Journal of Machine Learning Research 24 (2023) 1-40 Submitted 5/22; Revised 3/23; Published 6/23

GANs as Gradient Flows that Converge

Yu-Jui Huang YUJUI.HUANG@COLORADO.EDU
Department of Applied Mathematics
University of Colorado
Boulder, CO 80309-0526, USA

Yuchong Zhang YUCHONG.ZHANG@UTORONTO.CA
Department of Statistical Sciences
University of Toronto
Toronto, ON M5G 1Z5, Canada

Editor: Amos Storkey

Abstract

This paper approaches the unsupervised learning problem by gradient descent in the space of probability density functions. A main result shows that along the *gradient flow* induced by a distribution-dependent ordinary differential equation (ODE), the unknown data distribution emerges as the long-time limit. That is, one can uncover the data distribution by simulating the distribution-dependent ODE. Intriguingly, the simulation of the ODE is shown equivalent to the training of generative adversarial networks (GANs). This equivalence provides a new "cooperative" view of GANs and, more importantly, sheds new light on the divergence of GANs. In particular, it reveals that the GAN algorithm implicitly minimizes the mean squared error (MSE) between two sets of samples, and this MSE fitting alone can cause GANs to diverge. To construct a solution to the distribution-dependent ODE, we first show that the associated nonlinear Fokker-Planck equation has a unique weak solution, by the Crandall-Liggett theorem for differential equations in Banach spaces. Based on this solution to the Fokker-Planck equation, we construct a unique solution to the ODE, using Trevisan's superposition principle. The convergence of the induced gradient flow to the data distribution is obtained by analyzing the Fokker-Planck equation.

Keywords: unsupervised learning, generative adversarial networks, distribution-dependent ODEs, gradient flows, nonlinear Fokker-Planck equations

1. Introduction

A central theme in machine learning is to uncover the underlying distribution of observed data points. Mathematically, this can be stated as

$$\min_{\rho\in\mathcal{P}(\mathbb{R}^d)} d(\rho,\rho_{\rm d}), \tag{1}$$

where $\mathcal{P}(\mathbb{R}^d)$ is the set of probability density functions on $\mathbb{R}^d$, $\rho_{\rm d}\in\mathcal{P}(\mathbb{R}^d)$ is the *unknown* data distribution, and $d(\cdot,\cdot)$ is a metric (or semi-metric) on $\mathcal{P}(\mathbb{R}^d)$. Unlike traditional parameter fitting by maximum likelihood methods, *generative adversarial networks* (GANs) in Goodfellow et al. (2014) view (1) as a two-player non-cooperative game: a *generator* actively produces samples similar to the real data and a *discriminator* strives to distinguish between the real and fake samples.

In theory, GANs take the form of a min-max game where the discriminator chooses $D : \mathbb{R}^d \to [0, 1]$ to maximize its chance of differentiating real samples from fake ones, while the generator, who observes its opponent's choice D, selects $G : \mathbb{R}^n \to \mathbb{R}^d$ (for a fixed $n \leq d$) to neutralize the effect of D. In Goodfellow et al. (2014), this minimax problem is shown equivalent to (1) with $d(\cdot, \cdot) = \mathrm{JSD}(\cdot, \cdot)$, the Jensen-Shannon divergence; further, the generator's optimal strategy $G^* : \mathbb{R}^n \to \mathbb{R}^d$ exists and it reveals the data distribution ρ_{d}.

What is *not* answered by the theory of GANs is whether (and how) an initial guess $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ (or, an initial $G : \mathbb{R}^n \to \mathbb{R}^d$) can evolve gradually towards $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$ (or, $G^* : \mathbb{R}^n \to \mathbb{R}^d$). The GAN algorithm (Goodfellow et al., 2014, Algorithm 1) serves to find this "path to ρ_{d}" numerically: by using two artificial neural networks to represent the functions G and D, it updates their parameters repetitively, through a recursive optimization performed sequentially by the two players. While the GAN algorithm has brought significant improvements to artificial intelligence applications (see e.g., Denton et al. (2015), Vondrick et al. (2016), Ledig et al. (2017), Wiese et al. (2020)), it is *not* as stable as we wish—while a "path" can always be computed, it often diverges away from ρ_{d}.

Attempts to improve the stability of GANs are numerous, such as modifying training procedures (e.g., Salimans et al. (2016), Heusel et al. (2017), Arjovsky and Bottou (2017)), adding regularizing terms (e.g., Gulrajani et al. (2017), Miyato et al. (2018)), or using different metrics $d(\cdot, \cdot)$ (e.g., Wasserstein distance in Arjovsky et al. (2017), maximum mean discrepancy in Li et al. (2017) and Bińkowski et al. (2018), characteristic function distance in Ansari et al. (2020)). All the studies approached the stability issue at the *algorithmic* level, i.e., by inspecting the details of the recursive optimization procedures.

In this paper, we give a complete characterization of the "path to ρ_{d}" at the *theoretic* level. In a nutshell, this "path" will be characterized as a *gradient flow* in the space $\mathcal{P}(\mathbb{R}^d)$, whose evolution is governed by a *distribution-dependent* ordinary differential equation (ODE); see Theorem 7 and Proposition 8, the main results of this paper. This theoretic characterization has important practical implications.

First, it sheds new light on the divergence of GANs. By the gradient-flow identification of GANs, we are able to decompose the GAN algorithm and discover an unapparent fact: the algorithm implicitly minimizes the mean squared error (MSE) between two sets of samples in $\mathbb{R}^d$. While the MSE fitting demands *point-wise* similarity (i.e., the i^{th} sample in one set is similar to the i^{th} sample in the other set), what we actually need is only *set-wise* similarity (i.e., the distribution on $\mathbb{R}^d$ of the samples in one set is similar to that of the samples in the other set). As it imposes too stringent a criterion, the MSE fitting may keep producing inferior $G : \mathbb{R}^n \to \mathbb{R}^d$, such that ρ_{d} is never approached. This observation suggests a new potential route to alleviate instability: replacing the MSE fitting in the GAN algorithm by a measure of set-wise similarity; see Section 3.3 for detailed explanations.

Our main results, in addition, yield a new interpretation for GANs: the *non-cooperative* game between the generator and discriminator can be equivalently viewed as a *cooperative* game between a *navigator* and a *calibrator*. The navigator aims to navigate across the space $\mathcal{P}(\mathbb{R}^d)$ following the aforementioned distribution-dependent ODE, and the calibrator serves to "calibrate" the ODE's dynamics. To the best of our knowledge, a possible "cooperative" view of GANs was only briefly touched on in Goodfellow (2016). We materialize this idea with great specifics: the two players collaborate precisely to simulate an ODE that ensures smooth sailing to ρ_{d}; see the discussion below Proposition 8 for details.

1.1 Our Methodology

As in Goodfellow et al. (2014), we take $d(\cdot,\cdot) = \mathrm{JSD}(\cdot,\cdot)$ in (1). Our first observation is that $J(\rho) := \mathrm{JSD}(\rho, \rho_{\mathrm{d}})$ is *strictly convex* on $\mathcal{P}(\mathbb{R}^d)$, which suggests that (1) can potentially be solved by (deterministic) gradient descent in $\mathcal{P}(\mathbb{R}^d)$. By taking the "gradient of J at $\rho \in \mathcal{P}(\mathbb{R}^d)$" to be $\nabla \frac{\delta J}{\delta \rho}(\rho, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$, where $\frac{\delta J}{\delta \rho} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ is the linear functional derivative of J (Definition 4) and ∇ is the usual gradient operator in $\mathbb{R}^d$, we derive the gradient-descent ODE for (1), which is given by

$$dY_t = -\frac{1}{2}\left(\frac{\nabla \rho^{Y_t}(Y_t)}{\rho^{Y_t}(Y_t)} - \frac{\nabla \rho_{\mathrm{d}}(Y_t) + \nabla \rho^{Y_t}(Y_t)}{\rho_{\mathrm{d}}(Y_t) + \rho^{Y_t}(Y_t)}\right) dt, \quad \rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \tag{2}$$

This ODE is distribution-dependent in a distinctive way. Unlike a typical McKean-Vlasov stochastic differential equation (SDE) that depends explicitly on $\mathcal{L}(Y_t)$, the law of the state process Y at time t, (2) depends on the density $\rho^{Y_t} \in \mathcal{P}(\mathbb{R}^d)$ of Y_t and its gradient $\nabla \rho^{Y_t}$. The classical framework of interacting particle systems, crucial for the existence of solutions to McKean-Vlasov SDEs, does not easily accommodate this kind of dependence.

In view of this, we take an unconventional approach motivated by Barbu and Röckner (2020): by finding a weak solution $u(t, y)$ to the nonlinear Fokker-Planck equation associated with ODE (2), one can in turn construct a solution Y to (2), made specifically from $u(t, y)$.

To solve the nonlinear Fokker-Planck equation, i.e., (25) below, we transform it into a nonlinear Cauchy problem in a suitable Banach space, i.e., (36) below. By showing that the involved operator is "accretive" (Lemma 14), we are able to find a weak solution $u(t, y)$ by Crandall-Liggett's theorem for partial differential equations (PDEs) in Banach spaces; see Proposition 17. The uniqueness of solutions is also established by generalizing Brézis and Crandall (1979), which characterizes $u(t, y)$ as the unique weak solution in Theorem 21.

By substituting $u(t, \cdot)$ for the density $\rho^{Y_t}(\cdot)$ in (2), we obtain a standard ODE *without* distribution dependence, i.e., (49) below. To construct a solution to (2), we aim at finding a process Y such that (i) $t \mapsto Y_t$ satisfies the above standard ODE, and (ii) the density of Y_t exists and coincides with $u(t, \cdot)$, i.e., $\rho^{Y_t}(\cdot) = u(t, \cdot) \in \mathcal{P}(\mathbb{R}^d)$ for all $t \geq 0$. Thanks to the superposition principle of Trevisan (2016), we can choose a probability measure $\mathbb{P}$ on the canonical path space, so that the canonical process $Y_t(\omega) := \omega(t)$ fulfills (i) and (ii) above. This immediately gives a solution to (2) (Proposition 25) and it is in fact unique up to time-marginal distributions under suitable regularity conditions (Proposition 28).

Let us stress that while (2) takes the form of an ODE, it is nontrivial to find a suitable notion of solutions, due to two kinds of randomness intertwined at time 0; see the second paragraph in Section 5. Ultimately, we define a solution to (2) using a random selection of deterministic paths, represented by a probability measure $\mathbb{P}$ on the canonical path space (Definition 22). This formulation perfectly fits Trevisan's superposition principle, and is reminiscent of L.C. Young's generalized curves in the deterministic theory and weak solutions to SDEs (despite our restriction to the path space with only $\mathbb{P}$ varying).

With the unique solution Y to ODE (2) characterized, it is time to check if our "gradient descent" idea for (1) works. Two key findings stem from a detailed analysis of the transformed Fokker-Planck equation (i.e., (36) below). First, along the path of Y, the map $t \mapsto J(\rho^{Y_t}) = \mathrm{JSD}(\rho^{Y_t}, \rho_{\mathrm{d}})$ is nonincreasing (Proposition 30), i.e., ρ^{Y_t} moves closer to ρ_{d} continuously over time. Second, $\rho^{Y_{t_n}} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$, at least along a sequence $\{t_n\}$ in time

with $t_n \uparrow \infty$ (Corollary 32). Putting these together gives the ultimate convergence result: $\rho^{Y_t} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$ as $t \to \infty$; see Theorem 7. That is to say, we can uncover ρ_{d} along the *gradient flow* $\{\rho^{Y_t}\}_{t\geq 0}$ in $\mathcal{P}(\mathbb{R}^d)$, specified by the gradient-descent ODE (2).

Algorithm 1 is designed to simulate ODE (2). Intriguingly, it is shown equivalent to the GAN algorithm (Proposition 8), which yields the "cooperative" view of GANs mentioned above: the generator and discriminator now take the roles of a navigator and calibrator, respectively, working closely to simulate ODE (2); see the exposition in Section 3.1.

1.2 Relations to the Literature

We work with the general space $\mathcal{P}(\mathbb{R}^d)$ under the total variation distance, which differs from the classical setting—$\mathcal{P}_2(\mathbb{R}^d)$ under the second-order Wasserstein distance $\mathcal{W}_2$, where $\mathcal{P}_2(\mathbb{R}^d)$ is the set of elements in $\mathcal{P}(\mathbb{R}^d)$ with finite second moments. There are two main reasons for our non-standard setup. First, working with $\mathcal{P}_2(\mathbb{R}^d)$ implicitly assumes $\rho_{\mathrm{d}} \in \mathcal{P}_2(\mathbb{R}^d)$, while ρ_{d}, as the underlying data distribution, generally lies in $\mathcal{P}(\mathbb{R}^d)$. Second, using the total variation distance ensures the "right" kind of convergence. As we chose $\mathrm{JSD}(\cdot,\cdot)$ in the first place to measure the distance between densities, it should be kept consistently throughout our analysis. As shown in Remark 2, convergence in $\mathcal{P}(\mathbb{R}^d)$ under $\mathrm{JSD}(\cdot,\cdot)$ is equivalent to that under total variation (or, under $L^1(\mathbb{R}^d)$). Hence, "$\rho^{Y_t} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$" in Theorem 7 implies "$\mathrm{JSD}(\rho^{Y_t}, \rho_{\mathrm{d}}) \to 0$". That is, convergence to $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$ is not only achieved, but achieved under the *desired* distance function originally chosen.

Working with $\mathcal{P}(\mathbb{R}^d)$ (under total variation) deprives us of the full-fledged theory of $\mathcal{P}_2(\mathbb{R}^d)$ (under $\mathcal{W}_2$), developed in e.g., Ambrosio et al. (2005), Cardaliaguet et al. (2015), Carmona and Delarue (2018a), and Delarue et al. (2019). First, it is unclear how to define the "gradient" of $J(\rho) = \mathrm{JSD}(\rho, \rho_{\mathrm{d}})$ in $\mathcal{P}(\mathbb{R}^d)$, as the standard Lions and Wasserstein derivatives are only defined in $\mathcal{P}_2(\mathbb{R}^d)$ (or in $\mathcal{P}_q(\mathbb{R}^d)$ for $q \in [1,\infty)$, which consists of elements in $\mathcal{P}(\mathbb{R}^d)$ with finite q^{th} moments). Since $\nabla \frac{\delta J}{\delta \rho}$ admits a gradient-type property (Proposition 5) and it is well-defined in $\mathcal{P}(\mathbb{R}^d)$ without relying on the $\mathcal{P}_2(\mathbb{R}^d)$ structure (Lemma 6), we take it to be the proper "gradient" in $\mathcal{P}(\mathbb{R}^d)$. While $\nabla \frac{\delta J}{\delta \rho}$ is more general, it coincides with the Lions and Wasserstein derivatives when restricted to $\mathcal{P}_2(\mathbb{R}^d)$; see Section 2.2. On the other hand, to compute the evolution $t \mapsto J(\rho^{Y_t})$ and show its convergence to $J(\rho_{\mathrm{d}}) = 0$, we do not employ any Itô-type formula or compactness argument (which are unique to $\mathcal{P}_2(\mathbb{R}^d)$), but rely on the aforementioned analysis of the transformed Fokker-Planck equation (Proposition 30 and Corollary 32); see Remarks 31 and 33 for details.

Gradient flows have recently been used in many implicit generative models; see e.g., Gao et al. (2019), Gao et al. (2020), Ansari et al. (2021), and Mroueh and Nguyen (2021). Their algorithms are advocated as alternatives to GANs—in particular, the minimization part of GANs is replaced by a kind of gradient descent. Contrary to the common practice "modifying GANs to form a gradient-flow algorithm", Proposition 8 shows that the GAN algorithm itself, without any modification, already computes gradient flows. What sets us apart is a subtle difference in algorithm design. While our gradient-flow algorithm (i.e., Algorithm 1) resembles those in Gao et al. (2019) and Gao et al. (2020), we particularly coordinate the estimation of D with the gradient-descent update of G, so that the *zero-sum* game setup of GANs can be recovered. The algorithms in Gao et al. (2019) and Gao et al. (2020), by contrast, represent *non-zero-sum* games.

In the recent development of gradient-flow algorithms mentioned above, a common (yet unnoticed) issue is the "*inconsistent use of distance functions*": one proposes to solve (1) with a specific choice of $d(\cdot,\cdot)$ (e.g., an f-divergence), but carries out theoretic proofs under a different distance function (e.g., $\mathcal{W}_2$). This can be problematic: even if $\mathcal{W}_2(\rho^{Y_t},\rho_{\rm d})\to 0$ is proved, since the metric $\mathcal{W}_2$ is *not* equivalent to an f-divergence, ρ^{Y_t} need not converge to $\rho_{\rm d}$ under an f-divergence, the *desired* distance function initially chosen. As noted above, Theorem 7 settles this issue, as the target distance function (i.e., $\mathrm{JSD}(\cdot,\cdot)$) is equivalent to the actual distance function in use (i.e., the total variation distance).

1.3 Organization of the Paper and Notation

Section 2 formulates a proper "gradient" in $\mathcal{P}(\mathbb{R}^d)$ and derives the gradient-descent ODE (2). Section 3 presents the main results (Theorem 7 and Proposition 8) and discusses their implications, including the identification of GANs with gradient flows and the cause of divergence of GANs. Section 4 derives the nonlinear Fokker-Planck equation associated with ODE (2) and establishes the existence of a unique weak solution $u(t,y)$. Based on $u(t,y)$, Section 5 constructs a solution $t\mapsto Y_t$ to ODE (2) and shows that it is unique up to time-marginal distributions. Section 6 proves Theorem 7, i.e., the density of Y_t converges to $\rho_{\rm d}$, as $t\to\infty$. Appendices collect some proofs.

Throughout the paper, $\mathcal{P}(\mathbb{R}^d)$ denotes the set of all probability density functions on $\mathbb{R}^d$. For any $q\in[1,\infty)$, let $\mathcal{P}_q(\mathbb{R}^d)$ be the subset of $\mathcal{P}(\mathbb{R}^d)$ that contains density functions with finite q^{th} moments. We denote by $\mu_{\rm d}$ the probability measure on $\mathbb{R}^d$ induced by the underlying data distribution $\rho_{\rm d}\in\mathcal{P}(\mathbb{R}^d)$. For any $\mathbb{R}^d$-valued random variable Y, we denote by $\rho^Y\in\mathcal{P}(\mathbb{R}^d)$ the density function of Y (if it exists). Consider the second-order Wasserstein distance $\mathcal{W}_2(\rho_1,\rho_2):=(\inf\{\mathbb{E}[|X_1-X_2|^2]:\rho^{X_1}=\rho_1,\rho^{X_2}=\rho_2\})^{1/2}$, for $\rho_1,\rho_2\in\mathcal{P}_2(\mathbb{R}^d)$.

Given $n\in\mathbb{N}$, $S\subseteq\mathbb{R}^n$, and a measure ν on $\mathbb{R}^n$, we let $L^p(S,\nu)$, $p\in[1,\infty)$, be the set of $f:S\to\mathbb{R}$ with $\|f\|^p_{L^p(S,\nu)}:=\int_S|f|^p d\nu<\infty$, and $L^\infty(S,\nu)$ be the set of $f:S\to\mathbb{R}$ that are bounded ν-a.e. Also, we denote by $W^{1,p}(S,\nu)$ the set of $f:S\to\mathbb{R}$ whose weak derivatives of first order exist and lie in $L^p(S,\nu)$, and by $W^{1,p}_0(S,\nu)$ the closure of $C^\infty_c(S)$ (the set of infinitely differentiable $f:S\to\mathbb{R}$ with compact supports) under the $W^{1,p}(S,\nu)$-norm. For $p=2$, we write $H^1(S,\nu)=W^{1,2}(S,\nu)$ and $H^1_0(S,\nu)=W^{1,2}_0(S,\nu)$. When ν is the Lebesgue measure (denoted by Leb), we shorten the notation to $L^p(S)$, $W^{1,p}(S)$, and $H^1(S)$, etc. We denote by ∇ the Euclidean gradient operator with respect to (w.r.t.) $y\in\mathbb{R}^n$.

Let $\mathcal{X}$ be an arbitrary collection of $f:S\to\mathbb{R}$. For any $u:[0,\infty)\to\mathcal{X}$, we will often abuse the notation by writing $u(t,\cdot)=(u(t))(\cdot)\in\mathcal{X}$ for all $t\ge 0$ and treat $u(t,y)=(u(t))(y)$ as a generic function on $[0,\infty)\times S$.

2. Preliminaries

We study (1) with $d(\cdot,\cdot)$ therein taken to be the *Jensen-Shannon divergence*, i.e.,

$$\mathrm{JSD}(\rho,\bar\rho):=\frac{1}{2}D_{\rm KL}\left(\bar\rho\,\middle\|\,\frac{\bar\rho+\rho}{2}\right)+\frac{1}{2}D_{\rm KL}\left(\rho\,\middle\|\,\frac{\bar\rho+\rho}{2}\right),\quad\forall\rho,\bar\rho\in\mathcal{P}(\mathbb{R}^d),\tag{3}$$

where $D_{\rm KL}$ denotes the *Kullback-Leibler divergence*, defined by

$$D_{\rm KL}(\rho\|\bar\rho):=\int_{\mathbb{R}^d}\rho(x)\ln\left(\frac{\rho(x)}{\bar\rho(x)}\right)dx,\quad\forall\rho,\bar\rho\in\mathcal{P}(\mathbb{R}^d).\tag{4}$$

As shown in Endres and Schindelin (2003) and Theorem 1 in Österreicher and Vajda (2003), $\sqrt{\mathrm{JSD}(\cdot,\cdot)}$ defines a metric on $\mathcal{P}(\mathbb{R}^d)$, so that $\mathrm{JSD}(\cdot,\cdot)$ is a semi-metric. The problem (1) can now be stated as

$$\text{minimize} \quad J(\rho) := \mathrm{JSD}(\rho, \rho_{\mathrm{d}}) \quad \text{over } \mathcal{P}(\mathbb{R}^d). \tag{5}$$

Remark 1 *For any $\rho, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d)$, $0 \le \mathrm{JSD}(\rho, \bar{\rho}) \le \ln(2)$. This follows directly from*

$$D_{\mathrm{KL}}(\rho\|\bar{\rho}) = -\int_{\mathbb{R}^d} \rho(x) \ln\left(\frac{\bar{\rho}(x)}{\rho(x)}\right) dx \ge -\int_{\mathbb{R}^d} \rho(x)\left(\frac{\bar{\rho}(x)}{\rho(x)} - 1\right) dx = 0,$$
$$D_{\mathrm{KL}}\left(\rho \,\middle\|\, \frac{\rho+\bar{\rho}}{2}\right) = \int_{\mathbb{R}^d} \rho(x) \ln\left(\frac{2\rho(x)}{\rho(x)+\bar{\rho}(x)}\right) dx \le \int_{\mathbb{R}^d} \rho(x) \ln\left(\frac{2\rho(x)}{\rho(x)}\right) dx = \ln(2),$$

where we used the fact $\ln(x) \le x - 1$ $\forall x > 0$ in the first inequality.

Remark 2 *The metric $\sqrt{\mathrm{JSD}(\cdot,\cdot)}$ on $\mathcal{P}(\mathbb{R}^d)$ is equivalent to the total variation distance*

$$\mathrm{TV}(\rho, \bar{\rho}) := \sup_{A \in \mathcal{B}(\mathbb{R}^d)} \left| \int_A \rho(x)dx - \int_A \bar{\rho}(x)dx \right| \quad \forall \rho, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d). \tag{6}$$

This follows from $\mathrm{TV}(\rho, \bar{\rho}) = \frac{1}{2}\|\rho - \bar{\rho}\|_{L^1(\mathbb{R}^d)}$ (Lemma 34) and the estimate

$$\phi\left(\frac{1}{2}\|\rho - \bar{\rho}\|_{L^1(\mathbb{R}^d)}\right) \le 2\,\mathrm{JSD}(\rho, \bar{\rho}) \le \ln(2)\|\rho - \bar{\rho}\|_{L^1(\mathbb{R}^d)}, \quad \forall \rho, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d), \tag{7}$$

where $\phi : [0,1] \to \mathbb{R}_+$ is strictly increasing with $\phi(0) = 0$; see Theorem 2 (with $\beta = 1$ therein) in Österreicher and Vajda (2003).

Remark 3 *For any fixed $\bar{\rho} \in \mathcal{P}(\mathbb{R}^d)$, $\mathrm{JSD}(\cdot, \bar{\rho}) : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is strictly convex. Indeed, by Nielsen and Nock (2014) (see Table I therein), $\mathrm{JSD}(\cdot,\cdot)$ can be expressed as*

$$\mathrm{JSD}(\rho_1, \rho_2) = \int_{\mathbb{R}^d} f\left(\frac{\rho_1}{\rho_2}\right) \rho_2 dy, \quad \forall \rho_1, \rho_2 \in \mathcal{P}(\mathbb{R}^d), \tag{8}$$

with $f : [0,\infty) \to \mathbb{R}$ given by $f(x) := \frac{1}{2}\big[(x+1)\ln(\frac{2}{x+1}) + x\ln x\big]$. For any $\rho_1, \rho_2, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d)$ and $\lambda \in (0,1)$, from (8) and the strict convexity of f, we quickly obtain $\mathrm{JSD}(\lambda\rho_1 + (1-\lambda)\rho_2, \bar{\rho}) < \lambda\,\mathrm{JSD}(\rho_1, \bar{\rho}) + (1-\lambda)\,\mathrm{JSD}(\rho_2, \bar{\rho})$, i.e., $\mathrm{JSD}(\cdot, \bar{\rho})$ is strictly convex.

2.1 Gradient Descent for Functions on $\mathcal{P}(\mathbb{R}^d)$

For a strictly convex $f : \mathbb{R}^d \to \mathbb{R}$, it is well-known that the global minimizer $y^* \in \mathbb{R}^d$ of f, if it exists, can be found efficiently by (deterministic) gradient descent in the space $\mathbb{R}^d$. Specifically, for any initial point $y \in \mathbb{R}^d$, the ODE

$$dY_t = -\nabla f(Y_t)dt, \quad Y_0 = y \in \mathbb{R}^d, \tag{9}$$

converges to y^* as $t \to \infty$. Given that $\mathrm{JSD}(\cdot, \rho_{\mathrm{d}}) : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is strictly convex (Remark 3), it is natural to ask if its minimizer, i.e., $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$, can also be found by gradient descent,

now in the space $\mathcal{P}(\mathbb{R}^d)$. The ultimate question is how we should modify ODE (9) to accommodate a function defined on $\mathcal{P}(\mathbb{R}^d)$—particularly, how its gradient should be defined.

Given a strictly convex $G : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, we expect that the corresponding gradient-descent ODE takes the form

$$dY_t = -\partial_\rho G(\rho^{Y_t}, Y_t)dt, \quad \rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d), \tag{10}$$

where $\partial_\rho G(\rho^{Y_t}, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ denotes the "gradient of G at $\rho^{Y_t} \in \mathcal{P}(\mathbb{R}^d)$", which remains to be defined. As Y_0 is specified by a density $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ (but *not* as a fixed state $y \in \mathbb{R}^d$ in (9)), Y_t is a random variable, with density $\rho^{Y_t} \in \mathcal{P}(\mathbb{R}^d)$, for all $t \geq 0$. That is, (10) is now a *distribution-dependent* ODE. At any time $t \geq 0$, given the current density $\rho^{Y_t} \in \mathcal{P}(\mathbb{R}^d)$, $-\partial_\rho G(\rho^{Y_t}, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ dictates the direction along which each $y \in \mathbb{R}^d$ (i.e., a realization of Y_t) moves forward. Specifically, if we discretize ODE (10) with a fixed time step $\varepsilon > 0$, a realization $y \in \mathbb{R}^d$ of Y_t is transported to

$$\bar{y} := y - \varepsilon \partial_\rho G(\rho^{Y_t}, y) \in \mathbb{R}^d, \tag{11}$$

which represents a realization of $Y_{t+\varepsilon}$. It is highly nontrivial, however, to define the "gradient" $\partial_\rho G(\rho^{Y_t}, \cdot)$. As the domain $\mathcal{P}(\mathbb{R}^d)$ of G is not even a vector space, differentiation cannot be easily defined in the usual Fréchet or Gateaux sense.

A natural idea is to rely on the convexity of $\mathcal{P}(\mathbb{R}^d)$: for any $\rho, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d)$, we can move from ρ to $\bar{\rho}$ along the line segment $\{(1-\lambda)\rho + \lambda\bar{\rho} : \lambda \in [0,1]\}$ in $\mathcal{P}(\mathbb{R}^d)$ and study how G changes along the way. This leads to the notion of a *linear functional derivative*.

Definition 4 *A linear functional derivative of $G : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is a function $\frac{\delta G}{\delta \rho} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ that satisfies*

$$G(\bar{\rho}) - G(\rho) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta G}{\delta \rho}\big((1-\lambda)\rho + \lambda\bar{\rho}, y\big)(\bar{\rho} - \rho)(y)dyd\lambda, \quad \forall \rho, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d). \tag{12}$$

Under appropriate growth and continuity conditions of $\frac{\delta G}{\delta \rho}$, (12) implies

$$\lim_{\varepsilon \to 0} \frac{G(\rho + \varepsilon(\bar{\rho} - \rho)) - G(\rho)}{\varepsilon} = \int_{\mathbb{R}^d} \frac{\delta G}{\delta \rho}\big(\rho, y\big)(\bar{\rho} - \rho)(y)dy.$$

That is, $\frac{\delta G}{\delta \rho}(\rho, \cdot)$ can be viewed as the "gradient of G" if one moves along straight lines in $\mathcal{P}(\mathbb{R}^d)$ from one density to another.

The evolution of densities in ODE (10), nonetheless, is much more involved. At any time $t \geq 0$, $-\partial_\rho G(\rho^{Y_t}, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ serves as a vector field that moves any current point $y \in \mathbb{R}^d$ to a new location. Put differently, the transportation in (11) is of the general form

$$\bar{y} := y + \varepsilon\xi(y) = (I + \varepsilon\xi)(y) \in \mathbb{R}^d, \quad \text{for some } \xi : \mathbb{R}^d \to \mathbb{R}^d, \tag{13}$$

where I denotes the identity map on $\mathbb{R}^d$. Under (13), an initial probability measure μ on $\mathbb{R}^d$ is transformed into the pushforward measure

$$\mu_\varepsilon^\xi(\cdot) := \mu\left((I + \varepsilon\xi)^{-1}(\cdot)\right). \tag{14}$$

As the next result shows, when μ is induced by a density $\rho \in \mathcal{P}(\mathbb{R}^d)$, μ_ε^ξ also has a density $\rho_\varepsilon^\xi \in \mathcal{P}(\mathbb{R}^d)$; moreover, when one moves (nonlinearly) in $\mathcal{P}(\mathbb{R}^d)$ from ρ to ρ_ε^ξ, the appropriate "gradient of G at $\rho \in \mathcal{P}(\mathbb{R}^d)$" turns out to be a simple functional of $\frac{\delta G}{\delta \rho}(\rho, \cdot)$.

Proposition 5 *Let $G : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ has a linear functional derivative $\frac{\delta G}{\delta \rho} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$. Consider $\rho \in \mathcal{P}(\mathbb{R}^d)$ and the measure $\mu(\cdot) := \int_{\cdot} \rho(y)dy$ on $\mathbb{R}^d$. If ρ is of $C^1(\mathbb{R}^d)$, for any $\xi \in C_c^2(\mathbb{R}^d; \mathbb{R}^d)$, the measure μ_ε^ξ in (14) has a density $\rho_\varepsilon^\xi \in \mathcal{P}(\mathbb{R}^d)$, for $\varepsilon > 0$ small. Moreover, if $\frac{\delta G}{\delta \rho}(\rho, \cdot)$ is locally integrable on $\mathbb{R}^d$ and $\sup_{\lambda \in [0,1]} |\frac{\delta G}{\delta \rho}((1-\lambda)\rho + \lambda\bar{\rho}, \cdot) - \frac{\delta G}{\delta \rho}(\rho, \cdot)| \to 0$ locally uniformly as $\bar{\rho} \to \rho$ uniformly, then*

$$\lim_{\varepsilon \to 0} \frac{G(\rho_\varepsilon^\xi) - G(\rho)}{\varepsilon} = \int_{\mathbb{R}^d} \nabla \frac{\delta G}{\delta \rho}\big(\rho, y\big) \cdot \xi(y) d\mu(y). \tag{15}$$

Proposition 5, whose proof is relegated to Section A.2, conveys important messages. First, when every $y \in \mathbb{R}^d$ moves along the direction $\xi(y) \in \mathbb{R}^d$ (as in (13)), the original density $\rho \in \mathcal{P}(\mathbb{R}^d)$ is recast into $\rho_\varepsilon^\xi \in \mathcal{P}(\mathbb{R}^d)$. Furthermore, as (15) demonstrates, $\nabla \frac{\delta G}{\delta \rho}(\rho, y)$ specifies how moving along $\xi(y)$ changes the value of G. That is, $\nabla \frac{\delta G}{\delta \rho}(\rho, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is the proper "gradient of G at $\rho \in \mathcal{P}(\mathbb{R}^d)$" for the transportation in (13), which is the discretization of an ODE like (10). As a result, from now on we will write (10) as

$$dY_t = -\nabla \frac{\delta G}{\delta \rho}(\rho^{Y_t}, Y_t)dt, \quad \rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \tag{16}$$

2.2 Connection to the Lions and Wasserstein Derivatives

If we restrict ourselves to the subset $\mathcal{P}_2(\mathbb{R}^d)$ of $\mathcal{P}(\mathbb{R}^d)$, our gradient $\nabla \frac{\delta G}{\delta \rho}$ in fact coincides with the Lions and Wasserstein derivatives in the literature under suitable conditions. Let us briefly recall the definitions of these two kinds of derivatives in $\mathcal{P}_2(\mathbb{R}^d)$.

In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathbb{P}$ is an atomless measure, every $\mathbb{R}^d$-valued random variable $Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ has a density $\rho^Y \in \mathcal{P}_2(\mathbb{R}^d)$. Hence, we can associate to each $G : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ a lifted function $g : L^2(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ defined by $g(Y) := G(\rho^Y)$. As Fréchet differentiation is well-defined in the Banach space $L^2(\Omega, \mathcal{F}, \mathbb{P})$, the "derivative of G at $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$" can be defined as $Dg(Y_0)$, the Fréchet derivative of g at $Y_0 \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ with $\rho^{Y_0} = \rho_0$. By identifying $L^2(\Omega, \mathcal{F}, \mathbb{P})$ with its dual (as it is reflexive), we have $Dg(Y_0) \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Thanks to Proposition 5.25 in Carmona and Delarue (2018a), this map $Dg : L^2(\Omega, \mathcal{F}, \mathbb{P}) \to L^2(\Omega, \mathcal{F}, \mathbb{P})$ can be identified with a Borel function $\xi_{\rho_0} : \mathbb{R}^d \to \mathbb{R}^d$, i.e., $Dg(Y) = \xi_{\rho_0}(Y)$ a.s. for all $Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ with $\rho^Y = \rho_0$. This Borel function ξ_{ρ_0} is then called the *Lions derivative* of $G : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ at $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ (Carmona and Delarue, 2018a, Definition 5.22), and we will denote it by $\partial_\rho^L G(\rho_0, \cdot) := \xi_{\rho_0}(\cdot)$.

On the other hand, when $\mathcal{P}_2(\mathbb{R}^d)$ is equipped with the Wasserstein distance $\mathcal{W}_2$, we can borrow the idea from convex analysis to define the subdifferential (and superdifferential) of $G : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ at each $\rho \in \mathcal{P}_2(\mathbb{R}^d)$; see e.g., Definition 5.62 in Carmona and Delarue (2018a) and Definition 10.1.1 in Ambrosio et al. (2005). The *Wasserstein derivative* of G at $\rho \in \mathcal{P}_2(\mathbb{R})$, denoted by $\partial_\rho^W G(\rho, \cdot)$, is then defined as the unique element (if it exists) that resides in both the sub- and superdifferential of G at $\rho \in \mathcal{P}_2(\mathbb{R}^d)$.

Now, consider $G : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ such that $\nabla \frac{\delta G}{\delta \rho}(\rho, \cdot)$ is well-defined for all $\rho \in \mathcal{P}_2(\mathbb{R}^d)$. By Proposition 5.48 and Theorem 5.64 in Carmona and Delarue (2018a), under additional continuity and growth conditions on $\nabla \frac{\delta G}{\delta \rho}(\rho, \cdot)$, if the Lions derivative $\partial_\rho^L G(\rho, \cdot)$ exists and is continuous, then all three kinds of derivatives in discussion coincide, i.e.,

$$\partial_\rho^L G(\rho, \cdot) = \partial_\rho^W G(\rho, \cdot) = \nabla \frac{\delta G}{\delta \rho}(\rho, \cdot), \quad \forall \rho \in \mathcal{P}_2(\mathbb{R}^d). \tag{17}$$

Despite this, $\nabla \frac{\delta G}{\delta \rho}(\rho, \cdot)$ is more general in that it can be defined on $\mathcal{P}(\mathbb{R}^d)$, without the need of the $\mathcal{P}_2(\mathbb{R}^d)$ structure. It thus suits our study particularly; see the last two paragraphs in Section 2.3.

2.3 Problem Formulation

We aim at solving the problem (5) through ODE (16), with $G(\cdot)$ therein taken to be $J(\cdot) = \mathrm{JSD}(\cdot, \rho_{\mathrm{d}})$. The first task is to find the linear functional derivative $\frac{\delta J}{\delta \rho}$, which is characterized explicitly in the next result (whose proof is relegated to Appendix A.3).

Lemma 6 *$J : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ defined in (5) has a linear functional derivative given by*

$$\frac{\delta J}{\delta \rho}(\rho, y) = \frac{1}{2} \ln \left(\frac{2\rho(y)}{\rho_{\mathrm{d}}(y) + \rho(y)} \right), \quad \forall \rho \in \mathcal{P}(\mathbb{R}^d) \text{ and } y \in \mathbb{R}^d. \tag{18}$$

Thanks to (18), once we replace G by J in ODE (16), we obtain ODE (2). Our goal is show that this distribution-dependent ODE has a (unique) solution Y and the induced density flow $\{\rho^{Y_t}\}_{t \geq 0}$ converges in $\mathcal{P}(\mathbb{R}^d)$ to the data distribution ρ_{d}.

Let us stress that we will work with the general space $\mathcal{P}(\mathbb{R}^d)$ (under the total variation distance), instead of its subset $\mathcal{P}_2(\mathbb{R}^d)$ (under the $\mathcal{W}_2$ distance), although the latter is more tractable with a comprehensive theory well-developed (see e.g., Ambrosio et al. (2005)). Such a non-standard setup is essential to our study, for two reasons.

First, working with $\mathcal{P}_2(\mathbb{R}^d)$ implicitly assumes $\rho_{\mathrm{d}} \in \mathcal{P}_2(\mathbb{R}^d)$, while ρ_{d}, as the underlying data distribution, generally lies in $\mathcal{P}(\mathbb{R}^d)$. Second, using the total variation distance ensures that our study is consistent inherently. Since we chose in the first place to measure the distance between densities by $\mathrm{JSD}(\cdot, \cdot)$ in (3), such a distance function should be kept consistently throughout our analysis. Particularly, the desired convergence "$\rho^{Y_t} \to \rho_{\mathrm{d}}$" should be established as $\mathrm{JSD}(\rho^{Y_t}, \rho_{\mathrm{d}}) \to 0$. As convergence in JSD is equivalent to that in total variation (Remark 2), using the total variation distance is a valid, consistent choice.

3. Main Results and Contributions

We will impose the following standing assumption on $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$.

Assumption 6.1 *$\rho_{\mathrm{d}} > 0$ on $\mathbb{R}^d$, $\ln \rho_{\mathrm{d}} \in L^\infty_{\mathrm{loc}}(\mathbb{R}^d) \cap H^1(\mathbb{R}^d, \mu_{\mathrm{d}})$.*

The local boundedness of $\ln \rho_{\mathrm{d}}$ ensures that the weighted Sobolev space $W^{1,p}(\mathbb{R}^d, \mu_{\mathrm{d}})$ is a Banach space for all $p \geq 1$. Also, $\ln \rho_{\mathrm{d}} \in H^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ entails $\int_{\mathbb{R}^d} \frac{|\nabla \rho_{\mathrm{d}}|^2}{\rho_{\mathrm{d}}} dy < \infty$, i.e., the Fisher information of μ_{d} (relative to the Lebesgue measure) is finite.

Let us present below the main theoretic result of this paper.

Theorem 7 *There exists a solution Y to ODE (2) such that $\eta(t) := \rho^{Y_t} \in \mathcal{P}(\mathbb{R}^d)$ fulfills*

$$\eta \in C([0, \infty); L^1(\mathbb{R}^d)), \quad \eta / \rho_{\mathrm{d}} \in L^\infty_+([0, \infty) \times \mathbb{R}^d), \quad \text{and} \quad \nabla \eta \in L^1_{\mathrm{loc}}([0, \infty) \times \mathbb{R}^d). \tag{19}$$

Moreover, for any such a solution Y, $\rho^{Y_t} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$ as $t \to \infty$.

Theorem 7 constructs an "amenable" solution Y to the gradient-descent ODE (2), such that ρ^{Y_t} evolves continuously in t, bounded by a multiple of ρ_{d}, and its (weak) derivative is locally integrable. Moreover, its density flow $\{\rho^{Y_t}\}_{t\geq 0}$ leads us to the data distribution ρ_{d}. This suggests that one can uncover ρ_{d} by simulating ODE (2) and a corresponding algorithm is developed in Section 3.1 below. The proof of Theorem 7, based on all the developments in Sections 4, 5, and 6, is relegated to the end of Section 6.

It is worth noting that gradient flows have recently been used in many implicit generative models (see Gao et al. (2019), Gao et al. (2020), Ansari et al. (2021), and Mroueh and Nguyen (2021), among others). Several mathematical issues, however, persist in this line of research. A common, and perhaps most critical, one is the "*inconsistent use of distance functions*": one proposes to minimize the distance (e.g., an f-divergence) between the current density ρ_t and ρ_{d} along a gradient flow, but carries out theoretic proofs (partially or entirely) under a different distance function (e.g., $\mathcal{W}_2$).

As the choice of a distance function may not be arbitrary but dependent on the actual data set (see McCulloch and Wagner (2020)), this can be problematic: even if $\mathcal{W}_2(\rho_t, \rho_{\mathrm{d}}) \to 0$ is proved, simply because the metric $\mathcal{W}_2$ is *not* equivalent to an f-divergence, ρ_t need not get close to ρ_{d} under an f-divergence. That is, convergence under the *desired* distance function (i.e., the initially chosen one) remains in question. Also, the use of $\mathcal{W}_2$ readily requires ρ_{d} to lie in $\mathcal{P}_2(\mathbb{R}^d)$, where $\mathcal{W}_2$ is well-defined, while ρ_{d} generally belongs to $\mathcal{P}(\mathbb{R}^d)$.

Theorem 7, remarkably, overcomes all the mathematical issues. By taking the *target* distance function to be the Jensen-Shannon divergence (i.e., JSD in (3), a specific f-divergence), we work directly with the general space $\mathcal{P}(\mathbb{R}^d)$. The *actual* distance function in use is the total variation distance (i.e., the L^1 distance), which is equivalent to JSD. Hence, "$\rho^{Y_t} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$" established in Theorem 7 readily implies "$\mathrm{JSD}(\rho^{Y_t}, \rho_{\mathrm{d}}) \to 0$" along the same gradient flow. That is, convergence to $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$ is not only achieved, but achieved under the originally chosen target distance function.

This does not come at no cost. Working generally in $\mathcal{P}(\mathbb{R}^d)$ (under total variation) deprives us of the full-fledged gradient-flow theory in $\mathcal{P}_2(\mathbb{R}^d)$ (under $\mathcal{W}_2$), as elaborated in Ambrosio et al. (2005). The first major challenge is the existence of a solution to ODE (2). While we can view (2) as a degenerate McKean-Vlasov SDE without the diffusion term, the involved distribution-dependence is unusual. A McKean-Vlasov SDE typically involves $\mathcal{L}(Y_t)$, the law of the state process Y at time t. In general, an interacting particle system can be constructed so that the particles' empirical measure approximates $\mathcal{L}(Y_t)$, which leads to a solution to the SDE in the limit. This classical framework does not easily accommodate other forms of distribution-dependence. Specifically, (2) depends *not* on $\mathcal{L}(Y_t)$ explicitly, but rather on the Radon-Nikodym derivative of $\mathcal{L}(Y_t)$ w.r.t. the Lebesgue measure (i.e., $\rho^{Y_t} \in \mathcal{P}(\mathbb{R}^d)$) and its Euclidean derivative (i.e., $\nabla\rho^{Y_t}$). To overcome this, we first derive the nonlinear Fokker-Planck equation associated with (2) and show that it has a unique weak solution u, by the Crandall-Liggett theorem for differential equations in Banach spaces (Section 4). Next, from this function u, we build a solution $\{Y_t\}_{t\geq 0}$ to (2) that satisfies (19), using the superposition principle for SDEs (Section 5).

Another hurdle to Theorem 7 is the lack of suitable stochastic calculus. In $\mathcal{P}(\mathbb{R}^d)$ (under total variation), the dynamics of $t \mapsto \mathrm{JSD}(\rho^{Y_t}, \rho_{\mathrm{d}})$ does not follow from the standard Itô formula for a flow of measures, which is established only in $\mathcal{P}_2(\mathbb{R}^d)$ (under $\mathcal{W}_2$). We instead

carry out a detailed analysis of the nonlinear Fokker-Planck equation, which reveals how ρ^{Y_t} evolves in $\mathcal{P}(\mathbb{R}^d)$ (Section 6).

3.1 Simulating ODE (2)

To simulate ODE (2), one quickly realizes the challenge that its dynamics depends on ρ_{d}, which is unknown and simply what we intend to uncover. To circumvent this, we choose *not* to estimate $\rho^{Y_t} \in \mathcal{P}(\mathbb{R}^d)$ in (2) explicitly, *but* to find a function $G_t : \mathbb{R}^n \to \mathbb{R}^d$ (for a fixed $n \leq d$) such that $G_t(Z)$ approximates Y_t, where Z is a fixed $\mathbb{R}^n$-valued random variable (independent of t) that can be easily simulated (e.g., the multivariate Gaussian). By substituting $\rho^{G_t(Z)}$ for ρ^{Y_t} in (2), a direct calculation shows that (2) takes the form

$$dY_t = \frac{\nabla D_t(Y_t)}{2(1 - D_t(Y_t))}dt, \quad \textit{where} \quad D_t(y) := \frac{\rho_{\mathrm{d}}(y)}{\rho_{\mathrm{d}}(y) + \rho^{G_t(Z)}(y)} \ \ \forall y \in \mathbb{R}^d. \tag{20}$$

Remarkably, when G_t is given, one can estimate D_t *without* the knowledge of ρ_{d} or $\rho^{G_t(Z)}$. Indeed, by Proposition 1 in Goodfellow et al. (2014), D_t is the unique maximizer, among all $D : \mathbb{R}^d \to [0,1]$, of $\mathbb{E}_{y\sim\rho_{\mathrm{d}}}[\ln D(y)] + \mathbb{E}_{z\sim\rho^Z}\left[\ln\left(1 - D\left(G_t(z)\right)\right)\right]$. Therefore, it can be numerically approximated through data sampling.

In Algorithm 1 below, we take $G : \mathbb{R}^n \to \mathbb{R}^d$ and $D : \mathbb{R}^d \to [0,1]$ to be artificial neural networks and update their parameters (denoted by θ_G and θ_D) recursively; namely, (G, D) represents the time-varying (G_t, D_t) introduced above. Specifically, (i) we follow the first half of Algorithm 1 in Goodfellow et al. (2014) to estimate D with G given (Lines 2-4); (ii) plug D into (20) to simulate Y over a small time step $\varepsilon > 0$, resulting in the new points $\{y^{(i)}\}_{i=1}^m$ (Lines 5-6); (iii) update G by reducing the mean squared error (MSE) between $\{G(z^{(i)})\}_{i=1}^m$ and $\{y^{(i)}\}_{i=1}^m$ (Line 7). In (22), $\{G(z^{(i)})\}_{i=1}^m$ represents the realizations of Y_t at some $t \geq 0$. Through ODE (20), $\{G(z^{(i)})\}_{i=1}^m$ are transferred to the points $\{y^{(i)}\}_{i=1}^m$, which represent the realizations of $Y_{t+\varepsilon}$. Hence, the purpose of (iii) above is to modify G so that $\{G(z^{(i)})\}_{i=1}^m$ can properly represent $Y_{t+\varepsilon}$, instead of the previous Y_t.

Algorithm 1 is a specific, straightforward way to simulating ODE (2). Modifications can potentially be made based on more sophisticated simulation techniques for ODEs.

3.2 Connection to GANs

GANs in Goodfellow et al. (2014) encode the competition between a generator and a discriminator. The generator produces fake data points by sampling $G(Z)$, where Z is a fixed random variable with a (simple) density $\rho^Z \in \mathcal{P}(\mathbb{R}^n)$ and $G : \mathbb{R}^n \to \mathbb{R}^d$ is a (complicated) function chosen by the generator. To each data point $y \in \mathbb{R}^d$ (which can be real or fake), the discriminator assigns $D(y) \in [0,1]$, her subjective probability of y being real, where $D : \mathbb{R}^d \to [0,1]$ is chosen by the discriminator. The resulting min-max game is given by

$$\min_G \max_D \left\{\mathbb{E}_{Y\sim\rho_{\mathrm{d}}}[\ln D(Y)] + \mathbb{E}_{Z\sim\rho^Z}\left[\ln\left(1 - D\left(G(Z)\right)\right)\right]\right\}. \tag{24}$$

Thanks to Theorem 1 in Goodfellow et al. (2014), this minimax problem is equivalent to $\min_{G:\mathbb{R}^n\to\mathbb{R}^d} \mathrm{JSD}(\rho^{G(Z)}, \rho_{\mathrm{d}})$ and it admits a minimizer G^* such that $\rho^{G^*(Z)} = \rho_{\mathrm{d}}$. The GAN algorithm (i.e., Algorithm 1 in Goodfellow et al. (2014)) is proposed to find G^* (and thus ρ_{d}) numerically, through recursive optimization between the two players.

Algorithm 1 Simulating ODE (2)

Require: $m \in \mathbb{N}$, $\varepsilon > 0$

1: **for** number of training iterations **do**

2: • Sample minibatch of m noise samples $\{z^{(1)}, ..., z^{(m)}\}$ from noise prior ρ^Z.

3: • Sample minibatch of m examples $\{x^{(1)}, ..., x^{(m)}\}$ from the data distribution $\rho_{\rm d}$.

4: • Update $D : \mathbb{R}^d \to [0, 1]$ by ascending its stochastic gradient:

$$\nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \left[\ln D(x^{(i)}) + \ln \left(1 - D\left(G(z^{(i)}) \right) \right) \right]. \tag{21}$$

5: • Sample minibatch of m noise samples $\{z^{(1)}, ..., z^{(m)}\}$ from noise prior ρ^Z.

6: • Set $Y = \{y^{(1)}, ..., y^{(m)}\}$ by

$$y^{(i)} := G(z^{(i)}) + \frac{\nabla D(G(z^{(i)}))}{2(1 - D(G(z^{(i)})))} \varepsilon, \quad \forall i = 1, 2, ..., m. \tag{22}$$

7: • Update $G : \mathbb{R}^n \to \mathbb{R}^d$ by descending its stochastic gradient:

$$\nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} |G(z^{(i)}) - y^{(i)}|^2. \tag{23}$$

8: **end for**

Somewhat surprisingly, Algorithm 1 above, designed to simulate ODE (2), is equivalent to the GAN algorithm in Goodfellow et al. (2014).

Proposition 8 *Algorithm 1 above is equivalent to the GAN algorithm (i.e., Algorithm 1 in Goodfellow et al. (2014)), up to an adjustment of the learning rates.*

Proof As the update of D in (21) is the same as that in the first half of the GAN algorithm, it suffices to show that the update of G via (22)-(23) is equivalent to that in the second half of the GAN algorithm. By a direct calculation of (23), we get

$$\begin{aligned} \nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} |G(z^{(i)}) - y^{(i)}|^2 &= \frac{2}{m} \sum_{i=1}^{m} (G(z^{(i)}) - y^{(i)}) \cdot \nabla_{\theta_G} G(z^{(i)}) \\ &= \frac{2}{m} \sum_{i=1}^{m} \left(\frac{-\nabla D(G(z^{(i)}))}{2(1 - D(G(z^{(i)})))} \varepsilon \right) \cdot \nabla_{\theta_G} G(z^{(i)}) = \varepsilon \nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \ln \left(1 - D(G(z^{(i)})) \right), \end{aligned}$$

where the second equality follows from (22). The result follows by noting that the right-hand side above is precisely the update of G specified in the GAN algorithm multiplied by $\varepsilon > 0$; see the last equation in (Goodfellow et al., 2014, Algorithm 1). ■

Proposition 8 yields an intriguing interpretation for GANs: the *non-cooperative* game between the generator and discriminator can be equivalently viewed as a *cooperative* game

between a *navigator* and a *calibrator*. The navigator aims to sail across the space $\mathcal{P}(\mathbb{R}^d)$ following ODE (2). At any initial location in $\mathcal{P}(\mathbb{R}^d)$ (represented by an initial $G : \mathbb{R}^n \to \mathbb{R}^d$), he finds that he cannot set off right away, as the dynamics of (2) involves the unknown ρ_{d}. The calibrator comes into play to "calibrate" the dynamics of (2)—namely, estimate $D : \mathbb{R}^d \to [0, 1]$ in (20) using (21), which is the discriminator's optimization in the GAN algorithm. Once the calibration is done, the navigator travels a short distance in $\mathcal{P}(\mathbb{R}^d)$ following (2), updates $G : \mathbb{R}^n \to \mathbb{R}^d$ to reflect his new location, and asks the calibrator to re-calibrate. This corresponds to (22)-(23), equivalent to the generator's optimization in the GAN algorithm (by Proposition 8). To the best of our knowledge, a possible "cooperative" view of GANs was only briefly touched on, without details, at the end of p. 21 in Goodfellow (2016). Proposition 8 materializes this idea with great specifics: the two players cooperate precisely to simulate the gradient-descent ODE (2).

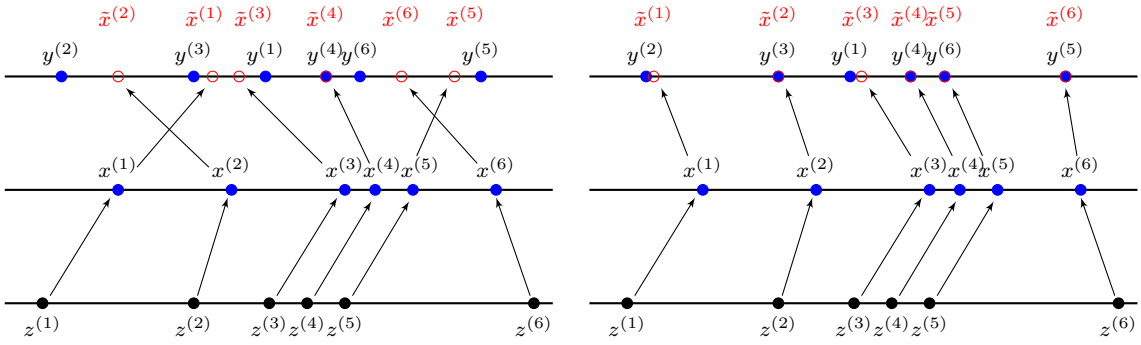
Proposition 8 also contrasts with a common practice in implicit generative modeling. Many algorithms have replaced the minimization part of a given GAN algorithm by a kind of gradient descent. For example, Gao et al. (2019), Ansari et al. (2021), and Gao et al. (2020) replace the minimization part of f-GAN by gradient descent in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$, and Mroueh and Nguyen (2021) replaces the minimization part of MMD GAN by gradient descent on a statistical manifold. The resulting gradient-flow algorithms are then advocated as alternatives (and improvements) to the initially-given GAN algorithms. Contrary to the common practice "modifying GANs to form a gradient-flow algorithm", Proposition 8 shows that the (vanilla) GAN algorithm itself, without any modification, already computes gradient flows. As detailed in Section 3.3 below, this gradient-flow identification of GANs crucially reveals an unapparent fact: the (vanilla) GAN algorithm implicitly involves an MSE fitting, which can cause or aggravate the divergence of GANs.

Note that Algorithm 1 resembles the gradient-flow algorithms in Gao et al. (2019) and Gao et al. (2020), but a subtle difference sets them apart. In Algorithm 1, the estimation of D (or equivalently, $\rho^{G(Z)}/\rho_{\mathrm{d}}$) is coordinated with the gradient-descent update of G, so that the *zero-sum* game (24) of GANs can be recovered. In Gao et al. (2019) and Gao et al. (2020), as the estimation of $\rho^{G(Z)}/\rho_{\mathrm{d}}$ and the update of G are attached to different objectives, their algorithms represent *non-zero-sum* games.

3.3 Implication for the Divergence of GANs

Mathematically, the recursive optimization in the GAN algorithm may compute either the desired min-max game value (24) or the corresponding max-min game value. If it computes the latter, the algorithm easily diverges in the form of *mode collapse*, i.e., the generator keeps producing $G : \mathbb{R}^n \to \mathbb{R}^d$ that maps many distinct $z \in \mathbb{R}^n$ to similar $y \in \mathbb{R}^d$ and fails to recover the entirety of ρ_{d}; see Section 5.1.1 of Goodfellow (2016). Beyond this typical "max-min game" diagnosis, our gradient-flow analysis zooms in on the GAN algorithm and points to an additional cause for the divergence of GANs.

By the proof of Proposition 8, the update of G in the GAN algorithm can be decomposed into two steps: generating new points $\{y^{(i)}\}_{i=1}^m$ along ODE (2) (i.e., (22)) and fitting $\{G(z^i)\}_{i=1}^m$ to $\{y^{(i)}\}_{i=1}^m$ by minimizing the MSE (i.e., (23)). While the MSE fitting demands *point-wise* similarity (i.e., $G(z^{(i)})$ is close to $y^{(i)}$, for $i = 1, 2, ..., m$), what we truly need is only *set-wise* similarity (i.e., the distribution of $\{G(z^{(i)})\}_{i=1}^m$ is similar to that of $\{y^{(i)}\}_{i=1}^m$).



(a) Point-wise similarity (b) Set-wise similarity

Figure 1: An initial $G : \mathbb{R} \to \mathbb{R}$ maps samples $\{z^{(i)}\}_{i=1}^6$ of Z to $\{x^{(i)}\}_{i=1}^6$. Two distinct $\tilde{G} : \mathbb{R} \to \mathbb{R}$ serve to relocate $\{x^{(i)}\}_{i=1}^6$ to approximate $\{y^{(i)}\}_{i=1}^6$, the new points obtained in (22), under the criterion of point-wise similarity and set-wise similarity, respectively. The red circles represent $\tilde{x}^{(i)} := \tilde{G}(z^{(i)})$, $i = 1, ..., 6$.

In Figure 1 (with $n = d = 1$), for an initial $G : \mathbb{R} \to \mathbb{R}$ and samples $\{z^{(i)}\}_{i=1}^6$ of Z, each $x^{(i)} := G(z^{(i)})$ is transported to $y^{(i)}$ via (22). Our goal is to modify G such that $\{x^{(i)}\}_{i=1}^6$ becomes similar to $\{y^{(i)}\}_{i=1}^6$. To achieve point-wise similarity, the MSE fitting (23) moves each $x^{(i)} = G(z^{(i)})$ toward $y^{(i)}$. But since the distance from $x^{(i)}$ to $y^{(i)}$ can be quite large, the movement of $x^{(i)}$ will stop halfway, once the gradient updates ∇_{θ_G} in (23) come to a halt near a local minimizer or saddle point in the parameter space for θ_G. In Figure 1(a), $x^{(i)}$ stops halfway at $\tilde{x}^{(i)} := \tilde{G}(z^{(i)})$ before reaching $y^{(i)}$ (except $i = 4$, for which $x^{(4)}$ and $y^{(4)}$ are close), where $\tilde{G}$ is the updated G obtained from (23). As the distribution of $\{\tilde{x}^{(i)}\}_{i=1}^6$ is distinct from that of $\{y^{(i)}\}_{i=1}^6$, we go astray from the gradient flow toward $\rho_{\rm d}$.

The takeaway is that the MSE fitting *alone* can cause GANs to diverge. When $\rho^{G(Z)}$ is distinct from $\rho_{\rm d}$, the term $\frac{\nabla D(y)}{2(1-D(y))}$ in ODE (20) can be large, so that $G(z^{(i)})$ and $y^{(i)}$ in (22) can be far apart. In this case, as explained above, the MSE fitting (23) tends to drive us away from the gradient flow toward $\rho_{\rm d}$. More precisely, the updated G moves $\rho^{G(Z)}$ *off* the gradient flow, such that it does not move closer to $\rho_{\rm d}$ in any reasonable sense. As the updated $\rho^{G(Z)}$ remains distinct from $\rho_{\rm d}$, the same argument above indicates that another training iteration is unlikely to help: it will simply update $\rho^{G(Z)}$ again to another distribution distinct from $\rho_{\rm d}$. That is, Algorithm 1 (equivalent to the GAN algorithm) will generate indefinitely one distribution after another, all of which stay distinct from $\rho_{\rm d}$. This is reminiscent of mode collapse in Figure 2 of Metz et al. (2017), where $\rho^{G(Z)}$ rotates among several distributions distinct from $\rho_{\rm d}$ and never converges.

In fact, if $x^{(i)} = G(z^{(i)})$ is allowed to move toward a nearby $y^{(j)}$, with j possibly different from i, the update of G can be made more effective. In Figure 1(b), as each $x^{(i)}$ only needs to travel a short distance to a nearby $y^{(j)}$, it is likely to reach (or get very close to) $y^{(j)}$, before the gradient updates ∇_{θ_G} come to a halt near a local minimizer or saddle point. The resulting distribution on $\mathbb{R}$ of $\{\tilde{G}(z^{(i)})\}_{i=1}^6$ is then very similar to that of $\{y^{(i)}\}_{i=1}^6$. This suggests a new potential route to encourage convergence of GANs: replacing the MSE (a

measure of point-wise similarity) in (23) by a measure of set-wise similarity. This warrants a detailed analysis in itself and will be left for future research.

4. The Nonlinear Fokker-Planck Equation

If there is a solution Y to ODE (2) (which remains to be proved), a heuristic calculation shows that the density flow $u(t,\cdot) := \rho^{Y_t}(\cdot) \in \mathcal{P}(\mathbb{R}^d)$, $t \geq 0$, satisfies the nonlinear Fokker-Planck equation

$$\frac{\partial u}{\partial t}(t,y) = \frac{1}{2}\operatorname{div}\left(\left(\frac{\nabla u(t,y)}{u(t,y)} - \frac{\nabla \rho_{\mathrm{d}}(y) + \nabla u(t,y)}{\rho_{\mathrm{d}}(y) + u(t,y)}\right) u(t,y)\right), \quad u(0,y) = \rho_0(y). \tag{25}$$

In view of this, we will construct a solution to (2) in a *backward* manner. In this section, we will establish the existence of a unique solution u to (25). The function u will then be used, in Section 5 below, to construct a solution Y to (2). This backward method was recently introduced in Barbu and Röckner (2020), for a McKean-Vlasov SDE depending on ρ^{Y_t}. As we will see, our case is markedly more involved than Barbu and Röckner (2020), due to the additional dependence on $\nabla\rho^{Y_t}$ in (2).

4.1 Preparations

For any fixed density flow $\{\rho_t\}_{t\geq 0}$ in $\mathcal{P}(\mathbb{R}^d)$, whenever $\rho_t(\cdot)$ is weakly differentiable, we introduce the infinitesimal generator

$$\mathcal{L}_{\rho_t}\varphi(t,y) := -\frac{1}{2}\left(\frac{\nabla \rho_t(t,y)}{\rho_t(t,y)} - \frac{\nabla \rho_{\mathrm{d}}(y) + \nabla \rho_t(t,y)}{\rho_{\mathrm{d}}(y) + \rho_t(t,y)}\right)\cdot \nabla\varphi, \quad \forall \varphi \in C^{1,1}((0,\infty)\times\mathbb{R}^d). \tag{26}$$

Thanks to integration by parts, we define a weak solution to (25) as follows.

Definition 9 *We say $u : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ is a weak solution to the nonlinear Fokker-Planck equation (25), if $u(t,\cdot)$ is weakly differentiable for a.e. $t \geq 0$ such that*

$$\int_0^\infty \int_{\mathbb{R}^d} \left(\varphi_t(t,y) + \mathcal{L}_{u(t,\cdot)}\varphi(t,y)\right) u(t,y)dydt = 0, \quad \forall \varphi \in C_c^{1,2}((0,\infty)\times\mathbb{R}^d). \tag{27}$$

Remark 10 *As we assume $\varphi(t,\cdot) \in C^2(\mathbb{R}^d)$ in (27), using integration by parts twice gives*

$$\int_{\mathbb{R}^d} \left(\mathcal{L}_{u(t,\cdot)}\varphi(t,y)\right) u(t,y)dy = \int_{\mathbb{R}^d} \left(\frac{1}{2}\frac{\nabla \rho_{\mathrm{d}}(y) + \nabla \rho_t(y)}{\rho_{\mathrm{d}}(y) + \rho_t(y)}\cdot\nabla\varphi + \frac{1}{2}\Delta\varphi\right) u(t,y)dy.$$

Hence, (27) can be equivalently stated as

$$\int_0^\infty \int_{\mathbb{R}^d} \left(\varphi_t(t,y) + \widetilde{\mathcal{L}}_{u(t,\cdot)}\varphi(t,y)\right) u(t,y)dydt = 0, \quad \forall \varphi \in C_c^{1,2}((0,\infty)\times\mathbb{R}^d), \tag{28}$$

where the generator $\widetilde{\mathcal{L}}_{\rho_t}$ is defined, for any fixed density flow $\{\rho_t\}_{t\geq 0}$ in $\mathcal{P}(\mathbb{R}^d)$, by

$$\widetilde{\mathcal{L}}_{\rho_t}\varphi(t,y) := \frac{1}{2}\frac{\nabla \rho_{\mathrm{d}}(y) + \nabla \rho_t(y)}{\rho_{\mathrm{d}}(y) + \rho_t(y)}\cdot\nabla\varphi + \frac{1}{2}\Delta\varphi, \quad \forall \varphi \in C^{1,2}(\mathbb{R}_+\times\mathbb{R}^d). \tag{29}$$

That is to say, the Fokker-Planck equation (25) is equivalent (in the weak sense) to

$$\frac{\partial u}{\partial t}(t,y) = -\operatorname{div}\left(\frac{1}{2}\frac{\nabla \rho_{\mathrm{d}}(y) + \nabla u(t,y)}{\rho_{\mathrm{d}}(y) + u(t,y)} u(t,y)\right) + \frac{1}{2}\Delta u(t,y), \quad u(0,y) = \rho_0(y). \tag{30}$$

To find a weak solution u to (25), or equivalently (30), let us turn (30) into a more amenable PDE. As $\rho_{\mathrm{d}} > 0$ (Assumption 6.1), we consider the ansatz

$$v(t,y) := \frac{u(t,y)}{\rho_{\mathrm{d}}(y)}, \quad \forall (t,y) \in [0,\infty) \times \mathbb{R}^d.$$

It follows that $u_t = \rho_{\mathrm{d}} v_t$, so that

$$\begin{aligned} -\frac{1}{2}\operatorname{div}\left(\frac{\nabla(\rho_{\mathrm{d}}+u)}{\rho_{\mathrm{d}}+u}u\right) + \frac{1}{2}\Delta u &= -\frac{1}{2}\operatorname{div}\left(\nabla(\rho_{\mathrm{d}}+u) - \frac{\nabla(\rho_{\mathrm{d}}(1+v))}{1+v}\right) + \frac{1}{2}\Delta u \\ &= -\frac{1}{2}\operatorname{div}\left(\nabla\rho_{\mathrm{d}} - \frac{\rho_{\mathrm{d}}\nabla(1+v) + (1+v)\nabla\rho_{\mathrm{d}}}{1+v}\right) \\ &= \frac{1}{2}\operatorname{div}(\rho_{\mathrm{d}}\nabla\ln(1+v)) = \frac{1}{2}\left(\rho_{\mathrm{d}}\Delta\ln(1+v) + \nabla\rho_{\mathrm{d}}\cdot\nabla\ln(1+v)\right). \end{aligned}$$

Hence, (30) can be written in terms of v as

$$v_t = \frac{1}{2}\Delta_{\mu_{\mathrm{d}}}\ln(1+v), \quad v(0) = \rho_0/\rho_{\mathrm{d}}. \tag{31}$$

where the operator $\Delta_{\mu_{\mathrm{d}}}$ is defined by

$$\Delta_{\mu_{\mathrm{d}}} := \Delta + \nabla\ln\rho_{\mathrm{d}}\cdot\nabla. \tag{32}$$

Remark 11 *Thanks to integration by parts,*

$$\int_{\mathbb{R}^d}(\Delta_{\mu_{\mathrm{d}}}w)\varphi d\mu_{\mathrm{d}} = \int_{\mathbb{R}^d}(\Delta w + \nabla\ln\rho_{\mathrm{d}}\cdot\nabla w)\varphi d\mu_{\mathrm{d}} = -\int_{\mathbb{R}^d}\nabla w\cdot\nabla\varphi d\mu_{\mathrm{d}}, \quad \forall w,\varphi \in C_c^\infty(\mathbb{R}^d).$$

That is, we can view $\Delta_{\mu_{\mathrm{d}}}$ as the "Laplacian" in the weighted Sobolev space $W^{1,2}(\mathbb{R}^d,\mu_{\mathrm{d}})$.

In light of Remark 11, we define weak solutions to (31) as below.

Definition 12 *We say $v : [0,\infty) \to L^1(\mathbb{R}^d,\mu_{\mathrm{d}})$ is a weak solution to* (31) *w.r.t. μ_{d}, if*

$$\int_0^\infty\int_{\mathbb{R}^d}\left(v(t,y)\varphi_t(t,y) + \frac{1}{2}\ln(1+v(t,y))\Delta_{\mu_{\mathrm{d}}}\varphi(t,y)\right)d\mu_{\mathrm{d}}dt = 0, \quad \forall\varphi \in C_c^{1,2}((0,\infty)\times\mathbb{R}^d). \tag{33}$$

Note that a weak solution to (31) is defined *not* under the standard Lebesgue integration on $\mathbb{R}^d$ (in contrast to (27) and the classical definition of weak solutions), but under the integration w.r.t. μ_{d}, the probability measure on $\mathbb{R}^d$ induced by $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$.

Now, let us recast the transformed Fokker-Planck equation (31) as a nonlinear Cauchy problem in the Banach space $L^1(\mathbb{R}^d,\mu_{\mathrm{d}})$. Specifically, under the condition

$$\rho_0 \le \beta\rho_{\mathrm{d}} \quad \text{for some} \quad \beta > 0, \tag{34}$$

we define the operator $A : D(A) \subseteq L^1(\mathbb{R}^d,\mu_{\mathrm{d}}) \to L^1(\mathbb{R}^d,\mu_{\mathrm{d}})$ by

$$Av := -\frac{1}{2}\Delta_{\mu_{\mathrm{d}}}\ln(1+v) = -\frac{1}{2}\Delta\ln(1+v) - \frac{1}{2}\nabla\ln\rho_{\mathrm{d}}\cdot\nabla\ln(1+v), \tag{35}$$

where

$$D(A) := \{v \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}}) \cap H_0^1(\mathbb{R}^d, \mu_{\mathrm{d}}) : 0 \le v \le \beta,\ Av \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}})\}$$

and the differentiation in (35) is understood in the weak sense. Then, (31) can be expressed as the Cauchy problem

$$v_t + Av = 0, \quad v(0) = \rho_0/\rho_{\mathrm{d}}. \tag{36}$$

Remark 13 *Condition* (34) *states that the Radon-Nikodym derivative of the initial probability measure (induced by $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$) w.r.t. the underlying measure μ_{d} is bounded. This covers all bounded and compactly supported $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$. Also, note that $\beta \ge 1$ necessarily.*

Our first finding is that both operators A and $-\Delta_{\mu_{\mathrm{d}}}$ are "accretive" in $L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$. To properly state this property, we will denote by I the identity map on $L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ (i.e., $I(v) = v$ for all $v \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$) in the next result, whose proof is relegated to Appendix B.1.

Lemma 14 *(i) The operator $A : D(A) \to L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ is accretive. That is,*

$$\|v_1 - v_2\|_{L^1(\mathbb{R}^d, \mu_{\mathrm{d}})} \le \|(I + \lambda A)v_1 - (I + \lambda A)v_2\|_{L^1(\mathbb{R}^d, \mu_{\mathrm{d}})}, \quad \forall v_1, v_2 \in D(A),\ \lambda > 0.$$

(ii) The operator $-\Delta_{\mu_{\mathrm{d}}} : D(-\Delta_{\mu_{\mathrm{d}}}) \to L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$, with

$$D(-\Delta_{\mu_{\mathrm{d}}}) := \{w \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}}) \cap H_0^1(\mathbb{R}^d, \mu_{\mathrm{d}}) : -\Delta_{\mu_{\mathrm{d}}} w \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}})\},$$

is accretive. That is, for all $w_1, w_2 \in D(-\Delta_{\mu_{\mathrm{d}}})$ and $\lambda > 0$,

$$\|w_1 - w_2\|_{L^1(\mathbb{R}^d, \mu_{\mathrm{d}})} \le \|(I - \lambda\Delta_{\mu_{\mathrm{d}}})w_1 - (I - \lambda\Delta_{\mu_{\mathrm{d}}})w_2\|_{L^1(\mathbb{R}^d, \mu_{\mathrm{d}})}.$$

As shown in Section 4.2 below, the accretiveness of A and $-\Delta_{\mu_{\mathrm{d}}}$ will play a crucial role in proving the existence of a unique weak solution to (36). To get a glimpse of the potential use of the accretiveness, let us look at the equation $(I + \lambda A)v = f$.

Definition 15 *For any $\lambda > 0$ and $f \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$, we say $v \in D(A)$ is a weak solution to $(I + \lambda A)v = f$ w.r.t. μ_{d} if*

$$\int_{\mathbb{R}^d} \left(v(y)\varphi(y) + \frac{\lambda}{2} \ln(1 + v(y))\Delta_{\mu_{\mathrm{d}}}\varphi(y) \right) d\mu_{\mathrm{d}} = \int_{\mathbb{R}^d} f(y)\varphi(y) d\mu_{\mathrm{d}}, \quad \forall \varphi \in C_c^2(\mathbb{R}^d). \tag{37}$$

Corollary 16 *For any $\lambda > 0$ and $f \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$, $(I + \lambda A)v = f$ has at most one weak solution $v \in D(A)$ w.r.t. μ_{d}. Such a solution, if it exists, satisfies $\int_{\mathbb{R}^d} v d\mu_{\mathrm{d}} = \int_{\mathbb{R}^d} f d\mu_{\mathrm{d}}$.*

Proof The uniqueness part follows from Lemma 14 (i). Let $v \in D(A)$ be the unique weak solution to $(I + \lambda A)v = f$ w.r.t. μ_{d}. Since $C_c^2(\mathbb{R}^d)$ is dense in $H_0^1(\mathbb{R}^d, \mu_{\mathrm{d}})$, (37) in fact holds for all $\varphi \in H_0^1(\mathbb{R}^d, \mu_{\mathrm{d}})$. Taking $\varphi \equiv 1 \in H_0^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ in (37) yields $\int_{\mathbb{R}^d} v d\mu_{\mathrm{d}} = \int_{\mathbb{R}^d} f d\mu_{\mathrm{d}}$. ■

4.2 Existence and Uniqueness

Thanks to the accretiveness of A in Lemma 14, we can invoke the Crandall-Liggett theorem to identify a sufficient condition, i.e., (38) below, for (31) to admit a weak solution w.r.t. $\mu_{\rm d}$. To properly state the result, we denote by $\overline{D(A)}$ the closure of $D(A)$ under the $L^1(\mathbb{R}^d,\mu_{\rm d})$-norm. Also, for any $\lambda>0$, we introduce the range of $I+\lambda A$ in $L^1(\mathbb{R}^d,\mu_{\rm d})$, defined by

$$R(I+\lambda A):=\{f\in L^1(\mathbb{R}^d,\mu_{\rm d}):(I+\lambda A)v=f \text{ has a weak solution } v\in D(A) \text{ w.r.t. } \mu_{\rm d}\}.$$

By Corollary 16, for each $f\in R(I+\lambda A)$, the corresponding weak solution $v\in D(A)$ is unique. Hence, we will constantly write $v=(I+\lambda A)^{-1}f$.

Proposition 17 *Let $\rho_0\in\mathcal{P}(\mathbb{R}^d)$ satisfy (34). If*

$$\overline{D(A)}\subseteq R(I+\lambda A)\quad \textit{for } \lambda>0 \textit{ small enough}, \tag{38}$$

there exists a weak solution $v:[0,\infty)\to D(A)$ to (31) w.r.t. $\mu_{\rm d}$ (cf. Definition 12). Moreover, v is continuous (i.e., $v\in C([0,\infty);L^1(\mathbb{R}^d,\mu_{\rm d}))$) and satisfies

$$v(t)=\lim_{n\to\infty}\left(I+\frac{t}{n}A\right)^{-n}v(0)\quad \textit{in}\ \ L^1(\mathbb{R}^d,\mu_{\rm d}),\quad \forall t\geq 0, \tag{39}$$

where the convergence is uniform in t on compact intervals.

Proof By definition, $\{v\in L^1(\mathbb{R}^d,\mu_{\rm d}):v\in C_c^\infty(\mathbb{R}^d),\ 0\leq v\leq\beta\}\subseteq D(A)$. As $C_c^\infty(\mathbb{R}^d)$ is dense in $L^1(\mathbb{R}^d,\mu_{\rm d})$, we have $\{v\in L^1(\mathbb{R}^d,\mu_{\rm d}):0\leq v\leq\beta\}\subseteq\overline{D(A)}$. By (34), this implies $\rho_0/\rho_{\rm d}\in\overline{D(A)}$. Due to the accretiveness of A in Lemma 14 (i), (Barbu, 2010, p.99, (3.5)) is satisfied with $\omega=0$; that is, A is ω-accretive with $\omega=0$, in the terminology of Barbu (2010). Under the ω-accretiveness of A, (38), and $\rho_0/\rho_{\rm d}\in\overline{D(A)}$, the Crandall-Liggett theorem (see e.g., Theorem 4.3 in Barbu (2010)) asserts that (36), or equivalently (31), has a unique *mild solution* $v\in C([0,\infty);L^1(\mathbb{R}^d,\mu_{\rm d}))$ (in the sense of Definition 4.3 in Barbu (2010))), which fulfills (39). Note that the operator $(I+\varepsilon A)^{-1}:\overline{D(A)}\to D(A)$ is well-defined for $\varepsilon>0$ small, thanks to (38) and Corollary 16, so that the right-hand side of (39) is well-defined. It remains to show that the mild solution v is a weak solution w.r.t. $\mu_{\rm d}$.

As stated below Theorem 4.3 in Barbu (2010), the mild solution $v\in C([0,\infty);L^1(\mathbb{R}^d,\mu_{\rm d}))$ satisfies the following: for $\varepsilon>0$ small enough, by taking $v^0:=\rho_0/\rho_{\rm d}$ and $v^i=(I+\varepsilon A)^{-1}v^{i-1}\in D(A)$ for all $i\in\mathbb{N}$, we have

$$v(t)=\lim_{\varepsilon\to 0}v_\varepsilon(t)\ \ \text{in } L^1(\mathbb{R}^d,\mu_{\rm d}),\ \ \text{uniformly in } t \text{ on compact intervals}, \tag{40}$$

where $v_\varepsilon:[0,\infty)\to L^1(\mathbb{R}^d,\mu_{\rm d})$ is defined by $v_\varepsilon(t):=v^{i-1}$ for $t\in[(i-1)\varepsilon,i\varepsilon)$, $i\in\mathbb{N}$. Observe from the construction of v_ε that

$$\frac{v_\varepsilon(t+\varepsilon,y)-v_\varepsilon(t,y)}{\varepsilon}-\frac{1}{2}\Delta_{\mu_{\rm d}}\ln(1+v_\varepsilon)(t+\varepsilon,y)=0.$$

Hence, for any $\varphi \in C_c^{1,2}((0,\infty)\times\mathbb{R}^d)$, by using the calculation in Remark 11 twice, we get

$$\begin{aligned}0 &= \int_{\mathbb{R}^d}\left(\frac{v_\varepsilon(t+\varepsilon,y)-v_\varepsilon(t,y)}{\varepsilon}\varphi(t+\varepsilon,y) - \frac{1}{2}\ln(1+v_\varepsilon)(t+\varepsilon,y)\Delta_{\mu_{\mathrm{d}}}\varphi(t+\varepsilon,y)\right)d\mu_{\mathrm{d}}\\ &= \int_{\mathbb{R}^d}\left(\frac{(v_\varepsilon\varphi)(t+\varepsilon,y)-(v_\varepsilon\varphi)(t,y)}{\varepsilon} - v_\varepsilon(t,y)\frac{\varphi(t+\varepsilon,y)-\varphi(t,y)}{\varepsilon}\right.\\ &\qquad\left. - \frac{1}{2}\ln(1+v_\varepsilon)(t+\varepsilon,y)\Delta_{\mu_{\mathrm{d}}}\varphi(t+\varepsilon,y)\right)d\mu_{\mathrm{d}}.\end{aligned}$$

Integrating the above equation over $t\in(0,\infty)$ yields

$$\begin{aligned}&\int_0^\infty\int_{\mathbb{R}^d}\left(v_\varepsilon(t,y)\frac{\varphi(t+\varepsilon,y)-\varphi(t,y)}{\varepsilon} + \frac{1}{2}\ln(1+v_\varepsilon)(t+\varepsilon,y)\Delta_{\mu_{\mathrm{d}}}\varphi(t+\varepsilon,y)\right)d\mu_{\mathrm{d}}dt\\ &\qquad = -\frac{1}{\varepsilon}\int_0^\varepsilon\int_{\mathbb{R}^d}(v_\varepsilon\varphi)(t,y)d\mu_{\mathrm{d}}dt.\end{aligned}$$

As $\varepsilon\to 0$, by $\varphi\in C_c^{1,2}((0,\infty)\times\mathbb{R}^d)$ and $0\le v_\varepsilon\le\beta\ \forall\varepsilon>0$ small (as $v_\varepsilon\in\overline{D(A)}$ by construction), the left-hand side tends to $\int_0^\infty\int_{\mathbb{R}^d}\left(v\varphi_t+\frac{1}{2}\ln(1+v)\Delta_{\mu_{\mathrm{d}}}\varphi\right)d\mu_{\mathrm{d}}dt$ and the right-hand side vanishes. That is, (33) is satisfied, i.e., v is a weak solution to (31) w.r.t. μ_{d}. ■

To establish the sufficient condition (38), we need to show that as $\lambda>0$ is small enough, there exists a weak solution $v\in D(A)$ to $(I+\lambda A)v=f$ w.r.t. μ_{d}, for any $f\in\overline{D(A)}$. To this end, we consider the change of variable $w:=\ln(1+v)$ and rewrite $(I+\lambda A)v=f$ as

$$-\frac{\lambda}{2}\Delta_{\mu_{\mathrm{d}}}w = -\frac{\lambda}{2}\Delta w - \frac{\lambda}{2}\nabla\ln\rho_{\mathrm{d}}\cdot\nabla w = f+1-e^w. \tag{41}$$

To find a weak solution to (41), we use the method of subsolutions and supersolutions, motivated by Evans (Evans, 1998, Section 9.3). Notably, the arguments therein needs to be extended from the integration w.r.t. Leb over a bounded domain to the integration w.r.t. μ_{d} over $\mathbb{R}^d$. The detailed derivation is relegated to Appendix B.2 and it leads to, in fact, a *stronger* version of (38).

Lemma 18 *$\overline{D(A)}\subseteq R(I+\lambda A)$ for all $\lambda>0$. That is, for any $\lambda>0$ and $f\in\overline{D(A)}$, there exists a weak solution $v\in D(A)$ to $(I+\lambda A)v=f$ w.r.t. μ_{d}.*

We now present the existence result for the nonlinear Fokker-Planck equation (25).

Corollary 19 *If $\rho_0\in\mathcal{P}(\mathbb{R}^d)$ satisfies (34), then the following hold.*

(i) There exists a weak solution $v\in C([0,\infty);L^1(\mathbb{R}^d,\mu_{\mathrm{d}}))$ to (31) w.r.t. μ_{d} (see Definition 12). Moreover, $0\le v\le\beta$, with $\beta>0$ as in (34), and satisfies (39).

(ii) There exists a weak solution $u:[0,\infty)\to\mathcal{P}(\mathbb{R}^d)$ to (25) (see Definition 9), given by

$$u(t):=\rho_{\mathrm{d}}v(t)\quad\forall t\ge 0,\quad \text{with } v:[0,\infty)\to L^1(\mathbb{R}^d,\mu_{\mathrm{d}}) \text{ taken from (i).} \tag{42}$$

Moreover, u is continuous in the sense that $u\in C([0,\infty);L^1(\mathbb{R}^d))$.

Proof By Lemma 18 and Proposition 17, we obtain (i) immediately. For each $t \geq 0$, thanks to Corollary 16, $v^n(t) := (I + \frac{t}{n}A)^{-n}v(0)$ is well-defined and satisfies $\int_{\mathbb{R}^d} v^n(t)d\mu_{\rm d} = \int_{\mathbb{R}^d} v(0)d\mu_{\rm d}$ for all $n \in \mathbb{N}$. Since $v^n(t) \to v(t)$ in $L^1(\mathbb{R}^d, \mu_{\rm d})$ (recall (39)) and $\rho_0 = u(0) = \rho_{\rm d}v(0)$, we get

$$\int_{\mathbb{R}^d} v(t)d\mu_{\rm d} = \int_{\mathbb{R}^d} v(0)d\mu_{\rm d} = \int_{\mathbb{R}^d} \rho_0 dy = 1. \tag{43}$$

Using (42) again then gives $\int_{\mathbb{R}^d} u(t)dy = \int_{\mathbb{R}^d} \rho_{\rm d}v(t)dy = \int_{\mathbb{R}^d} v(t)d\mu_{\rm d} = 1$. This, along with $u(t) = \rho_{\rm d}v(t) \geq 0$, shows that $u(t) \in \mathcal{P}(\mathbb{R}^d)$. Now, for any $\{t_n\}$ in $[0, \infty)$ with $t_n \to t \geq 0$,

$$\int_{\mathbb{R}^d} |u(t_n) - u(t)|dy = \int_{\mathbb{R}^d} |\rho_{\rm d}v(t_n) - \rho_{\rm d}v(t)|dy = \int_{\mathbb{R}^d} |v(t_n) - v(t)|d\mu_{\rm d} \to 0 \quad \text{as } n \to \infty, \tag{44}$$

where the convergence is due to $v \in C([0, \infty); L^1(\mathbb{R}^d, \mu_{\rm d}))$. Hence, $u \in C([0, \infty); L^1(\mathbb{R}^d))$. Finally, for any $\varphi \in C_c^{1,2}((0, \infty) \times \mathbb{R}^d)$, observe from (42) and integration by parts that

$$\begin{aligned}
\int_0^\infty \int_{\mathbb{R}^d} \left(\varphi_t + \widetilde{\mathcal{L}}_{u(t)}\varphi\right) u dy dt &= \int_0^\infty \int_{\mathbb{R}^d} \left(u\varphi_t - \frac{\rho_{\rm d} + u}{2} \operatorname{div}\left(\frac{u}{\rho_{\rm d} + u}\nabla\varphi\right) + \frac{1}{2}u\Delta\varphi\right) dy dt \\
&= \int_0^\infty \int_{\mathbb{R}^d} \left(v\varphi_t - \frac{1}{2}(1+v)\operatorname{div}\left(\frac{v}{1+v}\nabla\varphi\right) + \frac{1}{2}v\Delta\varphi\right) d\mu_{\rm d}dt \\
&= \int_0^\infty \int_{\mathbb{R}^d} \left(v\varphi_t - \frac{\nabla v}{2(1+v)} \cdot \nabla\varphi\right) d\mu_{\rm d}dt = \int_0^\infty \int_{\mathbb{R}^d} \left(v\varphi_t - \frac{1}{2}\nabla \ln(1+v) \cdot \nabla\varphi\right) d\mu_{\rm d}dt \\
&= \int_0^\infty \int_{\mathbb{R}^d} \left(v\varphi_t + \frac{1}{2}\ln(1+v)\Delta_{\mu_{\rm d}}\varphi\right) d\mu_{\rm d}dt = 0,
\end{aligned} \tag{45}$$

where the fifth equality stems from Remark 11 and the last equality holds as v satisfies (37). As such, u satisfies (28), and thus (27) by Remark 10, and is then a weak solution to (25). ■

To establish the uniqueness of solutions v to (31), we observe that (31) is an initial-value problem that involves the *modified* Laplacian $\Delta_{\mu_{\rm d}}$, which is applied to v through a nonlinear function (i.e., $\ln(1+v)$). In view of this, the uniqueness arguments in Brézis and Crandall (1979), for initial-value problems where the usual Laplacian Δ is applied in a nonlinear way, are expected to be useful. In Appendix B.3, a delicate plan is carried out to generalize the arguments in Brézis and Crandall (1979) to our setting where Δ and the Lebesgue measure are replaced by $\Delta_{\mu_{\rm d}}$ and $\mu_{\rm d}$. It ultimately shows the following.

Proposition 20 *If $v_1, v_2 \in C([0, \infty); L^1(\mathbb{R}^d, \mu_{\rm d})) \cap L^\infty_+([0, \infty) \times \mathbb{R}^d, \mu_{\rm d})$ are weak solutions to (31) w.r.t. $\mu_{\rm d}$, then*

$$\int_0^\infty \int_{\mathbb{R}^d} (v_1 - v_2)\varphi d\mu_{\rm d}dt = 0, \quad \forall \varphi \in C_c^{1,2}((0, \infty) \times \mathbb{R}^d). \tag{46}$$

Now, we are ready to present the main result of this section.

Theorem 21 *Let $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ satisfy (34). Then, $u : [0, \infty) \to \mathcal{P}(\mathbb{R}^d)$ in (42) is the unique weak solution to the Fokker-Planck equation (25) among the class of functions*

$$\mathcal{C} := \left\{\eta \in C([0, \infty); L^1(\mathbb{R}^d)) : \eta/\rho_{\rm d} \in L^\infty_+([0, \infty) \times \mathbb{R}^d)\right\}. \tag{47}$$

Specifically, if $\bar{u} : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ is a weak solution to (25) that belongs to $\mathcal{C}$, then $\bar{u}(t,\cdot) = u(t,\cdot)$ Leb-a.e. on $\mathbb{R}^d$ for all $t \geq 0$.

Proof We know from Corollary 19 that $u : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ given in (42) is a weak solution to (25) and it belongs to $\mathcal{C}$. Suppose that there exists another weak solution $\bar{u} : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ to (25) that belongs to $\mathcal{C}$. Then, $\bar{v} := \bar{u}/\rho_{\mathrm{d}}$ is nonnegative and bounded. By the same calculation in (44) (with u, v replaced by $\bar{u}, \bar{v}$), we see that $\bar{u} \in C(\mathbb{R}_+; L^1(\mathbb{R}^d))$ implies $\bar{v} \in C(\mathbb{R}_+; L^1(\mathbb{R}^d, \mu_{\mathrm{d}}))$. Moreover, by repeating the arguments in (45), but in a backward manner, we see that "$\bar{u}$ fulfills (28)" (as $\bar{u}$ is a weak solution to (25)) implies "$\bar{v}$ fulfills (33)", i.e., $\bar{v}$ is a weak solution to (31) w.r.t. μ_{d}. Now, in view of $\bar{u} = \rho_{\mathrm{d}}\bar{v}$ and (42), we conclude from Proposition 20 that

$$\int_0^\infty \int_{\mathbb{R}^d} (\bar{u} - u)\varphi dy dt = \int_0^\infty \int_{\mathbb{R}^d} (\bar{v} - v)\varphi d\mu_{\mathrm{d}} dt = 0, \quad \forall \varphi \in C_c^{1,2}((0,\infty) \times \mathbb{R}^d),$$

where v is given as in Corollary 19 (i). This implies $\bar{u}(t,\cdot) = u(t,\cdot)$ Leb-a.e. on $\mathbb{R}^d$, for Leb-a.e. $t \geq 0$. In the following, we show that this equality can be strengthened to hold for *all* $t \geq 0$. Recall $u \in C([0,\infty); L^1(\mathbb{R}^d))$ from Corollary 19. For any $t \geq 0$, this implies that

$$\lim_{s \to t} \int_{\mathbb{R}^d} \psi(y) u(s,y) dy = \int_{\mathbb{R}^d} \psi(y) u(t,y) dy \quad \forall \psi \in C_b(\mathbb{R}^d). \tag{48}$$

Similarly, as $\bar{u} \in C([0,\infty); L^1(\mathbb{R}^d))$, (48) also holds with u replaced by $\bar{u}$. Now, take $\{t_n\}$ in $[0,\infty)$ such that $t_n \to t$ and $\bar{u}(t_n,\cdot) = u(t_n,\cdot)$ Leb-a.e. on $\mathbb{R}^d$ for all $n \in \mathbb{N}$. Then, for any $\psi \in C_b(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} \psi(y)\bar{u}(t,y)dy = \lim_{n\to\infty} \int_{\mathbb{R}^d} \psi(y)\bar{u}(t_n,y)dy = \lim_{n\to\infty} \int_{\mathbb{R}^d} \psi(y)u(t_n,y)dy = \int_{\mathbb{R}^d} \psi(y)u(t,y)dy,$$

which readily shows $\bar{u}(t,\cdot) = u(t,\cdot)$ Leb-a.e. on $\mathbb{R}^d$. ■

5. Solutions to Density-Dependent ODE (2)

In this section, we focus on constructing a solution Y to ODE (2). The unique weak solution $u : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ to the Fokker-Planck equation (25) (Theorem 21) already gives important clues: as mentioned above (25), if (2) admits a solution Y, the density flow $\{\rho^{Y_t}\}_{t\geq 0}$ should (heuristically) coincide with $\{u(t)\}_{t\geq 0}$. The question now is how we can actually construct such a solution Y from the knowledge of $\{u(t)\}_{t\geq 0}$.

To answer this, the first challenge is what constitutes a "solution" to (2). As opposed to standard ODEs, (2) involves *two* different levels of randomness at time 0 and they trickle down as time passes by (through the otherwise deterministic dynamics), leaving Y_t a random variable for all $t \geq 0$. To see this, we follow the idea "$\rho^{Y_t} = u(t)$ for all $t \geq 0$", where $u : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ is the unique weak solution to (25), and rewrite (2) as

$$dY_t = -\frac{1}{2}\left(\frac{\nabla u(t,Y_t)}{u(t,Y_t)} - \frac{\nabla \rho_{\mathrm{d}}(Y_t) + \nabla u(t,Y_t)}{\rho_{\mathrm{d}}(Y_t) + u(t,Y_t)}\right) dt, \quad \rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \tag{49}$$

Clearly, there is the randomness of the initial point $y \in \mathbb{R}^d$, stated explicitly through $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$. Once an initial point $y \in \mathbb{R}^d$ is sampled, because the coefficient of the ODE in (49) is not necessarily Lipschitz, there can be multiple continuous paths $\omega : [0, \infty) \to \mathbb{R}^d$ with $\omega(0) = y$ such that $t \mapsto Y_t := \omega(t)$ solves the ODE (with the initial condition $Y_0 = y$). In other words, there remains the randomness of which continuous path to pick among those who solve the ODE in (49) (with $Y_0 = y$ fixed). In fact, these two levels of randomness can be jointly expressed by a probability measure $\mathbb{P}$ defined on the space of continuous paths

$$(\Omega, \mathcal{F}) := (C([0, \infty); \mathbb{R}^d), \mathcal{B}(C([0, \infty); \mathbb{R}^d))),$$

where $\mathcal{B}(C([0, \infty); \mathbb{R}^d))$ denotes the Borel σ-algebra of $C([0, \infty); \mathbb{R}^d)$.

Definition 22 *A process $Y : [0, \infty) \times \Omega \to \mathbb{R}^d$ is said to be a solution to (2) if*

$$Y_t(\omega) := \omega(t) \quad \forall (t, \omega) \in [0, \infty) \times \Omega, \tag{50}$$

and there exists a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ under which

(i) the density function $\eta_t \in \mathcal{P}(\mathbb{R}^d)$ of $Y_t : \Omega \to \mathbb{R}^d$ exists and is weakly differentiable for all $t \geq 0$, and $\eta_0 = \rho_0$ Leb-a.e.;

(ii) the collection of paths

$$\left\{ \omega \in \Omega : \omega(t) = \omega(0) - \frac{1}{2} \int_0^t \left(\frac{\nabla \eta_s(\omega(s))}{\eta_s(\omega(s))} - \frac{\nabla \rho_{\mathrm{d}}(\omega(s)) + \nabla \eta_s(\omega(s))}{\rho_{\mathrm{d}}(\omega(s)) + \eta_s(\omega(s))} \right) ds, \ t \geq 0 \right\}$$

has probability one.

Remark 23 *In Definition 22, $\mathbb{P}$ serves to sample continuous paths $\omega : [0, \infty) \to \mathbb{R}^d$ from the set in (ii), i.e., among those who fulfill ODE (2), in a way that the density of $\omega(0)$ coincides with $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$.*

To show that a solution to (2) exists (in the sense of Definition 22), we will resort to Trevisan's superposition principle (Theorem 2.5 in Trevisan (2016)). To this end, appropriate integrability needs to be checked first.

Lemma 24 *Let $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ satisfy (34). Then, the weak solution $v \in C([0, \infty); L^1(\mathbb{R}^d, \mu_{\mathrm{d}}))$ to (31) w.r.t. μ_{d}, specified as in Corollary 19 (i), satisfies*

$$\int_0^\infty \|\nabla v\|^2_{L^2(\mathbb{R}^d, \mu_{\mathrm{d}})} dt \leq (1 + \beta)\beta^2. \tag{51}$$

Proof Recall that v is constructed using the Crandall-Liggett theorem (see Proposition 17) and thus satisfies (40), which involves $v_\varepsilon : [0, \infty) \to L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ defined by $v_\varepsilon(t) := (I + \varepsilon A)^{-(i-1)} v(0)$ for $t \in [(i-1)\varepsilon, i\varepsilon)$ and $i \in \mathbb{N}$. By writing $z_i = v_\varepsilon(i\varepsilon)$ for $i \in \mathbb{N} \cup \{0\}$, we have $(I + \varepsilon A) z_i = z_{i-1}$ for $i \in \mathbb{N}$. Let us multiply this equation by z_i and integrate it w.r.t. μ_{d}. By Remark 11, we get

$$\int_{\mathbb{R}^d} z_i^2 d\mu_{\mathrm{d}} + \frac{\varepsilon}{2} \int_{\mathbb{R}^d} \nabla \ln(1 + z_i) \cdot \nabla z_i d\mu_{\mathrm{d}} = \int_{\mathbb{R}^d} z_{i-1} z_i d\mu_{\mathrm{d}}, \quad \forall i \in \mathbb{N}.$$

Thanks to $0 \le z_i \le \beta$ (as $z_i \in D(A)$) and Young's inequality, the above yields

$$\|z_i\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} + \frac{\varepsilon}{2(1+\beta)}\|\nabla z_i\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} \le \frac{1}{2}\|z_{i-1}\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} + \frac{1}{2}\|z_i\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}, \quad \forall i \in \mathbb{N}.$$

That is to say,

$$\varepsilon\|\nabla z_i\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} \le (1+\beta)\left(\|z_{i-1}\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} - \|z_i\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}\right), \quad \forall i \in \mathbb{N}.$$

By summing up the above relation over all $i \in \mathbb{N}$, we have

$$\int_\varepsilon^\infty \|\nabla v_\varepsilon(t)\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}dt = \sum_{i=1}^\infty \varepsilon\|\nabla z_i\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} \le (1+\beta)\|z_0\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} \le (1+\beta)\beta^2,$$

where the last inequality follows from $0 \le z_0 \le \beta$, as $z_0 = v_\varepsilon(0) = v(0) = \rho_0/\rho_{\rm d} \in \overline{D(A)}$; see the proof of Proposition 17. As the above estimate holds for all $\varepsilon > 0$ small enough, we conclude that for any fixed $\bar\varepsilon > 0$ small,

$$\int_{\bar\varepsilon}^\infty \|\nabla v_\varepsilon(t)\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}dt \le (1+\beta)\beta^2, \quad \forall \varepsilon \in (0,\bar\varepsilon]. \tag{52}$$

This shows that for any $j \in \{1,\dots,d\}$, $\{\partial_{y_j} v_\varepsilon\}_{\varepsilon\in(0,\bar\varepsilon]}$ is bounded in $L^2((\bar\varepsilon,\infty)\times\mathbb{R}^d, {\rm Leb}\otimes\mu_{\rm d})$; hence, $\partial_{y_j} v_\varepsilon$ converges weakly along a subsequence (without relabeling) to some $\eta_j \in L^2((\bar\varepsilon,\infty)\times\mathbb{R}^d, {\rm Leb}\otimes\mu_{\rm d})$. We claim that for Leb-a.e. $t \ge \bar\varepsilon$, the weak derivative $\partial_{y_j} v(t,\cdot)$ exists and equals $\eta_j(t,\cdot) \in L^2(\mathbb{R}^d,\mu_{\rm d})$. For any $\psi \in L^\infty((\bar\varepsilon,\infty))$ compactly supported and $\varphi \in C_c^\infty(\mathbb{R}^d)$, we deduce from (40), the bounded convergence theorem, and integration by parts that

$$\begin{aligned}\int_{\bar\varepsilon}^\infty \psi \int_{\mathbb{R}^d} v\varphi_{y_j} dydt &= \lim_{\varepsilon\to 0}\int_{\bar\varepsilon}^\infty \psi \int_{\mathbb{R}^d} v_\varepsilon\varphi_{y_j} dydt = -\lim_{\varepsilon\to 0}\int_{\bar\varepsilon}^\infty \psi \int_{\mathbb{R}^d} \partial_{y_j} v_\varepsilon \varphi dydt \\ &= -\lim_{\varepsilon\to 0}\int_{\bar\varepsilon}^\infty \int_{\mathbb{R}^d} \partial_{y_j} v_\varepsilon \frac{\psi\varphi}{\rho_{\rm d}} d\mu_{\rm d}dt = -\int_{\bar\varepsilon}^\infty \int_{\mathbb{R}^d} \eta_j \frac{\psi\varphi}{\rho_{\rm d}} d\mu_{\rm d}dt = -\int_{\bar\varepsilon}^\infty \psi\int_{\mathbb{R}^d}\eta_j\varphi dydt.\end{aligned}$$

Note that the fourth equality above follows from $\psi\varphi/\rho_{\rm d} \in L^2((\bar\varepsilon,\infty)\times\mathbb{R}^d, {\rm Leb}\otimes\mu_{\rm d})$ and $\partial_{y_j} v_\varepsilon \to \eta_j$ weakly in $L^2((\bar\varepsilon,\infty)\times\mathbb{R}^d, {\rm Leb}\otimes\mu_{\rm d})$. By taking $\psi \in L^\infty((\bar\varepsilon,\infty))$ in the above equation as $\psi(t) = {\rm sgn}\big(\int_{\mathbb{R}^d}(v(t)\varphi_{y_j} + \eta_j(t)\varphi)dy\big)1_E(t)$ with $E \subset (\bar\varepsilon,\infty)$, we see that

$$\int_{\mathbb{R}^d} v(t)\varphi_{y_j} dy = -\int_{\mathbb{R}^d}\eta_j(t)\varphi dy \quad \text{for Leb-a.e. } t \ge \bar\varepsilon.$$

As $\varphi \in C_c^\infty(\mathbb{R}^d)$ is arbitrary, this readily shows that for Leb-a.e. $t \ge \bar\varepsilon$, the weak derivative $\partial_{y_j} v(t,\cdot)$ is $\eta_j(t,\cdot) \in L^2(\mathbb{R}^d,\mu_{\rm d})$, i.e., the claim is proved. As a result,

$$\begin{aligned}\int_{\bar\varepsilon}^\infty \|\nabla v(t)\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}dt = \int_{\bar\varepsilon}^\infty \|\eta(t)\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}dt &\le \int_{\bar\varepsilon}^\infty \liminf_{\varepsilon\to 0}\|\nabla v_\varepsilon\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}dt \\ &\le \liminf_{\varepsilon\to 0}\int_{\bar\varepsilon}^\infty \|\nabla v_\varepsilon\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}dt \le (1+\beta)\beta^2,\end{aligned}$$

where the first inequality stems from $\partial_{y_j} v_\varepsilon(t,\cdot) \to \eta_j(t,\cdot)$ weakly in $L^2(\mathbb{R}^d, \mu_{\rm d})$ for Leb-a.e. $t \geq \bar{\varepsilon}$ and all $j \in \{1,2,...,d\}$ and the norm $\|\cdot\|_{L^2(\mathbb{R}^d,\mu_{\rm d})}$ being weakly lower semicontinuous, the second inequality is due to Fatou's lemma, and the last inequality holds because of (52). As $\bar{\varepsilon} > 0$ can be chosen to be arbitrarily small, the desired result follows. ■

A solution Y to (2) can now be constructed.

Proposition 25 *Let $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ satisfy (34). Then, there exists a solution Y to (2).*

Proof Let $u : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ be specified as in Theorem 21. Consider the collection of probabilities $\nu = \{\nu_t\}_{t\geq 0}$ on $\mathbb{R}^d$, with $d\nu_t(y) := u(t,y)dy$ for all $t \geq 0$. Recall that u satisfies (48), which directly implies that ν is narrowly continuous. Also, as u is a weak solution to the Fokker-Planck equation (25), ν solves (2.2) in Trevisan (2016) in the weak sense. Hence, to apply the "superposition principle" (Theorem 2.5 in Trevisan (2016)), it remains to verify (2.3) in Trevisan (2016), which in our case takes the form

$$\int_0^T \int_{\mathbb{R}^d} \left| \frac{\nabla u(t,y)}{u(t,y)} - \frac{\nabla \rho_{\rm d}(y) + \nabla u(t,y)}{\rho_{\rm d}(y) + u(t,y)} \right| u(t,y) dy dt < \infty, \quad \forall T > 0. \tag{53}$$

Indeed, "a_t" and "b_t" in (2.3) of Trevisan (2016) correspond to the diffusion and drift coefficients in (49), which are zero and $-\frac{1}{2}\big(\frac{\nabla u(t,y)}{u(t,y)} - \frac{\nabla \rho_{\rm d}(y)+\nabla u(t,y)}{\rho_{\rm d}(y)+u(t,y)}\big)$ respectively. Now, on strength of $u = \rho_{\rm d} v$ and $v \geq 0$ (see (42)), a direct calculation shows

$$\int_{\mathbb{R}^d} \left| \frac{\nabla u}{u} - \frac{\nabla \rho_{\rm d} + \nabla u}{\rho_{\rm d} + u} \right| u dy = \int_{\mathbb{R}^d} \left| \frac{\nabla v}{1+v} \right| d\mu_{\rm d} \leq \int_{\mathbb{R}^d} |\nabla v| \, d\mu_{\rm d} \leq \|\nabla v\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})},$$

where the last inequality follows from Hölder's inequality. Consequently,

$$\int_0^T \int_{\mathbb{R}^d} \left| \frac{\nabla u(t,y)}{u(t,y)} - \frac{\nabla \rho_{\rm d}(y) + \nabla u(t,y)}{\rho_{\rm d}(y) + u(t,y)} \right| u(t,y) dy dt \leq \int_0^T \|\nabla v(t)\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} dt < \infty, \ \forall T > 0,$$

where the finiteness follows from Lemma 24. That is, (53) is satisfied.

Now, consider the coordinate mapping process Y in (50). Thanks to Theorem 2.5 of Trevisan (2016), there exists a probability $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that (i) $\mathbb{P} \circ (Y_t)^{-1} = \nu_t$ for all $t \geq 0$, and (ii) for any $f \in C^{1,2}((0,\infty) \times \mathbb{R}^d)$, the process

$$t \mapsto f(t,\omega(t)) - f(0,\omega(0)) - \int_0^t \left(\partial_s f - \frac{1}{2} \left(\frac{\nabla u}{u} - \frac{\nabla \rho_{\rm d} + \nabla u}{\rho_{\rm d} + u} \right) \cdot \nabla f \right)(s,\omega(s)) ds \tag{54}$$

is a local martingale under $\mathbb{P}$. Note that (i) readily implies $\rho^{Y_t} = u(t) \in \mathcal{P}(\mathbb{R}^d)$ under $\mathbb{P}$ for all $t \geq 0$, i.e., Definition 22 (i) is satisfied. For each $i \in \{1,2,...,d\}$, by taking $f(t,y) = y_i$ for $y = (y_1, y_2, ..., y_d) \in \mathbb{R}^d$ in (54), we see that

$$t \mapsto M_t^{(i)}(\omega) := (\omega(t))_i - (\omega(0))_i + \frac{1}{2} \int_0^t \left(\frac{\partial_{y_i} u}{u} - \frac{\partial_{y_i} \rho_{\rm d} + \partial_{y_i} u}{\rho_{\rm d} + u} \right)(s,\omega(s)) ds$$

is a local martingale under $\mathbb{P}$. By the same arguments on p. 315 of Karatzas and Shreve (1991), we find that the quadratic variation of $M^{(i)}$ is constantly zero, i.e., $\langle M^{(i)} \rangle_t = 0$ for

all $t \geq 0$, thanks to the lack of a diffusion coefficient in (49). As $M^{(i)}$ is a local martingale with zero quadratic variation under $\mathbb{P}$, $M_t^{(i)} = 0$ for all $t \geq 0$ $\mathbb{P}$-a.s. That is,

$$(\omega(t))_i = (\omega(0))_i - \frac{1}{2}\int_0^t \left(\frac{\partial_{y_i} u}{u} - \frac{\partial_{y_i}\rho_{\mathrm{d}} + \partial_{y_i} u}{\rho_{\mathrm{d}} + u}\right)(s, \omega(s))ds, \quad \text{for } \mathbb{P}\text{-a.e. } \omega \in \Omega.$$

Because this holds for all $i \in \{1, 2, ..., d\}$, we conclude that

$$\omega(t) = \omega(0) - \frac{1}{2}\int_0^t \left(\frac{\nabla u}{u} - \frac{\nabla\rho_{\mathrm{d}} + \nabla u}{\rho_{\mathrm{d}} + u}\right)(s, \omega(s))ds, \quad \text{for } \mathbb{P}\text{-a.e. } \omega \in \Omega,$$

which shows that Definition 22 (ii) is also satisfied. Hence, Y is a solution to (2). ■

Remark 26 *In the proof above, the probability $\mathbb{P}$ on $(\Omega, \mathcal{F})$ can be decomposed into the two levels of randomness discussed above Definition 22. Let $\{Q_y\}_{y\in\mathbb{R}^d}$ be a regular conditional probability for $\mathcal{F}$ given $Y_0 = \omega(0)$; that is, $\{Q_y\}_{y\in\mathbb{R}^d}$ is a set of probability measures on $(\Omega, \mathcal{F})$ such that for any $A \in \mathcal{F}$, $y \mapsto Q_y(A)$ is Borel and $Q_y(A) = \mathbb{P}(A \mid Y_0 = y)$ for $\rho_0(x)dx$-a.e. $y \in \mathbb{R}^d$. Such $\{Q_y\}_{y\in\mathbb{R}^d}$ indeed exists in the path space $(\Omega, \mathcal{F})$; see e.g., Theorem 5.3.18 in Karatzas and Shreve (1991). Thus, we can express $\mathbb{P}$ as*

$$\mathbb{P}(A) = \int_{\mathbb{R}^d} \rho_0(y) Q_y(A) dy, \quad \forall A \in \mathcal{F}. \tag{55}$$

This explicitly shows that the randomness of the initial point $y \in \mathbb{R}^d$ is governed by $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$. Once an initial point $y \in \mathbb{R}^d$ is sampled, the randomness of picking a solution to the ODE in (49) (with $Y_0 = y$ fixed) is dictated by the probability measure Q_y.

Remark 27 *In view of Definition 22 and Proposition 25, the way we construct a solution to (2)—using a random selection of deterministic paths—is in line with L.C. Young's theory of generalized curves and close to the formulation in Ambrosio (2004). By taking $\rho_0 \equiv 1$ in (55) (i.e., replacing the density $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ by the Lebesgue measure), one recovers the form of (5.3) in Ambrosio (2004).*

On the strength of the uniqueness result for the nonlinear Fokker-Planck equation (25) (see Theorem 21), the uniqueness of solutions to (2) can be established accordingly. Recall the class of functions $\mathcal{C}$ in (47).

Proposition 28 *Let Y be a solution to (2) such that $\eta(t, y) := \rho^{Y_t}(y)$ satisfies*

$$\eta \in \mathcal{C} \quad \textit{and} \quad \nabla\eta \in L^1_{\mathrm{loc}}([0, \infty) \times \mathbb{R}^d). \tag{56}$$

Then, $\eta(t, \cdot) = u(t, \cdot)$ Leb-a.e. on $\mathbb{R}^d$ for all $t \geq 0$, where $u : [0, \infty) \to \mathcal{P}(\mathbb{R}^d)$ is the unique weak solution to the Fokker-Planck equation (25) specified as in (42).

Proof Let $\mathbb{P}$ be the probability measure on $(\Omega, \mathcal{F})$ associated with Y (see Definition 22) and $\mathbb{E}$ be the expectation under $\mathbb{P}$. For any $T > 0$ and compact subset K of $\mathbb{R}^d$, by Fubini-Tonelli's theorem,

$$\begin{aligned}\mathbb{E}\left[\int_0^T \left|\frac{\nabla\eta}{\eta} - \frac{\nabla\rho_{\mathrm{d}} + \nabla\eta}{\rho_{\mathrm{d}} + \eta}\right|(t, Y_t)1_K(Y_t)dt\right] &\le \int_0^T \int_K \left|\frac{\nabla\eta}{\eta} - \frac{\nabla\rho_{\mathrm{d}} + \nabla\eta}{\rho_{\mathrm{d}} + \eta}\right|\eta dydt \\ &\le \int_0^T \int_K |\nabla\eta| + \left|\frac{\eta\nabla\rho_{\mathrm{d}}}{\rho_{\mathrm{d}}}\right| dydt < \infty, \qquad (57)\end{aligned}$$

where the finiteness follows from (19) and Assumption 6.1. Now, for any $\varphi \in C_c^{1,2}((0,\infty) \times \mathbb{R}^d)$, whenever $T > 0$ is large enough, we have

$$0 = \varphi(T, Y_T) - \varphi(0, Y_0) = \int_0^T \left(\partial_t\varphi - \frac{1}{2}\nabla\varphi \cdot \left(\frac{\nabla\eta}{\eta} - \frac{\nabla\rho_{\mathrm{d}} + \nabla\eta}{\rho_{\mathrm{d}} + \eta}\right)\right)(t, Y_t)dt,$$

where the first equality is due to $\varphi(T, \cdot) = \varphi(0, \cdot) = 0$ and the second equality follows from (2). Taking expectation on both sides and using Fubini's theorem (applicable due to (57)) yield

$$0 = \int_0^T \int_{\mathbb{R}^d} \left(\partial_t\varphi - \frac{1}{2}\nabla\varphi \cdot \left(\frac{\nabla\eta}{\eta} - \frac{\nabla\rho_{\mathrm{d}} + \nabla\eta}{\rho_{\mathrm{d}} + \eta}\right)\right)\eta dydt.$$

In view of (26), taking $T \to \infty$ in the above shows that η fulfills (27), i.e., η is a weak solution to the nonlinear Fokker-Planck equation (25). Thanks to $\eta \in \mathcal{C}$, we conclude from Theorem 21 that $\eta(t, \cdot) = u(t, \cdot)$ Leb-a.e. on $\mathbb{R}^d$ for all $t \ge 0$. ■

Remark 29 *The uniqueness in Proposition 28 is weaker than the standard uniqueness of weak solutions to an SDE. The latter requires the law (or, the finite-dimensional distribution) of every solution Y to an SDE to be identical, while the former requires only the marginal distribution of Y_t to be the same for all $t \ge 0$.*

The weaker notion of uniqueness in Proposition 28 well serves our purpose. Since we intend to uncover the data distribution $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$ through gradient descent in $\mathcal{P}(\mathbb{R}^d)$ (governed by a solution Y to (2)), we are curious about whether the marginal distribution ρ^{Y_t} converges to ρ_{d} as $t \to \infty$. An affirmative answer is provided in the next section.

6. Convergence to Data Distribution ρ_{d}

This section is devoted to the main convergence theorem of this paper: for an appropriate solution Y to (2), $\rho^{Y_t} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$ as $t \to \infty$. As the first step towards the final result, we show that the gradient descent feature encoded in (2) indeed works out: the distance between ρ^{Y_t} and ρ_{d}, i.e., $J(\rho^{Y_t}) = \mathrm{JSD}(\rho^{Y_t}, \rho_{\mathrm{d}})$, decreases over time.

Proposition 30 *Let Y be a solution to (2) such that $\eta(t, y) := \rho^{Y_t}(y)$ satisfies (19). Then, for any $0 \le t_1 < t_2$,*

$$J(\rho^{Y_{t_2}}) - J(\rho^{Y_{t_1}}) \le -\int_{t_1}^{t_2} \int_{\mathbb{R}^d} \left|\nabla\frac{\delta J}{\delta\rho}(\rho^{Y_t}, y)\right|^2 \rho^{Y_t}(y)dydt \le 0. \qquad (58)$$

Proof As Y is a solution to (2) that fulfills (19), by Proposition 28, we have $\rho^{Y_t} = u(t) \in \mathcal{P}(\mathbb{R}^d)$ for all $t \geq 0$, where $u : [0, \infty) \to \mathcal{P}(\mathbb{R}^d)$ is the unique weak solution to the Fokker-Planck equation (25) specified as in (42). Recall $u \in C([0, \infty); L^1(\mathbb{R}^d))$ from Corollary 19. Also, by Remark 2, $J(\cdot) = \mathrm{JSD}(\cdot, \rho_{\mathrm{d}})$ is continuous under the $L^1(\mathbb{R}^d)$-norm. Hence, we may assume without loss of generality that $t_1, t_2 \in \mathbb{Q}$.

As $J(\cdot) = \mathrm{JSD}(\cdot, \rho_{\mathrm{d}})$, by (8) and the fact $u = \rho_{\mathrm{d}} v$ from (42), we have

$$J(\rho^{Y_t}) = J(u(t)) = \ln 2 + \frac{1}{2} \int_{\mathbb{R}^d} v(t) \ln v(t) - (1 + v(t)) \ln(1 + v(t)) d\mu_{\mathrm{d}}, \quad \forall t \geq 0,$$

On the other hand, by Lemma 6,

$$\nabla \frac{\delta J}{\delta \rho}(\rho^{Y_t}, y) = \frac{1}{2} \left(\frac{\nabla u}{u} - \frac{\nabla(\rho_{\mathrm{d}} + u)}{\rho_{\mathrm{d}} + u} \right) = \frac{1}{2} \left(\frac{\nabla v}{v} - \frac{\nabla v}{1 + v} \right) = \frac{1}{2} \frac{\nabla v}{v(1 + v)}.$$

Thus, (58) can be rewritten as

$$\left[\int_{\mathbb{R}^d} v(t) \ln v(t) - (1 + v(t)) \ln(1 + v(t)) d\mu_{\mathrm{d}} \right]_{t=t_1}^{t=t_2} \leq -\frac{1}{2} \int_{t_1}^{t_2} \int_{\mathbb{R}^d} \frac{|\nabla v|^2}{v(1 + v)^2} d\mu_{\mathrm{d}} dt. \tag{59}$$

Now, recall that v is constructed using the Crandall-Liggett theorem (see Proposition 17) and thus satisfies (40), which involves $v_\varepsilon : [0, \infty) \to L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ defined by $v_\varepsilon(t) := (I + \varepsilon A)^{-(i-1)} v(0)$ for $t \in [(i-1)\varepsilon, i\varepsilon)$ and $i \in \mathbb{N}$. By writing $z_i = v_\varepsilon(i\varepsilon)$ for $i \in \mathbb{N} \cup \{0\}$, we have $(I + \varepsilon A) z_i = z_{i-1}$ for $i \in \mathbb{N}$. For any $\gamma \in (0, 1)$, let us multiply the equation $z_{i-1} - z_i = \varepsilon A z_i$ by $\ln(\frac{\gamma + z_i}{1 + z_i})$ and integrate it w.r.t. μ_{d}. This yields

$$\begin{aligned} \int_{\mathbb{R}^d} (z_{i-1} - z_i) \ln \left(\frac{\gamma + z_i}{1 + z_i} \right) d\mu_{\mathrm{d}} &= \frac{\varepsilon}{2} \int_{\mathbb{R}^d} \nabla \ln(1 + z_i) \cdot \nabla \ln \left(\frac{\gamma + z_i}{1 + z_i} \right) d\mu_{\mathrm{d}} \\ &= \frac{\varepsilon}{2} \int_{\mathbb{R}^d} \frac{(1 - \gamma) |\nabla z_i|^2}{(\gamma + z_i)(1 + z_i)^2} d\mu_{\mathrm{d}}. \end{aligned} \tag{60}$$

Note that the integrand on the left hand side can be written as $F_{i-1} - F_i + R_i$, with

$$\begin{aligned} &F_i := (\gamma + z_i) \ln(\gamma + z_i) - (1 + z_i) \ln(1 + z_i), \\ &R_i := f(z_{i-1}, z_i) \quad \text{where} \quad f(a, b) := (\gamma + a) \ln \left(\frac{\gamma + b}{\gamma + a} \right) + (1 + a) \ln \left(\frac{1 + a}{1 + b} \right) \end{aligned} \tag{61}$$

Observe that $f(a, b) \leq 0$ for all $a, b \geq 0$. Indeed, by direct calculations, $f_a(a, b) < 0$ and $f_b(a, b) > 0$ for $a > b$, and $f_a(a, b) > 0$ and $f_b(a, b) < 0$ for $a < b$; this readily shows that f is maximized as $a = b$, giving the value $f(a, a) = 0$. With $R_i = f(z_{i-1}, z_i) \leq 0$, (60) implies

$$\int_{\mathbb{R}^d} (F_{i-1} - F_i) d\mu_{\mathrm{d}} \geq \frac{\varepsilon}{2} \int_{\mathbb{R}^d} \frac{(1 - \gamma) |\nabla z_i|^2}{(\gamma + z_i)(1 + z_i)^2} d\mu_{\mathrm{d}}.$$

Fix $\delta > 0$. As $t_1, t_2 \in \mathbb{Q}$, there exists $0 < \varepsilon < \delta$ small enough such that t_1, t_2 are both integer multiples of ε. Set $i_1 := t_1/\varepsilon$ to $i_2 := t_2/\varepsilon$. Summing up the previous inequality from $i = i_1 + 1$ to $i = i_2$ leads to

$$\int_{\mathbb{R}_d} F_{i_1} d\mu_{\mathrm{d}} - \int_{\mathbb{R}_d} F_{i_2} d\mu_{\mathrm{d}} \geq \frac{\varepsilon}{2} \sum_{i=i_1+1}^{i_2} \int_{\mathbb{R}^d} \frac{(1 - \gamma) |\nabla z_i|^2}{(\gamma + z_i)(1 + z_i)^2} d\mu_{\mathrm{d}}.$$

By (61), $z_i = v_\varepsilon(i\varepsilon)$, and the definition of v_ε, the above can be stated in terms of v_ε as

$$\begin{aligned}&\left[\int_{\mathbb{R}^d}(\gamma+v_\varepsilon(t))\ln(\gamma+v_\varepsilon(t))-(1+v_\varepsilon(t))\ln(1+v_\varepsilon(t))d\mu_{\mathrm{d}}\right]_{t=t_1}^{t=t_2}\\ \leq&-\frac{1}{2}\int_{t_1+\varepsilon}^{t_2+\varepsilon}\int_{\mathbb{R}^d}\frac{(1-\gamma)|\nabla v_\varepsilon(t)|^2}{(\gamma+v_\varepsilon(t))(1+v_\varepsilon(t))^2}d\mu_{\mathrm{d}}dt\leq-\frac{1}{2}\int_{t_1+\delta}^{t_2}\int_{\mathbb{R}^d}\frac{(1-\gamma)|\nabla v_\varepsilon(t)|^2}{(\gamma+v_\varepsilon(t))(1+v_\varepsilon(t))^2}d\mu_{\mathrm{d}}dt,\end{aligned}\tag{62}$$

where the last inequality holds as the integrand is nonnegative (recall $\gamma\in(0,1)$ and $v_\varepsilon\geq 0$). By the proof of Lemma 24 (under (52) particularly), we know that ∇v_ε converges weakly to ∇v, along a subsequence, in $L^2([t_1+\delta,t_2]\times\mathbb{R}^d,\mathrm{Leb}\otimes\mu_{\mathrm{d}})$. This, along with (40), implies

$$\frac{\nabla v_\varepsilon}{\sqrt{\gamma+v_\varepsilon}(1+v_\varepsilon)}\to\frac{\nabla v}{\sqrt{\gamma+v}(1+v)}\quad\text{weakly in } L^2([t_1+\delta,t_2]\times\mathbb{R}^d,\mathrm{Leb}\otimes\mu_{\mathrm{d}}).$$

Since the $L^2([t_1+\delta,t_2]\times\mathbb{R}^d,\mathrm{Leb}\otimes\mu_{\mathrm{d}})$-norm is weakly lower semicontinous, we have

$$\liminf_{\varepsilon\to0}\int_{t_1+\delta}^{t_2}\int\frac{|\nabla v_\varepsilon(t)|^2}{(\gamma+v_\varepsilon(t))(1+v_\varepsilon(t))^2}d\mu_{\mathrm{d}}dt\geq\int_{t_1+\delta}^{t_2}\int\frac{|\nabla v(t)|^2}{(\gamma+v(t))(1+v(t))^2}d\mu_{\mathrm{d}}dt.$$

Hence, as $\varepsilon\to0$ in (62), by (40), the boundedness of v_ε, and the above inequality, we obtain

$$\begin{aligned}&\left[\int_{\mathbb{R}^d}(\gamma+v(t))\ln(\gamma+v(t))-(1+v(t))\ln(1+v(t))d\mu_{\mathrm{d}}\right]_{t=t_1}^{t=t_2}\\ &\qquad\leq-\frac{1}{2}\int_{t_1+\delta}^{t_2}\int\frac{(1-\gamma)|\nabla v(t)|^2}{(\gamma+v(t))(1+v(t))^2}d\mu_{\mathrm{d}}dt.\end{aligned}$$

As $\delta>0$ is arbitrary, the above relation also holds for $\delta=0$. Then, as $\gamma\to0$, by using the dominated convergence theorem on the left-hand side and Fatou's lemma on the right-hand side, we obtain (59), as desired. ■

Remark 31 *It it tempting to conclude from Itô's formula for a flow of measures (see e.g., Theorem 4.14 in Carmona and Delarue (2018b)) and the identity* (17) *that* (58) *holds, and in fact with an equality. Indeed, this approach is taken in Theorem 2.9 of Hu et al. (2021), where a similar equality (instead of an inequality) is established for the minimization of a free energy functional. Note that Theorem 4.14 in Carmona and Delarue (2018b) and Theorem 2.9 in Hu et al. (2021) both require the smaller space $\mathcal{P}_2(\mathbb{R}^d)$ and appropriate continuity and growth conditions under the 2-nd Wasserstein distance. As these requirements are not met in our case, we take a fairly different approach in Proposition 30: we analyze $t\mapsto J(\rho^{Y_t})$ through PDE* (31)*, relying particularly on the approximation of v in* (40)*.*

By Proposition 30, ρ^{Y_t} moves closer to ρ_{d} continuously over time. The question now is whether ρ^{Y_t} will in fact converge to ρ_{d}. This indeed holds, at least along a subsequence.

Corollary 32 *Let Y be a solution to* (2) *such that $\eta(t,y):=\rho^{Y_t}(y)$ satisfies* (19)*. Then, there exists $\{t_n\}_{n\in\mathbb{N}}$ in $[0,\infty)$ with $t_n\uparrow\infty$ such that $\rho^{Y_{t_n}}\to\rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$.*

Proof As argued at the beginning of the proof of Proposition 30, $\rho^{Y_t} = u(t) = \rho_{\mathrm{d}} v(t) \in \mathcal{P}(\mathbb{R}^d)$ for all $t \geq 0$, where u and v are specified as in Corollary 19. In view of (51), there exists $\{t_n\}_{n\in\mathbb{N}}$ in $[0,\infty)$ with $t_n \uparrow \infty$ such that $\|\nabla v(t_n)\|_{L^2(\mathbb{R}^d,\mu_{\mathrm{d}})} \to 0$. This, along with $0 \leq v(t) \leq \beta$ for all $t \geq 0$, implies that the sequence $\{v(t_n)\}_{n\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^d,\mu_{\mathrm{d}})$. Hence, v_n converges weakly in $H^1(\mathbb{R}^d,\mu_{\mathrm{d}})$, possibly along a subsequence (without relabeling), to some $v_\infty \in H^1(\mathbb{R}^d,\mu_{\mathrm{d}})$. Note that "$v(t_n) \to v_\infty$ weakly in $H^1(\mathbb{R}^d,\mu_{\mathrm{d}})$" readily implies "$\nabla v(t_n) \to \nabla v_\infty$ weakly in $L^2(\mathbb{R}^d,\mu_{\mathrm{d}})$". As we already know $\nabla v(t_n) \to 0$ strongly in $L^2(\mathbb{R}^d,\mu_{\mathrm{d}})$, we must have $\nabla v_\infty = 0$, i.e., v_∞ is constant on $\mathbb{R}^d$. On the other hand, by Sobolev embedding, we have $v(t_n) \to v_\infty$ strongly in $L^2_{\mathrm{loc}}(\mathbb{R}^d,\mu_{\mathrm{d}})$. By the uniform boundedness of $\{v(t_n)\}_{n\in\mathbb{N}}$, this can be upgraded to "$v(t_n) \to v_\infty$ strongly in $L^p(\mathbb{R}^d,\mu_{\mathrm{d}})$, for all $p \in [1,\infty)$". Recall from (43) that $\int_{\mathbb{R}^d} v(t_n) d\mu_{\mathrm{d}} = 1$ for all $n \in \mathbb{N}$. This, together with $v_n \to v_\infty$ in $L^1(\mathbb{R}^d,\mu_{\mathrm{d}})$ and v_∞ being constant on $\mathbb{R}^d$, entails $v_\infty \equiv 1$. It follows that $\|\rho^{Y_{t_n}} - \rho_{\mathrm{d}}\|_{L^1(\mathbb{R}^d)} = \|u(t_n) - \rho_{\mathrm{d}}\|_{L^1(\mathbb{R}^d)} = \|v(t_n) - 1\|_{L^1(\mathbb{R}^d,\mu_{\mathrm{d}})} \to 0$. ■

Remark 33 *In the language of dynamical systems (see e.g., Section 4.3 in Henry (1981)), Corollary 32 states that the "omega limit set", defined by*

$$\Theta := \{\rho \in \mathcal{P}(\mathbb{R}^d) : \exists\{s_n\} \text{ in } [0,\infty) \text{ with } s_n \uparrow \infty \text{ s.t. } \rho^{Y_{s_n}} \to \rho \text{ in } L^1(\mathbb{R}^d)\},$$

must contain ρ_{d}. A result of this kind, in Step 1 in the proof of (Hu et al., 2021, Theorem 2.11), was obtained by a compactness argument, applicable under Lipschitz and growth conditions on a McKean-Vlasov SDE. As no such conditions are met in our case, compactness is elusive and we instead rely on (51)*, a property of the solution v to PDE* (31)*.*

Now, we are ready to prove Theorem 7, the main theoretic result of this paper.

Proof [Proof of Theorem 7] By Proposition 30, $t \mapsto J(\rho^{Y_t}) \geq 0$ is nonincreasing. This implies $\ell := \lim_{t\to\infty} J(\rho^{Y_t}) \geq 0$ is well-defined; moreover, for any $\{s_n\}_{n\in\mathbb{N}}$ in $[0,\infty)$ with $s_n \uparrow \infty$, we have $\lim_{n\to\infty} J(\rho^{Y_{s_n}}) = \ell$. In view of Corollary 32, there exist $\{t_n\}_{n\in\mathbb{N}}$ in $[0,\infty)$ with $t_n \uparrow \infty$ such that $\rho^{Y_{t_n}} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$. Thanks to the continuity of $J(\cdot)$ under the $L^1(\mathbb{R}^d)$-norm (see Remark 2), we get $\ell = \lim_{n\to\infty} J(\rho^{Y_{t_n}}) = J(\rho_{\mathrm{d}}) = 0$. Now, for any $\{s_n\}_{n\in\mathbb{N}}$ in $[0,\infty)$ with $s_n \uparrow \infty$, we have

$$\lim_{n\to\infty} \mathrm{JSD}(\rho^{Y_{s_n}}, \rho_{\mathrm{d}}) = \lim_{n\to\infty} J(\rho^{Y_{s_n}}) = \ell = 0.$$

By Remark 2 again, this implies $\rho^{Y_{s_n}} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$. As the sequence $\{s_n\}_{n\in\mathbb{N}}$ is arbitrarily picked, we conclude that $\rho^{Y_t} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$ as $t \to \infty$. ■

Acknowledgments

Yu-Jui Huang acknowledges support from National Science Foundation (NSF grant DMS-2109002). Yuchong Zhang acknowledges support from Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant RGPIN-2020-06290).

Appendix A. Proofs for Section 2

A.1 An Auxiliary Result

Lemma 34 *For any $\rho, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d)$, $\mathrm{TV}(\rho, \bar{\rho}) = \frac{1}{2}\|\rho - \bar{\rho}\|_{L^1(\mathbb{R}^d)}$.*

Proof Consider $A_* := \{x \in \mathbb{R}^d : \rho(x) > \bar{\rho}(x)\}$ and observe from (6) that

$$\mathrm{TV}(\rho, \bar{\rho}) = \max\bigg\{ \int_{A_*} (\rho(x) - \bar{\rho}(x))\, dx, \int_{A_*^c} (\bar{\rho}(x) - \rho(x))\, dx \bigg\}.$$

By a direct calculation,

$$\begin{aligned}\int_{A_*} (\rho(x) - \bar{\rho}(x))\, dx &= \frac{1}{2}\bigg\{ \int_{A_*} (\rho(x) - \bar{\rho}(x))\, dx + \bigg(1 - \int_{A_*^c} \rho(x)dx\bigg) - \bigg(1 - \int_{A_*^c} \bar{\rho}(x)dx\bigg)\bigg\} \\ &= \frac{1}{2}\bigg\{ \int_{A_*} (\rho(x) - \bar{\rho}(x))dx + \int_{A_*^c} (\bar{\rho}(x) - \rho(x))dx \bigg\} = \frac{1}{2}\|\rho - \bar{\rho}\|_{L^1(\mathbb{R}^d)}.\end{aligned}$$

As we similarly have $\int_{A_*^c} (\bar{\rho}(x) - \rho(x))dx = \frac{1}{2}\|\rho - \bar{\rho}\|_{L^1(\mathbb{R}^d)}$, the desired result follows. ■

A.2 Proof of Proposition 5

Given $\xi \in C_c^2(\mathbb{R}^d; \mathbb{R}^d)$, there exists $\theta \in C_c^3(\mathbb{R}^d; \mathbb{R})$ such that $\nabla\theta = \xi$. For any $\varepsilon > 0$, consider $\psi : \mathbb{R}^d \to \mathbb{R}$ defined by $\psi(x) = \frac{1}{2}|x|^2 + \varepsilon\theta(x)$ for all $x \in \mathbb{R}^d$. As $\nabla\xi$ is bounded (by $\xi \in C_c^2(\mathbb{R}^d; \mathbb{R}^d)$), we can take $\varepsilon > 0$ small enough such that ψ is strictly convex, or equivalently, $det(I + \varepsilon\nabla\xi) > 0$. This in turn implies that $\nabla\psi = I + \varepsilon\xi$ is one-to-one, such that the inverse $(I+\varepsilon\xi)^{-1}$ is well-defined. Hence, by Lemma 5.5.3 in Ambrosio et al. (2005), the measure μ_ε^ξ in (14) has a density $\rho_\varepsilon^\xi \in \mathcal{P}(\mathbb{R}^d)$, which is given by

$$\rho_\varepsilon^\xi(y) := \rho\left((I + \varepsilon\xi)^{-1}(y)\right)/det(\mathrm{Id} + \varepsilon\nabla\xi), \quad \forall y \in \mathbb{R}^d, \tag{63}$$

where Id denotes the $d \times d$ identity matrix. It can be shown that $(y, \varepsilon) \mapsto \rho_\varepsilon^\xi(y)$ is C^1 on $\mathbb{R}^d \times [0, 1]$ and satisfies $\rho_\varepsilon^\xi(y) \mid_{\varepsilon=0} = \rho(y)$ and $\frac{\partial\rho_\varepsilon^\xi(y)}{\partial\varepsilon} \mid_{\varepsilon=0} = -\nabla \cdot (\rho(y)\xi(y))$ (see equation (10.4.7) in Ambrosio et al. (2005)). Now, by (12),

$$\begin{aligned}\lim_{\varepsilon\to 0} \frac{G(\rho_\varepsilon^\xi) - G(\rho)}{\varepsilon} &= \lim_{\varepsilon\to 0} \int_0^1 \int_{\mathbb{R}^d} \frac{\delta G}{\delta\rho}\big((1-\lambda)\rho + \lambda\rho_\varepsilon^\xi, y\big)\frac{\rho_\varepsilon^\xi - \rho}{\varepsilon}(y)dyd\lambda \\ &= \lim_{\varepsilon\to 0} \int_0^1 \int_K \frac{\delta G}{\delta\rho}\big((1-\lambda)\rho + \lambda\rho_\varepsilon^\xi, y\big)\frac{\partial\rho_\varepsilon^\xi(y)}{\partial\varepsilon} \mid_{\varepsilon=\varepsilon_0(y)} dyd\lambda, \end{aligned} \tag{64}$$

where $\varepsilon_0(y) \in [0, \varepsilon]$ is obtained from the mean value theorem, and $K \subseteq \mathbb{R}^d$ is the compact support of ξ. The continuity of $\frac{\partial\rho_\varepsilon^\xi(y)}{\partial\varepsilon}$ in (y, ε) implies that $\frac{\partial\rho_\varepsilon^\xi(y)}{\partial\varepsilon} \mid_{\varepsilon=\varepsilon_0(y)}$ is bounded on K. Since $\rho_\varepsilon^\xi \to \rho$ uniformly, $\frac{\delta G}{\delta\rho}((1-\lambda)\rho + \lambda\rho_\varepsilon^\xi, y)$, as a function of (λ, y), converges uniformly to $\frac{\delta G}{\delta\rho}(\rho, y)$ on $[0, 1] \times K$ as $\varepsilon \to 0$ by assumption. Dominated convergence theorem then allows us to exchange the limit with the integral, yielding

$$\lim_{\varepsilon\to 0} \frac{G(\rho_\varepsilon^\xi) - G(\rho)}{\varepsilon} = -\int_0^1 \int_{\mathbb{R}^d} \frac{\delta G}{\delta\rho}(\rho, y)\nabla \cdot (\rho(y)\xi(y))dyd\lambda = \int_{\mathbb{R}^d} \nabla\frac{\delta G}{\delta\rho}(\rho, y) \cdot \xi(y)d\mu(y),$$

where the last equality is due to integration by parts.

A.3 Proof of Lemma 6

Recall from (8) that we can write $J(\rho) = \mathrm{JSD}(\rho, \rho_{\mathrm{d}}) = \int_{\mathbb{R}^d} f\big(\frac{\rho(y)}{\rho_{\mathrm{d}}(y)}\big)\rho_{\mathrm{d}}(y)dy$, where $f(x) := \frac{1}{2}\big[(x+1)\ln(\frac{2}{x+1}) + x\ln x\big]$, $\forall x \in [0,\infty)$, is a convex function satisfying $0 \le f(x) \le \frac{\ln 2}{2}(x+1)$. Fix $\rho, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d)$. For any $\lambda \in [0,1]$, set $\rho_\lambda := (1-\lambda)\rho + \lambda\bar{\rho} \in \mathcal{P}(\mathbb{R}^d)$, and define

$$g_\lambda := f'\left(\frac{\rho_\lambda}{\rho_{\mathrm{d}}}\right)(\bar{\rho} - \rho).$$

Using the convexity of f, one can show that g_λ is increasing in $\lambda \in [0,1]$ and integrable on $\mathbb{R}^d$ if $\lambda \in (0,1)$. Indeed, the monotonicity is implied from $\frac{d}{d\lambda}g_\lambda = f''\left(\frac{\rho_\lambda}{\rho_{\mathrm{d}}}\right)\frac{(\bar{\rho}-\rho)^2}{\rho_{\mathrm{d}}} \ge 0$, and the integrability is a consequence of the relation

$$-\frac{\ln 2}{2\lambda}(\rho + \rho_{\mathrm{d}}) \le \frac{f\left(\frac{\rho_\lambda}{\rho_{\mathrm{d}}}\right) - f\left(\frac{\rho}{\rho_{\mathrm{d}}}\right)}{\lambda}\rho_{\mathrm{d}} \le g_\lambda \le \frac{f\left(\frac{\bar{\rho}}{\rho_{\mathrm{d}}}\right) - f\left(\frac{\rho_\lambda}{\rho_{\mathrm{d}}}\right)}{1-\lambda}\rho_{\mathrm{d}} \le \frac{\ln 2}{2(1-\lambda)}(\bar{\rho} + \rho_{\mathrm{d}}).$$

Now, let $0 < \lambda_1 < \lambda_2 < 1$. By the fundamental theorem of calculus and Fubini's theorem,

$$\begin{aligned} J(\rho_{\lambda_2}) - J(\rho_{\lambda_1}) &= \int_{\mathbb{R}^d}\left[f\left(\frac{\rho_{\lambda_2}(y)}{\rho_{\mathrm{d}}(y)}\right) - f\left(\frac{\rho_{\lambda_1}(y)}{\rho_{\mathrm{d}}(y)}\right)\right]\rho_{\mathrm{d}}(y)dy \\ &= \int_{\mathbb{R}^d}\int_{\lambda_1}^{\lambda_2} g_\lambda(y)d\lambda dy = \int_{\lambda_1}^{\lambda_2}\int_{\mathbb{R}^d} g_\lambda(y)dyd\lambda. \end{aligned}$$

Letting $\lambda_1 \to 0$ and $\lambda_2 \to 1$ yields

$$J(\bar{\rho}) - J(\rho) = \int_0^1\int_{\mathbb{R}^d} g_\lambda(y)dyd\lambda = \int_0^1\int_{\mathbb{R}^d} f'\left(\frac{\rho_\lambda(y)}{\rho_{\mathrm{d}}(y)}\right)(\bar{\rho}(y) - \rho(y))dyd\lambda,$$

where the integral over λ may be interpreted as an improper integral if either $\int_{\mathbb{R}^d} g_0(y)dy = -\infty$ or $\int_{\mathbb{R}^d} g_1(y)dy = \infty$. This readily implies $\frac{\delta J}{\delta\rho}(\rho, y) = f'\big(\frac{\rho(y)}{\rho_{\mathrm{d}}(y)}\big) = \frac{1}{2}\ln\big(\frac{2\rho(y)}{\rho(y)+\rho_{\mathrm{d}}(y)}\big)$.

Appendix B. Proofs for Section 4

B.1 Derivation of Lemma 14

Lemma 35 *Let $E \subseteq \mathbb{R}$ be convex and $g : E \to \mathbb{R}$ be a concave, strictly increasing function such that for any $v : \mathbb{R}^d \to \mathbb{R}$ with $v(\mathbb{R}^d) \subseteq E$, the following hold:*

(i) if $v \in H_0^1(\mathbb{R}^d, \mu_d)$, then $g(v) \in H_0^1(\mathbb{R}^d, \mu_d)$;

(ii) if $v \in L^1(\mathbb{R}^d, \mu_d)$, then $1/(g'_+)(v) \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ (where g'_+ is the right derivative of g).

Then, the operator $-\Delta_{\mu_{\mathrm{d}}} g : D(-\Delta_{\mu_{\mathrm{d}}} g) \to L^1(\mathbb{R}^d, \mu_d)$, with

$$D(-\Delta_{\mu_{\mathrm{d}}} g) := \{v \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}}) \cap H_0^1(\mathbb{R}^d, \mu_d) : v(\mathbb{R}^d) \subseteq E,\ -\Delta_{\mu_{\mathrm{d}}} g(v) \in L^1(\mathbb{R}^d, \mu_d)\},$$

is accretive in $L^1(\mathbb{R}^d, \mu_d)$. That is, for any $v_1, v_2 \in D(-\Delta_{\mu_{\mathrm{d}}} g)$ and $\lambda > 0$,

$$\|v_1 - v_2\|_{L^1(\mathbb{R}^d, \mu_{\mathrm{d}})} \le \|(v_1 - \lambda\Delta_{\mu_{\mathrm{d}}} g(v_1)) - (v_2 - \lambda\Delta_{\mu_{\mathrm{d}}} g(v_2))\|_{L^1(\mathbb{R}^d, \mu_{\mathrm{d}})}.$$

Proof Fix $v_1, v_2 \in D(-\Delta_{\mu_\mathrm{d}} g)$ and $\lambda > 0$. Set $f_i := v_i - \lambda \Delta_{\mu_\mathrm{d}} g(v_i)$ for $i = 1, 2$. For any $\varphi \in H^1(\mathbb{R}^d, \mu_\mathrm{d}) \cap L^\infty(\mathbb{R}^d, \mu_\mathrm{d})$, by multiplying the equation $v_i - \lambda \Delta_{\mu_\mathrm{d}} g(v_i) = f_i$ by φ and using integration by parts as in Remark 11, we get

$$\int_{\mathbb{R}^d} v_i \varphi d\mu_\mathrm{d} + \lambda \int_{\mathbb{R}^d} \nabla g(v_i) \cdot \nabla \varphi d\mu_\mathrm{d} = \int_{\mathbb{R}^d} f_i \varphi d\mu_\mathrm{d}, \quad i = 1, 2,$$

Note that under integration by parts, the boundary term for the second integral above vanishes because $g(v_i) \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$ by condition (i). Let us write $v := v_1 - v_2$, $h := g(v_1) - g(v_2)$ and $f := f_1 - f_2$. We obtain from the above equation (by subtraction) that

$$\int_{\mathbb{R}^d} v \varphi d\mu_\mathrm{d} + \lambda \int_{\mathbb{R}^d} \nabla h \cdot \nabla \varphi d\mu_\mathrm{d} = \int_{\mathbb{R}^d} f \varphi d\mu_\mathrm{d}. \tag{65}$$

Now, for each $\varepsilon > 0$, consider the function

$$\chi_\varepsilon(x) := 1_{\{|x| \le \varepsilon\}} x/\varepsilon + 1_{\{|x| > \varepsilon\}} \mathrm{sgn}(x) \quad \text{for } x \in \mathbb{R},$$

which is a bounded, continuous, and weakly differentiable approximation of the $\mathrm{sgn}(\cdot)$ function. Taking $\varphi = \chi_\varepsilon(h) \in H^1(\mathbb{R}^d, \mu_\mathrm{d}) \cap L^\infty(\mathbb{R}^d, \mu_\mathrm{d})$ in (65) yields

$$\int_{\mathbb{R}^d} v \chi_\varepsilon(h) \, d\mu_\mathrm{d} + \lambda \int_{\mathbb{R}^d} |\nabla h|^2 \chi_\varepsilon'(h) \, d\mu_\mathrm{d} = \int_{\mathbb{R}^d} f \chi_\varepsilon(h) \, d\mu_\mathrm{d}.$$

As χ_ε is nondecreasing, the middle integral above is nonnegative. Consequently, we have

$$\|v\|_{L^1(\mathbb{R}^d, \mu_\mathrm{d})} + \int_{\{|h| \le \varepsilon\}} \left(v \frac{h}{\varepsilon} - |v| \right) d\mu_\mathrm{d} \le \int_{\mathbb{R}^d} f \chi_\varepsilon(h) \, d\mu \le \|f\|_{L^1(\mathbb{R}^d, \mu_\mathrm{d})},$$

where we used the fact $\mathrm{sgn}(v) = \mathrm{sgn}(h)$. Finally, observe that the integral on the left-hand side above converges to zero as $\varepsilon \to 0$. Indeed,

$$\int_{\{|h| \le \varepsilon\}} \left| v \frac{h}{\varepsilon} - |v| \right| d\mu_\mathrm{d} \le \int_{\{|h| \le \varepsilon\}} 2|v| d\mu_\mathrm{d} \le 2\varepsilon \int_{\mathbb{R}^d} \frac{1}{g_+'(\max(v_1, v_2))} d\mu_\mathrm{d} \to 0,$$

where the second inequality follows from $|h| = |g(v_1) - g(v_2)| \ge |v_1 - v_2| g_+'(\max(v_1, v_2))$, thanks to the concavity of g, and the convergence follows from condition (ii). We therefore conclude that $\|v\|_{L^1(\mathbb{R}^d, \mu_\mathrm{d})} \le \|f\|_{L^1(\mathbb{R}^d, \mu_\mathrm{d})}$, as desired. ■

Proof [Proof of Lemma 14] Applying Lemma 35 with $E = [0, \beta]$ and $g(x) = \frac{1}{2} \ln(1 + x)$ (resp. $E = \mathbb{R}$ and $g(x) = x$) yields (i) (resp. (ii)). ■

B.2 Derivation of Lemma 18

Let $\langle \cdot, \cdot \rangle$ denote the inner product in $L^2(\mathbb{R}^d, \mu_\mathrm{d})$. For any $\lambda > 0$, consider the bilinear form

$$\mathcal{B}[w, \varphi] := \frac{\lambda}{2} \int_{\mathbb{R}^d} \nabla w \cdot \nabla \varphi d\mu_\mathrm{d} = -\frac{\lambda}{2} \int_{\mathbb{R}^d} (\Delta_{\mu_\mathrm{d}} w) \varphi d\mu_\mathrm{d}, \quad \forall \, w, \varphi \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d}),$$

where the last equality follows from Remark 11.

Definition 36 *We say* $w \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d}) \cap L^\infty(\mathbb{R}^d, \mu_\mathrm{d})$ *is a*

(i) weak solution to (41) *w.r.t.* μ_d *if* $\mathcal{B}[w, \varphi] = \langle f + 1 - e^w, \varphi\rangle$ *for all* $\varphi \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$*;*

(ii) weak subsolution to (41) *w.r.t.* μ_d *if* $\mathcal{B}[w, \varphi] \le \langle f + 1 - e^w, \varphi\rangle$ $\forall \varphi \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$, $\varphi \ge 0$;

(iii) weak supersolution to (41) *w.r.t.* μ_d *if* $\mathcal{B}[w, \varphi] \ge \langle f + 1 - e^w, \varphi\rangle$ $\forall \varphi \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$, $\varphi \ge 0$.

The proof below relies on the construction of a weak solution w^* to (41) through a sequence of weak sub- and supersolutions.

Proof [Proof of Lemma 18] As $f \in \overline{D(A)}$, we have $0 \le f \le \beta$. It can then be verified directly that $\underline{w}_0 \equiv 0$ is a weak subsolution and $\overline{w}_0 \equiv \ln(1+\beta)$ is a weak supersolution to (41) w.r.t. μ_d. Starting from $\underline{w}_0$ and $\overline{w}_0$, we will construct recursively a sequence $\{\underline{w}_n, \overline{w}_n\}_{n\in\mathbb{N}}$ such that $\underline{w}_n$ and $\overline{w}_n$, for each $n \in \mathbb{N}$, are the unique weak solutions in $H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$ w.r.t. μ_d to the linear PDEs

$$\alpha \underline{w}_n - \frac{\lambda}{2}\Delta \underline{w}_n - \frac{\lambda}{2}\nabla \ln \rho_\mathrm{d} \cdot \nabla \underline{w}_n = \alpha \underline{w}_{n-1} + 1 - e^{\underline{w}_{n-1}} + f, \tag{66}$$

$$\alpha \overline{w}_n - \frac{\lambda}{2}\Delta \overline{w}_n - \frac{\lambda}{2}\nabla \ln \rho_\mathrm{d} \cdot \nabla \overline{w}_n = \alpha \overline{w}_{n-1} + 1 - e^{\overline{w}_{n-1}} + f, \tag{67}$$

where the constant α is chosen to satisfy $\alpha \ge 1 + \beta$, so that $x \mapsto g(x) := \alpha x - e^x$ is strictly increasing on $(-\infty, \ln(1+\beta)]$.

First, given that $\underline{w}_0, \overline{w}_0 \in L^\infty(\mathbb{R}^d, \mu_\mathrm{d})$, the existence of unique solutions $\underline{w}_1, \overline{w}_1 \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$ to (66)-(67) w.r.t. μ_d follows directly from the Lax-Milgram theorem applied to the bilinear form $\alpha\langle w, \varphi\rangle + \mathcal{B}[w, \varphi]$. Moreover, we claim that $\underline{w}_1, \overline{w}_1$ again belong to $L^\infty(\mathbb{R}^d, \mu_\mathrm{d})$ under the relation $\underline{w}_0 \le \underline{w}_1 \le \overline{w}_1 \le \overline{w}_0$ μ_d-a.e. As $\overline{w}_0$ is a weak supersolution to (41) w.r.t. μ_d, we have

$$\alpha \overline{w}_1 - \frac{\lambda}{2}\Delta \overline{w}_1 - \frac{\lambda}{2}\nabla \ln \rho_\mathrm{d} \cdot \nabla \overline{w}_1 = \alpha \overline{w}_0 + 1 - e^{\overline{w}_0} + f \le \alpha \overline{w}_0 - \frac{\lambda}{2}\Delta \overline{w}_0 - \frac{\lambda}{2}\nabla \ln \rho_\mathrm{d} \cdot \nabla \overline{w}_0$$

in the weak form. Setting $w := \overline{w}_1 - \overline{w}_0 \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$, we get $\alpha w - \frac{\lambda}{2}\Delta w - \frac{\lambda}{2}\nabla \ln \rho_\mathrm{d} \cdot \nabla w \le 0$ in the weak form. That is, $\alpha\langle w, \varphi\rangle + \mathcal{B}[w, \varphi] \le 0$, for all $\varphi \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$, $\varphi \ge 0$. Taking $\varphi = w^+ \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$ in the ineqaulity leads to the following implication:

$$\int_{\{w>0\}} \left(\alpha w^2 + \frac{\lambda}{2}|\nabla w|^2\right) d\mu_\mathrm{d} \le 0 \quad \Longrightarrow \quad w \le 0 \ \mu_\mathrm{d}\text{-a.e.} \tag{68}$$

We then conclude $\overline{w}_1 \le \overline{w}_0$ μ_d-a.e. Similarly, by the weak subsolution property of $\underline{w}_0$, we may again use (68) (with $w := \underline{w}_0 - \underline{w}_1$) to get $\underline{w}_1 \ge \underline{w}_0$ μ_d-a.e. Finally, thanks to $\underline{w}_0 \le \overline{w}_0$,

$$\begin{aligned}\alpha \underline{w}_1 - \frac{\lambda}{2}\Delta \underline{w}_1 - \frac{\lambda}{2}\nabla \ln \rho_\mathrm{d} \cdot \nabla \underline{w}_1 &= \alpha \underline{w}_0 + 1 - e^{\underline{w}_0} + f \\ &\le \alpha \overline{w}_0 + 1 - e^{\overline{w}_0} + f = \alpha \overline{w}_1 - \frac{\lambda}{2}\Delta \overline{w}_1 - \frac{\lambda}{2}\nabla \ln \rho_\mathrm{d} \cdot \nabla \overline{w}_1.\end{aligned}$$

Applying (68) again (with $w := \underline{w}_1 - \overline{w}_1$) gives $\underline{w}_1 \le \overline{w}_1$ $\mu_{\rm d}$-a.e. This finishes the proof of our claim that $\underline{w}_0 \le \underline{w}_1 \le \overline{w}_1 \le \overline{w}_0$ $\mu_{\rm d}$-a.e. As the same arguments above hold true for each iterate $n \in \mathbb{N}$, we obtain $\underline{w}_{n-1} \le \underline{w}_n \le \overline{w}_n \le \overline{w}_{n-1}$ $\mu_{\rm d}$-a.e. for all $n \in \mathbb{N}$.

Now, for each $n \in \mathbb{N}$, thanks to $\underline{w}_{n-1} \le \underline{w}_n$,

$$\begin{aligned}\alpha\underline{w}_{n+1} - \frac{\lambda}{2}\Delta\underline{w}_{n+1} - \frac{\lambda}{2}\nabla \ln \rho_{\rm d} \cdot \nabla \underline{w}_{n+1} &= \alpha\underline{w}_n + 1 - e^{\underline{w}_n} + f \\ &\ge \alpha\underline{w}_{n-1} + 1 - e^{\underline{w}_{n-1}} + f = \alpha\underline{w}_n - \frac{\lambda}{2}\Delta\underline{w}_n - \frac{\lambda}{2}\nabla \ln \rho_{\rm d} \cdot \nabla \underline{w}_n.\end{aligned}$$

Hence, applying (68) (with $w := \underline{w}_n - \underline{w}_{n+1}$) gives $\underline{w}_n \le \underline{w}_{n+1}$ $\mu_{\rm d}$-a.e. A similar argument shows $\overline{w}_{n+1} \le \overline{w}_n$ $\mu_{\rm d}$-a.e. That is, $\{\underline{w}_n\}_{n\in\mathbb{N}}$ is monotonically increasing, $\{\overline{w}_n\}_{n\in\mathbb{N}}$ is monotonically decreasing, and all the functions are $[0, \ln(1+\beta)]$-valued. Thus, $w^* := \lim_{n\to\infty} \overline{w}_n$ $\mu_{\rm d}$-a.e. is well-defined. We claim that w^* is a weak solution to (41) w.r.t. $\mu_{\rm d}$.[1]

Note that the operator $L := (\alpha I - \frac{\lambda}{2}(\Delta + \nabla \ln \rho_{\rm d} \cdot \nabla))^{-1}$ is continuous from $L^2(\mathbb{R}^d, \mu_{\rm d})$ to $H_0^1(\mathbb{R}^d, \mu_{\rm d})$. Indeed, given $f \in L^2(\mathbb{R}^d, \mu_{\rm d})$, by taking the test function $\varphi = w \in H_0^1(\mathbb{R}^d, \mu_{\rm d})$ in the weak form of $\alpha w - \frac{\lambda}{2}(\Delta w + \nabla \ln \rho_{\rm d} \cdot \nabla w) = f$, we deduce from Remark 11 that

$$\alpha \|w\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} + \frac{\lambda}{2}\|\nabla w\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} = \int_{\mathbb{R}^d} f w d\mu_{\rm d} \le \frac{1}{2\alpha}\|f\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})} + \frac{\alpha}{2}\|w\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})},$$

where the inequality follows from Young's inequality. It follows that

$$\min(\alpha, \lambda)\|w\|^2_{H_0^1(\mathbb{R}^d,\mu_{\rm d})} \le \frac{1}{\alpha}\|f\|^2_{L^2(\mathbb{R}^d,\mu_{\rm d})}, \tag{69}$$

which implies the desired continuity of L. Now, by rewriting (67) as $\overline{w}_n = L(\alpha\overline{w}_{n-1} + 1 - e^{\overline{w}_{n-1}} + f)$, we conclude from $w^* = \lim_{n\to\infty} \overline{w}_n$ $\mu_{\rm d}$-a.e., the bounded convergence theorem, and the continuity of L that $\overline{w}_n \to \overline{w} := L(\alpha w^* + 1 - e^{w^*} + f)$ in $H_0^1(\mathbb{R}^d, \mu_{\rm d})$. By passing to a subsequence (without relabeling), we have $\overline{w}_n \to \overline{w}$ $\mu_{\rm d}$-a.e., which entails $w^* = \overline{w}$ $\mu_{\rm d}$-a.e. Hence, we have $w^* \in H_0^1(\mathbb{R}^d, \mu_{\rm d})$ and $w^* = L(\alpha w^* + 1 - e^{w^*} + f)$, i.e. w^* is a weak solution to (41) w.r.t. $\mu_{\rm d}$. By a direct calculation, $v := e^{w^*} - 1 \in H_0^1(\mathbb{R}^d, \mu_{\rm d})$ is a weak solution to $(I + \lambda A)v = f$ w.r.t. $\mu_{\rm d}$. Finally, as $w^* = \lim_{n\to\infty} \overline{w}_n$ by construction satisfies $0 \le w^* \le \ln(1+\beta)$, v fulfills $0 \le v \le \beta$ and thus lies in $D(A)$. ■

B.3 Proof of Proposition 20

Fix $\varepsilon > 0$. By Remark 11, we may conclude from Lax-Milgram's theorem that for any $f \in L^2(\mathbb{R}^d, \mu_{\rm d})$, there is a unique weak solution $w \in H_0^1(\mathbb{R}^d, \mu_{\rm d})$ to $(\varepsilon I - \Delta_{\mu_{\rm d}})w = f$ w.r.t. $\mu_{\rm d}$. That is, the operator $\Gamma_\varepsilon := (\varepsilon I - \Delta_{\mu_{\rm d}})^{-1} : L^2(\mathbb{R}^d, \mu_{\rm d}) \to H_0^1(\mathbb{R}^d, \mu_{\rm d})$ is well-defined. As $\ln \rho_{\rm d} \in H^1(\mathbb{R}^d, \mu_{\rm d})$ (by Assumption 6.1), let us take $\{b_n\}$ in $C_c^\infty(\mathbb{R}^d)$ such that $b_n \to \nabla \ln \rho_{\rm d}$ in $L^2(\mathbb{R}^d, \mu_{\rm d})$. Then, for each $f \in L^2(\mathbb{R}^d, \mu_{\rm d})$, we consider the smoothed problem

$$(\varepsilon I - \Delta^n_{\mu_{\rm d}})w = f \quad \text{with} \quad \Delta^n_{\mu_{\rm d}} := \Delta + b_n \cdot \nabla, \tag{70}$$

and denote by $\Gamma^n_\varepsilon f := (\varepsilon I - \Delta^n_{\mu_{\rm d}})^{-1} f$ the unique weak solution $w \in H_0^1(\mathbb{R}^d, \mu_{\rm d})$ w.r.t. $\mu_{\rm d}$.

1. One could alternatively show that $\lim_{n\to\infty} \underline{w}_n$ is a weak solution to (41) w.r.t. $\mu_{\rm d}$.

Step 1: We show that for any $f \in L^\infty(\mathbb{R}^d, \mu_\mathrm{d})$,

$$\Gamma_\varepsilon^n f \to \Gamma_\varepsilon f \quad \text{in} \quad L^1(\mathbb{R}^d, \mu_\mathrm{d}). \tag{71}$$

Take $\{f_k\}$ in $C_c^\infty(\mathbb{R}^d)$ such that $f_k \to f$ in $L^1(\mathbb{R}^d, \mu_\mathrm{d})$. Without loss of generality, we may assume $|f_k| \le \|f\|_{L^\infty(\mathbb{R}^d,\mu_\mathrm{d})}$. For each $k \in \mathbb{N}$, by noting $\Gamma_\varepsilon^n f_k \in C^\infty(\mathbb{R}^d)$, we conclude from the maximum principle (Krylov, 1996, Theorem 2.9.2) that

$$\|\Gamma_\varepsilon^n f_k\|_{L^\infty(\mathbb{R}^d,\mu_\mathrm{d})} \le \frac{1}{\varepsilon}\|f_k\|_{L^\infty(\mathbb{R}^d,\mu_\mathrm{d})} \le \frac{1}{\varepsilon}\|f\|_{L^\infty(\mathbb{R}^d,\mu_\mathrm{d})}. \tag{72}$$

Now, in view of (70) and (32), the equation $(\varepsilon I - \Delta_{\mu_\mathrm{d}}^n)\Gamma_\varepsilon^n f_k = f_k$ can be written as

$$\big(\varepsilon I - \Delta_{\mu_\mathrm{d}} - (b_n - \nabla \ln \rho_\mathrm{d}) \cdot \nabla\big)\Gamma_\varepsilon^n f_k = f_k.$$

Let us multiply this equation by $\Gamma_\varepsilon^n f_k$ and integrate it w.r.t. μ_d. With the aid of Remark 11, Young's inequality, (72), and Hölder's inequality, we get

$$\begin{aligned}
&\varepsilon\|\Gamma_\varepsilon^n f_k\|_{L^2(\mu_\mathrm{d})}^2 + \|\nabla\Gamma_\varepsilon^n f_k\|_{L^2(\mu_\mathrm{d})}^2 = \int_{\mathbb{R}^d} f_k \Gamma_\varepsilon^n f_k d\mu_\mathrm{d} + \int_{\mathbb{R}^d} \Gamma_\varepsilon^n f_k (b_n - \nabla \ln \rho_\mathrm{d}) \cdot \nabla \Gamma_\varepsilon^n f_k d\mu_\mathrm{d} \\
&\le \frac{1}{2\varepsilon}\|f_k\|_{L^2(\mu_\mathrm{d})}^2 + \frac{\varepsilon}{2}\|\Gamma_\varepsilon^n f_k\|_{L^2(\mu_\mathrm{d})}^2 + \frac{1}{\varepsilon}\|f\|_{L^\infty(\mu_\mathrm{d})}\|b_n - \nabla \ln \rho_\mathrm{d}\|_{L^2(\mu_\mathrm{d})}\|\nabla\Gamma_\varepsilon^n f_k\|_{L^2(\mu_\mathrm{d})} \qquad (73)\\
&\le \frac{1}{2\varepsilon}\|f\|_{L^\infty(\mu_\mathrm{d})}^2 + \frac{\varepsilon}{2}\|\Gamma_\varepsilon^n f_k\|_{L^2(\mu_\mathrm{d})}^2 + \frac{1}{2\varepsilon^2}\|f\|_{L^\infty(\mu_\mathrm{d})}^2\|b_n - \nabla \ln \rho_\mathrm{d}\|_{L^2(\mu_\mathrm{d})}^2 + \frac{1}{2}\|\nabla\Gamma_\varepsilon^n f_k\|_{L^2(\mu_\mathrm{d})}^2.
\end{aligned}$$

It follows that

$$\varepsilon\|\Gamma_\varepsilon^n f_k\|_{L^2(\mu_\mathrm{d})}^2 + \|\nabla\Gamma_\varepsilon^n f_k\|_{L^2(\mu_\mathrm{d})}^2 \le \frac{1}{\varepsilon}\|f\|_{L^\infty(\mu_\mathrm{d})}^2 + \frac{1}{\varepsilon^2}\|f\|_{L^\infty(\mu_\mathrm{d})}^2\|b_n - \nabla \ln \rho_\mathrm{d}\|_{L^2(\mu_\mathrm{d})}^2 < \infty.$$

This shows that $\{\Gamma_\varepsilon^n f_k\}_{k\in\mathbb{N}}$ is bounded in $H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$. Hence, by passing to a subsequence (without relabeling), $\Gamma_\varepsilon^n f_k$ converges to some $\hat{w}$ weakly in $H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$ as $k \to \infty$. This implies that if we let $k \to \infty$ in the weak form of $(\varepsilon I - \Delta_{\mu_\mathrm{d}}^n)\Gamma_\varepsilon^n f_k = f_k$, we will get $(\varepsilon I - \Delta_{\mu_\mathrm{d}}^n)\hat{w} = f$ in the weak form, which indicates $\hat{w} = \Gamma_\varepsilon^n f$. That is, we have $\Gamma_\varepsilon^n f_k \to \Gamma_\varepsilon^n f$ weakly in $H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$ as $k \to \infty$. By taking the test function $\varphi = \Gamma_\varepsilon^n f \in H_0^1(\mathbb{R}^d, \mu_\mathrm{d})$ in the weak form of $(\varepsilon I - \Delta_{\mu_\mathrm{d}}^n)\Gamma_\varepsilon^n f = f$, we can repeat the gradient estimate (73), with f_k replaced by f, thereby obtaining

$$\varepsilon\|\Gamma_\varepsilon^n f\|_{L^2(\mu_\mathrm{d})}^2 + \|\nabla\Gamma_\varepsilon^n f\|_{L^2(\mu_\mathrm{d})}^2 \le \frac{1}{\varepsilon}\|f\|_{L^\infty(\mu_\mathrm{d})}^2 + \frac{1}{\varepsilon^2}\|f\|_{L^\infty(\mu_\mathrm{d})}^2\|b_n - \nabla \ln \rho_\mathrm{d}\|_{L^2(\mu_\mathrm{d})}^2. \tag{74}$$

By subtracting the equation $(\varepsilon I - \Delta_{\mu_\mathrm{d}})\Gamma_\varepsilon f = f$ from $(\varepsilon I - \Delta_{\mu_\mathrm{d}}^n)\Gamma_\varepsilon^n f = f$, we get

$$\varepsilon(\Gamma_\varepsilon^n f - \Gamma_\varepsilon f) - \Delta(\Gamma_\varepsilon^n f - \Gamma_\varepsilon f) - \nabla \ln \rho_\mathrm{d} \cdot \nabla(\Gamma_\varepsilon^n f - \Gamma_\varepsilon f) + (\nabla \ln \rho_\mathrm{d} - b_n) \cdot \nabla\Gamma_\varepsilon^n f = 0.$$

This implies $(\varepsilon I - \Delta_{\mu_\mathrm{d}})(\Gamma_\varepsilon^n f - \Gamma_\varepsilon f) = (b_n - \nabla \ln \rho_\mathrm{d}) \cdot \nabla\Gamma_\varepsilon^n f$, i.e. $\Gamma_\varepsilon((b_n - \nabla \ln \rho_\mathrm{d}) \cdot \nabla\Gamma_\varepsilon^n f) = \Gamma_\varepsilon^n f - \Gamma_\varepsilon f$. As $\Gamma_\varepsilon 0 = 0$ trivially, we deduce from Lemma 14 (ii) that

$$\begin{aligned}
\varepsilon\|\Gamma_\varepsilon^n f - \Gamma_\varepsilon f\|_{L^1(\mathbb{R}^d,\mu_\mathrm{d})} &\le \|(b_n - \nabla \ln \rho_\mathrm{d}) \cdot \nabla\Gamma_\varepsilon^n f\|_{L^1(\mathbb{R}^d,\mu_\mathrm{d})} \\
&\le \|b_n - \nabla \ln \rho_\mathrm{d}\|_{L^2(\mathbb{R}^d,\mu_\mathrm{d})}\|\nabla\Gamma_\varepsilon^n f\|_{L^2(\mathbb{R}^d,\mu_\mathrm{d})} \to 0 \quad \text{as } n \to \infty, \qquad (75)
\end{aligned}$$

where the convergence follows from $b_n \to \nabla \ln \rho_{\rm d}$ in $L^2(\mathbb{R}^d, \mu_{\rm d})$ and the boundedness of $\{\nabla \Gamma_\varepsilon^n f\}_{n\in\mathbb{N}}$ in $L^2(\mathbb{R}^d, \mu_{\rm d})$ (thanks to (74)). This readily gives (71).

Step 2: Let $v_1, v_2 \in C([0,\infty); L^1(\mathbb{R}^d, \mu_{\rm d})) \cap L^\infty_+([0,\infty)\times\mathbb{R}^d, \mu_{\rm d})$ be two weak solutions to (31) w.r.t. $\mu_{\rm d}$. By setting $v := v_1 - v_2$ and $h := \ln \frac{1+v_1}{1+v_2}$, we aim to show that $(\Gamma_\varepsilon v)_t = \frac{1}{2}(\varepsilon\Gamma_\varepsilon h - h)$ in the weak form, i.e.

$$\int_0^\infty \int_{\mathbb{R}^d} \left(\psi_t \Gamma_\varepsilon v + \frac{1}{2}\psi(\varepsilon\Gamma_\varepsilon h - h)\right) d\mu_{\rm d} dt = 0 \quad \forall\, \psi \in C_c^\infty((0,\infty)\times\mathbb{R}^d). \tag{76}$$

By construction, we have $v(0) = 0$ and $v_t = \frac{1}{2}\Delta_{\mu_{\rm d}} h$ in the weak form, i.e.

$$\int_0^\infty \int_{\mathbb{R}^d} \left(v\varphi_t + \frac{1}{2} h \Delta_{\mu_{\rm d}} \varphi\right) d\mu_{\rm d} dt = 0, \quad \forall \varphi \in C_c^\infty((0,\infty)\times\mathbb{R}^d), \tag{77}$$

where we use the calculation in Remark 11 twice. For any $\psi \in C_c^\infty((0,\infty)\times\mathbb{R}^d)$, define $\varphi(t,\cdot) := \Gamma_\varepsilon^n \psi(t,\cdot)$ for all $t \geq 0$. That is, $\varphi(t,\cdot)$ is the solution to the (linear) smoothed problem (70) (with $f(\cdot) = \psi(t,\cdot) \in C_c^\infty(\mathbb{R}^d)$), which implies $\varphi(t,\cdot) \in C^\infty(\mathbb{R}^d)$. By differentiating the equation $(\varepsilon I - \Delta_{\mu_{\rm d}}^n)\Gamma_\varepsilon^n \psi(t,\cdot) = \psi(t,\cdot)$ with respect to t, we observe that

$$\frac{\partial^m}{\partial t^m}\varphi(t,\cdot) = \frac{\partial^m}{\partial t^m}\Gamma_\varepsilon^n \psi(t,\cdot) = \Gamma_\varepsilon^n\left(\frac{\partial^m}{\partial t^m}\psi(t,\cdot)\right), \quad \forall m \in \mathbb{N}. \tag{78}$$

As a result, $\varphi \in C^\infty((0,\infty)\times\mathbb{R}^d)$. We will further show that φ is compactly supported. Let $\mathrm{supp}(\psi) \subseteq (0,\infty)\times\mathbb{R}^d$ be the compact support of ψ, $H \subset (0,\infty)$ be a bounded open interval containing the projection of $\mathrm{supp}(\psi)$ onto $(0,\infty)$, and $B \subset \mathbb{R}^d$ be an open ball containing the projection of $\mathrm{supp}(\psi)$ onto $\mathbb{R}^d$. For each $t \in H$, consider the Dirichlet problem $(\varepsilon I - \Delta_{\mu_{\rm d}}^n) w(\cdot) = \psi(t,\cdot)$ in B and $w = 0$ on ∂B, and denote by $\hat\varphi(t,\cdot) \in C^\infty(B)$ the unique solution. Now, we extend the domain of $\hat\varphi$ to $(0,\infty)\times\mathbb{R}^d$ by setting $\hat\varphi(t,y) := 0$ whenever $t \in (0,\infty)\setminus H$ or $y \in B^c$. Then, $\hat\varphi$ is compactly supported and for any $t > 0$, $\hat\varphi(t,\cdot) \in H_0^1(\mathbb{R}^d, \mu_{\rm d})$ satisfies $(\varepsilon I - \Delta_{\mu_{\rm d}}^n)\hat\varphi(t,\cdot) = \psi(t,\cdot)$ on $\mathbb{R}^d$. By the uniqueness of solutions to (70) in $H_0^1(\mathbb{R}^d, \mu_{\rm d})$, we must have $\varphi(t,\cdot) = \hat\varphi(t,\cdot)$ for all $t > 0$. Consequently, $\varphi = \Gamma_\varepsilon^n \psi$ is compactly supported with $\mathrm{supp}(\varphi) \subseteq H \times B$ for all $n \in \mathbb{N}$. With $\varphi = \Gamma_\varepsilon^n \psi \in C_c^\infty((0,\infty)\times\mathbb{R}^d)$, we obtain from (77) that

$$\int_0^\infty \int_{\mathbb{R}^d} \left(v(\Gamma_\varepsilon^n\psi)_t + \frac{1}{2} h \Delta_{\mu_{\rm d}}(\Gamma_\varepsilon^n \psi)\right) d\mu_{\rm d} dt = 0, \quad \forall \psi \in C_c^\infty((0,\infty)\times\mathbb{R}^d). \tag{79}$$

In view of (32), (70), and the equation $(\varepsilon I - \Delta_{\mu_{\rm d}}^n)\Gamma_\varepsilon^n \psi = \psi$, we have

$$\Delta_{\mu_{\rm d}}(\Gamma_\varepsilon^n\psi) = \Delta_{\mu_{\rm d}}^n(\Gamma_\varepsilon^n\psi) - (b_n - \nabla \ln \rho_{\rm d})\cdot\nabla\Gamma_\varepsilon^n\psi = \varepsilon\Gamma_\varepsilon^n\psi - \psi - (b_n - \nabla\ln\rho_{\rm d})\cdot\nabla\Gamma_\varepsilon^n\psi.$$

This, together with $(\Gamma_\varepsilon^n\psi)_t = \Gamma_\varepsilon^n \psi_t$ (by (78) with $m = 1$), allows us to rewrite (79) as

$$\int_0^\infty \int_{\mathbb{R}^d} \left(v\Gamma_\varepsilon^n\psi_t + \frac{1}{2}h(\varepsilon\Gamma_\varepsilon^n\psi - \psi)\right) d\mu_{\rm d} dt = \int_0^\infty \int_{\mathbb{R}^d} \frac{1}{2} h (b_n - \nabla\ln\rho_{\rm d})\cdot\nabla\Gamma_\varepsilon^n\psi \; d\mu_{\rm d} dt.$$

As h is bounded and $\mathrm{supp}(\nabla\Gamma_\varepsilon^n\psi) \subset H \times E$ for all $n \in \mathbb{N}$, we see that the right-hand side above vanishes as $n \to \infty$, by using Hölder's inequality and arguing as in (75) for the convergence. It follows that as $n \to \infty$, the previous equation becomes

$$\int_0^\infty \int_{\mathbb{R}^d} \left(v\Gamma_\varepsilon\psi_t + \frac{1}{2}h(\varepsilon\Gamma_\varepsilon\psi - \psi)\right) d\mu_{\rm d} dt = 0, \tag{80}$$

where the convergence of the left-hand side follows from the boundedness of v and h, $\mathrm{supp}(\Gamma_\varepsilon^n \psi) \subset H \times E$ for all $n \in \mathbb{N}$, and (71). Finally, observe that Γ_ε is symmetric in the sense that

$$\langle \Gamma_\varepsilon f, g\rangle = \langle f, \Gamma_\varepsilon g\rangle, \quad \forall f, g \in L^2(\mathbb{R}^d, \mu_{\mathrm{d}}). \tag{81}$$

Indeed, by taking the test function $\varphi \in H^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ as $\Gamma_\varepsilon g$ (resp. $\Gamma_\varepsilon f$) in the weak form of $f = (\varepsilon I - \Delta_{\mu_{\mathrm{d}}})\Gamma_\varepsilon f$ (resp. $g = (\varepsilon I - \Delta_{\mu_{\mathrm{d}}})\Gamma_\varepsilon g$), we get

$$\langle f, \Gamma_\varepsilon g\rangle = \varepsilon\langle \Gamma_\varepsilon f, \Gamma_\varepsilon g\rangle + \langle \nabla\Gamma_\varepsilon f, \nabla\Gamma_\varepsilon g\rangle = \langle g, \Gamma_\varepsilon f\rangle, \tag{82}$$

where we use the calculation in Remark 11. Hence, (80) is equivalent to (76), as desired.

Step 3: Now we set out to prove (46). Define $g^\varepsilon : [0,\infty) \to \mathbb{R}$ by

$$g^\varepsilon(t) := \langle \Gamma_\varepsilon v(t), v(t)\rangle = \varepsilon\|\Gamma_\varepsilon v(t)\|^2_{L^2(\mathbb{R}^d,\mu_{\mathrm{d}})} + \|\nabla\Gamma_\varepsilon v(t)\|^2_{L^2(\mathbb{R}^d,\mu_{\mathrm{d}})}, \tag{83}$$

where the last equality follows by taking $f = g = v(t)$ in (82). Rearranging the equation $(\varepsilon I - \Delta_{\mu_{\mathrm{d}}})\Gamma_\varepsilon h(t) = h(t)$ yields $\Delta_{\mu_{\mathrm{d}}}(\Gamma_\varepsilon h(t)) = \varepsilon\Gamma_\varepsilon h(t) - h(t)$. By a direct calculation, this in turn implies $(\varepsilon I - \Delta_{\mu_{\mathrm{d}}})(\varepsilon\Gamma_\varepsilon h - h) = \Delta_{\mu_{\mathrm{d}}} h$, i.e. $\Gamma_\varepsilon(\Delta_{\mu_{\mathrm{d}}} h) = \varepsilon\Gamma_\varepsilon h - h$. Hence, by recalling $(\Gamma_\varepsilon v)_t = \Gamma_\varepsilon v_t$ and $v_t = \frac{1}{2}\Delta_{\mu_{\mathrm{d}}} h$, we have

$$\begin{aligned}
(g^\varepsilon)' = \langle (\Gamma_\varepsilon v)_t, v\rangle + \langle \Gamma_\varepsilon v, v_t\rangle &= \left\langle \frac{1}{2}(\varepsilon\Gamma_\varepsilon h - h), v \right\rangle + \left\langle \Gamma_\varepsilon v, \frac{1}{2}\Delta_{\mu_{\mathrm{d}}} h \right\rangle \\
&= \left\langle \frac{1}{2}(\varepsilon\Gamma_\varepsilon h - h), v \right\rangle + \left\langle v, \frac{1}{2}(\varepsilon\Gamma_\varepsilon h - h) \right\rangle \\
&= \langle \varepsilon\Gamma_\varepsilon h - h, v\rangle = \langle \varepsilon h, \Gamma_\varepsilon v\rangle - \langle h, v\rangle \le \varepsilon\langle h, \Gamma_\varepsilon v\rangle,
\end{aligned}$$

where the third and fifth equalities follow from (81) and the inequality is due to $hv = \ln(\frac{1+v_1}{1+v_2})(v_1 - v_2) \ge 0$. This, together with $g^\varepsilon(0) = 0$ (as $v(0) = 0$), gives

$$\begin{aligned}
g^\varepsilon(t) \le \int_0^t \varepsilon\langle h(s), \Gamma_\varepsilon v(s)\rangle ds &\le \frac{\varepsilon}{2}\|v\|_{L^\infty(\mathbb{R}^d,\mu_{\mathrm{d}})} \int_0^t \left(1 + \|\Gamma_\varepsilon v(s)\|^2_{L^2(\mathbb{R}^d,\mu_{\mathrm{d}})}\right) ds \\
&\le \frac{1}{2}\|v\|_{L^\infty(\mathbb{R}^d,\mu_{\mathrm{d}})} \left(\varepsilon t + \int_0^t g^\varepsilon(s) ds\right),
\end{aligned}$$

where the second inequality follows from $|h| \le |v|$ (due to concavity of $\ln(\cdot)$ and $v_i \ge 0$) and Young's inequality, and the third inequality stems from (83). Then, by (83) and Grönwall's inequality,

$$\varepsilon\|\Gamma_\varepsilon v\|^2_{L^2(\mathbb{R}^d,\mu_{\mathrm{d}})} + \|\nabla\Gamma_\varepsilon v\|^2_{L^2(\mathbb{R}^d,\mu_{\mathrm{d}})} = g^\varepsilon(t) \le \frac{\varepsilon t}{2}\|v\|_{L^\infty(\mathbb{R}^d,\mu_{\mathrm{d}})} \exp\left(\frac{t}{2}\|v\|_{L^\infty(\mathbb{R}^d,\mu_{\mathrm{d}})}\right),$$

where the right-hand side above converges to 0 as $\varepsilon \to 0$. This implies $\varepsilon\Gamma_\varepsilon v, \nabla\Gamma_\varepsilon v \to 0$ in $L^2(\mathbb{R}^d, \mu_{\mathrm{d}})$, uniformly in t on compact intervals. Thus, taking any test function $\varphi \in C_c^\infty((0,\infty)\times\mathbb{R}^d)$ in the weak form of $v = (\varepsilon I - \Delta_{\mu_{\mathrm{d}}})\Gamma_\varepsilon v$ gives

$$\int_0^\infty \int_{\mathbb{R}^d} v\varphi d\mu_{\mathrm{d}} dt = \int_0^\infty \int_{\mathbb{R}^d} \left(\varepsilon\Gamma_\varepsilon v\varphi + \nabla\Gamma_\varepsilon v \cdot \nabla\varphi\right) d\mu_{\mathrm{d}} dt \to 0 \quad \text{as } \varepsilon \to 0,$$

which gives (46).

References

Luigi Ambrosio. Transport equation and Cauchy problem for BV vector fields. *Invent. Math.*, 158(2):227–260, 2004.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2005.

Abdul Fatir Ansari, Jonathan Scarlett, and Harold Soh. A characteristic function approach to deep implicit generative modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Abdul Fatir Ansari, Ming Liang Ang, and Harold Soh. Refining deep generative models via discriminator gradient flow. In *International Conference on Learning Representations*, 2021.

Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, volume 70, pages 214–223, 06–11 Aug 2017.

Viorel Barbu. *Nonlinear differential equations of monotone types in Banach spaces*. Springer Monographs in Mathematics. Springer, New York, 2010.

Viorel Barbu and Michael Röckner. From nonlinear Fokker-Planck equations to solutions of distribution dependent SDE. *Ann. Probab.*, 48(4):1902–1920, 2020.

Mikołlaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs, 2018. Preprint, available at https://arxiv.org/abs/1801.01401.

Haïm Brézis and Michael G. Crandall. Uniqueness of solutions of the initial-value problem for $u_t - \Delta\varphi(u) = 0$. *J. Math. Pures Appl. (9)*, 58(2):153–163, 1979.

Pierre Cardaliaguet, François Delarue, Jean-Michel Lasry, and Pierre-Louis Lions. The master equation and the convergence problem in mean field games. 2015. Preprint, available at https://arxiv.org/abs/1509.02505.

René Carmona and François Delarue. *Probabilistic theory of mean field games with applications. I.* Springer, Cham, 2018a.

René Carmona and François Delarue. *Probabilistic theory of mean field games with applications. II.* Springer, Cham, 2018b.

François Delarue, Daniel Lacker, and Kavita Ramanan. From the master equation to mean field game limit theory: a central limit theorem. *Electron. J. Probab.*, 24:Paper No. 51, 54, 2019.

Emily L Denton, Soumith Chintala, arthur szlam, and Rob Fergus. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

Yuan Gao, Yuling Jiao, Yang Wang, Yao Wang, Can Yang, and Shunkang Zhang. Deep generative learning via variational gradient flow. In *International Conference on Machine Learning*, pages 2093–2101, 2019.

Yuan Gao, Jian Huang, Yuling Jiao, and Jin Liu. Learning implicit generative models with theoretical guarantees. 2020. Preprint. Available at https://arxiv.org/abs/2002.02862.

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

Ian J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. 2016. Available at https://arxiv.org/abs/1701.00160.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Daniel Henry. *Geometric theory of semilinear parabolic equations*, volume 840 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin-New York, 1981.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kaitong Hu, Zhenjie Ren, David Šiška, and Ł ukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. *Ann. Inst. Henri Poincaré Probab. Stat.*, 57 (4):2043–2065, 2021.

Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991.

N. V. Krylov. *Lectures on elliptic and parabolic equations in Hölder spaces*, volume 12 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1996.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabas Poczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems 30*, pages 2203–2213. 2017.

Josie McCulloch and Christian Wagner. On the choice of similarity measures for type-2 fuzzy sets. *Information Sciences*, 510:135–154, 2020.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. 2018. Preprint, available at https://arxiv.org/abs/1802.05957.

Youssef Mroueh and Truyen Nguyen. On the convergence of gradient descent in GANs: MMD GAN as a gradient flow. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1720–1728, 2021.

Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2014.

Ferdinand Österreicher and Igor Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Ann. Inst. Statist. Math.*, 55(3):639–653, 2003.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Dario Trevisan. Well-posedness of multidimensional diffusion processes with weakly differentiable coefficients. *Electron. J. Probab.*, 21:Paper No. 22, 41, 2016.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant GANs: deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, 2020.