Journal of Machine Learning Research 24 (2023) 1-100 Submitted 12/20; Revised 2/23; Published 4/23

Euler-Lagrange Analysis of Generative Adversarial Networks

Siddarth Asokan SIDDARTHA@IISC.AC.IN
Robert Bosch Centre for Cyber-Physical Systems
Indian Institute of Science
Bengaluru-560012, India

Chandra Sekhar Seelamantula CSS@IISC.AC.IN
Department of Electrical Engineering
Indian Institute of Science
Bengaluru-560012, India

Editor: Shakir Mohamed

Abstract

We consider Generative Adversarial Networks (GANs) and address the underlying *functional* optimization problem *ab initio* within a variational setting. Strictly speaking, the optimization of the generator and discriminator *functions* must be carried out in accordance with the Euler-Lagrange conditions, which become particularly relevant in scenarios where the optimization cost involves regularizers comprising the derivatives of these functions. Considering Wasserstein GANs (WGAN) with a gradient-norm penalty, we show that the optimal discriminator is the solution to a Poisson differential equation. In principle, the optimal discriminator can be obtained in closed form without having to train a neural network. We illustrate this by employing a Fourier-series approximation to solve the Poisson differential equation. Experimental results based on synthesized Gaussian data demonstrate superior convergence behavior of the proposed approach in comparison with the baseline WGAN variants that employ weight-clipping, gradient or Lipschitz penalties on the discriminator on low-dimensional data. We also analyze the truncation error of the Fourier-series approximation and the estimation error of the Fourier coefficients in a high-dimensional setting. We demonstrate applications to real-world images considering latent-space prior matching in Wasserstein autoencoders and present performance comparisons on benchmark datasets such as MNIST, SVHN, CelebA, CIFAR-10, and Ukiyo-E. We demonstrate that the proposed approach achieves comparable reconstruction error and Fréchet inception distance with faster convergence and up to two-fold improvement in image sharpness.

Keywords: Generative adversarial networks, Calculus of variations, Euler-Lagrange conditions, Fourier-series approximation, Wasserstein autoencoder.

1 Introduction

The optimization of a generative adversarial network (GAN), originally proposed by Goodfellow et al. (2014), (Standard GAN, or SGAN) is a *min-max* game between two players — a generator (G) and a discriminator (D). The role of the generator is to create fake samples that mimic the ones coming from the training data distribution. The discriminator D is tasked with telling apart the real samples from the fake ones. The optimal G is the one that *outsmarts* D into confusing the fake samples for real. The SGAN optimization comprises the generator and the discriminator with respective loss functions. The generator

G accepts high-dimensional noise $\boldsymbol{z} \sim p_\ell$ as input and generates fake samples $G(\boldsymbol{z}) \sim p_g$. The discriminator D accepts an input $\boldsymbol{x}$, which could come from either the data distribution p_d, or the generator distribution p_g, and outputs a value $D(\boldsymbol{x})$. Effectively, the generator must learn a mapping from the noise distribution to the data distribution, whereas the discriminator must learn the optimal two-class classifier. Over the past decade, numerous variants of GANs have been proposed with several successful applications. Almost all known GAN flavors minimize either a divergence metric or an integral probability metric. In the following, we review important GANs under each category.

Divergence-minimizing GANs: GANs were originally designed to minimize the divergence between the *true* data distribution p_d and the generator distribution p_g. For instance, the SGAN formulation minimizes the Jensen-Shannon divergence, whereas the least-squares GAN (LSGAN) (Mao et al., 2017) is optimized for the Pearson-χ^2 divergence. The f-GAN formulation (Nowozin et al., 2016) is a generalization that includes a chosen f-divergence. In the original SGAN, the discriminator output is constrained to the interval $[0, 1]$, representing the probability of the input sample being real or fake, whereas LSGAN requires the discriminator output to match the chosen class-labels. In f-GANs, the discriminator output is real-valued, but the activation function maps it to a desired interval.

Integral probability metric GANs: In certain GAN flavors, the *discriminator* is replaced with a *real-valued critic* C that differentiates between the generator and data distributions in terms of an integral probability metric (IPM) defined over the class of critics to choose from. The choice of the class of critics gives rise to variants such as the Wasserstein GAN (WGAN) (Arjovsky et al., 2017) with a Lipschitz-1 critic, the minimum-mean discrepancy GAN (MMD-GAN) (Li et al., 2017) where the critic is bounded by a ball in a reproducing-kernel Hilbert space, or the Fisher GAN (Mroueh and Sercu, 2017) in which the second-order moments of the critic are constrained to be bounded. Sobolev GANs (Mroueh et al., 2018) favor critics with a finite energy in the gradient. The critic is a neural network similar to the discriminator, where the constraints are enforced appropriately, either by means of an adjustment of the network weights (Arjovsky et al., 2017; Roth et al., 2019; Wang and Liu, 2016), or through a suitable penalty incorporated into the loss function (Gulrajani et al., 2017; Roth et al., 2017; Mescheder et al., 2018). In our formulation, we use the term *discriminator* $D(\boldsymbol{x})$ to refer to either a divergence-based discriminator or the IPM-based critic with the context resolving any ambiguity.

In this paper, we will primarily focus on the IPM loss as considered in the context of WGANs. Within this framework, the regularized optimization problem takes the form

$$\min_{p_g} \max_{D} \left\{ \mathbb{E}_{\boldsymbol{x} \sim p_d}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x} \sim p_g}[D(\boldsymbol{x})] + \Omega(D(\boldsymbol{x})) \right\},$$

where $\mathbb{E}$ denotes the expectation operator and Ω is a suitable regularizer on D. The objective of the min-max optimization is to ensure that the optimal generator distribution $p_g^*(\boldsymbol{x})$ matches the data distribution $p_d(\boldsymbol{x})$. Typically, one considers gradient-based regularizers, which enforce smoothness on the discriminator.

The stupendous success of GANs in generating realistic images has resulted in significant efforts trying to explain them analytically. While divergence-minimizing GANs are analyzed in a probabilistic setting, IPM based GANs have been analyzed within the framework of optimal transport (Sanjabi et al., 2018; Bousquet et al., 2017; Lei et al., 2019). The energy

based GAN (EBGAN) (Zhao et al., 2017) and related works (Finn et al., 2016; Che et al., 2020) interpret the GAN loss as an energy function that assigns large values to regions of the manifold where there is a mismatch between the generator and data distributions. The min-max game could also be viewed as an instance of *imitation learning* (Finn et al., 2016; Ho and Ermon, 2016; Grnarova et al., 2018), drawing parallels to reinforcement and online learning paradigms. GANs have also been analyzed in a game-theoretic setting with convergence to the Nash equilibrium (Oliehoek et al., 2019; Fedus et al., 2018; Gao and Tembine, 2018), and in an information-theoretic setting, for instance, in the context of entropy minimization for interpolation in the latent space (Chen et al., 2016), or stochastic procedures to estimate ratios of densities and functions thereof (Mohamed and Lakshminarayanan, 2016). The convergence guarantees of various GAN training algorithms have been analyzed in a series of contributions (Salimans et al., 2016; Gulrajani et al., 2017; Roth et al., 2017; Kodali et al., 2017; Mescheder et al., 2018; Li et al., 2018).

On GANs and Partial Differential Equations: GANs have been employed to solve stochastic differential equations and various classical partial differential equations (PDEs) encountered in the context of harmonic oscillation, nonlinear oscillation (Yang et al., 2019), and more recently, infectious disease modelling (Randle et al., 2020), to name a few. On the flip side, ordinary differential equations (ODEs) have been used for GAN training. More specifically, the stability of GANs has been shown to improve when the discrete gradient-descent based network updates were replaced with a numerical solver for the corresponding ODE (Qin et al., 2020). The discriminator and generator are first parametrized as neural networks and subsequently solved for via an ODE. In contrast with these approaches, we view the GAN training problem as a regularized functional optimization problem, which cannot always be addressed using point-wise optimization. The argument becomes more compelling when gradient-based penalties are incorporated. The proposed variational approach is generic and subsumes the unregularized formulations, for instance, the original GAN formulation of Goodfellow et al. (2014). The optimization requires one to solve a PDE. This formalism yields new insights into the interplay between the generator and the discriminator. The Sobolev GAN formulation (Mroueh et al., 2018), which employs the IPM comes closest to our approach. Preliminary connections between the Sobolev GAN and the Fokker-Planck PDE were established in Mroueh et al. (2018). However, an in-depth analysis of the PDE was not carried out, and the connection was not leveraged to optimize the GAN more efficiently. Sobolev GAN implementation ultimately relies on an empirical approach for computing the gradient-based penalty in the optimization (Gulrajani et al., 2017). We show how the PDE connection can be leveraged to make the GAN optimization more efficient and insightful.

1.1 Our Contributions

In this paper, we analyze GANs within a variational framework by enforcing the Euler-Lagrange (EL) conditions to determine the optimum. In scenarios where the GAN loss does not involve derivative terms, the EL conditions degenerate to performing point-wise optimization. Unlike the existing results in the GAN literature, we explicitly enforce essential conditions, namely non-negativity and area under the density equal to unity, which the optimal generator distribution must satisfy in order to qualify as a valid density. We carry out the analysis for several GAN flavors: SGAN, LSGAN, and f-GANs (Appendix A).

To concretely demonstrate the importance and efficacy of the Euler-Lagrange variational framework, we consider the Wasserstein GAN loss proposed by Arjovsky et al. (2017), but with a difference — we consider a gradient-norm penalty on the discriminator, which is a novel variant of the penalty proposed by Mroueh et al. (2018). The penalized Wasserstein GAN loss necessitates Euler-Lagrange analysis because the new optimization objective involves gradient terms. The chosen gradient-norm penalization has an interesting consequence — the optimal discriminator, given the generator, turns out to be the solution to the Poisson PDE. In principle, the optimal discriminator could be determined in a single shot. By analyzing the PDE, first in 1-D, and subsequently, generalizing to higher dimensions, we show that the corresponding optimal generator, given the optimal discriminator, learns the desired target distribution (Sections 3 - 4). Our formulation also allows one to determine the optimal Lagrange multiplier in the gradient-norm-penalized Wasserstein GAN loss.

Since the GAN optimization alternates between the discriminator and generator, one could start with an initial generator distribution and determine the corresponding optimal discriminator single-shot by solving the Poisson PDE. In the next step, we solve for the generator distribution using the optimal discriminator. Our analysis also shows that the optimal generator distribution coincides with the data distribution.

In a real-world scenario, we only have discrete data, and solving the PDE becomes impractical. In such circumstances, we resort to an approximate solution by considering a Fourier-series representation of the discriminator. The choice of the Fourier bases is motivated by the fact that they are eigenfunctions of linear differential operators. Also, the separability and orthogonality properties of the multidimensional Fourier bases give rise to a computationally elegant approach to finding the solution, which obviates the need to optimize a neural network. The underlying formulation remains continuous, while the computations are carried out in discrete. We use the Fourier-series approximation mainly to illustrate the point and to serve as a proof of concept, although in principle, one could employ alternative and possibly more parsimonious bases expansions. In order to substantiate the developments, we provide experimental validations employing simulated unimodal and multimodal Gaussian data. Since the Fourier-series model complexity increases exponentially with the dimension of data, we resort to truncation and sampling of the Fourier coefficients. The superior performance of the Fourier-series approximations in lower dimensions motivates us to consider the Wasserstein autoencoder (WAE), where we replace the neural network discriminator with a Fourier-series solver operating in the latent space (Section 6).

We present results obtained by training the WAE on several datasets such as MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), CelebA (Liu et al., 2015), Ukiyo-E (Pinkney and Adler, 2020), and CIFAR-10 (Krizhevsky, 2009). The experiments demonstrate that the Fourier-series based discriminator leads to a faster and stabler convergence of the GAN component in WAE measured in terms of the Fréchet inception distance (FID) (Heusel et al., 2017) and reconstruction error, while also generating substantially sharper images on the datasets considered. The notable aspect is that these advantages accrue even without having to train a discriminator neural network. While our objective is not to avoid training a neural network for the discriminator, the proposed approach gives valuable insights into how closely coupled the discriminator and the generator optimization are, and gives us a deeper understanding of what exactly the neural-network based discriminator is trying to achieve.

The contributions of this paper may be summarized as follows. In the context of gradient-norm penalized WGAN, we show that the optimal discriminator, given the generator, solves a Poisson PDE. The solution relates to potential functions between the generator and data distributions in high dimensions. The solution is obtained using a truncated Fourier-series model, whose coefficients are obtained in closed form. This readily allows one to determine the optimal discriminator given the generator, while training only the generator network. We also show that the optimal value of the Lagrange multiplier can also be computed in closed form using a primal-dual approach. The advantage is that tracking the optimal Lagrange multiplier becomes a viable alternative for measuring training convergence in practical settings. We derive bounds on the errors introduced by the Fourier series truncation and sample estimation. Experimental validations on synthetic Gaussian and real-world datasets show that training a GAN with the proposed Fourier-series based discriminator outperforms baseline methods that consider a neural network for the discriminator. The applicability of the proposed framework is demonstrated for variants of divergence-minimizing GAN losses, with and without regularizers.

2 Mathematical Preliminaries

The cornerstone of our analysis is the Euler-Lagrange (EL) framework, which is at the heart of *Calculus of Variations* (Gelfand and Fomin, 1964; Mesterton-Gibbons, 2009). The EL conditions are of fundamental importance in solving several problems in physics (Goldstine, 1980; Ferguson, 2004).

Consider the functional optimization of a cost $\mathcal{L}$ defined as

$$\mathcal{L}\left(y(x), y'(x)\right) = \int_a^b \mathcal{F}\left(x, y(x), y'(x)\right)\ \mathrm{d}x, \tag{1}$$

with respect to $y(x)$, $x \in [a, b]$, which is assumed to be continuously differentiable or at least continuous with a piecewise-smooth derivative $y'(x)$, with finite Dirichlet boundary conditions. Let $y^*(x)$ denote the optimizer of $\mathcal{L}$. The first variation of $\mathcal{L}$ at the optimum y^*, is defined as the Gateaux derivative $\delta\mathcal{L}(y^*, \eta) = \left.\dfrac{\partial \mathcal{L}_\epsilon(y^*)}{\partial \epsilon}\right|_{\epsilon=0}$, where

$$\mathcal{L}_\epsilon(y^*) = \mathcal{L}\left(y^*(x) + \epsilon\,\eta(x), y^{*'}(x) + \epsilon\,\eta'(x)\right) = \int_a^b \mathcal{F}\left(x, y^*(x) + \epsilon\,\eta(x), y^{*'}(x) + \epsilon\,\eta'(x)\right)\ \mathrm{d}x,$$

where, in turn, $\eta(x)$ is a family of compactly supported, infinitely differentiable functions that are identically zero at the boundaries $x = a$ and $x = b$. Setting the first variation to zero and invoking the fundamental lemma of Calculus of Variations gives rise to the Euler-Lagrange condition. The *fundamental lemma of Calculus of Variations* states that if a function $f(x)$ satisfies the condition

$$\int_a^b f(x)\,\eta(x)\ \mathrm{d}x = 0$$

for all compactly supported, infinitely differentiable functions $\eta(x)$, then f must be identically zero almost everywhere in $[a, b]$.

The Euler-Lagrange condition that the optimizer $y^*(x)$ must satisfy is given as follows:

$$\frac{\partial \mathcal{F}}{\partial y} - \frac{\partial}{\partial x}\left(\frac{\partial \mathcal{F}}{\partial y'}\right)\Bigg|_{y=y^*(x)} = 0. \tag{2}$$

In the special case where the cost $\mathcal{L}$ does not involve the derivative of y, the EL condition reduces to the degenerate version:

$$\frac{\partial \mathcal{F}}{\partial y}\Bigg|_{y=y^*(x)} = 0,$$

which simply corresponds to a point-wise optimization of y over $x \in [a, b]$.

In the multivariate case, that is, $\boldsymbol{x} \in \mathbb{R}^n$, the cost is of the type

$$\mathcal{L}\left(y\left(\boldsymbol{x}\right), \{y_i'\}_{i=1}^n\right) = \int\limits_{\mathcal{X}\subseteq\mathbb{R}^n} \mathcal{F}\left(\boldsymbol{x}, y, \{y_i'\}_{i=1}^n\right)\, \mathrm{d}\boldsymbol{x},$$

where $\mathcal{X}$ is the domain of integration and y_i' denotes the partial derivative of $y(\boldsymbol{x})$ w.r.t. the i^{th} entry of $\boldsymbol{x}$, that is, x_i. The corresponding EL condition is

$$\frac{\partial \mathcal{F}}{\partial y} - \sum_{i=1}^{N}\left[\frac{\partial}{\partial x_i}\left(\frac{\partial \mathcal{F}}{\partial y_i'}\right)\right]\Bigg|_{y=y^*(\boldsymbol{x})} = 0. \tag{3}$$

The EL condition is a first-order condition and enforcing it yields the optimum. Whether the optimum corresponds to a minimizer or maximizer of the cost must be checked by invoking the second-order condition, more specifically the Legendre-Clebsch necessary condition for a minimizer. In the 1-D case, the condition is given by $\dfrac{\partial^2 \mathcal{F}}{\partial y'^2} \geq 0$. In the multivariate setting, this condition translates to the positive-semi-definiteness (p.s.d.) of the Hessian matrix $\mathbb{H}$ of the Hamiltonian $\mathcal{H}$, computed with respect to $\{y_i'(\boldsymbol{x})\}_{i=1}^n$ and evaluated at $y(\boldsymbol{x}) = y^*(\boldsymbol{x})$: $\mathbb{H}_{y,\mathcal{H}}\Big|_{y=y^*} \succ 0$, where $\succ$ denotes the p.s.d. property. The Hamiltonian is given by

$$\mathcal{H} = \sum_{i=1}^{n}\left(y_i'\frac{\partial \mathcal{F}}{\partial y_i'}\right) - \mathcal{F},$$

and the entries of the Hessian are given by

$$[\mathbb{H}_{y,\mathcal{H}}]_{i,j} = \frac{\partial^2 \mathcal{H}}{\partial y_i' \partial y_j'}.$$

We now apply the EL conditions to analyze Wasserstein GANs (WGANs) subject to the gradient-norm penalty in Section 3, and present similar analysis for divergence minimizing f-GAN variants in Appendix A and other gradient-regularized GAN losses in Appendix F.

3 Wasserstein GANs

The WGAN minimizes *earth mover's distance* (EMD) between the generator and the target data distributions, p_g and p_d, respectively. Earth mover's distance is a special case of the Wasserstein distance between two distributions. Through Kantorovich-Rubinstein duality, the WGAN optimization is specified via the min-max problem:

$$\min_{p_g} \left\{ \max_{D} \left\{ \mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] \right\} \right\},$$

which is equivalent to the sequential minimization:

$$D^*(\boldsymbol{x}, p_g) = \arg \min_{D:\ \|D\|_L \leq 1} \mathcal{L}_D^{\text{WGAN}}, \quad \text{where } \mathcal{L}_D^{\text{WGAN}} = -\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})], \text{ and}$$
$$p_g^*(\boldsymbol{x}) = \arg \min_{p_g} \mathcal{L}_G^{\text{WGAN}}, \quad \text{where } \mathcal{L}_G^{\text{WGAN}} = \mathbb{E}_{\boldsymbol{x}\sim p_d}[D^*(\boldsymbol{x}, p_g)] - \mathbb{E}_{\boldsymbol{x}\sim p_g}[D^*(\boldsymbol{x}, p_g)]$$

where in turn, $\|D(\boldsymbol{x})\|_L \leq 1$ denotes the Lipschitz constraint on the discriminator and $D^*(\boldsymbol{x}, p_g)$ is the optimal discriminator for a given generator distribution p_g. The optimal discriminator D^* is the one that penalizes regions of the input space where p_g differs from p_d, while satisfying the Lipschitz constraint. The constraint is typically imposed by clipping the weights of the discriminator network.

An alternative to weight-clipping is spectral normalization of the weights (Roth et al., 2019). Subsequent works (Gulrajani et al., 2017; Petzka et al., 2018; Terjék, 2020) replaced the Lipschitz constraint with a gradient penalty to avoid exploding gradients in a neural-network setting. For example, Gulrajani et al. (2017) replaced the Lipschitz-1 penalty with the gradient penalty (WGAN-GP): $(\|\nabla D(\boldsymbol{x})\|_2 - 1)^2 = 0$. It is well-known that a function whose gradient has a bounded norm satisfies the Lipschitz constraint (Adler and Lunz, 2018).

Table 1 lists a few important gradient-based regularizers proposed in the WGAN literature, which are considered in this paper. The original WGAN-GP empirically evaluated the discriminator gradient on samples drawn from the interpolated distribution $\alpha p_g + (1-\alpha)p_d$, $0 \leq \alpha \leq 1$, and penalizes values far away from 1 in the norm-squared sense. Petzka et al. (2018) incorporated a *one-sided hinge-like* penalty in the WGAN-LP formulation (LP stands for Lipschitz penalty). The gradient magnitude is upper-bounded by 1, by penalizing the discriminator only when the gradient magnitude exceeds 1. The gradients were evaluated empirically on an interpolated distribution as in the case of WGAN-GP. In the adversarial Lipschitz regularization proposed in WGAN-ALP (Terjék, 2020), for a sample drawn from either the data or generator distributions, the regularizer was evaluated along the *adversarial* penalty direction $\boldsymbol{r}_{adv}$ — the one along which the Lipschitz constraint is maximally violated. Mroueh et al. (2018) considered a gradient-norm penalty in the Sobolev GAN formulation, where they bounded the energy in the gradient of the discriminator, evaluated with respect to a base measure $\nu_p(\boldsymbol{x})$. From an implementation standpoint, they considered two base measures: (a) The midpoint distribution $\nu_p(\boldsymbol{x}) = \frac{p_d + p_g}{2}$, which is a special case of the WGAN-GP penalty (Gulrajani et al., 2017); and (b) A noise-convolved version of p_d, also considered in DRAGAN (Kodali et al., 2017). Mescheder et al. (2018) employed gradient penalties evaluated independently over real data (WGAN-R$_d$), over the generated data (WGAN-R$_g$), or a weighted combination of both (WGAN-R$_d$R$_g$) which can be seen as special cases of the Sobolev GAN penalty. Subsequent works extended the Wasserstein-1 distance

WGAN flavor	Discriminator loss
WGAN	$\mathcal{L}_D^{\text{WGAN}} = -\,\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})]$
WGAN-GP	$\mathcal{L}_D^{\text{WGAN}} + \lambda\, \mathbb{E}_{\boldsymbol{x}\sim \alpha p_g + (1-\alpha)p_d}\left[(\|\nabla D(\boldsymbol{x})\|_2 - 1)^2\right];\, 0 \leq \alpha \leq 1$
WGAN-R$_d$R$_g$	$\mathcal{L}_D^{\text{WGAN}} + \frac{\lambda_1}{2}\, \mathbb{E}_{\boldsymbol{x}\sim p_d}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right] + \frac{\lambda_2}{2}\, \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right]$
Sobolev GAN	$\mathcal{L}_D^{\text{WGAN}} + \lambda\, \mathbb{E}_{\boldsymbol{x}\sim \nu_p(\boldsymbol{x})}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right],$ where $\nu_p(\boldsymbol{x}) \geq 0;\ \int_{\mathcal{X}} \nu_p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$
WGAN-LP	$\mathcal{L}_D^{\text{WGAN}} + \lambda\, \mathbb{E}_{\boldsymbol{x}\sim \alpha p_g + (1-\alpha)p_d}\left[(\max(\|\nabla D(\boldsymbol{x})\|_2 - 1, 0))^2\right];\, 0 \leq \alpha \leq 1$
WGAN-ALP	$\mathcal{L}_D^{\text{WGAN}} + \lambda\, \mathbb{E}_{\boldsymbol{x}\sim p_d}\left[\left(\max\left(\frac{D(\boldsymbol{x}) - D(\boldsymbol{x}+\boldsymbol{r}_{adv})}{\|\boldsymbol{r}_{adv}\|_2} - 1, 0\right)\right)^2\right],$ where $\boldsymbol{r}_{adv} = \max_{\boldsymbol{r}:\|\boldsymbol{r}\|_2>0}\left\{\frac{D(\boldsymbol{x}) - D(\boldsymbol{x}+\boldsymbol{r})}{\|\boldsymbol{r}\|_2}\right\}$
WGAN-GNP (Proposed)	$\mathcal{L}_D^{\text{WGAN}} + \lambda_d \int\limits_{\boldsymbol{x}\in\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right)\, \mathrm{d}\boldsymbol{x}$

Table 1: Discriminator loss functions corresponding to various WGAN variants considered in the literature alongside the proposed WGAN with gradient-norm penalty (WGAN-GNP). The key difference lies in how the Lipschitz penalty is enforced on the discriminator. While the vanilla WGAN clips the discriminator network weights, the other WGAN flavors, including ours, consider gradient-based regularization.

based GAN to general L_p-norm spaces (Adler and Lunz, 2018) or propose solving the primal problem through differentiable Sinkhorn fixed-point iterations (Genevay et al., 2018).

3.1 WGAN with Gradient-norm Penalty

Let $\mathcal{X}$ denote the convex hull that contains the supports of p_d and p_g. In this work, we consider the following gradient-norm penalty (GNP) for the WGAN:

$$\Omega_D : \int_{\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right) \mathrm{d}\boldsymbol{x}. \tag{4}$$

In WGAN-GP, the gradients are evaluated over an interpolated distribution. As in the case of WGAN-R$_d$ or WGAN-R$_g$, the proposed penalty can be viewed as a particular case of the penalty considered in the Sobolev GAN formulation. While WGAN-R$_d$ and WGAN-R$_g$ enforce the penalty on the supports of p_d and p_g, respectively, the proposed WGAN-GNP considers a uniform distribution on $\mathcal{X}$, resulting in a closed-form solution to the discriminator, given the generator. In the WGAN-GNP setting, we constrain the generator and data distributions as follows:

Assumption 1 ($\mathcal{C}^1(\mathcal{X})$ **distributions**). *The generator and data distributions are compactly supported and continuously differentiable functions.*

Incorporating Ω_D into the WGAN discriminator cost results in the following optimization problem for the discriminator:

$$D^*(\boldsymbol{x}, p_d, p_g) = \arg\min_D \underbrace{\left\{ -\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] + \lambda_d \int_{\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right) \mathrm{d}\boldsymbol{x} \right\}}_{\mathcal{L}_D}. \tag{5}$$

The generator optimization is then given by

$$\begin{aligned} p_g^*(\boldsymbol{x}) &= \arg\min_{p_g} \{\mathcal{L}_G\}, \text{ where} \\ \mathcal{L}_G &= \mathbb{E}_{\boldsymbol{x}\sim p_d}[D^*(\boldsymbol{x}, p_d, p_g)] - \mathbb{E}_{\boldsymbol{x}\sim p_g}[D^*(\boldsymbol{x}, p_d, p_g)] + \lambda_p \left(\int_{\mathcal{X}} p_g(\boldsymbol{x}) \ \mathrm{d}\boldsymbol{x} - 1 \right) \\ &\quad + \int_{\mathcal{X}} \mu_p(\boldsymbol{x}) p_g(\boldsymbol{x}) \ \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{6}$$

where λ_p and $\mu_p(\boldsymbol{x})$ are the Karush-Kunh-Tucker (KKT) multipliers for the integral constraint $\Omega_{p_g} : \int_{\mathcal{X}} p_g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$, and the non-negativity constraint $\Phi_{p_g} : p_g(\boldsymbol{x}) \geq 0$, respectively. These constraints explicitly enforce p_g to be a valid p.d.f.

We analyze WGAN-GNP in the one-dimensional setting first and subsequently extend the analysis to higher dimensions.

3.2 WGAN-GNP and Euler-Lagrange Conditions in 1-D

In the 1-D setting, the gradient norm penalty in Equation (4) takes the following form:

$$\Omega_D : \ \int_{\mathcal{X}} \left(|D'(x)|^2 - 1\right) \ \mathrm{d}x,$$

where D' denotes the first derivative of D. The WGAN-GNP discriminator loss is given by

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_d}[D(x)] + \mathbb{E}_{x\sim p_g}[D(x)] + \lambda_d \int_{\mathcal{X}} \left(|D'(x)|^2 - 1\right) \ \mathrm{d}x. \tag{7}$$

The optimal discriminator in 1-D is given in the following result.

Theorem 1. ***Optimal WGAN-GNP discriminator (1-D)****: Consider the optimization of the one-dimensional WGAN-GNP discriminator loss given in Equation* (7)*. The optimal discriminator $D^*(x)$, given the generator p_g, is a solution to one-dimensional Poisson's second-order differential equation:*

$$D''(x) = \frac{p_g(x) - p_d(x)}{2\lambda_d}, \ \forall \ x \in \mathcal{X}, \tag{8}$$

and is given by the closed-form solution involving the twice-iterated antiderivatives:

$$D^*(x) = \frac{1}{2\lambda_d} \int \left(\int \left(p_g(x) - p_d(x)\right) \ \mathrm{d}x \right) \mathrm{d}x \pm x, \ \forall \ x \in \mathcal{X}, \tag{9}$$

where λ_d is the Lagrange multiplier corresponding to the gradient penalty, and $\int$ denotes the antiderivative.

Proof. The integrand in $\mathcal{L}_D$ in Equation (7) is given by

$$\mathcal{F}\left(x, D, D'\right) = D(x)(p_g(x) - p_d(x)) + \lambda_d\big(|D'(x)|^2 - 1\big).$$

The Euler-Lagrange condition in Equation (2), when applied to $\mathcal{F}$ results in the ordinary differential equation (ODE) given in Equation (8). The homogeneous solution to the 1-D Laplace equation $D'' = 0$ takes the form $D_h^*(x) = a_1 x + a_0$, while the particular solution $D_p^*(x, p_g, p_d)$ is twice-iterated antiderivative of the right-hand side of Equation (8). The optimal discriminator $D^*(x, p_g, p_d)$ is the sum of the homogeneous and particular solutions:

$$\begin{aligned} D^*(x, p_g, p_d) &= D_p^*(x, p_g, p_d) + D_h^*(x) \\ &= \frac{1}{2\lambda_d} \int \left(\int (p_g(x) - p_d(x)) \ \mathrm{d}x \right) \mathrm{d}x + a_1 x + a_0. \end{aligned}$$

The constants a_0 and a_1 must be estimated based on the boundary conditions. Upon convergence of the GAN, the optimal generator distribution must match the data distribution, that is, $p_g^* = p_d$. In this scenario, the optimal discriminator $D_{\text{opt}}^*(x) = D^*\left(x, p_g^*, p_d\right)$, which implies that the particular solution $D_p^* = 0$ and $D_{\text{opt}}^*(x)$ is the solution to Laplace's equation, that is, $D_{\text{opt}}^*(x) = D_h^*(x)$. Enforcing the gradient-norm penalty: $|(D_{\text{opt}}^*)'|^2 = 1$ yields $a_1 = \pm 1$. The choice of a_0 remains free, as the gradient-norm penalty is satisfied for all $a_0 \in \mathbb{R}$, which merely offsets D^* by a constant. Without loss of generality, we set $a_0 = 0$. Thus, we obtain the optimal closed-form WGAN-GNP discriminator for a given generator, given in Equation (9). □

While the homogeneous component can be estimated based on the form of Ω_D, we will show (in Theorem 2) that the generator optimization is independent of D_h. Therefore, the choice of a_1 and a_0 are inconsequential to the GAN training. While $D_h^*(x)$ is independent of the distributions, we note that the particular solution $D_p^*\left(x, p_g, p_d\right)$ and thereby, the sum $D^*(x, p_g, p_d)$ depend on the data being modelled, and the generator distribution from which the samples are provided. Henceforth, we use the notation $D_p^*(x)$ and $D^*(x)$ for brevity, while bearing in mind that the optimal discriminator is always determined for a given pair of generator and data distributions.

We now present an alternative, but equivalent, formulation of the optimal discriminator involving the fundamental solution of $D''(x) = \delta(x)$, where $\delta(x)$ denotes the Dirac-delta distribution. The reason for providing the alternative solution is that it readily generalizes to higher dimensions.

Lemma 1. ***Fundamental WGAN-GNP discriminator****: Given the 1-D Laplace equation in the distributional setting $D''(x) = \delta(x)$ together with its fundamental solution $\phi(x) = r(x) + b_1 x + b_0$, where $r(x)$ is the one-sided ramp function $r(x) = x$ for $x \geq 0$, and 0 elsewhere. The solution to the second-order ODE in Equation* (8) *is given by the convolution integral:*

$$D^*(x) = \frac{1}{2\lambda_d} \left(\phi * (p_g - p_d)\right)(x) + a_1 x + a_0. \ \forall \ x \in \mathcal{X}. \tag{10}$$

Proof. The fundamental solution is obtained by taking the second-order antiderivative of $\delta(x)$ by interpreting the derivative in the distributional sense. The antiderivative of $\delta(x)$ is

the Heaviside unit-step function, and in turn, the antiderivative of the unit-step function is the one-sided ramp function $r(x)$. This yields $\phi(x) = r(x) + b_1 x + b_0$. The symmetric fundamental solution $\phi(x) = \frac{1}{2}|x|$ can be obtained as a special case by setting $b_1 = 0.5$. The solution to Equation (8) can be obtained by convolving the right-hand side of (8) with the fundamental solution $\phi(x)$ (Evans, 2010). Including the homogeneous solution $a_1 x + a_0$ results in Equation (10). □

We would like to remark that the convolutional form of $D^*(x)$, which solves Poisson's PDE, is a special case of the Riemann-Liouville integral (Stein, 1970) given by

$$\mathrm{I}_1^\alpha[f](x) = c_1^\alpha \int_0^x (x-y)^{\alpha-1} f(y)\ \mathrm{d}y,$$

where α is the order of the derivative of f, $\alpha \in [0,1]$, and $D_p^*(x) = \mathrm{I}_1^2[p_g - p_d](x)$.

Lemma 2. ***Optimal Lagrange multiplier λ_d^* (1-D)****: The optimal Lagrange multiplier for the one-dimensional discriminator function given in Equation* (10) *is*

$$\lambda_d^* = \frac{1}{4}\sqrt{\frac{1}{|\mathcal{X}|}\int_{\mathcal{X}} \left(\left(\mathrm{sgn} * (p_g - p_d)\right)(x) + a_1\right)^2\ \mathrm{d}x}\,, \tag{11}$$

where $\mathrm{sgn}(x) = \frac{x}{|x|}$ *denotes the signum function,* $|\mathcal{X}|$ *is the length of the interval* $\mathcal{X}$ *in 1-D, and the positive root results in a* $D^*(x)$ *that minimizes* $\mathcal{L}_D$*.*

Proof. Enforcing the gradient-norm penalty on $D^*(x)$ yields the optimal Lagrange multiplier λ_d^*. The positive root is chosen based on the Legendre-Clebsch second-order condition (cf. Section 2), which yields $2\lambda_d^* > 0$ for D^* to be a minimizer of $\mathcal{L}_D$. The proof is provided in Appendix B.1. □

Given the optimal discriminator D^*, recall the Lagrangian of the WGAN generator cost given in Equation (6):

$$\mathcal{L}_G = \mathbb{E}_{\boldsymbol{x}\sim p_d}[D^*(x)] - \mathbb{E}_{\boldsymbol{x}\sim p_g}[D^*(x)] + \int_{\mathcal{X}} \left(\lambda_p + \mu_p(x)\right) p_g(x)\ \mathrm{d}x - \lambda_p,$$

where λ_p and $\mu_p(x)$ are the KKT multipliers associated with the integral and non-negativity constraints, respectively. The following result presents the optimal WGAN-GNP generator given the optimal discriminator $D^*(x)$.

Theorem 2. ***Optimal WGAN-GNP generator (1-D)****: Consider the optimization of the integral cost* $\mathcal{L}_G$ *given by*

$$\mathcal{L}_G = \int_{\mathcal{X}} \left(D^*(x)(p_d(x) - p_g(x)) + (\lambda_p + \mu_p(x))p_g(x)\right)\ \mathrm{d}x - \lambda_p,$$

where $D^*(x)$ *is given in Equation* (10)*, and* λ_p *and* $\mu_p(x)$ *are the KKT multipliers satisfying* $-\infty < \mu_p(x) \le 0$*,* $\mu_p(x)p_g(x) = 0,\ \forall\ x \in \mathcal{X}$*, and* λ_p *is a finite real value. Minimization of* $\mathcal{L}_G$ *yields*

$$p_g^*(x) = p_d(x), \quad \textit{and} \quad \mu_p^*(x) = 0,\ \forall\ x \in \mathcal{X},$$

and the solution is optimal for all finite real values of λ_p*.*

Proof. $\mathcal{L}_G$ involves a convolution term inside the integrand. Hence, it is not in a form where the Euler-Lagrange conditions can be readily applied. Early manifestations of such cost functions was in the context of elastodynamics (Gurtin, 1964a) and initial-value problems (Gurtin, 1964b). In order to solve such problems, Gurtin introduced variational analysis of convolutional costs starting from first principles and developed the corresponding counterpart of the fundamental lemma of calculus of variations. We adopt a similar approach by evaluating the first variation of $\mathcal{L}_G$ and then applying the fundamental lemma of calculus of variations (cf. Section 2). The result is the following equation:

$$(\phi * (p_d - p_g^*))(x) = \frac{\mu_p(x) + \lambda_p + a_1\, x}{2},$$

where ϕ is the fundamental solution to the Laplace equation. Considering the Laplacian on both sides gives $p_g^*(x) = p_d(x) - c\mu_p''(x)$. Enforcing the integral and non-negativity constraints on $p_g^*(x)$ results in the desired optimum $p_g^*(x) = p_d(x)$. The optimal solution p_g^* is independent of the homogeneous component of the discriminator $D_h(x)$. The detailed derivation is provided in Appendix B.2. □

3.3 Constraint Space of the Discriminator (1-D)

The appropriate class of functions that the discriminator must be drawn from depends on the choice of the regularizer. For gradient-based regularizers, Sobolev spaces are most appropriate (Mroueh and Sercu, 2017; Mroueh et al., 2018). The first-order L_2-normed Sobolev Space $W^{1,2}(\mathcal{X}, \nu)$ consists of finite-energy functions whose first-order derivative is also of finite energy with respect to measure ν defined over $\mathcal{X}$. Consider the discriminator $D \in W^{1,2}$ with the Sobolev norm

$$\|D\|_{W^{1,2}} = \sqrt{\|D\|_{2,\nu}^2 + \|D'\|_{2,\nu}^2} = \sqrt{\int_{\mathcal{X}} |D(x)|^2 \, \mathrm{d}\nu + \int_{\mathcal{X}} |D'(x)|^2 \, \mathrm{d}\nu} < \infty.$$

In Sobolev GAN, Mroueh et al. (2018) consider the space $W_0^{1,2}$, which consists of functions from $W^{1,2}$ that disappear on the boundary of a compact domain $\mathcal{X}$. Consequently, invoking the Poincaré inequality, it suffices to consider functions with finite energy in the gradient (Sobolev, 1963) and the semi-norm $\|D\|_{W_0^{1,2}} \leq \mathfrak{r}\|\nabla D\|_{2,\nu}$, for some positive constant $\mathfrak{r}$. The WGAN-GNP discriminator can therefore be interpreted as belonging to $W_0^{1,2}(\mathcal{X}, \mathcal{U}_{\mathcal{X}})$, where $\mathcal{U}_{\mathcal{X}}$ denotes the uniform measure over $\mathcal{X}$.

3.4 Fourier-series Approximation

The closed-form optimal discriminator given by Equations (9) or (10) involves evaluating an integral. Further, in a practical GAN setting, we do not have access to p_d and p_g in closed form. Hence, in practice, one must resort to alternative approaches to solve the ODE in Equation (8). Typical alternatives include basis function expansions or discretization of the differential operators, which gives rise to a finite-difference equation.

The existing GAN approaches, in particular, the WGAN approaches, employ a neural network for the discriminator and optimize it, agnostic to the underlying differential equation formulation of the discriminator. On the other hand, we prefer to solve the differential

equation, even if approximately, instead of employing a neural network. We use a Fourier bases expansion to solve the differential equation to serve as a proof of concept, although other basis expansions could be considered as well. Our approach is motivated by that of Fourier himself who introduced the series expansion to solve the heat equation in a metal (Fourier, 1807). The choice of Fourier bases is motivated by the fact that they are eigenfunctions of the Laplace operator, and the multidimensional Fourier bases can be expressed as a tensor product of the univariate counterparts, which simplifies computation of the Fourier coefficients. We refer to WGAN-GNP with the Fourier solver as WGAN-FS. From Assumption 1 and Section 3.3, the data and generator distributions, and the discriminator functions admit valid Fourier series expansions.

Consider the Fourier-series expansions of the data density, generator density, and the discriminator, respectively

$$p_d(x) = \sum_{m\in\mathbb{Z}} \alpha_m e^{\mathrm{j}\omega_o m x}, \quad p_g(x) = \sum_{m\in\mathbb{Z}} \beta_m e^{\mathrm{j}\omega_o m x}, \quad \text{and} \quad D(x) = \frac{1}{\lambda_d}\sum_{m\in\mathbb{Z}} \gamma_m e^{\mathrm{j}\omega_o m x},$$

respectively. The Fourier-series model assumes periodicity and the fundamental frequency ω_o has to be specified. Strictly speaking, although p_d, p_g, and $D(x)$ are not periodic, we are concerned with these expansions only over a certain domain of interest. This is also considered as the fundamental period, which could be determined based on prior knowledge of the data being modelled. Substituting the Fourier expansions in Equation (8) gives

$$\begin{aligned} -\frac{(\omega_o m)^2}{\lambda_d}\gamma_m \exp(\mathrm{j}\omega_o m x) &= \left(\frac{\beta_m - \alpha_m}{2\lambda_d}\right)\exp(\mathrm{j}\omega_o m x),\ \forall\ m \neq 0, \\ \Rightarrow \gamma_m &= \frac{\alpha_m - \beta_m}{2\omega_o^2 m^2},\ \forall\ m \neq 0, \end{aligned} \tag{12}$$

which are the Fourier coefficients corresponding to the discriminator, except γ_0. While the value of γ_0 can be determined based on the boundary conditions on $D(x)$, we observed experimentally that it merely introduces a constant offset in the optimal discriminator, and can therefore be ignored when training the generator. The Fourier-series approximation specifies the discriminator by relating its Fourier coefficients to those of p_d and p_g. The advantage of the Fourier expansion of a function is that the derivatives also admit an expansion in the same bases, so long as they have finite L_2 norm. Clubbing the homogeneous solution $a_1\, x + a_0$ with the particular solution obtained above gives the general solution. Without loss of generality, we set $a_0 = 0$ and $a_1 = 1$. The optimal WGAN-FS discriminator, given the generator (equivalently, its Fourier coefficients), takes the form:

$$D^*_{FS}(x) = x + \frac{1}{\lambda^*_{FS}} \sum_{m\in\mathbb{Z}-\{0\}} \gamma_m e^{\mathrm{j}\omega_o m x},$$

where λ^*_{FS} is the optimal Lagrange multiplier corresponding to the Fourier-series discriminator. The value of λ^*_{FS} must be determined by enforcing additional conditions such as the gradient-norm penalty. This aspect will be discussed in Section 3.6.

3.5 Fourier Series of a Probability Density Function

Consider the Fourier-series expansion of a compactly supported density $p(x)$ over $[0, T]$:

$$p(x) = \sum_{m \in \mathbb{Z}} a_m e^{\mathrm{j}\omega_o m x},\ x \in [0, T].$$

The Fourier coefficient a_m is given by

$$a_m = \frac{1}{T}\int_0^T p(x) e^{-\mathrm{j}\omega_o m x}\,\mathrm{d}x \ = \ \frac{1}{T}\underset{x\sim p}{\mathbb{E}}\left[e^{-\mathrm{j}\omega_o m x}\right] \ = \ \frac{1}{T}\varphi_p^*(\omega_o m),$$

where $\varphi_p^*(\omega)$ denotes the complex conjugate of the characteristic function $\varphi_p(\omega) = \mathbb{E}[e^{\mathrm{j}\omega x}]$. Effectively, the coefficient a_m is determined by uniformly sampling the characteristic function at $\omega = \omega_o m, m \in \mathbb{Z}$. Similarly, $\alpha_m = \frac{1}{T}\varphi_{p_d}^*(\omega_o m)$ and $\beta_m = \frac{1}{T}\varphi_{p_g}^*(\omega_o m)$, where φ_{p_d} and φ_{p_g} are the characteristic functions of p_d and p_g, respectively. For an infinitely supported density, such as the Gaussian, one could consider a sufficiently large interval ($T > 10\sigma$) about the mean for truncation. Then, the above expansion will hold only approximately, and in the limiting sense ($T \to \infty$), the series approximation approaches the Fourier transform.

Thus, the discriminator can be computed in closed-form given the generator and data distributions. The differential equation formalism and the Fourier-series approximation obviate the need for training a neural-network-based discriminator.

3.6 Practical Considerations in 1-D

For distributions whose characteristic function can be evaluated in closed-form, the computation of the Fourier coefficients and thereby the discriminator is straightforward. In practice, the infinite-order summations are truncated. Further, the characteristic function is not often available in closed form and instead, only data is given. In this scenario, we replace the Fourier coefficients $a_m = \frac{1}{T}\varphi_p^*(\omega_o m)$ with their sample estimates as follows:

$$\bar{a}_m = \frac{1}{NT}\sum_{x_k \in \mathcal{D}} e^{-\mathrm{j}\omega_o n x_k},$$

$\mathcal{D} = \{x_k \mid k = 1, 2, \ldots, N\}$ being the given data that follows the distribution $p(x)$.
The convergence properties of empirical characteristic functions have been extensively explored in the literature (Giardina and Chirlian, 1972; Feuerverger and Mureika, 1977). A 1-D function $p(x) \in \mathcal{C}^1(\mathcal{X})$ supported over $\mathcal{X}$ is of bounded variation $\mathrm{V}_{\mathcal{X}}[p(x)] = \int_{\mathcal{X}} |p'(x)|\,\mathrm{d}x \leq B$. The mean-squared error ϵ_p^2 in truncating the series to M terms is bounded above by $\frac{B^2}{\pi\omega_o M}$ (Giardina and Chirlian, 1972). For $D_{FS}^*(x)$, the mean-squared error is given by $\epsilon_D^2 \leq \frac{\mathfrak{c}}{\omega_o^3 M^3}$, for some positive constant $\mathfrak{c}$. The proof is included in Appendix D.1. We observe that the error in approximating D decays as $\frac{1}{M^3}$. Therefore, we expect that the truncated series will yield accurate approximations even for moderate M. Experimental results in Appendix E.1 support this claim.

Further, from an implementation perspective, as in the case of TensorFlow (Abadi et al., 2016), one could avoid complex arithmetic by using a trigonometric Fourier-series expansion

of the type

$$\tilde{p}(x) = \frac{\bar{a}_0}{2} + \sum_{m=1}^{M} \bar{a}_m^r \cos(\omega_o m x) + \bar{a}_m^i \sin(\omega_o m x),$$

where $\bar{a}_m^r = \text{Real}\{\bar{a}_m\}$ and $\bar{a}_m^i = \text{Imag}\{\bar{a}_m\}$. Considering truncated trigonometric Fourier-series expansions of p_g and p_d, and accounting for the homogeneous solution, we obtain

$$\tilde{D}_{FS}^*(x) = x + \frac{1}{\lambda_{FS}^*} \sum_{m=1}^{M} \gamma_m^r \cos(\omega_o m x) + \gamma_m^i \sin(\omega_o m x).$$

It has been reported in the literature that employing the ideal discriminator for a given generator prevents stable training of the generator due to vanishing gradients (Liu et al., 2017), while a very coarse approximation might not capture the modes present in the data distribution (Daskalakis et al., 2018). Hence, a smooth approximation of the ideal discriminator is considered a good compromise — this is precisely what the truncated Fourier-series approximation implicitly also achieves.

Enforcing the gradient-norm penalty on $\tilde{D}_{FS}^*(x)$ enables one to determine the optimal λ_{FS}^* associated with the Fourier-series discriminator. In terms of the data $\mathcal{D}$, λ_{FS}^* can be approximated as follows:

$$\lambda_{FS}^* \approx \sqrt{(2M+1)\left(\sum_{m=1}^{M} (\tau_m^i + \tau_m^r) + \frac{1}{N} \sum_{x_k \in \mathcal{D}} \sum_{m=1}^{M} (\tau_m^i - \tau_m^r) \cos(2\omega_o m x_k)\right)},$$

where $\tau_m^r = \frac{1}{2}(\gamma_m^r \omega_o m)^2$, and $\tau_m^i = \frac{1}{2}(\gamma_m^i \omega_o m)^2$. The derivation is provided in Appendix B.3.

Similar to the Fourier coefficients of $\tilde{D}_{FS}^*(x)$, we observe that τ_m^r and τ_m^i are functions of γ_m^r and γ_m^i, respectively, which in turn depend on the Fourier coefficients of p_d and p_g. As the generator distribution converges toward the optimal solution $p_g^* = p_d$, λ_d^* in Equation (11) as well as λ_{FS}^* converge to zero. Therefore, monitoring λ_{FS}^* in a WGAN-FS training scenario serves as a practical alternative to computing the divergence between p_g and p_d. As shown in Section 6.2 and Appendix E.1, this alternative is particularly useful when dealing with real-world images with no closed-form representation for the underlying distribution.

3.7 Illustration Using Synthetic 1-D Data

As a preliminary validation, we compare the performance of the proposed WGAN-FS on the 1-D Gaussian learning task.

Baselines: We compare WGAN-FS with the following two categories of baselines: (i) WGAN and its variants with different penalties, such as the gradient penalty (WGAN-GP), Lipschitz penalty (WGAN-LP), Sobolev GAN and stable alternatives to GP, such as WGAN-R$_d$ and WGAN-R$_g$; and (ii) base WGAN with variations of the proposed gradient-norm penalty (GNP), evaluated empirically on sample points drawn from the two datasets. WGAN-GNP implements the WGAN-GP algorithm with the GNP cost. While we compute the optimal Lagrange multiplier λ_d in closed-form in WGAN-FS, in Sobolev GANs, λ_d is

optimized to maximize the discriminator loss through stochastic gradient-descent (Mroueh et al., 2018). Recently, multi-layer networks with periodic sinusoidal activations (SIREN) have been shown to achieve state-of-the-art performance in learning image, sound and wavefield representations (Sitzmann et al., 2020). We therefore adopt two variants of SIREN for the discriminator: (a) A three-layer fully connected network with sin activation, called WGAN-GNP (3S); and (b) A single-layer fully connected network with sin activation and the same number of nodes as terms in the Fourier-series expansion, called WGAN-GNP (1S). Training WGAN-GNP (1S) is equivalent to learning the Fourier coefficients in the WGAN-FS formulation.

Experimental setup: The generator in all GAN variants is considered to be a linear transformation of the input: $y = w\,z + b$. Gaussian training data is drawn from $\mathcal{N}(10, 1)$, while noise z that is input to the generator is sampled from the standard Gaussian $\mathcal{N}(0, 1)$. While WGAN-FS uses a closed-form Fourier-series discriminator, the baselines use a three-layer fully connected discriminator network with leaky ReLU activation. The batch size is 500. For the baseline techniques, each training step involves 5 iterations of the discriminator network optimization followed by one iteration of the generator. WGAN-FS, on the other hand, uses a single-shot discriminator during each training step. Based on additional experiments conducted in Appendix E.1, we set the period $T = \frac{2\pi}{\omega_o}$ to 15 and the truncation frequency M to 10. The Adam optimizer (Kingma and Ba, 2015) is used with a learning rate $\eta = 0.05$, and the exponential decay parameters for the first and second moments are $\beta_1 = 0.5$ and $\beta_2 = 0.999$, respectively. The implementation was carried out using TensorFlow 2.0 (Abadi et al., 2016).

Results: Figures 1(a) and (b) show the discriminators learnt by the various GANs under consideration. The optimal classifier between the two Gaussians is also plotted for the sake of reference. All classifier outputs are rescaled to $[-0.5, 0.5]$ to facilitate comparison. While the discriminator peak values depend on the network architecture for the baselines, it is a function of the period T and the constant a_0 in WGAN-FS. We observe that the discriminator in WGAN-FS is closer to the optimal classifier than the ones learnt by the baseline WGAN variants. In the case of GNP variants, WGAN-GNP is comparable to the best case baseline WGAN-LP. The WGAN-GNP (3S) model is able to localize the target distribution, but the mismatch between the periodicity of the data and the activation function results in undesirable harmonics in $D(x)$. On the other hand, WGAN-GNP (1S), due to its Fourier-series structure, comes closest to the WGAN-FS discriminator. This experiment shows that solving the differential equation single-shot, even if approximately, is a more accurate alternative to training a network for the discriminator.

Figures 1(c) and (d) compare the Wasserstein-2 distance ($\mathcal{W}^{2,2}$) between the target and generator Gaussians. These plots show that WGAN-FS converges much faster than the baseline techniques. This is a direct consequence of solving the differential equation within the training process. The poor performance of WGAN-GNP (3S) and WGAN-GNP (1S) is attributed to the mismatch between the fundamental frequency of the sinusoid and the assumed periodicity. Figures 1(d) and (e) compare the Kullback-Leibler (KL) divergence between the generator and true data distributions. We observe similar performance improvements in WGAN-FS compared with the baselines, as in the $\mathcal{W}^{2,2}$ case.

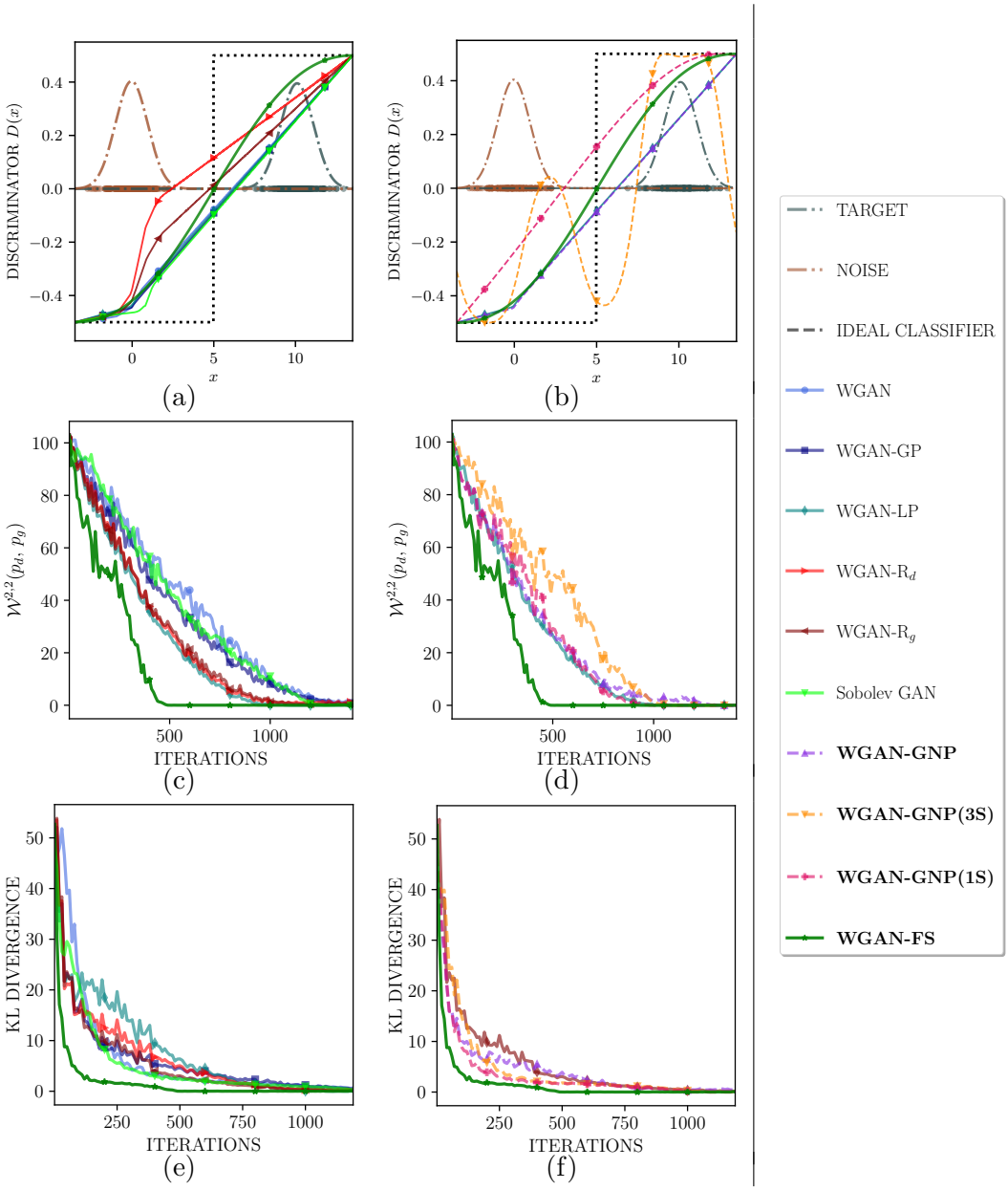


Figure 1: (Color online) Experiments on 1-D Gaussian data: (a) & (b) Discriminator learnt by WGAN-FS in comparison with those learnt by the baselines and WGAN-GNP variants. Baseline implementation with empirical gradient estimates learn piecewise linear discriminators, while WGAN-FS and variants of WGAN-GNP with sinusoidal activations learn a smoother function. The WGAN-FS discriminator is closest to the ideal classifier, while being smooth. Wasserstein-2 distance between p_d and p_g ($\mathcal{W}^{2,2}(p_d, p_g)$) versus iterations for (c) WGAN-FS and the baseline WGANs and (d) WGAN-FS and WGAN-GNP variants. Kullback-Leibler divergence versus iterations for (e) WGAN-FS and the baseline WGANs and (f) WGAN-FS and WGAN-GNP variants. WGAN-FS attained the lowest (best) values in terms of both metrics substantially faster than the baselines.

4 Multivariate WGAN-GNP

In the n-dimensional WGAN-GNP scenario, the discriminator loss takes the form:

$$\begin{aligned}\mathcal{L}_D &= -\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] + \lambda_d \int_{\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right)\, \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{X}} \left(D(\boldsymbol{x})\left(p_g(\boldsymbol{x}) - p_d(\boldsymbol{x})\right) + \lambda_d \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right)\right)\, \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{13}$$

We now determine the optimal discriminator corresponding to the loss given in Equation (13).

Theorem 3. ***Optimal WGAN-GNP discriminator in*** n***-D****: Consider the n-D WGAN discriminator loss subject to the gradient-norm penalty as given by Equation* (13)*. The optimizer of $\mathcal{L}_D$ solves Poisson's partial differential equation given by*

$$-\Delta D\left(\boldsymbol{x}\right) = \frac{p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})}{2\lambda_d}, \tag{14}$$

where $\Delta = \nabla.\nabla = \left(\partial_{x_1}^2 + \partial_{x_2}^2 + \ldots + \partial_{x_n}^2\right)$ denotes the Laplacian operator, with x_i being the i^{th} entry of $\boldsymbol{x}$, and $\partial_{x_i}^2 = \frac{\partial^2}{\partial x_i^2}$. The closed-form particular solution is given by the multidimensional convolution integral

$$D_p^*(\boldsymbol{x}) = \frac{1}{2\lambda_d} \int_{\mathcal{X}} \phi(\boldsymbol{x} - \boldsymbol{y})\left(p_d(\boldsymbol{y}) - p_g(\boldsymbol{y})\right)\, \mathrm{d}\boldsymbol{y}, \tag{15}$$

where $\phi(\boldsymbol{x})$ denotes the fundamental solution to the Laplace equation: $-\Delta D(x) = \delta(\boldsymbol{x})$. The fundamental solution is given by

$$\phi(\boldsymbol{x}) = \begin{cases} -\dfrac{1}{2\pi}\ln(\|\boldsymbol{x}\|), & \text{for } n = 2, \text{ and} \\ \dfrac{1}{n(n-2)\mathfrak{v}(n)}\dfrac{1}{\|\boldsymbol{x}\|^{n-2}}, & \text{for } n \geq 3, \end{cases} \tag{16}$$

where $\|\boldsymbol{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$ and $\mathfrak{v}(n)$ is the volume of the unit sphere in $\mathbb{R}^n$ given by $\mathfrak{v}(n) = \pi^{\frac{n}{2}}\left(\Gamma\left(\frac{n}{2}+1\right)\right)^{-1}$, with $\Gamma(n)$ denoting the gamma function.

Proof. Consider the integrand in Equation (13): $D(\boldsymbol{x})(p_g(\boldsymbol{x}) - p_d(\boldsymbol{x})) + \lambda_d\|\nabla D(\boldsymbol{x})\|_2^2$. Applying the Euler-Lagrange condition from Equation (3) for obtaining the optimum results in Poisson's partial differential equation (PDE) given in Equation (14).

A closed-form solution to Poisson's equation is obtained similar to the 1-D case. Solving the n-D inhomogeneous differential equation $-\Delta D(\boldsymbol{x}) = \delta(\boldsymbol{x})$ in polar coordinates yields the fundamental solution $\phi(\boldsymbol{x})$ given in Equation (16) (Evans, 2010). The solution to Poisson's equation $-\Delta D(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})}{2\lambda_d}$ is the convolution between $\phi(\boldsymbol{x})$ and $\frac{p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})}{2\lambda_d}$, which results in Equation (15). □

For the specific case $n = 2$, we obtain

$$D_p^*(\boldsymbol{x}) = \frac{-1}{4\pi\lambda_d} \int_{\mathcal{X}} \ln(\|\boldsymbol{x} - \boldsymbol{y}\|)\left(p_d(\boldsymbol{y}) - p_g(\boldsymbol{y})\right)\, \mathrm{d}\boldsymbol{y}. \tag{17}$$

For $n \geq 3$, we obtain

$$D_p^*(\boldsymbol{x}) = \frac{1}{2\lambda_d n(n-2)\mathfrak{v}(n)} \int_{\mathcal{X}} \frac{1}{\|\boldsymbol{x}-\boldsymbol{y}\|^{n-2}} \left(p_d(\boldsymbol{y}) - p_g(\boldsymbol{y})\right) \, \mathrm{d}\boldsymbol{y}. \tag{18}$$

Including the family of homogeneous solutions $D_h^*(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle + \text{constant}$, the general solution becomes

$$D^*(\boldsymbol{x}) = D_p^*(\boldsymbol{x}) + \langle \boldsymbol{a}, \boldsymbol{x} \rangle + \text{constant}. \tag{19}$$

As in the 1-D case, upon convergence of the GAN, we expect $p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x})$. The optimal discriminator in this scenario is given by $D_{opt}^*(\boldsymbol{x}) = D_h^*(\boldsymbol{x})$. Enforcing the gradient-norm penalty, we obtain the condition $\|\boldsymbol{a}\| = 1$, with the constant term merely resulting in an offset of $D^*(\boldsymbol{x})$.

Equations (17) and (18) are specific instances of the Calderon-Zygmund singular integral (Stein, 1970), with kernels $K_{D,2}(\boldsymbol{x}, \boldsymbol{y}) = \ln(\|\boldsymbol{x}-\boldsymbol{y}\|)$ and $K_{D,n}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|^{2-n}$, respectively, with singularities along $\boldsymbol{x} = \boldsymbol{y}$. The integrals are evaluated in the Cauchy principal-value sense. $D_p^*(\boldsymbol{x})$ for $n \geq 3$ is also a specific instance of the Riesz potential (Riesz, 1949; Landkof, 1972) given by

$$\mathrm{I}_n^\alpha[f](\boldsymbol{x}) = c_n^\alpha \int\limits_{\mathcal{X} \subseteq \mathbb{R}^n} \|\boldsymbol{x}-\boldsymbol{y}\|^{\alpha-n} f(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}, \qquad 0 < \alpha < n,$$

which, in turn, is an n-dimensional generalization of the Riemann-Liouville integral (Stein, 1970). In the present context, $D_p^*(\boldsymbol{x}) = \mathrm{I}_n^2[p_g - p_d](\boldsymbol{x})$.

In the language of electrostatics, $D_p^*(\boldsymbol{x})$ could be interpreted as the difference between the Newtonian potentials of the functions p_g and p_d. The charge-free space scenario corresponds to $p_g^* = p_d$, which results in $D_p^*(\boldsymbol{x}) = 0$.

A similar elliptic differential equation, defined via the Stein operator (Oates et al., 2017), was encountered in the context of Sobolev GANs derived using the Sobolev integral probability metric (Mroueh et al., 2018). Our choice of a uniform prior in Ω_D results in the Laplacian operator, and subsequently, Poisson's PDE, which is relatively easier to solve than Stein's operator based elliptic PDE.

The optimal Lagrange multiplier λ_d^* associated with optimal WGAN-GNP discriminator $D^*(\boldsymbol{x})$ is presented next.

Lemma 3. ***Optimal Lagrange multiplier*** λ_d^* ***(n-D)*** *: Consider the n-dimensional discriminator function given by Equation* (19)*. The associated optimal Lagrange multiplier is given by*

$$\lambda_d^* = \sqrt{\frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} \sum_{i=1}^{n} \left(\left(K_{n,i}^\lambda * (p_g - p_d) \right)(\boldsymbol{x}) + a_i \right)^2 \mathrm{d}\boldsymbol{x}},$$

where $|\mathcal{X}|$ denotes the volume of the support $\mathcal{X}$, and $K_{n,i}^\lambda$ is a singular convolutional kernel given by

$$K_{n,i}^\lambda(\boldsymbol{x}) = \frac{\partial K_{D,n}(\boldsymbol{x})}{\partial x_i} = \begin{cases} \dfrac{2}{\kappa_2} \left(\dfrac{x_i}{\|\boldsymbol{x}\|} \right), & \text{for } n = 2, \text{ and} \\ \dfrac{2-n}{\kappa_n} \left(\dfrac{x_i}{\|\boldsymbol{x}\|^n} \right), & \text{for } n \geq 3. \end{cases} \tag{20}$$

The positive square-root of λ_d^* results in a $D^*(\boldsymbol{x})$ that minimizes the cost function $\mathcal{L}_D$ given in Equation (13)*.*

Proof. As in the 1-D case, the optimal Lagrange multiplier λ_d^* can be found by enforcing the gradient-norm penalty Ω_D on $D^*(\boldsymbol{x})$.

The choice of the sign of the square-root is given by the second-order (Legendre-Clebsch) condition for a minimizer: $2\lambda_d^*\mathbb{I}_n \succ 0$, where $\mathbb{I}_n$ is the n-dimensional identity matrix and $\succ$ indicates positive-definiteness. The positive root minimizes the cost, whereas the negative root maximizes it. A detailed proof is presented in Appendix C.1. □

The kernel $K_{n,i}^{\lambda}(\boldsymbol{x})$ in Equation (20) is closely related to the Riesz kernels given by $K_{R_j}(\boldsymbol{x},\boldsymbol{y}) = \frac{x_j - y_j}{\|\boldsymbol{x}-\boldsymbol{y}\|^{n+1}}$. More precisely, $K_{n,i}^{\lambda}(\boldsymbol{x},\boldsymbol{y}) = K_{R_i}(\boldsymbol{x},\boldsymbol{y})\|\boldsymbol{x}-\boldsymbol{y}\|$.

Given the optimal discriminator and the Lagrange multiplier, consider the optimization of the generator cost. Similar to the 1-D case, the Lagrangian $\mathcal{L}_G$ of the WGAN-GNP cost in $\mathbb{R}^n$ is given as follows:

$$\mathcal{L}_G = \mathcal{L}_G^{\text{WGAN}} + \lambda_p \left(\int_{\mathcal{X}} p_g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - 1 \right) + \int_{\mathcal{X}} \mu_p(\boldsymbol{x}) p_g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x},$$

where λ_p and $\mu_p(\boldsymbol{x})$ are the KKT multipliers. The following result gives the optimal generator distribution.

Theorem 4. ***Optimal WGAN-GNP generator (n-D):*** *Consider the n-dimensional generator loss $\mathcal{L}_G$ subject to the integral constraint Ω_{p_g} and non-negativity constraint Φ_{p_g}, given by*

$$\mathcal{L}_G = \int_{\mathcal{X}} \left(D^*(\boldsymbol{x}) \left(p_d(\boldsymbol{x}) - p_g(\boldsymbol{x}) \right) + \left(\lambda_p + \mu_p(\boldsymbol{x}) \right) p_g(\boldsymbol{x}) \right) \, \mathrm{d}\boldsymbol{x} - \lambda_p,$$

where $D^*(\boldsymbol{x})$ is given by Equation (19)*, and the KKT multipliers satisfy $-\infty < \mu_p(\boldsymbol{x}) \leq 0$, $\mu_p(\boldsymbol{x}) p_g(\boldsymbol{x}) = 0$, and λ_p is a finite real value. Then, the optimal solution set is*

$$p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x}), \quad \text{and} \quad \mu_p^*(\boldsymbol{x}) = 0, \ \forall \ \boldsymbol{x} \in \mathcal{X},$$

and the solution is optimal for all finite real values of λ_p.

Proof. The proof is similar to the 1-D case and is given in Appendix C.2. □

4.1 Constraint Space of the Discriminator (n-D)

Consider the multivariate setting of the first-order L_2-normed Sobolev Space $W^{1,2}(\mathcal{X},\nu)$, with the norm given by

$$\|D\|_{W^{1,2}} = \sqrt{\|D\|_{2,\nu}^2 + \|\nabla D\|_{2,\nu}^2} = \sqrt{\int_{\mathcal{X}} \|D(\boldsymbol{x})\|^2 \, \mathrm{d}\nu + \int_{\mathcal{X}} \|\nabla D(\boldsymbol{x})\|^2 \, \mathrm{d}\nu} < \infty.$$

As in the 1-D case, since we do not explicitly enforce a bound on the energy of the discriminator, by virtue of the Poincaré inequality $\|D\|_{W_0^{1,2}} \leq \mathfrak{r}\|\nabla D\|_{2,\nu}$, the WGAN-GNP discriminator can be seen as coming from the semi-normed space $W_0^{1,2}(\mathcal{X},\mathcal{U}_{\mathcal{X}})$, where $\mathcal{U}_{\mathcal{X}}$ denotes the uniform measure over $\mathcal{X}$.

4.2 Multi-dimensional Fourier-series Solution

As in the 1-D case, we solve the discriminator PDE in Equation (14) using a Fourier-series expansion, but this time considering the multivariate counterparts:

$$p_d(\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \alpha_{\boldsymbol{m}} e^{\mathrm{j}\langle \boldsymbol{\omega}_{\boldsymbol{m}}, \boldsymbol{x}\rangle},\ p_g(\boldsymbol{x}) = \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \beta_{\boldsymbol{m}} e^{\mathrm{j}\langle \boldsymbol{\omega}_{\boldsymbol{m}}, \boldsymbol{x}\rangle},\ \text{and}\ D_{FS}(\boldsymbol{x}) = \frac{1}{\lambda_d} \sum_{\boldsymbol{m}\in\mathbb{Z}^n} \gamma_{\boldsymbol{m}} e^{\mathrm{j}\langle \boldsymbol{\omega}_{\boldsymbol{m}}, \boldsymbol{x}\rangle}, \tag{21}$$

with frequency harmonics $\boldsymbol{\omega}_{\boldsymbol{m}} = [m_1\omega_1, m_2\omega_2, \ldots, m_n\omega_n]^{\mathrm{T}}$. Substituting the Fourier-series expansions in (14) and comparing terms on both sides gives

$$\gamma_{\boldsymbol{m}} = \frac{1}{2}\left(\frac{\alpha_{\boldsymbol{m}} - \beta_{\boldsymbol{m}}}{\|\boldsymbol{\omega}_{\boldsymbol{m}}\|^2}\right), \quad \boldsymbol{m} \in \mathbb{Z}^n - \{\boldsymbol{0}\}. \tag{22}$$

The value of $\gamma_{\boldsymbol{0}}$ introduces a DC offset in $D_{FS}(\boldsymbol{x})$, and without loss of generality, we set $\gamma_{\boldsymbol{0}} = 0$. Similar to the 1-D case, we have $\alpha_{\boldsymbol{m}} = \left(\frac{1}{T}\right)^n \varphi^*_{p_d}(\boldsymbol{\omega}_{\boldsymbol{m}})$ and $\beta_{\boldsymbol{m}} = \left(\frac{1}{T}\right)^n \varphi^*_{p_g}(\boldsymbol{\omega}_{\boldsymbol{m}})$, where φ^* represents the complex conjugate of the characteristic function of the corresponding distribution. We now present results on applying the WGAN-FS discriminator for 2-D Gaussian and Gaussian mixture learning problems. As in the 1-D case, we truncate the Fourier series to M terms along each dimension, and replace the complex Fourier series with its trigonometric counterpart.

4.3 Illustration Using Synthetic 2-D Data

Experimental setup: We conduct experiments on 2-D Gaussian and 8-component Gaussian mixture models (GMM). We draw Gaussian data from $\mathcal{N}(0.75\mathbf{1}_2, 0.1\mathbb{I}_2)$, where $\mathbf{1}_2$ denotes a 2-D vector with both entries equal to 1, and $\mathbb{I}_2$ denotes the 2×2 identity matrix. The noise that is input to the generator is drawn from a Gaussian $\mathcal{N}(\mathbf{0}_2, \mathbb{I}_2)$. The choice of baselines, training parameters (learning rate, batch size, Fourier-series parameters, and the Adam optimizer decay rates) and the architectures of the generator and the discriminator are identical to the 1-D scenario (cf. Section 3.7).

In the 8-component GMM experiment, isotropic Gaussians are considered with standard deviation 0.05 and means lying in $[0, 1] \times [0, 1]$. The noise that is input to the generator is drawn from $\mathcal{N}(\mathbf{0}_{100}, \mathbb{I}_{100})$. The generator architecture for all WGAN models under consideration consists of three fully connected layers of 128, 64, and 32 nodes with LeakyReLU activation in each layer. The output layer has two nodes and a sigmoid activation. The discriminator network for the baseline models is a three-layer fully connected network as in the 1-D case. The training parameters are the same as in the 1-D case.

Results: Figures 2(a) and (b) show the Wasserstein-2 distance $\mathcal{W}^{2,2}(p_d, p_g)$ between the generator and true data distributions as a function of the iterations for the WGAN and WGAN-GNP flavors under consideration, respectively, for 2-D Gaussian data. The Wasserstein-2 distance decays much faster in the case of WGAN-FS compared with the baseline variants. As in the 1-D case, we observe that replacing the baseline gradient penalty with that of WGAN-GNP results in a performance on par with the best-case baseline. Similarly, training a single-layer discriminator with a sinusoidal activation function to approximately learn

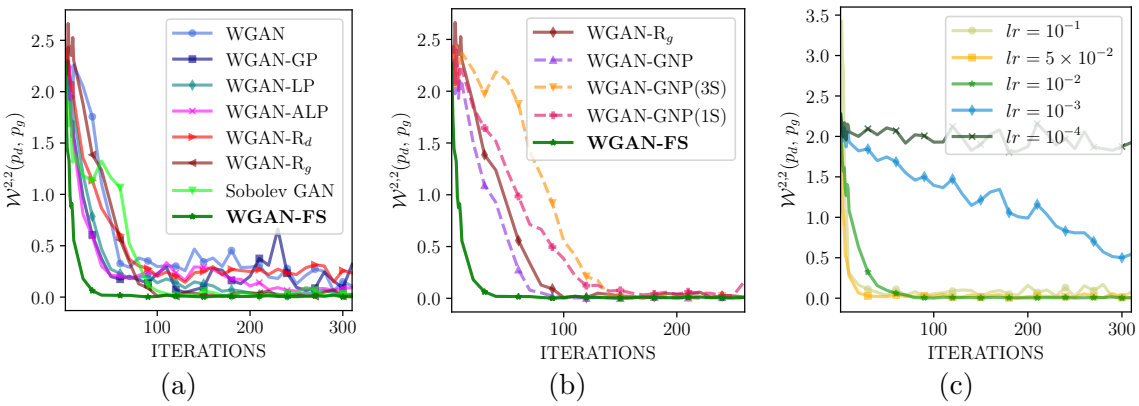


Figure 2: (Color online) Experiments on 2-D Gaussian data: (a) & (b) Wasserstein-2 distance ($\mathcal{W}^{2,2}(p_d, p_g)$) between WGAN-FS and (a) baseline WGAN variants, (b) trainable variants of the proposed WGAN-GNP. The closed-form Fourier-series approach to enforcing the gradient-norm penalty converges an order faster than the baselines and trainable variants of the same loss. (c) Wasserstein-2 distance ($\mathcal{W}^{2,2}(p_d, p_g)$) for WGAN-FS trained with different learning rates for the generator. WGAN-FS is robust to changes in the learning rate, and converges stably in terms of $\mathcal{W}^{2,2}$ for learning rates lr lower than 10^{-1}.

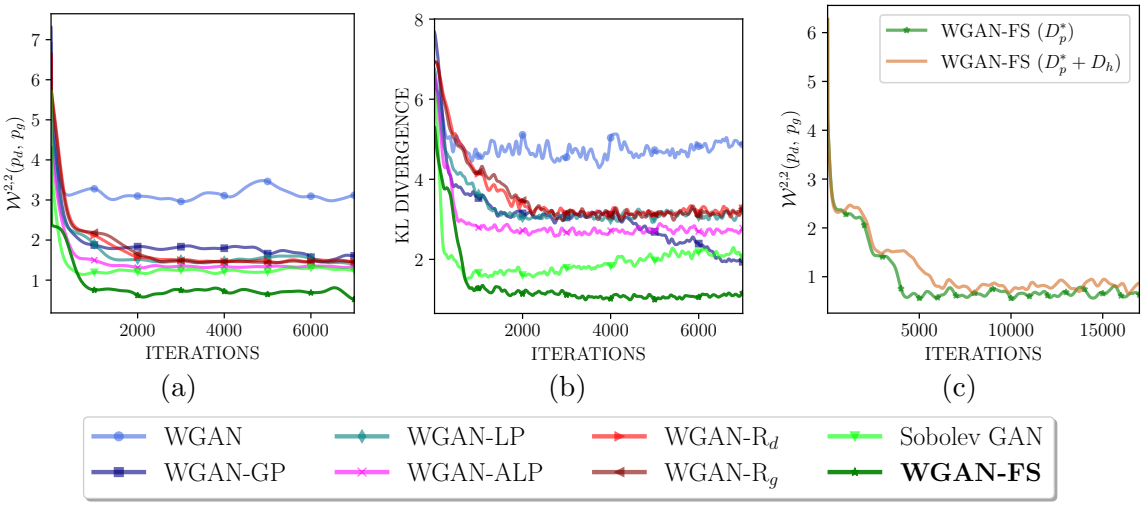


Figure 3: (Color online) Experiments on 2-D Gaussian-mixture data: Comparison of (a) Wasserstein-2 distance ($\mathcal{W}^{2,2}(p_d, p_g)$), and (b) Kullback-Leibler divergence between the data and generator distributions for WGAN-FS and baseline WGANs. WGAN-FS converges to a lower (better) value than the baselines in terms of both metrics. (c) Comparison of $\mathcal{W}^{2,2}(p_d, p_g)$ versus iterations for WGAN-FS with and without the homogeneous solution $D_h(\boldsymbol{x})$. The convergence of the WGAN-FS generator is relatively unaffected by the homogeneous component.

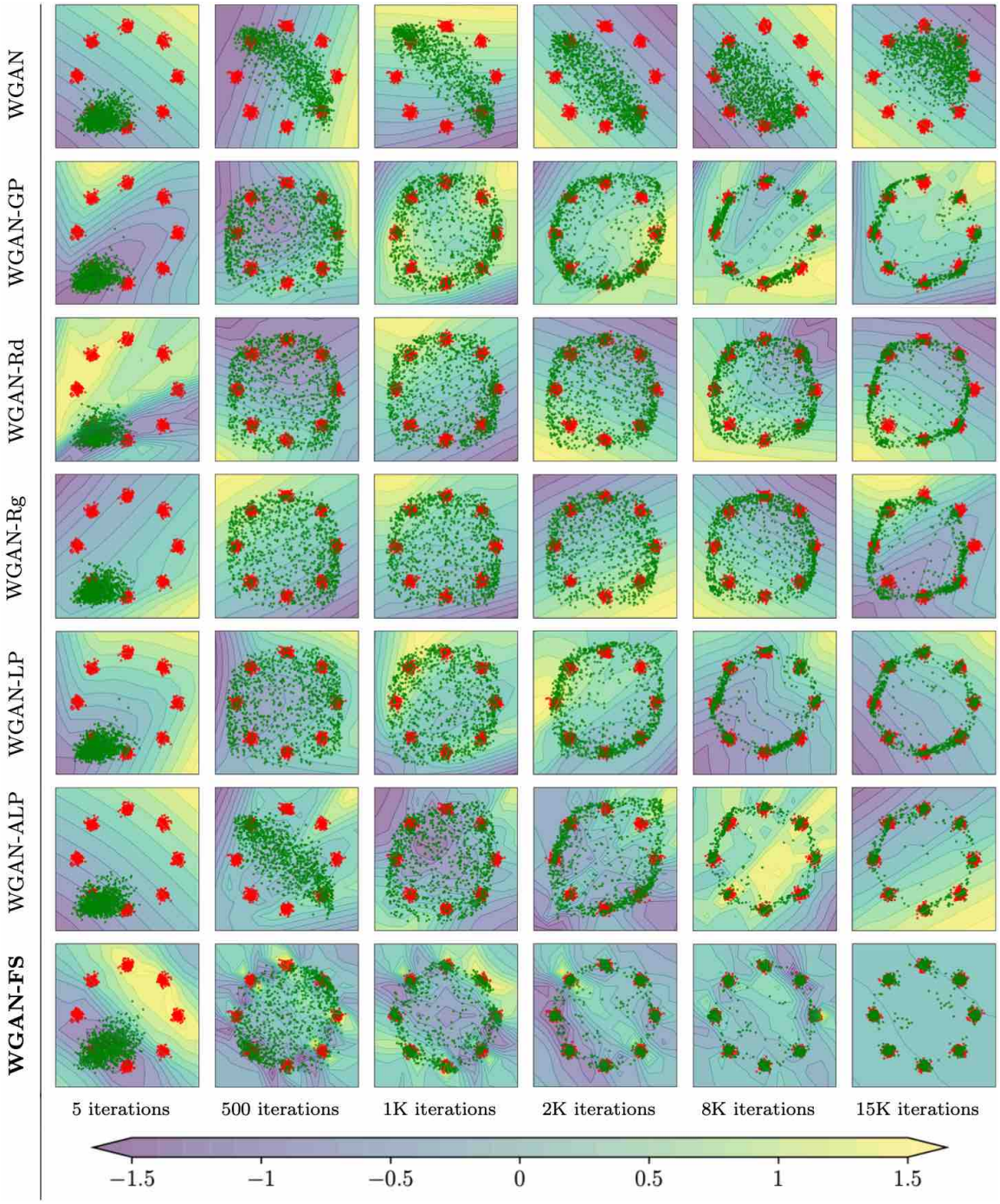


Figure 4: (Color online) Convergence of generator distribution (*green*) to the target multimodal Gaussian data (*red*) on the considered WGAN variants. The heat map represents the values taken by discriminator. The ideal $D(\boldsymbol{x})$ is the one that takes larger values at locations where $p_d > p_g$ and vice versa, converging to a constant after p_g^* approaches p_d. The Fourier-series approximation of WGAN-FS approach leads to a better representation of the discriminator during the initial iterations than the baselines, leading to faster convergence. 1K = 1000.

the Fourier coefficients results in poorer performance compared with WGAN-FS, as the suboptimal coefficients cannot represent the distributions p_d or p_g accurately. Figure 2(c) shows $\mathcal{W}^{2,2}(p_d, p_g)$ for WGAN-FS as the iterations progress considering several learning rates in the generator network. The plot indicates that learning rates in the range of 10^{-2} to 10^{-3} are optimal for smooth convergence. For lower learning rates, the convergence is not smooth as evidenced by the noise in $\mathcal{W}^{2,2}$. Learning rates larger than 0.1 resulted in the generator weights blowing up.

Figures 3(a) and (b) depict the $\mathcal{W}^{2,2}$ metric and KL divergence, respectively, as a function of iterations for the WGAN baseline models and the proposed WGAN-FS on the GMM learning task. The KL divergence is estimated parametrically by binning batches of samples to form histograms. The Wasserstein-2 distance is computed as a sample estimate using the publicly released *Python optimal transport* library (Flamary et al., 2021). We observe that, for the given choice of parameters, the baseline WGAN and WGAN-GP models latched on to different modes of the GMM at different stages of the optimization, failing to capture the entire distribution. We observe that WGAN-FS converges to lower values of the metrics compared with the baselines. Figure 4 shows the convergence of the generator distribution to the target data distribution in each case, while the associated heat-map represents the level-set of $D^*(\boldsymbol{x})$ at the given iteration. We observe that, during the initial iterations of training, WGAN-FS learns a significantly better representation of the underlying distributions compared with the baselines. This is evident from the fact that, while the baselines require optimizing a neural network for the discriminator, WGAN-FS provides the optimal discriminator for a given generator in closed form/single-shot at each iteration. Figure 3(c) compares the difference in performance of WGAN-FS with and without the homogeneous solution included. The generator optimization is independent of the homogeneous solution, with nearly identical performance in both cases, which is in accordance with the theoretical results.

5 Implementing the Fourier Series in Higher Dimensions

In higher dimensions, there is a combinatorial explosion in the number of terms — given data in $\mathbb{R}^n$, a Fourier-series expansion comprising M harmonics would have n^M terms. To get a feel for the kind of computational challenge that we are faced with, consider the MNIST dataset (LeCun et al., 1998) with data in $\mathbb{R}^{784}$. Even if one were to consider a truncated Fourier-series approximation with a mere 50 terms along each dimension, the total number of Fourier coefficients would be 784^{50}, which is of the order of 10^{144}. To gauge how big this number is, consider the following fact: According to an estimate, there are 10^{80} atoms in the known observable universe (Fermi-LAT Collaboration, 2018).

In this section, we derive bounds on the truncation error and discuss implementation related issues. Experiments on synthetic Gaussian and real-world image data validating these results are provided in Appendix E

5.1 Fourier Series in n-D

Given the Fourier-series expansions as in Equation (21), consider the case where the fundamental frequency ω_o is the same along all dimensions. Further, consider the following assumptions on p_d and p_g:

Assumption 2 (**Generator and data distributions**). *The generator and data distributions are compactly supported, ℓ-times continuously differentiable functions ($p_g, p_d \in \mathcal{C}^\ell(\mathbb{R}^n)$), with bounded energy in their gradients up to, and including order k ($p_g, p_d \in W^{k,2}(\mathcal{X})$), and vanish on the boundary of $\mathcal{X}$, i.e., we have $p_g, p_d \in \mathcal{C}^\ell(\mathcal{X}) \bigcap W^{k,2}(\mathcal{X})$, where $\ell > k$.*

It is known that such functions have rapidly decaying Fourier coefficients (Sobolev, 1963; Fefferman, 1973). Similar assumptions on the generator and data distributions were required when deriving the convergence rate of the training algorithms for Sobolev GANs (Liang, 2021). We now derive the bound on the mean-squared error incurred while truncating the Fourier series of the discriminator and p.d.f.s with the square partial sum:

$$\tilde{p}_d(\boldsymbol{x}) = \sum_{\boldsymbol{m} \in [M]^n} \alpha_{\boldsymbol{m}} e^{\mathrm{j}\omega_o \langle \boldsymbol{m}, \boldsymbol{x} \rangle}, \qquad \tilde{p}_g(\boldsymbol{x}) = \sum_{\boldsymbol{m} \in [M]^n} \beta_{\boldsymbol{m}} e^{\mathrm{j}\omega_o \langle \boldsymbol{m}, \boldsymbol{x} \rangle}, \text{ and}$$

$$\tilde{D}_{FS}(\boldsymbol{x}) = \frac{1}{\lambda_d} \sum_{\boldsymbol{m} \in [M]^n} \gamma_{\boldsymbol{m}} e^{\mathrm{j}\omega_o \langle \boldsymbol{m}, \boldsymbol{x} \rangle}, \tag{23}$$

where $[M]^n$ denotes the Cartesian product space $\{-M, -M+1, \ldots, M-1, M\}^n$.

Theorem 5. ***Bounds on the truncation error for the discriminator**: Consider the generator and data distributions coming from $\mathcal{C}^\ell(\mathcal{X}) \bigcap W^{k,2}(\mathcal{X})$, $\ell > k$, for finite k, and the infinite and truncated Fourier-series expansions defined in Equations* (21) *and* (23)*, respectively, where the coefficients $\gamma_{\boldsymbol{m}}$ are given by Equation* (22)*. The mean-squared error in truncation can be bounded as follows:*

$$\epsilon_D^2 = \|D_{FS}(\boldsymbol{x}) - \tilde{D}_{FS}(\boldsymbol{x})\|_2^2 \leq \mathfrak{C}_{n,T} \left(\frac{(M^2 n)^{-(k-2)}}{k-2} \right), \tag{24}$$

where $\mathfrak{C}_{n,T}$ is a positive constant that depends only on the dimensionality of the data (n), and the period (T).

The proof is provided in Appendix D.2. While the bound given in Theorem 5 is valid for finite k, given the truncation order M and the dimensionality of the data n, *smoother* functions (larger k) result in tighter bounds. While the ambient dimension of images is large (for example, MNIST in $\mathbb{R}^{784}$ or CelebA in $\mathbb{R}^{10^6}$), thanks to the *Manifold Hypothesis* (Kelley, 2017), it is reasonable to assume that images reside in lower-dimensional manifolds, or unions thereof (Lui et al., 2017; Khayatkhoei et al., 2018; Lei et al., 2019). We therefore propose to perform the Fourier-series approximation in learning latent representations of the data, where the bound in Equation (24) is more likely to be useful. This is effectively latent-space matching akin to that considered in Wasserstein autoencoders (WAEs) (Tolstikhin et al., 2018). The generator samples are low-dimensional representations of images learnt by an autoencoder, and the target distribution is a truncated Gaussian. We present comparisons on learning multivariate Gaussians using WGAN-FS in Appendix E.2, and learning image-space distributions with WGAN-FS in Appendix E.3.

While truncating the Fourier series is one-side of the approximation, the other is that of estimating the coefficients. Consider the Fourier-series representation of the data distribution

p_d, where $\alpha^r_{\boldsymbol{m}}$ and $\bar{\alpha}^r_{\boldsymbol{m}}$ are the true Fourier coefficient and its N-sample estimate, given by

$$\alpha^r_{\boldsymbol{m}} = \int_{\mathcal{X}} \cos(\omega_o \langle \boldsymbol{m}, \boldsymbol{x} \rangle) p_d(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \quad \text{and} \quad \bar{\alpha}^r_{\boldsymbol{m}} = \frac{1}{N} \sum_{\substack{k=1 \\ \boldsymbol{x}_k \sim p_d}}^{N} \cos(\omega_o \langle \boldsymbol{m}, \boldsymbol{x}_k \rangle), \tag{25}$$

respectively. Then, the expected error in approximating p_d through the sample estimate is given by the following theorem:

Theorem 6. ***Bound on the Fourier series approximation error for the data distribution:*** *Let the Fourier-series representation of the data distribution p_d be as given in Equation (21). Consider the true and N-sample estimates of the Fourier coefficients given in Equation (25). For finite k, the mean-squared error in approximating the truncated Fourier-series expansion of p_d can be bounded as follows:*

$$\mathbb{E}_{\boldsymbol{x}} \left[\epsilon^2_{p_d} \right] \leq \underbrace{\frac{M^n}{N} \left(1 - \frac{\mathfrak{m}_{p_d}}{n^{k+\frac{n+1}{2}}} \right)}_{\varepsilon_{stat}} + \underbrace{\mathfrak{M}_{p_d} \mathfrak{C}'_{n,k} \left(\frac{1}{M^{2k+1}} \right)}_{\varepsilon_{trunc}}, \tag{26}$$

where $\mathfrak{m}_{p_d} < \mathfrak{M}_{p_d}$ are two positive constants, $\mathfrak{C}'_{n,k}$ is a positive constant whose value depends on the dimensionality of the data n and the Sobolev order k, and ε_{stat} and ε_{trunc} represent the statistical and deterministic components of the error, respectively.

The proof is provided in Appendix D.3. A similar bound can also be derived for the generator distribution p_g. The key takeaway from Theorem 6 is that there exists a trade-off between minimizing the truncation error $\varepsilon_{\text{trunc}}$, and the statistical error $\varepsilon_{\text{stat}}$. For the approximation error $\varepsilon_{\text{stat}}$ to decay, given M and n, the batch size must increase at least at a rate of $N \approx M^{n+1}$. This results in a trade-off between discarding high-frequency components versus inaccurately estimating them due to finite sample size, which result in Gibbs' oscillations. Experiments illustrating this phenomenon on 1-D data are presented in Appendix E.1. We discuss the choice of the truncation length M in Section 5.2.

5.2 Practical Considerations in Truncated Fourier-series

Motivated by the result in Theorems 5 and 6, we make certain reasonable and simplifying assumptions on the Fourier-series expansion to circumvent the computational barrier.

In order to reduce the number of terms in the summation, we consider the fundamental frequency ω_o to be the same along all dimensions. We consider two truncation frequencies, M_{low} and M_{high}. Since images have a significant low-frequency content, we deterministically include all low-frequency components along each dimension to M_{low}. To improve the representation of high-frequency components, we perform uniform random sampling in the coefficient space between M_{low} and M_{high}. We consider a tiny subset of harmonics from the set of n^M harmonic frequencies. We pick L frequencies uniformly at random from $\mathcal{U}[M_{low}, M_{high}]$. The matrix of sampled frequencies is given as follows:

$$\mathbb{M} = \begin{bmatrix} \mathbb{U} \, , \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1} \, , \left\{ \begin{bmatrix} j & 1 & \dots & 1 \\ 1 & j & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & j \end{bmatrix}_{n \times n} \right\}_{j=2}^{M_{low}} \end{bmatrix}_{n \times L_1}, \tag{27}$$

where $\mathbb{U}$ is an $n \times L$ matrix whose elements are drawn from $(M_{low}, M_{high}]$. These simplifications together with a trigonometric Fourier expansion as in the 1-D case give rise to the following sampled Fourier-series approximation of the optimal discriminator:

$$D^*_{FS}(\boldsymbol{x}) \approx \frac{1}{\lambda^*_{FS}} \left(\frac{\gamma_0}{2} + \sum_{\boldsymbol{m} \in \mathcal{M}} \gamma^r_{\boldsymbol{m}} \cos(\omega_o \langle \boldsymbol{m}, \boldsymbol{x} \rangle) + \sum_{\boldsymbol{m} \in \mathcal{M}} \gamma^i_{\boldsymbol{m}} \sin(\omega_o \langle \boldsymbol{m}, \boldsymbol{x} \rangle) \right), \tag{28}$$

where $\mathcal{M}$ is a set comprising the columns of $\mathbb{M}$. Additionally, as in the 1-D case, the Fourier coefficients of p_d and p_g are obtained using their sample estimates computed over batches of size N:

$$\bar{\alpha}^r_{\boldsymbol{m}} \approx \frac{1}{NT} \sum_{\substack{k=1 \\ \boldsymbol{x}_k \sim p_d}}^{N} \cos(\omega_o \langle \boldsymbol{m}, \boldsymbol{x}_k \rangle), \qquad \bar{\alpha}^i_{\boldsymbol{m}} \approx \frac{1}{NT} \sum_{\substack{k=1 \\ \boldsymbol{x}_k \sim p_d}}^{N} \sin(\omega_o \langle \boldsymbol{m}, \boldsymbol{x}_k \rangle),$$

$$\bar{\beta}^r_{\boldsymbol{m}} \approx \frac{1}{NT} \sum_{\substack{k=1 \\ \boldsymbol{x}_k \sim p_g}}^{N} \cos(\omega_o \langle \boldsymbol{m}, \boldsymbol{x}_k \rangle), \text{ and } \bar{\beta}^i_{\boldsymbol{m}} \approx \frac{1}{NT} \sum_{\substack{k=1 \\ \boldsymbol{x}_k \sim p_g}}^{N} \sin(\omega_o \langle \boldsymbol{m}, \boldsymbol{x}_k \rangle).$$

Enforcing the gradient-norm penalty Ω_D on (28) results in λ^*_{FS} in n-D. The worst-case value of λ^*_{FS} satisfies:

$$\lambda^*_{FS} \approx \sqrt{(2|\mathcal{M}|+1) \left(\sum_{\boldsymbol{m} \in \mathcal{M}} (\tau^i_{\boldsymbol{m}} + \tau^r_{\boldsymbol{m}}) + \frac{1}{N} \sum_{k=1}^{N} \sum_{\boldsymbol{m} \in \mathcal{M}} (\tau^i_{\boldsymbol{m}} - \tau^r_{\boldsymbol{m}}) \cos(2\omega_o \langle \boldsymbol{m}, \boldsymbol{x}_k \rangle) \right)},$$

where

$$\tau^r_{\boldsymbol{m}} = \frac{1}{2} (\gamma^r_{\boldsymbol{m}})^2 \omega_o^2 \|\boldsymbol{m}\|^2, \text{ and } \tau^i_{\boldsymbol{m}} = \frac{1}{2} (\gamma^i_{\boldsymbol{m}})^2 \omega_o^2 \|\boldsymbol{m}\|^2,$$

and the samples $\boldsymbol{x}_k$ are drawn from the uniform mixture of p_d and p_g. The derivation is included in Appendix C.3. The quality of the sample estimates, measured in terms of the variance and mean-squared approximation error are presented in Appendix D.3. The trade-off between the truncation error (caused by discarding harmonics above a truncation order M), and the approximation error (caused by estimating the Fourier coefficients with N-sample averages), suggest that including low-frequency components improves the overall quality of estimation. Based on multiple experiments on synthetic learning tasks, we set $M_{low} = 2$, $M_{high} = 10$, and $L = 10^3$ (cf. Appendix E.2).

6 Wasserstein Adversarial Autoencoder

We extend the Fourier-series based WGAN to high-dimensional latent space matching based on Wasserstein autoencoders (WAEs) (Tolstikhin et al., 2018) and adversarial autoencoders (AAE) (Makhzani et al., 2015). In WAE, a vanilla autoencoder's (Hinton and Zemel, 1994; Schmidhuber, 2014) latent space representation is required to conform to a given *prior* distribution, usually a Gaussian or a mixture of Gaussians, through an auxiliary network that minimizes the distance between the two distributions. The *encoder-decoder* pair is

trained by minimizing an appropriate distance measured between the input data distribution and that of the reconstructed samples. In the WAE-GAN setting, the encoder of the WAE plays the role of the GAN generator, which is trained using a discriminator network to force the latent space distribution to match the prior using a suitable GAN loss. Considering the Euclidean distance metric between a data sample $\boldsymbol{x}$ and the corresponding reconstruction $\tilde{\boldsymbol{x}}$, and the SGAN loss, we obtain the AAE formulation. The vanilla WAE-GAN formulation employs the Euclidean loss for the reconstruction in combination with the KL divergence for the GAN loss. Sliced WAE (Kolouri et al., 2019) extended the framework to accommodate the sliced Wasserstein loss (Deshpande et al., 2018). As an alternative to the adversarial formulation, maximum mean discrepancy (MMD) based variations of the WAE have also gained popularity. The most recent of which, the Cramèr-Wold autoencoder (CWAE) (Knop et al., 2020) presents a characteristic kernel that allows for closed-form computation of the distance between the latent distribution and a standard Gaussian. Optimal transport based approaches either approximate the semi-discrete latent-space transportation map with the continuous Bernier potential by drawing links between the latent-space matching and the Monge-Ampère equation (Lei et al., 2019; An et al., 2020) or determine the Kantorovich potential in a two-step manner to learn a mapping from the source/prior distribution to the target using a WGAN-GP discriminator (Liu et al., 2018). Adding regularizers to the discriminator cost in AAEs has shown to improve the interpolation capabilities for the autoencoder component (Berthelot et al., 2019).

We introduce WAEFR, which is the Fourier-series representation of the discriminator integrated within the WAE-GAN framework. The block diagram is presented in Figure 5. Within WAEFR, the roles of the target and generator distributions are swapped, compared to WGAN-FS. In WAEFR, the target is the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbb{I})$, while the latent space distribution is optimized to match the target. In WAEFR, we use the mean-squared error for training the encoder-decoder pair:

$$\mathcal{L}_{AE}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2^2,$$

where $\tilde{\boldsymbol{x}} = \text{Decoder}(\text{Encoder}(\boldsymbol{x}))$ is the reconstruction of $\boldsymbol{x}$. The encoder-discriminator pair is trained using the WGAN-FS algorithm described in Section 4.2. We first solve for the Fourier coefficients $\bar{\beta}_{\boldsymbol{m}}$ and $\bar{\alpha}_{\boldsymbol{m}}$ corresponding to the latent space distribution, $p_{d_\ell} = \text{Encoder}(p_d)$, and the latent space prior p_ℓ, respectively, which gives us the closed-form discriminator. Subsequently, the encoder is trained with the WGAN loss:

$$\mathcal{L}_G = \mathbb{E}_{\tilde{\boldsymbol{z}} \sim p_{d_\ell}}[D(\tilde{\boldsymbol{z}})] - \mathbb{E}_{\boldsymbol{z} \sim p_\ell}[D(\boldsymbol{z})].$$

The training procedure for WAEFR is summarized in Algorithm 1.

6.1 Fourier-series Discriminator

To implement $D(\boldsymbol{x})$ such that the gradients flow to train the generator, the TensorFlow data handling pipeline was used gainfully by representing Equation (28) as a static two-layer network with an intermediate Fourier-series solver that computes the network weights in each iteration. Figure 6 depicts the network architecture. The operation $\omega_o \langle \boldsymbol{m}, \boldsymbol{x} \rangle$, with the associated weight matrix $\omega_o \mathbb{M}$ as given in Equation (27), constitutes the first layer and is followed by cosine/sine activations. The output of the first layer is used over batches of data

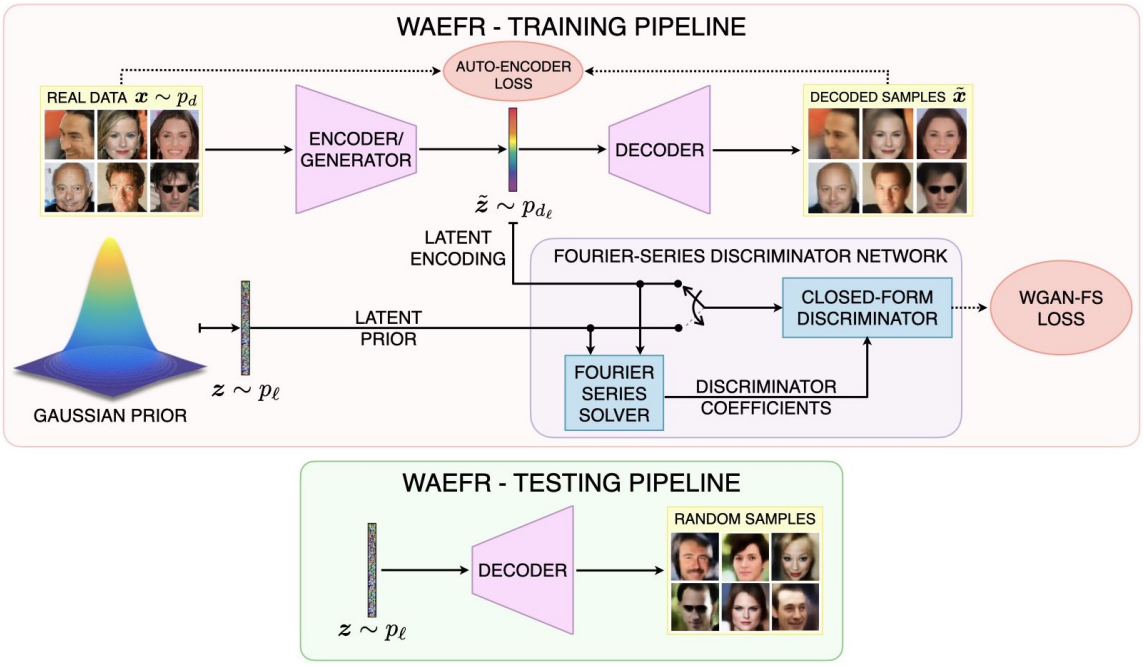


Figure 5: (Color online) The training and testing pipelines of WAEFR, the Wasserstein autoencoder with the Fourier-series representation of the discriminator. The discriminator function is evaluated in closed form based on the Fourier coefficients determined from the latent space distribution and the desired prior.

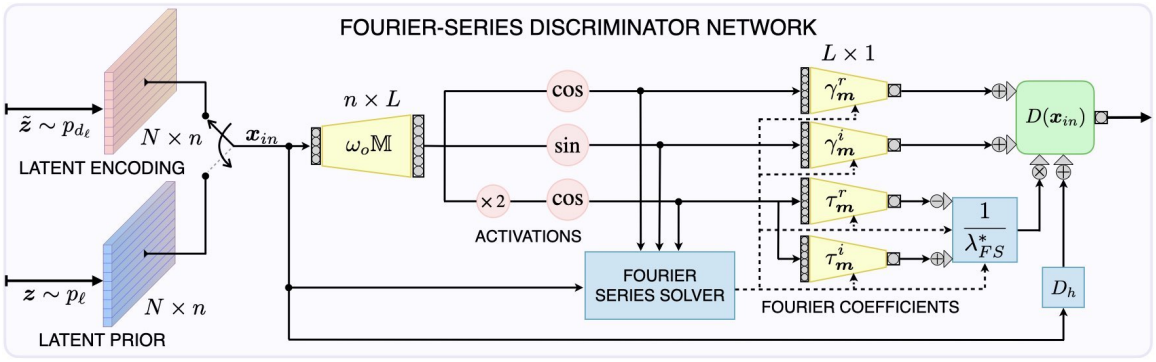


Figure 6: (Color online) Fourier-series based discriminator architecture. The latent representations are in $\mathbb{R}^n$, the data is input to the network in batches of size N, and the Fourier-series summation is truncated to L terms.

of size N to estimate the Fourier-series coefficients, from which the discriminator weights $\gamma^r_{\boldsymbol{m}}$ and $\gamma^i_{\boldsymbol{m}}$, and the parameters $\tau^r_{\boldsymbol{m}}$ and $\tau^i_{\boldsymbol{m}}$ that determine the Lagrange multiplier are evaluated (Section 5.1). Dense network connections with these weights, together with the evaluation of the homogeneous component D_h and Lagrange multiplier λ^*_{FS}, form the second layer of the discriminator network. While the first layer is fixed throughout the training of the GAN, the weights and biases of the second layer are evaluated for each batch of data.

Algorithm 1 WAEFR − Training the Wasserstein Autoencoder with a Fourier-series discriminator.

Inputs: Training data $\boldsymbol{x} \sim p_d$, prior distribution $\mathcal{N}(\mu_{\boldsymbol{z}}, \Sigma_{\boldsymbol{z}})$, batch size N, learning rate η, number of GAN pre-training epochs n_{GAN}
Models: Encoder/Generator Enc_θ; Decoder Dec_ψ; Fourier-series discriminator D^*_{FS}.
GAN pre-training:
for n_{GAN} iterations **do**
Sample: $\boldsymbol{x} \sim p_d$ − A batch of N real data samples.
Sample: $\tilde{\boldsymbol{z}} = \text{Enc}_\theta(\boldsymbol{x})$ − Latent encoding of real data.
Sample: $\boldsymbol{z} \sim \mathcal{N}(\mu_{\boldsymbol{z}}, \Sigma_{\boldsymbol{z}})$ − A batch of N prior distribution samples.
Compute: Fourier coefficients $\alpha_{\boldsymbol{m}}$ and $\beta_{\boldsymbol{m}}$
Compute: Discriminator coefficients $\gamma_{\boldsymbol{m}}$
Compute: Optimal Lagrange multiplier λ^*_{FS}
Evaluate: **WGAN-FS loss** $\mathcal{L}_G(D^*_{FS}(\tilde{\boldsymbol{z}}), D^*_{FS}(\boldsymbol{z}))$
Update: **Generator** $\text{Enc}_\theta \leftarrow \eta \nabla_\theta[\mathcal{L}_G]$
end for
WAEFR training:
while $\text{Enc}_\theta, \text{Dec}_\psi$ not converged **do**
Sample: $\boldsymbol{x} \sim p_d$ − A batch of N real samples.
Sample: $\tilde{\boldsymbol{z}} = \text{Enc}_\theta(\boldsymbol{x})$ − Latent encoding of real samples.
Sample: $\tilde{\boldsymbol{x}} = \text{Dec}_\psi(\tilde{\boldsymbol{z}})$ − Reconstructed samples.
Evaluate: **Autoencoder Loss**: $\mathcal{L}_{AE}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$
Update: **Autoencoder** $\text{Enc}_\theta \leftarrow \eta \nabla_\theta[\mathcal{L}_{AE}]$; $\text{Dec}_\psi \leftarrow \eta \nabla_\psi[\mathcal{L}_{AE}]$
Sample: $\boldsymbol{z} \sim \mathcal{N}(\mu_{\boldsymbol{z}}, \Sigma_{\boldsymbol{z}})$ − A batch of N prior distribution samples.
Compute: Fourier coefficients $\bar{\alpha}_{\boldsymbol{m}}$, $\bar{\beta}_{\boldsymbol{m}}$, and $\gamma_{\boldsymbol{m}}$
Compute: Optimal Lagrange multiplier λ^*_{FS}
Evaluate: **WGAN-FS loss** $\mathcal{L}_G(D^*_{FS}(\tilde{\boldsymbol{z}}), D^*_{FS}(\boldsymbol{z}))$
Update: **Generator** $\text{Enc}_\theta \leftarrow \eta \nabla_\theta[\mathcal{L}_G]$
end while
Output: Reconstructed random prior samples: $\text{Dec}_\psi(\boldsymbol{z})$

6.2 Experiments on Image Datasets

To illustrate the performance of WAEFR, we consider a learning task on several standard datasets: MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), CelebA (Liu et al., 2015), CIFAR-10 (Krizhevsky, 2009), and Ukiyo-E faces (Pinkney and Adler, 2020). On CIFAR-10, we consider both multi-class and single-class learning tasks.

Experimental setup: The convolutional autoencoder model proposed by Tolstikhin et al. (2018) is employed for both the baseline WAEs and WAEFR. The prior distribution is a 16-D Gaussian for MNIST, and 64-D Gaussian for the other datasets. In WAEFR, the Fourier-series period is set to $T = 15$, and the latent representation is passed through a linear activation with saturation (clipping) of the latent vector amplitudes beyond $[-10, 10]$ in order to prevent latching on to an aliased Fourier representation. Based on the analysis presented in Appendix E.2, the Fourier-series summation is truncated with $M_{low} = 2$ and $M_{high} = 10$ with $L = 10^2$ randomly sampled high-frequency terms. While the baseline WAEs uses the

deep convolutional discriminators (Tolstikhin et al., 2018), the WAEFR discriminator is as described in Section 6.1 where the weights are determined *single-shot*. This facilitates faster training of the encoder compared against an out-of-loop evaluation of $D(\boldsymbol{x})$. A batch size of 150 is used. The networks are trained using the Adam optimizer (Kingma and Ba, 2015). A learning rate of 2×10^{-4} is used for all the variants. The models are trained on 2×10^4 batches for MNIST, 5×10^4 batches for CIFAR-10, 7×10^4 batches for SVHN and 10^5 batches for CelebA and Ukiyo-E. We consider the following baselines:

- The Jensen-Shannon divergence GAN loss, which is equivalent to the base WAE configuration (Tolstikhin et al., 2018).
- The KL-divergence based Wasserstein Autoencoder (WAE-KL) (Tolstikhin et al., 2018)
- The WGAN loss, corresponding to a Wasserstein adversarial autoencoder (WAAE) with the Lipschitz penalty (Petzka et al., 2018) (WAAE-LP).
- The WGAN loss with adversarial Lipschitz penalty (Terjék, 2020) (WAAE-ALP).
- The Cramér-Wold autoencoder (CWAE) (Knop et al., 2020), which computes the Cramér-Wold distance between the latent-space and target Gaussian distributions.

The autoencoder loss is the mean-squared error in all the cases. Additionally, for all WAE baselines, the discriminator is updated thrice for every update of the generator. Pre-training the GAN component (Encoder-discriminator pair) for 10 epochs was found to result in faster convergence across all WAE-GAN variants. CelebA and Ukiyo-E images are resized to $64 \times 64 \times 3$. Pixel intensities are rescaled to $[-1, 1]$ in all experiments.

Evaluation metrics: The WAE variants are evaluated on the following metrics:

- The adversarial network's ability to match the latent and prior distributions, in terms of the **Fréchet inception distance** (FID) (Heusel et al., 2017) evaluated on batches of images decoded from prior sample vectors.
- The quality of the autoencoder's reconstructed samples, measured in terms of the **average reconstruction error** $\langle RE \rangle$ on unseen *test* set images, and defined as follows: $\langle RE \rangle = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_2^2$, where $\{\boldsymbol{x}_i\}$ are the samples and $\{\tilde{\boldsymbol{x}}_i\}$ are the corresponding reconstructions.
- The **continuity of the latent space**, demonstrated visually by decoding the interpolated points between the latent representations of two target data samples.
- The **sharpness of the decoded images** measured using the variance of the Laplacian of the image as proposed by Tolstikhin et al. (2018).

Additional details on the computation of these metrics are included in Appendix E.4. We analyze the FID and average reconstruction error as a function of the iterations. To demonstrate the continuity of the latent space, we linearly interpolate between the latent representations of the test set images and present the decoded interpolated images. We tabulate FID scores and $\langle RE \rangle$ for the converged models. For the case of single-class learning on CIFAR-10, all metrics are averaged across results obtained from training the models on

each of the ten classes, while images are presented for the *Boat* class. For WAEFR, we also plot λ^*_{FS} as a function of iterations to quantify the convergence of p_{d_ℓ} to p_ℓ.

Results: Figure 7 presents the generator loss $\mathcal{L}_G$ and the optimal Lagrange multiplier λ^*_{FS} as a function of iterations when training WAEFR on each of the datasets considered. We observe that, in each case, λ^*_{FS} converges in less than 200 iterations. This indicates that the GAN component of WAEFR converges early in the training, with the latent space of the generator taking the form of the desired prior, while subsequent training improves the accuracy of the autoencoder's mapping from the latent space to the target images.

Figure 8 presents reconstructed samples of test images for all the variants under consideration. We observe that the reconstructed image quality of WAEFR is on par with that of the baseline approaches. From the $\langle RE \rangle$ versus iterations plots shown in Figure 9, we observe that the reconstruction error of WAEFR after convergence is lower than that of the baselines on all datasets. Further, $\langle RE \rangle$ of WAEFR converges more smoothly compared with the baselines. The jitter in case of the baseline models may be attributed to the switching between the GAN and the autoencoder components of the WAE. On the other hand, since WAEFR considers a closed-form evaluation of the discriminator, the convergence behavior of the GAN, and consequently, the autoencoder, is smoother and superior. Figure 10 shows images generated by decoding randomly drawn samples from the prior distribution. While WAEFR generates images of visually comparable quality, the samples from CelebA are sharper and more diverse than the baseline models. All the variants generated more realistic images on single-class learning than on the multi-class task in CIFAR-10. Figure 11 shows the convergence of the FID score as the iterations progress. WAEFR outperforms the baselines by a significant margin when the latent space dimension is small, as in the case of MNIST, or when training with limited data, such as the Ukiyo-E dataset. WAEFR is on par with WAAE-LP and CWAE baselines on high-dimensional data.

Figures 12–17 show the images obtained by decoding interpolated points in the latent space. The first and last images in each case depict the ground truth reference images. The interpolations in WAEFR are on par with the baselines on MNIST. On the SVHN, CelebA, and Ukiyo-E datasets, WAEFR generates sharper images than the baselines. In the case of multi-class CIFAR-10, all variants failed to generate a realistic interpolation. This may be attributed to the large inter-class diversity in CIFAR-10. Table 2 presents the best-case FID scores, $\langle RE \rangle$ values, and the sharpness metric of the various models. Sharpness is evaluated in two scenarios: images obtained by decoding the latent-space interpolation between images; and the decoded samples drawn randomly following a prior distribution. WAEFR outperforms the baselines in terms of FID on CelebA, SVHN, and Ukiyo-E datasets, while achieving comparable performance on MNIST and CIFAR-10. In all the cases, WAEFR achieved the lowest reconstruction error. WAEFR also achieves up to two-fold improvement in image sharpness on face image datasets such as CelebA and Ukiyo-E in comparison with the baselines.

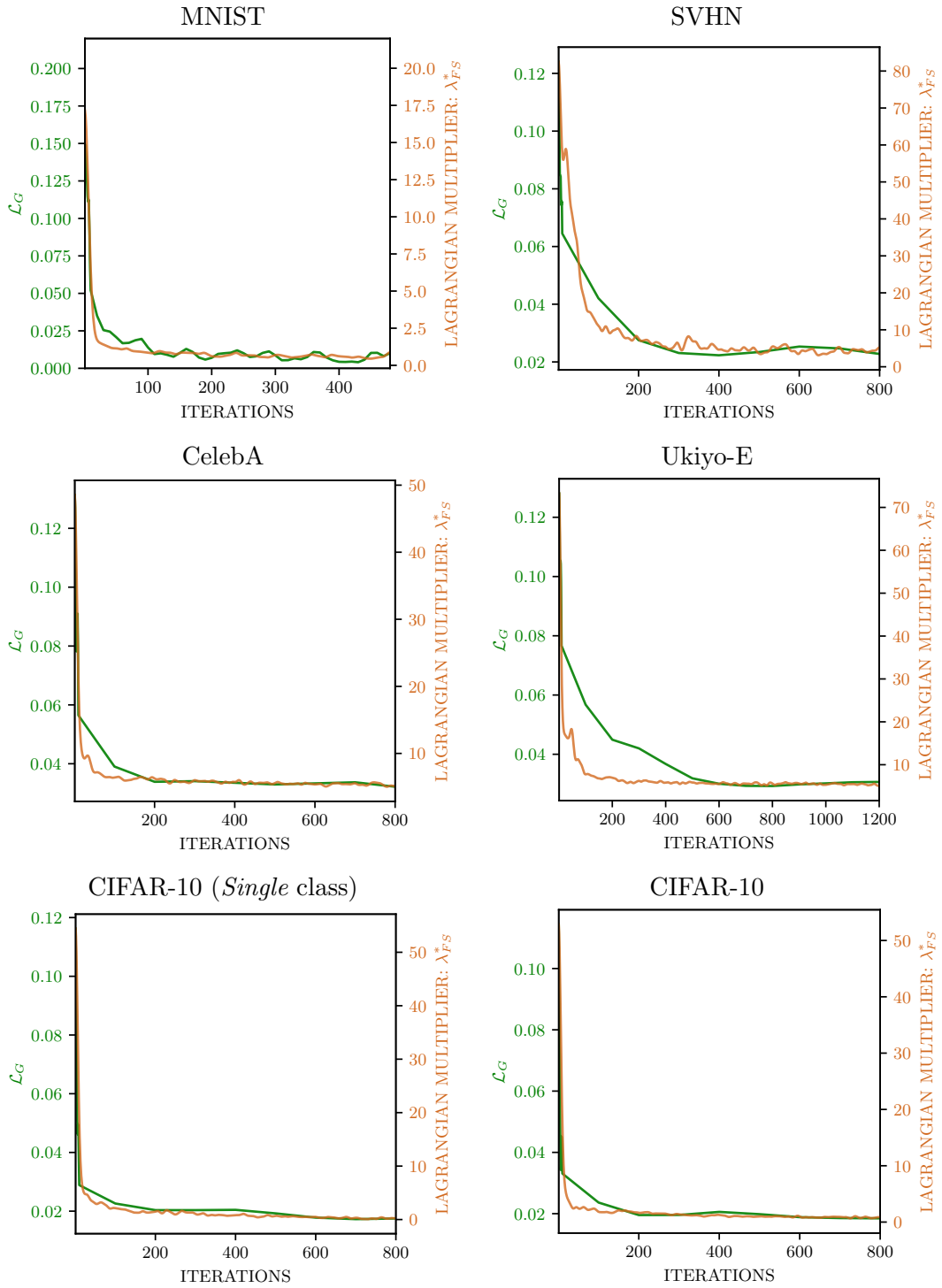


Figure 7: (Color online) Convergence of the optimal Lagrange multiplier λ^*_{FS} and generator loss $\mathcal{L}_G$ versus iterations for WAEFR trained on various datasets. The Lagrange multiplier is a measure of how quickly and stably the GAN component of the WAE converges. We observe that in all cases, the latent-space distribution matching is achieved by the GAN component in fewer than 200 iterations.

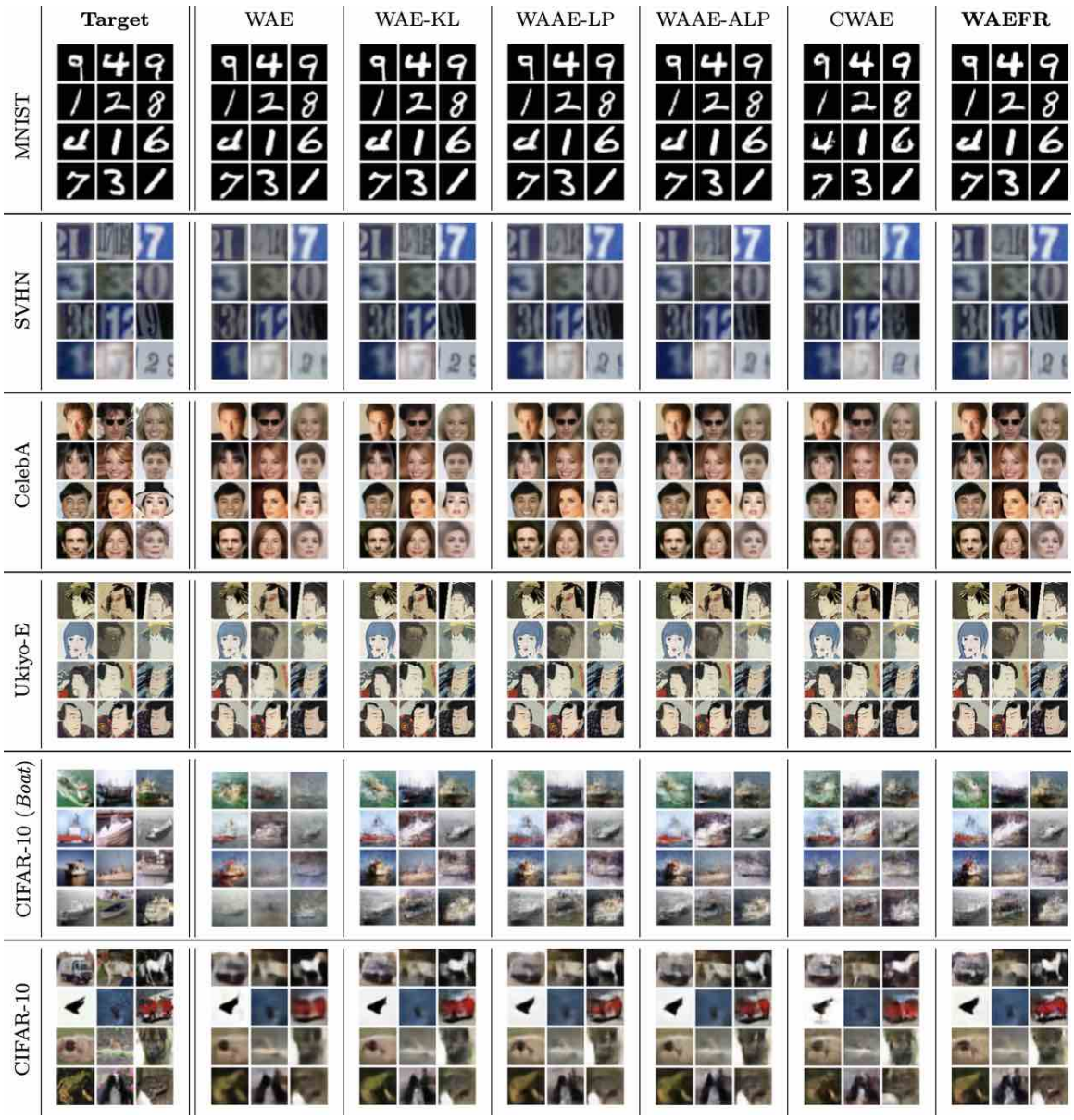


Figure 8: (● Color online) Reconstructed images from WAE, WAE-KL, WAAE-LP, WAAE-ALP, CWAE and WAEFR. While WAEFR generates sharper and more detailed reconstructions on face image datasets such as CelebA or natural image datasets such as CIFAR-10, its performance on the other datasets is on par with the baselines.

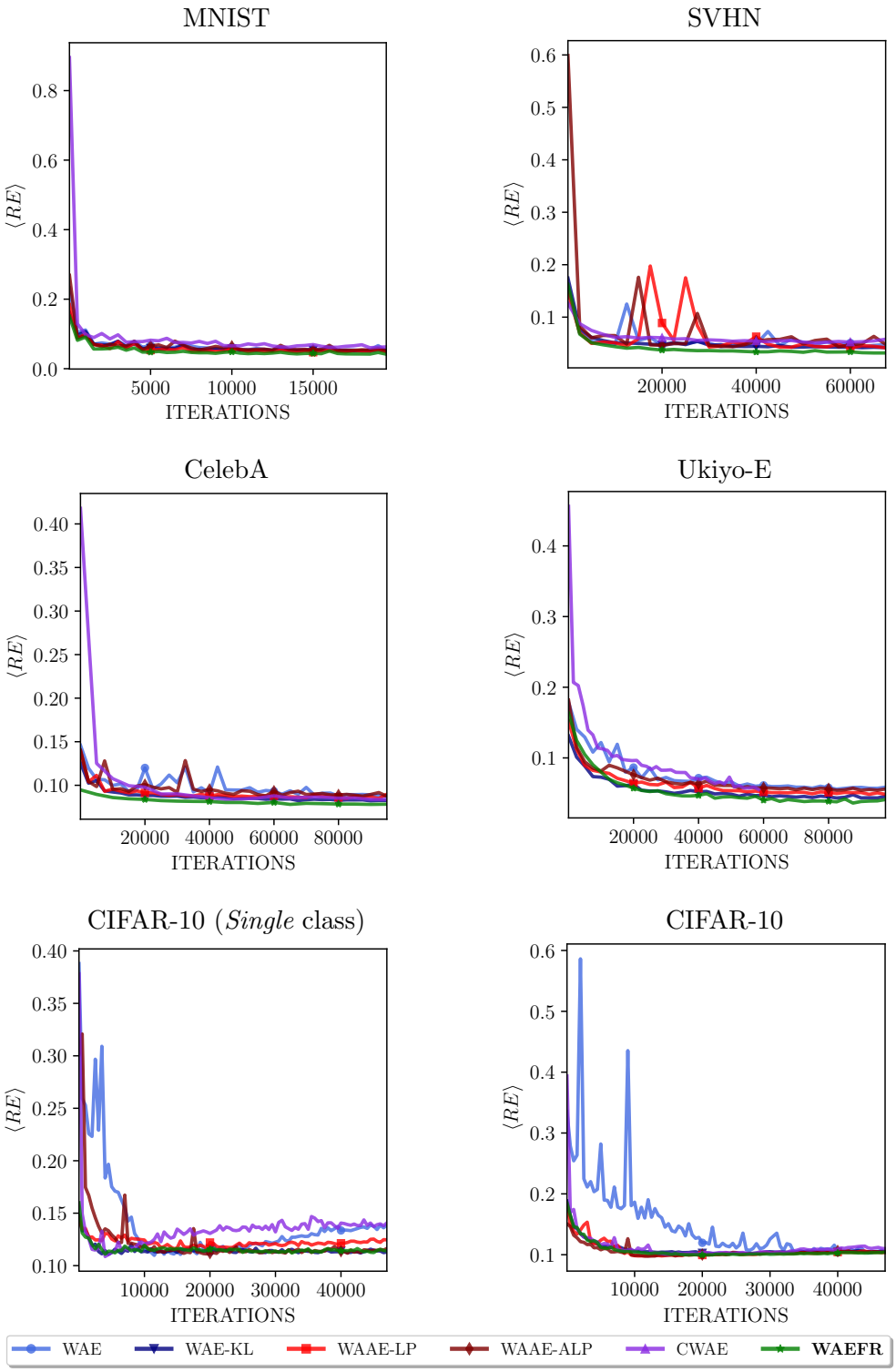


Figure 9: (Color online) Average reconstruction error $\langle RE \rangle$ versus iterations for various WAE GAN approaches considered. WAEFR converges to a lower $\langle RE \rangle$ in all the cases considered and its convergence is also smoother than the baseline variants.

Figure 10: (Color online) Images generated by WAE, WAE-KL, WAAE-LP, WAAE-ALP, CWAE and WAEFR by decoding random samples drawn from the prior distribution. WAEFR generates images of comparable quality on MNIST and CIFAR-10, while producing more diverse and sharper images on the SVHN, CelebA, and Ukiyo-E datasets.

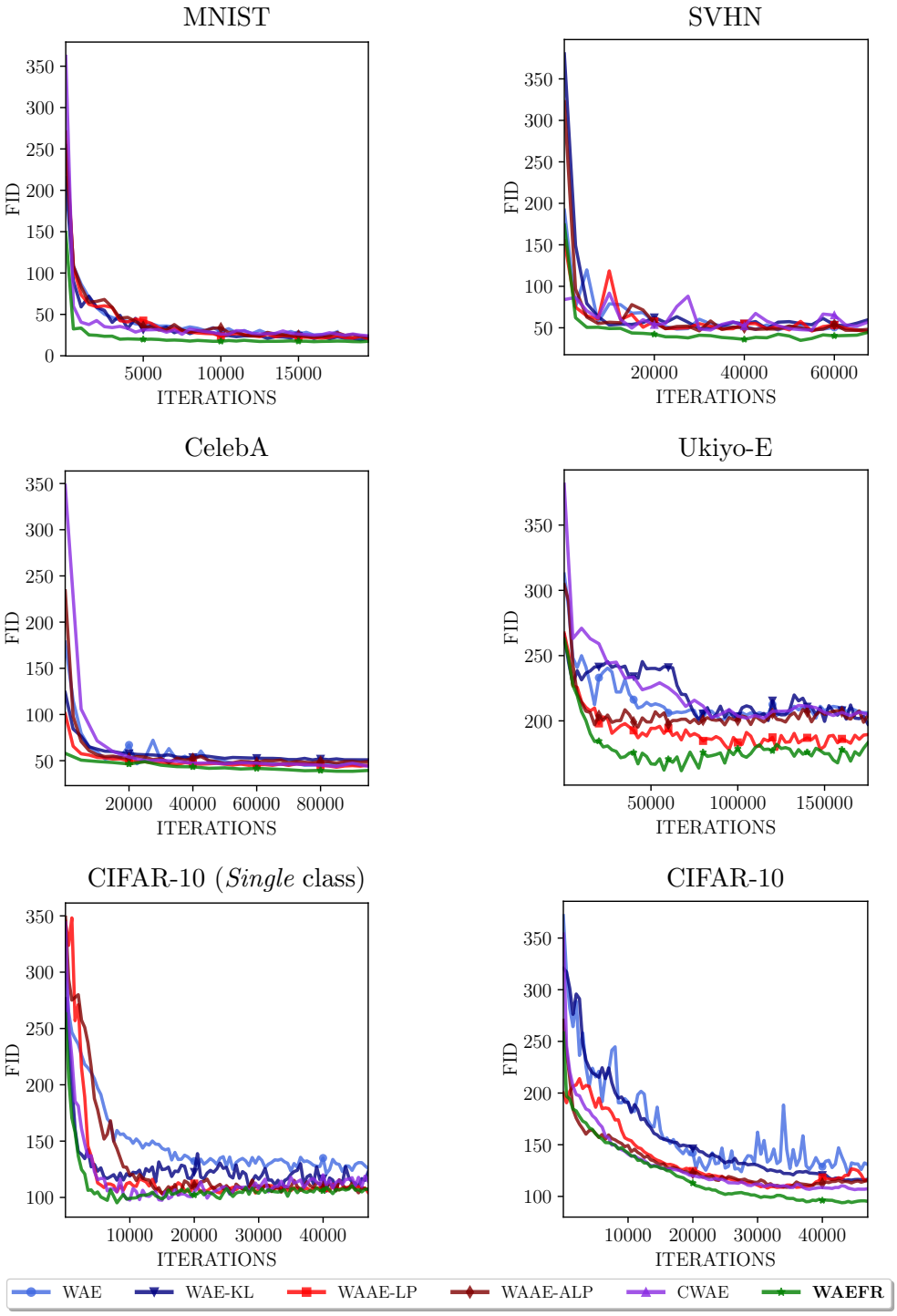


Figure 11: (Color online) FID vs. iterations for the various WAE GAN flavors considered, when evaluated on images generated by decoding randomly drawn samples following a prior distribution. WAEFR exhibits faster convergence on lower-dimensional latent-space representations (as in the case of MNIST) and comparable convergence for the higher-dimensional ones (the remaining datasets).

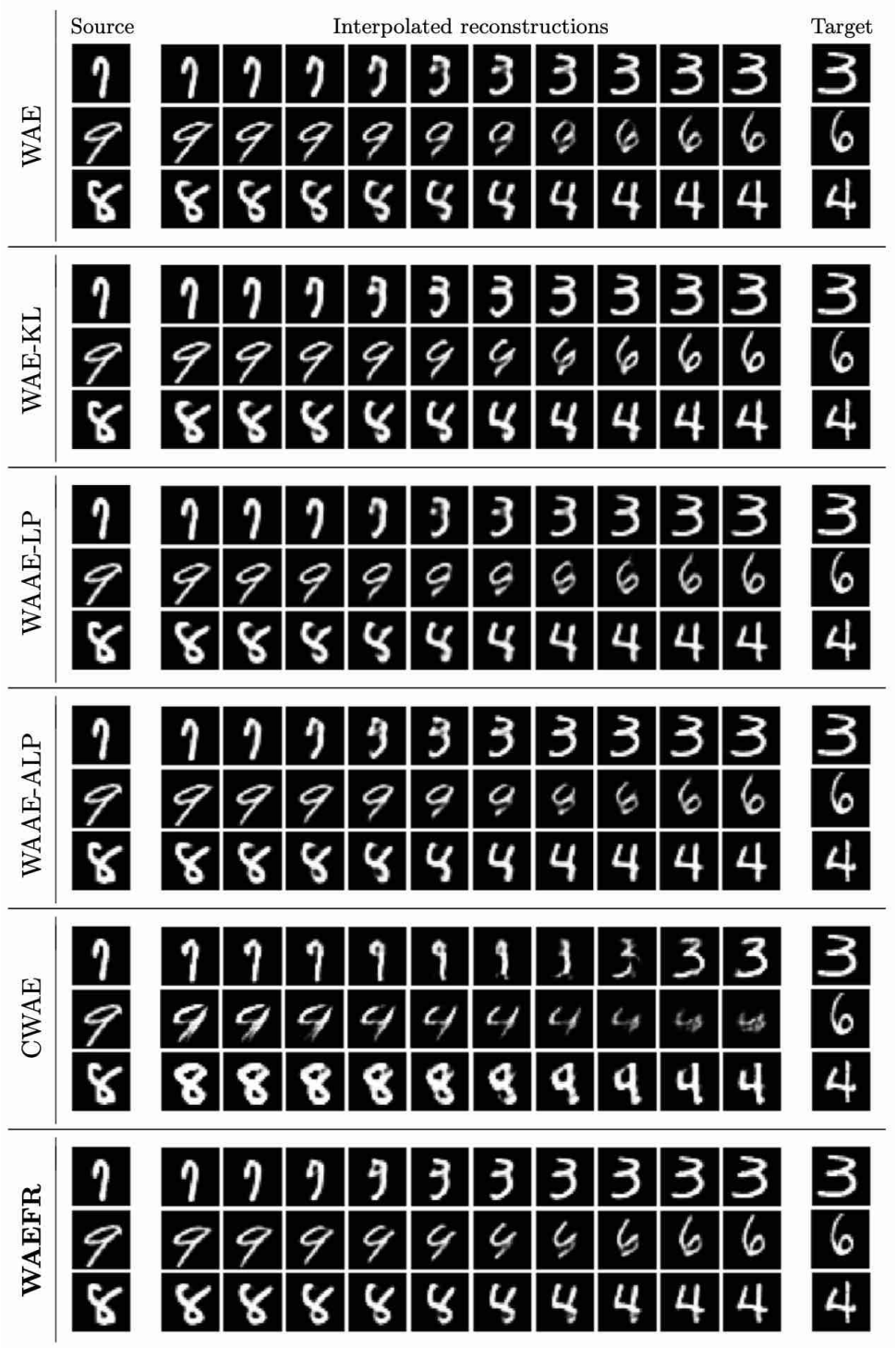


Figure 12: Images generated by decoding interpolated points between the latent space representations of two test images from the MNIST dataset. WAEFR interpolation gives rise to sharper images and superior convergence to the target image.

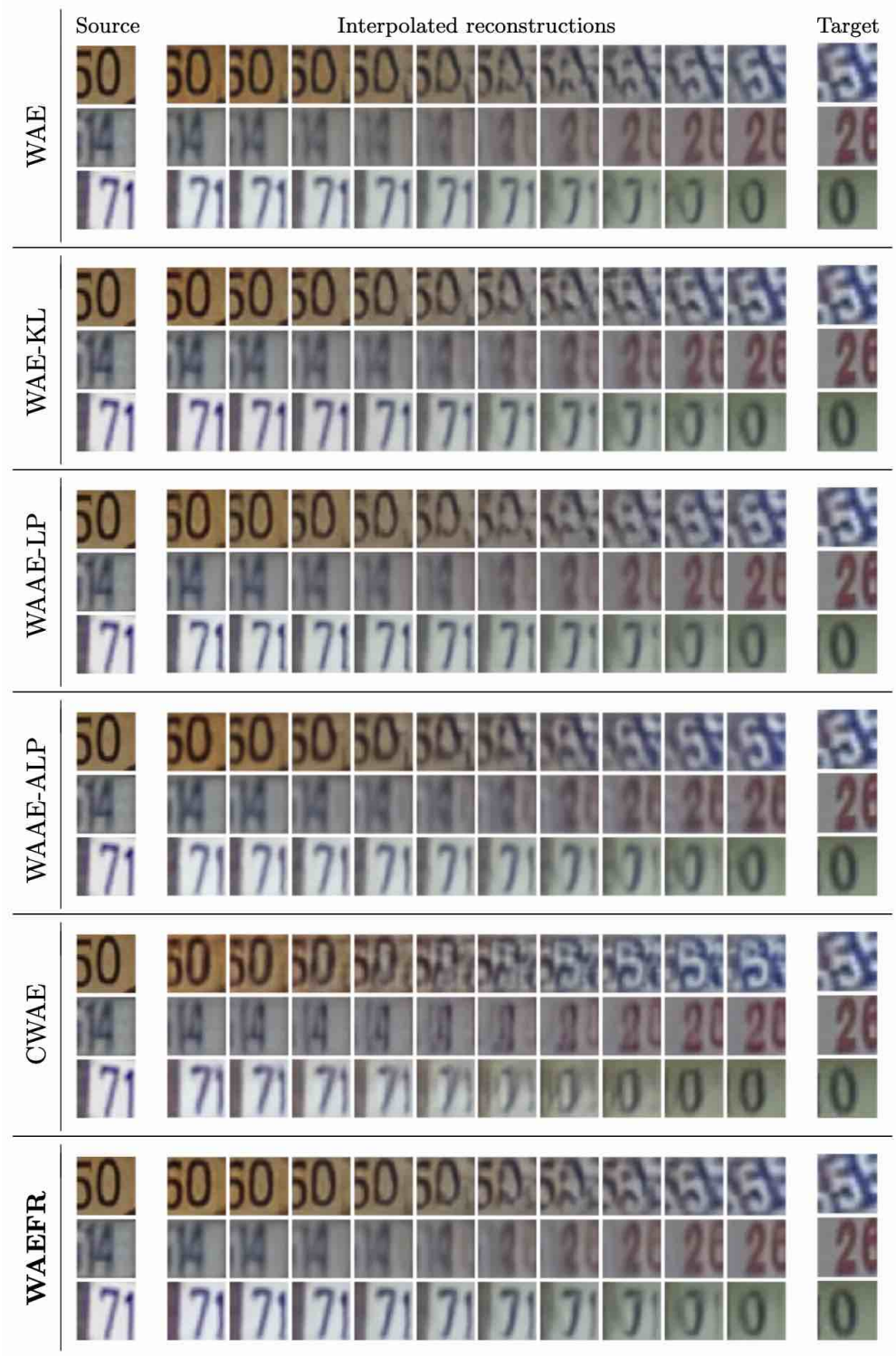


Figure 13: (Color online) Images generated by decoding interpolated points between the latent space representations of two validation set images on SVHN. Interpolations in WAEFR are sharper. CWAE fails to learn accurate reconstructions. Reconstructions of interpolated points produce visually sharper images in WAEFR.

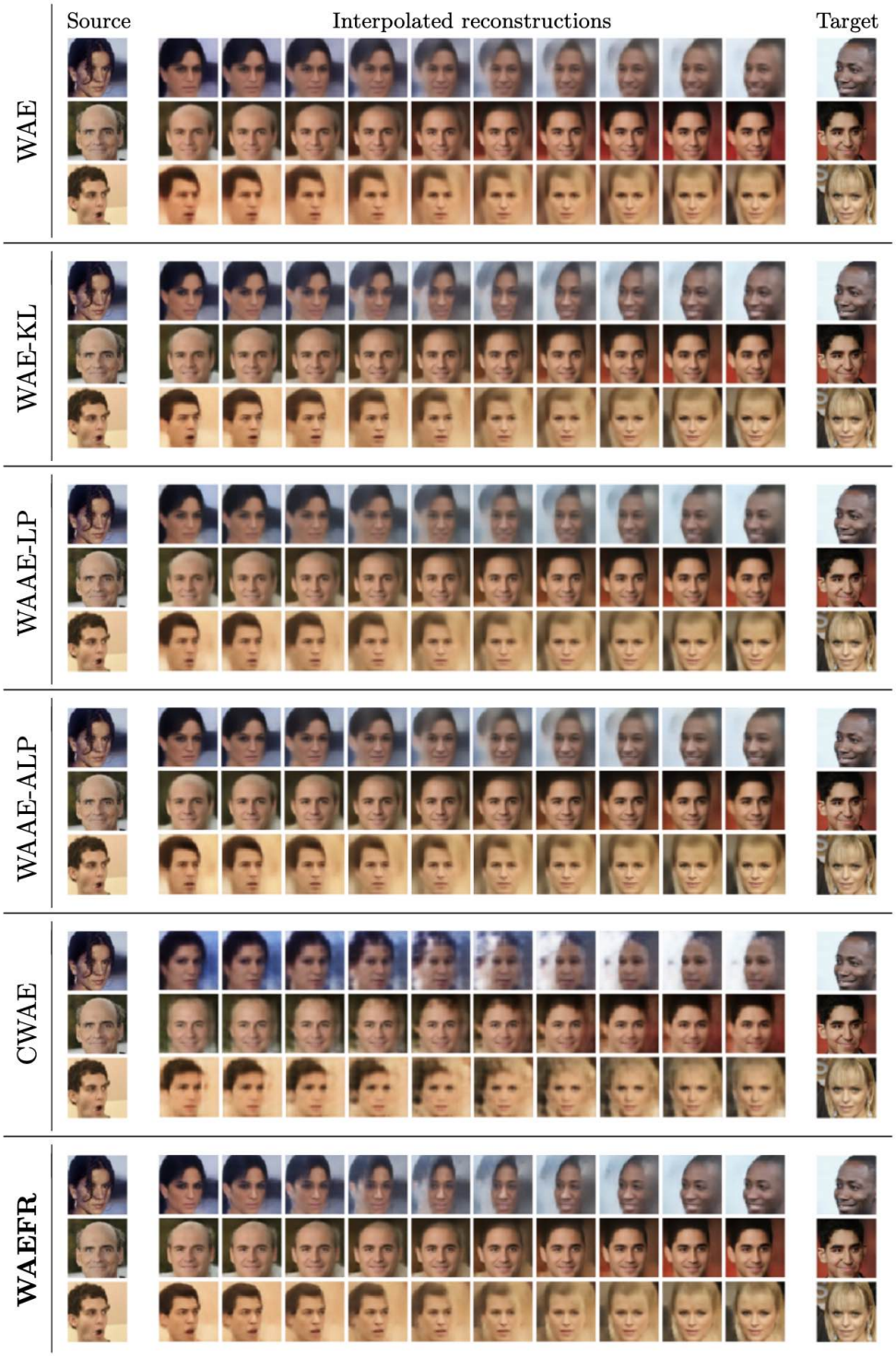


Figure 14: (Color online) Interpolation on the CelebA dataset. The images generated by WAEFR are sharper, preserve more details, and are closer to the ground truth indicating that the representations learnt by WAEFR are more accurate.

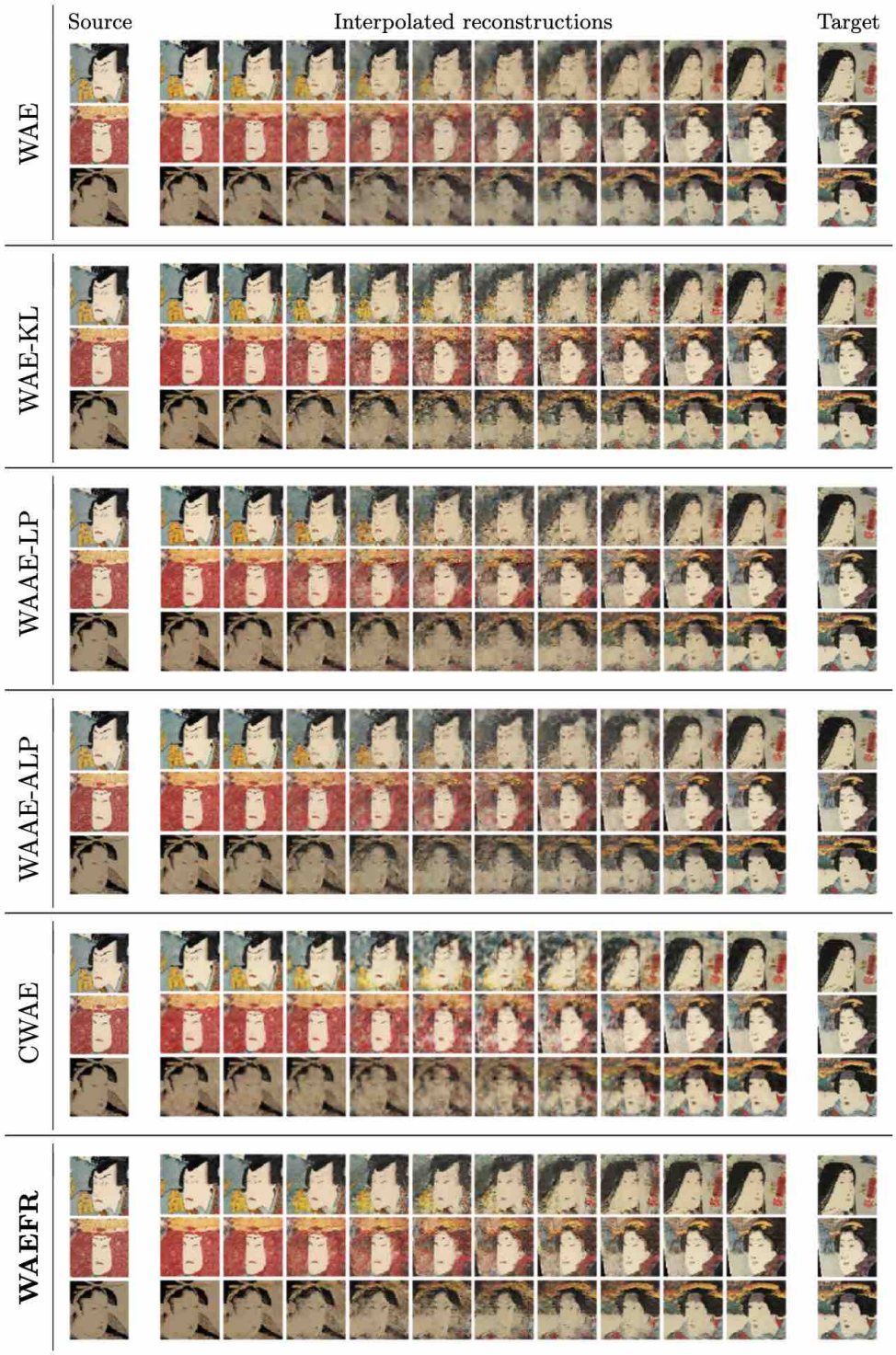


Figure 15: (Color online) Images generated by decoding interpolated points between the latent space representations of two validation set images from the Ukiyo-E faces dataset. WAEFR results in much sharper images than the baselines.

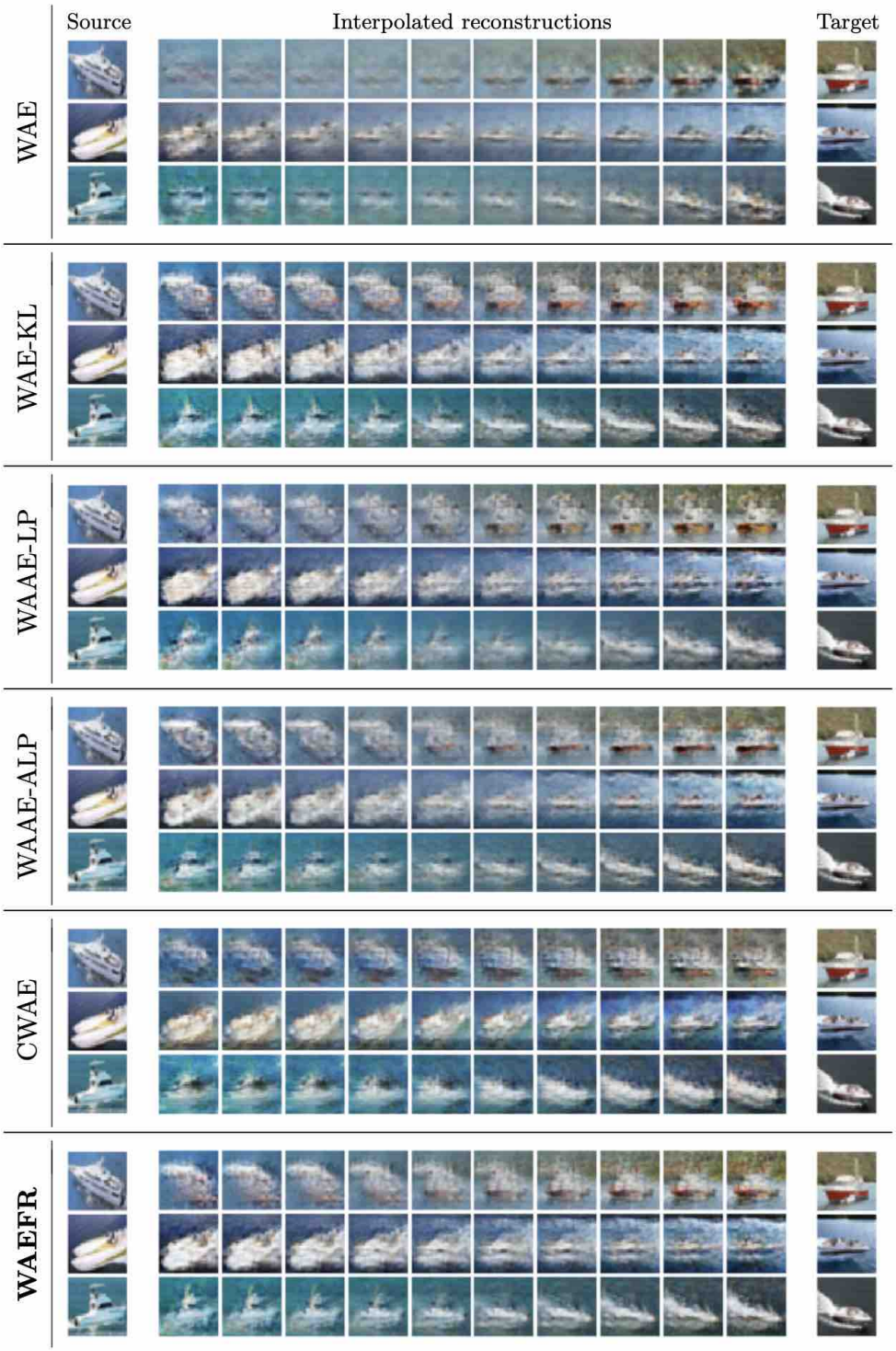


Figure 16: (Color online) Images generated by decoding interpolated points between the latent representations of two test set images from the *Boat* class of CIFAR-10. The performance of WAEFR is comparable to that of the baseline models.

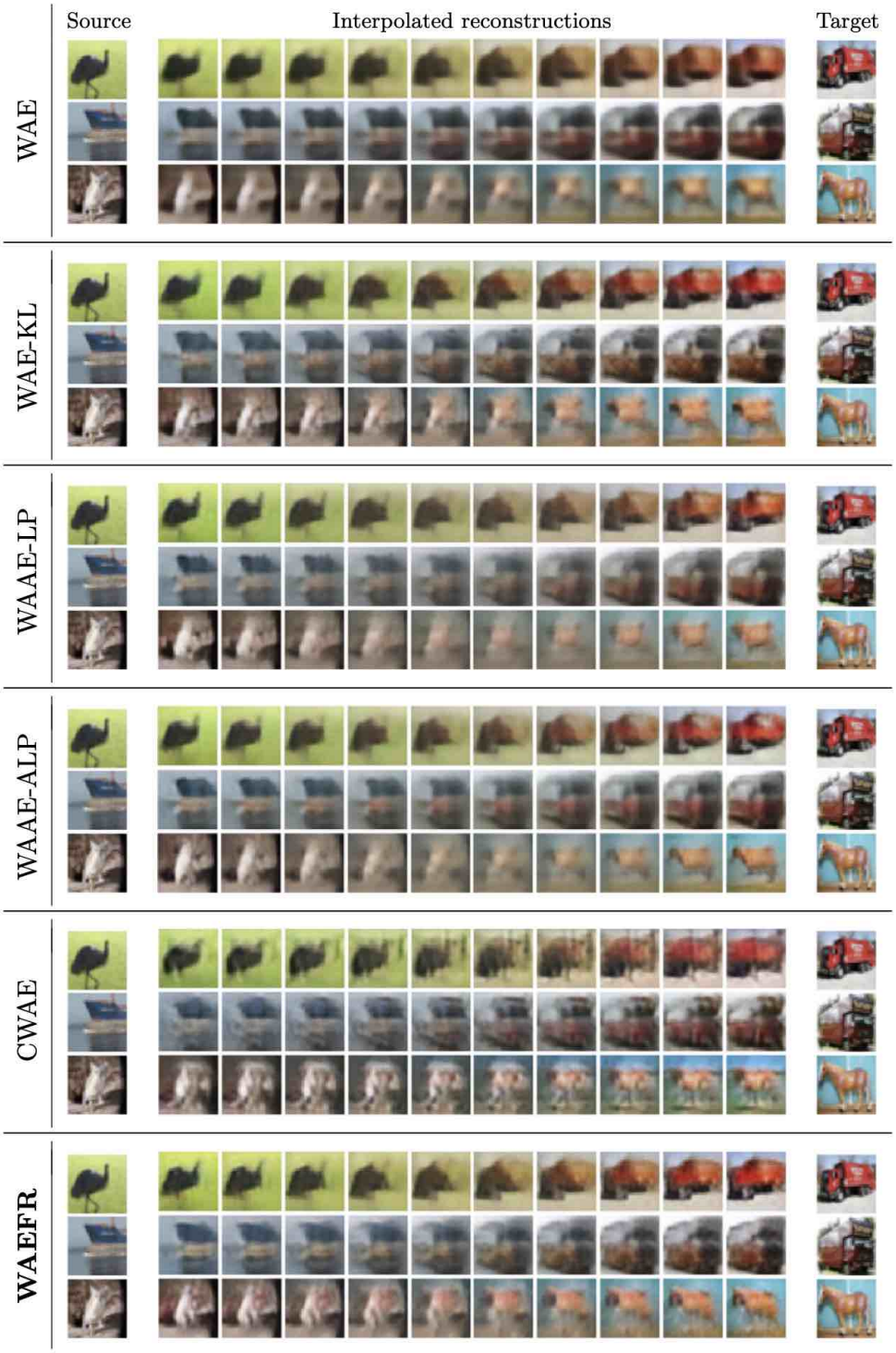


Figure 17: (Color online) Interpolation across image classes from CIFAR-10 dataset. In all the cases, the interpolated images are unrealistic, which indicates that the inter-class variability is too high to produce a semantically meaningful interpolation.

		GAN flavor	MNIST	SVHN	CelebA	Ukiyo-E	CIFAR-10 (Averaged)	CIFAR-10
FID ↓		WAE	21.676	46.083	42.943	204.446	124.165	123.88
		WAE-KL	26.231	59.717	59.223	215.013	112.650	115.96
		WAAE-LP	20.240	47.332	43.509	195.133	108.512	108.95
		WAAE-ALP	22.306	48.128	45.628	200.330	**107.509**	110.223
		CWAE	22.125	46.757	47.963	207.350	114.689	102.062
		WAEFR	**19.793**	**44.811**	**37.195**	**192.049**	108.804	**100.754**
$\langle RE \rangle \downarrow$		WAE	0.0827	0.0425	0.0939	0.0520	0.1786	0.125
		WAE-KL	0.0707	0.0380	0.0776	0.0421	0.1254	0.116
		WAAE-LP	0.0747	0.0353	0.0868	0.0429	0.1382	0.117
		WAAE-ALP	0.0836	0.0377	0.0956	0.0479	0.1294	0.119
		CWAE	0.0735	0.0478	0.0852	0.0831	0.1729	0.112
		WAEFR	**0.0693**	**0.0310**	**0.0762**	**0.0417**	**0.1227**	**0.107**
Sharpness	Random	WAE	0.1567	0.0018	0.0015	0.1210	0.0625	0.0011
		WAE-KL	0.1317	0.0014	0.0018	0.1255	0.0039	0.0032
		WAAE-LP	0.1520	0.0017	0.0044	**0.1566**	0.0155	0.0029
		WAAE-ALP	0.1609	0.0017	0.0035	0.1441	0.0150	0.0039
		CWAE	0.1703	0.0019	0.0036	0.0821	0.0158	0.0086
		WAEFR	**0.1717**	**0.0028**	**0.0084**	0.2275	**0.0194**	**0.0110**
	Interpolation	WAE	0.1681	0.0022	0.0032	0.0270	0.0035	0.0027
		WAE-KL	0.1629	0.0022	0.0044	0.0229	0.0053	0.0054
		WAAE-LP	0.1706	0.0024	0.0044	**0.0383**	0.0036	0.0041
		WAAE-ALP	0.1031	0.0019	0.0038	0.0345	0.0061	0.0031
		CWAE	0.1387	0.0028	0.0034	0.0136	0.0061	0.0045
		WAEFR	**0.1746**	**0.0029**	**0.0077**	0.0330	**0.0068**	**0.0064**
		Benchmark	0.1950	0.0051	0.0318	0.1805	0.0278	0.0361

Table 2: A comparison of various GAN flavors in terms of the performance metrics across several standard datasets. The best values are highlighted in boldface. CIFAR-10 (Averaged) corresponds to the metric computed per class followed by averaging across classes. The FID and $\langle RE \rangle$ scores (lower the better, indicated by ↓) are the best for WAEFR for almost all datasets. *Sharpness (Random)* is the sharpness computed based on random samples drawn from the prior distribution, whereas *Sharpness (Interpolation)* is the sharpness computed based on the interpolated images. *Sharpness (Benchmark)* is the sharpness value obtained over the entire dataset. Closer the sharpness to the benchmark, better is the model. We observe that WAEFR achieves almost two-fold improvement in sharpness values over the best-case baselines on CelebA.

7 Conclusion and Discussion

In this paper, we considered the functional optimization of standard GAN losses in a variational setting, by enforcing the Euler-Lagrange (EL) conditions to determine the optimum. While this approach subsumes point-wise optimization when the integral costs do not contain terms involving the gradients of the discriminator D or generator distribution p_g, the EL conditions become indispensable when gradient penalties or constraints are enforced. To truly appreciate the importance of Euler-Lagrange analysis, we considered the Wasserstein GAN subjected to a novel variant of the gradient-norm penalty (WGAN-GNP). This resulted in the optimal discriminator being the solution to a second-order partial differential equation (PDE). In principle, solving the PDE obviates the need for learning a neural network based discriminator. We showed analytically, both in the univariate as well as the multivariate settings, that the WGAN-GNP formulation results in the optimal generator distribution that matches with the desired data distribution. We did so not by assuming p_g to be a distribution, but by enforcing it through constraints of point-wise non-negativity and unit area under the curve. In addition, the discriminator PDE allows us to obtain a closed-form expression for the optimal Lagrange multiplier λ_d^* corresponding to the gradient-norm penalty, obviating the need to perform a hyperparameter search for the optimal λ_d based on empirical evidence.

The PDE connection for the optimal discriminator provides a novel viewpoint for GAN optimization. By employing a Fourier-series approximation, we showed that a *single-shot* solution can be obtained for the discriminator, given the generator. The solution relies on the estimates of the characteristic functions of the data and generator distributions. The superior performance of this novel approach was demonstrated in the univariate as well as low-dimensional multivariate Gaussian settings. A shortcoming, however, is that the approach does not scale well with the dimensionality of the data. In order to overcome this hurdle, we proposed several approximations in the high-dimensional scenario: Fourier-series model-order truncation, sample estimates of the characteristic function, and random sampling of the high-frequency harmonics. We presented bounds on the approximation error in each of these cases, which brought to light a trade-off between the truncation error and sample estimation error. While including higher-order terms improves the quality of the approximation, poorly estimating the high-frequency coefficients due to limited batches of data increases the estimation error. All of these approximations operating in the latent space of the high-dimensional data, as discovered by a Wasserstein autoencoder, resulted in a tractable model, as also demonstrated through experimental results on several standard datasets. It is important to mention that, despite several simplifications in the Fourier-series model, the proposed *single-shot* optimal discriminator in WGAN-FS and WAEFR resulted in a performance that is on par with the more sophisticated neural network counterparts, not only achieving faster convergence, but also up to two-fold improvement in the sharpness measure on face image datasets such as CelebA and Ukiyo-E.

Future Scope: The choice of Fourier bases was motivated by the specific PDE to be solved, which has a natural connection to harmonic functions, by virtue of the eigenfunction property. Owing to the orthogonality property of the Fourier bases, determining the coefficients also became significantly simpler. The Fourier approach serves as a proof of concept, and alternative, and potentially more parsimonious, bases representations (for

instance, wavelets) could also be employed for function approximation. In the context of WGAN-FS, the need for continuously differentiable distributions gives insights into potential generator architectures. Networks that can approximate the Laplacian of the discriminator function, or those with infinitely differentiable activation functions are potential directions for further exploration. Analogous to the WGAN-FS model, one could also consider bases expansions in the context of other GANs, for instance, the Stein operator (Oates et al., 2017) in Sobolev GANs. Euler-Lagrange analysis can also be employed to GAN losses that cannot be accommodated within the standard divergence minimization framework. While we presented results for non-saturating SGAN (Appendix A.3) and LSGAN (Appendix A.4), other variants such as the relativistic discriminator based GANs (Jolicoeur-Martineau, 2019) or the cycle consistent GAN (Zhu et al., 2017) could also be analyzed in the Euler-Lagrange framework. The framework would also be applicable in the scenario where the GAN loss includes derivatives of the generator distribution as well. It would also be suitable for several other regularized GAN variants (Kodali et al., 2017; Roth et al., 2017; Mescheder et al., 2018; Arbel et al., 2018) as illustrated in Appendix F.

Acknowledgements

This work is supported by the Microsoft Research Ph.D. Fellowship 2018, Qualcomm Innovation Fellowship 2019, 2021 and 2022, Robert Bosch Center for Cyber-Physical Systems Ph.D. Fellowships (2020-2021; 2021-2022), and the Science and Engineering Research Board (SERB) – Core Research Grant (CoRE), Department of Science and Technology.

GitHub Repository

The source code for the TensorFlow 2.0 (Abadi et al., 2016) implementations and the comparisons presented in this paper, the pre-trained models, and high-resolution images of the results are available at `https://github.com/DarthSid95/ELF_GANs`.

Appendix

Table of Contents

An Overview of the Appendices

The appendices are structured as follows: Appendix A presents an analysis of divergence minimizing GANs and their variants within the Euler-Lagrange framework. Appendix B provides proofs of theorems in 1-D, while extensions to n-D are presented in Appendix C. We analyze the various sources of error in the Fourier-series approximation in Appendix D. In Appendix E, we present additional experiments on learning multivariate synthetic Gaussians and image-space distributions with WGAN-FS. Appendix F contains the analysis of the Wasserstein GAN subject to the other gradient penalties proposed by Gulrajani et al. (2017) and Mescheder et al. (2018), and the SGAN and LSGAN subjected to the proposed gradient-norm penalty.

Appendix A. Euler-Lagrange Analysis of Divergence Minimizing GANs

We analyze the various divergence minimizing GANs within the variational framework and show that the degenerate Euler-Lagrange condition applies to determine the optimum in these GANs. Table 3 shows the discriminator and generator loss functions of various GANs that fall under this category. The standard GAN (SGAN) proposed by Goodfellow et al. (2014) considers both saturating and non-saturating losses. The term saturation refers to the generator gradients vanishing during training. The vanilla SGAN (employing the saturating loss) results in a min-max zero-sum game where the optimal generator minimizes the Jensen-Shannon divergence between p_d and p_g. On the contrary, the SGAN with a non-saturating loss (SGAN-NS) does not readily map to a divergence, but is preferred in a neural-network implementation as it provides better gradients at the cost of increased sensitivity to hyperparameters (Fedus et al., 2018). In the least-squares GAN (LSGAN), one minimizes the squared distance between the discriminator prediction and the class labels (a, b, c) for fake, real, and generated samples, respectively, with the optimization objective minimizing the Pearson-χ^2 divergence when $b-c=1$ and $b-a=2$ (Mao et al., 2017).

Nowozin et al. (2016) considered f-divergences of the form:

$$\mathfrak{D}_f(p_d\|p_g) = \int_{\mathcal{X}} p_g(\boldsymbol{x}) f\left(\frac{p_d(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right) \,\mathrm{d}\boldsymbol{x},$$

where the divergence function $f : \mathbb{R}_+ \to \mathbb{R}$ is convex, lower-semicontinuous and satisfies $f(1) = 0$, and $\mathcal{X}$ is a suitable domain of integration. Nowozin et al. (2016) demonstrated that, in a GAN setting, the minimization of $\mathfrak{D}_f(p_d\|p_g)$ is equivalent to the sequential minimization of the discriminator and generator losses:

$$\mathcal{L}_D^f = -\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d}[T(\boldsymbol{x})] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g}[f^c(T(\boldsymbol{x}))], \text{ and} \tag{29}$$

$$\mathcal{L}_G^f = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d}[T^*(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g}[f^c(T^*(\boldsymbol{x}))], \tag{30}$$

with respect to D and p_g, respectively, where $T(\boldsymbol{x})$ is the output of the discriminator subjected to activation g, that is, $T(\boldsymbol{x}) = g(D(\boldsymbol{x}))$ and $T^*(\boldsymbol{x}) = g(D^*(\boldsymbol{x}))$, where $D^*(\boldsymbol{x})$ is the optimal discriminator, and f^c is the Fenchel conjugate of f. The choice of the divergence f and the activation g gives rise to GANs that minimize divergences such as Kullback-Leibler (KL), reverse KL, Pearson-χ^2, squared Hellinger, Jensen-Shannon, etc.

GAN	Discriminator loss $\mathcal{L}_D$	Generator loss $\mathcal{L}_G$
SGAN	$\mathcal{L}_D^{\mathrm{S}}=-\mathbb{E}_{\boldsymbol{x}\sim p_d}[\ln D(\boldsymbol{x})]-\mathbb{E}_{\boldsymbol{x}\sim p_g}[\ln(1-D(\boldsymbol{x}))]$	$\mathcal{L}_G^{\mathrm{S}}=\mathbb{E}_{\boldsymbol{x}\sim p_d}[\ln D^*(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{x}\sim p_g}[\ln(1-D^*(\boldsymbol{x}))]$
SGAN-NS	$\mathcal{L}_D^{\mathrm{NS}}=-\mathbb{E}_{\boldsymbol{x}\sim p_d}[\ln D(\boldsymbol{x})]-\mathbb{E}_{\boldsymbol{x}\sim p_g}[\ln(1-D(\boldsymbol{x}))]$	$\mathcal{L}_G^{\mathrm{NS}}=-\mathbb{E}_{\boldsymbol{x}\sim p_d}[\ln(1-D^*(\boldsymbol{x}))]-\mathbb{E}_{\boldsymbol{x}\sim p_g}[\ln D^*(\boldsymbol{x})]$
LSGAN	$\mathcal{L}_D^{\mathrm{LS}}=\mathbb{E}_{\boldsymbol{x}\sim p_d}[(D(\boldsymbol{x})-b)^2]+\mathbb{E}_{\boldsymbol{x}\sim p_g}[(D(\boldsymbol{x})-a)^2]$	$\mathcal{L}_G^{\mathrm{LS}}=\mathbb{E}_{\boldsymbol{x}\sim p_d}[(D^*(\boldsymbol{x})-c)^2]+\mathbb{E}_{\boldsymbol{x}\sim p_g}[(D^*(\boldsymbol{x})-c)^2]$
f-GAN	$\mathcal{L}_D^{f}=-\mathbb{E}_{\boldsymbol{x}\sim p_d}[g(D(\boldsymbol{x}))]+\mathbb{E}_{\boldsymbol{x}\sim p_g}[f^c(g(D(\boldsymbol{x})))]$	$\mathcal{L}_G^{f}=\mathbb{E}_{\boldsymbol{x}\sim p_d}[g(D^*(\boldsymbol{x}))]-\mathbb{E}_{\boldsymbol{x}\sim p_g}[f^c(g(D^*(\boldsymbol{x})))]$

Table 3: A summary of various divergence minimizing GAN losses considered in this paper. The SGAN and f-GAN losses are symmetric and lead to a min-max optimization problem, whereas the LSGAN and SGAN-NS are not symmetric.

A.1 f-GANs and the Euler-Lagrange Condition

We now reformulate f-GANs subject to the non-negativity and integral constraints within the Euler-Lagrange framework. The results obtained are consistent with those available in the literature.

Theorem 7. ***Optimality of f-GANs:*** *Consider the optimization of the f-GAN losses $\mathcal{L}_D^f$ and $\mathcal{L}_G^f$ given in Equations (29) and (30), respectively. Let p_g be subject to the integral constraint $\Omega_{p_g}: \int_{\mathcal{X}} p_g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1$, and the non-negativity constraint $\Phi_{p_g}: p_g(\boldsymbol{x})\geq 0$. The f-GAN optimization is formulated as:*

$$\max_D \left\{ \int_{\mathcal{X}} \mathcal{F}_f(\boldsymbol{x}, T, p_g)\ \mathrm{d}\boldsymbol{x} \right\}, \quad \textit{and} \tag{31a}$$

$$\min_{p_g} \left\{ \int_{\mathcal{X}} \mathcal{F}_f(\boldsymbol{x}, T^*, p_g) + (\lambda_p + \mu_p(\boldsymbol{x}))\, p_g(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} \right\}, \tag{31b}$$

where

$$\mathcal{F}_f(\boldsymbol{x}, T, p_g) = \left(T(\boldsymbol{x})p_d(\boldsymbol{x}) - f^c(T(\boldsymbol{x}))p_g(\boldsymbol{x})\right), \quad \textit{and} \quad T^*(\boldsymbol{x}) = g(D^*(\boldsymbol{x})),$$

D^* being the optimal discriminator function, and λ_p and $\mu_p(\boldsymbol{x})$ being the Karush-Kuhn-Tucker (KKT) multipliers, such that $\mu_p(\boldsymbol{x})\leq 0$ and $\mu_p(\boldsymbol{x})p_g(\boldsymbol{x})=0,\ \forall \boldsymbol{x}\in\mathcal{X}$. The integrals are assumed to be well-defined over the support $\mathcal{X}$. The optimal discriminator $D^*(\boldsymbol{x})$ for a given p_g satisfies:

$$\left.\frac{\partial f^c}{\partial T}\right|_{T=T^*} = \frac{p_d}{p_g}, \tag{32}$$

and the optimal generator $p_g^*(\boldsymbol{x})$, given $D^*(\boldsymbol{x})$, satisfies:

$$\left. f^c(T^*)\right|_{p_g=p_g^*} = \lambda_p^* + \mu_p^*(\boldsymbol{x}), \tag{33}$$

where λ_p^* and $\mu_p^*(\boldsymbol{x})$ are the optimal KKT multipliers.

Proof. The proof proceeds by applying the Euler-Lagrange conditions to the costs given in Equations (31a) and (31b). Consider the f-GAN discriminator and generator losses (cf. Equations (29) and (30)):

$$\mathcal{L}_D^f = -\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d}[T(\boldsymbol{x})] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g}[f^c(T(\boldsymbol{x}))], \quad \text{and} \quad \mathcal{L}_G^f = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d}[T^*(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g}[f^c(T^*(\boldsymbol{x}))],$$

respectively. Expressing the expectations in integral form gives

$$\mathcal{L}_D^f = \int_{\mathcal{X}} (T(\boldsymbol{x})p_d(\boldsymbol{x}) + f^c(T(\boldsymbol{x}))p_g(\boldsymbol{x}))\ \mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}} \mathcal{F}_D^f(\boldsymbol{x}, T(\boldsymbol{x}))\ \mathrm{d}\boldsymbol{x}.$$

As the optimization of $\mathcal{L}_D^f$ over $D(\boldsymbol{x})$ does not involve gradient terms, the Euler-Lagrange condition applies point-wise: $\frac{\partial \mathcal{F}_D^f}{\partial D} = 0$, which yields

$$\left(p_d - p_g \frac{\partial f^c(T)}{\partial T}\right) \frac{\partial T}{\partial D}\bigg|_{T=T^*} = 0, \tag{34}$$

where

$$T = g(D) \quad \Rightarrow \quad \frac{\partial T}{\partial D} = g'(D),$$

g being the activation function at the output of the discriminator network. Based on the f-GAN formulations of Nowozin et al. (2016) (cf. Column 2 of Table 4), we observe that $g'(D) \neq 0,\ \forall\ \boldsymbol{x}$ such that $D(\boldsymbol{x}) \neq 0$. This yields the first result of Theorem 7:

$$\frac{\partial f^c(T)}{\partial T}\bigg|_{T=T^*} = \frac{p_d}{p_g}.$$

The optimal discriminator $D^*(\boldsymbol{x})$ is the one that satisfies the above equation with the corresponding output function $T^* = g(D^*)$. Given T^*, the generator optimization is carried out on $\mathcal{L}_G^f$, subject to the integral constraint Ω_{p_g} and non-negativity constraint Φ_{p_g}. The Lagrangian of the cost becomes

$$\mathcal{L}_G^f = \int_{\mathcal{X}} (T^* p_d + f^c(T^*)p_g + \lambda_p p_g + \mu_p(\boldsymbol{x})p_g)\ \mathrm{d}\boldsymbol{x}, \tag{35}$$

where λ_p and $\mu_p(\boldsymbol{x})$ are the KKT multipliers. Applying the EL condition to $\mathcal{L}_G^f$ gives:

$$\left(p_d - p_g \frac{\partial f^c(T^*)}{\partial T^*}\right) \frac{\partial T^*}{\partial D^*} \frac{\partial D^*}{\partial p_g} - f^c(T^*) + \lambda_p + \mu_p(\boldsymbol{x}) = 0.$$

Using (34) gives the condition that the optimal generator p_g^* must satisfy:

$$f^c(T^*)\big|_{p_g=p_g^*} = \lambda_p + \mu_p(\boldsymbol{x}).$$

The feasible KKT multipliers satisfy the integral and non-negativity constraints when enforced on p_g^*:

$$\int_{\mathcal{X}} p_g^*(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} = 1, \quad \text{and} \quad p_g^*(\boldsymbol{x}) \geq 0,\ \forall\ \boldsymbol{x} \in \mathcal{X}.$$

f-divergence	$g(D)$	$f^c(T)$	$D^*(\boldsymbol{x})$	$p_g^*(\boldsymbol{x})$	(λ_p^*, μ_p^*)
Kullback-Leibler (KL)	D	e^{T-1}	$1 + \ln\left(\frac{p_d}{p_g}\right)$	$\frac{p_d(\boldsymbol{x})}{\lambda_p^*+\mu_p^*}$	$(1, 0)$
Reverse KL	$-e^{-D}$	$-1 - \ln(-T)$	$\ln\left(\frac{p_d}{p_g}\right)$	$\frac{p_d(\boldsymbol{x})}{e^{\lambda_p^*+\mu_p^*+1}}$	$(-1, 0)$
Pearson-χ^2	D	$\frac{1}{4}T^2 + T$	$2\left(\frac{p_d-p_g}{p_g}\right)$	$\frac{p_d(\boldsymbol{x})}{\sqrt{\lambda_p^*+\mu_p^*+1}}$	$(0, 0)$
Squared-Hellinger	$1 - e^{-D}$	$\frac{T}{1-T}$	$\frac{1}{2}\ln\left(\frac{p_d}{p_g}\right)$	$\frac{p_d(\boldsymbol{x})}{(\lambda_p^*+\mu_p^*+1)^2}$	$(-2, 0)$
SGAN	$\ln\left(\frac{e^D}{e^D+1}\right)$	$\ln\left(\frac{1}{1-e^T}\right)$	$\ln\left(\frac{p_d}{p_g}\right)$	$\frac{p_d(\boldsymbol{x})}{e^{\lambda_p^*+\mu_p^*}-1}$	$(\ln 2, 0)$

Table 4: The optimal discriminator $D^*(\boldsymbol{x})$ and generator p_g^* for various f-GANs (Nowozin et al., 2016) given the activation function g and the Fenchel conjugate f^c. The optima $D^*(\boldsymbol{x})$ and p_g^* are the solutions to Equations (32) and (33), respectively.

In addition, $\mu_p(\boldsymbol{x}) \leq 0$ and it satisfies the complementary slackness condition

$$\mu_p(\boldsymbol{x}) p_g^*(\boldsymbol{x}) = 0, \ \forall \ \boldsymbol{x} \in \mathcal{X},$$

which gives $\mu_p(\boldsymbol{x}) = 0$ whenever $p_g^*(\boldsymbol{x}) > 0$, and $\mu_p(\boldsymbol{x}) \leq 0$ whenever $p_g^*(\boldsymbol{x}) = 0$. For all $\boldsymbol{x}$ such that $p_d(\boldsymbol{x}) = 0$, the generator cost evaluated at the optimal generator distribution $\mathcal{L}_G^f(p_g^*, \lambda_p, \mu_p(\boldsymbol{x}))$ becomes zero, and subsequently, the choice of $\mu_p^*(\boldsymbol{x})$ over this set is immaterial. Subsequent optimization of $\mathcal{L}_G^f(p_g^*, \lambda_p, \mu_p(\boldsymbol{x}))$, $\forall \ \boldsymbol{x}$ such that $p_d(\boldsymbol{x}) > 0$, over the KKT multipliers gives us the optimal multipliers λ_p^* and $\mu_p^*(\boldsymbol{x})$. This yields the second result of Theorem 7:

$$f^c(T^*)\big|_{p_g=p_g^*} = \lambda_p^* + \mu_p^*(\boldsymbol{x}).$$

□

When applied to any f-GAN variant (Nowozin et al., 2016), Theorem 7 yields the optimal discriminator D^* and optimal generator p_g^*, as listed in Table 4, which are consistent with the results known in f-GAN literature. We provide the proofs in Appendix A.2. Enforcing the complementary slackness condition: $\mu_p^*(\boldsymbol{x}) p_g^*(\boldsymbol{x}) = 0, \ \forall \ \boldsymbol{x} \in \mathcal{X}$, in addition to the integral and non-negativity constraints yields the optimal λ_p^* and $\mu_p^*(\boldsymbol{x})$ for each f-GAN. We observe from Table 4 that the optimal generator distribution $p_g^*(\boldsymbol{x})$ is non-negative for all f-GAN variants. Consequently, the optimal KKT multiplier $\mu_p^*(\boldsymbol{x}) = 0, \ \forall \ \boldsymbol{x}$ such that $p_d(\boldsymbol{x}) > 0$. The calculations show that it suffices to enforce only the integral constraint and the optimal solution automatically satisfies the non-negativity constraint. For the f-GAN variants considered, the optimal generator is $p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x})$, which is indeed the desired solution of the GAN optimization.

A.2 The Optimality of f-GANs

We now consider each of the f-GAN variants proposed by Nowozin et al. (2016) and present the optimal discriminator and generator functions in each case.

KL divergence: As an illustration, consider the f-GAN formulation with Kullback-Leibler divergence, which corresponds to $g(D) = D$ and $f^c(T) = e^{T-1}$. Following Theorem 7, the optimal discriminator and generator are given by:

$$D^*(\boldsymbol{x}) = 1 + \ln\left(\frac{p_d(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right) \text{ and } p_g^*(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x})}{\lambda_p + \mu_p(\boldsymbol{x})},$$

respectively. The support of the solution is restricted to $\text{supp}(p_d)$. We split the support $\mathcal{X}$ into two disjoint sets: $\mathcal{X}_+ = \{\boldsymbol{x} \mid p_d(\boldsymbol{x}) > 0\}$ and $\mathcal{X}_0 = \{\boldsymbol{x} \mid p_d(\boldsymbol{x}) = 0\}$. The loss $\mathcal{L}_G(p_g^*, \lambda_p, \mu_p(\boldsymbol{x}))$ vanishes everywhere outside $\mathcal{X}_+$, and hence the optimization is undefined over $\mathcal{X}_0$. Within $\mathcal{X}_+$, enforcing the complementary slackness condition on μ_p gives:

$$\mu_p^*(\boldsymbol{x}) = 0, \ \forall \ \boldsymbol{x} \text{ such that } p_d(\boldsymbol{x}) > 0.$$

Enforcing the integral constraint yields $\lambda_p^* = 1$ as the only feasible solution. This gives us the optimal generator distribution:

$$p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x}), \ \forall \ \boldsymbol{x} \in \mathcal{X}.$$

In summary, the optimal generator perfectly agrees with the data distribution.

Reverse-KL divergence: The EL analysis of the reverse-KL divergence based f-GAN closely follows the analysis for the KL-divergence based GAN. We have $g(D) = -e^{-D}$ and $f^c(T) = -1 - \ln(-T)$. Following Theorem 7, we obtain the optimal discriminator and generator functions as

$$D^*(\boldsymbol{x}) = \ln\left(\frac{p_d(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right) \quad \text{and} \quad p_g^*(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x})}{e^{\lambda_p + \mu_p(\boldsymbol{x}) + 1}},$$

respectively. Enforcing the complementary slackness condition $\mu_p(\boldsymbol{x}) p_g^*(\boldsymbol{x}) = 0$, we obtain the condition: $\frac{\mu_p(\boldsymbol{x})}{e^{\lambda_p + \mu_p(\boldsymbol{x}) + 1}} = 0, \ \forall \ \boldsymbol{x}$ such that $p_d(\boldsymbol{x}) > 0$. In conjunction with $\mu_p(\boldsymbol{x}) \leq 0$, we obtain $\mu_p^*(\boldsymbol{x}) = 0$ as the only feasible solution. As in the case of the KL-divergence f-GAN, the EL analysis is applicable only when the integrand in the cost is non-zero, that is, $p_g(\boldsymbol{x}) > 0$. Enforcing the integral constraint results in $\lambda_p = -1$ as the only solution. This gives us the optimal KKT multipliers $\mu_p^*(\boldsymbol{x}) = 0, \ \forall \ \boldsymbol{x} \in \mathcal{X}_+$, and $\lambda_p^* = -1$. The corresponding optimal generator distribution is $p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x})$.

Jensen-Shannon divergence: In this case, the f-GAN formulation considers $g(D) = \ln(2) - \ln\left(1 + e^{-D}\right)$ and $f^c(T) = -\ln\left(2 - e^{T}\right)$. From Theorem 7, we have

$$D^*(\boldsymbol{x}) = \ln\left(\frac{p_d(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right) \quad \text{and} \quad p_g^*(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x})}{2e^{(\lambda_p + \mu_p(\boldsymbol{x}))} - 1}.$$

Enforcing the integral, non-negativity, and complementary slackness constraints, we obtain the only feasible (and therefore optimal) set of KKT multipliers: $\mu_p^*(\boldsymbol{x}) = 0,\ \forall\ \boldsymbol{x} \in \mathcal{X}_+$ and $\lambda_p^* = 0$. Since both KKT multipliers are zero, it can be verified that unconstrained optimization over p_g also yields the same solution. Therefore, for this choice of divergence, the constraints are automatically satisfied.

SGAN divergence: The f-divergence of the SGAN is closely related to the Jensen-Shannon divergence, but for the $\ln(2)$ term (Goodfellow et al., 2014). The SGAN f-divergence formulation considers $g(D) = -\ln\left(1 + e^{-D}\right)$ and $f^c(T) = -\ln\left(1 - e^{T}\right)$. Applying Theorem 7 gives

$$D^*(\boldsymbol{x}) = \ln\left(\frac{p_d(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right), \quad \text{and} \quad p_g^*(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x})}{e^{(\lambda_p + \mu_p(\boldsymbol{x}))} - 1}. \tag{36}$$

The optimal output function corresponding to $D^*(\boldsymbol{x})$ is given by

$$T^*(\boldsymbol{x}) = g(D^*(\boldsymbol{x})) = \ln\left(\frac{p_d(\boldsymbol{x})}{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right) = \ln\left(D^*_{\text{SGAN}}(\boldsymbol{x})\right),$$

where D^*_{SGAN} is the optimal discriminator corresponding to the SGAN formulation proposed by Goodfellow et al. (2014). As in the KL and reverse KL divergence based f-GANs, the only feasible KKT multiplier associated with the non-negativity constraint is $\mu_p^*(\boldsymbol{x}) = 0,\ \forall\ \boldsymbol{x}$ such that $p_d(\boldsymbol{x}) > 0$, and the associated multiplier for the equality constraint is given by $\lambda_p = \ln(2)$. In summary, the optimal KKT multipliers are: $\mu_p^*(\boldsymbol{x}) = 0,\ \forall\ \boldsymbol{x} \in \mathcal{X}_+$, and $\lambda_p^* = \ln(2)$.

Pearson-χ^2 divergence: The Pearson-χ^2 divergence corresponds a special case of the LSGAN loss. The f-GAN formulation considers $g(D) = D$ and $f^c(T) = \frac{1}{4}T^2 + T$. The associated optimal discriminator and generator functions, derived following Theorem 7, are given by

$$D^*(\boldsymbol{x}) = 2\left(\frac{p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right) \quad \text{and} \quad p_g^*(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x})}{\sqrt{\lambda_p + \mu_p(\boldsymbol{x}) + 1}},$$

respectively. Applying the integral constraint Ω_{p_g}, we obtain $\lambda_p + \mu_p(\boldsymbol{x}) = 0$. In addition, enforcing the non-negativity and complementary slackness conditions yields $\mu_p(\boldsymbol{x}) = 0,\ \forall \boldsymbol{x}$, such that $p_d(\boldsymbol{x}) > 0$, as the only feasible solution. Similar to the Jensen-Shannon divergence case, we obtain the degenerate case of KKT multipliers: $\mu_p^*(\boldsymbol{x}) = 0,\ \forall \boldsymbol{x} \in \mathcal{X}_+$ and $\lambda_p^* = 0$, corresponding to the solution obtained in the unconstrained optimization problem.

Squared Hellinger divergence: The f-GAN associated with the squared Hellinger divergence has $g(D) = 1 - e^{-D}$ and $f^c(T) = \frac{T}{1-T}$. Applying Theorem 7, we obtain

$$D^*(\boldsymbol{x}) = \frac{1}{2}\ln\left(\frac{p_d(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right) \quad \text{and} \quad p_g^*(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x})}{(\lambda_p + \mu_p(\boldsymbol{x}) + 1)^2}, \tag{37}$$

respectively. Enforcing the integral constraints, we obtain the conditions:

$$\lambda_p + \mu_p(\boldsymbol{x}) = 0 \quad \text{or} \quad \lambda_p + \mu_p(\boldsymbol{x}) = -2.$$

The optimal KKT multiplier corresponding to the inequality constraint, as in the previous cases, is $\mu_p^*(\boldsymbol{x}) = 0,\ \forall\ \boldsymbol{x} \in \mathcal{X}_+$. The corresponding feasible set for λ_p is $\{-2,\, 0\}$. To find the optimal λ_p, we consider the dual optimization problem associated with only the integral constraint:

$$\begin{aligned}\lambda_p^* &= \arg \max_{\lambda_p \in \{-2,\, 0\}} \left\{ g(\lambda_p) = \inf_{p_g} \mathcal{L}_G(p_g, D^*) \right\}, \\ &= \arg \max_{\lambda_p \in \{-2, 0\}} \mathcal{L}_G(p_g^*, D^*).\end{aligned}$$

Substituting for $D^*(\boldsymbol{x})$ and $p_g^*(\boldsymbol{x})$ from Equation (37), considering $\mu_p^*(\boldsymbol{x}) = 0$, we obtain:

$$\lambda_p^* = \arg \max_{\lambda_p \in \{-2,\, 0\}} \left\{ -\frac{\lambda_p^2}{\lambda_p + 1} \right\}.$$

By inspection, $\lambda_p^* = -2$. Hence, the optimal KKT multipliers are $\mu_p^*(\boldsymbol{x}) = 0,\ \forall\ \boldsymbol{x} \in \mathcal{X}_+$ and $\lambda_p^* = -2$.

In addition to the f-divergence based GAN losses considered above, there exist two closely related variants: (i) The non-saturating SGAN (SGAN-NS) (Goodfellow et al., 2014) that alleviates the vanishing gradient problem in a practical GAN setting, and (ii) The general least-squares GAN (Mao et al., 2017) setting with class labels (a, b, c).

A.3 Euler-Lagrange Analysis of the Non-saturating SGAN

The non-saturating SGAN (SGAN-NS) proposed by Goodfellow et al. (2014) is a practical alternative to the SGAN loss, which ensures that the gradients do not vanish while training the generator and discriminator networks. However, it does not fit within the framework of divergence based GANs. Therefore, a straightforward divergence minimization argument does not apply, nevertheless a variational analysis can be carried out. Since the discriminator losses in both SGAN and SGAN-NS are the same, the optimal discriminator given by $D^*(\boldsymbol{x}) = \dfrac{p_d(\boldsymbol{x})}{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})}$ remains unchanged. Given the optimal discriminator, the Lagrangian of the generator loss, taking into account the integral constraint Ω_{p_g}, is given as follows:

$$\begin{aligned}\mathcal{L}_G^{\text{SGAN}-\text{NS}} &= -\,\mathbb{E}_{\boldsymbol{x} \sim p_g}[\ln(D^*(\boldsymbol{x}))] - \mathbb{E}_{x \sim p_d}[\ln\left(1 - \ln(D^*(\boldsymbol{x}))\right) + \lambda_p \left(\int_{\mathcal{X}} p_g(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} - 1 \right), \\ &= \int_{\mathcal{X}} \left(-p_g(\boldsymbol{x}) \ln(D^*(\boldsymbol{x})) - p_d(\boldsymbol{x}) \ln\left(1 - D^*(\boldsymbol{x})\right) + \lambda_p p_g(\boldsymbol{x}) \right)\ \mathrm{d}\boldsymbol{x} - \lambda_p.\end{aligned}$$

Applying the EL condition yields:

$$\ln(D^*(\boldsymbol{x})) + \frac{D^*(\boldsymbol{x})}{1 - D^*(\boldsymbol{x})}\bigg|_{p_g = p_g^*} = \lambda_p + 1, \tag{38}$$

$$\Rightarrow \left(\frac{p_d(\boldsymbol{x})}{p_d(\boldsymbol{x}) + p_g^*(\boldsymbol{x})} \right) \exp\left(\frac{p_d(\boldsymbol{x})}{p_g^*(\boldsymbol{x})} \right) = e^{\lambda_p + 1}. \tag{39}$$

The optimal generator distribution $p_g^*(\boldsymbol{x})$ is the one that solves the above transcendental equation. While no closed-form approaches exist for solving Equation (39), one could solve it approximately.

A second alternative suggested by Goodfellow et al. (2014) involves the removal of the expectation term associated with p_d in $\mathcal{L}_G^{\text{SGAN}-\text{NS}}$, since this term does not contribute toward the training of the generator network in practice. Incorporating this modification gives us the following Lagrangian of the generator loss:

$$\begin{aligned}\mathcal{L}_G^{\text{SGAN}-\text{NS}} &= -\,\mathbb{E}_{\boldsymbol{x}\sim p_g}[\ln(D^*(\boldsymbol{x}))] + \lambda_p\left(\int_{\mathcal{X}} p_g(\boldsymbol{x})\;\mathrm{d}\boldsymbol{x} - 1\right),\\ &= \int_{\mathcal{X}} p_g(\boldsymbol{x})\big(-\ln(D^*(\boldsymbol{x})) + \lambda_p\big)\;\mathrm{d}\boldsymbol{x} - \lambda_p.\end{aligned}$$

Applying the EL condition gives the following transcendental equation:

$$D^*(\boldsymbol{x})e^{D^*(\boldsymbol{x})}\big|_{p_g=p_g^*} = e^{\lambda_p+1}, \tag{40}$$

$$\Rightarrow \left(\frac{p_d(\boldsymbol{x})}{p_d(\boldsymbol{x})+p_g^*(\boldsymbol{x})}\right)\exp\left(\frac{p_d(\boldsymbol{x})}{p_d(\boldsymbol{x})+p_g^*(\boldsymbol{x})}\right) = e^{\lambda_p+1}. \tag{41}$$

Equation (41) can be solved through the principal branch of the Lambert-W function $W_0(\cdot)$ (Lambert, 1758; Corless et al., 1996; Bateman and Erdelyi, 1953), where the equation $ye^y = z$ for $y, z \in \mathbb{R}$ has a solution, if and only if $z \geq -1/e$. For $z \geq 0$, the solution is unique and is given by $y = W_0(z)$. Noting that the right-hand size of Equation (41) is always non-negative, we write:

$$\begin{aligned}\frac{p_d(\boldsymbol{x})}{p_d(\boldsymbol{x})+p_g^*(\boldsymbol{x})} &= W_0\left(e^{\lambda_p+1}\right)\\ \Rightarrow p_g^*(\boldsymbol{x}) &= \frac{W_0\left(e^{\lambda_p+1}\right)}{1-W_0\left(e^{\lambda_p+1}\right)}p_d(\boldsymbol{x}).\end{aligned} \tag{42}$$

The optimal generator $p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x})$ requires $W_0\left(e^{\lambda_p+1}\right) = 0.5$, which is achieved when $\lambda_p \approx -1.1935$. A similar link with the Lambert-W function was observed in the context analyzing SGAN-NS with infinite-width discriminators employing the neural tangent kernel (Franceschi et al., 2021)

In existing implementations, the integral constraint on p_g is not imposed explicitly. Instead, the generator network in trained to *learn* the inversion mapping implicitly. It was observed that, in practice, while SGAN-NS alleviates the problem of vanishing gradients, the training procedure is more sensitive to hyper-parameter tuning than the training of SGAN (Fedus et al., 2018). We attribute the sensitivity to the implicit inversion of the transcendental equations (38) or (40), in comparison with the linear mapping present in the saturating SGAN variant (cf. Equation (36)).

A.4 Euler-Lagrange Analysis of the Least-squares GAN

As a generalization to the Pearson-χ^2 divergence based f-GAN, we consider the least-squares GAN formulation presented by Mao et al. (2017). Consider the LSGAN discriminator and generator costs given by

$$\begin{aligned}\mathcal{L}_D^{\text{LSGAN}} &= \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_d}[(D(\boldsymbol{x})-b)^2] + \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_g}[(D(\boldsymbol{x})-a)^2],\ \text{and}\\ \mathcal{L}_G^{\text{LSGAN}} &= \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_d}[(D^*(\boldsymbol{x})-c)^2] + \frac{1}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_g}[(D^*(\boldsymbol{x})-c)^2] + \lambda_p\left(\int_{\mathcal{X}} p_g(\boldsymbol{x})\;\mathrm{d}\boldsymbol{x} - 1\right),\end{aligned}$$

respectively, where $D^*(\boldsymbol{x})$ is the minimizer of $\mathcal{L}_D^{\text{LSGAN}}$. The discriminator learns a regression model onto the target labels a and b for the generated and true data samples, respectively. The generator learns to output images that are classified by the discriminator with target label c. The generator loss is subjected to only the integral constraint, as the analysis in Appendix A.2 shows that the non-negativity constraint is met automatically. To optimize the discriminator loss, we consider $\mathcal{L}_D^{\text{LSGAN}}$ in its integral form:

$$\mathcal{L}_D^{\text{LSGAN}} = \frac{1}{2}\int_{\mathcal{X}} \left((D(\boldsymbol{x})-b)^2 p_d(\boldsymbol{x}) + (D(\boldsymbol{x})-a)^2 p_g(\boldsymbol{x})\right)\, \mathrm{d}\boldsymbol{x}.$$

Applying the EL conditions yields the optimal discriminator function given a generator distribution:

$$D^*(\boldsymbol{x}) = \frac{a\, p_g(\boldsymbol{x}) + b\, p_d(\boldsymbol{x})}{p_g(\boldsymbol{x}) + p_d(\boldsymbol{x})}.$$

Given the optimal discriminator, consider the generator loss:

$$\mathcal{L}_G^{\text{LSGAN}} = \frac{1}{2}\int_{\mathcal{X}} \left((D^*(\boldsymbol{x})-c)^2 \left(p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})\right) + \lambda_p p_g(\boldsymbol{x})\right)\, \mathrm{d}\boldsymbol{x} - \lambda_p.$$

Applying the EL condition gives

$$(D^*(\boldsymbol{x})-c)^2 + 2(D^*(\boldsymbol{x})-c)(p_g(\boldsymbol{x})+p_d(\boldsymbol{x}))\frac{\partial D^*(\boldsymbol{x})}{\partial p_g} + \lambda_p = 0.$$

Algebraic simplification results in the following quadratic in p_g:

$$p_g^2(\boldsymbol{x})\left((a-c)^2+\lambda_p\right) + 2p_g(\boldsymbol{x})p_d(\boldsymbol{x})\left((a-c)^2+\lambda_p\right) + p_d^2(\boldsymbol{x})\left((b-c)(2a-b-c)+\lambda_p\right) = 0.$$

Solving for the optimal generator distribution, we obtain:

$$p_g^*(\boldsymbol{x}) = -p_d(\boldsymbol{x}) \pm p_d(\boldsymbol{x})\sqrt{\frac{(a-b)^2}{(a-c)^2+\lambda_p}}.$$

Only the positive root yields a valid solution. Applying the integral constraint gives the optimal Lagrange multiplier

$$\lambda_p^* = \frac{(a-b)^2}{4} - (a-c)^2.$$

The choice of $b-c=1$ and $b-a=2$ proposed by Mao et al. (2017) gives $\lambda_p^* = -1$, and the optimal generator $p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x})$. Similarly, setting $b-a=2$ and $a-c=1$ yields $\lambda_p^* = 0$, and the corresponding GAN loss minimizes the Pearson-χ^2 divergence.

While in theory, infinitely many sets of (a,b,c) solve for the desired optimum $p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x})$, those labels that correspond to the Pearson-χ^2 loss are preferred as the corresponding p_g^* automatically satisfies the integral constraint without requiring an explicit penalty term.

Appendix B. Optimality of WGAN-GNP (1-D)

In this appendix, we present the proofs for the optimal WGAN-GNP Lagrange multiplier and generator function in the 1-D setting.

B.1 Optimal Lagrange Multiplier (1-D)

Consider the optimal discriminator function in 1-D given by

$$D^*(x) = \frac{1}{2\lambda_d}\left(\phi * (p_g - p_d)\right)(x) + a_1 x + a_0.$$

Its derivative is given by

$$\begin{aligned}\frac{\mathrm{d}D^*(x)}{\mathrm{d}x} &= \frac{1}{4\lambda_d}\int_{\mathcal{Y}} \frac{x-y}{|x-y|}\left(p_g(y) - p_d(y)\right)\ \mathrm{d}y + a_1 \\ &= \frac{1}{4\lambda_d}(\mathrm{sgn} * (p_g - p_d))(x) + a_1,\end{aligned}$$

where $\mathrm{sgn}(x) = \frac{x}{|x|}$ denotes the signum function. Enforcing the constraint that the gradient must have unit norm, we get the optimal Lagrange multiplier in 1-D:

$$\lambda_d^* = \frac{1}{4}\sqrt{\frac{1}{|\mathcal{X}|}\int_{\mathcal{X}} \left(\left(\frac{x}{|x|} * (p_g - p_d)\right)(x) + a_1\right)^2\ \mathrm{d}x}. \tag{43}$$

To analyze the second-order condition, consider the WGAN integral cost:

$$\mathcal{L}\left(D(x), D'(x)\right) = \int_{\mathcal{X}} \mathcal{F}\left(x, D(x), D'(x)\right)\ \mathrm{d}x.$$

The Legendre-Clebsch condition states that a minimizer must satisfy the second-order partial-differential condition

$$\mathcal{F}_{D'D'} = \frac{\partial^2 \mathcal{F}}{\partial D'^2} \geq 0,\ \forall\ x \in \mathcal{X}.$$

Consider the integrand $\mathcal{F}$:

$$\begin{aligned}&\mathcal{F}\left(x, D, D'\right) = D(x)(p_g(x) - p_d(x)) + \lambda_d\big(|D'(x)|^2 - 1\big), \\ &\quad \Rightarrow \mathcal{F}_{D'D'} = 2\lambda_d,\end{aligned}$$

which implies that a positive value for λ_d results in a minimizer.

In summary, considering the positive square-root in the expression for the optimal Lagrange multiplier in (43) results in a minimizer of $\mathcal{L}_D$.

We now analyze λ_d^* as a function of p_d and p_g. First, consider the convolution term inside the integral of Equation (43):

$$\begin{aligned}\left(\mathrm{sgn} * (p_g - p_d)\right)(x) &= \int_{-\infty}^{\infty} \left(p_g(y) - p_d(y)\right)\mathrm{sgn}(x-y)\mathrm{d}y \\ &= \int_{y=-\infty}^{x} \left(p_g(y) - p_d(y)\right)\mathrm{d}y - \int_{y=x}^{\infty} \left(p_g(y) - p_d(y)\right)\mathrm{d}y \\ &= F_{p_g}(x) - F_{p_d}(x) - \big(1 - F_{p_g}(x) - (1 - F_{p_d}(x))\big) \\ &= 2\left(F_{p_g}(x) - F_{p_d}(x)\right),\end{aligned}$$

where F_{p_g} and F_{p_d} denote the cumulative density functions (CDFs) of p_g and p_d, respectively. Substituting the above into Equation (43), we get:

$$\lambda_d^* = \frac{1}{4}\sqrt{\frac{1}{|\mathcal{X}|}\int_{\mathcal{X}} \left(2\left(F_{p_g}(x) - F_{p_d}(x)\right) + a_1\right)^2 \, \mathrm{d}x}.$$

In a practical setting, as the training of the GAN progresses, the Fourier coefficients of $p_g(x)$ converge to those of $p_d(x)$. As the distributions become closer to one another in the L_2 sense, their difference, and therefore the difference between their CDFs also reduces. Therefore, upon convergence of the GAN, ideally, we have $\alpha_m = \beta_m \; \forall \; m \in \mathbb{Z}$, and consequently $p_g^*(x) = p_d(x)$, yielding

$$\lambda_d^*\Big|_{p_g^*=p_d} = \frac{1}{4}\sqrt{\frac{1}{|\mathcal{X}|}\int_{\mathcal{X}} a_1^2 \, \mathrm{d}x} = \frac{|a_1|}{4}.$$

As observed in Theorem 2, the convergence of p_g to p_d is independent of the choice of a_1, which is part of the homogeneous solution. Therefore, without loss of generality, we set $a_1 = 0$, which gives us the favorable property that, $\lambda_d^* \to 0$ as $p_g(x) \to p_d(x)$, which can be used as a proxy for tracking the convergence of the GAN training.

B.2 Optimal WGAN-GNP Generator (1-D)

Consider the Lagrangian of the generator cost in 1-D:

$$\mathcal{L}_G = \int_{\mathcal{X}} \left((p_d(x) - p_g(x))D^*(x) + (\lambda_p + \mu_p(x))p_g(x)\right) \, \mathrm{d}x - \lambda_p,$$

with $D^*(x)$ given by Equation (19):

$$D^*(x) = \frac{1}{2\lambda_d^*}\int_{\mathcal{X}} \phi(x-y)\left(p_d(y) - p_g(y)\right)\mathrm{d}y + a_1\, x + a_0,$$

where $\mathcal{X}$ is the convex hull containing the supports of p_d and p_g. Without loss of generality, we consider the symmetric fundamental solution $\phi(x) = \frac{1}{2}|x| + b_0$. Since the integrand involves a convolution integral, it is not in the standard form considered in Equation (1). Hence, the EL conditions cannot be applied directly. Starting from first principles, we evaluate the first variation of $\mathcal{L}_G$ and set it to zero to obtain the optimizer. Consider an ϵ-perturbation of the loss $\mathcal{L}_G$ about the optimal generator p_g^*, denoted by $\mathcal{L}_{G,\epsilon}(p_g) = \mathcal{L}_G(p_g^*(x) + \epsilon\eta(x))$, where $\eta(x)$ is a family of compactly supported, absolutely integrable, infinitely differentiable functions that are identically zero at the boundaries of $\mathcal{X}$. The corresponding perturbed discriminator is represented by $D_\epsilon^*(x)$. Consider

$$\mathcal{L}_{G,\epsilon}(p_g) = \int_{\mathcal{X}} \Big(D_\epsilon^*(x)(p_d(x)) - p_g^*(x) - \epsilon\eta(x)) + (\lambda_p + \mu_p(x))(p_g^*(x) + \epsilon\eta(x))\Big) \, \mathrm{d}x,$$

where

$$\begin{aligned} D_\epsilon^*(x) &= \int_{\mathcal{X}} \phi(x-y)\left(\frac{p_g^*(y) + \epsilon\eta(y) - p_d(y)}{2\lambda_d^*}\right)\mathrm{d}y + a_1\, x + a_0 \\ &= \left(\phi * (p_g^* + \epsilon\eta - p_d)\right)(x) + a_1\, x + a_0. \end{aligned}$$

Consider the following derivatives with respect to ϵ:

$$\frac{\mathrm{d}\mathcal{L}_{G,\epsilon}}{\mathrm{d}\epsilon} = \int_{\mathcal{X}} \left(\frac{\mathrm{d}D^*_\epsilon(x)}{\mathrm{d}\epsilon}(p_d(x)) - p^*_g(x) - \epsilon\eta(x)) - D^*_\epsilon\eta(x) + (\lambda_p + \mu_p(x))\eta(x) \right) \mathrm{d}x, \text{ and}$$
$$\frac{\mathrm{d}D^*_\epsilon(x)}{\mathrm{d}\epsilon} = \frac{1}{2\lambda^*_d} \int_{\mathcal{X}} \phi(x-y)\eta(y)\ \mathrm{d}y.$$

Substituting for $D^*_\epsilon(x)$ and its derivative into $\frac{\mathrm{d}\mathcal{L}_{G,\epsilon}}{\mathrm{d}\epsilon}$, and evaluating it at $\epsilon = 0$ gives the first variation in $\mathcal{L}_G$:

$$\begin{aligned}
\partial\mathcal{L}_G &= \int_{\mathcal{X}}\int_{\mathcal{X}} \phi(x-y)\eta(y)(p^*_g(x) - p_d(x))\ \mathrm{d}y\ \mathrm{d}x \\
&\quad + \int_{\mathcal{X}} \left(2\alpha_d(\lambda_p + \mu_p(x)) + a_1 x - \left(\phi * (p_d - p^*_g)\right)(x) \right)\eta(x)\ \mathrm{d}x, \\
&= T_1 + T_2,
\end{aligned}$$

where $\alpha_d = \dfrac{\lambda^*_d}{4}$. Next, consider

$$T_1 = \int_{\mathcal{X}}\int_{\mathcal{X}} \phi(x-y)\eta(y)(p^*_g(x) - p_d(x))\ \mathrm{d}y\ \mathrm{d}x.$$

Considering a compact domain of integration and the functions ϕ, p^*_g, and p_d to be absolutely integrable over the domain, we apply Fubini's theorem and interchange the order of integration to obtain

$$T_1 = \int_{\mathcal{X}} \eta(y) \int_{\mathcal{X}} \left(p^*_g(x) - p_d(x)\right)\phi(x-y)\ \mathrm{d}x\ \mathrm{d}y.$$

Since ϕ is symmetric, T_1 simplifies to

$$T_1 = \int_{\mathcal{X}} \eta(y)(\phi * (p^*_g - p_d))(y)\ \mathrm{d}y.$$

Substituting for T_1 in the first variation yields

$$\partial\mathcal{L}_G = \int_{\mathcal{X}} \left(2\alpha_d(\lambda_p + \mu_p(x)) + a_1 x - 2\left(\phi * (p_d - p^*_g)\right)(x)\right)\ \eta(x)\ \mathrm{d}x.$$

Setting $\partial\mathcal{L}_G$ to zero and invoking the fundamental lemma of calculus of variations (cf. Section 2) gives rise to the following condition

$$\left(\phi * (p_d - p^*_g)\right)(x) = \alpha_d(\lambda_p + \mu_p(x)) + \frac{1}{2}a_1 x, \tag{44}$$

which the optimal generator p^*_g must satisfy. Rearranging (44) yields

$$\alpha_d\mu_p(x) = \left(\phi * (p_d - p^*_g)\right)(x) - \alpha_d\lambda_p - \frac{1}{2}a_1 x,$$

from which we obtain the Laplacian:

$$\alpha_d \mu_p''(x) = p_d(x) - p_g^*(x),$$

which gives the optimal generator distribution in terms of the KKT multiplier $\mu_p(x)$:

$$p_g^*(x) = p_d(x) - \alpha_d \mu_p''(x). \tag{45}$$

An alternative approach using Fourier analysis: The Fourier transform proves to be an efficient tool to arrive at Equation (45) starting from Equation (44). Recall $\phi(x) = \frac{1}{2}|x|$. Its Fourier transform cannot be defined in the L_1 or L_2 sense, but only in the sense of distributions: $\mathscr{F}\{\phi(x)\} = -\frac{1}{\omega^2}$, $\omega \in \mathbb{R} - \{0\}$. The Fourier transform of x is also given in the distributional sense as $\mathscr{F}\{x\} = -\mathrm{j}\delta'(\omega)$, where $\delta'(\omega)$ is the derivative of the Dirac delta, and must be understood in a distributional sense by its action on a test function $f \in \mathcal{C}^\infty(\mathbb{R})$: $\langle \delta'(\omega), f(\omega) \rangle = -f'(0)$. Writing (44) in the Fourier domain gives

$$\begin{aligned}\frac{1}{\omega^2}\left(\hat{p}_g^*(\omega) - \hat{p}_d(\omega)\right) &= \alpha_d \lambda_p \delta(\omega) + \alpha_d \hat{\mu}_p(\omega) - \mathrm{j}\frac{a_1}{2}\delta'(\omega) \\ \Rightarrow \hat{p}_g^*(\omega) &= \hat{p}_d(\omega) + \alpha_d\left(\lambda_p \omega^2 \delta(\omega) + \omega^2 \hat{\mu}_p(\omega)\right) - \mathrm{j}\frac{a_1}{2}\omega^2\delta'(\omega),\end{aligned}$$

where $\hat{p}_g^*(\omega)$ and $\hat{p}_d(\omega)$ are the Fourier transforms of $p_g^*(x)$ and $p_d(x)$, respectively. From the properties of the Dirac delta, it follows that $\omega^2\delta(\omega) = 0$ and $\omega^2\delta'(\omega) = 0$. Hence, we obtain

$$\hat{p}_g^*(\omega) = \hat{p}_d(\omega) + \alpha_d \omega^2 \hat{\mu}_p(\omega). \tag{46}$$

Invoking the differentiation property of the Fourier transform, we have

$$\mu_p''(x) \overset{\mathscr{F}}{\longleftrightarrow} -\omega^2 \hat{\mu}_p(\omega).$$

Taking the inverse Fourier transform on both sides of Equation (46) gives

$$p_g^*(x) = p_d(x) - \alpha_d \mu_p''(x),$$

which is identical to Equation (45).

We observe that p_g^* is independent of λ_p, while the optimal $\mu_p^*(x)$ must be determined, which can be done by enforcing the integral constraint:

$$\begin{aligned}\int_{\mathcal{X}} p_g^*(x)\ \mathrm{d}x = \int_{\mathcal{X}} \left(p_d(x) + \alpha_d \mu_p''(x)\right)\ \mathrm{d}x &= 1 \\ \Rightarrow \int_{\mathcal{X}} \mu_p''(x)\ \mathrm{d}x &= 0.\end{aligned} \tag{47}$$

Recall from Theorem 2 that $\mu_p(x) \leq 0,\ \forall\ x \in \mathcal{X}$, in order to satisfy the non-negativity constraint. Let us now split $\mathcal{X}$ into two disjoint sets:

$$\mathcal{X}_0 = \{x \mid p_d(x) = 0\}, \quad \text{and} \quad \mathcal{X}_+ = \{x \mid p_d(x) > 0\}.$$

Consider the complementary slackness condition:

$$\mu_p(x) p_g^*(x) = \mu_p(x) p_d(x) - \alpha_d \mu_p(x) \mu_p''(x) = 0,\ \forall\ x \in \mathcal{X}.$$

For $x \in \mathcal{X}_+$, we have either $\mu_p(x) = 0$ or $\mu_p''(x) = \dfrac{p_d}{\alpha_d}$ as feasible solutions, but in view of the condition in Equation (47), the latter becomes invalid. Therefore, for $x \in \mathcal{X}_+$, $\mu_p(x) = 0$ is the only solution. For $x \in \mathcal{X}_0$, the complementary slackness condition requires that $\mu_p(x)\mu_p''(x) = 0$. Consequently, either $\mu_p(x) = 0$ or $\mu_p(x) = c_1\, x + c_0$. The former solution implies that $\mu_p(x) = 0,\ \forall\ x \in \mathcal{X}$. If we consider the latter and enforce that $\mu_p(x) \leq 0,\ \forall\ x \in \mathcal{X}$, then c_1 is necessarily zero, while c_0 is a finite negative value.

Consolidating, we get the following optimal solutions for the Lagrangian parameter:

$$\mu_p^*(x) = 0,\ \forall\ x \in \mathcal{X}, \quad \text{or} \quad \mu_p^*(x) = \begin{cases} 0, & \forall\ x \in \mathcal{X}_+, \text{ and} \\ -\infty < c_0 \leq 0, & \forall\ x \in \mathcal{X}_0. \end{cases} \tag{48}$$

Both are equally good optima as far as satisfying the constraints $\mu_p(x) \leq 0$ and Equation (47) go, because in both cases, the loss function $\mathcal{L}_G$ evaluates to zero, rendering the specific choice of c_0 irrelevant. Without loss of optimality, we set $\mu_p^*(x) = 0,\ \forall\ x \in \mathcal{X}$.

The preceding analysis strongly suggests that $p_g^*(x) = p_d(x),\ \forall\ x \in \mathcal{X}$ is the optimal solution. From Equation (45), we observe that the optimality of p_g^* does not depend on the value of λ_p. For the loss $\mathcal{L}_G$ to be finite, λ_p must, however, be a finite real number.

Further, the optimality of the generator is unaffected by the homogeneous component of the optimal discriminator.

In summary, the optima are given as follows:

$$p_g^*(x) = p_d(x), \quad \mu_p^*(x) = 0,\ \forall x \in \mathcal{X}, \quad \text{and} \quad \lambda_p \in (-\infty, \infty).$$

B.3 Optimal Lagrange Multiplier in WGAN-FS (1-D)

Consider the truncated Fourier-series approximation for the discriminator

$$D_{FS}^*(x) = \frac{1}{\lambda_{FS}^*}\left(a_1 x + a_0 + \sum_{m=1}^{M}\left(\gamma_m^r \cos(\omega_o m x) + \gamma_m^i \sin(\omega_o m x)\right)\right).$$

In order to enforce the gradient-norm penalty, we require the square of the derivative:

$$\left(\frac{\mathrm{d}D_{FS}^*}{\mathrm{d}x}\right)^2 = \frac{1}{\lambda_{FS}^{*2}}\left(a_1 - \sum_{m=1}^{M}\left(\gamma_m^r \omega_o m \sin(\omega_o m x) + \gamma_m^i \omega_o m \cos(\omega_o m x)\right)\right)^2.$$

By Cauchy-Schwartz inequality, we have $\left(\sum_{i=1}^{n} u_i.1\right)^2 \leq n\sum_{i=1}^{n} u_i^2$. Using this inequality allows us to place the following bound:

$$\begin{aligned}\left(\frac{\mathrm{d}D_{FS}^*}{\mathrm{d}x}\right)^2 &\leq \frac{2M+1}{\lambda_{FS}^{*2}}\left(a_1^2 + \sum_{m=1}^{M}\omega_o^2 m^2\left(\gamma_m^{r^2}\sin^2(\omega_o m x) + \gamma_m^{i^2}\cos^2(\omega_o m x)\right)\right) \\ &= \frac{2M+1}{\lambda_{FS}^{*2}}\left(a_1^2 + \sum_{m=1}^{M}\left(\left(\tau_m^i + \tau_m^r\right) + \left(\tau_m^i - \tau_m^r\right)\cos(2\omega_o m x)\right)\right),\end{aligned}$$

where

$$\tau_m^r = \frac{1}{2}(\gamma_m^r)^2\omega_o^2 m^2, \quad \text{and} \quad \tau_m^i = \frac{1}{2}(\gamma_m^i)^2\omega_o^2 m^2.$$

Enforcing the gradient-norm penalty gives

$$0 \leq \int_{\mathcal{X}} \left((2M+1)\left(a_1^2 + \sum_{m=1}^{M} \left((\tau_m^i + \tau_m^r) + (\tau_m^i - \tau_m^r)\cos(2\omega_o m x) \right) \right) - \lambda_{FS}^{*2} \right) \, \mathrm{d}x.$$

Simplifying, we obtain:

$$\begin{aligned}\lambda_{FS}^{*2} &\leq \frac{(2M+1)}{|\mathcal{X}|} \int_{\mathcal{X}} \left(a_1^2 + \sum_{m=1}^{M} \left((\tau_m^i + \tau_m^r) + (\tau_m^i - \tau_m^r)\cos(2\omega_o m x) \right) \right) \, \mathrm{d}x. \\ &= (2M+1)\left(a_1^2 + \sum_{m=1}^{M} (\tau_m^i + \tau_m^r) \right) + \frac{(2M+1)}{|\mathcal{X}|} \sum_{m=1}^{M} (\tau_m^i - \tau_m^r) \left(\int_{\mathcal{X}} \cos(2\omega_o m x) \, \mathrm{d}x \right).\end{aligned}$$

In practice, we have data, and computing a sample estimate of the upper bound over a batch of size N gives the following

$$\lambda_{FS}^{*} \leq \sqrt{(2M+1)\left(a_1^2 + \sum_{m=1}^{M} (\tau_m^i + \tau_m^r) + \frac{1}{N}\sum_{k=1}^{N}\sum_{m=1}^{M} (\tau_m^i - \tau_m^r)\cos(2\omega_o m x_k) \right)}.$$

Recall that $a_1 = \pm 1$ for the optimal discriminator to satisfy the gradient-norm penalty when $p_g^* = p_d$ (cf. Section 3.2). We observed that, in practice, the contribution of a_1 is negligible and the upper bound could be used for λ_{FS}^*.

Appendix C. Optimality of WGAN-GNP (n-D)

In this appendix, we present the multivariate counterparts of the 1-D proofs presented in Appendix B.

C.1 Optimal Lagrange Multiplier (n-D)

Consider the optimal WGAN discriminator in n-D ($n \geq 3$):

$$D^*(\boldsymbol{x}) = \frac{\kappa_n}{2\lambda_d} \int_{\mathcal{X}} \frac{1}{\|\boldsymbol{x} - \boldsymbol{y}\|^{n-2}} \left(p_g(\boldsymbol{y}) - p_d(\boldsymbol{y}) \right) \, \mathrm{d}\boldsymbol{y} + \langle \boldsymbol{a}, \boldsymbol{x} \rangle + \text{constant},$$

where $\boldsymbol{a} = [a_1, a_2, \ldots, a_n]^{\mathrm{T}}$, $\kappa_n = (n(n-2)\mathfrak{v}(n))^{-1}$, and $\mathcal{X}$ is the convex hull of the supports of p_d and p_g, where in turn, $\mathfrak{v}(n)$ is the volume of the unit sphere in $\mathbb{R}^n$ given by $\mathfrak{v}(n) = \pi^{\frac{n}{2}} \left(\Gamma\left(\frac{n}{2}+1\right) \right)^{-1}$. Consider the partial derivative with respect to x_i, the i^{th} element of $\boldsymbol{x}$:

$$\frac{\partial D^*}{\partial x_i} = \frac{1}{2n\mathfrak{v}(n)\lambda_d} \int_{\mathcal{X}} \frac{x_i - y_i}{\|\boldsymbol{x} - \boldsymbol{y}\|^n} \left(p_g(\boldsymbol{y}) - p_d(\boldsymbol{y}) \right) \, \mathrm{d}\boldsymbol{y} + a_i.$$

Squaring and summing over all i gives

$$\|\nabla D^*\|_2^2 = \frac{1}{4n^2\mathfrak{v}(n)^2\lambda_d^2}\sum_{i=1}^{n}\left(\int_{\mathcal{X}}\frac{x_i-y_i}{\|\boldsymbol{x}-\boldsymbol{y}\|^n}\left(p_g(\boldsymbol{y})-p_d(\boldsymbol{y})\right)\,\mathrm{d}\boldsymbol{y}+a_i\right)^2.$$

Enforcing the gradient-norm penalty: $\int_{\mathcal{X}}\left(\|\nabla D^*\|_2^2-1\right)\,\mathrm{d}\boldsymbol{x}=0$ gives us the following condition on the optimal Lagrange multiplier λ_d^* for a given optimal discriminator:

$$\lambda_d^* = \pm\sqrt{\frac{1}{4n^2\mathfrak{v}(n)^2|\mathcal{X}|}\int_{\mathcal{X}}\left(\sum_{i=1}^{n}\left(\int_{\mathcal{X}}\frac{x_i-y_i}{\|\boldsymbol{x}-\boldsymbol{y}\|^n}\left(p_g(\boldsymbol{y})-p_d(\boldsymbol{y})\right)\,\mathrm{d}\boldsymbol{y}+a_i\right)^2\right)\,\mathrm{d}\boldsymbol{x}}, \tag{49}$$

where $|\mathcal{X}|=\int_{\mathcal{X}}1\,\mathrm{d}\boldsymbol{x}$ denotes the volume of $\mathcal{X}$.

We next have to determine the appropriate sign of λ_d^*, which is obtained by considering the second-order necessary conditions for optimality. Consider the n-D cost:

$$\mathcal{L}_D\left(D(\boldsymbol{x}),\{D_i'(\boldsymbol{x})\}_{i=1}^n\right)=\int_{\mathcal{X}}\mathcal{F}\left(\boldsymbol{x},D(\boldsymbol{x}),\{D_i'(\boldsymbol{x})\}_{i=1}^n\right)\,\mathrm{d}\boldsymbol{x},$$

where $D_i'(x)=\frac{\partial D}{\partial x_i}$. Recall that the Legendre-Clebsch condition (cf. Section 2) in the multidimensional case translates to positive-definiteness of the Hessian matrix $\mathbb{H}$ of the Hamiltonian $\mathcal{H}$ of the cost $\mathcal{L}_D$ computed with respect to $\{D_i'\}_{i=1}^n$, evaluated at $\lambda_d=\lambda_d^*, D(\boldsymbol{x})=D^*(\boldsymbol{x})$:

$$\mathbb{H}_{D,\mathcal{H}}\Big|_{\lambda_d=\lambda_d^*;\,D(\boldsymbol{x})=D^*(\boldsymbol{x})}\succ 0,$$

where $\succ$ denotes positive-definiteness. The entries of $\mathbb{H}_{D,\mathcal{H}}$ are given by

$$[\mathbb{H}_{D,\mathcal{H}}]_{i,j}=\frac{\partial^2\mathcal{H}}{\partial D_i'\partial D_j'},\quad\text{where the Hamiltonian}\quad\mathcal{H}=\sum_{i=1}^{n}\left(D_i'\frac{\partial\mathcal{F}}{\partial D_i'}\right)-\mathcal{F}.$$

Considering the integrand $\mathcal{F}$ of the WGAN-GNP cost given in Equation (13), the Hamiltonian turns out to be

$$\mathcal{H}=\lambda_d\left(\sum_{i=1}^{n}\left(D_i'(\boldsymbol{x})\right)^2\right)-D(\boldsymbol{x})(p_g(\boldsymbol{x})-p_d(\boldsymbol{x})).$$

Evaluating the Hessian with respect to D_i' yields the following:

$$[\mathbb{H}_{D',\mathcal{H}}]_{i,j}=\begin{cases}2\lambda_d, & \text{for } i=j,\text{ and}\\ 0, & \text{for } i\neq j.\end{cases}$$

$$\Rightarrow\mathbb{H}_{D',\mathcal{H}}\Big|_{\lambda_d=\lambda_d^*;\,D(\boldsymbol{x})=D^*(\boldsymbol{x})}=2\,\lambda_d^*\,\mathbb{I}_n,$$

where $\mathbb{I}_n$ is the $n\times n$ identity matrix. This condition is analogous to the 1-D case, where, picking the positive square-root for λ_d^* in Equation (49) results in $D^*(\boldsymbol{x})$ being a minimizer of the chosen cost.

C.2 Optimal WGAN-GNP Generator (n-D)

The derivation of the optimal generator p_g^* proceeds along the lines of the first-variation analysis in 1-D, taking into account the fact that the generator cost does not involve terms containing the derivatives of p_g. Consider the Lagrangian

$$\mathcal{L}_G = \int_{\mathcal{X}} \left((p_d(\boldsymbol{x}) - p_g(\boldsymbol{x}))D^*(\boldsymbol{x}) + (\lambda_p + \mu_p(\boldsymbol{x}))p_g(\boldsymbol{x})\right)\, \mathrm{d}\boldsymbol{x} - \lambda_p,$$

where $D^*(\boldsymbol{x})$ is as given in Equation (19):

$$D^*(\boldsymbol{x}) = \frac{1}{\lambda_d}\int_{\mathcal{X}} \phi(\boldsymbol{x}-\boldsymbol{y})\left(p_d(\boldsymbol{y}) - p_g(\boldsymbol{y})\right)\, \mathrm{d}\boldsymbol{y} + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \text{ constant},$$

with $\phi(\boldsymbol{x}) = \kappa_n \|\boldsymbol{x}\|^{2-n}$, where $\boldsymbol{x} \in \mathbb{R}^n,\ n \geq 3$, $\kappa_n = \frac{1}{n(n-2)\mathfrak{v}(n)}$, $\mathfrak{v}(n)$ is the volume of the unit sphere in $\mathbb{R}^n$, and $\mathcal{X}$ is the convex hull of the supports of p_d and p_g. Denote the optimal generator as $p_g^*(\boldsymbol{x})$. Consider the perturbation $p_g^*(\boldsymbol{x}) + \epsilon\,\eta(\boldsymbol{x})$, where $\eta(\boldsymbol{x})$ is the n-dimensional counterpart of $\eta(x)$ defined in Appendix B.2. The first variation $\partial\mathcal{L}_G$ is given by

$$\begin{aligned}\partial\mathcal{L}_G &= \int_{\mathcal{X}}\int_{\mathcal{X}} \phi(\boldsymbol{y})\eta(\boldsymbol{x}-\boldsymbol{y})(p_g^*(\boldsymbol{x}) - p_d(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{x} \\ &\quad + \int_{\mathcal{X}} \Big(2\alpha_d\left(\lambda_p + \mu_p(\boldsymbol{x})\right) + \langle \boldsymbol{a}, \boldsymbol{x}\rangle - \left(\phi * \left(p_d - p_g^*\right)\right)(\boldsymbol{x})\Big)\eta(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &= T_1 + T_2,\end{aligned}$$

where $\alpha_d = \dfrac{\lambda_d^*}{4}$. The term T_1 involves a convolution with a singular kernel $\phi(\boldsymbol{y})$, with the singularity at the origin. The integrals therefore have to be evaluated in the Cauchy principal-value sense. We make the interpretation explicit by defining:

$$\text{p.v.}\int_{\mathcal{X}} (\ \cdot\ )\, \mathrm{d}\boldsymbol{x} = \lim_{\xi\to 0}\int_{\mathcal{X}^\xi} (\ \cdot\ )\, \mathrm{d}\boldsymbol{x},$$

where $\mathcal{X}^\xi = \mathcal{X} - \mathcal{B}(0,\xi)$, which is formed by removing a ball of radius ξ centered at the origin. Recall that $\mathcal{X}$ is assumed to be compactly supported, and hence $\mathcal{X}^\xi$ is compactly supported as well. Consider η to be absolutely integrable over $\mathcal{X}^\xi$. Applying Fubini's theorem to T_1 yields

$$\begin{aligned}T_1 &= \lim_{\xi\to 0}\int_{\mathcal{X}^\xi} \phi(\boldsymbol{y}) \int_{\mathcal{X}^\xi} \left(p_g^*(\boldsymbol{x}) - p_d(\boldsymbol{x})\right)\eta(\boldsymbol{x}-\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}, \\ &= \lim_{\xi\to 0}\int_{\mathcal{X}^\xi}\int_{\mathcal{X}^\xi} \phi(\boldsymbol{y})\left(p_g^*(\boldsymbol{x}+\boldsymbol{y}) - p_d(\boldsymbol{x}+\boldsymbol{y})\right)\eta(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}.\end{aligned}$$

Swapping the order of integration yields

$$T_1 = \lim_{\xi\to 0}\int_{\mathcal{X}^\xi} \eta(\boldsymbol{x})\left(\phi * \left(p_g^* - p_d\right)(\boldsymbol{x})\right)\, \mathrm{d}\boldsymbol{x},$$

since ϕ is radially symmetric. Substituting T_1 back into $\partial\mathcal{L}_G$, setting it to zero, and invoking the fundamental lemma of calculus of variations (cf. Section 2), we obtain the condition

$$\left(\phi * (p_g^* - p_d)\right)(\boldsymbol{x}) = \alpha_d(\lambda_p + \mu_p(\boldsymbol{x})) + \frac{1}{2}\langle \boldsymbol{a}, \boldsymbol{x}\rangle, \tag{50}$$

which the optimal generator p_g^* must satisfy. Applying the Laplacian operator Δ on both sides of Equation (50) and noting that $-\Delta\phi(\boldsymbol{x}) = \delta(\boldsymbol{x})$ from Theorem 3, we get

$$p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x}) + \alpha_d \Delta\mu_p(\boldsymbol{x}). \tag{51}$$

An alternative approach to obtaining Equation (51) from (50) involves the use of the Fourier transform of distributions. The Fourier transform of $\langle \boldsymbol{a}, \boldsymbol{x}\rangle$ can be defined in the distributional sense as follows:

$$\langle \boldsymbol{a}, \boldsymbol{x}\rangle = \sum_{i=1}^{n} a_i x_i \overset{\mathscr{F}}{\longleftrightarrow} \sum_{i=1}^{n} \mathrm{j}\ a_i \underbrace{\left(\delta'(\omega_i)\prod_{k\neq i}\delta(\omega_k)\right)}_{\delta_i'(\boldsymbol{\omega})},$$

where $\delta'(\omega_i)$ denotes the derivative of the Dirac delta considered along ω_i, which is the i^{th} element of $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_n]^{\mathrm{T}}$. Similarly, consider the n-dimensional radially symmetric function:

$$f_\tau(\boldsymbol{x}) = 2^{-\frac{\tau}{2}} \frac{\|\boldsymbol{x}\|^\tau}{\Gamma\left(\frac{n+\tau}{2}\right)},$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $\Gamma(\cdot)$ denotes the gamma function. Gelfand and Shilov (1958) showed that the Fourier transform of $f_\tau(\boldsymbol{x})$ is also radially symmetric, and has an expression given again in terms of f_τ as follows:

$$\mathscr{F}\{f_\tau(\boldsymbol{x})\} = (2\pi)^{\frac{n}{2}} f_{-n-\tau}(\boldsymbol{\omega}). \tag{52}$$

Multiplying Equation (50) by $2^{\frac{n-2}{2}}$ gives

$$\left(f_{2-n}(\boldsymbol{x}) * (p_d - p_g^*)\right)(\boldsymbol{x}) = \frac{2^{\frac{n-2}{2}}\alpha_d}{\kappa_n}\left(\lambda_p + \mu_p(\boldsymbol{x})\right) + \left(\frac{2^{\frac{n-4}{2}}}{\kappa_n}\right)\langle \boldsymbol{a}, \boldsymbol{x}\rangle.$$

Taking the Fourier transform on both sides, we get

$$f_{-2}(\boldsymbol{\omega})\left(\hat{p}_d(\boldsymbol{\omega}) - \hat{p}_g^*(\boldsymbol{\omega})\right) = \left(\frac{2^{\frac{n-2}{2}}\alpha_d}{\kappa_n}\right)\left(\lambda_p\delta(\boldsymbol{\omega}) + \hat{\mu}_p(\boldsymbol{\omega})\right) + \left(\frac{2^{\frac{n-4}{2}}}{\kappa_n}\right)\sum_{i=1}^{n}\left(\mathrm{j}\ a_i\delta_i'(\boldsymbol{\omega})\right), \tag{53}$$

where $\hat{p}_g^*(\boldsymbol{\omega})$, $\hat{p}_d(\boldsymbol{\omega})$, and $\hat{\mu}_p(\boldsymbol{\omega})$ are the n-dimensional Fourier transforms of $p_g^*(\boldsymbol{x})$, $p_d(\boldsymbol{x})$, and $\mu_p(\boldsymbol{x})$, respectively. Rearranging the terms, we get

$$\begin{aligned}\hat{p}_g^*(\boldsymbol{\omega}) = \hat{p}_d(\boldsymbol{\omega}) &- \left(\frac{2^{\frac{n-4}{2}}\Gamma\left(\frac{n-2}{2}\right)\alpha_d}{\kappa_n}\right)\left(\lambda_p\delta(\boldsymbol{\omega}) + \hat{\mu}_p(\boldsymbol{\omega})\right)\|\boldsymbol{\omega}\|^2 \\ &+ \left(\frac{2^{\frac{n-6}{2}}\Gamma\left(\frac{n-2}{2}\right)}{\kappa_n}\right)\sum_{i=1}^{n}\left(\mathrm{j}\ a_i\delta_i'(\boldsymbol{\omega})\right)\|\boldsymbol{\omega}\|^2.\end{aligned}$$

It can be verified that $\|\boldsymbol{\omega}\|^2\delta(\boldsymbol{\omega}) = 0$, and $\|\boldsymbol{\omega}\|^2\delta_i'(\boldsymbol{\omega}) = 0$. Consequently, the above equation simplifies to

$$\hat{p}_g^*(\boldsymbol{\omega}) = \hat{p}_d(\boldsymbol{\omega}) + \left(\frac{2^{\frac{n-4}{2}}\Gamma\left(\frac{n-2}{2}\right)\alpha_d}{\kappa_n}\right)(-\|\boldsymbol{\omega}\|_2^2)\hat{\mu}_p(\boldsymbol{\omega}). \tag{54}$$

Invoking the derivative properties of the n-dimensional Fourier transform, we have

$$\frac{\partial^2\mu_p(\boldsymbol{x})}{\partial x_i^2} \overset{\mathscr{F}}{\longleftrightarrow} -\omega_i^2\,\hat{\mu}_p(\boldsymbol{\omega}),$$

$$\Rightarrow \Delta\mu_p(\boldsymbol{x}) = \sum_{i=1}^{n}\frac{\partial^2\mu_p(\boldsymbol{x})}{\partial x_i^2} \overset{\mathscr{F}}{\longleftrightarrow} -\|\boldsymbol{\omega}\|_2^2\,\hat{\mu}_p(\boldsymbol{\omega}).$$

Taking the inverse transform on both sides of (54) gives the optimal generator

$$p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x}) + \left(\frac{2^{\frac{n-4}{2}}\Gamma\left(\frac{n-2}{2}\right)\alpha_d}{(2\pi)^{\frac{n}{2}}\kappa_n}\right)\Delta\mu_p(\boldsymbol{x}).$$

Recall that $\kappa_n = \frac{1}{n(n-2)\mathfrak{v}(n)}$, where $\mathfrak{v}(n) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}$, which is the volume of the unit hypersphere in $\mathbb{R}^n$. Substituting into the above yields

$$p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x}) + \alpha_d\Delta\mu_p(\boldsymbol{x}),$$

which is in agreement with the solution obtained in (51).

The next step would be to determine the optimal KKT multipliers λ_p and $\mu_p(\boldsymbol{x})$. The analysis follows analogously to the 1-D case, by replacing the second derivative operator with the Laplacian operator. Consider splitting $\mathcal{X}$ into disjoint sets $\mathcal{X}_+ = \{\boldsymbol{x} \mid p_d(\boldsymbol{x}) > 0\}$ and $\mathcal{X}_0 = \{\boldsymbol{x} \mid p_d(\boldsymbol{x}) = 0\}$. Enforcing the integral, non-negativity, and complementary slackness constraints on $p_g^*(\boldsymbol{x})$ yields the following choices for $\mu_p^*(\boldsymbol{x})$:

$$\mu_p^*(\boldsymbol{x}) = 0,\ \forall\ \boldsymbol{x} \in \mathcal{X}, \quad \text{or} \quad \mu_p^*(\boldsymbol{x}) = \begin{cases} 0, & \forall\ \boldsymbol{x} \in \mathcal{X}_+,\ \text{and} \\ -\infty < c_0 \leq 0, & \forall\ \boldsymbol{x} \in \mathcal{X}_0. \end{cases} \tag{55}$$

Either choice of $\mu_p^*(\boldsymbol{x})$ results in the generator loss $\mathcal{L}_G$ evaluating to zero, and the optimal generator distribution matching the data distribution: $p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x})$. Further, the optimality of the generator is independent of the value of λ_p.

Summarizing, the optima are:

$$p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x});\ \mu_p^*(\boldsymbol{x}) = 0,\ \forall\ \boldsymbol{x} \in \mathcal{X} \quad \text{and} \quad \lambda_p \in (-\infty, \infty)$$

The preceding derivation considered the n-dimensional case with $n \geq 3$. The $n = 1$ case was presented in Appendix B.2. The analysis for $n = 2$ is presented next. The analysis follows analogously to the $n \geq 3$ case up until Equation (50). Thereafter, the difference lies in the Fourier transform of $\phi(\boldsymbol{x})$. We have

$$\phi(\boldsymbol{x}) = \begin{cases} -\frac{1}{2\pi}\ln(\|\boldsymbol{x}\|),\ \boldsymbol{x} \in \mathbb{R}^n - \{\boldsymbol{0}\}, & n = 2 \\ \kappa_n\|\boldsymbol{x}\|^{2-n} & n \geq 3 \end{cases}$$

To determine the Fourier transform of $\phi(\boldsymbol{x}) = -\dfrac{1}{2\pi}\ln(\|\boldsymbol{x}\|)$, we must pay attention to the singularity at the origin. The Fourier transform must be defined in the distributional sense. Vladimirov (1984) showed that the Fourier transform of $\dfrac{1}{\|\boldsymbol{x}\|^2},\ \boldsymbol{x} \in \mathbb{R}^n - \{\mathbf{0}\}$ is given in the distributional sense as follows:

$$\mathscr{F}\left\{\frac{1}{\|\boldsymbol{x}\|^2}\right\} = -2\pi\ln(\|\boldsymbol{\omega}\|) - 2\pi C_0, \quad \boldsymbol{\omega} \in \mathbb{R}^n - \{\mathbf{0}\},$$

where

$$C_0 = \int_0^1 \frac{1 - J_0(u)}{u}\,\mathrm{d}u - \int_1^\infty \frac{J_0(u)}{u}\,\mathrm{d}u,$$

where in turn, $J_0(u)$ is the zeroth-order Bessel function of the first kind (Abramowitz, 1974). From the duality property of the Fourier transform, we have

$$\mathscr{F}\left\{\phi(\boldsymbol{x})\right\} = \mathscr{F}\left\{-\frac{1}{2\pi}\ln(\|\boldsymbol{x}\|)\right\} = \frac{1}{2\pi}\frac{1}{\|\boldsymbol{\omega}\|^2} + \frac{C_0}{2\pi}\delta(\boldsymbol{\omega}).$$

From here on, the rest of the analysis corresponding to the optimal WGAN-GNP generator proceeds as in the $n \geq 3$ case in particular, Equation (53) onward, where $f_{-2}(\boldsymbol{\omega})$ is replaced with the above Fourier transform. Ultimately, the optimal solution, $p_g^*(\boldsymbol{x}) = p_d(\boldsymbol{x})$, remains unchanged.

C.3 Optimal Lagrange Multiplier in WGAN-FS (n-D)

Consider the Fourier-series (FS) discriminator $D_{FS}^*(\boldsymbol{x})$ in the multivariate case:

$$D_{FS}^*(\boldsymbol{x}) \approx \frac{1}{\lambda_{FS}^{*2}}\left(\langle \boldsymbol{a}, \boldsymbol{x}\rangle + \text{constant} + \sum_{\boldsymbol{m}\in\mathcal{M}}\left(\gamma_{\boldsymbol{m}}^r \cos(\omega_o\langle \boldsymbol{m}, \boldsymbol{x}\rangle) + \gamma_{\boldsymbol{m}}^i \sin(\omega_o\langle \boldsymbol{m}, \boldsymbol{x}\rangle)\right)\right).$$

Taking the derivative with respect to x_ℓ and squaring, we get:

$$\left(\frac{\partial D_{FS}^*}{\partial x_\ell}\right)^2 = \frac{1}{\lambda_{FS}^{*2}}\left(a_\ell - \sum_{\boldsymbol{m}\in\mathcal{M}}\left(\gamma_{\boldsymbol{m}}^r \omega_o m_\ell \sin(\omega_o\langle \boldsymbol{m}, \boldsymbol{x}\rangle) + \gamma_{\boldsymbol{m}}^i \omega_o m_\ell \cos(\omega_o\langle \boldsymbol{m}, \boldsymbol{x}\rangle)\right)\right)^2.$$

Using the Cauchy-Schwartz inequality:

$$\left(\sum_{\ell=1}^n u_\ell.1\right)^2 \ \leq\ n\sum_{\ell=1}^n u_\ell^2,$$

we obtain the following bound:

$$\left(\frac{\partial D_{FS}^*}{\partial x_\ell}\right)^2 \leq \frac{2|\mathcal{M}|+1}{\lambda_{FS}^{*2}}\left(a_\ell^2 + \sum_{\boldsymbol{m}\in\mathcal{M}}\omega_o^2 m_\ell^2\left(\gamma_{\boldsymbol{m}}^{r^2}\sin^2(\omega_o\langle \boldsymbol{m}, \boldsymbol{x}\rangle) + \gamma_{\boldsymbol{m}}^{i^2}\cos^2(\omega_o\langle \boldsymbol{m}, \boldsymbol{x}\rangle)\right)\right),$$

where $|\mathcal{M}|$ is the cardinality of the set $\mathcal{M}$ of the selected harmonics. Summing over ℓ yields:

$$\|\nabla D^*\|_2^2 \leq \frac{2|\mathcal{M}|+1}{\lambda_{FS}^{*2}} \left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \omega_o^2 \|\boldsymbol{m}\|^2 \left(\gamma_{\boldsymbol{m}}^{r^2} \sin^2(\omega_o \langle \boldsymbol{m}, \boldsymbol{x}\rangle) + \gamma_{\boldsymbol{m}}^{i^2} \cos^2(\omega_o \langle \boldsymbol{m}, \boldsymbol{x}\rangle) \right) \right)$$

$$\Rightarrow \|\nabla D^*\|_2^2 \leq \frac{2|\mathcal{M}|+1}{\lambda_{FS}^{*2}} \left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \left(\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r\right) \cos(2\omega_o \langle \boldsymbol{m}, \boldsymbol{x}\rangle) \right) \right),$$

$$\text{where } \tau_{\boldsymbol{m}}^r = \frac{1}{2}(\gamma_{\boldsymbol{m}}^r)^2 \omega_o^2 \|\boldsymbol{m}\|^2, \quad \text{and} \quad \tau_{\boldsymbol{m}}^i = \frac{1}{2}(\gamma_{\boldsymbol{m}}^i)^2 \omega_o^2 \|\boldsymbol{m}\|^2.$$

Enforcing the gradient-norm penalty: $\int_{\mathcal{X}} \left(\|\nabla D^*\|_2^2 - 1 \right) \, \mathrm{d}\boldsymbol{x} = 0$, gives

$$0 \leq \int_{\mathcal{X}} \left((2|\mathcal{M}|+1) \left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \left(\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r\right) \cos(2\omega_o \langle \boldsymbol{m}, \boldsymbol{x}\rangle) \right) \right) - \lambda_{FS}^{*2} \right) \mathrm{d}\boldsymbol{x}.$$

Simplifying the above gives the condition on the optimal Lagrange multiplier:

$$\lambda_{FS}^{*2} \leq \frac{(2|\mathcal{M}|+1)}{|\mathcal{X}|} \int_{\mathcal{X}} \left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \left(\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r\right) \cos(2\omega_o \langle \boldsymbol{m}, \boldsymbol{x}\rangle) \right) \right) \, \mathrm{d}\boldsymbol{x}.$$

$$= (2|\mathcal{M}|+1) \left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r\right) + \sum_{\boldsymbol{m}\in\mathcal{M}} \left(\left(\frac{\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r}{|\mathcal{X}|} \right) \int_{\mathcal{X}} \cos(2\omega_o \langle \boldsymbol{m}, \boldsymbol{x}\rangle) \mathrm{d}\boldsymbol{x} \right) \right).$$

Given the data

$$\mathcal{D} = \{\boldsymbol{x}_k\} = \{\boldsymbol{x}_d, \text{ s.t. } \boldsymbol{x}_d \sim p_d\} \bigcup \{\boldsymbol{x}_g, \text{s.t. } \boldsymbol{x}_g \sim p_g\}$$

of cardinality $|\mathcal{D}| = N$, we can estimate the upper bound on λ_{FS}^* as follows:

$$\lambda_{FS}^* \leq \sqrt{(2|\mathcal{M}|+1) \left(\|\boldsymbol{a}\|^2 + \sum_{\boldsymbol{m}\in\mathcal{M}} (\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r) + \frac{1}{N} \sum_{k=1}^{N} \sum_{\boldsymbol{m}\in\mathcal{M}} (\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r) \cos(2\omega_o \langle \boldsymbol{m}, \boldsymbol{x}_k\rangle) \right)}.$$

Recall that $\|\boldsymbol{a}\| = 1$ for the optimal discriminator to satisfy the gradient-norm penalty Ω_D when $p_g^* = p_d$ (cf. Section 4). In practice, the contribution of $\|\boldsymbol{a}\|$ was found to be negligible in comparison with the other terms. The worst-case choice for the Lagrange multiplier is

$$\lambda_{FS}^* = \sqrt{(2|\mathcal{M}|+1) \left(\sum_{\boldsymbol{m}\in\mathcal{M}} (\tau_{\boldsymbol{m}}^i + \tau_{\boldsymbol{m}}^r) + \frac{1}{N} \sum_{k=1}^{N} \sum_{\boldsymbol{m}\in\mathcal{M}} (\tau_{\boldsymbol{m}}^i - \tau_{\boldsymbol{m}}^r) \cos(2\omega_o \langle \boldsymbol{m}, \boldsymbol{x}_k\rangle) \right)}.$$

Appendix D. Fourier-series Error Analysis

In this appendix, we analyze the various sources of error in approximating the infinite Fourier series of the generator and data distributions, and the discriminator, and derive upper bounds for the mean-squared error when truncating the Fourier series, and when computing the Fourier coefficients through sample estimates. Our analysis is inspired by the 1-D analysis reported in Giardina and Chirlian (1972). We also generalize the results to higher dimensions, which is pertinent to the present discussion.

D.1 Fourier-series Truncation Error (1-D)

Consider the infinite and truncated Fourier series expansions of p_d, p_g and $D^*(x)$, given as follows:

$$p_d(x) = \sum_{m\in\mathbb{Z}} \alpha_m e^{\mathrm{j}\omega_o m x}, \qquad \text{and} \qquad \tilde{p}_d(x) = \sum_{m=-M}^{M} \alpha_m e^{\mathrm{j}\omega_o m x},$$

$$p_g(x) = \sum_{m\in\mathbb{Z}} \beta_m e^{\mathrm{j}\omega_o m x}, \qquad \text{and} \qquad \tilde{p}_g(x) = \sum_{m=-M}^{M} \beta_m e^{\mathrm{j}\omega_o m x}, \quad \text{and}$$

$$D_{FS}(x) = \frac{1}{\lambda_d}\sum_{m\in\mathbb{Z}} \gamma_m e^{\mathrm{j}\omega_o m x}, \qquad \text{and} \qquad \tilde{D}_{FS}(x) = \frac{1}{\lambda_d}\sum_{m=-M}^{M} \gamma_m e^{\mathrm{j}\omega_o m x}.$$

Considering the Fourier expansion of p_d, the mean-squared error in truncation is given by

$$\epsilon_{p_d}^2 = \|p_d(x) - \tilde{p}_d(x)\|_2^2 = \int_{-\infty}^{\infty} \left(p_d(x) - \tilde{p}_d(x)\right)^2 \, \mathrm{d}x.$$

Applying Parseval's identity, we get

$$\epsilon_{p_d}^2 = (2T) \sum_{m=M+1}^{\infty} |\alpha_m|^2.$$

A continuously differentiable and compactly supported function $p_d \in \mathcal{C}_c^1$ is of bounded variation over the support $\mathcal{X}$, denoted by $V_{\mathcal{X}}\left[p_d\right] = \int_{\mathcal{X}} |p_d'(x)| \, \mathrm{d}x \leq B_d$. Consider the modulus of the Fourier coefficient α_m:

$$\begin{aligned} |\alpha_m| &= \frac{1}{T}\left|\int_{-\frac{T}{2}}^{\frac{T}{2}} p_d(x) e^{-\mathrm{j}\omega_o m x} \mathrm{d}x\right| \\ &= \frac{1}{\omega_o m T}\left|\int_{-\frac{T}{2}}^{\frac{T}{2}} p_d(x) \, \mathrm{d}(e^{-\mathrm{j}\omega_o m x})\right|. \end{aligned}$$

Integrating by parts, and noting that $p_d(x)e^{-\mathrm{j}\omega_o m x}$ is T-periodic, we have:

$$\begin{aligned} |\alpha_m| &= \frac{1}{\omega_o m T}\left|\int_{-\frac{T}{2}}^{\frac{T}{2}} e^{-\mathrm{j}\omega_o m x} \, \mathrm{d}\left(p_d(x)\right)\right| \\ &\leq \frac{1}{\omega_o m T} V_{\mathcal{X}}\left[p_d(x)\right]. \end{aligned} \tag{56}$$

This gives us the bound $|\alpha_m| \leq \frac{B_d}{2\pi m}$. Substituting this result into the mean-squared error yields

$$\epsilon_{p_d}^2 \leq \frac{B_d^2}{\pi\omega_o} \sum_{m=M+1}^{\infty} \frac{1}{m^2}.$$

Bounding the tail sum with an integral results in the following:

$$\epsilon_{p_d}^2 \leq \frac{B_d^2}{\pi\omega_o}\int_{y=M}^{\infty}\frac{1}{y^2}\mathrm{d}y$$
$$\Rightarrow \epsilon_{p_d}^2 \leq \left(\frac{B_d^2}{\pi\omega_o}\right)\frac{1}{M}.$$

Similarly, the mean-squared error in the truncation of p_g is bounded as follows:

$$\epsilon_{p_g}^2 \leq \left(\frac{B_g^2}{\pi\omega_o}\right)\frac{1}{M},$$

where B_g is the bound on the variation of p_g.
These truncation errors result in the following error in the truncated Fourier series of the discriminator:

$$\epsilon_D^2 = (2T)\sum_{m=M+1}^{\infty}|\gamma_m|^2 = \frac{2T}{\omega_o^2}\sum_{m=M+1}^{\infty}\frac{|\alpha_m-\beta_m|^2}{m^2}.$$

Using the inequality $(a-b)^2 \leq 2(a^2+b^2)$ gives

$$\epsilon_D^2 \leq \frac{4T}{\omega_o^2}\sum_{m=M+1}^{\infty}\frac{|\alpha_m|^2+|\beta_m|^2}{m^2}.$$

From the result in Equation (56), we have

$$\epsilon_D^2 \leq \frac{4(B_d^2+B_g^2)}{\omega_o^3}\sum_{m=M+1}^{\infty}\frac{1}{m^4}.$$

Bounding the tail sum with an integral gives

$$\epsilon_D^2 \leq \frac{4(B_d^2+B_g^2)}{\omega_o^3}\int_{y=M}^{\infty}\frac{1}{y^4}\mathrm{d}y$$
$$= \frac{4(B_d^2+B_g^2)}{3\omega_o^3M^3},$$

which is the desired bound on the truncation error of the Fourier-series expansion of the discriminator.

D.2 Fourier-series Truncation Error (n-D)

The infinite and truncated complex Fourier-series expansions of $D(\boldsymbol{x})$ are given by

$$D_{FS}(\boldsymbol{x}) = \frac{1}{\lambda_d}\sum_{\boldsymbol{m}\in\mathbb{Z}^n}\gamma_{\boldsymbol{m}}e^{\mathrm{j}\omega_o\langle\boldsymbol{m},\boldsymbol{x}\rangle}, \text{ and} \qquad \tilde{D}_{FS}(\boldsymbol{x}) = \frac{1}{\lambda_d}\sum_{\boldsymbol{m}\in[M]^n}\gamma_{\boldsymbol{m}}e^{\mathrm{j}\omega_o\langle\boldsymbol{m},\boldsymbol{x}\rangle},$$

respectively, where $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{m} = [m_1, m_2, \ldots, m_n]^{\mathrm{T}} \in \mathbb{Z}_+^n$, $[M]^n$ denotes the Cartesian product space $\{-M, -M+1, \ldots, M-1, M\}^n$ and ω_o is the fundamental frequency common to all the dimensions. The mean-squared error in truncation is given by

$$\epsilon_D^2 = \|D_{FS}(\boldsymbol{x}) - \tilde{D}_{FS}(\boldsymbol{x})\|_2^2 = \int_{-\infty}^{\infty} \Big(D_{FS}(\boldsymbol{x}) - \tilde{D}_{FS}(\boldsymbol{x})\Big)^2 \, \mathrm{d}\boldsymbol{x}.$$

From Parseval's identity, we have:

$$\begin{aligned}\epsilon_D^2 &= (2T)^n \sum_{m_1=M+1}^{\infty} \sum_{m_2=M+1}^{\infty} \ldots \sum_{m_n=M+1}^{\infty} |\gamma_{\boldsymbol{m}}|^2 \\ &= (2T)^n \sum_{\boldsymbol{m} \in \mathbb{Z}_+^n \setminus [M]_+^n} |\gamma_{\boldsymbol{m}}|^2,\end{aligned}$$

where $[M]_+^n$ denotes the Cartesian product space $\{1, 2, \ldots, M-1, M\}^n$. Substituting for $\gamma_{\boldsymbol{m}}$ from Equation (22) gives

$$\epsilon_D^2 = \frac{(2T)^n}{4\omega_o^4} \sum_{\boldsymbol{m} \in \mathbb{Z}_+^n \setminus [M]_+^n} \frac{|\alpha_{\boldsymbol{m}} - \beta_{\boldsymbol{m}}|^2}{\|\boldsymbol{m}\|^4}.$$

From Cauchy-Schwartz inequality, we have $|\alpha_{\boldsymbol{m}} - \beta_{\boldsymbol{m}}|^2 \leq 2\left(\alpha_{\boldsymbol{m}}^2 + \beta_{\boldsymbol{m}}^2\right)$. Substituting into the above equation, we have

$$\epsilon_D^2 \leq \frac{(2T)^n}{\omega_o^4} \sum_{\boldsymbol{m} \in \mathbb{Z}_+^n \setminus [M]_+^n} \frac{|\alpha_{\boldsymbol{m}}|^2 + |\beta_{\boldsymbol{m}}|^2}{\|\boldsymbol{m}\|^4}. \tag{57}$$

The right-hand side of the truncation error can be improved by invoking additional smoothness assumptions on p_d and p_g. Consider p_d and p_g to be in $\mathcal{C}^\ell(\mathcal{X}) \bigcap W^{k,2}(\mathcal{X})$; $\ell > k$ (cf. Assumption 2). The fact that the functions are in $\mathcal{C}^\ell(\mathcal{X})$ ensures that the derivatives of p_d and p_g are well-defined in classical sense (as opposed to the weak derivatives, that are considered in the case of Sobolev spaces). Recall the definition of the Sobolev-k space $W^{k,2}$ that subsumes $\mathcal{C}^k(\mathcal{X})$ functions:

$$\begin{aligned}W^{k,2}(\mathcal{X}) &= \left\{ f \; ; \|f\|_{W^{k,2}} = \left(\|f\|_2^2 + \sum_{i=1}^{k} \|f^{(i)}\|_2^2 \right) < \infty \right\} \\ &\equiv \left\{ f \; ; \|f\|'_{W^{k,2}} = \|f\|_2^2 + \|f^{(k)}\|_2^2 < \infty \right\},\end{aligned}$$

where $f^{(i)}$ denotes the vector consisting of all i^{th} partial derivatives of f, and the equivalence holds in the sense of the topology induced by the norms $\|f\|_{W^{k,2}}$ and $\|f\|'_{W^{k,2}}$ (Sobolev, 1963). Furthermore, when considering the Fourier-series expansion of functions in $W^{k,2}$, the following equivalences hold in terms of the Fourier coefficients holds (Sobolev, 1963):

$$W^{k,2}(\mathcal{X}) = \left\{ f \in L_2(\mathcal{X}) \; ; \text{ s.t. } \sum_{\boldsymbol{m} \in \mathbb{Z}^n} \left(1 + \|\boldsymbol{m}\|^2 + \|\boldsymbol{m}\|^4 + \ldots + \|\boldsymbol{m}\|^{2k}\right) |f_{\boldsymbol{m}}|^2 < \infty \right\} \tag{58}$$

where $\{f_{\boldsymbol{m}}\}$ denote the Fourier coefficients of f. From the bound in Equation (58), for the infinite sum to converge, given data in $\mathbb{R}^n$, we require the individual terms $|f_{\boldsymbol{m}}|^2$ to decay at a rate greater than $\|\boldsymbol{m}\|^{-(2k+n)}$. Therefore, there exists a constant $\mathfrak{M}_f < \infty$ such that, for finite k, we have

$$|f_{\boldsymbol{m}}|^2 \leq \mathfrak{M}_f \|\boldsymbol{m}\|^{-(2k+n+1)}, \quad \forall\ \boldsymbol{m}.$$

While a similar bound can be derived using the exponent $-(2k+n+\tau),\ \tau = 1, 2, \ldots$, we set $\tau = 1$ to obtain a tight bound. This bound on $|f_{\boldsymbol{m}}|$ is adequate for the infinite sum in Equation (58) to be bounded, as shown in Appendix D.2.1. Substituting the above bound into Equation (57) yields

$$\epsilon_D^2 \leq \left(\frac{(2T)^n(\mathfrak{M}_{p_d} + \mathfrak{M}_{p_g})}{\omega_o^4}\right) \sum_{\boldsymbol{m} \in \mathbb{Z}_+^n \setminus [M]_+^n} \|\boldsymbol{m}\|^{-(2k+n+4)}.$$

The sum can be bounded by considering an appropriate integral as shown below:

$$\sum_{\boldsymbol{m} \in \mathbb{Z}^n \setminus [M]^n} \|\boldsymbol{m}\|^{-(2k+n+4)} \leq \int_{y_1=M}^{\infty} \int_{y_2=M}^{\infty} \cdots \int_{y_n=M}^{\infty} \|\boldsymbol{y}\|^{-(2k+n+4)}\ \mathrm{d}y_1 \mathrm{d}y_2 \ldots \mathrm{d}y_n,$$

where y_i is the i^{th} entry in $\boldsymbol{y}$. Converting to n-dimensional spherical coordinates and simplifying, we get

$$\sum_{\boldsymbol{m} \in \mathbb{Z}^n \setminus [M]^n} \|\boldsymbol{m}\|^{-(2k+n+4)} \leq \mathcal{S}_{n-1} \int_{r=M\sqrt{n}}^{\infty} r^{-(2k+5)}\ \mathrm{d}r,$$

where $\mathcal{S}_{n-1} = \pi^{\frac{n}{2}} \left(\Gamma\left(\frac{n}{2}\right)\right)^{-1}$ is the hyper-surface area of the n-dimensional unit sphere. Evaluating the integral on the right-hand side yields the following bound:

$$\sum_{\boldsymbol{m} \in \mathbb{Z}^n \setminus [M]^n} \|\boldsymbol{m}\|^{-(2k+n+4)} \leq \frac{\mathcal{S}_{n-1}}{(2k+4)}\ (M\sqrt{n})^{-(2k+4)}.$$

Using this bound gives us the desired result of Theorem 5:

$$\epsilon_D^2 \leq \underbrace{\frac{(2T)^n(\mathfrak{M}_{p_d} + \mathfrak{M}_{p_g})\mathcal{S}_{n-1}}{2\omega_o^4}}_{\mathfrak{C}_{n,T}} \left(\frac{(M^2 n)^{-(k+2)}}{(k+2)}\right) = \mathfrak{C}_{n,T} \left(\frac{(M^2 n)^{-(k+2)}}{(k+2)}\right).$$

Extension to Infinitely Differentiable Functions: The above analysis can be extended to the case when p_d and p_g belong to $\mathcal{C}^\infty(\mathcal{X})$, i.e., they are infinitely differentiable. In this case, the definition in Equation (58) holds for all k, and we have:

$$|f_{\boldsymbol{m}}|^2 \approx \mathfrak{M}'_f e^{-\|\boldsymbol{m}\|^2},$$

where $\mathfrak{M}'_{f_{\boldsymbol{m}}}$ is a constant dependent on f, that is, the Fourier coefficients decay exponentially. Substituting the above in ϵ_{MS}^2 and bounding the sum by the n-dimensional integral in

hyperspherical coordinates gives

$$\epsilon_D^2 \leq \left(\frac{(2T)^n(\mathfrak{M}'_{p_d}+\mathfrak{M}'_{p_g})}{\omega_o^4}\right)\mathcal{S}_{n-1}\int_{r=M\sqrt{n}}^{\infty} e^{-r^2}r^{n-1}\,\mathrm{d}r$$
$$= \left(\frac{(2T)^n(\mathfrak{M}'_{p_d}+\mathfrak{M}'_{p_g})}{\omega_o^4}\right)\mathcal{S}_{n-1}\int_{s=M^2n}^{\infty} e^{-s}s^{\frac{n}{2}-1}\,\mathrm{d}s,$$

where the equality results from the change of variable $s = r^2$. The integral is the upper incomplete gamma function $\Gamma\left(\frac{n}{2}, M^2\,n\right)$, This gives us the bound

$$\epsilon_D^2 \leq \underbrace{\left(\frac{(2T)^n(\mathfrak{M}'_{p_d}+\mathfrak{M}'_{p_g})}{\omega_o^4}\right)\mathcal{S}_{n-1}}_{\mathfrak{C}'_{n,T}}\Gamma\left(\frac{n}{2}, M^2\,n\right).$$

Asymptotically, as $M \to \infty$, the upper incomplete gamma function satisfies the property: $\Gamma\left(\frac{n}{2}, M^2\,n\right) \to (M^2 n)^{\frac{n}{2}}e^{-M^2 n}$.

D.2.1 Bound on the Fourier Coefficients

Consider the Sobolev-k space $W^{k,2}$. For functions drawn from $W^{k,2}(\mathcal{X})$, the coefficients of the Fourier-series expansion $\{f_{\boldsymbol{m}}\}$ must satisfy (Sobolev, 1963):

$$\mathrm{S}_k = \sum_{\boldsymbol{m}\in\mathbb{Z}^n}\left(\left(1+\|\boldsymbol{m}\|^2+\|\boldsymbol{m}\|^4+\ldots+\|\boldsymbol{m}\|^{2k}\right)|f_{\boldsymbol{m}}|^2\right)$$
$$= |f_{\boldsymbol{0}}|^2 + \sum_{\boldsymbol{m}\in\mathbb{Z}^n\setminus\{\boldsymbol{0}\}}\left(\left(1+\|\boldsymbol{m}\|^2+\|\boldsymbol{m}\|^4+\ldots+\|\boldsymbol{m}\|^{2k}\right)|f_{\boldsymbol{m}}|^2\right) < \infty. \tag{59}$$

Consider the bound on the Fourier coefficients for finite k, given by:

$$|f_{\boldsymbol{m}}|^2 \leq \mathfrak{M}_f\|\boldsymbol{m}\|^{-(2k+n+1)}, \quad \forall\ \boldsymbol{m}.$$

Since we are working with Fourier-series representations of p.d.f.s, we have $f_{\boldsymbol{0}} = 1$. Therefore

$$\mathrm{S}_k \leq 1 + \mathfrak{M}_f\sum_{\boldsymbol{m}\in\mathbb{Z}^n\setminus\{\boldsymbol{0}\}}\left(\left(1+\|\boldsymbol{m}\|^2+\|\boldsymbol{m}\|^4+\ldots+\|\boldsymbol{m}\|^{2k}\right)\|\boldsymbol{m}\|^{-(2k+n+1)}\right)$$
$$= 1 + \mathfrak{M}_f\sum_{\boldsymbol{m}\in\mathbb{Z}^n\setminus\{\boldsymbol{0}\}}\left(\|\boldsymbol{m}\|^{-(2k+n+1)}+\|\boldsymbol{m}\|^{-((2k-2)+n+1)}+\ldots+\|\boldsymbol{m}\|^{-(n+1)}\right)$$
$$\leq 1 + (k+1)\,\mathfrak{M}_f\sum_{\boldsymbol{m}\in\mathbb{Z}^n\setminus\{\boldsymbol{0}\}}\|\boldsymbol{m}\|^{-(n+1)},$$

where the second inequality is obtained by approximating each term in the summation by the largest one. Further bounding the infinite sum by an appropriate integral as follows:

$$\mathrm{S}_k \leq 1 + 2(k+1)\,\mathfrak{M}_f\int_{y_1=1}^{\infty}\ldots\int_{y_n=1}^{\infty}\|\boldsymbol{y}\|^{-(n+1)}\,\mathrm{d}y_1\ldots\mathrm{d}y_n,$$

and converting to n-dimensional spherical coordinates, we get

$$\mathrm{S}_k \leq 1 + (k+1)\,\mathfrak{M}_f\,\mathcal{S}_{n-1}\int_{r=\sqrt{n}}^{\infty} r^{-2}\,\mathrm{d}r,$$

where $\mathcal{S}_{n-1}$ is the hyper-surface area of the n-D unit sphere. Evaluating the integral yields

$$\mathrm{S}_k \leq 1 + (k+1)\,\mathfrak{M}_f\,\mathcal{S}_{n-1} n^{-\frac{1}{2}},$$

which is finite.

D.3 Error in the Sample Estimation of Fourier Coefficients

In Appendix D.2, we analyzed the effect of truncating the Fourier series with a rectangular sum of M terms along each dimension. In practice, in addition to truncating the Fourier series, there is also error arising out of approximating the Fourier coefficients of p_d and p_g. Consider the trigonometric Fourier expansion of p_d:

$$p_d(\boldsymbol{x}) = \frac{\alpha_{\mathbf{0}}}{2} + \sum_{\boldsymbol{m}\in\mathbb{Z}_+^n} \alpha^r_{\boldsymbol{m}}\cos(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle) + \alpha^i_{\boldsymbol{m}}\sin(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)$$

and its M^{th}-order approximation:

$$\tilde{p}_d(\boldsymbol{x}) = \frac{\bar{\alpha}_{\mathbf{0}}}{2} + \sum_{\boldsymbol{m}\in[M]_+^n} \bar{\alpha}^r_{\boldsymbol{m}}\cos(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle) + \bar{\alpha}^i_{\boldsymbol{m}}\sin(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle),$$

where $\boldsymbol{x}\in\mathbb{R}^n$, $\mathbb{Z}_+^n$ denotes the set of positive non-zero integers, $[M]_+^n$ represents the product space $\{1,2,\ldots,M\}^n$, $\boldsymbol{m}=[m_1,m_2,\ldots,m_n]\in\mathbb{Z}_+^n$, and $\alpha^r_{\boldsymbol{m}}$ and $\bar{\alpha}^r_{\boldsymbol{m}}$ are the true Fourier coefficient and its N-sample estimate given by

$$\alpha^r_{\boldsymbol{m}} = \int_{\mathcal{X}} \cos(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)p_d(\boldsymbol{x})\;\mathrm{d}\boldsymbol{x} = \mathbb{E}_{\boldsymbol{x}\sim p_d}\left[\cos(\omega_o\langle \boldsymbol{m},\boldsymbol{x}\rangle)\right], \text{and}$$
$$\bar{\alpha}^r_{\boldsymbol{m}} = \frac{1}{N}\sum_{\substack{k=1\\ \boldsymbol{x}_k\sim p_d}}^{N} \cos(\omega_o\langle \boldsymbol{m},\boldsymbol{x}_k\rangle),$$

respectively. Considering independent and identically distributed (*i.i.d.*) samples $\{\boldsymbol{x}_i\}$, it can be shown that

$$\mathbb{E}_{\boldsymbol{x}}\left[\bar{\alpha}^r_{\boldsymbol{m}}\right] = \frac{1}{N}\sum_{\substack{k=1\\ \boldsymbol{x}_k\sim p_d}}^{N} \mathbb{E}_{\boldsymbol{x}}\left[\cos(\omega_o\langle \boldsymbol{m},\boldsymbol{x}_k\rangle)\right] = \alpha^r_{\boldsymbol{m}}$$

Similarly, the variance of the estimate is given by

$$\mathrm{Var}\left(\bar{\alpha}^r_{\boldsymbol{m}}\right) = \mathbb{E}_{\boldsymbol{x}}\left[\left(\bar{\alpha}^r_{\boldsymbol{m}}\right)^2\right] - \left(\mathbb{E}_{\boldsymbol{x}}\left[\bar{\alpha}^r_{\boldsymbol{m}}\right]\right)^2 = \mathbb{E}_{\boldsymbol{x}}\left[\left(\bar{\alpha}^r_{\boldsymbol{m}}\right)^2\right] - \left(\alpha^r_{\boldsymbol{m}}\right)^2. \tag{60}$$

Expanding the first term on the right-hand side yields

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[(\bar{\alpha}^r_{\boldsymbol{m}})^2\right] &= \frac{1}{N^2}\,\mathbb{E}_{\boldsymbol{x}}\left[\left(\sum_{i=1}^{N}\cos\left(\omega_o\langle\boldsymbol{m},\boldsymbol{x}_i\rangle\right)\right)^2\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{x}}\left[\cos^2\left(\omega_o\langle\boldsymbol{m},\boldsymbol{x}_i\rangle\right)\right] \\ &\quad + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\mathbb{E}_{\boldsymbol{x}}\left[\cos\left(\omega_o\langle\boldsymbol{m},\boldsymbol{x}_i\rangle\right)\cos\left(\omega_o\langle\boldsymbol{m},\boldsymbol{x}_j\rangle\right)\right].\end{aligned}$$

For *i.i.d.* samples $\{\boldsymbol{x}_i\}$, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[(\bar{\alpha}^r_{\boldsymbol{m}})^2\right] &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{x}}\left[\cos^2\left(\omega_o\langle\boldsymbol{m},\boldsymbol{x}_i\rangle\right)\right] \\ &\quad + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\mathbb{E}_{\boldsymbol{x}}\left[\cos\left(\omega_o\langle\boldsymbol{m},\boldsymbol{x}_i\rangle\right)\right]\mathbb{E}_{\boldsymbol{x}}\left[\cos\left(\omega_o\langle\boldsymbol{m},\boldsymbol{x}_j\rangle\right)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{x}}\left[\cos^2\left(\omega_o\langle\boldsymbol{m},\boldsymbol{x}_i\rangle\right)\right] + \left(\frac{N^2-N}{N^2}\right)\left(\alpha^r_{\boldsymbol{m}}\right)^2.\end{aligned}$$

Applying the half-angle trigonometric formulae gives

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[(\bar{\alpha}^r_{\boldsymbol{m}})^2\right] &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{x}}\left[\frac{1}{2}+\frac{1}{2}\cos\left(2\omega_o\langle\boldsymbol{m},\boldsymbol{x}_i\rangle\right)\right] + \left(\frac{N^2-N}{N^2}\right)\left(\alpha^r_{\boldsymbol{m}}\right)^2, \\ &= \frac{1}{2N} + \frac{1}{2N}\,\mathbb{E}_{\boldsymbol{x}}\left[\cos\left(2\omega_o\langle\boldsymbol{m},\boldsymbol{x}_i\rangle\right)\right] + \left(\frac{N^2-N}{N^2}\right)\left(\alpha^r_{\boldsymbol{m}}\right)^2, \\ \Rightarrow \mathbb{E}_{\boldsymbol{x}}\left[(\bar{\alpha}^r_{\boldsymbol{m}})^2\right] &= \frac{1}{2N} + \frac{1}{2N}\alpha^r_{2\boldsymbol{m}} + \left(\frac{N^2-N}{N^2}\right)\left(\alpha^r_{\boldsymbol{m}}\right)^2.\end{aligned}$$

Substituting the above into Equation (60) yields:

$$\mathrm{Var}\left(\bar{\alpha}^r_{\boldsymbol{m}}\right) = \frac{1}{N}\left(\frac{1}{2}+\frac{1}{2}\alpha^r_{2\boldsymbol{m}} - \left(\alpha^r_{\boldsymbol{m}}\right)^2\right). \tag{61}$$

A similar analysis for $\bar{\alpha}^i_{\boldsymbol{m}}$ gives

$$\mathbb{E}_{\boldsymbol{x}}\left[\bar{\alpha}^i_{\boldsymbol{m}}\right] = \alpha^i_{\boldsymbol{m}}, \quad \text{and} \quad \mathrm{Var}\left(\bar{\alpha}^i_{\boldsymbol{m}}\right) = \frac{1}{N}\left(\frac{1}{2}-\frac{1}{2}\alpha^r_{2\boldsymbol{m}} - \left(\alpha^r_{\boldsymbol{m}}\right)^2\right). \tag{62}$$

We first bound the mean-squared error between the target expansion $p_d(\boldsymbol{x})$ and the approximation $\tilde{p}_d(\boldsymbol{x})$ in general, and subsequently, specialize the result for infinitely differentiable functions. Consider the error

$$\epsilon^2_{p_d} = \|p_d(\boldsymbol{x}) - \tilde{p}_d(\boldsymbol{x})\|_2^2 = \int_{\mathcal{X}}\left(p_d(\boldsymbol{x}) - \tilde{p}_d(\boldsymbol{x})\right)^2\,\mathrm{d}\boldsymbol{x}.$$

From Parseval identity, we have:

$$\epsilon_{p_d}^2 = \frac{|\bar{\alpha}_{\mathbf{0}} - \alpha_{\mathbf{0}}|^2}{2} + \sum_{\boldsymbol{m}\in[M]_+^n} \left(|\bar{\alpha}_{\boldsymbol{m}}^r - \alpha_{\boldsymbol{m}}^r|^2 + |\bar{\alpha}_{\boldsymbol{m}}^i - \alpha_{\boldsymbol{m}}^i|^2\right) + \sum_{\boldsymbol{m}\in\mathbb{Z}_+^n\setminus[M]_+^n} \left(|\alpha_{\boldsymbol{m}}^r|^2 + |\alpha_{\boldsymbol{m}}^i|^2\right).$$

The analysis for $\boldsymbol{m} = \mathbf{0}$ can be accounted for in $\bar{\alpha}_{\boldsymbol{m}}^r$, and we have $\mathbb{E}_x[\bar{\alpha}_{\mathbf{0}}^r] = \alpha_{\mathbf{0}}^r = 1$ with $\mathrm{Var}(\bar{\alpha}_{\mathbf{0}}^r) = 0$. Taking the expected value of ϵ_{MS}^2 with respect to $\boldsymbol{x}$, we get

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{p_d}^2\right] &= \sum_{\boldsymbol{m}\in[M]_+^n} \mathbb{E}_{\boldsymbol{x}}\left[(\bar{\alpha}_{\boldsymbol{m}}^r - \alpha_{\boldsymbol{m}}^r)^2\right] + \mathbb{E}_{\boldsymbol{x}}\left[\left(\bar{\alpha}_{\boldsymbol{m}}^i - \alpha_{\boldsymbol{m}}^i\right)^2\right] + \sum_{\mathbb{Z}_+^n\setminus[M]_+^n} |\alpha_{\boldsymbol{m}}^r|^2 + |\alpha_{\boldsymbol{m}}^i|^2 \\ &= \underbrace{\sum_{\boldsymbol{m}\in[M]_+^n} \mathrm{Var}(\bar{\alpha}_{\boldsymbol{m}}^r) + \mathrm{Var}(\bar{\alpha}_{\boldsymbol{m}}^i)}_{T_1} + \underbrace{\sum_{\mathbb{Z}_+^n\setminus[M]_+^n} |\alpha_{\boldsymbol{m}}^r|^2 + |\alpha_{\boldsymbol{m}}^i|^2}_{T_2},\end{aligned}$$

where T_1 the statistical component of the error, caused by the error in approximating the Fourier coefficient by replacing expectations with their samples estimates, and T_2 is deterministic, given the choice of the truncation frequency M. Substituting for the variance terms and simplifying, we get:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{p_d}^2\right] &= \frac{1}{N}\sum_{\boldsymbol{m}\in[M]_+^n} \left(1 - |\alpha_{\boldsymbol{m}}^r|^2 - |\alpha_{\boldsymbol{m}}^i|^2\right) + \sum_{\mathbb{Z}_+^n\setminus[M]_+^n} \left(|\alpha_{\boldsymbol{m}}^r|^2 + |\alpha_{\boldsymbol{m}}^i|^2\right), \\ &= \frac{M^n}{N} - \sum_{\boldsymbol{m}\in[M]_+^n} |\alpha_{\boldsymbol{m}}|^2 + \sum_{\mathbb{Z}_+^n\setminus[M]_+^n} |\alpha_{\boldsymbol{m}}|^2, \end{aligned} \tag{63}$$

where $\alpha_{\boldsymbol{m}} = \alpha_{\boldsymbol{m}}^r + \mathrm{j}\,\alpha_{\boldsymbol{m}}^i$ are the coefficients of the exponential Fourier series. Akin to the analysis in Appendix D.2, consider $\mathfrak{m}_{p_d}\|\boldsymbol{m}\|^{-(2k+n+1)} \leq |\alpha_{\boldsymbol{m}}|^2 \leq \mathfrak{M}_f\|\boldsymbol{m}\|^{-(2k+n+1)}$, where $\mathfrak{m}_{p_d} < \mathfrak{M}_{p_d}$. Employing these bounds yields

$$\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{MS}^2\right] \leq \frac{M^n}{N} - \frac{\mathfrak{m}_{p_d}}{N}\underbrace{\sum_{\boldsymbol{m}\in[M]_+^n} \|\boldsymbol{m}\|^{-(2k+n+1)}}_{\mathrm{S}_1} + \mathfrak{M}_{p_d}\underbrace{\sum_{\boldsymbol{m}\in\mathbb{Z}_+^n\setminus[M]_+^n} \|\boldsymbol{m}\|^{-(2k+n+1)}}_{\mathrm{S}_2}.$$

We can bound the elements in sum S_1 considering $\boldsymbol{m} = [1, 1, \ldots, 1]^{\mathrm{T}}$. S_2 can be bounded by the integral in hyperspherical coordinates similar to the procedure employed in Appendix D.2:

$$\begin{aligned}\mathrm{S}_2 &= \sum_{\boldsymbol{m}\in\mathbb{Z}_+^n\setminus[M]_+^n} \|\boldsymbol{m}\|^{-(2k+n+1)} \\ &\leq \int_{y_1=M}^{\infty}\int_{y_2=M}^{\infty}\cdots\int_{y_n=M}^{\infty} \|\boldsymbol{y}\|^{-(2k+n+1)}\ \mathrm{d}y_1\mathrm{d}y_2\ldots\ \mathrm{d}y_n \\ &= \mathcal{S}_{n-1}\int_{r=M\sqrt{n}}^{\infty} r^{-(2k+2)}\ \mathrm{d}r,\end{aligned}$$

where $\mathcal{S}_{n-1}$ is the surface area of the n-D unit hypersphere. For finite k, we have

$$\mathrm{S}_2 \leq \underbrace{\left(\frac{\mathcal{S}_{n-1}}{(2k+1)\,n^{(k+\frac{1}{2})}}\right)}_{\mathfrak{C}'_{n,k}}\frac{1}{M^{2k+1}} = \mathfrak{C}'_{n,k}\frac{1}{M^{2k+1}}.$$

Substituting for S_1 and S_2 gives

$$\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{p_d}^2\right] \leq \underbrace{\frac{M^n}{N}\left(1-\frac{\mathfrak{m}_{p_d}}{n^{k+\frac{n+1}{2}}}\right)}_{\varepsilon_{\text{stat}}} + \underbrace{\mathfrak{M}_{p_d}\mathfrak{C}'_{n,k}\frac{1}{M^{2k+1}}}_{\varepsilon_{\text{trunc}}}, \tag{64}$$

where $\varepsilon_{\text{stat}}$ and $\varepsilon_{\text{trunc}}$ denote the statistical and deterministic contributions to the error, respectively. Observe that as the dimensionality of data increases, the batch size must increase at a rate of $N \approx M^{n+1}$ for the approximation error $\varepsilon_{\text{stat}}$ to decay. For a given N, increasing M results in poorer estimates of the Fourier coefficients. One requires more samples to estimate the high-frequency components accurately, failing which, undesirable oscillations will appear in the representation. Experimental illustrations of this oscillation phenomenon will be presented in Appendix E.1. The contribution of $\varepsilon_{\text{stat}}$ associated with the sample estimation of $\alpha_{\boldsymbol{m}}$ is larger than the truncation error $\varepsilon_{\text{trunc}}$ for most M. This results in a trade-off between discarding high-frequency components versus poorly estimating them due to insufficient samples. The relative effect of $\varepsilon_{\text{stat}}$ and $\varepsilon_{\text{trunc}}$ indicates that it is indeed better to discard the high-frequency terms in these scenarios. We restrict our Fourier-series expansions in all experimentation to include up to $M_{low} = 2$ harmonics along all dimensions. For data in n-D, this results in an exponential blow-up of terms at the rate of $n^{M_{low}}$ for larger M. For instance, with $M_{low} = 3$ for 64-D data, there would be $64^3 \sim 2 \times 10^5$ terms in the Fourier expansion. To improve the representation of high-frequency components, we also uniformly randomly sample $L_1 = \mathcal{O}(10^3)$ harmonics between $M_{low} = 2$ and $M_{high} = 10$.

Extension to Infinitely Differentiable Functions: We now extend the result in Equation (64) to the case when p_d and p_g are infinitely differentiable. For $\mathcal{C}^\infty(\mathcal{X})$ functions, each term in the Fourier series is approximately $\mathcal{O}(e^{-\|\boldsymbol{m}\|^2})$. Similar to case when k is finite, there exist two constants $\mathfrak{m}'_{p_d} < \mathfrak{M}'_{p_d}$ such that:

$$\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_D^2\right] \sim \frac{M^n}{N} - \frac{\mathfrak{m}'_{p_d}}{N}\underbrace{\sum_{\boldsymbol{m}\in[M]_+^n} e^{-\|\boldsymbol{m}\|^2}}_{\mathrm{S}_1} + \mathfrak{M}'_{p_d}\underbrace{\sum_{\boldsymbol{m}\in\mathbb{Z}_+^n\setminus[M]_+^n} e^{-\|\boldsymbol{m}\|^2}}_{\mathrm{S}_2}.$$

Each term in S_1 can be bound by the largest value, as in the $\mathcal{C}_c^k$ case. The sum in S_2 can be bounded by the hyperspherical integral as shown in Appendix D.2, which yields

$$\begin{aligned}\mathrm{S}_2 = \sum_{\boldsymbol{m}\in\mathbb{Z}_+^n\setminus[M]_+^n} e^{-\|\boldsymbol{m}\|^2} &\leq \int_{y_1=M}^{\infty}\int_{y_2=M}^{\infty}\cdots\int_{y_n=M}^{\infty} e^{-\|\boldsymbol{y}\|^2}\,\mathrm{d}y_1\mathrm{d}y_2\ldots\mathrm{d}y_n \\ &= \mathcal{S}_{n-1}\int_{r=M\sqrt{n}}^{\infty} e^{-r^2}r^{n-1}\mathrm{d}r,\end{aligned}$$

where, as before, $\mathcal{S}_{n-1}$ denotes the surface area of a unit hypersphere in $\mathbb{R}^n$. As in Appendix D.2, the above integral represents the upper incomplete Gamma function, which gives the bound:

$$\mathrm{S}_2 \leq \mathcal{S}_{n-1}\Gamma\left(\frac{n}{2}, M^2\,n\right).$$

Substituting back for S_1 and S_2 gives:

$$\mathbb{E}_{\boldsymbol{x}}\left[\epsilon_{p_d}^2\right] \sim \underbrace{\frac{M^n}{N}\left(1-\mathfrak{m}'_{p_d}e^{-n}\right)}_{\varepsilon_{\text{stat}}} + \underbrace{\mathfrak{M}'_{p_d}\mathcal{S}_{n-1}\Gamma\left(\frac{n}{2}, M^2\,n\right)}_{\varepsilon_{\text{trunc}}},$$

where $\varepsilon_{\text{stat}}$ and $\varepsilon_{\text{trunc}}$ are the statistical and deterministic components of the approximation error, as discussed for the $\mathcal{C}^k(\mathcal{X})$ case, which is the desired bound on the approximation for $\mathcal{C}^\infty(\mathcal{X})$ functions. This gives us a bound on the error when approximating the Fourier series of truncated Gaussian distributions, such as in the case of latent-space matching in Wasserstein autoencoders.

Appendix E. Additional Experimentation

In this appendix, we present additional experiments and results on univariate and multivariate synthetic Gaussian data, and on learning the image-space distributions with WGAN-FS. We also provide additional details on the evaluation metrics used.

E.1 Additional Experiments on 1-D and 2-D Gaussians

To begin with, we present results on learning 1-D and 2-D Gaussians and Gaussian mixtures with the WGAN-FS algorithm.

Accuracy of the Fourier-series approximation: The experimental setup is as described in Section 3.7. The fundamental period T is set to 7 in all the experiments. In Figure 18, we present the target distribution p_d and its Fourier-series approximation for various choices of truncation order M and batch size N to illustrate the trade-off between truncating the Fourier series at low frequencies, and the error in approximating high-frequency coefficients with sparse samples. We observe that, when M is small (e.g., $M = 5$), introducing additional samples does not improve the quality of the approximation. This is a manifestation of the truncation error ($\varepsilon_{\text{trunc}}$) seen in Equation (64). For larger M, (e.g., $M \geq 25$), we observe that, in line with the theory, the high-frequency terms have a larger variance in their estimate and require larger N to be estimated accurately. This is the statistical component of the error, ($\varepsilon_{\text{stat}}$), which can be reduced by increasing N. As inferred from Equation (64), the artifacts can be suppressed from the approximation by setting $N > M^{n+1}$ (for example, with $N = 500$ for $M = 10$ and $N = 1000$ for $M = 25$). We observe similar performance trade-offs in the case of learning a bimodal Gaussian mixture in 1-D, as shown in Figure 19. Additionally, when N and M are both small, the Fourier-series approximation fails to capture the smaller mode. Based on these observations, we expect WGAN-FS to perform relatively better with lower M even in the high-dimensional setting.

Choosing the fundamental period T: We next present results on varying the assumed period T, given truncation order M and batch size N. Based on the previous experiments, we set $M = 10$ and $N = 100$. We consider the 1-D Gaussian learning scenario as described in Section 3.7. The target is a Gaussian $\mathcal{N}(5, 1)$, while the noise distribution is $\mathcal{N}(0, 1)$. We compare results for various choices of the time period $T \in \{2, 5, 7, 11, 25, 75\}$. Figure 20 compares the quality of the Fourier-series approximation of the target distribution for each value of T. Since a Gaussian is infinitely supported, there will be aliasing in the Fourier representation no matter what the choice of the period is. In order to capture maximum area

under the curve, to keep the aliasing error small, and to prevent the generator from latching on to an aliased version of the target density, we choose T to encompass 12σ supports of both the generator and the target densities in the fundamental period (for example, $T \geq 6$ for the standard normal distribution). A good choice of the fundamental period T is one that is centered around the generator distribution, but also encompasses the target distribution. For the scenario where the standard normal $\mathcal{N}(0,1)$ is chosen as the noise distribution when learning a target $\mathcal{N}(\mu,\sigma)$ we observe that $T \approx \max\{6, \mu + 6\sigma\}$ results in a superior quality of the Fourier-series approximation of the target.

Figures 21(a) and (b) plot the Wasserstein-2 distance $\mathcal{W}^{2,2}$ and generator loss $\mathcal{L}_G$, respectively, as a function of iterations for various T. We observe that, for small T, the generator latches on to an aliased version of the target, resulting in a large value for $\mathcal{W}^{2,2}$, although the loss $\mathcal{L}_G$ converges to zero. Choosing a large value of T makes the distribution appear like a spike (high-frequency) in the fundamental period and therefore, an accurate representation requires a larger value of M. For large M, although the Fourier-series approximation is not accurate, the generator samples converge to the desired target samples in terms of $\mathcal{W}^{2,2}$ and $\mathcal{L}_G$ by virtue of uniqueness of the Fourier representation for a given set of samples. Figure 21(c) shows the learnt discriminator for various choices of T. For small T, the learnt discriminator is unable to classify the target and generator distributions accurately. By virtue of the truncated Fourier-series approximation, the discriminator always learns a smooth approximation of the target classifier.

Convergence of the optimal Lagrange multiplier: We next illustrate the suitability of the optimal Lagrange multiplier λ^*_{FS} to serve as a proxy to measure convergence of the GAN generator during training. Figure 22 shows λ^*_{FS} and the Wasserstein-2 distance ($\mathcal{W}^{2,2}$) between p_d and p_g as a function of iterations. We observe that, for higher learning rates ($lr \approx 10^{-1}$), λ^*_{FS} does not converge to zero, which may be attributed to the fact that the $\mathcal{W}^{2,2}$ metric measures the convergence only between the first- and second-order statistics, while λ^*_{FS} measures the coefficient-wise convergence between the Fourier-series of p_d and p_g, which indirectly measures the L_2 error between the generator and target densities. This suggests that, while the models converge in the Wasserstein-2 sense for higher learning rates, convergence in the L_2 sense occurs for lower rates (here, $lr \leq 10^{-2}$). Based on these results, we set the learning rate to 10^{-3} for the generator in the subsequent experiments.

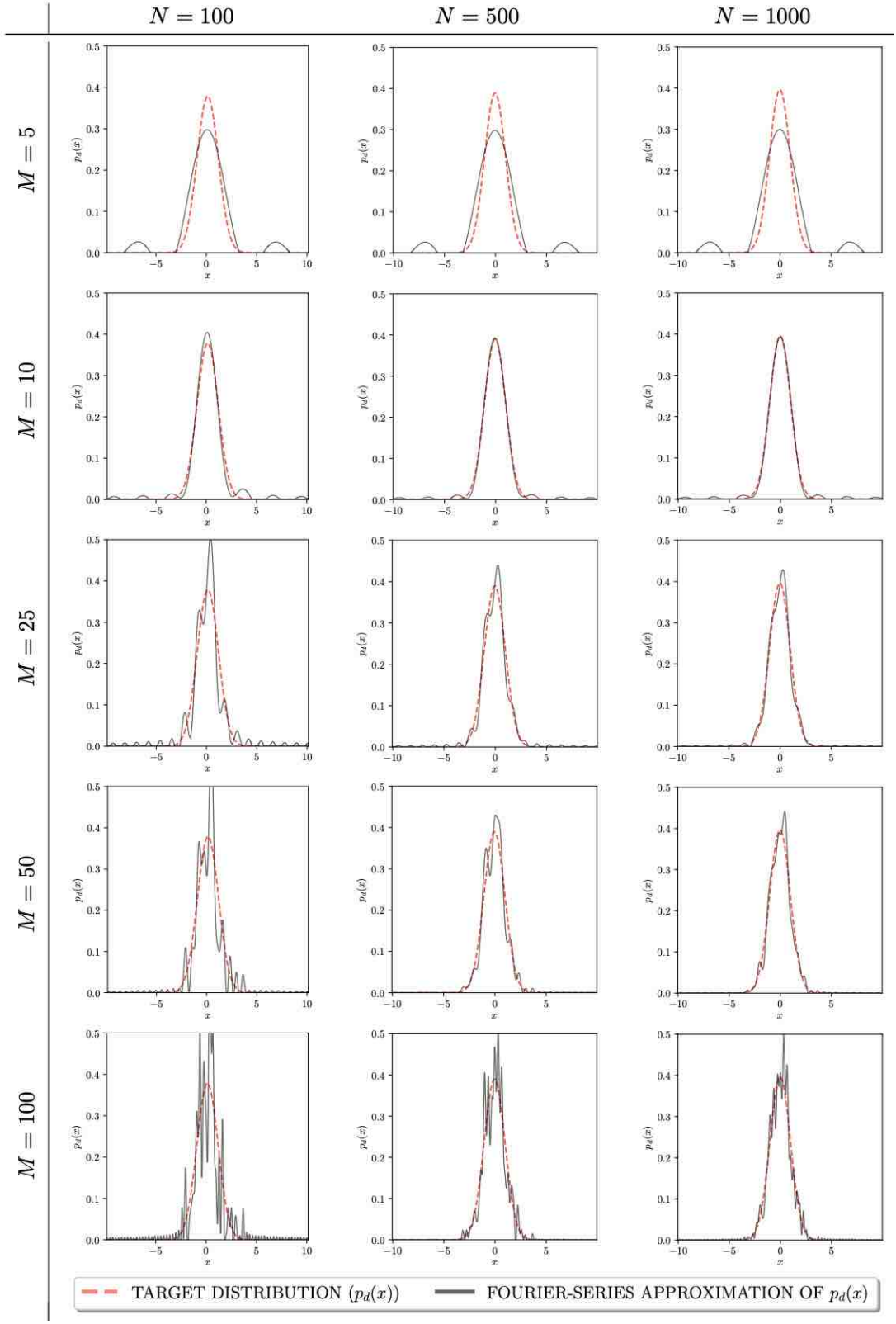


Figure 18: (Color online) Comparison of the quality of the Fourier-series approximation of a Gaussian $p_d(x)$ for various batch sizes N and truncation frequencies M.

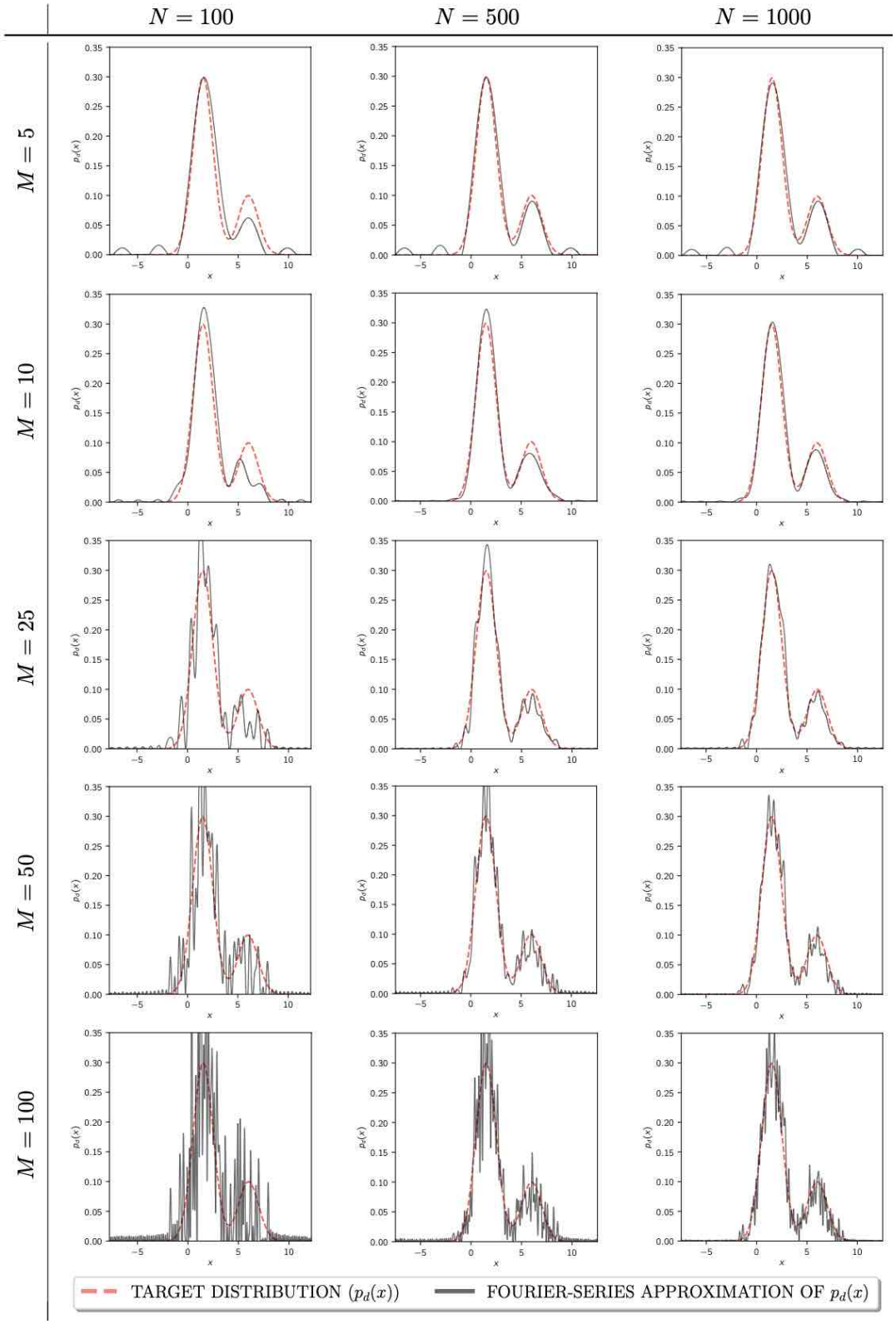


Figure 19: (Color online) Comparison of the quality of the Fourier-series approximation of a bimodal Gaussian p_d for various batch sizes N and truncation frequencies M.

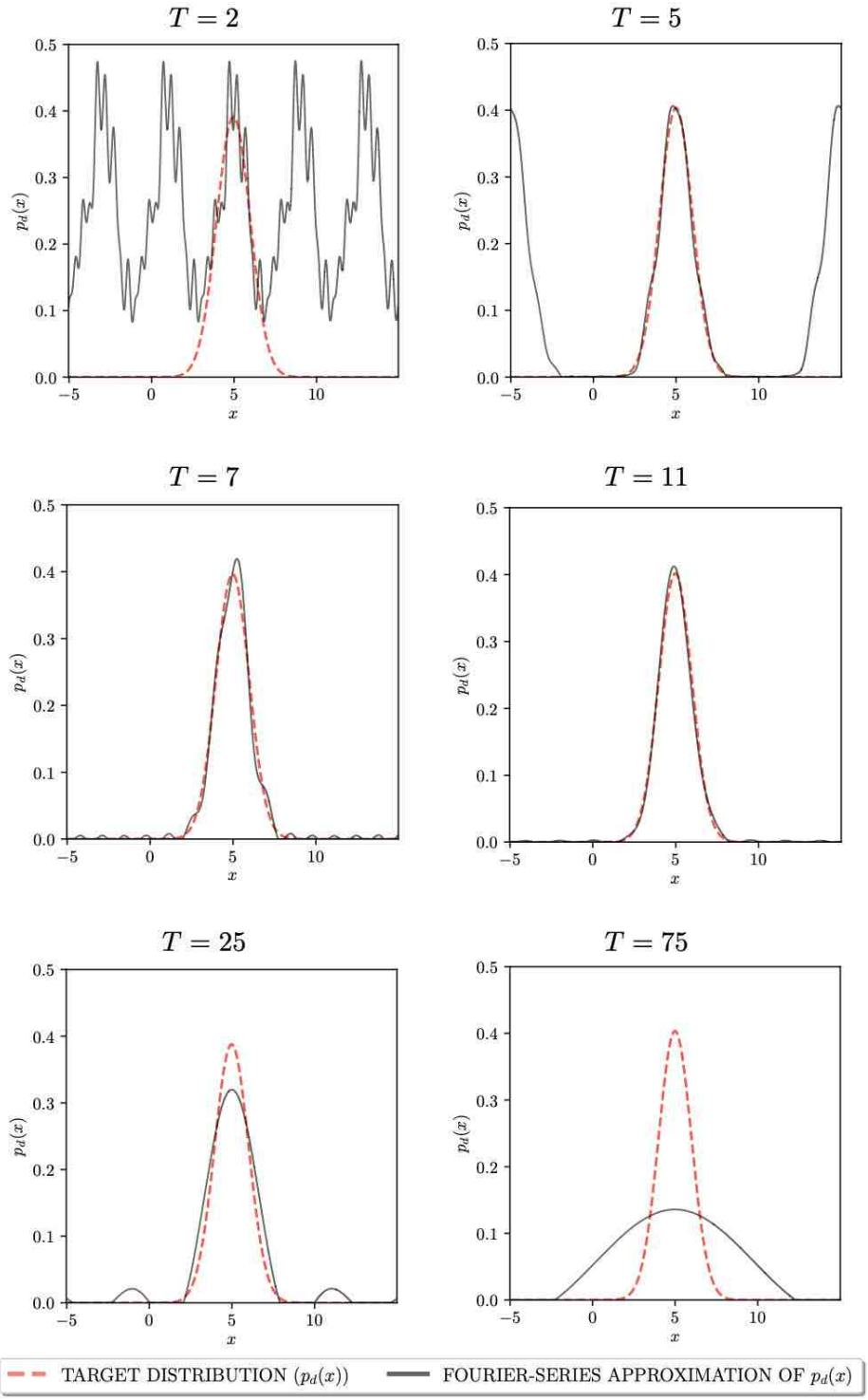


Figure 20: (Color online) Comparison of the quality of the 10-component Fourier-series approximation of a Gaussian $p_d(x)$ for various choices of the fundamental period T. Underestimating the time period results in aliasing, while overestimating it results in worse approximations of the distribution and requires additional high-frequency components in the expansion to improve upon the quality.

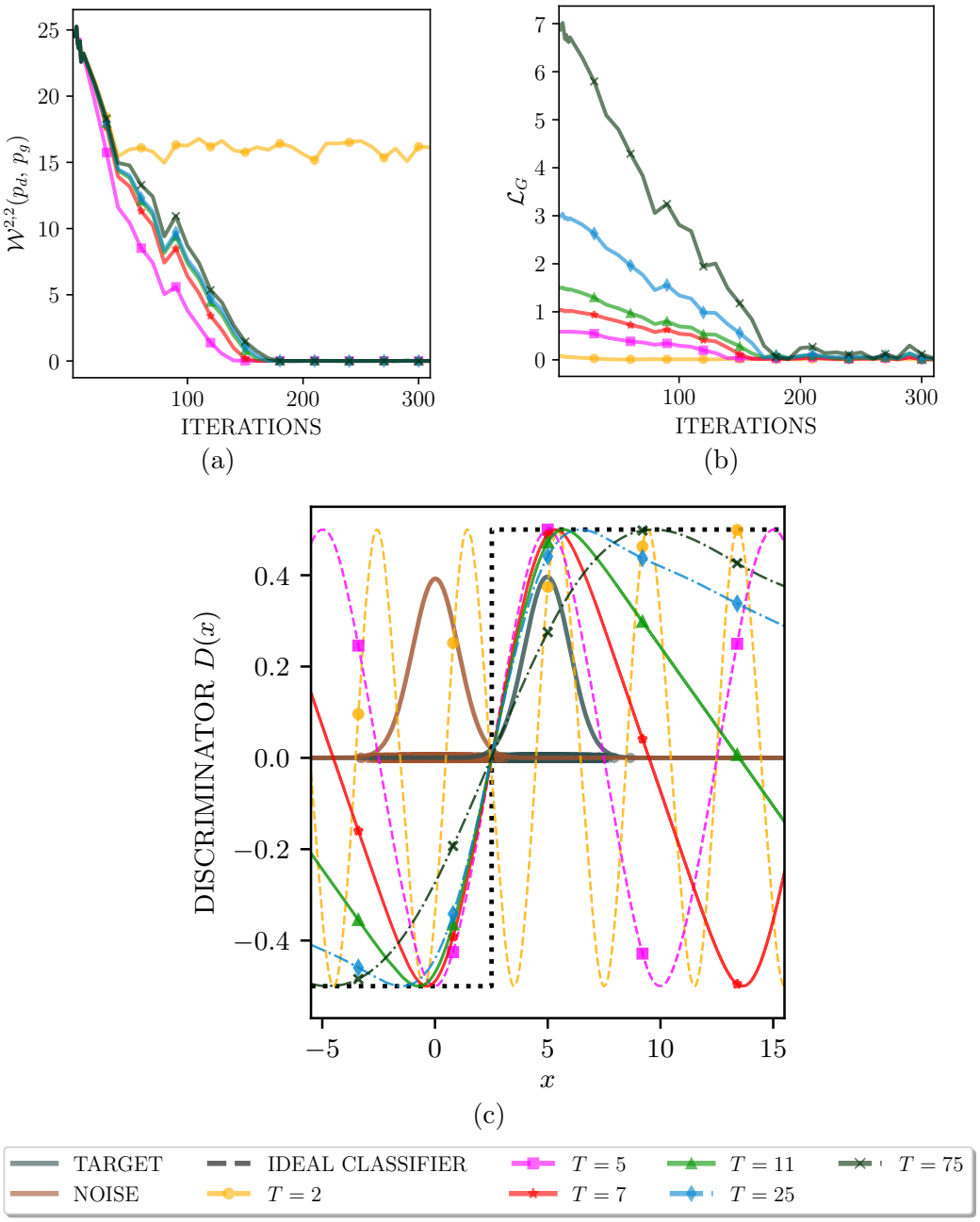


Figure 21: (Color online) Experiments on 1-D Gaussian data: Comparison of (a) Wasserstein-2 distance $\mathcal{W}^{2,2}(p_d, p_g)$; and (b) Generator loss $\mathcal{L}_G$ as a function of iterations when training WGAN-FS for various choices of T. For small T, the generator latches on to periodic replicas of the target, resulting in higher $\mathcal{W}^{2,2}$ values but low $\mathcal{L}_G$. (c) Comparison of the learnt discriminator when training WGAN-FS for various choices of T. WGAN-FS learns a smooth approximation of the true classifier for all T that contain 12σ windows of the generator and target distribution, thereby avoiding aliasing.

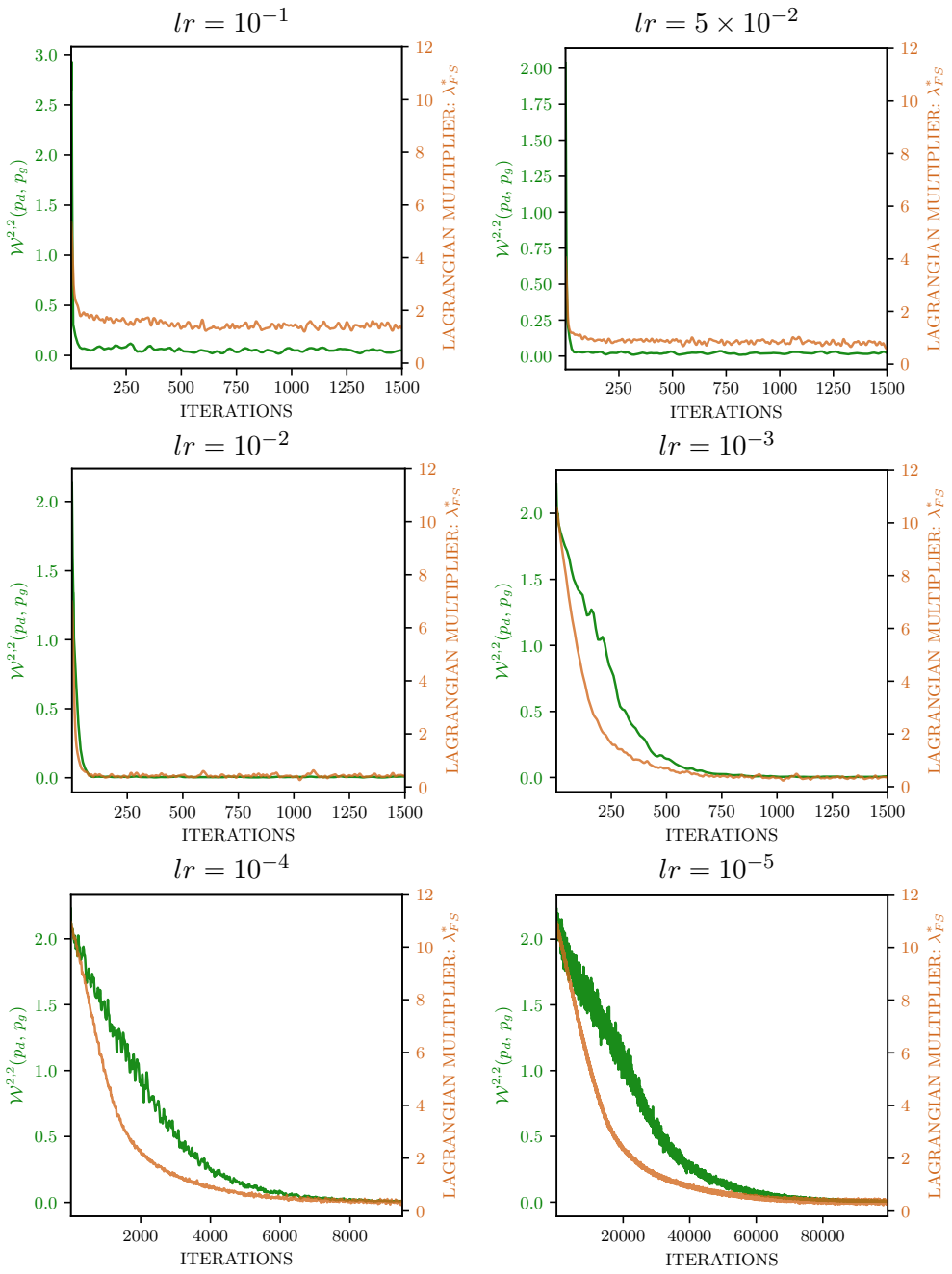


Figure 22: (Color online) Convergence of the optimal Lagrange multiplier λ^*_{FS} alongside Wasserstein-2 distance between p_d and p_g ($\mathcal{W}^{2,2}(p_d, p_g)$) for various learning rates. For higher learning rates, while the model appears to converge in the sense of $\mathcal{W}^{2,2}(p_d, p_g)$, which is a measure only up to second-order statistics, we observe from λ^*_{FS} that the distributions converge in the L_2 sense (the Fourier representation of p_g converging to that of p_d) only for learning rates lower than 10^{-2}. For very low rates (such as 10^{-5}), the convergence is not smooth. Therefore, we use learning rates in the range $[10^{-2}, 10^{-4}]$ in the subsequent experiments.

E.2 Experiments on n-dimensional Gaussians

We now present experimental results on learning multivariate Gaussian data with truncated Fourier-series expansions for WGAN-FS based on the sampling scheme described in Section 5.1.

Experimental Setup: The experiments are conducted on n-D Gaussian data drawn from $\mathcal{N}(0.75\mathbf{1}_n, 0.2\mathbb{I}_n)$, where $\mathbf{1}_n$ denotes an n-dimensional vector with all entries equal to 1, and $\mathbb{I}_n$ is the n-dimensional identity matrix. The input to the generator is 100-D Gaussian noise. To simulate the scenario of training on real-world images with the WAE Encoder (Tolstikhin et al., 2018), the noise input is provided to a fully connected layer with $32 \times 32 \times 3$ nodes, whose output is reshaped to $(32, 32, 3)$. Subsequently, the reshaped noise vectors are provided as input to a network consisting of four convolution layers with 1024, 256, 128, and 64 filters in successive layers. The output of the convolution layers is flattened and provided to a fully connected layer with n output nodes. The learning rate is set to 10^{-2}, and batch size to $N = 100$. Recall that the Fourier-series expansion consists of two levels of approximation, one for the low-frequency part and the other for the high-frequency part. We consider all harmonics up to M_{low}, and a set of L distinct uniformly drawn/sampled harmonics between M_{low} and M_{high}. We pick $10 \leq n \leq 256$ to represent different latent space dimensions used in standard autoencoder architectures for images (Tolstikhin et al., 2018).

Results: Figure 23 shows the Wasserstein-2 metric $\mathcal{W}^{2,2}$, generator loss $\mathcal{L}_G$ and Lagrange multiplier λ^*_{FS} as a function of iterations, when training WGAN-FS to learn 10-D Gaussian data. We set $M_{low} = 2$ and $M_{high} = 10$. We experiment on multiple choices of the sample size: $L \in \{5, 10, 20, 100, 500, 1000, 10000, 25000\}$. We observe from Figure 23(a) that the model converges faster for smaller L (for example $L \leq 500$ in the experiments). However, as seen in Figure 23(b), for small L, the value of $\mathcal{L}_G$ is higher. From Figure 23(c), we see that for large L (such as $L > 10^3$), the convergence of the model in terms of λ^*_{FS} is slower. We attribute this to the slower convergence of the high-frequency components in the Fourier-series expansions due to increased variance in estimating these components for a given batch size N. This disparity is more pronounced when λ^*_{FS} is plotted on the logarithmic scale, as seen in Figure 23(b). We therefore chose $10^2 \leq L \leq 10^4$ to be a good compromise between achieving lower values of the generator loss and faster convergence of the model. The findings were similar when training the WGAN-FS model on 64-D and 128-D Gaussians (cf. Figures 24 and 25, respectively).

In order to motivate the need for latent space matching, we compare the performance of WGAN-FS for various n, given the sampling parameters $M_{low} = 2$, $M_{high} = 10$ and $L = 1000$. From Figure 26, we observe that, as n increases, both $\mathcal{W}^{2,2}$ and λ^*_{FS} exhibit poorer convergence (saturation to higher values). There is also increased jitter in the convergence of the loss and λ^*_{FS} as n increases. From these results, it is evident that the WGAN-FS discriminator exhibits superior performance on low-dimensional distribution matching, such as in the case of latent-space matching in adversarial or Wasserstein autoencoders (Makhzani et al., 2015; Tolstikhin et al., 2018).

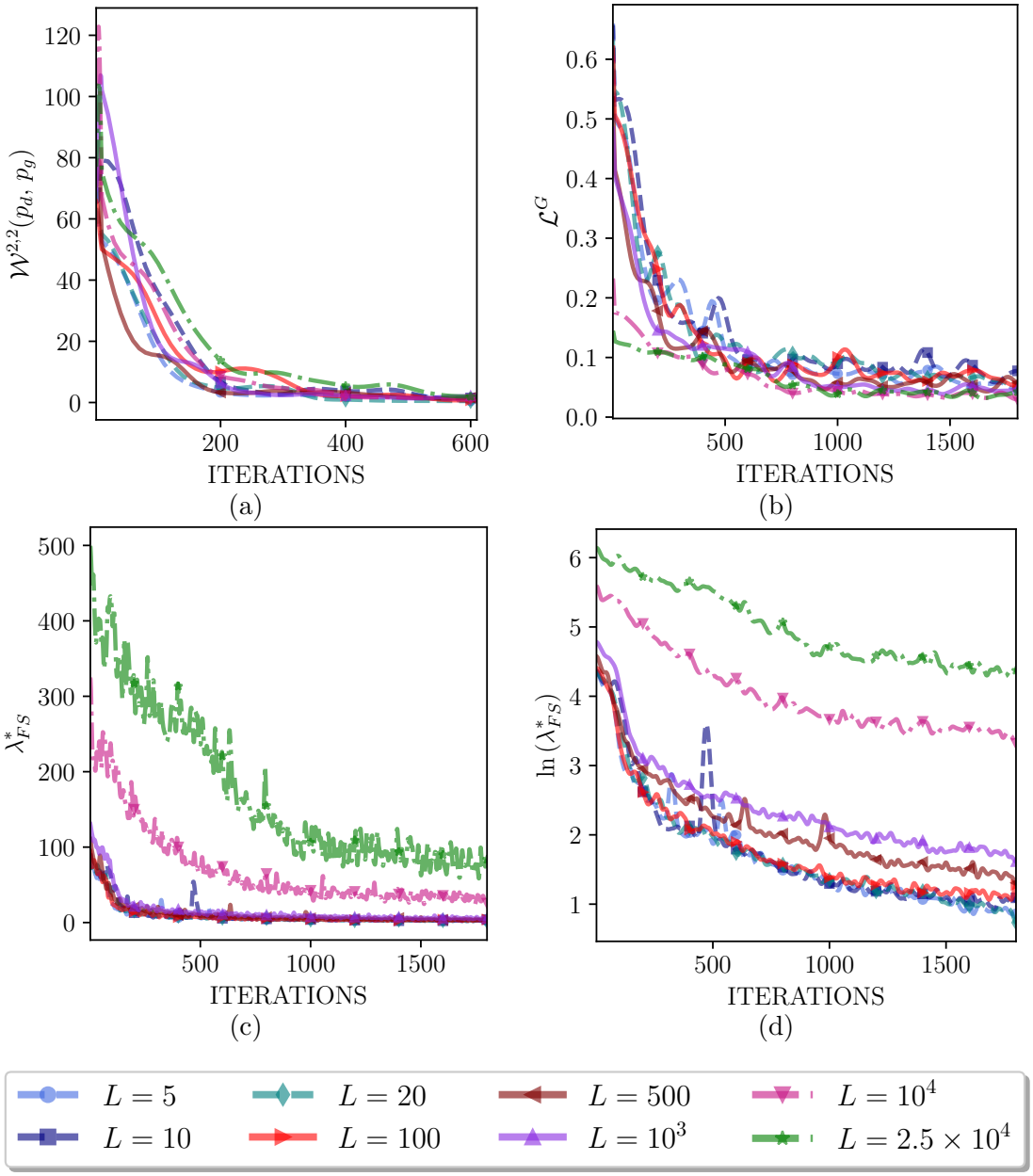


Figure 23: (Color online) Experiments on 10-D Gaussian data: Plots comparing the convergence of: (a) Wasserstein-2 distance $\mathcal{W}^{2,2}$; (b) Generator loss $\mathcal{L}_G$; (c) Optimal Lagrange multiplier λ^*_{FS}, and (d) the natural logarithm of λ^*_{FS} as a function of iterations when training WGAN-FS with L randomly sampled high-frequency components. The convergence is slower for large L as the error in estimating the coefficients increases with an increase in the number of high frequency terms.

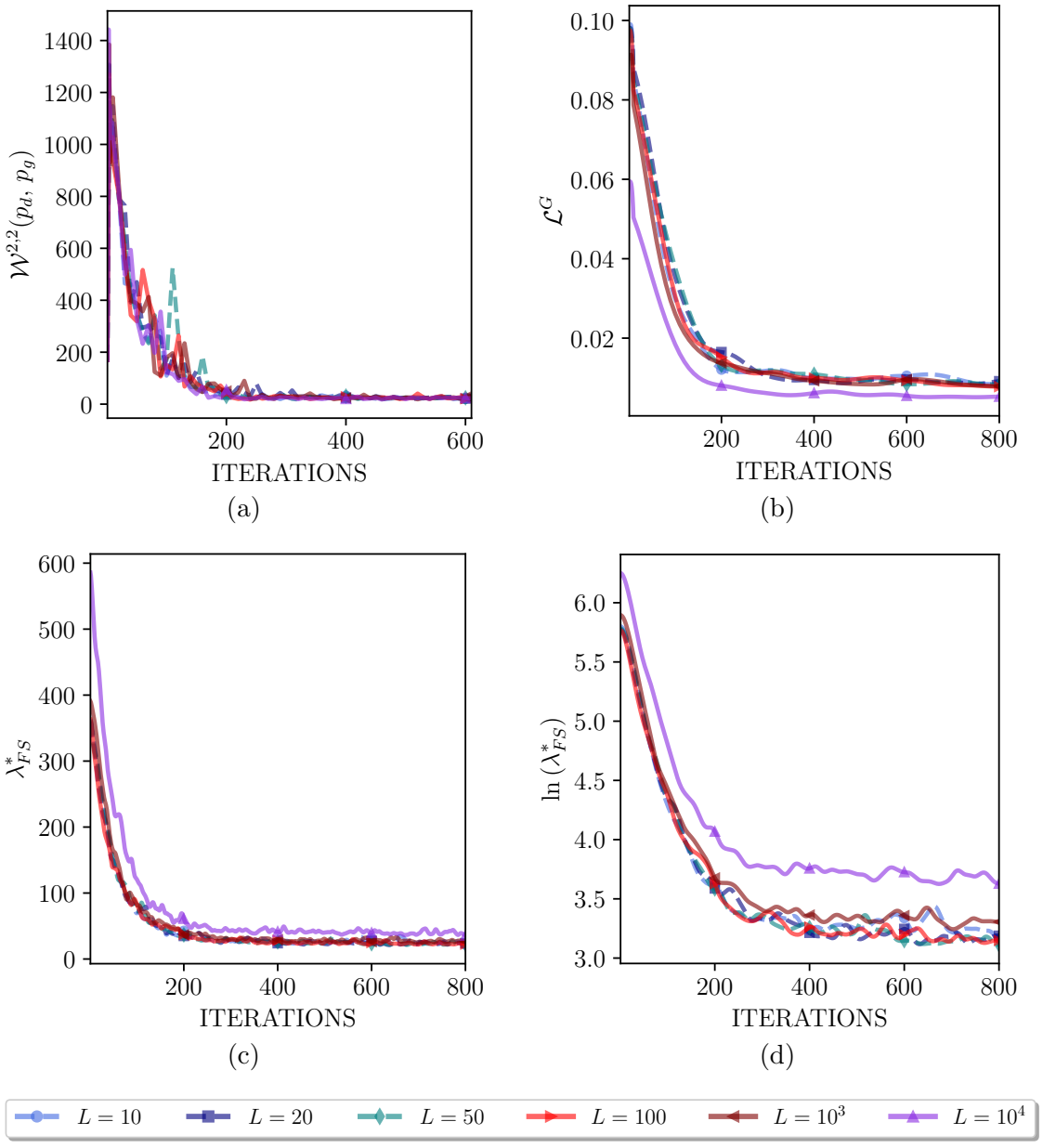


Figure 24: (Color online) Experiments on 64-D Gaussian data: Plots comparing the convergence of: (a) Wasserstein-2 distance $\mathcal{W}^{2,2}$; (b) Generator loss $\mathcal{L}_G$; (c) Optimal Lagrange multiplier λ^*_{FS}, and (d) the natural logarithm of λ^*_{FS} when training WGAN-FS on 64-dimensional Gaussian data for various number of sampled high-frequency coefficients, L. We observe that λ^*_{FS} converges to a worse (higher) value for larger L, while Wasserstein-2 distance $\mathcal{W}^{2,2}(p_d, p_g)$ and generator loss $\mathcal{L}_G$ are worse for small L. Setting L to be around 10^3 is a viable compromise.

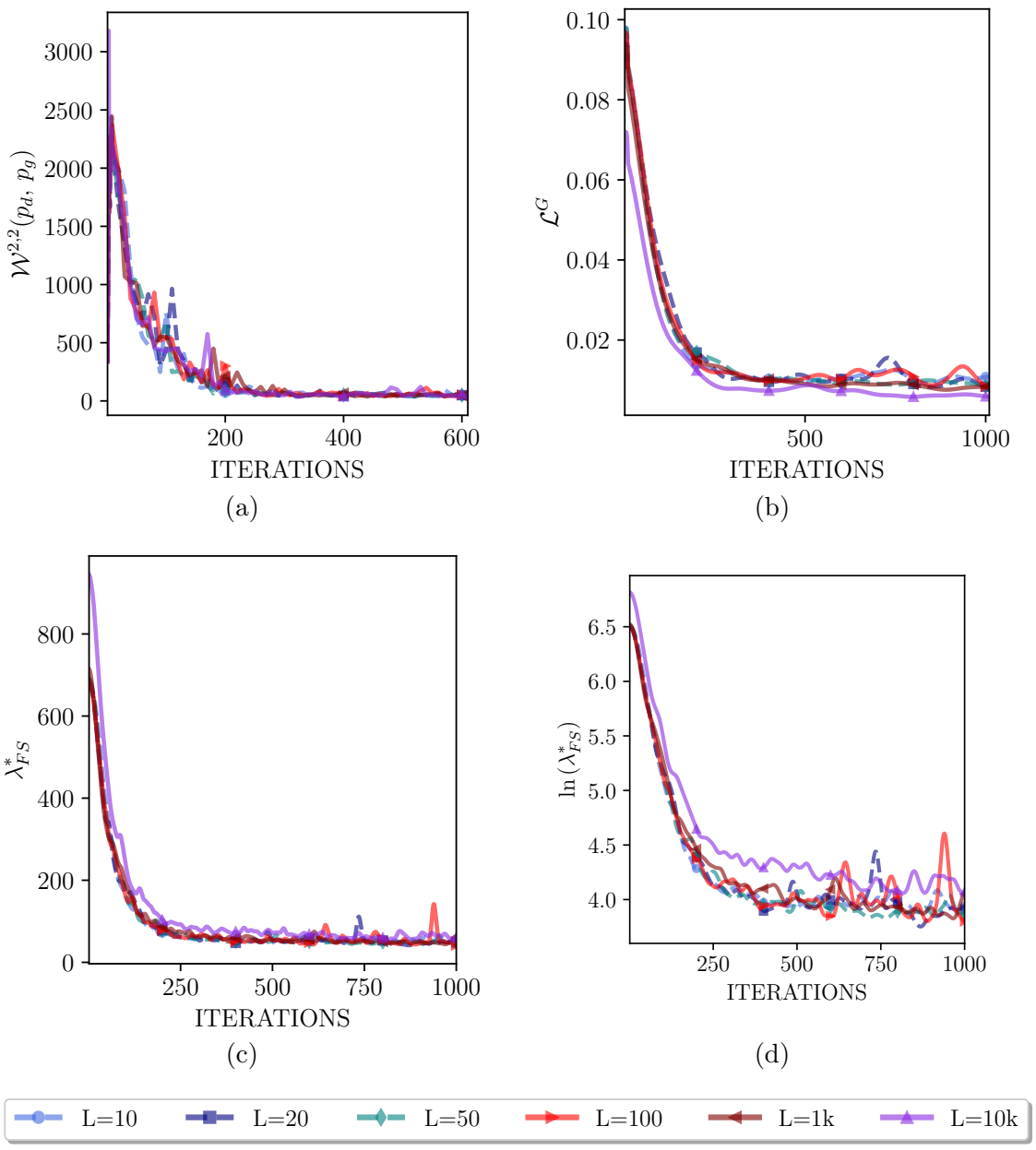


Figure 25: (Color online) Experiments on 128-D Gaussian data: Plots comparing the convergence of: (a) Wasserstein-2 distance $\mathcal{W}^{2,2}$; (b) Generator loss $\mathcal{L}_G$; (c) Optimal Lagrange multiplier λ^*_{FS}, and (d) the natural logarithm of λ^*_{FS} when training WGAN-FS on 128-dimensional Gaussian data for various number of sampled high-frequency coefficients, L. We observe that the models converge to worse (higher) values of λ^*_{FS} as L increases. This suggests that Fourier-series-based discriminator performs better when fewer high-frequency components are included in the approximation.

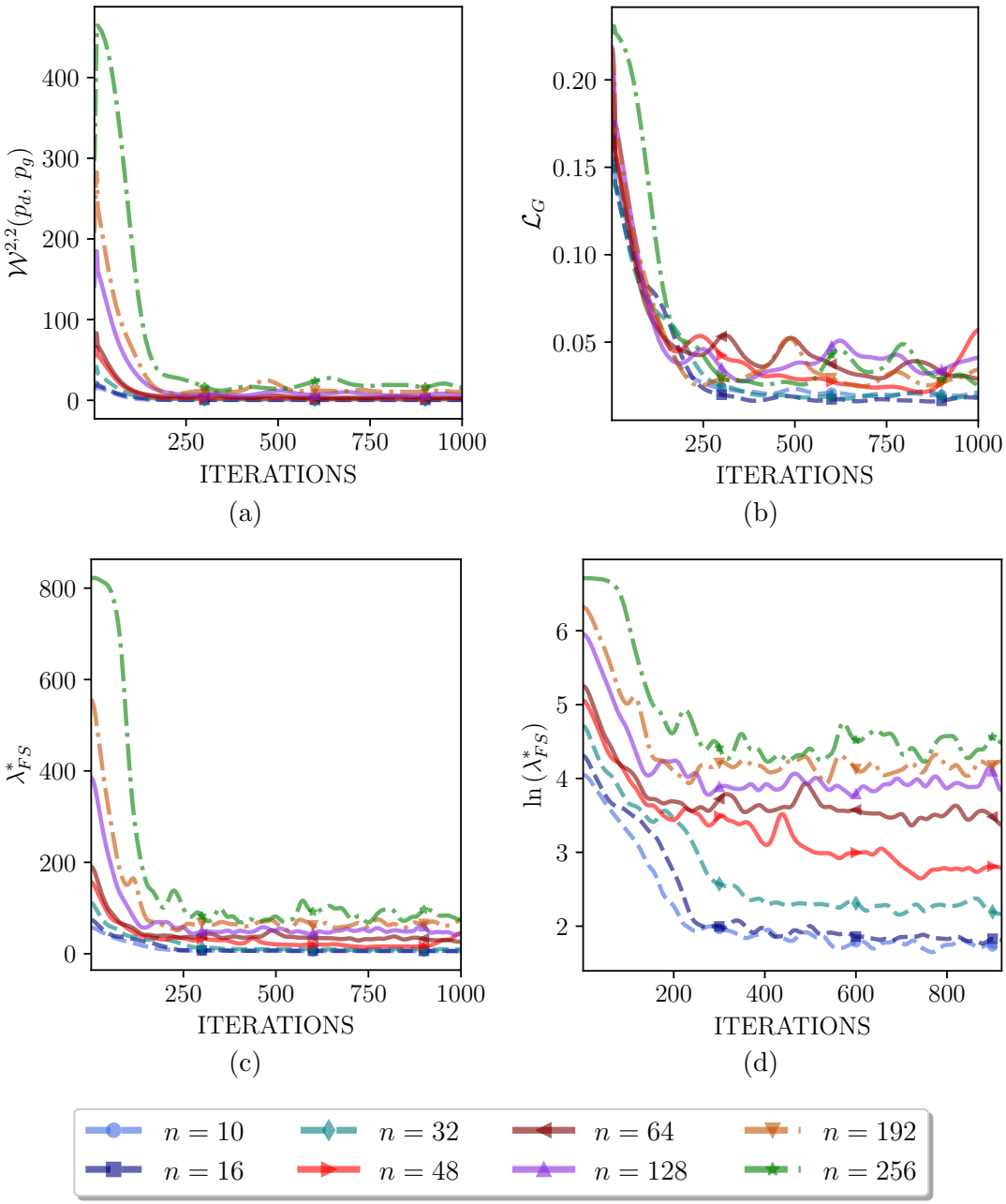
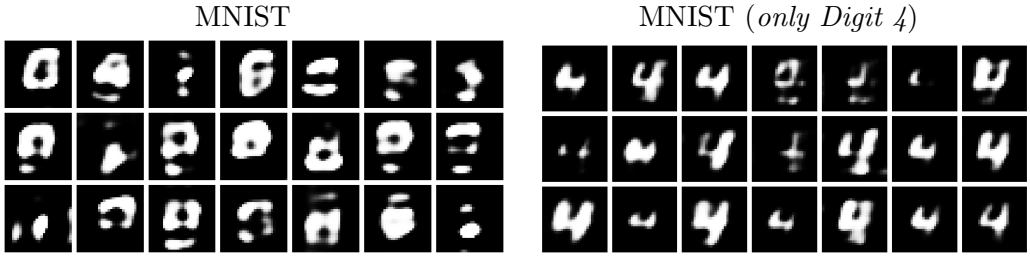


Figure 26: (Color online) Plots comparing the convergence of (a) Wasserstein-2 distance $\mathcal{W}^{2,2}(p_d, p_g)$, (b) Generator loss $\mathcal{L}_G$, (c) the optimal Lagrange multiplier λ^*_{FS}, and (d) the natural logarithm of λ^*_{FS} when training WGAN-FS on n-dimensional data, for various n. Across all three metrics, we observe that the models converge to worse (higher) values as the dimensionality of the data increases. This suggests that Fourier-series-based discriminator performs better on lower-dimensional latent-space matching.



(a) (b)

Figure 27: Images generated by training WGAN-FS on 784-dimensional data consisting of vectorized images drawn from (a) the entire MNIST dataset; and (b) only the *Digit 4* class of MNIST. WGAN-FS fails to converge when trained on the complex MNIST multimodal data due to errors in estimating the full distribution with a truncated Fourier series and a small batch size. While the performance is better in the case of single-class learning, the images are sub-par compared to the baseline GANs.

E.3 Image-space Matching with WGAN-FS

In an n-dimensional setting, the Fourier-series approximation, and thereby, the WGAN-FS approach require that the underlying distributions are at least $\lceil \frac{n}{2} \rceil$-times continuously differentiable for the truncation error derived in Appendix D.2 to be finite. However, it is widely accepted that image datasets lie in unions of low-dimensional manifolds in a higher dimensional space (Lui et al., 2017) resulting in a multimodal p_d. This can lead to poorer estimates of the Fourier-series coefficients of p_d and p_g, and the WGAN-FS generator will not learn the data distribution accurately. To validate this, we trained the WGAN-FS model on vectorized MNIST data, with $\boldsymbol{x} \in \mathbb{R}^{784}$. The generator consists of a 3-layer fully connected network with 128, 256, and 512 nodes, in successive layers with the $hyperbolic-tan$ activation. The output layer consists of 784 nodes with a *sigmoid* activation and the input to the network is a 100-dimensional Gaussian vector. The fundamental period of the Fourier series is set to 2. The sampling scheme described in Appendix E.2 is used, with $M_{low} = 2$, $M_{high} = 10$ and $L = 1000$. The model is trained with the Adam (Kingma and Ba, 2015) optimizer with a learning rate of 10^{-3} and a batch size of 250.

Figure 27(a) shows the images generated by WGAN-FS when trained with the entire MNIST dataset. We observe that the model has failed to learn the data distribution accurately. Instead, it learns only the average statistics (such as the mean digit) of the dataset. Figure 27(b) shows that the model performs better on a single-class learning task, where the data distribution is more structured. However, the visual quality of the generated images is below par than those generated by the baseline GAN variants (Arjovsky et al., 2017; Gulrajani et al., 2017; Terjék, 2020). We attribute this poor performance to inaccuracies in estimating the Fourier-series coefficients from a small batch size and insufficient harmonics, given that the ambient dimension of the data is 784. We also note that the WGAN-FS model did not converge when trained on 3072-dimensional CIFAR-10 or SVHN images or other high-resolution datasets. These results further motivate the need for training the Fourier-series model on latent space representations of the data.

E.4 Additional Details on Evaluation Metrics

In this appendix, we provide additional details on the evaluation metrics used in Sections 3.7, 4.3, and 6 of the main manuscript:

- **Wasserstein-2 distance:** For Gaussian generator and target distributions, the Wasserstein-2 distance has a closed-form expression:

$$\mathcal{W}^{2,2}\left(\mathcal{N}(\mu_d,\Sigma_d),\mathcal{N}(\mu_g,\Sigma_g)\right) = \|\mu_d-\mu_g\|_2^2 + \text{Trace}\left(\Sigma_d+\Sigma_g-2\left(\Sigma_d^{\frac{1}{2}}\Sigma_g\Sigma_d^{\frac{1}{2}}\right)^{\frac{1}{2}}\right).$$

- **Fréchet Inception Distance (FID):** The FID was introduced by Heusel et al. (2017) to measure the visual quality of images generated by GANs. FID is highly correlated with human evaluation of such images. In this setting, we first consider the InceptionV3 model (Szegedy et al., 2015) without the topmost layer, loaded with pre-trained weights for ImageNet (Deng et al., 2009) classification, to generate embeddings of the real and generated data. Next, we assume that the embeddings of the real and generated data are distributed as $\mathcal{N}(\mu_d,\Sigma_d)$ and $\mathcal{N}(\mu_g,\Sigma_g)$, respectively, and compute FID as the Fréchet distance between them:

$$\mathcal{F}_r\left(\mathcal{N}(\mu_d,\Sigma_d),\mathcal{N}(\mu_g,\Sigma_g)\right) = \|\mu_d-\mu_g\|_2^2 + \text{Trace}\left(\Sigma_d+\Sigma_g-2\left(\Sigma_d\Sigma_g\right)^{\frac{1}{2}}\right).$$

  The InceptionV3 model accepts input dimensions in the range of $76\times76\times3$ to $229\times229\times3$. For consistency with the literature, color images are upscaled to $229\times229\times3$ using bilinear interpolation. Grayscale images are upscaled to 229×229 and replicated across the color channels. The FID score is measured over batches of 10^4 images. The best-case FID scores of the converged models are measured using 5×10^4 samples drawn from both the target dataset and the WAE in all cases except Ukiyo-E and single-class CIFAR-10, where the entire target class (about 5×10^3 images) is used. We use the publicly available TensorFlow implementation of *clean-fid* (Parmar et al., 2021) to compute the metric. Our implementation of CWAE (Knop et al., 2020) and WAE (Tolstikhin et al., 2018) produced FID scores that are consistent with the literature on CIFAR-10 and CelebA datasets. FID scores for MNIST, SVHN and Ukiyo-E datasets were not reported in the baselines.

- **Image Sharpness**: We employ the metric proposed by Tolstikhin et al. (2018) to characterize image sharpness. The sharpness metric is measured on two sets of data: (i) Images obtained by decoding sample vectors drawn from the prior distribution; and (ii) Images obtained by decoding interpolated latent vectors. The test images are rescaled to have pixel intensities in $[0,1]$, and convolved with the Laplacian kernel $\begin{bmatrix}-1&-1&-1\\-1&8&-1\\-1&-1&-1\end{bmatrix}$ to emphasize edges. The variance of the pixel intensities in the Laplacian of the image is evaluated and averaged over a batch of 10^3 images to measure sharpness. Blurred images possess fewer distinct edges thereby resulting in a lower variance than sharper images.

Appendix F. Other Gradient-Regularized GANs

In this appendix, we consider Wasserstein GAN with the gradient penalty (WGAN-GP) (Gulrajani et al., 2017) and the centered WGAN-R$_d$ and WGAN-R$_g$ gradient penalties (Mescheder et al., 2018) within the Euler-Lagrange framework. We also consider the SGAN and LSGAN subject to the considered gradient-norm penalty, resulting in SGAN-GNP and LSGAN-GNP, respectively.

WGAN-GP: Consider the WGAN-GP discriminator loss given by

$$\begin{aligned}\mathcal{L}_D &= -\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] + \lambda\,\mathbb{E}_{\boldsymbol{x}\sim \alpha p_g+(1-\alpha)p_d}\left[(\|\nabla D(\boldsymbol{x})\|_2-1)^2\right]\\ &= \int_{\mathcal{X}} \left(D(\boldsymbol{x})\,(p_g(\boldsymbol{x})-p_d(\boldsymbol{x})) + (p_g(\boldsymbol{x})+(1-\alpha)p_d(\boldsymbol{x}))\,(\|\nabla D(\boldsymbol{x})\|_2-1)^2\right)\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Applying the EL condition for optima yields the following condition that the optimal discriminator $D^*(\boldsymbol{x})$ must satisfy:

$$\begin{aligned}&(\alpha p_g(\boldsymbol{x})+(1-\alpha)p_d(\boldsymbol{x}))\left(1-\|\nabla D(\boldsymbol{x})\|^{-1}\right)\Delta D(\boldsymbol{x}) + \left(\frac{p_d(\boldsymbol{x})-p_g(\boldsymbol{x})}{2\lambda}\right)\\ &\quad+\left(\alpha\langle\nabla p_g(\boldsymbol{x}),\nabla D(\boldsymbol{x})\rangle + (1-\alpha)\langle\nabla p_d(\boldsymbol{x}),\nabla D(\boldsymbol{x})\rangle\right)\left(1-\|\nabla D(\boldsymbol{x})\|^{-1}\right)\\ &\quad+\left.(\alpha p_g(\boldsymbol{x})+(1-\alpha)p_d(\boldsymbol{x}))\,\|\nabla D(\boldsymbol{x})\|^{-3}\left\langle(\nabla D(\boldsymbol{x})).^2,\mathrm{diag}(\mathbb{H}_D)\right\rangle\right|_{D=D^*}=0,\end{aligned}$$

where

$$\nabla D = [D'_1, D'_2, \ldots, D'_n]^{\mathrm{T}}, \quad \text{with} \quad D'_i = \frac{\partial D}{\partial x_i}$$

is the gradient vector, $(\nabla D).^2$ represents element-wise squaring, and

$$\mathrm{diag}(\mathbb{H}_D) = [D''_{11}, D''_{22}, \ldots, D''_{nn}]^{\mathrm{T}}, \quad \text{with} \quad D''_{ii} = \frac{\partial^2 D}{\partial x_i^2}$$

is the vector formed by the diagonals of the Hessian matrix of D. The result is an intractable, non-linear, second-order differential equation.

WGAN-R_dR_g: The WGAN-R$_d$R$_g$ loss presented by Mescheder et al. (2018) is:

$$\begin{aligned}\mathcal{L}_D &= -\mathbb{E}_{\boldsymbol{x}\sim p_d}[D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[D(\boldsymbol{x})] + \frac{\lambda_1}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_d}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right] + \frac{\lambda_2}{2}\,\mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\|\nabla D(\boldsymbol{x})\|_2^2\right]\\ &= \int_{\mathcal{X}}\left(D(\boldsymbol{x})\,(p_g(\boldsymbol{x})-p_d(\boldsymbol{x})) + \frac{1}{2}\,(\lambda_1 p_d(\boldsymbol{x})+\lambda_2 p_g(\boldsymbol{x}))\,\|\nabla D(\boldsymbol{x})\|_2^2\right)\,\mathrm{d}\boldsymbol{x},\end{aligned}$$

where the EL condition for optimality results in the following differential equation:

$$\Delta D + \frac{\langle\lambda_1\nabla p_d + \lambda_2\nabla p_g, \nabla D\rangle}{\lambda_1 p_d + \lambda_2 p_g} = \frac{p_g - p_d}{\lambda_1 p_d + \lambda_2 p_g}.$$

Unlike WGAN-GP, this formulation results in a second-order elliptic differential equation with a variable coefficient. This is an instance of the Stein operator considered by Mroueh

et al. (2018), where the gradient penalty is evaluated with respect to the measure $\lambda_1 p_d + \lambda_2 p_g$. Fourier transform based techniques could be employed to solve the PDE.

SGAN-GNP: Based on a finding that the gradient penalty improves the performance of WGANs (Gulrajani et al., 2017), the same penalty was used to improve the performance of several f-GAN variants by Roth et al. (2017). The expectations are evaluated with respect to an interpolated distribution between $p_d(\boldsymbol{x})$ and $p_g(\boldsymbol{x})$ by Roth et al. (2017) and between $p_d(\boldsymbol{x})$ and the standard Gaussian by Kodali et al. (2017).

We consider the application of the gradient-norm penalty introduced in the context of WGANs in this paper to SGAN and LSGAN. Consider the optimization of the SGAN discriminator in 1-D with the gradient-norm penalty:

$$\min_{D(x)} \mathcal{L}_{\mathrm{D}}^{\mathrm{SGAN}} \text{ s.t. } \int_{\mathcal{X}} \left(|D'(x)|^2 - 1\right) \, \mathrm{d}x = 0.$$

The integrand in the Lagrangian of the constrained discriminator loss is

$$\mathcal{F}\left(x, D, D'\right) = \ln\left(D(x)\right) p_d(x) + \ln\left(1 - D(x)\right) p_g(x) + \lambda_d |D'(x)|^2.$$

Following the Euler-Lagrange condition, the optimal discriminator must satisfy the following differential equation:

$$D''(x) = \frac{p_d(x) - \left(p_d(x) + p_g(x)\right) D(x)}{2\lambda_d D(x)(1 - D(x))}. \tag{65}$$

A similar solution can be obtained in the multidimensional case as well, where the SGAN optimization problem becomes

$$\min_{D(\boldsymbol{x})} \mathcal{L}_{\mathrm{D}}^{\mathrm{SGAN}} \text{ s.t. } \int_{\mathcal{X}} \left(\|\nabla D(\boldsymbol{x})\|_2^2 - 1\right) \, \mathrm{d}\boldsymbol{x} = 0.$$

The Euler-Lagrange condition gives the second-order non-linear PDE:

$$\Delta D(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x}) - \left(p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})\right) D(\boldsymbol{x})}{2\lambda_d D(\boldsymbol{x})(1 - D(\boldsymbol{x}))}. \tag{66}$$

Equations (65) and (66) are not amenable to a closed-form solution. A practical alternative would be to use a numerical solver.

LSGAN-GNP: Finally, consider LSGAN with the gradient-norm penalty in 1-D:

$$\min_{D(x)} \mathcal{L}_{\mathrm{D}}^{\mathrm{LSGAN}} \text{ s.t. } \int_{\mathcal{X}} \left(|D'(x)|^2 - 1\right) \, \mathrm{d}x = 0.$$

The integrand in the Lagrangian of the constrained discriminator loss is

$$\mathcal{F}\left(x, D, D'\right) = (D(x) - b)^2 p_d + (D(x) - a)^2 p_g + \lambda_d |D'(x)|^2.$$

Applying the Euler-Lagrange condition gives

$$-\underbrace{D''(x)}_{\mathrm{T}_1}+\underbrace{\left(\frac{p_g(x)+p_d(x)}{\lambda_d}\right)D(x)}_{\mathrm{T}_2}=\underbrace{\frac{ap_g(x)+bp_d(x)}{\lambda_d}}_{\mathrm{T}_3}. \tag{67}$$

The Fourier-series expansions of $p_g(x)$, $p_d(x)$, and $D(x)$ as defined in Section 3.4 simplify T_1 and T_3 readily, whereas T_2 could be simplified using the convolution property:

$$\mathrm{T}_2=\frac{1}{\lambda_d}\sum_{n=-\infty}^{\infty}\chi_n e^{\mathrm{j}\omega_o nx},\quad \text{where}\quad \chi_n=\sum_{\ell=-\infty}^{\infty}\gamma_\ell(\alpha_{n-\ell}+\beta_{n-\ell}).$$

The sequence $\{\chi_n\}$ is a sum of two convolutions, one between the sequences $\{\alpha_\ell\}$ and $\{\gamma_\ell\}$, and the other between $\{\beta_\ell\}$ and $\{\gamma_\ell\}$. The sequence $\{\gamma_\ell\}$ has to be determined in order to arrive at the discriminator. Simplifying (67) in view of the Fourier-series representations gives the following optimality conditions in terms of the Fourier coefficients:

$$\lambda_d\omega_o^2n^2\gamma_n-\big(a\alpha_n+b\beta_n\big)+\sum_{\ell=-\infty}^{\infty}\gamma_\ell(\alpha_{n-\ell}+\beta_{n-\ell})=0,\ \forall\ n\in\mathbb{Z}-\{0\}, \tag{68}$$

which is an infinite system of linear equations. One approach may be to consider a truncated Fourier-series expansion, which would give rise to a finite system of linear equations that can be solved using iterative algorithms or the Moore-Penrose pseudo-inverse.

Let us now consider the discriminator loss in n dimensions:

$$\min_{D(\boldsymbol{x})}\ \mathcal{L}_{\mathrm{D}}^{\mathrm{LSGAN}}\ \text{s.t.}\ \int_{\mathcal{X}}\big(\|\nabla D(\boldsymbol{x})\|_2^2-1\big)\ \mathrm{d}\boldsymbol{x}=0.$$

The EL condition applied to the Lagrangian of the discriminator loss results in the following:

$$-\Delta D(\boldsymbol{x})+\left(\frac{p_d(\boldsymbol{x})+p_g(\boldsymbol{x})}{\lambda_d}\right)D(\boldsymbol{x})=\frac{ap_g(\boldsymbol{x})+bp_d(\boldsymbol{x})}{\lambda_d}.$$

The above PDE is of the form $-\Delta D(\boldsymbol{x})+\beta_1(\boldsymbol{x})D(\boldsymbol{x})=\beta_0(\boldsymbol{x})$. As in the 1-D case, one could consider Fourier-series expansions and take advantage of the multidimensional convolution property. However, the computational complexity would increase exponentially with n.

References

M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint, arXiv:1603.04467*, Mar. 2016. URL `https://arxiv.org/abs/1603.04467`.

M. Abramowitz. *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Dover Publications, Inc., 1974.

J. Adler and S. Lunz. Banach Wasserstein GAN. In *Advances in Neural Information Processing Systems 31*, pages 6754–6763. 2018.

D. An, Y. Guo, N. Lei, Z. Luo, S.-T. Yau, and X. Gu. AE-OT: A new generative model based on extended semi-discrete optimal transport. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

M. Arbel, D. Sutherland, M. Bińkowski, and A. Gretton. On gradient regularizers for MMD GANs. In *Advances in Neural Information Processing Systems 31*, pages 6700–6710, 2018.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

H. Bateman and A. Erdelyi. *Higher Transcendental Functions*. McGraw-Hill, New York, 1953.

D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

O. Bousquet, S. Gelly, I. Tolstikhin, C. J. Simon-Gabriel, and B. Schoelkopf. From optimal transport to generative modeling: the VEGAN cookbook. *arXiv preprint, arXiv:1705.07642*, May 2017. URL `https://arxiv.org/abs/1705.07642`.

T. Che et al. Your GAN is secretly an energy-based model and you should use discriminator driven latent sampling. *arXiv preprint, arXiv:2003.06060*, June 2020. URL `https://arxiv.org/abs/2003.06060`.

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29*, pages 2180–2188, 2016.

R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert-W function. In *Advances in Computational Mathematics*, pages 329–359, 1996.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

I. Deshpande, Z. Zhang, and A. Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 2010.

W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

C. Fefferman. Pointwise convergence of Fourier series. *Annals of Mathematics*, 98(3):551–571, 1973. ISSN 0003486X. URL `http://www.jstor.org/stable/1970917`.

J. Ferguson. A brief survey of the history of the calculus of variations and its applications. *arXiv preprint, arXiv:math/0402357*, Feb. 2004. URL `arxiv.org/abs/math/0402357`.

Fermi-LAT Collaboration. A gamma-ray determination of the universe's star formation history. *Science*, 362:1031–1034, 2018.

A. Feuerverger and R. A. Mureika. The empirical characteristic function and its applications. *The Annals of Statistics*, 5(1):88 – 97, 1977.

C. Finn, P. F. Christiano, P. Abbeel, and S. Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint, arXiv:1611.03852*, Nov. 2016. URL `https://arxiv.org/abs/1611.03852`.

R. Flamary et al. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

J. B. J. Fourier. Mémoire sur la propagation de la chaleur dans les corps solides. *Présenté le 21 décembre 1807 à l'Institut national - Nouveau Bulletin des sciences par la Sociétéphilomatique de Paris*, pages 215–221, 1807.

J.-Y. Franceschi, E. de Bézenac, I. Ayed, M. Chen, S. Lamprier, and P. Gallinari. A neural tangent kernel perspective of GANs. *arXiv preprint, arXiv:2106.05566*, abs/2106.05566, June 2021. URL `https://arxiv.org/abs/2106.05566`.

J. Gao and H. Tembine. Bregman learning for generative adversarial networks. In *Proceedings of the Chinese Control And Decision Conference*, pages 82–89, 2018.

I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Prentice-Hall, 1964.

I. M. Gelfand and G. E. Shilov. *Generalized Functions, Vol. 1: Properties and Operations*. American Mathematical Society, 1958.

A. Genevay, G. Peyre, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617. PMLR, 2018.

C. Giardina and P. Chirlian. Bounds on the truncation error of periodic signals. *IEEE Transactions on Circuit Theory*, 19(2):206–207, 1972.

H. H. Goldstine. *A History of the Calculus of Variations from the 17th Through the 19th Century*. Springer, New York, 1980.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

P. Grnarova, K. Y. Levy, A. Lucchi, T. Hofmann, and A. Krause. An online learning approach to generative adversarial networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, pages 5767–5777. 2017.

M. E. Gurtin. Variational principles for linear elastodynamics. *Archive for Rational Mechanics and Analysis*, 16(1):34–50, 1964a.

M. E. Gurtin. Variational principles for linear initial-value problems. *Quarterly of Applied Mathematics*, 22(3):252–256, 1964b.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems 30*, volume 30, 2017.

G. E. Hinton and R. Zemel. Autoencoders, minimum description length and Helmholtz free energy. In *Advances in Neural Information Processing Systems 6*, volume 6. Morgan-Kaufmann, 1994.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29*, pages 4565–4573. 2016.

A. Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard GAN. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

J. L. Kelley. *General Topology*. Courier Dover Publications, Inc., 2017.

M. Khayatkhoei, M. K. Singh, and A. Elgammal. Disconnected manifold learning for generative adversarial networks. In *Advances in Neural Information Processing Systems 32*, volume 31, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

S. Knop, P. Spurek, J. Tabor, I. Podolak, M. Mazur, and S. Jastrzębski. Cramér-Wold autoencoder. *Journal of Machine Learning Research*, 21(164):1–28, 2020. URL `http://jmlr.org/papers/v21/19-560.html`.

N. Kodali, J. D. Abernethy, J. Hays, and Z. Kira. On convergence and stability of GANs. *arXiv preprint, arXiv:1705.07215*, May 2017. URL `http://arxiv.org/abs/1705.07215`.

S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced Wasserstein auto-encoders. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

A. Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009. URL `https://ci.nii.ac.jp/naid/20001706980/en/`.

J. H. Lambert. Observationes variae in mathesin puram. *Acta Helveticae physico-mathematico-anatomico-botanico-medica*, pages 128–168, 1758.

N. S. Landkof. *Foundations of Modern Potential Theory*. Springer-Verlag, Berlin, New York, 1972.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

N. Lei, Y. Guo, D. An, X. Qi, Z. Luo, S. T. Yau, and X. Gu. Mode collapse and regularity of optimal transportation maps. *arXiv preprints, arXiv:1902.02934*, Feb. 2019. URL `https://arxiv.org/abs/1902.02934`.

C. L. Li, W. C. Chang, Y. Cheng, Y. Yang, and B. Poczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems 30*, pages 2203–2213. 2017.

J. Li, A. Madry, J. Peebles, and L. Schmidt. On the limitations of first-order approximation in GAN dynamics. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3005–3013, 2018.

T. Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021. URL `jmlr.org/papers/v22/20-911.html`.

H. Liu, Y. Guo, N. Lei, Z. Shu, S. T. Yau, D. Samaras, and X. Gu. Latent space optimal transport for generative models. *arXiv preprint, arXiv:1809.05964*, abs/1809.05964, Sep. 2018. URL `https://arxiv.org/abs/1809.05964`.

S. Liu, S. Bousquet, and K. Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, 2015.

K. Y.-C. Lui, Y. Cao, M. Gazeau, and K. S. Zhang. Implicit manifold learning on generative adversarial networks. *The International Conference on Machine Learning Workshop on Implicit Models*, 2017.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprints, arXiv:1511.05644*, Nov. 2015. URL `https://arxiv.org/abs/1511.05644`.

X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *Proceedings of International Conference on Computer Vision*, 2017.

L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3481–3490, 2018.

M. Mesterton-Gibbons. *A Primer on the Calculus of Variations and Optimal Control Theory*. American Mathematical Society, 2009.

S. Mohamed and B. Lakshminarayanan. Learning in implicit generative models. *arXiv preprints, arXiv:1610.03483*, Oct. 2016. URL `https://arxiv.org/abs/1610.03483`.

Y. Mroueh and T. Sercu. Fisher GAN. In *Advances in Neural Information Processing Systems 30*, pages 2513–2523. 2017.

Y. Mroueh, C. Li, T. Sercu, A. Raj, and Y. Cheng. Sobolev GAN. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29*, pages 271–279. 2016.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

F. A. Oliehoek, R. Savani, J. Gallego, E. van der Pol, and R. Groß. Beyond local Nash equilibria for adversarial networks. In *Artificial Intelligence, BNAIC*, volume 1021, pages 73–89, 2019.

G. Parmar, R. Zhang, and J.-Y. Zhu. On buggy resizing libraries and surprising subtleties in FID calculation. *arXiv preprint, arXiv:2104.11222*, abs/2104.11222, April 2021. URL `https://arxiv.org/abs/2104.11222`.

H. Petzka, A. Fischer, and D. Lukovnikov. On the regularization of Wasserstein GANs. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

J. N. M. Pinkney and D. Adler. Resolution dependent GAN interpolation for controllable image synthesis between domains. *arXiv preprint, arXiv:2010.05334*, Oct. 2020. URL `https://arxiv.org/abs/2010.05334`.

C. Qin, Y. Wu, J. Springenberg, A. Brock, J. Donahue, T. Lillicrap, and P. Kohli. Training generative adversarial networks by solving ordinary differential equations. In *Advances in Neural Information Processing Systems 34*. 2020.

D. Randle, P. Protopapas, and D. Sondak. Unsupervised learning of solutions to differential equations with generative adversarial networks. *arXiv preprint, arXiv:2007.11133*, July 2020. URL `https://arxiv.org/abs/2007.11133`.

M. Riesz. L'intégrale de Riemann-Liouville et le problème de Cauchy. *Acta Mathematica*, 81:1–222, 1949.

K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems 30*, pages 2015–2025, 2017.

K. Roth, Y. Kilcher, and T. Hofmann. Adversarial training generalizes data-dependent spectral norm regularization. *arXiv preprint, arXiv:1906.01527*, June 2019. URL `https://arxiv.org/abs/1906.01527`.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems 29*, pages 2234–2242, 2016.

M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee. On the convergence and robustness of training GANs with regularized optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 7091–7101. 2018.

J. Schmidhuber. Deep learning in neural networks: An overview. *arXiv preprint, arXiv:1404.7828*, abs/1404.7828, Aug. 2014. URL `http://arxiv.org/abs/1404.7828`.

V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473, 2020.

S. L. Sobolev. *Applications of Functional Analysis in Mathematical Physics*. American Mathematical Society, 1963.

E. M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, 1970.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint, arXiv:1512.00567*, Dec. 2015. URL `https://arxiv.org/abs/1512.00567`.

D. Terjék. Adversarial Lipschitz regularization. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

I. O. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

V. S. Vladimirov. *Equations of Mathematical Physics*. Mir Publishers, 1984.

D. Wang and Q. Liu. Learning to draw samples: with application to amortized MLE for generative adversarial learning. *arXiv preprint, arXiv:1611.01722*, Nov. 2016. URL `https://arxiv.org/abs/1611.01722`.

L. Yang et al. Highly-scalable, physics-informed GANs for learning solutions of stochastic PDEs. In *Proceedings of the IEEE/ACM Third Workshop on Deep Learning on Supercomputers*, pages 1–11, 2019.

J. J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

J-Y Zhu, T Park, P Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision*, 2017.