Statistically Optimal Generative Modeling with Maximum Deviation from the Empirical Distribution

Elen Vardanyan [1 *] **Sona Hunanyan** [1 *] **Tigran Galstyan** [1 2 *] **Arshak Minasyan** [3] **Arnak Dalalyan** [3]

Abstract

This paper explores the problem of generative modeling, aiming to simulate diverse examples from an unknown distribution based on observed examples. While recent studies have focused on quantifying the statistical precision of popular algorithms, there is a lack of mathematical evaluation regarding the non-replication of observed examples and the creativity of the generative model. We present theoretical insights into this aspect, demonstrating that the Wasserstein GAN, constrained to left-invertible push-forward maps, generates distributions that not only avoid replication but also significantly deviate from the empirical distribution. Importantly, we show that left-invertibility achieves this without compromising the statistical optimality of the resulting generator. Our most important contribution provides a finite-sample lower bound on the Wasserstein-1 distance between the generative distribution and the empirical one. We also establish a finite-sample upper bound on the distance between the generative distribution and the true data-generating one. Both bounds are explicit and show the impact of key parameters such as sample size, dimensions of the ambient and latent spaces, noise level, and smoothness measured by the Lipschitz constant.

1. Introduction

Generative modeling is a widely-used machine learning technique that has found applications in various scientific and industrial domains, including health (Yan et al., 2018; Nie et al., 2017), climate (Gagne et al., 2020), finance (Wiese et al., 2020), energy (Fekri et al., 2019), physics (Paganini et al., 2018), chemistry (Maziarka et al., 2020), and biology (Repecka et al., 2021). The primary goal of generative models is to simulate new examples by learning from training data, while ensuring diversity and avoiding the replication of examples from the training set.

Assessing the performance of a generative model can be done qualitatively by evaluating the realism of the generated examples, which we refer to as accuracy. However, accuracy should be balanced with another crucial property: the diversity of generated examples and their difference from the training examples. This property, referred to as the generator's creativity (Li et al., 2024), is essential to avoid overfitting (producing examples that are slight modifications of those in the training set). Qualitative evaluation of diversity is challenging due to the large size of training sets, making it impossible to retain all the examples they contain. Nonetheless, diversity is as important as accuracy, particularly in applications where generative models aim to enrich datasets in cases where data acquisition is expensive or infeasible. Generating examples that closely resemble the observed data diminishes the utility of such algorithms.

The success of deep neural nets in generative modeling has attracted significant attention from the machine learning community. The number of proposed methods in recent years, following the influential work by (Goodfellow et al., 2014), is extensive, making it impractical to cite all of them here[1]. While many of these methods have been empirically validated and justified using heuristics, a more comprehensive mathematical quantification of their strengths and limitations is often lacking.

The importance of diversity in generated examples is well-acknowledged but presents practical challenges. Numerous papers have empirically studied the issue of limited diversity in learned distributions, proposing some solutions (Dumoulin et al., 2017; Arora et al., 2017; 2018; Srivastava et al., 2017). However, the majority of studies on diversity have primarily targeted the mitigation of mode collapse. This phenomenon occurs when certain examples in the train-

[*]Equal contribution [1]Department of Mathematics, Yerevan State University (YSU), Armenia [2]YerevaNN, Armenia [3]CREST, GENES, Institut Polytechnique de Paris, France. Correspondence to: Elen Vardanyan <evardanyan@aua.am>, Arnak Dalalyan <arnak.dalalyan@ensae.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]For a comprehensive list, refer to https://github.com/hindupuravinash/the-gan-zoo

ing/testing set are overlooked by the learned distribution; as a result, some test set examples are markedly dissimilar to those generated by the learned distribution.

In this paper, we explore another aspect of diversity within the learned distribution. We aim to ensure that the examples it generates are not mere copies of the samples in the training set but rather novel instances. To achieve this, we seek to understand to what extent the learned distribution can deviate from the empirical distribution of observations while still closely adhering to the true underlying law.

From a theoretical standpoint, endeavors to address "mode collapse" primarily involve proposing methodological enhancements that aim to achieve a better precision, where the precision is understood as the closeness of the learned distribution (in Wasserstein or KL, for instance) to the true underlying law. The idea is that if the learned distribution closely aligns with the true distribution in these metrics, it is less likely to miss important modes of the true distribution. However, none of these approaches quantifies the dissimilarity between the learned distribution and the empirical distribution of the training set. This is a critical consideration, as the training process typically involves fitting the empirical distribution with a parametric class. The ultimate goal, however, is to generate examples that differ from those in the training set.

Contributions The key insight of this paper is the following: if the learned distribution is defined as the push-forward of the uniform distribution by a smooth map, then ensuring the push-forward map has a smooth left inverse prevents overfitting to the empirical distribution and promotes creativity. Moreover, when left invertibility is imposed on an already statistically optimal generator, it seems to retain its optimality. This claim, though speculative, is supported by our study of Wasserstein GANs (WGAN) (Arjovsky et al., 2017; Gulrajani et al., 2017). We introduce LIPERM (Left-Inverse Penalized Empirical Risk Minimizer), a penalized version of WGAN that favors left invertibility of the push-forward map. Our main result establishes a lower bound on the Wasserstein-1 distance between the learned distribution and its empirical counterpart. We then establish an upper bound on the precision measured by an integral probability metrics. It takes the form of a finite-sample risk bound describing the behavior of the learned generator as a function of the sample size n and the dimension d of the latent space. Importantly, these bounds are independent of the ambient dimension D and are rate-optimal, as confirmed by lower bounds in (Schreuder et al., 2021; Tang and Yang, 2023).

When the latent dimension $d \geqslant 2$, we prove that LIPERM's separation from any distribution concentrated on the training sample is at least of order $n^{-1/d}$, while its precision is established to be at most of order $n^{-1/d}$. Notably, $n^{-1/d}$ is the rate of approximating the true data-generating distribution with its empirical counterpart (Dudley, 1969; Boissard and Gouic, 2014; Niles-Weed and Rigollet, 2022). Consequently, a generative model separated from the empirical distribution by a distance larger than $n^{-1/d}$ would have suboptimal precision. Thus, LIPERM, being sufficiently distant from distributions replicating observed examples, ensures diverse and novel example generation.

In Figure 1, we illustrate the explored framework on the task of generating points on a 2D spiral using a 1D latent space. In the left panel, the map $g : [0, 1] \to \mathbb{R}^2$ corresponds to a generator with high diversity and a small left-inverse penalty (LIP); the generated examples exhibit a wide range of variations and deviations from the training distribution. In the second panel, the map g is inferior to the first one in terms of diversity, resulting in less varied generated examples, and it has a higher LIP. A closer examination reveals that for certain points on the spiral that are close to each other, their preimages under g are located far apart. This means that the left-inverse of g has a large Lipschitz norm. Finally, in the rightmost panel, g generates only a few examples, producing a restricted set of outputs. It has a LIP equal to $+\infty$. This illustration highlights the connection between diversity and the left-inverse penalty.

Prior work In recent years, a notable surge in papers has focused on mathematical aspects of generative models. Some treat generative models as tools for distribution and density estimation, establishing their optimality (Liang, 2021; Belomestny et al., 2023; Biau et al., 2021; 2020; Uppal et al., 2019; Kwon and Chae, 2024; Chae et al., 2023). The convergence rate of adversarial generative models—under the manifold assumption—independent of the ambient space dimension, is highlighted in (Huang et al., 2022; Schreuder et al., 2021; Tang and Yang, 2023; Stéphanovitch et al., 2023) using integral probability metrics. The intricate relationship between minimax optimality and distribution learning is explored in (Chen et al., 2022).

The analysis of diffusion-based generative models is presented in (De Bortoli et al., 2021; Bortoli et al., 2022), while the investigation of autoencoders and their relation to the Langevin process is carried out in (Block et al., 2020). A regularization scheme for training GANs, based on adding a penalty on the weighted gradient-norm of the discriminator, is introduced in (Roth et al., 2017); see also (Petzka et al., 2018). The theoretical characterization of the mode-seeking behavior of general f-divergences and Wasserstein distances, along with a guarantee for mixtures, is provided in (Li and Farnia, 2023). In the context of generator invertibility and mode collapse in GANs, (Bai et al., 2018) suggest that invertible generators might effectively alleviate mode collapse. (Xi and Bloem-Reddy, 2023) propose a theoretical framework for analyzing the indeterminacies of latent variable models.

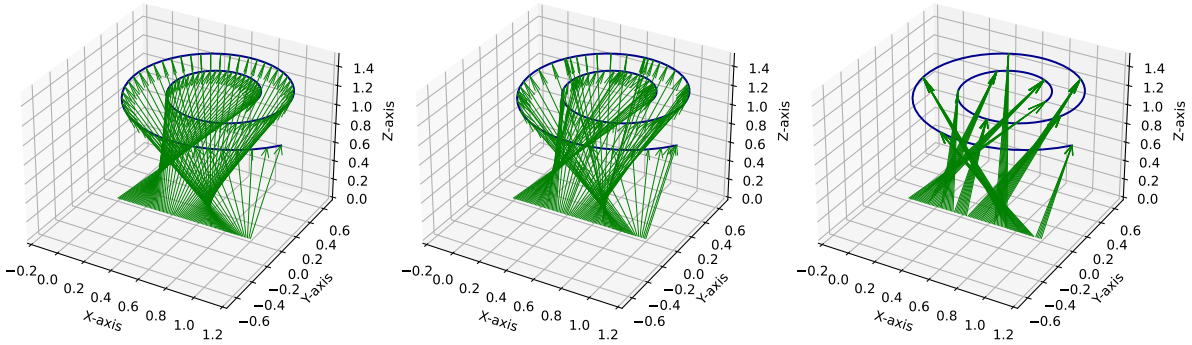


Figure 1. Illustration of the framework of this paper: generating points on a 2D spiral using a 1D latent space. The green arrows represent the mapping $g : [0, 1] \to \mathbb{R}^2$. Each arrow indicates how points from the latent space are mapped to positions in the 2D spiral.

The issue of generative models memorizing training data and potentially copying from it has been emphasized in early works such as (Nagarajan et al., 2018) and (Gulrajani et al., 2019). Recent observations have further revealed instances of duplicated training examples, particularly in diffusion models for image generation and large language models (Somepalli et al., 2023a;b; Daras et al., 2023; Carlini et al., 2021; 2023; Jagielski et al., 2023). Advancing theoretical understanding in this realm could contribute to the development of algorithms mitigating the risk of replication.

Notation For every integer $d > 1$, we denote by $\mathcal{U}_d$ the uniform distribution on $[0, 1]^d$. The norm $\|\boldsymbol{x}\|$ of an element $\boldsymbol{x}$ from an Euclidean space is always the Euclidean norm. We denote by $\mathbb{E}[\boldsymbol{X}]$ the expectation of a random variable. If necessary, we write $\mathbb{E}_P[\boldsymbol{X}]$ to stress that the expectation is considered under the condition that $\boldsymbol{X}$ is drawn from P. For a random vector $\boldsymbol{X}$ and a real number $q \geqslant 1$, we use the notation $\|\boldsymbol{X}\|_{\mathbb{L}_q} = \mathbb{E}^{1/q}[\|\boldsymbol{X}\|^q]$. For two subsets A and B of some Euclidean spaces, and a positive number L, we say that a function $f : A \to B$ is L-Lipschitz-continuous, if $\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| \leqslant L\|\boldsymbol{x} - \boldsymbol{x}'\|$ for every $\boldsymbol{x}, \boldsymbol{x}' \in A$. The set of all the L-Lipschitz continuous functions from A to B is denoted by $\mathrm{Lip}_L(A \to B)$. The Dirac mass at a point $\boldsymbol{x}$ is denoted by $\delta_{\boldsymbol{x}}$. The notation $P_{n,Z}$ is often used to design the empirical distribution $(1/n)\sum_{i=1}^n \delta_{\boldsymbol{Z}_i}$ (Z can be replaced by other letters). We set $\mathsf{V}_d = \pi^{d/2}/\Gamma(1 + d/2)$ to be the volume of the unit ball. Notation Id_d stands for the identity mapping $\mathsf{Id}_d(\boldsymbol{x}) = \boldsymbol{x}$ on $\mathbb{R}^d$, or any subset of it. For two sets of functions $\mathcal{G}$ and $\mathcal{H}$, we set $\mathcal{G}_{\mathcal{H}}$ the subset of $\mathcal{G}$ the elements of which admit a left inverse function in $\mathcal{H}$.

2. Left-Inverse-Penalized Empirical Risk

Let P^* be the distribution of training data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ in the D-dimensional Euclidean space $\mathbb{R}^D$ equipped with the Borel σ-algebra $\mathscr{B}(\mathbb{R}^D)$. While D is typically large, we assume that the examples $\boldsymbol{X}_i$ may originate from latent variables in a lower-dimensional space of dimension d. This is closely related to the so-called "manifold assumption" (Fefferman et al., 2016; Narayanan and Mitter, 2010). To capture this, several popular algorithms such as GAN or WGAN, choose an integer $d > 0$ much smaller than D and seek to learn a generative distribution $\widehat{P}_n$ as a smooth transformation of the uniform distribution in $[0, 1]^d$. This is the setting considered in this paper: the trained generative distribution is chosen of the form $P_g = g\sharp\mathcal{U}_d$, where $g : [0, 1]^d \to \mathbb{R}^D$ is called the push-forward map and P_g defined by $P_g(A) = P(g^{-1}(A))$, $\forall A \in \mathscr{B}(\mathbb{R}^D)$, is the push-forward distribution.

In this paper, we study learned distributions obtained by minimizing the penalized empirical risk using a suitable penalty. Let $\mathcal{H}$ and $\mathcal{G}$ be two functional classes such that

$$\mathcal{H} \subset \mathrm{Lip}_{L_{\mathcal{H}}}([0, 1]^D \to [0, 1]^d),$$
$$\mathcal{G} \subset \mathrm{Lip}_{L_{\mathcal{G}}}([0, 1]^d \to [0, 1]^D).$$

Let d be a distance on the space of probability distributions. For $q \geqslant 1$, we define the left inverse penalty

$$\begin{aligned}\mathrm{pen}_{\mathcal{H}}(g) &= \min_{h\in\mathcal{H}} \int_{[0,1]^d} \|h \circ g(\boldsymbol{u}) - \boldsymbol{u}\|^q \mathrm{d}\boldsymbol{u} \\ &= \min_{h\in\mathcal{H}} \|h \circ g - \mathsf{Id}_d\|^q_{\mathbb{L}_q}. \end{aligned} \tag{LIP}$$

We then define the penalized empirical risk

$$\widehat{L}_n^{\mathsf{d},\mathcal{H}}(g) = \mathsf{d}(g\sharp\mathcal{U}_d, P_{n,X}) + \lambda\,\mathrm{pen}_{\mathcal{H}}(g) \tag{1}$$

for a tuning parameter $\lambda > 0$. The learned generator is the push-forward distribution $\widehat{g}_n\sharp\mathcal{U}_d$, with $\widehat{g}_n = \widehat{g}_n(\lambda, \mathcal{G}, \mathsf{d}, \mathcal{H})$ being a solution to the minimization problem

$$\widehat{g}_n \in \operatorname*{arg\,min}_{g\in\mathcal{G}} \widehat{L}_n^{\mathsf{d},\mathcal{H}}(g). \tag{LIPERM}$$

The choices of the distance d and of λ are important. For d we will mainly use the Wasserstein-1 distance W_1 or a more general integral probability metric (IPM) defined by

$$\mathsf{d}_{\mathcal{F}}(P, Q) = \sup_{f\in\mathcal{F}} \big|\mathbb{E}_P[f(\boldsymbol{X})] - \mathbb{E}_Q[f(\boldsymbol{X})]\big|, \tag{IPM}$$

where $\mathcal{F}$ is a set of test functions $\mathbb{R}^D \to \mathbb{R}$, and P, Q are two probability measures on $([0,1]^D, \mathscr{B}([0,1]^D))$. The Wasserstein-1 distance corresponds to an IPM with $\mathcal{F}$ being the set of all 1-Lipschitz-continuous functions.

As for λ, the case $\lambda = 0$ corresponds to the standard setting of minimal distance estimator including GAN, WGAN, and their variants. The other extreme case $\lambda = +\infty$ corresponds to minimizing the training error $\mathsf{d}(g\sharp\mathcal{U}_d, P_{n,X})$ under the constraint $g \in \mathcal{G}_{\mathcal{H}}$. Although this constrained estimator has some attractive properties, its computation might be more challenging than that of penalized one.

Note that the sets $\mathcal{G}$ and $\mathcal{H}$ are entirely determined by the user's choice, as are d and λ. We refer to $\widehat{g}_n$ defined by (LIPERM) as the left-inverse-penalized empirical risk minimizer. Although most of our results will refer to the general form (1) of the loss function, it might be useful for the reader to keep in mind the central example corresponding to the case $\mathsf{d} = \mathsf{W}_1$ and $q = 2$ leading to the min-max problem

$$\min_{g\in\mathcal{G}} \max_{h\in\mathcal{H}} \Big\{ \mathsf{W}_1(g\sharp\mathcal{U}_d, P_{n,X}) + \lambda \int \|h\circ g(\boldsymbol{u}) - \boldsymbol{u}\|^2 \, \boldsymbol{u} \Big\}.$$

The rationale behind considering this learning procedure is as follows: When training a generative model, the objective is to produce a distribution that (a) is easy to sample from, (b) is close to the distribution of the training data, and (c) avoids replicating the examples in the training dataset. The cost function in (LIPERM) consists of two terms, each contributing to one of the last two desired properties. If the term $\mathsf{d}(g\sharp\mathcal{U}_d, P_{n,X})$ is small, the generator $g\sharp\mathcal{U}_d$ closely approximates the empirical distribution. Additionally, if $\mathrm{pen}_{\mathcal{H}}(g)$ is small, g is nearly invertible with a smooth inverse. Intuitively, when g possesses a smooth inverse, it disperses the unit hypercube $[0,1]^d$ across a large region in $[0,1]^D$ rather than concentrating around a small neighborhood of the observations from the training set. In particular, the following simple fact holds true.

Lemma 1. *Any distribution $P_g = g\sharp\mathcal{U}_d$ defined by a push-forward map $g \in \mathcal{G}_{\mathcal{H}}$ has no atom. In particular, it satisfies $P_g(\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}) = 0$.*

This lemma implies, in particular, that if the learned distribution is defined by LIPERM with $\lambda = +\infty$, then the probability of generating an example that was present in the training set is equal to zero.

Proof of Lemma 1. Clearly, for any $\boldsymbol{x} \in \mathbb{R}^D$, $\widehat{P}_n(\{\boldsymbol{x}\}) = \mathrm{Leb}_d\big(g^{-1}(\boldsymbol{x})\big)$. Since g has a left inverse, $g^{-1}(\boldsymbol{x})$ contains at most 1 point, resulting in a Lebesgue measure of zero. □

Motivated by practical convenience, instead of an exact solution of LIPERM, we will consider an ε-approximate solution satisfying

$$\widehat{L}_n^{\mathsf{d},\mathcal{H}}(\widehat{g}_{n,\varepsilon}) \leqslant \min_{g\in\mathcal{G}} \widehat{L}_n^{\mathsf{d},\mathcal{H}}(g) + \varepsilon, \qquad \text{(LIPERMe)}$$

where $\epsilon > 0$ is a small number. It is clear that

$$L^{\mathsf{d},\mathcal{H}}(\widehat{g}_n) \leqslant L^{\mathsf{d}\mathcal{H}}(\widehat{g}_{n,\varepsilon}) \leqslant L^{\mathsf{d},\mathcal{H}}(\widehat{g}_n) + \varepsilon.$$

The subsequent sections aim to mathematically characterize the properties of (LIPERMe).

The left-inverse penalty in our work is similar to methods in deep learning, such as variational autoencoders (Kingma and Welling, 2014) and Cycle-GAN (Zhu et al., 2017). Variational autoencoders learn a compact latent representation and generate new examples by sampling from the learned latent space, using the penalty $\|g\circ h - \mathsf{Id}_D\|_{\mathbb{L}_2}$ instead of (LIP). Cycle-GAN focuses on style transfer and image-to-image translation, enforcing cyclic consistency similar to our LIP. Specifically, Cycle-GAN aims to satisfy the conditions $F\circ G(x) \approx x$ and $G\circ F(y) \approx y$ using two neural networks that perform style transfer in opposite directions.

The resemblance between our left-inverse penalty and methods yielding favorable empirical results suggests that minimizing the penalized empirical risk, though challenging, is feasible. We leverage this resemblance in the accompanying implementation, leading to experimental results reported in Section 6.

3. Main Result: Deviation from the Empirical Distribution

If the class $\mathcal{G}$ is rich, it is likely to contain a function $\widehat{g}$ that overfits the training data: the distance $\mathsf{W}_1(\widehat{g}\sharp\mathcal{U}_d, P_{n,X})$ might be very small or even zero. This type of overfitting has been observed in practice, as highlighted in (Somepalli et al., 2023a;b; Daras et al., 2023; Carlini et al., 2021; 2023; Jagielski et al., 2023). This behavior is undesirable for most generative modeling applications, such as image or music generation. Simply resampling examples from the training set is not the intended outcome.

The main finding in this paper is that overfitting to the empirical distribution can be mitigated or substantially restrained by imposing constraints on admissible generators, specifically requiring them to have a smooth left inverse. This holds true, particularly for the learned distribution defined by (LIPERM) with $\lambda = \infty$, referred to as the hard constraint case. In this setting, we establish that the learned distribution is significantly distant from the empirical distribution. Furthermore, we extend this finding to generators that are nearly left invertible, such as the generator defined by (LIPERMe) with a $\lambda > 0$.

To demonstrate that the learned generator, $\widehat{g}_n$, does not replicate the examples from the training set, we examine the distance between the probability measure induced by the generator, $\widehat{g}_n\sharp\mathcal{U}_d$, and any distribution Q satisfying $Q(\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}) = 1$. A larger distance indicates a greater dissimilarity, which is desirable.

3.1. Warm-up: The Case of Hard Constraint ($\lambda = \infty$)

We first consider the generator $\widehat{g}_n$ obtained by imposing the hard constraint $h \circ g = \mathrm{Id}_d$ on feasible solutions. This means that $\widehat{g}_n$ minimizes the distance, d, between $g\sharp\mathcal{U}_d$ and the empirical distribution $P_{n,X}$, over the set of all $g \in \mathcal{G}$ for which there exists $h \in \mathcal{H}$ satisfying $h \circ g = \mathsf{Id}_d$.

While our main results apply to more general IPMs, we primarily focus on the W_1 distance. This choice of d is particularly relevant due to its interpretation as an optimal transport distance. It has found successful applications in various fields, such as computer vision, economics and biology (Ollivier et al., 2014; Peyré and Cuturi, 2019).

Proposition 1. *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in [0,1]^D$. For any $g : [0,1]^d \to [0,1]^D$ having an $L_{\mathcal{H}}$-Lipschitz-continuous left inverse (that is there exists $h : \mathbb{R}^D \to \mathbb{R}^d$ such that $h \circ g = \mathsf{Id}_d$), it holds that*

$$\mathsf{W}_1(g\sharp\mathcal{U}_d, Q) \geqslant \frac{1}{2L_{\mathcal{H}}(1 + (2\mathsf{V}_d n)^{1/d})}, \qquad \text{(LB-hard)}$$

where Q is any probability satisfying $Q(\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}) = 1$ and[2] $\mathsf{V}_d = \pi^{d/2}/\Gamma(1 + d/2)$.

It is well-known (Dudley, 1969) that the W_1 distance between the empirical and the true distribution is $O(n^{-1/d})$. It has also been recently proved that the rate $n^{-1/d}$ is optimal (Tang and Yang, 2023), in the sense that it cannot be improved by any other learned distribution. Thus, if a generator $\tilde{g}_n$ is rate-optimal, the triangle inequality yields $\mathsf{W}_1(\tilde{g}_n\sharp\mathcal{U}_d, P_{n,X}) \leqslant \mathsf{W}_1(\tilde{g}_n\sharp\mathcal{U}_d, P^*) + \mathsf{W}_1(P^*, P_{n,X}) = O(n^{-1/d})$. Thus, if a learned generator is rate-optimal, its maximal distance from the empirical distribution $P_{n,X}$ of the training set is of order $n^{-1/d}$. This means, in view of our result, that if a rate-optimal learned generator has a smooth left inverse, it lays at a maximal distance from $P_{n,X}$.

To the best of our knowledge, Proposition 1 provides the first mathematical quantification of the diversity of examples generated by the generator. It not only demonstrates that the generator deviates maximally from the empirical distribution but also shows that it maintains the same minimum distance from any distribution concentrated on n points.

3.2. The Case of Soft Constraint: $\lambda \in (0, \infty)$

We now turn our attention to measuring the dissimilarity between the empirical distribution and the generator obtained by LIPERMe when $\lambda < +\infty$. We refer to this scenario as the case of a soft constraint on left-invertibility. Recall that the introduction of a penalized version of the optimization problem, as opposed to the hard constraint, aims to enhance computational tractability and facilitate implementation using available tools. On the downside, the lower bound on the distance from the empirical distribution, as presented in the following theorem, is slightly weaker than that for the constrained generator.

Theorem 1. *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in [0,1]^D$ and $\mathcal{H} \subseteq \mathrm{Lip}_{L_{\mathcal{H}}}([0,1]^D \to [0,1]^d)$. For any $\lambda > 0$,*

$$\mathsf{W}_1(\widehat{g}_{n,\varepsilon}\sharp\mathcal{U}_d, P_{n,X}) \geqslant \frac{1}{2L_{\mathcal{H}}(1 + (2\mathsf{V}_d n)^{1/d})} - \frac{1}{L_{\mathcal{H}}\lambda^{1/q}}\Big(\inf_{g \in \mathcal{G}_{\mathcal{H}}} \mathsf{W}_1(g\sharp\mathcal{U}_d, P_{n,X}) + \varepsilon\Big)^{1/q},$$

where $\mathcal{G}_{\mathcal{H}} = \{g \in \mathcal{G} : \mathrm{pen}(g) = 0\}$.

Corollary 1. *If λ is chosen so that the inequality*

$$\lambda \geqslant 8^q(1 + (2\mathsf{V}_d n)^{q/d}) \inf_{g \in \mathcal{G}_{\mathcal{H}}} \mathbb{E}\big[\mathsf{W}_1(g\sharp\mathcal{U}_d, P_{n,X})\big], \quad (2)$$

holds true, then

$$\mathbb{E}\big[\mathsf{W}_1(\widehat{g}_n\sharp\mathcal{U}_d, P_{n,X})\big] \geqslant \frac{1}{4L_{\mathcal{H}}(1 + (2\mathsf{V}_d n)^{1/d})}.$$

The corollary tells us that if the penalty λ is not too small, any generator satisfying (LIPERMe) has a deviation of the order $n^{-1/d}$ from the empirical distribution. To gain some understanding of how restrictive this constraint on λ is, let us note that if there is a L^*-Lipschitz-continuous $g^* \in \mathcal{G}_{\mathcal{H}}$ such that $\mathsf{W}_1(P^*, g^*\sharp\mathcal{U}_d) \leqslant \sigma^*$, one can check that[3]

$$\inf_{g \in \mathcal{G}_{\mathcal{H}}} \mathbb{E}\big[\mathsf{W}_1(g\sharp\mathcal{U}_d, P_{n,X})\big] \leqslant \frac{cL^*\sqrt{d}}{n^{1/d}} + \sigma^* \quad (3)$$

where c is a universal constant.

Remark 1. In view of (2) and (3), when $\sigma^* = 0$ and $q = 2$, choosing the penalty parameter λ larger than $C_d n^{1/d}$— for a constant C_d that depends only on the dimension of the latent space— is enough to guarantee that the generator LIPERMe will significantly deviate from the empirical distribution.

Before closing this section, let us note that the inspection of the proof shows that the claim of the last theorem holds true if we replace W_1 by any other distance dominating W_1. In particular, the claim is true for W_2 and for $\mathsf{d}_{\mathcal{F}}$ with any $\mathcal{F}$ containing all the 1-Lipschitz functions.

4. Precision of Left-Inverse-Penalized ERM

In this section, we assess the precision of a generator satisfying (LIPERMe) for $\mathsf{d} = \mathsf{d}_{\mathcal{F}}$, the integral probability metric (IPM) based on a set of test functions $\mathcal{F}$. We introduce two parameters σ^* and L^*, that quantify the "manifold assumption". More precisely, we say that P^* satisfies assumption $\mathsf{A}(L^*, \sigma^*)$ if

$$\exists\, g^* \in \mathrm{Lip}_{L^*}\big([0,1]^d \to [0,1]^D\big) \text{ such that } \mathsf{W}_1(g^*\sharp\mathcal{U}_d, P^*) \leqslant \sigma^*. \qquad \mathsf{A}(L^*, \sigma^*)$$

[2] V_d is the volume of the unit ball in $\mathbb{R}^d$.

[3] See Appendix B.4

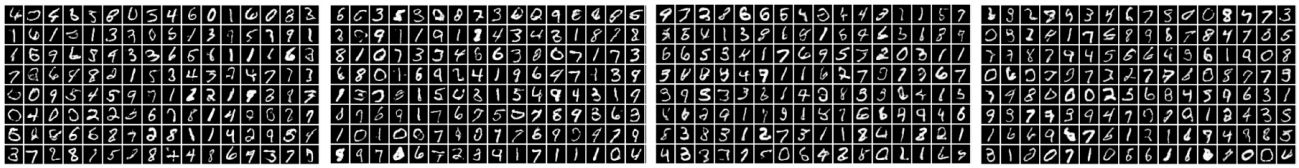


Figure 2. Handwritten digits generated by LIPERM, from left to right: $\lambda = 0, 1, 4, 8$. See Fig. 9 for higher-resolution images.

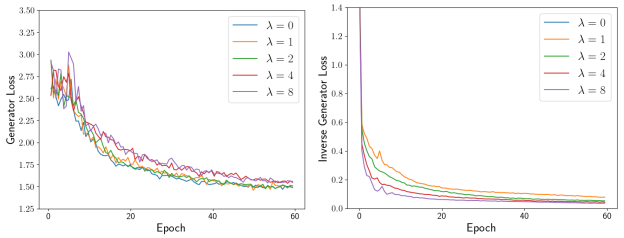


Figure 3. LIPERM on MNIST data. The behavior of the generator loss and of the left-inverse penalty across the iterations.

This assumption accommodates both well-specified and misspecified scenarios. The former corresponds to $\mathsf{A}(L^*, 0)$, where P^* is the push-forward of $\mathcal{U}_d$ by an L^*-Lipschitz-continuous function. The latter corresponds to the case $\mathsf{A}(L^*, \sigma^*)$ with $\sigma^* > 0$, the parameter σ^* indicating the degree of the departure from a well-specified setting.

Theorem 2. *Assume that $\mathcal{F} \subseteq \mathrm{Lip}_1(\mathbb{R}^D \to \mathbb{R})$ and the dimension of the latent space satisfies $d > 2$. If the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d., drawn from P^* satisfying $\mathsf{A}(L^*, \sigma^*)$ for some L^*, σ^*, then* LIPERMe *$\widehat{g}_{n,\varepsilon}$ satisfies*

$$\mathbb{E}[\mathsf{d}_{\mathcal{F}}(\widehat{g}_{n,\varepsilon\sharp}\mathcal{U}_d, P^*)] \leqslant \inf_{g\in\mathcal{G}} \left\{\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P^*) + \lambda\,\mathrm{pen}_{\mathcal{H}}(g)\right\} + 4\sigma^* + \varepsilon + \frac{cL^*\sqrt{d}}{n^{1/d}}, \tag{UB}$$

where $c > 0$ is a universal constant.

Remark 2. If we assume that the oracle g^* appearing in $\mathsf{A}(L^*, \sigma^*)$ has a left-inverse that is $L_{\mathcal{H}}$-Lipschitz-continuous, the upper bound provided by Theorem 2 becomes $5\sigma^* + cL^*\sqrt{d}\,n^{-1/d}$. As proven in (Schreuder et al., 2021; Tang and Yang, 2023), this upper bound is minimax-rate-optimal. Notably, the ambient dimension does not appear in the upper bound.

Remark 3. The assumption $d > 2$ is not crucial for deriving an upper bound similar to (UB). In case of $d \in \{1, 2\}$, slight modifications occur in the last term: for $d = 1$, the denominator becomes $n^{1/2}$, and for $d = 2$, an additional $\log n$ factor appears in the numerator. These adjustments are derived by combining our proof, detailed in Appendix A.1, with the corresponding approximation bounds for the uniform distribution in W_1 distance, addressing the cases of $d \in \{1, 2\}$.

Remark 4. When the true distribution P^* exhibits low sample diversity, LIPERMe struggles to provide a precise approximation. In this scenario, the upper bound in Theorem 2 tends to be large, as indicated by the constant σ^*.

5. Handling Functional Approximations

The generator (LIPERM) and its approximate version (LIPERMe) rely on the functional classes $\mathcal{G}$ and $\mathcal{H}$. Opting for smaller, parametric classes enhances computational efficiency. Yet, for minimizing the bias term in (UB), it is essential to choose $\mathcal{G}$ and $\mathcal{H}$ as large as possible. This prompts the question: what if, during training, we substitute $\mathcal{G}$ and $\mathcal{H}$ with smaller sets $\mathcal{G}_0$ and $\mathcal{H}_0$ possessing good approximation properties? Neural networks, acknowledged as universal approximators for smooth functions (Yarotsky, 2017; Petersen and Voigtlaender, 2018; Nakada and Imaizumi, 2020), are compelling candidates for $\mathcal{G}_0$ and $\mathcal{H}_0$.

This consideration extends to the functional class $\mathcal{F}$. While we aim to gauge the precision of a generator using an IPM with a broad $\mathcal{F}$, replacing $\mathcal{F}$ with a smaller set during the training process decreases computational complexity. The next result shows the impact of replacing $\mathcal{F}, \mathcal{G}$ and $\mathcal{H}$ in (LIPERMe) by smaller approximation classes.

Proposition 2. *Let $\mathcal{F} \subseteq \mathrm{Lip}_1(\mathbb{R}^D \to \mathbb{R})$ and $d > 2$. Let* pen *be defined by (LIP) with $q = 2$. Assume that observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. and drawn from P^*, satisfying $\mathsf{A}(L^*, \sigma^*)$ for some $L^*, \sigma^* > 0$. Let $\mathcal{F}_0 \subseteq \mathcal{F}$, $\mathcal{G}_0 \subseteq \mathcal{G}$ and $\mathcal{H}_0 \subseteq \mathcal{H}$ be functional classes satisfying*

$$\sup_{f\in\mathcal{F}} \inf_{f_0\in\mathcal{F}_0} \|f_0 - f\|_\infty \leqslant \delta_{\mathcal{F}}, \tag{4}$$

$$\sup_{g\in\mathcal{G}} \inf_{g_0\in\mathcal{G}_0} \|g_0 - g\|_\infty \leqslant \delta_{\mathcal{G}},$$

$$\sup_{h\in\mathcal{H}} \inf_{h_0\in\mathcal{H}_0} \|h_0 - h\|_\infty \leqslant \delta_{\mathcal{H}}. \tag{5}$$

Then, any learned generator $\widehat{g}^0 = \widehat{g}^0_{n,\varepsilon}$ satisfying

$$\widehat{L}_n^{\mathsf{d}_{\mathcal{F}_0},\mathcal{H}_0}(\widehat{g}^0_{n,\varepsilon}) \leqslant \min_{g\in\mathcal{G}_0} \widehat{L}_n^{\mathsf{d}_{\mathcal{F}_0},\mathcal{H}_0}(g) + \varepsilon,$$

also satisfies $\widehat{L}_n^{\mathsf{d}_{\mathcal{F}},\mathcal{H}}(\widehat{g}^0_{n,\varepsilon}) \leqslant \min_{g\in\mathcal{G}} \widehat{L}_n^{\mathsf{d}_{\mathcal{F}},\mathcal{H}}(g) + \varepsilon + \delta$, where $\delta = 2\delta_{\mathcal{F}} + 2\sqrt{d}\,\delta_{\mathcal{H}} + \big(1 + 2\lambda\sqrt{d}\,L_{\mathcal{H}}\big)\delta_{\mathcal{G}}$.

Combining the results of this proposition with Theorem 2, we obtain the following property: if $\mathcal{F}_0$, $\mathcal{G}_0$ and $\mathcal{H}_0$ are chosen so that $\delta_{\mathcal{F}} \leqslant \varepsilon/6$, $\delta_{\mathcal{G}} \leqslant \varepsilon/3(1 + 2\lambda\sqrt{d}\,L_{\mathcal{H}})^{-1}$ and $\delta_{\mathcal{H}} \leqslant d^{-1/2}\varepsilon/6$, see (4-5), then any ε-minimizer $\widehat{g}^0$ of

$\widehat{L}_n^{\mathsf{d}_{\mathcal{F}_0},\mathcal{H}_0}$ over $\mathcal{G}_0$ satisfies

$$\mathbb{E}[\mathsf{d}_{\mathcal{F}}(\widehat{g}^0\sharp\mathcal{U}_d, P^*)] \leqslant \inf_{g\in\mathcal{G}} \left\{\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P^*) + \lambda\,\mathrm{pen}_{\mathcal{H}}(g)\right\} + 4\sigma^* + \frac{cL^*\sqrt{d}}{n^{1/d}} + 2\varepsilon.$$

This suggests that with appropriately chosen approximation classes, the learned generator can approach the performance of the best generator from the class of all Lipschitz-continuous functions g with a Lipschitz-continuous left-inverse h. Notably, utilizing a set of "critics" $\mathcal{F}$ consisting of neural networks that are δ-approximations of 1-Lipschitz functions results in an error within 2δ of the error obtained when trained with $\mathcal{F} = \mathrm{Lip}_1$. Note that according to (Yarotsky, 2017, Theorem 1), one can achieve an error $\delta_{\mathcal{F}}$ for $\mathcal{F}$ the functions with derivatives bounded by 1 using ReLU activated neural networks of depth $O(\log(1/\delta_{\mathcal{F}}))$ and of the number of weights and units $O(\delta_{\mathcal{F}}^{-D}\log(1/\delta_{\mathcal{F}}))$. Furthermore, in view of Theorem 1,

$$\mathsf{W}_1(\widehat{g}^0\sharp\mathcal{U}_d, P_{n,X}) \geqslant \frac{1}{2L_{\mathcal{H}}(1+(2\mathsf{V}_d n)^{1/d})} - \frac{1}{L_{\mathcal{H}}\lambda^{1/2}}\Big(\inf_{g\in\mathcal{G}_{\mathcal{H}}} \mathsf{W}_1(g\sharp\mathcal{U}_d, P_{n,X}) + 2\varepsilon\Big)^{1/2},$$

where $c > 0$ is a universal constant.

Finally, note that the last proposition can be used with $\mathcal{G} = \mathcal{G}_0$, a parametric set defined by neural networks with a given architecture. Then $\delta_{\mathcal{G}} = 0$, which conveniently simplifies the expression of δ. In addition, the approximation error may be directly bounded using results from (Yang et al., 2022; Lu and Lu, 2020).

6. Numerical Experiments

In this section, we aim to evaluate the performance of the left-inverse-penalized WGANs. Our implementation[4] follows the pseudo-code presented in Algorithm 1, and is inspired by the code accompanying (Gulrajani et al., 2017), where WGANs with gradient penalty on the discriminator/critic network is discussed. We add the LIPERM penalization to the objective function of the WGANs. All the functional classes $\mathcal{F}_0$, $\mathcal{G}_0$, and $\mathcal{H}_0$ are represented by neural networks, the architectures of which are presented in the supplementary material. In all our experiments, we chose $n_{critic} = 5$ and $\gamma = 1$.

We conducted experiments on three widely used datasets: Swiss Roll, MNIST and CIFAR 10. The results are briefly summarized in this section. The main messages of these experiments are that (a) it is possible to implement the LIPERM algorithm and to get generators that are nearly invertible, that is they have a small left inverse penalty, (b) the visual quality of the results does not deteriorate when the penalty parameter λ is increased. As for the non replication, it seems that with the architectures and optimizers used in the standard data sets considered in this section, the replication or the lack of creativity is not an issue. Therefore, we could not observe an increase in creativity due to introducing the left inverse penalty.

[4]Our code uses the framework of (Varuna Jayasiri, 2020) and is available here.

Algorithm 1 WGAN-LIPERM. We take $\lambda \in \{0, 1, 4, 8\}$,

Require: LIP coefficient λ, gradient penalty coefficient γ, number of iterations N_{iter}, number of critic iterations per generator iteration n_{critic}, batch size m
Require: initial critic and generator parameters $(\boldsymbol{w}_0, \boldsymbol{\theta}_0)$, initial left inverse network parameters $\boldsymbol{\phi}_0$, $k = 0$.

repeat
  $k \leftarrow k + 1$
  for $t = 1, ..., n_{\text{critic}}$ **do**
    for $i = 1, ..., m$ **do**
      Draw $\boldsymbol{x} \sim P_{n,X}$ (true examples)
      Draw $\boldsymbol{u} \sim \mathcal{U}_d$ (latent variables)
      Draw $\epsilon \sim \mathcal{U}[0, 1]$
      $\tilde{\boldsymbol{x}} \leftarrow G_{\boldsymbol{\theta}}(\boldsymbol{u})$ (generated examples)
      $\widehat{\boldsymbol{x}} \leftarrow \epsilon\boldsymbol{x} + (1-\epsilon)\tilde{\boldsymbol{x}}$
      $L_d \leftarrow F_{\boldsymbol{w}}(\tilde{\boldsymbol{x}}) - F_{\boldsymbol{w}}(\boldsymbol{x})$
      $L_d^{(i)} \leftarrow L_d + \gamma(\|\nabla_{\widehat{\boldsymbol{x}}} F_{\boldsymbol{w}}(\widehat{\boldsymbol{x}})\|^2 - 1)^2$
    end for
    $\boldsymbol{w} \leftarrow \mathrm{Adam}(\nabla_w \frac{1}{m}\sum_{i=1}^m L_d^{(i)}, \{\boldsymbol{w}\})$
  end for
  for $i = 1, ..., m$ **do**
    Draw $\boldsymbol{u} \sim \mathcal{U}_d$
    $L_g^{(i)} \leftarrow -F_{\boldsymbol{w}}(G_{\boldsymbol{\theta}}(\boldsymbol{u})) + \lambda\|H_{\boldsymbol{\phi}}(G_{\boldsymbol{\theta}}(\boldsymbol{u})) - \boldsymbol{u}\|^2$
  end for
  $(\boldsymbol{\theta}, \boldsymbol{\phi}) \leftarrow \mathrm{Adam}(\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \frac{1}{m}\sum_{i=1}^m L_g^{(i)}, \{(\boldsymbol{\theta}, \boldsymbol{\phi})\})$
until $k > N_{\text{iter}}$

Swiss Roll (Marsland, 2009): The Swiss Roll dataset consists of 2D points arranged in a rolled structure, corrupted by a 2D Gaussian noise with some standard deviation σ. We used three values of $\sigma = 3/2, 3/4, 3/8$ We used a training set of size 1000, a batch-size of 200 and run experiments for each value of λ from $\{0, 1, 4, 8\}$. Fig. 4 and Fig. 8 in Appendix show the training set and the points generated by the learned distribution. These results are consistent with the theory: for increasing but not very large values of λ the accuracy of the generator is preserved. Note that it is known (Srivastava et al., 2017) that training a generator for this data is highly unstable. Plots in Fig. 7 confirm this instability and show that, unfortunately, the left-inverse penalty does not alleviate it. We also conducted additional unreported experiments, training for over 4,000 epochs (up to 20,000 epochs), but observed no improvement in the results.

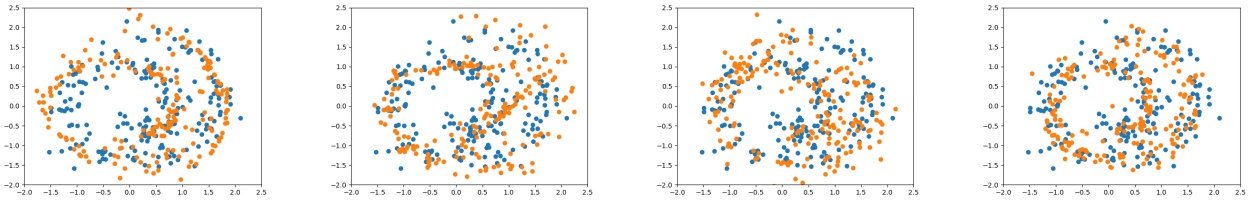


Figure 4. Swiss Roll: samples generated by LIPERM WGAN (blue) and original data (orange) for $\lambda = 0, 1, 4, 8$ (left to right).

MNIST (LeCun, 1998): MNIST dataset is commonly used for handwritten digit recognition. In this paper, we train WGAN plus LIPERM penalty on MNIST and generate images for different values of LIPERM coefficient $\lambda \in \{0, 1, 4, 8\}$. As shown in Fig. 3, the training process seems to converge after nearly 60 epochs, when the batch size is 512. The architectures of the networks used in this experiment are presented in Table 2 in Appendix.

The results depicted in Fig. 2 suggest that the generated images when $\lambda = 1, 4, 8$ are as good as those for the vanilla WGAN ($\lambda = 0$). Furthermore, the right plot of Fig. 3 shows that the implementation we used effectively minimizes the LIP and that the final result is almost left invertible.

CIFAR-10 (Krizhevsky, 2009): Finally, to demonstrate that our method can also be applied on real-world images, we perform experiments on CIFAR-10 dataset. For Generator network we employ ResNet-style architecture and for Discriminator and Inverse-Generator we use simpler convolutional networks. For more architectural details, please refer to our implementation.

Results and generated samples are presented in Figure 5 and Figure 6. One can observe that increasing λ does not decrease the quality of generated samples. Indeed, the inception scores as well as the generator losses depicted in Figure 5 seem to show that the choice of λ has almost no impact on the accuracy. However, according to the right plot of the same figure, the generator corresponding to $\lambda = 8$ has the smallest penalty and, therefore, is closer to be left invertible.

7. Summary, Conclusion and Limitations

In this paper, we have presented a theoretical analysis of training generative models with two key properties: avoidance of replication of the observed examples and convergence to the true distribution at a minimax optimal rate. Our main contribution is that the existence of a smooth left-inverse implies the first one of these properties. We further introduced the left-inverse-penalized empirical risk minimization LIPERM framework, with a penalty encouraging the generator to possess a smooth inverse. We showed, both theoretically and empirically, that LIPERM and its approximate version LIPERMe enjoy the mentioned desirable properties.

Limitations The incorporation of left invertibility poses certain computational challenges. While our numerical experiments indicate that these difficulties are not insurmountable, optimization errors will likely be dominant in most applications. Our work does not explicitly handle choosing d, the dimension of the latent space. One approach to address this is considering the Bourgain theorem (Bourgain, 1985), as in (Xiao et al., 2018). The choice of d may also influence the Lipschitz constant L^*, which plays a pivotal role. The interplay between these quantities needs to be better understood; the methodology developed in (Jordan and Dimakis, 2020; 2021; Wang and Manchester, 2023) might help in this task. Extending our analysis to functional classes with higher smoothness and diffusion models is non-trivial. On a related note, considering distances between distributions that are not IPMs, such as the Sinkhorn divergence (Genevay et al., 2018; Luise et al., 2020) cannot be done using the methodology of this paper. These challenges pose interesting questions for future research.

Impact Statement

This paper presents work that aims to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

This work was supported by the grant ANR-11-IDEX0003/Labex Ecodec/ANR-11-LABX-0047, and the ADVANCE Research Grant provided by the Foundation for Armenian Science and Technology (FAST) and Yerevan State University (YSU).

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

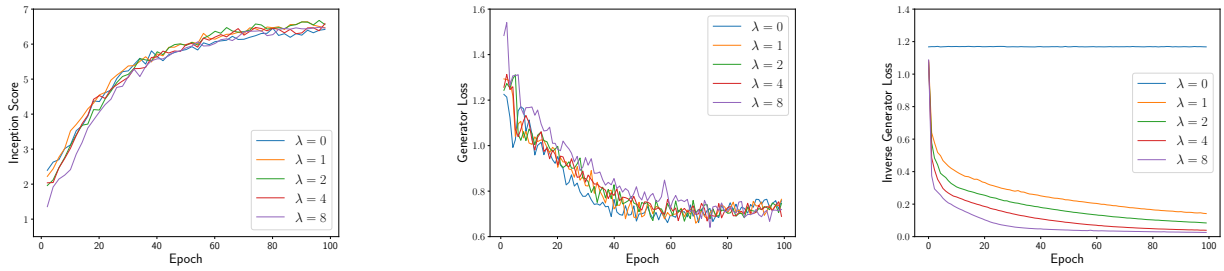


Figure 5. Experimental results on CIFAR-10 data set. Left: the evolution of the Inception Score across the iterations. Middle: the evolution of the generator loss across the iterations, for various values of λ. Right: the evolution of the left-inverse penalty across the iterations, for various values of λ.

Figure 6. Generated image samples on CIFAR-10 for different values of LIPERMe $\lambda = 0$; $\lambda = 1$; $\lambda = 4$, and $\lambda = 8$

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and Equilibrium in Generative Adversarial Nets (GANs). In *ICML, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 224–232. PMLR, 2017.

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs Learn the Distribution? Some Theory and Empirics. In *ICLR, Vancouver, BC, Canada, April 30 - May 3, 2018*, 2018.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of Discriminators Implies Diversity in GANs. *ArXiv*, abs/1806.10586, 2018.

Denis Belomestny, Eric Moulines, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Rates of Convergence for Density Estimation with Generative Adversarial Networks, 2023.

Gérard Biau, Benoît Cadre, Maxime Sangnier, and Ugo Tanielian. Some Theoretical Properties of GANs. *The Annals of Statistics*, 48(3):1539 – 1566, 2020. doi: 10.1214/19-AOS1858.

Gérard Biau, Maxime Sangnier, and Ugo Tanielian. Some Theoretical Insights into Wasserstein GANs. *J. Mach. Learn. Res.*, 22:119:1–119:45, 2021.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative Modeling with Denoising Auto-Encoders and Langevin Sampling. *CoRR*, abs/2002.00107, 2020.

Emmanuel Boissard and Thibaut Le Gouic. On the Mean Speed of Convergence of Empirical and Occupation Measures in Wasserstein Distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539 – 563, 2014.

Valentin De Bortoli, Emile Mathieu, MJ Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian Score-Based Generative Modelling. *Advances in Neural Information Processing Systems*, page 46, 2022.

Jean Bourgain. On Lipschitz Embedding of Finite Metric Spaces in Hilbert space. *Israel J. Math.*, 52(1-2):46–52, 1985.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Minwoo Chae, Dongha Kim, Yongdai Kim, and Lizhen Lin. A Likelihood Approach to Nonparametric Estimation of a Singular Distribution Using Deep Generative Models. *Journal of Machine Learning Research*, 24(77): 1–42, 2023.

Sitan Chen, Jerry Li, Yuanzhi Li, and Raghu Meka. Minimax Optimality (Probably) Doesn't Imply Distribution Learning for GANs. *arXiv preprint arXiv:2201.07206*, 2022.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient Diffusion: Learning Clean Distributions from Corrupted Data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17695–17709. Curran Associates, Inc., 2021.

Richard M. Dudley. The Speed of Mean Glivenko-Cantelli Convergence. *The Annals of Mathematical Statistics*, 40 (1):40 – 50, 1969. doi: 10.1214/aoms/1177697802.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martín Arjovsky, Olivier Mastropietro, and Aaron C. Courville. Adversarially Learned Inference. In *ICLR, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the Manifold Hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Mohammad Navid Fekri, Ananda Mohon Ghosh, and Katarina Grolinger. Generating Energy Data for Machine Learning with Recurrent Generative Adversarial Networks. *Energies*, 13(1):130, 2019.

David John Gagne, Hannah M Christensen, Aneesh C Subramanian, and Adam H Monahan. Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz'96 Model. *Journal of Advances in Modeling Earth Systems*, 12(3):e2019MS001896, 2020.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *AISTATS*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 27: 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved Training of Wasserstein GANs. In *NIPS*, pages 5767–5777, 2017.

Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards GAN Benchmarks Which require Generalization. In *ICLR (Poster)*. OpenReview.net, 2019.

Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An Error Analysis of Generative Adversarial Networks for Learning Distributions. *Journal of Machine Learning Research*, 23:1–43, 2022.

Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring Forgetting of Memorized Training Examples. In *The Eleventh International Conference on Learning Representations*, 2023.

Matt Jordan and Alex Dimakis. Provable Lipschitz Certification for Generative Models. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5118–5126. PMLR, 2021.

Matt Jordan and Alexandros G Dimakis. Exactly Computing the Local Lipschitz Constant of ReLU Networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7344–7353. Curran Associates, Inc., 2020.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *2nd International Conference on Learning Representations, ICLR*, 2014.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009.

Hyeok Kyu Kwon and Minwoo Chae. Minimax Optimal Density Estimation using a Shallow Generative Model with a One-Dimensional Latent Variable. pages 469–477, 2024.

Yann LeCun. The MNIST Database of Handwritten Digits. 1998. URL http://yann.lecun.com/exdb/mnist/.

Cheuk Ting Li and Farzan Farnia. Mode-Seeking Divergences: Theory and Applications to GANs. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8321–8350. PMLR, 25–27 Apr 2023.

Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.

Tengyuan Liang. How Well Generative Adversarial Networks Learn Distributions. *J. Mach. Learn. Res.*, 22(1), jan 2021.

Yulong Lu and Jianfeng Lu. A Universal Approximation Theorem of Deep Neural Networks for Expressing Probability Distributions. *Advances in neural information processing systems*, 33:3094–3105, 2020.

Giulia Luise, Massimiliano Pontil, and Carlo Ciliberto. Generalization properties of optimal transport gans with latent distribution learning. *CoRR*, abs/2007.14641, 2020.

Stephen Marsland. *Machine Learning - An Algorithmic Perspective.* Chapman and Hall / CRC machine learning and pattern recognition series. CRC Press, 2009.

Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, Tomasz Danel, and Michał Warchoł. Mol-CycleGAN: A Generative Model for Molecular Optimization. *Journal of Cheminformatics*, 12(1):1–18, 2020.

Vaishnavh Nagarajan, Colin Raffel, and Ian J Goodfellow. Theoretical Insights into Memorization in GANs. *Neural Information Processing Systems (NeurIPS) 2017 - Integration of Deep Learning Theories Workshop*, 2018.

Ryumei Nakada and Masaaki Imaizumi. Adaptive Approximation and Generalization of Deep Neural Network with Intrinsic Dimensionality. *The Journal of Machine Learning Research*, 21(1):7018–7055, 2020.

Hariharan Narayanan and Sanjoy K. Mitter. Sample Complexity of Testing the Manifold Hypothesis. In *NIPS*, pages 1786–1794. Curran Associates, Inc., 2010.

Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical Image Synthesis with Context-Aware Generative Adversarial Networks. In *Medical Image Computing and Computer Assisted Intervention, 2017, Proceedings, Part III 20*, pages 417–425. Springer, 2017.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of Wasserstein Distances in the Spiked Transport Model. *Bernoulli*, 28(4):2663 – 2688, 2022. doi: 10.3150/21-BEJ1433.

Yann Ollivier, Hervé Pajot, and Cedric Villani. *Optimal Transport: Theory and Applications*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2014.

Michela Paganini, Luke de Oliveira, and Benjamin Nachman. CaloGAN: Simulating 3D High Energy Particle Showers in Multilayer Electromagnetic Calorimeters with Generative Adversarial Networks. *Physical Review D*, 97 (1):014021, 2018.

Philipp Petersen and Felix Voigtlaender. Optimal Approximation of Piecewise Smooth Functions using Deep ReLU Neural Networks. *Neural Networks*, 108:296–330, 2018.

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the Regularization of Wasserstein GANs. In *International Conference on Learning Representations*, 2018.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, Otto Savolainen, Rolandas Meškys, Martin Engqvist, and Aleksej Zelezniak. Expanding Functional Protein Sequence Spaces using Generative Adversarial Networks. *Nature Machine Intelligence*, 3(4):324–333, 2021.

Kevin Roth, Aurélien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing Training of Generative Adversarial Networks through Regularization. In *NIPS*, pages 2018–2028, 2017.

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak S. Dalalyan. Statistical Guarantees for Generative Models without Domination. In *ALT*, volume 132 of *Proceedings of Machine Learning Research*, pages 1051–1071. PMLR, 2021.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and Mitigating Copying in Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, and Charles Sutton. VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning. In *Neural Information Processing Systems*, 2017.

Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Wasserstein GANs are Minimax Optimal Distribution Estimators, 2023.

Rong Tang and Yun Yang. Minimax Rate of Distribution Estimation on Unknown Submanifold under Adversarial Losses, 2023.

Ananya Uppal, Shashank Singh, and Barnabás Póczos. Nonparametric Density Estimation & Convergence Rates for GANs under Besov IPM Losses. *Advances in neural information processing systems*, 32, 2019.

Nipun Wijerathne Varuna Jayasiri. labml.ai annotated paper implementations, 2020. URL https://nn.labml.ai/.

Ruigang Wang and Ian R. Manchester. Direct Parameterization of Lipschitz-Bounded Deep Networks. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 36093–36110. PMLR, 2023.

Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant GANs: Deep Generation of Financial Time Series. *Quantitative Finance*, 20(9):1419–1440, 2020.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in Generative Models: Characterization and Strong Identifiability. *arXiv preprint arXiv:2206.00801*, 2023.

Chang Xiao, Peilin Zhong, and Changxi Zheng. BourGAN: Generative Networks with Metric Embeddings. *Advances in neural information processing systems*, 31, 2018.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. DeepLesion: Automated Mining of Large-Scale Lesion Annotations and Universal Lesion Detection with Deep Learning. *Journal of medical imaging*, 5(3):036501–036501, 2018.

Yunfei Yang, Zhen Li, and Yang Wang. On the Capacity of Deep Generative Networks for Approximating Distributions. *Neural networks*, 145:144–154, 2022.

Dmitry Yarotsky. Error Bounds for Approximations with Deep ReLU Networks. *Neural Networks*, 94:103–114, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2868–2876, 2017.

Contents

Appendix

The purpose of this appendix is twofold: to present the proofs of all the mathematical claims presented in the main paper and to provide some additional experimental results. We illustrate the effect of the penalization parameter λ on the trained generator using the LIPERMe framework. The code that can be used to reproduce all the experiments can be found here https://github.com/TigranGalstyan/LIPERM_annotated_deep_learning_paper_implementations/tree/liperm/labml_nn/gan/wasserstein/gradient_penalty/liperm.

A. Proof of the upper bound on the risk

The proof of Theorem 2, provided below, relies on repeated use of the triangle inequality, the near-minimization property of LIPERMe, as well as on exploiting Lipschitz continuity assumptions. The final bound is derived by combining these inequalities with the approximation bound in the Wasserstein-1 distance of the uniform distribution by its empirical counterpart.

This result demonstrates that under mild assumptions, the generator trained by LIPERMe achieves the optimal rate for any choice of λ. Therefore, there is flexibility in selecting λ to enforce dissimilarity with the training examples without compromising accuracy.

A.1. Proof of Theorem 2

Let g be an arbitrary element from $\mathcal{G}$. The proof begins by using the triangle inequality multiple times, resulting in the following sequence of inequalities:

$$\begin{aligned}
\mathsf{d}_{\mathcal{F}}(\widehat{g}_{n,\epsilon}\sharp\mathcal{U}_d, P^*) &\leqslant \mathsf{d}_{\mathcal{F}}(\widehat{g}_{n,\epsilon}\sharp\mathcal{U}_d, P_{n,X}) + \mathsf{d}_{\mathcal{F}}(P_{n,X}, P^*) \\
&= \mathsf{d}_{\mathcal{F}}(\widehat{g}_{n,\epsilon}\sharp\mathcal{U}_d, P_{n,X}) + \lambda\,\mathrm{pen}_{\mathcal{H}}(\widehat{g}_{n,\epsilon}) - \lambda\,\mathrm{pen}_{\mathcal{H}}(\widehat{g}_{n,\epsilon}) + \mathsf{d}_{\mathcal{F}}(P_{n,X}, P^*) \\
&\leqslant \mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P_{n,X}) + \lambda\,\mathrm{pen}_{\mathcal{H}}(g) + \epsilon - \lambda\,\mathrm{pen}_{\mathcal{H}}(\widehat{g}_{n,\epsilon}) + \mathsf{d}_{\mathcal{F}}(P_{n,X}, P^*) \\
&\leqslant \mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P^*) + \lambda\,\mathrm{pen}_{\mathcal{H}}(g) + \epsilon - \lambda\,\mathrm{pen}_{\mathcal{H}}(\widehat{g}_{n,\epsilon}) + 2\mathsf{d}_{\mathcal{F}}(P_{n,X}, P^*).
\end{aligned}$$

In the second inequality, we use the fact that $\widehat{g}_{n,\varepsilon}$ is an ε-minimizer of (LIPERM), whereas in the last line, the triangle

inequality is applied. As the left-inverse penalty is always nonnegative, the last display implies:

$$\begin{aligned}\mathsf{d}_{\mathcal{F}}(\widehat{g}_{n,\epsilon}\sharp\mathcal{U}_d, P^*) &\leqslant \inf_{g\in\mathcal{G}} \left\{\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P^*) + \lambda\,\mathrm{pen}_{\mathcal{H}}(g)\right\} + 2\mathsf{d}_{\mathcal{F}}(P_{n,X}, P^*) + \epsilon \\ &\leqslant \inf_{g\in\mathcal{G}} \left\{\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P^*) + \lambda\,\mathrm{pen}_{\mathcal{H}}(g)\right\} + 2\mathsf{d}_{\mathcal{F}}(P_{n,X}, g^*\sharp\mathcal{U}_d) + 2\sigma^* + \epsilon. \end{aligned} \tag{6}$$

The last line follows from $\mathsf{A}(L^*, \sigma^*)$ and the triangle inequality. By applying $\mathsf{A}(L^*, \sigma^*)$ again, we can establish the existence of independent random variables $\boldsymbol{U}_i \sim \mathcal{U}_d$ such that $\mathbb{E}[\|\boldsymbol{X}_i - g(\boldsymbol{U}_i)\|] \leqslant \sigma^*$, for every $i \in [n]$. This implies that

$$\begin{aligned}\mathsf{d}_{\mathcal{F}}(P_{n,X}, g^*\sharp\mathcal{U}_d) &= \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i) - \mathbb{E}[f\circ g^*(\boldsymbol{U})]\right| \\ &= \sup_{f\in\mathcal{F}} \left\{\left|\frac{1}{n}\sum_{i=1}^n \big(f(\boldsymbol{X}_i) - f\circ g^*(\boldsymbol{U}_i)\big) + \frac{1}{n}\sum_{i=1}^n \big(f\circ g^*(\boldsymbol{U}_i) - \mathbb{E}[f\circ g^*(\boldsymbol{U})]\big)\right|\right\} \\ &\leqslant \frac{1}{n}\sum_{i=1}^n \big\|\boldsymbol{X}_i - g^*(\boldsymbol{U}_i)\big\| + \sup_{\psi\in\mathrm{Lip}_{L^*}} \left|\frac{1}{n}\sum_{i=1}^n \big(\psi(\boldsymbol{U}_i) - \mathbb{E}[\psi(\boldsymbol{U})]\big)\right|. \end{aligned}$$

Here, we used the fact that $\mathcal{F} \subseteq \mathrm{Lip}_1([0,1]^D \to \mathbb{R})$ and that the composition of a 1-Lipschitz-continuous and an L^*-Lipschitz-continuous functions is itself L^*-Lipschitz continuous. Taking the expectation and employing the dual formulation of the Wasserstein-1 distance, we arrive at

$$\begin{aligned}\mathbb{E}[\mathsf{d}_{\mathcal{F}}(P_{n,X}, g^*\sharp\mathcal{U}_d)] &\leqslant \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\big\|\boldsymbol{X}_i - g^*(\boldsymbol{U}_i)\big\|] + L^*\mathbb{E}[\mathsf{W}_1(P_{n,U}, \mathcal{U}_d)] \\ &\leqslant \sigma^* + L^*\mathbb{E}[\mathsf{W}_1(P_{n,U}, \mathcal{U}_d)]. \end{aligned} \tag{7}$$

Here, $\widehat{P}_{n,U}$ is the empirical distribution of the sample $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$. Finally, using the well-known bound on the error of the empirical distribution in the Wasserstein-1 distance (for example, see (Niles-Weed and Rigollet, 2022, Proposition 1)), we obtain:

$$\mathbb{E}[\mathsf{W}_1(P_{n,U}, \mathcal{U}_d)] \leqslant \frac{c\sqrt{d}}{n^{1/d}}.$$

Combining this bound with (6) and (7), we get the stated upper bound.

B. Proofs for the deviation of the generative distribution from the empirical distribution

B.1. Lower bounding the distance between the uniform distribution and any discrete distribution

Proposition 3. *Assume that $\mathcal{F}$ contains all the 1-Lipschitz-continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$. For any set of points $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in [0,1]^d$ and any set of weights $w_1, \ldots, w_n \geqslant 0$ summing to one, we have*

$$\mathsf{d}_{\mathcal{F}}\Big(\mathcal{U}_d, \sum_{i=1}^n w_i\delta_{\boldsymbol{a}_i}\Big) \geqslant \frac{1}{2 + 2(2\mathsf{V}_d n)^{1/d}},$$

where $\mathcal{U}_d$ is the uniform distribution on $[0,1]^d$ with $\mathsf{V}_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ being the volume of the unit ball[5] in $\mathbb{R}^d$.

Proof. Let $A = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$ and define the function

$$g_A : [0,1]^d \to [0,1], \qquad g_A(\boldsymbol{x}) = \min\big(1, \min_{\boldsymbol{a}\in A} \|\boldsymbol{x} - \boldsymbol{a}\|\big),$$

We start the proof by noticing that for any $A \subseteq \mathbb{R}^d$ the function g_A is 1-Lipschitz. Therefore, we have $\{g_A : A \subseteq \mathbb{R}^d\} \subseteq \mathrm{Lip}_1 \subseteq \mathcal{F}$. Therefore,

$$\mathsf{d}_{\mathcal{F}}\Big(\mathcal{U}_d, \sum_{i=1}^n w_i\delta_{\boldsymbol{a}_i}\Big) \geqslant \sup_A \left(\int_{[0,1]^d} g_A(\boldsymbol{x})\,d\boldsymbol{x} - \sum_{i=1}^n w_i g_A(\boldsymbol{a}_i)\right) = \sup_A \int_{[0,1]^d} g_A(\boldsymbol{x})\,d\boldsymbol{x}.$$

[5] Γ is Euler's gamma function.

We denote by $B(\boldsymbol{x}_0, r)$ the ball in $\mathbb{R}^d$ with center $\boldsymbol{x}_0$ and radius r, $B(\boldsymbol{x}_0, r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_0\| \leqslant r\}$. Let us define $A_i = B(\boldsymbol{a}_i, r)$, for $i = 1, \dots, n$. Using this notation we arrive at

$$\begin{aligned}\int_{[0,1]^d} g_A(\boldsymbol{x})\, d\boldsymbol{x} &\geqslant \int_{(\cup_{i=1}^n A_i)^c \cap [0,1]^d} \min\big(1, \min_{\boldsymbol{y}\in A} \|\boldsymbol{x} - \boldsymbol{y}\|\big)\, d\boldsymbol{x} \\ &\geqslant \int_{(\cup_{i=1}^n A_i)^c \cap [0,1]^d} \min\big(1, r\big)\, d\boldsymbol{x}. \end{aligned} \tag{8}$$

We now state an auxiliary lemma, the proof of which is deferred to the end of the section.

Lemma 2. *Let $S = [0,1]^d$ with $d \in \mathbb{N}$ and let μ be a probability measure on $[0,1]^d$ admitting a density (with respect to the Lebesgue measure) bounded by some constant $b < \infty$. For any $r > 0$ and $k \in \mathbb{N}$, if $B_1, \dots, B_k \subseteq \mathbb{R}^d$ are balls of radius r such that*

$$\mu\big(B_1 \cup \ldots \cup B_k\big) > 1/2$$

then,

$$r^d > \frac{1}{2bk\,\mathsf{V}_d} \cdot r^{-d}, \quad \textit{where} \quad \mathsf{V}_d = \mathrm{Vol}(B(\mathbf{0}, 1)). \tag{9}$$

Let us choose $r = (2n\mathsf{V_d})^{-1/d}$. This value of r does not satisfy (9). In view of Lemma 2, this implies that

$$\mu\big(A_1 \cup \ldots \cup A_n\big) \leqslant 1/2.$$

Since $\mu = \mathcal{U}_d$ is a probability measure on $[0,1]^d$, we get

$$\mu\big((A_1 \cup \ldots \cup A_n)^c \cap [0,1]^d\big) > 1/2.$$

Combining this with (8), we arrive at

$$\mathsf{d}_{\mathcal{F}}\Big(\mathcal{U}_d, \sum_{i=1}^n w_i \delta_{a_i}\Big) \geqslant (1/2)\min(1, (2\mathsf{V}_d n)^{-1/d}).$$

In other terms, if $n \geqslant (2\mathsf{V}_d)^{-1}$, then

$$\mathsf{d}_{\mathcal{F}}\Big(\mathcal{U}_d, \sum_{i=1}^n w_i \delta_{\boldsymbol{a}_i}\Big) \geqslant \frac{1}{2}(2\mathsf{V}_d n)^{-1/d} = \frac{1}{2^{(d+1)/d}}(\mathsf{V}_d n)^{-1/d}$$

otherwise

$$\mathsf{d}_{\mathcal{F}}\Big(\mathcal{U}_d, \sum_{i=1}^n w_i \delta_{\boldsymbol{a}_i}\Big) \geqslant \frac{1}{2}.$$

Putting together the inequalities from the last two displays concludes the proof of the desired lower bound. □

Proof of Lemma 2. Let ν be the Lebesgue measure on $[0,1]^d$. We know that $\mu(A) = \int_A \varphi(x)\, \nu(dx)$ with a probability density function φ satisfying $0 \leqslant \varphi(x) \leqslant b$ for all $x \in [0,1]^d$. Therefore,

$$\frac{1}{2} < \mu(B_1 \cup \ldots \cup B_k) \leqslant \sum_{j=1}^k \mu(B_j) = \sum_{j=1}^k \int_{B_j} \varphi(x)\, \nu(dx) \leqslant \sum_{j=1}^k b \cdot \nu(B_j) \tag{10}$$

Moreover, we know that $\nu(B_j) = \mathsf{V}_d r^d$ for all $j = 1, \dots k$. Combining this inequality with (10), we get

$$\frac{1}{2} < kb\,\mathsf{V}_d\, r^d.$$

This yields $r^d > \frac{1}{2bk\,\mathsf{V}_d}$. □

B.2. Proof of Proposition 1

Recall that

$$\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P_{n,X}) = \sup_{f\in\mathcal{F}} \left| \int_{[0,1]^d} f(g(\boldsymbol{u}))\,\mathrm{d}\boldsymbol{u} - \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i)\right|.$$

In addition, there exists an $L_{\mathcal{H}}$-Lipschitz-continuous function h such that $h\circ g = \mathsf{Id}_d$ and

$$\|h(\boldsymbol{x}) - h(\boldsymbol{x}')\| \leqslant L_{\mathcal{H}}\|\boldsymbol{x}-\boldsymbol{x}'\| \qquad \forall \boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^D.$$

In order to establish the lower bound on $\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P_{n,X})$, we use the fact that $\mathcal{F}$ contains all the functions of the form $\psi\circ h$, where $\psi:[0,1]^d\to\mathbb{R}$ is any $(1/L_{\mathcal{H}})$-Lipschitz-continuous function. Indeed, since h is $L_{\mathcal{H}}$-Lipschitz, the function $\psi\circ h$ belongs to $\mathrm{Lip}_1\subseteq\mathcal{F}$. Therefore,

$$\begin{aligned}\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P_{n,X}) &\geqslant \sup_{\psi\in\mathrm{Lip}_{L_{\mathcal{H}}^{-1}}} \left\{ \int_{[0,1]^d} (\psi\circ \underbrace{h\circ g}_{=\mathsf{Id}_d})(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} - \frac{1}{n}\sum_{i=1}^n (\psi\circ h)(\boldsymbol{X}_i)\right\}\\ &= \sup_{\psi\in\mathrm{Lip}_{L_{\mathcal{H}}^{-1}}} \left\{ \int_{[0,1]^d} \psi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \frac{1}{n}\sum_{i=1}^n \psi(\boldsymbol{Z}_i)\right\},\end{aligned}$$

where we have used the notation $\boldsymbol{Z}_i = h(\boldsymbol{X}_i)$, for $i=1,\dots,n$. Clearly, $\psi\in\mathrm{Lip}_{L_{\mathcal{H}}^{-1}}$ is equivalent to $L_{\mathcal{H}}\psi\in\mathrm{Lip}_1$. This implies that

$$\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P_{n,X}) \geqslant \frac{1}{L_{\mathcal{H}}}\sup_{\psi\in\mathrm{Lip}_1}\left\{ \int_{[0,1]^d}\psi(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} - \frac{1}{n}\sum_{i=1}^n\psi(\boldsymbol{Z}_i)\right\}.$$

The right-hand side of the inequality above is precisely the W_1 distance between the uniform measure on $[0,1]^d$ and the empirical distribution of $\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_n$. Therefore, we arrive at

$$\mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P_{n,X}) \geqslant \frac{1}{L_{\mathcal{H}}}\mathsf{W}_1\Big(\mathcal{U}_d, \frac{1}{n}\sum_{i=1}^n\delta_{\boldsymbol{Z}_i}\Big) \geqslant \frac{1}{2L_{\mathcal{H}}(1+(2\mathsf{V}_d n)^{1/d})},$$

where the last inequality follows from Proposition 3.

B.3. Proof of Theorem 1

Let $\widehat{h}_n$ be a function from $\mathcal{H}$ attaining the minimum $\min_{h\in\mathcal{H}}\|h\circ\widehat{g}_{n,\epsilon}-\mathsf{Id}_d\|_{\mathbb{L}_q}^q$. Since $\mathcal{F}$ contains all 1-Lipschitz continuous functions, it also contains all the functions of the form $f=\psi\circ\widehat{h}_n$, for $\psi\in\mathrm{Lip}(1/L_{\mathcal{H}})$. Using the notation $\boldsymbol{Z}_i=\widehat{h}_n(\boldsymbol{X}_i)$, this implies that

$$\begin{aligned}\mathsf{d}_{\mathcal{F}}(\widehat{g}_{n,\epsilon}\sharp\mathcal{U}_d, P_{n,X}) &= \sup_{f\in\mathcal{F}}\left|\int_{[0,1]^d} f(\widehat{g}_{n,\epsilon}(\boldsymbol{x}))d\boldsymbol{x} - \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i)\right|\\ &\geqslant \sup_{\psi\in\mathrm{Lip}_{L_{\mathcal{H}}^{-1}}}\left\{\int_{[0,1]^d}(\psi\circ\widehat{h}_n\circ\widehat{g}_{n,\epsilon})(\boldsymbol{x})\,d\boldsymbol{x} - \frac{1}{n}\sum_{i=1}^n(\psi\circ\widehat{h}_n)(\boldsymbol{X}_i)\right\}\\ &\geqslant \frac{1}{L_{\mathcal{H}}}\sup_{\psi\in\mathrm{Lip}_1}\left\{\int_{[0,1]^d}(\psi\circ\underbrace{\widehat{h}_n\circ\widehat{g}_{n,\epsilon},}_{\approx\mathsf{Id}_d})(\boldsymbol{u})\,d\boldsymbol{u} - \frac{1}{n}\sum_{i=1}^n\psi(\boldsymbol{Z}_i)\right\}.\end{aligned}$$

By adding and subtracting the term $\int_{[0,1]^d}\psi(\boldsymbol{u})\,d\boldsymbol{u}$, we arrive at

$$\begin{aligned}\mathsf{d}_{\mathcal{F}}(\widehat{g}_{n,\epsilon}\sharp\mathcal{U}_d, P_{n,X}) &\geqslant \frac{1}{L_{\mathcal{H}}}\sup_{\psi\in\mathrm{Lip}_1}\left|\int_{[0,1]^d}\psi(\boldsymbol{u})\,d\boldsymbol{u} - \frac{1}{n}\sum_{i=1}^n\psi(\boldsymbol{Z}_i)\right| - \frac{1}{L_{\mathcal{H}}}\sup_{\psi\in\mathrm{Lip}_1}\big\|\psi(\widehat{h}_n\circ\widehat{g}_{n,\epsilon})-\psi\big\|_{\mathbb{L}_1}\\ &\geqslant \frac{1}{L_{\mathcal{H}}}\Big(\mathsf{W}_1(\mathcal{U}_d,\widehat{P}_{n,Z}) - \big\|\widehat{h}_n\circ\widehat{g}_{n,\epsilon}-\mathsf{Id}_d\big\|_{\mathbb{L}_1}\Big)\\ &\geqslant \frac{1}{L_{\mathcal{H}}}\Big(\mathsf{W}_1(\mathcal{U}_d,\widehat{P}_{n,Z}) - \mathrm{pen}_{\mathcal{H}}(\widehat{g}_{n,\epsilon})^{1/q}\Big). \qquad (11)\end{aligned}$$

Here, the second inequality follows from the Lipschitz-continuity of ψ, whereas the last inequality is a consequence of the facts that the $\mathbb{L}_1$ norm on $[0,1]^d$ is dominated by the $\mathbb{L}_q$ norm given $q \geqslant 1$, and that $\widehat{h}_n$ is a minimizer of $\|h \circ \widehat{g}_n - \mathsf{Id}_d\|_q$ over $\mathcal{H}$. We use the same lower bound here

$$\mathsf{W}_1(\mathcal{U}_d, \widehat{P}_{n,Z}) \geqslant \frac{1}{2 + 2(2\mathsf{V}_d n)^{1/d}}. \tag{12}$$

To complete the proof we need to find an upper bound on the second term of the right-hand side of (11). Given (LIPERM), we know that for any $g \in \mathcal{G}$ it holds

$$\mathsf{d}_{\mathcal{F}}(\widehat{g}_{n,\epsilon}\sharp\mathcal{U}_d, P_{n,X}) + \lambda\,\mathrm{pen}_{\mathcal{H}}(\widehat{g}_{n,\epsilon}) \leqslant \mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P_{n,X}) + \lambda\,\mathrm{pen}_{\mathcal{H}}(g) + \epsilon.$$

Since the distance $\mathsf{d}_{\mathcal{F}}$ is always nonnegative, by choosing g from $\mathcal{G}_0$ the last term of the last display vanishes and we get

$$\lambda\,\mathrm{pen}_{\mathcal{H}}(\widehat{g}_{n,\epsilon}) \leqslant \inf_{g\in\mathcal{G}_0} \mathsf{d}_{\mathcal{F}}(g\sharp\mathcal{U}_d, P_{n,X}) + \epsilon. \tag{13}$$

Combining inequalities (11), (12) and (13) we obtain the claim of the theorem.

B.4. Proof of inequality (3)

Using the fact that $g^* \in \mathcal{G}_{\mathcal{H}}$ and the triangle inequality, we get

$$\begin{aligned}\inf_{\mathcal{G}_{\mathcal{H}}} \mathbb{E}\big[\mathsf{W}_1(g\sharp\mathcal{U}_d, P_{n,X})\big] &\leqslant \mathbb{E}\big[\mathsf{W}_1(g^*\sharp\mathcal{U}_d, P_{n,X})\big] \\ &\leqslant \mathbb{E}\big[\mathsf{W}_1(g^*\sharp\mathcal{U}_d, g^*\sharp P_{n,U})\big] + \mathbb{E}\big[\mathsf{W}_1(g^*\sharp P_{n,U}, P_{n,X})\big].\end{aligned} \tag{14}$$

Let $\boldsymbol{U}_i \sim \mathcal{U}_d$ be n iid random vectors drawn from the uniform distribution (they will be defined more specifically later in the proof). Let $P_{n,U}$ be the empirical distribution of $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$. Recall that (see, for example, (Niles-Weed and Rigollet, 2022, Proposition 1))

$$\mathbb{E}[\mathsf{W}_1(\mathcal{U}_d, P_{n,U})] \leqslant \frac{c\sqrt{d}}{n^{1/d}}.$$

This implies that

$$\mathbb{E}\big[\mathsf{W}_1(g^*\sharp\mathcal{U}_d, g^*\sharp P_{n,U})\big] \leqslant L^*\mathbb{E}\big[\mathsf{W}_1(\mathcal{U}_d, P_{n,U})\big] \leqslant \frac{cL^*\sqrt{d}}{n^{1/d}}. \tag{15}$$

On the other hand, it is clear that

$$\begin{aligned}\mathbb{E}[\mathsf{W}_1(g^*\sharp P_{n,U}, P_{n,X})] &\leqslant \mathbb{E}\left[(1/n)\sum_{i=1}^n \|g^*(\boldsymbol{U}_i) - \boldsymbol{X}_i\|\right] \\ &= \mathbb{E}\big[\|g^*(\boldsymbol{U}_1) - \boldsymbol{X}_1\|\big].\end{aligned} \tag{16}$$

If we assume now that $\boldsymbol{U}_1$ is chosen in such a way that the joint distribution of $g^*(\boldsymbol{U}_1)$ and $\boldsymbol{X}_1$ is the optimal coupling between the marginal distributions of these two random vectors, we get

$$\mathbb{E}\big[\|g^*(\boldsymbol{U}_1) - \boldsymbol{X}_1\|\big] = \mathsf{W}_1(g^*\sharp\mathcal{U}_d, P^*) \leqslant \sigma^*. \tag{17}$$

Combining (16) and (17), we get

$$\mathbb{E}[\mathsf{W}_1(g^*\sharp P_{n,U}, P_{n,X})] \leqslant \sigma^*. \tag{18}$$

Finally, inequalities (15) and (18), in conjunction with (14), yield

$$\inf_{\mathcal{G}_{\mathcal{H}}} \mathbb{E}\big[\mathsf{W}_1(g\sharp\mathcal{U}_d, P_{n,X})\big] \leqslant \frac{cL^*\sqrt{d}}{n^{1/d}} + \sigma^*.$$

C. Proof of the risk bound when trained with approximated functional classes

Although Proposition 2 was stated for the case $q = 2$ only, we provide the proof for any $q \geqslant 1$. Replacing q by 2 in the final expression of this proof leads to the claim of the proposition.

C.1. Proof of Proposition 2

We consider a trained generator $\widehat{g}^0_{n,\varepsilon}$ satisfying

$$\mathsf{d}_{\mathcal{F}_0}(\widehat{g}^0_{n,\varepsilon}\sharp\, U_d, P_{n,X}) + \mathrm{pen}_{\mathcal{H}_0}(\widehat{g}^0_{n,\varepsilon}) \leqslant \mathsf{d}_{\mathcal{F}_0}(g\sharp\, U_d, P_{n,X}) + \mathrm{pen}_{\mathcal{H}_0}(g) + \varepsilon, \qquad \text{for all } g \in \mathcal{G}_0 \tag{19}$$

and our goal is to upper bound the expression $\mathsf{d}_{\mathcal{F}}(\widehat{g}^0_{n,\varepsilon}\sharp\, U_d, P_{n,X})$. Using Lemma 3, we have

$$\mathsf{d}_{\mathcal{F}}(\widehat{g}^0_{n,\varepsilon}\sharp\, U_d, P_{n,X}) \leqslant \mathsf{d}_{\mathcal{F}_0}(\widehat{g}^0_{n,\varepsilon}\sharp\, U_d, P_{n,X}) + 2\delta_{\mathcal{F}}. \tag{20}$$

The first term of the right-hand side can be upper bounded as follows:

$$\begin{aligned}
\mathsf{d}_{\mathcal{F}_0}(\widehat{g}^0_{n,\varepsilon}\sharp\, U_d, P_{n,X}) &\leqslant \mathsf{d}_{\mathcal{F}_0}(\widehat{g}^0_{n,\varepsilon}\sharp\, U_d, P_{n,X}) + \lambda \mathrm{pen}_{\mathcal{H}_0}(\widehat{g}^0_{n,\varepsilon}) \\
&\leqslant \inf_{g_0\in\mathcal{G}_0} \Big(\mathsf{d}_{\mathcal{F}_0}(g_0\sharp\, U_d, P_{n,X}) + \lambda \mathrm{pen}_{\mathcal{H}_0}(g_0)\Big) + \varepsilon \\
&\leqslant \inf_{g_0\in\mathcal{G}_0} \Big(\mathsf{d}_{\mathcal{F}_0}(g_0\sharp\, U_d, P_{n,X}) + \lambda \mathrm{pen}_{\mathcal{H}}(g_0)\Big) + \varepsilon + q d^{(q-1)/2}\delta_{\mathcal{H}} \\
&\leqslant \inf_{g_0\in\mathcal{G}_0} \Big(\mathsf{d}_{\mathcal{F}}(g_0\sharp\, U_d, P_{n,X}) + \lambda \mathrm{pen}_{\mathcal{H}}(g_0)\Big) + \varepsilon + q d^{(q-1)/2}\delta_{\mathcal{H}}.
\end{aligned} \tag{21}$$

where we used the positiveness of the penalty function for the first inequality, inequality (19) for the second inequality, Lemma 5 for the third inequality and the fact that $\mathcal{F}_0 \subseteq \mathcal{F}$ for the fourth inequality.

The last step is to use Lemma 6, which allows to upper bound the inf over $\mathcal{G}_0$ by the inf over $\mathcal{G}$, modulo an additive error term proportional to $\delta_{\mathcal{G}}$. More precisely,

$$\inf_{g_0\in\mathcal{G}_0} \Big(\mathsf{d}_{\mathcal{F}}(g_0\sharp\, U_d, P_{n,X}) + \lambda \mathrm{pen}_{\mathcal{H}}(g_0)\Big) \leqslant \inf_{g\in\mathcal{G}} \Big(\mathsf{d}_{\mathcal{F}}(g\sharp\, U_d, P_{n,X}) + \lambda \mathrm{pen}_{\mathcal{H}}(g)\Big) + \big(1 + \lambda q d^{(q-1)/2} L_{\mathcal{H}}\big)\delta_{\mathcal{G}}. \tag{22}$$

Combining (20), (21) and (22), we get the inequality

$$\mathsf{d}_{\mathcal{F}}(\widehat{g}^0_{n,\varepsilon}\sharp\, U_d, P_{n,X}) \leqslant \inf_{g\in\mathcal{G}} \Big(\mathsf{d}_{\mathcal{F}}(g\sharp\, U_d, P_{n,X}) + \lambda \mathrm{pen}_{\mathcal{H}}(g)\Big) + \varepsilon + \underbrace{2\delta_{\mathcal{F}} + q d^{(q-1)/2}\delta_{\mathcal{H}} + \big(1 + \lambda q d^{(q-1)/2} L_{\mathcal{H}}\big)\delta_{\mathcal{G}}}_{=:\delta}.$$

This completes the proof.

C.2. Impact of approximating $\mathcal{F}$ on the IPM

Lemma 3. *If $\mathcal{F}_0$ is such that $\inf_{f_0\in\mathcal{F}_0} \|f - f_0\| \leqslant \delta$ for every $f \in \mathcal{F}$, then*

$$\mathsf{d}_{\mathcal{F}}(P, Q) - \mathsf{d}_{\mathcal{F}_0}(P, Q) \leqslant 2\delta \quad \text{for all distributions} \quad P, Q.$$

Proof. Recall the definition of $\mathsf{d}_{\mathcal{F}}(P, Q) = \sup_{f\in\mathcal{F}} \big|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\big|$. This implies that

$$\begin{aligned}
\mathsf{d}_{\mathcal{F}}(P, Q) - \mathsf{d}_{\mathcal{F}_0}(P, Q) &= \sup_{f\in\mathcal{F}} \inf_{f_0\in\mathcal{F}_0} \Big(\underbrace{\big|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\big|}_{\text{independent of } f_0} - \underbrace{\big|\mathbb{E}_P[f_0(X)] - \mathbb{E}_Q[f_0(X)]\big|}_{\text{independent of } f} \Big) \\
&\leqslant \sup_{f\in\mathcal{F}} \inf_{f_0\in\mathcal{F}_0} \big|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] - \mathbb{E}_P[f_0(X)] - \mathbb{E}_Q[f_0(X)]\big| \quad (|a| - |b| \leqslant |a - b|) \\
&\leqslant \sup_{f\in\mathcal{F}} \inf_{f_0\in\mathcal{F}_0} \Big(\big|\mathbb{E}_P[f(X)] - \mathbb{E}_P[f_0(X)]\big| + \big|\mathbb{E}_Q[f(X)] - \mathbb{E}_Q[f_0(X)]\big|\Big) \qquad \text{(triangle ineq.)} \\
&\leqslant \sup_{f\in\mathcal{F}} \inf_{f_0\in\mathcal{F}_0} \Big(\mathbb{E}_P[\underbrace{|f(X) - f_0(X)|}_{\leqslant \|f - f_0\|_\infty}] + \mathbb{E}_Q[\underbrace{|f(X) - f_0(X)|}_{\leqslant \|f - f_0\|_\infty}]\Big) \\
&\leqslant \sup_{f\in\mathcal{F}} \inf_{f_0\in\mathcal{F}_0} \|f - f_0\|_\infty \leqslant 2\delta.
\end{aligned}$$

This completes the proof of the lemma. □

C.3. Impact of approximating $\mathcal{H}$ on the Left-Inverse-Penalty

Before analyzing the sensitivity of the left-inverse-penalty to the deviations from $\mathcal{H}$, we need an auxiliary lemma.

Lemma 4. *If a, b are arbitrary numbers from some interval $[0, C]$, and $q \geqslant 1$, then*

$$|a^q - b^q| \leqslant qC^{q-1}|b - a|.$$

Proof. Let us first assume that $c \in [0, 1]$. For any $q \geqslant 1$,

$$|c^q - 1| \leqslant q|c - 1|.$$

Then for $a, b \in \mathbb{R}$ such that $0 \leqslant a \leqslant b$, we have

$$|a^q - b^q| = b^q\big|(a/b)^q - 1\big| \leqslant qb^q\Big|\frac{a}{b} - 1\Big| = qb^{q-1}|b - a|.$$

The claim of the lemma follows by upper bounding b by C. □

Lemma 5. *If $\mathcal{H}_0$ is such that $\min\limits_{h_0 \in \mathcal{H}_0} \|h - h_0\|_\infty \leqslant \delta$ for all $h \in \mathcal{H}$, then*

$$\mathrm{pen}_{\mathcal{H}_0}(g) - \mathrm{pen}_{\mathcal{H}}(g) \leqslant qd^{(q-1)/2}\delta, \qquad \textit{for all } g \in \mathcal{G}.$$

Proof. Recall that $\mathrm{pen}_{\mathcal{H}}(g) = \min_{h \in \mathcal{H}} \|h \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}^q$. This yields

$$\begin{aligned}
\mathrm{pen}_{\mathcal{H}_0}(g) - \mathrm{pen}_{\mathcal{H}}(g) &= \min_{h_0 \in \mathcal{H}_0} \|h_0 \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}^q - \min_{h \in \mathcal{H}} \|h \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}^q \\
&= \max_{h \in \mathcal{H}} \min_{h_0 \in \mathcal{H}_0} \Big(\|h_0 \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}^q - \|h \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}^q\Big)
\end{aligned}$$

We apply Lemma 4 with

$$\begin{aligned}
a &= \|h_0 \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q} = \Big(\int_{[0,1]^d} \|\underbrace{h_0(g(u))}_{\in [0,1]^d} - \underbrace{u}_{\in [0,1]^d}\|^q \,\mathrm{d}u\Big)^{1/q} \leqslant \sqrt{d} \\
b &= \|h \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q} \leqslant \sqrt{d}.
\end{aligned}$$

This leads to

$$\begin{aligned}
\|h_0 \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}^q - \|h \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}^q &\leqslant qd^{(q-1)/2}\Big|\|h_0 \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q} - \|h \circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}\Big| \quad (|\|a\| - \|b\|| \leqslant \|a - b\|) \\
&\leqslant qd^{(q-1)/2}\|h_0 \circ g - \mathsf{Id}_d - h \circ g + \mathsf{Id}_d\|_{\mathbb{L}_q} \\
&\leqslant qd^{(q-1)/2}\|h_0 \circ g - h \circ g\|_{\mathbb{L}_q} \\
&= qd^{(q-1)/2}\|h_0 - h\|_\infty \leqslant qd^{(q-1)/2}\delta.
\end{aligned}$$

This completes the proof of the lemma □

C.4. Impact of approximating $\mathcal{G}$ on the Penalized Empirical Risk

As in the main text, here also we assume that the elements of the set $\mathcal{H}$ are all Lipschitz-continuous with a Lipschitz constant bounded by $L_{\mathcal{H}}$.

Lemma 6. *Let P be an arbitrary distribution and $\mathcal{F} \subseteq \mathrm{Lip}_1(\mathbb{R}^D \to \mathbb{R})$. If $\mathcal{G}_0$ is such that $\min_{g_0 \in \mathcal{G}_0} \|g - g_0\|_\infty \leqslant \delta$ for every $g \in \mathcal{G}$, then the following is true:*

$$\min_{g_0 \in \mathcal{G}_0} \Big(\mathsf{d}_{\mathcal{F}}(g_0 \sharp U_d, P) + \lambda \mathrm{pen}_{\mathcal{H}}(g_0)\Big) \leqslant \inf_{g \in \mathcal{G}} \mathsf{d}_{\mathcal{F}}(g \sharp U_d, P) + \lambda \mathrm{pen}_{\mathcal{H}}(g) + \big(1 + \lambda q d^{(q-1)/2} L_{\mathcal{H}}\big)\delta.$$

Proof. Let g be any function from $\mathcal{G}$ and g_0 be the element of $\mathcal{G}_0$ satisfying $\|g - g_0\|_\infty \leqslant \delta$. On the one hand, for this function g_0, we have

$$\begin{aligned}
\mathsf{d}_{\mathcal{F}}(g_0\sharp U_d, P) - \mathsf{d}_{\mathcal{F}}(g\sharp U_d, P) &= \sup_{f\in\mathcal{F}} \Big|\mathbb{E}_{\mathcal{U}_d}[f(g_0(U))] - \mathbb{E}_P[f(X)]\Big| - \sup_{f\in\mathcal{F}} \Big|\mathbb{E}_{\mathcal{U}_d}[f(g(U))] - \mathbb{E}_P[f(X)]\Big| \\
&\leqslant \sup_{f\in\mathcal{F}} \Big|\mathbb{E}_{\mathcal{U}_d}\Big[f(g_0(U)) - f(g(U))\Big]\Big| \\
&\leqslant \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{U}_d}\Big[\Big|f(g_0(U)) - f(g(U))\Big|\Big] \\
&\leqslant \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{U}_d}\Big[\big\|g_0(U) - g(U)\big\|\Big] \leqslant \delta. \qquad (23)
\end{aligned}$$

Here, the inequality of the second line follows from $\sup|F| - \sup|G| \leqslant \sup(|F| - |G|) \leqslant \sup|F - G|$, while the first inequality of the last line is a consequence of the assumption that the functions from $\mathcal{F}$ are all 1-Lipschitz.

On the other hand,

$$\begin{aligned}
\mathrm{pen}_{\mathcal{H}}(g_0) - \mathrm{pen}_{\mathcal{H}}(g) &= \min_{h\in\mathcal{H}} \|h\circ g_0 - \mathsf{Id}_d\|^q_{\mathbb{L}_q} - \min_{h\in\mathcal{H}} \|h\circ g - \mathsf{Id}_d\|^q_{\mathbb{L}_q} \\
&\leqslant \|h^*\circ g_0 - \mathsf{Id}_d\|^q_{\mathbb{L}_q} - \|h^*\circ g - \mathsf{Id}_d\|^q_{\mathbb{L}_q},
\end{aligned}$$

where h^* is the minimizer of $\|h\circ g - \mathsf{Id}_d\|^q_{\mathbb{L}_q}$ over $\mathcal{H}$. Combining with Lemma 4, we get

$$\begin{aligned}
\mathrm{pen}_{\mathcal{H}}(g_0) - \mathrm{pen}_{\mathcal{H}}(g) &\leqslant qd^{(q-1)/2}\Big|\|h^*\circ g_0 - \mathsf{Id}_d\|_{\mathbb{L}_q} - \|h^*\circ g - \mathsf{Id}_d\|_{\mathbb{L}_q}\Big| \\
&\leqslant qd^{(q-1)/2}\|h^*\circ g_0 - \mathsf{Id}_d - h^*\circ g + \mathsf{Id}_d\|_{\mathbb{L}_q} \\
&= qd^{(q-1)/2}\|h^*\circ g_0 - h^*\circ g\|_{\mathbb{L}_q} \\
&\leqslant qd^{(q-1)/2}\|h^*\circ g_0 - h^*\circ g\|_\infty \\
&\leqslant qd^{(q-1)/2}L_{\mathcal{H}}\|g_0 - g\|_\infty \\
&\leqslant qd^{(q-1)/2}L_{\mathcal{H}}\delta. \qquad (24)
\end{aligned}$$

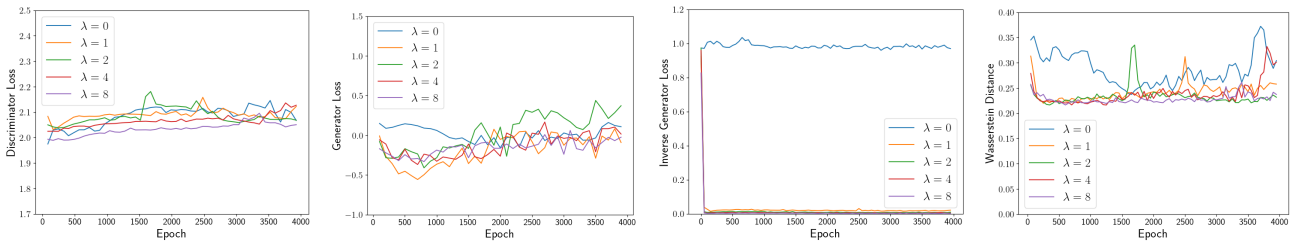
Combining (23) and (24), we get that for any $g\in\mathcal{G}$, there is a $g_0\in\mathcal{G}_0$ such that

$$\mathsf{d}_{\mathcal{F}}(g_0\sharp U_d, P) + \lambda\mathrm{pen}_{\mathcal{H}}(g_0) \leqslant \mathsf{d}_{\mathcal{F}}(g\sharp U_d, P) + \lambda\mathrm{pen}_{\mathcal{H}}(g) + \big(1 + \lambda qd^{(q-1)/2}L_{\mathcal{H}}\big)\delta.$$

This completes the proof of the lemma. □

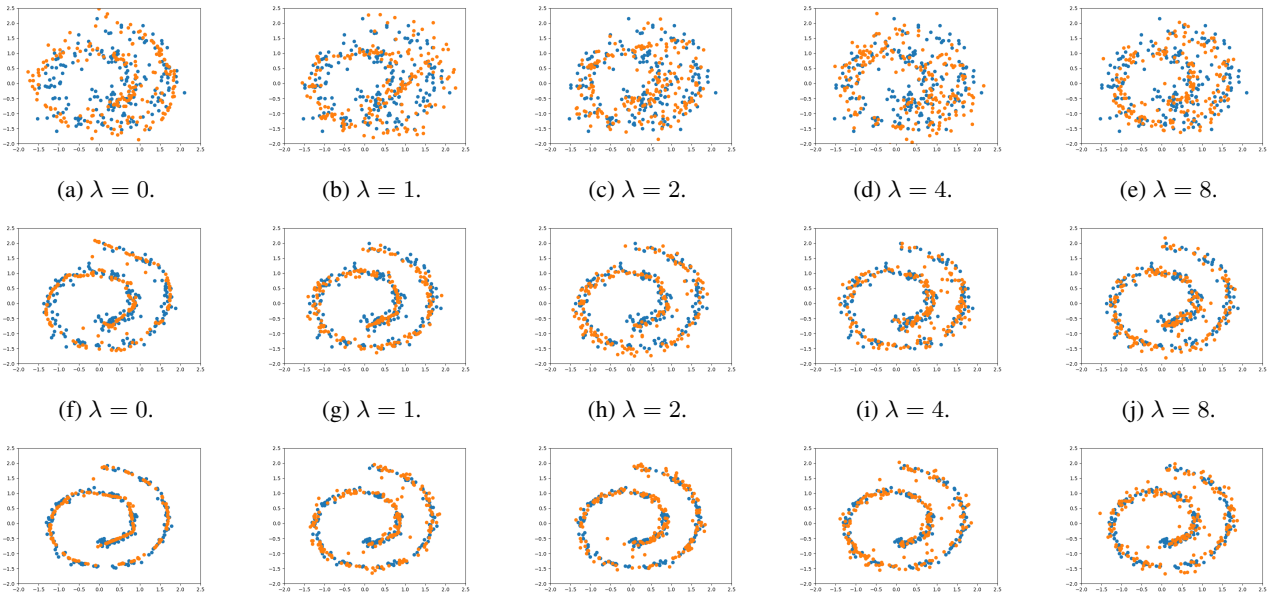
D. Numerical experiments

This section contains some tables and plots that we could not include in the main paper due to the space restrictions. Recall that our experiments were done on three data sets: Swiss roll, MNIST and CIFAR-10. In these experiments, we trained a distribution using the Wasserstein GAN with a left inverse penalty.



(a) Discriminator Loss (b) Generator Loss (c) Inverse Generator Loss (d) Wasserstein distance

Figure 7. Evolution of various losses across the iterations in the experiment on Swiss Roll data generated with the noise magnitude $\sigma = 1.5$.

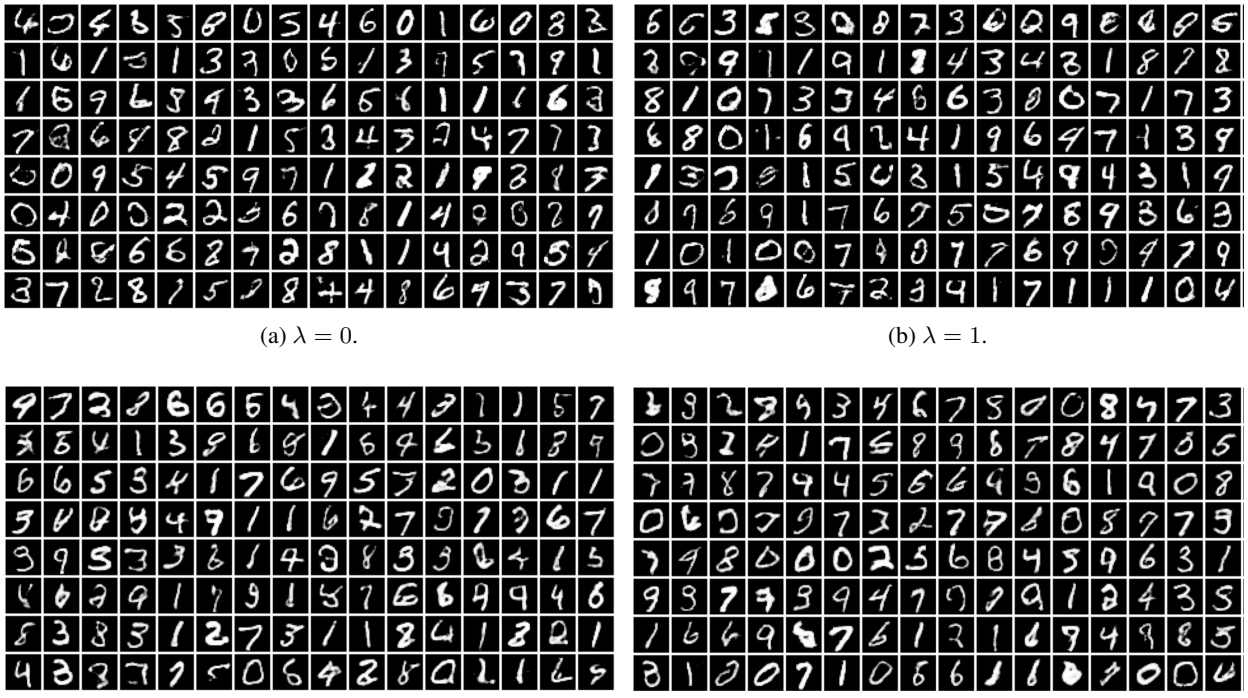


(a) $\lambda = 0$. (b) $\lambda = 1$. (c) $\lambda = 2$. (d) $\lambda = 4$. (e) $\lambda = 8$.

(f) $\lambda = 0$. (g) $\lambda = 1$. (h) $\lambda = 2$. (i) $\lambda = 4$. (j) $\lambda = 8$.

(k) $\lambda = 0$. (l) $\lambda = 1$. (m) $\lambda = 2$. (n) $\lambda = 4$. (o) $\lambda = 8$.

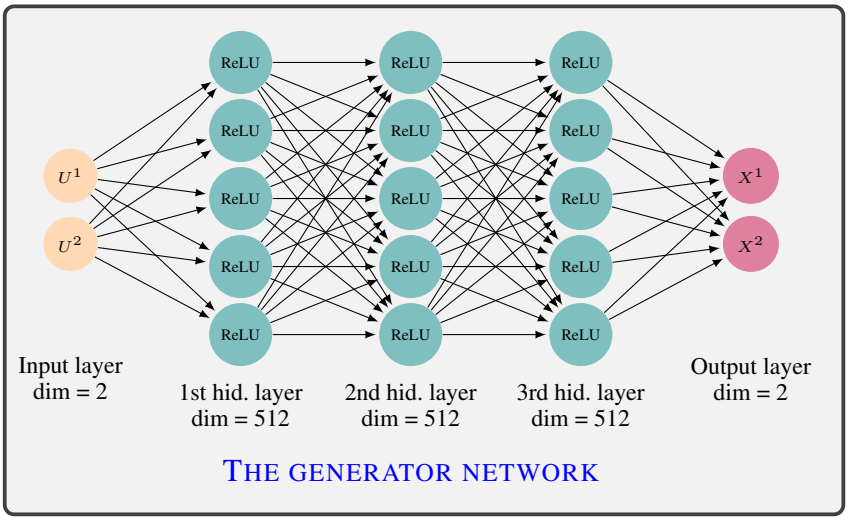
Figure 8. Swiss Roll data: training set and examples generated from the learned distribution, when the noise magnitude is $\sigma = 3/2$ (top row), $\sigma = 3/4$ middle row and $\sigma = 3/8$ (bottom row). LIPERM was trained with $\lambda = 0, 1, 2, 4, 8$.



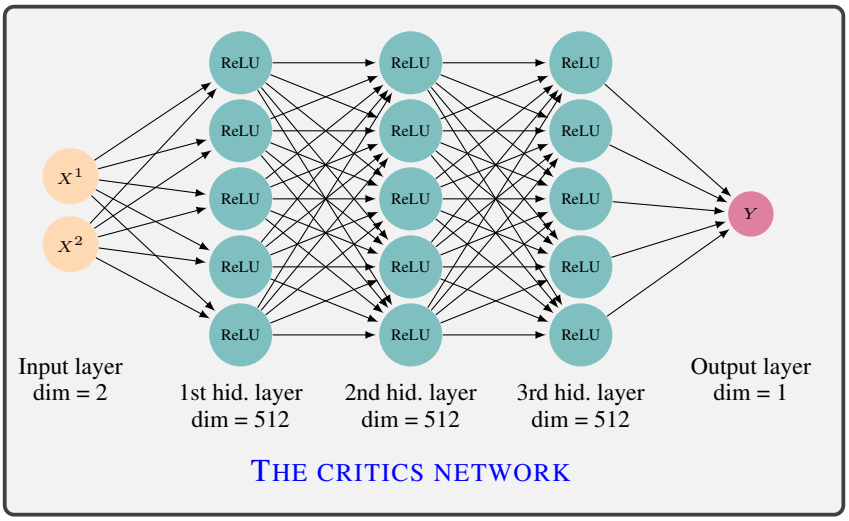
(a) $\lambda = 0$. (b) $\lambda = 1$.

(c) $\lambda = 4$. (d) $\lambda = 8$.

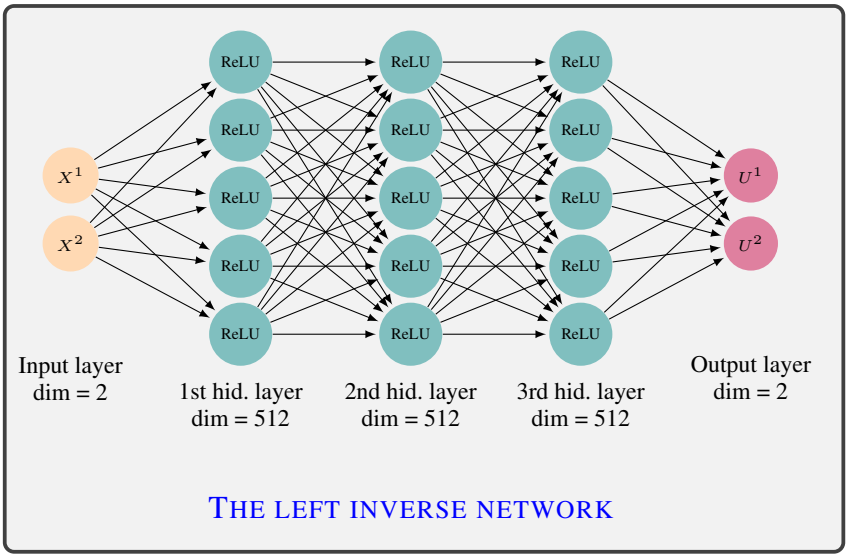
Figure 9. Handwritten digits generated by LIPERM for different values of λ.



Layer	Operation
1	Linear + ReLU latent_DIM → DIM
2	Linear + ReLU DIM → DIM
3	Linear + ReLU DIM → DIM
Out	Linear DIM → out_DIM



Layer	Operation
1	Linear + ReLU out_DIM → DIM
2	Linear + ReLU DIM → DIM
3	Linear + ReLU DIM → DIM
Out	Linear DIM → 1



Layer	Operation
1	Linear + LeakyReLU out_DIM → DIM
2	Linear + LeakyReLU DIM → DIM
3	Linear + LeakyReLU DIM → DIM
Out	Linear DIM → latent_DIM

Table 1. Neural network architectures for the generator g, the critic f and the left inverse h used in the experiments conducted on Swiss Roll datasets. In this case, latent_DIM = out_DIM = 2, DIM = 512.

(a) Discriminator Network Architecture used in our experiments on MNIST data set.

Layer	Conv2d + LeakyReLu	Conv2d + BatchNorm2d + LeakyRelu	Conv2d + BatchNorm2d + LeakyRelu	Conv2d
Input dim	28×28	14×14	7×7	3×3
Nb input channels	1	256	512	1024
Kernel size	4	4	3	3
Stride	2	2	2	1
Padding	1	1	0	0
Output dim	14×14	7×7	3×3	1×1
Nb output channels	256	512	1024	1

(b) Generator Network Architecture used in our experiments on MNIST data set.

Layer	ConvTranspose2d +BatchNorm2d +ReLu	ConvTranspose2d +BatchNorm2d +Relu	ConvTranspose2d +BatchNorm2d +Relu	ConvTranspose2d
Input dim	1×1	3×3	7×7	3×3
Nb input channels	100	1024	512	256
Kernel size	3	3	4	4
Stride	1	2	2	2
Padding	0	0	1	1
Output dim	3×3	7×7	14×14	28×28
Nb output channels	1024	512	256	1

(c) Inverse Generator Network Architecture used in our experiments on MNIST data set.

Layer	Conv2d + LeakyReLu	Conv2d + BatchNorm2d + LeakyRelu	Conv2d + BatchNorm2d + LeakyRelu	Conv2d	Linear
Input dim	28×28	14×14	7×7	3×3	100
Nb input channels	1	256	512	1024	1
Kernel size	4	4	3	3	-
Stride	2	2	2	1	-
Padding	1	1	0	0	-
Output dim	14×14	7×7	3×3	1×1	100
Nb output channels	256	512	1024	100	1

Table 2. Neural network architectures for the generator g, the critic f, and the lest inverse h used in the experiments conducted on MNIST dataset. The negative slope parameter of the LeakyReLU is set to 0.2.