Generative Adversarial Networks: Dynamics

Matias G. Delgadino matias.delgadino@math.utexas.edu
Department of Mathematics
University of Texas at Austin
Austin, TX 78712, USA

Bruno B. Suassuna bruno.b.suassuna@mat.puc-rio.br
Departamento de Matematica
Pontifícia Universidade Católica do Rio de Janeiro
Rio De Janeiro, RJ 22451-900, Brazil

Rene Cabrera rene.cabrera@math.utexas.edu
Department of Mathematics
University of Texas at Austin
Austin, TX 78712, USA

Abstract

We study quantitatively the overparametrization limit of the original Wasserstein-GAN algorithm. Effectively, we show that the algorithm is a stochastic discretization of a system of continuity equations for the parameter distributions of the generator and discriminator. We show that parameter clipping to satisfy the Lipschitz condition in the algorithm induces a discontinuous vector field in the mean field dynamics, which gives rise to blow-up in finite time of the mean field dynamics. We look into a specific toy example that shows that all solutions to the mean field equations converge in the long time limit to time periodic solutions, this helps explain the failure to converge.

Keywords: GAN, Aggregation Equation, blow-up

1 Introduction

Generative algorithms are at the forefront of the machine learning revolution we are currently experiencing. Some of the most famous types are diffusion models Sohl-Dickstein et al. (2015), generative language models Radford et al. (2018) and Generative Adversarial Networks (GAN) Goodfellow et al. (2014). GAN was one of the first algorithms to successfully produce synthetically realistic images and audio and is the topic of this article.

A guiding assumption for GAN is that the support of the distribution can be well approximated by a lower dimensional object. That is to say although $P_* \in \mathcal{P}(\mathbb{R}^K)$, we expect that the inherent correlations in data, like values of neighboring pixels in an image, drastically reduce the dimensionality of the problem. In broad terms, we expect that, in some non-specified sense, the effective dimension of the support of P_* is less or equal than a latent dimension $L \ll K$. The GAN algorithm tries to find an easy way to evaluate a continuous function from $G : \mathbb{R}^L \to \mathbb{R}^K$, which we call the generator. The objective is to make $G(Z)$ to be approximately distributed like P_*, where Z is distributed like the standard Gaussian $\mathcal{N}(0,1) \in \mathcal{P}(\mathbb{R}^L)$. To get an idea of orders of magnitude, Karras et al. (2017) creates realistic looking high resolution images of faces with $K = 1024 \times 1024 \times 3 = 3145728$ and $L = 512$.

As the word adversarial in its name suggests, the algorithm pits two Neural Networks against each other, the generator network G and the discriminator network D. The discriminator network tries to discern from the synthetic samples $G(Z)$ and the real samples $X \sim P_*$. For this purpose, the optimization over the discriminator network D is the dual formulation of a metric between

the associated synthetic data distribution $G\#\mathcal{N}$ and the real data distribution P_*. The original algorithm Goodfellow et al. (2014) used Jensen-Shannon divergence. The version we analyze in detail here is the Wasserstein-GAN (WGAN) Arjovsky et al. (2017) which uses the 1-Wasserstein distance instead. The behavior of GAN is known to be directly tied to the choice of the metric, see Section 3 for more details.

The architecture of the Neural Networks (NN) which parametrize the generator and discriminator also plays a large role in the success of the algorithms. The paradigm for architectures at the time of the first prototypes of GANs was to use Convolutional Neural Networks (CNNs) which exploit the natural spatial correlations of pixels, see for example AlexNet introduced in Krizhevsky et al. (2017). Currently, the paradigm has changed with the advent of attention networks which are more parallelizable and outperform CNNs in most benchmarks, see Vaswani et al. (2017). In this paper, we forego the interesting question of the role of NN architecture to understand in more detail the induced dynamics, see Section 2.1 for more details.

To understand the dynamics, we will follow the success of understanding the overparametrized limit in the supervised learning problem for shallow one hidden layer NN architectures Mei et al. (2018); Chizat and Bach (2018); Rotskoff and Vanden-Eijnden (2022), see also Fernández-Real and Figalli (2022); Wojtowytsch and E (2020) for reviews of these results. In a nutshell, to the first order these articles relate Stochastic Gradient Descent (SGD) parameter training to a stochastic discretization of an associated aggregation equation Bertozzi et al. (2011); Carrillo et al. (2011), and to a second order to an aggregation diffusion equation Carrillo et al. (2006). In probabilistic terms, this is akin to the law of large numbers Sirignano and Spiliopoulos (2020a) and the central limit theorem Sirignano and Spiliopoulos (2020b).

Our contribution, which is novel even in the supervised learning case, is to quantify this type of analysis. First, we show a quantitative result for the stability of the limiting aggregation equation in the 2-Wasserstein metric, see Theorem 5. The difficulty of the stability in our case is not the regularity of the activation function Chizat and Bach (2018), but instead the growth of the Lipschitz constant with respect to the size of the parameters themselves. Next, we show a quantitative convergence of the empirical process to the solution to the mean field PDE, to our knowledge this is the first of its kind in terms of a strong metric like the 2-Wasserstein metric, see Theorem 6 and Corollary 8.

Moreover, the WGAN algorithm clips the discriminator parameters after every training iteration. In the follow up work Gulrajani et al. (2017) observed numerically that it created undesirable behavior. In terms of the mean field PDE (7), the clipping of parameters induces an associated discontinuous vector field. This explains from a mathematical viewpoint the pathology mentioned before. In a nutshell, the parameter distribution can blow-up in finite time, and after that time the discriminator network loses the universal approximation capabilities, see Secion 2.4.

Failure to converge is a known problem of GAN. For instance, Karras et al. (2017) introduces a progressive approach to training higher and higher resolution pictures, effectively having hot start of the algorithm at every step. By looking at an enlightening simplified example, we can explicitly understand the long time behavior of the algorithm. In this example, any initialization eventually settles to a time periodic orbit, which implies that the generator oscillates forever, see Section 3.

1.1 Outline of the paper

The rest of the paper is organized as follows. Section 2 contains the notation and the main results: the well posedness of the mean field PDE system (7) Theorem 5, and the quantified mean field convergence Theorem 6. Section 3 contains an enlightening example of the dynamics of WGAN. Section 4 contains the proof of Theorem 5. Section 5 contains the proof of Theorem 6. Section 6

presents the conclusions and discusses some future directions for research. Appendix A recalls well posedness and approximation of differential inclusions.

2 Set up and main results

We consider a cloud of data points $\{x_i\}_{i\in I} \subset \mathbb{R}^K$, which we assume to be generated by an underlying probability measure $P_* \in \mathcal{P}(\mathbb{R}^K)$. Although we do not have direct access to P_*, we assume the cloud of data is large enough so that we can readily sample $x_i \sim P_*$ without any inherent bias.

The task is to generate approximate samples of the distribution P_* from a base underlying probability measure which is easy to sample. We consider the Gaussian distribution $\mathcal{N}(0,1) \in \mathcal{P}(\mathbb{R}^L)$ in a latent space $\mathbb{R}^L$, where L is the dimension of the latent space, which is to be chosen by the user. We will try to approximate P_* by the push forward of said base distribution $G_\Theta \# \mathcal{N}$, where $G_\Theta : \mathbb{R}^L \to \mathbb{R}^K$ is a parametric function, which is usually chosen to be a Neural Network.

To choose the parameters Θ, whose dimensionality we will set later, we consider the following optimization problem

$$\inf_\Theta d_1(G_\Theta \# \mathcal{N}, P_*),$$

where d_1 is the 1-Wasserstein distance. Although this problem seems rather straight forward, the Wasserstein distance is notorious for being difficult to calculate in high dimensions, and we do not have direct access to P_*; hence, in practice a proxy of said distance is chosen. More specifically, we approximate the dual problem

$$d_1(G_\Theta \# \mathcal{N}, P_*) = \sup_{D\in \mathrm{Lip}_1} \int_{\mathbb{R}^L} D(G_\Theta(z))\, d\mathcal{N}(z) - \int_{\mathbb{R}^K} D(x)\, dP_*(x), \tag{1}$$

by replacing the Lip_1 class of functions by the parametric function $D_\Omega : \mathbb{R}^K \to \mathbb{R}$,

$$d_1(G_\Theta \# \mathcal{N}, P_*) \sim \sup_\Omega \int_{\mathbb{R}^L} D_\Omega(G_\Theta(z))\, d\mathcal{N}(z) - \int_{\mathbb{R}^K} D_\Omega(x)\, dP_*(x).$$

The parametric function D_Ω will also be considered as a Neural Network and the parameters Ω are restricted to a compact convex set. The precise definition of G_Θ and D_Ω as Neural Networks with a single hidden layer is given bellow, letting $\sigma : \mathbb{R} \to \mathbb{R}$ denote the activation function. Since the parameters Ω are restricted to a compact set, if σ is C^1 bounded the family $\{D_\Omega\}$ is uniformly Lipschitz.

Remark 1 *The original GAN Goodfellow et al. (2014) utilizes the Jensen-Shannon divergence, which in terms of Legendre-Fenchel dual can be written as*

$$\mathbb{JS}(G_\Theta \# \mathcal{N}, P_*) = \sup_{D\in C_b(\mathbb{R}^K)} \int_{\mathbb{R}^L} \log D(G_\Theta(z))\, d\mathcal{N}(z) + \int_{\mathbb{R}^K} \log(1 - D(x))\, dP_*(x).$$

2.1 Neural Networks

For both the generator G_Θ and discriminator D_Ω, we consider the simple case of a single hidden layer, which has the universal approximation property, see Cybenko (1989). That is to say

$$G_\Theta(z) = \left(\frac{1}{N}\sum_{i=1}^N \alpha_i^1 \sigma(\beta_i^1 \cdot z + \gamma_i^1),\ \dots,\ \frac{1}{N}\sum_{i=1}^N \alpha_i^K \sigma(\beta_i^K \cdot z + \gamma_i^K) \right),$$

where the array $\Theta = (\theta_1, \cdots, \theta_N) \in ((\mathbb{R} \times \mathbb{R}^L \times \mathbb{R})^K)^N$ is given by $\theta_i = (\alpha_i^j, \beta_i^j, \gamma_i^j)_{1 \leq j \leq K}$, and D_Ω defined by

$$D_\Omega(x) = \frac{1}{M} \sum_{i=1}^{M} a_i \sigma(b_i \cdot x + c_i),$$

where the array $\Omega = (\omega_1, \cdots, \omega_M) \in (\mathbb{R} \times \mathbb{R}^K \times \mathbb{R})^M$ is given by $\omega_i = (a_i, b_i, c_i)$. To obtain rigorous quantitative estimates, throughout the paper we consider activation functions $\sigma : \mathbb{R} \to \mathbb{R}$ that are bounded in $C^2(\mathbb{R})$. The typical example being the sigmoid function

$$\sigma(u) = \frac{1}{1 + e^{-u}}.$$

Simplifying notation, we denote

$$\alpha^j \sigma(\beta^j \cdot z + \gamma^j) = \sigma(z; \theta^j) \qquad \text{with} \qquad \theta^j = (\alpha^j, \beta^j, \gamma^j) \in \mathbb{R} \times \mathbb{R}^L \times \mathbb{R}, \qquad 1 \leq j \leq K,$$

and

$$a\sigma(b \cdot x + c) = \sigma(x; \omega) \qquad \text{with} \qquad \omega = (a, b, c) \in \mathbb{R} \times \mathbb{R}^K \times \mathbb{R}.$$

Remark 2 *The mean field analysis of two hidden layers NN is also possible, see for instance Sirignano and Spiliopoulos (2022).*

2.2 Training the parameters by SGD

We follow a simplified version of parameter training algorithm which is given in the original reference Arjovsky et al. (2017), the only difference is that for comprehensibility we consider stochastic gradient descent instead of RMSProp (Tieleman (2012)), see Remark 4. We use n as the full step indexing, and l for the sub-index related to the extra training for the Discriminator's parameters. We initialize the parameters chaotically:

$$\Omega^{1,1} \sim \nu_{in}^{\otimes M} \qquad \text{and} \qquad \Theta^1 \sim \mu_{in}^{\otimes N},$$

where

$$\Omega^{1,1} \in (\mathbb{R} \times \mathbb{R}^K \times \mathbb{R})^M \qquad \text{and} \qquad \Theta^1 \in ((\mathbb{R} \times \mathbb{R}^L \times \mathbb{R})^K)^N,$$

and the initial distributions

$$\nu_{in} \in \mathcal{P}\left(\mathbb{R} \times \mathbb{R}^K \times \mathbb{R}\right) \qquad \text{and} \qquad \mu_{in} \in \mathcal{P}\left((\mathbb{R} \times \mathbb{R}^L \times \mathbb{R})^K\right)$$

are fixed independent of N and M. Of course, correlations in parameter initialization and N and M dependent initial conditions can be introduced if they were desirable.

Iteratively in n until convergence, and iteratively for $l = 2, ..., n_c$ with n_c a user defined parameter, we define

$$\Omega^{n,l} = \text{clip}\left(\Omega^{n,l-1} + h\nabla_\Omega \left(D_{\Omega^{n,l-1}}(G_{\Theta^n}(z_l^n)) - D_{\Omega^{n,l-1}}(x_l^n)\right)\right),$$

$$\Omega^{n+1,1} = \Omega^{n,n_c},$$

and

$$\Theta^{n+1} = \Theta^n - h\nabla_\Theta D_{\Omega^{n+1,1}}(G_{\Theta^n}(z_{n_c+1}^n)),$$

where the function clip stands for the projection onto $[-1, 1] \times [-1, 1]^K \times [-1, 1]$, and $h > 0$ is the learning rate which is a user chosen parameter. The families $\{x_l^n\}_{n \in \mathbb{N}, l \in \{1,...,n_c\}}$, $\{z_l^n\}_{n \in \mathbb{N}, l \in \{1,...,n_c+1\}}$ are independent $\mathbb{R}^K$ and $\mathbb{R}^L$ valued random variables distributed by P_* and $\mathcal{N}$, respectively.

Remark 3 *The clipping of the parameter is made to ensure that the discriminator network is uniformly bounded in the Lipschitz norm, to approximate Kantorovich's duality formulation* (1)*. With this in mind, we should notice that the clipping of all parameter is slightly indiscriminate. For instance, the dependence of the discriminator function with respect to the parameter a is bounded by our assumption on the activation function σ, and would not need to be clipped.*

Remark 4 *We should note that other versions of SGD like Adam or RMSProp (see Kingma and Ba (2014) and Tieleman (2012)) are preferred by users as they are considered to outperform SGD. They introduce adaptive time stepping and momentum in an effort to avoid metastability of plateaus, and falling into shallow local minima. These tweaks of SGD add another layer of complexity which we will not analyze in this paper.*

2.3 Associated Measures

Associated to each family of parameters at the iteration step n we consider the empirical measures,

$$\begin{aligned} \mu_N^n &= \frac{1}{N}\sum_{i=1}^{N} \delta_{\Theta_i^n} \in \mathcal{P}\left((\mathbb{R}\times\mathbb{R}^L\times\mathbb{R})^K\right) \\ \nu_M^n &= \frac{1}{M}\sum_{i=1}^{M} \delta_{\Omega_i^{n,1}} \in \mathcal{P}\left(\mathbb{R}\times\mathbb{R}^K\times\mathbb{R}\right). \end{aligned}$$

Abusing notation slightly and for general probability measures $\mu \in \mathcal{P}\left((\mathbb{R}\times\mathbb{R}^L\times\mathbb{R})^K\right)$ and $\nu \in \mathcal{P}\left(\mathbb{R}\times\mathbb{R}^K\times\mathbb{R}\right)$, define

$$G_\mu(z) = \left(\int_{\mathbb{R}\times\mathbb{R}^L\times\mathbb{R}} \sigma(z;\theta_1)\, d\mu_1(\theta_1),\ \ldots,\ \int_{\mathbb{R}\times\mathbb{R}^L\times\mathbb{R}} \sigma(z;\theta_K)\, d\mu_K(\theta_K)\right) \tag{2}$$

and

$$D_\nu(x) = \int_{\mathbb{R}\times\mathbb{R}^K\times\mathbb{R}} \sigma(x;\omega) d\nu(\omega), \tag{3}$$

where μ_i, for $i = 1, ..., K$, denotes the i-th marginal of μ. We should note that due to the exchangeability of the parameters, there is no loss of information from considering the pair (Θ^n, Ω^n) versus the pair (μ_N^n, ν_M^n). In fact, using the previous notations we have

$$G_{\Theta^n} = G_{\mu_N^n} \qquad \text{and} \qquad D_{\Omega^n} = D_{\nu_M^n}.$$

Hence, to understand the behavior of the algorithm in the overparameterization limit, we will center our attention on the evolution of the empirical measures. More specifically, we consider the curves $\mu \in C\left([0,\infty); \mathcal{P}\left((\mathbb{R}\times\mathbb{R}^L\times\mathbb{R})^K\right)\right)$ and $\nu \in C\left([0,\infty); \mathcal{P}\left(\mathbb{R}\times\mathbb{R}^K\times\mathbb{R}\right)\right)$ to be, respectively, the linear interpolation of μ_N^n and ν_M^n at the time values $t_n = n(h/N)$.

The choice of the scale $\Delta t = h/N$ is arbitrary, and could also be expressed in terms of M. The relationship between N, M and n_c gives rise to different mean field limits

$$n_c \frac{N}{M} \to \gamma_c \sim \begin{cases} +\infty \\ 1 \\ 0, \end{cases} \tag{4}$$

and we will obtain different behavior in terms of limiting dynamics. In this paper, we address the intermediate limit $\gamma_c \sim 1$, but we should notice that in practice it is also interesting to study when

$\gamma_c = \infty$, which assumes that the discriminator has been trained to convergence, see Section 3 for an illustrative example. For notational simplicity, we write the proof for $N = M$ and $n_c = 1$, but our methods are valid for any finite value of $\gamma_c \sim 1$.

Explicitly, for any $t \in [0, \infty)$, we find $n \in \mathbb{N}$ and $s \in [0, 1)$ such that

$$(1 - s)t_n + st_{n+1} = t$$

and set the intermediate value as the 2-Wasserstein geodesics:

$$\mu_N(t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{(1-s)\theta_i^n + s\theta_i^{n+1}} \qquad \text{and} \qquad \nu_N(t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{(1-s)\omega_i^{n,1} + s\omega_i^{n+1,1}}. \tag{5}$$

2.4 Identifying the limit

For a given pair of measures μ and ν, consider the energy functional:

$$E[\mu, \nu] = \int_{\mathbb{R}^L} D_\nu(G_\mu(z))\, d\mathcal{N}(z) - \int_{\mathbb{R}^K} D_\nu(x)\, dP_*(x). \tag{6}$$

The evolution of the limit can be characterized by the gradient descent of E on μ and gradient ascent on ν, the latter restricted to $\mathcal{P}([-1, 1] \times [-1, 1]^K \times [-1, 1])$. In terms of equations we consider

$$\begin{cases} \partial_t \mu - \nabla_\theta \cdot \left(\mu\, \nabla_\theta \frac{\delta E}{\delta \mu}[\mu, \nu] \right) = 0, \\ \partial_t \nu + \gamma_c \nabla_\omega \cdot \left(\nu\, \mathrm{Proj}_{\pi_Q} \nabla_\omega \frac{\delta E}{\delta \nu}[\mu, \nu] \right) = 0, \\ \mu(0) = \mu_{in}, \qquad \nu(0) = \nu_{in}, \end{cases} \tag{7}$$

where we define $Q = [-1, 1] \times [-1, 1]^K \times [-1, 1]$ and the first variations are

$$\begin{aligned} \frac{\delta E}{\delta \mu}[\mu, \nu](\theta) &= \int_{\mathbb{R}^L} \int_Q \nabla_1 \sigma(G_\mu(z); \omega) \cdot (\sigma(z; \theta_1), \ldots, \sigma(z; \theta_K))\, d\nu(\omega)\, d\mathcal{N}(z), \\ \frac{\delta E}{\delta \nu}[\mu, \nu](\omega) &= \int_{\mathbb{R}^L} \sigma(G_\mu(z); \omega)\, d\mathcal{N}(z) - \int_{\mathbb{R}^K} \sigma(x; \omega)\, dP_*(x) \end{aligned}$$

and $\mathrm{Proj}_{\pi_Q} : Q \times (\mathbb{R} \times \mathbb{R}^K \times \mathbb{R}) \to \mathbb{R} \times \mathbb{R}^K \times \mathbb{R}$ is the projection onto the tangent cone $\pi_Q(\omega)$. In the present case the projection can be defined by components as follows:

$$\mathrm{Proj}_{\pi_Q}(\omega, V)_l = \begin{cases} V_l, & \omega_l \in (-1, 1) \\ V_l \frac{1 - \mathrm{sign}(V_l \omega_l)}{2}, & \omega_l \in \{-1, 1\} \end{cases}. \tag{8}$$

We should notice in fact that the projection is trivial away from the boundary, or if the vector field at the boundary points into the domain. Effectively, the projection does not allow for mass to exit the domain. We do note that this can easily make mass collapse onto the boundary and flatten the support of the distribution ν into less dimensions, see Section 3 for a further discussion.

In the context of ODEs, the projection onto convex sets was considered by Henry (1973), which we recall and expand on Appendix A. For Hilbert spaces setting, we mention the more general sweeping processes introduced by Mureau Moreau (1977). Recently, projections of solutions to the continuity equation onto semi-convex subsets have been considered as models of pedestrian dynamics with density constraints, see for instance Di Marino et al. (2016); Santambrogio (2018); De Philippis et al. (2016).

2.5 Main Result

We start by showing that the mean field parameter dynamics with a discontinuous vector fields are well defined and stable. We quantify all the results with respect to the Wasserstein distance, with d_2 and d_4 representing the standard 2-Wasserstein and 4-Wasserstein distance, respectively.

Theorem 5 *Given initial conditions $(\mu_{in}, \nu_{in}) \in \mathcal{P}\left((\mathbb{R} \times \mathbb{R}^L \times \mathbb{R})^K\right) \times \mathcal{P}(Q)$ such that for some $\delta > 0$*

$$\int e^{\delta|\alpha|^2} \, d\mu_{in} < \infty, \tag{9}$$

there exists a unique absolutely continuous weak solution to the mean field system (7).

Moreover, we have the following stability estimate: For any $T \in [0, \infty)$, there exists $C > 1$ such that

$$\sup_{t\in[0,T]} d_2((\mu_1(t), \nu_1(t)), (\mu_2(t), \nu_2(t))) \leq C d_4^2((\mu_{1,in}, \nu_{1,in}), (\mu_{2,in}, \nu_{2,in})), \tag{10}$$

for any pair of weak solutions (μ_1, ν_1) and (μ_2, ν_2).

The proof of Theorem 5 is given in Section 4, see Proposition 10 for a precise dependence of the constants. Our main result is the following estimate on the continuous time approximation of parameter dynamics.

Theorem 6 *Let $(\mu_N(t), \nu_N(t))$ be the empirical measures associated to the continuous time interpolation of the parameter values, assumed to be initialized by independent samples from $(\mu_{\text{in}}, \nu_{in})$ given by (5). Consider $(\hat{\mu}_N(t), \hat{\nu}_N(t))$ the unique solution to the PDE (7) with random initial conditions $(\mu_N(0), \nu_N(0))$. If μ_{in} has bounded double exponential moments on α, that is to say for some $\delta > 0$*

$$\mathbb{E}_{\mu_{in}}\left[e^{e^{\delta|\alpha|^2}}\right] < \infty, \tag{11}$$

then for any fixed time horizon $T \in [0, \infty)$ there exists $C > 0$ such that

$$\sup_{t\in[0,T]} \mathbb{E} d_2^2((\mu_N(t), \nu_N(t)), (\hat{\mu}_N(t), \hat{\nu}_N(t))) \leq \frac{C}{N}. \tag{12}$$

Remark 7 *The need for (11) stems from the linear dependence of the Lipschitz constant of the mean field vector field with respect to the size of the parameters, see Lemma 13.*

The proof of Theorem 6 is presented in Section 5. Using the convergence Theorem 6 and the stability of the mean field Theorem 5, we can obtain a convergence rate estimate which suffers the curse of dimensionality.

Corollary 8 *Under the hypotheses of Theorem 5 and Theorem 6, for any fixed $T > 0$, there exists $C > 0$ such that*

$$\max_{t\in[0,T]} \mathbb{E} d_2^2((\mu(t), \nu(t)), (\mu_N(t), \nu_N(t))) \leq \frac{C}{N^{\frac{2}{K(L+2)}}} \tag{13}$$

where (μ, ν) is the unique solution of (7) and (μ_N, ν_N) is the curve of interpolated empirical measures associated to the parameter training (5).

Remark 9 *We should note that the difference between the results of Theorem 6 and Corollary 8 is that the estimate (12) does not suffer from the curse of dimensionality, while the stronger estimate (13) does. The later dependence on dimension is typical and sharp for the approximation of the*

Wasserstein distance with sampled empirical measures, see Dudley (1978); Fournier and Guillin (2015); Bolley et al. (2007). This stiff dependence on dimension suggests that studying the long time behavior of the mean field dynamics of smooth initial data $(\mu(t), \nu(t))$ is not necessarily applicable in practice. Instead, the focus should be to show that with high probability that discrete mean field trajectories $(\hat{\mu}_N(t), \hat{\nu}_N(t))$ converge to a desirable saddle point of the dynamics. See Section 3 for an explicit example of long time behavior.

Proof [Proof of Corollary 8] We consider the auxiliary pair of random measure-valued paths $(\hat{\mu}_N, \hat{\nu}_N)$ which are a solution to (7) with stochastic initial conditions $(\mu_N(0), \nu_N(0))$, that is

$$\hat{\mu}_N(0) = \mu_N(0) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_{i,in}} \qquad \text{and} \qquad \hat{\nu}_N(0) = \nu_N(0) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\omega_{i,in}},$$

where $\theta_{i,in}$ and $\omega_{i,in}$ are N independent samples from μ_{in} and ν_{in}, respectively.

By the large deviation estimate in Fournier and Guillin (2015), for q large enough we have

$$\mathbb{E}[d_4^4((\hat{\mu}_N(0), \hat{\nu}_N(0)), (\mu_{in}, \nu_{in}))] \leq C M_q^{\frac{4}{q}} \left(\frac{1}{N^{\frac{4}{K(L+2)}}} + \frac{1}{N^{\frac{q-4}{q}}} \right),$$

where M_q denotes the q-th moment of $\mu_{in} \otimes \nu_{in}$. By Theorem 5, taking q large enough, and using that $\mu_{in} \otimes \nu_{in}$ has finite moments of all orders we have

$$\mathbb{E}[d_2^2((\hat{\mu}_N(t), \hat{\nu}_N(t)), (\mu(t), \nu(t))] \leq C\mathbb{E}[d_4^2((\hat{\mu}_N(0), \hat{\nu}_N(0)), (\mu(0), \nu(0)))] \leq \frac{C}{N^{\frac{2}{K(L+2)}}}.$$

By the triangle inequality,

$$d_2((\mu(t), \nu(t)), (\mu_N(t), \nu_N(t))) \leq d_2((\mu_N(t), \nu_N(t)), (\hat{\mu}_N(t), \hat{\nu}_N(t)) + d_2((\hat{\mu}_N(t), \hat{\nu}_N(t)), (\mu(t), \nu(t)),$$

so the result follows by the previous estimate and Theorem 6. ■

3 Mode Collapse and Oscillatory Behavior

A standard problem of GANs is known as mode collapse, Srivastava et al. (2017); Metz et al. (2016); Thanh-Tung and Tran (2020). This can be broadly described as the generator outputting only a small subset of the types of clusters that are present in the original distribution. Although the generator outputs a convincing sample if considered individually, the overall distribution of samples is off. An extreme example is when the generator outputs almost identical samples for any value of the latent variable z.

An explanation of this behavior for the original GAN algorithm is the use of Jensen-Shannon divergence $\mathbb{JS}(G\#\mathcal{N}, P_*)$, see Remark 1. More specifically, if the measures are mutually singular $G\#\mathcal{N} \perp P_*$, then $\mathbb{JS}(G\#\mathcal{N}, P_*) = \log(2)$ independently of how close the supports are to each other. Namely, the gradient of the associated loss function vanishes and there are no local incentives for the generator to keep learning. As we are not expecting for the support of these measures to be absolutely continuous, in fact we are postulating that in some sense the dimension of the support of $G\#\mathcal{N}$ is smaller than $L \ll K$, this case is more likely to be normal than the exception.

The W-GAN Arjovsky et al. (2017) and its improved variant Gulrajani et al. (2017) try to fix this by considering 1-Wasserstein distance instead which does not suffer from the vanishing gradient

problem. Still, training the Generator to get useful outputs is not an easy task, it requires a lot of computation time and more often than not it fails to converge. For instance to produce realistic looking images Karras et al. (2017) took 32 days of GPU compute time, and the networks are trained on progressively higher and higher resolution images to help with convergence.

3.1 An explicit example

Consider a bimodal distribution as the toy example:

$$P_* = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1 \in \mathcal{P}(\mathbb{R}).$$

We consider the simplest network that can approximate this measures perfectly. We consider the generator, depending on a single parameter $g \in \mathbb{R}$ to be given by

$$G(z, g) = \begin{cases} -1 & x < g \\ 1 & x > g. \end{cases}$$

Although, this generator architecture seems far from our assumptions 2.1. This type of discontinuity arises naturally as a limit when the parameters go to infinity. Namely, if we take b, $c \to \infty$ in such a way that $c/b \to g \in \mathbb{R}$, then

$$\sigma(bx + c) \to \begin{cases} 0 & x < g \\ 1 & x > g, \end{cases}$$

where σ is the sigmoid. The generator G can then be recovered as a linear combination of two such limits. The generated distribution is given by

$$G_g \# P = \Phi(g)\delta_{-1} + (1 - \Phi(g))\delta_1,$$

where $\Phi(g) = P(\{z < g\})$ is the cumulative distribution function of the prior distribution $P \in \mathcal{P}(\mathbb{R})$, which we can chose. We make the choice of the cumulative distribution

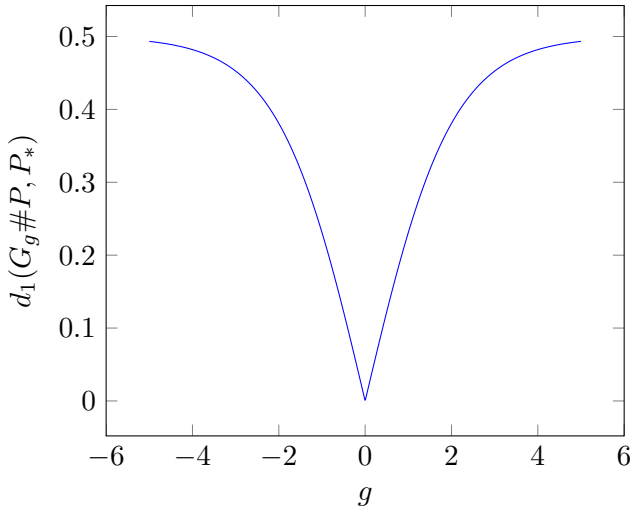
$$\Phi(g) = \frac{1}{1 + e^{-g}} \qquad \text{for } g \in \mathbb{R}$$

to simplify the calculations. Under this choice for $g = 0$, we have that $G_g \# P = P_*$, hence the network can approximate the target measure perfectly.

Moreover, we can explicitly compute the 1-Wasserstein distance,

$$d_1(G_g \# P, P_*) = \left| \frac{1}{2} - \Phi(g) \right|,$$

see the figure below.



We can clearly see that this function has a unique minimum at $g_* = 0$, and also that this function is concave in g away from $g = g_*$. The concavity of the functional makes the problem more challenging from the theoretical perspective and it will explain the oscillatory behavior of the algorithm close to the minimizer g_*.

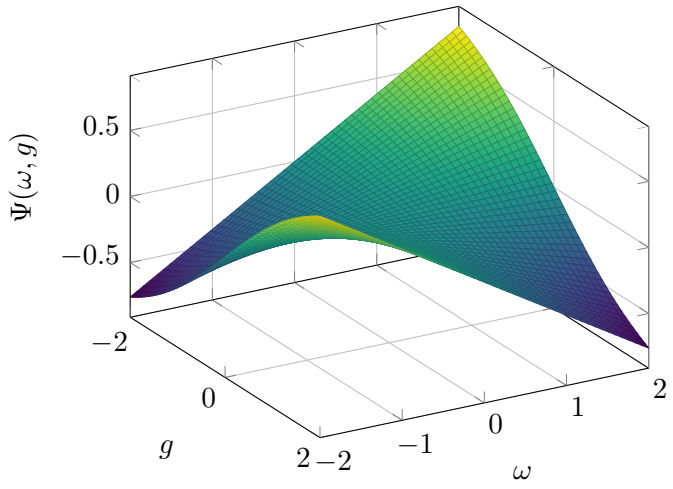
For the discriminator, we consider a ReLU activation given by

$$D(x;\omega) = (\omega x)_+$$

with $\omega \in [-1, 1]$. We note that taking a single parameter, instead of a distribution, for the discriminator is supported by the mean field dynamics (7). In the sense that under a bad initialization of parameters, the parameters of the discriminator can blow up in finite time to $\nu = \delta_{\omega(t)}$.

We consider the joint dependence function

$$\begin{aligned}
\Psi(\omega, g) &= \int_{\mathbb{R}} D_\omega(G_g(z))\, dP(z) - \int_{\mathbb{R}} D_\omega(x)\, dP_*(x) \\
&= \Phi(g)(-\omega)_+ + (1 - \Phi(g))(\omega)_+ - \frac{1}{2}(-\omega)_+ - \frac{1}{2}(\omega)_+ \\
&= \left(\frac{1}{2} - \Phi(g)\right)\omega.
\end{aligned}$$

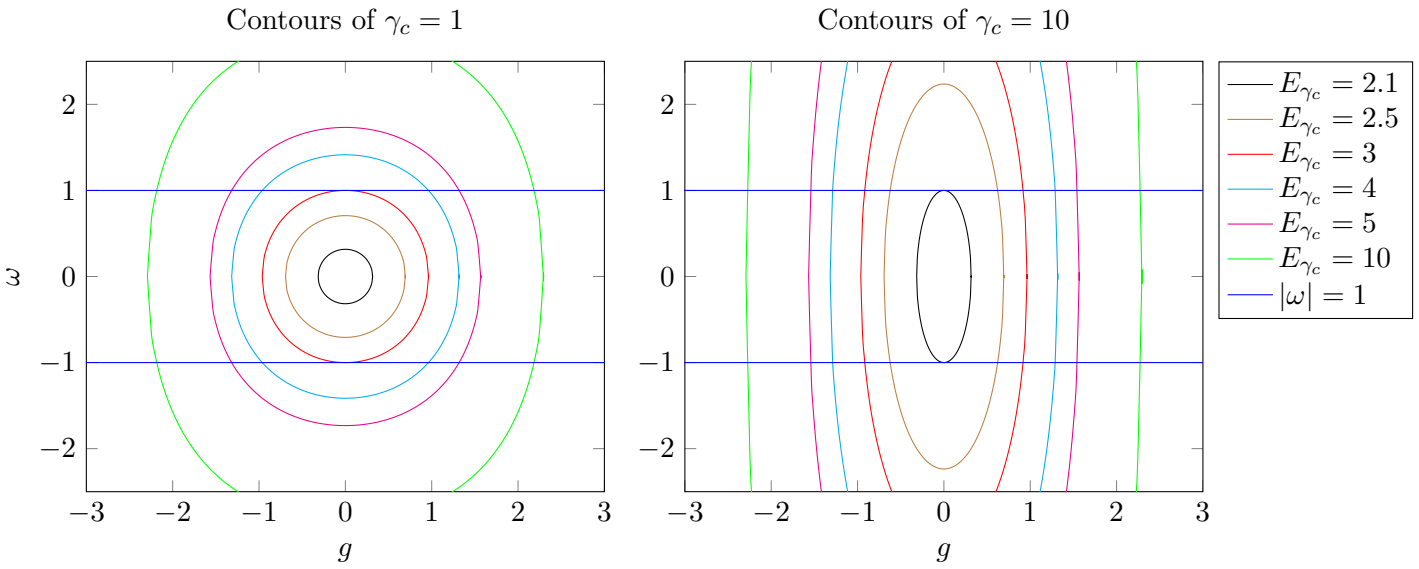


Ignoring, for now, the projection onto $\omega \in [-1, 1]$, we have the dynamics

$$\begin{cases} \dot{g}(t) &= -\nabla_g \Psi[g, \omega] &= \dfrac{1}{2}\dfrac{e^{-g}}{(1+e^{-g})^2}\omega \\ \dot{\omega}(t) &= \gamma_c \nabla_\omega \Psi[g, \omega] &= \dfrac{\gamma_c}{2}\dfrac{e^{-g}-1}{1+e^{-g}}, \end{cases}$$

where γ_c is the critics speed up (4). These dynamics can be integrated perfectly, to obtain that

$$E_{\gamma_c}(\omega(t), g(t)) = 2\cosh(g(t)) + \frac{|\omega(t)|^2}{\gamma_c} = 2\cosh(g_{in}) + \frac{|\omega_{in}|^2}{\gamma_c} = E_{\gamma_c}(\omega_{in}, g_{in}).$$



In the figure above, we plot the level sets of E_{γ_c} as well as the restriction of $|\omega| \leq 1$. We notice that, given the value of γ_c there exists a unique level set

$$E_*(\gamma_c) = 2 + \frac{1}{\gamma_c}$$

such that the level set $\{E_{\gamma_c} = E_*\}$ is tangent to the restriction $|\omega| \leq 1$.

Now, we consider the dynamics with the restriction $|\omega| \leq 1$. We notice that for any initial conditions (ω_{in}, g_{in}) satisfying $E_{\gamma_c}(\omega_{in}, g_{in}) \leq E_*(\gamma_c)$ the trajectory of parameters is unaffected by the restriction $|\omega| \leq 1$ and it is time periodic. On the other hand, if we consider initial conditions (ω_{in}, g_{in}) satisfying $E_{\gamma_c}(\omega_{in}, g_{in}) > E_*(\gamma_c)$ and $|\omega_{in}| \leq 1$, the trajectory will follow the unconstrained dynamics until it hits the boundary of the restriction $\omega(t) \in \partial Q = \{|\omega| = 1\}$. Then it follows on the boundary $\omega(t) \in \partial Q = \{|\omega| = 1\}$ until it reaches the point $(\omega(t_*), g(t_*)) = (\pm 1, 0)$ on the tangential level set $\{E_{\gamma_c} = E_*\}$ and start following this trajectory becoming time periodic. Hence, there exists $t_* = t(E_{\gamma_c}(\omega_{in}, g_{in}))$ large enough, such that the trajectory $(\omega(t), g(t)) \in \{E_{\gamma_c}(\omega_{in}, g_{in}) = E_*\}$ for $t > t_*$. Therefore, we can conclude that

$$|g(t)| \leq \cosh^{-1}\left(1 + \frac{1}{2\gamma_c}\right) \qquad \forall t > t_*.$$

Looking back at the figure, we can see that for $\gamma_c = 1$ that the limiting trajectory is $\{E_1 = 3\}$, and that the generator parameter oscillates in the range $|g(t)| \leq 0.96$ for $t > t_*$. While for $\gamma_c = 10$,

we obtain that the limiting trajectory is $\{E_{10} = 2.1\}$ and the limiting oscillations are smaller $|g(t)| \leq 0.31$ for $t > t_*$.

We do notice that regardless of the parameter γ_c and the initial configuration, the limiting trajectory is always periodic in time. In fact, we expect that every trajectory of the mean field dynamics settles into a periodic solution.

4 Properties of the mean field

One of the main theoretical obstructions to understand the well-posedness of this flow is that the projection operator Proj_{π_Q} induces a discontinuous vector field in (7). Nevertheless, the convexity of the domain $Q = [-1, 1] \times [-1, 1]^K \times [-1, 1]$ can be leveraged to obtain a stability estimate.

Given a time dependent continuous vector field $V : [0, \infty) \times Q \to \mathbb{R} \times \mathbb{R}^K \times \mathbb{R}$, its projection $\text{Proj}_{\pi_Q} V_t$ is a Borel measurable vector field which is square integrable in space and time for any finite time horizon $T > 0$ and curve of probability measures $\nu \in C([0, \infty), \mathcal{P}(Q))$,

$$\int_0^T \left(\int_Q |\text{Proj}_{\pi_Q} V_t|^2 \, d\nu_t \right) \, dt < \infty.$$

Hence, as long as the underlying velocity field inducing the motion is continuous, we can consider the notion of weak solution for the continuity equation given by (Ambrosio et al., 2005, Chapter 8).

With this in mind, we first notice the Lipschitz continuity properties of the vector fields that induce the motion (7). More specifically, we denote by

$$V^{\Theta}_{(\mu,\nu)}(\theta) = -\nabla_\theta \frac{\delta E}{\delta \mu}[\mu, \nu](\theta) = \mathbb{E}_z v^{\Theta}_{(\mu,\nu)}(\theta, z) \tag{14}$$

and

$$V^{\Omega}_{(\mu,\nu)}(\omega) = \nabla_\omega \frac{\delta E}{\delta \nu}[\mu, \nu](\omega) = \mathbb{E}_z \mathbb{E}_x v^{\Omega}_{(\mu,\nu)}(\omega, z, x), \tag{15}$$

where we define the vector fields

$$v^{\Theta}_{(\mu,\nu)}(\theta, z) = -\nabla_\theta \int_{[-1,1]^{1+K+1}} \nabla_1 \sigma(G_\mu(z); \omega) \cdot (\sigma(z; \theta_1), \, ..., \sigma(z; \theta_K)) \, d\nu(\omega) \tag{16}$$

and

$$v^{\Omega}_{(\mu,\nu)}(\omega, z, x) = \nabla_\omega [\sigma(G_\mu(z); \omega) \; - \sigma(x; \omega)]. \tag{17}$$

In Lemma 13 below, we show that $V^{\Theta}_{(\mu,\nu)}(\theta)$ and $V^{\Omega}_{\mu,\nu}(\omega)$ are Lipschitz continuous with respect to the dependence of arguments θ, ω as well as the measure arguments (μ, ν). Notice that V^{Ω} and v^{Ω} do not depend on ν, only on μ.

By (Ambrosio et al., 2005, Theorem 8.2.1), any continuous solution to the continuity equation (7) is supported over solutions of the associated characteristic field. Using the classical theory Henry (1973) for projected ODE flows, we can show that the characteristic equations

$$\begin{cases} \frac{d}{dt}(\theta, \omega) = (V^{\Theta}_{(\mu,\nu)}(\theta_i), \text{Proj}_{\pi_Q(\omega_i)} V^{\Omega}_{(\mu,\nu)}(\omega_i)) \\ (\theta, \omega)(0) = (\theta_{in}, \omega_{in}) \end{cases} \tag{18}$$

have a unique solution. More specifically, an absolutely continuous curve $(\mu, \nu) \in AC([0, \infty); \mathcal{P}\left((\mathbb{R}^{L+2})^K\right) \times \mathcal{P}(Q))$ is a weak solution to (7), if it is given as the image of the initial distributions (μ_{in}, ν_{in}) through the unique projected ODE flow. That is to say,

$$(\mu, \nu)(t) = \Phi^t_{(\mu,\nu)} \# (\mu_{in}, \nu_{in}) \tag{19}$$

the family of continuous mappings $\Phi^t_{\mu,\nu} : (\mathbb{R} \times \mathbb{R}^L \times \mathbb{R})^K \times Q \to (\mathbb{R} \times \mathbb{R}^L \times \mathbb{R})^K \times Q$ given by

$$\Phi^t_{(\mu,\nu)}(\omega_{in}, \theta_{in}) = (\omega(t), \theta(t)), \tag{20}$$

where $(\omega(t), \theta(t))$ is the unique Lipschitz solution to (18).

One of the main technical hurdles is that the vector fields inducing the motion are only locally Lipschitz. The Lipschitz constant depends itself on the size of α, which is the first variable of θ. Hence, to obtain a stability estimates we need to measure the distance of the initial condition in a p-Wasserstein distance with $p > 2$. The choice of $p = 4$ in the following result is arbitrary.

Proposition 10 (Stability) *Assume $(\mu_1, \nu_1), (\mu_2, \nu_2) \in AC\left([0,\infty); \mathcal{P}\left((\mathbb{R}^{L+2})^K\right) \times \mathcal{P}(Q)\right)$ are weak solutions to (7) which satisfies (19). Assume that the initial distribution has bounded exponential moments in the following sense: there exists $\delta > 0$ such that*

$$\int e^{\delta|\alpha|^2}\, d\mu_{1,in}, \int e^{\delta|\alpha|^2}\, d\mu_{2,in} < \infty.$$

Then for any $t > 0$ we have the bound

$$d_2((\mu_1(t), \nu_1(t)), (\mu_2(t), \nu_2(t))) \leq A(t) e^{B(t)} d_4^2((\mu_{1,in}, \nu_{1,in}), (\mu_{2,in}, \nu_{2,in})), \tag{21}$$

where

$$A(t) = e^{C\left(t^2 + t\Lambda\right)} \left(\int e^{Ct|\alpha|} d\mu_{1,\text{in}} + \int e^{Ct|\alpha|} d\mu_{2,\text{in}} \right)^{1/2},$$

and

$$B(t) = CtA(t)\left(t + \Lambda\right),$$

with

$$\Lambda = 1 + \left(\int |\alpha|^2 d\mu_{1,\text{in}} \right)^{1/2} + \left(\int |\alpha|^2 d\mu_{2,\text{in}} \right)^{1/2}$$

and $C > 0$ a constant that only depends on $\|\sigma\|_{C^2}$.

Remark 11 *The double exponential growth on the estimate is related to the dependence Lipschitz constant of the vector field with respect to the size of the parameters themselves, see Lemma 13 for the specific estimates.*

For discrete initial conditions, existence to (7) follows from applying the results in Appendix A. Using stability, we can then approximate the initial condition by taking discrete approximations of it.

Proposition 12 (Existence) *For any initial condition $(\mu_{in}, \nu_{in}) \in \mathcal{P}\left((\mathbb{R}^{L+2})^K\right) \times \mathcal{P}(Q)$ satisfying that there exists $\delta > 0$ such that*

$$\int e^{\delta|\alpha|^2}\, d\mu_{in} < \infty,$$

there exists $(\mu, \nu) \in AC\left([0,\infty); \mathcal{P}\left((\mathbb{R}^{L+2})^K\right) \times \mathcal{P}(Q)\right)$ a weak solution to (7) which satisfies the mild formulation (19).

Proof [Proof of Proposition 12] For any $L \in \mathbb{N}$ we consider a deterministic discretization

$$\mu^L_{in} = \sum_{i=1}^L w_i \delta_{\theta^i_L} \qquad \text{and} \qquad \nu^L_{in} = \sum_{i=1}^L v_i \delta_{\omega^i_L}$$

of the initial conditions μ_{in}, ν_{in}, where w_i and v_i are weights which add up to 1. The main properties we need from this discretization is that

$$\lim_{L\to\infty} d_4((\mu_{in}^L,\nu_{in}^L),(\mu_{in},\nu_{in}))=0 \qquad \text{and} \qquad \int e^{\delta|\alpha|^2}\,d\mu_{in}^L \le \int e^{\delta|\alpha|^2}\,d\mu_{in}.$$

Such a discretization can be given by the following procedure. For simplicity we consider $R = 2^{k(L+2)K}$, we divide the box $[-\log R,\log R]^{(L+2)K}$ into equal sized boxes $\{B_i\}_{i=1}^L$. We assign θ^i to be the the point with the smallest norm of the box B_i, and the weights are given by $w_i=\mu_{in}(B_i)$. We add any leftover mass on $([-\log R,\log R]^{(L+2)K})^c$ to the delta at the origin. We do the same to produce ν_{in}^L.

By Appendix A, for any $L\in\mathbb{N}$ there exists a unique solution to the projected ODE associated to the solution of the mean field equations with initial conditions given by (μ_{in}^L,ν_{in}^L). Hence, we can construct a global weak solution to the PDE $(\mu^L(t),\nu^L(t))$. By the stability result, we know that for any finite time horizon $T>0$, $\{(\mu^L,\nu^L)\}_L$ form a Cauchy sequence in $AC([0,T],\mathcal{P}\left((\mathbb{R}^{L+2})^K\right)\times \mathcal{P}(Q))$. Hence, there exists $(\mu,\nu)\in AC\left([0,\infty);\mathcal{P}\left((\mathbb{R}^{L+2})^K\right)\times\mathcal{P}(Q)\right)$, such that for any fixed time horizon T

$$\lim_{L\to\infty}\sup_{t\in[0,T]} d_2^2((\mu(t),\nu(t)),(\mu^L(t),\nu^L(t))=0.$$

By Lemma 14, μ^L satisfies the growth condition (26), and so does μ. By Lemma 15, we have that the associated projected ODE flows also converge

$$\lim_{L\to\infty}\sup_{t\in[0,T]} |\Phi^t_{(\mu^L,\nu^L)}(\theta,\omega)-\Phi^t_{(\mu,\nu)}(\tilde\theta,\tilde\omega)|^2 \le e^{C(\Lambda+|\alpha|+|\tilde\alpha|)}|(\theta,\omega)-(\tilde\theta,\tilde\omega)|^2. \tag{22}$$

Using that

$$(\mu^L(t),\nu^L(t))=\Phi^t_{(\mu^L,\nu^L)}\#(\mu_{in}^L\nu_{in}^L)$$

and the uniform exponential integrability of (μ_{in}^L,ν_{in}^L), we can conclude that

$$(\mu(t),\nu(t))=\Phi^t_{(\mu,\nu)}\#(\mu_{in},\nu_{in}),$$

which in turn implies that (μ,ν) is a weak solution to (7) satisfying (19). ■

For the next lemma we use the notation $\theta=(\theta_1,\dots,\theta_K)$ with $\theta_i=(\alpha_i,\beta_i,\gamma_i)\in\mathbb{R}\times\mathbb{R}^L\times\mathbb{R}$, and $\alpha=(\alpha_1,\dots,\alpha_K)\in\mathbb{R}^K$.

Lemma 13 *There exists $C\in\mathbb{R}$ depending on $\|\sigma\|_{C^1}$ such that the vector fields* (14), (15), (16) *and* (17) *satisfy the bounds*

$$\left\|V^\Omega_{(\mu,\nu)}\right\|_\infty \le C\left(1+\int|\alpha|d\mu\right), \qquad \left\|v^\Omega_{(\mu,\nu)}(\cdot,z,x)\right\|_\infty \le C\left(1+|x|+\int|\alpha|d\mu\right), \tag{23}$$

and

$$\begin{aligned}\left\|\left(V_j^\Theta\right)_r\right\|_\infty &\le \begin{cases} C & \text{for } r=1\\ C|\alpha_j| & \text{for } r\neq 1,\end{cases}\\ \left\|\left(v_j^\Theta\right)_r\right\|_\infty &\le \begin{cases} C & \text{for } r=1\\ C|\alpha_j|\,(1+|z|) & \text{for } r\neq 1,\end{cases}\end{aligned} \tag{24}$$

where $(v_j)_r$ denotes the r-th component of the j-th position.

Moreover, we have the following Lipschitz estimate. There exists $C \in \mathbb{R}$ depending on $\|\sigma\|_{C^2}$, such that

$$\left|V^{\Theta}_{(\mu_1,\nu_1)}(\theta) - V^{\Theta}_{(\mu_2,\nu_2)}(\tilde{\theta})\right| \leq C(|\alpha| + |\tilde{\alpha}| + A(\mu_1,\mu_2)) \left(d_2((\mu_1,\nu_1),(\mu_2,\nu_2)) + |\theta - \tilde{\theta}|\right)$$

$$|V^{\Omega}_{(\mu_1,\nu_1)}(\omega_1) - V^{\Omega}_{(\mu_2,\nu_2)}(\omega_2)| \leq CA(\mu_1,\mu_2)\big(|\omega_1 - \omega_2| + d_2(\mu_1,\mu_2)\big),$$

and

$$\begin{aligned}&|v^{\Theta}_{(\mu_1,\nu_1)}(\theta,z) - v^{\Theta}_{(\mu_2,\nu_2)}(\tilde{\theta},z)| \\ &\quad\leq C\left(|\alpha| + |\tilde{\alpha}| + A(\mu_1,\mu_2) + |z|\right)\left(d_2((\mu_1,\nu_1),(\mu_2,\nu_2)) + |\theta - \tilde{\theta}|\right),\end{aligned}$$

$$|v^{\Omega}_{(\mu_1,\nu_1)}(\omega_1,z,x) - v^{\Omega}_{(\mu_2,\nu_2)}(\omega_2,z,x)| \leq C\left(A(\mu_1,\mu_2) + |x| + |z|\right)\big(|\omega_1 - \omega_2| + d_2(\mu_1,\mu_2)\big),$$

where

$$A(\mu_1,\mu_2) = 1 + \left(\int |\alpha|^2 d\mu_1\right)^{1/2} + \left(\int |\alpha|^2 d\mu_2\right)^{1/2},$$

and α_1, α_2 are the first components of θ_1, θ_2, respectively.

Proof [Proof of Lemma 13] Throughout the proof, we use the notation $\theta = (\theta_1, ..., \theta_K)$ with $\theta_i = (\alpha_i, \beta_i, \gamma_i) \in \mathbb{R} \times \mathbb{R}^L \times \mathbb{R}$, $\alpha = (\alpha_1, ..., \alpha_K) \in (\mathbb{R}^L)^K$, and $\omega = (a, b, c) \in Q$. We begin by explicitly writing out the vector fields

$$v^{\Omega}_{(\mu,\nu)}(\omega,z,x) = \begin{pmatrix} \sigma(b \cdot G_\mu(z) + c) - \sigma(b \cdot x + c) \\ aG_\mu(z)\sigma'(b \cdot G_\mu(z) + c) - ax\sigma'(b \cdot x + c) \\ a\sigma'(b \cdot G_\mu(z) + c) - a\sigma'(b \cdot x + c) \end{pmatrix},$$

and $v^{\Theta}_{(\mu,\nu)}(\theta,z) = \left(v^{\Theta}_{(\mu,\nu);1}(\theta_1,z), \cdots, v^{\Theta}_{(\mu,\nu);K}(\theta_K,z)\right)$ with for $1 \leq j \leq K$:

$$v^{\Theta}_{(\mu,\nu);j}(\theta_j,z) = -\int_{[-1,1]^{1+K+1}} \begin{pmatrix} ab_j\sigma(\beta_j \cdot z + \gamma_j)\sigma(b \cdot G_\mu(z) + c) \\ ab_j\alpha_j z\sigma'(\beta_j \cdot z + \gamma_j)\sigma(b \cdot G_\mu(z) + c) \\ ab_j\alpha_j\sigma'(\beta_j \cdot z + \gamma_j)\sigma(b \cdot G_\mu(z) + c) \end{pmatrix} d\nu(\omega).$$

Bounding the generator (2), we have

$$|G_\mu(z)| \leq \int |\sigma(z;\theta)| d\mu(\theta) \leq C\left(\int |\alpha| d\mu(\theta)\right). \tag{25}$$

Using (25), and that $|a|$, $|b|$, $|c| \leq 1$, we readily obtain (23) and (24). Applying the mean value theorem,

$$\begin{aligned}&\nabla_\omega\sigma(x_1;\omega_1) - \nabla_\omega\sigma(x_2;\omega_2) \\ &\quad= \begin{pmatrix} \sigma'(\xi_0)[b_1 \cdot x_1 - b_2 \cdot x_2 + c_1 - c_2] \\ a_1x_1\,\sigma''(\xi_1)[(b_1 \cdot x + c_1) - (b_2 \cdot y + c_2)] + (x_1(a_1 - a_2) + (x_1 - x_2)a_2)\sigma'(b_2 \cdot y + c_2) \\ a_1\sigma''(\xi_1)[(b_1 \cdot x_1 + c_1) - (b_2 \cdot x_2 + c_2)] + (a_1 - a_2)\sigma'(b_2 \cdot x_2 + c_2) \end{pmatrix},\end{aligned}$$

where ξ_0, ξ_1 are points in between $b_1 \cdot x + c_1$ and $b_2 \cdot y + c_2$. To obtain the estimate for v^Ω, we consider the difference above in two instances $x_1 = x_2 = x$, and taking $x_1 = G_{\mu_1}(z)$ and $x_2 = G_{\mu_2}(z)$. Using the triangle inequality, and $\|\sigma\|_{C^2} < \infty$, we can conclude

$$\begin{aligned}&|v^{\Omega}_{(\mu_1,\nu_1)}(\omega_1,z,x) - v^{\Omega}_{(\mu_2,\nu_2)}(\omega_2,z,x)| \\ &\quad\leq C\,\left(1 + |x| + G_{\mu_1}(z) + G_{\mu_2}(z)\right)\,\left(|\omega_1 - \omega_2| + |G_{\mu_1}(z) - G_{\mu_2}(z)|\right).\end{aligned}$$

To estimate $G_{\mu_1}(z) - G_{\mu_2}(z)$, we consider π a coupling between μ_1 and μ_2, and notice that the difference is given by

$$|G_{\mu_1}(z) - G_{\mu_2}(z)| = \left| \int \sigma(z;\theta) - \sigma(z;\tilde{\theta}) d\pi(\theta,\tilde{\theta}) \right| \leq \int |\sigma(z;\theta) - \sigma(z;\tilde{\theta})| \, d\pi.$$

Estimating,

$$|\sigma(z,\theta) - \sigma(z,\tilde{\theta})| \leq C\big(1 + (|\alpha| + |\tilde{\alpha}|)(1+|z|)\big)|\theta - \tilde{\theta}|.$$

Applying the Cauchy-Schwarz inequality,

$$|G_{\mu_1}(z) - G_{\mu_2}(z)|^2 \leq C\left(1 + \left(\int |\alpha|^2 \, d\mu_1 + \int |\tilde{\alpha}|^2 \, d\mu_2\right)(1+|z|^2)\right) \int |\theta - \tilde{\theta}|^2 \, d\pi.$$

Taking π to be the optimal coupling with respect to the d_2 distance, and using (25), we conclude

$$\begin{aligned} &|v^{\Omega}_{(\mu_1,\nu_1)}(\omega_1, z) - v^{\Omega}_{(\mu_2,\nu_2)}(\omega_2, z)|^2 \\ &\quad \leq C\left(1 + |x|^2 + |z|^2 + \textstyle\sum_{i=1,2} \int |\alpha|^2 \, d\mu_i\right)\big(|\omega_1 - \omega_2|^2 + d_2^2(\mu_1,\mu_2)\big). \end{aligned}$$

For v^{Θ}, apply the same argument as above to obtain a bound that also depends on the size of $|\alpha|$. ∎

Lemma 14 *Let $(\mu,\nu) \in AC([0,T];\mathcal{P}((\mathbb{R}^{L+2})^K) \times \mathcal{P}(Q))$ a weak solution to (7), then*

$$\int |\alpha|^2 d\mu_t \leq C\left(\int |\alpha|^2 d\mu_{\text{in}} + t^2\right). \tag{26}$$

Proof By the bound $\|(V_j^{\Theta})_1\|_{\infty} \leq C$, we conclude that $|\alpha(t,\theta_{\text{in}})| \leq |\alpha_{\text{in}}| + Ct$, which implies the desired bound. ∎

A key step in the proof of existence and uniqueness, Proposition 12 and Proposition 10, is the stability of the projected ODE flow.

Lemma 15 *We consider (μ_1,ν_1), $(\mu_2,\nu_2) \in AC([0,T];\mathcal{P}((\mathbb{R}^{L+2})^K)\times\mathcal{P}(Q))$ that satisfy the growth condition (26). The associated flow maps (20) satisfy the bounds*

$$\begin{aligned} |\Phi^t_{(\mu_1,\nu_1)}(\theta_1,\omega_1) - \Phi^t_{(\mu_2,\nu_2)}(\theta_2,\omega_2)|^2 &\leq e^{C(\Lambda + |\alpha_1| + |\alpha_2|)t} e^{Ct^2} |(\theta_1,\omega_1) - (\theta_2,\omega_2)|^2 \\ &\quad + Ce^{C(\Lambda+|\alpha_1|+|\alpha_2|)t} e^{Ct^2} \int_0^t C(r) d_2^2((\mu_1,\nu_1)(r),(\mu_2,\nu_2)(r)) dr, \end{aligned}$$

where

$$\begin{aligned} \Lambda &= 1 + \left(\int |\alpha_1|^2 d\mu_{1,\text{in}}\right)^{1/2} + \left(\int |\alpha_2|^2 d\mu_{2,\text{in}}\right)^{1/2}; \\ C(r) &= e^{-C|\alpha_!|r} e^{-C|\alpha_2|r} e^{-C\Lambda r} e^{-Cr^2/2}(\Lambda + r + |\alpha_1| + |\alpha_2|) \end{aligned}$$

for some constant $C > 0$ depending on T.

Proof Recall the Lipschitz bounds on Lemma 13 are given by

$$C_\Theta(\theta_1,\theta_2) = C\left(1+|\alpha_1|+|\alpha_2|+\left(\int|\alpha|^2 d\mu_1\right)^{1/2}+\left(\int|\alpha|^2 d\mu_2\right)^{1/2}\right),$$
$$C_\Omega = C\left(1+\left(\int|\alpha|^2 d\mu_1\right)^{1/2}+\left(\int|\alpha|^2 d\mu_2\right)^{1/2}\right).$$

By the bound $\|(V_j^\Theta)_1\|_\infty \le C$, we conclude that

$$|\alpha(t,\theta_{\text{in}})| \le |\alpha_{\text{in}}| + Ct.$$

Combining this with the growth assumption (26), we have

$$C_\Theta(\theta_1(t),\theta_2(t)) \le C\left(C_\Theta(\theta_{1,\text{in}},\theta_{2,\text{in}})+t\right) \quad \text{and} \quad C_\Omega(t) \le C\left(C_\Omega(\mu_{1,\text{in}},\mu_{2,\text{in}})+t\right).$$

Taking the derivative of the distance, we find

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}|\theta_1(t)-\theta_2(t)|^2 &= \left\langle \theta_1(t)-\theta_2(t), V^\Theta_{(\mu_{t,1},\nu_{t,1})}(\theta_1(t)) - V^\Theta_{(\mu_{t,2},\nu_{t,2})}(\theta_2(t))\right\rangle \\
&\le C_\Theta(\theta_1(t),\theta_2(t))\left(|\theta_1(t)-\theta_2(t)|^2 + d_2(\mu_{1,t},\mu_{2,t})^2 + d_2(\nu_{1,t},\nu_{2,t})^2\right), \\
\frac{1}{2}\frac{d}{dt}&|\omega_1(t)-\omega_2(t)|^2 \\
&= \left\langle \omega_1(t)-\omega_2(t), \text{Proj}_{\pi(Q)} V^\Omega_{(\mu_{t,1},\nu_{t,1})}(\omega_1(t)) - \text{Proj}_{\pi(Q)} V^\Omega_{(\mu_{t,2},\nu_{t,2})}(\omega_2(t))\right\rangle \\
&\le \left\langle \omega_1(t)-\omega_2(t), V^\Omega_{(\mu_{t,1},\nu_{t,1})}(\omega_1(t)) - V^\Omega_{(\mu_{t,2},\nu_{t,2})}(\omega_2(t))\right\rangle \\
&\le C_\Omega(t)\left(|\omega_1(t)-\omega_2(t)|^2 + d_2(\mu_{1,t},\mu_{2,t})^2\right),
\end{aligned}$$

where we have used the non-expansiveness property of the projection.

Let $\Lambda_\Omega = 1+\left(\int|\alpha|^2 d\mu_{1,\text{in}}\right)^{1/2}+\left(\int|\alpha|^2 d\mu_{2,\text{in}}\right)^{1/2}$ and $\Lambda_\Theta = \Lambda_\Omega + |\alpha_1(0)| + |\alpha_2(0)|$. The estimates above can then be written as

$$\begin{aligned}
\frac{d}{dt}|\theta_1(t)-\theta_2(t)|^2 &\le C(\Lambda_\Theta+t)\left(|\theta_1(t)-\theta_2(t)|^2 + d_2(\mu_{1,t},\mu_{2,t})^2 + d_2(\nu_{1,t},\nu_{2,t})^2\right), \\
\frac{d}{dt}|\omega_1(t)-\omega_2(t)|^2 &\le C(\Lambda_\Omega+t)\left(|\omega_1(t)-\omega_2(t)|^2 + d_2(\mu_{1,t},\mu_{2,t})^2\right),
\end{aligned}$$

which by Gronwall's inequality implies that

$$\begin{aligned}
|\theta_1(t)-\theta_2(t)|^2 &\le e^{C(\Lambda_\Theta t+t^2)}|\theta_{1,\text{in}}-\theta_{2,\text{in}}|^2 \\
&\quad + C\int_0^t e^{C(\Lambda_\Theta t+t^2-\Lambda_\Theta r-r^2)}(\Lambda_\Theta+r)\left(d_2^2((\mu_1,\nu_1),(\mu_2,\nu_2)\right)dr. \\
|\omega_1(t)-\omega_2(t)|^2 &\le e^{C(\Lambda_\Omega t+t^2/2)}|\omega_{1,\text{in}}-\omega_{2,\text{in}}|^2 \\
&\quad + C\int_0^t e^{C(\Lambda_\Omega t+t^2/2-\Lambda_\Omega r-r^2)}(\Lambda_\Omega+r)d_2^2(\mu_1,\mu_2)dr.
\end{aligned}$$

Putting both inequalities together, we arrive at the desired result. ■

We now use this ODE estimate to prove Proposition 10.

Proof [Proof of Proposition 10] Let

$$d(t) = d_2^2((\mu_1, \nu_1)(t), (\mu_2, \nu_2)(t)),$$

and notice that for any coupling Π_* between $\mu_{1,\text{in}} \otimes \nu_{1,\text{in}}$ and $\mu_{2,\text{in}} \otimes \nu_{2,\text{in}}$

$$d(t) \leq \int |(\theta_1(t), \omega_1(t)) - (\theta_2(t), \omega_2(t))|^2 d\Pi_*((\theta_{1,\text{in}}, \omega_{1,\text{in}}), (\theta_{2,\text{in}}, \omega_{2,\text{in}})),$$

since the push-forward of Π_* along the ODE flow at time t is a coupling between $\mu_{1,t} \otimes \nu_{1,t}$ and $\mu_{2,t} \otimes \nu_{2,t}$. Using Lemma 15, we obtain that

$$\begin{aligned} d(t) \leq & \underbrace{\int e^{C(\Lambda+|\alpha_1|+|\alpha_2|)t} e^{Ct^2} |(\theta_1, \omega_1)(0) - (\theta_2, \omega_2)(0)|^2 d\Pi_*}_{I} \\ & + C \int_0^t d(r) \underbrace{\left(\int e^{C(\Lambda+|\alpha_1|+|\alpha_2|)(t-r)} e^{C(t^2-r^2)} (\Lambda + r + |\alpha_1| + |\alpha_2|) d\Pi_* \right)}_{II} dr. \end{aligned}$$

For I we apply the Cauchy-Schwarz and Cauchy's inequality, and take Π_* as the optimal coupling with respect to the 4-Wasserstein to get the bound,

$$\begin{aligned} I &\leq e^{C(\Lambda t + t^2)} \left(\int e^{C|\alpha|t} d\mu_{1,\text{in}} + \int e^{C|\alpha|t} d\mu_{2,\text{in}} \right)^{1/2} \left(\int |(\theta_1, \omega_1)(0) - (\theta_2, \omega_2)(0)|^4 d\Pi_* \right)^{1/2} \\ &\leq e^{C(\Lambda t + t^2)} \left(\int e^{C|\alpha|t} d\mu_{1,\text{in}} + \int e^{C|\alpha|t} d\mu_{2,\text{in}} \right)^{1/2} d_4^2((\mu_{\text{in},1}, \nu_{\text{in},1}), (\mu_{\text{in},2}, \nu_{\text{in},2})). \end{aligned}$$

We bound II from above uniformly in r by the Cauchy-Schwarz inequality,

$$II \leq C \left(\int e^{C|\alpha|t} d\mu_{1,\text{in}} + \int e^{C|\alpha|t} d\mu_{2,\text{in}} \right)^{1/2} (\Lambda_\Omega + t).$$

Therefore, we find that for every $t > 0$

$$d(t) \leq A(t) d_4^2((\mu_{\text{in},1}, \nu_{\text{in},1}), (\mu_{\text{in},2}, \nu_{\text{in},2})) + B(t) \int_0^t d(r) dr,$$

where we define $B(t) = CA(t)(\Lambda + t)$ and

$$A(t) = e^{C(\Lambda t + t^2)} \left(\int e^{C|\alpha|t} d\mu_{1,\text{in}} + \int e^{C|\alpha|t} d\mu_{2,\text{in}} \right)^{1/2}.$$

Gronwall's inequality implies that for all $t \geq 0$

$$d_2^2((\mu_{t,1}, \nu_{t,1}), (\mu_{t,2}, \nu_{t,2})) \leq A(t) e^{tB(t)} d_4^2((\mu_{\text{in},1}, \nu_{\text{in},1}), (\mu_{\text{in},2}, \nu_{\text{in},2})),$$

using the monotonicity of $A(t)$ and $B(t)$. ■

5 Continuous time approximation of parameter dynamics

Proof [Proof of Theorem 6] We consider the parameter training algorithm with learning rate $h > 0$ and a single hidden layer of N neurons for both the generator and the discriminator neural networks. We denote the parameter values at step n by $(\theta_i^n, \omega_i^n)_{i=1,\dots,N}$ and the parameter dynamics is

$$\begin{cases} \theta_i^{n+1} = \theta_i^n + \frac{h}{N} v^{\Theta}_{\mu_N^n, \nu_N^n}(\theta_i^n, z_n) \\ \omega_i^{n+1} = \mathrm{Proj}_Q(\omega_i^n + \frac{h}{N} v^{\Omega}_{\mu_N^n, \nu_N^n}(\omega_i^n, z_n, x_n)), \end{cases}$$

where at each step we sample $x_n \sim P_*$ and $z_n \sim \mathcal{N}$ independently, μ_N^n denotes the empirical measure associated to $\theta_1^n, \dots, \theta_N^n$ and ν_N^n the empirical measure associated to $\omega_1^n, \dots, \omega_N^n$. The parameters are assumed to be initialized by independently (θ_i^0, ω_i^0) by sampling $\mu_{in} \otimes \nu_{in}$. The linear interpolation of the parameters to a continuous time variable $t > 0$ with time step $\Delta t = h/N$ will be denoted by (θ_i, ω_i), where we let $\theta_i(t_n) = \theta_i^n$ and $\omega_i(t_n) = \omega_i^n$, with $t_n = n\Delta t = nh/N$. We let μ and ν be the empirical measures associated to $\theta_1, \dots, \theta_N$ and $\omega_1, \dots, \omega_N$. We suppress the dependence on N of the measures for notational simplicity.

We consider the mean field ODE system defined by the expectation of the vector fields over z and x

$$\begin{cases} \frac{d}{dt}\hat{\theta}_i = V^{\Theta}_{(\hat{\mu},\hat{\nu})}(\hat{\theta}_i) \\ \frac{d}{dt}\hat{\omega}_i = \mathrm{Proj}_{\pi_Q(\omega_i)} V^{\Omega}_{(\hat{\mu},\hat{\nu})}(\hat{\omega}_i), \end{cases}$$

where $\hat{\mu}$ and $\hat{\nu}$ are the empirical measures associated to $\hat{\theta}_1, \dots, \hat{\theta}_N$ and $\hat{\omega}_1, \dots, \hat{\omega}_N$, respectively, and the initial conditions are coupled to the parameter training by $\hat{\theta}_i(0) = \theta_i^0$ and $\hat{\omega}_i(0) = \omega_i^0$. More clearly, the probability measures $\hat{\mu}$ and $\hat{\nu}$ are the solutions of the PDE (7) with random initial conditions chosen as $(\hat{\mu}(0), \hat{\nu}(0)) = (\mu_N(0), \nu_N(0))$.

To simplify the arguments, we first consider the distance between mean field ODE system and the discrete projected forward Euler algorithm

$$\begin{cases} \hat{\theta}_i^{n+1} = \hat{\theta}_i^n + \Delta t\, V^{\Theta}_{(\hat{\mu}^n, \hat{\nu}^n)}(\hat{\theta}_i) \\ \hat{\omega}_i^{n+1} = \mathrm{Proj}_Q\left(\hat{\omega}_i^n + \Delta t\, V^{\Omega}_{(\hat{\mu}^n, \hat{\nu}^n)}(\hat{\omega}_i^n)\right) \end{cases},$$

where we let $T > 0$ be a fixed time horizon and consider $\Delta t = h/N$, where $h > 0$ is the user defined learning rate. To estimate the difference between the continuum and the discrete approximation, we can use a similar argument to Theorem 17, taking into consideration the bound on the Lipschitz constant of the vector fields given by Lemma 13. We can obtain the bound

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N |\hat{\theta}_i^{\Delta t} - \hat{\theta}|^2\right] \leq \Delta t C\left(1 + \mathbb{E}_{\mu^{in}}\left[e^{Ce^{C|\alpha|}}\right]\right).$$

The argument is simpler than the argument below, so we skip it to avoid burdensome repetition.

We define

$$e_i^n = |\hat{\theta}_i^n - \theta_i^n|^2 + |\hat{\omega}_i^n - \omega_i^n|^2 \qquad \text{and} \qquad e^n = \frac{1}{N}\sum_{i=1}^N e_i^n,$$

and notice the inequality

$$d_2^2((\mu^n, \nu^n), (\hat{\mu}^n, \hat{\nu}^n)) \leq e^n.$$

Using a step in either algorithm

$$\begin{aligned} e_i^{n+1} &= |\hat{\theta}_i^n + \Delta t V^{\Theta}_{(\hat{\mu}^n, \hat{\nu}^n)}(\hat{\theta}_i^n) - (\theta_i^n + \Delta t v^{\Theta}_{(\mu^n, \nu^n)}(\theta_i^n))|^2 \\ &\quad + |\mathrm{Proj}_Q(\hat{\omega}_i^n + \Delta t V^{\Omega}_{(\hat{\mu}^n, \hat{\nu}^n)}(\hat{\omega}_i^n)) - \mathrm{Proj}_Q(\omega_i^n + \Delta t v^{\Omega}_{(\mu^n, \nu^n)}(\omega_i^n))|^2. \end{aligned}$$

Using that the projection is contractive, expanding the square and bounding we obtain

$$e_i^{n+1} \leq e_i^n + \Delta t(A_i^n + B_i^n) + (\Delta t)^2 C_i^n,$$

where

$$\begin{aligned} A_i^n &= -2\langle \hat{\theta}_i^n - \theta_i^n, V^{\Theta}_{(\hat{\mu}^n,\hat{\nu}^n)}(\hat{\theta}_i^n) - V^{\Theta}_{(\mu^n,\nu^n)}(\theta_i^n)\rangle + \langle \hat{\omega}_i^n - \omega_i^n, V^{\Omega}_{(\hat{\mu}^n,\hat{\nu}^n)}(\hat{\omega}_i^n) - V^{\Omega}_{(\mu^n,\nu^n)}(\omega_i^n)\rangle, \\ B_i^n &= -2\langle \hat{\theta}_i^n - \theta_i^n, V^{\Theta}_{(\hat{\mu}^n,\hat{\nu}^n)}(\theta_i^n) - v^{\Theta}_{(\mu^n,\nu^n)}(\theta_i^n, z_n)\rangle \\ &\quad +2\langle \hat{\omega}_i^n - \omega_i^n, V^{\Omega}_{(\hat{\mu}^n,\hat{\nu}^n)}(\omega_i^n) - v^{\Omega}_{(\mu^n,\nu^n)}(\omega_i^n, z_n, x_n)\rangle, \end{aligned}$$

and

$$C_i^n = 2\left(|V^{\Theta}_{(\hat{\mu}^n,\hat{\nu}^n)}(\hat{\theta}_i^n)|^2 + |v^{\Theta}_{(\mu^n,\nu^n)}(\theta_i^n)|^2 + |V^{\Omega}_{(\hat{\mu}^n,\hat{\nu}^n)}(\hat{\omega}_i^n)|^2 + |v^{\Omega}_{(\mu^n,\nu^n)}(\omega_i^n)|^2\right).$$

Using the bounds Lemma 13, we get the growth bound

$$|\alpha_i^n|,\ |\hat{\alpha}_i^n| \leq |\alpha_{i,in}| + Cn\Delta t,$$

and the estimates for $n\Delta t < T$

$$A_i^n \leq K_i(e_i^n + e^n) \qquad \text{and} \qquad C_i^n \leq (1 + |x^n|^2 + |z_n|^2)K_i^2,$$

where

$$K_i = C\left(1 + \left(\frac{1}{N}\sum_j |\alpha_{j,in}|^2\right)^{1/2} + |\alpha_{i,in}|\right).$$

Using $e_0 = 0$ and a telescopic sum, we get

$$e_i^{n+1} \leq \Delta t K_i \sum_{r=0}^{n}(e_i^r + e^r) + \Delta t \sum_{r=0}^{n} B_i^r + \Delta t K_i^2 \sum_{r=0}^{n}(1 + |x^r|^2 + |z_r|^2).$$

Next, we will take the conditional expectation with respect to the variables $\{\alpha_{j,in}\}$. To this end, we notice the bound

$$\begin{aligned} \mathbb{E}\left[\sum_{r=0}^{n} B_i^r \middle| \{\alpha_{j,in}\}\right] &\leq \mathbb{E}\left[\left|\sum_{r=0}^{n} B_i^r\right|^2 \middle| \{\alpha_{j,in}\}\right]^{1/2} \\ &= \left(\sum_{r=0}^{n} \mathbb{E}[|B_i^r|^2 || \{\alpha_{j,in}\}] + 2\sum_{r_1=0}^{n}\sum_{r_2=r_1+1}^{n} \mathbb{E}[B_i^{r_1} B_i^{r_2} | \{\alpha_{j,in}\}]\right)^{1/2} \\ &\leq \left(K_i^2 \sum_{r=0}^{n} \mathbb{E}[e_i^r + e^r | \{\alpha_{j,in}\}]\right)^{1/2} \\ &\leq K_i\left(1 + \sum_{r=0}^{n} \mathbb{E}[e_i^r + e^r | \{\alpha_{j,in}\}]\right), \end{aligned}$$

where we have used that by Lemma 13

$$|B_i^r|^2 \leq K_i^2(e_i^r + e^r)$$

and that

$$\mathbb{E}[B_i^{r_1} B_i^{r_2} | \{\alpha_{j,in}\}] = 0,$$

which follows by using the law of iterated expectation with the sigma algebra $\mathcal{F}^{r_2}$ generated by $\{(\theta_{in}^i, \omega_{in}^i\}_{i=1}^N$, $\{x^r\}_{r=0}^{r_2-1}$ and $\{z^r\}_{r=0}^{r_2-1}$. Namely,

$$\mathbb{E}[B_i^{r_1} B_i^{r_2} | \{\alpha_{j,in}\}] = \mathbb{E}[B_i^{r_1} \mathbb{E}[B_i^{r_2} | \mathcal{F}^{r_2}] | \{\alpha_{j,in}\}],$$

where we have used that each $B_i^{r_1}$ is a measure with respect to $\mathcal{F}^{r_2}$ as $r_1 < r_2$. Finally, using that z^{r_2} and x^{r_2} are independent with respect $\mathcal{F}^{r_2}$, and that $(\hat{\theta}_i^{r_2-1}, \hat{\omega}_i^{r_2-1})$ and $(\theta_i^{r_2-1}, \omega_i^{r_2-1})$ are measurable with respect $\mathcal{F}^{r_2}$, we have

$$\begin{aligned}
\mathbb{E}[B_i^{r_2} | \mathcal{F}^{r_2}] \quad &= -2\mathbb{E}_{z^{r_2}}[\langle \hat{\theta}_i^{r_2-1} - \theta_i^{r_2-1}, V^{\Theta}_{(\hat{\mu}^{r_2-1}, \hat{\nu}^{r_2-1})}(\theta_i^{r_2-1}) - v^{\Theta}_{(\mu^{r_2-1}, \nu^{r_2-1})}(\theta_i^{r_2-1}, z^{r_2}) \rangle] \\
&\quad +2\mathbb{E}_{z^{r_2}, x^{r_2}}[\langle \hat{\omega}_i^{r_2-1} - \omega_i^{r_2-1}, V^{\Omega}_{(\hat{\mu}^{r_2-1}, \hat{\nu}^{r_2-1})}(\omega_i^n) - v^{\Omega}_{(\mu^{r_2-1}, \nu^{r_2-1})}(\omega_i^{r_2-1}, z^{r_2}, x^{r_2}) \rangle] \\
&\quad = 0.
\end{aligned}$$

Using the previous bound, that the distributions for x^r and z^r have finite second moments, and that K_i is a deterministic function of $\{\alpha_{j,in}\}$, we obtain up to a change of constant

$$\mathbb{E}[e_i^{n+1} | \{\alpha_{j,in}\}] \leq \Delta t K_i \sum_{r=0}^{n} \mathbb{E}[e_i^r + e^r | \{\alpha_{j,in}\}] + K_i^2 \Delta t.$$

Applying a discrete version of Gromwal's inequality, we have

$$\mathbb{E}[e_i^{n+1} | \{\alpha_{j,in}\}] \leq \Delta t K_i e^{TK_i} \sum_{r=0}^{n} \mathbb{E}[e^r | \{\alpha_{j,in}\}] + \Delta t K_i^2 e^{TK_i}.$$

Summing over i, we obtain

$$\mathbb{E}[e^{n+1} | \{\alpha_{j,in}\}] \leq \Delta t K \sum_{r=0}^{n} \mathbb{E}[e^r | \{\alpha_{j,in}\}] + \Delta t \frac{1}{N} \sum_{i=1}^{N} K_i^2 e^{TK_i},$$

where

$$K = \frac{1}{N} \sum_{i=1}^{N} K_i e^{TK_i}.$$

Using discrete Gromwall's inequality one last time we have the estimate

$$\mathbb{E}[e^{n+1} | \{\alpha_{j,in}\}] \leq \Delta t e^{TK} \frac{1}{N} \sum_{i=1}^{N} K_i^2 e^{TK_i}. \tag{27}$$

Taking expectation, we can bound

$$\mathbb{E}[e^{n+1}] \leq \Delta t \left(\mathbb{E} e^{2TK} + \frac{1}{N} \sum_{i=1}^{N} K_i^4 + \frac{1}{N} \sum_{i=1}^{N} e^{2TK_i} \right).$$

Hence, up to changing constants we have the bound

$$\begin{aligned}
\mathbb{E}[e^{n+1}] \quad &\leq \quad \Delta t C \left(1 + \mathbb{E}\left[e^{C \frac{1}{N} \sum_{i=1}^{N} e^{C|\alpha_i|}} \right] \right) \\
&= \quad \Delta t C \left(1 + \mathbb{E}_{\mu^{in}}\left[e^{\frac{C}{N} e^{C|\alpha|}} \right]^N \right) \\
&\leq \quad \Delta t C \left(1 + \mathbb{E}_{\mu^{in}}\left[e^{C e^{C|\alpha|}} \right] \right) \leq C \Delta t = C \frac{h}{N}.
\end{aligned} \tag{28}$$

The desired bound (12) follows from using the bound (11) to show that the right hand side above is finite. ■

6 Conclusions and future directions

We showed rigorously and quantitatively that the Wasserstein-GAN algorithm is a stochatic discretization of the well-posed PDE system given by (7). Here, we use the insight gained from the dynamics to explain some of the pitfalls of W-GAN Arjovsky et al. (2017) that help explain why is the algorithm finicky to converge. We center in two salient points: the discontinuity of the vector field for the parameters of the discriminator network and the long time behavior of the mean field dynamics.

We noticed that the clipping of the parameters induces that the dynamics are given by a discontinuous vector field, which forces the dynamics into a box Q. In essence, the parameters of the discriminator move within the box Q without anticipating its boundary and crash into ∂Q. This is akin to birds flying into a window. This produces blow-up of the distribution of discriminator parameters in finite time. Still, the measure valued solution is well defined for all times $t > 0$. Most noticeably, for this solution once the dimension of the support of the measure is reduced, it will never fatten back up. In an extreme case, the dynamics can lead to the distribution of the discriminator parameters being $\nu(t) = \delta_{\omega(t)}$ for any $t > t_*$.

In the follow up work Gulrajani et al. (2017), finite time blow-up was already observed in toy numerical examples. Gulrajani et al. (2017) improves the original W-GAN algorithm by enforcing 1-Lipschitz condition with a penalization. With respect to the underlying energy functional, this is equivalent for the mean field dynamics to considering

$$\begin{aligned} E[\mu,\nu] &= \int_{\mathbb{R}^L} D_\nu(G_\mu(z))\, d\mathcal{N}(z) - \int_{\mathbb{R}^K} D_\nu(x)\, dP_*(x) \\ &\quad +\lambda \int_0^1 \int_{\mathbb{R}^L} \int_{\mathbb{R}^K} ||\nabla D_\nu((1-s)G_\mu(z) + sx)| - 1|^2 \; dP_*(x) d\mathcal{N}(z) ds, \end{aligned}$$

with λ being a user chosen penalization parameter. The evolution of the mean field limit can be formally characterized as the gradient descent of E on μ and gradient ascent on ν. In terms of equations we consider

$$\begin{cases} \partial_t \mu - \nabla_\theta \cdot \left(\mu \, \nabla_\theta \frac{\delta E}{\delta \mu}[\mu,\nu] \right) = 0, \\ \partial_t \nu + \gamma_c \nabla_\omega \cdot \left(\nu \, \nabla_\omega \frac{\delta E}{\delta \nu}[\mu,\nu] \right) = 0, \\ \mu(0) = \mu_{in}, \qquad \nu(0) = \nu_{in}. \end{cases}$$

Understanding, the difference in the dynamics for these improved algorithms is an interesting open problem.

For the long time behavior of the dynamics (7), we refer to Section 3 for intuition where we show in a toy example of ODEs that for any initial conditions the dynamics stabilize to a limiting periodic orbit. Generalizing this to absolutely continuous initial data is quite complicated, we mention the recent work for the Euler equation, in Hassainia et al. (2023) the authors construct vortex patches that replicate the motion of leapfrogging vortex points. Moreover, for the general system, we expect that the dynamics will always converge to some limiting periodic orbit. Showing this rigorously is a challenging PDE problem.

In terms of the curse of dimensionality exhibited in Corollary 8, an alternative would be to quantify the convergence of the algorithm in a Reproducing Kernel Hilbert Space (RKHS). In PDE

terms, this would mean to show well posedness of the PDE in a negative Sobolev space like H^{-s} with $s > d/2$.

Acknowledgments

We would like to acknowledge Justin Sirignano, Yao Yao and Federico Camara Halac for useful conversations at the beginning of this project. MGD would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the program *Frontiers in Kinetic Theory* where part of the work on this paper was undertaken. This work was supported by EPSRC grant no EP/R014604/1. The research of MGD was partially supported by NSF-DMS-2205937 and NSF-DMS RTG 1840314. The research of RC was partially supported by NSF-DMS RTG 1840314.

Appendix A.

Following the ideas of Henry (1973), in this section we prove the existence, uniqueness and stability to a class of ODEs with discontinuous forcing given by a projection. We also show quantitative convergence of the projected forward Euler algorithm, for which we could not find a good reference for.

Before we present the main result, we introduce some notation that we need. For any closed convex subset $Q \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$ there exists an unique $\text{Proj}_Q x \in Q$ such that

$$\|\text{Proj}_Q x - x\| = \min_{q \in Q} \|q - x\|.$$

The map Proj_Q is non-expansive, which means that for all x, $y \in \mathbb{R}^d$:

$$\|\text{Proj}_Q(x) - \text{Proj}_Q(y)\| \leq \|x - y\|.$$

We denote by $\pi_Q(x) \subset \mathbb{R}^d$ the tangent cone of Q at $x \in Q$,

$$\pi_Q(x) = \overline{\{v \in \mathbb{R}^d \,|\, \exists\, \epsilon > 0, \quad x + \epsilon v \in Q\}} = \left\{ v \in \mathbb{R}^d \;\middle|\; \lim_{h \to 0^+} \frac{d(x + hv, Q)}{h} = 0 \right\},$$

which is a closed convex cone. The map $\text{Proj}_{\pi_Q(x)} : \mathbb{R}^d \to \mathbb{R}^d$ denotes the projection onto $\pi_Q(x) \subset \mathbb{R}^d$. We notice that for a smooth vector field $V(x) : Q \to \mathbb{R}^d$, the mapping $x \in \mathbb{R}^n \mapsto \text{Proj}_{\pi_Q(x)}(V(x))$ is discontinuous at points x such that $V(x) \notin \pi_Q(x)$.

Theorem 16 (Henry (1973)) *Let $Q \subset \mathbb{R}^d$ be a closed and convex subset of $\mathbb{R}^d$ and $V : \mathbb{R}^d \to \mathbb{R}^d$ a C^1 vector field, which satisfies that there exists $C > 0$, such that $|V(x)|, |\nabla V(x)| \leq C$. Then, for any initial condition $x_{in} \in Q$ there exists a unique absolutely continuous curve $x : [0, \infty) \to Q$ such that*

$$\begin{cases} \dot{x} = \text{Proj}_{\pi_Q(x)} V(x), \\ x(0) = x_{in}, \end{cases} \tag{29}$$

with the equality satisfied for almost every t. Moreover, the solutions are also stable with respect to the initial condition x_{in}:

$$\|x_1(t) - x_2(t)\| \leq e^{\|\nabla V\|_\infty t} \|x_1(0) - x_2(0)\|,$$

where x_1 and x_2 are two solution to (29).

Moreover, we can approximate these solutions by a projected forward Euler algorithm.

Theorem 17 *Let $x_{\Delta t} : [0,\infty) \to Q$ be the linear interpolation at times $n\Delta t$ of $\{x^n_{\Delta t}\}$ defined by the projected Euler algorithm*

$$\begin{cases} x^{n+1}_{\Delta t} = Proj_Q(x^n_{\Delta t} + \Delta t\, V(x^n_{\Delta t})) \\ x^0_{\Delta t} = x_{in}. \end{cases} \tag{30}$$

Then, for any time horizon $T > 0$, as $\Delta t \to 0$ we have

$$\|x_{\Delta t} - x\| \le e^{(1+\|\nabla V\|_\infty)t} \left(2(\Delta t)^{1/2} \|V\|_\infty + \Delta t \|V\|_\infty \|\nabla V\|_\infty \right),$$

where x is the unique solution of (29).

Proof [Proof of Theorem 16 and Theorem 17] For each $x \in Q$, we define the normal cone $N_Q(x)$ as

$$N_Q(x) = \{n \in \mathbb{R}^d \,|\, \forall q \in Q : \quad \langle n, q - x\rangle \le 0\},$$

or equivalently the set of vectors $n \in \mathbb{R}^d$ such that $\langle n, w\rangle \le 0$ for all $w \in \pi_Q(x)$. It follows directly from the projection property the following useful result.

Lemma 18 *For any $v \in \mathbb{R}^d$ and $x \in Q$, the vector $n_x = v - \mathrm{Proj}_{\pi_Q(x)} v$ is orthogonal to $\mathrm{Proj}_{\pi_Q(x)} v$ and $n_x \in N_Q(x)$. Conversely, if $w \in \pi_Q(x)$ is that $n_x = v - w \in N_Q(x)$ and $\langle w, v - w\rangle = 0$, then $w = \mathrm{Proj}_{\pi_Q(x)} v$.*

Uniqueness and Stability. Suppose that $x_1 : [0,T] \to Q$ and $x_2 : [0,T] \to Q$ are solutions of (29). Then,

$$\frac{d}{dt}\frac{1}{2}\|x_1 - x_2\|^2 = \langle x_1 - x_2, \mathrm{Proj}_{\pi_Q(x_1)} V(x_1) - \mathrm{Proj}_{\pi_Q(x_2)} V(x_2)\rangle.$$

Using the property of the projection we have

$$\begin{aligned} \langle x_1 - x_2, \mathrm{Proj}_{\pi_Q(x_1)} V(x_1) - \mathrm{Proj}_{\pi_Q(x_2)} V(x_2)\rangle &\le \langle x_1 - x_2, V(x_1) - V(x_2)\rangle \\ &\le \|\nabla V\|_\infty \|x_1 - x_2\|^2, \end{aligned}$$

where we have used Lemma 18 for the first inequality, and the Lipschitz propety for the second inequality. Grownwall's inequality applied to $\|x_1 - x_2\|^2$ gives:

$$\|x_1(t) - x_2(t)\| \le e^{\|\nabla V\|_\infty t} \|x_1(0) - x_2(0)\|,$$

which shows the uniqueness and stability of solutions with respect to the initial condition.

Equivalence with a relaxed problem. Using $N_Q(x)$, we now introduce a relaxed problem which we prove is equivalent to the ODE (29). For each $x \in Q$ we define the compact convex set $\mathcal{V}(x) \subset \mathbb{R}^d$ by

$$\mathcal{V}(x) = \{V(x) - n_x \,|\, n_x \in N_Q(x), \quad \|n_x\|^2 \le V(x) \cdot n_x\}.$$

The relaxed problem is finding an absolutely continuous curve $x : [0,T] \to Q$ such that

$$\begin{cases} \dot{x}(t) \in \mathcal{V}(x(t)) \\ x(0) = x_{in}, \end{cases} \tag{31}$$

for almost every $t \in [0,T]$. To show the equivalence between (29) and (31), we need the following Lemma.

Lemma 19 *For all $x \in Q$ we have $V(x) \in \mathcal{V}(x)$, $\mathrm{Proj}_{\pi_Q(x)}V(x) \in \mathcal{V}(x)$ and*

$$\mathcal{V}(x) \cap \pi_Q(x) = \{\mathrm{Proj}_{\pi_Q(x)}V(x)\}.$$

Proof [Proof of Lemma 19] Taking $n_x = 0$ in the definition of $\mathcal{V}(x)$ gives that $V(x) \in \mathcal{V}(x)$. Writing $n_x = V(x) - \mathrm{Proj}_{\pi_Q(x)}V(x)$, we recall from Lemma 18 that $n_x \in N_Q(x)$ and

$$\langle n_x, \mathrm{Proj}_{\pi_Q(x)}V(x)\rangle = 0,$$

so $\|n_x\|^2 = \langle n_x, V(x)\rangle$ and we conclude $\mathrm{Proj}_{\pi_Q(x)}V(x) \in \mathcal{V}(x)$. Now note that if $V(x) - n_x \in \pi_Q(x)$ with $n_x \in N_Q(x)$, then $\langle V(x) - n_x, n_x\rangle \leq 0$ with equality only if $V(x) - n_x = \mathrm{Proj}_{\pi_Q(x)}V(x)$, as we have noted above. So if $V(x) - n_x \in \pi_Q(x) \cap \mathcal{V}(x)$ then $V(x) - n_x = \mathrm{Proj}_{\pi_Q(x)}V(x)$. ■

An absolutely continuous curve $x : [0, T] \to Q$ is such that

$$\lim_{h\to 0^+} \frac{x(t+h) - x(t) - \dot{x}(t)\,h}{h} = 0,$$

for almost every t. Since $x(t) \in Q$ for all $t \in [0, T]$,

$$0 = \lim_{h\to 0^+} \frac{d(x(t+h), Q)}{h} = \lim_{h\to 0^+} \frac{d(x(t) + h\dot{x}(t), Q)}{h},$$

which shows $\dot{x}(t) \in \pi_Q(x(t))$. If we have a solution to the relaxed problem, then the differential inclusion $\dot{x}(t) \in \mathcal{V}(x(t))$ is satisfied almost everywhere, therefore we have $\dot{x}(t) = \mathrm{Proj}_{\pi_Q(x)}V(x)$ since by Lemma 19 $\mathcal{V}(x) \cap \pi_Q(x) = \{\mathrm{Proj}_{\pi_Q(x)}V(x)\}$, and we conclude that (31) and (29) are equivalent.

Existence. Consider $\nu_Q(x^{n+1}_{\Delta t}) \in N_Q(x^{n+1}_{\Delta t})$ unit vectors and $0 \leq \lambda \leq 1$ such that

$$x^{n+1}_{\Delta t} = x^n_{\Delta t} + \Delta t\, V(x^n_{\Delta t}) - \Delta t\, \lambda\, (V(x^n_{\Delta t}) \cdot \nu_Q(x^{n+1}_{\Delta t}))_+ \nu_Q(x^{n+1}_{\Delta t}),$$

which follows directly from the properties of the projection. For each $n \geq 0$ we consider the discrete velocity

$$u^n_{\Delta t} = \frac{x^{n+1}_{\Delta t} - x^n_{\Delta t}}{\Delta t},$$

which we re-write as

$$\begin{aligned} u^n_{\Delta t} \;=\; & V(x^{n+1}_{\Delta t}) - \lambda (V(x^{n+1}_{\Delta t}) \cdot \nu_Q(x^{n+1}_{\Delta t}))_+ \nu_Q(x^{n+1}_{\Delta t}) + \underbrace{V(x^n_{\Delta t}) - V(x^{n+1}_{\Delta t})}_{I} \\ & + \underbrace{\lambda\, ((V(x^n_{\Delta t}) \cdot \nu_Q(x^{n+1}_{\Delta t}))_+ - (V(x^{n+1}_{\Delta t}) \cdot \nu_Q(x^{n+1}_{\Delta t}))_+)\nu_Q(x^{n+1}_{\Delta t})}_{II}. \end{aligned}$$

We notice the bounds

$$|I|,\ |II| \leq \|\nabla V\|_\infty |x^{n+1}_{\Delta t} - x^n_{\Delta t}| \leq \Delta t \|\nabla V\|_\infty \|V\|_\infty.$$

Therefore, letting B_1 denote the unit ball centred at the origin,

$$u^n_{\Delta t} \in \mathcal{V}(x^{n+1}_{\Delta t}) + \|\nabla V\|_\infty \|V\|_\infty \Delta t\, B_1.$$

Hence, for any $\Delta t > 0$ we can conclude that for a.e. t

$$(x_{\Delta t}(t), \dot{x}_{\Delta t}(t)) \in \mathrm{Graph}(\mathcal{V}) + \Delta t(\|V\|_\infty B_1 \times \|\nabla V\|_\infty \|V\|_\infty B_1), \tag{32}$$

Noting that $x_{\Delta t}$ is uniformly Lipschitz with constant less that $\|V\|_\infty$, we get up to subsequence there exists a Lipschitz function $X : [0, \infty) \to Q$ such that $x_{\Delta t} \to X$ uniformly at compact subintervals, by Arzerlà-Ascoli. We conclude using Mazur's Lemma that the derivative of x belongs almost everywhere to the upper limit of the convex hull of the values of $\dot{x}_{\Delta t}(t)$,

$$\dot{x}(t) \in \limsup_{\epsilon \to 0^+} \text{co}(\dot{x}_{\Delta t}(t)_{0<\Delta t<\epsilon}).$$

Using that $\mathcal{V}(x(t))$ is convex and closed, we conclude

$$(x(t), \dot{x}(t)) \in \text{Graph}(\mathcal{V}),$$

which implies that x is a solution to the relaxed problem, and therefore a solution to the original (29).

Quantitative Estimate. We differentiate the distance, between X and $x_{\Delta t}$ to obtain

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}|X - x_{\Delta t}|^2 &= \langle X - x_{\Delta t}, \dot{X} - \dot{x}_{\Delta t} \rangle \\
&\leq \langle X - x_{\Delta t}, V(X) - V(x_{\Delta t}) \rangle + \Delta t \|V\|_\infty |\dot{X} - \dot{x}_{\Delta t}| \\
&\quad + \Delta t \|V\|_\infty \|\nabla V\|_\infty |X - x_{\Delta t}| \\
&\leq (1 + \|\nabla V\|_\infty)|X - x_{\Delta t}|^2 + 2\Delta t \|V\|_\infty^2 + (\Delta t)^2 \|V\|_\infty^2 \|\nabla V\|_\infty^2,
\end{aligned}$$

where we have used estimate (32) and the contraction to property. Using Gromwall's inequality and that $|X - x_{\Delta t}|^2 = 0$, we obtain

$$|X - x_{\Delta t}|^2 \leq e^{2(1+\|\nabla V\|_\infty)t} \left(2\Delta t \|V\|_\infty^2 + (\Delta t)^2 \|V\|_\infty^2 \|\nabla V\|_\infty^2\right).$$

■

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Andrea L Bertozzi, Thomas Laurent, and Jesús Rosado. Lp theory for the multidimensional aggregation equation. *Communications on Pure and Applied Mathematics*, 64(1):45–83, 2011.

François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137:541–593, 2007.

JA Carrillo, M DiFrancesco, A Figalli, T Laurent, and C Slepcev. Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Mathematical Journal*, 156(2):229–271, 2011.

José A Carrillo, Robert J McCann, and Cédric Villani. Contractions in the 2-wasserstein length space and thermalization of granular media. *Archive for Rational Mechanics and Analysis*, 179: 217–263, 2006.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Guido De Philippis, Alpár Richárd Mészáros, Filippo Santambrogio, and Bozhidar Velichkov. Bv estimates in optimal transportation and applications. *Archive for Rational Mechanics and Analysis*, 219:829–860, 2016.

Simone Di Marino, Bertrand Maury, and Filippo Santambrogio. Measure sweeping processes. *Journal of Convex Analysis*, 23(2):567–601, 2016.

Richard M Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978.

Xavier Fernández-Real and Alessio Figalli. The continuous formulation of shallow neural networks as wasserstein-type gradient flows. In *Analysis at Large: Dedicated to the Life and Work of Jean Bourgain*, pages 29–57. Springer, 2022.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, August 2015. URL `https://hal.science/hal-00915365`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Zineb Hassainia, Taoufik Hmidi, and Nader Masmoudi. Rigorous derivation of the leapfrogging motion for planar euler equations. *arXiv preprint arXiv:2311.15765*, 2023.

Claude Henry. An existence theorem for a class of differential equations with multivalued right-hand side. *Journal of Mathematical Analysis and Applications*, 41(1):179–186, 1973.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled Generative Adversarial Networks. In *International Conference on Learning Representations*, 2016.

Jean Jacques Moreau. Evolution problem associated with a moving convex set in a hilbert space. *Journal of differential equations*, 26(3):347–374, 1977.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

Filippo Santambrogio. Crowd motion and evolution pdes under density constraints. *ESAIM: Proceedings and Surveys*, 64:137–157, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020a.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020b.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE, 2020.

Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Stephan Wojtowytsch and Weinan E. Can shallow neural networks beat the curse of dimensionality? a mean field training perspective. *IEEE Transactions on Artificial Intelligence*, 1(2):121–129, 2020. doi: 10.1109/TAI.2021.3051357.