Why Does Differential Privacy with Large ε Defend Against Practical Membership Inference Attacks?

Andrew Lowy[1], **Zhuohang Li**[2], **Jing Liu**[3], **Toshiaki Koike-Akino**[3], **Kieran Parsons**[3], **Ye Wang**[3]

[1]University of Wisconsin-Madison
[2]Vanderbilt University
[3]Mitsubishi Electric Research Laboratories
alowy@wisc.edu, zhuohang.li@vanderbilt.edu, {jiliu, koike, parsons, yewang}@merl.com

Abstract

For "small" privacy parameter ε (e.g. $\varepsilon < 1$), ε-differential privacy (DP) provides a strong *worst-case* guarantee that no *membership inference attack* (MIA) can succeed at determining whether a person's data was used to train a machine learning model. The guarantee of DP is *worst-case* because: a) it holds even if the attacker already knows the records of all but one person in the data set; and b) it holds uniformly over all data sets. In practical applications, such a worst-case guarantee may be overkill: *practical* attackers may lack exact knowledge of (nearly all of) the private data, and our data set might be easier to defend, in some sense, than the worst-case data set. Such considerations have motivated the industrial deployment of DP models with large privacy parameter (e.g. $\varepsilon \geqslant 7$), and it has been observed empirically that DP with large ε can successfully defend against state-of-the-art MIAs. Existing DP theory cannot explain these empirical findings: e.g., the theoretical privacy guarantees of $\varepsilon \geqslant 7$ are essentially vacuous. In this paper, we aim to close this gap between theory and practice and understand *why a large DP parameter can prevent practical MIAs*. To tackle this problem, we propose a new privacy notion called *practical membership privacy* (PMP). PMP models a practical attacker's uncertainty about the contents of the private data. The PMP parameter has a natural interpretation in terms of the success rate of a practical MIA on a given data set. We quantitatively analyze the PMP parameter of two fundamental DP mechanisms: the exponential mechanism and Gaussian mechanism. Our analysis reveals that a large DP parameter often translates into a much smaller PMP parameter, which guarantees strong privacy against *practical* MIAs. Using our findings, we offer principled guidance for practitioners in choosing the DP parameter.

Introduction

Machine learning (ML) systems, such as large language models (LLMs), have the potential to transform various facets of society and industry. However, the growing ubiquity of these systems raises privacy concerns and a long line of work has demonstrated how to *attack* ML models and uncover private details about individuals whose data was used to train the model. For example, (Carlini et al. 2021) extracted individual training examples by querying an LLM.

Membership inference attacks (MIAs) (Shokri and Shmatikov 2015; Dwork et al. 2015) are a fundamental class of privacy attacks. An MIA receives a trained model $\mathcal{A}(D)$ and a target data point x as inputs and aims to infer whether or not the target point was used to train the model (i.e. whether or not $x \in D$). In this paper, we focus on white-box attackers who know the (randomized) algorithm $\mathcal{A}$. MIAs can violate people's privacy: for example, genomic data sets may contain information about people with a particular medical diagnosis, and knowing that someone is in the data set reveals that they have the diagnosis (Homer et al. 2008). Moreover, MIAs are often used as building blocks for other attacks, such as training data extraction attacks (Carlini et al. 2021). Thus, if we can prevent MIAs, we can often also prevent other attacks.

For sufficiently small δ and $\varepsilon \geqslant 0$, (ε, δ)-*differential privacy* (DP) (Dwork et al. 2006) guarantees that no MIA can succeed with high probability, by requiring that the output distribution of the ML model be insensitive to the presence or absence of any individual data point (see Definition 1). Pure ε-DP bounds the probability that an arbitrary MIA can succeed by $1/(1 + e^{-\varepsilon})$. Thus, for example, $\varepsilon \leqslant 0.1$ implies that no MIA can do much better than randomly guessing (52.5%) whether or not a target data point was used to train the model. However, the guarantee of DP degrades rapidly with ε, e.g., if $\varepsilon \geqslant 7$, then an ε-DP algorithm is potentially vulnerable to MIAs that succeed with probability $\geqslant 0.999$. On the other hand, large ε values of $\varepsilon \geqslant 7$ are often deployed in industrial applications (Apple 2016; Úlfar Erlingsson, Pihur, and Korolova 2014; Ding, Kulkarni, and Yekhanin 2017; Desfontaines 2021). Moreover, values of $\varepsilon \geqslant 8$ have been shown empirically to be highly effective at thwarting state-of-the-art MIAs (Carlini et al. 2022). Existing theory cannot adequately explain the empirical success of ε-DP with large ε at defending against MIAs.

This paper aims to bridge this gap between theory and practice. We rigorously address the following question:

Why does DP with large ε defend against practical MIAs?

Contributions To answer this question, we begin with the observation that differential privacy provides a guarantee against a *worst-case* MIA, which holds uniformly for all data sets. Namely, *DP ensures that even an attacker with knowledge of $n-1$ data points $D\backslash\{x\}$ cannot infer whether*

or not the target point x was used as input to $\mathcal{A}$. On the other hand, *practical* attackers typically do not have such fine-grained knowledge of the underlying data set as the worst-case attacker that DP models. Indeed, the literature on MIAs typically assumes that the attacker has some knowledge of the data *distribution* (e.g. query access to the distribution or knowledge of a subpopulation from which the data was randomly drawn), but does not know any of the points in the given data set with certainty. The attacker must rely on the output of the training algorithm $\mathcal{A}(D)$ and distributional knowledge to infer membership of the target point x.

We model this practical MIA setting in our definition of *practical membership privacy* (PMP, Definition 2). We show that PMP is a useful notion of privacy: PMP is weaker than the strong worst-case notion of DP (Proposition 5), but strong enough to guarantee that no practical MIA can succeed with high probability (Lemma 6). Moreover, PMP is not susceptible to the blatant privacy breaches that afflict other weakenings of DP that have been defined in the literature (see Related Work and Appendix).

We analyze the relationship between the PMP and DP parameters for two popular DP mechanisms: the exponential mechanism (McSherry and Talwar 2007) and the Gaussian mechanism (Dwork et al. 2006). We show that the PMP parameter can be much smaller than the DP parameter for these mechanisms, e.g., the ε-DP exponential mechanism satisfies $\varepsilon/75$-PMP for certain subpopulations. This helps explain why large values of ε can provide strong protection against practical MIAs: for example, the $(\varepsilon = 7.5)$-DP exponential mechanism lacks meaningful privacy guarantees against a worst-case attacker, but the resulting $(\varepsilon/75 = 0.1)$-PMP guarantee ensures that no practical MIA can succeed with probability much higher than random guessing (52.5%). We conclude by discussing the implications of our results for practitioners in choosing the DP parameter, and highlighting interesting directions for future work.

Differential Privacy

Definition 1 (Differential Privacy (Dwork et al. 2006))**.** *Let $\varepsilon \geqslant 0$, $\delta \in [0, 1)$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Z}$ is (ε, δ)-differentially private (DP) if for all pairs of adjacent data sets $D, D' \in \mathcal{X}^n$ and all measurable subsets $S \subseteq \mathcal{Z}$, we have*

$$\mathbb{P}(\mathcal{A}(D) \in S) \leqslant e^{\varepsilon}\mathbb{P}(\mathcal{A}(D') \in S) + \delta,$$

where the probability is solely over the randomness of $\mathcal{A}$. If $\delta = 0$, we say that $\mathcal{A}$ satisfies "pure DP" and write ε-DP. If $\delta > 0$, we say "approximate DP" and write (ε, δ)-DP.

Practical Membership Privacy

In this section, we define a privacy notion—called practical membership privacy (PMP)—that models the practical MIA setting.

PMP models a membership inference attacker who does not know any elements of D^* with certainty, but has some distributional knowledge of D. Specifically, we assume that the attacker knows *a "parent set" $X \in \mathcal{X}^{2n}$ from which D was drawn uniformly at random*[1]. One can interpret the parent set X as representing a subpopulation from which the data was known to be drawn (e.g., health insurance customers or hospital patients) or a dataset (e.g., MNIST) consisting of training samples D and test samples $X \backslash D$. PMP ensures that such an attacker cannot succeed in correctly determining membership of any target point with high probability:

Definition 2 (Practical Membership Privacy[2])**.** *Let $\varepsilon \geqslant 0$, $\delta \in [0, 1)$ and $X \in \mathcal{X}^{2n}$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Z}$ satisfies (ε, δ)-practical membership privacy (PMP) with respect to X if for all $x \in X$ and all measurable subsets $S \subseteq \mathcal{Z}$, we have*

$$\begin{aligned} e^{-\varepsilon}\left(\mathbb{P}(\mathcal{A}(D) \in S | x \notin D) - \delta\right) \\ \leqslant \mathbb{P}(\mathcal{A}(D) \in S | x \in D) \\ \leqslant e^{\varepsilon}\mathbb{P}(\mathcal{A}(D) \in S | x \notin D) + \delta, \end{aligned}$$

*where the probability is taken both over the random draw of $D \sim$ **Unif**$(\{E \subset X : |E| = n\})$ and the randomness of $\mathcal{A}$. $\mathcal{A}$ is (ε, δ)-PMP if $\mathcal{A}$ is (ε, δ)-PMP with respect to X for all $X \in \mathcal{X}^{2n}$. To denote "pure PMP", when $\delta = 0$, we will simply write ε-PMP as a shorthand for $(\varepsilon, 0)$-PMP.*

The key differences between the PMP model and the DP model are: 1) our (practical) attacker only has partial information about the other $n - 1$ samples in D, whereas DP allows the (worst-case) attacker to know the other $n-1$ samples with certainty; and 2) our definition is dependent on the parent data set X, whereas DP holds uniformly over all data sets. Our assumption on the attacker's knowledge is more realistic than the DP assumption in many private data analysis settings: In practice, it is uncommon that an attacker knows $n - 1$ points in a data set (but not the n-th point). However, it is often the case that an attacker knows that the data set was drawn from some sub-population $X \subset \mathcal{X}$; Definition 2 models an attacker with this knowledge.

The following lemma provides alternative characterizations of PMP:

Lemma 3. *Let $X \in \mathcal{X}^{2n}$, $x \in X$, $X_{in}(x) := \{D \subset X : |D| = n, x \in X\}$, and $X_{out}(x) = \{D \subset X : |D| = n, x \notin X\}$. Let $S \subset \mathcal{Z}$ be a measurable set. If*

$$\begin{aligned} e^{-\varepsilon}\left(\mathbb{P}(x \notin D | \mathcal{A}(D) \in S) - \delta\right) \\ \leqslant \mathbb{P}(x \in D | \mathcal{A}(D) \in S) \\ \leqslant e^{\varepsilon}\mathbb{P}(x \notin D | \mathcal{A}(D) \in S) + \delta, \end{aligned} \tag{1}$$

then

$$\begin{aligned} e^{-\varepsilon}\left(\mathbb{P}(\mathcal{A}(D) \in S | x \notin D) - 2\delta\right) \\ \leqslant \mathbb{P}(\mathcal{A}(D) \in S | x \in D) \\ \leqslant e^{\varepsilon}\mathbb{P}(\mathcal{A}(D) \in S | x \notin D) + 2\delta. \end{aligned} \tag{2}$$

[1]The choice of $2n$ as the size of the parent set is for analytical convenience. Our analysis extends to the case where, e.g., X contains αn points for some $\alpha > 1$.

[2]To simplify some of our analyses, we will assume that X consists of $2n$ distinct points, w.l.o.g: If there are repeated points, then we can re-define X without repeats for some smaller n.

Also, (2) *holds iff*

$$e^{-\varepsilon}\left(\frac{1}{N}\sum_{D'\in X_{out}(x)}\mathbb{P}_{\mathcal{A}}(\mathcal{A}(D')\in S)-\delta\right)$$
$$\leqslant \frac{1}{N}\sum_{D\in X_{in}(x)}\mathbb{P}_{\mathcal{A}}(\mathcal{A}(D)\in S)$$
$$\leqslant e^{\varepsilon}\left(\frac{1}{N}\sum_{D'\in X_{out}(x)}\mathbb{P}_{\mathcal{A}}(\mathcal{A}(D')\in S)\right)+\delta, \quad (3)$$

where $N := |X_{in}(x)| = |X_{out}(x)| = \binom{2n}{n}/2$ *and the probabilities in* (3) *are taken solely over the randomness of* $\mathcal{A}$.

Moreover, if $\delta = 0$*, then* (1) *holds iff* (2) *holds iff* (3) *holds. Thus,* $\mathcal{A}$ *is* ε*-PMP w.r.t.* X *iff any of these three inequalities holds for all* $x \in X$ *and all* $S \subset \mathcal{Z}$.

Proofs are deferred to the Appendix. A consequence of the equivalence between (2) and (3) is that if $n = 1$, then ε-PMP and ε-DP are equivalent—and satisfy ε*-local differential privacy* (Kasiviswanathan et al. 2011):

Corollary 4. *If* $n = 1$*, then* $\mathcal{A}$ *is* (ε, δ)*-DP iff* $\mathcal{A}$ *is* $(\varepsilon, 2\delta)$*-PMP w.r.t.* X *for every* $X \in \mathcal{X}^{2n}$.

For $n > 1$, PMP is weaker than DP. For simplicity, we present this result for $\delta = 0$:

Proposition 5. *If* $\mathcal{A}$ *is* ε*-DP, then* $\mathcal{A}$ *is* ε*-PMP. Moreover, if* $n > 2$*, then there exists an* $\ln(2)$*-PMP* $\mathcal{A}$ *that is not* ε'*-DP for any* $\varepsilon' < \infty$.

Intuitively, Proposition 5 is true because the inequalities in (3) involve averages over data points in X, rather than the worst-case supremum appearing in the definition of DP. Moreover, the data set might not be worst case for PMP. The averages correspond to the practical attacker's uncertainty about which samples are in D, which makes it harder to infer membership of x than the worst-case DP attacker. Also, the $\ln(2)$-PMP parameter in Proposition 5 is not tight, as our construction for $n = 3$ can be extended to get an ε-PMP algorithm with $\varepsilon < \ln(2)$ for $n > 3$, e.g., one can get $\varepsilon \leqslant \ln(40/37) < 0.08$ for $n = 6$.

Next, we bound the success probability of a *practical MIA* (as defined at the beginning of this section) in terms of the PMP parameter:

Lemma 6. *Let* $\mathcal{A}$ *be* ε*-PMP with respect to* X *and* $\mathcal{M}$ *be any practical MIA. Then, the probability that* $\mathcal{M}$ *successfully infers membership, for any* $x \in X$*, never exceeds* $1/(1 + e^{-\varepsilon})$.

Analogously, it is well-known that ε-DP ensures that that success probability of the *worst-case* attacker (who knows all but one sample of D) never exceeds $1/(1 + e^{-\varepsilon})$.

In the Appendix, we record additional basic properties of PMP, such as post-processing.

Related Work

Some prior works have sought to understand why large ε effectively prevents practical privacy attacks from various different angles. Most of these approaches seek to weaken the assumptions on the attacker in some respect. See Ghazi et al. (2022, Section 7) for a thorough discussion of different directions in which weaker assumptions on the attacker may be imposed. Below, we list these directions and cite a few related works for each.

Assumptions about the attacker's capabilities: DP assumes that the attacker has unlimited computational resources and is capable of executing any sort of attack. Some relaxations of DP, such as computational DP (Mironov et al. 2009), model an attacker with limited computational resources. Other privacy notions (e.g., k-anonymity) model an attacker that only executes a specific type of attack (e.g., record-linkage attack). In contrast to these works, our PMP notion models an attacker with the same vast capabilities as the DP attacker.

Assumptions about the attacker's goals: DP protects against membership inference attacks, which is equivalent (up to a factor of 2 in ε) to an attacker learning an arbitrary one-bit function of the target individual's data. Some works have considered a modified attacker with more ambitious goals (e.g., training data reconstruction (Hayes, Mahloujifar, and Balle 2023)). Other works have relaxed the DP definition to consider an attacker that only aims to extract certain bits of information from the target individual, e.g., attribute-level partial DP (Ghazi et al. 2022). In contrast to these works, our work considers an attacker with the same goals as the DP attacker. Thus, the attacker that we model is stronger along the "goals" axis than these prior works.

Assumptions about the attacker's knowledge: DP permits an attacker to know everything about the data set except for one private bit that they aim to infer. Several works have sought to model the uncertainty that a practical attacker has about the contents of the data set, e.g., (Bassily et al. 2013; Li et al. 2013; Yeom et al. 2018; Sablayrolles et al. 2019; Humphries et al. 2020; Izzo et al. 2022; Leemann, Pawelczyk, and Kasneci 2023).

Similarly, our PMP notion models an attacker with weaker knowledge than the DP attacker. PMP has advantages over previously proposed privacy notions that model the attacker's uncertainty. For example, as we discuss in the Appendix, many previously proposed definitions can be satisfied by algorithms that leak the data of some members of the data set and are therefore not (intuitively) private. By contrast, PMP is not susceptible to these blatant privacy violations. Moreover, the focus of our work—on precisely understanding the risk of a privacy breach with a practical (uncertain) attacker against specific DP algorithms—is different from these prior works.

In the Appendix, we discuss prior works seeking to weaken assumptions about the attacker's knowledge in more detail. We highlight pathologies with previously proposed definitions, in which algorithms that clearly leak an individual's data can still satisfy these other definitions. Also, in contrast to some other works, PMP does not impose any distributional or independence assumptions on the underlying data. Instead, we allow for data to be drawn from an arbitrary subpopulation X. This makes our analysis harder, but also makes our definition and results stronger. Finally,

we reiterate that prior works did not provide the quantitative interpretations of practical privacy guarantees of concrete DP mechanisms that our work provides. In this work, **we give quantitative bounds relating the DP parameter ε to the PMP parameter** and a **precise interpretation of the guarantees of our PMP notion against any practical attacker** (Lemma 6). Together, these results enable a **rigorous interpretation of the privacy guarantees of ε-DP against a practical (less knowledgeable) attacker**.

Practical Privacy Guarantees of the Exponential Mechanism

In this section, we characterize the practical membership privacy of one of the most powerful and versatile differentially private algorithms: the *exponential mechanism* (McSherry and Talwar 2007). To define the exponential mechanism, let $\mathcal{W}$ be a finite set of objects.[3] Let $\ell : \mathcal{W} \times \mathcal{X}^n \to \mathbb{R}$ be some loss function. Given data D, our goal is to privately select an object $w \in \mathcal{W}$ that approximately minimizes the loss function.

Definition 7 (Exponential Mechanism)**.** *Given inputs $D, \mathcal{W}, \ell$, the exponential mechanism $\mathcal{A}_E$ selects and outputs some object $w \in \mathcal{W}$. The probability that a particular w is selected is proportional to* $\exp\left(\frac{-\varepsilon \ell(w,D)}{2\Delta_\ell}\right)$*, where* $\Delta_\ell = \max_{w\in\mathcal{W}} \sup_{D\sim D';D,D'\in\mathcal{X}^n} |\ell(w,D) - \ell(w,D')|$.

Lemma 8. *(McSherry and Talwar 2007) The exponential mechanism is ε-DP.*

The following proposition gives an exact description of the PMP parameter as a function of the DP parameter ε:

Proposition 9. *Let $X \in \mathcal{X}^{2n}$. The ε-DP exponential mechanism is $\tilde{\varepsilon}(X)$-PMP with respect to X if and only if*

$$\tilde{\varepsilon}(X) \geqslant \ln\left[\frac{\sum_{D\in X_{in}(x)} c(D)\exp\left(-\frac{\varepsilon}{2\Delta_\ell}\ell(w,D)\right)}{\sum_{D'\in X_{out}(x)} c(D')\exp\left(-\frac{\varepsilon}{2\Delta_\ell}\ell(w,D')\right)}\right]$$

for all $w \in \mathcal{W}$ and $x \in X$, where $c(D) = \left[\sum_{w'\in\mathcal{W}} \exp\left(\frac{-\varepsilon\ell(w',D)}{2\Delta_\ell}\right)\right]^{-1}$ and $c(D')$ is defined similarly.

For a given loss function ℓ and subpopulation X, Proposition 9 allows us to compute the PMP paramater $\tilde{\varepsilon}(X)$ of the exponential mechanism as a function of ε. In combination with Lemma 6, this will allow us to interpret ε in terms of the success rate of an arbitrary practical membership inference attacker.

Numerical Simulations We investigate the PMP parameter $\tilde{\varepsilon}$ vs. the DP parameter ε for different subpopulations X. We fix the loss function: $\ell(w, D) = \frac{1}{n}\sum_{i=1}^{n} \|w - D_i\|_2$, which is a convex empirical risk minimization problem corresponding to the geometric median. Our goal is to understand the *ratio* $\tilde{\varepsilon}(X)/\varepsilon$ that we get for different X, and different factors that affect the ratio (e.g., the distribution and dimension of the data). We choose $\mathcal{W} = \{w_1, \ldots, w_m\}$ to be a set of m random standard normal unit vectors in $\mathbb{R}^d$, standardized to have unit ℓ_2-norm. We then draw $X \sim \mathcal{N}(w_1, \sigma^2)^{2n\times d}$ and clip the ℓ_2 norm of each data point, so $\|x_i\|_2 \leqslant C$ for all $i \in [2n]$, where C is the clip threshold.

Recall that there are two key differences between PMP and DP: one difference lies in the attacker's knowledge/uncertainty about the data, and the second is that PMP is defined with respect to a subpopulation X, whereas DP is worst-case over all X. In order to disentangle these two effects, we plot two curves in each experiment: the (average, over T trials) ratios $\tilde{\varepsilon}(X)/\varepsilon$ and the ratio$_X$ $\tilde{\varepsilon}(X)/\varepsilon(X)$. Here $\varepsilon(X)$ is defined as in Definition 1 except that that we only require the inequality to hold for adjacent data sets D, D' that are subsets of X, rather than $\mathcal{X}^n = \{x \in \mathbb{R}^d : \|x\|_2 \leqslant C\}^n$. The ratio$_X$ $\tilde{\varepsilon}(X)/\varepsilon(X)$ controls for the effect of the data and just describes the effect of the practical attacker's uncertainty compared to the worst-case DP attacker's certainty about members of $D\backslash\{x\}$. The ratio $\tilde{\varepsilon}(X)/\varepsilon$ captures the role of both the attacker's knowledge and the data being potentially easier to defend than the worst-case data set.

Figure 1 shows the ratios $\tilde{\varepsilon}(X)/\varepsilon$ and $\tilde{\varepsilon}(X)/\varepsilon(X)$ vs. the standard deviation σ of the data. Note that for small σ, the data is easier to defend/harder to attack because everyone in the data set looks similar: the attacker cannot easily distinguish between the output distribution of the algorithm when the target $x \in D$ vs. when $x \notin D$. Conversely, large σ makes it likely that some "outlier" x that is easier for the attacker to identify will be in X. Thus, the ratio $\tilde{\varepsilon}(X)/\varepsilon$ increases with σ. On the other hand, the ratio $\tilde{\varepsilon}(X)/\varepsilon(X)$ does not significantly depend on σ.

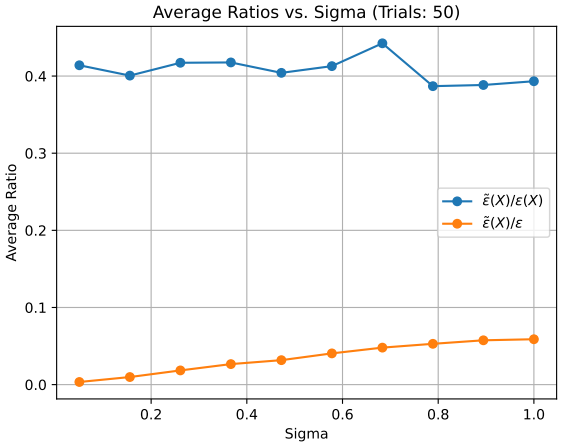

Figure 1: Ratios vs. σ, with 1-dim. data, $n = 6$, $m = 10$, $C = 10$, $\varepsilon(X) = 5$.

For example, when $\sigma = 1$ (standard normal data), the ratio $\tilde{\varepsilon}(X)/\varepsilon \approx 0.075$, which mostly reflects the fact that this data set is far from worst case. In this case, $\varepsilon(X) = 5$ and $\varepsilon \approx 28.5$, which does not afford any meaningful privacy guarantees under classical DP theory. However, the PMP pa-

[3] If $\mathcal{W}$ is infinite, then the exponential mechanism can still be applied after discretizing $\mathcal{W}$.

rameter $\tilde{\varepsilon}(X) \approx 2.14$, which provides a meaningful guarantee against practical MIAs on this particular subpopulation X, by Lemma 6. Moreover, for small $\sigma < 1$, the smaller ratios imply stronger PMP guarantees for fixed values of ε: e.g. for $\sigma < .1$, the PMP parameter $\tilde{\varepsilon}(X)$ approaches zero.

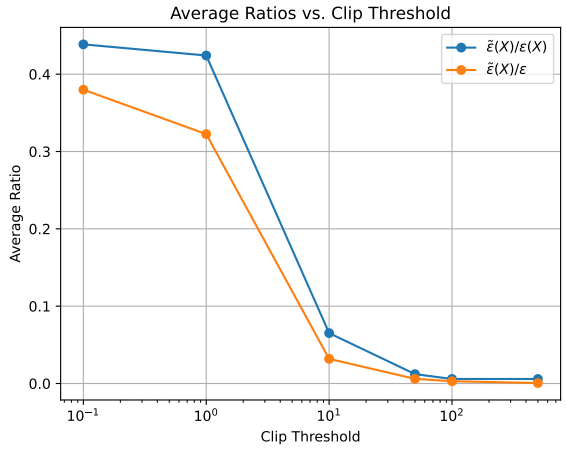

Figure 2: Ratios vs. Clip threshold C, with 5-dim. data, $n = 6$, 2 outliers, $m = 32$, $\sigma = 1$, $\varepsilon(X) = 10$.

Figure 2 shows the effect of clip threshold on the ratios. A small clip threshold C reduces the effect of outlier data points, while a large clip threshold C permits more outliers in the data set. To amplify the effect of outliers, we choose 2 points in X at random and multiply them by 100. These extreme outliers cause $\varepsilon(X)$ (and ε) to be much larger than $\tilde{\varepsilon}(X)$, since the worst-case DP attacker who knows an outlier in $D\backslash\{x\}$ can use this information to easily infer membership of x. By contrast, the practical attacker cannot use outliers to launch an MIA as effectively because they are uncertain about which other points are in D. For example, when $C = 50$, both ratios are less than 0.0123. This means that a 10-DP algorithm with no meaningful privacy guarantee against a worst-case attacker satisfies 0.123-PMP and hence can defend against any practical attacker almost perfectly ($1/(1+e^{-.123}) \approx 0.53$). Moreover, the DP parameter $\varepsilon \approx \varepsilon(X)$ in the presence of extreme outliers because such X is nearly worst-case from a privacy perspective. In this experiment, $n = 6$ since the runtime of computing $\tilde{\varepsilon}(X)$ is exponential in n. We would expect the ratios to become even smaller for larger n because the practical attacker's uncertainty would increase.

Finally, Figure 3 shows that the ratios become smaller as the dimension of the data increases. This can be attributed to the particular choice of loss function and Euclidean geometry in higher dimensions. In general, the effect of dimension on the ratios will depend on the loss function/problem.

Practical Privacy Guarantees of the Gaussian Mechanism

This section analyzes the practical membership privacy of one of the most widely used (ε, δ)-DP algorithms: the *Gaussian Mechanism*. Given a function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian mechanism simply adds isotropic Gaussian noise to the output of q:

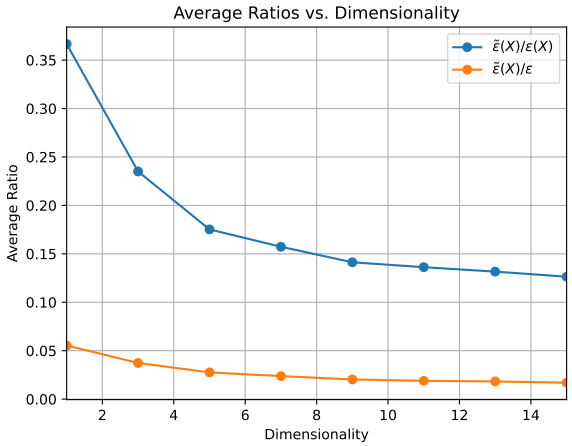

Figure 3: Ratios vs. Dimension of data d, with $n = 6$, $m = 10$, $\sigma = 1$, $\varepsilon(X) = 2$.

$$\mathcal{A}_G(D) := q(D) + \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right).$$

Denote the cumulative distribution function of $Y \sim \mathcal{N}(0,1)$ by Φ.

Lemma 10. *(Balle and Wang 2018) Let $q : \mathcal{X}^n \to \mathbb{R}^d$ be a function with global ℓ_2-sensitivity $\Delta = \sup_{D \sim D'} \|q(D) - q(D')\|_2$. For any $\varepsilon \geqslant 0$ and $\delta \in [0,1]$, the Gaussian mechanism $\mathcal{A}_G$ is (ε, δ)-DP if and only if*

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) \leqslant \delta. \qquad (4)$$

For $\mathcal{A} := \mathcal{A}_G$ and $x \in X$, define the mixture distributions $\mathbb{P}_{\text{in},x}(S) := \frac{1}{|X_{\text{in}}(x)|}\sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S)$ and $\mathbb{P}_{\text{out},x}(S) := \frac{1}{|X_{\text{out}}(x)|}\sum_{D \in X_{\text{out}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S)$. Our analysis will utilize the following characterizations of DP and PMP, which are immediate from the definitions:

Lemma 11. *Denote the hockey-stick divergence between random variables P and Q by $D_{e^\varepsilon}(P\|Q) := \int_{\mathbb{R}} \max\{0, p(t) - e^\varepsilon q(t)\}dt$, where p and q denote the probability density or mass functions of P and Q respectively. Then, $\mathcal{A}$ is (ε, δ)-DP if and only if $\max\{D_{e^\varepsilon}(\mathcal{A}(D)\|\mathcal{A}(D')), D_{e^\varepsilon}(\mathcal{A}(D')\|\mathcal{A}(D))\} \leqslant \delta$ for all $D \sim D'$. Moreover, $\mathcal{A}$ is (ε, δ)-PMP w.r.t. X if and only if $\max\{D_{e^\varepsilon}(\mathbb{P}_{in,x}\|\mathbb{P}_{out,x}), D_{e^\varepsilon}(\mathbb{P}_{out,x}\|\mathbb{P}_{in,x})\} \leqslant \delta$ for all $x \in X$.*

The following technical result will be crucial in our analysis.

Proposition 12. *Let $\mathcal{A}_G$ be the (ε, δ)-DP Gaussian mecha-*

nism. Then, for any $x \in X$,

$$\max\{D_{e^\varepsilon}(\mathbb{P}_{in,x}||\mathbb{P}_{out,x}), D_{e^\varepsilon}(\mathbb{P}_{out,x}||\mathbb{P}_{in,x})\} \quad (5)$$
$$\leqslant \frac{1}{n|X_{in}(x)|} \sum_{D\in X_{in}(x)} \sum_{D'\in X_{out}(x), D'\sim D} \left[\Phi\left(\frac{\|q(D)-q(D')\|}{2\sigma} - \frac{\varepsilon\sigma}{\|q(D)-q(D')\|}\right) - e^\varepsilon \Phi\left(-\frac{\|q(D)-q(D')\|}{2\sigma} - \frac{\varepsilon\sigma}{\|q(D)-q(D')\|}\right)\right].$$

The main tools used in the proof of Proposition 12 are joint convexity of the hockey-stick divergence (which holds since D_{e^ε} is an f-divergence) and a bound on $D_{e^\varepsilon}(\mathcal{A}_G(D)||\mathcal{A}_G(D')$ due to (Balle and Wang 2018).

By Proposition 12 and Lemma 11, $\mathcal{A}_G$ is (ε, δ)-PMP if the right-hand side of inequality 5 is upper-bounded by δ. The differences between this sufficient condition for PMP and the condition (4) for DP is that (4) is worst-case over all pairs of adjacent data sets in $\mathcal{X}^n$, whereas PMP only requires an average-case bound over all adjacent subsets of X.

Our Approach Our approach for analyzing the PMP parameter $\tilde{\varepsilon}(X)$ for the (ε, δ)-DP Gaussian mechanism is as follows:

1. Given target DP parameters (ε, δ), find the approximately smallest σ such that the Gaussian mechanism is (ε, δ) via Lemma 10 and (Balle and Wang 2018, Algorithm 1).
2. Upper bound the hockey-stick divergence between $\mathbb{P}_{\text{in},x}$ and $\mathbb{P}_{\text{out},x}$ in Proposition 12.
3. Using the value of σ obtained in step 1), find the approximately smallest $\tilde{\varepsilon}(X)$ such that our upper bound in Proposition 12 is $\leqslant \delta$ for all $x \in X$: this ensures that the Gaussian mechanism is $(\tilde{\varepsilon}(X), \delta)$-PMP w.r.t. X, by Lemma 11.

Note that a naive implementation of step 3 would run in exponential (in n) time. To execute step 3 efficiently, we observe that the right-hand-side of Inequality 5 can be greatly simplified when the function is of the form $q(D) = \sum_{x\in D} f(x)$, where f is some sample-wise function. Since, the summation is constrained to be over D containing x and D' that is adjacent to D, where x is replaced with a different x', the value of $q(D) - q(D')$ is equal to $f(x) - f(x')$. Thus, the terms of the summation are a function of only x' (given that x is fixed), with each possible $x' \neq x$ repeatedly appearing an equal number of times. Hence, instead of dealing with the average-case over all adjacent datasets, we can compute an equivalent average over all choices of $x' \neq x$, given by

$$\frac{1}{2n-1}\sum_{x'\neq x}\left[\Phi\left(\frac{\|f(x)-f(x')\|}{2\sigma} - \frac{\varepsilon\sigma}{\|f(x)-f(x')\|}\right) - e^\varepsilon\Phi\left(-\frac{\|f(x)-f(x')\|}{2\sigma} - \frac{\varepsilon\sigma}{\|f(x)-f(x')\|}\right)\right].$$

Numerical Simulations For our simulations, we consider empirical mean estimation: $q(D) = \sum_{x\in D} x/n$. The goals of these simulations are the same as in the simulations of the previous section: to quantify the ratios $\tilde{\varepsilon}(X)/\varepsilon$ and $\tilde{\varepsilon}(X)/\varepsilon(X)$ and understand the factors that cause these ratios to be large or small. We draw an i.i.d. Gaussian data set $X \sim \mathcal{N}(0, \sigma^2)^{2n\times d}$ and clip the ℓ_2 norm of each data point, so $\|x_i\|_2 \leqslant C$ for all $i \in [2n]$, in order to bound global sensitivity of q.

Figure 4 shows the ratios vs. the DP parameter $\varepsilon(X)$. First, note that the ratio $\tilde{\varepsilon}(X)/\varepsilon$ is small for all values of $\varepsilon(X)$. For example, even when $\varepsilon(X) = 10$ and $\tilde{\varepsilon}(X)/\varepsilon$ is at its largest, we still have a small PMP parameter $\tilde{\varepsilon}(X) < 0.9$. Second, we see that there is a large gap between the two (orange and blue) curves, especially when $\varepsilon(X)$ is large. This indicates that the worst-case DP parameter ε is significantly bigger than the subpopulation-specific DP parameter $\varepsilon(X)$ in this experiment. Thus, X is far from being worst-case. Third, the ratios increase with the DP parameter $\varepsilon(X)$.

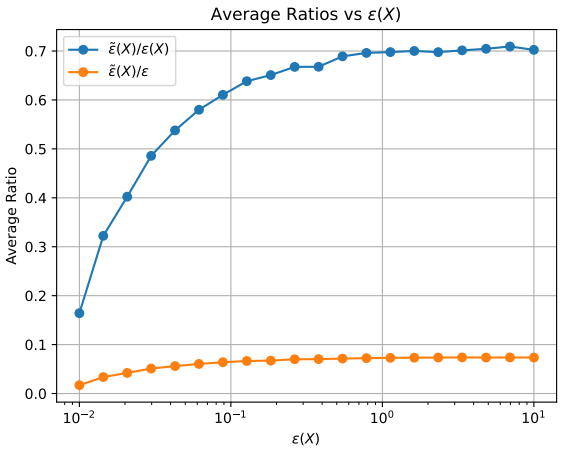

Figure 4: Ratios vs. $\varepsilon(X)$, with $n = 100$, $d = 20$, $C = 50$, no outliers, $\sigma = 1$, $\delta = 10^{-2}$.

Figure 5 shows the effect of the clip threshold C on the ratios in the presence of outliers. We produce outliers by choosing 2 points at random and scaling them by a factor of 10. Similar to Figure 2, we see that the ratios shrink as the clip threshold C increases. For example, for large $C = 100$, a DP parameter of $\varepsilon = 5$ would translate into a much smaller PMP parameter of $\tilde{\varepsilon}(X) = 1$. One difference between Figure 5 and Figure 2 is that the gap between the blue and orange curves is larger in Figure 5 than in Figure 2. The reason is that the data X is relatively easier to keep private in the experiment that was used to produce 5, whereas X was nearly worst-case in Figure 2. This is due to differences in the outlier scaling, dimension, σ, and the loss function/learning problem.

Figure 6 shows that the ratios increase with the dimension of the data. In combination with Figure 3, we see that the effect of dimension on the ratios may differ substantially for different learning problems. Thus, practitioners may want to

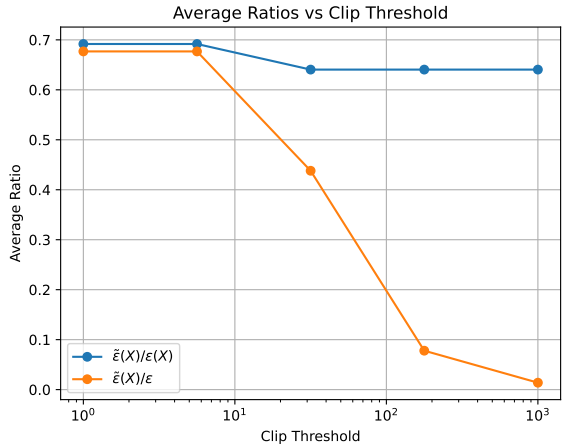

Figure 5: Ratios vs. Clip threshold C, with $n = 100$, $d = 10$, 2 outliers, $\sigma = 5$, $\delta = 10^{-2}$.

apply problem-specific context to guide the choice of ε.

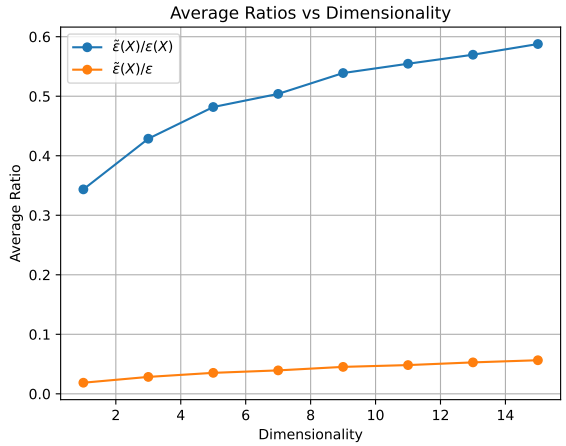

Figure 6: Ratios vs. Dimensionality, with ε with $n = 100$, $C = 50$, no outliers, $\sigma = 1$, $\delta = 10^{-2}$.

Discussion and Conclusion

In this paper, we analyzed the risk of data leakage of DP algorithms against a practical attacker who lacks certainty about the contents of the data set. At a high level, our results are encouraging: we rigorously show that even at larger ε, DP mechanisms can actually provide guaranteed defense against practical MIAs.

We also gleaned more granular insights. For example, Figure 1 indicates that if a data analyst has *a priori* knowledge that the subpopulation from which data is drawn is approximately i.i.d./homogeneous, then they can afford to choose larger ε: homogeneous data is easier to keep private. Also, data sets containing extreme outliers make it relatively much easier for a worst-case MIA to attack than for a practical MIA (e.g., see Figure 2). Strategies like aggressive clipping can be used to mitigate the negative effects of outliers on privacy. Practitioners can use our code (which we plan to make available online) to help choose an appropriate ε for their particular problem/data population, while aiming to get a small corresponding PMP parameter, e.g., $\tilde{\varepsilon}(X) \leqslant 0.1$.

We emphasize that our motivation for studying the notion of PMP was to better understand DP; we do not advocate for using PMP as a substitute for DP. PMP has certain shortcomings: As discussed in Ghazi et al. (2022, Section 7), an attacker's level of uncertainty may decrease over time, e.g., due to subsequent releases of information. Consequently, PMP does not satisfy the same sequential composition property that DP satisfies. We hope that by providing clearer interpretations of the DP parameter in terms of vulnerability to practical MIAs, our work facilitates more widespread use of DP algorithms in industry and government.

References

Apple. 2016. Differential Privacy Overview.

Balle, B.; and Wang, Y.-X. 2018. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, 394–403. PMLR.

Bassily, R.; Groce, A.; Katz, J.; and Smith, A. 2013. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 439–448. IEEE.

Bhaskar, R.; Bhowmick, A.; Goyal, V.; Laxman, S.; and Thakurta, A. 2011. Noiseless database privacy. In *Advances in Cryptology–ASIACRYPT 2011: 17th International Conference on the Theory and Application of Cryptology and Information Security, Seoul, South Korea, December 4-8, 2011. Proceedings 17*, 215–232. Springer.

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramer, F. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1897–1914. IEEE.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T. B.; Song, D.; Erlingsson, U.; et al. 2021. Extracting Training Data from Large Language Models. In *USENIX Security Symposium*, volume 6.

Desfontaines, D. 2021. A list of real-world uses of differential privacy. https://desfontain.es/privacy/real-world-differential-privacy.html. Ted is writing things (personal blog).

Ding, B.; Kulkarni, J.; and Yekhanin, S. 2017. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30.

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

Dwork, C.; Smith, A.; Steinke, T.; Ullman, J.; and Vadhan, S. 2015. Robust traceability from trace amounts. In *2015*

IEEE 56th Annual Symposium on Foundations of Computer Science, 650–669. IEEE.

Ghazi, B.; Kumar, R.; Manurangsi, P.; and Steinke, T. 2022. Algorithms with More Granular Differential Privacy Guarantees. *arXiv preprint arXiv:2209.04053*.

Hayes, J.; Mahloujifar, S.; and Balle, B. 2023. Bounding Training Data Reconstruction in DP-SGD. *arXiv preprint arXiv:2302.07225*.

Homer, N.; Szelinger, S.; Redman, M.; Duggan, D.; Tembe, W.; Muehling, J.; Pearson, J. V.; Stephan, D. A.; Nelson, S. F.; and Craig, D. W. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8): e1000167.

Humphries, T.; Oya, S.; Tulloch, L.; Rafuse, M.; Goldberg, I.; Hengartner, U.; and Kerschbaum, F. 2020. Investigating membership inference attacks under data dependencies. *arXiv preprint arXiv:2010.12112*.

Izzo, Z.; Yoon, J.; Arik, S. O.; and Zou, J. 2022. Provable Membership Inference Privacy. *arXiv preprint arXiv:2211.06582*.

Kasiviswanathan, S. P.; Lee, H. K.; Nissim, K.; Raskhodnikova, S.; and Smith, A. 2011. What can we learn privately? *SIAM Journal on Computing*, 40(3): 793–826.

Kifer, D.; and Machanavajjhala, A. 2012. A rigorous and customizable framework for privacy. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, 77–88.

Leemann, T.; Pawelczyk, M.; and Kasneci, G. 2023. Gaussian Membership Inference Privacy. *arXiv preprint arXiv:2306.07273*.

Li, N.; Qardaji, W.; and Su, D. 2012. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, 32–33.

Li, N.; Qardaji, W.; Su, D.; Wu, Y.; and Yang, W. 2013. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 889–900.

Long, Y.; Bindschaedler, V.; and Gunter, C. A. 2017. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136*.

Mahloujifar, S.; Sablayrolles, A.; Cormode, G.; and Jha, S. 2022. Optimal membership inference bounds for adaptive composition of sampled gaussian mechanisms. *arXiv preprint arXiv:2204.06106*.

McSherry, F.; and Talwar, K. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94–103. IEEE.

Mironov, I.; Pandey, O.; Reingold, O.; and Vadhan, S. 2009. Computational differential privacy. In *Annual International Cryptology Conference*, 126–142. Springer.

Sablayrolles, A.; Douze, M.; Schmid, C.; Ollivier, Y.; and Jégou, H. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, 5558–5567. PMLR.

Sason, I.; and Verdú, S. 2016. f-divergence inequalities. *arXiv preprint arXiv:1508.00335*.

Shokri, R.; and Shmatikov, V. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1310–1321.

Úlfar Erlingsson; Pihur, V.; and Korolova, A. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security*.

Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, 268–282. IEEE.

More details on related works

In this Appendix, we discuss prior works seeking to weaken assumptions about the attacker's knowledge in more detail. We highlight pathologies with previously proposed definitions, in which algorithms that clearly leak an individual's data can still satisfy these other definitions. Also, in contrast to some other works, PMP does not impose any distributional or independence assumptions on the underlying data. Instead, we allow for data to be drawn from an arbitrary subpopulation X. This makes our analysis harder, but also makes our definition and results stronger. Finally, we re-iterate that prior works did not provide the quantitative interpretations of practical privacy guarantees of concrete DP mechanisms that our work provides.

The work of Bassily et al. (2013) was motivated by similar goals to our own. They propose *distributional DP* (DDP), a special case of their more general "coupled-worlds privacy" framework. DDP utilizes a "simulator" in its definition, requiring that the output distribution of the algorithm $\mathcal{A}(D)$ be (ε, δ)-indistinguishable from the output distribution of some simulator run on the "scrubbed" data $\textit{Sim}(D_{-i})$, for all $i \in [n]$. The paper shows that certain noiseless protocols (e.g. real-valued summation, histograms, stable functions) satisfy DDP w.r.t. certain distribution classes $\mathcal{P}$. Moreover, they discuss the relation of DDP to previous notions of privacy—namely, Pufferfish privacy (Kifer and Machanavajjhala 2012) and noiseless privacy (Bhaskar et al. 2011).

However, the DDP definition suffers from a shortcoming, which does not occur with our PMP definition. Under fairly mild distribution classes, DDP can permit pathological algorithms that simply release the entire dataset. Let $\mathcal{P}$ consist of distributions on datasets D of size n, where knowledge of any $n-1$ points reveals the remaining point. In such a case, the algorithm $\mathcal{A}(D) = D$ becomes permissible, as it is perfectly ($\epsilon = \delta = 0$) indistinguishable from a simulator $\textit{Sim}(D_{-i})$ that can also output the entire dataset, by recovering the missing point. For example, let $D_1, \ldots, D_n$ be uniformly drawn from binary sequences with even parity, i.e., for any $i \in [n]$, we have $D_i = \oplus_{j \in [n]: j \neq i} D_j$. Note that this is a simple modification of Example 1 from Bassily et al. (2013), but with the auxiliary side information Z (meant to reveal the parity of the binary sequence) omitted, and instead the fixed parity is incorporated into the distribution of the binary sequence. This also applies to generalizations of this example with distributions where the sum (or mean) of the dataset is fixed and known.

Li et al. (2013) proposes a notion of membership privacy that is similar in spirit to DDP. Roughly speaking, an algorithm satisfies the (positive) membership privacy notion of (Li et al. 2013) w.r.t. a family of distributions $\mathcal{P}$ on $\mathcal{X}^n$ if $\mathbb{P}_{P,\mathcal{A}}(x \in X | \mathcal{A}(X) \in S) \approx \mathbb{P}_P(x \in X)$ for all $P \in \mathcal{P}, x \in \mathcal{X}, S \subset \mathbb{A}$. Conceptually, the essential differences between this definition and our definition of PMP are: that our definition is parameterized by a parent data set, whereas theirs is parameterized by a family of distributions that correspond to the attacker's prior knowledge; also, their definition is non-symmetric in positive vs. negative membership inference, whereas our definition is symmetric. To address this latter limitation, (Li et al. 2013) introduce a second definition of negative membership privacy that protects against attacks that determine that someone was *not* a member of the training data. Having two definitions seems unnecessary and our framework eliminates this need. They prove post-processing property of their membership privacy notion. The paper concludes by giving different instantiations of membership privacy for different choices of $\mathcal{P}$ and recovering prior notions of privacy (including DP) along the way. In particular, Li et al. (2013, Theorem 5.10) shows that their membership privacy definition recovers "DP under sampling" (Li, Qardaji, and Su 2012) for the distribution family $\mathcal{P}_\beta$ consisting of distributions such that $P(x) \in \{0, \beta\}$ for some choice of β. An algorithm satisfies "DP under sampling" if it is DP when composed with the subsampling operation that first samples each point in the data set with probability β and then executes the algorithm on the subsampled data set. A drawback of (Li et al. 2013) is that the privacy parameter of their definition is not analyzed carefully or related to the DP parameter. We address this drawback in our work.

The work of Long, Bindschaedler, and Gunter (2017) proposes *differential training privacy* (DTP) to empirically estimate the privacy risk of publishing a classifier. Their DTP definition is specifically given for classifiers that output a vector of probabilities for predicted labels y and features x: $p_{\mathcal{A}(T)}(y|x)$, where T is the training data set. Thus, their DTP notion is also data set-specific. Essentially, their definition requires that the predicted label probabilities of $\mathcal{A}$ do not change too much when any single point in the training data set is removed: $p_{\mathcal{A}(T)}(y|x) \leqslant e^{\varepsilon} p_{\mathcal{A}(T \setminus z)}(y|x)$ should hold for all $z \in T$ and all feature-label pairs (x, y) in the universe. Thus, their definition seems to be conceptually more similar to DP than it is to our definition of PMP. They provide an efficiently computable approximation of DTP that they compute in empirical case studies. They use these case studies to reason about the privacy risks of non-DP classifiers trained on certain data sets. No theoretical treatment of their DTP notion is provided.

The work of Yeom et al. (2018) proposes a different distribution-dependent definition of membership privacy based on the following membership experiment: data $S \sim P^n$ is drawn i.i.d. from some distribution and a learning algorithm $\mathcal{A}(S)$ is run on the training data. Then a random bit $b \sim Ber(1/2)$ is drawn. If $b = 0$, then we draw a point $z \in S$ at random. If $b = 1$, then we draw a random point $z \sim P$. The attacker observes the target point z and the output of the algorithm $\mathcal{A}(S)$ (and implicitly has knowledge of P) and tries to guess the value of b (i.e., membership of z). They define the membership advantage of an attacker in terms of its success rate, and say an algorithm is membership private (w.r.t. P) if every attacker has small membership advantage. Note that this membership experiment is the one that Carlini et al. (2022) assume in their attack model. Compared to our PMP notion, a critical difference is that their definition only protects the privacy of the people in the data set *on average* (over the random draw of z from S). By contrast, our definition provides a stronger *worst-case* (over $z \in S$) guarantee, ensuring

that the data of *every person* in S remains private. Another difference is that Yeom et al. (2018) uses a parent distribution P, whereas we use a parent data set X. Yeom et al. (2018)'s definition is conceptually similar to DDP, but the precise way it is measured (in terms of advantage) differs and also it is framed as an experiment with an MIA.

Yeom et al. (2018) shows that DP implies bounded membership advantage and studies the connection between overfitting and membership advantage. Additionally, they look at the connection between membership inference and attribute inference.

The work of Humphries et al. (2020) proposed a variation of the definition in (Yeom et al. 2018) to deal with a specific limitation of (Yeom et al. 2018)'s definition. Namely, Humphries et al. (2020) argues that the i.i.d. data assumption is problematic because DP guarantees become much weaker in the presence of data dependencies and because the assumption may not be satisfied in practice. Thus, they modify the definition in Yeom et al. (2018) by assuming that P is a mixture of K distributions: first, $k \sim [K]$ is drawn uniformly and then $S \sim P_k^n$ is drawn (conditionally i.i.d. given k). If $b = 1$, then the target point z is drawn from the mixture distribution: first $k' \sim [K]$ is drawn and then $z \sim P_{k'}$. Note that this modification allows for data dependencies.

Humphries et al. (2020) provides tighter bounds on the relation between DP and membership advantage, compared with (Yeom et al. 2018). They also empirically evaluate membership inference with data dependencies. Again, the main difference between our notion and Humphries et al. (2020) is that we use a parent set instead of a parent distribution. Note that our definition also permits data dependencies, since the parent data set may consist of dependent data.

The work of Sablayrolles et al. (2019) defines a training algorithm that returns a parameter θ as being (ε, δ)-membership private w.r.t. a loss function $\ell(\theta, z)$. Their definition Sablayrolles et al. (2019, Definition 3) essentially requires a membership private algorithm to satisfy $\ell(\theta, z_1) \approx \int_w \ell(t, z_1) p_T(w) dw$ with high probability over the random draw of the training data set $T = (z_1, \ldots, z_n)$. Here $p_T(w)$ is the posterior density of the parameter w given $(z_2, \ldots, z_n)$, which is assumed to take a particular form given in (Sablayrolles et al. 2019, Definition 12). Roughly speaking, it is assumed that w depends on the data through an "exponential mechanism"-like training algorithm. An immediate problem with their definition is the dependence on the loss function, which greatly reduces the generality and flexibility of the definition. (Sablayrolles et al. 2019) characterize the optimal MIA under certain assumptions discussed above. They show that DP implies a bound on the membership advantage. They run experiments showing that their attack—based on the theoretically optimal attack under their assumption on the posterior—performs well.

The work of Mahloujifar et al. (2022) is motivated by the desire to get tighter bounds on membership inference privacy for existing algorithms. They measure membership inference privacy by using a very strong definition of membership privacy that is similar to DP in that it assumes (implicitly) that the attacker knows the other $n - 1$ points in the training set. As we argue, this assumption is usually unrealistic and a major benefit of our PMP definition is that it relaxes this assumption by modeling the adversary's uncertainty about the training data set.

The recent work of Izzo et al. (2022) works towards a theory of membership inference privacy (MIP). Their notion of η-MIP is similar to our notion of ε-PMP in terms of being average case over the uniformly random draw of the training data, and worst-case over outcomes. However, their MIP notion is fundamentally weaker than our PMP notion. In particular, it is easy to see that the following blatantly non-private algorithm satisfies η-MIP but does not satisfy ε-PMP for any $\varepsilon < \infty$: $\mathcal{A}$ releases a training example D_1 with probability $n\eta$ and otherwise outputs *NULL*. Thus, their MIP notion may not be strong enough to offer the meaningful and intuitive membership privacy guarantees that we desire. Moreover, PMP implies MIP, as the following lemma shows:

Lemma 13. *If $\mathcal{A}$ is ε-PMP, then $\mathcal{A}$ is $\frac{1-e^{-\varepsilon}}{2}$-MIP.*

Proof. Let $\Delta = 1 - e^{-\varepsilon} \in [0, 1)$. Assume for concreteness that $\mathcal{A}$ is discrete. (A similar argument works if $\mathcal{A}$ is continuous.) Then since $\mathcal{A}$ is ε-PMP, we have

$$1 - \Delta \leqslant \frac{\mathbb{P}(x \in D | \mathcal{A}(D) = a)}{\mathbb{P}(x \notin D | \mathcal{A}(D) = a)} \leqslant \frac{1}{1 - \Delta}$$

for almost every $a \in \mathcal{Z}$, where $\mathcal{Z}$ denotes the range of $\mathcal{A}(D)$. By the proof of (Izzo et al. 2022, Theorem 7), we get

$$\max\left(\mathbb{P}(x \in D | \mathcal{A}(D) = a), \mathbb{P}(x \notin D | \mathcal{A}(D) = a)\right) \leqslant \frac{1 + \Delta}{2}$$

and

$$\begin{aligned}\int_{\mathcal{Z}} \max\left(\mathbb{P}(x \in D | \mathcal{A}(D) = a), \mathbb{P}(x \notin D | \mathcal{A}(D) = a)\right) \mathbb{P}(\mathcal{A}(D) = a) &\leqslant \frac{1 + \Delta}{2} \int_{\mathcal{Z}} P(\mathcal{A}(D) = a) \\ &= \frac{1 + \Delta}{2}.\end{aligned}$$

By the definition of η-MIP, the above inequality implies that $\mathcal{A}$ is η-MIP for $\eta = \Delta/2 = (1 - e^{-\varepsilon})/2$. □

Finally, the concurrent and independent work of Leemann, Pawelczyk, and Kasneci (2023) proposes a Gaussian-DP analog of the membership inference privacy (MIP) notion. They show how to implement their Gaussian MIP with noisy SGD and give a novel MIA based on their MIP notion.

Proofs of Theoretical Results

In this Appendix, we re-state and prove our theoretical results. First, we show that PMP satisfies post-processing.

Lemma 14 (Post-processing property of PMP)**.** *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Z}$ be (ε, δ)-PMP. If $f : \mathcal{Z} \to \mathcal{Y}$ is any function, then $f \circ \mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-PMP.*

Proof. Let $S \subset \mathcal{Y}$ be measurable, $X \in \mathcal{X}^{2n}$, $x \in X$, $N = |X_{\text{in}}(x)| = |X_{\text{out}}(x)|$, where $X_{\text{in}}(x)$ and $X_{\text{out}}(x)$ are defined in Lemma 3. Assume w.l.o.g. that f is deterministic. (If f is randomized, then we can reduce to the deterministic case by considering convex combinations.) Let $T_S := \{z \in \mathcal{Z} : f(z) \in S\} = f^{-1}(S)$. Note that for any $D \in X_{\text{in}}(x)$, there exists a $D' \in X_{\text{out}}(x)$ that is adjacent to D: if $x = D_i$, take $D' = (D_1, \ldots, D_{i-1}, x', D_{i+1}, \ldots, D_n)$ for some $x' \in X \backslash D$. Then, by Lemma 3, we have

$$\begin{aligned}
\frac{1}{N} \sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(f \circ \mathcal{A}(D) \in S) &= \frac{1}{N} \sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in T_S) \\
&\leqslant \frac{1}{N} \sum_{D' \in X_{\text{out}}(x)} e^{\varepsilon} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in T_S) + \frac{\delta}{2} \\
&= \delta/2 + e^{\varepsilon} \frac{1}{N} \sum_{D' \in X_{\text{out}}(x)} \mathbb{P}_{\mathcal{A}}(f \circ \mathcal{A}(D') \in S).
\end{aligned}$$

By Lemma 3, we conclude that $f \circ \mathcal{A}$ is (ε, δ)-PMP. □

Lemma 15 (Re-statement of Lemma 3)**.** *Let $X \in \mathcal{X}^{2n}$, $x \in X$, $X_{in}(x) := \{D \subset X : |D| = n, x \in X\}$, and $X_{out}(x) = \{D \subset X : |D| = n, x \notin X\}$. Let $S \subset \mathcal{Z}$ be a measurable set. If*

$$e^{-\varepsilon} \left(\mathbb{P}(x \notin D | \mathcal{A}(D) \in S) - \delta\right) \leqslant \mathbb{P}(x \in D | \mathcal{A}(D) \in S) \leqslant e^{\varepsilon} \mathbb{P}(x \notin D | \mathcal{A}(D) \in S) + \delta, \tag{6}$$

then

$$e^{-\varepsilon} \left(\mathbb{P}(\mathcal{A}(D) \in S | x \notin D) - 2\delta\right) \leqslant \mathbb{P}(\mathcal{A}(D) \in S | x \in D) \leqslant e^{\varepsilon} \mathbb{P}(\mathcal{A}(D) \in S | x \notin D) + 2\delta. \tag{7}$$

Also, (7) *holds iff*

$$e^{-\varepsilon} \left(\frac{1}{N} \sum_{D' \in X_{out}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S) - \delta \right) \leqslant \frac{1}{N} \sum_{D \in X_{in}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S) \leqslant e^{\varepsilon} \left(\frac{1}{N} \sum_{D' \in X_{out}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S) \right) + \delta, \tag{8}$$

where $N := |X_{in}(x)| = |X_{out}(x)| = \binom{2n}{n}/2$ and the probabilities in (8) *are taken solely over the randomness of $\mathcal{A}$.*

Moreover, if $\delta = 0$, then (6) *holds iff* (7) *holds iff* (8) *holds. Thus, $\mathcal{A}$ is ε-PMP w.r.t. X iff any of these three inequalities holds for all $x \in X$ and all $S \subset \mathcal{Z}$.*

Proof. Suppose (6) holds. Then, by Bayes' rule and the fact that $\mathbb{P}(x \in D) = \mathbb{P}(x \notin D) = 1/2$, we have

$$e^{-\varepsilon} \left(-\delta + \frac{\mathbb{P}(\mathcal{A}(D) \in S | x \notin D)}{2\mathbb{P}(\mathcal{A}(D) \in S)} \right) \leqslant \frac{\mathbb{P}(\mathcal{A}(D) \in S | x \in D)}{2\mathbb{P}(\mathcal{A}(D) \in S)} \leqslant e^{\varepsilon} \frac{\mathbb{P}(\mathcal{A}(D) \in S | x \notin D)}{2\mathbb{P}(\mathcal{A}(D) \in S)} + \delta. \tag{9}$$

Multiplying (9) by $2\mathbb{P}(\mathcal{A}(D) \in S)$ and using the fact that $\mathbb{P}(\mathcal{A}(D) \in S) \in [0, 1]$ yields (7).

Next we prove the equivalence between (7) and (8). Observe that

$$\begin{aligned}
\mathbb{P}(\mathcal{A}(D) \in S | x \in D) &= \mathbb{P}(\mathcal{A}(D) \in S | D \in X_{\text{in}}(x)) && (10) \\
&= \frac{\mathbb{P}(\mathcal{A}(D) \in S, D \in X_{\text{in}}(x))}{\mathbb{P}(D \in X_{\text{in}}(x))} && (11) \\
&= \frac{\frac{1}{N} \sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S)}{1/2} && (12) \\
&= \frac{2}{N} \sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S). && (13)
\end{aligned}$$

Similarly, $\mathbb{P}(\mathcal{A}(D) \in S | x \notin D) = \frac{2}{N} \sum_{D' \in X_{\text{out}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S)$. Substituting these equalities into (7) and then dividing by 2 yields (8).

Now suppose $\delta = 0$. Then we have already shown that (6) implies (7) and that (7) is equivalent to (8). Conversely, if (7) holds, then by Bayes rule and the fact that $\mathbb{P}(x \in D) = \mathbb{P}(x \notin D) = 1/2$, we get

$$e^{-\varepsilon} 2\mathbb{P}(x \notin D | \mathcal{A}(D) \in S)\mathbb{P}(\mathcal{A}(D) \in S) \leqslant 2\mathbb{P}(x \in D | \mathcal{A}(D) \in S)\mathbb{P}(\mathcal{A}(D) \in S) \leqslant e^{\varepsilon} 2\mathbb{P}(x \notin D | \mathcal{A}(D) \in S)\mathbb{P}(\mathcal{A}(D) \in S).$$

If $\mathbb{P}(\mathcal{A}(D) \in S) > 0$, then dividing the above by $2\mathbb{P}(\mathcal{A}(D) \in S)$ implies that (6) holds. This completes the proof. □

Corollary 16 (Re-statement of Corollary 4)**.** *If $n = 1$, then $\mathcal{A}$ is (ε, δ)-DP iff $\mathcal{A}$ is $(\varepsilon, 2\delta)$-PMP w.r.t. X for every $X \in \mathcal{X}^{2n}$.*

Proof. If $n = 1$, then $N = 1$ and the sums in (3) are each only over one term. Thus, (3) holds for all $X = \{x, x'\} \in \mathcal{X}^2$ iff

$$e^{-\varepsilon}(\mathbb{P}(\mathcal{A}(x') \in S) - \delta) \leqslant \mathbb{P}(\mathcal{A}(x) \in S) \leqslant e^{\varepsilon}\mathbb{P}(\mathcal{A}(x') \in S) + \delta \tag{14}$$

iff $\mathcal{A}$ is (ε, δ)-DP. By Theorem 3, this condition is also equivalent to $\mathcal{A}$ being $(\varepsilon, 2\delta)$-PMP w.r.t. X for every $X \in \mathcal{X}^2$. □

Proposition 17 (Re-statement of Proposition 5)**.** *If $\mathcal{A}$ is ε-DP, then $\mathcal{A}$ is ε-PMP. Moreover, if $n > 2$, then there exists an $\ln(2)$-PMP $\mathcal{A}$ that is not ε'-DP for any $\varepsilon' < \infty$.*

Proof. The first statement is a consequence of Lemma 3 and uses arguments from the proof of (Izzo et al. 2022, Proposition 6). Let $X \in \mathcal{X}^{2n}$ consist of $2n$ distinct points, let $x \in X$ and $S \subset \mathcal{Z}$. Let $\mathcal{A}$ be ε-DP. By Lemma 3, we have

$$\frac{\mathbb{P}(\mathcal{A}(D) \in S | x \notin D)}{\mathbb{P}(\mathcal{A}(D) \in S | x \in D)} = \frac{\sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S)}{\sum_{D' \in X_{\text{out}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S)}.$$

Now, for any $D = (D_1, \ldots, D_n) \in X_{\text{in}}(x)$, there is a unique $i \in [n]$ such that $D_i = x$. Let $x' \in X \backslash D$ and $D' := (D_1, \ldots, D_{i-1}, x', D_{i+1}, \ldots, D_n)$, which is a neighboring data set of D (i.e. $D \sim D'$) and $D' \in X_{\text{out}}(x)$. Note that there are n choices for x'. Thus, we can see that D has n neighboring data sets in $X_{\text{out}}(x)$. Similarly, every $D' \in X_{\text{out}}(x)$ has n neighbors in $X_{\text{in}}(x)$. Thus,

$$n \sum_{D' \in X_{\text{out}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S) = \sum_{D \in X_{\text{in}}(x)} \sum_{\substack{D' \in X_{\text{out}}(x) \\ D' \sim D}} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S).$$

This implies

$$\begin{aligned}
\frac{\mathbb{P}(\mathcal{A}(D) \in S | x \notin D)}{\mathbb{P}(\mathcal{A}(D) \in S | x \in D)} &= \frac{\sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S)}{\sum_{D' \in X_{\text{out}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S)} \\
&= \frac{\sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S)}{\frac{1}{n} \sum_{D \in X_{\text{in}}(x)} \sum_{D' \in X_{\text{out}}(x), D' \sim D} \mathbb{P}(\mathcal{A}(D') \in S)} \\
&\leqslant \frac{e^{\varepsilon} \sum_{D \in X_{\text{in}}(x)} \min_{D' \in X_{\text{out}}(x), D' \sim D} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S)}{\sum_{D \in X_{\text{in}}(x)} \text{Average}_{D' \in X_{\text{out}}(x), D' \sim D} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S)} \\
&\leqslant e^{\varepsilon}.
\end{aligned}$$

A similar argument proves the other inequality.

For the second statement, assume for simplicity that $n = 3$. It will be easy to see that our construction extends to $n > 3$ (and indeed the PMP parameter can be reduced for $n > 3$, giving a stronger result). Let $\mathcal{X} = \{0, 1, 2, 3, 4, 5\}$, and $X = (0, 1, 2, 3, 4, 5)$. Define $\mathcal{A}(D) = \text{sum}(D) \pmod 6$ as the modular addition operator. First, $\mathcal{A}$ is clearly not ε'-DP for any $\varepsilon' < \infty$ since $\mathcal{A}$ is not randomized. Concretely, if $\mathcal{A}(D) = 0$ for some $D \in \mathcal{X}^n$, then replacing D_1 by $D_1' = D_1 + 1 \pmod 6$ and letting $D' = (D_1', D_2, \ldots, D_n)$ implies that $\mathcal{A}(D') = 1$; hence the privacy loss is infinite.

Next, $\mathcal{A}$ is $\ln(2)$-PMP. To see this, let $x = 0$ and compute $\max_{a \in \mathcal{X}} \sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) = a) = \max_{a \in \mathcal{X}} |\{D \in X_{\text{in}}(x) : \sum(D) = a \pmod 6\}| = 2$. On the other hand, $\min_{a \in \mathcal{X}} \sum_{D \in X_{\text{out}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) = a) = \min_{a \in \mathcal{X}} |\{D \in X_{\text{out}}(x) : \sum(D) = a \pmod 6\}| = 1$. By symmetry and Lemma 3, $\mathcal{A}$ is ε-PMP with respect to X if and only if

$$\left| \ln \frac{\sum_{D \in X_{\text{in}}(x)} \mathbb{P}(\mathcal{A}(D) = a)}{\sum_{D' \in X_{\text{out}}(x)} \mathbb{P}(\mathcal{A}(D') = a)} \right| \leqslant \varepsilon$$

for all $a \in \mathcal{X}$. By the above computations, we see that this holds only if $\varepsilon \geqslant \ln(2)$. □

Lemma 18 (Re-statement of Lemma 6)**.** *Let $\mathcal{A}$ be ε-PMP with respect to X and $\mathcal{M}$ be any practical MIA. Then, the probability that $\mathcal{M}$ successfully infers membership, for any $x \in X$, never exceeds $1/(1 + e^{-\varepsilon})$.*

Proof. Suppose $\mathbb{P}(\mathcal{A}(D) \in S | x \in D) = e^{\varepsilon}\mathbb{P}(\mathcal{A}(D') \in S | x \notin D')$ for some $S \subset \mathcal{Z}$ and $x \in X$. Then any practical MIA's success probability is upper bounded by $\max\left(\mathbb{P}(x \in D|\mathcal{A}(D) \in S), \mathbb{P}(x \notin D|\mathcal{A}(D) \in S)\right)$, which corresponds to the success probability of the Bayes optimal practical MIA, $\mathcal{M}^*$. Assume w.l.o.g. that $\max\left(\mathbb{P}(x \in D|\mathcal{A}(D) \in S), \mathbb{P}(x \notin D|\mathcal{A}(D) \in S)\right) = \mathbb{P}(x \in D|\mathcal{A}(D) \in S)$. Then,

$$\begin{aligned}
\mathbb{P}(\mathcal{M}^* \text{ is correct}) &\leqslant \mathbb{P}(x \in D|\mathcal{A}(D) \in S) \\
&= \frac{\mathbb{P}(\mathcal{A}(D) \in S|x \in D)\mathbb{P}(x \in D)}{\mathbb{P}(\mathcal{A}(D) \in S)} \\
&= \frac{e^{\varepsilon}\mathbb{P}(\mathcal{A}(D) \in S|x \notin D) \times 1/2}{(1/2)\left(\mathbb{P}(\mathcal{A}(D) \in S|x \in D) + \mathbb{P}(\mathcal{A}(D) \in S|x \notin D)\right)} \\
&= \frac{e^{\varepsilon}}{1+e^{\varepsilon}} = \frac{1}{1+e^{-\varepsilon}}.
\end{aligned}$$

□

Proposition 19 (Re-statement of Proposition 9)**.** *Let $X \in \mathcal{X}^{2n}$. The ε-DP exponential mechanism is $\tilde{\varepsilon}(X)$-PMP with respect to X if and only if*

$$\tilde{\varepsilon}(X) \geqslant \ln\left[\frac{\sum_{D \in X_{in}(x)} c(D)\exp\left(-\frac{\varepsilon}{2\Delta_\ell}\ell(w,D)\right)}{\sum_{D' \in X_{out}(x)} c(D')\exp\left(-\frac{\varepsilon}{2\Delta_\ell}\ell(w,D')\right)}\right]$$

for all $w \in \mathcal{W}$ and $x \in X$, where $c(D) = \left[\sum_{w' \in \mathcal{W}}\exp\left(\frac{-\varepsilon\ell(w',D)}{2\Delta_\ell}\right)\right]^{-1}$ and $c(D')$ is defined similarly.

Proof. By Lemma 9, the ε-DP exponential mechanism is $\tilde{\varepsilon}(X)$-PMP with respect to X if and only if

$$\begin{aligned}
\tilde{\varepsilon} &\geqslant \max_{w \in \mathcal{W}, x \in X} \ln\left[\frac{\sum_{D \in X_{\text{in}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) = w)}{\sum_{D' \in X_{\text{out}}(x)} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') = w}\right] \\
&= \max_{w \in \mathcal{W}, x \in X} \ln\left[\frac{\sum_{D \in X_{\text{in}}(x)} c(D)\exp\left(-\frac{\varepsilon}{2\Delta_\ell}\ell(w,D)\right)}{\sum_{D' \in X_{\text{out}}(x)} c(D')\exp\left(-\frac{\varepsilon}{2\Delta_\ell}\ell(w,D')\right)}\right].
\end{aligned}$$

□

Proposition 20 (Re-statement of Proposition 12)**.** *Let $\mathcal{A}_G$ be the (ε, δ)-DP Gaussian mechanism. Then, for any $x \in X$,*

$$\begin{aligned}
\max\left\{D_{e^{\varepsilon}}(\mathbb{P}_{in,x}||\mathbb{P}_{out,x}), D_{e^{\varepsilon}}(\mathbb{P}_{out,x}||\mathbb{P}_{in,x})\right\} \leqslant \frac{1}{n|X_{in}(x)|} \sum_{D \in X_{in}(x)} \sum_{\substack{D' \in X_{out}(x) \\ D' \sim D}} \\
\left[\Phi\left(\frac{\|q(D)-q(D')\|}{2\sigma} - \frac{\varepsilon\sigma}{\|q(D)-q(D')\|}\right) - e^{\varepsilon}\Phi\left(-\frac{\|q(D)-q(D')\|}{2\sigma} - \frac{\varepsilon\sigma}{\|q(D)-q(D')\|}\right)\right].
\end{aligned}$$

Proof. Let $x \in X$ and $N := |X_{\text{in}}(x)| = |X_{\text{out}}(x)|$. Let $P(S) := \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D) \in S)$, $P'(S) := \mathbb{P}_{\mathcal{A}}(\mathcal{A}(D') \in S)$, and denote the density functions of these distributions by p and p' respectively. (The distributions P and P' are parameterized by specific data sets D and D', but we omit the dependence to reduce notational clutter.) Note that $p(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(q(D)-t)^2}{2\sigma^2}\right)$ and $p'(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(q(D')-t)^2}{2\sigma^2}\right)$. Recall that the hockey-stick divergence $D_{e^{\varepsilon}}$ is an f-divergence, with $f(t) = f_{\varepsilon}(t) = \max(t - e^{\varepsilon}, 0)$ (Sason and Verdú 2016). By joint convexity,

$$\begin{aligned}
D_{e^{\varepsilon}}(\mathbb{P}_{\text{in},x}||\mathbb{P}_{\text{out},x}) &\leqslant \frac{1}{N}\sum_{D \in X_{\text{in}}(x)} D_{e^{\varepsilon}}\left(P\middle\|\frac{1}{n}\sum_{\substack{D' \in X_{\text{out}}(x) \\ D' \sim D}} P'\right) \\
&\leqslant \frac{1}{N}\sum_{D \in X_{\text{in}}(x)} \frac{1}{n}\sum_{\substack{D' \in X_{\text{out}}(x) \\ D' \sim D}} D_{e^{\varepsilon}}\left(P||P'\right). \qquad (15)
\end{aligned}$$

Now,

$$\begin{aligned} D_{e^\varepsilon}(P||P') &= \int \max\left(0, q(t) - e^\varepsilon q'(t)\right) dt \\ &= \int_{t:q(t)\geqslant e^\varepsilon q'(t)} [q(t) - e^\varepsilon q'(t)] dt, \end{aligned}$$

and by (Balle and Wang 2018), we have

$$\begin{aligned} D_{e^\varepsilon}(P||P') &= \mathbb{P}_{y\sim\mathcal{A}(D)|D}\left[\log\frac{p(y)}{p'(y)} > \varepsilon\right] - e^\varepsilon \mathbb{P}_{z\sim\mathcal{A}(D')|D'}\left[\log\frac{p'(y)}{p(y)} < -\varepsilon\right] \\ &= \Phi\left(\frac{\|q(D) - q(D')\|}{2\sigma} - \frac{\varepsilon\sigma}{\|q(D) - q(D')\|}\right) - e^\varepsilon \Phi\left(-\frac{\|q(D) - q(D')\|}{2\sigma} - \frac{\varepsilon\sigma}{\|q(D) - q(D')\|}\right). \end{aligned}$$

Plugging this identity into the inequality 15 completes the proof. □