Training Generative Models From Privatized Data via Entropic Optimal Transport

Daria Reshetova, Wei-Ning Chen, *Graduate Student Member, IEEE*, and Ayfer Özgür, *Senior Member, IEEE*

***Abstract*—Local differential privacy is a powerful method for privacy-preserving data collection. In this paper, we develop a framework for training Generative Adversarial Networks (GANs) on differentially privatized data. We show that entropic regularization of optimal transport – a popular regularization method in the literature that has often been leveraged for its computational benefits – enables the generator to learn the raw (unprivatized) data distribution even though it only has access to privatized samples. We prove that at the same time this leads to fast statistical convergence at the parametric rate. This shows that entropic regularization of optimal transport uniquely enables the mitigation of both the effects of privatization noise and the curse of dimensionality in statistical convergence. We provide experimental evidence to support the efficacy of our framework in practice.**

***Index Terms*—Privacy, GANs, entropic optimal transport.**

I. Introduction

LOCAL differential privacy (LDP) [1], [2] has emerged as a popular criterion to provide privacy guarantees on individuals' personal data and has been recently deployed by major technology organizations for privacy-preserving data collection from peripheral devices. In this framework, the user data is locally randomized (e.g., by the addition of noise) before it is transferred to the data curator, so the privacy guarantee does not rely on a trusted centralized server. Mathematically provable guarantees on the randomization mechanism ensure that any adversary that gets access to the privatized data will be unable to learn too much about the user's personal information. This directly alleviates many of the systematic privacy and security challenges associated with traditional data collection. Learning from privatized data, however, requires rethinking machine learning methods to extract accurate and useful population-level models from the privatized (noisy) data.

In this paper, we consider the problem of training generative models from locally privatized user data. In recent years, deep-learning-based generative models, known as Generative Adversarial Networks (GANs), have become a popular framework for learning data distributions and sampling, and have achieved impressive results in various domains [3], [4], [5]. GANs aim to learn a mapping $G(\cdot)$, called the generator, which comes from a set of functions $\mathcal{G} \subseteq \{G : \mathcal{Z} \to \mathcal{X}\}$ usually modeled as a neural network, and maps a latent random variable $Z \in \mathcal{Z}$ with some known simple distribution to a random variable $G(Z) \in \mathcal{X}$, with distribution $P_{G(Z)}$ that is close to the target probability measure P_X. For example, by using the popular p-Wasserstein distance as a discrepancy measure between the generated and target distributions the GAN optimization problem becomes,

$$\min_{G\in\mathcal{G}} W_p^p\big(P_{G(Z)}, P_X\big). \tag{1}$$

In practice, the target distribution P_X is represented by its samples $\{X_i\}_{i=1}^n \sim P_X^{\otimes n}$ and the optimization problem is solved by replacing P_X in (1) with the empirical distribution Q_X^n of the samples, i.e.,

$$\min_{G\in\mathcal{G}} W_p^p\big(P_{G(Z)}, Q_X^n\big). \tag{2}$$

How can we use the GAN framework above to learn a generative model for P_X when we have only access to samples $\{Y_i = M(X_i)\}_{i=1}^n$ privatized by an LDP mechanism $M : \mathcal{X} \to \mathcal{Y}$? For example, Y_i can represent a privatized image obtained from X_i by adding sufficient Gaussian or Laplace noise independently to each pixel. Simply replacing the target distribution P_X in (1) with the empirical distribution Q_Y^n of the privatized samples,

$$\min_{G\in\mathcal{G}} W_p^p\big(P_{G(Z)}, Q_Y^n\big), \tag{3}$$

will result in a generative model for $P_Y = M\#P_X$, the push-forward distribution of P_X through the privatization mechanism M, rather than the original distribution P_X. In other words, we will learn to generate the *privatized* data (e.g., noisy images) instead of learning to generate the original (raw) data.

In this paper, we show that a simple but non-intuitive modification of the objective in (3) – the addition of an entropic regularization term – allows one to provably learn the original distribution P_X from the privatized samples Y_i under de-facto privatization mechanisms such as the Laplace or Gaussian mechanism, i.e., $P_{G_n(Z)} \to P_X$ for the minimizer G_n of the entropically regularized version of (3) provided that the generator class $\mathcal{G}$ is expressive enough to generate P_X. More generally, we show that the original distribution P_X can be recovered under any privatization mechanism M by entropic regularization of optimal transport with a suitably

Manuscript received 29 October 2023; revised 11 February 2024 and 30 March 2024; accepted 6 April 2024. Date of publication 16 April 2024; date of current version 3 May 2024. This work was supported in part by NSF under Award CCF-2213223. *(Corresponding author: Daria Reshetova.)*

The authors are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94205 USA (e-mail: resh@stanford.edu; wnchen@stanford.edu; aozgur@stanford.edu).

Digital Object Identifier 10.1109/JSAIT.2024.3387463

2641-8770 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

chosen cost function given by the negative log-likelihood of the privatization mechanism. Note that the fact that we can learn the population distribution P_X from which the original samples have been generated does not imply that we can learn the original samples X_i, (i.e., somehow "denoise" the observed privatized samples Y_i), and indeed the post-processing property of DP ensures that the DP guarantee on the samples Y_i translates to the learned model G_n as well as any new samples generated from this model.

Entropic regularization for Optimal Transport GANs has been of significant interest in the prior literature, albeit for different reasons. Historically, it has been leveraged primarily for its computational benefits, enabling an efficient approximation of the optimal transport problem [6]. More recently, [7] (also see [8] and [9]) has shown that it facilitates rapid convergence of GANs and circumvents the curse of dimensionality. In particular, without regularization the solution of the empirical problem in (2) converges to the solution of population problem in (1) as $\Omega(n^{-2/d})$, where d is the dimension of the target distribution (P_X), while [7] shows that for $p = 2$ entropic regularization enables convergence at the parametric rate $O(1/\sqrt{n})$. In this paper, we prove similar convergence guarantees for the privatized setting for both $p = 1$ and $p = 2$. In the non-privatized setting of [7] entropic regularization of (1) is needed to facilitate convergence albeit introducing undesirable regularization bias that changes the solution (i.e., the generated distribution does not converge to the target distribution P_X). In the privatized setting, we show that entropic regularization has the unique advantage of both mitigating the effects of privatization noise and facilitating convergence. Therefore, our framework can be potentially useful even in the unprivatized setting as a way of facilitating convergence without biasing the solution. The contributions of our paper are summarized as follows:

- *LDP Framework for Optimal Transport GANs*: We propose a novel framework for training GANs from differentially privatized data based on entropic regularization of Optimal Transport. Previous approaches to privatization in GANs exclusively focus on privatizing the training process, for example, by using DP-SGD methods. In contrast, in our framework, privatization occurs exclusively at the data level and hence it is particularly suitable for user-generated data, e.g., federated learning, where each user can locally privatize its data before sending it to the service provider or data collector. The training of the model from privatized samples is indistinguishable from training a non-privatized GAN (with entropic regularization), which enables the immediate use of existing entropic optimal transport libraries.
- *Sample Complexity Bounds*: We prove convergence guarantees for our LDP framework with entropic optimal transport, including the convergence results for Laplace and Gaussian privatization mechanisms. These results show that entropic regularization uniquely mitigates both the effects of privatization noise and the curse of dimensionality and provides a clear understanding of the trade-offs involved between privacy, accuracy, and the volume of data. In the non-privatization setting, previous convergence results have been limited to the entropic 2-Wasserstein distance setting.
- *Empirical Validation:* We supplement our theoretical contributions with a comprehensive set of experiments designed to validate our claims. These experiments demonstrate the efficacy of our approach in practical scenarios and provide empirical evidence of the superior performance of our method.

A. Connection to Rate-Distortion Theory

In this section, we illustrate how the main idea of our paper is inherently connected to rate-distortion theory. Consider the special case of (3) for $p = 2$ with the empirical distribution Q_Y^n of the privatized samples replaced by their true distribution $P_Y = M\#P_X$, in which case we can explicitly write it as (see also (6):

$$\min_{G\in\mathcal{G}} \inf_{\pi\in\Pi(P_{G(Z)},P_Y)} \mathbb{E}_{(G(Z),Y)\sim\pi}\Big[\|G(Z)-Y\|^2\Big],$$

where $\pi \in \Pi(P_{G(Z)}, P_Y)$ represents the set of all joint distributions on $\mathcal{X}\times\mathcal{Y}$ with marginals $P_{G(Z)}$ and P_Y. The entropic regularization we advocate in this paper transforms this problem to the following problem (see also (8)):

$$\min_{G\in\mathcal{G}} \inf_{\pi\in\Pi(P_{G(Z)},P_Y)} \Big\{\mathbb{E}_{(G(Z),Y)\sim\pi}\Big[\|G(Z)-Y\|^2\Big] + \lambda I_\pi(G(Z),Y)\Big\}, \tag{4}$$

where $I_\pi(G(Z), Y)$ is the mutual information between $G(Z)$ and Y as dictated by the joint distribution π and $\lambda \in \mathbb{R}$ is the regularization parameter that we can choose. Assuming the set of functions $\mathcal{G}$ is rich enough to generate any distribution $P_{G(Z)}$ on $\mathcal{X}$ (we relax this condition and make it more precise in Theorem 1) and relabeling $P_{G(Z)} = P_{\hat{X}}$ for simplicity, we can rewrite (4) as:

$$\inf_{P_{\hat{X}|Y}} \mathbb{E}\Big[\|\hat{X}-Y\|^2\Big] + \lambda I\Big(\hat{X},Y\Big).$$

One can recognize this as the Lagrangian form of the following rate-distortion problem under mean-squared error:

$$\inf_{P_{\hat{X}|Y}:\mathbb{E}\left[\|\hat{X}-Y\|^2\right]\le D} I\Big(\hat{X},Y\Big),$$

where Y can be interpreted as the source variable and $\hat{X}$ as its reconstruction. For general P_Y, there is no explicit characterization of the solution of this problem. Our paper leverages the observation that this problem is easy to solve in one special case: when $Y = X + N$, for arbitrary $X \sim P_X$ and $N \sim \mathcal{N}(0, D)$ independent of X. In this case, the optimal conditional distribution $P_{\hat{X}|Y}$ (or equivalently the test channel $P_{Y|\hat{X}}$) is such that

$$Y = \hat{X} + W, \qquad \hat{X} \sim P_X, \qquad W \sim \mathcal{N}(0, D).$$

This can be observed from the standard characterization of the rate-distortion function for a Gaussian source under mean-squared error; see proof of [10, Th. 10.3.2] or see Theorem 1 where we prove a more general result. Note that this implies that the reconstruction $\hat{X}$ of Y has distribution P_X which

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

corresponds to the unprivatized distribution in our setting. Note that this conclusion holds only if the desired compression rate D matches the distribution of the Gaussian component $N \sim \mathcal{N}(0, D)$ of Y. This corresponds to a specific choice of the regularization parameter λ in our framework in (4).

B. Related Work

Estimation, inference, and learning problems under local differential privacy (LDP) constraints have been of significant interest in the recent literature with emphasis on two canonical tasks: discrete distribution and mean estimation [11], [12], [13], [14], [15], [16], [17]. However, insights from these solutions do not extend to training generative models with high-dimensional data under LDP constraints. The understanding of learning problems under LDP constraints is relatively limited, and even less so in the non-interactive setting when the data is accessed only once as in our setting, in which case training can be exponentially harder as shown in [2], [16].

The exploration of differentially private learning in generative models has primarily been focused on introducing privacy during the training phase, e.g., by adding noise to the gradients during training [18], [19], [20], [21]. However, noisy gradients can amplify inherent instability during GANs' training process [22], [23]. These methods can be applied in a federated learning setting by locally privatizing the gradients at each user and transmitting them to the server at every iteration of the learning algorithm [24]. However, this leads to a large communication overhead. In contrast, there is only one round of communication in our LDP setting; users send their locally privatized data to the server. The training of the GAN from this privatized data is effectively indistinguishable from the non-private case.

Entropic regularization of optimal transport has been initially proposed as a computationally efficient approximation of optimal transport [6], [25]. Subsequently, [8], [26] have shown statistical convergence benefits of entropic regularization when estimating optimal transport from empirical samples. More recently, [7] have shown that these statistical benefits extend to the entropic 2-Wasserstein GAN setting, where both the generated distribution and the target distribution depend on the empirical samples. We extend those results to the privatized setting, showing fast convergence in both the entropic 1- and 2-Wasserstein GAN settings.

II. Background and Problem Setup

To formally state the problem, we first introduce the necessary concepts of privacy.

A. Local Differential Privacy

A local randomizer $\mathcal{A} : \mathcal{X} \to \mathcal{Z}$ acting on the data domain $\mathcal{X}$ satisfies ϵ-LDP for $\epsilon \geq 0$ if for any $S \subseteq \mathcal{Z}$ and for any pair of inputs $x, x' \in \mathcal{X}$, it holds that

$$P(\mathcal{A}(x) \in S) \leq e^{\epsilon} P\big(\mathcal{A}\big(x'\big) \in S\big)$$

LDP ensures that the input to $\mathcal{A}$ cannot be accurately inferred from its output. To achieve LDP, one common approach is via the following Laplace mechanism.

Laplace Mechanism [1]: For any $\epsilon > 0$ and any function $f : \mathcal{X} \to \mathbb{R}^k$ such that $\|f(x) - f(x')\|_1 \leq \Delta$ for any $x, x' \in \mathcal{X}$, the randomized mechanism $\mathcal{A}(x) = f(x) + (s_1, \ldots, s_k)$ with $s_i \sim \text{Laplace}(0, \Delta/\epsilon)$ independent of $s_j, j \neq i$ satisfies ϵ-DP and is called the Laplace Mechanism. We will call ϵ/Δ the noise scale of the mechanism (also called noise multiplier). Oftentimes, in ML applications, the (pure) LDP constraint may be too stringent, so the following relaxation on pure LDP is often adopted.

Approximate Local Differential Privacy: A local randomized algorithm $\mathcal{A} : \mathcal{X} \to \mathcal{Z}$ acting on the data domain $\mathcal{X}$ satisfies (ϵ, δ)-(approximate) LDP for $\epsilon \geq 0, \delta \in (0, 1)$, if for any $S \subseteq \mathcal{Z}$ and for any pair of inputs $x, x' \in \mathcal{X}$, it holds that

$$P(\mathcal{A}(x) \in S) \leq e^{\epsilon} P\big(\mathcal{A}\big(x'\big) \in S\big) + \delta$$

(ϵ, δ)-LDP is very similar to pure LDP, but it allows the privacy requirement to be violated with (small) probability δ. One of the most versatile mechanisms to achieve (ϵ, δ)-DP is the following Gaussian Mechanism.

Gaussian Mechanism [27], [28]: For any $\epsilon > 0$, $\delta \in (0, 0.5)$, and any function $f : \mathcal{X} \to \mathbb{R}^k$ such that $\|f(x) - f(x')\|_2 \leq \Delta$ for any $x, x' \in \mathcal{X}$, the randomized mechanism $\mathcal{A}(x) = f(x) + (s_1, \ldots, s_k)$ with $s_i \sim \mathcal{N}(0, \sigma^2)$ independent of $s_j, j \neq i$ is called the Gaussian Mechanism and satisfies (ϵ, δ)-DP if

$$\sigma > \frac{c + \sqrt{c^2 + \epsilon}}{\epsilon\sqrt{2}} \Delta, \text{ where } c^2 = \ln \frac{2}{\sqrt{16\delta + 1} - 1}. \quad (5)$$

Similar to the Laplace mechanism, we will call σ the noise scale of the Gaussian mechanism.

B. Optimal Transport GANs

Optimal Transport GANs minimize the distance between the target and generated distributions. Contrary to the Jensen-Shannon divergence, which was first introduced as a loss function for generative models, and many other popular distances on probability measures (total variation distance, KL-divergences), optimal transport (OT) is defined through a cost function in the sample space and thus is meaningful for distributions with non-overlapping supports. Moreover, for certain costs, OT defines a distance between distributions and metrizes weak convergence on distributions with finite moments.

Optimal Transport: Let $c : \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$ be a cost function taking non-negative values and $\mathcal{P}(\mathcal{U})$ be the set of all probability measures with support $\mathcal{U} \subseteq \mathbb{R}^d$. Then for $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^d$ and $P_U \in \mathcal{P}(\mathcal{U}), P_V \in \mathcal{P}(\mathcal{V})$, two probability measures on $\mathcal{U}, \mathcal{V}$ respectively, the Optimal Transport between P_U and P_V is

$$OT_c(P_U, P_V) = \inf_{\pi \in \Pi(P_U, P_V)} \mathbb{E}_{(U,V)\sim\pi}[c(U, V)],$$

where $\Pi(P_U, P_V) = \{\pi \in \mathcal{P}(\mathcal{U} \times \mathcal{V})$: $\int_{\mathcal{V}} \pi(u, v)dv = P_U(U), \int_{\mathcal{U}} \pi(u, v)du = P_V(v)\}$ is the set of all couplings of P_U and P_V, i.e., all joint probability measures with marginal distributions P_U and P_V.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

p-Wasserstein distance: When the cost is $c(x, y) = \|x - y\|_p^p$ the optimal transport becomes the p-Wasserstein distance between P_U, P_V (raised to power p):

$$W_p^p(P_U, P_V) = \inf_{\pi \in \Pi(P_U, P_V)} \mathbb{E}_{(U,V)\sim\pi}\Big[\|U - V\|_p^p\Big]. \quad (6)$$

Optimal Transport GAN: The main objective of GANs is to find a mapping $G(\cdot)$, called a generator, that comes from a set of functions $\mathcal{G} \subseteq \{G : \mathcal{Z} \to \mathcal{X}\}$ (usually modeled as a neural network) and maps a latent random variable $Z \in \mathcal{Z}$ with some known distribution to a variable $X \in \mathcal{X}$ with some target probability measure P_X approximated by the empirical distribution Q_X^n of n samples $\{X_i\}_{i=1}^n \sim P_X^{\otimes n}$. Using the optimal transport to measure the dissimilarity between the generated $P_{G(Z)}$ and target distribution leads to the following learning problem of GAN:

$$\min_{G\in\mathcal{G}} OT_c\big(P_{G(Z)}, Q_X^n\big). \quad (7)$$

Note that when the cost function is a distance raised to power p as in (1) and (2), the GAN is also known as p-Wasserstein GAN [29], [30].

Entropic Optimal Transport GAN: Solving the formulation in (7) involves solving for the optimal transport plan π – a joint distribution over the real and generated sample spaces, which is a difficult optimization problem with very slow convergence. Adding entropic regularization to the objective makes the problem strongly convex and thus solvable in linear time [25].

Formally, the entropy-regularized optimal transport, also known as Sinkhorn distance [6], is defined as

$$S_c(P_U, P_V) = \inf_{\pi \in \Pi(P_U, P_V)} \mathbb{E}_{(U,V)\sim\pi}[c(U, V)] + I_\pi(U, V),$$

where $I_\pi(U, V) = \int \log\Big(\frac{d\pi(u,v)}{dP_U(u)dP_Y(V)}\Big)d\pi(u, v)$ is the mutual information between U and V under the coupling π. In case $c(x, y) = \|x - y\|_p^p/\lambda$, the Sinkhorn distance is proportional to the entropy-regularized Wasserstein distance

$$\begin{aligned}\lambda S_c(P_U, P_V) &= W_{p,\lambda}(P_U, P_V)\\ &= \inf_{\pi \in \Pi(P_U, P_V)} \mathbb{E}_{(U,V)\sim\pi}\Big[\|U - V\|_p^p\Big] + \lambda I_\pi(U, V), \quad (8)\end{aligned}$$

The objective of an entropic optimal transport GAN is the entropy-regularized optimal transport between the generated distribution $G\#P_Z = P_{G(Z)}$ for some latent noise Z and the empirical approximation Q_X^n of the target distribution:

$$\min_{G\in\mathcal{G}} S_c\big(P_{G(Z)}, Q_X^n\big)$$

and the objective of an entropic p-Wasserstein GAN is

$$\min_{G\in\mathcal{G}} W_{p,\lambda}\big(P_{G(Z)}, Q_X^n\big). \quad (9)$$

It is worth mentioning that both non-regularized and regularized optimal transport formulations admit a dual formulation with optimization over functions of the input random variables. We note that the dual formulation for regularized optimal transport is unconstrained and hence easier to use, while the constraints for the unregularized counterpart are usually harder to enforce (e.g., Lipschitzness [29] or convexity [30]).

C. Optimal Transport GANs With LDP Data

Let $M : \mathcal{X} \to \mathcal{Y}$ be a randomized privacy-preserving mechanism: $Y = M(x) \sim P_M(y|x)$. For example, $P_{M(X)|X}(y|x)$ can be the Laplace pdf at $y - x$ for the Laplace mechanism or the Gaussian pdf at $y - x$ for the Gaussian mechanism. Let $P_Y = M\#P_X$ denote the distribution of Y, i.e., the push-forward distribution of P_X through the privatization mechanism M. The goal of learning a GAN from privatized samples is to reconstruct $G(Z) \approx X$ in distribution from a sample $S = \{Y\}_{i=1}^n \sim P_Y^{\otimes n}$ with empirical distribution $Q_Y^n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$.

III. Main Results

First, we focus on the population setting where the distribution of the privatized samples P_Y is known and show that by tailoring the cost function for optimal transport to the privatization mechanism M, the GAN learning problem with entropic optimal transport can recover the raw data distribution P_X.

Theorem 1: Let $X \sim P_X$ and $Y = M(X) \sim P_{M(X)|X}(\cdot|X)$. Assume that the privatisation mechanism M and the set of generator functions $\mathcal{G}$ is such that for any $G \in \mathcal{G}$,

$$D_{KL}\big(P_X\|P_{G(Z)}\big) > 0 \implies D_{KL}\big(P_{M(X)}\|P_{M(G(Z))}\big) > 0. \quad (10)$$

Let $c(x, y) = -\log P_{M(X)|X}(y|x)$ and

$$G^* = \arg\min_{G\in\mathcal{G}} S_c\big(P_{G(Z)}, P_Y\big). \quad (11)$$

If $P_X \in \{P_{G(Z)}|G \in \mathcal{G}\}$, i.e., P_X is realizable with the set of generator functions $\mathcal{G}$, then $P_{G^*(Z)} = P_X$.

Proof: Fix some $G \in \mathcal{G}$ and assume that Y is a continuous random variable. Denote $\mathcal{P} = \Pi(P_{G(Z)}, P_Y)$, $U = G(Z)$ Then by the definition of mutual information:

$$\begin{aligned}S_c\big(P_{G(Z)}, P_Y\big) &= \inf_{\pi\in\mathcal{P}} \mathbb{E}_{(U,Y)\sim\pi}\big[-\log p_M(Y|U)\big] + I_\pi(U, Y)\\ &= \inf_{\pi\in\mathcal{P}} \int\left(-\log p_M(y|u) + \log\frac{\pi(u, y)}{P_U(u)P_Y(y)}\right)\pi(u, y)dudy\end{aligned}$$

Notice that the terms on the RHS can be rearranged into the Kullback-Leibler divergence:

$$\begin{aligned}&S_c\big(P_{G(Z)}, P_Y\big)\\ &= \inf_{\pi\in\mathcal{P}} \int\left(-\log(P_Y(y)) + \log\frac{\pi_{Y|U}(y|u)}{p_M(y|u)}\right)\pi(u, y)dudy\\ &= h(Y) + \inf_{\pi\in\mathcal{P}} \mathbb{E}_{U\sim P_{G(Z)}} D_{KL}\big(\pi_{Y|U}(\cdot|U)\|p_M(\cdot|U)\big) \quad (12)\end{aligned}$$

The right-hand side is minimized when $\pi_{Y|U}(y|u) = p_M(y|u)$ for any $u \in \text{supp}(P_{G(Z)})$, $y \in \text{supp}(P_Y)$, which is a feasible coupling only if $P_{M(G(Z))} = P_Y$. By the realizability assumption $P_Y = P_{M(X)} = P_{M(G^*(Z))}$, so $P_{M(G(Z))} = P_Y \iff P_{M(G(Z))} = P_{M(G^*(Z))}$, or equivalently $D_{KL}(P_{M(G(Z))}\|P_{M(G^*(Z))}) = 0$. In addition to that, (10) implies that whenever the privatized distributions are equal $D_{KL}(P_{M(G(Z))}\|P_{M(G^*(Z))}) = 0$, it has to be that the original data distributions are equal too, so we conclude that $D_{KL}(P_{G(Z)}\|P_{G^*(Z)}) = 0$, or equivalently $P_{G(Z)} = P_{G^*(Z)} = P_X$. The same set of equalities also holds in the case when Y is discrete with differential entropy changing to entropy. ■

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

The theorem indicates that the optimal solution to the GAN optimization problem (11) generates the target distribution P_X, assuming that the generator class $\mathcal{G}$ is expressive enough to generate the target distribution. Note that the cost function in the definition of the entropic optimal transport $S_c(P_{G(Z)}, P_Y)$ in (11) has to be chosen as $c(x, y) = -\log P_{M(X)|X}(y|x)$ to match the privatization mechanism M. Thus provided that there are enough privatized samples, the generator will output the raw target distribution. Assumption (10) ensures that privatizing any generated distribution other than P_X will result in a distribution different from P_Y, or equivalently $P_X \neq P_{G(Z)}$ implies that $P_{M(X)} \neq P_{M(G(Z))}$, where the $\neq$ is in terms of the distribution functions.

This condition is needed to eliminate degenerate cases of privatization mechanisms, for example $M(X) = 0$. Note that if $P_{M(X)} = P_{M(G(Z))}$ and $P_{G(Z)} \neq P_X$, it is not possible to differentiate between them since the learning framework only has access to the privatized distribution. Moreover, we note that the condition is satisfied for any additive noise privatization mechanism provided that the noise characteristic function is non-zero everywhere, which holds for Laplace and Gaussian privatization mechanisms.

We next show that when the privatization mechanism is given by the popular Laplace or the Gaussian mechanisms, the entropic OT problem reduces to the entropic 1- and 2-Wasserstein GAN problems respectively.

Corollary 1: Under the conditions of Theorem 1, if $\sup_{x\in\mathcal{X}} \|x\|_1 \leq \Delta_1$, $p = 1$, and $Y = M(X)$ is the Laplace mechanism with noise scale ϵ/Δ_1, then training a GAN with loss $W_{1,\epsilon/\Delta_1}(P_{G(Z)}, P_Y)$ is ϵ-LDP, and recovers the target distribution: $P_{G^*(Z)} = P_X$.

Corollary 2: Under the conditions of Theorem 1, if $\sup_{x\in\mathcal{X}} \|x\|_2 \leq \Delta_2$, $p = 2$, and $Y = M(X)$ is the Gaussian mechanism with noise scale σ defined in (5), then training a GAN with loss $W_{2,2\sigma^2}(P_{G(Z)}, P_Y)$ is (ϵ, δ)-LDP, and recovers the target distribution: $P_{G^*(Z)} = P_X$.

We note that first [31] showed that projection with respect to entropic optimal transport is maximum likelihood deconvolution, and Corollary 1 and 2 can be viewed as the population case of [31] when the data distribution comes from a convolution model. While [31] does not draw a connection between LDP and entropic optimal transport and is not concerned with proposing an entropy-regularized GAN framework for privatized data, their result has a similar flavor to our results in Corollary 1 and 2. However, we note that proving that entropic projection is maximum likelihood deconvolution as done in [31] is more involved, while our Theorem1, which holds for general privatization mechanisms and not only under additive noise ones as in [31], simply follows from the non-negativity of KL divergence and is inherently related to the characterization of the rate distortion function as discussed in Section I-A.

The results so far are only applicable to the realizable case $P_X \in \{P_{G(Z)}|G \in \mathcal{G}\}$, namely when the true data distribution P_X can be generated. However, this is not always the case in practice, and the approximation error of the class $\mathcal{G}$ given by $\min_{G\in\mathcal{G}} W_p^p(P_{G(Z)}, P_X)$ can be non-zero. The following lemma can be used in this case. Note that it holds for $p \in \{1, 2\}$ which correspond to the Laplace and Gaussian mechanisms respectively.

Lemma 1: Let $X \sim P_X$ and $Y = M(X) = X + N$, where $N \sim f_N(z) \propto e^{-\|z\|_p^p/(p\sigma^p)}$, $p \in \{1, 2\}$, and

$$G^* = \arg\min_{G\in\mathcal{G}} W_{p,p\sigma^p}\big(P_{G(Z)}, P_Y\big).$$

If $P_X \notin \{P_{G(Z)}|G \in \mathcal{G}\}$:

$$D_{KL}\big(P_{G^*(Z)+N}\|P_{X+N}\big) \leq \min_{G\in\mathcal{G}} W_p^p\big(P_{G(Z)}, P_X\big)$$

where $D_{KL}(P\|Q) = \int P \log \frac{dP}{dQ}$ is the KL-divergence.

Lemma 1 bounds the KL Divergence between the push-forwards of the generated and target distributions. This is sometimes called the smoothed KL divergence between P_X and $P_{G^*(Z)}$ [32]. It ensures that if ϵ is the approximation error in p-Wasserstein distance of the class $\mathcal{G}$, then $P_{G*(Z)}$ is ϵ-close to the target distribution P_X in smoothed KL-divergence. We prove the lemma in Section V-A.

Lemma 1, Theorem 1, and Corollaries 1, 2 have been established in the population setting where we work directly with P_Y. In practice, P_Y is approximated by the empirical distribution Q_Y^n of its samples $\{Y\}_{i=1}^n \sim P_Y^{\otimes n}$. We next investigate how fast the solution of the empirical problem converges to the population solution and first we establish a new result on convergence of the entropic optimal transport that is suited to our framework.

Theorem 2 (Entropic Optimal Transport GAN Excess Risk Bound): Let the target data distribution P_X be a probability measure with bounded support, namely $\operatorname{supp} P_X \subseteq \mathcal{X} \subset \mathbb{R}^d$ and $\sup_{x\in\mathcal{X}} \|x\|_\infty = D < \infty$, and let the set of admissible generators be $\mathcal{G} \subseteq \{G : \mathcal{Z} \to \mathcal{X}\}$. Let the cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be measurable with respect to the product measure: $\mathbb{E}_{(X,G(Z))\sim P_X\times P_{G(Z)}} c(X, G(Z)) < \infty$ for any $G \in \mathcal{G}$ and non-negative: $c(x, y) \geq 0$ for any $x \in \mathcal{X}, y \in \mathbb{R}^d$. If the exponentiated negative cost function is a Mercer kernel, that is $k(x, y) = e^{-c(x,y)}$ is

- continuous
- symmetric: $k(x, y) = k(y, x)$
- positive definite: for any number n and any set of points $\{x_i\}_{i=1}^n \subset \mathcal{X}$ the matrix with entries $K_{ij} = K(x_i, x_j)$ is positive semi-definite

then for

$$G^* = \arg\min_{G\in\mathcal{G}} S_c\big(P_{G(Z)}, P_Y\big),\ G_n = \arg\min_{G\in\mathcal{G}} S_c\big(P_{G(Z)}, Q_Y^n\big),$$

where Q_Y^n is the empirical distribution of $\{Y_i\}_{i=1}^n - n$ i.i.d. samples from P_Y it holds that

$$\mathbb{E}\big[S_c\big(P_{G_n(Z)}, P_Y\big) - S_c\big(P_{G^*(Z)}, P_Y\big)\big] \leq 4\mathbb{E} \sup_{x\in\mathcal{X}} e^{2c(x,Y)}/\sqrt{n}.$$

The detailed proof of the theorem is given in Section V-B and is based on two main ideas: one is a simple decomposition of the dual function of entropic optimal transport similar to [33] and the other one is a Rademacher complexity bound for one of the dual potential, which we obtain through the Mercer's decomposition of the conditional probability distribution. We defer the proof to Section V, while we provide a discussion and give two important corollaries – for a general privatization mechanism and the Laplace mechanism.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

The above theorem is easily adjusted to the privatization framework by plugging in $c(x,y) = -\log P_{M(X)|X}(y|x)$, which results in the following corollary.

Corollary 3: Let the target distribution P_X be a probability measure with bounded support, namely $\operatorname{supp} P_X \subseteq \mathcal{X} \subset \mathbb{R}^d$ and $\sup_{x\in\mathcal{X}} \|x\| < \infty$, and let the set of admissible generators be $\mathcal{G} \subseteq \{G : \mathcal{Z} \to \mathcal{X}\}$. Additionally, let the distribution function of the privatization mechanism be a Mercer kernel, that is $k(x,y) = P_{M(X)|X}(y|x)$ is

- continuous
- symmetric: $k(x,y) = k(y,x)$
- positive definite: for any number n and any set of points $\{x_i\}_{i=1}^n \subset \mathcal{X}$ the matrix with entries $K_{ij} = K(x_i,x_j)$ is positive semi-definite.

Then for $Y = M(X) \sim p_{M(X)|X}$, $c(x,y) = -\log P_{Y|X}(y|x)$,

$$G^* = \arg\min_{G\in\mathcal{G}} S_c\big(P_{G(Z)}, P_Y\big),\ G_n = \arg\min_{G\in\mathcal{G}} S_c\big(P_{G(Z)}, Q_Y^n\big),$$

where Q_Y^n is the empirical distribution of $\{Y_i\}_{i=1}^n - n$ i.i.d. samples from P_Y it holds that

$$\begin{aligned}&\mathbb{E}_{Y\sim P_Y}\big[S_c\big(P_{G_n(Z)}, P_Y\big) - S_c\big(P_{G^*(Z)}, P_Y\big)\big]\\ &\quad\le 4\mathbb{E}\sup_{x\in\mathcal{X}} P_{M(x)|X}(Y|x)^{-2}/\sqrt{n}\end{aligned} \tag{13}$$

One can now apply the corollary to the Laplace mechanism, whose pdf is a Mercer kernel. However, the direct application of the theorem will not result in a meaningful upper bound since the RHS of (13) is infinite, but one can still use Theorem 2 after noting that the cost function decomposes into $c(x,y) \propto \|x-y\|_1 = a(y) + \tilde{c}(x,y)$, where the function $\tilde{c}$ is bounded and the term $a(y)$ only depends on y and not x, so it can be factored out of the entropic optimal transport. This leads to the following corollary, where we only require the support of the data distribution to be bounded.

Corollary 4: If $Y = M(X)$ is the Laplace mechanism with noise scale σ, the support of the data distribution is bounded in ∞ norm: $\sup_{x:P_X(x)>0}\|x\|_\infty \le D$ as well as the output of the generator functions: $\forall G \in \mathcal{G}$ the ∞ norm of the output does not exceed D: $\sup_{G\in\mathcal{G}}\|G\|_\infty \le D$ then for

$$G^* = \arg\min_{G\in\mathcal{G}} W_{1,\sigma}\big(P_{G(Z)}, P_Y\big),\ G_n = \arg\min_{G\in\mathcal{G}} W_{1,\sigma}\big(P_{G(Z)}, Q_Y^n\big),$$

where Q_Y^n is the empirical distribution of n i.i.d. samples from P_Y it holds that

$$\mathbb{E}\big[W_{1,\sigma}\big(P_{G_n(Z)}, P_Y\big) - W_{1,\sigma}\big(P_{G^*(Z)}, P_Y\big)\big] \le 4\sigma e^{4dD/\sigma}/\sqrt{n}.$$

Proof: For the Laplace mechanism $c(x,y) = d\log(2\sigma) + \|x-y\|_1/\sigma$. Let $(b(y))_i = \max\{\min\{y_i, D\}, -D\}$, namely $b(y)$ clips y to the interval $[-D, D]$ coordinate-wise. Then $|x_i - y_i| = |x_i - b(y_i)| + |b(y_i) - y_i|$, denote $\tilde{c}(x,y) = c(x,b(y))$. Since $c(x,y) = \tilde{c}(x,y) + \|y - b(y)\|_1$, where the term $\|b(y)-y\|_1$ does not depend on x and, therefore, the coupling one can write

$$S_c\big(P_{G(Z)}, P_Y\big) = S_{\tilde{c}}\big(P_{G(Z)}, P_Y\big) + \mathbb{E}_{Y\sim P_Y}\|Y - b(Y)\|_1,$$

which leads to the following excess risk bound

$$\begin{aligned}&\mathbb{E}\big[S_c\big(P_{G_n(Z)}, P_Y\big) - S_c\big(P_{G^*(Z)}, P_Y\big)\big]\\ &\quad= \mathbb{E}\big[S_{\tilde{c}}\big(P_{G_n(Z)}, P_Y\big) - S_{\tilde{c}}\big(P_{G^*(Z)}, P_Y\big)\big]\end{aligned}$$

Theorem 2 can now be applied to $S_{\tilde{c}}$ to result in

$$\begin{aligned}\mathbb{E}\big[S_c\big(P_{G_n(Z)}, P_Y\big) - S_c\big(P_{G^*(Z)}, P_Y\big)\big] &\le 4\mathbb{E}\sup_{x\in\mathcal{X}} e^{2c(x,b(y))}/\sqrt{n}\\ &\le 4\sup_{x\in\mathcal{X}, y:\|y\|_\infty\le D} e^{2\|x-y\|_1/\sigma}/\sqrt{n} \le 4e^{4dD/\sigma}/\sqrt{n}.\end{aligned}$$

Expressing the Wasserstein distance in terms of optimal transport using (8) leads to

$$\mathbb{E}\big[W_{1,\sigma}\big(P_{G_n(Z)}, P_Y\big) - W_{1,\sigma}\big(P_{G^*(Z)}, P_Y\big)\big] \le 4\sigma e^{4dD/\sigma}/\sqrt{n}$$

■

Note that to achieve ϵ-differential privacy one needs to choose $\sigma \ge \epsilon/\sup_{x:P_X(x)>0}\|x\|_1$ and if, for example, the data is supported on $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_\infty \le D\}$ then to ensure ϵ-LDP on needs to choose $\sigma = \sup_{x\in\mathcal{X}}\|x\|_1/\epsilon = dD/\epsilon$, which leads to

$$\mathbb{E}\big[W_{1,\sigma}\big(P_{G_n(Z)}, P_Y\big) - W_{1,\sigma}\big(P_{G^*(Z)}, P_Y\big)\big] \le 4dDe^{4\epsilon}/\epsilon\sqrt{n}.$$

Note that in this case, the excess risk not only does not suffer from the curse of dimensionality as in the case of unregularized Wasserstein distance but scales linearly with the dimension. However, it is important to note that in our framework we have changed the loss function from unregularized Wasserstein distance to entropic Wasserstein distance and the excess risk bound here is in terms of entropic Wasserstein distance. In particular, this result does not imply a parametric rate of convergence in terms of unregularized Wasserstein distance.

We next state the generalization result for the Gaussian mechanism which is proven in Section V-C. To formally state the sample complexity results, let us first recall some definitions. A distribution P_X supported on a d-dimensional set $\mathcal{X}$ is σ^2 sub-Gaussian for $\sigma \ge 0$ if $\mathbb{E}\exp\big(\|X\|^2/(2d\sigma^2)\big) \le 2$. Let $\sigma^2(X) = \min\{\sigma \ge 0 \big| \mathbb{E}\exp(\|X\|^2/(2d\sigma^2)) \le 2\}$ denote the sub-Gaussian parameter of the distribution of X. A set of generators $\mathcal{G}$ is said to be star-shaped with a center at 0 if a line segment between 0 and $G \in \mathcal{G}$ also lies in $\mathcal{G}$, i.e.,

$$G \in \mathcal{G} \Rightarrow \alpha G \in \mathcal{G}, \forall \alpha \in [0,1]. \tag{14}$$

Note that these conditions are not very restricting. For example, the set of all linear generators, the set of linear functions with a bounded norm or a fixed dimension, the set of all L-Lipschitz functions or neural networks with a relu ($f(x) = \max(0,x)$) activation function at the last layer all satisfy (14).

Theorem 3 (Excess Risk of the Gaussian Mechanism): Let P_Z and P_X be sub-Gaussian, the support of P_X be d-dimensional, and the generator set $\mathcal{G}$ consist of L-Lipschitz functions, namely $\|G(Z_1) - G(Z_2)\| \le L\|Z_1 - Z_2\|$ for any $Z_1, Z_2 \in \mathcal{Z}$. Assume additionally that $\mathcal{G}$ satisfies (14). If $Y = M(X) = X + N$ is the Gaussian mechanism with noise scale σ then for

$$G^* = \arg\min_{G\in\mathcal{G}} W_{2,2\sigma^2}\big(P_{G(Z)}, P_Y\big), G_n = \arg\min_{G\in\mathcal{G}} W_{2,2\sigma^2}\big(P_{G(Z)}, Q_Y^n\big),$$

where Q_Y^n is the empirical distribution of n i.i.d. samples from P_Y it holds that

$$\begin{aligned}&\mathbb{E}\big[W_{2,2\sigma^2}\big(P_{G_n(Z)}, P_Y\big) - W_{2,2\sigma^2}\big(P_{G^*(Z)}, P_Y\big)\big]\\ &\quad\le C_d\sigma^2 n^{-1/2}\left(1 + \left(\tau^2(1+\sigma(X)/\sigma)^2\right)^{\lceil 5d/4\rceil+3}\right),\end{aligned}$$

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

where $\tau = \max\{L\sigma(Z)/\sigma(X), 1\}$ and C_d is a dimension dependent constant.

The theorems show that the excess risk diminishes at the parametric rate (of order $1/\sqrt{n}$), which breaks the curse of dimensionality (convergence of order $n^{-\Omega(1/d)}$), often attributed to GANs. We also observe that the excess risk is approximately linear in σ^2, the privatization noise scale, beyond a certain threshold ($\sigma^2 > \sigma(X)^2$)). This implies that convergence for larger σ^2, corresponding to higher privacy, can be achieved by increasing the number of samples n.

The above results show that the value of the loss function under the empirical solution G_n converges to the value of the loss function under the population solution G^*. However, this result does not directly relate $P_{G_n(Z)}$ to P_X. Next, we use Theorem 3 and Corollary 5 to upper bound the smoothed KL-divergence between P_X and $P_{G_n(Z)}$.

Corollary 5: If the target distribution can be generated, that is $P_X \in \{P_{G(Z)} | G \in \mathcal{G}\}$, then

- under the conditions of Corollary 4 one has

$$\mathbb{E}\big[D_{KL}\big(P_Y \| P_{G^n(Z)+N}\big)\big] \le 4e^{4dD/\sigma}/\sqrt{n},$$

  where $N \sim f_N$ is the privatization noise of the Laplace mechanism, $f_N(x) \propto e^{-\|x\|_1/\sigma}$
- under the conditions of Theorem 3

$$\mathbb{E}\big[D_{KL}\big(P_Y \| P_{G^n(Z)+N}\big] \le C_d n^{-1/2}\left(1 + \left(\tau^2(1+\sigma(X)/\sigma)^2\right)^{\lceil 5d/4\rceil+3}\right)$$

Note that the parametric convergence of the smoothed KL-divergence results in the convergence of the smoothed Wasserstein distance [32], which is, in turn, a distance metrizing weak convergence similar to W_p.

We note that known results on the sample complexity of entropic optimal transport are either only applicable to $c(x, y) \propto \|x - y\|_2^2$ [8] or require the cost function to be ∞-differentiable [26], [34], both of which assumptions do not apply to the Laplace setting. Reference [33] proves sample complexity bounds for entropic optimal transport but with bounded cost, and most importantly, extending their result to the GAN setting where one of the distributions depends on the sample would require restricting the set of the generators to have a small complexity (VC-dimension or Rademacher complexity for example), which is not common in practice and is hard to compute. Our result, on the other hand, does not depend on the complexity of the set of generators, provided that their output has a bounded norm. We achieve this by bounding the Rademacher complexity of the set of one of the dual potentials. Similar to [26], we invoke the bound involving the RKHS norms, but instead of using the smoothness of the dual potentials and the RKHS of the Sobolev space (which requires at least $\lceil d/2 \rceil$ continuous differentiability), we rely on the fact that in the case of Laplace distribution the privatization mechanism's pdf is a Mercer kernel, which allows us to use its RKHS norms without requiring smoothness. This technique can also be used to provide sample complexity bounds for entropic 1-Wasserstein distance. Moreover, the properties of ℓ_1 norm allow us to eliminate the dependence on the tails of the privatized data distribution, which is only sub-exponential and does not concentrate as well as sub-Gaussian distributions. The proof of the theorem as well as all other proofs are deferred to Section V.

IV. Experimental Results

We first describe the approach we used to privatize the data and train the GAN, and then present the experimental results.[1]

Data Privatization: We conduct experiments for both the Laplace and Gaussian data privatization mechanisms. We set $f(x) = \text{vec}(x)$, where $\text{vec}(x)$ – is the vectorization of x, specifically if $x \in \mathbb{R}^d$ is a vector then $f(x) = x$ and if it is a matrix $x \in \mathbb{R}^{d_1 \times d_2}$ then $f(x) \in \mathbb{R}^d$, $d = d_1 d_2$ in the definitions of the Laplace and Gaussian privatization mechanisms.

For the Laplace mechanism and LDP budget ϵ we set the ℓ_1 sensitivity to be $\Delta = \sup_{x,x' \in \mathcal{X}} \|f(x) - f(x')\|_1$ and the noise scale $\sigma = \Delta/\epsilon$, so $Y = f(X) + Z$, where $Z_i \sim \text{Laplace}(0, \sigma)$ i.i.d. for each coordinate $i \in \{1, \ldots, d\}$. Similarly, for the Gaussian mechanism and LDP budget ϵ we set the ℓ_2 sensitivity to be $\Delta = \sup_{x \in \mathcal{X}} \|f(x) - f(x')\|_2$ and the noise scale σ that satisfies (5) so $Y = f(X) + Z$, where $Z_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. for each coordinate $i \in \{1, \ldots, d\}$.

GAN training: For training the Sinkhorn GAN we follow the work of [35] using Sinkhorn-Knopp algorithm [36] to approximate the optimal transport plan π in (9) from mini-batches of size b for the generated and privatized training data. The algorithm is stated below, where θ stands for the parameter of the Generator, i.e., $\mathcal{G} = \{G_\theta : \mathcal{Z} \to \mathcal{X} | \theta \in \Theta\}$.

Dataset and architecture: We train our models on synthetic data as well as MNIST data [37], consisting of 60000 grayscale images of handwritten digits. We do not use the labels to mimic a fully unsupervised training scenario. The generator model for MNIST is DCGAN from [38] with latent space dimension 100. All the losses were used in the primal formulations (6), (8) with optimization over the coupling matrix.

Remark 1: Note that since the DP noise is added to the training data, even if the training algorithm is an iterative process, the final privacy guarantee does not depend on (1) the number of rounds and (2) the specific privacy accountant or composition theorem used.

A. Synthetic Data

We first test our method on synthetic data. In this experiment, we sample data uniformly from a two-dimensional manifold shaped as a half-circle of radius 1 and we assume the support $\mathcal{X}$ is known to be the half circle, so that the ℓ_1 sensitivity is 2 and the ℓ_2 sensitivity is $\sqrt{2}$. The 400000 sampled points are privatized with Laplace or Gaussian noise and then the entropic $p-$Wasserstein GAN for the corresponding p is trained on the privatized data using Algorithm 1. We used 2-dimensional latent noise uniform in $[-1, 1]^2$ and a small Neural Network with 2 hidden layers and 256 neurons on each hidden layer. We trained it with batch gradient descent

[1] Additional experiments with synthetic data, and the dependence of the loss function on the number of samples and privacy budget can be found on arxiv: https://arxiv.org/abs/2306.09547.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Training GAN With $W_{p,p\sigma^p}$

Input: θ_0, $\mathcal{D} = \{y_i\}_{i=1}^n$ (the privatized training data), b (batch size), L (number of Sinkhorn iterations), α (learning rate)

Output: θ

$\theta \leftarrow \theta_0$

for $t = 1, 2, \ldots$ **do**

Sample $\{y_i\}_{i=1}^b$ i.i.d. from the dataset $\mathcal{D}$, $Q_Y^b := \frac{1}{b}\sum_{i=1}^b \delta_{y_i}$

Sample $\{z_i\}_{i=1}^b \overset{\text{i.i.d}}{\sim} P_{\mathcal{Z}}$, $Q_G^b := \frac{1}{b}\sum_{i=1}^b \delta_{G_\theta(z_i)}$

Calculate the optimal transport plan for $S_c(Q_G^b, Q_Y^b)$ with L Sinkhorn-Knopp steps

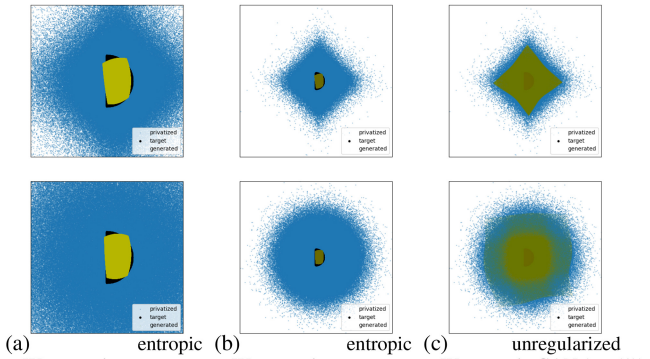
$\pi \approx \underset{\pi \in \Pi(Q_G^b, Q_Y^b)}{\arg\min} \mathbb{E}_{(U,Y)\sim\pi}[c(U,Y)] + I_\pi(U), Y)$

$C_{ij} \leftarrow c(y_i, G_\theta(z_j))$ for $i, j = 1, \ldots, b$

$g_t \leftarrow \nabla_\theta \langle \pi, C \rangle$

$\theta \leftarrow \theta - \alpha g_t$.

end for

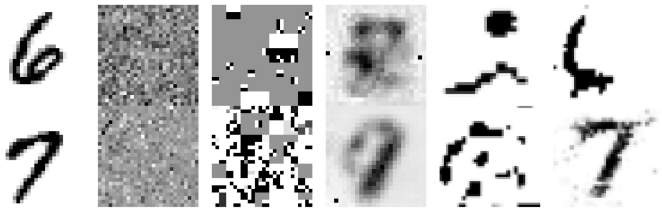


(a) entropic p-Wasserstein GAN (eq. (9)) (b) entropic p-Wasserstein GAN (eq. (9)), enlarged (c) unregularized p-Wasserstein GAN (eq. (1))

Fig. 1. Learning data from the half-circle-shaped manifold privatized with Laplace mechanism $\epsilon = 5$ (top) and Gaussian mechanism $\epsilon = 5, \delta = 10^{-4}$ (bottom). Columns (a) and (b) show the manifold learned with entropic p-Wasserstein GAN (9), and column (c) shows the manifold learned with unregularized p-Wasserstein GAN (7). Note $p = 1$ for Laplace mechanism and $p = 2$ for Gaussian mechanism.

using RMSprop optimizer with a learning rate of 10^{-4}. The entropy-regularized Wasserstein distance was calculated with geomloss library [25] for the full dataset in each iteration ($b = n$ in Algorithm 1) with scaling parameter set to 0.99. Figure 1 shows the learned manifolds with data privatized with Laplace mechanism $\epsilon = 5$ and entropic 1-Wasserstein loss (9) (top) and with data privatized with the Gaussian mechanism $\epsilon = 5, \delta = 10^{-4}$ and entropic 2-Wasserstein loss (9) (bottom). We note that for both the Laplace and Gaussian mechanisms entropic regularization allows to recover the original domain of the data (columns (a) and (b)), even-though noise in both cases appears to be large enough to completely obfuscate the data domain (column (b)). Without regularization (column (c)), the model generates the privatized distribution and fails to recover the original domain.

We note that the $\epsilon = 5$ local differential privacy guarantee obtained with the Laplace mechanism can be translated to a central privacy guarantee by leveraging privacy amplification by shuffling. Since the distribution of the output of



(a) (b) (c) (d) (e) (f)

Fig. 2. Learning with LDP privatized images: (a) raw images from MNIST dataset, (b) image samples privatized with Gaussian $(\epsilon, \delta) = (35, 10^{-4})$ (top) and Laplace mechanism $\epsilon = 196$ (bottom), (c) images from the second column denoised with the wavelet transform and (d) the Noise2Void [41], (e) random samples from the output distribution of the unregularized p-Wasserstein GAN (2), and (f) generated samples from the entropic p-Wasserstein GAN trained on privatized data (picked to be the same digit as in column (a)).

Algorithm 1 does not depend on the order of the samples in the privatized data (due to the random sampling) and because the local privatization mechanism only depends on the data at the client and no auxiliary input, one can assume that the data is shuffled before privatization, which allows to apply [39, Th. 3.2], resulting in a $(\epsilon_c = 0.353, \delta_c = 10^{-6})$ central differential privacy guarantee for the Laplace mechanism.

B. MNIST: Comparison With Denoising

We next provide our experimental results with MNIST [37] and DCGAN [38] generator. The pixel values of the images were rescaled to $[-1, 1]$ leading to $\Delta = 2 \times 28^2$ ℓ_1 sensitivity and $\Delta = 56$ ℓ_2 sensitivity.

We use 100-dimensional Uniform $[0, 1]$ noise at the input to the generator ($\mathcal{P}_z = \text{Unif}[0,1]^{100}$). In Figure 2, we show two raw samples from the MNIST dataset (column (a)) and the corresponding privatized images (column (b)). In column (c), we denoise the privatized images in the second column with wavelet transform [40]; the results indicate that the wavelet transform can not be used to recover the images. Here, the wavelet transform parameters for denoising (the wavelet basis, the level and reconstruction thresholds) were optimized to minimize the average distance between the reconstructed and original image under the particular noise instance, thus providing better results than one would expect in a fully privatized setting. In column (d), we used the noise2void [41] image denoising mechanism with parameters as given in the paper and trained it on the whole dataset of privatized images, and showed that it also failed to reconstruct the images. These experiments suggest that the privatization noise is large enough to preclude reconstruction of the original images. Column (e) shows samples generated by GANs trained with p-Wasserstein loss (without entropic regularization) that fails to learn from privatized data. Finally, in (f) we provide samples obtained by our method stated in Algorithm 1 in Figure 2. Note that while the generated images by the entropic $p-$Wasserstein GAN are not perfect for the chosen privacy levels, the results do suggest that the model has learned to generate new images of digits. For training the entropic p-Wasserstein GAN we used 400 Sinkhorn-Knopp iterations and the Adam optimizer with learning rate 10^{-4} for 100 epochs.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

Fig. 3. Entropic 1-Wasserstein GAN on MNIST trained on data privatized with the Laplace mechanism achieving ϵ-LDP $\epsilon = 35$ (left) and $\epsilon = 25$ (right).

Fig. 4. Entropic 2-Wasserstein GAN on MNIST trained on data privatized with the Gaussian mechanism achieving (ϵ, δ)-LDP with $\delta = 10^{-4}$ and $\epsilon = 30$ (left) and $\epsilon = 25$ (right).

C. MNIST: Higher Privacy Samples

In this section, we further investigate the performance of entropic p-Wasserstein GAN on locally privatized data. We set the number of sinkhorn steps $L = 400$ and the batch size to be $b = 400$ and we performed optimization with Adam optimizer [42] and learning rate varied in $\{0.005, 10^{-4}, 5 \times 10^{-5}\}$. We optimized for 150 epochs. For $p = 1$ we first took the discrete cosine transform of the images and clipped the coefficients below 0.8 quantile to preserve more information, and then to control the sensitivity, we projected each training image onto an ℓ_1 ball with radius 140 (the parameters were chosen based on 1 held-out image in a way that it would not visually distort the image beyond recognition). We also applied DCT transform to the generator output before plugging it into the loss function. Similarly, for $p = 2$ we projected each training image onto an ℓ_2 ball with radius 20, but we did not apply any transforms (since ℓ_2-norm does not change under multiplication by an orthonormal matrix).

The results are presented on Figure 3 for $p = 1$ and Figure 4 for $p = 2$. They indicate the effectiveness of our model at higher privacy regimes. However, smaller ϵ values still produced a lot of noise in the generated samples or eroded the images significantly. This can be potentially mitigated by increasing the number of samples as suggested in Theorem 3; however the relatively small size of MNIST limits the privacy levels that can be achieved. Note that the privacy parameters chosen for the images are rather large, however, our method still performs reasonably despite the exponential blowup in the excess risk. We believe that the dependence of the excess risk on the privacy budget ϵ can thus be improved and is a direction for future work.

We provide additional samples for different privacy levels and report 400 randomly sampled digits on Figure 5 for the Laplace and Gaussian mechanisms.

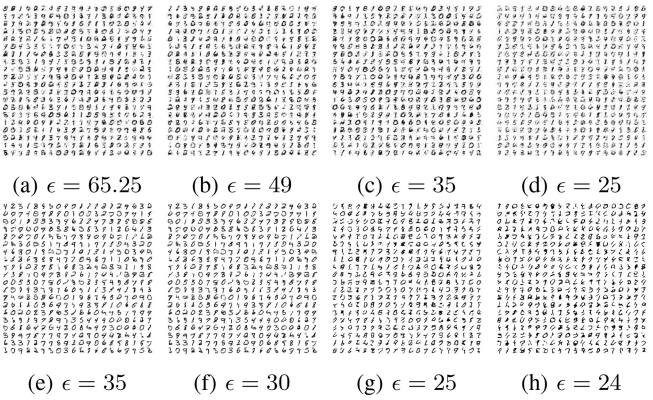
D. Comparison With Noisy Wasserstein GAN

We next provide our experimental results with MNIST and FashionMNIST [43], which is a set of 60000 grayscale images of clothing items of size 28×28.

Here we fix $p = 1$ and compare entropic 1-Wasserstein GAN to a noisy 1-Wasserstein GAN applied to data privatized with the Laplace mechanism.

The noisy 1-Wasserstein GAN is a GAN trained with the following unregularized loss function:

$$W_1\left(P_{M(G(Z))}, Q_Y^n\right), \tag{15}$$



(a) $\epsilon = 65.25$ (b) $\epsilon = 49$ (c) $\epsilon = 35$ (d) $\epsilon = 25$

(e) $\epsilon = 35$ (f) $\epsilon = 30$ (g) $\epsilon = 25$ (h) $\epsilon = 24$

Fig. 5. Laplace mechanism with different privacy budgets ϵ and clipping of the discrete cosine transform, entropic 1-Wasserstein GAN (top) and Gaussian mechanism with different privacy budgets ϵ for $\delta = 10^{-4}$ and clipping the Euclidean norm of images, entropic 2-Wasserstein GAN (bottom).

(a) 1-Wasserstein GAN (15), FID = 318

(b) entropic 1-Wasserstein GAN, 9, FID=208

Fig. 6. Fashion MNIST samples generated with 1-Wasserstein GAN and the addition of noise to the generator (a) and entropic 1-Wasserstein GAN (b).

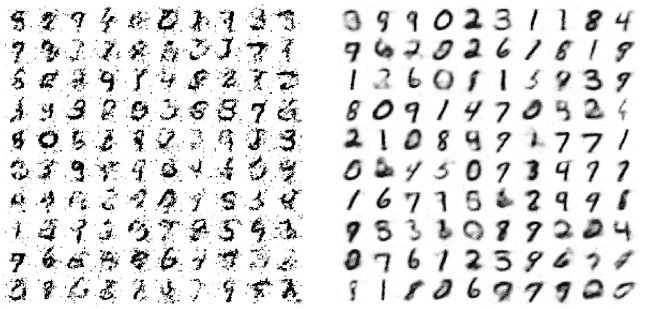
where $M(\cdot)$ is the privatization mechanism. Note that this loss function also satisfies Theorem 1, but suffers from the curse of dimensionality, namely $W_1(P_{M(G(Z))}, Q_Y^n) = \Omega(n^{-1/d})$. However, Wasserstein GANs have been successful in practice [29], [44] and use a different benchmark, so we compare their performance with the performance of the entropic 1-Wasserstein GAN.

We again use a DCGAN as a generator, but we removed the batch normalization layers and used a larger batch size of 6000 as suggested by [45]. As in Section IV-B, experiments in this section do not involve projecting onto the ℓ_1/ℓ_2 balls. The pixel values of the images were rescaled to $[-1, 1]$ leading to $\Delta_1 = 2 \times 28^2$ ℓ_1 sensitivity and $\Delta_2 = 56$ ℓ_2 sensitivity. For FashionMNIST we choose the noise scale $\sigma = 7$, which results in $\epsilon = 224$-LDP, and for MNIST we choose the noise scale $\sigma = 3$, which results in $\epsilon = 97$-LDP. We also report Frechet Inception Distance (FID) [46] calculated between the generated and the validation set as a quantitative measure of performance. We report the FID measures and random samples in Figure 6 for FashionMNIST and Figure 7 for MNIST.

The experiments clearly show that the images generated by the entropic Wasserstein GAN still look like clothing items, while the images generated by the unregularized Wasserstein GAN with noise added to the generator look like noise. The closeness of the distributions is also validated by the closeness in FID scores.

Note that the images for the entropic 1-Wasserstein GAN are still corrupted by noise even given the very high privacy budget. This is due to the large ℓ_1 sensitivity of the data. In the next section we address that problem with clipping the image norms before applying privatization noise.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.



(a) 1-Wasserstein GAN (15), FID = 223 (b) entropic 1-Wasserstein GAN, 9, FID=28

Fig. 7. MNIST samples generated with 1-Wasserstein GAN and the addition of noise to the generator (a) and entropic 1-Wasserstein GAN (b).

(a) $\epsilon = 100$, FID = 183 (b) $\epsilon = 78, 6$, FID = 214

Fig. 8. Entropic 1-Wasserstein GAN trained on FashionMNIST data privatized with the Laplace mechanism.

(a) $\epsilon = 35, \delta = 10^{-4}$, FID = 207 (b) $\epsilon = 25$, FID = 237

Fig. 9. Entropic 2-Wasserstein GAN trained on FashionMNIST data privatized with the Gaussian mechanism.

E. Fashion MNIST: Higher Privacy Samples

To achieve a smaller privacy budget we clipped the norm of the images to have sensitivity $\Delta_1 = 700$ and $\Delta_1 = 550$. Adding Laplace noise with $\sigma = 7$ then results in $\epsilon = 100$ and $\epsilon = 78, 6$ -LDP. The results are presented in Figure 8.

Similarly, for $p = 2$ we clipped the ℓ_2 norm to have sensitivity $\Delta_2 = 40$. Adding Gaussian noise with $\sigma = 9.17$ or $\sigma = 7.24$ results in $\epsilon = 25$ and $\epsilon = 35$ privacy for $\delta = 10^{-4}$, see Figure 9.

The results indicate the effectiveness of our model at higher privacy regimes. However, smaller ϵ values still produced a lot of noise in the generated samples or eroded the images significantly. This can be potentially mitigated by increasing the number of samples as suggested in Theorem 3; however the relatively small size of MNIST limits the privacy levels that can be achieved.

V. Detailed Proofs

In this section we prove the theorems from Section III.

A. Proof of Lemma 1

We first prove the following lemma, which is used in this proof and the proof of Corollary 5 (Corollary 5 is a direct consequence of the lemma and excess risk results – Theorem 3 and Corollary 4).

Lemma 2: Let $X \sim P_X$ and $Y = M(X) = X+N$, where $N = (N_1, \ldots, N_d) \sim f_N$ independent of X and $f_N(x) \propto e^{-\|x\|^p/(p\sigma^p)}$ Then

$$D_{KL}\big(P_Y \| P_{G(Z)+N}\big) \leq \big(W_{2,2\sigma^2}\big(P_Y, P_{G(Z)}\big) - W_{2,2\sigma^2}(P_Y, P_X)\big)/p\sigma^p, \quad (16)$$

where $D_{KL}(P\|Q)$ is the KL-divergence ($D_{KL}(P\|Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$ for continuous P_X and $D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$ for discrete X).

Proof: We start by noting that $c(x, y) = -\log f_N(x - y) = \|x - y\|_p^p/(p\sigma^p) + C$, where C is a constant. Thus, for two probability measures μ, ν

$$S_c(\mu, \nu) = W_{p,p\sigma^p}(\mu, \nu)/p\sigma^p + C. \quad (17)$$

Additionally, since $P_Y = P_{M(X)}$, the KL-divergence in (12) is zero for $P_{G(Z)} = P_X$, so $S_c(P_X, P_Y) = h(Y)$. We can now rewrite the difference on the RHS of (16) using (17) and (12):

$$\begin{aligned}&\big(W_{p,p\sigma^p}\big(P_{G(Z)}, P_Y\big) - W_{p,p\sigma^p}(P_X, P_Y)\big)/p\sigma^p \\ &= S_c\big(P_{G(Z)}, P_Y\big) - S_c(P_X, P_Y) \\ &= \inf_{\pi\in\Pi(P_{G(Z)},P_Y)} \mathbb{E}_{U\sim P_{G(Z)}} D_{KL}\big(\pi_{Y|U}(\cdot|U)\|p_M(\cdot|U)\big), \quad (18)\end{aligned}$$

where $p_M(\cdot|u) = f_N(\cdot - U)$ is the conditional pdf of the privatization mechanism.

We then use the chain rule for KL-divergence, which states that for any two joint distributions $Q^1 \ll Q^2$ with marginals (Q_X^1, Q_Y^1) and (Q_X^2, Q_Y^2) correspondingly, it holds that

$$\begin{aligned}D_{KL}\Big(Q^1\|Q^2\Big) &= D_{KL}\Big(Q_X^1\|Q_X^2\Big) \\ &\quad + \mathbb{E}_{X\sim Q_X^1} D_{KL}\Big(Q^1(\cdot|X)\|Q^2(\cdot|X)\Big) \quad (19)\end{aligned}$$

Setting $Q^1(u, y) = \pi(u, y)$ and $Q^2 = P_G(u)p_M(y|u) = P_{G,M(G)}$, we can rewrite the D_{KL} term in (18) as

$$\begin{aligned}&\big(W_{p,p\sigma^p}\big(P_{G(Z)}, P_Y\big) - W_{p,p\sigma^p}(P_X, P_Y)\big)/p\sigma^p \\ &= \inf_{\pi\in\Pi(P_{G(Z)},P_Y)} D_{KL}\big(\pi\|P_{G(Z),M(G(Z))}\big). \quad (20)\end{aligned}$$

Finally, (19) also shows that the KL-divergence between two joint distributions dominates the KL-divergence between the corresponding the marginals, namely $D_{KL}(\pi\|P_{G,M(G)}) \geq D_{KL}(P_Y\|P_{M(G(Z))})$, so continuing from (20) we get

$$\begin{aligned}&\big(W_{p,p\sigma^p}\big(P_{G(Z)}, P_Y\big) - W_{p,p\sigma^p}(P_X, P_Y)\big)/p\sigma^p \\ &\geq D_{KL}\big(P_Y\|P_{M(G(Z))}\big) = D_{KL}\big(P_Y\|P_{G(Z)+N}\big).\end{aligned}$$ ∎

We next prove Lemma 1. We first show that

$$\begin{aligned}&W_{p,p\sigma^p}\big(P_{G(Z)}, P_Y\big) - W_{p,p\sigma^p}(P_X, P_Y) \\ &\leq \begin{cases} W_2^2\big(P_{G(Z)}, P_X\big) & \text{if } p = 2, \\ p2^{p-1} W_p\big(P_{G(Z)}, P_X\big)\big(\sigma^p + W_p\big(P_{G(Z)}, P_X\big)\big)^{1-1/p} & \text{if } p \geq 1. \end{cases} \quad (21)\end{aligned}$$

We continue from (18): fix some coupling $\pi \in \Pi(P_{G(Z)}, P_X)$ and let the joint distribution of $U = G(Z), X$, and Y be $(U, X, Y) \sim \gamma(u, x, y) = \pi(u, x)f_N(y - x)$, or equivalently, $U - X - Y$ is a Markov chain with $(U, X) \sim \pi$ and $Y = X+N$ with N independent of (X, U). Then $\gamma_{UY}(u, y) \in \Pi(P_{G(Z)}, P_Y)$ and

$$\begin{aligned}\gamma_{Y|U}(y|u) &= \int \gamma(y|x)\gamma(x|u)\, dx \\ &= \int f_N(y - x)\pi(x|u)\, dx = \mathbb{E}_{X\sim\pi_{X|U=u}} f_N(y - X).\end{aligned}$$

Plugging this into (18) gives

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

$$\begin{aligned}&\big(W_{p,p\sigma^p}\big(P_{G(Z)},P_Y\big)-W_{p,p\sigma^p}(P_X,P_Y)\big)/p\sigma^p\\&\le \mathbb{E}_{U\sim P_{G(Z)}}D_{KL}\big(\gamma_{Y|U}(\cdot|U)\|f_N(\cdot-U)\big)\\&=\mathbb{E}_{U\sim P_{G(Z)}}D_{KL}\big(\mathbb{E}_{X\sim\pi_{X|U}(\cdot|U)}f_N(\cdot-X)\|f_N(\cdot-U)\big)\\&\le \mathbb{E}_{U\sim P_{G(Z)}}\mathbb{E}_{X\sim\pi_{X|U}(\cdot|U)}D_{KL}(f_N(\cdot-X)\|f_N(\cdot-U)),\quad(22)\end{aligned}$$

where (22) follows from the convexity of KL-divergence and Jensen's inequality. We can now plug in the definition of KL-divergence leading to

$$\begin{aligned}&\big(W_{p,p\sigma^p}\big(P_{G(Z)},P_Y\big)-W_{p,p\sigma^p}(P_X,P_Y)\big)/p\sigma^p\\&\le \mathbb{E}_{(U,X)\sim\pi}D_{KL}(f_N(\cdot-X)\|f_N(\cdot-U))\\&=\mathbb{E}_{(U,X)\sim\pi}\int f_N(z-X)\log(f_N(z-X)/f_N(z-U))dz\\&=\mathbb{E}_{(U,X)\sim\pi,N\sim f_N}\Big[\|N+X-U\|_p^p-\|N\|_p^p\Big]/p\sigma^p.\quad(23)\end{aligned}$$

In the special case of $p=2$ it follows that

$$W_{p,p\sigma^p}\big(P_{G(Z)},P_Y\big)-W_{p,p\sigma^p}(P_X,P_Y)\le\mathbb{E}_{(U,X)\sim\pi}\Big[\|X-U\|_2^2\Big],$$

and taking the infimum over $\pi\in\Pi(P_X,P_{G(Z)})$ leads to (21). When $p\neq 2, p\ge 1$ we can upper bound the RHS of (23) using the convexity of $f(x)=(x)^p$ for $x\ge 0$, which states that $f(x+\delta)-f(x)\le f'(x+\delta)\delta$ leads to

$$\begin{aligned}&\mathbb{E}_{(U,X)\sim\pi,N\sim f_N}\Big[\|N+X-U\|_p^p-\|N\|_p^p\Big]\\&\le\mathbb{E}_{(U,X)\sim\pi,N\sim f_N}\Big[\big(\|N\|_p+\|X-U\|_p\big)^p-\|N\|_p^p\Big]\\&\le p\mathbb{E}_{(U,X)\sim\pi,N\sim f_N}\Big[\|X-U\|_p\big(\|N\|_p+\|X-U\|_p\big)^{p-1}\Big]\end{aligned}$$

We then use Hölder's inequality $\mathbb{E}[XY]\le(\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^{p/(p-1)})^{1-1/p}$ to get

$$\begin{aligned}&\mathbb{E}_{(U,X)\sim\pi,N\sim f_N}\Big[\|N+X-U\|_p^p-\|N\|_p^p\Big]\\&\le p\mathbb{E}\Big[\|X-U\|_p^p\Big]^{1/p}\mathbb{E}\Big[\big(\|N\|_p+\|X-U\|_p\big)^p\Big]^{1-1/p}\\&\le p2^{p-1}\mathbb{E}\Big[\|X-U\|_p^p\Big]^{1/p}\mathbb{E}\Big[\|N\|_p^p+\|X-U\|_p^p\Big]^{1-1/p}\\&=p2^{p-1}\mathbb{E}\Big[\|X-U\|_p^p\Big]^{1/p}\Big(\sigma^p+\mathbb{E}\Big[\|X-U\|_p^p\Big]\Big)^{1-1/p}\end{aligned}$$

We can now take the infimum over the couplings $\pi\in\Pi(P_X,P_{G(Z)})$ and arrive at (21) for $p\neq 2$ case. By Lemma 2 and (21) choosing $G^*=\arg\min_{G\in\mathcal{G}}W_{p,p\sigma^p}(P_{G(Z)},P_Y)$ and $G^W=\arg\min_{G\in\mathcal{G}}W_p(P_{G(Z)},P_Y)$ we get:

$$\begin{aligned}&p\sigma^pD_{KL}\big(P_Y\|P_{G^*(Z)+N}\big)\\&\le W_{p,p\sigma^p}\big(P_{G^*(Z)},P_Y\big)-W_{p,p\sigma^p}(P_X,P_Y)\\&\le W_{p,p\sigma^p}\big(P_{G^W(Z)},P_Y\big)-W_{p,p\sigma^p}(P_X,P_Y)\\&\le\begin{cases}W_2^2\big(P_{G(Z)},P_X\big) & \text{if } p=2,\\ p2^{p-1}W_p\big(P_{G(Z)},P_X\big)\big(\sigma^p+W_p\big(P_{G(Z)},P_X\big)\big)^{1-1/p} & \text{if } p\ge 1.\end{cases}\end{aligned}$$

B. Proof of Theorem 2

We will be using the dual formulation of entropic optimal transport, so we begin by providing some related results. We denote the dual objective of $S_c(\mu,\nu)$ for probability measures μ,ν of support $\mathrm{supp}(\mu),\mathrm{supp}(\nu)\subseteq\mathbb{R}^d$:

$$\begin{aligned}\Phi(f,g;\mu,\nu)=&\mathbb{E}_{X\sim\mu}f(X)+\mathbb{E}_{Y\sim\nu}g(Y)\\&-\mathbb{E}_{(X,Y)\sim\mu\times\nu}\Big[e^{f(X)+g(Y)-c(X,Y)}\Big]+1.\end{aligned}$$

Here $f:\mathrm{supp}(\mu)\to\mathbb{R}$ and $g:\mathrm{supp}(\nu)\to\mathbb{R}$ are called dual potentials are real-valued functions from the support of μ and ν and $f\in L^1(\mu), g\in L^1(\nu)$, where for a probability measure μ we denote the set of absolutely integratable functions w.r.t. μ as $L^1(\mu)=\{f:\mathrm{supp}(\mu)\to\mathbb{R}|\mathbb{E}_{X\sim\mu}[|f(X)|]<\infty\}$.

The dual function yields a lower bound on the optimal transport: for any $f,g\in L^1(\mu)\times L^1(\nu)$, $S_c(\mu,\nu)\ge\Phi(f,g;\mu,\nu)$. Strong duality is guaranteed to hold by [47, corollary 3.1] (case (B)) for any μ,ν that satisfy $\mathbb{E}_{X,Y\sim\mu\times\nu}[c(X,Y)]<\infty$, which holds for any combination of Q_Y^n,P_Y,P_X and $P_{G(Z)}$ for any $G\in\mathcal{G}$ by the assumption. Strong duality means that

$$S_c(\mu,\nu)=\max_{f,g\in L^1(\mu)\times L^1(\nu)}\Phi(f,g;\mu,\nu).\quad(24)$$

The optimality conditions for the dual problem $\max_{f,g\in L^1(\mu)\times L^1(\nu)}\Phi(f,g;\mu,\nu)$ yield for any $x,y\in\mathbb{R}^d$ (only the values of the dual potentials for $x\in\mathrm{supp}(\mu)$ and $y\in\mathrm{supp}(\nu)$ affect the problem value, but we extend them to $x,y\in\mathbb{R}^d$, this is known as the canonical extension, see [48])

$$f(x)=-\log\mathbb{E}_{Y\sim\mu}e^{g(Y)-c(x,Y)},\ g(y)=-\log\mathbb{E}_{X\sim\nu}e^{f(X)-c(X,y)}\quad(25)$$

Note that the dual potentials are defined up to an additive constant, that is $f(x)+c, g(x)-c$ is also a pair of optimal dual potentials. The optimality conditions yield $\Phi(f,g;\mu,\nu)=\mathbb{E}_{X\sim\mu}f(X)+\mathbb{E}_{Y\sim\nu}g(Y)=S_c(\mu,\nu)$, so we assume the optimal potentials are chosen to have

$$\mathbb{E}_{X\sim\mu}f(X)=\mathbb{E}_{Y\sim\nu}g(Y)=S_c(\mu,\nu)/2\ge 0.\quad(26)$$

We can now proceed to bound the excess risk:

$$\begin{aligned}&S_c\big(P_{G_n(Z)},P_Y\big)-S_c\big(P_{G^*(Z)},P_Y\big)\\&=\underbrace{S_c\big(P_{G_n(Z)},P_Y\big)-S_c\big(P_{G_n(Z)},Q_Y^n\big)}_{A}\\&\quad+\underbrace{S_c\big(P_{G_n(Z)},Q_Y^n\big)-S_c\big(P_{G^*(Z)},P_Y\big)}_{B}\end{aligned}\quad(27)$$

We start by bounding the first term on the RHS in (27) using the dual formulation: denoting the optimal dual potentials for the population optimal transport with generator G_n

$$f_n,g_n=\arg\max_{f,g\in L^1(P_{G_n(Z)})\times L^1(P_Y)}\Phi\big(f,g;P_{G_n(Z)},P_Y\big).$$

Then rewriting (27) in the dual formulation gives

$$\begin{aligned}A=&\Phi\big(f_n,g_n;P_{G_n(Z)},P_Y\big)\\&-\max_{f,g\in L^1(P_{G_n(Z)})\times L^1(Q_Y^n)}\Phi\big(f,g;P_{G_n(Z)},Q_Y^n\big)\\\le&\Phi\big(f_n,g_n;P_{G_n(Z)},P_Y\big)-\Phi\big(f_n,g_n;P_{G_n(Z)},Q_Y^n\big),\end{aligned}$$

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

We can next plug in the definition (24) for Φ to get

$$A \le \mathbb{E}_{Y\sim P_Y} g_n(Y) - \frac{1}{n}\sum_{i=1}^{n} g_n(Y_i) - \mathbb{E}_{Y\sim P_Y}\mathbb{E}_{X\sim P_{G_n(Z)}}\left[e^{f_n(X)+g_n(Y)-c(X,Y)}\right] \quad (28)$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X\sim P_{G_n(Z)}}\left[e^{f_n(X)+g_n(Y_i)-c(X,Y_i)}\right]. \quad (29)$$

Recall the optimality condition (25) for g_n, which asserts that for any $y \in \mathbb{R}^d$: $\mathbb{E}_{X\sim P_{G_n(Z)}}\left[e^{f_n(X)+g_n(Y)-c(X,y)}\right] = 1$, so the last summands in (28) and (29) are equal to -1 and 1 respectively and cancel out, leaving

$$A \le \mathbb{E}_{Y\sim P_Y} g_n(Y) - \frac{1}{n}\sum_{i=1}^{n} g_n(Y_i). \quad (30)$$

Bounding B in (27) is simpler than bounding A because by the optimality of G_n:

$$B \le S_c\left(P_{G^*(Z)}, Q_Y^n\right) - S_c\left(P_{G^*(Z)}, P_Y\right),$$

and now standard results for the sample complexity of entropic optimal transport like Theorem 2 from [33] can be used to bound it since G^* does not depend on the sample. However, the known results will require additional assumptions on the cost function/privatization mechanism, which we would like to avoid, so we proceed by bounding B in a fashion similar to A. Denote

$$\hat{f}^*, \hat{g}^* = \arg\max_{f,g\in L^1(P_{G^*(Z)})\times L^1(Q_Y^n)} \Phi\left(f, g; P_{G_n(Z)}, Q_Y^n\right),$$

the optimality of these potentials and strong duality results in the following bound similar to the one for A:

$$B \le \Phi\left(\hat{f}^*, \hat{g}^*; P_{G^*(Z)}, Q_Y^n\right) - \Phi\left(\hat{f}^*, \hat{g}^*; P_{G_n(Z)}, P_Y\right) \le \frac{1}{n}\sum_{i=1}^{n}\hat{g}^*(Y_i) - \mathbb{E}_{Y\sim P_Y}\hat{g}^*(Y). \quad (31)$$

Bounding the expected values of A and B can now be done in the same way as bounding the excess risk in classic learning problems, which can be achieved through Rademacher complexity.

Definition 1 (Rademacher Complexity): For a family of functions $\mathcal{F}$ and a fixed sample $S = \{Y_i\}_{i=1}^n$ the empirical Rademacher complexity of $\mathcal{F}$ with respect to the sample S is defined as:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Y_i)\right],$$

where the expectation is taken with respect to $\sigma = (\sigma_1, \ldots, \sigma_n)$ with σ_i being independent uniform random variables taking values in $\{\pm 1\}$.

For any integer $n \ge 1$ the Rademacher complexity of $\mathcal{F}$ is the expectation of the empirical Rademacher complexity over all samples of size n:

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{S\sim P_Y^{\otimes n}}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Y_i)\right].$$

Rademacher complexity is one of the key tools to bound suprema of empirical processing, with the following lemma connecting the two, which appears in [49] in the proof of [49, Th. 3.3, eq. (3.13)] (we removed the unused assumptions):

Lemma 3: For a set of functions $\mathcal{F}$ mapping $\mathcal{X}$ to $\mathbb{R}$ and a sample $S = \{Y_i\} \sim P_Y^{\otimes n}$:

$$\mathbb{E}_S \sup_{f\in\mathcal{F}}\left(\mathbb{E}\left[f(Y)\right] - \frac{1}{n}\sum_{i=1}^{n} f(Y_i)\right) \le 2\mathfrak{R}_n(\mathcal{F})$$

Fix $S = \{Y_i\}_{i=1}^n$, applying this lemma to A in (30) gives

$$\mathbb{E}[A] \le \mathbb{E}\left[\mathbb{E}_{Y\sim P_Y} g_n(Y) - \frac{1}{n}\sum_{i=1}^{n} g_n(Y_i)\right] \le \mathbb{E}\sup_{g\in\mathcal{H}_n}\left(\mathbb{E}_{Y\sim P_Y} g(Y) - \frac{1}{n}\sum_{i=1}^{n} g(Y_i)\right) \le 2\mathfrak{R}_n(\mathcal{H}_n),$$

where $\mathcal{H}_n$ is the set of dual potentials g for all the admissible generators $g \in \mathcal{G}$, that is $\mathcal{H}_n = \{g : \mathbb{R}^d \to \mathbb{R} | \exists f, \exists G \in \mathcal{G} : \Phi(f, g; P_{G(Z)}, P_Y) = S_c(P_{G(Z)}, P_Y)\}$. Note that here we again assume that the optimal dual potentials are extended to $\mathbb{R}^d$ and (26) holds. Similarly for B in (31):

$$\mathbb{E}[B] \le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{g}(Y_i) - \mathbb{E}_{Y\sim P_Y}\hat{g}(Y)\right] \le 2\mathfrak{R}_n\left(\hat{\mathcal{H}}\right),$$

where $\hat{\mathcal{H}}$ is the set of dual potentials g for all the admissible generators $g \in \mathcal{G}$ for the empirical problem, that is $\hat{\mathcal{H}} = \{g : \mathbb{R}^d \to \mathbb{R} | \exists f, \exists G \in \mathcal{G} : \Phi(f, g; P_{G(Z)}, Q_Y^n) = S_c(P_{G(Z)}, Q_Y^n)\}$. The excess risk is finally bounded by

$$\mathbb{E}\left[S_c\left(P_{G_n(Z)}, P_Y\right) - S_c\left(P_{G^*(Z)}, P_Y\right)\right] \le 2\mathfrak{R}_n(\mathcal{H}_n) + 2\mathfrak{R}_n(\hat{\mathcal{H}}) \quad (32)$$

The rest of the proof bounds the Rademacher complexities using the properties of the dual potentials. We start by bounding the optimal dual potential from $\mathcal{H}_n$, the bound for $\hat{\mathcal{H}}$ is identical.

Let f, g be the maximizes of $\Phi(f, g; P_{G(Z)}, P_Y)$. The optimality conditions (25) and our convention (26) together with Jensen's inequality for $-\log(x)$ yield for any $x, y \in \mathbb{R}^d$:

$$f(x) = -\log\mathbb{E}_{Y\sim P_Y}\left[e^{g(Y)-c(x,Y)}\right] \le \mathbb{E}_{Y\sim P_Y}\left[c(x,Y) - g(Y)\right] \le \mathbb{E}_{Y\sim P_Y}[c(x,Y)] \quad (33)$$

$$g(y) = -\log\mathbb{E}_{X\sim P_{G(Z)}}\left[e^{f(X)-c(X,y)}\right] \le \mathbb{E}_{X\sim P_{G(Z)}}\left[c(X,y) - f(X)\right] \le \mathbb{E}_{X\sim P_{G(Z)}}\left[c(X,y)\right] \quad (34)$$

Note that for any function $h(y)$ it holds that $\hat{\mathfrak{R}}_n(\mathcal{H}_n) = \hat{\mathfrak{R}}_n(\mathcal{H}_n \oplus h(y))$, namely, adding or subtracting a specific function from all the functions in a set does not change the Rademacher complexity, so let $h(y) = \sup_{x\in\mathcal{X}} c(x, y)$ and $u(y) = e^{h(y)-g(y)}$. By the upper bound on $g(y)$ (34):

$$u(y) \ge e^{\sup_{x\in\mathcal{X}} c(x,y) - \mathbb{E}_{X\sim P_{G(Z)}}[c(X,y)]} \ge 1,$$

so the function $f(x) = -\log x$ is 1-Lipschitz on the range of u. By Talagrand's lemma [49, Lemma 5.7], composition with a 1-Lipschitz function cannot increase the Rademacher complexity

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

of a function set. Denoting $\mathcal{U}_n = \{u(y) = e^{\sup_{x\in\mathcal{X}} c(x,y)-g(y)} | g \in \mathcal{H}_n\}$ we arrive at

$$\hat{\mathfrak{R}}_n(\mathcal{H}_n) = \hat{\mathfrak{R}}_n((-\log)\circ\mathcal{U} \oplus h(y)) \le \hat{\mathfrak{R}}_S(\mathcal{U}).$$

To further bound the Rademacher complexity of $\mathcal{U}$ we use the following result for positive definite symmetric kernels that is a direct consequence of Mercer's theorem [50].

Theorem 4 ([49, Ths. 6.8 and 6.12]): Let $\mathcal{Z} \subset \mathbb{R}^d$ and $K : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$ be a positive definite symmetric kernel. Then there exists a Hilbert space $\mathbb{H}$ and a mapping $\Phi : \mathcal{Z} \to \mathbb{H}$ such that: $\forall x, x' \in \mathcal{X}, \quad K(x, x') = \langle\Phi(x), \Phi(x')\rangle$. $\mathbb{H}$ is called a reproducing kernel Hilbert space (RKHS) associated to K. and let $\Phi : \mathcal{Z} \to \mathbb{H}$ be a feature mapping associated to k. Let $S = \{z_i\}_{i=1}^n \subset \mathcal{Z}$ be a sample of size n, and let $\mathcal{H} = \{x \mapsto \langle \mathbf{w}, \Phi(x)\rangle : \|\mathbf{w}\|_{\mathbb{H}} \le \Lambda\}$ for some $\Lambda \ge 0$. Then

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \le \Lambda\left(\sum_{i=1}^n K(z_i, z_i)\right)^{1/2}/n$$

First, we check the conditions of the theorem: fix some $u \in \mathcal{U}$, then there exists a $G \in \mathcal{G}$ and a pair f, g of optimal dual potentials maximizing $\Phi(f, g; P_{G(Z),P_Y})$ such that

$$u(y) = e^{\sup_{x\in\mathcal{X}} c(x,y)-g(y)} = e^{\sup_{x\in\mathcal{X}} c(x,y)}\mathbb{E}_{X\sim P_{G(Z)}} e^{f(X)-c(X,y)}.$$

To simplify the notation denote $v(y) = e^{\sup_{x\in\mathcal{X}} c(x,y)}$ and note that $K(x, y) = v(x)v(y)e^{-c(x,y)}$ is a positive definite symmetric kernel (as a product of two kernels), so applying Theorem 4 to kernel K leads to

$$\begin{aligned} u(y) &= \mathbb{E}_{X\sim P_{G(Z)}}\left[K(X, y)e^{f(X)}/v(X)\right] \\ &= \left\langle \mathbb{E}_{X\sim P_{G(Z)}}\left[\Phi(X)e^{f(X)}/v(X)\right], \Phi(y)\right\rangle = \langle w, \Phi(y)\rangle. \end{aligned}$$

So $u(y)$ is indeed a linear function in RKHS and its associated norm is

$$\begin{aligned} \|w\|_H^2 &= \mathbb{E}_{X,X'\sim P^2_{G(Z)}} e^{f(X')+f(X)-c(X,X')} \\ &\le \mathbb{E}_{X,X'\sim P^2_{G(Z)}} e^{\mathbb{E}_{Y\sim P_Y}[c(X',Y)+c(X,Y)-c(X,X')]} \\ &\le e^{2\sup_{x\in\mathcal{X}} \mathbb{E}_{Y\sim P_Y}[c(x,Y)]}, \end{aligned}$$

where we used the upper bound on $f(x)$ given in (34). Combining this with Theorem 4 leads to

$$\begin{aligned} \hat{\mathfrak{R}}_S(\mathcal{H}_n) \le \hat{\mathfrak{R}}_S(\mathcal{U}) &\le e^{\sup_{x\in\mathcal{X}} \mathbb{E}_{Y\sim P_Y}[c(x,Y)]}\frac{\sqrt{\sum_{i=1}^n v(Y_i)^2}}{n} \\ &= e^{\sup_{x\in\mathcal{X}} \mathbb{E}_{Y\sim P_Y}[c(x,Y)]}\frac{\sqrt{\sum_{i=1}^n e^{2\sup_{x\in\mathcal{X}} c(x,Y_i)}}}{n} \end{aligned}$$

To get the bound for the Rademacher complexity we take the expectation of both sides and apply the Jensen's inequality, which leads to

$$\begin{aligned} \mathfrak{R}_n(\mathcal{H}_n) &\le e^{\sup_{x\in\mathcal{X}} \mathbb{E}_{Y\sim P_Y}[c(x,Y)]}\sqrt{\mathbb{E}e^{2\sup_{x\in\mathcal{X}} c(x,Y)}/n} \\ &\le \mathbb{E}e^{2\sup_{x\in\mathcal{X}} c(x,Y)}/\sqrt{n} \end{aligned} \tag{35}$$

The derivation for $\hat{\mathfrak{R}}_S(\hat{\mathcal{H}})$ follows the same lines with the only change being the use of Q_Y^n instead of P_Y leading to

$$\hat{\mathfrak{R}}_S(\hat{\mathcal{H}}) \le e^{\sup_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n c(x,Y_i)}\left(\sum_{i=1}^n e^{2\sup_{x\in\mathcal{X}} c(x,Y_i)}\right)^{1/2}/n.$$

Taking the expectation of both sides and applying the Cauchy-Schwartz inequality leads to

$$\begin{aligned} \mathfrak{R}_n(\hat{\mathcal{H}}) &\le \left(\mathbb{E}\Big[\sup_{x\in\mathcal{X}} e^{\frac{2}{n}\sum_{i=1}^n c(x,Y_i)}\Big]\mathbb{E}\Big[e^{2\sup_{x\in\mathcal{X}} c(x,Y)}\Big]/n\right)^{1/2} \\ &\le \mathbb{E}e^{2\sup_{x\in\mathcal{X}} c(x,Y)}/\sqrt{n} \end{aligned}$$

which combined with (35) and (32) gives

$$\mathbb{E}\big[S_c\big(P_{G_n(Z)}, P_Y\big) - S_c\big(P_{G^*(Z)}, P_Y\big)\big] \le 4\mathbb{E}e^{2\sup_{x\in\mathcal{X}} c(x,Y)}/\sqrt{n}$$

C. Proof of Theorem 3

Theorem 3 follows from [7, Th. 6].

Theorem 5 [7, Th. 6]: Let P_Z and P_Y be sub-Gaussian and the set of generators $\mathcal{G}$ consist of L-Lipschitz functions, i.e., $\|G(Z_1) - G(Z_2)\| \le L\|Z_1 - Z_2\|$ for any Z_1, Z_2 in the support of P_Z and let $\mathcal{G}$ satisfy (14). Then for $\tau^2 = \max\{L^2\sigma^2(Z), \sigma^2(Y)\}$ the generalization error for entropic GAN with $p = 2$ (9) can be upper bounded as

$$\begin{aligned} &\mathbb{E}\big[W^2_{2,\lambda}\big(P_{G^n(Z)}, P_Y\big) - W^2_{2,\lambda}\big(P_{G^*(Z)}, P_Y\big)\big] \\ &\quad\le C_d\lambda n^{-1/2}\left(1 + \left(2\tau^2/\lambda\right)^{\lceil 5d/4\rceil+3}\right). \end{aligned}$$

In our case $\lambda = 2\sigma^2$ and $Y = X + N$ with $N \sim \mathcal{N}(0, \sigma^2 I)$, so $\sigma(Y) \le \sigma(X) + \sigma(N)$, where $\sigma(N) = \sigma^2$. Thus, plugging it into the theorem we get

$$\begin{aligned} &\mathbb{E}\big[W^2_{2,\lambda}\big(P_{G^n(Z)}, P_Y\big) - W^2_{2,\lambda}\big(P_{G^*(Z)}, P_Y\big)\big] \\ &\quad\le C_d\sigma^2 n^{-1/2}\left(1 + \left(\max\{L\sigma(Z), \sigma(X) + \sigma\}/\sigma\right)^{2\left\lceil\frac{5d}{4}\right\rceil+6}\right) \end{aligned}$$

Letting $\tau = \max\{L\sigma(Z)/\sigma(X), 1\}$ we get $(\sigma(X) + \sigma)\tau \ge \max\{L\sigma(Z), \sigma(X) + \sigma\}$, which leads to

$$\begin{aligned} &\mathbb{E}\big[W^2_{2,\lambda}\big(P_{G^n(Z)}, P_Y\big) - W^2_{2,\lambda}\big(P_{G^*(Z)}, P_Y\big)\big] \\ &\quad\le C_d\sigma^2 n^{-1/2}\left(1 + \left(\tau^2\big(1 + \sigma(X)/\sigma\big)^2\right)^{\left\lceil\frac{5d}{4}\right\rceil+3}\right). \end{aligned}$$

VI. Discussion and Conclusion

We have proposed and analyzed a new framework for locally differentially private training of GANs. Our analysis indicates that the addition of mutual information to the objective of the optimal transport GAN can act as a deconvolution operator provided the right choice of the cost function. The method not only recovers the original distribution in the population setting but also converges at a parametric rate and can be easily combined with non-privatized training methods as a black box in practice since the modifications do not influence the training process. We believe understanding how to train ML models from privatized data and improving the privacy/utility trade-offs is of paramount importance for the future of privacy-preserving machine learning.

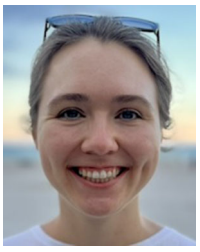
References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr.*, 2006, pp. 265–284.

[2] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, 2011.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

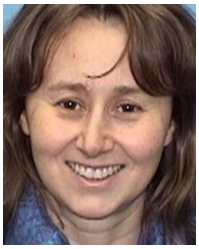
[3] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your GAN: Interactive point-based manipulation on the generative image manifold," in *Proc. ACM Conf. Proc.*, 2023, pp. 1–11.
[4] W. W. Booker, D. D. Ray, and D. R. Schrider, "This population does not exist: Learning the distribution of evolutionary histories with generative adversarial networks," *Genetics*, vol. 224, no. 2, 2023, Art. no. iyad063.
[5] E. R. Chan et al., "Efficient geometry-aware 3D generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16123–16133.
[6] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 2292–2300.
[7] D. Reshetova, Y. Bai, X. Wu, and A. Özgür, "Understanding entropic regularization in GANs," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2021, pp. 825–830.
[8] G. Mena and J. Niles-Weed, "Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4541–4551.
[9] S. Feizi, F. Farnia, T. Ginart, and D. Tse, "Understanding GANs in the LQG setting: Formulation, generalization and stability," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 304–311, May 2020.
[10] T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1999.
[11] R. Bassily, K. Nissim, U. Stemmer, and A. G. Thakurta, "Practical locally private heavy hitters," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.
[12] M. Bun, J. Nelson, and U. Stemmer, "Heavy hitters and the structure of local privacy," *ACM Trans. Algorithms*, vol. 15, no. 4, pp. 1–40, 2019.
[13] W.-N. Chen, P. Kairouz, and A. Ozgur, "Breaking the dimension dependence in sparse distribution estimation under communication constraints," in *Proc. 34th Annu. Conf. Learn. Theory*, 2021, pp. 1028–1059.
[14] W.-N. Chen, P. Kairouz, and A. Ozgur, "Breaking the communication-privacy-accuracy trilemma," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3312–3324.
[15] A. T. Suresh, X. Y. Felix, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3329–3337.
[16] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018, *arXiv:1812.00984*.
[17] Y. Han, P. Mukherjee, A. Ozgur, and T. Weissman, "Distributed statistical estimation of high-dimensional and nonparametric distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2018, pp. 506–510.
[18] D. Chen, T. Orekondy, and M. Fritz, "GS-WGAN: A gradient-sanitized approach for learning differentially private generators," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12673–12684.
[19] T. Cao, A. Bie, A. Vahdat, S. Fidler, and K. Kreis, "Don't generate me: Training differentially private generative models with sinkhorn divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12480–12492.
[20] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," 2018, *arXiv:1802.06739*.
[21] X. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model (technical report)," 2018, *arXiv:1801.01594*.
[22] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*.
[23] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3481–3490.
[24] A. Mansbridge, G. Barbour, D. Piras, C. Frye, I. Feige, and D. Barber, "Learning to noise: Application-agnostic data sharing with local differential privacy," 2020, *arXiv:2010.12464*.
[25] G. Peyré and M. Cuturi, "Computational optimal transport: With applications to data science," *Found. Trends® Mach. Learn.*, vol. 11, nos. 5–6, pp. 355–607, 2019.
[26] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré, "Sample complexity of sinkhorn divergences," in *Proc. 22nd Int. Conf. Artif. Intell. Stat.*, 2019, pp. 1574–1583.
[27] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptogr. Techn.*, 2006, pp. 486–503.
[28] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends® Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.
[29] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[30] A. Korotin, V. Egiazarian, A. Asadulaev, A. Safin, and E. Burnaev, "Wasserstein-2 generative networks," 2019, *arXiv:1909.13082*.
[31] P. Rigollet and J. Weed, "Entropic optimal transport is maximum-likelihood deconvolution," *Comptes Rendus Mathematique*, vol. 356, no. 11, pp. 1228–1235, 2018.
[32] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy, "Convergence of smoothed empirical measures with applications to entropy estimation," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4368–4391, Jul. 2020.
[33] A. J. Stromme, "Minimum intrinsic dimension scaling for entropic optimal transport," 2023, *arXiv:2306.03398*.
[34] G. Luise, M. Pontil, and C. Ciliberto, "Generalization properties of optimal transport GANs with latent distribution learning," 2020, *arXiv:2007.14641*.
[35] A. Genevay, G. Peyré, and M. Cuturi, "Learning generative models with sinkhorn divergences," in *Proc. Int. Conf. Artif. Intell. Stati.*, 2018, pp. 1608–1617.
[36] R. Flamary et al., "POT: Python optimal transport," *J. Mach. Learn. Res.*, vol. 22, no. 78, pp. 1–8, 2021.
[37] Y. LeCun. "The MNIST database of handwritten digits," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/
[38] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[39] V. Feldman, A. McMillan, and K. Talwar, "Stronger privacy amplification by shuffling for rényi and approximate differential privacy," in *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2023, pp. 4966–4981.
[40] S. Mallat, *A Wavelet Tour of Signal Processing*. Amsterdam, The Netherlands: Elsevier, 1999.
[41] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2Void-learning denoising from single noisy images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2129–2137.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[43] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
[44] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5769–5779.
[45] A. Bie, G. Kamath, and G. Zhang, "Private GANs, revisited," 2023, *arXiv:2302.02936*.
[46] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6629–6640.
[47] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.
[48] A.-A. Pooladian and J. Niles-Weed, "Entropic estimation of optimal transport maps," 2021, *arXiv:2109.12004*.
[49] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.
[50] J. Mercer, "XVI. Functions of positive and negative type, and their connection the theory of integral equations," *Philos. Trans. Royal Soc. London. Ser. A, Contain. Papers Math. Phys. Character*, vol. 209, nos. 441–458, pp. 415–446, 1909.



Daria Reshetova received the B.Sc. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology in 2016, and the Ph.D. degree in electrical engineering from Stanford University in 2023. Her research interests include statistical learning theory, information theory, generative models, and differential privacy. She was a recipient of the Stanford Graduate Fellowship.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.

Wei-Ning Chen (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering and mathematics and the M.S. degree in communication engineering from National Taiwan University in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Stanford University. His research interests include information theory, statistics, and theoretical machine learning, with applications in differential privacy and federated learning. He was a recipient of the Stanford Graduate Fellowship.



Ayfer Özgür (Senior Member, IEEE) is currently an Associate Professor with the Department of Electrical Engineering, Stanford University, where she is the Chambers Faculty Scholar with the School of Engineering. Her research interests include information theory, wireless communications, statistics, and machine learning. She received the EPFL Best Ph.D. Thesis Award in 2010, the NSF CAREER Award in 2013, the Okawa Foundation Research Grant, the Faculty Research Awards from Google and Facebook, and the IEEE Communication Theory Technical Committee Early Achievement Award in 2018. She was selected as the Inaugural Goldsmith Lecturer of the IEEE ITSoc in 2020.

Authorized licensed use limited to: Stanford University Libraries. Downloaded on February 03,2025 at 18:44:39 UTC from IEEE Xplore. Restrictions apply.