldsymbol{0}, \tau^2 \boldsymbol{I}\right)$
10:      **Update parameters**
11:      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \tilde{\boldsymbol{g}}$
12: **end for**
13: Return $\boldsymbol{\theta}$

B   Additional Information on Related Work

Privacy Auditing. Our research is also linked to the literature focusing on the validation of theoretical privacy guarantees, also known as privacy auditing. This literature involves the assessment of privacy breaches in private algorithms [18]. Typically, privacy auditing entails the utilization of membership inference attacks [31, 41], wherein the effectiveness of the attack is translated into an empirical approximation of the privacy level, denoted as $\hat{\varepsilon}$. While most existing privacy auditing methods, such as those by Jagielski et al. [18], Nasr et al. [26, 27] and Zanella-Béguelin et al. [42] are computationally extensive as they require multiple shadow models to be fitted to conduct privacy audits, more recent works suggest privacy audits based on a single model fit [2, 24, 34]. Similar to this recent line of work, our privacy notion can be audited by a single model fit. However, our work differs in the sense that our auditing algorithm precisely evaluates our suggested privacy notion f-MIP.

Extended comparison to privacy attacks. There is a long line of prior work developing [5, 7, 15, 23, 29, 30, 31] or analyzing [35, 36, 37] privacy attacks on machine learning models. A common class of attacks called *membership inference attacks* focus on determining if a given instance is present in the training data of a particular model [5, 8, 9, 29, 30, 31, 32, 40, 41]. Most of these attacks typically exploit the differences in the distribution of model confidence on the true label (or the loss) between the instances that are in the training set and those that are not [5, 30, 31, 40]. For example, Shokri et al. [31] proposed a loss-based membership inference attack which determines if an instance is in the training set by testing if the loss of the model for that instance is less than a specific threshold. Other membership inference attacks are also predominantly loss-based attacks where the calibration of the threshold varies from one proposed attack to the other [5, 30, 40]. Some

14

Info	Loss [41]	CFD [29]	Loss LRT [5]	CFD LRT [29]	Gradient LRT
Query access to $f_{\hat{\theta}}$	✓	×	✓	×	×
Query access to $\nabla f_{\hat{\theta}}$	×	×	×	×	✓
Query access to $\mathcal{R}$	×	✓	×	✓	×
Known loss function	✓	×	✓	×	×
Access to $\mathcal{D}^N$	×	×	✓	✓	✓
Access to true labels	✓	×	✓	×	×
Analytical	×	×	×	×	✓
Shadow models	×	×	✓	✓	×

Table 2: Summarizing the assumptions underlying the different MI attacks. The recourse based attacks do not require access to the true labels nor do they need to know the correct loss functions, but they additionally require access to a recourse generating API $\mathcal{R}$. To the best our knowledge, our gradient attack is the only one for which analytical results exist.

works leverage different information that goes beyond the loss functions to do membership inference attacks. For instance, Shokri et al. [32] and Pawelczyk et al. [29] leverage model explanations to orchestrate membership inference attacks.

Comparison to existing attacks. In Table 2, we summarize the assumptions underlying different membership inference attacks. Note that our attack does not require the training of multiple shadow models on data from the data distribution $\mathcal{D}^N$. Instead, we derive the distributions of the LRT test statistic under the null and alternative hypotheses in closed form (see Appendix E), which drops the requirement of training (appropriately parameterized) shadow models to approximate these two distributions. These shadow models can be trained since the attacker is allowed access to the general data distribution $\mathcal{D}$. Similar to other LRT attacks, our attack also requires access to $\mathcal{D}$ to approximate the parameters $\boldsymbol{\Sigma}, \boldsymbol{\mu}$ and K required for the construction and verification of our likelihood ratio based attack. As opposed to other attacks, our attack is based on the requirement that the attacker has access to model gradients which is a realistic assumption in many federated learning scenarios [20]. Appendix C.3 summarizes our gradient based LRT attack in more detail.

Type	Dataset	# Samples (N)	# Parameters (d)	Batch size (n)	Epochs	C	τ^2	Architecture
I	CIFAR-10	500	650	500	5	10.0	0.0	ResNet56
T	Purchase	1970	2580	1970	5	10.0	0.0	3 layer DNN
T	Adult	790	1026	790	5	10.0	0.0	Random feature NN

Table 3: The parameters for the verification experiment are chosen so that the analytical privacy levels from Figure 2 are $\mu_{\text{step}} = 1.13$ and $\mu = 2.54$, respectively. Note that "I" denotes image and "T" denotes tabular.

Type	Dataset	# Samples (N)	# Parameters (d, =K)	Batch size (n)	Epochs	C	Architecture
I	CIFAR-10	48000	650	400	10	500.0	ResNet56
T	Purchase	54855	2580	795	3	2000.0	3 layer DNN
T	Adult	43000	1026	1000	20	800.0	Random feature NN

Table 4: Parameters for the utility experiment from Figure 3. Note that "I" denotes image and "T" denotes tabular. The dataset size were chosen to make them divisible by the batch size. The required noise τ^2 is determined by the required privacy level μ.

Comparing the threat models underlying f-DP and f-MIP. We note that the underlying threat model in membership inference (MI) attacks features several key differences to the threat model underlying DP, which controls an attacker's capacity to distinguish *any* two neighboring datasets D and D'.

First, in MI attacks, the datasets are sampled from the distribution $\mathcal{D}$, whereas DP protects all datasets which corresponds to granting the attacker the capacity of full dataset manipulation. Thereby the MI attack model is sensible in cases where the attacker cannot manipulate the dataset through injection of malicious samples ("canaries"). Instead, the notion of MI attacks is more realistic in cases when an attacker only has API access or access to the trained model but cannot interfere during training.

15

$\mu =$	0.40	0.52	0.66	0.86	1.11	1.43	1.84	2.37	3.05	3.94	5.08	6.55	8.44	10.88	14.03	18.09	23.33	30.08	38.78	50.00
CIFAR-10 (MIP)	2.84	2.44	2.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CIFAR-10 (DP)	2.84	2.44	2.13	1.89	1.70	1.55	1.42	1.32	1.24	1.17	1.11	1.06	1.02	0.98	0.94	0.91	0.88	0.85	0.83	0.81
Purchase (MIP)	4.72	4.14	3.68	3.32	3.04	2.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Purchase (DP)	4.72	4.14	3.68	3.32	3.04	*2.81*	2.62	2.46	2.32	2.21	2.11	2.02	1.94	1.87	1.81	1.75	1.70	1.65	1.61	1.57
Adult (MIP)	3.38	2.77	2.30	1.93	1.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Adult (DP)	3.38	2.77	2.30	1.93	1.65	1.43	1.26	1.13	1.02	0.94	0.87	0.81	0.77	0.73	0.69	0.66	0.63	0.61	0.59	0.57

Table 5: Values of τ obtained for the Utility Experiment. We observe that for the higher values of μ there is often no need to add any noise to the gradients to obtain MIP, whereas substantial noise still needs to be added in the case of DP, which results in the reduced utility observed in Figure 3.

Typically membership inference attacks are a fundamental ingredient in crafting data extraction attacks [6], and hence we expect a privacy notion based on the membership inference threat model to be broadly applicable. Second, the samples that are protected under this notion are also drawn from the distribution. Consequently, MI primarily protects typical samples. In most cases, the distribution covers the data that the model is conceived to handle in practice, such that protecting against extreme outliers may be overconstraining. Finally, the goals of the attackers in both threat models are different. Instead of being able to tell apart two datasets, the MI attacker is interested in inferring whether a given sample was part of the model's training set. As the sample is already known and only a binary response is required, this goal is weaker than other types of attacks such as full reconstruction attacks [6, 15]. Therefore, the MI threat model covers many goals of realistic attackers. For the reader's convenience, we replicate the tabular overview over these key differences from Table 1 below.

C   Experimental Details

C.1   Hyperparameters

In this section, we summarize the parameter settings for our experiments. In Table 3, we provide details on the verification experiment shown in Figure 2. In Table 4, we summarize the hyperparameters used in our utility experiment shown in Figure 3.

To compute the analytical privacy levels in Figure 2, we use Corollary 5.1 with $K = d$ and $\tau = 0$, resulting in $\mu_{\text{step}} = \sqrt{\frac{2d}{2n+1}} \approx 1.14$ with the batch sizes and model parameters in Table 3. For the last Figure, we use the result shown in Appendix D.3, indicating that the combined privacy level when performing k steps of SGD without subsampling and individual privacy level μ_{step}, is given by $\mu = \sqrt{k}\mu_{\text{step}} = \sqrt{5} \cdot \mu_{\text{step}} \approx 2.54$ (calculation was performed prior to rounding).

For the utility experiment, we chose the noise level τ according to Equation (10) when we use MIP. For Differential Privacy, we use the result by Dong et al. [11, Corollary 4]. However our τ has the following relation to the σ by Dong et al., $\tau = \frac{2C}{n}\sigma$ so that we plug in $\sigma = \frac{n}{2C}\tau$ in Corollary 4 of Dong et al. [11]. We solve both Equation (10) and Corollary 4 of [11] numerically to obtain the level τ_{MIP} required to obtain μ-GMIP and τ_{DP} for μ-GDP for 20 values of μ between 0.4 and 50 that are linearly spaced in logspace. Due to Theorem 4.2, we never need to add more noise for μ-GMIP than for μ-GDP. Therefore, we set $\tau_{\text{MIP}} \leftarrow \min\{\tau_{\text{MIP}}, \tau_{\text{DP}}\}$, i.e., we take the minimum of the noise levels required for GDP, GMIP when we would like to guarantee GMIP. We obtain the values given in Table 5.

C.2   Ablation studies

We show results of several ablation studies in Figure 4. The correspond to the verification setup used in Figure 2a with simulated gradients and confirm that the approximations made in our analysis even hold for very small values of $d = 2, n = 5$ or $C = 1$ and regardless of the gradients' distribution. We provide log-log scale plots for the setups corresponding to Figure 2 in Figure 5.

C.3   Gradient Likelihood Ratio (GLiR) Attack

We follow the common approach and trace the information flow from the data through the training process of stochastic gradient descent [1, 33]. We follow [1, 33] and make the standard assumption that only the mean over the individual gradients $m = \frac{1}{n}\sum_{i=1}^{n} \theta_i$, where $\theta_i \in \mathbb{R}^d$ is a sample gradient is used to update the model (or is published directly). Consistent with the definition of the

16

(a) Changing n (b) Changing n (unif. dist.)   (c) Changing C (d) Changing d

(e) Changing n (f) Changing n (unif. dist.)   (g) Changing C (h) Changing d

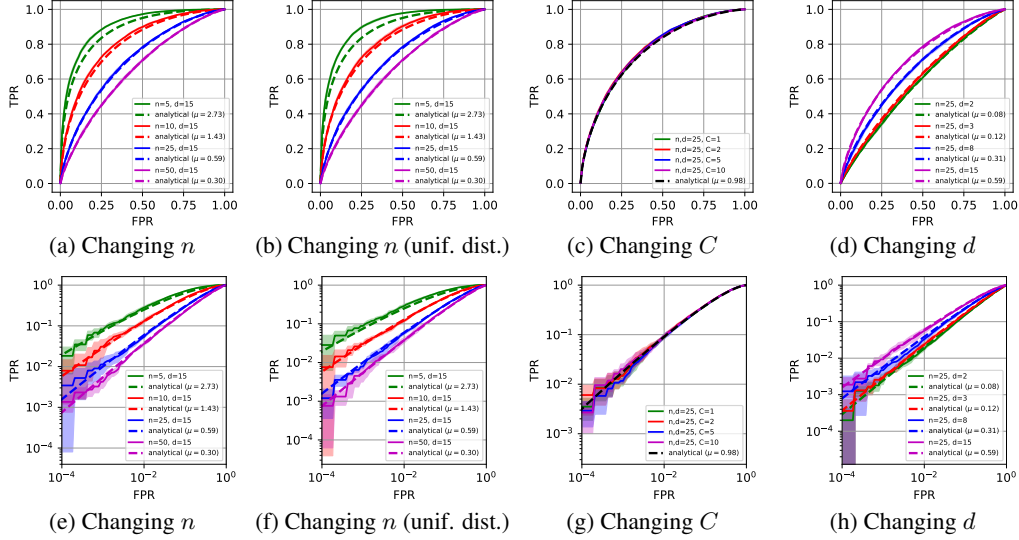



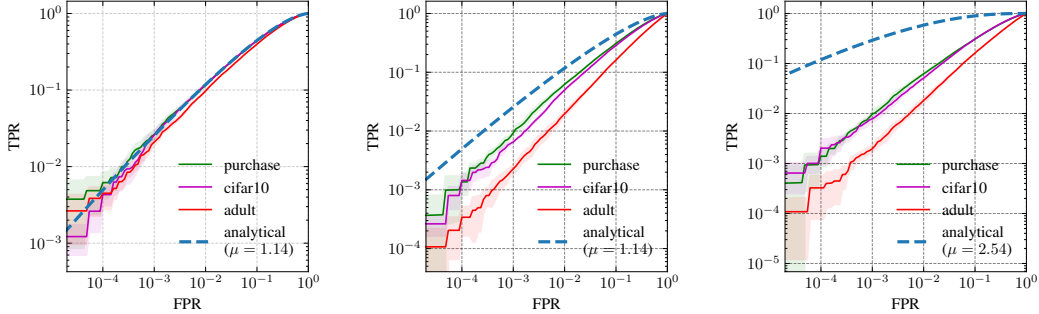
Figure 4: **Ablation studies on approximation quality; Log-log plots in bottom row. (a, e)**: Decreasing the batch size n starts showing an effect when the batch size becomes as low as $n = 5$, which is a batch size rarely used in practice. **(b, f)**: If we change the gradient distribution to be Uniform, there is no significant difference. This is expected as the CLT also holds for means of variables with bounded support. **(c, g)**: Changing the cropping threshold C has no effect on the empirical predictions ($\mathbb{E}[\|\theta\|] = 5$ in this example) **(d, h)**: Changing the gradient dimension d only has a minor effect when $d = 2$, which is a gradient dimension unlikely to be used in practice.

(a) Single step of simulated gradient distribution with known parameters.   (b) Single step with real model gradients and estimated parameters.   (c) As in (b), but now composition of 5 steps for real model gradients.

Figure 5: **Observed trade-off curves for the gradient attacks when $\tau^2 = 0$, Loglog-Scale**. We show trade-off curves when the gradient distribution is known (left) and when the gradients are obtained from a trained model that was finetuned on various data sets (center, right). The analytical solutions are computed with a value of $K = d$.

membership inference game, the attacker now tries to predict whether a specific gradient $\boldsymbol{\theta}'$ was part of the set $\{\boldsymbol{\theta}_i\}_i$ that was used to compute the mean gradient $\boldsymbol{m}$ or not.

An important requirement in the construction of the gradient likelihood ratio (GLiR) attack is the estimation of the true gradient mean $\boldsymbol{\mu}$ and the true inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ since these quantities are essential parts of both the test statistic $S = (\boldsymbol{m} - \boldsymbol{\theta}')^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{m} - \boldsymbol{\theta}')$ and the true gradient susceptibility term K (see Proof of Theorem 5.1). Here, we briefly summarize the attack algorithm (see Algorithm 1 for pseudo code):

1. The attacker uses their access to the data distribution, which is standard for membership inference attacks (see e.g., [5, 29]), to obtain estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$, which we refer to as $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the true means and covariances of the gradient distributions.

17

2. Given a gradient $\boldsymbol{\theta}'$, the attacker uses $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$, and estimates $\hat{K}$.

3. Given a gradient $\boldsymbol{\theta}'$, under the hypothesis that $\boldsymbol{\theta}'$ is part of the test set, the attacker uses $\hat{K}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ to computes the quantiles of the non-central chi-squared distribution and compares them to the test statistic S, resulting in p-values.

4. This procedure is repeated for several steps. The p-values can be aggregated through different means, where one strategy would be a simple multiplication of p-values as when assuming independence (corresponding to a sum of the $\log p$-values). However the threshold would have to be adjusted to compensate for multiple testing.

5. Finally, the attacker uses the p-values to determine whether a given gradient $\boldsymbol{\theta}'$ was part of the training set or not. The full trade-off curve can then be obtained by varying the thresholds over the p-values.

C.4 Additional Membership Inference attacks

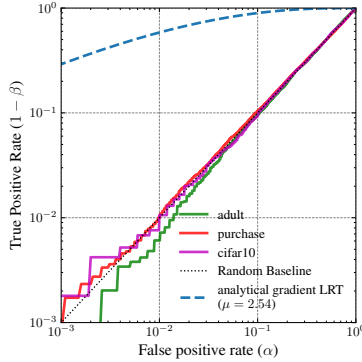
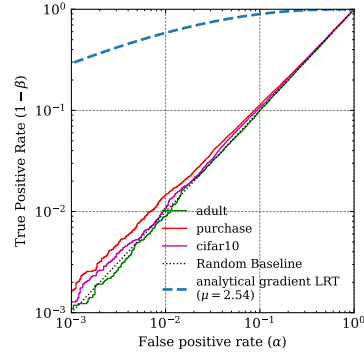
We provide the same plots as in Figure 2 using a loglog-scale in Figure 5 to show that our bounds also hold for the low FPR regime. Further, we can directly compare the analytical LRT gradient based attacks with the empirical loss based LRT attacks [5]: we see that, at the false positive rate of 10^{-2}, the loss-based attacks work substantially less reliably than our proposed analytical gradient based attacks (compare Figures 5c and 6a). The loss-based LRT attacks do not work at all when the model is trained for 5 steps only while our gradient attacks work up to 10 times more reliably. In Figure 6b, we plot loss-based attacks on models that were trained for more steps than in Figure 6a. In particular, these models were trained using the same number of epochs than the models from the utility experiment of Figure 3. Now, we see that the loss-based attacks slowly start working.

(a) Loss LRT attacks on the same models as in the verification experiment from Figures 2 and 5.

(b) Loss LRT attacks on the same models as in the utility experiment from Figure 3.

Figure 6: **Observed trade-off curves for the empirical loss based LRT attacks by Carlini et al. [5] when $\tau^2 = 0$, Loglog-Scale**. We compare the analytical trade-off curves from the gradient attack to the trade-off curves obtained from the empirical loss-based LRT attacks. The analytical solutions are computed as in Figure 5.

D Proof of Theorem 4.1 and Results for General Hypothesis Test Calculus [11]

D.1 Properties of the stochastic composition operator (Theorem 4.1)

Theorem 4.1 (Stochastic composition of trade-off functions). *The stochastic composition $\bigotimes_{\boldsymbol{x}\sim\mathcal{D}} h(\boldsymbol{x})$ of trade-off functions $h(\boldsymbol{x})$ maintains the characteristics of a trade-off function, i.e., (1) it is convex, (2) non-increasing, (3) $\left(\bigotimes_{\boldsymbol{x}\sim\mathcal{D}} h(\boldsymbol{x})\right)(r) \leq 1 - r$ for all $r \in [0,1]$, and (4) it is continuous at $r = 0$.*

Proof. **(1) convexity.** We start by proving convexity (1). Let $0 \leq a \leq b \leq 1$ and let $\lambda \in [0,1]$ and define $H(r) := \left(\bigotimes_{\boldsymbol{x}\sim\mathcal{D}} h(\boldsymbol{x})\right)(r)$ for brevity.

18

We have

$$H(a) = \min_{\bar{\alpha} \in \mathcal{E}(a,\mathcal{D})} \{\beta_h(\bar{\alpha})\} \tag{11}$$

and we denote the test-specific FPR function (TS-FPR) that reaches this minimum by $\bar{\alpha}_a(\boldsymbol{x}) \in \mathcal{E}(a,\mathcal{D})$ such that

$$H(a) = \beta_h(\bar{\alpha}_a) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [h(\boldsymbol{x})(\bar{\alpha}_a(\boldsymbol{x}))] . \tag{12}$$

We can do the same for b and find a TS-FPR function $\bar{\alpha}_b(\boldsymbol{x}) \in \mathcal{E}(b,\mathcal{D})$ such that

$$H(b) = \beta_h(\bar{\alpha}_b) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [h(\boldsymbol{x})(\bar{\alpha}_b(\boldsymbol{x}))] . \tag{13}$$

For any $\lambda \in [0,1]$, we can define the convex combination of the TS-FPR functions $\bar{\alpha}_{\lambda,a,b}(\boldsymbol{x}) = \lambda\bar{\alpha}_a(\boldsymbol{x}) + (1-\lambda)\bar{\alpha}_b(\boldsymbol{x})$ and see that

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [\lambda\bar{\alpha}_a(\boldsymbol{x}) + (1-\lambda)\bar{\alpha}_b(\boldsymbol{x})] = \lambda\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [\bar{\alpha}_a(\boldsymbol{x})] + (1-\lambda) [\bar{\alpha}_b(\boldsymbol{x})] \tag{14}$$
$$= \lambda a + (1-\lambda)b \tag{15}$$

which implies that

$$\bar{\alpha}_{\lambda,a,b} \in \mathcal{E}(\lambda a + (1-\lambda)b, \mathcal{D}), \tag{16}$$

i.e., $\bar{\alpha}_{\lambda,a,b}$ is a valid TS-FPR function for a global type 1 error of $\lambda a + (1-\lambda)b$. We can now chose the function $\bar{\alpha}_{\lambda,a,b}$ to bound the minimum which allows to complete the proof

$$H(\lambda a + (1-\lambda)b) = \min_{\bar{\alpha} \in \mathcal{E}(\lambda a + (1-\lambda)b, \mathcal{D})} \{\beta_h(\bar{\alpha})\} \tag{17}$$
$$\leq \beta_h(\bar{\alpha}_{\lambda,a,b}) \tag{18}$$
$$= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [h(\boldsymbol{x}) (\lambda\bar{\alpha}_a(\boldsymbol{x}) + (1-\lambda)\bar{\alpha}_b(\boldsymbol{x}))] \tag{19}$$
$$\leq \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [\lambda h(\boldsymbol{x}) (\bar{\alpha}_a(\boldsymbol{x})) + (1-\lambda)h(\boldsymbol{x}) (\bar{\alpha}_b(\boldsymbol{x}))] \tag{20}$$
$$= \lambda\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [h(\boldsymbol{x}) (\bar{\alpha}_a(\boldsymbol{x}))] + (1-\lambda)\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [h(\boldsymbol{x}) (\bar{\alpha}_b(\boldsymbol{x}))] \tag{21}$$
$$= \lambda H(a) + (1-\lambda)H(b). \tag{22}$$

In this derivation we use the convexity of the trade-off function $h(\boldsymbol{x})$ to arrive at Equation (20).

(3) upper bounded by $H(r) \leq 1 - r$. We prove property (3) next. For $r \in [0,1]$, we have

$$H(r) = \min_{\bar{\alpha} \in \mathcal{E}(r,\mathcal{D})} \{\beta_h(\bar{\alpha})\} \tag{23}$$

where we can bound

$$\beta_h(\bar{\alpha}) = \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}} [h(\boldsymbol{x})(\bar{\alpha}(\boldsymbol{x}))] \leq \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}} [1 - \bar{\alpha}(\boldsymbol{x})] = 1 - \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [\bar{\alpha}(\boldsymbol{x})] = 1 - r, \tag{24}$$

where we use the fact that $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [\bar{\alpha}(\boldsymbol{x})] = r$ for $\bar{\alpha} \in \mathcal{E}(r,\mathcal{D})$. Therefore,

$$H(r) = \min_{\bar{\alpha} \in \mathcal{E}(r,\mathcal{D})} \{\beta_h(\bar{\alpha})\} \leq 1 - r. \tag{25}$$

(2) non-increasing. Take any two points $0 \leq a < b \leq 1$. We can establish that $H(1) = 0$ by verifying that $H(r) \geq 0$ and the upper bound property (3). As $a < b \leq 1$, we can express

$$b = \lambda a + (1-\lambda) \cdot 1, \tag{26}$$

for some $\lambda \in [0,1)$. From the convexity property of H, we infer that

$$H(b) = H(\lambda a + (1-\lambda) \cdot 1) \leq \lambda H(a) + (1-\lambda)H(1) = \lambda H(a) \leq H(a), \tag{27}$$

or $H(a) \geq H(b)$. We therefore conclude that H is non-increasing.

(4) Continuity at $r = 0$. We will use the common ϵ,δ-criterion of continuity. Thus, we will have to show that for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $|x - 0| < \delta$ (as the support of $H(x)$ is [0,1], this means $x < \delta$), we have $|H(\delta) - H(0)| < \epsilon$.

We first denote the value of the composed trade-off at 0 by $H_0 := H(0)$ and $h_0(\boldsymbol{x}) := h(\boldsymbol{x})(0)$. Denote the TS-FPR function that takes the minimum again by $\bar{\alpha}$. Because $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [\bar{\alpha}] = 0$ together with $\bar{\alpha} \geq 0$ implies $\bar{\alpha} = 0$ almost everywhere, we can equivalently express $H_0 = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} [h_0(\boldsymbol{x})]$.

19

Now let $1 \geq \epsilon > 0$ (for $\epsilon > 1$, we can choose $\delta = 1$) and let $\epsilon' := \epsilon/4$. We first note that the individual trade-off functions $h(\boldsymbol{x})$ are increasing and continuous at $r=0$ themselves, which means that for every $h(\boldsymbol{x}), \epsilon_0$ there exists a $d(\boldsymbol{x}, \epsilon_0) > 0$ such that for all $y < d(\boldsymbol{x}, \epsilon_0)$ we have $h(\boldsymbol{x})(0) - h(\boldsymbol{x})(y) = h_0(\boldsymbol{x}) - h(\boldsymbol{x})(y) < \epsilon_0$. Absolute values are not required because $h(\boldsymbol{x})$ is monotonously decreasing. As $\boldsymbol{x}$ is stochastic, these d's also follow a certain distribution.

We now choose $\delta(\epsilon) = d_{\epsilon'}$ such that

$$P_{\boldsymbol{x} \sim \mathcal{D}}\left(h_0(\boldsymbol{x}) - h(\boldsymbol{x})(d_{\epsilon'}/\epsilon') \geq \epsilon'\right) < \epsilon'. \tag{28}$$

This means that the probability that the change in the trade-off function when going from 0 to $d_{\epsilon'}/\epsilon'$ will exceed ϵ', is bounded by ϵ'. We can find such a $d_{\epsilon'} > 0$ for every $\epsilon' > 0$ by conducting the following steps: Finding a $d(\boldsymbol{x}, \epsilon')$ for each $\boldsymbol{x}$. This is possible due to the continuity of the individual trade-offs. However, the values of $d(\boldsymbol{x}, \epsilon')$ can grow arbitrarily large or small for some $\boldsymbol{x}$. Therefore, we select the $1 - \epsilon'$-quantile $q_{1-\epsilon'}$ of the distribution of $d(\boldsymbol{x}, \epsilon')$ for $\boldsymbol{x} \sim \mathcal{D}$ for which we certainly have $0 < q_{1-\epsilon'} < 1$. We then choose $d_{\epsilon'} = q_{1-\epsilon'}\epsilon' > 0$. Having found such value $d_{\epsilon'}$ with the characteristic in Equation (28) implies

$$\mathbb{E}_{\boldsymbol{x}}\left[H_0 - h(\boldsymbol{x})(d_{\epsilon'}/\epsilon')\right] = \mathbb{E}_{\boldsymbol{x}}\left[h_0(\boldsymbol{x}) - h(\boldsymbol{x})(d_{\epsilon'}/\epsilon')\right] < 2\epsilon', \tag{29}$$

as with a probability smaller than ϵ', the change can only be bounded by 1, for the other values, it is bounded by ϵ'. We now bound $H_0 - H(y)$ for $y < d_{\epsilon'}$ to show that $H_0 - H(y) < \epsilon$. First we note that to obtain a global true positive rate of $d_{\epsilon'}$, $P(\bar{\alpha} > d_{\epsilon'}/\epsilon') \leq \epsilon'$:

$$H_0 - H(y) \leq H_0 - H(d_{\epsilon'}) \tag{30}$$
$$\leq P(\bar{\alpha} \leq d_{\epsilon'}/\epsilon)\mathbb{E}_{\boldsymbol{x}}[h_0(\boldsymbol{x}) - h(\boldsymbol{x})(\bar{\alpha}(\boldsymbol{x}))|\bar{\alpha} \leq d_{\epsilon'}/\epsilon]+ \tag{31}$$
$$P(\bar{\alpha} > d_{\epsilon'}/\epsilon)\mathbb{E}_{\boldsymbol{x}}[h_0(\boldsymbol{x}) - h(\boldsymbol{x})(\bar{\alpha}(\boldsymbol{x}))|\bar{\alpha} > d_{\epsilon'}/\epsilon] \tag{32}$$
$$\leq 1 \cdot \mathbb{E}_{\boldsymbol{x}}[H_0 - h(\boldsymbol{x})(\bar{\alpha}(\boldsymbol{x}))|\bar{\alpha} \leq d_{\epsilon'}/\epsilon] + \epsilon' \tag{33}$$

We also note that:

$$\mathbb{E}_{\boldsymbol{x}}[h_0(\boldsymbol{x}) - h(\boldsymbol{x})(d_{\epsilon'}/\epsilon)|\bar{\alpha} \leq d_{\epsilon'}/\epsilon] \tag{34}$$
$$= \frac{\mathbb{E}_{\boldsymbol{x}}[h_0(\boldsymbol{x}) - h(\boldsymbol{x})(d_{\epsilon'}/\epsilon)] - P(\bar{\alpha} > d_{\epsilon'}/\epsilon)\mathbb{E}_{\boldsymbol{x}}[h_0(\boldsymbol{x}) - h(\boldsymbol{x})(d_{\epsilon'}/\epsilon)|\bar{\alpha} \leq d_{\epsilon'}/\epsilon]}{P(\bar{\alpha} \leq d_{\epsilon'}/\epsilon)} \tag{35}$$
$$< \frac{2\epsilon'}{1 - \epsilon'} < 2\epsilon'\frac{4}{3} < 3\epsilon'. \tag{36}$$

In total we arrive at $H(0) - H(y) \leq H(0) - H(d_{\epsilon'}) < 3\epsilon' + \epsilon' = 4\epsilon' = \epsilon$. $\square$

D.2 A composition lemma for individual tests

We repeat a lemma from Dong et al. [11].

Definition D.1. *The tensor product of two trade-off functions $f = Test(P; Q)$ and $g = Test(P'; Q')$ where P, P', Q, Q' are distributions is defined as*

$$f \otimes g := Test(P \times P'; Q \times Q'). \tag{37}$$

Thus, the trade-off function f of a test that is composed of independent dimension-wise tests with trade-off functions $f_1, \ldots, f_n$ can be written as $f = f_1 \otimes f_2 \otimes \ldots \otimes f_n$. We reiterate the following result:

Lemma D.1. *If there are two trade-off functions $f_1 \geq f_2$, i.e., the test f_1 is uniformly at least as hard as f_2, for any other trade-off function g:*

$$f_1 \otimes g \geq f_2 \otimes g. \tag{38}$$

Thus, by making an individual test harder, the functional composition (the tensor product, see below) will also be uniformly harder or maintain its hardness. This lemma corresponds to Lemma C.2. of Dong et al. [11, Appendix C] where the corresponding proof can be found.

20

D.3 Composition results for f-MIP derived from results for f-DP

The result given in Lemma 5.1 follows from Theorem 11 and Lemma 3 in Dong et al. [11], which provide functional composition results (we refer to the subsequent execution of two algorithms as *functional* composition and the to the stochastic selection of a test as in Definition 4.1 as *stochastic* composition) for general hypotheses tests. In particular, the following result [11, Theorem 4] can be restated:

Theorem D.1. *Let $A_i : D \times D_1 \times \ldots \times D_{i-1} \to D_i$ be a series of f_i-DP algorithms ("mechansisms" in [11]) for all inputs $x \in D, y_1 \in D_1, \ldots, y_{i-1} \in D_{i-1}$, for $i = 1, \ldots, r$. Then the r-fold composed mechanism $M : D \to D_1 \times \ldots \times D_r$, defined as $M = (A_1(x), A_2(x, A_1(x)), \ldots, A_r(x, \ldots))$ is $f_1 \otimes \ldots \otimes f_r$-DP.*

This and other composition results from Dong et al. also apply to the MIP bounds derived for our analysis of DP-SGD and for a stricter from of f-MIP, where additionally each Test$(A_0, A_1(x')) > f$ (in Definition 4.2) is bounded by f for each x'. We can then replace f-DP with f-MIP in the results.

Proof Scheme. This can be seen as follows: Suppose we have several steps $i = 1, \ldots, k$ each being naturally f_i-MIP as each of the tests in Definition 4.2 is bounded by f_i for each x'. We can apply the functional composition results from Dong et al. for each x' independently, as they hold for hypotheses tests in general. We therefore bound the functional composition for each individual x' through the composition result FuncComp$(f_1, \ldots, f_k)$. When we finally perform the stochastic composition to obtain f-MIP, we use the result stated below (Theorem D.2) which tells us that if each individual test in the stochastic composition is bounded through some trade-off, this in an upper bound on the entire stochastic composition as well. In our case each individual test is bounded by FuncComp$(f_1, \ldots, f_k)$, which will thus bound the stochastic composition result and the level of f-MIP as well.

In our analysis of SGD, such a worst-case exists and boils down to the test for the highest value of K considered (as this results in the highest μ). Using these insights, the results by Dong et al. [11] can generally be transferred without further ramifications. Instead of using a value of $\frac{1}{\sigma}$ for the privacy level of each step, we plug in μ_{step} and arrive at Lemma 5.1.

Example: Composing μ-GMIP algorithms. As an alternative example for a composition results, we note that Dong et al. [11, Corollary 2] provide a result which explicitly states that the r-fold composition of steps which are μ_i-GDP each is $\sqrt{\mu_1 + \ldots + \mu_r}$-GDP. This result also holds for μ-GMIP as well, if the stochastic composition operator can again be bounded by the trade-off g_{μ_i} for every x', such as in the SGD case we study in this work. We use this result to arrive at the privacy level shown in Figure 2c, where we conduct k steps of SGD, resulting in a combined privacy level of $\mu = \sqrt{k}\mu_{\text{step}}$.

We now consider the concluding remark. We first note that it is possible to make the functional composition harder by using smaller batch sizes than allowed in the theorem (see Dong et al. [11, Theorem 9, Fact 1] to verify that the trade-off function will be at least as hard when the subsampling ratio is smaller). This is what we do for the steps up to T. For the subsequent steps $t > T$, we would be allowed so use the same mechanism with μ_{step}-MIP but at a smaller batch size (this would usually require additional noise to be added). However, we do not perform these steps, thereby again making the functional composition of the tests harder.

D.4 Worst-case bound for stochastic composition

Theorem D.2 (Worst-case bounds for stochastic composition). *Let $h : \mathcal{X} \to \mathcal{F}$ denote a mapping from the input space to a set of trade-off functions $\mathcal{F}$. Suppose there is a trade-off function f^* such that every other trade-off function $f \in \mathcal{F}$ is uniformly at least as hard as f^*,*

$$f \geq f^*, \forall f \in \mathcal{F}. \tag{39}$$

Then, the stochastic composition of trade-off functions will also be uniformly at least as hard as f^*, i.e.

$$\left(\bigotimes_{x \sim \mathcal{D}} h(x) \right) \geq f^* \tag{40}$$

21

regardless of the choice of h or the distribution $\mathcal{D}$.

Proof. Denote $H(r) := \left(\bigotimes_{\boldsymbol{x} \sim \mathcal{D}} h(\boldsymbol{x}) \right)(r)$ again for brevity.

$$H(r) = \min_{\bar{\alpha} \in \mathcal{E}(r, \mathcal{D})} \{\beta_h(\bar{\alpha})\} \tag{41}$$

where for every $\bar{\alpha} \in \mathcal{E}(r, \mathcal{D})$ we can bound

$$\beta_h(\bar{\alpha}) = \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}}\left[h(\boldsymbol{x})(\bar{\alpha}(\boldsymbol{x}))\right] \geq \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}}\left[f^*\left(\bar{\alpha}(\boldsymbol{x})\right)\right] \geq f^*\left(\mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}}\left[\bar{\alpha}(\boldsymbol{x})\right]\right) = f^*(r). \tag{42}$$

We use the Jensens inequality to derive that $\mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}}\left[f^*\left(\bar{\alpha}(\boldsymbol{x})\right)\right] \geq f^*\left(\mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}}\left[\bar{\alpha}(\boldsymbol{x})\right]\right)$ because f^* is convex. Therefore, we conclude that for every $r \in [0, 1]$

$$H(r) = \min_{\bar{\alpha} \in \mathcal{E}(r, \mathcal{D})} \{\beta_h(\bar{\alpha})\} \geq f^*(r). \tag{43}$$

$\square$

E Proof of Theorem 5.1 and Corollary 5.1

For the sake of better readability, we summarize our setting before we proceed with the formal proof:

- $\Theta = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n], \boldsymbol{\theta}_i \sim P$, where $\boldsymbol{\theta}_i \in \mathbb{R}^d$. P can be any distribution with finite mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ (in our application, θ_i are gradients of the samples)

- The sample mean $\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta}_i$ is published

- $b \in \{0, 1\}$ is drawn uniformly at random. If $b = 0, \boldsymbol{x}' \sim \Theta$, if $b = 1, \boldsymbol{x}' \sim P$. $\boldsymbol{x}'$ is published.

- The attacker $\mathcal{A}(\boldsymbol{m}, \boldsymbol{x}') = b'$ attempts to predict the value of b, i.e., whether $\boldsymbol{x}'$ was in the training set or not.

The following hypothesis test succinctly summarizes the attacker's problem in the this setting[3]:

$$H_0 : \boldsymbol{x}' \text{ was drawn from } \Theta \qquad\qquad H_1 : \boldsymbol{x}' \text{ was drawn from } P. \tag{44}$$

Based on this testing setup, the attacker constructs an attack based on a likelihood ratio test. Next, we summarize the individual proof steps before we give the formal proof:

1. Derive the distributions of $\boldsymbol{m}$ under the null and the alternative hypothesis for a given $\boldsymbol{x}'$.

2. Given these two distributions, we can setup the likelihood ratio, which will yield the test statistic S. This statistic will be optimal due to the Neyman-Pearson fundamental testing lemma;

3. Given the test statistic S and the distribution under the null hypothesis, we derive the rejection region of the likelihood ratio test for a given level α;

4. Finally, we derive the false negative rate $\beta(\alpha)$ of the likelihood ratio test. This trade-off function depends on the CDF and inverse CDF of non-central χ^2 distribution;

5. In a final step, we provide approximations to the trade-off function for large d.

6. Steps 1-5 initially prove f-membership inference privacy of a single SGD step when we add Gaussian noise with covariance $\hat{\tau}^2 \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is the covariance of the gradients that we would like to privatize (see Appendix E.1). We then use this result to bound the privacy of SGD with unit noise in Appendix E.2. There we show that the result from Appendix E.1 can be used in combination with a scaled noise level to guarantee membership inference privacy when we add Gaussian noise with covariance $\tau^2 \boldsymbol{I}$.

---

[3]Note that the hypotheses are interchanged with respect to the main paper here. Following the remarks after Corollary 2 of Dong et al., the trade-off function is inverted by when interchanging H_0, H_1. To arrive at the trade-off in the main paper, we will later invert the trade-off function derived here.

22

E.1 Proof of Theorem 5.1 and Corollary 5.1 with data dependent noise

In this section, we consider the effect that averaging with Gaussian noise has on membership inference privacy. We first add Gaussian noise with covariance $\hat{\tau}^2 \Sigma$, where Σ is the covariance of the gradients that we would like to privatize. In the next section, we will consider the case of independent unit noise $\tau^2 I$, which can be derived from the result presented here. The proof in this subsection follows the steps outlined in the previous section.

Proof. **Step 1: Deriving the distributions of m under H_0 and H_1.** First, we derive the distributions of m for both cases of interest. We suppose that the number of averaged samples is sufficiently large such that we can apply the Central Limit Theorem. Note that this does not restrict the form of the distribution P, besides having finite variance. Below, we start with the distribution of m under H_0 (with no additional noise yet):

$$m \sim \mathcal{N}\left(\frac{1}{n}x' + \frac{n-1}{n}\mu, \frac{(n-1)}{n^2}\Sigma\right) = \mathcal{N}\left(\mu + \frac{1}{n}(x'-\mu), \frac{(n-1)}{n^2}\Sigma\right). \tag{45}$$

Moreover, under the alternative hypothesis H_1, we have:

$$m \sim \mathcal{N}\left(\mu, \frac{1}{n}\Sigma\right). \tag{46}$$

Instead of testing the distributions of m we directly, we can equivalently test $m - x'$ by subtracting x' from both means. Adding Gaussian noise $Y \sim \mathcal{N}(0, \hat{\tau}^2 \Sigma)$ under both hypotheses results in the test that we provide below:

$$\bigotimes_{x' \sim \mathcal{D}} \mathrm{Test}\left[\mathcal{N}\left(\frac{n-1}{n}\mu, \frac{(n-1)}{n^2}\Sigma\right) - \frac{n-1}{n}x' + Y, \mathcal{N}\left(\mu, \frac{1}{n}\Sigma\right) - x' + Y\right]. \tag{47}$$

Again, we now consider the test for a fixed x'. If we can show that there is one x' that makes the test harder than any other $x' \in \mathcal{X}$, we can apply Theorem D.2 and show that the composed test is uniformly at least as hard as for x'.

We conduct the following reformulations (note that the hardness of a test remains unaffected by by invertible transforms, e.g., linear transforms):

$$\mathrm{Test}\left[\mathcal{N}\left(\frac{n-1}{n}\mu, \frac{(n-1)}{n^2}\Sigma\right) - \frac{n-1}{n}x' + Y, \mathcal{N}\left(\mu, \frac{1}{n}\Sigma\right) - x' + Y\right] \tag{48}$$

$$\iff \mathrm{Test}\left[\mathcal{N}\left(\frac{n-1}{n}(\mu - x'), \frac{(n-1)}{n^2}\Sigma + \hat{\tau}^2\Sigma\right), \mathcal{N}\left(\mu - x', \frac{1}{n}\Sigma + \hat{\tau}^2\Sigma\right)\right] \tag{49}$$

$$\iff \mathrm{Test}\left[\mathcal{N}\left(-\frac{1}{n}(\mu - x'), \left(\frac{(n-1)}{n^2} + \hat{\tau}^2\right)\Sigma\right), \mathcal{N}\left(0, \left(\frac{1}{n} + \hat{\tau}^2\right)\Sigma\right)\right] \tag{50}$$

$$\iff \mathrm{Test}\left[\mathcal{N}\left(-\frac{1}{n\sqrt{n^{-1} + \hat{\tau}^2}}\Sigma^{-\frac{1}{2}}(\mu - x'), \left(\frac{n-1}{n^2} + \hat{\tau}^2\right)\left(\frac{1}{n} + \hat{\tau}^2\right)^{-1}I\right), \mathcal{N}(0, I)\right] \tag{51}$$

$$\iff \mathrm{Test}\left[\mathcal{N}\left(-\frac{1}{n\sqrt{n^{-1} + \hat{\tau}^2}}\tilde{\delta}, \frac{n(1 + \hat{\tau}^2 n) - 1}{n(1 + \hat{\tau}^2 n)}I\right), \mathcal{N}(0, I)\right], \tag{52}$$

where $\tilde{\delta} = \Sigma^{-\frac{1}{2}}(\mu - x') = \tilde{\mu} - \tilde{x}'$ and the tilde indicates the corresponding quantities transformed by $\Sigma^{-\frac{1}{2}}$. For instance, to arrive at Equation (50) and Equation (51) the random variable $m - x'$ is transformed by subtracting $(\mu - x')$ and multiplied by $\frac{1}{\sqrt{n^{-1} + \hat{\tau}^2}}\Sigma^{-\frac{1}{2}} = \sqrt{\frac{n^2}{n + n^2 \hat{\tau}^2}}\Sigma^{-\frac{1}{2}}$, respectively. This yields the following transformed likelihood ratio test for the transformed random variable

$$Q := \sqrt{\frac{n^2}{n + n^2 \hat{\tau}^2}}\Sigma^{-\frac{1}{2}}((m - x') - (\mu - x')) = \sqrt{\frac{n^2}{n + n^2 \hat{\tau}^2}}\Sigma^{-\frac{1}{2}}(m - \mu) \tag{53}$$

$$= \sqrt{\frac{n^2}{n + n^2 \hat{\tau}^2}}(\tilde{m} - \tilde{\mu}). \tag{54}$$

23

Step 2: Deriving the test statistic. By the Neyman-Pearson Lemma [22, Theorem 3.2.1.], conducting the likelihood ratio test will be most powerful test at a given false positive rate. The corresponding likelihood ratio is given as follows:

$$\text{LR} = \frac{p_0(Q)}{p_1(Q)} = \frac{\mathcal{N}\left(Q; \boldsymbol{\mu}_1, \sigma_1^2 \boldsymbol{I}\right)}{\mathcal{N}\left(Q; \boldsymbol{\mu}_2, \sigma_2^2 \boldsymbol{I}\right)} \tag{55}$$

$$= \frac{\mathcal{N}\left(Q; -\frac{1}{\sqrt{n+n^2\hat{\tau}^2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\mu} - \boldsymbol{x}'\right), \frac{n+n^2\hat{\tau}^2-1}{n+n^2\hat{\tau}^2}\boldsymbol{I}\right)}{\mathcal{N}\left(Q; \boldsymbol{0}, \boldsymbol{I}\right)} \tag{56}$$

$$= c_2 \mathcal{N}\left(Q; \boldsymbol{d}, \boldsymbol{D}\right). \tag{57}$$

We can use identities for the ratio of two normal distributions (with $\boldsymbol{\mu}_1 = -\frac{1}{\sqrt{n(1+n\hat{\tau}^2)}}\tilde{\boldsymbol{\delta}}, \boldsymbol{\mu}_2 = \boldsymbol{0}, \boldsymbol{\Sigma}_1 = \frac{n(1+\hat{\tau}^2 n)-1}{n(1+\hat{\tau}^2 n)}\boldsymbol{I}, \boldsymbol{\Sigma}_2 = \boldsymbol{I}$) and obtain

$$\boldsymbol{D} = (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})^{-1} = \left(\frac{n(1+\hat{\tau}^2 n)-1}{n(1+\hat{\tau}^2 n)} - \frac{n(1+\hat{\tau}^2 n)}{n(1+\hat{\tau}^2 n)}\right)^{-1}\boldsymbol{I} = (n + \hat{\tau}^2 n^2 - 1)\boldsymbol{I} \tag{58}$$

and

$$\boldsymbol{d} = \boldsymbol{D}\left(\boldsymbol{\Sigma}_1^{-1}\mu_1 - \boldsymbol{\Sigma}_2^{-1}\mu_2\right) = (n + \hat{\tau}^2 n^2 - 1)\frac{n(1+\hat{\tau}^2 n)}{n(1+\hat{\tau}^2 n)-1}\left(-\frac{1}{\sqrt{n+\hat{\tau}^2 n^2}}\tilde{\boldsymbol{\delta}}\right) \tag{59}$$

$$= -\sqrt{n + \hat{\tau}^2 n^2}\tilde{\boldsymbol{\delta}} = -\sqrt{n + \hat{\tau}^2 n^2}\left(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{x}}'\right). \tag{60}$$

For a Gaussian likelihood ratio of the form in Equation (57), i.e., $S = c_2 \exp\left(-\frac{1}{2}(\boldsymbol{s}' - \boldsymbol{d})^\top \boldsymbol{D}^{-1}(\boldsymbol{s}' - \boldsymbol{d})\right)$, where $\boldsymbol{s}' = \sqrt{\frac{n^2}{n+n^2\hat{\tau}^2}}(\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}})$ it suffices to use the inner argument as a test statistic, as exp is an invertible transform. Therefore, we can use the following as a test statistic:

$$S = \sqrt{\frac{n^2}{n+n^2\hat{\tau}^2}}\left((\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) + \frac{n+\hat{\tau}^2 n^2}{n}\tilde{\boldsymbol{\delta}}\right)^\top \frac{1}{(n+n^2\hat{\tau}^2)-1}\boldsymbol{I}\sqrt{\frac{n^2}{n+n^2\hat{\tau}^2}}\left((\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) + \frac{n+\hat{\tau}^2 n^2}{n}\tilde{\boldsymbol{\delta}}\right) \tag{61}$$

$$= \frac{n^2}{((n+n^2\hat{\tau}^2)-1)(n+n^2\hat{\tau}^2)}\left((\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) + \frac{n+\hat{\tau}^2 n^2}{n}\tilde{\boldsymbol{\delta}}\right)^\top\left((\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) + \frac{n+\hat{\tau}^2 n^2}{n}\tilde{\boldsymbol{\delta}}\right), \tag{62}$$

for which we can derive the closed-form distributions under both the null and alternative hypotheses.

Step 3: Deriving the distributions of S under H_0 and H_1. From above, we know that, under the respective hypotheses we have:

$$H_0 : \sqrt{\frac{n^2}{n+n^2\hat{\tau}^2}}(\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) \sim \mathcal{N}\left(-\frac{1}{\sqrt{n+n^2\hat{\tau}^2}}\tilde{\boldsymbol{\delta}}, \frac{n+n^2\hat{\tau}^2-1}{n+n^2\hat{\tau}^2}\boldsymbol{I}\right) \tag{63}$$

$$H_1 : \sqrt{\frac{n^2}{n+n^2\hat{\tau}^2}}(\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{64}$$

and

$$H_0 : (\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) \sim \mathcal{N}\left(-\frac{1}{n}\tilde{\boldsymbol{\delta}}, \frac{n+n^2\hat{\tau}^2-1}{n^2}\boldsymbol{I}\right) \tag{65}$$

$$H_1 : (\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) \sim \mathcal{N}\left(\boldsymbol{0}, \frac{n+n^2\hat{\tau}^2}{n^2}\boldsymbol{I}\right). \tag{66}$$

and hence, for $\tilde{\boldsymbol{l}} = (\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) + \frac{n+n^2\hat{\tau}^2}{n}\tilde{\boldsymbol{\delta}} = (\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{\mu}}) + \frac{n+n^2\hat{\tau}^2}{n}(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{x}}')$ the distributions are given by:

$$H_0 : \tilde{\boldsymbol{l}} \sim \mathcal{N}\left(\frac{n-1}{n}(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{x}}') + n\hat{\tau}^2(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{x}}'), \frac{n+n^2\hat{\tau}^2-1}{n^2}\boldsymbol{I}\right) \tag{67}$$

$$H_1 : \tilde{\boldsymbol{l}} \sim \mathcal{N}\left((\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{x}}') + n\hat{\tau}^2(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{x}}'), \frac{n+n^2\hat{\tau}^2}{n^2}\boldsymbol{I}\right). \tag{68}$$

24

Next, note that if $\tilde{l} \sim \mathcal{N}(l, \kappa I)$, and $\frac{\tilde{l}}{\sqrt{\kappa}} \sim \mathcal{N}(\frac{l}{\sqrt{\kappa}}, I)$ then we have that

$$\|\tilde{l}\|_2^2 = \kappa \sum_{j=1}^{d} (U_j + b_j)^2, \tag{69}$$

where U_j are standard normal variables. Hence, $\|\tilde{l}\|_2^2$ follows the law of a (scaled) non-central chi-squared distribution with d degrees of freedom and $\boldsymbol{b} = \frac{1}{\sqrt{\kappa}}l$, i.e.,

$$\frac{\|\tilde{l}\|_2^2}{\kappa} \sim \chi_d'^2(\gamma), \tag{70}$$

where $\gamma = \sum_{j=1}^{d} b_j^2 = \sum_{j=1}^{d} \left(\frac{1}{\sqrt{\kappa}}l_j\right)^2 = \frac{1}{\kappa}\sum_{j=1}^{d} l_j^2$.

Under the null hypothesis, we obtain $\kappa_0 = \frac{n+n^2\hat{\tau}^2-1}{n^2}$ and $\boldsymbol{l}_0 = \frac{n-1+n^2\hat{\tau}^2}{n}(\tilde{\mu} - \tilde{x}')$. Therefore, the distribution of the test statistic under the null hypothesis is given by the scaled, non-central chi-squared random variable:

$$H_0 : S = \frac{n^2\kappa_0}{(n+n^2\hat{\tau}^2-1)(n+n^2\hat{\tau}^2)} \frac{\|\tilde{l}\|_2^2}{\kappa_0} \sim \frac{n^2\kappa_0}{(n+n^2\hat{\tau}^2-1)(n+n^2\hat{\tau}^2)}\chi_d'^2(\gamma_0) \tag{71}$$

$$\sim \frac{1}{n+n^2\hat{\tau}^2}\chi_d'^2(\gamma_0). \tag{72}$$

In summary,

$$(n+n^2\hat{\tau}^2)S \sim \chi_d'^2(\gamma_0), \tag{73}$$

where $\gamma_0 = \|\boldsymbol{l}_0\|_2^2/\kappa_0$.

Step 4: Deriving the rejection region for any α. Therefore, the rejection region of the null hypothesis at a significance level of α can be formulated as:

$$\left\{(n+\hat{\tau}^2n^2)S \geq \text{CDF}^{-1}_{\chi_d'^2(\gamma_0)}(1-\alpha)\right\} = \left\{S \geq \frac{1}{n+\hat{\tau}^2n^2}\text{CDF}^{-1}_{\chi_d'^2(\gamma_0)}(1-\alpha)\right\}. \tag{74}$$

Step 5: Deriving the false negative rate $\beta(\alpha)$ at any α. To compute the type two error rate β (the null hypothesis is accepted, but the alternative H_1 is true), we compute the probability of mistakenly accepting it. To this end, note that $\kappa_1 = \frac{n+n^2\hat{\tau}^2}{n^2}$ and that $\boldsymbol{l}_1 = (1+ +n\hat{\tau}^2)(\tilde{\mu} - \tilde{x}')$. Now, we can derive the distribution of the test statistic under the alternative hypothesis:

$$H_1 : S = \frac{n^2\kappa_1}{(n+n^2\hat{\tau}^2-1)(n+n^2\hat{\tau}^2)} \frac{\|\tilde{l}\|_2^2}{\kappa_1} \sim \frac{1}{n+n^2\hat{\tau}^2-1}\chi_d'^2(\gamma_1), \tag{75}$$

where $\gamma_1 = \|\boldsymbol{l}_1\|_2^2/\kappa_1$. Note that the form of this distribution,

$$(n+n^2\hat{\tau}^2-1)S \sim \chi_d'^2(\gamma_1), \tag{76}$$

where $\gamma_1 = n\|\tilde{\mu} - \tilde{x}'\|_2^2 := nK$ when no noise is added ($\hat{\tau} = 0$) is the statistic used in the GLiR attack (Algorithm 1). The type two error is given by:

$$\beta = P_1\left(S \leq \frac{1}{n+n^2\hat{\tau}^2}\text{CDF}^{-1}_{\chi_d'^2(\gamma_0)}(1-\alpha)\right) \tag{77}$$

$$= P_1\left((n+n^2\hat{\tau}^2-1)S \leq \frac{n+n^2\hat{\tau}^2-1}{n+n^2\hat{\tau}^2}\text{CDF}^{-1}_{\chi_d'^2(\gamma_0)}(1-\alpha))\right)$$

$$= P\left(X \leq \frac{n+n^2\hat{\tau}^2-1}{n+n^2\hat{\tau}^2}\text{CDF}^{-1}_{\chi_d'^2(\gamma_0)}(1-\alpha))\right)$$

$$= \text{CDF}_{\chi_d'^2(\gamma_1)}\left(\frac{n+n^2\hat{\tau}^2-1}{n+n^2\hat{\tau}^2}\text{CDF}^{-1}_{\chi_d'^2(\gamma_0)}(1-\alpha))\right)$$

$$= \text{CDF}_{\chi_d'^2(\gamma_1)}\left(\frac{n+n^2\hat{\tau}^2-1}{n+n^2\hat{\tau}^2}\text{CDF}^{-1}_{\chi_d'^2(\gamma_0)}(1-\alpha))\right), \tag{78}$$

where $X \sim \chi_d'^2(\gamma_1)$.

25

Remark E.1. *Computing the inverse trade-off, i.e., solving for α would result in*

$$\alpha = 1 - CDF_{\chi_d'^2(\gamma_0)}\left(\frac{n + n^2\hat{\tau}^2}{n + n^2\hat{\tau}^2 - 1}CDF^{-1}_{\chi_d'^2(\gamma_1)}(\beta)\right), \tag{79}$$

giving the trade-off curve for the hypotheses as stated in main paper and resulting in Theorem 5.1. For the subsequent analysis this interchange of H_0, H_1 does not play a role as the trade-off function is symmetric for large d, n (it is its own inverse).

Step 6: Large d, n approximations. The following fact will be useful. Let $V \sim \chi_d'^2(\gamma)$, then $\frac{V - (d+\gamma)}{\sqrt{2(d+2\gamma)}} \to \mathcal{N}(0,1)$ when $d \to \infty$. The trade-off function for our hypothesis test can thus be expressed through the normal CDF Φ as:

$$\beta \approx \Phi\left(\frac{\left(\frac{n+\hat{\tau}^2 n^2 - 1}{n+\hat{\tau}^2 n^2}CDF^{-1}_{\chi_d'^2(\gamma_0)}(1-\alpha)\right) - (d+\gamma_1)}{\sqrt{2(d+2\gamma_1)}}\right) \tag{80}$$

$$\approx \Phi\left(\frac{\left(\frac{n+\hat{\tau}^2 n^2 - 1}{n+\hat{\tau}^2 n^2}\left(\sqrt{2(d+2\gamma_0)}\Phi^{-1}(1-\alpha) + (d+\gamma_0)\right)\right) - (d+\gamma_1)}{\sqrt{2(d+2\gamma_1)}}\right) \tag{81}$$

$$= \Phi\left(\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\alpha)\frac{(n+n^2\hat{\tau}^2)s_1 - (n+n^2\hat{\tau}^2 - 1)s_0}{(n+n^2\hat{\tau}^2)s_1}\right.$$
$$\left. - \frac{(n+n^2\hat{\tau}^2)m_1 - (n+n^2\hat{\tau}^2 - 1)m_0}{(n+n^2\hat{\tau}^2)s_1}\right), \tag{82}$$

where $s_1 = \sqrt{2(d+2\gamma_1)}$, $s_0 = \sqrt{2(d+2\gamma_0)}$, $m_0 = d + \gamma_0$ and $m_1 = d + \gamma_1$. Thus we have:

$$\beta \approx \Phi\left(\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\alpha)\left(1 - \frac{n+\hat{\tau}^2 n^2 - 1}{n+\hat{\tau}^2 n^2}\sqrt{\frac{d+2\gamma_0}{d+2\gamma_1}}\right)\right.$$
$$\left. - \frac{(n+\hat{\tau}^2 n^2)(d+\gamma_1) - (n+\hat{\tau}^2 n^2 - 1)(d+\gamma_0)}{(n+\hat{\tau}^2 n^2)\sqrt{2d+4\gamma_1}}\right), \tag{83}$$

For large n, $\frac{n+\hat{\tau}^2 n^2 - 1}{n+\hat{\tau}^2 n^2}\sqrt{\frac{d+2\gamma_0}{d+2\gamma_1}} \approx 1$ and the second term can be dropped in the approximation. Next, recall that $\gamma_0 = \frac{n^2\|l_0\|^2}{n+n^2\hat{\tau}^2 - 1}$ and $\gamma_1 = \frac{n^2\|l_1\|^2}{n+n^2\hat{\tau}^2}$:

$$\beta \approx \Phi\left(\Phi^{-1}(1-\alpha) - \frac{(n+\hat{\tau}^2 n^2)(d + \frac{n^2\|l_1\|^2}{n+n^2\hat{\tau}^2}) - (n+\hat{\tau}^2 n^2 - 1)(d + \frac{n^2\|l_0\|^2}{n+n^2\hat{\tau}^2 - 1})}{(n+\hat{\tau}^2 n^2)\sqrt{2d + 4\frac{n^2\|l_1\|^2}{n+n^2\hat{\tau}^2}}}\right) \tag{84}$$

$$= \Phi\left(\Phi^{-1}(1-\alpha) - \frac{(n+\hat{\tau}^2 n^2)d + n^2\|l_1\|^2 - (n+\hat{\tau}^2 n^2 - 1)d - n^2\|l_0\|^2}{(n+\hat{\tau}^2 n^2)\sqrt{2d + 4\frac{n^2}{n+\hat{\tau}^2 n^2}\|l_1\|^2}}\right) \tag{85}$$

$$= \Phi\left(\Phi^{-1}(1-\alpha) - \frac{d + n^2(\|l_1\|^2 - \|l_0\|^2)}{(n+\hat{\tau}^2 n^2)\sqrt{2d + 4\frac{n^2}{n+\hat{\tau}^2 n^2}\|l_1\|^2}}\right) \tag{86}$$

$$= \Phi\left(\Phi^{-1}(1-\alpha) - \frac{d + n^2(\|l_1\|^2 - \|l_0\|^2)}{(n+\hat{\tau}^2 n^2)\sqrt{2d + 4\frac{n^2}{n+\hat{\tau}^2 n^2}\|l_1\|^2}}\right). \tag{87}$$

We further obtain $\|l_0\|^2 = \frac{(n-1+\hat{\tau}^2 n^2)^2}{n^2}K$ and $\|l_1\|^2 = \frac{(n+\hat{\tau}^2 n^2)}{n^2}K$ where $K = \|\tilde{\mu} - \tilde{x}'\|^2$. Therefore, we have

$$\beta \approx \Phi\left(\Phi^{-1}(1-\alpha) - \frac{d + (2(n+\hat{\tau}^2 n^2) - 1)K}{(n+\hat{\tau}^2 n^2)\sqrt{2d + 4(n+\hat{\tau}^2 n^2)K}}\right). \tag{88}$$

26

Thus, for large n, d the trade-off curve can be well approximated by the μ-GMIP trade-off with:

$$\mu = \frac{d + \left(2(n + \hat{\tau}^2 n^2) - 1\right)K}{(n + \hat{\tau}^2 n^2)\sqrt{2d + 4(n + \hat{\tau}^2 n^2)K}}. \tag{89}$$

We observe that the hardness of the test decreases with K. Therefore, the stochastic composition of tests is uniformly at least as hard as the test with the largest K possible (Theorem D.2). We obtain the result shown in the paper by replacing the data-dependent noise $\hat{\tau}$ by data-independent noise τ at a ratio of $\hat{\tau}^2 = \tau^2/C^2$ as detailed in the next section. $\square$

E.2 Proof of Theorem 5.1 and Corollary 5.1 with data independent noise

In the previous section, we have assumed that $Y \sim \hat{\tau}^2 \boldsymbol{\Sigma}$. This assumption does not quite match common practice; in practice, the Gaussian mechanism adds noise of the form: $Y \sim \tau^2 \boldsymbol{I}$. Hence, the following subsection investigates the effect of adding unit noise $Y \sim \tau^2 \boldsymbol{I}$.

Proof. We first study the case of data-dependent noise using an eigenspace transform. Denoting an eigenvalue composition $\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ and performing the corresponding transform (multiplication by $\boldsymbol{Q}^\top$) in Equation (47), we obtain:

$$\iff \text{Test}\left[\mathcal{N}\left(\frac{n-1}{n}(\boldsymbol{\mu} - \boldsymbol{x}'), \frac{(n-1)}{n^2}\boldsymbol{\Sigma} + \hat{\tau}^2\boldsymbol{\Sigma}\right), \mathcal{N}\left(\boldsymbol{\mu} - \boldsymbol{x}', \frac{1}{n}\boldsymbol{\Sigma} + \hat{\tau}^2\boldsymbol{\Sigma}\right)\right] \tag{90}$$

$$\iff \text{Test}\left[\mathcal{N}\left(-\frac{1}{n}\boldsymbol{Q}^\top(\boldsymbol{\mu} - \boldsymbol{x}'), \left(\frac{(n-1)}{n^2} + \hat{\tau}^2\right)\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}\right), \mathcal{N}\left(\boldsymbol{0}, \left(\frac{1}{n} + \hat{\tau}^2\right)\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}\right)\right] \tag{91}$$

$$\iff \text{Test}\left[\mathcal{N}\left(-\frac{1}{n}\boldsymbol{Q}^\top(\boldsymbol{\mu} - \boldsymbol{x}'), \left(\frac{(n-1)}{n^2} + \hat{\tau}^2\right)\boldsymbol{\Lambda}\right), \mathcal{N}\left(\boldsymbol{0}, \left(\frac{1}{n} + \hat{\tau}^2\right)\boldsymbol{\Lambda}\right)\right]. \tag{92}$$

We see that the test decomposes in a series of d dimension-wise 1D hypotheses tests of the form:

$$\text{Test}\left[\mathcal{N}\left(-\frac{1}{n}\boldsymbol{Q}^\top(\boldsymbol{\mu} - \boldsymbol{x}')_i, \frac{n-1}{n^2}\lambda_i + \hat{\tau}^2\lambda_i\right), \mathcal{N}\left(0, \frac{1}{n}\lambda_i + \hat{\tau}^2\lambda_i\right)\right]. \tag{93}$$

Here, λ_i are the eigenvalues of $\boldsymbol{\Sigma}$ that are on the diagonal of the matrix $\boldsymbol{\Lambda}$. On the other hand, when adding unit noise of magnitude τ, i.e., setting $Y = \mathcal{N}(\boldsymbol{0}, \tau^2 \boldsymbol{I})$ in Equation (47):

$$\text{Test}\left[\mathcal{N}\left(\frac{n-1}{n}(\boldsymbol{\mu} - \boldsymbol{x}'), \frac{(n-1)}{n^2}\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}\right), \mathcal{N}\left(\boldsymbol{\mu} - \boldsymbol{x}', \frac{1}{n}\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}\right)\right] \tag{94}$$

$$\iff \text{Test}\left[\mathcal{N}\left(-\frac{1}{n}\boldsymbol{Q}^\top(\boldsymbol{\mu} - \boldsymbol{x}'), \left(\frac{(n-1)}{n^2}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q} + \tau^2\boldsymbol{Q}^\top\boldsymbol{I}\boldsymbol{Q}\right)\right), \tag{95}$$

$$\mathcal{N}\left(\boldsymbol{0}, \left(\frac{1}{n}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q} + \tau^2\boldsymbol{Q}^\top\boldsymbol{I}\boldsymbol{Q}\right)\right)\right] \tag{96}$$

$$\iff \text{Test}\left[\mathcal{N}\left(-\frac{1}{n}\boldsymbol{Q}^\top(\boldsymbol{\mu} - \boldsymbol{x}'), \frac{(n-1)}{n^2}\boldsymbol{\Lambda} + \tau^2\boldsymbol{I}\right), \mathcal{N}\left(\boldsymbol{0}, \frac{1}{n}\boldsymbol{\Lambda} + \tau^2\boldsymbol{I}\right)\right]. \tag{97}$$

This test also decomposes in d 1D tests of the form:

$$\text{Test}\left[\mathcal{N}\left(\underbrace{-\frac{1}{n}\boldsymbol{Q}^\top(\boldsymbol{\mu} - \boldsymbol{x}')_i}_{\mu_i}, \underbrace{\frac{n-1}{n^2}\lambda_i + \tau^2}_{\sigma_1^2}\right), \mathcal{N}\left(0, \underbrace{\frac{1}{n}\lambda_i + \tau^2}_{\sigma_2^2}\right)\right]. \tag{98}$$

The test is therefore equally hard as testing d independent normal random variables because the covariance matrices shown are diagonal. We will show that by setting

$$\tau^2 = \hat{\tau}^2 \cdot \max_i\{\lambda_i\} \tag{99}$$

each individual, dimension-wise test is made strictly harder than the corresponding dimension-wise test in the data-dependent noise case and therefore the composed test is also harder (this is shown in Lemma D.1).

27

We will do so by computing the power function of the test and showing that is is montonically decreasing in τ, which means that for each individual test, that higher effective noise in that dimension makes the test harder. To show this, note that the likelihood ratio for this test results in

$$S = \frac{(m - \hat{\mu})^2}{\hat{\sigma}^2} \tag{100}$$

where

$$\hat{\sigma}^2 = \left(\frac{\sigma_2^2 - \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right)^{-1} = \frac{\left(\frac{n-1}{n^2} \lambda_i + \tau^2 \right) \left(\frac{1}{n} \lambda_i + \tau^2 \right) n^2}{\lambda_i}, \tag{101}$$

$$\hat{\mu} = \hat{\sigma}^2 \left(\frac{1}{\sigma_1^2} \boldsymbol{\mu}_1 - \frac{1}{\sigma_2^2} \boldsymbol{\mu}_2 \right) \tag{102}$$

$$= \hat{\sigma}^2 \frac{1}{\sigma_2^2 \sigma_1^2} \left(\sigma_2^2 \boldsymbol{\mu}_1 - \sigma_1^2 \boldsymbol{\mu}_2 \right) \tag{103}$$

$$= \frac{\sigma_2^2 \sigma_1^2}{\sigma_2^2 - \sigma_1^2} \frac{1}{\sigma_2^2 \sigma_1^2} \left(\sigma_2^2 \boldsymbol{\mu}_1 - \sigma_1^2 \boldsymbol{\mu}_2 \right) \tag{104}$$

$$= \frac{1}{\sigma_2^2 - \sigma_1^2} \left(\sigma_2^2 \boldsymbol{\mu}_1 - \sigma_1^2 \boldsymbol{\mu}_2 \right) \tag{105}$$

$$= \frac{n^2}{\lambda_i} \left(-\frac{1}{n} \boldsymbol{Q}^\top (\boldsymbol{\mu} - \boldsymbol{x}') \left(\frac{1}{n} \lambda_i + \tau^2 \right) \right) \tag{106}$$

$$= -\boldsymbol{Q}^\top (\boldsymbol{\mu} - \boldsymbol{x}')_i \left(\frac{n^2}{\lambda_i} \left(\frac{\lambda_i}{n^2} + \frac{\tau^2}{n} \right) \right) \tag{107}$$

$$= -\boldsymbol{Q}^\top (\boldsymbol{\mu} - \boldsymbol{x}')_i \left(1 + \frac{n\tau^2}{\lambda_i} \right) = n\mu_i \left(1 + \frac{n\tau^2}{\lambda_i} \right). \tag{108}$$

Distribution under the null hypothesis: Again, as before we derive the distribution under the null hypothesis:

$$m - \hat{\mu} \sim \mathcal{N} \left(\mu_i \left(1 - n \left(1 + \frac{n\tau^2}{\lambda_i} \right) \right), \frac{(n-1)}{n^2} \lambda_i + \tau^2 \right) \tag{109}$$

$$\frac{m - \hat{\mu}}{\sigma_1} \sim \mathcal{N} \left(\mu_i \frac{1 - n - \frac{n^2 \tau^2}{\lambda_i}}{\sqrt{\frac{(n-1)}{n^2} \lambda_i + \tau^2}}, 1 \right) \tag{110}$$

$$\frac{m - \hat{\mu}}{\sigma_1} \sim \mathcal{N} \left(-\mu_i \frac{n - 1 + \frac{n^2 \tau^2}{\lambda_i}}{\sqrt{\frac{\lambda_i}{n^2} \left(n - 1 + \frac{n^2 \tau^2}{\lambda_i} \right)}}, 1 \right) \tag{111}$$

$$\frac{m - \hat{\mu}}{\sigma_1} \sim \mathcal{N} \left(-\mu_i \sqrt{n - 1 + \frac{n^2 \tau^2}{\lambda_i}} \sqrt{\frac{n^2}{\lambda_i}}, 1 \right) \tag{112}$$

$$\frac{m - \hat{\mu}}{\sigma_1} \sim \mathcal{N} \left(-\mu_i \left(n \sqrt{\frac{n-1}{\lambda_i} + \frac{n^2 \tau^2}{\lambda_i^2}} \right), 1 \right) \tag{113}$$

$$\left(\frac{m - \hat{\mu}}{\sigma_1} \right)^2 \sim \chi_1'^2 \left(\mu_i^2 n^2 \left(\frac{n-1}{\lambda_i} + \frac{n^2 \tau^2}{\lambda_i^2} \right) \right) = \chi_1'^2 (\gamma_0), \tag{114}$$

where $\chi_d(\gamma)$ again denotes the non-central χ-distribution with d degrees of freedom.

Rejection region: Therefore, the rejection region of the null hypothesis at a significance level of α can be formulated as:

$$\left\{ \frac{\hat{\sigma}^2}{\sigma_1^2} S \geq \mathrm{CDF}_{\chi_1'^2(\gamma_0)}^{-1} (1 - \alpha) \right\} = \left\{ S \geq \frac{\sigma_1^2}{\hat{\sigma}^2} \mathrm{CDF}_{\chi_1'^2(\gamma_0)}^{-1} (1 - \alpha) \right\}, \tag{115}$$

28

where $\frac{\sigma_1^2}{\hat{\sigma}^2} = \frac{(\frac{n-1}{n}\lambda_i + \tau^2)\lambda_i}{n^2(\frac{n-1}{n}\lambda_i+\tau^2)(\frac{1}{n}\lambda_i+\tau^2)} = \frac{\lambda_i}{n^2(\frac{1}{n}\lambda_i+\tau^2)} = \frac{\lambda_i}{n\lambda_i + n^2\tau^2}$.

Distribution under the alternative hypothesis: As before, we also need to derive the distribution of the test statistic under the alternative hypothesis:

$$m - \hat{\mu} \sim \mathcal{N}\left(-n\mu_i\left(1 + \frac{n\tau^2}{\lambda_i}\right), \frac{1}{n}\lambda_i + \tau^2\right) \tag{116}$$

$$\frac{m - \hat{\mu}}{\sigma_2} \sim \mathcal{N}\left(-n\mu_i\frac{1 + \frac{n\tau^2}{\lambda_i}}{\sqrt{\frac{1}{n}\lambda_i + \tau^2}}, 1\right) \tag{117}$$

$$\frac{m - \hat{\mu}}{\sigma_2} \sim \mathcal{N}\left(-n\mu_i\frac{1 + \frac{n\tau^2}{\lambda_i}}{\sqrt{\frac{\lambda_i}{n}\left(1 + \frac{n\tau^2}{\lambda_i}\right)}}, 1\right) \tag{118}$$

$$\frac{m - \hat{\mu}}{\sigma_2} \sim \mathcal{N}\left(-\mu_i n\sqrt{\frac{n}{\lambda_i}\left(1 + \frac{n\tau^2}{\lambda_i}\right)}, 1\right) \tag{119}$$

$$\left(\frac{m - \hat{\mu}}{\sigma_2}\right)^2 \sim \mathcal{X}_1'^2\left(\mu_i^2 n^2\left(\frac{n}{\lambda_i} + \frac{n^2\tau^2}{\lambda_i^2}\right)\right) \tag{120}$$

Therefore,

$$\frac{\hat{\sigma}^2}{\sigma_2^2}S \sim \mathcal{X}_1'^2\left(\mu_i^2 n^2\left(\frac{n}{\lambda_i} + \frac{n^2\tau^2}{\lambda_i^2}\right)\right) = \mathcal{X}_1'^2(\gamma_1). \tag{121}$$

False negative rate: Finally, 1 - power of the test is given by:

$$\beta(\alpha) = P\left\{S \leq \frac{\sigma_1^2}{\hat{\sigma}^2}\text{CDF}^{-1}_{\mathcal{X}_1'^2(\gamma_0)}(1 - \alpha)\right\} \tag{122}$$

$$= P\left\{\frac{\hat{\sigma}^2}{\sigma_2^2}S \leq \frac{\hat{\sigma}^2}{\sigma_2^2}\frac{\sigma_1^2}{\hat{\sigma}^2}\text{CDF}^{-1}_{\mathcal{X}_1'^2(\gamma_0)}(1 - \alpha)\right\} \tag{123}$$

$$= \text{CDF}_{\mathcal{X}_1'^2(\gamma_1)}\left(\frac{\sigma_1^2}{\sigma_2^2}\text{CDF}^{-1}_{\mathcal{X}_1'^2(\gamma_0)}(1 - \alpha)\right). \tag{124}$$

We will now introduce two quantities on which the power of this test depends. In particular

$$q := \frac{n + \frac{n^2\tau^2}{\lambda_i} - 1}{n + \frac{n^2\tau^2}{\lambda_i}}, \tag{125}$$

which has the intriguing property that

$$\frac{\gamma_0}{\gamma_1} = \frac{\sigma_1^2}{\sigma_2^2} = q. \tag{126}$$

Using this insight, the trade-off function can be expressed as

$$\beta(\alpha, \gamma_1(\lambda_i, \mu_i), q(\lambda_i)) = \text{CDF}_{\mathcal{X}_1'^2(\gamma_1)}\left(q\text{CDF}^{-1}_{\mathcal{X}_1'^2(q\gamma_1)}(1 - \alpha)\right). \tag{127}$$

We determine the hardest test by showing that when considering a fixed μ_i, α, we obtain

$$\frac{\partial\beta}{\partial(\tau^2)} > 0, \tag{128}$$

which indicates that at a higher level of noise, the type 2 error rate will increase, making the test harder (more privacy). The derivative computation can be found in Appendix E.2.1. When we set $\tau^2 = \hat{\tau}^2 \cdot \max_i\{\lambda_i\}$, the individual tests in the case of data-independent noise are thus harder (μ_i, n, d stays the same) than the respective dimension-wise test for the data-dependent noise. Therefore the corresponding bound derived for $\hat{\tau}$ will hold for data independent noise of strength τ as well.

29

The largest possible eigenvalue λ_i of the covariance matrix Σ when cropping each vector at norm C is given by C^2. Therefore, we can set $\tau^2 = \hat{\tau}^2 C^2$ to obtain a harder test with data independent noise. On the converse, applying data-independent noise of level τ^2 results in a harder test than applying data-dependent noise with $\hat{\tau}^2 = \frac{\tau^2}{C^2}$. Plugging this result in Equation (89) requires to only replace $\hat{\tau}^2 = \frac{\tau^2}{C^2}$ and introducing the quantity $n_{\text{effective}}$ as

$$n_{\text{effective}} = n + \frac{n^2 \tau^2}{C^2}. \tag{129}$$

$\square$

E.2.1 Derivatives of the type 2 error

In this section we calculate the derivatives of the type 2 error function

$$\beta(\alpha, \gamma_1(\lambda_i, \mu_i), q(\lambda_i)) = \text{CDF}_{\mathcal{X}_1'^2(\gamma_1)}\left(q\text{CDF}^{-1}_{\mathcal{X}_1'^2(q\gamma_1)}(1-\alpha)\right). \tag{130}$$

Considering a fixed μ_i we show that

$$\frac{\partial \beta}{\partial(\tau^2)} > 0, \tag{131}$$

which indicates that with larger added noise, the type 2 error rate will decrease, making the test uniformly harder (more privacy). Thus the hardest test is the one with maximum τ^2. To emphasize that we are deriving by (τ^2) (the squared quantity), we write $\tau_2 := \tau^2$.

We perform the following calculations:

Derivatives of CDFs and inverse CDFs. We will start by deriving some useful derivatives of the CDFs and inverse CDFs. We start with the inverse CDF. Initially, Let $u = \sqrt{\text{CDF}^{-1}_{\mathcal{X}_1'^2(\gamma_0)}(1-\alpha)}$ or $u^2 = \text{CDF}^{-1}_{\mathcal{X}_1'^2(\gamma_0)}(1-\alpha)$ (note that the quantile function here is always > 0):

$$\Phi(u - \sqrt{\gamma_0}) - \Phi(-(u + \sqrt{\gamma_0})) = \alpha. \tag{132}$$

Keeping α constant, we can use the implicit function derivative theorem with $F(\gamma_0, u(\gamma_0)) = \Phi(u - \sqrt{\gamma_0}) - \Phi(-(u + \sqrt{\gamma_0})) - \alpha = 0$, we obtain

$$\frac{\partial u}{\partial \gamma_0} = -\frac{-\phi(u - \sqrt{\gamma_0})\frac{1}{2\sqrt{\gamma}} + \phi(-u - \sqrt{\gamma_0})\frac{1}{2\sqrt{\gamma}}}{\phi(u - \sqrt{\gamma_0}) + \phi(-u - \sqrt{\gamma_0})} \tag{133}$$

$$= \frac{1}{2\sqrt{\gamma_0}}\frac{\phi(u - \sqrt{\gamma_0}) - \phi(-u - \sqrt{\gamma_0})}{\phi(u - \sqrt{\gamma_0}) + \phi(-u - \sqrt{\gamma_0})} \tag{134}$$

We furthermore have

$$\frac{\text{CDF}^{-1}_{\mathcal{X}_1'^2(\gamma_0)}(1-\alpha)}{\partial \gamma_0} = 2\sqrt{\text{CDF}^{-1}_{\mathcal{X}_1'^2(\gamma_0)}(1-\alpha)}\frac{\partial u}{\partial \gamma_0}. \tag{135}$$

We note that

$$\text{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s) = \Phi(\sqrt{s} - \sqrt{\gamma_1}) - \Phi(-\sqrt{s} - \sqrt{\gamma_1})), \tag{136}$$

where s is the argument of the CDF. The derivative

$$\frac{\partial \text{CDF}_{\mathcal{X}_1'^2(\gamma_1)}}{\partial \gamma_1} = -\frac{1}{2\sqrt{\gamma_1}}\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) - \phi(-\sqrt{s} - \sqrt{\gamma_1})\right) \leq 0 \tag{137}$$

$$\frac{\partial \text{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial s} = \frac{1}{2\sqrt{s}}\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) + \phi(-\sqrt{s} - \sqrt{\gamma_1})\right) > 0. \tag{138}$$

30

Derivative with respect to μ_i. We compute the derivative with respect to μ_i first, which will be handy in the subsequent derivation of the derivative by τ_2. To denote that we consider a fixed α, we write β_α. We have

$$\beta_\alpha(\gamma_1(\mu_i, \lambda_i)), q(\lambda_i)) = \mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}\left(q\mathrm{CDF}^{-1}_{\mathcal{X}_1'^2(q\gamma_1)}(1 - \alpha)\right). \tag{139}$$

Thus, we can calculate

$$\frac{\partial \beta_\alpha}{\partial \mu_i} = \frac{\partial \beta_\alpha}{\partial \gamma_1}\frac{\partial \gamma_1}{\partial \mu_i} = \underbrace{\left(\frac{\partial \mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial \gamma_1} + \frac{\partial \mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial s}\frac{\partial s}{\partial \gamma_1}\right)}_{<0?}\underbrace{\frac{\partial \gamma_1}{\partial \mu_i}}_{>0}, \tag{140}$$

where $s = q\mathrm{CDF}^{-1}_{\mathcal{X}_1'^2(q\gamma_1)}(1 - \alpha)$. We can confirm that $\frac{\partial \gamma_1}{\partial \mu_i} > 0$ because

$$\gamma_1 = \mu_i^2 n^2 \left(\frac{n}{\lambda_i} + \frac{n^2\tau^2}{\lambda_i^2}\right) \tag{141}$$

is monotonically increasing in μ_i, because all eigenvalues $\lambda_i > 0$ and the factor after μ_i^2 is positive overall. To establish that the overall derivative is negative, we therefore have to confirm that the first factor is always negative:

$$\frac{\partial \mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial \gamma_1} = -\frac{1}{2\sqrt{\gamma_1}}\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) - \phi(-\sqrt{s} - \sqrt{\gamma_1})\right) \tag{142}$$

$$\frac{\partial s}{\partial \gamma_1} = q\frac{\mathrm{CDF}^{-1}_{\mathcal{X}_1'^2(q\gamma_1)}(1 - \alpha)}{\partial \gamma_1} = 2q^2\sqrt{\mathrm{CDF}^{-1}_{\mathcal{X}_1'^2(q\gamma_1)}(1 - \alpha)}\frac{\partial u}{\partial \gamma_0}\Big|_{\gamma_0 = q\gamma_1}. \tag{143}$$

Let $s = q\mathrm{CDF}^{-1}_{\mathcal{X}_1'^2(qL)}(1 - \alpha)$

$$\frac{\partial \mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial s}\frac{\partial s}{\partial \gamma_1} = \frac{\mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial s}2q^2\sqrt{\mathrm{CDF}^{-1}_{\mathcal{X}_1'^2(q\gamma_1)}(1 - \alpha)}\frac{\partial u}{\partial \gamma_0}|_{\gamma_0 = q\gamma_1} \tag{144}$$

$$= \frac{\mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial s}2q^2\sqrt{\frac{s}{q}}\frac{\partial u}{\partial \gamma_0}|_{\gamma_0 = q\gamma_1} \tag{145}$$

$$= \frac{\mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial s}2q^2\sqrt{\frac{s}{q}}\frac{1}{2\sqrt{\gamma_0}}\frac{\phi(u - \sqrt{\gamma_0}) - \phi(-u - \sqrt{\gamma_0})}{\phi(u - \sqrt{\gamma_0}) + \phi(-u - \sqrt{\gamma_0})}. \tag{146}$$

We can plug in $\gamma_0 = \gamma_1 q$, and $u = \frac{\sqrt{s}}{\sqrt{q}}$

$$= \frac{\mathrm{CDF}_{\mathcal{X}_1'^2(\gamma_1)}(s)}{\partial s}2q^2\sqrt{\frac{s}{q}}\frac{1}{2\sqrt{\gamma_1 q}}\frac{\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) - \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})}{\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) + \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})} \tag{147}$$

$$= \frac{1}{2\sqrt{s}}\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) + \phi(-\sqrt{s} - \sqrt{\gamma_1})\right)\frac{q\sqrt{s}}{\sqrt{\gamma_1}}\frac{\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) - \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})}{\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) + \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})} \tag{148}$$

$$= \frac{1}{2}\frac{q}{\sqrt{\gamma_1}}\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) + \phi(-\sqrt{s} - \sqrt{\gamma_1})\right)\frac{\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) - \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})}{\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) + \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})} \tag{149}$$

In total we obtain

$$\frac{\partial \beta_\alpha(\gamma_1)}{\partial \gamma_1} = -\frac{1}{2\sqrt{\gamma_1}}\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) - \phi(-\sqrt{s} - \sqrt{\gamma_1})\right) \tag{150}$$

$$+ \frac{1}{2}\frac{q}{\sqrt{\gamma_1}}\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) + \phi(-\sqrt{s} - \sqrt{\gamma_1})\right)\frac{\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) - \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})}{\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) + \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})} \tag{151}$$

$$= \frac{1}{2\sqrt{\gamma_1}}\left(-S_1 + qS_2\frac{T_1}{T_2}\right). \tag{152}$$

31

By inspection of the terms, we see that

$$S_1 = \phi(\sqrt{s} - \sqrt{\gamma_1}) - \phi(-\sqrt{s} - \sqrt{\gamma_1}) \geq 0 \tag{153}$$

because $|\sqrt{s} - \sqrt{\gamma_1}| \leq |-\sqrt{s} - \sqrt{\gamma_1}|$ and the normal density ϕ decreases with the absolute value of its argument. We can therefore rewrite the expression as

$$\frac{\partial \beta_\alpha(\gamma_1)}{\partial \gamma_1} = \frac{S_1}{2\sqrt{\gamma_1}} \left(-1 + q\frac{S_2 T_1}{S_1 T_2} \right) \tag{154}$$

and the sign is determined by the second factor (the first fraction $\frac{S_1}{2\sqrt{\gamma_1}}$ will be non-negative).

$$\frac{S_2 T_1}{S_1 T_2} = \frac{\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) + \phi(-\sqrt{s} - \sqrt{\gamma_1})\right)}{\left(\phi(\sqrt{s} - \sqrt{\gamma_1}) - \phi(-\sqrt{s} - \sqrt{\gamma_1})\right)} \frac{\left(\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) - \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})\right)}{\left(\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) + \phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})\right)} \tag{155}$$

$$= \frac{\Gamma - \left(\phi(\sqrt{s} - \sqrt{\gamma_1})\phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) - \phi(-\sqrt{s} - \sqrt{\gamma_1})\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) \right)}{\Gamma + \left(\phi(\sqrt{s} - \sqrt{\gamma_1})\phi(-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) - \phi(-\sqrt{s} - \sqrt{\gamma_1})\phi(\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q}) \right)} \tag{156}$$

$$= \frac{\Gamma - \Delta}{\Gamma + \Delta}, \tag{157}$$

where we have:

$$\Delta = \frac{1}{2\pi} \left(\exp\left(-\frac{(\sqrt{s} - \sqrt{\gamma_1})^2 + (-\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})^2}{2} \right) - \exp\left(-\frac{(-\sqrt{s} - \sqrt{\gamma_1})^2 + (\frac{\sqrt{s}}{\sqrt{q}} - \sqrt{\gamma_1 q})^2}{2} \right) \right) \tag{158}$$

$$= \frac{1}{2\pi} \left(\exp\left(-\frac{s - 2\sqrt{s}\sqrt{\gamma_1} + \gamma_1 + \frac{s}{q} + 2\frac{\sqrt{s}\sqrt{\gamma_1 q}}{\sqrt{q}} + \gamma_1 q}{2} \right) \right. \tag{159}$$

$$\left. - \exp\left(-\frac{s + 2\sqrt{s}\sqrt{\gamma_1} + \gamma_1 + \frac{s}{q} - 2\frac{\sqrt{s}\sqrt{\gamma_1 q}}{\sqrt{q}} + \gamma_1 q}{2} \right) \right) \tag{160}$$

$$= \frac{1}{2\pi} \left(\exp\left(-\frac{s + \gamma_1 + \frac{s}{q} + \gamma_1 q}{2} \right) - \exp\left(-\frac{s + \gamma_1 + \frac{s}{q} + \gamma_1 q}{2} \right) \right) \tag{161}$$

$$= 0. \tag{162}$$

Bringing this result ($\frac{S_2 T_1}{S_1 T_2} = 1$) back, we obtain

$$\frac{\partial \beta_\alpha(\gamma_1)}{\partial \gamma_1} = -\frac{S_1}{2\sqrt{\gamma_1}} (1 - q) \leq 0. \tag{163}$$

Derivative w.r.t. τ^2. We denote $\tau^2 = \tau_2$ to emphasize that we are deriving with respect to the square of τ:

$$\frac{\partial \beta_\alpha}{\partial \tau_2} = \underbrace{\frac{\partial \beta_\alpha}{\partial \gamma_1}}_{<0} \underbrace{\frac{\partial \gamma_1}{\partial \tau_2}}_{>0} + \underbrace{\frac{\partial \beta_\alpha}{\partial q}}_{>0} \underbrace{\frac{\partial q}{\partial \tau_2}}_{>0}. \tag{164}$$

We perform the following calculations:

$$\frac{\partial \beta_\alpha}{\partial \gamma_1} = -\frac{S_1}{2\sqrt{\gamma_1}} (1 - q) < 0 \tag{165}$$

$$\frac{\partial \gamma_1}{\partial \tau_2} = \frac{n^4 \mu^2}{\lambda_i^2} > 0 \tag{166}$$

$$\frac{\partial \beta_\alpha}{\partial q} = \frac{S_2 \sqrt{s}}{2q} + \frac{S_1 \sqrt{\gamma_1}}{2} > 0 \tag{167}$$

$$\frac{\partial q}{\partial \tau_2} = \frac{n^2}{\lambda_i (n + r)^2} > 0, \tag{168}$$

32

where $r = \frac{n^2 \tau_2}{\lambda_i}$. Plugging the terms together, we see that

$$\frac{\partial \beta_\alpha}{\partial \tau_2} = -\frac{S_1}{2\sqrt{\gamma_1}} (1-q) \frac{n^4 \mu^2}{\lambda_i^2} + \left(\frac{S_2 \sqrt{s}}{2q} + \frac{S_1 \sqrt{\gamma_1}}{2} \right) \frac{n^2}{\lambda_i (n+r)^2} \tag{169}$$

$$= \frac{n^2}{2\lambda_i} \left(-\frac{S_1}{\sqrt{\gamma_1}} (1-q) \frac{n^2 \mu^2}{\lambda_i} + \left(\frac{S_2 \sqrt{s}}{q} + S_1 \sqrt{\gamma_1} \right) \frac{1}{(n+r)^2} \right), \tag{170}$$

using $1 - q = \frac{1}{n + \frac{n^2 \tau^2}{\lambda_i}} = \frac{1}{n+r}$ we obtain

$$\frac{\partial \beta_\alpha}{\partial \tau_2} = \frac{n^2}{2\lambda_i (n+r)} \left(-\frac{S_1}{\sqrt{\gamma_1}} \frac{n^2 \mu^2}{\lambda_i} + \left(\frac{S_2 \sqrt{s}}{q} + S_1 \sqrt{\gamma_1} \right) \frac{1}{(n+r)} \right). \tag{171}$$

Further using $\frac{n^2 \mu^2}{\lambda_i^2} = \frac{\gamma_1}{n+r}$ we obtain

$$\frac{\partial \beta_\alpha}{\partial \tau_2} = \frac{n^2}{2\lambda_i (n+r)^2} \left(-S_1 \sqrt{\gamma_1} + \frac{S_2 \sqrt{s}}{q} + S_1 \sqrt{\gamma_1} \right) \tag{172}$$

$$= \frac{S_2 n^2 \sqrt{s}}{2q\lambda_i (n+r)^2} > 0. \tag{173}$$

F On the Relation between f-MIP and f-DP

F.1 Proof of Theorem 4.2

We first restate the theorem:

Theorem 4.2 (f-DP implies f-MIP). *Let an algorithm $\mathcal{A} : D^n \to \mathbb{R}^d$ be f-differentially private [11]. Then, the algorithm $\mathcal{A}$ will also be f-membership inference private.*

Proof. If the (probabilistic) algorithm $\mathcal{A}$ is f-DP, this requires that for all neighboring datasets S and S' (where only a single instance is changed), we have [11]

$$\text{Test}\left(A(S), A(S')\right) \geq f. \tag{174}$$

This includes any randomly sampled dataset $\mathcal{X} \in D^{n-1}$ plus an instance $\boldsymbol{x}$ as well as the dataset $\mathcal{X}$ plus the additional instance $\boldsymbol{x}'$ that is known to the attacker. Therefore setting

$$S = \mathcal{X} \cup \{\boldsymbol{x}\} \tag{175}$$

$$S' = \mathcal{X} \cup \{\boldsymbol{x}'\} \tag{176}$$

we immediately see that for every $\mathcal{X}, \boldsymbol{x}, \boldsymbol{x}'$, under f-DP, we have

$$\text{Test}\left(\mathcal{A}(\mathcal{X} \cup \{\boldsymbol{x}\}), \mathcal{A}(\mathcal{X} \cup \{\boldsymbol{x}'\})\right) \geq f. \tag{177}$$

Consulting our hypothesis test formulation of MI attacks, we note that the sets S, S' correspond to the inputs to A which result in output distributions A_0 and $A_1(\boldsymbol{x}')$ in the formulation of Equation (3), but for a specific $\boldsymbol{x}, \mathcal{X}$.

Additionally, as opposed to DP, in the membership inference attack scenario, the attacker has no knowledge of $\boldsymbol{x}, \mathcal{X}$. As the privacy constraint in Equation (177) holds for each individual choice of $\boldsymbol{x}, \mathcal{X}$, the test with randomly sampled instances cannot be easier than the test in Equation (177) (in practice, the test will be much harder when $\boldsymbol{x}, \mathcal{X}$ are stochastic as well), i.e.,

$$\forall \boldsymbol{x}' : \text{Test}\left(A_0; A_1(\boldsymbol{x}')\right) \geq f. \tag{178}$$

We can now leverage Theorem D.2 which highlights that when stochastically drawing from tests that are bounded by a certain trade-off function, the composed test can also be bounded by this function, which results in

$$\bigotimes_{\boldsymbol{x}' \sim \mathcal{D}} \text{Test}\left(A_0; A_1(\boldsymbol{x}')\right) \geq f, \tag{179}$$

the definition of f-MIP. $\qquad \square$

33

F.2 On the Correspondence between μ-DP and μ-MIP through noisy SGD

Our results on noisy SGD in Theorem 5.1 allow to translate between the parameters μ_{DP} of μ-GDP and μ_{MIP} of μ-GMIP, *when our algorithm is used to guarantee these privacy notions*. This relation can be made explicit through the following Corollary:

Corollary F.1 (Translating μ values between DP and MIP). *For one step of DP-SGD and the usual setting of $K = d$, we can convert the privacy parameters μ_{DP} and μ_{MIP} as follows:*

1. *If the step is μ_{DP}-GDP it will be μ_{MIP}-GMIP with*

$$\mu_{MIP} = \min\left\{ \sqrt{\frac{d}{n + \frac{4C^2}{\mu_{DP}^2} + \frac{1}{2}}}, \ \ \mu_{DP} \right\}.$$

2. *If the step is μ_{MIP}-GMIP it will be μ_{DP}-GDP with*

$$\mu_{DP} = \begin{cases} \dfrac{2}{\sqrt{\frac{d}{\mu_{MIP}^2} - n - \frac{1}{2}}}, & \text{if } \mu_{MIP} < \sqrt{\frac{2d}{2n+1}} \\ \infty, & \text{else.} \end{cases}$$

Proof. The result can be seen by solving for the required noise level τ. When guaranteeing μ_{DP}-GDP, we have

$$\mu_{\text{DP}} = \frac{1}{\sigma} = \frac{2C}{n\tau} \tag{180}$$

by [11]. Additionally, for MIP ($K = d$) we have

$$\mu_{\text{MIP}} = \frac{2n_{\text{effective}}d}{n_{\text{effective}}\sqrt{2d + 4n_{\text{effective}}d}} = \sqrt{\frac{2d}{2n_{\text{effective}} + 1}}. \tag{181}$$

(1) We can solve Equation (180) for τ arriving at

$$\tau = \frac{2C}{n\mu_{\text{DP}}} \tag{182}$$

and plug it into Equation (181) to arrive at the first term of the $\min$ in statement 1. We note that due to Theorem 4.2, the level of μ_{MIP} cannot be higher than the level of μ_{DP} and thus take the min of both terms.

(2) To arrive at statement 2, we perform the opposite conversion and solve Equation (181) for τ. Solving for $n_{\text{effective}}$, we obtain

$$n_{\text{effective}} = n + \frac{n^2\tau^2}{C^2} = \frac{d}{\mu_{\text{MIP}}^2} - \frac{1}{2}. \tag{183}$$

As we have $\tau^2 > 0$ and therefore require $n_{\text{effective}} > n$, no solution exists for $\mu_{\text{MIP}}^2 < \frac{2d}{2n+1}$. For the remaining values, we obtain

$$\tau^2 = C^2 \frac{\frac{d}{\mu_{\text{MIP}}^2} - \frac{1}{2} - n}{n^2} \tag{184}$$

which, plugged into Equation (180), gives rise to the result in statement 2. $\qquad\square$

34