Deciphering the Interplay between Local Differential Privacy, Average Bayesian Privacy, and Maximum Bayesian Privacy

Xiaojin Zhang * XIAOJINZHANG@HUST.EDU.CN
Huazhong University of Science and Technology, China

Yulin Fei FEI_YULIN@HUST.EDU.CN
Huazhong University of Science and Technology, China

Wei Chen † LEMURIA_CHEN@HUST.EDU.CN
Huazhong University of Science and Technology, China

Editor:

Abstract

The swift evolution of machine learning has led to emergence of various definitions of privacy due to the threats it poses to privacy, including the concept of local differential privacy (LDP). Although widely embraced and utilized across numerous domains, this conventional approach to measure privacy still exhibits certain limitations, spanning from failure to prevent inferential disclosure to lack of consideration for the adversary's background knowledge. In this comprehensive study, we introduce Bayesian privacy and delve into the intricate relationship between LDP and its Bayesian counterparts, unveiling novel insights into utility-privacy trade-offs. We introduce a framework that encapsulates both attack and defense strategies, highlighting their interplay and effectiveness. The relationship between LDP and Maximum Bayesian Privacy (MBP) is first revealed, demonstrating that under uniform prior distribution, a mechanism satisfying ξ-LDP will satisfy ξ-MBP and conversely ξ-MBP also confers 2ξ-LDP. Our next theoretical contribution are anchored in the rigorous definitions and relationships between Average Bayesian Privacy (ABP) and Maximum Bayesian Privacy (MBP), encapsulated by equations $\epsilon_{p,a} \leq \frac{1}{\sqrt{2}}\sqrt{(\epsilon_{p,m}+\epsilon)\cdot(e^{\epsilon_{p,m}+\epsilon}-1)}$. These relationships fortify our understanding of the privacy guarantees provided by various mechanisms. Our work not only lays the groundwork for future empirical exploration but also promises to facilitate the design of privacy-preserving algorithms, thereby fostering the development of trustworthy machine learning solutions.

Keywords: Differential Privacy, Bayesian Privacy, Privacy-Utility Trade-offs, Adversarial Model, Privacy-Preserving Algorithms

1 Introduction

Burgeoning fields in machine learning continually bring people's concern—jeopardizing privacy when training models with sensitive data (Li and Li (2009); Alvim et al. (2012); Li et al. (2024)). Yet in the mean time, the pursuit of granular information risks compromising the privacy of those supplying the data. The mutual perennial goal of countless companies and

*. National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, China
†. Corresponding author

governments is to extract meaningful insights from data without compromising individual's privacy (Erlingsson et al. (2014); Cormode et al. (2018)). To accomplish this objective, certain scholars like McMahan et al. (2017) have suggested a distributed learning framework, for instance, federated learning. This marvelous design lets clients train their own model locally without transmitting their own private dataset. However, this method is not a panacea (Mammen (2021)). Geiping et al. (2020) and Nasr et al. (2019) found that potential attackers also can conduct gradient attack when clients' gradients are transmitted to central server. So we still need some ways to protect privacy more strictly, and the first question is how to measure it by a reasonable way. To respond to this challenge, a variety of privacy metrics have been proposed to measure privacy.

Among these, Local Differential Privacy (LDP), a concept formalized by Dwork et al. (2006), has attained the status of the benchmark. LDP provides strong privacy guarantees by ensuring that any two different data point will have very similar output of computation. The extent to which the output reveals information about the value of x is constrained by a function of ϵ. This restriction curtails the confidence that potential attackers can possess concerning the actual value of x. Despite its widespread adoption, LDP faces critical limitations, particularly in terms of lacking consideration for the adversary's background knowledge and the data distribution at play. According to du Pin Calmon and Fawaz (2012), it is contended that in a general context, differential privacy does not furnish privacy assurances regarding average or maximum information disclosure. Li et al. (2013) highlighted that the efficacy of differential privacy in averting inferential disclosures is lacking. The possibility remains for deducing potentially sensitive details about an individual with notable precision from differentially private outputs.

Recent research done by Zhu et al. (2019) and He et al. (2019) has begun to address these limitations by proposing privacy concepts aligned with the Bayesian inference perspective of adversary, such as Average Bayesian Privacy (ABP) and Maximum Bayesian Privacy (MBP). These concepts have emerged from an understanding that adversaries may possess prior knowledge and use their Bayesian reasoning capabilities to infer sensitive information from the output of machine learning algorithms. While these new paradigms offer a more tailored approach to privacy, there's a pressing need to thoroughly understand their relationships with traditional privacy measures like LDP.

Motivated by these, our work seeks to (a) present a rigorous framework that characterizes both adversarial attacks and defense mechanisms, critically examining the effectiveness of various strategies within this dichotomy, and (b) advance the field of machine learning privacy by exploring the theoretical relationship between Bayesian privacy measures and traditional privacy metrics like LDP. Our contributions are listed below:

- Firstly, we introduce a comprehensive privacy attacks and defenses framework in Section 3, which describes threat model by characterizing the adversary's objectives, capabilities, knowledge base, and methodologies employed in executing attacks, and outline the defender's strategic approaches foundational to the mechanisms designed for the protection of sensitive information.

- We establish ϵ-ABP and ϵ-MBP in Section 4 (refer to Definition 4.1 and Definition 4.2), which gauge privacy leakage through Jensen-Shannon (JS) divergence—a symmetric divergence metric with a square root that adheres to the triangle inequality. This

selection of divergence metric enables us to assess the trade-offs between data utility and privacy preservation, providing nuanced comprehension of privacy vulnerabilities

- Then in Section 5, we establish that algorithms satisfying ξ-LDP also provide ξ-MBP (see **Lemma** 5.1), thereby illustrating that local privacy protections translate to global privacy assurances in a Bayesian context. Conversely, we show that ξ-MBP yields 2ξ-LDP (see **Lemma** 5.2), highlighting the versatility of Bayesian privacy in protecting individual data points.

- The relationship between ABP and MBP is also dicussed, indicating that MBP implies ABP. A brief mathematical relation is present below (Further elaboration is available in **Theorem** 5.4):

$$\epsilon_{p,a} \leq \frac{1}{\sqrt{2}} \sqrt{(\epsilon_{p,m} + \epsilon) \cdot (e^{\epsilon_{p,m}+\epsilon} - 1)}, \tag{1}$$

These relationships enhance comprehension regarding the privacy assurances afforded by mechanisms aimed at safeguarding data privacy. Our analysis not only strengthens the theoretical foundations for privacy measures but also guides the design of advanced privacy-preserving algorithms.

For convenience, We use Table 1 to outline the key symbols utilized throughout this manuscript.

Table 1: List of Notation

Notation	Meaning
ϵ_p	privacy leakage of the protector
$\mathcal{L}$	loss function
Δ	extent of distortion
W	released parameter of model
$\mathcal{X}$	set of instances
$\mathcal{Y}$	labels set
P	instances $\mathcal{X}$'s distribution
$\breve{W}$	original parameter of model
d	dataset
$\mathrm{TV}(\cdot)$	total variation distance
$\mathcal{A}(\cdot)$	original model
$\mu/\mu(P)$	statistic of distribution P
d'	another dataset different from d
n	size of dataset d
$F^{\mathcal{A}}$	belief distribution after observing protected information
$F^{\mathcal{B}}$	belief distribution before observing protected information
$F^{\mathcal{O}}$	belief distribution after observing unprotected information
$f_{D\|W}(d\|w)$	posterior probability density function
$f_D^{\mathcal{O}}(d), f_D^{\mathcal{B}}(d), f_D^{\mathcal{A}}(d)$	density function of $F^{\mathcal{O}}, F^{\mathcal{B}}, F^{\mathcal{A}}$
$d^{(m)}$	m-th data in dataset
$f_{W\|D}(w\|d)$	likelihood function based on observed parameter W

ξ	maximum bayesian privacy constant
$f_D(d)$	prior distribution
κ_1	$f_D^{\mathcal{O}}(d)$
$\hat{\kappa_1}$	estimator of κ_1
κ_2	$f_D^{\mathcal{B}}(d)$
$\hat{\kappa_2}$	estimator of κ_2
κ_3	MBP, i.e., ξ
$\hat{\kappa_3}$	estimator of κ_3

2 Related Work

2.1 Metrics for Measuring Privacy

Differential Privacy (usually abbreviated as DP) stands as a rigorously formalized privacy framework pioneered by Dwork et al. (2006). Techniques for enforcing DP in deep neural network training were proposed by Abadi et al. (2016), encompassing to diverse optimization methods, including Momentum Rumelhart et al. (1986) and AdaGrad Duchi et al. (2011). An adaptation of DP known as Bayesian Differential Privacy (BDP), which incorporates considerations of data distribution, was introduced by Triastcyn and Faltings (2020a), along with a novel privacy assessment approach. By employing average total variation distance (abbreviated as TV distance) for privacy measurement and minimum mean-square error (MMSE), mutual information, and error probability to assess utility, Rassouli and Gündüz (2019) quantified trade-offs between utility and privacy.

Achieving pure DP can be challenging in practical learning scenarios. The loss of privacy was quantified by Eilat et al. (2020) through the difference between a designer's prior and posterior beliefs regarding an agent's type, utilizing Kullback-Leibler (KL) divergence. The constraints of obtaining privacy via posterior sampling were discussed by Foulds et al. (2016), while the benefits of Laplace mechanism against Bayesian inference were underscored both from the practical and theoretical perspectives.

Privacy measures akin to our approach have been considered in other works. du Pin Calmon and Fawaz (2012) introduced metrics for average information leakage and maximum information leakage, intended to balance privacy against utility constraints, with privacy leakage assessed using KL divergence by Gu et al. (2021). Conversely, our work adopts Jensen-Shannon (JS) divergence for measuring privacy leakage, in line with Zhang et al. (2022, 2023c). JS divergence is favored over KL divergence due to its symmetric nature and because its square root meets the triangle inequality (Nielsen (2019); Endres and Schindelin (2003)), which aids in quantifying privacy-utility trade-offs more effectively.

2.2 Relation between Distinct Privacy Measurements

Numerous algorithms have been developed to attain differential privacy. And inspired by advent of differential privacy, many scholars have tried to propose new privacy metrics. Concentrated differential privacy (Concentrated-DP), proposed by Dwork and Rothblum (2016), offers a relaxation of differential privacy's conventional notion. They defined the concept based on the mean of the privacy leakage of a randomized mechanism, stipulating that the mechanism satisfies that definition if it is small and subgaussian. Based on

Table 2: relation between various privacy metrics.

Reference	**Privacy Metric**[1]	**Relation To Other Metrics**
Dwork and Rothblum (2016)	Concentrated DP	DP⇒Concentrated DP
Mironov (2017)	Renyi-DP	Renyi-DP⇒ (ϵ, δ)-DP
Mironov et al. (2009)[2]	IND-CDP	SIM-CDP ⇒ $\text{SIM}_{\forall\exists}$-CDP[3]
	SIM-CDP	⇒ IND-CDP
Dong et al. (2019)	f-DP	Gaussian DP[4]⇒Renyi-DP
Blum et al. (2008)	Distributional Privacy	Distributional Privacy⇒DP
Triastcyn and Faltings (2020b)	Bayesian-DP	Bayesian DP ⇒ DP
Our Work	ξ-MBP	ξ-LDP⇒ ξ-MBP⇒2ξ-LDP
	ϵ-ABP	MBP⇒ABP

[1] All "DP"s below are shorthand for differential privacy.
[2] This article propose Computational Differential Privacy, abbreviated as **CDP**. SIM-CDP and IND-CDP are two concrete kinds of CDP actually.
[3] Definition of $\text{SIM}_{\forall\exists}$-CDP is extended by SIM-CDP.
[4] Gaussian DP is a specific family of f-DP.

Rényi divergence, Rényi Differential Privacy (RDP), was introduced by Mironov (2017) as a variant of differential privacy offering relaxation. RDP unifies concentrated-DP, ϵ-DP, and (ϵ, δ)-DP (Dwork and Rothblum (2016) and Bun and Steinke (2016)). Mironov et al. (2009) proposed computational differential privacy (CDP), measuring privacy against computationally-bounded adversaries. They introduced two variants, IND-CDP and SIM-CDP to adapt to the computational setting. The definition of IND-CDP originated from (ϵ, δ)-differential privacy. It was relaxed by ignoring some additive distinguishing advantage. SIM-CDP was further extended to two new definitions called $\text{SIM}_{\forall\exists}$-CDP and SIM^{+}-CDP. And it was proved that (a) a SIM-CDP algorithm naturally provides SIM∀∃-CDP, (b) a SIM∀∃-CDP naturally provides IND-CDP. Dong et al. (2019) gave a new way to relax the privacy definition, called f-differentially privacy (f-DP). They further proposed a single-parameter family of privacy in the class of f-DP called Gaussian differential privacy. Triastcyn and Faltings (2020a) introduced a novel variant of differential privacy, called Bayesian differential privacy (BDP), that takes into account the dataset's distribution. They also proposed a privacy accounting approach. Eilat et al. (2020) further contributed to this field by measuring privacy loss through the difference between designer's prior belief and posterior belief regarding type of agent, utilizing KL divergence. However, Foulds et al. (2016) highlighted constraints associated with utilizing posterior sampling for ensuring privacy in

practical scenarios. Despite this, the advantages of Laplace mechanism for Bayesian inference in privacy are demonstrated both from theoretical and practical viewpoints, as discussed by Triastcyn and Faltings (2020a). The concept of distributional privacy involves generating a completely new database from a specified distribution D. Furthermore, it has been established that distributional privacy offers a level of protection that is demonstrably stronger than traditional differential privacy. In situations where a trusted third party is unavailable and users perturb their inputs randomly to ensure their data's privacy, local differential privacy (LDP) emerges as a viable solution. LDP ensures privacy during the estimation of statistics and has garnered widespread adoption among various technology organizations. Cormode et al. (2018) elucidate key techniques of related systems. Additionally, Duchi et al. (2013) explore the delicate balance between privacy preservation and the rate of convergence in locally private settings. Furthermore, Ding et al. (2017) introduce innovative LDP mechanisms tailored for the repeated collection of counter data.

3 Privacy Attack and Defense Framework

This section delineates the adversarial landscape confronting data privacy in machine learning applications. We systematically outline the threat model, characterizing the adversary's objectives, capabilities, knowledge base, and methodologies for launching attacks. Concurrently, we describe the defender's strategies that underpin the protection mechanisms aimed at safeguarding sensitive information. It is worth noting that this privacy attack and defense framework has been widely investigated, including Zhang et al. (2023a,b); Kang et al. (2022); Zhang et al. (2023e); Asi et al. (2023); du Pin Calmon and Fawaz (2012).

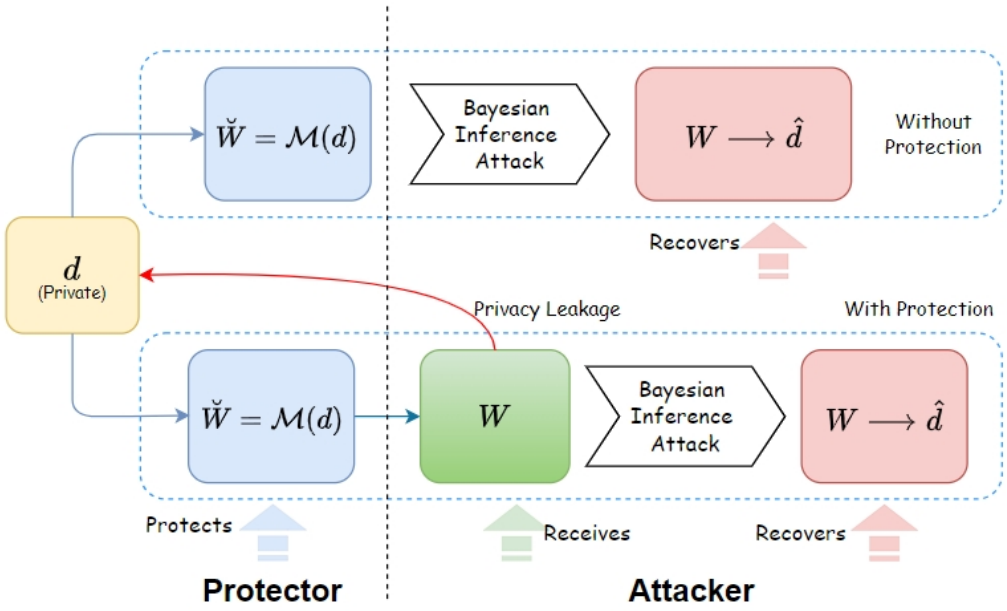

Figure 1: Schematic diagram of privacy attack and defense framework.

3.1 Privacy Attack Model

The threat model provides a structured perspective on the potential vulnerabilities and risks to client data privacy. Herein, we examine the various facets of an attacker's strategy, including their intents, resources, intelligence, and tactics.

Attacker's Objective Within our framework, we posit a scenario where a protector operates an algorithm $\mathcal{M}(\cdot)$, processing input data d to yield output $\breve{W}$. This result must then be conveyed to a recipient—who is presupposed to be the attacker. The attacker aspires to extract confidential details regarding d from the observable $\breve{W}$, while balancing the need to maintain utility within certain bounds. The underlying objective is to refine the attacker's probabilistic beliefs about d through analysis of $\breve{W}$.

Attacker's Capability Our model assumes a semi-honest (or "honest-but-curious") attacker. Such an adversary adheres to protocol specifications in terms of computations but seeks to glean additional private information by inferring it from $\breve{W}$. By performing statistical inference on the output $\breve{W}$, the attacker aims to ascertain insights into the protector's confidential data d.

Attacker's Knowledge Based on previous descriptions, the attacker has access to the output $\breve{W}$ (though post-transformation by the protector's mechanism, they receive a modified version W). Furthermore, the attacker may possess pre-existing knowledge about the private data d, potentially acquired through past breaches or leaks. Additionally, the attacker is cognizant of the utility constraints associated with their analyses.

Attacker's Method With knowledge of a distorted parameter W, we assume that the attacker adopts a Bayesian inference attack, i.e., find a data d that optimizes posterior belief, denoted by:

$$\arg\max_{d} \mathrm{IH}(d|w) = \arg\max_{d} [\mathrm{I}(w|d) + \mathrm{H}(d)], \tag{2}$$

where $\mathrm{IH}(d|w)$ represents the logarithm of the posterior probability density function $f_{D|W}(d|w)$, $\mathrm{I}(w|d)$ is the likelihood function's logarithm based on observed parameter w, and $\mathrm{H}(d)$ symbolizes the prior probability distribution's logarithm $f_D(d)$. By applying Bayesian theorem, Optimizing the logarithmic form of the posterior $f_{D|W}(d|w)$ with respect to D entails aggregating both $\log(f_{W|D}(w|d))$ and $\log(f_D(d))$.

3.2 Privacy Protection Model

We now turn our attention to the defensive measures employed by the protector whose design is to shield sensitive datasets from unauthorized exposure.

Protector's Objective To safeguard privacy post-computation by algorithm $\mathcal{M}(\cdot) : d \to \breve{W}$, the protector needs to devise a privacy-preserving mechanism, which is a stochastic transformation, having the ability to map $\breve{W} \to W$, where W is a less-informative derivative. This ensures compliance with privacy standards such as ϵ-differential privacy.

Protector's Capability It is incumbent upon the protector to transmit W in lieu of $\breve{W}$ to conceal the original data d. To this end, the protector must be adept at generating and communicating a perturbed variant W, which duly obscures the private dataset d. This task may be complicated by historical disclosures of d.

Protector's Knowledge The protector is privy to the private dataset d, the computed output $\breve{W}$, and the intricacies of the employed algorithm $\mathcal{M}(\cdot)$. They are also versed in the workings of the privacy-preserving transformation. Additionally, an understanding of the adversary's utility limitations is crucial for the protector to tailor the distortion of $\breve{W}$ into W effectively, satisfying the informational requirements of the adversary whilst preserving the sanctity of d.

Protector's Method In defence of privacy (i.e., the confidentiality of dataset d), the protector must execute the randomized privacy-preserving mechanism, which transforms $\breve{W}$ to W. Execution of the algorithm is expected to equip it with the requisite properties for safeguarding dataset d against intrusion while adhering to the adopted privacy standard. For instance, optimized algorithms for protecting privacy under the defense and attacking framework have been designed as detailed in the studies by Zhang et al. (2023e) and du Pin Calmon and Fawaz (2012), while Asi et al. (2023) choose to apply exponential mechanism(i.e., add random noise to $\breve{W}$) to protect privacy.

4 Privacy Measurements

In the following section, we outline various metrics used to assess privacy and robustness in the context of our proposed defense and attacking framework. Privacy measurement in machine learning has emerged as a critical area of study, with various approaches proposed to quantify and evaluate the privacy guarantees of different algorithms and models. This section provides a detailed exposition on the metrics used to quantify privacy under the previously discussed privacy attack and defense framework, including concepts such as Maximum Bayesian Privacy, Average Bayesian Privacy, and Local Differential Privacy. These approaches enable researchers to quantitatively evaluate the privacy risks associated with different models and algorithms. As privacy concerns continue to grow, further advancements in privacy measurement techniques are critical for developing robust and privacy-preserving machine learning solutions.

Information privacy, particularly through frameworks like ABP and MBP, offers several advantages over LDP. ABP and MBP provide a more nuanced approach to privacy by considering prior knowledge and the Bayesian perspective of an adversary, allowing for a tailored balance between privacy and utility. Unlike LDP, which applies noise uniformly to data, ABP and MBP can adjust the privacy guarantees based on the actual data distribution and the adversary's potential background knowledge, potentially leading to better data utility for the same privacy level. This is particularly beneficial in scenarios where it's crucial to maintain high data fidelity for analysis or machine learning purposes. Additionally, with the concept of "privacy loss budgets" tailored to individual data points, these models might reduce the overall quantity of noise necessary to achieve desired level of privacy compared to LDP.

First, we define the concept of average Bayesian privacy and explore its correlation with the notion of maximum Bayesian privacy in Section 5. These metrics serve as crucial benchmarks in evaluating the resilience of data to inference attacks. The average privacy leakage, defined in **Definition** 4.1, assesses the discrepancy between the adversary's initial and revised beliefs concerning sensitive information. Adversaries update their belief via Bayesian inference attack on protected model data. Hence, we term this kind of privacy

leakage as *Bayesian* privacy leakage formally. In certain situations, Incorporating protected model data has the probability to cause reduced utility and efficiency during federated model training.

Let $F^{\mathcal{O}}$, $F^{\mathcal{A}}$, and $F^{\mathcal{B}}$ denote the attacker's belief distributions regarding D when observing the original information, protected information, and no information respectively. We use $f_D^{\mathcal{A}}$, $f_D^{\mathcal{O}}$, and $f_D^{\mathcal{B}}$ to denote each one's probability density function, respectively. Specifically, $f_D^{\mathcal{A}}(d) = \int_{\mathcal{W}} f_{D|W}(d|w)dP^{\mathcal{D}}(w)$, $f_D^{\mathcal{O}}(d) = \int_{\mathcal{W}} f_{D|W}(d|w)dP^{\mathcal{O}}(w)$, and $f_D^{\mathcal{B}}(d) = f_D(d)$.

Instead of KL divergence, it is JS divergence which is utilized to quantify leakage of privacy. Compared to KL divergence, JS divergence's primary advantage lies in its symmetry, and a desirable property: its square root satisfies triangle inequality, proved by Endres and Schindelin (2003). This property simplifies assessment of trade-offs. Next, we introduce the concept of average Bayesian privacy for the protection mechanism.

Definition 4.1 (Average Bayesian Privacy) *Let $\mathcal{M}$ represent the protection mechanism that maps from the private information D to the model information W, and $\mathcal{G}$ represent the attacking mechanism. Let ϵ_p be average privacy leakage, which is defined as:*

$$\epsilon_{p,a}(\mathcal{M}, \mathcal{G}) = \sqrt{JS(F^{\mathcal{A}}||F^{\mathcal{B}})}, \tag{3}$$

*In this formula, $JS(F^{\mathcal{A}}||F^{\mathcal{B}}) = \frac{1}{2}\int_{\mathcal{D}} f_D^{\mathcal{A}}(d)\log\frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{M}}(d)}\boldsymbol{d}\mu(d) + \frac{1}{2}\int_{\mathcal{D}} f_D^{\mathcal{B}}(d)\log\frac{f_D^{\mathcal{B}}(d)}{f_D^{\mathcal{M}}(d)}\boldsymbol{d}\mu(d)$, Attacker's belief distribution about D upon observing the protected information and without observing is $F^{\mathcal{A}}$ and $F^{\mathcal{B}}$, and $f_D^{\mathcal{M}}(d) = \frac{1}{2}(f_D^{\mathcal{A}}(d) + f_D^{\mathcal{B}}(d))$. **We say $\mathcal{M}$ is ϵ-ABP when** $\epsilon_{p,a}(\mathcal{M}, \mathcal{G}) \leq \epsilon$.*

Remark: *We operate under the assumption that the private information D takes on discrete values, and then*

$$\begin{aligned} JS(F^{\mathcal{A}}||F^{\mathcal{B}}) &= \frac{1}{2}\left[KL\left(F^{\mathcal{A}}||F^{\mathcal{M}}\right) + KL\left(F^{\mathcal{B}}||F^{\mathcal{M}}\right)\right] \\ &= \frac{1}{2}\sum_{d\in\mathcal{D}} f_D^{\mathcal{A}}(d)\log\frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{M}}(d)} + \frac{1}{2}\sum_{d\in\mathcal{D}} f_D^{\mathcal{B}}(d)\log\frac{f_D^{\mathcal{B}}(d)}{f_D^{\mathcal{M}}(d)}. \end{aligned}$$

The Maximum Bayesian Privacy (denoted by $\epsilon_{p,m}$) encapsulates the highest degree of privacy assurance that can be sustained against an adversary with partial or complete knowledge about the data distribution. This metric is situated within the Bayesian privacy paradigm, which assesses the risk associated with the disclosure of private information. It underscores the importance of maintaining privacy when data undergoes sharing or utilization for designated tasks.

Definition 4.2 (ξ-Maximum Bayesian Privacy) *Let F denote the belief distribution held by the attacker regarding the private data, and let f indicate the associated probability density function. We define the system to possess ξ-Maximum Bayesian Privacy if $\forall w \in \mathcal{W}$, $\forall d \in \mathcal{D}$, the following is satisfied:*

$$-\xi \leq \log\left(\frac{f_{D|W}(d|w)}{f_D(d)}\right) \leq \xi.$$

The ξ-Maximum Bayesian Privacy quantifies the level of privacy protection afforded by measuring the disparities in the conditional and marginal distributions of the data. A smaller value of ξ implies stronger privacy guarantees against potential adversaries.

5 Theoretical Relationship between LDP, MBP and ABP

In the following, relations among Maximum Bayesian Privacy (MBP), Average Bayesian Privacy (ABP) and Local Differential Privacy (LDP) are investigated. Through mathematical definitions and theorems, we demonstrate how these concepts of privacy are interconnected and how one can transition from one privacy framework to another. Specifically, we will prove through theorems that an algorithm satisfying the requirements of LDP also satisfies those of MBP, and vice versa. Additionally, we will discuss the connection between MBP and ABP, along with how to convert from a mapping that provides maximum privacy protection to one that offers an average level of privacy protection. These theoretical results not only deepen our understanding of the connections between various privacy measures but also offer significant guidance for crafting and analyzing algorithms that preserve privacy.

LDP aims to safeguard individual data contributors' privacy, enabling the extraction of aggregate statistics or insights from the data. The concept of LDP is articulated as follows.

Definition 5.1 (ϵ-LDP) *A randomized function $\mathcal{M}$ defined as ϵ-local differentially private if, $\forall$ possible datasets d and d', $forall$ subset S of the output space, the following inequality is true*

$$\frac{\Pr[\mathcal{M}(d) \in S]}{\Pr[\mathcal{M}(d') \in S]} \leq e^{\epsilon}. \tag{4}$$

In this formula, ϵ is non-negative, which quantifies the level of privacy protection, and $\Pr[\mathcal{M}(d) \in S]$ represents the probability that the randomized mechanism $\mathcal{M}$ outputs a result in the set S when given the dataset d.

This definition states that the probability of any outcome being produced by the mechanism when given dataset d should be similar to the probability of the same outcome being produced when given a dataset d', smaller ϵ values indicate stronger privacy guarantees. LDP ensures that individual contributions to the dataset are protected by adding noise or randomness to the results of queries or computations, making it difficult to infer specific details about any individual in the dataset.

The following **Lemma** 5.1 implicates that if an algorithm provides ξ-LDP, then it also offers ξ-MBP. In other words, the same mechanism that provides local privacy protection, where individual data points are protected, also ensures a certain level of global privacy in a Bayesian sense. This means that the mechanism is not only useful for protecting individual data points but also for safeguarding the privacy of the dataset as a whole. Please check **Appendix A.1** for detailed analysis and proof.

Lemma 5.1 *Let MBP be defined using **Definition** 4.2. Consider a case where the adversary's prior belief distribution follows a uniform distribution. That is, $f_D(d) = f_D(d')$, $\forall d', d \in \mathcal{D}$. If mapping $f_{W|D}(\cdot|\cdot)$ satisfies that $\forall d', d \in \mathcal{D}$,*

$$\frac{f_{W|D}(w|d)}{f_{W|D}(w|d')} \in [e^{-\xi}, e^{\xi}],$$

then $f_{W|D}(\cdot|\cdot)$ constitutes a privacy-preserving transformation ensuring ξ-Maximum Bayesian Privacy, where for all $d \in \mathcal{D}$ and $w \in \mathcal{W}$,

$$\left|\log\left(\frac{f_{D|W}(d|w)}{f_D(d)}\right)\right| \leq \xi. \tag{5}$$

Conversely, the following lemma states that if an algorithm offers ξ-MBP, it also provides 2ξ-LDP. This suggests that mechanisms crafted to safeguard the privacy of the dataset as a whole, as evaluated by Bayesian privacy, also afford a degree of protection for individual data points within a local differential privacy framework.

Lemma 5.2 *Let MBP be defined using **Definition** 4.2. Consider a situation where the adversary's prior belief distribution follows a uniform distribution. Consequently, $f_D(d) = f_D(d')$ for any d, d'. If an algorithm is ξ-Maximum Bayesian Privacy, then it is (2ξ)-differentially private. Suppose privacy-preserving mapping $f_{W|D}(\cdot|\cdot)$ ensures ξ-maximum Bayesian privacy. That is, $\xi = \max_{w\in\mathcal{W}, d\in\mathcal{D}} \left|\log\left(\frac{f_{D|W}(d|w)}{f_D(d)}\right)\right|$. Then, the function $f_{W|D}(\cdot|\cdot)$ achieves (2ξ)-differential privacy*

$$\frac{f_{W|D}(w|d)}{f_{W|D}(w|d')} \in [e^{-2\xi}, e^{2\xi}].$$

This lemma establishes a bridge between local differential privacy and maximum Bayesian privacy, showing that mechanisms satisfying one privacy criterion also offer a certain level of privacy protection according to the other criterion. This provides insights into the relationship between different privacy definitions and can be useful when designing and analyzing privacy-preserving algorithms in various contexts.

With **Lemma** 5.1 and **Lemma** 5.2, we are now ready to derive the relationship between MBP and LDP.

Theorem 5.3 (Relationship between MBP and LDP) *Consider a situation in which the adversary's prior belief distribution follows a uniform distribution. Let MBP be defined using **Definition** 4.2. If an algorithm is ξ-LDP, then it is ξ-MBP. If an algorithm is ξ-MBP, then it is 2ξ-LDP.*

Remark: We might also explore the situation where the prior belief distribution of the adversary is ϵ-close to a uniform distribution and the analytical approach would be akin to the current scenario.

The first conclusion of **Theorem** 5.3 addresses a scenario in which the prior belief distribution of an adversary closely resembles a uniform distribution, meaning that the adversary has a relatively unbiased and uniform belief about the data points in the set d. The lemma introduces an algorithm with a privacy-preserving mapping denoted as $f_{W|D}(\cdot|\cdot)$. This mapping provides a maximum Bayesian privacy guarantee, where privacy is quantified by the parameter ξ. The parameter ξ quantifies the maximum privacy loss resulting from the mapping, determined by the logarithm of the ratio between the conditional posterior distribution of a data point d given an observation w and the prior distribution of d. In other words, ξ measures the maximum privacy leakage under this mechanism.

The second conclusion of **Theorem** 5.3 implies that if the algorithm provides ξ-maximum Bayesian privacy, then it also provides (2ξ)-differential privacy. This implies that the privacy-preserving mapping not only demonstrates robustness in terms of Bayesian privacy but also fulfills a corresponding concept of privacy, namely, differential privacy. The differential privacy guarantee of mapping is expressed in terms of the ratios of conditional probabilities. Specifically, the ratio of the conditional probabilities of observing w given two different data points d and d' falls within the range of $[e^{-2\xi}, e^{2\xi}]$. This range specifies the extent to which the mapping distorts the likelihood of observing w given different data points. The larger ξ is, the wider this range becomes, indicating stronger privacy protection. This conclusion shows a connection between maximum Bayesian privacy and differential privacy. It illustrates that a mapping ensuring maximum Bayesian privacy, particularly when the adversary's prior belief distribution approximates uniformity, also confers a degree of differential privacy. This finding holds significance in comprehending and quantifying the privacy assurances offered by such mappings within the realm of privacy-preserving data analysis.

The following theorem demonstrates the relationship between average Bayesian privacy and maximum Bayesian privacy.

Theorem 5.4 (Relationship Between MBP and ABP) *Assume that the prior belief of the attacker satisfies that*

$$\frac{f_D^{\mathcal{B}}(d)}{f_D(d)} \in [e^{-\epsilon}, e^{\epsilon}].$$

Let MBP be defined using ***Definition*** *4.2. For any $\epsilon_{p,m} \geq 0$, suppose privacy-preserving mapping $f_{W|D}(\cdot|\cdot)$ ensuring $\epsilon_{p,m}$-Maximum Bayesian Privacy. Namely,*

$$\frac{f_{D|W}(d|w)}{f_D(d)} \in [e^{-\epsilon_{p,m}}, e^{\epsilon_{p,m}}],$$

for any $w \in \mathcal{W}$. Then, we have that the ABP (denoted as $\epsilon_{p,a}$) is bounded by

$$\epsilon_{p,a} \leq \frac{1}{\sqrt{2}} \sqrt{(\epsilon_{p,m} + \epsilon) \cdot (e^{\epsilon_{p,m}+\epsilon} - 1)},$$

where $\epsilon_{p,a}$ is introduced in ***Definition*** *4.1.*

6 Conclusions

This paper delves into the core of attack and defense mechanism and construct a universal model for both. This model encompasses a variety of protection mechanisms and attack methods. Within this unified framework, the effectiveness of various attack and defense mechanisms can be assessed. We rigorously explore the conceptual and mathematical intricacies that interlink MBP with LDP—cornerstones in the quest for privacy-preserving machine learning systems. Our seminal contributions, articulated through the relationship between MBP and ABP, demonstrate a fact that MBP usually implies ABP, thus is a theoretically stricter metric. While our theoretical contributions push the envelope in privacy research, they also acknowledge limitations due to the reliance on Bayesian priors.

The significance of delineating the theoretical relationship between Bayesian privacy and LDP in the context of artificial intelligence lies in providing a formalized assurance that learning algorithms can generalize well on new data while simultaneously maintaining the confidentiality of the training data, which is essential for building trust in AI systems and fostering their adoption in privacy-sensitive applications. Future work is thus directed towards empirical validations, expansions to encompass a wider array of data distributions and adversarial strategies, and the adaptation of these theoretical tenets to real-world scenarios.

ACKNOWLEDGMENTS This work was supported by the National Science and Technology Major Project under Grant 2022ZD0115301.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, New York, NY, USA, 2016. ACM.

Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *Formal Aspects of Security and Trust: 8th International Workshop, FAST 2011, Leuven, Belgium, September 12-14, 2011. Revised Selected Papers 8*, pages 39–54. Springer, 2012.

Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. *arXiv preprint arXiv:2302.01855*, 2023.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, page 609–618, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580470. doi: 10.1145/1374376.1374464. URL `https://doi.org/10.1145/1374376.1374464`.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1655–1658, 2018.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *arXiv preprint arXiv:1712.01524*, 2017.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *2012 50th annual Allerton conference on communication, control, and computing (Allerton)*, pages 1401–1408. IEEE, 2012.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and minimax bounds: Sharp rates for probability estimation. *arXiv preprint arXiv:1305.6000*, 2013.

Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Ran Eilat, Kfir Eliaz, and Xiaosheng Mu. Bayesian privacy. *Ben Gurion University and Tel Aviv University and Columbia University*, 2020.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

Hanlin Gu, Lixin Fan, Bowen Li, Yan Kang, Yuan Yao, and Qiang Yang. Federated deep learning with bayesian privacy. *arXiv preprint arXiv:2109.13012*, 2021.

Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 148–162, 2019.

Yan Kang, Jiahuan Luo, Yuanqin He, Xiaojin Zhang, Lixin Fan, and Qiang Yang. A framework for evaluating privacy-utility trade-off in vertical federated learning. *arXiv preprint arXiv:2209.03885*, 2022.

Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 889–900, 2013.

Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526, 2009.

Yan Li, Hongyang Yan, Teng Huang, Zijie Pan, Jiewei Lai, Xiaoxue Zhang, Kongyang Chen, and Jin Li. Model architecture level privacy leakage in neural networks. *Science China Information Sciences*, 67(3):132101, 2024.

Priyanka Mary Mammen. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Ilya Mironov, Omkant Pandey, Omer Reingold, and Salil Vadhan. Computational differential privacy. In *Annual International Cryptology Conference*, pages 126–142. Springer, 2009.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

Borzoo Rassouli and Deniz Gündüz. Optimal utility-privacy trade-off with total variation distance as a privacy measure. *IEEE Transactions on Information Forensics and Security*, 15:594–603, 2019.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Aleksei Triastcyn and Boi Faltings. Bayesian differential privacy for machine learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9583–9592. PMLR, 13–18 Jul 2020a. URL `https://proceedings.mlr.press/v119/triastcyn20a.html`.

Aleksei Triastcyn and Boi Faltings. Bayesian differential privacy for machine learning. In *International Conference on Machine Learning*, pages 9583–9592. PMLR, 2020b.

Xiaojin Zhang, Hanlin Gu, Lixin Fan, Kai Chen, and Qiang Yang. No free lunch theorem for security and utility in federated learning. *arXiv preprint arXiv:2203.05816*, 2022.

Xiaojin Zhang, Kai Chen, and Qiang Yang. Towards achieving near-optimal utility for privacy-preserving federated learning via data generation and parameter distortion. *arXiv preprint arXiv:2305.04288*, 2023a.

Xiaojin Zhang, Lixin Fan, Siwei Wang, Wenjie Li, Kai Chen, and Qiang Yang. A game-theoretic framework for federated learning. *arXiv preprint arXiv:2304.05836*, 2023b.

Xiaojin Zhang, Yan Kang, Kai Chen, Lixin Fan, and Qiang Yang. Trading off privacy, utility and efficiency in federated learning. *ACM Trans. Intell. Syst. Technol.*, 2023c. ISSN 2157-6904.

Xiaojin Zhang, Yan Kang, Lixin Fan, Kai Chen, and Qiang Yang. A meta-learning framework for tuning parameters of protection mechanisms in trustworthy federated learning. *arXiv preprint arXiv:2305.18400*, 2023d.

Xiaojin Zhang, Wenjie Li, Kai Chen, Shutao Xia, and Qiang Yang. Theoretically principled federated learning for balancing privacy and utility. *arXiv preprint arXiv:2305.15148*, 2023e.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

Appendix A. Relationship between LDP and MBP

In this section, we derive the relationship between LDP and MBP.

A.1 LDP implies MBP

The adversary has a *prior belief distribution*, denoted as $f_D(d)$, that is very close (within ϵ) to a uniform distribution. In other words, the adversary doesn't have a strong prior bias towards any particular mini-batch data in the whole set $\mathcal{D}$.

There exists a *mapping function* $f_{W|D}(\cdot|\cdot)$, which takes a mini-batch data d and maps it to w such that for any two mini-batch data d and d' in $\mathcal{D}$, the ratio of the conditional probabilities of mapping to w remains within a certain range, specifically between $e^{-\xi}$ and e^{ξ}. This means that the mapping process is not overly sensitive to the specific input data.

The mapping function $f_{W|D}(\cdot|\cdot)$ is said to provide a "privacy-preserving" transformation that guarantees a *maximum Bayesian privacy* level of ξ. In other words, for any observed value w and any mini-batch data d in $\mathcal{D}$, the logarithm of the ratio of the conditional posterior distribution of d given w to the prior distribution of d (i.e., the Bayesian privacy leakage) does not exceed ξ in absolute value. This means that the mapping function provides strong privacy protection by limiting the information leakage about the original mini-batch data d when the adversary observes w.

The following lemma asserts that when the adversary's prior belief distribution is close to uniform, and a certain condition on the mapping function is met, the mapping function provides a high level of privacy protection by limiting the information leakage to a maximum of ξ in a Bayesian sense when the adversary observes a transformed w.

Lemma A.1 (LDP implies MBP) *Consider the scenario where the prior belief distribution of the adversary is a uniform distribution. That is, $f_D(d) = f_D(d')$ for any $d, d' \in \mathcal{D}$. If $f_{W|D}(\cdot|\cdot)$ is ξ-LDP. That is, the mapping $f_{W|D}(\cdot|\cdot)$ satisfies that*

$$\frac{f_{W|D}(w|d)}{f_{W|D}(w|d')} \in [e^{-\xi}, e^{\xi}],$$

then $f_{W|D}(\cdot|\cdot)$ is a privacy-preserving mapping that guarantees ξ-maximum Bayesian privacy. That is, $\left|\log\left(\frac{f_{D|W}(d|w)}{f_D(d)}\right)\right| \leq \xi$, $\forall w \in \mathcal{W}, d \in \mathcal{D}$.

Proof From the definition of ξ-LDP, we know that

$$\frac{f_{W|D}(w|d)}{f_{W|D}(w|d')} \in [e^{-\xi}, e^{\xi}],$$

for all $d, d' \in \mathcal{D}$. Notice that

$$\frac{f_{W|D}(w|d)}{f_{W|D}(w|d')} = \frac{f_{D|W}(d|w) f_D(d')}{f_{D|W}(d'|w) f_D(d)}.$$

It implies that $\frac{f_{D|W}(d|w)}{f_D(d)} \leq \frac{f_{D|W}(d'|w)}{f_D(d')} \cdot e^{\xi}$.

We assume that $f_D(d) = f_D(d')$. Notice that

$$\sum_{d \in \mathcal{D}} f_{D|W}(d|w) = 1.$$

For any fixed $\tilde{d} \in \mathcal{D}$, we have that

$$\begin{aligned}
1 &= \sum_{d \in \mathcal{D}} f_{D|W}(d|w) \\
&= f_{D|W}(\tilde{d}|w) + \sum_{d \in \mathcal{D} \setminus \tilde{d}} f_{D|W}(d|w) \\
&\leq f_{D|W}(\tilde{d}|w) + e^{\xi} \cdot f_{D|W}(\tilde{d}|w) \cdot (|\mathcal{D}| - 1) \\
&= (1 + e^{\xi} \cdot (|\mathcal{D}| - 1)) \cdot f_{D|W}(\tilde{d}|w).
\end{aligned}$$

Therefore, we have that

$$f_{D|W}(\tilde{d}|w) \geq \frac{1}{1 + e^{\xi} \cdot (|\mathcal{D}| - 1)}.$$

Similarly, we have that

$$\begin{aligned}
1 &= \sum_{d \in \mathcal{D}} f_{D|W}(d|w) \\
&= f_{D|W}(\tilde{d}|w) + \sum_{d \in \mathcal{D} \setminus \tilde{d}} f_{D|W}(d|w) \\
&\geq f_{D|W}(\tilde{d}|w) + e^{-\xi} \cdot f_{D|W}(\tilde{d}|w) \cdot (|\mathcal{D}| - 1) \\
&= (1 + e^{-\xi} \cdot (|\mathcal{D}| - 1)) \cdot f_{D|W}(\tilde{d}|w).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
f_{D|W}(\tilde{d}|w) &\leq \frac{1}{1 + e^{-\xi} \cdot (|\mathcal{D}| - 1)} \\
&\leq \frac{e^{\xi}}{|\mathcal{D}|}.
\end{aligned}$$

Therefore, $\forall \tilde{d} \in \mathcal{D}$, we have that

$$f_{D|W}(\tilde{d}|w) \leq e^{\xi} f_D(\tilde{d}).$$

We say a mechanism is ξ-MBP if $\forall w \in \mathcal{W}^d$,

$$\frac{f(d|w)}{f(d)} = \frac{f(w|d)f(d)}{f(w)f(d)} = \frac{f(w|d)}{f(w)} \leq e^{\xi}. \tag{6}$$

Therefore, $f_{D|W}(\cdot|\cdot)$ is ξ-MBP. ■

A.2 MBP implies LDP

The following lemma illustrates that if an algorithm is ξ-maximum Bayesian privacy, then it is 2ξ-Differentially Private.

Lemma A.2 (MBP implies LDP) *Consider the scenario where the prior belief distribution of the adversary is a uniform distribution. That is, $f_D(d) = f_D(d')$ for any $d, d' \in \mathcal{D}$. Let $f_{W|D}(\cdot|\cdot)$ be a privacy-preserving mapping that guarantees ξ-maximum Bayesian privacy. That is, $\xi = \max_{w\in\mathcal{W}, d\in\mathcal{D}} \left|\log\left(\frac{f_{D|W}(d|w)}{f_D(d)}\right)\right|$. Then, the mapping $f_{W|D}(\cdot|\cdot)$ is 2ξ-Local Differential Private:*

$$\frac{f_{W|D}(w|d)}{f_{W|D}(w|d')} \in [e^{-2\xi}, e^{2\xi}].$$

Proof From the definition of maximum Bayesian privacy leakage, we know that

$$\frac{f_{D|W}(d|w)}{f_D(d)} \in [e^{-\xi}, e^{\xi}],$$

for any $w \in \mathcal{W}$ and $d \in \mathcal{D}$. Therefore, for any $d, d' \in \mathcal{D}$ we have

$$\frac{f_{W|D}(w|d)}{f_{W|D}(w|d')} = \frac{f_{D|W}(d|w)f_D(d')}{f_{D|W}(d'|w)f_D(d)} \in [e^{-2\xi}, e^{2\xi}].$$

From the definition of Differential Privacy, we know that it is 2ξ-Differentially Private. ■

Combining the results in **Appendix A.1** and **Appendix A.2**, we are now ready to analyze the relationship between MBP and LDP, which is shown in the following theorem.

Theorem A.3 (Relationship between MBP and LDP) *Consider the scenario where the prior belief distribution of the adversary is a uniform distribution. That is, $f_D(d) = f_D(d')$ for any $d, d' \in \mathcal{D}$. If an algorithm is 2ξ-LDP, then it is 2ξ-MBP. If an algorithm is ξ-maximum Bayesian privacy, then it is 2ξ-LDP.*

Appendix B. Relationship Between Average Bayesian Privacy and Maximum Bayesian Privacy

In this section, we introduce the relationship between average Bayesian privacy and maximum Bayesian privacy.

Theorem B.1 (Relationship Between MBP and ABP) *Assume that the prior belief of the attacker satisfies that*

$$\frac{f_D^{\mathcal{B}}(d)}{f_D(d)} \in [e^{-\epsilon}, e^{\epsilon}].$$

Let MBP be defined using ***Definition*** *4.2. For any $\epsilon_{p,m} \geq 0$, let $f_{W|D}(\cdot|\cdot)$ be a privacy preserving mapping that guarantees $\epsilon_{p,m}$-maximum Bayesian privacy. That is,*

$$\frac{f_{D|W}(d|w)}{f_D(d)} \in [e^{-\epsilon_{p,m}}, e^{\epsilon_{p,m}}],$$

for any $w \in \mathcal{W}$. Then, we have that the ABP (denoted as $\epsilon_{p,a}$) is bounded by

$$\epsilon_{p,a} \leq \frac{1}{\sqrt{2}} \sqrt{(\epsilon_{p,m} + \epsilon) \cdot (e^{\epsilon_{p,m}+\epsilon} - 1)},$$

where $\epsilon_{p,a}$ is introduced in ***Definition*** *4.1.*

Proof Let $F^{\mathcal{M}} = \frac{1}{2}(F^{\mathcal{A}} + F^{\mathcal{B}})$. We have

$$\begin{aligned}
JS(F^{\mathcal{A}}||F^{\mathcal{B}}) &= \frac{1}{2}\left[KL\left(F^{\mathcal{A}}||F^{\mathcal{M}}\right) + KL\left(F^{\mathcal{B}}||F^{\mathcal{M}}\right)\right] \\
&= \frac{1}{2}\left[\int_{\mathcal{S}} f_D^{\mathcal{A}}(d) \log \frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{M}}(d)} \mathbf{d}\mu(d) + \int_{\mathcal{S}} f_D^{\mathcal{B}}(d) \log \frac{f_D^{\mathcal{B}}(d)}{f_D^{\mathcal{M}}(d)} \mathbf{d}\mu(d)\right] \\
&= \frac{1}{2}\left[\int_{\mathcal{S}} f_D^{\mathcal{A}}(d) \log \frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{M}}(d)} \mathbf{d}\mu(d) - \int_{\mathcal{S}} f_D^{\mathcal{B}}(d) \log \frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)} \mathbf{d}\mu(d)\right] \\
&\leq \frac{1}{2}\left[\int_{\mathcal{D}} \left|f_D^{\mathcal{A}}(d) - f_D^{\mathcal{B}}(d)\right| \left|\log \frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)}\right| \mathbf{d}\mu(d)\right],
\end{aligned}$$

where the inequality is due to $\frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{M}}(d)} \leq \frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)}$.

Therefore, we have that

$$JS(F^{\mathcal{A}}||F^{\mathcal{B}}) \leq \frac{1}{2}\left[\int_{\mathcal{D}} \left|f_D^{\mathcal{A}}(d) - f_D^{\mathcal{B}}(d)\right| \left|\log \frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)}\right| \mathbf{d}\mu(d)\right].$$

Let $w^\star = \arg\max_{w \in \mathcal{W}} f_{D|W}(d|w)$. Then we have that

$$\begin{aligned}
f_D^{\mathcal{A}}(d) &= \int_{\mathcal{W}} f_{D|W}(d|w) dP_W^{\mathcal{D}}(w) \\
&\leq f_{D|W}(d|w^\star) \int_{\mathcal{W}} dP_W^{\mathcal{D}}(w) \\
&\leq e^{\epsilon_{p,m}} \cdot f_D(d) \\
&\leq e^{\epsilon_{p,m}+\epsilon} \cdot f_D^{\mathcal{B}}(d). \qquad (7)
\end{aligned}$$

Similarly, let $w_l = \arg\min_{w \in \mathcal{W}} f_{D|W}(d|w)$. Then we have that

$$\begin{aligned}
f_D^{\mathcal{A}}(d) &= \int_{\mathcal{W}} f_{D|W}(d|w) dP_W^{\mathcal{D}}(w) \\
&\geq f_{D|W}(d|w_l) \int_{\mathcal{W}} dP_W^{\mathcal{D}}(w) \\
&\geq e^{-\epsilon_{p,m}} \cdot f_D(d) \\
&\geq e^{-(\epsilon_{p,m}+\epsilon)} \cdot f_D^{\mathcal{B}}(d). \qquad (8)
\end{aligned}$$

Combining Eq. (7) and Eq. (8), we have

$$\left|\log\left(\frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{B}}(d)}\right)\right| \leq \epsilon_{p,m} + \epsilon.$$

Now we analyze according to the following two cases.
Case 1: $f_D^{\mathcal{B}}(d) \leq f_D^{\mathcal{A}}(d)$. Then we have

$$\left|\log\left(\frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)}\right)\right| = \log\left(\frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)}\right) \leq \log\left(\frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{B}}(d)}\right) \leq \left|\log\left(\frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{B}}(d)}\right)\right| \leq \epsilon_{p,m} + \epsilon.$$

Case 2: $f_D^{\mathcal{B}}(d) \geq f_D^{\mathcal{A}}(d)$. Then we have

$$\left|\log\left(\frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)}\right)\right| = \log\left(\frac{f_D^{\mathcal{B}}(d)}{f_D^{\mathcal{M}}(d)}\right) \leq \log\left(\frac{f_D^{\mathcal{B}}(d)}{f_D^{\mathcal{A}}(d)}\right) \leq \left|\log\left(\frac{f_D^{\mathcal{A}}(d)}{f_D^{\mathcal{B}}(d)}\right)\right| \leq \epsilon_{p,m} + \epsilon.$$

Therefore,

$$\left|\log\left(\frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)}\right)\right| \leq \epsilon_{p,m} + \epsilon. \tag{9}$$

Combining Eq. (7), Eq. (8) and Eq. (9), we have that

$$\begin{aligned} JS(F^{\mathcal{A}}||F^{\mathcal{B}}) &\leq \frac{1}{2}\left[\int_{\mathcal{S}} \left|f_D^{\mathcal{A}}(d) - f_D^{\mathcal{B}}(d)\right| \left|\log\frac{f_D^{\mathcal{M}}(d)}{f_D^{\mathcal{B}}(d)}\right| \mathbf{d}\mu(d)\right] \\ &\leq \frac{1}{2}(\epsilon_{p,m} + \epsilon) \cdot \left(e^{\epsilon_{p,m}+\epsilon} - 1\right). \end{aligned}$$

Therefore,

$$\epsilon_{p,a} = \sqrt{JS(F^{\mathcal{A}}||F^{\mathcal{B}})} \leq \frac{1}{\sqrt{2}}\sqrt{(\epsilon_{p,m} + \epsilon) \cdot (e^{\epsilon_{p,m}+\epsilon} - 1)}.$$

■

Appendix C. Practical Estimation for Maximum Bayesian Privacy

ξ represents the maximum privacy leakage over all possible released information w from the set $\mathcal{W}$ and all the dataset d, expressed as

$$\xi = \max_{w \in \mathcal{W}, d} \left|\log\left(\frac{f_{D|W}(d|w)}{f_D(d)}\right)\right|. \tag{10}$$

When the extent of the attack is fixed, ξ becomes a constant. Given Assumption D.1, we can infer that

$$\xi = \max_{d} \left|\log\left(\frac{f_{D|W}(d|w^*)}{f_D(d)}\right)\right|. \tag{11}$$

We approximate ξ by the following estimator:

$$\hat{\xi} = \max_{d} \left| \log \left(\frac{\hat{f}_{D|W}(d|w^*)}{f_D(d)} \right) \right|. \tag{12}$$

We define κ_3 as the maximum Bayesian Privacy ξ, and let $\hat{\kappa}_3$ denote its estimate $\hat{\xi}$

$$\hat{\kappa}_3 = \hat{\xi}. \tag{13}$$

The task of estimating the maximum Bayesian Privacy ξ reduces to determining the value of $\hat{f}_{D|W}(d|w_m)$, which will be addressed in the subsequent section.

C.1 Estimation for $\hat{f}_{D|W}(d|w_m)$

We first generate a set of models $\{w_m\}_{m=1}^M$ stochastically: to generate the m-th model w_m, we run an algorithm (e.g., stochastic gradient descent), using a mini-batch $\mathcal{S}$ randomly sampled from $\mathcal{D} = \{z_1, \ldots, z_{|\mathcal{D}|}\}$.

The set of parameters $\{w_m\}_{m=1}^M$ is generated for a given dataset $\mathcal{D}$ by randomly initializing the parameters for each model, iterating a specified number of times, and updating the parameters with mini-batches sampled from the dataset using approaches such as gradient descent. After the iterations, the final set of parameters for each model is returned as $\{w_m\}_{m=1}^M$.

We then estimate $\hat{f}_{D|W}(d|w_m)$ with $\{w_m\}_{m=1}^M$. As mentioned in subsection 3.1, the attacker can conduct Bayesian inference attack, i.e., estimate dataset d with $\hat{d}$ that optimizes posterior belief, denoted by:

$$\hat{d} = \arg\max_{d} \log f_{D|W}(d|w) = \arg\max_{d} \log[f_{W|D}(w|d) \times f_D^{\mathcal{B}}(d)], \tag{14}$$

Empirically, $\hat{f}_D^{\mathcal{O}}(d)$ can be estimated by Eq. (17): $\hat{f}_D^{\mathcal{O}}(d) = \frac{1}{M} \sum_{m=1}^M \hat{f}_{D|W}(d|w_m)$. Then We determine the success of attack on an image using a function Ω, which measures the similarity between the distribution of recovered dataset and the original one (e.g. Ω can be Kullback–Leibler divergence or Cross Entropy). We consider the original dataset d to be successfully recovered if $\Omega(\hat{f}_D^{\mathcal{O}}(d), f_D(d)) > t$, where t, a pre-given constant, is the threshold for determining successful recovery of the original data.

This algorithm is used to estimate the conditional probability $f_{D|W}(d|w)$. It takes a mini-batch of data d from client k, with batch size S, and a set of parameters $\{w_m\}_{m=1}^M$ as input. The goal of the algorithm is to estimate the conditional probability $f_{D|W}(d|w)$ by iteratively updating the parameters w_m.

The key steps of the algorithm are as follows:

- **Compute the gradients ∇w_m on each parameter** w_m, which is done by calculating the gradient of the loss function $\mathcal{L}$ with respect to the parameter w_m. This is performed using samples from the mini-batch data d. Then **update the parameters** w_m **using gradient descent**, where w'_m represents the updated parameters obtained by subtracting the product of the learning rate η and the gradient ∇w_m from w_m.

- For each parameter w_m, iterate the following steps for T times: **Generate perturbed data** $\tilde{d}$ by applying the Backward Iterative Alignment (BIA) method with the parameters w'_m and the gradient ∇w_m.
- **Determine whether the original data z_d is recovered** by comparing it with the perturbed data $\tilde{z}_d$, based on a defined threshold Ω that measures the difference between them. Accumulate the count of recovered instances for each data sample z_d in the variable M_d^m.
- **Estimate the conditional probability** $\hat{f}_{D|W}(d|w_m)$ for each data sample z_d in the dataset d by dividing M_d^m by the product of the total number of iterations T and the batch size S, i.e., $\frac{\boldsymbol{M_d^m}}{\boldsymbol{S \cdot T}}$. Return the estimated conditional probabilities $\hat{f}_{D|W}(d|w_m)$ for each parameter w_m.

The core idea of this algorithm is to estimate the conditional probability by generating perturbed data and comparing it with the original data during the parameter update process. By performing multiple iterations and accumulating counts, the algorithm provides an estimation of the conditional probability.

C.2 Estimation Error of MBP

To streamline the error bound analysis, it is presumed that both the data and model parameters take on discrete values.

The following lemma states that the estimated value $\hat{\kappa}_1(d)$ is very likely to be close to the true value $\kappa_1(d)$. The probability of the estimate being within an error margin that is a small percentage (ϵ times) of the true value is very high—specifically, at least $1 - 2\exp(-\frac{\epsilon^2 T \kappa_1(d)}{3})$. This means that the larger the number T or the smaller the error parameter ϵ, the more confident we are that our estimate is accurate. For the detailed analysis, please refer to Zhang et al. (2023d).

Lemma C.1 *With probability at least* $1 - 2\exp\left(\frac{-\epsilon^2 T \kappa_1(d)}{3}\right)$*, we have that*

$$|\hat{\kappa}_1(d) - \kappa_1(d)| \leq \epsilon \kappa_1(d). \tag{15}$$

Appendix D. Practical Estimation for Average Bayesian Privacy

In the scenario of federated learning, Zhang et al. (2023d) derived the upper bound of privacy leakage based on the specific form of $\mathrm{TV}(P^{\mathcal{O}}||P^{\mathcal{D}})$ of $\mathcal{M}$ according to this equation: $\widetilde{\epsilon}_p = 2C_1 - C_2 \cdot \mathrm{TV}(P^{\mathcal{O}}||P^{\mathcal{D}})$. In this section, we elaborate on the methods we propose to estimate C_1 and C_2. First, we introduce the following assumption.

Assumption D.1 *We assume that the amount of Bayesian privacy leaked from the true parameter is maximum over all possible parameters. That is, for any* $w \in \mathcal{W}$*, we have that*

$$\frac{f_{D|W}(d|w^*)}{f_D(d)} \geq \frac{f_{D|W}(d|w)}{f_D(d)}, \tag{16}$$

where w^* *represents the true model parameter.*

D.1 Estimation for $C_1 = \sqrt{\mathbf{JS}(F^{\mathcal{O}}||F^{\mathcal{B}})}$

C_1 quantifies the average square root of the Jensen-Shannon divergence between the adversary's belief distribution about the private information before and after observing the unprotected parameter, independent of the employed protection mechanisms. The JS divergence is defined as:

$$\mathrm{JS}(F^{\mathcal{O}} \parallel F^{\mathcal{B}}) = \frac{1}{2}\left(\int_{\mathcal{D}} f_D^{\mathcal{O}}(d) \log \frac{f_D^{\mathcal{O}}(d)}{\frac{1}{2}(f_D^{\mathcal{O}}(d) + f_D^{\mathcal{B}}(d))}\, \mathbf{d}\mu(d) + \int_{\mathcal{D}} f_D(d) \log \frac{f_D(d)}{\frac{1}{2}(f_D^{\mathcal{O}}(d) + f_D^{\mathcal{B}}(d))}\, \mathbf{d}\mu(d)\right),$$

where $f_D^{\mathcal{O}}(d) = \int_{\mathcal{W}^{\mathcal{O}}} f_{D|W}(d|w)\, dP^{\mathcal{O}}(w) = \mathbb{E}_w[f_{D|W}(d|w)]$, and $f_D^{\mathcal{B}}(d) = f_D(d)$.

To estimate C_1, we first need to estimate the values of $f_D^{\mathcal{O}}(d)$ and $f_D^{\mathcal{B}}(d)$. It is approximated that

$$\hat{f}_D^{\mathcal{O}}(d) = \frac{1}{M}\sum_{m=1}^{M} \hat{f}_{D|W}(d|w_m), \tag{17}$$

where w_m signifies the model parameter observed by the attacker at the m-th attempt to infer data d given w_m.

Denoting $C_1 = \sqrt{\mathrm{JS}(F^{\mathcal{O}} \parallel F^{\mathcal{B}})}$, and letting $\kappa_1 = f_D^{\mathcal{O}}(d)$, $\kappa_2 = f_D^{\mathcal{B}}(d) = f_D(d)$, and $\hat{\kappa}_1 = \hat{f}_D^{\mathcal{O}}(d)$, we have

$$C_1^2 = \frac{1}{2}\int_{\mathcal{D}} \kappa_1 \log \frac{\kappa_1}{\frac{1}{2}(\kappa_1+\kappa_2)}\, d\mu(d) + \frac{1}{2}\int_{\mathcal{D}} \kappa_2 \log \frac{\kappa_2}{\frac{1}{2}(\kappa_1+\kappa_2)}\, d\mu(d).$$

With these estimates, the squared estimated value of C becomes

$$\hat{C}_1^2 = \frac{1}{2}\int_{\mathcal{D}} \hat{\kappa}_1 \log \frac{\hat{\kappa}_1}{\frac{1}{2}(\hat{\kappa}_1+\kappa_2)}\, d\mu(d) + \frac{1}{2}\int_{\mathcal{D}} \kappa_2 \log \frac{\kappa_2}{\frac{1}{2}(\hat{\kappa}_1+\kappa_2)}\, d\mu(d).$$

D.2 Estimation for $C_2 = \frac{1}{2}(e^{2\xi} - 1)$

Recall that the Maximum Bayesian Privacy, is defined as $\xi = \max_{w\in\mathcal{W}, d\in\mathcal{D}} \left|\log\left(\frac{f_{D|W}(d|w)}{f_D(d)}\right)\right|$ and represents the maximal privacy leakage across all possible information w released. The quantity ξ remains constant when the extent of the attack is fixed. Under Assumption D.1, we have that

$$\xi = \max_{d\in\mathcal{D}} \left|\log\left(\frac{f_{D|W}(d|w^*)}{f_D(d)}\right)\right|. \tag{18}$$

The approximation of ξ is given by:

$$\hat{\xi} = \max_{d\in\mathcal{D}} \left|\log\left(\frac{\hat{f}_{D|W}(d|w^*)}{f_D(d)}\right)\right|. \tag{19}$$

Letting $\kappa_3 = \xi$ and $\hat{\kappa}_3 = \hat{\xi}$, we compute C_2 as:

$$\hat{C}_2 = \frac{1}{2}\left(e^{2\hat{\kappa}_3} - 1\right). \tag{20}$$

Ultimately, the computation of C_1 and C_2 hinges on the estimation of $\hat{f}_{D|W}(d|w_m)$, which is further elaborated upon in Section C.1.

D.3 Estimation Error of ABP

In this section, we analyze the estimation error of ABP. To facilitate the analysis of the error bound, we consider that both the data and the model parameter are in discrete form. For the detailed analysis, please refer to Zhang et al. (2023d).

The following theorem establishes that a bounded estimation error for $\kappa_1(d)$—specifically, not exceeding ϵ times the value of $\kappa_1(d)$—implies a bounded estimation error for another variable, C_1. This bound for C_1 is defined by a specific formula incorporating ϵ and its logarithmic terms. In summary, the theorem provides a calculated limit for how large the estimation error for C_1 can be, given the known estimation error for $\kappa_1(d)$.

Theorem D.1 (The Estimation Error of C_1) *Assume that* $|\kappa_1(d)-\hat{\kappa}_1(d)| \leq \epsilon\kappa_1(d)$. *We have that*

$$|\hat{C}_1 - C_1| \leq \sqrt{\frac{(1+\epsilon)\log\frac{1+\epsilon}{1-\epsilon}}{2} + \frac{\epsilon + \max\left\{\log(1+\epsilon), \log\frac{1}{(1-\epsilon)}\right\}}{2}}. \tag{21}$$

The analysis for the estimation error of C_2 depends on the application scenario. To facilitate the analysis, we assume that C_2 is known in advance. With the estimators and estimation bound for C_1, we are now ready to derive the estimation error for privacy leakage.

The following theorem addresses the accuracy of an estimated privacy parameter, $\hat{\epsilon}_p$, in comparison to its true value, ϵ_p. The theorem provides a mathematical limit on how much the estimated value $\hat{\epsilon}_p$ can differ from the true privacy parameter ϵ_p. The difference between them is not to exceed a certain threshold, which is a complex expression involving the privacy parameter ϵ itself and logarithmic terms. This threshold is important in applications where it is crucial to know how accurately the privacy parameter can be estimated.

Theorem D.2 *We have that*

$$|\hat{\epsilon}_p - \epsilon_p| \leq \frac{3}{2} \cdot \sqrt{\frac{(1+\epsilon)\log\frac{1+\epsilon}{1-\epsilon}}{2} + \frac{\epsilon + \max\left\{\log(1+\epsilon), \log\frac{1}{(1-\epsilon)}\right\}}{2}}. \tag{22}$$