Wasserstein Differential Privacy

Chengyi Yang[1], **Jiayin Qi**[2*], **Aimin Zhou**[1]

[1]Shanghai Institute of AI for Education, School of Computer Science and Technology, and Key Laboratory of MEA (Ministry of Education), East China Normal University
[2]Cyberspace Institute of Advanced Technology, Guangzhou University
52265901027@stu.ecnu.edu.cn, qijiayin@139.com, amzhou@cs.ecnu.edu.cn

Abstract

Differential privacy (DP) has achieved remarkable results in the field of privacy-preserving machine learning. However, existing DP frameworks do not satisfy all the conditions for becoming metrics, which prevents them from deriving better basic private properties and leads to exaggerated values on privacy budgets. We propose Wasserstein differential privacy (WDP), an alternative DP framework to measure the risk of privacy leakage, which satisfies the properties of symmetry and triangle inequality. We show and prove that WDP has 13 excellent properties, which can be theoretical supports for the better performance of WDP than other DP frameworks. In addition, we derive a general privacy accounting method called Wasserstein accountant, which enables WDP to be applied in stochastic gradient descent (SGD) scenarios containing subsampling. Experiments on basic mechanisms, compositions and deep learning show that the privacy budgets obtained by Wasserstein accountant are relatively stable and less influenced by order. Moreover, the overestimation on privacy budgets can be effectively alleviated. The code is available at https://github.com/Hifipsysta/WDP.

Introduction

Differential privacy (Dwork et al. 2006b) is a mathematically rigorous definition of privacy, providing quantifiable descriptions of the risk on leaking sensitive information. In the early stage, researches on differential privacy mainly focused on the issue of statistical queries (SQ) (McSherry 2009; Kasiviswanathan et al. 2011). With the risk of privacy leakage being warned in machine learning (Wang, Si, and Wu 2015; Shokri et al. 2017; Zhu, Liu, and Han 2019), differential privacy has been gradually applied for privacy protection in deep learning (Shokri and Shmatikov 2015; Abadi et al. 2016; Phan et al. 2019; Cheng et al. 2022).

However, these techniques are always constructed on the postulation of standard DP (Dwork et al. 2006b), which only provides the worst-case scenario, and tends to overestimate privacy budgets under the measure of maximum divergence (Triastcyn and Faltings 2020). Although the most commonly applied approximate differential privacy (ε, δ-DP) (Dwork et al. 2006a) ignores extreme situations with small probabilities by introducing a relaxation term δ called failure probability, it is believed that (ε, δ)-DP cannot strictly handle composition problems (Mironov 2017; Dong, Roth, and Su 2022). To address the above issues, further researches have been considering the specific data distribution, which can be divided into two main directions: *the distribution of privacy loss* and *the distribution of unique difference*. For example, concentrated differential privacy (CDP) (Dwork and Rothblum 2016), zero-concentrated differential privacy (zCDP) (Bun and Steinke 2016), and truncated concentrated differential privacy (tCDP) (Bun et al. 2018) all assume that the mean of privacy loss follows subgaussian distribution. While Bayesian differential privacy (BDP) (Triastcyn and Faltings 2020) considers the distribution of the only different data entry x'. Nevertheless, they are all defined by the upper bound of divergence, which implies that their privacy budgets are overly pessimistic (Triastcyn and Faltings 2020).

In this paper, we introduce a variant of differential privacy from another perspective. We define the privacy budget through the upper bound of the Wasserstein distance between adjacent distributions, which is called *Wasserstein differential privacy* (WDP). From a semantic perspective, WDP also follows the concept of indistinguishability (Dwork et al. 2006b) in differential privacy. Specifically, for all possible adjacent databases D and D', WDP reflects the maximum variation of optimal transport (OT) cost between the distributions queried by an adversary before and after any data entry change in the database.

Intuitively speaking, the advantages of WDP can be divided into at least two aspects. (1) WDP focuses on individuals within the distribution, rather than focusing on the entire distribution like divergence, which is consistent with the original intention of differential privacy to protect individual private information from leakage. (2) More importantly, WDP satisfies all the conditions to become a metric, including non-negativity, symmetry and triangle inequality (see Proposition 1-3), which is not fully possessed by privacy loss under the definition of divergence, as divergence itself does not satisfy symmetry and triangle inequality (see Proposition 11 in the appendix of Mironov (2017)).

The combination of DP and OT has been taken into consideration in several existing works. Their contributions are essentially to provide privacy guarantees for computing Wasserstein distance between data domains (Tien, Habrard, and Sebban 2019), distributions (Rakotomamonjy

*Corresponding Author

and Ralaivola 2021) or graph embeddings (Jin and Chen 2022). However, our work is to compute privacy budgets through Wasserstein distance, and the contributions are summarized as follows:

Firstly, we propose an alternative DP framework called Wasserstein differential privacy (WDP), which satisfies three basic properties of a metric (non-negativity, symmetry and triangle inequality), and is easy to convert with other DP frameworks (see Proposition 9-11).

Secondly, we show that WDP has 13 excellent properties. More notably, basic sequential composition, group privacy among them and advanced composition are all derived from triangle inequality, which shows the advantages of WDP as a metric DP.

Thirdly, we derive advanced composition, privacy loss and absolute moment under WDP, and finally develop Wasserstein accountant to track and account privacy budgets in subsampling algorithms such as SGD in deep learning.

Fourthly, we conduct experiments to evaluate WDP on basic mechanisms, compositions and deep learning. Results show that applying WDP as privacy framework can effectively avoid overstating the privacy budgets.

Related Work

Pure differential privacy (ε-DP) (Dwork et al. 2006b) provides strict guarantees for all measured events through maximum divergence. To address the long tailed distribution generated by privacy mechanism, (ε, δ)-DP (Dwork et al. 2006a) ignores extremely low probability events through a relaxation term δ. However, (ε, δ)-DP is considered to an overly relaxed definition (Bun et al. 2018) and cannot effectively handle composition problems, such as leading to parameter explosion (Mironov 2017) or failing to capture correct hypothesis testing (Dong, Roth, and Su 2022). In view of this, CDP (Dwork and Rothblum 2016) applies a subgaussian assumption to the mean of privacy loss. zCDP (Bun and Steinke 2016) capture privacy loss is a subgaussian random variable through Rényi divergence. Rényi differential privacy (RDP) (Mironov 2017) proposes a more general definition of DP based on Rényi divergence. tCDP (Bun et al. 2018) further relaxes zCDP. BDP (Triastcyn and Faltings 2020) considers the distribution of unique different entries. Subspace differential privacy (Gao, Gong, and Yu 2022) and integer subspace differential privacy (Dharangutte et al. 2023) consider privacy computing scenarios with external constraints. However, these concepts are all based on divergence, so that their privacy loss does not have the property of metrics. Although f-DP and its special case Gaussian differential privacy (GDP) (Dong, Roth, and Su 2022) innovatively define privacy based on the trade-off function between two types of errors in hypothesis testing, they are difficult to associate with other DP frameworks.

Wasserstein Differential Privacy

In this section, we introduce the concept of Wasserstein distance and define our Wasserstein differential privacy.

Definition 1 (Wasserstein distance (Rüschendorf 2009)). For two probability distributions P and Q defined over $\mathcal{R}$, their μ-Wasserstein distance is

$$W_\mu(P,Q) = \left(\inf_{\gamma\in\Gamma(P,Q)} \int_{\mathcal{X}\times\mathcal{Y}} \rho(x,y)^\mu d\gamma(x,y)\right)^{\frac{1}{\mu}}. \quad (1)$$

Where $\rho(x,y) = \|x-y\|$ is the norm defined in probability space $\Omega = \mathcal{X}\times\mathcal{Y}$. $\Gamma(P,Q)$ is the set for all the possible joint distributions, and $\gamma(x,y) > 0$ satisfying $\int \gamma(x,y)\,dy = P(x)$ and $\int \gamma(x,y)\,dx = Q(y)$.

In practical sense, $\rho(x,y)$ can be regarded as the cost for one unit of mass transported from x to y. $\gamma(x,y)$ can be seen as a transport plan representing the share to be moved from P to Q, which measures how much mass must be transported in order to complete the transportation.

In particular, when μ is equal to 1, we can obtain the 1-Wasserstein distance applied in Wasserstein generative adversarial network (WGAN) (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017). The successful application of 1-Wasserstein distance in WGAN should be attributed to Kantorovich-Rubinstein duality, which effectively reduces the computational complexity of Wasserstein distance.

Definition 2 (Kantorovich-Rubinstein distance (Kantorovich and Rubinshten 1958)). According to the property of Kantorovich-Rubinstein duality, 1-Wasserstein distance can be equivalently expressed as Kantorovich-Rubinstein distance

$$K(P,Q) = \sup_{\|\varphi\|_L\le 1} \mathbb{E}_{x\sim P}[\varphi(x)] - \mathbb{E}_{y\sim Q}[\varphi(y)]. \quad (2)$$

Where $\varphi : \mathcal{X} \to \mathcal{R}$ is the so-called Kantorovich potential, giving the optimal transport map by a close-form formula. Where $\|\varphi\|_L$ is the Lipschitz bound of Kantorovich potential, $\|\varphi\|_L \le 1$ indicates that φ satisfies the 1-Lipschitz condition with

$$\|\varphi\|_L = \sup_{x\neq y} \frac{\rho(\varphi(x),\varphi(y))}{\rho(x,y)}. \quad (3)$$

Definition 3 ((μ,ε)-WDP). A randomized algorithm $\mathcal{M}$ is said to satisfy (μ,ε)-Wasserstein differential privacy if for any adjacent datasets $D, D' \in \mathcal{D}$ and all measurable subsets $S \subseteq \mathcal{R}$ the following inequality holds

$$W_\mu(Pr[\mathcal{M}(D)\in S], Pr[\mathcal{M}(D')\in S]) = \left(\inf_{\gamma\in\Gamma(Pr_{\mathcal{M}}(D),Pr_{\mathcal{M}}(D'))} \int_{\mathcal{X}\times\mathcal{Y}} \rho(x,y)^\mu d\gamma(x,y)\right)^{\frac{1}{\mu}} \le \varepsilon. \quad (4)$$

Where $\mathcal{M}(D)$ and $\mathcal{M}(D')$ represent two outputs when algorithm $\mathcal{M}$ respectively performs on dataset D and D'. $Pr[\mathcal{M}(D)\in S]$ and $Pr[\mathcal{M}(D')\in S]$ are the probability distributions, also denoted as $Pr_{\mathcal{M}}(D)$ and $Pr_{\mathcal{M}}(D')$ in this paper. The value of $W_\mu(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D'))$ is the privacy loss under (μ,ε)-WDP and its upper bound ε is called privacy budget.

Symbolic representations. WDP can also be represented as $W_\mu(\mathcal{M}(D),\mathcal{M}(D')) \le \varepsilon$. To emphasize the inputs are two probability distributions, we denote WDP as $W_\mu(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D')) \le \varepsilon$. To avoid confusion, we also represent RDP as $D_\alpha(Pr_{\mathcal{M}}(D)\|Pr_{\mathcal{M}}(D')) \le \varepsilon$, although

the representation $D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \leq \varepsilon$ implies that the results depend on the randomized algorithm and the queried data. They are both reasonable because $\mathcal{M}(D)$ can be seen as a random variable that satisfies $\mathcal{M}(D) \sim Pr_{\mathcal{M}}(D)$.

For the convenience on computation, we define Kantorovich Differential Privacy (KDP) as an alternative way to obtain privacy loss or privacy budget under $(1,\varepsilon)$-WDP.

Definition 4 (Kantorovich Differential Privacy). If a randomized algorithm $\mathcal{M}$ satisfies$(1,\varepsilon)$-WDP, which can also be written as the form of Kantorovich-Rubinstein duality

$$K\left(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D')\right) = \sup_{\|\varphi\|_L \leq 1} \mathbb{E}_{x\sim Pr_{\mathcal{M}}(D)}[\varphi(x)] - \mathbb{E}_{x\sim Pr_{\mathcal{M}}(D')}[\varphi(x)] \leq \varepsilon. \quad (5)$$

ε-KDP is equivalent to $(1,\varepsilon)$-WDP, and can be computed more efficiently through duality formula based on Kantorovich-Rubinstein distance.

Properties of WDP

Proposition 1 (Symmetry). Let $\mathcal{M}$ be a (μ,ε)-WDP algorithm, for any $\mu \geq 1$ and $\varepsilon \geq 0$ the following equation holds

$$\begin{aligned} &W_\mu(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D')) \\ &\quad = W_\mu\left(Pr_{\mathcal{M}}(D'), Pr_{\mathcal{M}}(D)\right) \leq \varepsilon. \end{aligned} \quad (6)$$

The symmetric property of (μ,ε)-WDP is implied in its definition. Specifically, the joint distribution $\Gamma(\cdot)$ satisfies $\Gamma(Pr_{\mathcal{M}}(D'), Pr_{\mathcal{M}}(D)) = \Gamma(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D'))$. In addition, Kantorovich differential privacy also satisfies this property and the proof is available in the appendix.

Proposition 2 (Triangle Inequality) Let $D_1, D_2, D_3 \in \mathcal{D}$ be three arbitrary datasets. Suppose there are fewer different data entries between D_1 and D_2 compared with D_1 and D_3, and the differences between D_1 and D_2 are included in the differences between D_1 and D_3. For any randomized algorithm $\mathcal{M}$ satisfies (μ,ε)-WDP with $\mu \geq 1$, we have

$$\begin{aligned} &W_\mu(Pr_{\mathcal{M}}(D_1), Pr_{\mathcal{M}}(D_3)) \\ &\quad \leq W_\mu(Pr_{\mathcal{M}}(D_1), Pr_{\mathcal{M}}(D_2)) \\ &\qquad + W_\mu(Pr_{\mathcal{M}}(D_2), Pr_{\mathcal{M}}(D_3)). \end{aligned} \quad (7)$$

The proof is available in the appendix, and Minkowski's inequality is applied in the deduction process. Proposition 2 can also be understood as the cost that converting from $Pr_{\mathcal{M}}(D_1)$ to $Pr_{\mathcal{M}}(D_2)$ and then to $Pr_{\mathcal{M}}(D_3)$ is not lower than the cost that converting from $Pr_{\mathcal{M}}(D_1)$ to $Pr_{\mathcal{M}}(D_3)$ directly. Triangle inequality is indispensable in proving several properties, such as basic sequential composition (see Proposition 6), group privacy (see Proposition 13) and advanced composition (see Theorem 1).

Proposition 3 (Non-Negativity). For $\mu \geq 1$ and any randomized algorithm $\mathcal{M}$, we have $W_\mu(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D')) \geq 0$.

Proof. See proof of Proposition 3 in the appendix.

Proposition 4 (Monotonicity). For $1 \leq \mu_1 \leq \mu_2$, we have $W_{\mu_1}(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D')) \leq W_{\mu_2}(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D'))$, or we can equivalently described this proposition as (μ_2,ε)-WDP implies (μ_1,ε)-WDP.

The proof is available in the appendix, and the derivation is completed with the help of Lyapunov's inequality.

Proposition 5 (Parallel Composition). Suppose a dataset D is divided into n parts disjointly which are denoted as $D_i, i = 1, 2, \cdots, n$. Each randomized algorithm $\mathcal{M}_i$ performed on different seperated datasets D_i respectively. If $\mathcal{M}_i : \mathcal{D} \to \mathcal{R}_i$ satisfies (μ,ε_i)-WDP for $i = 1, 2, \cdots, n$, then the set of randomized algorithms $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_n\}$ satisfies $(\mu, \max\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\})$-WDP.

Proof. See proof of Proposition 5 in the appendix.

Proposition 6 (Sequential Composition). Consider a series of randomized algorithms $\mathcal{M} = \{\mathcal{M}_1, \cdots, \mathcal{M}_i, \cdots, \mathcal{M}_n\}$ performed on a dataset sequentially. If any $\mathcal{M}_i : \mathcal{D} \to \mathcal{R}_i$ satisfies (μ,ε_i)-WDP, then $\mathcal{M}$ satisfies $(\mu, \sum_{i=1}^n \varepsilon_i)$-WDP.

Proof. See proof of Proposition 6 in the appendix.

Proposition 7 (Laplace Mechanism). If an algorithm $f : \mathcal{D} \to \mathcal{R}$ has sensitivity $\Delta_p f$ and the order $\mu \geq 1$, then the Laplace mechanism $\mathcal{M}_L = f(x) + Lap(0,\lambda)$ preserves $\left(\mu, \frac{1}{2}\Delta_p f \left(\sqrt{2\left[1/\lambda + \exp(-1/\lambda) - 1\right]}\right)^{\frac{1}{\mu}}\right)$-WDP.

Proof. See proof of Proposition 7 in the appendix.

Proposition 8 (Gaussian Mechanism). If an algorithm $f : \mathcal{D} \to \mathcal{R}$ has sensitivity $\Delta_p f$ and the order $\mu \geq 1$, then the Gaussian mechanism $\mathcal{M}_G = f(x) + \mathcal{N}\left(0, \sigma^2\right)$ preserves $\left(\mu, \frac{1}{2}\left(\Delta_p f/\sigma\right)^{\frac{1}{\mu}}\right)$-WDP.

The proof of Gaussian mechanism is available in the appendix. The relation between parameters and privacy budgets in Laplace mechanism and Gaussian mechanism are summarized in Table 1.

Proposition 9 (From DP to WDP) If $\mathcal{M}$ preserves ε-DP with sensitivity Δf , it also satisfies $\left(\mu, \frac{1}{2}\Delta_p f\left(2\varepsilon \cdot (e^\varepsilon - 1)\right)^{\frac{1}{2\mu}}\right)$-WDP.

Proof. See proof of Proposition 9 in the appendix.

Proposition 10 (From RDP to WDP) If $\mathcal{M}$ preserves (α,ε)-RDP with sensitivity $\Delta_p f$, it also satisfies $\left(\mu, \frac{1}{2}\Delta_p f\left(2\varepsilon\right)^{\frac{1}{2\mu}}\right)$-WDP.

Proof. See proof of Proposition 10 in the appendix.

Proposition 11 (From WDP to RDP and DP) Suppose $\mu \geq 1$ and $\log(p_{\mathcal{M}}(\cdot))$ is an L-Lipschitz function. If $\mathcal{M}$ preserves (μ,ε)-WDP with sensitivity $\Delta_p f$, it also satisfies $\left(\alpha, \frac{\alpha}{\alpha-1}L \cdot \varepsilon^{\mu/(\mu+1)}\right)$-RDP. Specifically, when $\alpha \to \infty$, $\mathcal{M}$ satisfies $\left(L \cdot \varepsilon^{\mu/(\mu+1)}\right)$-DP.

The proof is available in the appendix. Where $p_{\mathcal{M}}(\cdot)$ is the probability density function of distribution $Pr_{\mathcal{M}}(\cdot)$.

Proposition 12 (Post-Processing). Let $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ be a (μ,ε)-Wasserstein differentially private algorithm. Let $\mathcal{G} : \mathcal{R} \to \mathcal{R}'$ be an arbitrary randomized mapping. For any order $\mu \in [1,\infty)$ and all measurable subsets $S \subseteq \mathcal{R}$, $\mathcal{G}(\mathcal{M})(\cdot)$ is also (μ,ε)-Wasserstein differentially private, namely

$$W_\mu\left(Pr[\mathcal{G}(\mathcal{M}(D)) \in S], Pr[\mathcal{G}(\mathcal{M}(D')) \in S]\right) \leq \varepsilon. \quad (8)$$

proof. See proof of Proposition 12 in the appendix.

Proposition 13 (Group Privacy). Let $\mathcal{M} : \mathcal{D} \mapsto \mathcal{R}$ be a (μ,ε)-Wasserstein differentially private algorithm. Then for any pairs of datasets $D, D' \in \mathcal{D}$ differing in k data entries $x_1, \cdots, x_k$ for any $i = 1, \cdots, k$, $\mathcal{M}$ is $(\mu, k\varepsilon)$-Wasserstein differentially private.

Differential Privacy Framework	Laplace Mechanism	Gaussian Mechanism
DP	$1/\lambda$	∞
RDP for order α	$\alpha > 1$: $\frac{1}{\alpha-1}\log\left\{\frac{\alpha}{2\alpha-1}\exp\left(\frac{\alpha-1}{\lambda}\right)+\frac{\alpha-1}{2\alpha-1}\exp\left(-\frac{\alpha}{\lambda}\right)\right\}$ $\alpha = 1$: $1/\lambda + \exp(-1/\lambda) - 1$	$\alpha/(2\sigma^2)$
WDP for order μ	$\frac{1}{2}\Delta_p f\left(\sqrt{2\left[1/\lambda+\exp(-1/\lambda)-1\right]}\right)^{\frac{1}{\mu}}$	$\frac{1}{2}\left(\Delta_p f/\sigma\right)^{\frac{1}{\mu}}$

Table 1: Privacy budgets of DP, RDP and WDP for Basic Mechanisms. The Laplace mechanism and Gaussian mechanism of DP and RDP with sensitivity 1 are obtained from Table 2 in Mironov (2017). When it comes to WDP, the sensitivity $\Delta_p f$ can be an arbitrary positive constant.

Proof. See proof of Proposition 13 in the appendix.

Implementation in Deep Learning

Advanced Composition

To derive advanced composition under WDP, we first define generalized (μ, ε)-WDP.

Definition 5 (Generalized (μ, ε)-WDP) A randomized mechanism $\mathcal{M}$ is generalized (μ, ε)-Wasserstein differentially private if for any two adjacent datasets $D, D' \in \mathcal{D}$ holds that

$$Pr[W_\mu(Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D')) \geq \varepsilon] \leq \delta. \tag{9}$$

According to the above definition, we find that (μ, ε)-WDP can be regarded as a special case of generalized (μ, ε)-WDP when δ tends to zero.

Definition 5 is helpful for designing Wasserstein accountant applied in private deep learning, and we will deduce several necessary theorems based on this notion in the following.

Theorem 1 (Advanced Composition) Suppose a randomized algorithm $\mathcal{M}$ consists of a sequence of (μ, ε)-WDP algorithms $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_T$, which perform on dataset D adaptively and satisfy $\mathcal{M}_t : \mathcal{D} \to \mathcal{R}_t, t \in \{1, 2, \cdots, T\}$. $\mathcal{M}$ is generalized (μ, ε)-Wasserstein differentially private with $\varepsilon > 0$ and $\mu \geq 1$ if for any two adjacent datasets $D, D' \in \mathcal{D}$ hold that

$$\exp\left[\beta \sum_{t=1}^{T} \mathbb{E}(W_\mu(Pr_{\mathcal{M}_t}(D), Pr_{\mathcal{M}_t}(D'))) - \beta\varepsilon\right] \leq \delta. \tag{10}$$

Where β is a customization parameter that satisfies $\beta > 0$.

Proof. See proof of Theorem 1 in the appendix.

Privacy Loss and Absolute Moment

Theorem 2 Suppose an algorithm $\mathcal{M}$ consists of a sequence of private algorithms $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_T$ protected by Gaussian mechanism and satisfying $\mathcal{M}_t : \mathcal{D} \to \mathcal{R}$, $t = \{1, 2, \cdots, T\}$. If the subsampling probability, scale parameter and l_2-sensitivity of algorithm $\mathcal{M}_t$ are represented by $q \in [0, 1]$, $\sigma > 0$ and $d_t \geq 0$, then the privacy loss under WDP at epoch t is

$$W_\mu\left(Pr_{\mathcal{M}_t}(D), Pr_{\mathcal{M}_t}(D')\right) = \inf_{d_t}\left[\sum_{i=1}^{n}\mathbb{E}\left(|Z_{ti}|^\mu\right)\right]^{\frac{1}{\mu}},$$
$$Z_t \sim \mathcal{N}\left(qd_t, (2-2q+2q^2)\sigma^2\right). \tag{11}$$

Where $Pr_{\mathcal{M}_t}(D)$ is the outcome distribution when performing $\mathcal{M}_t$ on D at epoch t. $d_t = \|g_t - g_t'\|_2$ represents the l_2 norm between pairs of adjacent gradients g_t and g_t'. In addition, Z_t is a vector follows Gaussian distribution, and Z_{ti} represents the i-th component of Z_t.

Proof. See proof of Theorem 2 in the appendix.

Note that $\mathbb{E}\left(|Z_{ti}|^\mu\right)$ is the μ-order raw absolute moment of the Gaussian distribution $\mathcal{N}\left(qd_t, (2-2q+2q^2)\sigma^2\right)$. We know that the raw moment of a Gaussian distribution can be obtained by taking the μ-th order derivatives of the moment generating function with respect to z. Nevertheless, we do not adopt such an indirect approach. We successfully derive a direct formula, as shown in Lemma 1.

Lemma 1 (Raw Absolute Moment) Assume that $Z_t \sim \mathcal{N}(qd_t, (2-2q+2q^2)\sigma^2)$, we can obtain the raw absolute moment of Z as follow

$$\mathbb{E}\left(|Z_t|^\mu\right) = (2Var)^{\frac{\mu}{2}}\frac{GF\left(\frac{\mu+1}{2}\right)}{\sqrt{\pi}}\mathcal{K}\left(-\frac{\mu}{2}, \frac{1}{2}; -\frac{q^2d_t^2}{2Var}\right). \tag{12}$$

Where Var represents the Variance of random variable Z, and can be expressed as $Var = (2-2q+2q^2)\sigma^2$. $GF\left(\frac{\mu+1}{2}\right)$ represents Gamma function as follow

$$GF\left(\frac{\mu+1}{2}\right) = \int_0^\infty x^{\frac{\mu+1}{2}-1}e^{-x}dx, \tag{13}$$

and $\mathcal{K}\left(-\frac{\mu}{2}, \frac{1}{2}; -\frac{q^2d_t^2}{2Var}\right)$ represents Kummer's confluent hypergeometric function as

$$\sum_{n=0}^{\infty}\frac{q^{2n}{d_t}^{2n}}{n!\cdot 4^n(1-q+q^2)^n\sigma^{2n}}\prod_{i=1}^{n}\frac{\mu-2i+2}{1+2i-2}. \tag{14}$$

proof. Our mathematical deduction is based on the work from Winkelbauer (2012), and the proof is available in the appendix.

Wasserstein Accountant in Deep Learning

Next, we will deduce Wasserstein accountant applied in private deep learning. We obtain Theorem 3 based on the above preparations including advanced composition, privacy loss and absolute moment under WDP.

Theorem 3 (Tail Bound) Under the conditions described in Theorem 2, $\mathcal{M}$ satisfies (μ, ε)-WDP for

$$\log \delta = \beta \sum_{t=1}^{T} \inf_{d_t} \left[\mathbb{E} \sum_{i=1}^{n} \left(|Z_{ti}|^{\mu} \right) \right]^{\frac{1}{\mu}} - \beta\varepsilon. \quad (15)$$

Where $Z \sim \mathcal{N}\left(qd_t, (2-2q+2q^2)\sigma^2\right)$ and $d_t = \|g_t - g_t'\|_2$. The proof of Theorem 3 is available in the appendix. In another case, if we have determined δ and want to know the privacy budget ε, then we can utilize the result in Corollary 1.

Corollary 1 Under the conditions described in Theorem 2, $\mathcal{M}$ satisfies (μ, ε)-WDP for

$$\varepsilon = \sum_{t=1}^{T} \inf_{d_t} \left[\sum_{i=1}^{n} \mathbb{E}\left(|Z_{ti}|^{\mu} \right) \right]^{\frac{1}{\mu}} - \frac{1}{\beta} log\delta. \quad (16)$$

Corollary 1 is more commonly used than Theorem 3 since the total privacy budget generated by an algorithm plays a more important role in privacy computing.

Experiments

The experiments in this paper consist of four parts. Firstly, we test Laplace Mechanism and Gaussian Mechanism under RDP and WDP with ever-changing orders. Secondly, we carry out the experiments of composition and compare our Wasserstein accountant with Bayesian accountant and moments accountant. Thirdly, we consider the application scenario of deep learning, and train a convolutional neural network (CNN) optimized by differentially private stochastic gradient descent (DP-SGD) (Abadi et al. 2016) on the task of image classification. At last, we demonstrate the impact of hyperparameter variations on privacy budgets. All the experiments were performed on a single machine with Ubuntu 18.04, 40 Intel(R) Xeon(R) Silver 4210R CPUs @ 2.40GHz, and two NVIDIA Quadro RTX 8000 GPUs.

Basic Mechanisms

We conduct experiments to test Laplace Mechanism and Gaussian Mechnism under RDP and WDP. Our experiments are based on the results of Proposition 7, 8 and Table 1. We set the scale parameters of Laplace mechanism and Gaussian mechanism as 1, 2, 3 and 5 respectively. The order μ of WDP is allowed to varies from 1 to 10, and so is the order α of RDP. We plot the values of privacy budgets ε with increasing orders, and the results are shown in Figure 1.

We can observe that the privacy budgets of WDP increase with μ growing, which corresponds to our monotonicity property (see Proposition 4). More importantly, we find that the privacy budgets of WDP are not susceptible to the order μ, because their curves all exhibit slow upward trends. However, the privacy budgets of RDP experience a steep increase under Gaussian mechanism when the noise scale equals 1, simply because its order α increases. In addition, the slopes of RDP curves with different noise scales are significantly different. These phenomena lead users to confusion about order selection and risk assessment through privacy budgets when utilizing RDP.

Composition

For the convenience of comparison, we adopt the same settings as the composition experiment in Triastcyn and Faltings (2020). We imitate heavy-tailed gradient distributions by generating synthetic gradients from a Weibull distribution with 0.5 as its shape parameter and 50×1000 as its size.

The hyper-parameter σ remains unchanged after being set as 0.2, and the threshold of gradient clipping C is set to $\{0.05, 0.50, 0.75, 0.99\}$-quantiles of gradient norm in turns. To observe the original variations of their privacy budgets, we do not clip gradients. Thus, C only affects Gaussian noise with variance $C^2\sigma^2$ in DP-SGD (Abadi et al. 2016) in this experiment. In addition, we also provide the composition results with gradient clipping in the appendix for comparison.

In Figure 2, we have the following key observations. (1) The curves obtained from Wasserstein accountant (WA) almost replicate the changes and trends depicted by the curves obtained from moments accountant (MA) and Bayesian accountant (BA). (2) The privacy budgets under WA are always the lowest, and this advantage becomes more significant with C increasing.

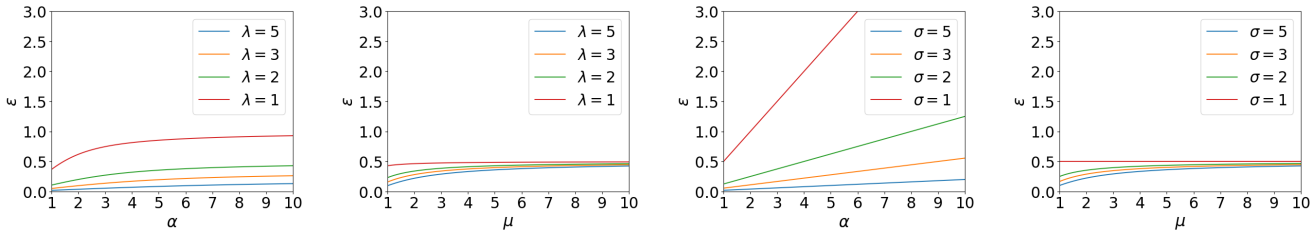
The above results show that Wasserstein accountant can retain the privacy features expressed by MA and BA at a lower privacy budget.

Deep Learning

We adopt DP-SGD (Abadi et al. 2016) as the private optimizer to obtain the privacy budgets under MA, BA and our WA when applying a CNN model designed by Triastcyn and Faltings (2020) to the task of image classification on four baseline datasets including MNIST (Lecun et al. 1998), CIFAR-10 (Krizhevsky and Hinton 2009), SVHN (Netzer et al. 2011) and Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017).

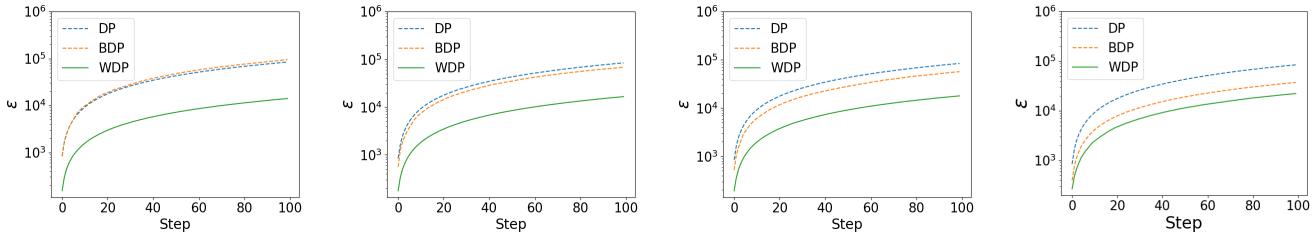
In the experiment of deep learning, we allow different DP frameworks to adjust the noise scale σ according to their own needs. The reasons are as follows: (1) MA supported by DP can easily lead to gradient explosion when the noise scale is small, thus σ can only take a relatively larger value to avoid this situation. However, an excessive noise limits the performance of BDP and WDP. (2) In addition, this setting enables our experimental results more convenient to compare with that in BDP (Triastcyn and Faltings 2020), because the deep learning experiment in BDP is also designed in this way.

Table 2 shows the results obtained under the above experimental settings. We can observe the following phenomenons: (1) WDP requires lower privacy budgets than DP and RDP to achieve the same level of test accuracy. (2) The convergence speed of the deep learning model under WA is



(a) LM for RDP (b) LM for WDP (c) GM for RDP (d) GM for WDP

Figure 1: Privacy buget curves of (μ, ε)-WDP and (α, ε)-RDP for Laplace mechanism (LM) and Gaussian mechanism (GM) with varying orders. Where λ and σ is the scale of LM and GM respectively. The sensitivities are set to 1 and remains unchanged.



(a) 0.05-quantile of $\|g_t\|$ (b) 0.50-quantile of $\|g_t\|$ (c) 0.75-quantile of $\|g_t\|$ (d) 0.99-quantile of $\|g_t\|$

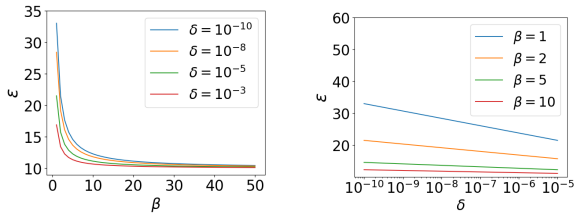
Figure 2: Privacy budgets over synthetic gradients obtained by moments accountant under DP, Bayesian accountant under BDP and Wasserstein accountant under WDP without gradient clipping.

faster than that of MA and BA. Taking the experiments on MNIST dataset as an example, DP and BDP need more than 100 epochs and 50 epochs of training respectively to achieve the accuracy of 96%. While our WDP can reach the same level after only 16 epochs of training.

BDP (Triastcyn and Faltings 2020) attributes its better performance than DP to considering the gradient distribution information. Similarly, we can also analyze the advantages of WDP from the following aspects. (1) From the perspective of definition, WDP also utilizes gradient distribution information through $\gamma \in (Pr_{\mathcal{M}}(D), Pr_{\mathcal{M}}(D'))$. From the perspective of Wasserstein accountant, the information of gradient distribution is included in d_t and Z_t. (2) More importantly, privacy budgets under WDP will not explode even under low noise conditions. Because Wasserstein distance is more stable than Renyi divergence or maximum divergence, which is similar to the reason why WGAN (Arjovsky, Chintala, and Bottou 2017) succeed to alleviate the problem of mode collapse by applying Wasserstein distance.

Effect of β and δ

We also conduct experiments to illustrate the relation between privacy budgets and related hyperparameters. Our experiments are based on the results from Theorem 3 and Corollary 1, which have been proved before. In Figure 3(a), the hyperparameter β in WDP are allowed to varies from 1 to 50, and the failure probability δ of WDP can only be $\{10^{-10}, 10^{-8}, 10^{-5}, 10^{-3}\}$. While in Figure 3(b), the failure probability δ is allowed to varies from 10^{-10} to 10^{-5}, and the hyperparameter β under WDP can only be



(a) ε varies with β (b) ε varies with δ

Figure 3: The impact of β and δ. The coordinates of horizontal axis in 3(b) are on a logarithmic scale.

$\{1, 2, 5, 10\}$. We observe that β has a clear effect on the value of ε in Figure 3(a). ε decreases quickly when β is less than 10, while very slowly when it is greater than 10. When it comes to 3(b), ε seems to be decreasing uniformly with the exponential growth of delta.

Discussion

Relations to Other DP Frameworks

We establish the bridges between WDP, DP and RDP through Proposition 9, 10 and 11. We know that ε-DP implies $\left(\mu, \frac{1}{2}\Delta_p f \left(2\varepsilon \cdot (e^{\varepsilon} - 1)\right)^{\frac{1}{2\mu}}\right)$-WDP and (α, ε)-RDP implies $\left(\mu, \frac{1}{2}\Delta_p f \left(2\varepsilon\right)^{\frac{1}{2\mu}}\right)$-WDP. In addition, (μ, ε)-WDP implies $\left(\alpha, \frac{\alpha}{\alpha-1} L \cdot \varepsilon^{\mu/(\mu+1)}\right)$-RDP or $\left(L \cdot \varepsilon^{\mu/(\mu+1)}\right)$-DP.

Dataset	Accuracy		Privacy		
	Non Private	Private	DP ($\delta = 10^{-5}$)	BDP ($\delta = 10^{-10}$)	WDP ($\delta = 10^{-10}$)
MNIST	99%	96%	2.2 (0.898)	0.95 (0.721)	**0.76 (0.681)**
CIFAR-10	86%	73%	8.0 (0.999)	0.76 (0.681)	**0.52 (0.627)**
SVHN	93%	92%	5.0 (0.999)	0.87 (0.705)	**0.40 (0.599)**
F-MNIST	92%	90%	2.9 (0.623)	0.91 (0.713)	**0.45 (0.611)**

Table 2: Privacy budgets accounted by DP, BDP and WDP on MNIST, CIFAR-10, SVHN and Fashion-MNIST (F-MNIST). The values in parentheses are the probability of potential attack success computed by $P(A) = 1/(1 + e^{-\varepsilon})$ (see Section 3 in Triastcyn and Faltings (2020)).

With the above basic conclusions, we can obtain more derivative relationships through RDP or DP. For example, we can obtain that (μ, ε)-WDP implies $\frac{1}{2}\left(L \cdot \varepsilon^{\mu/(\mu+1)}\right)^2$-zCDP (zero-concentrated differentially private) according to Proposition 1.4 in Bun and Steinke (2016),

Advantages from Metric Property

The privacy losses of DP, RDP and BDP are all non-negative but asymmetric, and do not satisfy triangle inequality (Mironov 2017). Several obvious advantages of WDP as a metric DP have been mentioned in the introduction (see Section) and verified in the experiments (see Section), and here we provide more additional details.

Triangle inequality. (1) Several properties including basic sequential composition, group privacy and advanced composition are derived from triangle inequality. (2) Properties in WDP are more comprehensible and easier to utilize than those in RDP. For example, RDP have to introduce additional conditions of 2^c-stable and $\alpha \geq 2^{c+1}$ to derive group privacy (see Proposition 2 in Mironov (2017)), where c is a constant. In contrast, our WDP utilizes its intrinsic triangle inequality to obtain group privacy without introducing any complex concepts or conditions.

Symmetry. We have considered that the asymmetry of privacy loss would not be transferred to the privacy budget. Specifically, even if $D_\alpha(Pr_{\mathcal{M}}(D)\|Pr_{\mathcal{M}}(D')) \neq D_\alpha(Pr_{\mathcal{M}}(D')\|Pr_{\mathcal{M}}(D))$, $D_\alpha(Pr_{\mathcal{M}}(D)\|Pr_{\mathcal{M}}(D')) \leq \varepsilon$ still implies $D_\alpha(Pr_{\mathcal{M}}(D')\|Pr_{\mathcal{M}}(D)) \leq \varepsilon$, because neighboring datasets D and D' can be all possible pairs. Even so, symmetrical privacy loss still has at least two advantages: (1) When computing privacy budgets, it can reduce the amount of computation for traversing adjacent datasets by half. (2) When proving properties, it is not necessary to exchange datasets and deduce it again like non-metric DP (e.g. see Proof of Theorem 3 in Triastcyn and Faltings (2020)).

Limitations

WDP has excellent mathematical properties as a metric DP, and can effectively alleviate exploding privacy budgets as an alternative DP framework. However, when the volume of data in the queried database is extremely small, WDP may release a much smaller privacy budget than other DP frameworks. Fortunately, this situation only occurs when there is very little data available in the dataset. WDP has great potential in deep learning that requires a large amount of data to train neural network models.

Additional Specifications

Other possibility. Symmetry can be obtained by replacing Rényi divergence with Jensen-Shannon divergence (JSD) (Rao and Nayak 1985). While JSD does not satisfy the triangle inequality unless we take its square root instead (Osán, Bussandri, and Lamberti 2018). Nevertheless, it still tends to exaggerate privacy budgets excessively, as it is defined based on divergence.

Comparability. Another question worth explaining is why the privacy budgets obtained by DP, RDP, and WDP can be compared. (1) Their process of computing privacy budgets follows the same mapping, namely $\mathcal{M} : \mathcal{D} \to \mathcal{R}$. (2) They are essentially measuring the differences in distributions between adjacent datasets, although their respective measurement methods are different. (3) Privacy budgets can be uniformly transformed into the probability of successful attacks (Triastcyn and Faltings 2020).

Computational problem. Although obtaining the Wasserstein distance requires relatively high computational costs (Dudley 1969; Fournier and Guillin 2015), we do not need to worry about this issue. Because WDP does not need to directly calculate the Wasserstein distance no matter in basic privacy mechanisms or Wasserstein accountant for deep learning (see Proposition 7-8 and Theorem 1-3).

Conclusion

In this paper, we propose an alternative DP framework called Wasserstein differential privacy (WDP) based on Wasserstein distance. WDP satisfies the properties of symmetry, triangle inequality and non-negativity that other DPs do not satisfy all, which enables the privacy losses under WDP to become real metrics. We prove that WDP has several excellent properties (see Proposition 1-13) through Lyapunov's inequality, Minkowski's inequality, Jensen's inequality, Markov's inequality, Pinsker's inequality and triangle inequality. We also derive advanced composition theorem, privacy loss and absolute moment under the postulation of WDP and finally obtain Wasserstein accountant to compute cumulative privacy budgets in deep learning (see Theorem 1-3 and Lemma 1). Our evaluations on basic mechanisms, compositions and deep learning show that WDP enables privacy budgets to be more stable and can effectively avoid the overestimation or even explosion on privacy.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 72293583, No. 72293580), Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901), Defense Industrial Technology Development Program (JCKY2019204A007) and Sino-German Research Network (GZ570).

References

Abadi, M.; Chu, A.; Goodfellow, I. J.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 308–318.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, 214–223.

Bobkov, S.; and Ledoux, M. 2019. One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances. *Memoirs of the American Mathematical Society*, 261(1259).

Bun, M.; Dwork, C.; Rothblum, G. N.; and Steinke, T. 2018. Composable and Versatile Privacy via Truncated CDP. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 74–86. ACM.

Bun, M.; and Steinke, T. 2016. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In *Theory of Cryptography Conference (TCC)*, volume 9985, 635–658.

Cheng, A.; Wang, J.; Zhang, X. S.; Chen, Q.; Wang, P.; and Cheng, J. 2022. DPNAS: Neural Architecture Search for Deep Learning with Differential Privacy. In *Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 6358–6366.

Clement, P.; and Desch, W. 2008. An Elementary Proof of the Triangle Inequality for the Wasserstein Metric. *Proceedings of the American Mathematical Society*, 136(1): 333–339.

Dharangutte, P.; Gao, J.; Gong, R.; and Yu, F. 2023. Integer Subspace Differential Privacy. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 7349–7357. AAAI Press.

Dong, J.; Roth, A.; and Su, W. J. 2022. Gaussian Differential Privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1): 3–37.

Dudley, R. M. 1969. The Speed of Mean Glivenko-Cantelli Convergence. *Annals of Mathematical Statistics*, 40: 40–50.

Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; and Naor, M. 2006a. Our Data, Ourselves: Privacy via Distributed Noise Generation. In Vaudenay, S., ed., *25th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, volume 4004, 486–503. Springer.

Dwork, C.; and Lei, J. 2009. Differential Privacy and Robust Statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, 371–380.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. D. 2006b. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference (TCC)*, volume 3876, 265–284. Springer.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theory Computer Science*, 9(3-4): 211–407.

Dwork, C.; and Rothblum, G. N. 2016. Concentrated Differential Privacy. *arXiv preprint arXiv:1603.01887*.

Erven, T. V.; and Harremoës, P. 2014. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions Information Theory*, 60(7): 3797–3820.

Fedotov, A. A.; Harremoës, P.; and Topsøe, F. 2003. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6): 1491–1498.

Fournier, N.; and Guillin, A. 2015. On the Rate of Convergence in Wasserstein Distance of the Empirical Measure. *Probability Theory and Related Fields*, 162: 707–738.

Gao, J.; Gong, R.; and Yu, F. 2022. Subspace Differential Privacy. In *Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 3986–3995.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 5767–5777.

Jin, H.; and Chen, X. 2022. Gromov-Wasserstein Discrepancy with Local Differential Privacy for Distributed Structural Graphs. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2115–2121.

Kantorovich, L. V.; and Rubinshten, G. S. 1958. On a Space of Completely Additive Functions. *Vestnik Leningrad Univ*, 13(7): 52–59.

Kasiviswanathan, S. P.; Lee, H. K.; Nissim, K.; Raskhodnikova, S.; and Smith, A. D. 2011. What Can We Learn Privately? *SIAM Journal on Computing*, 40(3): 793–826.

Krizhevsky, A.; and Hinton, G. 2009. Learning Multiple Layers of Features from Tiny Images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

McSherry, F. 2009. Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, 19–30.

Mironov, I. 2017. Rényi Differential Privacy. In *30th IEEE Computer Security Foundations Symposium (CSF)*, 263–275.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Osán, T. M.; Bussandri, D. G.; and Lamberti, P. W. 2018. Monoparametric Family of Metrics Derived from Classical Jensen–Shannon Divergence. *Physica A: Statistical Mechanics and its Applications*, 495: 336–344.

Panaretos, V. M.; and Zemel, Y. 2019. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6(1).

Phan, N.; Vu, M. N.; Liu, Y.; Jin, R.; Dou, D.; Wu, X.; and Thai, M. T. 2019. Heterogeneous Gaussian Mechanism: Preserving Differential Privacy in Deep Learning with Provable Robustness. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 4753–4759.

Rakotomamonjy, A.; and Ralaivola, L. 2021. Differentially Private Sliced Wasserstein Distance. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, 8810–8820.

Rao, C.; and Nayak, T. 1985. Cross entropy, Dissimilarity Measures, and Characterizations of Quadratic Entropy. *IEEE Transactions on Information Theory*, 31(5): 589–593.

Rüschendorf, L. 2009. Optimal Transport. Old and New. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 111(2): 18–21.

Shokri, R.; and Shmatikov, V. 2015. Privacy-Preserving Deep Learning. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 1310–1321.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (SP)*, 3–18.

Tien, N. L.; Habrard, A.; and Sebban, M. 2019. Differentially Private Optimal Transport: Application to Domain Adaptation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2852–2858.

Triastcyn, A.; and Faltings, B. 2020. Bayesian Differential Privacy for Machine Learning. In *International Conference on Machine Learning (ICML)*, 9583–9592.

Wang, Y.; Si, C.; and Wu, X. 2015. Regression Model Fitting under Differential Privacy and Model Inversion Attack. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1003–1009.

Winkelbauer, A. 2012. Moments and Absolute Moments of the Normal Distribution. *arXiv preprint arXiv:1209.4340*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep Leakage from Gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 14747–14756.