

PRIVATE MEASURES, RANDOM WALKS, AND SYNTHETIC DATA

MARCH BOEDIHARDJO, THOMAS STROHMER, AND ROMAN VERSHYNIN

ABSTRACT. Differential privacy is a mathematical concept that provides an information-theoretic security guarantee. While differential privacy has emerged as a de facto standard for guaranteeing privacy in data sharing, the known mechanisms to achieve it come with some serious limitations. Utility guarantees are usually provided only for a fixed, *a priori* specified set of queries. Moreover, there are no utility guarantees for more complex—but very common—machine learning tasks such as clustering or classification. In this paper we overcome some of these limitations. Working with metric privacy, a powerful generalization of differential privacy, we develop a polynomial-time algorithm that creates a *private measure* from a data set. This private measure allows us to efficiently construct private synthetic data that are accurate for a wide range of statistical analysis tools. Moreover, we prove an asymptotically sharp min-max result for private measures and synthetic data for general compact metric spaces. A key ingredient in our construction is a new *superregular random walk*, whose joint distribution of steps is as regular as that of independent random variables, yet which deviates from the origin logarithmically slowly.

CONTENTS

1. Introduction	1
2. Background and Notation	5
3. A superregular random walk	7
4. Metric privacy	14
5. A private measure on the line	15
6. The Traveling Salesman Problem	18
7. A private measure on a metric space	22
8. A lower bound	26
9. Examples and asymptotics	29
Acknowledgement	33
References	33

1. INTRODUCTION

1.1. Motivation. The right to privacy is enshrined in the Universal Declaration of Human Rights [7]. However, as artificial intelligence is more and more permeating our daily lives, data sharing is increasingly locking horns with data privacy concerns. Differential privacy (DP), a probabilistic mechanism that provides an information-theoretic privacy guarantee, has emerged as a de facto standard for implementing privacy in data sharing [23]. For instance, DP has been adopted by several tech companies [21] and will also be used in connection with the release of the Census 2020 data [3, 2].

Yet, current embodiments of DP come with some serious limitations [18, 26, 52]:

- (i) Utility guarantees are usually provided only for a fixed set of queries. This means that either DP has to be used in an interactive scenario or the queries have to be specified in advance.
- (ii) There are no utility guarantees for more complex—but very common—machine learning tasks such as clustering or classification.
- (iii) DP can suffer from a poor privacy-utility tradeoff, leading to either insufficient privacy protection or to data sets of rather low utility, thereby making DP of limited use in many applications [18].

Another approach to enable privacy in data sharing is based on the concept of synthetic data [9]. The goal of synthetic data is to create a dataset that maintains the statistical properties of the original data while not exposing sensitive information. The combination of differential privacy with synthetic data has been suggested as a best-of-both-world solutions [24, 9, 31, 35, 13]. While combining DP with synthetic data can indeed provide more flexibility and thereby partially address some of the issues in (i), in and of itself it is not a panacea for the aforementioned problems.

One possibility to construct differentially private synthetic datasets that are not tailored to a priori specified queries is to simply add independent Laplacian noise to each data point. However, the amount noise that has to be added to achieve sufficient DP is too large with respect to maintaining satisfactory utility even for basic counting queries [53], not to mention more sophisticated machine learning tasks.

This raises the fundamental question whether it is even possible to construct in a numerically efficient manner differentially private synthetic data that come with rigorous utility guarantees for a wide range of (possibly complex) queries, while achieving a favorable privacy-utility tradeoff? In this paper we will answer this question to the affirmative.

1.2. A private measure. A main objective of this paper is to construct a *private measure* on a given metric space (T, ρ) . Namely, we design an algorithm that transforms a probability measure μ on T into another probability measure ν on T , and such that this transformation is both private and accurate.

For clarity, let us first consider the special case of empirical measures, where our goal can be understood as creating *differentially private synthetic data*. Specifically, we are looking for a computationally tractable algorithm that transforms true input data $X = (X_1, \dots, X_n) \in T^n$ into synthetic output data $Y = (Y_1, \dots, Y_m) \in T^m$ for some m , and which is ε -differentially private (see Definition 2.1) and such that the empirical measures

$$\mu_X = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \quad \text{and} \quad \mu_Y = \frac{1}{m} \sum_{i=1}^m \delta_{Y_i}$$

are close to each other in the Wasserstein 1-metric (recalled in Section 2.2.2):

$$\mathbb{E} W_1(\mu_X, \mu_Y) \leq \gamma, \tag{1.1}$$

where $\gamma > 0$ is as small as possible.

The main result of this paper is a computationally effective private algorithm whose accuracy γ that is expressed in terms of the multiscale geometry of the metric space (T, ρ) . A consequence of this result, Theorem 9.6, states that if the metric space has Minkowski dimension $d \geq 1$, then, ignoring the dependence on ε and lower-order terms in the exponent, we have

$$\mathbb{E} W_1(\mu_X, \mu_Y) \sim n^{-1/d} \tag{1.2}$$

The dependence on n is optimal and quite intuitive. Indeed, if the true data X consists of n i.i.d. random points chosen uniformly from the unit cube $T = [0, 1]^d$, then the average spacing between these points is of the order $n^{-1/d}$. So our result shows that privacy can be achieved by a *microscopic perturbation*, one whose magnitude is roughly the same as the average spacing between the points.

Our more general result, Theorem 7.2, holds for arbitrary compact metric spaces (T, ρ) and, more importantly, for general input measures (not just empirical ones). To be able to work in such generality, we employ the notion of *metric privacy* which reduces to differential privacy when we specialize to empirical measures (Section 2.1).

1.3. Uniform accuracy over Lipschitz statistics. The choice of the Wasserstein 1-metric to quantify accuracy ensures that all Lipschitz statistics are preserved uniformly. Indeed, by Kantorovich-Rubinstein duality theorem, (1.1) yields

$$\mathbb{E} \sup_f \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{m} \sum_{i=1}^m f(Y_i) \right| \leq \gamma \quad (1.3)$$

where the supremum is over all 1-Lipschitz functions $f : T \rightarrow \mathbb{R}$.

Standard private synthetic data generation methods that come with rigorous accuracy guarantees do so with respect to a *predefined* set of linear queries, such as low-dimensional marginals, see e.g. [8, 44, 22, 13]. While this may suffice in some cases, there is no assurance that the synthetic data behave in the same way as the original data under more complex, but frequently employed, machine learning techniques. For instance, if we want to apply a clustering method to the synthetic data, we cannot be sure that the results we get are close to those for the true data. This can drastically limit effective and reliable analysis of synthetic data.

In contrast, since the synthetic data constructed via our proposed method satisfy a *uniform* bound (1.3), this provides data analysts with a vastly increased toolbox of machine learning methods for which one can expect outcomes that are similar for the original data and the synthetic data.

As concrete examples let us look at two of the most common tasks in machine learning, namely clustering and classification. While not every clustering method will satisfy a Lipschitz property, there do exist Lipschitz clustering functions that achieve state-of-the-art results, see e.g. [32, 55]. Similarly, there is distinct interest in Lipschitz function based classifiers, since they are more robust and less susceptible to adversarial attacks. This includes conventional classification methods such as support vector machines [51] as well as classifiers based on Lipschitz neural networks [50, 10]. These are just a few examples of complex machine learning tools that can be reliably applied to the synthetic data constructed via our private measure algorithm. Moreover, since our results hold for general compact metric spaces, this paves the way for creating private synthetic data for a wide range of data types. We will present a detailed algorithmic and numerical investigation of the proposed method in a forthcoming paper.

1.4. A superregular random walk. The most popular way to achieve privacy is by adding random noise, typically either by adding an appropriate amount of Laplacian noise or Gaussian noise (these methods are aptly referred to as *Laplacian mechanism* and *Gaussian mechanism*, respectively [23]). We, too, can try to make a probability measure μ on T private by discretizing T (replacing it with a finite set of points) and then adding random noise to the weights of the points. Going this route, however, yields suboptimal results. For example, it is

not difficult to check that if T is the interval $[0, 1]$, the accuracy of the Laplacian mechanism can not be better than $n^{-1/2}$, which is suboptimal compared to optimal accuracy n^{-1} in (1.2).

This loss of accuracy is caused by the accumulation of additive noise. Indeed, adding n independent random variables of unit variance produces noise of the order $n^{1/2}$. This prompts a basic probabilistic question: can we construct n random variables that are “close” to being independent, but whose partial sums cancel more perfectly than those of independent random variables? We answer this question affirmatively in Theorem 3.1, where we construct random variables Z_1, \dots, Z_n whose joint distribution is as regular as that of i.i.d. Laplacian random variables, yet whose partial sums grow *logarithmically* as opposed to $n^{1/2}$:

$$\mathbb{E} \max_{1 \leq k \leq n} |Z_1 + \dots + Z_k| = O(\log^2 n).$$

One can think of this as a random walk that is locally similar to the one with i.i.d. steps, but globally is much more bounded. Our construction is a nontrivial modification of Lévy’s construction of Brownian motion. It may be interesting and useful beyond applications to privacy.

1.5. Comparison to existing work. The numerically efficient construction of accurate differentially private synthetic data is highly non-trivial. As case in point, Ullman and Vadhan [45] showed (under standard cryptographic assumptions) that in general it is NP-hard to make private synthetic Boolean data which approximately preserve all two-dimensional marginals. There exists a substantial body of work for generating privacy-preserving synthetic data, cf. e.g. [4, 15, 1, 17, 36], but—unlike our work—without providing any rigorous privacy or accuracy guarantees. Those papers on synthetic data that do provide rigorous guarantees are limited to accuracy bounds for a finite set of a priori specified queries, see for example [8, 12, 44, 22, 13, 14], see also the tutorial [46]. As discussed before, this may suffice for specific purposes, but in general severely limits the impact and usefulness of synthetic data. In contrast, the present work provides accuracy guarantees for a wide range of machine learning techniques. Furthermore, our results hold for general compact metric spaces, as we establish metric privacy instead of just differential privacy.

A special example of the topic investigated in this paper is the publication of differentially private histograms, which is a well studied problem in the privacy literature, see e.g. [27, 40, 37, 54, 53, 38, 56, 2] and Chapter 4 in [34]. In the specific context of histograms, the Haar function based approach to construct a superregular random walk proposed in our paper is related to the wavelet-based method [53] and to other hierarchical histogram partitioning methods [27, 40, 56]. Like our approach, [27, 53] obtain consistency of counting queries across the hierarchical levels, owing to the specific way that noise is added. Also, the accuracy bounds obtained in [27, 53] are similar to ours, as they are also polylogarithmic (although we are able to obtain a smaller exponent). There are, however, several *key differences*. While our approach gives a convenient way to generate accurate and differentially private *synthetic data* Y from true data X , the methods of the aforementioned papers are not suited to create synthetic data. Instead, these methods release answers to queries. Moreover, accuracy is proven for just a *single* given range query and not simultaneously for all queries like we do. This limitation makes it impossible to create accurate synthetic data with the algorithms in [27, 53]. Moreover, unlike the aforementioned papers, our work allows the data to be quite general, since we prove metric privacy and not just differential privacy. Furthermore, our results apply to multi-dimensional data, and are not limited to the one-dimensional setting.

There exist several papers on the private estimation of density and other statistical quantities [28, 19], and sampling from distributions in a private manner is the topic of [41]. While definitely interesting, that line of work is not concerned with synthetic data, and thus there is little overlap with this work.

1.6. The architecture of the paper. The remainder of this paper is organized as follows. We introduce some background material and notation in Section 2, such as the concept of *metric privacy* which generalizes differential privacy. In Section 3 we construct a *superregular random walk* (Theorem 3.1). We analyze metric privacy in more detail in Section 4, where we also provide a link from the general private measure problem to private synthetic data (Lemma 4.1). In Section 5 we use the superregular random walk to construct a *private measure on the interval* $[0, 1]$ (Theorem 5.4). In Section 6 we use a link between the Traveling Salesman Problem and minimum spanning trees to devise a *folding technique*, which we apply in Section 7 to “fold” the interval into a space-filling curve to construct a *private measure on a general metric space* (Theorem 7.2). Postprocessing the private measure with quantization and splitting, we then generate *private synthetic data* in a general metric space (Corollary 7.4). In Section 8 we turn to *lower bounds* for private measures (Theorem 8.5) and synthetic data (Theorem 8.6) on a general metric space. We do this by employing a technique of Hardt and Talwar, which we present in a Proposition 8.1 that identifies general limitations for synthetic data. In Section 9 we illustrate our general results on a specific example of a metric space: the Boolean cube $[0, 1]^d$. We construct a private measure (Corollary 9.1) and private synthetic data (Corollary 9.2) on the cube, and show near optimality of these results in Corollary 9.3 and Corollary 9.4, respectively. Results similar to the ones for the d -dimensional cube hold for arbitrary metric space of Minkowski dimension d . For any such space, we prove an *asymptotically sharp min-max results* for private measures (Theorem 9.5) and synthetic data (Theorem 9.5).

2. BACKGROUND AND NOTATION

The motivation behind the concept of differential privacy is the desire to protect an individual’s data, while publishing aggregate information about the database [23]. Adding or removing the data of one individual should have a negligible effect on the query outcome, as formalized in the following definition.

Definition 2.1 (Differential Privacy [23]). *A randomized algorithm \mathcal{M} gives ε -differential privacy if for any input databases D and D' differing on at most one element, and any measurable subset $S \subseteq \text{range}(\mathcal{M})$, we have*

$$\frac{\mathbb{P}\{\mathcal{M}(D) \in S\}}{\mathbb{P}\{\mathcal{M}(D') \in S\}} \leq \exp(\varepsilon),$$

where the probability is with respect to the randomness of \mathcal{M} .

2.1. Defining metric privacy. While differential privacy is a concept of the discrete world (where datasets can differ in a *single element*), it is often desirable to have more freedom in the choice of input data. The following general notion (which seems to be known under slightly different, and somewhat less general, versions, see e.g. [5] and the references therein) extends the classical concept of differential privacy.

Definition 2.2 (Metric privacy). *Let (T, ρ) be a compact metric space and E be a measurable space. A randomized algorithm $\mathcal{A} : T \rightarrow E$ is called α -metrically private if, for any inputs*

$x, x' \in T$ and any measurable subset $S \subset E$, we have

$$\frac{\mathbb{P}\{\mathcal{A}(x) \in S\}}{\mathbb{P}\{\mathcal{A}(x') \in S\}} \leq \exp(\alpha \rho(x, x')). \quad (2.1)$$

To see how this metric privacy encompasses differential privacy, consider a product space $T = \mathcal{X}^n$ and equip it with the Hamming distance

$$\rho_H(x, x') = |\{i \in [n] : x_i \neq x'_i\}|. \quad (2.2)$$

The α -differential privacy of an algorithm $\mathcal{A} : \mathcal{X}^n \rightarrow E$ can be expressed as

$$\frac{\mathbb{P}\{\mathcal{A}(x) \in S\}}{\mathbb{P}\{\mathcal{A}(x') \in S\}} \leq \exp(\alpha) \quad \text{whenever } \rho_H(x, x') \leq 1. \quad (2.3)$$

Note that (2.3) is equivalent to (2.1) for $\rho = \rho_H$. Obviously, (2.1) implies (2.3). The converse implication can be proved by replacing one coordinate of x by the corresponding coordinate of x' and applying (2.3) $\rho_H(x, x')$ times, then telescoping. Let us summarize:

Lemma 2.3 (MP vs. DP). *Let \mathcal{X} be an arbitrary set. Then an algorithm $\mathcal{A} : \mathcal{X}^n \rightarrow E$ is α -differentially private if and only if \mathcal{A} is α -metrically private with respect to the Hamming distance (2.2) on \mathcal{X}^n .*

Unlike differential privacy, metric privacy goes beyond product spaces, and thus allows the data to be quite general. In this paper, for example, the input data are probability measures. Moreover, metric privacy does away with the assumption that the data sets D, D' be different in a *single element*. This assumption is sometimes too restrictive: general measures, for example, do not break down into natural single elements.

2.2. Distances between measures. This paper will use three classical notions of distance between measures.

2.2.1. Total variation. The total variation (TV) norm [20, Section III.1] of a signed measure μ on a measurable space (Ω, \mathcal{F}) is defined as¹

$$\|\mu\|_{\text{TV}} = \frac{1}{2} \sup_{\Omega = \bigcup_i A_i} \sum_i |\mu(A_i)| \quad (2.4)$$

where the supremum is over all partitions Ω into countably many parts $A_i \in \mathcal{F}$. If Ω is countable, we have

$$\|\mu\|_{\text{TV}} = \frac{1}{2} \sum_{\omega \in \Omega} |\mu(\{\omega\})|. \quad (2.5)$$

The TV distance between two probability measures μ and ν is defined as the TV norm of the signed measure $\mu - \nu$. Equivalently,

$$\|\mu - \nu\|_{\text{TV}} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|.$$

¹The factor 2 is chosen for convenience.

2.2.2. *Wasserstein distance.* Let (Ω, ρ) be a bounded metric space. We define the Wasserstein 1-distance (henceforth simply referred to as Wasserstein distance) between probability measures μ and ν on Ω as [49]

$$W_1(\mu, \nu) = \inf_{\gamma} \int_{\Omega \times \Omega} \rho(x, y) d\gamma(x, y) \quad (2.6)$$

where the infimum is over all *couplings* γ of μ and ν , or probability measures on $\Omega \times \Omega$ whose marginals on the first and second coordinates are μ and ν , respectively. In other words, $W_1(\nu, \mu)$ minimizes the *transportation cost* between the “piles of earth” μ and ν .

The Kantorovich-Rubinstein duality theorem [49] gives an equivalent representation:

$$W_1(\mu, \nu) = \sup_{\|h\|_{\text{Lip}} \leq 1} \left(\int h d\mu - \int h d\nu \right)$$

where the supremum is over all continuous, 1-Lipschitz functions $h : \Omega \rightarrow \mathbb{R}$.

For probability measures μ and ν on \mathbb{R} , the Wasserstein distance has the following representation, according to Vallender [47]:

$$W_1(\mu, \nu) = \|F_\mu - F_\nu\|_{L^1(\mathbb{R})}. \quad (2.7)$$

Here $F_\mu(x) = \mu((-\infty, x])$ is the cumulative distribution function of μ , and similarly for $F_\nu(x)$.

Vallender’s identity (2.7) can be used to define Wasserstein distance for *signed measures* on \mathbb{R} . Moreover, for signed measures on $[0, 1]$, the Wasserstein distance defined this way is always finite, and it defines a pseudometric.

3. A SUPERREGULAR RANDOM WALK

The classical random walk with independent steps of unit variance is not bounded: it deviates from the origin at the expected rate $\sim n^{1/2}$. Surprisingly, there exists a random walk whose joint distribution of steps is as regular as that of independent Laplacians, yet that deviates from the origin *logarithmically* slowly.

Theorem 3.1 (A superregular random walk). *For every $n \in \mathbb{N}$, there exists a probability density of the form $f(z) = \frac{1}{\beta} e^{-V(z)}$ on \mathbb{R}^n that satisfies the following two properties.*

(i) *(Regularity): the potential V is 1-Lipschitz in the ℓ^1 norm, i.e.*

$$|V(x) - V(y)| \leq \|x - y\|_1 \quad \text{for all } x, y \in \mathbb{R}^n. \quad (3.1)$$

(ii) *(Boundedness): a random vector $Z = (Z_1, \dots, Z_n)$ distributed according to the density f satisfies*

$$\mathbb{E} \max_{1 \leq k \leq n} |Z_1 + \dots + Z_k| \leq C \log^2 n, \quad (3.2)$$

and

$$\max_{1 \leq k \leq n} \mathbb{E} |Z_1 + \dots + Z_k| \leq C \log^{\frac{3}{2}} n, \quad (3.3)$$

where $C > 0$ is a universal constant.

3.1. Heuristics. The first candidate for the random walk could be a discretization of the standard Brownian motion on the interval $[0, 1]$. The reflection principle yields the boundedness property in the theorem, even without any logarithmic loss. However, the regularity property fails miserably.

To achieve regularity, we would like to grant Brownian motion more freedom. This will be done by modifying Levy's construction of the Brownian motion. In this construction, the path of a Brownian motion on an interval $[0, n]$ is defined as a random series with respect to the Schauder basis of the space of continuous functions, see also [11, Section IX.1].

To that end, we recall the definition of the *Schauder basis* of triangular functions of $C[0, 1]$. Let I_1 be the semi-open interval $[0, 1)$ and N_q the set of all integers j such that $2^{q-1} \leq j < 2^q$. For any integer j there exists a unique pair of integers (q_j, p_j) such that $j = 2^{q_j} + p_j$ where $0 \leq p_j \leq 2^{q_j-1} - 1$. Let

$$I_j = \left\{ t : \frac{p_j}{2^{q_j-1}} \leq t < \frac{p_j+1}{2^{q_j-1}} \right\}, \quad t_j = \frac{2p_j+1}{2^{q_j}},$$

and define $\bar{\phi}_1(t) = t$ for $t \in I_1$ and

$$\bar{\phi}_j = \begin{cases} 2^{q_j} \left(t - \frac{p_j}{2^{q_j-1}} \right) & \text{for } t \in I_{2j}, \\ 2^{q_j} \left(\frac{p_j+1}{2^{q_j-1}} - t \right) & \text{for } t \in I_{2j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

The modification of this definition from $C[0, 1]$ to $C[0, n]$ is obvious by dilation: $\phi_j(t) = \bar{\phi}_j(t/n)$.

Thus, the basis functions ϕ_j are defined by levels $\ell = 0, 1, 2, \dots$. At level $\ell = 0$, we have two functions ϕ_1 and ϕ_2 , and each level $\ell \geq 1$ contains 2^ℓ functions ϕ_j supported in disjoint intervals of length $n/2^\ell$. Throughout this section, $\ell(j)$ will denote the level the function ϕ_j belongs to, e.g. $\ell(1) = \ell(2) = 0$, $\ell(3) = \ell(4) = 1$, $\ell(5) = \ell(6) = \ell(7) = \ell(8) = 2$, etc. See Figure 1 for an illustration of these functions.

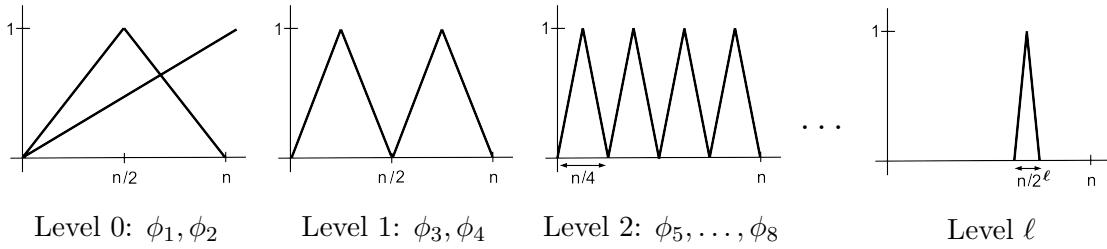


Figure 1. Schauder basis

Lévy's definition of the standard Brownian motion on the interval $[0, n]$ is

$$B_n(t) = \sum_{j=1}^{\infty} 2^{-\ell(j)/2} G_j \phi_j(t), \quad (3.4)$$

where G_j are i.i.d. standard normal random variables.

To grant more freedom to Brownian motion, we get rid of the suppressing factors $2^{-\ell(j)/2}$ in the Levy construction (3.4). The resulting series will be divergent, but we can truncate it

defining

$$W_n(t) = \sum_{j=1}^n G_j \phi_j(t). \quad (3.5)$$

The random walk in the theorem could then be defined as $Z_1 + \dots + Z_k = W_n(k)$. It is more volatile than the Brownian motion, but still can be shown to satisfy the boundedness assumption.

This is essentially the idea of the construction that yields Theorem 3.1. We make two minor modifications though. First, since regularity is defined using ℓ^1 norm, it is more natural to use the Laplacian distribution for G_j instead of the normal distribution. We will make G_j i.i.d. random variables with distribution² $\text{Lap}(\log n)$. Second, instead of defining the random walk and then taking its differences to define Z_i , it is more convenient to define the differences directly. This corresponds to working with the derivative of the random walk, i.e. with “white noise”

$$W'_n(t) = \sum_{j=1}^n G_j \phi'_j(t),$$

and set $Z_i = W'_n(i)$. The derivatives of the functions constituting the Schauder basis $\{\phi_j(t)\}$ form the Haar basis $\{\psi_j(t)\}$, cf. [42, 11]. For notational convenience, we denote

$$\psi_1(t) = \phi'_1(t) = 1 \quad \text{and} \quad \psi_i(t) = \frac{1}{2} \phi'_i(t), \quad i = 2, 3, \dots$$

The Haar basis is illustrated in Figure 2; it is an orthogonal basis of $\ell^2[n]$, see [11].

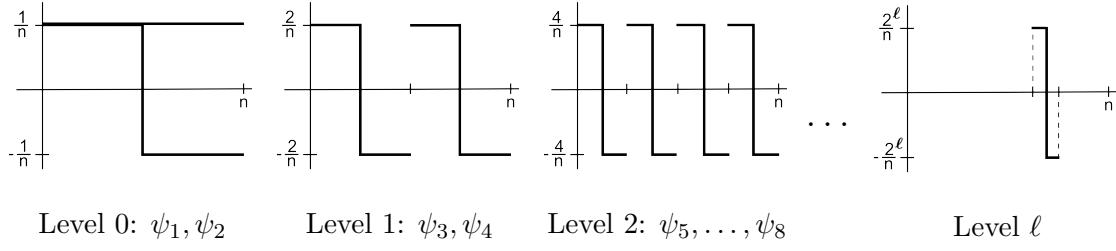


Figure 2. Haar basis

3.2. Formal construction. First observe that the regularity property (3.1) of a probability distribution on \mathbb{R}^n passes on to the marginal distributions. For example, regularity of a random vector $(X_1, X_2) \in \mathbb{R}^2$ means that

$$f_{(X_1, X_2)}(x_1, x_2) \leq \exp(-|x_1 - y_1| - |x_2 - y_2|) f_{(X_1, X_2)}(y_1, y_2),$$

for all $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$. In particular,

$$f_{(X_1, X_2)}(x_1, x_2) \leq \exp(-|x_2 - y_2|) f_{(X_1, X_2)}(x_1, y_2).$$

Taking integral with respect to x_1 on both sides yields

$$f_{X_2}(x_2) \leq \exp(-|x_2 - y_2|) f_{X_2}(y_2),$$

which is equivalent to the regularity of the random vector $X_2 \in \mathbb{R}^1$. The same argument works in higher dimensions.

²Define $X \sim \text{Lap}(\sigma)$ by $\mathbb{P} \{ |X| > t \} = \exp(-t/\sigma)$.

Thus, by dropping at most $n/2$ terms if necessary, we can assume without loss of generality that

$$n = 2^L \quad \text{for some } L \in \mathbb{N}. \quad (3.6)$$

Consider i.i.d. random variables $\Lambda_1, \dots, \Lambda_n \sim \text{Lap}(L+2)$ and the Haar basis ψ_1, \dots, ψ_n of $\ell^2[n]$ introduced in the previous subsection. Define the random function $Z : [n] \rightarrow \mathbb{R}$ by

$$Z = \sum_{j=1}^n \Lambda_j \psi_j.$$

Define the increments by $Z_i = Z(i)$ for $i = 1, \dots, n$. The construction is complete. It remains to check boundedness and regularity.

3.3. Boundedness. Fix $k \in [n]$. We would like to bound the partial sum

$$S_k := \sum_{i=1}^k Z_i = \sum_{i=1}^k Z(i) = \langle Z, \mathbf{1}_{[k]} \rangle = \sum_{j=1}^n \Lambda_j a_{kj} \quad \text{where } a_{kj} = \langle \psi_j, \mathbf{1}_{[k]} \rangle.$$

(Here we use the inner product in $\ell^2[n]$, and denote by $\mathbf{1}_{[k]}$ the indicator of the discrete interval $[k] = \{1, \dots, k\}$.)

For every $k, j \in [n]$ we have

$$|a_{kj}| \leq \|\psi_j\|_1 \|\mathbf{1}_{[1,k]}\|_\infty = 1.$$

Moreover, the random variables $\Lambda_j \sim \text{Lap}(L+2)$ are subexponential,³ and $\|\Lambda_j\|_{\psi_1} \lesssim L$. Hence

$$\|\Lambda_j a_{kj}\|_{\psi_1} \lesssim L \quad \forall k, j \in [n]. \quad (3.7)$$

Furthermore, we claim that for each $k \in [n]$, at most $L+2$ terms a_{kj} are nonzero. Indeed, let us fix k and recall that the definition of Haar functions yields

$$a_{kj} = \langle \psi_j, \mathbf{1}_{[k]} \rangle = 0 \quad \text{if } k \notin \text{supp}(\psi_j).$$

On any given level $\ell \in \{1, \dots, L\}$, the Haar functions ψ_j have disjoint support, so there is a single $j = j(k, \ell)$ for which $k \in \text{supp}(\psi_j)$. Therefore, for each level $\ell \in \{1, \dots, L\}$, there can be at most one nonzero coefficient a_{kj} . Two more nonzero coefficients can be on level $\ell = 0$, coming from the functions ϕ_1 and ϕ_2 . This proves our claim that, for each $k \in [n]$, the number of nonzero coefficients a_{kj} is bounded by $L+2$.

Summarizing, we showed that for each $k \in [n]$, the sum $S_k = \sum_{j=1}^n \Lambda_j a_{kj}$ is a sum of at most $L+2$ independent mean zero subexponential random variables that satisfy (3.7). Applying Bernstein's inequality (see [48, Theorem 2.8.1]), we obtain for every $k \in [n]$ and $t \geq 0$:

$$\mathbb{P} \{ |S_k| \geq t \} \leq 2 \exp \left(-c \min(t^2/L^3, t/L) \right).$$

Let $s \geq 1$ and apply this bound for $t = CsL^2$ where C is a sufficiently large absolute constant. We obtain

$$\mathbb{P} \{ |S_k| \geq CsL^2 \} \leq 2 \exp \left(-2 \min(s^2, s)L \right) \leq \frac{2}{n} \exp(-s),$$

where we used that $s \geq 1$ and (3.6) in the last step. Taking the union bound over $k \in [n]$, we get

$$\mathbb{P} \left\{ \max_{1 \leq k \leq n} |S_k| \geq CsL^2 \right\} \leq 2 \exp(-s) \quad \forall s \geq 1.$$

³For basic facts about subexponential random variables used in this argument, refer e.g. to [48, Section 2.8].

This implies that $\mathbb{E} \max_{1 \leq k \leq n} |S_k| \lesssim L^2 \lesssim \log^2 n$ and proves the first boundedness property (3.2).

The second boundedness property (3.3) follows similarly, and even in a simpler way, if we choose $t = CsL^{3/2}$ and bypass the union bound.

3.4. Regularity. Recall that the Haar functions ψ_1, \dots, ψ_n form an orthogonal basis of $\ell^2[n]$. However, this basis is not orthonormal, as the norm of each function ψ_j on level ℓ satisfies $\|\psi_j\|_2^2 = 2^\ell/n$. Thus, every function $x \in \ell^2[n]$ admits the orthogonal decomposition

$$x = \sum_{j=1}^n \lambda(x)_j \psi_j \quad \text{where } \lambda(x)_j = \frac{n}{2^\ell} \langle \psi_j, x \rangle.$$

The key property of the coefficient vector $\lambda(x)$ is its approximate sparsity, which we can express via the ℓ^1 norm.

Lemma 3.2 (Sparsity). *For any function $x \in \ell^2[n]$, the coefficient vector $\lambda(x)$ satisfies*

$$\|\lambda(x)\|_1 \leq (L+2)\|x\|_1.$$

Proof. First, let us prove the lemma for the indicator of any single point $k \in [n]$, i.e. for $x = \mathbf{1}_{\{k\}}$. Here we have

$$\lambda(x)_j = \frac{n}{2^\ell} \psi_j(k).$$

By construction, any function ψ_j on level ℓ takes on three values: 0 and $\pm 2^\ell/n$. Moreover, on any given level $\ell \geq 1$, the functions ψ_j have disjoint support, so there is a single well-defined $j = j(k, \ell)$ for which $k \in \text{supp}(\psi_j)$. Therefore, among all functions ψ_j on a given level $\ell \geq 1$, only one can make $\lambda(x)_j$ nonzero, namely the one with $j = j(k, \ell)$, and such a nonzero value always equals $(n/2^\ell)(\pm 2^\ell/n) = \pm 1$. Summarizing, the level $\ell = 0$ contributes two nonzero coefficients $\lambda(x)_j$, while each further level $\ell \in \{1, \dots, L\}$ contributes only one. Hence $\lambda(x)$ has $L+2$ nonzero coefficients $\lambda(x)_j$, each taking values ± 1 . Therefore, $\|\lambda(x)\|_1 \leq L+2$.

To extend this bound to a general function $x \in \ell^2[n]$, decompose it as $x = \sum_{k=1}^n x(k) \mathbf{1}_{\{k\}}$. Then, by linearity, $\lambda(x) = \sum_{k=1}^n x(k) \lambda(\mathbf{1}_{\{k\}})$, so

$$\|\lambda(x)\|_1 \leq \sum_{k=1}^n |x(k)| \|\lambda(\mathbf{1}_{\{k\}})\|_1.$$

The bound $\|\lambda(\mathbf{1}_{\{k\}})\|_1 \leq L+2$ from the first part of the argument completes the proof of the lemma. \square

We are ready to prove regularity. Consider the random function $Z = \sum_{j=1}^n \Lambda_j \psi_j$ constructed in Subsection 3.2. In our new notation, the coefficient vector of Z is $\lambda(Z) = (\Lambda_1, \dots, \Lambda_n) =: \Lambda$. We have for any $x, y \in \ell^2[n]$:

$$r(x, y) := \frac{\text{dens}_X(x)}{\text{dens}_X(y)} = \frac{\text{dens}_\Lambda(\lambda(x))}{\text{dens}_\Lambda(\lambda(y))}. \quad (3.8)$$

To see this, recall that the map $x \mapsto \lambda(x)$ is a linear bijection on $\ell^2[n]$. Hence for any $\varepsilon > 0$ and for the unit ball B of $\ell^2[n]$, we have

$$\frac{\mathbb{P}\{X \in x + \varepsilon B\}}{\mathbb{P}\{X \in y + \varepsilon B\}} = \frac{\mathbb{P}\{\Lambda \in \lambda(x) + \varepsilon \lambda(B)\}}{\mathbb{P}\{\Lambda \in \lambda(y) + \varepsilon \lambda(B)\}}.$$

Taking the limit on both sides as $\varepsilon \rightarrow 0_+$ and applying the Lebesgue differentiation theorem yield (3.8).

By construction, the coefficients Λ_i of the random vector $\Lambda \in \mathbb{R}^n$ are $\text{Lap}(L+2)$ i.i.d. random variables. Hence

$$\text{dens}_\Lambda(z) = \frac{1}{2(L+2)} \exp\left(-\frac{\|z\|_1}{L+2}\right), \quad z \in \mathbb{R}^n.$$

Thus,

$$r(x, y) = \exp\left(\frac{\|\lambda(y)\|_1 - \|\lambda(x)\|_1}{L+2}\right).$$

By the triangle inequality and Lemma 3.2, we have

$$\|\lambda(y)\|_1 - \|\lambda(x)\|_1 \leq \|\lambda(x) - \lambda(y)\|_1 \leq (L+2)\|x - y\|_1.$$

Thus

$$r(x, y) \leq \exp(\|x - y\|_1).$$

If we express the density in the form $\text{dens}_X(x) = \frac{1}{\beta} e^{-V}$, the bound we proved can be written as

$$\exp(V(y) - V(x)) \leq \exp(\|x - y\|_1),$$

or $V(y) - V(x) \leq \|x - y\|_1$. Swapping x with y yields $|V(x) - V(y)| \leq \|x - y\|_1$. The proof of Theorem 3.1 is complete. \square

3.5. Optimality. The reader might wonder if the logarithmic factors are necessary in Theorem 3.1. While we do not know if this is the case for the bound (3.3), the logarithmic factor can not be completely removed from the uniform bound (3.2):

Proposition 3.3 (Logarithm is needed in (3.2)). *Let $n \geq 4$ be a natural number and consider a probability density of the form $f(z) = \frac{1}{\beta} e^{-V(z)}$ on \mathbb{R}^n . Assume that the potential V is 1-Lipschitz in the ℓ^1 -norm, i.e. (3.1) holds. Then a random vector $Z = (Z_1, \dots, Z_n)$ distributed according to the density f satisfies*

$$\mathbb{E} \max_{1 \leq k \leq n} |Z_1 + \dots + Z_k| \geq \frac{1}{8} \log n.$$

Proof. Since triangle inequality yields $\|Z\|_\infty = \max_{1 \leq k \leq n} |Z_k| \leq 2 \max_{1 \leq k \leq n} |Z_1 + \dots + Z_k|$, it suffices to check that

$$\mathbb{E} \|Z\|_\infty \geq \frac{1}{4} \log n.$$

Assume for contradiction that this bound fails. Then, considering the cube

$$Q = \left\{ z \in \mathbb{R}^n : \|z\|_\infty < \frac{1}{2} \log n \right\},$$

we obtain by Markov's inequality that $\mathbb{P}\{Z \in Q\} > 1/2$.

Let e_1, \dots, e_n denote the standard basis vectors in \mathbb{R}^n and consider the following translates of the cube Q :

$$Q_i = Q + \frac{1}{2} \log(n) e_i, \quad i = 1, \dots, n.$$

Note the following two properties. First, the cubes Q_i are disjoint. Second, since V is 1-Lipschitz in the ℓ^1 -norm, for each $i = 1, \dots, n$, the densities of the random vectors Z and

$Z - \frac{1}{2} \log(n) e_i$ and (Z_1, \dots, Z_n) differ by a multiplicative factor of at most $\exp(\frac{1}{2} \log n) = \sqrt{n}$ pointwise. Therefore,

$$\mathbb{P}\{Z \in Q_i\} = \mathbb{P}\left\{Z - \frac{1}{2} \log(n) e_i \in Q\right\} \geq \frac{1}{\sqrt{n}} \mathbb{P}\{Z \in Q\} > \frac{1}{2\sqrt{n}}, \quad i = 1, \dots, n.$$

Hence, using these two properties we get

$$1 \geq \sum_{i=1}^n \mathbb{P}\{Z \in Q_i\} > \frac{\sqrt{n}}{2}.$$

It follows that $n < 4$, which contradicts the assumption of the lemma. The proof is complete. \square

3.6. Beyond the ℓ^1 norm? One may wonder why specifically the ℓ^1 norm appears in the regularity property of Theorem 3.1. As we will see shortly, the regularity with respect to the ℓ^1 norm is exactly what is needed in our applications to privacy. However, it might be interesting to see if there are natural extensions of Theorem 3.1 for general ℓ^p norms. The lemma below rules out one such avenue, showing that if a potential V is Lipschitz with respect to the ℓ^p norm for some $p > 1$, the corresponding random walk deviates at least *polynomially* fast (as opposed to logarithmically fast).

Proposition 3.4 (No boundedness for ℓ^p -regular potentials). *Let $n \in \mathbb{N}$ and consider a probability density of the form $f(z) = \frac{1}{\beta} e^{-V(z)}$ on \mathbb{R}^n . Assume that the potential V is 1-Lipschitz in the ℓ^p -norm. Then a random vector $Z = (Z_1, \dots, Z_n)$ distributed according to the density f satisfies*

$$\mathbb{E}|Z_1 + \dots + Z_n| \geq \frac{1}{4} n^{1-\frac{1}{p}}.$$

Proof. We can write $Z_1 + \dots + Z_n = \langle Z, u \rangle$ where $u = (1, \dots, 1)^\top$. Since $\|n^{-\frac{1}{p}} u\|_p = 1$ and V is 1-Lipschitz in the ℓ^p norm, the densities of the random vectors $Z + n^{-\frac{1}{p}} u$ and Z differ by a multiplicative factor of at most e pointwise. Therefore,

$$\begin{aligned} \mathbb{E}|\langle Z, u \rangle| &\geq e^{-1} \mathbb{E}|\langle Z + n^{-\frac{1}{p}} u, u \rangle| \\ &\geq e^{-1} \left(|\langle n^{-\frac{1}{p}} u, u \rangle| - \mathbb{E}|\langle Z, u \rangle| \right) \quad (\text{by triangle inequality}) \\ &= e^{-1} \left(n^{1-\frac{1}{p}} - \mathbb{E}|\langle Z, u \rangle| \right). \end{aligned}$$

Rearranging the terms, we deduce that

$$\mathbb{E}|\langle Z, u \rangle| \geq \frac{e^{-1}}{1 + e^{-1}} n^{1-\frac{1}{p}} \geq \frac{1}{4} n^{1-\frac{1}{p}},$$

which completes the proof. \square

In light of Theorem 3.1 and Proposition 3.4 it might be interesting to see if an obstacle remains for the density $f(z) = \frac{1}{\beta} e^{-V(z)^p}$ for $p > 1$.

4. METRIC PRIVACY

4.1. Private measures. The superregular random walk we just constructed will become the main tool in solving the following *private measure problem*. We are looking for a private and accurate algorithm \mathcal{A} that transforms a probability measure μ on a metric space (T, ρ) into another finitely-supported probability measure $\mathcal{A}(\mu)$ on (T, ρ) .

We need to specify what we mean by privacy and accuracy here. *Metric privacy* offers a natural framework for our problem. Namely, we consider Definition 2.2 for the space $(\mathcal{M}(T), \text{TV})$ of all probability measures on T equipped with the TV metric (recalled in Section 2.2.1). Thus, for any pair of input measures μ and μ' on T that are close in the TV metric, we would like the distributions of the (random) output measures $\mathcal{A}(\mu)$ and $\mathcal{A}(\mu')$ to be close:

$$\frac{\mathbb{P}\{\mathcal{A}(\mu) \in S\}}{\mathbb{P}\{\mathcal{A}(\mu') \in S\}} \leq \exp(\alpha \|\mu - \mu'\|_{\text{TV}}). \quad (4.1)$$

The accuracy will be measured via the *Wasserstein distance* (recalled in Section 2.2.2). We hope to make $W_1(\mathcal{A}(\mu), \mu)$ as small as possible. The reason for choosing W_1 as distance is that it allows us to derive accuracy guarantees for general Lipschitz statistics, as outlined below.

4.2. Synthetic data. The private measure problem has an immediate application for *differentially private synthetic data*. Let (T, ρ) be a compact metric space. We hope to find an algorithm \mathcal{B} that transforms the true data $X = (X_1, \dots, X_n) \in T^n$ into synthetic data $Y = (Y_1, \dots, Y_m) \in T^m$ for some m such that the empirical measures

$$\mu_X = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \quad \text{and} \quad \mu_Y = \frac{1}{m} \sum_{i=1}^m \delta_{Y_i}$$

are close in the Wasserstein distance, i.e. we hope to make $W_1(\mu_X, \mu_Y)$ small. This would imply that synthetic data accurately preserves all Lipschitz statistics, i.e.

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \approx \frac{1}{m} \sum_{i=1}^m f(Y_i)$$

for any Lipschitz function $f : T \rightarrow \mathbb{R}$.

This goal can be immediately achieved if we solve a version of the private measure problem, described in Section 4.1, with the additional requirement that $\mathcal{A}(\mu)$ be an empirical measure. Indeed, define the algorithm \mathcal{B} by feeding the empirical measure μ_X into \mathcal{A} , i.e. set $\mathcal{B}(X) = \mathcal{A}(\mu_X)$. The accuracy follows, and the differential privacy of \mathcal{B} can be seen as follows.

For any pair X, X' of input data that differ in a single element, the corresponding empirical measures differ by at most $1/n$ with respect to the TV distance, i.e.

$$\|\mu_X - \mu_{X'}\|_{\text{TV}} \leq \frac{1}{n}.$$

Then, for any subset S in the output space, we can use (4.1) to get

$$\frac{\mathbb{P}\{\mathcal{B}(X) \in S\}}{\mathbb{P}\{\mathcal{B}(X') \in S\}} = \frac{\mathbb{P}\{\mathcal{A}(\mu_X) \in S\}}{\mathbb{P}\{\mathcal{A}(\mu_{X'}) \in S\}} \leq \exp(\alpha \|\mu - \mu'\|_{\text{TV}}) \leq \exp(\alpha/n).$$

Thus, if $\alpha = \varepsilon n$, the algorithm \mathcal{B} is ε -differentially private. Let us record this observation formally.

Lemma 4.1 (Private measure yields private synthetic data). *Let (T, ρ) be a compact metric space. Let \mathcal{A} be an algorithm that inputs a probability measure on T , and outputs something. Define the algorithm \mathcal{B} that takes data $X = (X_1, \dots, X_n) \in T^n$ as an input, creates the empirical measure μ_X and feeds it into the algorithm \mathcal{A} , i.e. set $\mathcal{B}(X) = \mathcal{A}(\mu_X)$. If \mathcal{A} is α -metrically private in the TV metric and $\alpha = \varepsilon n$, then \mathcal{B} is ε -differentially private.*

Thus, our main focus from now on will be on solving the private measure problem; private synthetic data will follows as a consequence.

5. A PRIVATE MEASURE ON THE LINE

In this section, we construct a private measure on the interval $[0, 1]$. Later we will extend this construction to general metric spaces.

5.1. Discrete input space. Let us start with a somewhat restricted goal, and then work toward wider generality. In this subsection, we will (a) assume that the input measure μ is always supported on some fixed finite subset

$$\Omega = \{\omega_1, \dots, \omega_n\} \quad \text{where } 0 \leq \omega_1 \leq \dots \leq \omega_n \leq 1$$

and (b) allow the output $\mathcal{A}(\mu)$ to be a signed measure. We will measure accuracy with the Wasserstein distance.

5.1.1. Perturbing a measure by a superregular random walk. Apply the Superregular Random Walk Theorem 3.1 and rescale the random variables Z_i by setting $U_i = (2/\alpha)Z_i$. The regularity property of the random vector $U = (U_1, \dots, U_n)$ takes the form

$$\frac{\text{dens}_Y(x)}{\text{dens}_Y(y)} \leq \exp\left(\frac{\alpha}{2}\|x - y\|_1\right) \quad \text{for all } x, y \in \mathbb{R}^n, \quad (5.1)$$

and the boundedness property takes the form

$$\max_{1 \leq k \leq n} \mathbb{E}|U_1 + \dots + U_k| \leq \frac{C \log^{\frac{3}{2}} n}{\alpha}. \quad (5.2)$$

Let us make the algorithm \mathcal{A} perturb the measure μ on Ω by the weights U_i , i.e. we set

$$\mathcal{A}(\mu)(\omega_i) = \mu(\{\omega_i\}) + U_i, \quad i = 1, \dots, n. \quad (5.3)$$

5.1.2. Privacy. Any measure ν on Ω can be identified with the vector $\bar{\nu} \in \mathbb{R}^n$ by setting $\bar{\nu}_i = \nu(\{\omega_i\})$. Then, for any measure η on Ω , we have

$$\text{dens}_{\mathcal{A}(\mu)}(\eta) = \text{dens}_{\bar{\mu} + Y}(\bar{\eta}) = \text{dens}_Y(\bar{\eta} - \bar{\mu}). \quad (5.4)$$

Fix two measures μ and μ' on Ω . By above, we have

$$\begin{aligned} \frac{\text{dens}_{\mathcal{A}(\mu)}(\eta)}{\text{dens}_{\mathcal{A}(\mu')}(\eta)} &= \frac{\text{dens}_Y(\bar{\eta} - \bar{\mu})}{\text{dens}_Y(\bar{\eta} - \bar{\mu}')} \quad (\text{by (5.4)}) \\ &\leq \exp\left(\frac{\alpha}{2}\|\bar{\mu} - \bar{\mu}'\|_1\right) \quad (\text{by (5.1)}) \\ &= \exp\left(\alpha\|\mu - \mu'\|_{\text{TV}}\right) \quad (\text{by (2.5)}). \end{aligned}$$

This shows that the algorithm \mathcal{A} is α -metrically private in the TV metric.

5.1.3. *Accuracy.* If ν and μ are signed measures on Ω , then with the definition (2.7) of the Wasserstein metric for signed measures, we have

$$\begin{aligned} W_1(\nu, \mu) &= \int_0^1 |\nu([0, x]) - \mu([0, x])| dx \\ &= \int_0^1 \left| \sum_{i: \omega_i \leq x} (\nu(\{\omega_i\}) - \sum_{i: \omega_i \leq x} \mu(\{\omega_i\})) \right| dx \\ &= \sum_{k=1}^n (\omega_{k+1} - \omega_k) \left| \sum_{i=1}^k \nu(\{\omega_i\}) - \sum_{i=1}^k \mu(\{\omega_i\}) \right|, \end{aligned}$$

where we set $\omega_{n+1} = 1$.

Applying this general observation for $\nu = \mathcal{A}(\mu)$ and using (5.3), we obtain

$$W_1(\mathcal{A}(\mu), \mu) = \sum_{k=1}^n (\omega_{k+1} - \omega_k) |U_1 + \dots + U_k|.$$

Take expectation on both sides and use (5.2) to conclude that

$$\mathbb{E} W_1(\mathcal{A}(\mu), \mu) \leq \frac{C \log^{\frac{3}{2}} n}{\alpha}.$$

The following result summarizes what we have proved.

Proposition 5.1 (Input in discrete space, output signed measure). *Let Ω be finite subset of $[0, 1]$ and let $n = |\Omega|$. Let $\alpha > 0$. There exists a randomized algorithm \mathcal{A} that takes a probability measure μ on Ω as an input and returns a signed measure ν on Ω as an output, and with the following two properties.*

- (i) *(Privacy): the algorithm \mathcal{A} is α -metrically private in the TV metric.*
- (ii) *(Accuracy): for any input measure μ , the expected accuracy of the output signed measure ν in the Wasserstein distance is*

$$\mathbb{E} W_1(\nu, \mu) \leq \frac{C \log^{\frac{3}{2}} n}{\alpha}.$$

Let ν be the signed measure obtained in Proposition 5.1. Let $\widehat{\nu}$ be a probability measure on Ω that minimizes $W_1(\widehat{\nu}, \nu)$. (The minimizer could be non-unique.) Note that this is a convex problem, since by (??), this problem is equivalent to minimizing

$$\sum_{k=1}^n (\omega_{k+1} - \omega_k) \left| \sum_{i=1}^k \widehat{\nu}(\{\omega_i\}) - \sum_{i=1}^k \nu(\{\omega_i\}) \right|,$$

under the constraints $\widehat{\nu}(\{\omega_i\}) \geq 0$ and $\sum_{i=1}^k \widehat{\nu}(\{\omega_i\}) = 1$.

By minimality, $W_1(\widehat{\nu}, \nu) \leq W_1(\mu, \nu)$. So $W_1(\widehat{\nu}, \mu) \leq W_1(\widehat{\nu}, \nu) + W_1(\nu, \mu) \leq 2W_1(\mu, \nu)$.

Proposition 5.2 (Private measure on a finite subset of the interval). *Let Ω be finite subset of $[0, 1]$ and let $n = |\Omega|$. Let $\alpha > 0$. There exists a randomized algorithm \mathcal{B} that takes a probability measure μ on Ω as an input and returns a probability measure ν on Ω as an output, and with the following two properties.*

- (i) *(Privacy): the algorithm \mathcal{B} is α -metrically private in the TV metric.*

(ii) (Accuracy): for any input measure μ , the expected accuracy of the output measure ν in the Wasserstein distance is

$$\mathbb{E} W_1(\nu, \mu) \leq \frac{C \log^{\frac{3}{2}} n}{\alpha}.$$

5.2. Extending the input space to the interval. Next, we would like to extend our framework to a continuous setting, and allow measures to be supported by the entire interval $[0, 1]$. We can do this by quantization.

5.2.1. Quantization. Fix $n \in \mathcal{N}$ and let $\mathcal{N} = \{\omega_1, \dots, \omega_n\}$ be a $(1/n)$ -net of $[0, 1]$. Consider the proximity partition

$$[0, 1] = I_1 \cup \dots \cup I_n$$

where we put a point $x \in [0, 1]$ into I_i if x is closer to ω_i than to any other points in \mathcal{N} . (We break any ties arbitrarily.)

We can quantize any signed measure ν on $[0, 1]$ by defining

$$\nu_{\mathcal{N}}(\{\omega_i\}) = \nu(I_i), \quad i = 1, \dots, n. \quad (5.5)$$

Obviously, $\nu_{\mathcal{N}}$ is a signed measure on \mathcal{N} . Moreover, if ν is a measure, then so is $\nu_{\mathcal{N}}$. And if ν is a probability measure, then so is $\nu_{\mathcal{N}}$. In the latter case, it follows from the construction that

$$W_1(\nu, \nu_{\mathcal{N}}) \leq 1/n. \quad (5.6)$$

(By definition of the net, transporting any point x to the closest point ω_i covers distance at most $1/n$.)

Lemma 5.3 (Quantization is a contraction in TV metric). *Any signed measure ν on $[0, 1]$ satisfies*

$$\|\nu_{\mathcal{N}}\|_{\text{TV}} \leq \|\nu\|_{\text{TV}}.$$

Proof. Using (2.5), (5.5), and (2.4), we obtain

$$\|\nu_{\mathcal{N}}\|_{\text{TV}} = \frac{1}{2} \sum_{i=1}^n |\nu_{\mathcal{N}}(\{\omega_i\})| = \frac{1}{2} \sum_{i=1}^n |\nu(I_i)| \leq \|\nu\|_{\text{TV}}.$$

The lemma is proved. □

5.2.2. A private measure on the interval.

Theorem 5.4 (Private measure on the interval). *Let $\alpha \geq 2$. There exists a randomized algorithm \mathcal{A} that takes a probability measure μ on $[0, 1]$ as an input and returns a finitely-supported probability measure ν on $[0, 1]$ as an output, and with the following two properties.*

- (i) (Privacy): the algorithm \mathcal{A} is α -metrically private in the TV metric.
- (ii) (Accuracy): for any input measure μ , the expected accuracy of the output measure ν in the Wasserstein distance is

$$\mathbb{E} W_1(\nu, \mu) \leq \frac{C \log^{\frac{3}{2}} \alpha}{\alpha}.$$

Proof. Take a measure μ on $[0, 1]$, preprocess it by quantizing as in the previous subsection, and feed the quantized measure $\mu_{\mathcal{N}}$ into the algorithm \mathcal{B} of Proposition 5.2 for $\Omega = \mathcal{N}$.

The contraction property (Lemma 5.3) ensures that

$$\|\mu_{\mathcal{N}} - \mu'_{\mathcal{N}}\|_{\text{TV}} \leq \|\mu - \mu'\|_{\text{TV}}.$$

This and the privacy property of Proposition 5.2 for measures on \mathcal{N} guarantee that quantization does not destroy privacy, i.e. the algorithm $\mu \mapsto \mathcal{B}(\mu_{\mathcal{N}})$ is still α -metrically private as claimed.

As for the accuracy, Proposition 5.2 for the measure $\mu_{\mathcal{N}}$ gives

$$\mathbb{E} W_1(\mathcal{B}(\mu_{\mathcal{N}}), \mu_{\mathcal{N}}) \leq \frac{C \log^{\frac{3}{2}} n}{\alpha}.$$

Moreover, the accuracy of quantization (5.6) states that $W_1(\mu, \mu_{\mathcal{N}}) \leq 1/n$. By triangle inequality, we conclude that

$$\mathbb{E} W_1(\mathcal{B}(\mu_{\mathcal{N}}), \mu) \leq \frac{1}{n} + \frac{C \log^{\frac{3}{2}} n}{\alpha}.$$

Taking n to be the largest integer less than or equal to α yields the conclusion of the theorem. \square

6. THE TRAVELING SALESMAN PROBLEM

In order to extend the construction of the private measure on the interval $[0, 1]$ to a general metric space (T, ρ) , a natural approach would be to map the interval $[0, 1]$ onto some *space-filling curve* of T . Since a space filling curves usually are infinitely long, we should do this on the discrete level, for some δ -net of T rather than T itself. In this section, we will bound length of such discrete space-filling curve in terms of the metric geometry of T . In the next section, we will see how this bound determines the accuracy of a private measure in T .

A natural framework for this step is related to Traveling Salesman Problem (TSP), which is a central problem in optimization and computer science, and whose history goes back to at least 1832 [6].

Let $G = (V, E)$ be an undirected weighted connected graph. We occasionally refer to the weights of the edges as lengths. A *tour* of G is a connected walk on the edges that visits every vertex at least once, and returns to the starting vertex. The TSP is the problem of finding a tour of G with the shortest length. Let us denote this length by $\text{TSP}(G)$.

Although it is NP-hard to compute $\text{TSP}(G)$, or even to approximate it within a factor of 123/122 [30], an algorithm of Christofides and Serdyukov [16, 43] from 1976 gives a 3/2-approximation for TSP, and it was shown recently that the factor 3/2 can be further improved [29].

6.1. TSP in terms of the minimum spanning tree. Within a factor of 2, the traveling salesman problem is equivalent to another key problem, namely the problem of finding the *minimum spanning tree* (MST) of G . A spanning tree of G is a subgraph that is a tree and which includes all vertices of G . It always exists and can be found in polynomial time [33, 39]. A spanning tree of G with the smallest length is called the minimum spanning tree of G ; we denote its length by $\text{MST}(G)$. The following equivalence is a folklore.

Lemma 6.1. *Any undirected weighted connected graph G satisfies*

$$\text{MST}(G) \leq \text{TSP}(G) \leq 2 \text{MST}(G).$$

Proof. For the lower bound, it is enough to find a spanning tree of G of length bounded by $\text{TSP}(G)$. Consider the minimal tour of G of length $\text{TSP}(G)$ as a subgraph of G . Let T be a spanning tree of the tour. Since the tour contains all vertices of G , so does T , and thus T is a spanning tree of G . Since T is obtained by removing some edges of the tour, the length of T is bounded by of the tour, which is $\text{TSP}(G)$. The lower bound is proved.

For the upper bound, note that dropping any edges of G can only increase the value of TSP. Thus TSP of G is bounded by the TSP of its spanning tree T . Moreover, TSP of any tree T equals twice the sum of lengths of the edges of T . This can be seen by considering the *depth-first search* tour of T , which starts at the root and explores as deep as possible along each branch before backtracking, see Figure 3. \square

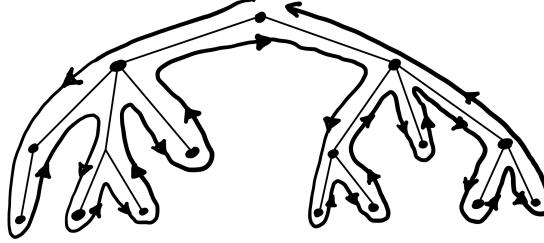


Figure 3. The depth-first search tour demonstrates that the TSP of a tree equals twice the sum of lengths of its edges.

6.2. Metric TSP. Let (T, ρ) be a finite metric space. We can consider T as a complete weighted graph, whose weights of edges are defined as the distances between the points. The TSP for (T, ρ) is known as *metric TSP*.

Although a tour can visit the same vertex of T multiple times, this can be prevented by skipping the vertices previously visited. The triangle inequality shows that skipping can only decrease the length of the tour. Therefore, the shortest tour in a complete graph is always a *Hamiltonian cycle*, a walk that visits all vertices of T exactly once before returning to the starting vertex. Let us record this observation:

Lemma 6.2. *The TSP of a finite metric space (T, ρ) equals the smallest length of a Hamiltonian cycle of T .*

6.3. A geometric bound on TSP. We would like to compute $\text{TSP}(T)$ in terms of the geometry of the metric space (T, ρ) . Here we will prove an upper bound on $\text{TSP}(T)$ in terms of the covering numbers. Recall that the *covering number* $N(T, \rho, \varepsilon)$ is defined as the smallest cardinality of an ε -net of T , or equivalently the smallest number of closed balls with centers in T and radii ε whose union covers T , see [48, Section 4.2].

Theorem 6.3 (TSP via covering numbers). *For any finite metric space (T, ρ) , we have*

$$\text{TSP}(T) \leq 16 \int_0^\infty (N(T, \rho, x) - 1) dx.$$

Proof. Step 1: constructing a spanning tree. Let us construct a small spanning tree T_0 of T and use Lemma 6.1. Let $\varepsilon_j = 2^{-j}$, $j \in \mathbb{Z}$, and let \mathcal{N}_j be ε_j -nets of T with cardinalities $|\mathcal{N}_j| = N(T, \rho, \varepsilon_j)$. Since T is finite, we must have $|\mathcal{N}_j| = 1$ for all sufficiently small j . Let j_0 be the largest integer for which $|\mathcal{N}_{j_0}| = 1$.

At the root of T_0 , let us put a single point that forms the net \mathcal{N}_{j_0} . At the next level, put all the points of the net \mathcal{N}_{j_0+1} , and connect them to the root by edges. The weights of these edges, which are defined as the distances of the points to the root, are all bounded by ε_{j_0} . At the next level, put all points of the net \mathcal{N}_{j_0+2} , and connect each such point to the closest point in the previous level \mathcal{N}_{j_0+1} . (Break any ties arbitrarily.) Since the latter set is

a ε_{j_0+1} -net, the weights of all these edges are bounded by ε_{j_0+1} . Repeat these steps until the levels do not grow anymore, i.e. until the level contains all the points in T ; see Figure 4 for illustration.

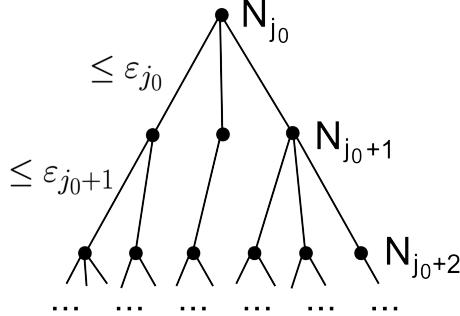


Figure 4. Chaining: construction of a spanning tree of a metric space.

If all the nets \mathcal{N}_j that make up the levels of the tree T_0 are disjoint, then T_0 is a spanning tree of T . Assume that this is the case for time being.

Step 2: bounding the length of the tree. For each of the levels $j = j_0 + 1, j_0 + 2, \dots$, the tree T_0 has $|\mathcal{N}_j|$ edges connecting the points of level j to the level $j - 1$, and each such edge has length (weight) bounded by ε_{j-1} . So $\text{MST}(T)$ is bounded by the sum of the lengths of the edges of T_0 , i.e.

$$\text{MST}(T) \leq \sum_{j=j_0+1}^{\infty} \varepsilon_{j-1} |\mathcal{N}_j|.$$

Step 3: bounding the sum by the integral. Our choice $\varepsilon_j = 2^{-j}$ yields $\varepsilon_{j-1} = 4(\varepsilon_j - \varepsilon_{j+1})$. Moreover, our choice of j_0 yields $|\mathcal{N}_j| \geq 2$ for all $j \geq j_0 + 1$, which implies $|\mathcal{N}_j| \leq 2(|\mathcal{N}_j| - 1)$ for such j . Therefore

$$\begin{aligned} \text{MST}(T) &\leq 8 \sum_{j=j_0+1}^{\infty} (\varepsilon_j - \varepsilon_{j+1}) (|\mathcal{N}_j| - 1) & (6.1) \\ &= 8 \sum_{j=j_0+1}^{\infty} \int_{\varepsilon_{j+1}}^{\varepsilon_j} (N(T, \rho, \varepsilon_j) - 1) dx \quad (\text{since } |\mathcal{N}_j| = N(T, \rho, \varepsilon_j)) \\ &\leq 8 \int_0^{\infty} (N(T, \rho, x) - 1) dx. \end{aligned}$$

An application of Lemma 6.1 completes the proof.

Step 4: splitting. The argument above assumes that all levels \mathcal{N}_j of the tree T_0 are disjoint. This assumption can be enforced by splitting the points of T . If, for example, a point $\omega \in \mathcal{N}_j$ is also used in \mathcal{N}_k for some $k < j$, add to T another a replica of ω – a point ω' that has zero distance to ω and the same distances to all other points as ω . Use ω in \mathcal{N}_j and ω' in \mathcal{N}_k . Preprocessing the metric space (T, ρ) by such splitting yields a pseudometric space (T', ρ) in which all levels \mathcal{N}_j are disjoint, and whose TSP is the same. \square

Remark 6.4 (Integrating up to the diameter). Note that $N(T, \rho, x) = 1$ for any $x > \text{diam}(T)$, since any single point makes an x -net of T for such x . Therefore, the integrand in Theorem 6.3

vanishes for such x , and we have

$$\text{TSP}(T) \leq 16 \int_0^{\text{diam}(T)} N(T, \rho, x) dx. \quad (6.2)$$

6.4. Folding. It is a simple observation that an interval of length $\text{TSP}(T)$ can be embedded, or “folded”, into T :

Proposition 6.5 (Folding). *For any finite metric space (T, ρ) there exists a finite subset Ω of the interval $[0, \text{TSP}(T)]$ and a 1-Lipschitz bijection $F : \Omega \rightarrow T$.*

Heuristically, the map F “folds” the interval $[0, \text{TSP}(T)]$ into the shortest Hamiltonian path of the metric space T , see Figure 5. We can think of this as a *space-filling curve* of T .

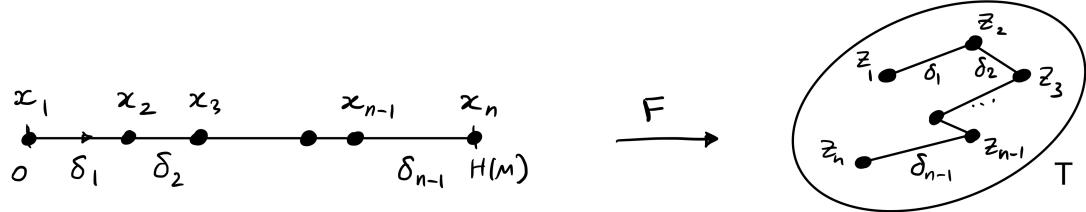


Figure 5. The map F folds an interval $[0, \text{TSP}(M)]$ into a Hamiltonian path (a “space-filling curve”) of the metric space T .

Proof. Let us exploit the heuristic idea of folding. Fix a Hamiltonian cycle in T of length $\text{TSP}(T)$, whose existence is given by Lemma 6.2. Formally, this means that we can label the elements of the space as $T = \{z_1, \dots, z_n\}$ in such a way that the lengths

$$\delta_i = \rho(z_{i+1}, z_i), \quad i = 1, \dots, n-1,$$

satisfy $\sum_{i=1}^{n-1} \delta_i \leq \text{TSP}(T)$. Define $\Omega = \{x_1, \dots, x_n\}$ by

$$x_1 = 0; \quad x_k = \sum_{i=1}^{k-1} \delta_i, \quad k = 2, \dots, n.$$

Then all $x_k \leq \text{TSP}(T)$, so $\Omega \subset [0, \text{TSP}(T)]$ as claimed.

Note that for every $k = 1, \dots, n-1$ we have

$$\rho(z_{k+1}, z_k) = \delta_k = x_{k+1} - x_k.$$

Then, for any integers $1 \leq k \leq k+j \leq n$, triangle inequality and telescoping give

$$\begin{aligned} \rho(z_{k+j}, z_k) &\leq \rho(z_{k+j}, z_{k+j-1}) + \rho(z_{k+j-1}, z_{k+j-2}) + \dots + \rho(z_{k+1}, z_k) \\ &= (x_{k+j} - x_{k+j-1}) + (x_{k+j-1} - x_{k+j-2}) + \dots + (x_{k+1} - x_k) \\ &= x_{k+j} - x_k. \end{aligned}$$

This shows that the folding map $F : x_k \mapsto z_k$ is a bijection that satisfies

$$\rho(F(x), F(y)) \leq |x - y| \quad \text{for all } x, y \in \Omega.$$

In other words, F is 1-Lipschitz. The proof is complete. \square

7. A PRIVATE MEASURE ON A METRIC SPACE

We are ready to construct a private measure on an arbitrary compact metric space (T, ρ) . We do this as follows: (a) discretize T replacing it with a finite δ -net; (b) fold an interval of length $\text{TSP}(T)$ onto T using Proposition 6.5; and (c) using this folding, pushforward onto T the private measure on the interval constructed in Section 5. The accuracy of the resulting private measure on T is determined by the length of the interval $\text{TSP}(T)$, which in turn can be expressed using the covering numbers of T (Theorem 6.3).

7.1. Finite metric spaces. Let us start by extending Proposition 5.2 from a finite subset on $[0, 1]$ to a finite subset of (T, ρ) .

Proposition 7.1 (Private measure on a finite metric space). *Let (T, ρ) be a finite metric space and let $n = |T|$. Let $\alpha > 0$. There exists a randomized algorithm \mathcal{B} that takes a probability measure μ on T as an input and returns a probability measure ν on T as an output, and with the following two properties.*

- (i) (Privacy): the algorithm \mathcal{B} is α -metrically private in the TV metric.
- (ii) (Accuracy): for any input measure μ , the expected accuracy of the output measure ν in the Wasserstein distance is

$$\mathbb{E} W_1(\nu, \mu) \leq \frac{C \log^{\frac{3}{2}} n}{\alpha} \text{TSP}(T).$$

Proof. Applying Folding Proposition 6.5, we obtain an n -element subset $\Omega \subset [0, \text{TSP}(T)]$ and a 1-Lipschitz bijection $F : \Omega \rightarrow T$. Applying Proposition 5.2 and rescaling by the factor $\text{TSP}(T)$, we obtain an α -metrically private algorithm \mathcal{B} that transforms a probability measure μ on Ω into a probability measure ν on Ω , and whose accuracy is

$$\mathbb{E} W_1(\nu, \mu) \leq \frac{C \log^{\frac{3}{2}} n}{\alpha} \text{TSP}(T). \quad (7.1)$$

Define a new metric $\bar{\rho}$ on Ω by $\bar{\rho}(x, y) = \rho(F(x), F(y))$. Since F is 1-Lipschitz, we have $\bar{\rho}(x, y) \leq |x - y|$. Note that the Wasserstein distance can only become smaller if the underlying metric is replaced by a smaller metric. Therefore, the bound (7.1), which holds with respect to the usual metric $|x - y|$ on Ω , automatically holds with respect to the smaller metric $\bar{\rho}(x, y)$.

It remains to note that $(\Omega, \bar{\rho})$ is isometric to (T, ρ) . So the accuracy result (7.1), which as we saw holds in $(\Omega, \bar{\rho})$, automatically transfers to (T, ρ) (by considering the pushforward measure). \square

7.2. General metric spaces. Quantization allows us to pass from discrete metric spaces to general spaces. A similar technique was used in Section 5.2 for the interval $[0, 1]$. We will repeat it here for a general metric space.

7.2.1. Quantization. Fix $\delta > 0$ and let $\mathcal{N} = \{\omega_1, \dots, \omega_n\}$ be a δ -net of T such that $n = |\mathcal{N}| = N(T, \rho, \delta)$. Consider the proximity partition

$$T = I_1 \cup \dots \cup I_n$$

where we put a point $x \in T$ into I_i if x is closer to ω_i than to any other points in \mathcal{N} . (We break any ties arbitrarily.)

We can quantize any signed measure ν on T by defining

$$\nu_{\mathcal{N}}(\{\omega_i\}) = \nu(I_i), i = 1, \dots, n.$$

Obviously, $\nu_{\mathcal{N}}$ is a signed measure on \mathcal{N} . Moreover, if ν is a measure, then so is $\nu_{\mathcal{N}}$. And if ν is a probability measure, then so is $\nu_{\mathcal{N}}$. In the latter case, it follows from the construction that

$$W_1(\nu, \nu_{\mathcal{N}}) \leq \delta. \quad (7.2)$$

(By definition of the net, transporting any point x to the closest point ω_i covers distance at most δ .) Furthermore, Lemma 5.3 easily generalizes and yields

$$\|\nu_{\mathcal{N}}\|_{\text{TV}} \leq \|\nu\|_{\text{TV}}. \quad (7.3)$$

Finally, let us bound the TSP of the net \mathcal{N} using Theorem 6.3. We trivially have $N(\mathcal{N}, \rho, x) \leq |\mathcal{N}| = N(T, \rho, \delta)$ for any $x > 0$. Moreover, since $\mathcal{N} \subset T$, we also have $N(\mathcal{N}, \rho, x) \leq N(T, \rho, x/2)$, see [48, Exercise 4.2.10]. Using the former bound for $x < 2\delta$ and the latter bound for $x \geq 2\delta$ and applying (6.2), we obtain

$$\begin{aligned} \text{TSP}(\mathcal{N}) &\lesssim \int_0^{\text{diam}(\mathcal{N})} N(\mathcal{N}, \rho, x) dx \\ &\leq 2\delta N(T, \rho, \delta) + \int_{2\delta}^{\text{diam}(T)} N(T, \rho, x/2) dx \\ &= 2 \left(\delta N(T, \rho, \delta) + \int_{\delta}^{\text{diam}(T)/2} N(T, \rho, x) dx \right) \\ &\leq 2 \left(2 \int_{\delta/2}^{\delta} N(T, \rho, x) dx + \int_{\delta}^{\text{diam}(T)/2} N(T, \rho, x) dx \right) \\ &\leq 4 \int_{\delta/2}^{\text{diam}(T)/2} N(T, \rho, x) dx. \end{aligned} \quad (7.4)$$

7.2.2. A private measure on a general metric space.

Theorem 7.2 (Private measure on a metric space). *Let (T, ρ) be a compact metric space. Let $\alpha, \delta > 0$. There exists a randomized algorithm \mathcal{A} that takes a probability measure μ on T as an input and returns a finitely-supported probability measure ν on T as an output, and with the following two properties.*

- (i) (Privacy): *the algorithm \mathcal{A} is α -metrically private in the TV metric.*
- (ii) (Accuracy): *for any input measure μ , the expected accuracy of the output measure ν in the Wasserstein distance is*

$$\mathbb{E} W_1(\nu, \mu) \leq 2\delta + \frac{C}{\alpha} \log^{\frac{3}{2}}(N(T, \rho, \delta)) \int_{\delta}^{\text{diam}(T)} N(T, \rho, x) dx.$$

Proof. Preprocess the input measure μ by quantizing as in the previous subsection, and feed the quantized measure $\mu_{\mathcal{N}}$ into the algorithm \mathcal{B} of Proposition 7.1 for the metric space (\mathcal{N}, ρ) .

The contraction property (7.3) ensures that

$$\|\mu_{\mathcal{N}} - \mu'_{\mathcal{N}}\|_{\text{TV}} \leq \|\mu - \mu'\|_{\text{TV}}$$

for any two input measures μ and μ' . This and the privacy property in Proposition 7.1 for measures on \mathcal{N} guarantee that quantization does not destroy privacy, i.e. the algorithm $\mathcal{A} : \mu \mapsto \mathcal{B}(\mu_{\mathcal{N}})$ is still α -metrically private as claimed.

Next, the accuracy property in Proposition 7.1 for the measure $\mu_{\mathcal{N}}$ on \mathcal{N} gives

$$\mathbb{E} W_1(\mathcal{B}(\mu_{\mathcal{N}}), \mu_{\mathcal{N}}) \leq \frac{C}{\alpha} \log^{\frac{3}{2}}(N(T, \rho, \delta)) \text{TSP}(\mathcal{N}).$$

Moreover, the accuracy of quantization (7.2) states that $W_1(\mu, \mu_N) \leq \delta$. By triangle inequality, we conclude that

$$\mathbb{E} W_1(\mathcal{B}(\mu_N), \mu) \leq \delta + \frac{C}{\alpha} \log^{\frac{3}{2}}(N(T, \rho, \delta)) \text{ TSP}(\mathcal{N}).$$

Thus, by (7.4),

$$\mathbb{E} W_1(\mathcal{B}(\mu_N), \mu) \leq \delta + \frac{C}{\alpha} \log^{\frac{3}{2}}(N(T, \rho, \delta)) \int_{\delta/2}^{\text{diam}(T)/2} N(T, \rho, x) dx.$$

Since $N(T, \rho, 2\delta) \leq N(T, \rho, \delta)$, replacing δ by 2δ completes the proof of the theorem. \square

7.3. Private synthetic data. The output of the algorithm \mathcal{A} in Theorem 7.2 is a finitely-supported probability measure ν on T . Quantization allows to transform ν into an *empirical measure*

$$\mu_Y = \frac{1}{m} \sum_{i=1}^m \delta_{Y_i} \quad (7.5)$$

where Y_1, \dots, Y_m is some finite sequence of elements of T , in which repetitions are allowed. In other words, we can make the output of our algorithm a *synthetic data* $Y = (Y_1, \dots, Y_m)$. Let us record this observation.

Corollary 7.3 (Outputting an empirical measure). *Let (T, ρ) be a compact metric space. Let $\alpha, \delta > 0$. There exists a randomized algorithm \mathcal{A} that takes a probability measure μ on T as an input and returns $Y = (Y_1, \dots, Y_m) \in T^m$ for some m as an output, and with the following two properties.*

- (i) *(Privacy): the algorithm \mathcal{A} is α -metrically private in the TV metric.*
- (ii) *(Accuracy): for any input measure μ , the expected accuracy of the empirical measure μ_Y in the Wasserstein distance is*

$$\mathbb{E} W_1(\mu_Y, \mu) \leq 3\delta + \frac{C}{\alpha} \log^{\frac{3}{2}}(N(T, \rho, \delta)) \int_{\delta}^{\text{diam}(T)} N(T, \rho, x) dx.$$

Proof. Since the output probability measure ν in Theorem 7.2 is finitely supported, it has the form

$$\nu = \sum_{i=1}^r w_i \delta_{Y_i}$$

for some natural number r , positive weights w_i and elements $R_i \in T$.

Let us quantize the weights w_i by the uniform quantizer with step $1/m$ where m is a large integer. Namely, set

$$q(w_i) := \frac{\lfloor m w_i \rfloor}{m}.$$

Obviously, the total quantization error satisfies

$$\kappa := \sum_{i=1}^r (w_i - q(w_i)) \in [0, r/m]. \quad (7.6)$$

To make the quantized weights a probability measure, let us add the total quantization error to any given weight, say the first. Thus, define

$$w'_1 := q(w_1) + \kappa \quad \text{and} \quad w'_i := q(w_i), \quad i = 2, \dots, r$$

and set

$$\nu' := \sum_{i=1}^r w'_i \delta_{Y_i}.$$

Note the three key properties of ν' . First, since the weights w'_i sum to one, ν' is a probability measure. Second, since ν' is obtained from ν by transporting a total mass of κ across the metric space T , we have

$$W_1(\nu, \nu') \leq \kappa \cdot \text{diam}(T) \leq \frac{r}{m} \cdot \text{diam}(T) \leq \delta$$

where the second inequality follows from (7.6) and the last one by choosing m large enough. Third, all quantized weights $q(w_i)$ belong to $\frac{1}{m}\mathbb{Z}$ by definition. Thus, $\kappa = 1 - \sum_{i=1}^r q(w_i)$ is also in $\frac{1}{m}\mathbb{Z}$. Therefore, all weights w'_i are in $\frac{1}{m}\mathbb{Z}$, too. Hence, $w'_i = m_i/m$ for some nonnegative integers m_i . In other words,

$$\nu' = \frac{1}{m} \sum_{i=1}^r m_i \delta_{Y_i}.$$

Since ν' is a probability measure, we must have $\sum_{i=1}^r m_i = m$. Redefine the sequence Y_1, \dots, Y_m by repeating each element Y_i of the sequence Y_1, \dots, Y_r exactly m_i times. Thus $\nu' = \frac{1}{m} \sum_{i=1}^m \delta_{Y_i}$, as required. \square

Corollary 7.3 allows us to transform any true data $X = (X_1, \dots, X_n)$ into a private synthetic data $Y = (Y_1, \dots, Y_m)$. To do this, feed the algorithm \mathcal{A} with the empirical measure on the true data $\mu_X = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$. Recall from Lemma 4.1 that if the algorithm \mathcal{A} is α -metrically private for $\alpha = \varepsilon n$, then the algorithm $X \mapsto Y = \mathcal{A}(\mu_X)$ yields ε -differential private synthetic data. Let us record this observation:

Corollary 7.4 (Differentially private synthetic data). *Let (T, ρ) be a compact metric space. Let $\varepsilon, \delta > 0$. There exists a randomized algorithm \mathcal{A} that takes true data $X = (X_1, \dots, X_n) \in T^n$ as an input and returns synthetic data $Y = (Y_1, \dots, Y_m) \in T^m$ for some m as an output, and with the following two properties.*

- (i) (Privacy): the algorithm \mathcal{A} is ε -differentially private.
- (ii) (Accuracy): for any true data X , the expected accuracy of the synthetic data Y is

$$\mathbb{E} W_1(\mu_Y, \mu_X) \leq 3\delta + \frac{C}{\varepsilon n} \log^{\frac{3}{2}}(N(T, \rho, \delta)) \int_{\delta}^{\text{diam}(T)} N(T, \rho, x) dx,$$

where μ_X and μ_Y denote the corresponding empirical measures.

An interested reader may now skip to Section 9.1 where we illustrate Corollary 7.4 for a specific example of the metric space, namely the d -dimensional cube $T = [0, 1]^d$.

Remark 7.5 (A computationally effective algorithm). We will present a detailed discussion of the algorithmic aspects of the proposed synthetic data generation method in a forthcoming paper. Here, we only mention that our algorithm works in polynomial time⁴ with respect to the cardinality of the dataset. To be more precise, assuming that the input measure μ is given by an oracle for any set A , the oracle gives us $\mu(A)$ and we need a polynomial number of calls to such an oracle.

⁴under the stipulation that an ε -net (of polynomial cardinality) can be constructed in polynomial time

8. A LOWER BOUND

This section is devoted to impossibility results, which yield *lower* bounds on the accuracy of any private measure on a general metric space (T, ρ) . While there may be a gap between our upper and lower bounds for general metric spaces, we will see in Section 9 that this gap vanishes asymptotically for spaces of Minkowski dimension d .

The proof of the lower bound uses the geometric method pioneered by Hardt and Talwar [25]. A lower bound is more convenient to express in terms of packing rather than covering numbers. Recall that the *packing number* $N_{\text{pack}}(T, \rho, \varepsilon)$ of a compact metric space (T, ρ) is defined as the largest cardinality of an ε -separated subset of T . The covering and packing numbers are equivalent up to a factor of 2:

$$N_{\text{pack}}(T, \rho, 2\varepsilon) \leq N(T, \rho, \varepsilon) \leq N_{\text{pack}}(T, \rho, \varepsilon), \quad (8.1)$$

see [48, Lemma 4.2.8]. Thus, in all results of this section, packing numbers can be replaced by covering numbers at the cost of changing absolute constants.

8.1. A master lower bound. We first prove a general result that establishes limitations of metric privacy. To understand this statement better, it may be helpful to assume that $\mathcal{M}_0 = \mathcal{M}_1$ and $\rho_0 = \rho_1$ in the first reading.

Proposition 8.1 (A master lower bound). *Let $\mathcal{M}_0 \subset \mathcal{M}_1$ be two subsets, and let ρ_i be a metric on \mathcal{M}_i , $i = 0, 1$. Assume that for some $t, \alpha > 0$ we have*

$$\text{diam}(\mathcal{M}_0, \rho_0) \leq 1 \quad \text{and} \quad N_{\text{pack}}(\mathcal{M}_0, \rho_1, t) > 2e^\alpha.$$

Then, for any randomized algorithm $\mathcal{A} : \mathcal{M}_0 \rightarrow \mathcal{M}_1$ that is α -metrically private with respect to the metric ρ_0 , there exists $x \in \mathcal{M}_0$ such that

$$\mathbb{E} \rho_1(\mathcal{A}(x), x) > t/4.$$

Proof. For contradiction, assume that

$$\mathbb{E} \rho_1(\mathcal{A}(x), x) \leq t/4, \quad (8.2)$$

for all $x \in \mathcal{M}_0$. Let \mathcal{N} be a t -separated subset of the metric space (\mathcal{M}_0, ρ_1) with cardinality

$$|\mathcal{N}| > 2e^\alpha. \quad (8.3)$$

The separation condition implies that the balls $B(y, \rho_1, t/2)$ centered at the points $y \in \mathcal{N}$ and with radii $t/2$ are all disjoint.

Fix any reference point $y \in \mathcal{M}_0$. The disjointness of the balls yields

$$\sum_{x \in \mathcal{N}} \mathbb{P} \{ \mathcal{A}(y) \in B(x, \rho_1, t/2) \} \leq 1. \quad (8.4)$$

On the other hand, by the definition of α -metric privacy, for each $x \in \mathcal{N}$ we have:

$$\mathbb{P} \{ \mathcal{A}(y) \in B(x, \rho_1, t/2) \} \geq \exp[-\alpha \rho_0(x, y)] \cdot \mathbb{P} \{ \mathcal{A}(x) \in B(x, \rho_1, t/2) \}.$$

The diameter assumption yields $\rho_0(x, y) \leq 1$. Furthermore, using the assumption (8.2) and Markov's inequality, we obtain

$$\mathbb{P} \{ \mathcal{A}(x) \in B(x, \rho_1, t/2) \} = \mathbb{P} \{ \rho_1(\mathcal{A}(x), x) \leq t/2 \} \geq \frac{1}{2}.$$

Combining the two bounds gives

$$\mathbb{P} \{ \mathcal{A}(y) \in B(x, \rho_1, t/2) \} \geq \frac{1}{2e^\alpha}.$$

Substitute this into (8.4) to get

$$\sum_{x \in \mathcal{N}} \frac{1}{2e^\alpha} \leq 1.$$

In other words, we conclude that $|\mathcal{N}| \leq 2e^\alpha$, which contradicts (8.3). The proof is complete. \square

8.2. Metric entropy of the space of probability measures. For a given compact metric space (T, ρ) , we denote by $\mathcal{M}(T)$ the collection of all Borel probability measures on T . We are going to apply Proposition 8.1 for $\mathcal{M}_0 = \mathcal{M}_1 = \mathcal{M}(T)$, for $\rho_1 =$ Wasserstein metric and $\rho_0 =$ TV metric. That proposition requires a lower bound on the packing number $N_{\text{pack}}(\mathcal{M}(T), W_1, t/3)$. In the next lemma, we relate this packing number to that of (T, ρ) . Essentially, it says that if T is large, then there are a lot of probability measures on T .

Proposition 8.2 (Metric entropy of the space of probability measures). *For any compact metric space (T, ρ) and every $t > 0$, we have*

$$N_{\text{pack}}(\mathcal{M}(T), W_1, t/3) \geq \exp(cN_{\text{pack}}(T, \rho, t)),$$

where $c > 0$ is a universal constant.

The proof will use the following lemma.

Lemma 8.3 (A lower bound on the Wasserstein distance). *Let (T, ρ) be a t -separated⁵ compact metric space. Then, for any pair of probability measures μ, ν on T , we have*

$$W_1(\mu, \nu) \geq \mu(B^c) t \quad \text{where } B = \text{supp}(\nu).$$

Proof. Suppose that γ is a coupling of μ and ν . Since ν is supported on B , we have $\gamma(B^c \times B^c) \leq \gamma(T \times B^c) = \nu(B^c) = 0$, which means that $\gamma(B^c \times B^c) = 0$. Therefore

$$\gamma(B^c \times B) = \gamma(B^c \times T) - \gamma(B^c \times B^c) = \mu(B^c).$$

Since the sets B^c and B are disjoint, the separation assumption implies that $\rho(x, y) > t$ for all pairs $x \in B^c$ and $y \in B$. Thus,

$$\int_{T \times T} \rho(x, y) d\gamma(x, y) \geq \int_{B^c \times B} \rho(x, y) d\gamma(x, y) \geq t\gamma(B^c \times B) = t\mu(B^c).$$

Since this holds for all coupling γ of μ and ν , the result follows. \square

Lemma 8.4 (Many different measures). *Let (\mathcal{N}, ρ) be a t -separated compact metric space, and assume that $|\mathcal{N}| \geq 2n$ for some $n \in \mathbb{N}$. Then there exists a family of at least $\exp(cn)$ empirical measures on n points of T that are pairwise $t/3$ -separated in the Wasserstein distance, where $c > 0$ is a universal constant.*

Proof. Let $\mu = n^{-1} \sum_{i=1}^n \delta_{X_i}$ and $\nu = n^{-1} \sum_{i=1}^n \delta_{Y_i}$ be two independent random empirical measures on T . Let us condition on ν and denote $B = \text{supp}(\nu)$. Then

$$\mu(B^c) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \in B^c\}}.$$

Now, $\mathbf{1}_{\{X_i \in B^c\}}$ are i.i.d. Bernoulli random variables that take value 1 with probability

$$\mathbb{P}\{X_i \in B^c\} = \frac{|B^c|}{|\mathcal{N}|} \geq \frac{1}{2},$$

⁵This means that the distance between any two distinct points in T is larger than t .

since by construction we have $|B| \leq n$ and by assumption $|\mathcal{N}| \geq 2n$. Then, applying Chernoff inequality (see [48, Exercise 2.3.2]), we conclude that $\mu(B^c) > 1/3$ with probability bigger than $1 - e^{-5cn}$, where $c > 0$ is a universal constant. Lemma 8.3 yields that $W_1(\mu, \nu) > t/3$.

Now consider a sequence μ_1, \dots, μ_K of independent random empirical measures on T . Using the result above and taking a union bound we conclude that, with probability at least $1 - \binom{K}{2}e^{-5cn}$, the inequality $W_1(\mu_i, \mu_j) > t/3$ holds for all pairs of distinct indices $i, j \in \{1, \dots, K\}$. Choosing $K = \lceil e^{cn} \rceil$ makes K between e^{cn} (as claimed) and e^{2cn} . Thus, the success probability is more than $1 - (e^{2cn})^2e^{-5cn}$, which is positive. The existence of the required family of measures follows. \square

Proof of Proposition 8.2. Let $\mathcal{N} \subset T$ be a t -separated subset of cardinality $|\mathcal{N}| = N_{\text{pack}}(T, \rho, t)$. Lemma 8.4 implies the existence of a set of at least $\exp(c|\mathcal{N}|)$ probability measures on T that is $(t/3)$ -separated in the Wasserstein distance. In other words, we have $N_{\text{pack}}(\mathcal{M}(T), W_1, t/3) \geq \exp(c|\mathcal{N}|)$. Proposition 8.2 is proved. \square

8.3. Lower bounds for private measures and synthetic data. Now we are ready to prove the two main lower bounds on the accuracy for (a) metrically private measures and (b) differential private data.

Theorem 8.5 (Private measure: a lower bound). *Let (T, ρ) be a compact metric space. Assume that for some $t > 0$ and $\alpha \geq 1$ we have*

$$N_{\text{pack}}(T, \rho, t) > C\alpha.$$

Then, for any randomized algorithm \mathcal{A} that takes a probability measure μ on T as an input and returns a probability measure ν on T as an output and that is α -metrically private with respect to the TV metric, there exists μ such that

$$\mathbb{E} W_1(\nu, \mu) > t/12.$$

Proof. The assumption on the packing number for a sufficiently large constant C and Proposition 8.2 yield

$$N_{\text{pack}}(\mathcal{M}(T), W_1, t/3) \geq e^{2\alpha} > 2e^\alpha.$$

Next, apply Proposition 8.1 with $t/3$ instead of t , and for $\mathcal{M}_0 = \mathcal{M}_1 = \mathcal{M}(T)$, setting ρ_1 and ρ_0 to be the Wasserstein and the TV metrics, respectively. The required conclusion follows. \square

Theorem 8.6 (Synthetic data: a lower bound). *There exists an absolute constant n_0 such that the following holds. Let (T, ρ) be a compact metric space. Assume that for some $t > 0$ and some integer $n > n_0$ we have*

$$N_{\text{pack}}(T, \rho, t) > 2n.$$

Then, for any c -differentially private randomized algorithm \mathcal{A} that takes true data $X = (X_1, \dots, X_n) \in T^n$ as an input and returns synthetic data $Y = (Y_1, \dots, Y_m) \in T^m$ for some m as an output, there exists input data X such that

$$\mathbb{E} W_1(\mu_Y, \mu_X) > t/12,$$

where μ_X and μ_Y denote the empirical measures on X and Y .

Proof. First note that a version of Proposition 8.2 holds for empirical measures. Namely, denote the set of all empirical measures on n points of T by $\mathcal{M}_n(T)$. If $N_{\text{pack}}(T, \rho, t) > 2n$ then we claim that

$$N_{\text{pack}}(\mathcal{M}_n(T), W_1, t/3) > 2e^{c_1 n}. \quad (8.5)$$

To see this, let $\mathcal{N} \subset T$ be a t -separated subset of cardinality $|\mathcal{N}| > 2n$. Lemma 8.4 implies the existence of a set of at least $e^{cn} \geq 2e^{c_1 n}$ members of $\mathcal{M}_n(T)$ that is $(t/3)$ -separated in the Wasserstein distance. The claim (8.5) follows.

In preparation to apply Proposition 8.1, consider the sets $\mathcal{M}_0 := T^n$ and $\mathcal{M}_1 := \bigcup_{k=1}^{\infty} T^k$. Consider the normalized Hamming metric

$$\rho_0(X, X') = \frac{1}{n} |\{i \in [n] : X_i \neq X'_i\}|$$

on \mathcal{M}_0 , and the Wasserstein metric

$$\rho_1(X, X') = W_1(\mu_X, \mu_{X'})$$

on \mathcal{M}_1 . Then we clearly have $\text{diam}(\mathcal{M}_0, \rho_0) \leq 1$, and (8.5) is equivalent to $N_{\text{pack}}(\mathcal{M}_0, \rho_1, t/3) > 2e^{c_1 n}$.

If $\mathcal{A} : \mathcal{M}_0 \rightarrow \mathcal{M}_1$ is a c -differentially private algorithm, then \mathcal{A} is (cn) -metrically private in the metric ρ_0 due to Lemma 2.3. Applying Proposition 8.1 with $t/3$ instead of t and $\alpha = c_1 n$, we obtain the required conclusion. \square

9. EXAMPLES AND ASYMPTOTICS

9.1. A private measure on the unit cube. Let us work out the bound of Theorem 7.2 for a concrete example: the d -dimensional unit cube equipped with the ℓ^∞ metric, i.e. $(T, \rho) = ([0, 1]^d, \|\cdot\|_\infty)$. The covering numbers satisfy

$$N(T, \|\cdot\|_\infty, x) \leq (1/x)^d, \quad x > 0,$$

since the set $x\mathbb{Z}^d \cap [0, 1]^d$ forms an x -net of T . Thus the accuracy is

$$\mathbb{E} W_1(\nu, \mu) \lesssim \delta + \frac{\log^{\frac{3}{2}}(1/\delta)}{\alpha} \int_{\delta}^1 (1/x)^d dx \lesssim \delta + \frac{\log^{\frac{3}{2}}(1/\delta)}{\alpha} \cdot (1/\delta)^{d-1}$$

if $d \geq 2$. Optimizing in δ yields

$$\mathbb{E} W_1(\nu, \mu) \lesssim \left(\frac{\log^{\frac{3}{2}} \alpha}{\alpha} \right)^{1/d},$$

which wonderfully extends Theorem 5.4 for $d = 1$. Combining the two results, for $d = 1$ and $d = 2$, we obtain the following general result:

Corollary 9.1 (Private measure on the cube). *Let $d \in \mathbb{N}$ and $\alpha \geq 2$. There exists a randomized algorithm \mathcal{A} that takes a probability measure μ on $[0, 1]^d$ as an input and returns a finitely-supported probability measure ν on $[0, 1]^d$ as an output, and with the following two properties.*

- (i) (Privacy): *the algorithm \mathcal{A} is α -metrically private in the TV metric.*
- (ii) (Accuracy): *for any input measure μ , the expected accuracy of the output measure ν in the Wasserstein distance is*

$$\mathbb{E} W_1(\nu, \mu) \leq C \left(\frac{\log^{\frac{3}{2}} \alpha}{\alpha} \right)^{1/d}.$$

Similarly, by invoking Corollary 7.4, we obtain ε -differential privacy for synthetic data:

Corollary 9.2 (Private synthetic data in the cube). *Let $d, n \in \mathbb{N}$ and $\varepsilon > 0$. There exists a randomized algorithm \mathcal{A} that takes true data $X = (X_1, \dots, X_n) \in ([0, 1]^d)^n$ as an input and returns synthetic data $Y = (Y_1, \dots, Y_m) \in ([0, 1]^d)^m$ for some m as an output, and with the following two properties.*

(i) (*Privacy*): the algorithm \mathcal{A} is ε -differentially private.
(ii) (*Accuracy*): for any true data X , the expected accuracy of the synthetic data Y is

$$\mathbb{E} W_1(\mu_Y, \mu_X) \leq C \left(\frac{\log^{\frac{3}{2}}(\varepsilon n)}{\varepsilon n} \right)^{1/d},$$

where μ_X and μ_Y denote the corresponding empirical measures.

The two results above are nearly sharp. Indeed, let us work out the lower bound for the cube, using Theorem 8.5. The covering numbers satisfy

$$N_{\text{pack}}(T, \|\cdot\|_\infty, x) \geq (c/x)^d, \quad x > 0,$$

which again can be seen by considering a rescaled integer grid. Setting $t = c/(2C\alpha)^{1/d}$ we get $N(T, \|\cdot\|_\infty, t) > C\alpha$. Hence

$$\mathbb{E} W_1(\nu, \mu) > t/12 \gtrsim (1/\alpha)^{1/d},$$

which matches the upper bound in Corollary 9.1 up to a logarithmic factor. Let us record this result.

Corollary 9.3 (Private measure on the cube: a lower bound). *Let $d \in \mathbb{N}$ and $\alpha \geq 2$. Then, for any randomized algorithm \mathcal{A} that takes a probability measure μ on $[0, 1]^d$ as an input and returns a probability measure ν on $[0, 1]^d$ as an output, and that is α -metrically private with respect to the TV metric, there exists μ such that*

$$\mathbb{E} W_1(\nu, \mu) > c \left(\frac{1}{\alpha} \right)^{1/d}.$$

In a similar way, by invoking the lower bound in Theorem 8.6, we obtain the following nearly matching lower bound for Corollary 9.2:

Corollary 9.4 (Private synthetic data in the cube: a lower bound). *Let $d, n \in \mathbb{N}$. Then, for any c -differentially private randomized algorithm \mathcal{A} that takes true data $X = (X_1, \dots, X_n) \in ([0, 1]^d)^n$ as an input and returns synthetic data $Y = (Y_1, \dots, Y_m) \in ([0, 1]^d)^m$ for some m as an output, there exists input data X such that*

$$\mathbb{E} W_1(\nu_Y, \mu_X) > c \left(\frac{1}{n} \right)^{1/d}.$$

where μ_X and μ_Y denotes the empirical measures on X and Y .

9.2. Asymptotic result. The only property of the cube $T = [0, 1]^d$ we used in the previous section is the behavior on its covering numbers,⁶ namely that

$$N(T, \rho, x) \asymp (1/x)^{-d}, \quad x > 0. \quad (9.1)$$

Therefore, the same results on private measures and synthetic data hold for any compact metric space (T, ρ) whose covering numbers behave this way. In particular, it follows that any probability measure μ on T can be transformed into a α -metrically private measure ν on T , with accuracy

$$\mathbb{E} W(\nu, \mu) \asymp (1/\alpha)^{1/d}. \quad (9.2)$$

(ignoring logarithmic factors), and this result is nearly sharp. Similarly, any true data $X \in T^n$ can be transformed into ε -differentially private synthetic data $Y \in T^m$ for some m , with accuracy

$$\mathbb{E} W(\mu_Y, \mu_X) \asymp (1/n)^{1/d}. \quad (9.3)$$

⁶The lower bound used packing numbers, but they are equivalent to covering numbers due to (8.1).

(ignoring logarithmic factors and dependence on ε), and this result is nearly sharp.

These intuitive observations can be formalized using the notion of *Minkowski dimension*. By definition, the metric space (T, ρ) has Minkowski dimension d if

$$\lim_{x \rightarrow 0} \frac{\log N(T, \rho, x)}{\log(1/x)} = d.$$

The following two asymptotic results combine upper and lower bounds, and essentially show that (9.2) and (9.3) hold in any space of dimension d .

Theorem 9.5 (Private measure, asymptotically). *Let (T, ρ) be a compact metric space of Minkowski dimension $d \geq 1$. Then*

$$\liminf_{\alpha \rightarrow \infty} \sup_{\mathcal{A}} \frac{\log(\mathbb{E} W_1(\mathcal{A}(\mu), \mu))}{\log \alpha} = -\frac{1}{d}.$$

Here the infimum is over randomized algorithms \mathcal{A} than input and output a probability measure on T and are α -metrically private with respect to the TV metric; the supremum is over all probability measures μ on T .

Proof. We deduce the upper bound from Theorem 7.2 and the lower bound from Theorem 8.5.

Upper bound. By rescaling, we can assume without loss of generality that $\text{diam}(T, \rho) = 1$. Fix any $\varepsilon > 0$. By definition of Minkowski dimension, there exists $\delta_0 > 0$ such that

$$N(T, \rho, x) \leq (1/x)^{d+\varepsilon} \quad \text{for all } x \in (0, \delta_0). \quad (9.4)$$

Then

$$\int_{\delta}^1 N(T, \rho, x) dx \leq \int_{\delta}^{\delta_0} (1/x)^{d+\varepsilon} dx + \int_{\delta_0}^1 N(T, \rho, x) dx \leq K(1/\delta)^{d+\varepsilon-1} + I(\delta_0)$$

where $K = 1/(d + \varepsilon - 1)$ and $I(\delta_0) = \int_{\delta_0}^1 N(T, \rho, x) dx$. The last step follows if we replace δ_0 by infinity and compute the integral.

If we let $\delta \downarrow 0$, we see that $K(1/\delta)^{d+\varepsilon-1} \rightarrow \infty$ while $I(\delta_0)$ stays the same since it does not depend on δ . Therefore, there exists $\delta_1 > 0$ such that $I(\delta_0) \leq K(1/\delta)^{d+\varepsilon-1}$ for all $\delta \in (0, \delta_1)$. Therefore,

$$\int_{\delta}^1 N(T, \rho, x) dx \leq 2K(1/\delta)^{d+\varepsilon-1} \quad \text{for all } \delta \in (0, \min(\delta_0, \delta_1)).$$

Applying Theorem 7.2 for such δ and using (9.4), we get

$$\inf_{\mathcal{A}} \sup_{\mu} \mathbb{E} W_1(\nu, \mu) \leq 2\delta + \frac{C}{\alpha} \log^{\frac{3}{2}} \left((1/\delta)^{d+\varepsilon} \right) \cdot 2K(1/\delta)^{d+\varepsilon-1}. \quad (9.5)$$

Optimizing in δ , we find that a good choice is

$$\delta = \delta(\alpha) = \left(\frac{\log^{\frac{3}{2}}(K\alpha)}{K\alpha} \right)^{\frac{1}{d+\varepsilon}}.$$

For any sufficiently large α , we have $\delta < \min(\delta_0, \delta_1)$ as required, and substituting $\delta = \delta(\alpha)$ into the bound in (9.5) we get after simplification:

$$\inf_{\mathcal{A}} \sup_{\mu} \mathbb{E} W_1(\nu, \mu) \leq (1 + 2CK)\delta(\alpha).$$

Furthermore, recalling that K does not depend on α , it is clear that

$$\lim_{\alpha \rightarrow \infty} \frac{\log((1 + 2CK)\delta(\alpha))}{\log \alpha} = -\frac{1}{d + \varepsilon}.$$

Thus

$$\limsup_{\alpha \rightarrow \infty} \frac{\log(\inf_{\mathcal{A}} \sup_{\mu} \mathbb{E} W_1(\nu, \mu))}{\log \alpha} \leq -\frac{1}{d + \varepsilon}.$$

Since $\varepsilon > 0$ is arbitrary, it follows that

$$\limsup_{\alpha \rightarrow \infty} \inf_{\mathcal{A}} \sup_{\mu} \frac{\log(\mathbb{E} W_1(\mathcal{A}(\mu), \mu))}{\log \alpha} \leq -\frac{1}{d}. \quad (9.6)$$

Lower bound. Fix any $\varepsilon > 0$. By definition of Minkowski dimension and the equivalence (8.1), there exists $\delta_0 > 0$ such that

$$N_{\text{pack}}(T, \rho, x) \geq N(T, \rho, x) > (1/x)^{d-\varepsilon} \quad \text{for all } x \in (0, \delta_0).$$

Set

$$x(\alpha) = \left(\frac{1}{C\alpha} \right)^{\frac{1}{d-\varepsilon}}.$$

Then, for any sufficiently large α , we have $x \in (0, \delta_0)$ and

$$N_{\text{pack}}(T, \rho, x(\alpha)) > C\alpha.$$

Applying Theorem 8.5, we get

$$\inf_{\mathcal{A}} \sup_{\mu} \mathbb{E} W_1(\nu, \mu) \geq x(\alpha)/20.$$

It is easy to check that

$$\lim_{\alpha \rightarrow \infty} \frac{\log(x(\alpha)/20)}{\log \alpha} = -\frac{1}{d - \varepsilon}.$$

Thus

$$\liminf_{\alpha \rightarrow \infty} \frac{\log(\inf_{\mathcal{A}} \sup_{\mu} \mathbb{E} W_1(\nu, \mu))}{\log \alpha} \geq -\frac{1}{d - \varepsilon}.$$

Since $\varepsilon > 0$ is arbitrary, it follows that

$$\liminf_{\alpha \rightarrow \infty} \inf_{\mathcal{A}} \sup_{\mu} \frac{\log(\mathbb{E} W_1(\mathcal{A}(\mu), \mu))}{\log \alpha} \geq -\frac{1}{d}.$$

Combining with the upper bound (9.6), we complete the proof. \square

In a similar way, we can deduce the following asymptotic result for private synthetic data. The argument is analogous; the upper bound follows from Corollary 7.4 and the lower bound from Theorem 8.6.

Theorem 9.6. *Let (T, ρ) be a compact metric space of Minkowski dimension $d \geq 1$. Then, for every $\varepsilon \in (0, c)$, we have*

$$\liminf_{n \rightarrow \infty} \inf_{\mathcal{A}} \sup_{\mu} \frac{\log(\mathbb{E} W_1(\mathcal{A}(\mu), \mu))}{\log n} = -\frac{1}{d}.$$

Here the infimum is over ε -differentially private randomized algorithms \mathcal{A} that take true data $X = (X_1, \dots, X_n) \in ([0, 1]^d)^n$ as an input and return synthetic data $Y = (Y_1, \dots, Y_m) \in ([0, 1]^d)^m$ for some m as an output; the supremum is over the input data X .

ACKNOWLEDGEMENT

M.B. acknowledges support from NSF DMS-2140592. T.S. acknowledges support from NSF DMS-2027248, NSF CCF-1934568 and a CeDAR Seed grant. R.V. acknowledges support from NSF DMS-1954233, NSF DMS-2027299, U.S. Army 76649-CS, and NSF+Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning.

REFERENCES

- [1] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 510–526. Springer, 2018.
- [2] John Abowd, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. *US Census Bureau*, 2019.
- [3] John M Abowd. The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.
- [4] John M Abowd and Simon D Woodcock. Disclosure limitation in longitudinal linked data. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 215277, 2001.
- [5] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geoindistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914, 2013.
- [6] David L Applegate, Robert E Bixby, and William J Vašek Chvátal. Cook. the traveling salesman problem: A computational study. *Princeton Series in Applied Mathematics*, 2007.
- [7] UN General Assembly et al. Universal declaration of human rights. *UN General Assembly*, 302(2):14–25, 1948.
- [8] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 273–282, 2007.
- [9] Steven M Bellovin, Preetam K Dutta, and Nathan Reitinger. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22:1, 2019.
- [10] Louis Béthune, Alberto González-Sanz, Franck Mamalet, and Mathieu Serrurier. The many faces of 1-Lipschitz neural networks. *arXiv preprint arXiv:2104.05097*, 2021.
- [11] Rabindra Nath Bhattacharya and Edward C Waymire. *A basic course in probability theory*, volume 69. Springer, 2007.
- [12] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):1–25, 2013.
- [13] March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Covariance’s Loss is Privacy’s Gain: Computationally Efficient, Private and Accurate Synthetic Data. *arXiv preprint arXiv:2107.05824*, 2021.
- [14] March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Private sampling: a noiseless approach for generating differentially private synthetic data. *SIAM Journal on Mathematics of Data Science*, to appear, 2021.
- [15] Jim Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13(4):321–327, 2003.
- [16] N Christofides. Worst-case analysis of a new heuristic for the traveling salesman problem. In *Proc. Symposium on New Directions and Recent Results in Algorithms and Complexity*, 1976.
- [17] Jessamyn Dahmen and Diane Cook. Synsys: A synthetic data generation system for healthcare applications. *Sensors*, 19(5):1181, 2019.
- [18] Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35, 2021.
- [19] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.
- [20] N. Dunford and J.T. Schwartz. *Linear Operators, Part I: General Theory*. Wiley, 1958.
- [21] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, 9(2), 2019.

- [22] Cynthia Dwork, Aleksandar Nikolov, and Kunal Talwar. Efficient algorithms for privately releasing marginals via convex relaxations. *Discrete & Computational Geometry*, 53(3):650–673, 2015.
- [23] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.
- [24] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *NIPS’12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, 2012.
- [25] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM symposium on Theory of computing, STOC ’10*, pages 705–714, New York, NY, USA, 2010.
- [26] Mathew E Hauer and Alexis R Santos-Lozada. Differential privacy in the 2020 census will distort covid-19 rates. *Socius*, 7:2378023121994014, 2021.
- [27] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially-private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1–2):1021–1032, 2010.
- [28] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.
- [29] Anna R Karlin, Nathan Klein, and Shayan Oveis Gharan. A (slightly) improved approximation algorithm for metric TSP. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 32–45, 2021.
- [30] Marek Karpinski, Michael Lampis, and Richard Schmied. New inapproximability bounds for TSP. *Journal of Computer and System Sciences*, 81(8):1665–1677, 2015.
- [31] Michael Kearns and Aaron Roth. How much still needs to be done to make algorithms more ethical. *URL: <https://www.shine.cn/opinion/2008214615/>*, 2020.
- [32] Leonid V Kovalev. Lipschitz clustering in metric spaces. *arXiv preprint arXiv:2108.03535*, 2021.
- [33] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.
- [34] Ninghui Li, Min Lyu, Dong Su, and Weining Yang. Differential privacy: From theory to practice. *Synthesis Lectures on Information Security, Privacy, & Trust*, 8(4):1–138, 2016.
- [35] Terrance Liu, Giuseppe Vietri, Thomas Steinke, Jonathan Ullman, and Steven Wu. Leveraging public data for practical private query release. In *International Conference on Machine Learning*, pages 6968–6977. PMLR, July 2021.
- [36] Ofer Mendelovich and Michael D Lesh. Fidelity and privacy of synthetic medical data. *arXiv preprint arXiv:2101.08658*, 2021.
- [37] Xue Meng, Hui Li, and Jiangtao Cui. Different strategies for differentially private histogram publication. *Journal of Communications and Information Networks*, 2(3):68–77, 2017.
- [38] Boel Nelson and Jenni Reuben. Chasing accuracy and privacy, and catching both: A literature survey on differentially private histogram publication. *arXiv preprint arXiv:1910.14028*, page 83, 2019.
- [39] Robert Clay Prim. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401, 1957.
- [40] Wahbeh Qardaji, Weining Yang, and Ninghui Li. Understanding hierarchical methods for differentially private histograms. *Proceedings of the VLDB Endowment*, 6(14):1954–1965, 2013.
- [41] Sofya Raskhodnikova, Satchit Sivakumar, Adam Smith, and Marika Swanberg. Differentially private sampling from distributions. *Advances in Neural Information Processing Systems*, 34, 2021.
- [42] Zbigniew Semadeni. *Schauder bases in Banach spaces of continuous functions*, volume 918. Springer, 2006.
- [43] AI Serdyukov. O nekotorykh ekstremal'nykh obkhodakh v grafakh. *Upravlyayemye sistemy*, 17:76–79, 1978.
- [44] Justin Thaler, Jonathan Ullman, and Salil Vadhan. Faster algorithms for privately releasing marginals. In *International Colloquium on Automata, Languages, and Programming*, pages 810–821. Springer, 2012.
- [45] Jonathan Ullman and Salil Vadhan. PCPs and the hardness of generating private synthetic data. In *Theory of Cryptography Conference*, pages 400–416. Springer, 2011.
- [46] Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.
- [47] SS Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.
- [48] Roman Vershynin. *High-dimensional probability. An introduction with applications in data science*. Cambridge University Press, 2018.

- [49] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.
- [50] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.
- [51] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *J. Mach. Learn. Res.*, 5(Jun):669–695, 2004.
- [52] Gus Wezerek and David Van Riper. Changes to the Census could make small towns disappear. *New York Times*, Feb. 6, 2020.
- [53] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on knowledge and data engineering*, 23(8):1200–1214, 2010.
- [54] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *The VLDB journal*, 22(6):797–822, 2013.
- [55] Lin Yang, Wentao Fan, and Nizar Bouguila. Clustering analysis via deep generative models with mixture models. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [56] Jun Zhang, Xiaokui Xiao, and Xing Xie. Privtree: A differentially private algorithm for hierarchical decompositions. In *Proceedings of the 2016 International Conference on Management of Data*, pages 155–170, 2016.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA IRVINE
Email address: `marchb@uci.edu`

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, DAVIS AND CENTER OF DATA SCIENCE
 AND ARTIFICIAL INTELLIGENCE RESEARCH, UC DAVIS
Email address: `strohmer@math.ucdavis.edu`

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA IRVINE
Email address: `rvershyn@uci.edu`