mathematics

Article

Dynamics of Fourier Modes in Torus Generative Adversarial Networks

Ángel González-Prieto [1,*,†], Alberto Mozo [2,†], Edgar Talavera [2,†] and Sandra Gómez-Canaval [2,†]

[1] Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Madrid, Spain
[2] Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain; a.mozo@upm.es (A.M.); e.talavera@upm.es (E.T.); sm.gomez@upm.es (S.G.-C.)
* Correspondence: angel.gonzalez.prieto@upm.es
† These authors contributed equally to this work.

Abstract: Generative Adversarial Networks (GANs) are powerful machine learning models capable of generating fully synthetic samples of a desired phenomenon with a high resolution. Despite their success, the training process of a GAN is highly unstable, and typically, it is necessary to implement several accessory heuristics to the networks to reach acceptable convergence of the model. In this paper, we introduce a novel method to analyze the convergence and stability in the training of generative adversarial networks. For this purpose, we propose to decompose the objective function of the adversary min–max game defining a periodic GAN into its Fourier series. By studying the dynamics of the truncated Fourier series for the continuous alternating gradient descend algorithm, we are able to approximate the real flow and to identify the main features of the convergence of GAN. This approach is confirmed empirically by studying the training flow in a 2-parametric GAN, aiming to generate an unknown exponential distribution. As a by-product, we show that convergent orbits in GANs are small perturbations of periodic orbits so the Nash equillibria are spiral attractors. This theoretically justifies the slow and unstable training observed in GANs.

Keywords: Generative Adversarial Networks; dynamical systems; machine learning; Morse theory; Nash equilibrium

check for updates

Citation: González-Prieto, Á.; Mozo, A.; Talavera, E.; Gómez-Canaval, S. Dynamics of Fourier Modes in Torus Generative Adversarial Networks. *Mathematics* **2021**, *9*, 325. https://doi.org/10.3390/math9040325

Academic Editor: János Sztrik
Received: 15 December 2020
Accepted: 1 February 2021
Published: 6 February 2021

1. Introduction

Since their very inception, Generative Adversarial Networks (GANs) have revolutionized the areas of machine learning and deep learning. They address very successfully one of the most outstanding problems in pattern recognition: given a collection of examples of a certain phenomenon that we want to replicate, construct a generative model able to create new completely synthetic instances following the same patterns as the original ones. Ideally, the goal would be to capture the underlying pattern so subtly that no external critic would be able to distinguish between real samples and synthesized instances.

The proposal of Goodfellow et al. [1] is to confront two neural networks in an adversary game to solve this problem. More precisely, they propose to consider a neural network G playing the role of a generator agent and a network D acting as the discriminator. The discriminator D is trained to distinguish as accurately as possible between real samples and fake/synthetic samples. On the other hand, G aim to generate synthetic instances of high quality in such a way that D is barely able to distinguish from real data. The two networks are, thus, in effective competition. When, as a by-product of this competition, the agents reach an optimal point, we obtain a generator able to generate almost indistinguishable synthetic samples as well as a discriminator very proficient in classifying real and fake instances.

The way in which these networks are trained to reach this optimal point is through a common objective function. Explicitly, in [1], it is proposed to consider the following function:

$$\mathcal{F}(\theta_D, \theta_G) = \mathbb{E}_\Omega \log\left[D_{\theta_D}(X)\right] + \mathbb{E}_\Lambda \log\left[1 - D_{\theta_D}(G_{\theta_G})\right],$$

where θ_D are the inner weights of D, θ_G are the weights of G, Ω is the probability space of the real data, and Λ is the latent probability space from which G samples the noise to be transformed into synthetic instances. In this manner, $\mathcal{F}$ is essentially the error that D suffers in the classification problem between real and fake examples so D tries to maximize it and G tries to minimize it. Hence, it gives rise to a non-convex min–max game and the goal of the training process is to reach a Nash equilibrium.

Several training approaches have been proposed to reach these Nash equlibria, but the most widely used method is the so-called Alternating Gradient Descend (AGD). Roughly speaking, the idea is to, alternatively, train D by tuning θ_D with cost function $\mathcal{F}$ and weights θ_G fixed and, after a certain amount of epochs, to reverse the roles and to update θ_G with cost function $-\mathcal{F}$ and weights θ_D fixed. This optimization procedure has led to astonishing results, particularly in the domain of image processing and generation. Using several architectures and sophisticated multi-level training, GANs are able to generate images with such a high quality that a human eye is not capable to distinguish them from real images [2].

Despite these achievements, the stability of the AGD algorithm for GANs is a major issue. In [3], the authors proved that the Nash equillibria for GANs are locally stable provided that some ideal conditions on the optimality of the equillibria are fulfilled. Nevertheless, these conditions may be unfeasible, as shown in [4], so actual convergence and stability are not guaranteed in real applications. In particular, one of the most challenging problems arising during the training of GANs is the so-called mode collapse [5]. This state is characterized by a generator that has degenerated into a network that is only able to generate a single synthetic sample (or a very small number of them) with almost no variation and such that the discriminator confuses it with a real sample (typically, because the synthetic sample is actually very close to a real one). In this state, the system is no longer a generative model but simply a copier of real data.

Furthermore, by construction, neural network-based GANs have some intrinsic constraints in their expressivity that lead to very unrealistic synthetic samples in the context far from image generation. For instance, neural networks produce a smooth output function, which provokes GANs having lots of difficulties in dealing with the generation of real samples drawn from a discrete distribution (e.g., according to an exponential distribution) [6] or with some drastic semantic restrictions (e.g., nonnegative values for counters) [7]. These scenarios do not typically appear in image generation but are common in other domains such as data augmentation for machine learning [8]. These problems lead to additional inconveniences for stable convergence and usually give rise to highly unstable models that require a very handcrafted stopping criteria and optimization heuristics.

A multitude of works have been oriented towards a deeper understanding of the instability of the training of GANs as well as to propose solutions. A thorough theoretical study of the sources of instability and their causes can be found in [9], and in [10,11], the authors analyzed the real capability of the GAN for learning the distribution both through a theoretical and an empirical approach. In addition, in order to mitigate the instability of the training in [12], the authors proposed a collection of heuristical methods through variations of the standard backpropagation algorithm that contribute to stabilizing the training process of GANs. Moreover, in [13], the use of regularization procedures was proposed to speed up the convergence.

Another very active research line is the proposal of alternative models for GANs that guarantee better convergence. It is well known that the key reason why GAN should capture the original distribution is because they implicitly optimize the Jensen–Shannon divergence (JSD) between the real underlying distribution and the generated distribution of the synthetic data [1]. In order to change this framework, in [14], the authors proposed to modify the cost function in such a way that the new GAN did not optimize JSD but an Earth-mover distance known as Wasserstein distance, giving rise to the celebrated Wasserstein Generative Adversarial Networks (WGANs). In a similar vein, in [15], it was proposed to use the f-divergence (a divergence in the spirit of the Kullback–Leibler divergence) as the criterion for training GANs. Even genetic algorithms have been used to stabilize the training process, as in [16], where the authors applied genetic programming to optimize the use of different adversarial training objectives and evolved a population of generators to adapt to the discriminator, which acts as the hostile environment driving evolution. Nevertheless, despite all these efforts, no master method is currently available, and hence, assuring a fast, or even effective, convergence of GANs is an open problem.

Our contribution. In this paper, we propose a novel method to analyze the convergence of GANs through Fourier analysis. Concretely, we propose to approximate the objective function $\mathcal{F}$ by its Fourier series, truncated with enough precision that the local dynamics of $\mathcal{F}$ can be understood by means of a trigonometric polynomial.

Recall that any function $\mathcal{F}(\theta) : \mathbb{T}^n \to \mathbb{C}$ defined on the n-dimensional torus $\mathbb{T}^n = (S^1)^n$ (equivalently, an n-periodic function on $\mathbb{R}^n$) can be decomposed into a series of complex exponential functions, known as its Fourier series:

$$\mathcal{F}(\theta) = \sum_{\mathbf{m} \in \mathbb{Z}^n} \alpha_{\mathbf{m}}\, e^{2\pi i \mathbf{m} \cdot \theta},$$

where the series is indexed by the so-called Fourier modes or frequencies, $\mathbf{m}$ defined on the rectangular lattice $\mathbb{Z}^n \subseteq \mathbb{R}^n$. In principle, the previous equality must be understood as a decomposition in the Hilbert space of square-integrable functions, $L^2(\mathbb{T}^n)$. However, if $\mathcal{F}$ has enough regularity, then the Fourier series on the right-hand side also converges uniformly to the original function $\mathcal{F}$. This implies that, taking enough Fourier modes, $\mathcal{F}$ can be effectively approximated by a truncated Fourier series. Moreover, if $\mathcal{F}$ is real-valued, expressing the complex exponential as a combination of sine and cosine functions, we obtain an approximation of $\mathcal{F}$ by a trigonometric polynomial, $\Theta(\mathcal{F})$.

This approximation can be applied to the study of the convergence of GANs as follows. The continuous version of the AGD algorithm can be the thought of as a path of weights, $(\theta_D(t), \theta_G(t))$, depending on the time parameter $t \in \mathbb{R}$. In particular, $(\theta_D(0), \theta_G(0))$ are the initial random weights of the GAN and $(\theta_D(t), \theta_G(t))$ determine the state of the networks after training for a time $t > 0$. In this manner, if we seek to increase $\mathcal{F}(\theta_D, \theta_G)$ in the direction θ_D and to decrease it in the direction θ_G, the AGD gives rise to a system of Ordinary Differential Equations (ODEs) given by

$$\begin{cases} \theta_D' = \nabla_D \mathcal{F}(\theta_D, \theta_G), \\ \theta_G' = -\nabla_G \mathcal{F}(\theta_D, \theta_G), \end{cases}$$

where θ_D' and θ_G' denote the derivatives of the functions $\theta_D(t)$ and $\theta_G(t)$ with respect to time t. This flow aims to converge to a Nash equilibrium of the objective function $\mathcal{F}$ of the GAN, and for this reason, we refer to it as the Nash flow.

However, in many interesting cases, the function $\mathcal{F}$ may be very involved and lacks an analytic closed expression that would enable an explicit analysis (e.g., even in the toy example of Equation (13), the cost function is intractable analytically). To address this problem, we propose to approximate $\mathcal{F}$ by its truncated Fourier series, $\Theta(\mathcal{F})$. In this way, at least locally, the dynamic of the original Nash flow can be read from the solutions to the simplified system

$$\begin{cases} \theta_D' = \nabla_D \Theta(\mathcal{F})(\theta_D, \theta_G), \\ \theta_G' = -\nabla_G \Theta(\mathcal{F})(\theta_D, \theta_G). \end{cases}$$

In order to analyze this system of ODEs, we propose a novel method focused on studying the dynamics of the Nash flow on Fourier basic functions and on subsequent further approximations. As we will see, for the Nash flow of a basic trigonometric function, the Nash equillibria are not attractors of the flow but centers, that is, they are surrounded by periodic functions that spin around the critical point. When we consider more Fourier modes in the Fourier expansion of $\mathcal{F}$, these periodic orbits may break, leading to spiral attractors or spiral repulsors. The conditions that bifurcate the centers into spiral sinks or sources can be given explicitly in terms of the combinatorics of the considered Fourier modes.

This provides a theoretical justification to the empirically observed instability of the GAN training: the convergent orbits towards a Nash equilibrium are mere perturbations of periodic orbits, falling slowly and spirally to the optimal point. For this reason, small variations in the training hyperparameters, such as the learning rate, the number of epochs, or the batch size, may lead to very different dynamics, which confers to training its characteristic instability. In addition, in this paper, we empirically evaluate this method against a GAN that aims to generate samples according to an unknown exponential distribution. To facilitate the visualization, we consider a simple GAN, with 1-dimensional parameter spaces in each network, in such a way that the Nash flow can be plotted as a planar path. We show that the proposed approach allows us to understand the simplified dynamics of the GAN and to extract qualitative information of the Nash flow.

It is worth mentioning that, in order to have a natural Fourier series, the considered objective function $\mathcal{F}$ of the GAN must be periodic. This may seem unrealistic in real-life GANs, but this is actually not a very strong condition. Usually, seeking to prove theoretical results about the convergence of GANs, most work forces $\mathcal{F}$ to have compact support (for instance, to assure that it is Lipschitz as in WGANs). In practice, this is accomplish by clipping the output of the generator and discriminator functions for large inputs. This provokes that, artificially, the objective function turns into a periodic function, and thus, it can be studied through the method introduced in this paper. We expect that this work will open the door to new methods for analyzing and quantifying the convergence of GANs by importing well-established techniques of harmonic analysis and dynamical systems on closed manifolds, as studied in global analysis.

The structure of this paper is as follows. In Section 2, we review the theoretical fundamentals of GANs and their associated objective function and training method. In Section 2.1, we sketch briefly some basic concepts of Morse theory, a very successful theory that allows us to relate the analytic properties of the function to be optimized with the topological properties of the underlying space. In Section 2.2, we introduce the Nash flow and discuss some of the arising problems for its convergence. In Section 3, we introduce torus GANs, and particularly, in Section 3.1 we explain how to perform Fourier analysis on the torus. Section 4 is devoted to the analysis of the Nash flow for truncated Fourier series both for basic function (Sections 4.1 and 4.2) and for more complicated combinations (Sections 4.3 and 4.4). In addition, in Section 5, the empirical testing of this method is performed, with comparisons between the real dynamic and the predicted ideal dynamic. Finally, in Section 7, we summarize some of the keys ideas of this paper and sketch some lines of future work.

2. GANs Dynamics

As introduced by Goodfellow in [1], a GAN network is a competitive model in which two intelligent agents (typically two neural networks) compete to improve their performance and to generate very precise samples according to a given distribution.

To be precise, let $X : \Omega \to \mathbb{R}^d$ be a d-dimensional random vector, defined on a certain probability space Ω. This random vector X should be understood as a very complex phenomenon whose samples we would like to replicate. For this purpose, we consider two functions:

$$D : \mathbb{R}^d \times \Theta_D \to \mathbb{R}, \quad G : \Lambda \times \Theta_G \to \mathbb{R}^d,$$

called the *discriminator* and the *generator*, respectively. Here, Λ is a probability space, called the latent space, and Θ_D, Θ_G are two given topological spaces. These functions should be seen as parametric families of functions $D_{\theta_D} : \mathbb{R}^d \to \mathbb{R}$ and $G_{\theta_G} : \Lambda \to \mathbb{R}^d$, parametrized by $\theta_D \in \Theta_D$ and $\theta_G \in \Theta_G$.

The aim of the GAN is to tune the parameters θ_D and θ_G is such a way that, given $x \in \mathbb{R}^d$, $D_{\theta_D}(x)$ intends to predict whether $x = X(\omega)$ for some $\omega \in \Omega$, i.e., whether x is compatible with being a real instance or it is a fake datum. Observe that, throughout this paper, we follow the convention that $D_{\theta_D}(x)$ is the probability of being a real instance; thus, $D_{\theta_D}(x) = 1$ means that D_{θ_D} is sure that x is real, and $D_{\theta_D}(x) = 0$ means that D_{θ_D} is sure that x is fake. On the other hand, the generative function, G_{θ_G}, is a d-dimensional random vector that seeks to converge in distribution to the original distribution X. Typically, the probability space Λ is $\mathbb{R}^l$ with a certain standard probability distribution λ, as the spherical normal distribution or a uniform distribution on the unit cube.

Remark 1. *In typical applications in machine learning, Ω is given by a finite set $\Omega = \{x_1, \ldots, x_N\}$, with $x_i \in \mathbb{R}^d$, and endowed with a discrete probability (typically, the uniform one) so X is just the identity function. In customary applications of GANs, we have that the instances x_i are images, represented by their pixel map, so the objective of the GAN is to generate new images as similar as possible to the ones in the dataset Ω.*

The competition appears because the agents D and G try to improve non-simultaneously satifactible objectives. On one hand, D tries to improve its performance in the classification problem, but on the other hand, G tries to generate as best results as possible to cheat D. To be precise, recall that perfect fit for the classification problem for D_{θ_D} is given by $D_{\theta_D}(x) = 1$ if x is an instance of X and $D_{\theta_D}(x) = 0$ if not. Hence, the L^1 error made by D_{θ_D} with respect to perfect classification is

$$\mathcal{E}(\theta_D, \theta_G) = \mathbb{E}_\Omega\left[1 - D_{\theta_D}(X)\right] + \mathbb{E}_\Lambda\left[D_{\theta_D}(G_{\theta_G})\right] = 1 - \mathbb{E}_\Omega\left[D_{\theta_D}(X)\right] + \mathbb{E}_\Lambda\left[D_{\theta_D}(G_{\theta_G})\right],$$

where $\mathbb{E}_\Omega$ and $\mathbb{E}_\Lambda$ denote the mathematical expectation on Ω and Λ, respectively. In this way, the objective of D_{θ_D} is to minimize $\mathcal{E}$, while the goal of G_{θ_G} is to maximize it. It is customary in the literature to consider the function $1 - \mathcal{E}$ as the objective and to weight the error with a certain smooth concave function $f : \mathbb{R} \to \mathbb{R}$. In this way, the final cost function is

$$\mathcal{F}(\theta_D, \theta_G) = \mathbb{E}_\Omega f\left[D_{\theta_D}(X)\right] + \mathbb{E}_\Lambda f\left[-D_{\theta_D}(G_{\theta_G})\right]. \tag{1}$$

Remark 2. *Typical choices for the weight function f are $f(s) = -\log(1 + \exp(-s))$, as in the original paper of Goodfellow [1], or $f(s) = s$, as in the Wasserstein GAN [9].*

However, in sharp contrast with what is typical in machine learning, the aim of the GAN is not to maximize/minimize $\mathcal{F}$. The objectives of the D and G agents are opposing: while D tries to maximize $\mathcal{F}$, the generator tries to minimize it. In this vein, the objective of the GAN is

$$\min_{\theta_G} \max_{\theta_D} \mathcal{F}(\theta_D, \theta_G) = \min_{\theta_G} \max_{\theta_D} \mathbb{E}_\Omega f\left[D_{\theta_D}(X)\right] + \mathbb{E}_\Lambda f\left[-D_{\theta_D}(G_{\theta_G})\right]. \tag{2}$$

In the case that the latent space Λ is naturally equipped with a topology (as in the case $\Lambda = (\mathbb{R}^l, \lambda)$), it is customary to require that $\mathcal{F} : \Theta_D \times \Theta_G \to \mathbb{R}$ is a continuous function. In addition, in our case, Θ_G and Θ_D are differentiable manifolds, so we require that both D and G are C^2 maps in both arguments, and thus, $\mathcal{F}$ is a differentiable function on $\Theta_D \times \Theta_G$.

To be precise, the algorithm proposed by Goodfellow [1] suggests to freeze the internal weights of G and to use it to generate a batch of fake examples from Λ. With this set of fake instances and another batch of real instances created using X (i.e., sampling randomly from the dataset of real instances), we train D to improve its accuracy in the classification problem with the usual backpropagation (i.e., gradient descent) method. Afterwards, we freeze the weights of D and we sample a batch of latent data of Λ (i.e., we randomly sample noise using the latent distribution) and we use it to train G using gradient descent for G with objective function $\theta_G \mapsto \mathbb{E}_\Lambda f(-D(G_{\theta_G}))$. Finally, we can alternate this process as many times as needed until we reach the desired results. Several metrics have been proposed to quantify this performance, specially regarding the domain of image generation, such as Inception Score (IS) [12], Fréchet Inception Distance (FID) [17], or perceptual similarity measures [18]. For a survey of these techniques, please refer to [19].

2.1. Review of Morse Theory

Let us suppose for a while that, instead of looking for solutions of (2), we were seeking the local maxima of $\mathcal{F}$. In this situation, the standard approach in machine learning is to consider the Morse flow, also known as gradient ascent flow. For it, let us fix riemannian metrics on Θ_D and Θ_G. Using them, we can compute the *gradient* of $\mathcal{F}$, $\nabla\mathcal{F} = (\nabla_D\mathcal{F}, \nabla_G\mathcal{F})$, where $\nabla_D\mathcal{F}, \nabla_G\mathcal{F}$ denote the gradient in the θ_D, θ_G directions, respectively. Then, the Morse flow is the differentiable flow on $\Theta_D \times \Theta_G$ generated by the vector field $\nabla\mathcal{F}$. Explicitly, it is given by the system of ODEs:

$$\begin{cases} \theta_D' = \nabla_D\mathcal{F}(\theta_D, \theta_G), \\ \theta_G' = \nabla_G\mathcal{F}(\theta_D, \theta_G). \end{cases} \tag{3}$$

This flow has been the objective of very intense studies in the context of differentiable geometry and geometric topology. For instance, it is the crucial tool used in Smale's proof of the Poincaré conjecture in high dimension [20] and has been successfully used to understand the topology of moduli spaces of solutions to highly nonlinear partial differential equations coming from theoretical physics [21], among others.

Obviously, the critical points of the system (3) are exactly the *critical points* of $\mathcal{F}$ in the sense that the differential $d\mathcal{F}|_{(\theta_D^0, \theta_G^0)} = 0$. In order to control the dynamics of this ODE around a critical point, a key concept is the notion of index of a point.

Definition 1. *Let (θ_D^0, θ_G^0) be a critical point of $\mathcal{F}$. The* Hessian *of $\mathcal{F}$ at (θ_D^0, θ_G^0) is the symmetric 2-form $H\mathcal{F}|_{\theta_D^0, \theta_G^0} \in Sym^2(T^*_{\theta_D^0}\Theta_D \oplus T^*_{\theta_G^0}\Theta_G)$ given by*

$$\mathrm{Hess}(\mathcal{F})|_{\theta_D^0, \theta_G^0}(v, w) = w(\tilde{v}(\mathcal{F})),$$

for $v \in T_{\theta_D^0}\Theta_D, w \in T_{\theta_G^0}\Theta_G$, and $\tilde{v}$, with any extension of v to an vector field in a small neighborhood of (θ_D^0, θ_G^0).

The point (θ_D^0, θ_G^0) is said to be non-degenerate *if* $\mathrm{Hess}(\mathcal{F})|_{\theta_D^0, \theta_G^0}$ *is non-degenerated in the* 2*-form. In that case, the* index *of the point, denoted $\lambda(\theta_D^0, \theta_G^0)$, is the number of negative eigenvalues of* $\mathrm{Hess}(\mathcal{F})|_{\theta_D^0, \theta_G^0}$. *A function $\mathcal{F}$ is said to be* Morse *if all its critical points are non-degenerate.*

More explicitly, let $\partial_D^1, \ldots, \partial_D^{d_D}$ be a basis of $T_{\theta_D^0}\Theta_D$ and $\partial_G^1, \ldots, \partial_G^{d_G}$ be a basis of $T_{\theta_G^0}\Theta_G$, where d_D and d_G are the dimensions of Θ_D and Θ_G respectively. Then, Hessian is the matrix of second derivatives:

$$\mathrm{Hess}(\mathcal{F}) = \begin{pmatrix} \frac{\partial^2\mathcal{F}}{\partial\theta_D^i\partial\theta_D^j} & \frac{\partial^2\mathcal{F}}{\partial\theta_D^i\partial\theta_G^j} \\ \frac{\partial^2\mathcal{F}}{\partial\theta_G^i\partial\theta_D^j} & \frac{\partial^2\mathcal{F}}{\partial\theta_G^i\partial\theta_G^j} \end{pmatrix}$$

If Θ_D and Θ_G are compact, Morse functions are known to form a dense open set of the space of continuous functions on $\Theta_D \times \Theta_D$ [20]. Moreover, the critical points of a Morse function are isolated in the sense that there exists an open neighborhood of each critical point that contains only that critical point. Indeed, the stability of a critical point (θ_D, θ_G) is fully determined by its index. Then, (θ_D, θ_G) is a sink in a hypersurface of dimension $\lambda(\theta_D, \theta_G)$ while it is a source in a hypersurface of dimension $d_D d_G - \lambda(\theta_D, \theta_G)$. In particular, the only sinks of the Morse flow are precisely the local maxima of $\mathcal{F}$, in which $\mathrm{Hess}(\mathcal{F})$ is negative-definite and, thus, $\lambda(\theta_D, \theta_G) = d_D d_G$.

Another important fact that we use is the following topological interpretation of the indices, known as the Poincaré–Hopf theorem. It claims that, if Θ_D and Θ_G are compact, then

$$\sum_{(\theta_D,\theta_G)\in \mathrm{Crit}(\mathcal{F})} (-1)^{\lambda(\theta_D,\theta_G)} = \chi(\Theta_D \times \Theta_G) = \chi(\Theta_D)\chi(\Theta_G). \tag{4}$$

Here, $\mathrm{Crit}(\mathcal{F})$ denotes the (finite) set of critical points of $\mathcal{F}$ and χ is the Euler characteristic of the space.

2.2. The Nash Flow

Now, let us come back to our optimization problem (2). Despite the simplicity of the formulation of the cost function, this problem is very far from being trivial. The best scenario would be to obtain a so-called Nash equilibrium.

Definition 2. *Let $\mathcal{F} : \Theta_D \times \Theta_G \to \mathbb{R}$ be a differentiable function. A point $(\theta_D^0, \theta_G^0) \in \Theta_D \times \Theta_G$ is said to be a* Nash equilibrium *if*

- *the function $\theta_D \mapsto \mathcal{F}(\theta_D, \theta_G^0)$ has a maximum at θ_D^0.*
- *the function $\theta_G \mapsto \mathcal{F}(\theta_D^0, \theta_G)$ has a minimum at θ_G^0.*

Remark 3. *A Nash equilibrium is in particular a critical point of $\mathcal{F}$.*

In this vein, it is natural to consider an analogous differentiable flow to (3) but converging to Nash equilibria. For this purpose, fix riemannian metrics on Θ_D and Θ_G as above and consider the gradient $\nabla\mathcal{F} = (\nabla_D\mathcal{F}, \nabla_G\mathcal{F})$. Now, we twist the gradient to consider the *Nash vector field*:

$$\mathcal{N}(\mathcal{F}) = (\nabla_D\mathcal{F}, -\nabla_G\mathcal{F}).$$

Definition 3. *The* Nash flow *is the differentiable flow on $\Theta_D \times \Theta_G$ generated by the Nash vector field $\mathcal{N}(\mathcal{F})$. Explicitly, it is the system of* ODEs*:*

$$\begin{cases} \theta_D' = \nabla_D\mathcal{F}(\theta_D, \theta_G), \\ \theta_G' = -\nabla_G\mathcal{F}(\theta_D, \theta_G). \end{cases} \tag{5}$$

This flow (or, more precisely, the associated discrete-time version known as the AGD flow) has been intensively used for training GANs from their very inception. Already in Goodfellow's seminar paper [1], this flow was proposed as a method for seeking Nash equilibriums of the game (2).

To understand the dynamics of the Nash flow, let us study it around a critical point. Working in a local chart around a critical point, with an adapted basis $\partial_D^1, \ldots, \partial_D^{d_D}, \partial_G^1, \ldots, \partial_G^{d_G}$ of $T_{\theta_D^0}\Theta_D \oplus T_{\theta_G^0}\Theta_G$, the differential of the Nash vector field is the Nash Hessian:

$$\mathcal{N}\mathrm{Hess}(\mathcal{F}) = (\mathcal{N}(\mathcal{F}))_* = \begin{pmatrix} \frac{\partial^2\mathcal{F}}{\partial\theta_D^i\partial\theta_D^j} & \frac{\partial^2\mathcal{F}}{\partial\theta_D^i\partial\theta_G^j} \\ -\frac{\partial^2\mathcal{F}}{\partial\theta_G^i\partial\theta_D^j} & -\frac{\partial^2\mathcal{F}}{\partial\theta_G^i\partial\theta_G^j} \end{pmatrix}$$

In this manner, in a small neighborhood of a critical point $(\theta_D^0, \theta_G^0) \in \Theta_D \times \Theta_G$ of $\mathcal{F}$ (in particular, around a Nash equilibrium), the dynamics are determined by the linearized version:

$$\begin{pmatrix} \theta_D' \\ \theta_G' \end{pmatrix} = \left.\begin{pmatrix} \frac{\partial^2\mathcal{F}}{\partial\theta_D^i\partial\theta_D^j} & \frac{\partial^2\mathcal{F}}{\partial\theta_D^i\partial\theta_G^j} \\ -\frac{\partial^2\mathcal{F}}{\partial\theta_G^i\partial\theta_D^j} & -\frac{\partial^2\mathcal{F}}{\partial\theta_G^i\partial\theta_G^j} \end{pmatrix}\right|_{(\theta_D^0,\theta_G^0)} \begin{pmatrix} \theta_D \\ \theta_G \end{pmatrix}$$

However, in sharp contrast with the Morse flow, even if $\mathcal{F}$ has non-degenerate critical points, it may happen that the Nash equilibria are not attractors. For instance, if the Nash Hessian has a vanishing diagonal (as in Section 4.2), then periodic orbits arise around the critical point and the flow is non-convergent.

Nonetheless, this behavior can be controlled. Suppose for simplicity that $d_D = d_G = 1$ (higher dimensional scenarios can be treated analogously by splitting the tangent space). In that case, the eigenvalues of $\mathcal{N}\mathrm{Hess}(\mathcal{F})$ are either both real or complex conjugated.

- If the eigenvalues are real around a Nash equilibrium, both eigenvalues must be nonnegative, since in the usual Hessian, they have different signs. Hence, the Nash equilibrium is a non-repulsor of the Nash flow. Moreover, if $\mathcal{F}$ is Morse, then its eigenvalues do not vanish and, thus, the Nash equilibrium is an attractor.
- If the eigenvalues are complex conjugated, say $\lambda, \overline{\lambda} \in \mathbb{C}$, then the dynamic is controlled by the real part of λ, $\mathrm{Re}(\lambda)$. There is an invariant way of computing this quantity as through the trace of $\mathcal{N}\mathrm{Hess}(\mathcal{F})$ since
$$2\mathrm{Re}(\lambda) = \lambda + \overline{\lambda} = \mathrm{tr}(\mathcal{N}\mathrm{Hess}(\mathcal{F})) = \frac{\partial^2 \mathcal{F}}{\partial \theta_D^2} - \frac{\partial^2 \mathcal{F}}{\partial \theta_G^2}.$$
Observe that this is nothing but the wave operator acting on $\mathcal{F}$. In the case that this trace is negative, the critical point is an attractor with spiral dynamic; if it is positive, it is a repulsor, and if it vanishes, it is a center with surrounding periodic orbits.

It is worth mentioning that, in the case of GANs, the function $\mathcal{F}$ of (2) to be optimized does not define a convex–concave game so, in general, the convergence of the usual training methods through Nash flow is not guaranteed [3]. Under some ideal assumptions on the behaviour of the game around the Nash equilibrium points, in [3], the authors proved that the Nash flow is locally asymptotically stable. However, the hypotheses needed to apply this result are quite strong and seem to be unfeasible in practice. For instance, in [4], the authors show an example of a very simple GAN, the so-called Dirac GAN, for which the usual gradient descend does not converge.

3. Torus GANs

From now on, let us focus on a very particular case of GAN that we call a *torus GAN*. Let us denote
$$\mathbb{T}^n = \underbrace{S^1 \times \ldots S^1}_{n \text{ times}}$$
as the n-dimensional torus. Then, we take as parameter spaces $\Theta_D = \mathbb{T}^{d_D}$ and $\Theta_G = \mathbb{T}^{d_G}$. In this way, the cost functional becomes a function:
$$\mathcal{F} : \mathbb{T}^{d_D} \times \mathbb{T}^{d_G} = \mathbb{T}^{d_D + d_G} \to \mathbb{R}.$$

Remark 4. *This particular choice is not as arbitrary as it may seem at a first sight. In the end, a torus GAN is any GAN in which the generator and discriminator are periodic functions on their parameters θ_D and θ_G for some large enough period. In standard neural network-based GANs, it is customary to clip the output of the neural network in order to prevent the internal weights from becoming arbitrarily large. This is particularly important in Wasserstein GANs, where the objective function is required to be Lipschitz, and this is achieved by forcing the cost function to have compact support. In this way, after clipping, both the generator and the discriminator agents are periodic functions, and thus, they define a torus GAN.*

Working on the torus has important consequences to the dynamics the Morse flow. Some of them are the following:

- Divergent orbits are not allowed. Since $\mathbb{T}^n$ is compact, standard results of prologability of solutions for a short time show that the orbits of any vector flow cannot blow up. Intuitively, they cannot escape by tending to infinity. In particular, if $\mathcal{F}$ is a Morse function, all the orbits in the Morse flow must converge to a critical point. This is a consequence of the fact that, along a non-constant orbit of the Morse flow, the function $\mathcal{F}$ is strictly increasing since

$$\frac{d}{dt}\mathcal{F}(\theta_D, \theta_G) = d\mathcal{F}(\theta_D', \theta_G') = d\mathcal{F}(\nabla\mathcal{F}) = ||\nabla\mathcal{F}||^2 > 0.$$

Thus, since $\mathcal{F}$ is bounded, the flow is forced to converge to a constant orbit, that is, to a critical point of $\mathcal{F}$. This prevents the appearance of periodic orbits in the Morse flow. In the Nash flow, this may no longer hold and periodic orbits may arise (as in Section 4.2).

- Topological restrictions: the Euler characteristic of $\mathbb{T}^n$ is $\chi(\mathbb{T}^n) = \chi(S^1)^n = 0$. Hence, Equation (4) implies that

$$\sum_{(\theta_D, \theta_G) \in \mathrm{Crit}(\mathcal{F})} = 0.$$

In other word, there is the same number of critical points of even index as of odd index. In particular, if $d_D = d_G = 1$, there are as many saddle points (which are points of index 1) as maxima and minima (which are points of index 2 or 0).

3.1. Fourier Analysis in the Torus

In order to understand the cost function $\mathcal{F}$ of a torus GAN, we apply techniques of harmonic analysis to it. We suppose that the reader is familiar with basic notions of Fourier and harmonic analysis, such as Hilbert spaces and orthogonal Schauder basis on them. Otherwise, please refer to [22].

Let us consider $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$ so that functions on $\mathbb{T}^n$ are n-periodic functions on the unit square. Recall that a fundamental result of Fourier analysis is that the space $L^2(\mathbb{T}^n)$ of complex-valued square-integrable functions on $\mathbb{T}^n$ is a Hilbert space with product given by

$$\langle \mathcal{F}, \mathcal{G} \rangle = \int_{\mathbb{T}^n} \mathcal{F}(\theta)\overline{\mathcal{G}(\theta)}\, d\theta.$$

Moreover, this space is spanned by the orthonormal basis of functions:

$$e_{\mathbf{m}}(\theta) = e^{2\pi i \mathbf{m}\cdot\theta},$$

where $\mathbf{m} = (m_1, \ldots, m_n) \in \mathbb{Z}^n$, $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{T}^n$ and $\mathbf{m}\cdot\theta = m_1\theta_1 + \ldots + m_n\theta_n$ is the standard inner product. In other words, any $\mathcal{F} \in L^2(\mathbb{T}^n)$ can be uniquely written as a sum:

$$\mathcal{F}(\theta) = \sum_{\mathbf{m}\in\mathbb{Z}^n} \alpha_{\mathbf{m}}\, e_{\mathbf{m}}(\theta) = \sum_{\mathbf{m}\in\mathbb{Z}^n} \alpha_{\mathbf{m}}\, e^{2\pi i \mathbf{m}\cdot\theta},$$

in the sense that this sum is convergent in $L^2(\mathbb{T}^n)$ and converges to $\mathcal{F}$. This expression is referred to as the *Fourier series* of $\mathcal{F}$. The coefficients $\alpha_{\mathbf{m}}$ are called the *Fourier coefficients* or the *Fourier modes* of $\mathcal{F}$. Using the orthogonality of the functions $e_{\mathbf{m}}(\theta)$, they can be obtained as

$$\alpha_{\mathbf{m}} = \langle \mathcal{F}, e_{\mathbf{m}}(\theta) \rangle = \int_{\mathbb{T}^n} \mathcal{F}(\theta) e^{-2\pi i \mathbf{m}\cdot\theta}\, d\theta.$$

In principle, the convergence of the Fourier series to $\mathcal{F}$ is only in the L^2 sense (c.f. [23] for a Fourier series of a continuous function not converging pointwise everywhere or [24] for an everywhere divergent Fourier series of a L^1 function). However, if $\mathcal{F}$ is C^1, since we are working on a compact space, it is automatically Hölder and, thus, its Fourier series converges uniformly [25]. This means that, for every $\epsilon > 0$

$$\left|\left| \mathcal{F} - \sum_{m_i=-N}^{N} \alpha_{\mathbf{m}}\, e_{\mathbf{m}} \right|\right|_\infty = \sup_{\theta\in\mathbb{T}^n} \left| \mathcal{F}(\theta) - \sum_{m_i=-N}^{N} \alpha_{\mathbf{m}}\, e^{2\pi i \mathbf{m}\cdot\theta} \right| < \epsilon,$$

for all N large enough. Similar approximations can be obtained for the k first derivatives of $\mathcal{F}$ if it has enough regularity (concretely, if it is C^{k+1}).

This approximation is very useful for estimating the associated flow. Recall that, using the Gronwall inequality [26], if X, Y are two Lipschitz vector fields, then there exists a constant $M > 0$ such that their associated flows $\theta(t)$ and $\vartheta(t)$ satisfy

$$|\theta(t) - \vartheta(t)| \leq \frac{e^{Mt} - 1}{M} ||X - Y||_\infty$$

for all t. In other words, for medium times, the flow of X may be approximated through the flow of Y.

Remark 5. *The previous estimation implies that, locally, the dynamics of the flows $\theta(t)$ and $\vartheta(t)$ are similar. In particular, this is useful for analyzing convergence around critical points. Nevertheless, the global dynamics of $\theta(t)$ and $\vartheta(t)$ may be quite different, say, they may have different numbers of critical points.*

In our context, this idea can be exploited as follows. Let us denote by

$$\Theta_N(\mathcal{F}) = \sum_{m_i=-N}^{N} \alpha_{\mathbf{m}}\, e_{\mathbf{m}}$$

the truncated Fourier series of $\mathcal{F}$. If $\mathcal{F}$ is C^2, then $\nabla\mathcal{F}$ and $\nabla\Theta_N(\mathcal{F})$ are close vector fields and, thus,

$$|\theta(t) - \theta_N(t)| \leq \frac{e^{Mt} - 1}{M} ||\nabla\mathcal{F} - \nabla\Theta_N(\mathcal{F})||_\infty \leq \epsilon(e^{Mt} - 1)$$

for N large enough, where $\theta(t)$ is the Morse flow for $\mathcal{F}$ and $\theta_N(t)$ is the Morse flow for $\Theta_N(\mathcal{F})$. Working verbatim with the Nash vector fields, we obtain similar estimates for the solutions of the Nash flow.

4. Dynamics of Fourier Basis

In this section, we focus on the Nash flow of truncated approximations of Fourier series of a C^2 function $\mathcal{F}$. As we mentioned above, these solutions approximate quite well the real Nash flow of $\mathcal{F}$ for short times (particularly, around critical points).

For the sake of simplicity, in this section, we focus on the 2-dimensional case in which $d_D = d_G = 1$ so that $\mathcal{F} = \mathcal{F}(\theta_1, \theta_2)$ is a function:

$$\mathcal{F} : \mathbb{T}^2 \to \mathbb{R}.$$

Moreover, we truncate the Fourier series at the level $N = 2$. Similar arguments can be carried out for higher dimension and more accurate precision of the Fourier series with similar results, but the calculations become more involved.

First, let us rewrite the Fourier series of $\mathcal{F}$ as a trigonometric polynomial. Recall that the trigonometric functions can be obtained from the complex exponential as

$$\cos(2\pi\theta) = \frac{e^{2\pi i\theta} + e^{-2\pi i\theta}}{2}, \quad \sin(2\pi\theta) = \frac{e^{2\pi i\theta} - e^{-2\pi i\theta}}{2i}.$$

Since the function $\mathcal{F}$ is real-valued, we can group the coefficients and obtain a formula for the Fourier series in term of trigonometric functions as

$$\begin{aligned}\mathcal{F}(\theta_1, \theta_2) = &\sum_{m_1,m_2=0}^{\infty} a^{0,0}_{m_1,m_2} \sin(2\pi m_1\theta_1)\sin(2\pi m_2\theta_2) + \sum_{m_1,m_2=0}^{\infty} a^{0,1}_{m_1,m_2} \sin(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2) \\ &+ \sum_{m_1,m_2=0}^{\infty} a^{1,0}_{m_1,m_2} \cos(2\pi m_1\theta_1)\sin(2\pi m_2\theta_2) + \sum_{m_1,m_2=0}^{\infty} a^{1,1}_{m_1,m_2} \cos(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2).\end{aligned}$$

The coefficients are real numbers that can be obtained as

$$a^{0,0}_{m_1,m_2} = \delta_{m_1,m_2}\langle \mathcal{F}, \sin(2\pi m_1\theta_1)\sin(2\pi m_2\theta_2)\rangle = \delta_{m_1,m_2}\int_{\mathbb{T}^2}\mathcal{F}(\theta_1,\theta_2)\sin(2\pi m_1\theta_1)\sin(2\pi m_2\theta_2)\,d\theta_1 d\theta_2,$$
$$a^{0,1}_{m_1,m_2} = \delta_{m_1,m_2}\langle \mathcal{F}, \sin(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2)\rangle = \delta_{m_1,m_2}\int_{\mathbb{T}^2}\mathcal{F}(\theta_1,\theta_2)\sin(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2)\,d\theta_1 d\theta_2,$$
$$a^{1,0}_{m_1,m_2} = \delta_{m_1,m_2}\langle \mathcal{F}, \cos(2\pi m_1\theta_1)\sin(2\pi m_2\theta_2)\rangle = \delta_{m_1,m_2}\int_{\mathbb{T}^2}\mathcal{F}(\theta_1,\theta_2)\cos(2\pi m_1\theta_1)\sin(2\pi m_2\theta_2)\,d\theta_1 d\theta_2,$$
$$a^{1,1}_{m_1,m_2} = \delta_{m_1,m_2}\langle \mathcal{F}, \cos(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2)\rangle = \delta_{m_1,m_2}\int_{\mathbb{T}^2}\mathcal{F}(\theta_1,\theta_2)\cos(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2)\,d\theta_1 d\theta_2,$$

where δ_{m_1,m_2} is a coefficient that $\delta_{m_1,m_2} = 1$ if $m_1 = m_2 = 0$; $\delta_{m_1,m_2} = 2$ if $m_1 = 0$ and $m_2 > 0$, $m_1 > 0$ and $m_2 = 0$; and $\delta_{m_1,m_2} = 4$ $m_1, m_2 > 0$.

To shorten notation, from now on, we denote

$$\Lambda^{0,0}_{m_1,m_2}(\theta_1,\theta_2) = \sin(2\pi m_1\theta_1)\sin(2\pi m_2\theta_2), \qquad \Lambda^{0,1}_{m_1,m_2}(\theta_1,\theta_2) = \sin(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2),$$
$$\Lambda^{1,0}_{m_1,m_2}(\theta_1,\theta_2) = \cos(2\pi m_1\theta_1)\sin(2\pi m_2\theta_2), \qquad \Lambda^{1,1}_{m_1,m_2}(\theta_1,\theta_2) = \cos(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2),$$

This notation is particularly useful because, for any $\alpha, \beta \in \mathbb{Z}_2$,

$$\frac{\partial}{\partial\theta_1}\Lambda^{\alpha,\beta}_{m_1,m_2} = (-1)^{\alpha}2\pi m_1\Lambda^{\alpha+1,\beta}_{m_1,m_2}, \quad \frac{\partial}{\partial\theta_2}\Lambda^{\alpha,\beta}_{m_1,m_2} = (-1)^{\beta}2\pi m_2\Lambda^{\alpha,\beta+1}_{m_1,m_2},$$

where the sum is interpreted as the sum in $\mathbb{Z}_2$.

From this expression of the Fourier series, we approximate the dynamics of the Nash flow for $\mathcal{F}$ by truncating the Fourier series. In particular, we sort the coefficients $a^{\alpha,\beta}_{m_1,m_2}$ by decreasing order of their absolute value. Looking only at the two largest coefficients and normalizing so that the leading coefficient is 1, we consider the approximation to $\mathcal{F}$:

$$\Theta(\mathcal{F}) = \Lambda^{\alpha,\beta}_{m_1,m_2} + \mu\Lambda^{\gamma,\delta}_{n_1,n_2}, \tag{6}$$

where $\alpha, \beta, \gamma, \delta \in \mathbb{Z}_2$, (m_1, m_2) are the leading Fourier modes and (n_1, n_2) are the second largest modes, and $|\mu| < 1$.

4.1. Nash Flow for Single Variable Fourier Basis

From now on, we aim to analyze the Nash flow for a truncated Fourier series. As we see in Section 5, from it, we can envisage the global dynamics of the Nash flow for the objective function of a GAN.

First, let us consider the simplest Fourier modes, namely with $m_1 = 0$ or $m_2 = 0$. In this case, the dynamics is quite simple and, in most cases, can be pulled apart. In the case of $\Lambda^{\alpha,\beta}_{0,0}(\theta_1,\theta_2) \equiv 1$, the Nash flow equations amount to

$$\begin{cases} \theta_1' = \frac{\partial}{\partial\theta_1}\Lambda^{\alpha,\beta}_{0,0}(\theta_1,\theta_2) = 0, \\ \theta_2' = -\frac{\partial}{\partial\theta_2}\Lambda^{\alpha,\beta}_{0,0}(\theta_1,\theta_2) = 0. \end{cases}$$

Therefore, the solutions are constant orbits $(\theta_1(t), \theta_2(t)) = (\theta_1^0, \theta_2^0)$ for some fixed $(\theta_1^0, \theta_2^0) \in \mathbb{T}^2$. For this reason, it does not contribute to the dynamics.

For Fourier modes of the form $\Lambda^{0,\beta}_{m_1,0}(\theta_1,\theta_2) = \sin(2\pi m_1\theta_1)$ or $\Lambda^{1,\beta}_{m_1,0}(\theta_1,\theta_2) = \cos(2\pi m_1\theta_1)$, the situation is also very simple. Now, the Nash flow is given by

$$\begin{cases} \theta_1' &= \frac{\partial}{\partial\theta_1}\Lambda^{\alpha,\beta}_{m_1,0}(\theta_1,\theta_2) = 2\pi m_1\Lambda^{\alpha+1,\beta}_{m_1,0}(\theta_1,\theta_2), \\ \theta_2' &= -\frac{\partial}{\partial\theta_2}\Lambda^{\alpha,\beta}_{m_1,0}(\theta_1,\theta_2) = 0. \end{cases}$$

The solution to this system has the form $(\theta_1(t), \theta_2(t)) = (f^{\alpha}_{m_1}(t), \theta_2^0)$ for some fixed θ_2^0, and $f^{\alpha}_{m_1}(t)$ is a differentiable function depending on m_1 and α (the explicit form of $f^{\alpha}_{m_1}(t)$

can be obtained by solving the 1-dimensional ODE for θ_1 by separation of variables). Thus, the flow is completely horizontal with $2m_1$ lines of critical points at the lines $\theta_1 = \frac{2k_1-\alpha+1}{4m_1}$ for $k_1 \in \mathbb{Z}$. Half of these critical lines are attractive, corresponding to the maxima of $f^{\alpha}_{m_1}$, and half of them are repulsive, corresponding to the minima.

The situation of the Fourier modes of the form $\Lambda^{\alpha,0}_{0,m_2}(\theta_1,\theta_2) = \sin(2\pi m_2\theta_2)$ or $\Lambda^{\alpha,1}_{0,m_2}(\theta_1,\theta_2) = \cos(2\pi m_2\theta_2)$ is completely symmetric. Now, the flow is vertical and the critical lines are at $\theta_2 = \frac{2k_2-\alpha+1}{4m_2}$ for $k_2 \in \mathbb{Z}$ (but the attractive ones correspond to the minima and the repulsive to the minima).

Furthermore, we can collect all the Fourier modes with a vanishing frequency into a single function. To be precise, decompose the Fourier series of $\mathcal{F}$ as

$$\mathcal{F} = \underbrace{\frac{a^{0,0}_{0,0}}{2} + \sum_{\substack{1\le m_1<\infty\\ \alpha=0,1}} a^{\alpha,0}_{m_1,0}\Lambda^{\alpha,0}_{m_1,0}}_{\Delta_1(\theta_1)} + \underbrace{\frac{a^{0,0}_{0,0}}{2} + \sum_{\substack{1\le m_2<\infty\\ \beta=0,1}} a^{0,\beta}_{m_1,0}\Lambda^{0,\beta}_{0,m_2}}_{\Delta_2(\theta_2)} + \underbrace{\sum_{m_1,m_2=1}^{\infty} a^{0,0}_{m_1,m_2}\Lambda^{\alpha,\beta}_{m_1,m_2}}_{\Theta(\theta_1,\theta_2)}.$$

Now, the superposition principle applied to (5) implies that any solution to the Nash flow has the following form:

$$(\theta_1(t),\theta_2(t)) = (\hat{\theta}_1(t),\theta_2^0) + (\theta_1^0,\hat{\theta}_2(t)) + \Phi(t),$$

where $(\hat{\theta}_1(t),\theta_2^0)$ is a horizontal flow corresponding to the solution of (5) for Δ_1 (explicitly, $\hat{\theta}_1$ is the solution to the equation $\hat{\theta}_1' = \frac{d}{d\theta_1}\Delta_1(\hat{\theta}_1))$, $(\theta_1^0,\hat{\theta}_2(t))$ is a vertical flow corresponding to the solution of (5) for Δ_2 (i.e., $\hat{\theta}_2$ is the solution to $\hat{\theta}_2' = -\frac{d}{d\theta_2}\Delta_2(\hat{\theta}_2)$), and Φ is the solution to the (coupled) system of Equation (5) for Θ.

For this reason, in many cases, the effect of the Δ_1 and the Δ_2 parts in the dynamics is negligible and can be ignored.

4.2. Nash Flow for Fourier Basis

In this section, we analyze the dynamics of the Nash flow for the remaining Fourier basis. For this purpose, let us consider the function $\Lambda^{\alpha,\beta}_{m_1,m_2}$ for some $\alpha,\beta \in \mathbb{Z}_2$ with $m_1,m_2 \geq 1$. The Nash vector field associated with it is

$$\mathcal{N}\left(\Lambda^{\alpha,\beta}_{m_1,m_2}\right) = 2\pi\left((-1)^{\alpha}m_1\Lambda^{\alpha+1,\beta}_{m_1,m_2}, (-1)^{\beta}m_2\Lambda^{\alpha,\beta+1}_{m_1,m_2}\right). \quad (7)$$

Recall that, if $(\theta_1,\theta_2) \in \mathbb{T}^2$ is a zero of $\Lambda^{\alpha,\beta}_{m_1,m_2}$, then it satisfies

$$4\theta_1 m_1 \equiv 2k_1 + \alpha \mod 4\mathbb{Z}, \quad \text{or} \quad 4\theta_2 m_2 \equiv 2k_2 + \beta \mod 4\mathbb{Z},$$

for some $k_1,k_2 \in \mathbb{Z}$. In other words, if we take into account the periodicity of the function $\Lambda^{\alpha,\beta}_{m_1,m_2}$, the zeros are given by

$$\theta_1 = \frac{2k_1+\alpha}{4m_1}, \quad \text{or} \quad \theta_2 = \frac{2k_2+\beta}{4m_2},$$

for $0 \leq k_1 < 2m_1$ and $0 \leq k_2 < 2m_2$. Observe that all these values are different, so $\Lambda^{\alpha,\beta}_{m_1,m_2}$ has $4m_1m_2$ zeros.

Coming back to Equation (7), we observe that, if $(\theta_1,\theta_2) \in \mathbb{T}^2$ is a critical point of the Nash vector field (i.e., a critical point of $\Lambda^{\alpha,\beta}_{m_1,m_2}$), then it satisfies one of the following two possibilities:

$$\begin{array}{lll} \text{(I)} & (4\theta_1 m_1, 4\theta_2 m_2) \equiv (2k_1-\alpha+1, 2k_2-\beta+1) & \mod 4\mathbb{Z}\times 4\mathbb{Z}, \\ \text{(II)} & (4\theta_1 m_1, 4\theta_2 m_2) \equiv (2k_1+\alpha, 2k_2+\beta) & \mod 4\mathbb{Z}\times 4\mathbb{Z}. \end{array}$$

Beware of the change in sign for the coefficient of α and β for points (I). This is just a matter of notational convenience, as shown below. Equivalently, the these conditions can be written explicitly as

$$\begin{aligned} &\text{(I)} \quad (\theta_1, \theta_2) = \left(\frac{2k_1 - \alpha + 1}{4m_1}, \frac{2k_2 - \beta + 1}{4m_2} \right), \quad \text{for } k_1, k_2 \in \mathbb{Z}, \\ &\text{(II)} \quad (\theta_1, \theta_2) = \left(\frac{2k_1 + \alpha}{4m_1}, \frac{2k_2 + \beta}{4m_2} \right), \quad \text{for } k_1, k_2 \in \mathbb{Z}. \end{aligned}$$

Thus, the Nash vector field has $8m_1m_2$ critical points: $4m_1m_2$ critical points of type (I) and $4m_1m_2$ of type (II).

Regarding the Nash Hessian, it is explicitly given by

$$\mathcal{N}\mathrm{Hess}\left(\Lambda^{\alpha,\beta}_{m_1,m_2}\right) = 4\pi^2 \begin{pmatrix} -m_1^2 \Lambda^{\alpha,\beta}_{m_1,m_2} & (-1)^{\alpha+\beta} m_1 m_2 \Lambda^{\alpha+1,\beta+1}_{m_1,m_2} \\ (-1)^{\alpha+\beta+1} m_1 m_2 \Lambda^{\alpha+1,\beta+1}_{m_1,m_2} & m_2^2 \Lambda^{\alpha,\beta}_{m_1,m_2} \end{pmatrix}$$

Therefore, evaluated at a critical point of the form (I), we get that

$$\mathcal{N}\mathrm{Hess}\left(\Lambda^{\alpha,\beta}_{m_1,m_2}\right)|_{\text{(I)}} = (-1)^{k_1+k_2} 4\pi^2 \begin{pmatrix} -m_1^2 & 0 \\ 0 & m_2^2 \end{pmatrix}.$$

These are all saddle points for the Nash flow, with an attractive direction and a repulsive direction.

On the other hand, the Nash Hessian evaluated at a critical point of the form (II) is

$$\begin{aligned} \mathcal{N}\mathrm{Hess}\left(\Lambda^{\alpha,\beta}_{m_1,m_2}\right)|_{\text{(II)}} &= (-1)^{k_1+k_2+\alpha+\beta} 4\pi^2 \begin{pmatrix} 0 & m_1 m_2 \\ -m_1 m_2 & 0 \end{pmatrix} \\ &\sim (-1)^{k_1+k_2+\alpha+\beta} 4\pi^2 m_1 m_2 \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}. \end{aligned}$$

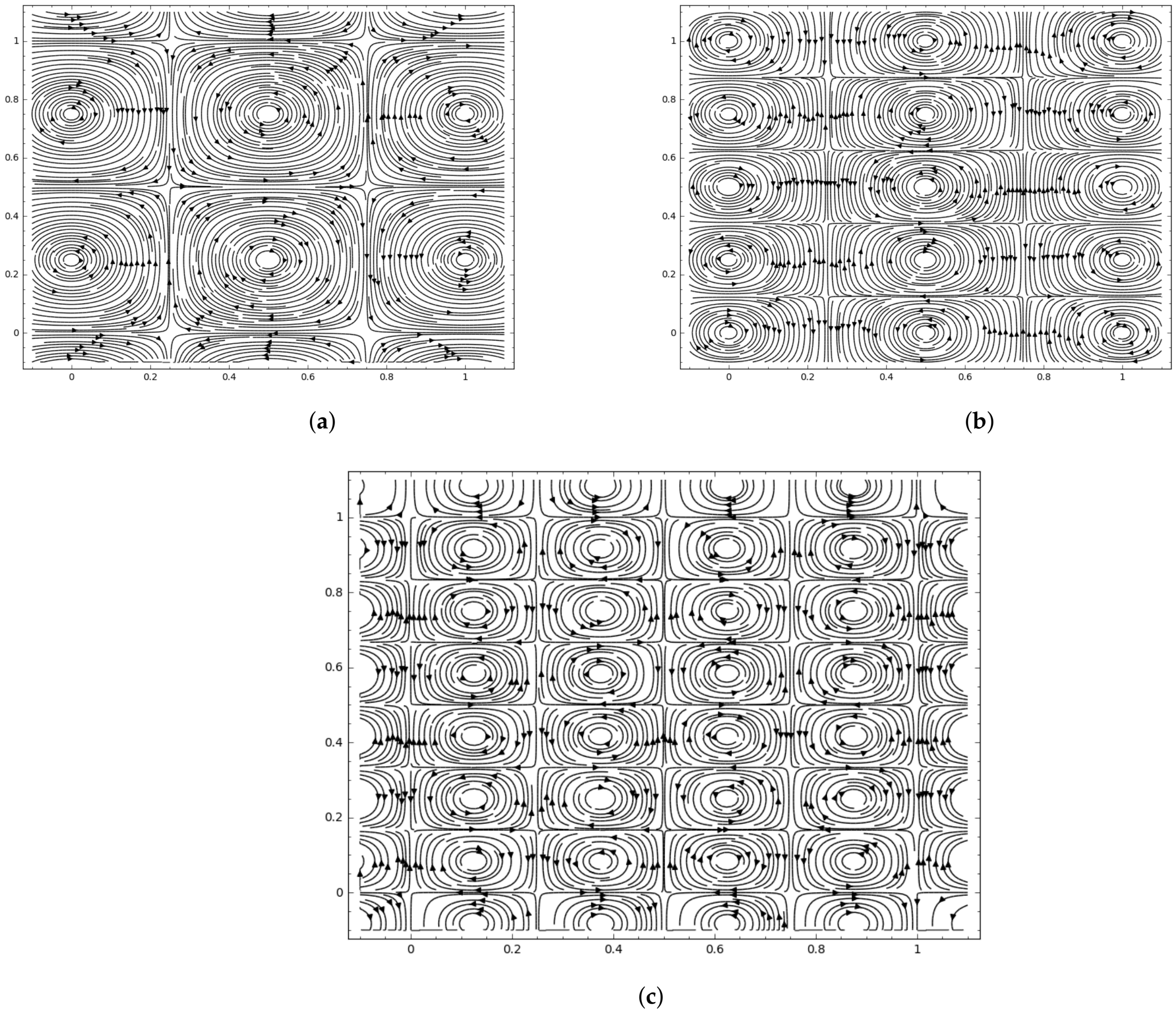
In this situation, we obtain a center critical point with periodic orbits around it and no convergent flow lines. This dynamic is depicted in Figure 1. Observe that, in this plot, the 2-dimensional torus $\mathbb{T}^2$ is represented as the square $[0,1] \times [0,1]$ with the boundaries identified in pairs, i.e., the left boundary $\{0\} \times [0,1]$ is identified with the right boundary $\{1\} \times [0,1]$ preserving the orientation and so are the bottom boundary $[0,1] \times \{0\}$ and the upper one $\{1\} \times [0,1]$).

Putting together these calculations, we have proven the following result.

Proposition 1. *The Nash flow for the Fourier basis function $\Lambda^{\alpha,\beta}_{m_1,m_2}$ has $8m_1m_2$ critical points, for which the dynamics are*

(I) *$4m_1m_2$ points are saddle points for the flow, half of them corresponding to the maxima of $\Lambda^{\alpha,\beta}_{m_1,m_2}$ and half of them to the minima.*

(II) *$4m_1m_2$ points are center points for the flow, surrounded by periodic orbits and corresponding to the saddle points of $\Lambda^{\alpha,\beta}_{m_1,m_2}$.*



(**a**)

(**b**)

(**c**)

Figure 1. Nash flow dynamics of Fourier basis functions: (**a**) $\Lambda^{0,1}_{1,1} = \sin(2\pi\theta_1)\cos(2\pi\theta_2)$, (**b**) $\Lambda^{0,0}_{1,2} = \sin(2\pi\theta_1)\sin(4\pi\theta_2)$, and (**c**) $\Lambda^{1,1}_{2,3} = \cos(4\pi\theta_1)\cos(6\pi\theta_2)$.

4.3. Nash Flow for Simplified Truncated Fourier Series

In [4], it is proven that, under some ideal conditions, the Nash flow associated with the cost function of a GAN has stable Nash equilibriums. For this reason, according to Proposition 1, these cost functions cannot be the basis functions of the Fourier series. In other words, its Fourier approximation (6) is nontrivial. Hence, in order to capture the actual dynamics of the GAN flow, let us consider a general truncated Fourier series of the following form:

$$\Theta = \Lambda^{\alpha,\beta}_{m_1,m_2} + \mu\Lambda^{\gamma,\delta}_{n_1,n_2},$$

for some $\alpha, \beta, \gamma, \delta \in \mathbb{Z}_2$, $-1 \leq \mu \leq 1$ and Fourier modes $m_1, m_2, n_1, n_2 \geq 1$.

In order to simplify the computations, in this section, we suppose that $m_1 = m_2 = 1$. After this case, the general setting is studied. In this simplified case, at a point $(\theta^0_1, \theta^0_2) = (k_1/2 + \alpha/4, k_2/2 + \beta/4)$ of the form (II), we have

$$\nabla\Theta|_{(\theta^0_1,\theta^0_2)} = 2\pi\mu((-1)^{\gamma} n_1 \Lambda^{\gamma+1,\delta}_{n_1,n_2}(\theta^0_1,\theta^0_2), (-1)^{\delta} n_2 \Lambda^{\gamma,\delta+1}_{n_1,n_2}(\theta^0_1,\theta^0_2)).$$

At this point, we have the following two options.

- If $\nabla\Theta|_{(\theta_1^0,\theta_2^0)} = 0$, then (θ_1^0,θ_2^0) is also a critical point of Θ. Hence, the dynamic of the Nash flow near (θ_1^0,θ_2^0) is determined by the Nash Hessian at that point. This Hessian is given by

$$\mathcal{N}\mathrm{Hess}(\Theta)|_{(\theta_1^0,\theta_2^0)} = (-1)^{k_1+k_2+\alpha+\beta}4\pi^2\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} + \mu\mathcal{N}\mathrm{Hess}\left(\Lambda^{\gamma,\delta}_{n_1,n_2}\right)|_{(\theta_1^0,\theta_2^0)}$$

Suppose that $(\gamma,\delta) = (\alpha+1,\beta+1)$ in $\mathbb{Z}_2\times\mathbb{Z}_2$. Set $\sigma = (-1)^{n_1k_1+n_2k_2+\alpha n_1/2+\beta n_2/2}$. Observe that $\Lambda^{\alpha,\beta}_{n_1,n_2}(\theta_1^0,\theta_2^0) = 0$ and $\Lambda^{\alpha+1,\beta+1}_{n_1,n_2}(\theta_1^0,\theta_2^0) = \sigma$, so we have that

$$\mathcal{N}\mathrm{Hess}\left(\Lambda^{\gamma,\delta}_{n_1,n_2}\right)|_{(\theta_1^0,\theta_2^0)} = 4\pi^2\mu\sigma\begin{pmatrix}-n_1^2 & 0\\ 0 & n_2^2\end{pmatrix}$$

With this calculation at hand, we observe the following. By continuity, for $|\mu|$ small, since $\mathcal{N}\mathrm{Hess}\left(\Lambda^{\alpha,\beta}_{1,1}\right)|_{(\theta_1^0,\theta_2^0)}$ has complex eigenvalues, then $\mathcal{N}\mathrm{Hess}(\Theta)|_{(\theta_1^0,\theta_2^0)}$ also has complex eigenvalues. In particular, they must be conjugated, say $\lambda,\overline{\lambda}\in\mathbb{C}$. In that case, the stability of a critical point at (θ_1^0,θ_2^0) is governed by the following trace:

$$2\mathrm{Re}(\lambda) = \lambda+\overline{\lambda} = \mathrm{tr}\mathcal{N}\mathrm{Hess}(\Theta)|_{(\theta_1^0,\theta_2^0)} = 4\pi^2\mu\sigma\left(n_2^2-n_1^2\right).$$

Hence, if $n_2 < n_1$ and $\mu\sigma = 1$, or $n_2 > n_1$ and $\mu\sigma = -1$ (respectively $n_2 > n_1$ and $\mu\sigma = 1$, or $n_2 < n_1$ and $\mu\sigma = -1$), any critical point nearby $(\theta_1,\theta_2)\in\mathbb{T}^2$ is an spiral attractor (respectively repulsor). In the case that $n_1 = n_2$, the eigenvalues are multiples of i and $-i$ so the point is still a center and the behaviour bifurcates depending on further Fourier modes.

On the other hand, if $\gamma = \alpha$ or $\delta = \beta$ in $\mathbb{Z}_2$, then we have that

$$\mathcal{N}\mathrm{Hess}\left(\Lambda^{\gamma,\delta}_{n_1,n_2}\right)|_{(\theta_1^0,\theta_2^0)} = \pm4\pi^2\mu\begin{pmatrix}0 & n_1n_2\\ -n_1n_2 & 0\end{pmatrix} \tag{8}$$

Therefore, $\mathcal{N}\mathrm{Hess}(\Theta)|_{(\theta_1^0,\theta_2^0)}$ is still an anti-diagonal matrix and the dynamics depends on further Fourier modes.

- If $\nabla\Theta|_{(\theta_1^0,\theta_2^0)} \neq 0$, then (θ_1^0,θ_2^0) is no longer a critical point of Θ. However, if $|\mu|$ is small, by the implicit function theorem, nearby (θ_1^0,θ_2^0), there must be a unique critical point $(\tilde{\theta}_1,\tilde{\theta}_2)\in\mathbb{T}^2$ of Θ. Again, by continuity, since $\mathcal{N}\mathrm{Hess}\left(\Lambda^{\alpha,\beta}_{1,1}\right)|_{(\theta_1,\theta_2)}$ has complex eigenvalues, then $\mathcal{N}\mathrm{Hess}(\Theta)|_{(\tilde{\theta}_1,\tilde{\theta}_2)}$ also has complex eigenvalues and their real part can be controlled through the trace.

Explicitly, the Nash Hessian is

$$\mathcal{N}\mathrm{Hess}(\Theta)|_{(\tilde{\theta}_1,\tilde{\theta}_2)} = 4\pi^2\left.\begin{pmatrix}-n_1^2\mu\Lambda^{\gamma,\delta}_{n_1,n_2} & \pm1\pm\mu n_1n_2\Lambda^{\gamma+1,\delta+1}_{n_1,n_2}\\ \mp1\mp\mu n_1n_2\Lambda^{\gamma+1,\delta+1}_{n_1,n_2} & n_2^2\mu\Lambda^{\gamma,\delta}_{n_1,n_2}\end{pmatrix}\right|_{(\tilde{\theta}_1,\tilde{\theta}_2)}$$

Therefore, its trace is given by

$$4\pi^2\mu\Lambda^{\gamma,\delta}_{n_1,n_2}(\tilde{\theta}_1,\tilde{\theta}_2)\left(n_2^2-n_1^2\right). \tag{9}$$

In particular, if $n_1 = n_2$, then the new critical point $(\tilde{\theta}_1,\tilde{\theta}_2)$ is still a center. Otherwise, the behaviour is determined by the sign of $\Lambda^{\gamma,\delta}_{n_1,n_2}(\tilde{\theta}_1,\tilde{\theta}_2)$. This sign can be read from the gradient and the Nash Hessian at (θ_1^0,θ_2^0).

To illustrate this idea, we consider a particular combination of signs. The other cases can be obtained analogously. Suppose that the first component of the gradient satisfies

$$\frac{\partial\Theta}{\partial\theta_1}(\theta_1^0,\theta_2^0) = 2\pi\mu(-1)^{\gamma} n_1 \Lambda^{\gamma+1,\delta}_{n_1,n_2}(\theta_1^0,\theta_2^0) > 0.$$

In addition, suppose that the entries of the first row of the Nash Hessian have signs

$$\begin{aligned}\left(\mathcal{N}\mathrm{Hess}\left(\Lambda^{\gamma,\delta}_{n_1,n_2}\right)|_{(\theta_1^0,\theta_2^0)}\right)_{1,1} &= -4\pi^2 n_1^2 \mu \Lambda^{\gamma,\delta}_{n_1,n_2}(\theta_1^0,\theta_2^0) > 0,\\ \left(\mathcal{N}\mathrm{Hess}\left(\Lambda^{\gamma,\delta}_{n_1,n_2}\right)|_{(\theta_1^0,\theta_2^0)}\right)_{1,2} &= \pm 1 \pm \mu n_1 n_2 \Lambda^{\gamma+1,\delta+1}_{n_1,n_2}(\theta_1^0,\theta_2^0) < 0.\end{aligned}$$

In that case, this means that $(\tilde{\theta}_1,\tilde{\theta}_2)$ has the form $(\tilde{\theta}_1,\tilde{\theta}_2) = (\theta_1^0 - \epsilon_1, \theta_2^0 + \epsilon_2)$ for a small $\epsilon_1,\epsilon_2 > 0$. Therefore, the sign of (9) is determined by the sign of $\Lambda^{\gamma,\delta}_{n_1,n_2}(\theta_1^0 - \epsilon_1, \theta_2^0 + \epsilon_2)$, which is a well-defined quantity that only depends on the particular point (θ_1^0,θ_2^0) and $\gamma,\delta \in \mathbb{Z}_2$.

4.4. Nash Flow for General Truncated Fourier Series

In the general case, the calculation is similar but more involved. To alleviate notation, let us consider the auxiliary functions:

$$\sigma^0(\theta) = \begin{cases} 0 & \text{if } \theta = 0 \text{ or } \frac{1}{2},\\ 1 & \text{if } 0 < \theta < \frac{1}{2},\\ -1 & \text{if } \frac{1}{2} < \theta < 1,\end{cases} \qquad \sigma^1(\theta) = \begin{cases} 0 & \text{if } \theta = \frac{1}{4} \text{ or } \frac{3}{4},\\ 1 & \text{if } 0 \leq \theta < \frac{1}{4} \text{ or } \frac{3}{4} \leq \theta < 1,\\ -1 & \text{if } \frac{1}{4} < \theta < \frac{3}{4}.\end{cases}$$

Notice that these maps are just the sign functions of the trigonometric functions $\sigma^0(\theta) = \mathrm{sign}(\sin(2\pi\theta))$ and $\sigma^1(\theta) = \mathrm{sign}(\cos(2\pi\theta))$, with the customary assumption that the sign function vanishes at zero. If needed, we may extend them to the whole real line by periodicity.

Now, let us consider a truncated Fourier series with arbitrary frequencies $m_1, m_2, n_1, n_2 \geq 1$ of the following form:

$$\Theta = \Lambda^{\alpha,\beta}_{m_1,m_2} + \mu\Lambda^{\gamma,\delta}_{n_1,n_2}.$$

Analogously to the previous case, the gradient of Θ at a point

$$(\theta_1^0,\theta_2^0) = \left(\frac{(2k_1+\alpha)n_1}{4m_1}, \frac{(2k_2+\beta)n_2}{4m_2}\right) \in \mathbb{T}^2$$

of the form (II) is

$$\nabla\Theta|_{(\theta_1^0,\theta_2^0)} = 2\pi\mu((-1)^{\gamma} n_1 \Lambda^{\gamma+1,\delta}_{n_1,n_2}(\theta_1^0,\theta_2^0), (-1)^{\delta} n_2 \Lambda^{\gamma,\delta+1}_{n_1,n_2}(\theta_1^0,\theta_2^0)).$$

Therefore, we again find a bifurcation of behaviour depending on whether $\nabla\Theta|_{(\theta_1^0,\theta_2^0)} = 0$. If $\nabla\Theta|_{(\theta_1^0,\theta_2^0)} = 0$, the Nash Hessian it is given by

$$\mathcal{N}\mathrm{Hess}(\Theta)|_{(\theta_1^0,\theta_2^0)} = (-1)^{k_1+k_2+\alpha+\beta} 4\pi^2 \begin{pmatrix} 0 & m_1 m_2 \\ -m_1 m_2 & 0 \end{pmatrix} + \mu\mathcal{N}\mathrm{Hess}\left(\Lambda^{\gamma,\delta}_{n_1,n_2}\right)|_{(\theta_1^0,\theta_2^0)}$$

As above, the character of this matrix depends some combinatorials of (α,β) and (γ,δ). Explicitly, we have that

$$\mathcal{N}\mathrm{Hess}(\Theta)|_{(\theta_1^0,\theta_2^0)} = 4\pi^2 \left.\begin{pmatrix} -n_1^2\mu\Lambda^{\gamma,\delta}_{n_1,n_2} & \pm m_1 m_2 \pm \mu n_1 n_2 \Lambda^{\gamma+1,\delta+1}_{n_1,n_2} \\ \mp m_1 m_2 \mp \mu n_1 n_2 \Lambda^{\gamma+1,\delta+1}_{n_1,n_2} & n_2^2\mu\Lambda^{\gamma,\delta}_{n_1,n_2} \end{pmatrix}\right|_{(\theta_1^0,\theta_2^0)}$$

When $|\mu|$ is small, $\mathcal{N}\mathrm{Hess}(\Theta)|_{(\theta_1^0,\theta_2^0)}$ has complex eigenvalues $\lambda, \overline{\lambda} \in \mathbb{C}$. Since $\lambda + \overline{\lambda} = 2\mathrm{Re}(\lambda)$, the dynamics are ruled by the real part $\mathrm{Re}(\lambda)$ which is given by the following trace:

$$4\pi^2\mu\Lambda^{\gamma,\delta}_{n_1,n_2}|_{(\theta_1^0,\theta_2^0)}\left(n_2^2 - n_1^2\right).$$

Its negativity (respectively positivity) can be controlled with the trigonometric sign functions as

$$\mu\sigma^{\gamma}(\theta_1^0 n_1)\sigma^{\delta}(\theta_2^0 n_2)\left(n_2^2 - n_1^2\right) < 0 \text{ (respectively} > 0).$$

Remark 6. *There are many cases in which this trace does not vanish. For instance, if $(\gamma,\delta) = (\alpha+1,\beta+1)$ in $\mathbb{Z}_2 \times \mathbb{Z}_2$, in general,*

$$\Lambda^{\alpha+1,\beta+1}_{n_1,n_2}\left(\frac{2k_1+\alpha}{4m_1}, \frac{2k_2+\beta}{4m_2}\right) \neq 0.$$

To be precise, given $n \in \mathbb{N}$, let us denote by $par(n)$ the unique integer such that $n = 2^{par(n)}n'$ with n' odd. In that case, we have that $\Lambda^{\alpha+1,\beta+1}_{n_1,n_2}\left(\frac{2k_1+\alpha}{4m_1}, \frac{2k_2+\beta}{4m_2}\right) = 0$ for some $k_1,k_2 \in \mathbb{Z}$ if and only if $par(m_1) = par(n_1) + (-1)^{\alpha}$ or $par(m_2) = par(n_2) + (-1)^{\beta}$. It would be interesting to study the relation between the behavior and the small divisors phenomena observed in Kolmogorov-Arnold-Moser (KAM) theory [27].

The case with $\nabla\Theta|_{(\theta_1^0,\theta_2^0)} \neq 0$ can be treated similarly, but now, we must not look at the Nash Hessian exactly at (θ_1^0,θ_2^0) but at a point nearby. Generalizing the argument of Section 4.3, set

$$A = (-1)^{\gamma}\mu n_1\sigma^{\gamma+1}(\theta_1^0 n_1)\sigma^{\delta}(\theta_2^0 n_2), \quad B_1 = \mu\sigma^{\gamma}(\theta_1^0 n_1)\sigma^{\delta}(\theta_2^0 n_2),$$

$$B_2 = (-1)^{k_1+k_2+\alpha+\beta}m_1m_2 + (-1)^{\delta+\gamma}\mu n_1 n_2\sigma^{\gamma+1}(\theta_1^0 n_1)\sigma^{\delta+1}(\theta_2^0 n_2).$$

Then, the unique critical point $(\tilde{\theta}_1,\tilde{\theta}_2)$ close to (θ_1^0,θ_2^0) has the following form:

$$(\tilde{\theta}_1,\tilde{\theta}_2) = \left(\theta_1^0 + \mathrm{sign}(AB_1)\epsilon_1, \theta_2^0 + \mathrm{sign}(AB_2)\epsilon_2\right),$$

for small enough $\epsilon_1,\epsilon_2 > 0$. Therefore, the dynamic of the critical point $(\tilde{\theta}_1,\tilde{\theta}_2)$ is determined by

$$\mu\sigma^{\gamma}((\theta_1^0 + \mathrm{sign}(AB_1)\epsilon_1)n_1)\sigma^{\delta}((\theta_2^0 + \mathrm{sign}(AB_2)\epsilon_2)n_2)\left(n_2^2 - n_1^2\right). \tag{10}$$

This quantity controls the the sign of the trace of the Nash Hessian in analogy with the analysis of Section 4.3. Therefore, if this last quantity is negative, then $(\tilde{\theta}_1,\tilde{\theta}_2)$ is a spiral attractor and, if it is positive, the point becomes a repulsor.

To illustrate the different bifurcation phenomena explained in this section, in Figure 2, the Nash follows some truncated series of low frequencies. Finally, summarizing this discussion, we obtained the following result.

Theorem 1. *For μ small enough, the truncated Fourier series*

$$\Theta = \Lambda^{\alpha,\beta}_{m_1,m_2} + \mu\Lambda^{\gamma,\delta}_{n_1,n_2},$$

has an attracting (respectively repulsive) spiral critical point at each of the points of the form (II),

$$\left(\theta_1^0,\theta_2^0\right) = \left(\frac{2k_1+\alpha}{4m_1}, \frac{2k_2+\beta}{4m_2}\right),$$

for $k_1,k_2 \in \mathbb{Z}$ provided the following:

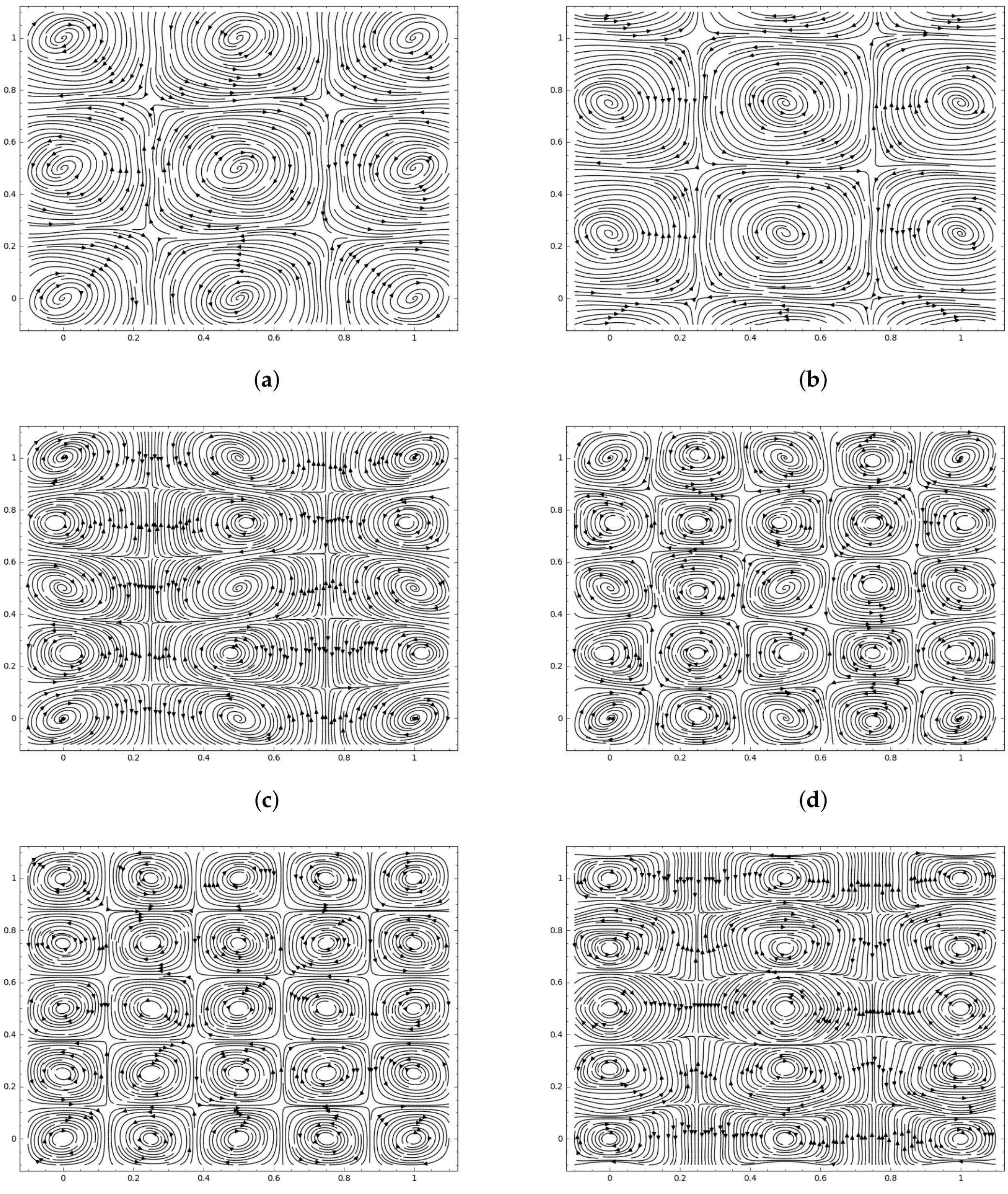
- *If* $\nabla\Theta|_{(\theta_1^0,\theta_2^0)} = 0$, *it must hold that*

$$\mu\sigma^{\gamma}(\theta_1^0)\sigma^{\delta}(\theta_2^0)\left(n_2^2 - n_1^2\right) < 0 \textit{ (respectively} > 0\textit{).}$$

- *If* $\nabla\Theta|_{(\theta_1^0,\theta_2^0)} \neq 0$, *it must hold that*

$$\mu\sigma^{\gamma}((\theta_1^0 + sign(AB_1)\epsilon_1)n_1)\sigma^{\delta}((\theta_2^0 + sign(AB_2)\epsilon_2)n_2)\left(n_2^2 - n_1^2\right) < 0 \textit{ (respectively} > 0\textit{).}$$

for $\epsilon_1, \epsilon_2 > 0$ *that is small enough.*



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 2. Nash flow dynamics of truncated Fourier series: cases (**a**–**d**) show breaking of the periodic orbits into spiral flow, and cases (**e**,**f**) preserve the periodic orbits. (**a**) $\Theta = \Lambda_{1,1}^{0,0} + 0.03\Lambda_{3,5}^{1,1}$. (**b**) $\Theta = \Lambda_{1,1}^{0,1} + 0.02\Lambda_{3,5}^{1,0}$. (**c**) $\Theta = \Lambda_{1,2}^{0,0} + 0.1\Lambda_{2,3}^{1,1}$. (**d**) $\Theta = \Lambda_{2,2}^{0,0} + 0.1\Lambda_{3,5}^{1,1}$. (**e**) $\Theta = \Lambda_{2,2}^{0,0} + 0.02\Lambda_{4,4}^{1,1}$. (**f**) $\Theta = \Lambda_{1,2}^{0,0} + 0.1\Lambda_{3,5}^{0,0}$.

Remark 7. *Even though half of the critical points near the points of the form (II) are attractors for the Nash flow of Θ, the dynamic is an small perturbation of a center. In this manner, the convergence is slow, highly spiralizing towards the Nash equilibrium. This theoretically justifies the slow and bad conditioned convergence observed in GANs networks.*

5. Empirical Analysis

In this section, we show empirically how these Fourier approximations can be useful for understanding the convergence in the training of GANs. For this purpose, in this section, we consider a simple model for a 2-parametric torus GAN (i.e., with $d_D = d_G = 1$) and we analyze its convergence by means of its truncated Fourier series.

In the notation of Section 3, we take $d = 1$ (1-dimensional real data) and the parameter spaces is $\Theta_D = \Theta_G = S^1$. The latent space is $\Lambda = [0,1] \subseteq \mathbb{R}$ with the uniform probability (standard Lebesgue measure). Fix a periodic functions $\chi : S^1 \to \mathbb{R}$. Choose a 1-parametric continuous distribution $\mathcal{D}_\xi$ depending on the parameter $\xi \in \mathbb{R}$, with cumulative distribution function F_ξ and probability density function f_ξ. Fix $\omega \in S^1$, and the real data X is sampled according to the distribution $X \sim \mathcal{D}_{\chi(\omega)}$.

As discriminator function, for $\theta_1 \in S^1$, we consider the function $D_{\theta_1} : \mathbb{R} \to \mathbb{R}$ given by

$$D_{\theta_1}(x) = \frac{f_{\chi(\omega)}(x)}{f_{\chi(\omega)}(x) + f_{\chi(\theta_1)}(x)}. \tag{11}$$

On the other hand, for $\theta_2 \in S^1$, the generator is the function $G_{\theta_2} : \Lambda = [0,1] \to \mathbb{R}$ given by

$$G_{\theta_2}(\lambda) = F^{-1}_{\chi(\theta_2)}(\lambda), \tag{12}$$

where $F^{-1}_{\chi(\theta_2)}$ is the quantile function of $\mathcal{D}_{\chi(\theta_2)}$.

With these choices of generator and discriminator and taking as weight function $f(t) = -\log(1 + \exp(-t))$, as in [1], the cost functional (1) is reduced to

$$\begin{aligned}\mathcal{F}(\theta_1, \theta_2) &= \mathbb{E}_\Omega \log\left[D_{\theta_1}(X)\right] + \mathbb{E}_\Lambda \log\left[1 - D_{\theta_1}(G_{\theta_2})\right] \\ &= \int_{\mathbb{R}} \log\left(\frac{f_{\chi(\omega)}(x)}{f_{\chi(\omega)}(x) + f_{\chi(\theta_1)}(x)}\right) f_{\chi(\omega)}(x)\, dx \\ &\quad + \int_0^1 \log\left(1 - \frac{f_{\chi(\omega)}\left(F^{-1}_{\chi(\theta_2)}(\lambda)\right)}{f_{\chi(\omega)}\left(F^{-1}_{\chi(\theta_2)}(\lambda)\right) + f_{\chi(\theta_1)}\left(F^{-1}_{\chi(\theta_2)}(\lambda)\right)}\right) d\lambda.\end{aligned} \tag{13}$$

Remark 8. *These choices of shapes for the discriminator and generator functions are justified by [1, Proposition 1]. There, it is proven that, for a fixed generator G with transformed probability density function f_G, the optimal discriminator $D_{\theta_1^0}$ is given by*

$$D_{\theta_1^0}(x) = \frac{f_{\chi(\omega)}(x)}{f_{\chi(\omega)}(x) + f_G(x)}. \tag{14}$$

On the other hand, recall that, if $\Lambda = [0,1]$ with the uniform probability, then $F_\xi^{-1} : \Lambda = [0,1] \to \mathbb{R}$ is a random variable with distribution $\mathcal{D}(\xi)$. Thus, in our case, G_{θ_2} is a random variable with distribution $\mathcal{D}_{\chi(\theta_2)}$ and, therefore, transformed density $f_{\chi(\theta_2)}$.

In this vein, the goal of the generator G given by (12) is to adjust θ_2 to reach the value $\theta_2 = \omega$, for which G generates exactly the real data. On the other side, for fixed parameter θ_2 for G, D given by (11) aims to tune θ_1 to the value $\theta_1 = \theta_2$, for which D is the perfect discriminator (14).

For the purposes of these experiments, we fix the underlying distribution $\mathcal{D}_\xi$ to be the exponential distribution with mean $1/\xi$ and $\chi(\theta) = \sin(\pi\theta)^2 + 1$. Recall that, in this situation, $f_\xi(x) = \xi e^{-\xi x}$ y $F_\xi(x) = 1 - e^{-\xi x}$. In this way, the discriminator function (11) and the generator (12) are given by

$$D_{\theta_1}(x) = \frac{e^{x\sin(\pi\theta_1)^2}}{\frac{\left(\sin(\pi\theta_1)^2+1\right)}{\left(\sin(\pi\omega)^2+1\right)} e^{x\sin(\pi\omega)^2} + e^{x\sin(\pi\theta_1)^2}}, \quad G_{\theta_2}(\lambda) = \frac{1}{\sin(\pi\theta_2)^2+1}\log\left(-\frac{1}{\lambda-1}\right). \tag{15}$$

Moreover, from now on, we fix $\omega = 1/4$, so that $\chi(\omega) = 3/4$. The resulting probability density and cumulative distribution functions of the real data are plotted in Figure 3.

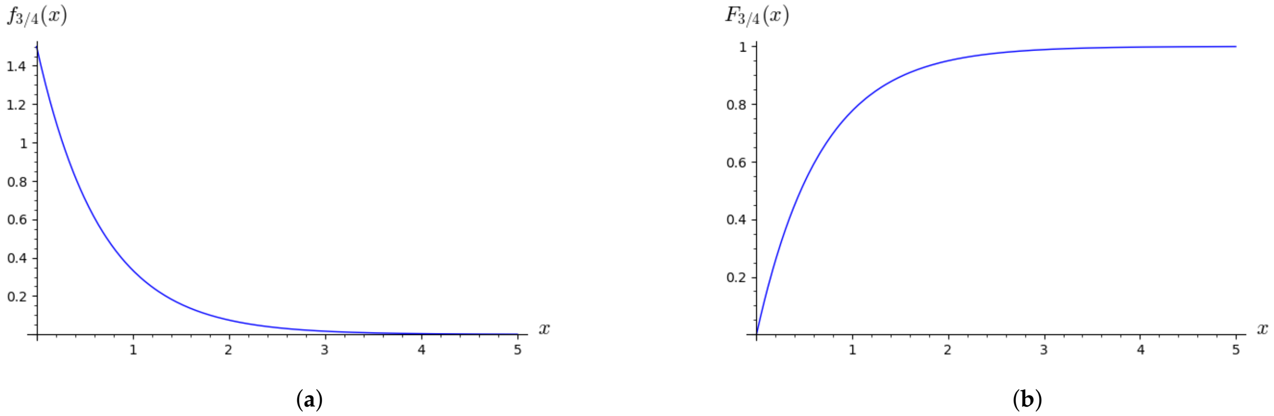


Figure 3. Distribution of the real data: (**a**) probability density function and (**b**) cumulative distribution function.

With this choice of real distribution, the generator function as well as the transformed probability density function are plotted in Figure 4 and the discriminator function is shown in Figure 5.

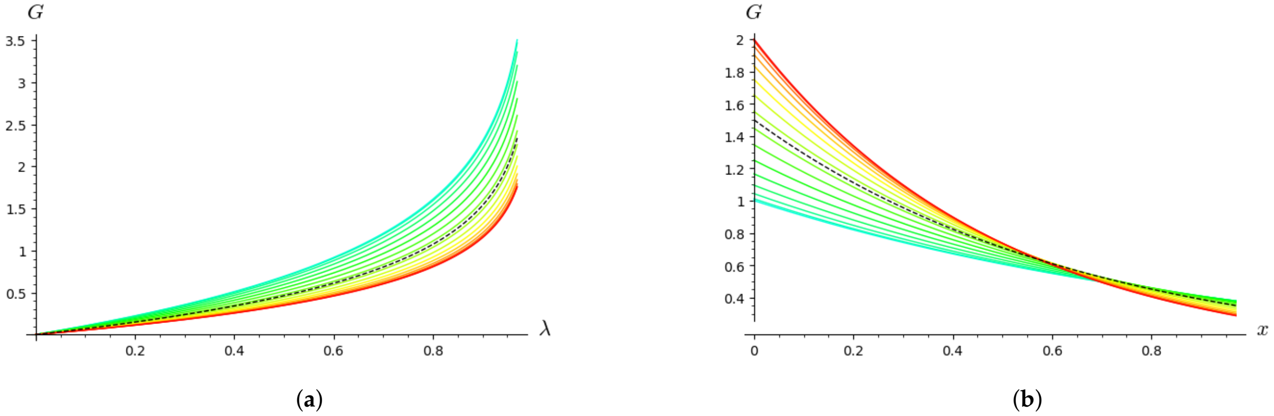


Figure 4. Generator functions for $0 \leq \theta_2 \leq \frac{1}{2}$: the warmer the plot, the bigger the value of θ_2. The dashed line corresponds to the real data. (**a**) Output of the function. (**b**) Transformed probability density function.

In addition, in Figure 6, we show graphically the cost function $\mathcal{F}(\theta_1, \theta_2)$ of (13) on $\mathbb{T}^2$. The numerical approximation of the integrals in (13) were carried out with the Simpson rule. The function was sampled at 225 knot points and subsequently interpolated by means of a multiquadratic radial basis interpolation. Observe that one of the Nash equilibria of $\mathcal{F}$ is at $(\theta_1, \theta_2) = (1/4, 1/4)$ (bottom corner of the plot). Moreover, by the symmetries of χ, the plot suggests that $(\theta_1, \theta_2) = (1/4, 3/4), (3/4, 1/4), (3/4, 3/4)$ are also Nash equilibria.

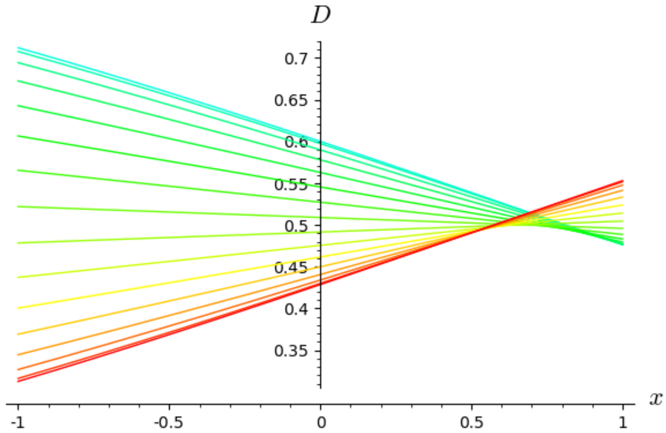


Figure 5. Discriminator functions for $0 \leq \theta_1 \leq \frac{1}{2}$: the warmer the plot, the larger the value of θ_1. For fixed generator parameter θ_2, the optimal value for θ_1 corresponds to the line with $\theta_1 = \theta_2$.

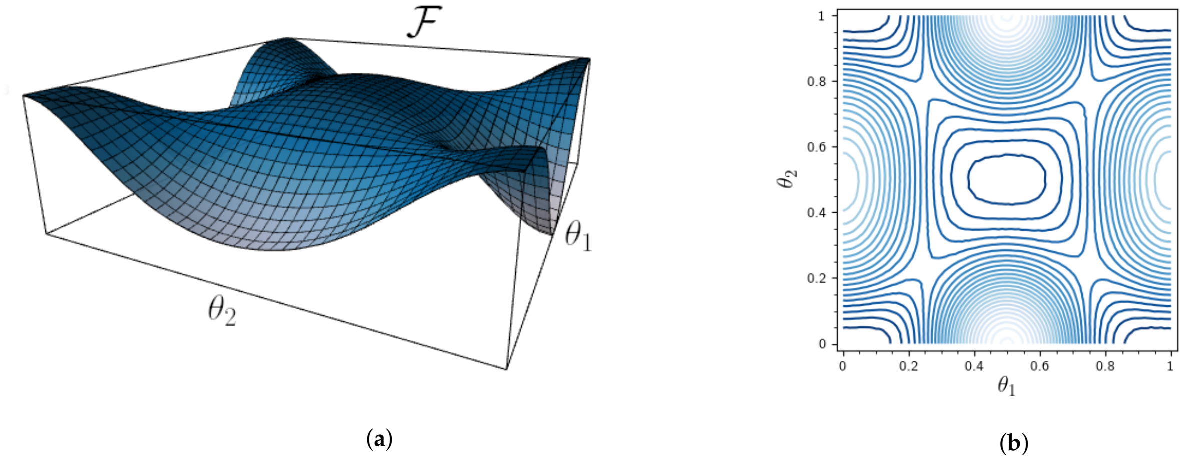


Figure 6. Graphical representation of the landscape of the cost function $\mathcal{F}(\theta_1, \theta_2) : \mathbb{T}^2 \to \mathbb{R}$. (**a**) Plot of the function $\mathcal{F}(\theta_1, \theta_2)$. The four saddle points lie near each of the four corners of the frame. (**b**) Contour plot of $\mathcal{F}(\theta_1, \theta_2)$.

In Figure 7, we show the Nash flow associated with the cost function $\mathcal{F} : \mathbb{T}^2 \to \mathbb{R}$. As can be checked in the image, the flow confirms that there exists four Nash equilibrium points, corresponding to $(\theta_1^0, \theta_2^0) = (1/4, 1/4), (1/4, 3/4), (3/4, 1/4)$, and $(3/4, 3/4)$, all of them being attractors for the Nash flow. Another four critical points of $\mathcal{F}$ can be observed in the figure: the points $(0, 0)$ and $(1/2, 1/2)$ correspond to the two maxima of $\mathcal{F}$, and the points $(0, 1/2)$ and $(1/2, 0)$ correspond to the two minima. Observe that these critical points are saddle points for the flow, with an attractive direction and a repulsive direction. Finally, notice that (4) is satisfied since the maxima and minima have even indices (2 and 0, respectively), and the Nash equilibria have odd indices.

Now, let us decompose $\mathcal{F}$ according to its Fourier series. In Table 1, we show the modes with the largest absolute Fourier coefficients. These coefficients have been computed using the formulae of Section 4 by applying rectangular quadrature as the numerical integration method and looking at the modes with $1 \leq m_1, m_2 \leq 10$.

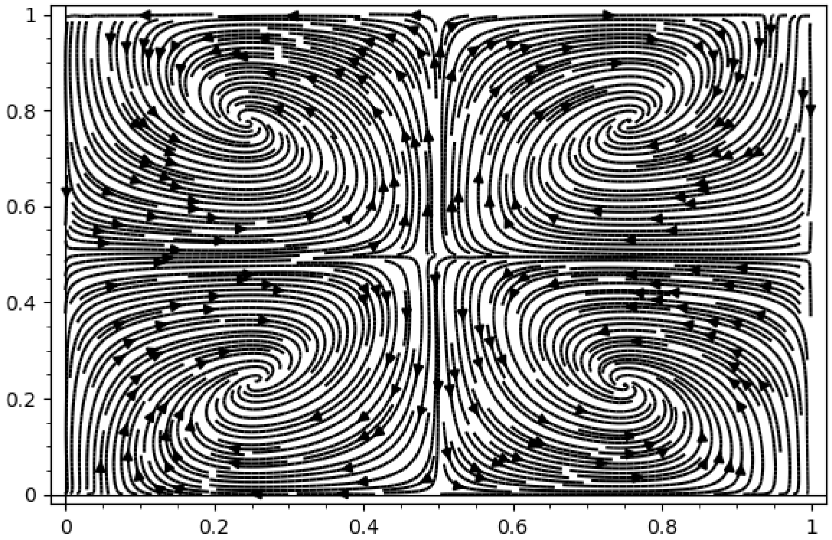


Figure 7. Dynamics of the Nash flow for the torus GAN: four attractive Nash equilibria can be observed.

Table 1. Fourier modes of the cost function for the torus GAN. The ten modes with the largest absolute value of their associated coefficient are shown. The last column shows the ratio between each Fourier coefficient and the largest coefficient.

m_1	m_2	α	β	$a^{\alpha,\beta}_{m_1,m_2}$	**Ratio**
1	1	1	1	0.06127	1.0000
1	2	1	1	0.01102	0.1800
2	1	1	1	−0.00503	−0.0822
2	2	1	1	−0.00404	−0.0660
2	3	1	1	−0.00325	−0.0532
2	4	1	1	−0.00308	−0.0504
2	5	1	1	−0.00305	−0.0499
2	7	1	1	−0.00304	−0.0497
2	9	1	1	−0.00304	−0.0496
2	10	1	1	−0.00304	−0.0496

From these results, we observe that the predominant Fourier modes of $\mathcal{F}$ are cosine basis functions, $\Lambda^{1,1}_{m_1,m_2}(\theta_1,\theta_2) = \cos(2\pi m_1\theta_1)\cos(2\pi m_2\theta_2)$. The largest coefficient corresponds to the mode $(m_1,m_2) = (1,1)$. Observe that this is not surprising: $(m_1,m_2) = (1,1)$ is the unique mode with four critical points of type (II), which correspond to the four Nash equilibria of Figure 7 (in other words, the four saddle points in Figure 6).

For $s \geq 0$, let us order the first s Fourier modes decreasingly according to the absolute value of their coefficient, $(m^0_1, m^0_2) = (1,1), (m^1_1, m^1_2), \dots, (m^s_1, m^s_2)$. Denote by $b^{1,1}_{m^i_i,m^i_2} = a^{1,1}_{m^i_i,m^i_2}/a^{1,1}_{m^0_i,m^0_2}$ the ratio of the Fourier coefficients. We can approximate the Nash flow of the cost function $\mathcal{F}$ by the truncated Fourier series:

$$\Theta_s(\theta_1,\theta_2) = \Lambda^{1,1}_{m^0_1,m^0_2}(\theta_1,\theta_2) + \sum_{i=1}^{s} b^{1,1}_{m^i_i,m^i_2}\Lambda^{1,1}_{m^i_i,m^i_2}(\theta_1,\theta_2).$$

The associated Nash flow is depicted in Figure 8. As can be checked there, the critical points nearby points of type (II) are (approximately) centers for $s \leq 3$. The reason for this behavior is twofold. In the following, let $(\theta_1^0, \theta_2^0) = (1/4, 1/4), (1/4, 3/4), (3/4, 1/4)$ or $(3/4, 3/4)$.

- For $s \leq 2$, we have that $\nabla\Theta_s|_{(\theta_1^0, \theta_2^0)} = 0$ since, in the gradient, there is always a term with a factor $\cos(2\pi\theta)$ that vanishes at these points. Hence, the critical point of Θ_s is exactly at (θ_1^0, θ_2^0). Nevertheless, since all the terms $\Lambda_{m_1,m_2}^{\alpha,\beta}$ appearing in the Fourier series have equal $(\alpha, \beta) = 1$, as mentioned in Section 4.3, we still have that the Nash Hessian has the form in (8) with vanishing diagonal entries. Hence, the critical point (θ_1^0, θ_2^0) is still a center.
- For $s = 3$, we find that $\nabla\Theta_3|_{(\theta_1^0, \theta_2^0)} \neq 0$, so a new critical point $(\tilde{\theta}_1, \tilde{\theta}_2)$ appears near (θ_1^0, θ_2^0). Nevertheless, for this new mode, we have that $m_1^3 = m_2^3 = 2$, so Equation (10) still vanishes, proving that the new critical point is still a center.

Finally, let us consider the case $s = 4$. In this situation, we also have $\nabla\Theta_4|_{(\theta_1^0, \theta_2^0)} \neq 0$, so a new critical point $(\tilde{\theta}_1, \tilde{\theta}_2)$ appears near (θ_1^0, θ_2^0). The dynamic around it is governed by Equation (10). To do so, we calculate the sign of the quantities A, B_1 and B_2 of Section 4.4 and we get

$$A > 0, \quad B_1 < 0, \quad B_2 < 0.$$

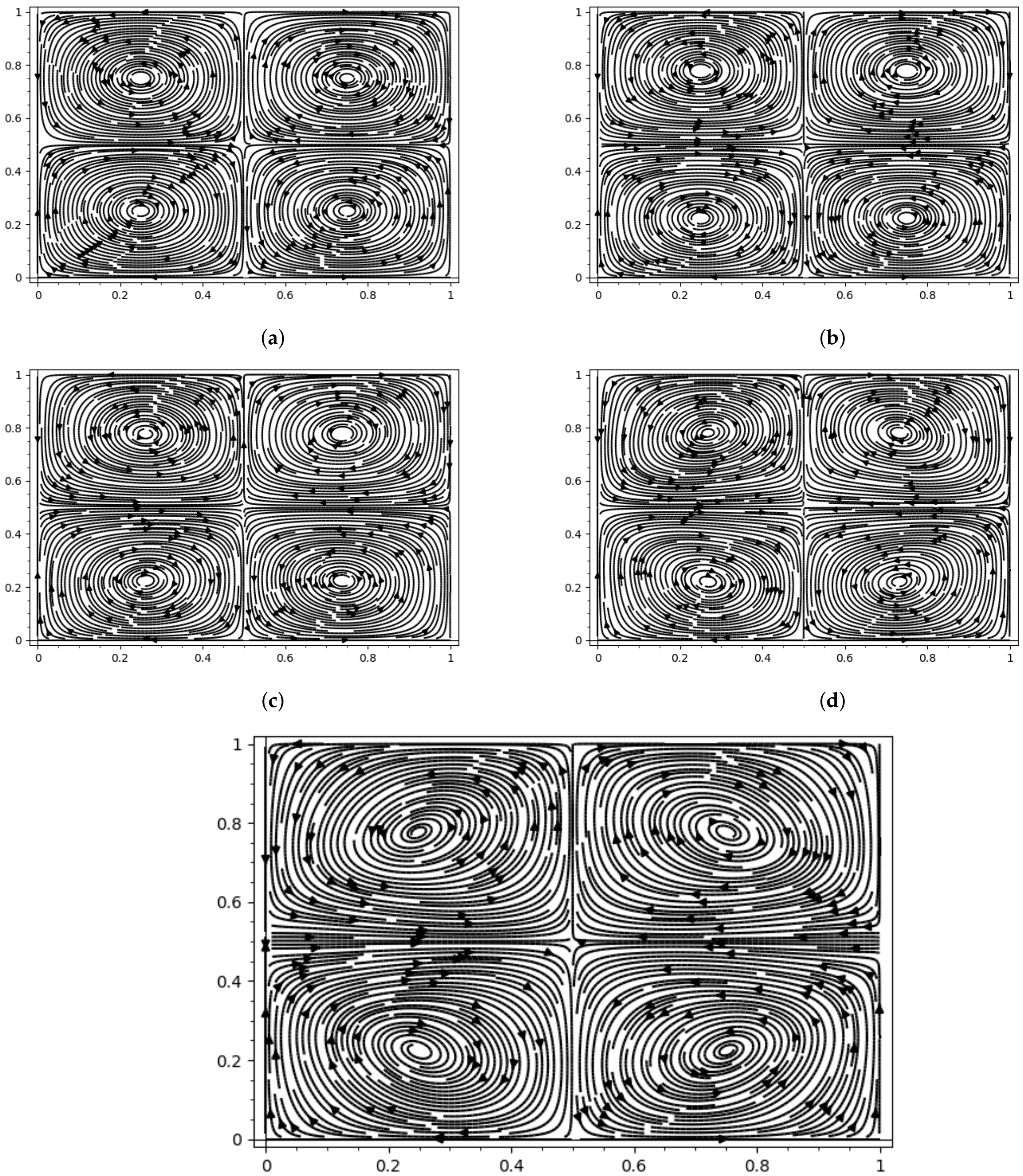
Hence, the new critical point has the form $(\tilde{\theta}_1, \tilde{\theta}_2) = (\theta_1^0 - \epsilon_1, \theta_2^0 - \epsilon_2)$ for a small $\epsilon_1, \epsilon_2 > 0$. For these values, we have that

$$\sigma^1(2(\theta_1^0 - \epsilon_1)) = -1, \quad \sigma^\delta(3(\theta_2^0 - \epsilon_2)) = -1.$$

Therefore, checking Equation (10), we get

$$\mu\sigma^1(n_1(\theta_1^0 - \epsilon_1))\sigma^\delta(n_2(\theta_2^0 - \epsilon_2)) \cdot \left(n_2^2 - n_1^2\right) = -0.003 \cdot (-1) \cdot (-1)(3^2 - 2^2) < 0.$$

Therefore, for $s = 4$, the trend changes and the centers turn into spiral attractor critical points. This is the attractive behavior observed in Figure 8e. Notice that this dynamic agrees with the real one observed in Figure 7, which empirically confirms the validity of our approach.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

Figure 8. Nash flow dynamics of truncated Fourier series approximations for the cost function of the torus GAN: (**a**) approximation Θ_0, (**b**) approximation Θ_1, (**c**) approximation Θ_2, (**d**) approximation Θ_3, and (**e**) approximation Θ_4.

6. Methodology for Practical Applications

The discussion of Sections 4 and 5 opens the door to practical application of the analysis techniques introduced in this paper to study convergence of real-world GANs. Observe that, in general, the knowledge of the underlying cost function $\mathcal{F}$ (c.f. Equation (1)) of a GAN is very limited. Indeed, several metrics have been proposed in the literature to screen the evolution of the training of the GAN. These metrics provide a way to measure indirectly the convergence of the GAN but definitely skip a thorough analysis of the cost function. Nevertheless, using the techniques introduced in this paper, we show that it is possible to methodically analyze the dynamics of the Nash flow for the GAN problem through partial sums of the Fourier series of the cost function. It is remarkable that this valuable information about the behaviour of the training process cannot be extracted from $\mathcal{F}$ itself.

In this section, we aim to organize the previous analysis into a precise methodology that can be applied in practice. As it will become clear, this process was implicit in the reasoning provided in Section 5. The proposed process of analysis comprises the following steps:

1. Evaluate cost function $\mathcal{F}(\theta_D, \theta_G)$ in a uniform grid for the parameters (θ_D, θ_G) (the weights of the two neural networks forming the GAN in the deep learning framework). Observe that, for these evaluations, it is not necessary to train the GAN networks. The sampling process amounts to fixing the weights of the networks and to computing the mean prediction error of the discriminant against real and synthetic instances. No optimization of the weights must be carried out.
2. Compute the Discrete Fourier Transform (DFT) of $\mathcal{F}$ by means of the obtained samples. This process can be done efficiently through the Fast Fourier Transform (FFT) algorithm.
3. Use the results of the DFT to estimate the Fourier modes and coefficients of $\mathcal{F}$. Sort the modes decreasingly according to the absolute value of their associated Fourier coefficient.
4. Consider a truncation level $s \geq 0$ (starting with $s = 0$). Compute the critical points of Θ_s, the truncated Fourier series of $\mathcal{F}$ with s terms. Using the techniques developed in Section 4 (see also Section 5), analyze the local dynamics of the Nash flow around the critical points of Θ_s.
5. While some of the critical points of Θ_s are a center, increase the truncation level by 1. Repeat the steps 4 and 5 until a truncation level s_0 is reached such that all the critical points of Θ_{s_0} are either attractors or repulsors.

After this process, we found a truncation level s_0 such that the local dynamics of Θ_{s_0} around the critical points are conjugated to the local dynamics of $\mathcal{F}$ around its Nash equilibria. This information can be exploited to analyze the training process of the GAN. For instance, if the convergence to the critical point is very slow, in the sense that the trace of the Nash Hessian is close to zero, then a hard convergence of the training process should be expected. This leads to remarkable instabilities during the learning process that may prevent the system from converging with a raw gradient descent optimization procedure. In that case, the obtained results strongly suggest that several heuristics for stabilizing the training process must be implemented. Additionally, since the equilibria are spiral attractors, if the learning rate of the gradient descend method is not small enough, the discrete time approximation may not converge. In that case, the information about the convergence rate in the simplified Fourier model can be used to properly anneal the learning rate, leading to a much stable convergence.

Despite the utility of the proposed methodology, it suffers several issues that must be addressed in future works to obtain an efficient analysis procedure. The first one is that the previous proposal has an obvious bottleneck: the sampling process of the cost function on the parameters (θ_D, θ_G) may require a huge number of samples due to the course of dimensionality. Nevertheless, it is important to mention that it is not necessary to use a very dense grid since we want to understand the Fourier modes of the cost function $\mathcal{F}$ and not to obtain a detailed picture of the landscape of $\mathcal{F}$. This largely alleviates the sampling process to make it feasible.

Another possible solution is to not sample on the whole (θ_D, θ_G) space but on a smaller dimensional subspace concentrating the flow. For that purpose, the GAN network can be trained and, after some epochs, the flow will have entered in a certain "convergence subspace" that encloses the long-time evolution of the flow. This subspace can be estimated by several methods, for instance by considering the subspace generated by the last $k \geq 1$ gradient vectors obtained in the training process. In that case, instead of working on the high dimensional (θ_D, θ_G)-space, we can restrict our analysis to the k-dimensional affine space generated by these vectors. This is a much smaller subspace in which the sampling process can be carried out. Nevertheless, proposing other efficient methods of sampling that enable accurate approximations of the Fourier series of $\mathcal{F}$ is an interesting topic for future work.

Another important remark is that the methodology proposed to estimate the Fourier series through the FFT is much more efficient than the quadrature methods used in Section 5. However, it also may lead to poorer estimations of the Fourier coefficients. This inaccuracy may produce errors when choosing the leading Fourier modes if their importance (absolute value of their Fourier coefficients) are similar. To avoid these problems, all the possible permutations of these similar modes (say, modes whose coefficients differ less than a fixed threshold) must be considered during the analysis of Nash flow of the Fourier series.

7. Conclusions

In this paper, we studied a novel approach to deeply analyze the converge of GAN networks on tori. This is an outstanding open problem in machine learning and deep learning that prevents GANs being suitable for use in arbitrary domains, as feature generation outside the world of image processing.

In this paper, we proposed to decompose the cost function of a GAN into its Fourier mode and to envisage the dynamics around the Nash equilibria through its truncated Fourier approximation. For that purpose, we performed a thorough analysis of the dynamics of trigonometric series with one and two terms. Roughly speaking, this analysis showed that, if we truncate the Fourier series at its first mode, all the critical points are centers surrounded by periodic orbits. When we add subtler Fourier modes to the approximation, this dynamic may be preserved or may bifurcate to give rise to spiral attractors or repulsors. This dynamic is essentially determined by the trace of the Nash Hessian of the cost function. Hence, following this idea, in this paper, we exhibited explicitly the bifurcation condition for the Nash flow of the truncated Fourier approximations. These conditions have an involved shape taking into account the monotonicity of the trigonometric functions on a neighborhood of the critical point, but eventually, the conditions are very explicit and can be easily checked. As byproduct of this analysis, we observed that, even though the Nash equilibria are stable points as proven in [4], the dynamic of the training process is close to a center and the convergence is slow and spiral.

To test this idea, we conducted an experimental analysis with a torus GAN toy-model. Through this example, we observed that the number and distribution of the critical points is determined by the first Fourier model. Nevertheless, it was necessary to reach the forth Fourier term to discover the attractive dynamics, as predicted in the GAN literature. Comparing the approximated flow with the real flow, we observed that the approximation is able to replicate not only the local but also the global dynamics of real GAN.

We expect that this work will be useful for quantifying the complexity and convergence properties of GAN. To show how this theoretical analysis can be put into practice, in Section 6, we proposed a methodology of analysis that enables a characterization of the training dynamics of real-world GANs by means of the techniques developed in this work. From the obtained information about the convergence of the learning process of the networks, several improvements for stabilizing the training can be implemented, such as a progressive reduction of the learning rate to adapt the geometry of the spiral flow.

It is worth mentioning that the results presented in this paper apply not only to torus toy-models but also to more realistic networks. It may seem at a first sight that standard GANs do not fulfil the periodicity requirement to be defined on a torus. However, in many cases, the outputs of the generator and the discriminator networks are clipped for large enough inputs. This fix is crucial to maintain several required analytic properties, as the Lipschitz condition for Wasserstein GANs [14]. After this clipping, the GAN does actually turn into a torus GAN since the generator and discriminator functions are periodic (with a large period). In this manner, most of the regular GANs used in image generation and feature generation fit in the framework introduced in this paper. This is crucial, since dynamics on a closed manifold are deeply related to the underlying topology, for instance, through the Poincaré–Hopf theorem or deeper Morse-like results.

Nevertheless, much work must be done before this project can be turned into a reality. First, in order to compute the Fourier series of the cost function, we had to sample the cost function of the GAN at a dense mesh of weights. Using this sampling, we were able to estimate the Fourier coefficients through standard quadrature techniques, as the Simpson rule. In shallow networks with few neurons, a similar approach can be applied, but for deeper networks, this dense sampling is unfeasible. For this reason, better methods for estimating the Fourier coefficients of the cost function are needed, maybe by exploding the analytical and harmonical properties of the trigonometric functions. In addition, to illustrate the method, in this paper, we carried out all the calculations on a 2-dimensional torus. The computation in higher dimensional tori may follow similar lines, but definitely a thorough analysis of the bifurcation conditions in the higher dimensional setting is not obvious.

Summarizing, in this paper, we introduced a novel method for understanding the dynamics of GANs through harmonic analysis. We showed that, despite the Nash equilibria of the GAN being stable, the convergence is a perturbation of a center and, thus, slow and complicated. The method allowed us to identify a simplified model of the dynamics that may be useful for tuning several hyperparameters of the used GANs as the learning rate of the number of epochs to be trained. We expect that this work will open the door to new methods of study of dynamics of GAN by using harmonic analysis and trascendental methods.

Author Contributions: Conceptualization, Á.G.-P.; methodology, Á.G.-P. and A.M.; software E.T. and S.G.-C.; validation, E.T. and S.G.-C.; formal analysis, Á.G.-P.; writing—original draft preparation, Á.G.-P.; writing—review and editing, A.M., E.T., and S.G.-C.; project administration, A.M.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under grant 833685 (SPIDER).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
2. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
3. Nagarajan, V.; Kolter, J.Z. Gradient descent GAN optimization is locally stable. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5585–5595.
4. Mescheder, L.M.; Geiger, A.; Nowozin, S. Which Training Methods for GANs do actually Converge? In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; pp. 3478–3487.
5. Goodfellow, I. NIPS 2016 tutorial: Generative adversarial networks. *arXiv* **2016**, arXiv:1701.00160.
6. Kusner, M.J.; Hernández-Lobato, J.M. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv* **2016**, arXiv:1611.04051.
7. Diesendruck, M.; Elenberg, E.R.; Sen, R.; Cole, G.W.; Shakkottai, S.; Williamson, S.A. Importance weighted generative networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin, Germany, 2019; pp. 249–265.
8. Antoniou, A.; Storkey, A.; Edwards, H. Data augmentation generative adversarial networks. *arXiv* **2017**, arXiv:1711.04340.
9. Arjovsky, M.; Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv* **2017**, arXiv:1701.04862.
10. Arora, S.; Ge, R.; Liang, Y.; Ma, T.; Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). *arXiv* **2017**, arXiv:1703.00573.

11. Arora, S.; Risteski, A.; Zhang, Y. Do GANs learn the distribution? some theory and empirics. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
12. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 2234–2242.
13. Roth, K.; Lucchi, A.; Nowozin, S.; Hofmann, T. Stabilizing training of generative adversarial networks through regularization. *arXiv* **2017**, arXiv:1705.09367v2.
14. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.
15. Nowozin, S.; Cseke, B.; Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv* **2016**, arXiv:1606.00709.
16. Wang, C.; Xu, C.; Yao, X.; Tao, D. Evolutionary generative adversarial networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 921–934. [CrossRef]
17. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv* **2017**, arXiv:1706.08500.
18. Snell, J.; Ridgeway, K.; Liao, R.; Roads, B.D.; Mozer, M.C.; Zemel, R.S. Learning to generate images with perceptual similarity metrics. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 4277–4281.
19. Borji, A. Pros and cons of gan evaluation measures. *Comput. Vis. Image Underst.* **2019**, *179*, 41–65. [CrossRef]
20. Milnor, J. *Lectures on the H-Cobordism Theorem*; Princeton University Press: Princeton, NJ, USA, 2015; Volume 2258.
21. Atiyah, M.F.; Bott, R. The yang-mills equations over riemann surfaces. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Sci.* **1983**, *308*, 523–615.
22. Rudin, W. *Real and Complex Analysis*; Tata McGraw-Hill Education: New York, NY, USA, 2006.
23. Du Bois-Reymond, P. Ueber die fourierschen reihen. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen* **1873**, *1873*, 571–584.
24. Kolmogorov, A. Une séries de Fourier-Lebesgue divergente partout. *CR Acad. Sci. Paris* **1926**, *183*, 1327–1328.
25. Zygmund, A. *Trigonometric Series*; Cambridge University Press: Cambridge, UK, 2002; Volume 1.
26. Gronwall, T.H. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Ann. Math.* **1919**, *20*, 292–296. [CrossRef]
27. Arnol'd, V.I. *Mathematical Methods of Classical Mechanics*; Springer Science & Business Media: Berlin, Germany, 2013; Volume 60.