A Fisher-Rao gradient flow for entropic mean-field min-max games

Anonymous authors
Paper under double-blind review

Abstract

Gradient flows play a substantial role in addressing many machine learning problems. We examine the convergence in continuous-time of a *Fisher-Rao* (Mean-Field Birth-Death) gradient flow in the context of solving convex-concave min-max games with entropy regularization. We propose appropriate Lyapunov functions to demonstrate convergence with explicit rates to the unique mixed Nash equilibrium.

1 Introduction

The rapid progress of machine learning (ML) techniques such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), adversarial learning (Madry et al., 2018), multi-agent reinforcement learning (Zhang et al., 2021) has propelled a surge of interest in the study of optimization problems on the space of probability measures in recent years. Particularly noteworthy are the numerous works, e.g., (Hsieh et al., 2019; Domingo-Enrich et al., 2020; Wang & Chizat, 2023; Lu, 2023; Trillos & Trillos, 2023; Kim et al., 2024), illustrating how training GANs and addressing adversarial robustness can be cast as min-max games over probability measures. In this setting, understanding the time evolution of the players' initial strategies to the equilibrium of the game leverages the use of gradient flows on the space of probability measures. A discussion about the applications of gradient flows in optimization and sampling can be found in a recent survey (Trillos et al., 2023).

The Fisher-Rao (FR) gradient flow has been recently studied in the context of mean-field optimization problems (Liu et al., 2023), accelerating Langevin-based sampling from multi-modal distributions (Lu et al., 2019; 2023), and training shallow neural networks in the mean-field regime (Rotskoff et al., 2019). The present paper aims to extend these results and focuses on the continuous-time convergence of the *Fisher-Rao* (FR) gradient flow to the unique mixed Nash equilibrium of an entropy-regularized min-max game.

1.1 Notation and setup

For any $\mathcal{Z} \subseteq \mathbb{R}^d$, we denote by $\mathcal{P}_{\mathrm{ac}}(\mathcal{Z})$ the space of probability measures on $\mathcal{Z}$ which are absolutely continuous with respect to the Lebesgue measure. Following a standard convention, elements in $\mathcal{P}_{\mathrm{ac}}(\mathcal{Z})$ denote probability measures as well as their densities. Let $\mathcal{X}$, $\mathcal{Y} \subseteq \mathbb{R}^d$ and fix two reference probability measures $\pi(\mathrm{d}x) \propto e^{-U^\pi(x)}\mathrm{d}x \in \mathcal{P}_{\mathrm{ac}}(\mathcal{X})$ and $\rho(\mathrm{d}y) \propto e^{-U^\rho(y)}\mathrm{d}y \in \mathcal{P}_{\mathrm{ac}}(\mathcal{Y})$, where $U^\pi : \mathcal{X} \to \mathbb{R}$ and $U^\rho : \mathcal{Y} \to \mathbb{R}$ are two measurable functions. The relative entropy $\mathrm{D}_{\mathrm{KL}}(\cdot|\pi) : \mathcal{P}(\mathcal{X}) \to [0, \infty]$ with respect to π is given by

$$\mathrm{D}_{\mathrm{KL}}(\nu|\pi) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{\nu(x)}{\pi(x)}\right)\nu(x)\mathrm{d}x, & \text{if } \nu \text{ is absolutely continuous with respect to } \pi, \\ +\infty, & \text{otherwise,} \end{cases}$$

and we define $\mathrm{D}_{\mathrm{KL}}(\mu|\rho)$ analogously for any $\mu \in \mathcal{P}(\mathcal{Y})$. Let $F : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ be a convex-concave (possibly non-linear) function and $\sigma > 0$ be a regularization parameter. We consider the min-max problem

$$\min_{\nu \in \mathcal{P}(\mathcal{X})} \max_{\mu \in \mathcal{P}(\mathcal{Y})} V^\sigma(\nu, \mu), \text{ with } V^\sigma(\nu, \mu) \coloneqq F(\nu, \mu) + \frac{\sigma^2}{2}\left(\mathrm{D}_{\mathrm{KL}}(\nu|\pi) - \mathrm{D}_{\mathrm{KL}}(\mu|\rho)\right). \tag{1}$$

In order to solve (1), one typically aims to identify *mixed Nash equilibria* (MNEs) (von Neumann et al., 1944; Nash, 1951), characterized by pairs of measures $(\nu_\sigma^*, \mu_\sigma^*) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ that satisfy

$$V^\sigma(\nu_\sigma^*, \mu) \leq V^\sigma(\nu_\sigma^*, \mu_\sigma^*) \leq V^\sigma(\nu, \mu_\sigma^*), \quad \text{for all } (\nu, \mu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}). \tag{2}$$

It is important to highlight that when F is bilinear and $\sigma = 0$, i.e., $V^0(\nu, \mu) = \int_{\mathcal{Y}} \int_{\mathcal{X}} f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, for some $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, measures characterized by (2) represent MNEs in the classical sense of a two-player zero-sum game. Results concerned with the existence and uniqueness of an MNE for (1) are presented in Appendix A in Lascu et al. (2023). Therein, it is also proved that V^σ Γ-converges to F as $\sigma \downarrow 0$, under mild assumptions on F, π and ρ.

In optimization, the monotonic decrease of the objective function along the gradient flow is key to proving convergence. However, for min-max games the monotonic decrease no longer holds due to the conflicting actions of the players. Hence, a suitable Lyapunov function is needed. A common choice is the so-called Nikaidò-Isoda (NI) error (Nikaidô & Isoda, 1955), which, for all $(\nu, \mu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, can be defined as

$$\mathrm{NI}(\nu, \mu) \coloneqq \max_{\mu' \in \mathcal{P}(\mathcal{Y})} V^\sigma(\nu, \mu') - \min_{\nu' \in \mathcal{P}(\mathcal{X})} V^\sigma(\nu', \mu).$$

From the saddle point condition (2), it follows that $\mathrm{NI}(\nu, \mu) \geq 0$ and $\mathrm{NI}(\nu, \mu) = 0$ if and only if (ν, μ) is a MNE. We will also consider an alternative Lyapunov function given by (6).

In what follows, we will introduce the FR gradient flow on the space $(\mathcal{P}_{\mathrm{ac}}(\mathcal{X}) \times \mathcal{P}_{\mathrm{ac}}(\mathcal{Y}), \mathrm{FR})$, where FR is the Fisher-Rao distance defined by (3).

1.2 Fisher-Rao (mean-field birth-death) gradient flow

In this work, we are mainly focusing on the the dynamical formulation of the Fisher-Rao metric (see e.g. Gallouët & Monsaingeon (2017); Liu et al. (2023)). For any $\lambda_0, \lambda_1 \in \mathcal{P}_{\mathrm{ac}}(\mathcal{M})$, with $\mathcal{M} \subseteq \mathbb{R}^d$, the variational representation of the FR distance is given by

$$\mathrm{FR}^2(\lambda_0, \lambda_1) \coloneqq \inf \left\{ \int_0^1 \int_{\mathcal{M}} \left| r_s(x) - \int_{\mathcal{M}} r_s(y)\lambda_s(\mathrm{d}y) \right|^2 \lambda_s(\mathrm{d}x)\mathrm{d}s : \partial_t \lambda_t = \lambda_t \left(r_t - \int_{\mathcal{M}} r_t(y)\lambda_t(\mathrm{d}y) \right) \right\}, \tag{3}$$

where the infimum is taken over all curves $[0, 1] \ni t \mapsto (\lambda_t, r_t) \in \mathcal{P}_{\mathrm{ac}}(\mathcal{M}) \times L^2(\mathcal{M}; \lambda_t)$ solving

$$\partial_t \lambda_t = \lambda_t \left(r_t - \int_{\mathcal{M}} r_t(y)\lambda_t(\mathrm{d}y) \right) \tag{4}$$

in the distributional sense, such that $t \mapsto \lambda_t$ is weakly continuous with endpoints λ_0 and λ_1. The reaction term $r_t(x) \in \mathbb{R}$ is a scalar that dictates how much mass is created/destroyed at time $t > 0$ and position $x \in \mathbb{R}$. The integral in (4) guarantees that λ_t is a probability measure for all $t \geq 0$.

Inspired by Liu et al. (2023), we consider the Fisher-Rao (mean-field birth-death) gradient flow on the space $(\mathcal{P}_{\mathrm{ac}}(\mathcal{X}) \times \mathcal{P}_{\mathrm{ac}}(\mathcal{Y}), \mathrm{FR})$ in the setting of (1). As opposed to the mean-field Best Response dynamics studied in Lascu et al. (2023), which is another flow that can be used to solve (1), and which relies on introducing a fixed point perspective on min-max games, the FR dynamics utilize a gradient flow $(\nu_t, \mu_t)_{t \geq 0}$ in the Fisher-Rao geometry. As a first attempt at defining a Fisher-Rao gradient flow for solving (1), consider

$$\begin{cases} \partial_t \nu_t(x) = -\frac{\delta V^\sigma}{\delta \nu}(\nu_t, \mu_t, x)\nu_t(x), \\ \partial_t \mu_t(y) = \frac{\delta V^\sigma}{\delta \mu}(\nu_t, \mu_t, y)\mu_t(y), \end{cases} \tag{5}$$

with initial condition $(\nu_0, \mu_0) \in \mathcal{P}_{\mathrm{ac}}(\mathcal{X}) \times \mathcal{P}_{\mathrm{ac}}(\mathcal{Y})$. We adopt the convention that

$$\int_{\mathcal{X}} \frac{\delta V^\sigma}{\delta \nu}(\nu_t, \mu_t, x)\nu_t(\mathrm{d}x) = \int_{\mathcal{Y}} \frac{\delta V^\sigma}{\delta \mu}(\nu_t, \mu_t, y)\mu_t(\mathrm{d}y) = 0$$

since the flat derivatives of V^σ are uniquely defined up to an additive shift (see Definition B.1 in Section B). Thus, the total mass 1 of probability measures is still preserved along the gradient flow.

1.2.1 Sketch of convergence proof for the FR gradient flow

For the sake of presenting an intuitive heuristic argument, we ignore here for now that the *flat derivatives* $(\nu,\mu,x)\mapsto \frac{\delta V^\sigma}{\delta \nu}(\nu,\mu,x)$ and $(\nu,\mu,y)\mapsto \frac{\delta V^\sigma}{\delta \mu}(\nu,\mu,y)$ may not exist for the V^σ defined in (1), due to the relative entropy term $\mathrm{D_{KL}}$ being only lower semicontinuous (for this reason, in our analysis in Section 2, we will replace these two derivatives with appropriately defined auxiliary functions a and b). Nevertheless, we will now demonstrate that choosing the flow $(\nu_t,\mu_t)_{t\geq 0}$ as in (5), makes the function

$$t\mapsto \mathrm{D_{KL}}(\nu_\sigma^*|\nu_t)+\mathrm{D_{KL}}(\mu_\sigma^*|\mu_t) \tag{6}$$

decrease in t, under the assumption of convexity-concavity of F. Let $(\nu,\mu)\in\mathcal{P}_{\mathrm{ac}}(\mathcal{X})\times\mathcal{P}_{\mathrm{ac}}(\mathcal{Y})$. Then assuming the existence of the flow $(\nu_t,\mu_t)_{t\geq 0}$ satisfying (5), and the differentiablity of the map $t\mapsto \mathrm{D_{KL}}(\nu|\nu_t)+\mathrm{D_{KL}}(\mu|\mu_t)$, we formally have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D_{KL}}(\nu|\nu_t)+\mathrm{D_{KL}}(\mu|\mu_t)\right)&=\int_{\mathcal{X}}\partial_t\left(\nu(x)\log\frac{\nu(x)}{\nu_t(x)}\right)\mathrm{d}x+\int_{\mathcal{Y}}\partial_t\left(\mu(y)\log\frac{\mu(y)}{\mu_t(y)}\right)\mathrm{d}y\\&=-\int_{\mathcal{X}}\left(\nu(x)-\nu_t(x)\right)\frac{\partial_t\nu_t(x)}{\nu_t(x)}\mathrm{d}x-\int_{\mathcal{Y}}\left(\mu(y)-\mu_t(y)\right)\frac{\partial_t\mu_t(y)}{\mu_t(y)}\mathrm{d}y\\&=\int_{\mathcal{X}}\frac{\delta V^\sigma}{\delta\nu}(\nu_t,\mu_t,x)(\nu-\nu_t)(\mathrm{d}x)-\int_{\mathcal{Y}}\frac{\delta V^\sigma}{\delta\mu}(\nu_t,\mu_t,y)(\mu-\mu_t)(\mathrm{d}y),\end{aligned}$$

where the second equality follows from the fact that $\int\partial_t\nu_t(x)\mathrm{d}x=\int\partial_t\mu_t(y)\mathrm{d}y=0$. Then, assuming that $\nu\mapsto F(\nu,\mu)$ and $\mu\mapsto F(\nu,\mu)$ are convex and concave, respectively (see Assumption 1), we observe that $\nu\mapsto V^\sigma(\nu,\mu)$ and $\mu\mapsto V^\sigma(\nu,\mu)$ are σ-strongly-convex and σ-strongly-concave relative to $\mathrm{D_{KL}}$, respectively (see Lemma 3.1 in Section A), that is

$$\begin{aligned}V^\sigma(\nu,\mu_t)-V^\sigma(\nu_t,\mu_t)&\geq\int_{\mathcal{X}}\frac{\delta V^\sigma}{\delta\nu}(\nu_t,\mu_t,x)(\nu-\nu_t)(\mathrm{d}x)+\frac{\sigma^2}{2}\mathrm{D_{KL}}(\nu|\nu_t),\\V^\sigma(\nu_t,\mu)-V^\sigma(\nu_t,\mu_t)&\leq\int_{\mathcal{Y}}\frac{\delta V^\sigma}{\delta\mu}(\nu_t,\mu_t,y)(\mu-\mu_t)(\mathrm{d}y)-\frac{\sigma^2}{2}\mathrm{D_{KL}}(\mu|\mu_t).\end{aligned}$$

Therefore, we obtain that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D_{KL}}(\nu|\nu_t)+\mathrm{D_{KL}}(\mu|\mu_t)\right)\leq V^\sigma(\nu,\mu_t)-V^\sigma(\nu_t,\mu_t)+V^\sigma(\nu_t,\mu_t)-V^\sigma(\nu_t,\mu)\\-\frac{\sigma^2}{2}\mathrm{D_{KL}}(\nu|\nu_t)-\frac{\sigma^2}{2}\mathrm{D_{KL}}(\mu|\mu_t).\end{aligned}\tag{7}$$

Setting $(\nu,\mu)=(\nu_\sigma^*,\mu_\sigma^*)$ in (7), using the the saddle point condition (2), i.e., $V^\sigma(\nu_\sigma^*,\mu_t)-V^\sigma(\nu_t,\mu_\sigma^*)\leq 0$, and applying Gronwall's inequality gives

$$\mathrm{D_{KL}}(\nu_\sigma^*|\nu_t)+\mathrm{D_{KL}}(\mu_\sigma^*|\mu_t)\leq e^{-\frac{\sigma^2}{2}t}\left(\mathrm{D_{KL}}(\nu_\sigma^*|\nu_0)+\mathrm{D_{KL}}(\mu_\sigma^*|\mu_0)\right).$$

Since $\mathrm{D_{KL}}(\nu_\sigma^*|\nu)+\mathrm{D_{KL}}(\mu_\sigma^*|\mu)\geq 0$ with equality if and only if $\nu=\nu_\sigma^*$ and $\mu=\mu_\sigma^*$, it follows that the unique MNE of (1) is achieved with exponential rate $e^{-\frac{\sigma^2}{2}t}$. To prove convergence in terms of the NI error, first observe that since

$$\frac{\sigma^2}{2}\left(\mathrm{D_{KL}}(\nu|\nu_t)+\mathrm{D_{KL}}(\mu|\mu_t)\right)\geq 0,$$

for all (ν,μ), (7) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D_{KL}}(\nu|\nu_t)+\mathrm{D_{KL}}(\mu|\mu_t)\right)\leq V^\sigma(\nu,\mu_t)-V^\sigma(\nu_t,\mu).$$

We integrate this inequality from 0 to $t>0$, divide by t and apply Jensen's inequality to $\nu\mapsto V^\sigma(\nu,\mu)$ and $\mu\mapsto V^\sigma(\nu,\mu)$, respectively. Lastly, maximizing over (ν,μ) gives

$$\mathrm{NI}\left(\frac{1}{t}\int_0^t\nu_s\mathrm{d}s,\frac{1}{t}\int_0^t\mu_s\mathrm{d}s\right)\leq\frac{1}{t}\left(\max_\nu\mathrm{D_{KL}}(\nu|\nu_0)+\max_\mu\mathrm{D_{KL}}(\mu|\mu_0)\right).$$

1.3 Our contribution

We prove the existence of the FR gradient flow $(\nu_t, \mu_t)_{t\geq 0}$ and show that it converges with rate $\mathcal{O}\left(e^{-\frac{\sigma^2}{2}t}\right)$ to the unique MNE of (1) with respect to the Lyapunov function $t \mapsto \mathrm{D}_{\mathrm{KL}}(\nu_\sigma^*|\nu_t) + \mathrm{D}_{\mathrm{KL}}(\mu_\sigma^*|\mu_t)$. We also show that $\mathrm{NI}\left(\frac{1}{t}\int_0^t \nu_s \mathrm{d}s, \frac{1}{t}\int_0^t \mu_s \mathrm{d}s\right)$ converges to zero with rate $\mathcal{O}\left(\frac{1}{t}\right)$.

1.4 Related works

Recent intensive research has been dedicated to examining the convergence of the Wasserstein gradient flow to MNEs within the specific formulation of game (1) in which F is bilinear ($F(\nu,\mu) = \int_{\mathcal{Y}}\int_{\mathcal{X}} f(x,y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$) and regularized by entropy rather than relative entropy D_{KL}. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be either compact smooth manifolds without boundary, embedded in the Euclidean space, or Euclidean tori, while f exhibits sufficient regularity, typically being at least continuously differentiable with Lipschitz conditions satisfied by $\nabla_x f$ and $\nabla_y f$. This line of research is explored in works such as Domingo-Enrich et al. (2020); Ma & Ying (2022); Lu (2023); Wang & Chizat (2023).

In this context, Ma & Ying (2022); Lu (2023) investigate the convergence of the Wasserstein gradient flow, proving exponential convergence to the MNE when the players' flows $(\nu_t)_{t\geq 0}$ and $(\mu_t)_{t\geq 0}$ converge at different rates. In Ma & Ying (2022), the analysis considers the scenario where one player's flow reaches equilibrium while the other remains governed by the Wasserstein gradient flow equation. Notably, Ma & Ying (2022, Theorem 5) shows the convergence (without explicit rate) of the flow $(\nu_t, \mu_t)_{t\geq 0}$ to the unique MNE under these separated dynamics.

On the other hand, Lu (2023) examines the situation where the players' flows evolve at varying speeds but with finite timescale separation, meaning that neither player has reached equilibrium. Here, Lu (2023, Theorem 2.1) proves exponential convergence of the finitely timescale-separated Wasserstein gradient flow to the unique MNE, with the convergence rate depending upon regularization and timescale separation parameters. In contrast to Ma & Ying (2022); Lu (2023), we prove that the Fisher-Rao gradient flow (5) converges exponentially fast to the unique MNE while players' dynamics converge at the same speed.

Other works focused on the convergence of the Wasserstein-Fisher-Rao (WFR) gradient flow, combining both the Wasserstein gradient flow (allowing particles to diffuse in space) and the Fisher-Rao gradient flow (forcing particles to evade low probability regions). Assuming that F is bilinear and $\sigma = 0$, Domingo-Enrich et al. (2020) investigates the Wasserstein-Fisher-Rao gradient flow's convergence (without giving explicit rates). For $t_0 > 0$ (dependent on the parameters governing the individual contributions of the Wasserstein and Fisher-Rao components in the WFR gradient flow), and under circumstances where the Fisher-Rao component predominates over the Wasserstein component, Domingo-Enrich et al. (2020, Theorem 2) establishes that the pair $\left(\frac{1}{t_0}\int_0^{t_0} \nu_s \mathrm{d}s, \frac{1}{t_0}\int_0^{t_0} \mu_s \mathrm{d}s\right)$ is an ϵ-approximate MNE of the game, i.e., $\mathrm{NI}\left(\frac{1}{t_0}\int_0^{t_0} \nu_s \mathrm{d}s, \frac{1}{t_0}\int_0^{t_0} \mu_s \mathrm{d}s\right) \leq \epsilon$, for any arbitrarily chosen $\epsilon > 0$.

Lastly, Wang & Chizat (2023) introduces a proximal point method that can be viewed as a discrete-time counterpart of the WFR gradient flow. Working within the framework of bilinear F, with $\sigma = 0$, and unique MNE, Wang & Chizat (2023, Theorem 2.2) establishes the local exponential convergence of the iterates to the unique MNE of the game with respect to the NI error and the WFR distance, provided that the initialization is done in close vicinity to the MNE.

2 Main results

As we explained in the introduction, we study the convergence of the FR gradient flow to the unique MNE of the entropy-regularized two-player zero-sum game given by (1), where $F : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ is a non-linear function and $\sigma > 0$. Throughout the paper, we have the following assumptions.

Assumption 1 (Convexity-concavity of F)**.** *Suppose F admits first order flat derivatives with respect to both ν and μ as stated in Definition B.1. Furthermore, suppose that F is convex in ν and concave in μ, i.e.,*

for any $\nu, \nu' \in \mathcal{P}(\mathcal{X})$ and any $\mu, \mu' \in \mathcal{P}(\mathcal{Y})$, we have

$$F(\nu', \mu) - F(\nu, \mu) \geq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, \mu, x)(\nu' - \nu)(\mathrm{d}x), \tag{8}$$

$$F(\nu, \mu') - F(\nu, \mu) \leq \int_{\mathcal{Y}} \frac{\delta F}{\delta \mu}(\nu, \mu, y)(\mu' - \mu)(\mathrm{d}y). \tag{9}$$

Assumption 2 (Boundedness of first order flat derivatives)**.** *There exist constants $C_\nu, C_\mu > 0$ such that for all $(\nu, \mu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ and for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have*

$$\left|\frac{\delta F}{\delta \nu}(\nu, \mu, x)\right| \leq C_\nu, \quad \left|\frac{\delta F}{\delta \mu}(\nu, \mu, y)\right| \leq C_\mu.$$

Assumption 3 (Boundedness of second order flat derivatives)**.** *Suppose F admits second order flat derivatives and that there exist constants $C_{\nu,\nu}, C_{\mu,\mu}, C_{\nu,\mu}, C_{\mu,\nu} > 0$ such that for all $(\nu, \mu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ and for all $(x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}$, we have*

$$\left|\frac{\delta^2 F}{\delta \nu^2}(\nu, \mu, x, x')\right| \leq C_{\nu,\nu}, \quad \left|\frac{\delta^2 F}{\delta \mu^2}(\nu, \mu, y, y')\right| \leq C_{\mu,\mu},$$
$$\left|\frac{\delta^2 F}{\delta \nu \delta \mu}(\nu, \mu, y, x)\right| \leq C_{\nu,\mu}, \quad \left|\frac{\delta^2 F}{\delta \mu \delta \nu}(\nu, \mu, x, y)\right| \leq C_{\mu,\nu}.$$

Note that the order of the flat derivatives in ν and μ can be interchanged due to Lascu et al. (2023, Lemma B.2). Using Assumption 3, it is straightforward to check that there exist constants $C'_\nu, C'_\mu > 0$ such that for all $(\nu, \mu) \in \mathcal{P}_{\mathrm{ac}}(\mathcal{X}) \times \mathcal{P}_{\mathrm{ac}}(\mathcal{Y})$, $(\nu', \mu') \in \mathcal{P}_{\mathrm{ac}}(\mathcal{X}) \times \mathcal{P}_{\mathrm{ac}}(\mathcal{Y})$ and all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have that

$$\left|\frac{\delta F}{\delta \nu}(\nu, \mu, x) - \frac{\delta F}{\delta \nu}(\nu', \mu', x)\right| \leq C'_\nu \left(\mathrm{TV}(\nu, \nu') + \mathrm{TV}(\mu, \mu')\right), \tag{10}$$

$$\left|\frac{\delta F}{\delta \mu}(\nu, \mu, y) - \frac{\delta F}{\delta \mu}(\nu', \mu', y)\right| \leq C'_\mu \left(\mathrm{TV}(\nu, \nu') + \mathrm{TV}(\mu, \mu')\right). \tag{11}$$

Remark 2.1. *An example of a function F which satisfies Assumptions 1, 2, 3 is $F(\nu, \mu) = \int_{\mathcal{Y}} \int_{\mathcal{X}} f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, for a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ which is bounded but possibly non-convex-non-concave. Indeed, Assumption 1 is trivially satisfied by such F, while Assumptions 2 and 3 hold due to the boundedness of f. This type of objective function F is prototypical in applications including the training of GANs (see, e.g., Arjovsky et al. (2017); Hsieh et al. (2019)) and distributionally robust optimization (see, e.g, Madry et al. (2018); Sinha et al. (2018)).*

Assumption 4 (Ratio condition)**.** *Suppose $(\nu_0, \mu_0) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ are absolutely continuous and comparable with π and ρ, respectively, in the sense that*

1. *There exist constants $r_\nu, r_\mu > 0$ such that*

$$\inf_{x \in \mathcal{X}} \frac{\nu_0(x)}{\pi(x)} \geq r_\nu, \quad \inf_{y \in \mathcal{Y}} \frac{\mu_0(y)}{\rho(y)} \geq r_\mu. \tag{12}$$

2. *There exist constants $R_\nu, R_\mu > 1$ such that*

$$\sup_{x \in \mathcal{X}} \frac{\nu_0(x)}{\pi(x)} \leq R_\nu, \quad \sup_{y \in \mathcal{Y}} \frac{\mu_0(y)}{\rho(y)} \leq R_\mu. \tag{13}$$

It can be proved (see, e.g., Lascu et al. (2023, Proposition A.1)) that the MNE $(\nu^*_\sigma, \mu^*_\sigma)$ of 1 is given implicitly by the equations

$$\nu^*_\sigma(x) = \frac{1}{Z(\nu^*_\sigma, \mu^*_\sigma)} \exp\left(-\frac{2}{\sigma^2} \frac{\delta F}{\delta \nu}(\nu^*_\sigma, \mu^*_\sigma, x) - U^\pi(x)\right), \tag{14}$$

$$\mu_\sigma^*(y) = \frac{1}{Z'(\nu_\sigma^*, \mu_\sigma^*)} \exp\left(\frac{2}{\sigma^2}\frac{\delta F}{\delta \mu}(\nu_\sigma^*, \mu_\sigma^*, y) - U^\rho(y)\right), \tag{15}$$

where $Z(\nu_\sigma^*, \mu_\sigma^*)$ and $Z'(\nu_\sigma^*, \mu_\sigma^*)$ are normalizing constants.

Combining (14) and (15) and Assumption 2, we deduce that Assumption 4 is equivalent to assuming that there exist constants $\bar{r}_\nu, \bar{r}_\mu > 0$ and $\bar{R}_\nu, \bar{R}_\mu > 1$ such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\bar{r}_\nu \leq \frac{\nu_0(x)}{\nu_\sigma^*(x)} \leq \bar{R}_\nu, \quad \bar{r}_\mu \leq \frac{\mu_0(y)}{\mu_\sigma^*(y)} \leq \bar{R}_\mu. \tag{16}$$

We emphasize that Assumption 4 is natural in the context of Fisher-Rao flows. From (5), we observe that the support of the measures along the gradient flow does not increase. Thus, it is essential that the initialization is comparable with the MNE (cf. (16)) as reflected in Assumption 4. It is observed in Liu et al. (2023), that Assumption 4 can be understood as a "warm start" type of condition.

Returning to the question of flat differentiability of V^σ, which was raised in Subsection 1.2, if we assume that F is flat differentiable with respect to both ν and μ (see Assumption 1), then the maps $(\nu, \mu, x) \mapsto a(\nu, \mu, x) \coloneqq \frac{\delta F}{\delta \nu}(\nu, \mu, x) + \frac{\sigma^2}{2}\log\left(\frac{\nu(x)}{\pi(x)}\right) - \frac{\sigma^2}{2}\,\mathrm{D_{KL}}(\nu|\pi)$ and $(\nu, \mu, y) \mapsto b(\nu, \mu, y) \coloneqq \frac{\delta F}{\delta \mu}(\nu, \mu, y) - \frac{\sigma^2}{2}\log\left(\frac{\mu(y)}{\rho(y)}\right) + \frac{\sigma^2}{2}\,\mathrm{D_{KL}}(\mu|\rho)$ are well-defined and formally correspond to the flat derivatives $\frac{\delta V^\sigma}{\delta \nu}(\nu, \mu, \cdot)$ and $\frac{\delta V^\sigma}{\delta \mu}(\nu, \mu, \cdot)$, respectively, for those measures ν and μ for which such derivatives exist (note that we will only need to consider a and b along our gradient flow $(\nu_t, \mu_t)_{t\geq 0}$, so our argument can be interpreted as stating that, while V^σ is not flat differentiable everywhere, it is indeed flat differentiable along our gradient flow). The relative entropy terms D_{KL} appear in the definition of a and b as normalizing constants to ensure that $\int_{\mathcal{X}} a(\nu, \mu, x)\nu(\mathrm{d}x) = 0$ and $\int_{\mathcal{Y}} b(\nu, \mu, y)\mu(\mathrm{d}y) = 0$, since we adopt the convention that the flat derivatives of F are uniquely defined up to an additive shift (see Definition B.1 in Section B). Motivated by this discussion, we define the Fisher-Rao gradient flow $(\nu_t, \mu_t)_{t\geq 0}$ on the space $(\mathcal{P}_{\mathrm{ac}}(\mathcal{X}) \times \mathcal{P}_{\mathrm{ac}}(\mathcal{Y}), \mathrm{FR})$ by

$$\begin{cases} \partial_t \nu_t(x) = -a(\nu_t, \mu_t, x)\nu_t(x), \\ \partial_t \mu_t(y) = b(\nu_t, \mu_t, y)\mu_t(y), \end{cases} \tag{17}$$

with initial condition $(\nu_0, \mu_0) \in \mathcal{P}_{\mathrm{ac}}(\mathcal{X}) \times \mathcal{P}_{\mathrm{ac}}(\mathcal{Y})$.

The following result extends Liu et al. (2023, Theorem 2.1) to the case of two-player zero-sum games by showing that the Fisher-Rao gradient flow (17) admits a unique solution $(\nu_t, \mu_t)_{t\geq 0}$.

Theorem 2.2. *Suppose that Assumption 2, 3 and condition* (13) *from Assumption 4 hold. Then the system of equations* (17) *has a unique solution* $(\nu_t, \mu_t)_{t\geq 0}$. *Moreover, for* $t \geq 0$,

$$\mathrm{D_{KL}}(\nu_t|\pi) \leq 2\log R_\nu + \frac{4}{\sigma^2}C_\nu, \quad \mathrm{D_{KL}}(\mu_t|\rho) \leq 2\log R_\mu + \frac{4}{\sigma^2}C_\mu \tag{18}$$

and there exist constants $R_{1,\nu}, R_{1,\mu} > 1$ *such that for all* $t \geq 0$,

$$\sup_{x\in\mathcal{X}} \frac{\nu_t(x)}{\pi(x)} \leq R_{1,\nu}, \quad \sup_{y\in\mathcal{Y}} \frac{\mu_t(y)}{\rho(y)} \leq R_{1,\mu}. \tag{19}$$

Additionally, if condition (12) *from Assumption 4 holds, then there exist constants* $r_{1,\nu}$, $r_{1,\mu} > 0$ *such that for all* $t \geq 0$,

$$\inf_{x\in\mathcal{X}} \frac{\nu_t(x)}{\pi(x)} \geq r_{1,\nu}, \quad \inf_{y\in\mathcal{Y}} \frac{\mu_t(y)}{\rho(y)} \geq r_{1,\mu}. \tag{20}$$

The bounds obtained in this theorem allow us to prove that the map $t \mapsto \mathrm{D_{KL}}(\nu_\sigma^*|\nu_t) + \mathrm{D_{KL}}(\mu_\sigma^*|\mu_t)$ is differentiable along the FR dynamics (17).

Next, we state the other main result of this paper. We prove two different types of convergence results for the FR dynamics (17) in min-max games given by (1), one in terms of the players' strategies and the other in terms of the payoff function. Whereas the proof of the existence of the flow (Theorem 2.2) follows a

route similar to Liu et al. (2023), the proof of convergence in Theorem 2.3 is significantly different from Liu et al. (2023) since, as indicated in the introduction, in the present paper, convergence has to be studied with respect to appropriately chosen Lyapunov functions and, moreover, we do not rely on the Polyak-Łojasiewicz inequality.

Theorem 2.3. *Suppose that Assumption 2, 3 and 4 hold and let* $(\nu, \mu) \in \mathcal{P}_{ac}(\mathcal{X}) \times \mathcal{P}_{ac}(\mathcal{Y})$. *Then the map* $t \mapsto \mathrm{D_{KL}}(\nu|\nu_t) + \mathrm{D_{KL}}(\mu|\mu_t)$ *is differentiable along the FR dynamics* (17). *Suppose furthermore that Assumption 1 holds and that* (ν_0, μ_0) *are chosen such that* $\max_{\nu \in \mathcal{P}_{ac}(\mathcal{X})} \mathrm{D_{KL}}(\nu|\nu_0) + \max_{\mu \in \mathcal{P}_{ac}(\mathcal{Y})} \mathrm{D_{KL}}(\mu|\mu_0) < \infty$. *Then, for all* $t \geq 0$, *we have*

$$\mathrm{D_{KL}}(\nu_\sigma^*|\nu_t) + \mathrm{D_{KL}}(\mu_\sigma^*|\mu_t) \leq e^{-\frac{\sigma^2}{2}t} \left(\mathrm{D_{KL}}(\nu_\sigma^*|\nu_0) + \mathrm{D_{KL}}(\mu_\sigma^*|\mu_0) \right),$$

$$\mathrm{NI}\left(\frac{1}{t}\int_0^t \nu_s \mathrm{d}s, \frac{1}{t}\int_0^t \mu_s \mathrm{d}s \right) \leq \frac{1}{t} \left(\max_{\nu \in \mathcal{P}_{ac}(\mathcal{X})} \mathrm{D_{KL}}(\nu|\nu_0) + \max_{\mu \in \mathcal{P}_{ac}(\mathcal{Y})} \mathrm{D_{KL}}(\mu|\mu_0) \right). \tag{21}$$

In game theoretic language, the first result of Theorem 2.3 says that convergence of the FR dynamics (17) to the unique MNE $(\nu_\sigma^*, \mu_\sigma^*)$ in terms of the strategies ν_t and μ_t is achieved with exponential rate depending on σ. On the other hand, the second result gives a convergence estimate in terms of the NI error, which in turn is expressed as a function of the payoff V^σ. More precisely, the NI error of the time-averaged flow converges to 0 with rate $\mathcal{O}(1/t)$ regardless of the value of σ.

Remark 2.4. *Exponential convergence of the single mean-field birth-death flow* $(\nu_t)_{t \geq 0}$ *with respect to* $\mathrm{D_{KL}}(\nu_\sigma^*|\nu_t)$ *can be shown to also hold in the setting of Liu et al. (2023), however it was not studied in Liu et al. (2023), which considered only convergence of* $V^\sigma(\nu_t)$ *for a convex energy function* $V^\sigma : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$.

Remark 2.5. *The FR gradient flow can also be interpreted as the continuous-time limit of mirror descent (MD) algorithm with* $\mathrm{D_{KL}}$ *divergence over the space of probability measures (see, e.g., Liu et al. (2023)). In the context of a minimization problem, assuming convexity of the objective function, Aubin-Frankowski et al. (2022, Theorem* 4*) proves that the MD algorithm converges with rate* $\mathcal{O}\left(\frac{1}{N}\right)$ *to the minimizer, where* N *is the algorithm's total number of iterations.*

It is crucial to emphasize that the rate $\mathcal{O}\left(\frac{1}{t}\right)$ *obtained in Theorem 2.3 is not immediately implicit from the optimization result. Although one might expect that the mirror descent-ascent (MDA) algorithm for a min-max game with convex-concave payoff function converges with rate* $\mathcal{O}\left(\frac{1}{N}\right)$, *this is not the case because, as stressed in Lascu et al. (2024), the dynamics of the game are greatly influenced by the order the players move: either simultaneously (players move at the same time) or sequentially (each player moves upon observing the opponents' moves). The continuous-time analysis does not capture these two situations. Consequently, Lascu et al. (2024, Theorem* 2.6, 2.16*) show that the convergence rates of the simultaneous and sequential MDA schemes to mixed Nash equilibria of convex-concave min-max games, measured in the NI error, are* $\mathcal{O}\left(N^{-1/2}\right)$ *and* $\mathcal{O}\left(N^{-2/3}\right)$, *respectively.*

Remark 2.6. *It is worth commenting that the second result* (21) *of Theorem 2.3 can be obtained by only requiring* F *to be convex-concave (see Assumption 1) and setting* $\sigma = 0$, *i.e., the proof of* (21) *does not use the fact that* V^σ *is regularized by* $\mathrm{D_{KL}}$ *with a positive* σ. *Although we are presently unaware of a way to prove the existence of the FR flow* (17) *for* $\sigma = 0$ *using the Picard iteration technique we employed in Theorem 2.2, we believe that a possible way to prove the existence of the FR flow for* $\sigma = 0$ *could be via a minimizing movement scheme approach in the spirit of Gallouët & Monsaingeon (2017). Indeed, the FR gradient flow can be viewed as the zero-step-size limit of a proximal gradient algorithm under the Fisher-Rao distance (see equation* (4.5) *in Gallouët & Monsaingeon (2017)). Furthermore, for absolutely continuous measures, one could explicitly write the geodesics minimizing* (3) *(see Subsection* 2.2 *in Gallouët & Monsaingeon (2017)) along which the iterates generated by the proximal gradient algorithm can be interpolated.*

3 Proof of Theorem 2.3

Before we present the proof of Theorem 2.3, we begin with a technical lemma, which extends Liu et al. (2023, Lemma 2.5) to the min-max games setup, and which is proved in Section A. The proof of Theorem 2.3 is done in two steps:

- First, we show that, for fixed (ν, μ), the map $t \mapsto \mathrm{D_{KL}}(\nu|\nu_t) + \mathrm{D_{KL}}(\mu|\mu_t)$ is differentiable when $(\nu_t, \mu_t)_{t\geq 0}$ is defined as the FR flow (17).
- Second, we show that $\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D_{KL}}(\nu_\sigma^*|\nu_t) + \mathrm{D_{KL}}(\mu_\sigma^*|\mu_t)\right)$ is bounded from above by $-\frac{\sigma^2}{2}\left(\mathrm{D_{KL}}(\nu_\sigma^*|\nu_t) + \mathrm{D_{KL}}(\mu_\sigma^*|\mu_t)\right)$, and then we apply Gronwall's inequality to obtain exponential convergence. Subsequently, we establish linear convergence for $t \mapsto \mathrm{NI}\left(\frac{1}{t}\int_0^t \nu_s \mathrm{d}s, \frac{1}{t}\int_0^t \mu_s \mathrm{d}s\right)$.

Lemma 3.1 (Relative σ-strong-convexity-concavity to $\mathrm{D_{KL}}$)**.** *For V^σ given by* (1)*, if Assumption 1 holds, then V^σ satisfies the following inequalities for all $(\nu, \mu), (\nu', \mu') \in \mathcal{P}_{ac}(\mathcal{X}) \times \mathcal{P}_{ac}(\mathcal{Y})$:*

$$V^\sigma(\nu', \mu) - V^\sigma(\nu, \mu) \geq \int_{\mathcal{X}} a(\nu, \mu, x)(\nu' - \nu)(\mathrm{d}x) + \frac{\sigma^2}{2}\mathrm{D_{KL}}(\nu'|\nu),$$

$$V^\sigma(\nu, \mu') - V^\sigma(\nu, \mu) \leq \int_{\mathcal{Y}} b(\nu, \mu, y)(\mu' - \mu)(\mathrm{d}y) - \frac{\sigma^2}{2}\mathrm{D_{KL}}(\mu'|\mu).$$

Proof of Theorem 2.3. Step 1: Differentiability of $\mathrm{D_{KL}}$ *with respect to the FR flow* (17)*:* Suppose that Assumption 2, 3 and 4 hold. In order to show the differentiability of $t \mapsto \mathrm{D_{KL}}(\nu|\nu_t)$ for fixed $\nu \in \mathcal{P}_{\mathrm{ac}}(\mathcal{X})$ with respect to (17), it suffices to show that there exists an integrable function $f : \mathcal{X} \to \mathbb{R}$ such that

$$\left|\partial_t\left(\nu(x)\log\frac{\nu(x)}{\nu_t(x)}\right)\right| \leq f(x),$$

for all $t \geq 0$. Indeed, using (17) and (33), we have that

$$\begin{aligned}\left|\partial_t\left(\nu(x)\log\frac{\nu(x)}{\nu_t(x)}\right)\right| = \left|\nu(x)\frac{\partial_t\nu_t(x)}{\nu_t(x)}\right| &= \left|\nu(x)\left(\frac{\delta F}{\delta\nu}(\nu_t, \mu_t, x) + \frac{\sigma^2}{2}\log\left(\frac{\nu_t(x)}{\pi(x)}\right) - \frac{\sigma^2}{2}\mathrm{D_{KL}}(\nu_t|\pi)\right)\right| \\ &\leq \left(3C_\nu + \frac{\sigma^2}{2}\left(\max\{|\log r_{1,\nu}|, \log R_{1,\nu}\} + 2\log R_\nu\right)\right)\nu(x) \coloneqq f(x).\end{aligned}$$

An identical argument gives the differentiability of $t \mapsto \mathrm{D_{KL}}(\mu|\mu_t)$ for fixed $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathcal{Y})$. Then, we have that the map $t \mapsto \mathrm{D_{KL}}(\nu|\nu_t) + \mathrm{D_{KL}}(\mu|\mu_t)$ is differentiable.

Step 2: Convergence of the FR flow: Since $t \mapsto \mathrm{D_{KL}}(\nu|\nu_t) + \mathrm{D_{KL}}(\mu|\mu_t)$ is differentiable, we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D_{KL}}(\nu|\nu_t) + \mathrm{D_{KL}}(\mu|\mu_t)\right) &= \int_{\mathcal{X}}\partial_t\left(\nu(x)\log\frac{\nu(x)}{\nu_t(x)}\right)\mathrm{d}x + \int_{\mathcal{Y}}\partial_t\left(\mu(y)\log\frac{\mu(y)}{\mu_t(y)}\right)\mathrm{d}y \\ &= -\int_{\mathcal{X}}\left(\nu(x) - \nu_t(x)\right)\frac{\partial_t\nu_t(x)}{\nu_t(x)}\mathrm{d}x - \int_{\mathcal{Y}}\left(\mu(y) - \mu_t(y)\right)\frac{\partial_t\mu_t(y)}{\mu_t(y)}\mathrm{d}y \\ &= \int_{\mathcal{X}} a(\nu_t, \mu_t, x)(\nu - \nu_t)(\mathrm{d}x) - \int_{\mathcal{Y}} b(\nu_t, \mu_t, y)(\mu - \mu_t)(\mathrm{d}y),\end{aligned}$$

where in the second equality we used the fact that $\int_{\mathcal{X}}\partial_t\nu_t(x)\mathrm{d}x = \int_{\mathcal{Y}}\partial_t\mu_t(y)\mathrm{d}y = 0$.

If $\sigma > 0$ and Assumption 1 holds then, using Lemma 3.1 in the equation above, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D_{KL}}(\nu|\nu_t) + \mathrm{D_{KL}}(\mu|\mu_t)\right) \leq V^\sigma(\nu, \mu_t) - V^\sigma(\nu_t, \mu) - \frac{\sigma^2}{2}\mathrm{D_{KL}}(\nu|\nu_t) - \frac{\sigma^2}{2}\mathrm{D_{KL}}(\mu|\mu_t). \tag{22}$$

Setting $(\nu, \mu) = (\nu_\sigma^*, \mu_\sigma^*)$ in (22) and using the saddle point condition (2) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D_{KL}}(\nu_\sigma^*|\nu_t) + \mathrm{D_{KL}}(\mu_\sigma^*|\mu_t)\right) \leq -\frac{\sigma^2}{2}\mathrm{D_{KL}}(\nu_\sigma^*|\nu_t) - \frac{\sigma^2}{2}\mathrm{D_{KL}}(\mu_\sigma^*|\mu_t).$$

Hence, by Gronwall's inequality, we obtain that

$$\mathrm{D_{KL}}(\nu_\sigma^*|\nu_t) + \mathrm{D_{KL}}(\mu_\sigma^*|\mu_t) \leq e^{-\frac{\sigma^2}{2}t}\left(\mathrm{D_{KL}}(\nu_\sigma^*|\nu_0) + \mathrm{D_{KL}}(\mu_\sigma^*|\mu_0)\right).$$

On the other hand, since $\frac{\sigma^2}{2}\left(\mathrm{D_{KL}}(\nu|\nu_t)+\mathrm{D_{KL}}(\mu|\mu_t)\right)\geq 0$, for all (ν,μ), (22) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D_{KL}}(\nu|\nu_t)+\mathrm{D_{KL}}(\mu|\mu_t)\right)\leq V^\sigma(\nu,\mu_t)-V^\sigma(\nu_t,\mu).$$

Hence, integrating this inequality from 0 to $t>0$ and dividing by t, it follows that

$$\begin{aligned}\frac{1}{t}\int_0^t\left(V^\sigma(\nu_s,\mu)-V^\sigma(\nu,\mu_s)\right)\mathrm{d}s &\leq \frac{1}{t}\left(\mathrm{D_{KL}}(\nu|\nu_0)+\mathrm{D_{KL}}(\mu|\mu_0)-\mathrm{D_{KL}}(\nu|\nu_t)-\mathrm{D_{KL}}(\mu|\mu_t)\right)\\ &\leq \frac{1}{t}\left(\mathrm{D_{KL}}(\nu|\nu_0)+\mathrm{D_{KL}}(\mu|\mu_0)\right). \quad (23)\end{aligned}$$

Then, using the fact that $\nu\mapsto V^\sigma(\nu,\mu)$ is convex and $\mu\mapsto V^\sigma(\nu,\mu)$ is concave, it follows from Jensen's inequality that

$$\frac{1}{t}\int_0^t V^\sigma(\nu_s,\mu)\mathrm{d}s-\frac{1}{t}\int_0^t V^\sigma(\nu,\mu_s)\mathrm{d}s\geq V^\sigma\left(\frac{1}{t}\int_0^t\nu_s\mathrm{d}s,\mu\right)-V^\sigma\left(\nu,\frac{1}{t}\int_0^t\mu_s\mathrm{d}s\right). \quad (24)$$

Combining (23) with (24) and taking maximum over (ν,μ), we obtain

$$\mathrm{NI}\left(\frac{1}{t}\int_0^t\nu_s\mathrm{d}s,\frac{1}{t}\int_0^t\mu_s\mathrm{d}s\right)\leq\frac{1}{t}\left(\max_\nu\mathrm{D_{KL}}(\nu|\nu_0)+\max_\mu\mathrm{D_{KL}}(\mu|\mu_0)\right).$$

□

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17263–17275. Curran Associates, Inc., 2022.

René A. Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*. Springer International Publishing, 2018.

Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, and Joan Bruna. A mean-field analysis of two-player zero-sum games. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20215–20226. Curran Associates, Inc., 2020.

Thomas O. Gallouët and Léonard Monsaingeon. A JKO splitting scheme for Kantorovich-Fisher-Rao gradient flows. *SIAM J. Math. Anal.*, 49(2):1100–1130, 2017. ISSN 0036-1410. doi: 10.1137/16M106666X.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed Nash equilibria of generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2810–2819. PMLR, 09–15 Jun 2019.

Juno Kim, Kakei Yamamoto, Kazusato Oko, Zhuoran Yang, and Taiji Suzuki. Symmetric mean-field langevin dynamics for distributional minimax problems. In *The Twelfth International Conference on Learning Representations*, 2024.

Razvan-Andrei Lascu, Mateusz B. Majka, and Łukasz Szpruch. Entropic mean-field min-max problems via best response flow, 2023. arXiv:2306.03033.

Razvan-Andrei Lascu, Mateusz B. Majka, and Łukasz Szpruch. Mirror descent-ascent for mean-field min-max problems, 2024. arXiv:2402.08106.

Linshan Liu, Mateusz B. Majka, and Łukasz Szpruch. Polyak–łojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes. *Applied Mathematics and Optimization*, 87(3), March 2023. ISSN 0095-4616. doi: 10.1007/s00245-022-09962-0.

Yulong Lu. Two-scale gradient descent ascent dynamics finds mixed nash equilibria of continuous games: A mean-field perspective. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating Langevin Sampling with Birth-death, May 2019. arXiv:1905.09863.

Yulong Lu, Dejan Slepčev, and Lihan Wang. Birth–death dynamics for sampling: global convergence, approximations and their asymptotics. *Nonlinearity*, 36:5731–5772, 09 2023.

Chao Ma and Lexing Ying. Provably convergent quasistatic dynamics for mean-field two-player zero-sum games. In *International Conference on Learning Representations*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951. ISSN 0003486X.

Hukukane Nikaidô and Kazuo Isoda. Note on non-cooperative convex game. *Pacific Journal of Mathematics*, 5:807–815, 1955.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5508–5517. PMLR, 09–15 Jun 2019.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Camilo Garcia Trillos and Nicolas Garcia Trillos. On adversarial robustness and the use of wasserstein ascent-descent dynamics to enforce it, 2023. arXiv:2301.03662.

Nicolás García Trillos, Bahram Hosseini, and Daniel Sanz-Alonso. From optimization to sampling through gradient flows. *Notices of the American Mathematical Society*, 2023.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN 9780387790527.

John von Neumann, Oskar Morgenstern, and Ariel Rubinstein. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 1944. ISBN 9780691130613.

David Šiška and Łukasz Szpruch. Gradient Flows for Regularized Stochastic Control Problems. *arXiv e-prints*, art. arXiv:2006.05956, June 2020.

Guillaume Wang and Lénaïc Chizat. An exponentially converging particle method for the mixed nash equilibrium of continuous games, 2023. arXiv:2211.01280.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms, 2021. arXiv:1911.10635.

A Technical results and proofs

In this section, we present the proofs of the remaining results formulated in Section 2 of the paper.

A.1 Proof of Lemma 3.1.

In this subsection, we present the proof of Lemma 3.1.

Proof of Lemma 3.1. Using (8) from Assumption 1, it follows that

$$\begin{aligned}
V^\sigma(\nu',\mu) - V^\sigma(\nu,\mu) &\geq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu,\mu,x)(\nu'-\nu)(\mathrm{d}x) + \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\nu'|\pi) - \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\nu|\pi) \\
&= \int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu,\mu,x) + \frac{\sigma^2}{2}\log\left(\frac{\nu(x)}{\pi(x)}\right)\right)(\nu'-\nu)(\mathrm{d}x) - \frac{\sigma^2}{2}\int_{\mathcal{X}} \log\left(\frac{\nu(x)}{\pi(x)}\right)(\nu'-\nu)(\mathrm{d}x) \\
&\qquad + \frac{\sigma^2}{2}\int_{\mathcal{X}} \log\left(\frac{\nu'(x)}{\pi(x)}\right)\nu'(\mathrm{d}x) - \frac{\sigma^2}{2}\int_{\mathcal{X}} \log\left(\frac{\nu(x)}{\pi(x)}\right)\nu(\mathrm{d}x) \\
&= \int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu,\mu,x) + \frac{\sigma^2}{2}\log\left(\frac{\nu(x)}{\pi(x)}\right)\right)(\nu'-\nu)(\mathrm{d}x) + \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\nu'|\nu) \\
&= \int_{\mathcal{X}} a(\nu,\mu,x)(\nu'-\nu)(\mathrm{d}x) + \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\nu'|\nu).
\end{aligned}$$

Similarly, using (9) from Assumption 1, it follows that

$$\begin{aligned}
V^\sigma(\nu,\mu') - V^\sigma(\nu,\mu) &\leq \int_{\mathcal{Y}} \frac{\delta F}{\delta \mu}(\nu,\mu,y)(\mu'-\mu)(\mathrm{d}y) - \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\mu'|\rho) + \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\mu|\rho) \\
&= \int_{\mathcal{Y}} \left(\frac{\delta F}{\delta \mu}(\nu,\mu,y) - \frac{\sigma^2}{2}\log\left(\frac{\mu(y)}{\rho(y)}\right)\right)(\mu'-\mu)(\mathrm{d}y) + \frac{\sigma^2}{2}\int_{\mathcal{Y}} \log\left(\frac{\mu(y)}{\rho(y)}\right)(\mu'-\mu)(\mathrm{d}y) \\
&\qquad - \frac{\sigma^2}{2}\int_{\mathcal{Y}} \log\left(\frac{\mu'(y)}{\rho(y)}\right)\mu'(\mathrm{d}y) + \frac{\sigma^2}{2}\int_{\mathcal{Y}} \log\left(\frac{\mu(y)}{\rho(y)}\right)\mu(\mathrm{d}y) \\
&= \int_{\mathcal{Y}} \left(\frac{\delta F}{\delta \mu}(\nu,\mu,y) - \frac{\sigma^2}{2}\log\left(\frac{\mu(y)}{\rho(y)}\right)\right)(\mu'-\mu)(\mathrm{d}y) - \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\mu'|\mu) \\
&= \int_{\mathcal{Y}} b(\nu,\mu,y)(\mu'-\mu)(\mathrm{d}y) - \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\mu'|\mu).
\end{aligned}$$

□

A.2 Existence and uniqueness of the FR flow

In this subsection, we present the proof of our main result concerning the existence and uniqueness of the Fisher-Rao (FR) flow, i.e., Theorem 2.2. We construct a Picard iteration which is proved to be well-defined in Lemma A.1. Lemma A.2 shows that the Picard iteration is contractive in an appropriate metric. Then in order to conclude the proof of Theorem 2.2 we show the ratio condition (19).

The proof of Theorem 2.2 is an adaptation of the proof of Liu et al. (2023, Theorem 2.1) to the min-max setting (1).

Proof of Theorem 2.2. Step 1: Existence of the gradient flow and bound (18) *on* $[0,T]$. In order to prove the existence of a solution $(\nu_t,\mu_t)_{t\geq 0}$ to

$$\begin{cases}
\partial_t \nu_t(x) = -\left(\frac{\delta F}{\delta \nu}(\nu_t,\mu_t,x) + \frac{\sigma^2}{2}\log\left(\frac{\nu_t(x)}{\pi(x)}\right) - \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\nu_t|\pi)\right)\nu_t(x), \\
\partial_t \mu_t(y) = \left(\frac{\delta F}{\delta \mu}(\nu_t,\mu_t,y) - \frac{\sigma^2}{2}\log\left(\frac{\mu_t(y)}{\rho(y)}\right) + \frac{\sigma^2}{2}\,\mathrm{D}_{\mathrm{KL}}(\mu_t|\rho)\right)\mu_t(y),
\end{cases} \tag{25}$$

we first notice that (25) is equivalent to

$$\begin{cases} \partial_t \log \nu_t(x) = -\left(\frac{\delta F}{\delta \nu}(\nu_t, \mu_t, x) + \frac{\sigma^2}{2} \log\left(\frac{\nu_t(x)}{\pi(x)} \right) - \frac{\sigma^2}{2} \mathrm{D}_{\mathrm{KL}}(\nu_t | \pi) \right), \\ \partial_t \log \mu_t(y) = \left(\frac{\delta F}{\delta \mu}(\nu_t, \mu_t, y) - \frac{\sigma^2}{2} \log\left(\frac{\mu_t(y)}{\rho(y)} \right) + \frac{\sigma^2}{2} \mathrm{D}_{\mathrm{KL}}(\mu_t | \rho) \right). \end{cases} \tag{26}$$

By Duhamel's formula, (26) is equivalent to

$$\begin{cases} \log \nu_t(x) = e^{-\frac{\sigma^2}{2}t} \log \nu_0(x) - \int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \left(\frac{2}{\sigma^2} \frac{\delta F}{\delta \nu}(\nu_s, \mu_s, x) - \log \pi(x) - \mathrm{D}_{\mathrm{KL}}(\nu_s | \pi) \right) \mathrm{d}s, \\ \log \mu_t(y) = e^{-\frac{\sigma^2}{2}t} \log \mu_0(x) + \int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \left(\frac{2}{\sigma^2} \frac{\delta F}{\delta \nu}(\nu_s, \mu_s, y) + \log \rho(y) + \mathrm{D}_{\mathrm{KL}}(\mu_s | \rho) \right) \mathrm{d}s. \end{cases}$$

Based on these formulas, we will define a Picard iteration scheme. To this end, let us first fix $T > 0$ and choose a pair of flows of probability measures $(\nu_t^{(0)}, \mu_t^{(0)})_{t \in [0,T]}$ such that

$$\int_0^T \mathrm{D}_{\mathrm{KL}}(\nu_s^{(0)} | \pi) \mathrm{d}s < \infty, \quad \int_0^T \mathrm{D}_{\mathrm{KL}}(\mu_s^{(0)} | \rho) \mathrm{d}s < \infty.$$

For each $n \geq 1$, we fix $\nu_0^{(n)} = \nu_0^{(0)} = \nu_0$ and $\mu_0^{(n)} = \mu_0^{(0)} = \mu_0$ (with ν_0 and μ_0 satisfying condition (13) from Assumption 4) and define $(\nu_t^{(n)}, \mu_t^{(n)})_{t \in [0,T]}$ by

$$\begin{aligned} \log \nu_t^{(n)}(x) &= e^{-\frac{\sigma^2}{2}t} \log \nu_0(x) \\ &\quad - \int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \left(\frac{2}{\sigma^2} \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x) - \log \pi(x) - \mathrm{D}_{\mathrm{KL}}(\nu_s^{(n-1)} | \pi) \right) \mathrm{d}s, \end{aligned} \tag{27}$$

$$\begin{aligned} \log \mu_t^{(n)}(y) &= e^{-\frac{\sigma^2}{2}t} \log \mu_0(y) \\ &\quad + \int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \left(\frac{2}{\sigma^2} \frac{\delta F}{\delta \mu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, y) + \log \rho(y) + \mathrm{D}_{\mathrm{KL}}(\mu_s^{(n-1)} | \rho) \right) \mathrm{d}s. \end{aligned} \tag{28}$$

We have the following result.

Lemma A.1. *The sequence of flows* $\left((\nu_t^{(n)}, \mu_t^{(n)})_{t \in [0,T]} \right)_{n=0}^{\infty}$ *given by* (27) *and* (28) *is well-defined and such that for all* $n \geq 1$ *and all* $t \in [0, T]$ *we have*

$$\mathrm{D}_{\mathrm{KL}}(\nu_t^{(n)} | \pi) \leq 2 \log R_\nu + \frac{4}{\sigma^2} C_\nu, \quad \mathrm{D}_{\mathrm{KL}}(\mu_t^{(n)} | \rho) \leq 2 \log R_\mu + \frac{4}{\sigma^2} C_\mu.$$

Proof of Lemma A.1. The proof follows from the same induction argument used to prove Liu et al. (2023, Lemma 3.1). □

For fixed $T > 0$, we consider the sequence of flows $\left((\nu_t^{(n)}, \mu_t^{(n)})_{t \in [0,T]} \right)_{n=0}^{\infty}$ in $\left(\mathcal{P}(\mathcal{X})^{[0,T]} \times \mathcal{P}(\mathcal{Y})^{[0,T]}, \mathcal{TV}^{[0,T]} \right)$, where, for any $(\nu_t, \mu_t)_{t \in [0,T]} \in \mathcal{P}(\mathcal{X})^{[0,T]} \times \mathcal{P}(\mathcal{Y})^{[0,T]}$, the distance $\mathcal{TV}^{[0,T]}$ is defined by

$$\mathcal{TV}^{[0,T]}\left((\nu_t, \mu_t)_{t \in [0,T]}, (\nu_t', \mu_t')_{t \in [0,T]} \right) := \int_0^T \mathrm{TV}(\nu_t, \nu_t') \mathrm{d}t + \int_0^T \mathrm{TV}(\mu_t, \mu_t') \mathrm{d}t \,.$$

Since $(\mathcal{P}(\mathcal{X}), \mathrm{TV})$ is complete, we can apply the argument from Šiška & Szpruch (2020, Lemma A.5) with $p = 1$ to conclude that $\left(\mathcal{P}(\mathcal{X})^{[0,T]}, \int_0^T \mathrm{TV}(\nu_t, \nu_t') \mathrm{d}t \right)$ and $\left(\mathcal{P}(\mathcal{Y})^{[0,T]}, \int_0^T \mathrm{TV}(\mu_t, \mu_t') \mathrm{d}t \right)$ are complete. Therefore, one can deduce that $\left(\mathcal{P}(\mathcal{X})^{[0,T]} \times \mathcal{P}(\mathcal{Y})^{[0,T]}, \mathcal{TV}^{[0,T]} \right)$ is also complete. We consider the Picard iteration mapping $\phi\left((\nu_t^{(n-1)}, \mu_t^{(n-1)})_{t \in [0,T]} \right) := (\nu_t^{(n)}, \mu_t^{(n)})_{t \in [0,T]}$ defined via (27) and (28) and show that ϕ is contractive in $\left(\mathcal{P}(\mathcal{X})^{[0,T]} \times \mathcal{P}(\mathcal{Y})^{[0,T]}, \mathcal{TV}^{[0,T]} \right)$. Then the Banach fixed point theorem will give us the existence and uniqueness of the solution to (17).

Lemma A.2. *The mapping* $\phi\left((\nu_t^{(n-1)}, \mu_t^{(n-1)})_{t\in[0,T]}\right) := (\nu_t^{(n)}, \mu_t^{(n)})_{t\in[0,T]}$ *defined via* (27) *and* (28) *is contractive in* $\left(\mathcal{P}(\mathcal{X})^{[0,T]} \times \mathcal{P}(\mathcal{Y})^{[0,T]}, \mathcal{TV}^{[0,T]}\right)$.

Proof of Lemma A.2. From (27), we have

$$\log \nu_t^{(n)}(x) - \log \nu_t^{(n-1)}(x) = -\int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \times \\ \times \left[\frac{2}{\sigma^2}\left(\frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-2)}, \mu_s^{(n-2)}, x)\right) - \mathrm{D_{KL}}(\nu_s^{(n-1)}|\pi) + \mathrm{D_{KL}}(\nu_s^{(n-2)}|\pi)\right] \mathrm{d}s\,.$$

Multiplying both sides by $\nu_t^{(n)}(x)$ and integrating with respect to x, we obtain

$$\mathrm{D_{KL}}(\nu_t^{(n)}|\nu_t^{(n-1)}) = -\int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \left[\frac{2}{\sigma^2}\int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-2)}, \mu_s^{(n-2)}, x)\right) \nu_t^{(n)}(\mathrm{d}x) \right. \\ \left. - \mathrm{D_{KL}}(\nu_s^{(n-1)}|\pi) + \mathrm{D_{KL}}(\nu_s^{(n-2)}|\pi)\right] \mathrm{d}s\,. \quad (29)$$

Moreover, note that

$$\int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-2)}, \mu_s^{(n-2)}, x)\right) \nu_t^{(n)}(\mathrm{d}x) \\ = \int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-2)}, x) + \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-2)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-2)}, \mu_s^{(n-2)}, x)\right) \nu_t^{(n)}(\mathrm{d}x) \\ = \int_{\mathcal{X}}\int_{\mathcal{Y}}\int_0^1 \frac{\delta^2 F}{\delta \mu \delta \nu}\left(\nu_s^{(n-1)}, \mu_s^{(n-2)} + \lambda\left(\mu_s^{(n-1)} - \mu_s^{(n-2)}\right), x, w\right) \mathrm{d}\lambda \left(\mu_s^{(n-1)} - \mu_s^{(n-2)}\right)(\mathrm{d}w) \nu_t^{(n)}(\mathrm{d}x) \\ + \int_{\mathcal{X}}\int_{\mathcal{X}}\int_0^1 \frac{\delta^2 F}{\delta \nu^2}\left(\nu_s^{(n-2)} + \lambda\left(\nu_s^{(n-1)} - \nu_s^{(n-2)}\right), \mu_s^{(n-2)}, x, z\right) \mathrm{d}\lambda \left(\nu_s^{(n-1)} - \nu_s^{(n-2)}\right)(\mathrm{d}z) \nu_t^{(n)}(\mathrm{d}x).$$

Similarly, again from (27) we have

$$\log \nu_t^{(n-1)}(x) - \log \nu_t^{(n)}(x) = -\int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \times \\ \times \left[\frac{2}{\sigma^2}\int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu_s^{(n-2)}, \mu_s^{(n-2)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x)\right) \nu_t^{(n)}(\mathrm{d}x) \right. \\ \left. - \mathrm{D_{KL}}(\nu_s^{(n-2)}|\pi) + \mathrm{D_{KL}}(\nu_s^{(n-1)}|\pi)\right] \mathrm{d}s.$$

Multiplying both sides by $\nu_t^{(n-1)}(x)$ and integrating with respect to x, we obtain

$$\mathrm{D_{KL}}(\nu_t^{(n-1)}|\nu_t^{(n)}) = -\int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \left[\frac{2}{\sigma^2}\int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu_s^{(n-2)}, \mu_s^{(n-2)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x)\right) \nu_t^{(n-1)}(\mathrm{d}x) \right. \\ \left. - \mathrm{D_{KL}}(\nu_s^{(n-2)}|\pi) + \mathrm{D_{KL}}(\nu_s^{(n-1)}|\pi)\right] \mathrm{d}s. \quad (30)$$

Similarly as before, we note that

$$\begin{aligned}
&\int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu_s^{(n-2)}, \mu_s^{(n-2)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x) \right) \nu_t^{(n-1)}(\mathrm{d}x) \\
&= -\int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-2)}, x) + \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-2)}, x) - \frac{\delta F}{\delta \nu}(\nu_s^{(n-2)}, \mu_s^{(n-2)}, x) \right) \nu_t^{(n-1)}(\mathrm{d}x) \\
&= -\int_{\mathcal{X}} \int_{\mathcal{Y}} \int_0^1 \frac{\delta^2 F}{\delta \mu \delta \nu} \left(\nu_s^{(n-1)}, \mu_s^{(n-2)} + \lambda \left(\mu_s^{(n-1)} - \mu_s^{(n-2)} \right), x, w \right) \times \mathrm{d}\lambda \left(\mu_s^{(n-1)} - \mu_s^{(n-2)} \right) (\mathrm{d}w) \nu_t^{(n-1)}(\mathrm{d}x) \\
&\quad - \int_{\mathcal{X}} \int_{\mathcal{X}} \int_0^1 \frac{\delta^2 F}{\delta \nu^2} \left(\nu_s^{(n-2)} + \lambda \left(\nu_s^{(n-1)} - \nu_s^{(n-2)} \right), \mu_s^{(n-2)}, x, z \right) \times \mathrm{d}\lambda \left(\nu_s^{(n-1)} - \nu_s^{(n-2)} \right) (\mathrm{d}z) \nu_t^{(n-1)}(\mathrm{d}x).
\end{aligned}$$

Combining (29) and (30), we obtain

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(\nu_t^{(n)} | \nu_t^{(n-1)}) + \mathrm{D}_{\mathrm{KL}}(\nu_t^{(n-1)} | \nu_t^{(n)}) = -\int_0^t e^{-\frac{\sigma^2}{2}(t-s)} \times \Bigg[\\
&\int_{\mathcal{X}} \int_{\mathcal{Y}} \int_0^1 \frac{\delta^2 F}{\delta \mu \delta \nu} \left(\nu_s^{(n-1)}, \mu_s^{(n-2)} + \lambda \left(\mu_s^{(n-1)} - \mu_s^{(n-2)} \right), x, w \right) \mathrm{d}\lambda \left(\mu_s^{(n-1)} - \mu_s^{(n-2)} \right) (\mathrm{d}w) \left(\nu_t^{(n)} - \nu_t^{(n-1)} \right) (\mathrm{d}x) \\
&+ \int_{\mathcal{X}} \int_{\mathcal{X}} \int_0^1 \frac{\delta^2 F}{\delta \nu^2} \left(\nu_s^{(n-2)} + \lambda \left(\nu_s^{(n-1)} - \nu_s^{(n-2)} \right), \mu_s^{(n-2)}, x, z \right) \mathrm{d}\lambda \left(\nu_s^{(n-1)} - \nu_s^{(n-2)} \right) (\mathrm{d}z) \left(\nu_t^{(n)} - \nu_t^{(n-1)} \right) (\mathrm{d}x) \Bigg] \mathrm{d}s.
\end{aligned}$$

Hence, due to Assumption 3, we get

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(\nu_t^{(n)} | \nu_t^{(n-1)}) + \mathrm{D}_{\mathrm{KL}}(\nu_t^{(n-1)} | \nu_t^{(n)}) \\
&\qquad \leq \mathrm{TV}(\nu_t^{(n)}, \nu_t^{(n-1)}) \int_0^t e^{-\frac{\sigma^2}{2}(t-s)} \left(C_{\mu,\nu} \, \mathrm{TV}(\mu_s^{(n-1)}, \mu_s^{(n-2)}) + C_{\nu,\nu} \, \mathrm{TV}(\nu_s^{(n-1)}, \nu_s^{(n-2)}) \right) \mathrm{d}s \\
&\qquad \leq \max\{C_{\mu,\nu}, C_{\nu,\nu}\} \, \mathrm{TV}(\nu_t^{(n)}, \nu_t^{(n-1)}) \int_0^t e^{-\frac{\sigma^2}{2}(t-s)} \left(\mathrm{TV}(\mu_s^{(n-1)}, \mu_s^{(n-2)}) + \mathrm{TV}(\nu_s^{(n-1)}, \nu_s^{(n-2)}) \right) \mathrm{d}s.
\end{aligned}$$

By the Pinsker-Csizsar inequality, $\mathrm{TV}^2(\nu_t^{(n)}, \nu_t^{(n-1)}) \leq \frac{1}{2} \mathrm{D}_{\mathrm{KL}}(\nu_t^{(n)} | \nu_t^{(n-1)})$, and hence

$$\begin{aligned}
4 \, \mathrm{TV}^2(\nu_t^{(n)}, \nu_t^{(n-1)}) \leq \max\{C_{\mu,\nu}, C_{\nu,\nu}\} \, \mathrm{TV}(\nu_t^{(n)}, \nu_t^{(n-1)}) \int_0^t e^{-\frac{\sigma^2}{2}(t-s)} \Big(& \mathrm{TV}(\mu_s^{(n-1)}, \mu_s^{(n-2)}) \\
& + \mathrm{TV}(\nu_s^{(n-1)}, \nu_s^{(n-2)}) \Big) \mathrm{d}s,
\end{aligned}$$

which gives

$$\mathrm{TV}(\nu_t^{(n)}, \nu_t^{(n-1)}) \leq \frac{1}{4} \max\{C_{\mu,\nu}, C_{\nu,\nu}\} \int_0^t e^{-\frac{\sigma^2}{2}(t-s)} \Big(\mathrm{TV}(\mu_s^{(n-1)}, \mu_s^{(n-2)}) + \mathrm{TV}(\nu_s^{(n-1)}, \nu_s^{(n-2)}) \Big) \mathrm{d}s.$$

An almost identical argument leads to

$$\mathrm{TV}(\mu_t^{(n)}, \mu_t^{(n-1)}) \leq \frac{1}{4} \max\{C_{\nu,\mu}, C_{\mu,\mu}\} \int_0^t e^{-\frac{\sigma^2}{2}(t-s)} \Big(\mathrm{TV}(\mu_s^{(n-1)}, \mu_s^{(n-2)}) + \mathrm{TV}(\nu_s^{(n-1)}, \nu_s^{(n-2)}) \Big) \mathrm{d}s.$$

If we set $C_{\max} \coloneqq \max\{C_{\mu,\nu}, C_{\nu,\mu}\} + \max\{C_{\nu,\mu}, C_{\mu,\mu}\}$, and add the previous two inequalities, we obtain

$$\begin{aligned}
&\mathrm{TV}(\nu_t^{(n)}, \nu_t^{(n-1)}) + \mathrm{TV}(\mu_t^{(n)}, \mu_t^{(n-1)}) \leq \frac{C_{\max}}{4} \int_0^t e^{-\frac{\sigma^2}{2}(t-s)} \Big(\mathrm{TV}(\mu_s^{(n-1)}, \mu_s^{(n-2)}) + \mathrm{TV}(\nu_s^{(n-1)}, \nu_s^{(n-2)}) \Big) \mathrm{d}s. \\
&\leq \left(\frac{C_{\max}}{4}\right)^{n-1} e^{-\frac{\sigma^2}{2}t} \int_0^t \int_0^{t_1} \dots \int_0^{t_{n-2}} e^{\frac{\sigma^2}{2}t_{n-1}} \Big(\mathrm{TV}(\nu_{t_{n-1}}^{(1)}, \nu_{t_{n-1}}^{(0)}) + \mathrm{TV}(\mu_{t_{n-1}}^{(1)}, \mu_{t_{n-1}}^{(0)}) \Big) \mathrm{d}t_{n-1} \dots \mathrm{d}t_2 \mathrm{d}t_1 \\
&\leq \left(\frac{C_{\max}}{4}\right)^{n-1} e^{-\frac{\sigma^2}{2}t} \frac{t^{n-2}}{(n-2)!} \int_0^t e^{\frac{\sigma^2}{2}t_{n-1}} \Big(\mathrm{TV}(\nu_{t_{n-1}}^{(1)}, \nu_{t_{n-1}}^{(0)}) + \mathrm{TV}(\mu_{t_{n-1}}^{(1)}, \mu_{t_{n-1}}^{(0)}) \Big) \mathrm{d}t_{n-1} \\
&\leq \left(\frac{C_{\max}}{4}\right)^{n-1} \frac{t^{n-2}}{(n-2)!} \int_0^t \Big(\mathrm{TV}(\nu_{t_{n-1}}^{(1)}, \nu_{t_{n-1}}^{(0)}) + \mathrm{TV}(\mu_{t_{n-1}}^{(1)}, \mu_{t_{n-1}}^{(0)}) \Big) \mathrm{d}t_{n-1},
\end{aligned}$$

where in the third inequality we bounded $\int_0^{t_{n-2}} dt_{n-1} \leq \int_0^t dt_{n-1}$ and in the fourth inequality we bounded $e^{\frac{\sigma^2}{2}t_{n-1}} \leq e^{\frac{\sigma^2}{2}t}$. Hence, we obtain

$$\begin{aligned}
&\int_0^T \mathrm{TV}(\nu_t^{(n)}, \nu_t^{(n-1)}) \mathrm{d}t + \int_0^T \mathrm{TV}(\mu_t^{(n)}, \mu_t^{(n-1)}) \mathrm{d}t \\
&\quad \leq \left(\frac{C_{\max}}{4}\right)^{n-1} \frac{T^{n-1}}{(n-1)!} \left(\int_0^T \mathrm{TV}(\nu_{t_{n-1}}^{(1)}, \nu_{t_{n-1}}^{(0)}) \mathrm{d}t_{n-1} + \int_0^T \mathrm{TV}(\mu_{t_{n-1}}^{(1)}, \mu_{t_{n-1}}^{(0)}) \mathrm{d}t_{n-1} \right).
\end{aligned}$$

For sufficiently large n, the constant on the right hand side becomes less than 1 and the proof is complete. □

By Lemma A.2, for any $T > 0$ we obtain the existence of a pair of flows $(\nu_t, \mu_t)_{t\in[0,T]}$ satisfying (25). Moreover, for Lebesgue-almost all $t \in [0, T]$ we have

$$\mathrm{TV}(\nu_t^{(n)}, \nu_t) \to 0, \quad \mathrm{TV}(\mu_t^{(n)}, \mu_t) \to 0 \qquad \text{as } n \to \infty\,,$$

which implies

$$\nu_t^{(n)} \to \nu_t, \quad \mu_t^{(n)} \to \mu_t \qquad \text{weakly as } n \to \infty\,.$$

Hence, using the lower semi-continuity of the entropy, we obtain

$$\mathrm{D_{KL}}(\nu_t|\pi) \leq \liminf_{n\to\infty} \mathrm{D_{KL}}(\nu_t^{(n)}|\pi) \leq 2\log R_\nu + \frac{4}{\sigma^2} C_\nu\,, \tag{31}$$

$$\mathrm{D_{KL}}(\mu_t|\rho) \leq \liminf_{n\to\infty} \mathrm{D_{KL}}(\mu_t^{(n)}|\rho) \leq 2\log R_\mu + \frac{4}{\sigma^2} C_\mu\,, \tag{32}$$

where both second inequalities follow from Lemma A.1. In order to ensure that the solution $(\nu_t, \mu_t)_{t\in[0,T]}$ can be extended to all $t \geq 0$, we first need to prove the bound on the ratios ν_t/π and μ_t/ρ in (19).

Step 2: Ratio condition (19). Using (27) and (28), we see that for any $t \in [0, T]$ we have

$$\log \frac{\nu_t^{(n)}(x)}{\pi(x)} = e^{-\frac{\sigma^2}{2}t} \log \frac{\nu_0(x)}{\pi(x)} - \int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \left(\frac{2}{\sigma^2} \frac{\delta F}{\delta \nu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, x) - \mathrm{D_{KL}}(\nu_s^{(n-1)}|\pi) \right) \mathrm{d}s,$$

$$\log \frac{\mu_t^{(n)}(y)}{\rho(y)} = e^{-\frac{\sigma^2}{2}t} \log \frac{\mu_0(y)}{\rho(y)} + \int_0^t \frac{\sigma^2}{2} e^{-\frac{\sigma^2}{2}(t-s)} \left(\frac{2}{\sigma^2} \frac{\delta F}{\delta \mu}(\nu_s^{(n-1)}, \mu_s^{(n-1)}, y) + \mathrm{D_{KL}}(\mu_s^{(n-1)}|\rho) \right) \mathrm{d}s.$$

Using Assumption 2, (13), (31) and (32) we obtain

$$\log \frac{\nu_t(x)}{\pi(x)} \leq \log R_\nu + C_\nu + \frac{\sigma^2}{2} \left(2\log R_\nu + \frac{4}{\sigma^2} C_\nu \right),$$

$$\log \frac{\mu_t(y)}{\rho(y)} \leq \log R_\mu + C_\mu + \frac{\sigma^2}{2} \left(2\log R_\mu + \frac{4}{\sigma^2} C_\mu \right).$$

Hence we can choose $R_{1,\nu} := 1 + \exp\left(\log R_\nu + C_\nu + \frac{\sigma^2}{2}\left(2\log R_\nu + \frac{4}{\sigma^2}C_\nu\right)\right)$ and $R_{1,\mu} := 1 + \exp\left(\log R_\mu + C_\mu + \frac{\sigma^2}{2}\left(2\log R_\mu + \frac{4}{\sigma^2}C_\mu\right)\right)$. Note $R_{1,\nu}, R_{1,\mu} > 1$ are conveniently chosen so that $\log R_{1,\nu}, \log R_{1,\mu} > 0$ in our subsequent calculations. Obtaining a lower bound on $\frac{\nu_t(x)}{\pi(x)}$ and $\frac{\mu_t(y)}{\rho(y)}$ follows similarly, by using (12) instead of (13).

Step 3: Existence of the gradient flow on $[0,\infty)$. In order to complete our proof, note that the unique solution $(\nu_t, \mu_t)_{t\in[0,T]}$ to (25) can also be expressed as

$$\nu_t(x) = \nu_0(x)\exp\left(-\int_0^t \left(\frac{\delta F}{\delta \nu}(\nu_s, \mu_s, x) + \frac{\sigma^2}{2}\log\left(\frac{\nu_s(x)}{\pi(x)}\right) - \frac{\sigma^2}{2}\,\mathrm{D_{KL}}(\nu_s|\pi)\right)\mathrm{d}s\right),$$

$$\mu_t(y) = \mu_0(y)\exp\left(\int_0^t \left(\frac{\delta F}{\delta \mu}(\nu_s, \mu_s, y) - \frac{\sigma^2}{2}\log\left(\frac{\mu_s(y)}{\rho(y)}\right) + \frac{\sigma^2}{2}\,\mathrm{D_{KL}}(\mu_s|\rho)\right)\mathrm{d}s\right).$$

From 2, (31), (32) and (19), we obtain for any $t \in [0,T]$

$$\left|\frac{\delta F}{\delta \nu}(\nu_t, \mu_t, x) + \frac{\sigma^2}{2}\log\left(\frac{\nu_t(x)}{\pi(x)}\right) - \frac{\sigma^2}{2}\,\mathrm{D_{KL}}(\nu_t|\pi)\right| \leq 3C_\nu + \frac{\sigma^2}{2}\left(\max\{|\log r_{1,\nu}|, \log R_{1,\nu}\} + 2\log R_\nu\right) =: C_{V,\nu}, \quad (33)$$

$$\left|\frac{\delta F}{\delta \mu}(\nu_t, \mu_t, y) - \frac{\sigma^2}{2}\log\left(\frac{\mu_t(y)}{\rho(y)}\right) + \frac{\sigma^2}{2}\,\mathrm{D_{KL}}(\mu_t|\rho)\right| \leq 3C_\mu + \frac{\sigma^2}{2}\left(\max\{|\log r_{1,\mu}|, \log R_{1,\mu}\} + 2\log R_\mu\right) =: C_{V,\mu}.$$

This gives $\|\nu_t\|_{TV} \leq \|\nu_0\|_{TV} e^{C_{V,\nu}t}$ and $\|\mu_t\|_{TV} \leq \|\mu_0\|_{TV} e^{C_{V,\mu}t}$, and shows that ν_t and μ_t do not explode in any finite time, hence we obtain a global solution $(\nu_t, \mu_t)_{t\in[0,\infty)}$. In particular, the bounds in (31), (32), (19) and (20) hold for all $t \geq 0$. □

B Notation and definitions

In this section we recall some important definitions. Following Carmona & Delarue (2018, Definition 5.43), we start with the notion of differentiability on the space of probability measure that we utilize throughout the paper.

Definition B.1. *Fix $p \geq 0$. For any $\mathcal{M} \subseteq \mathbb{R}^d$, let $\mathcal{P}_p(\mathcal{M})$ be the space of probability measures on $\mathcal{M}$ with finite p-th moments. A function $F : \mathcal{P}_p(\mathcal{M}) \to \mathbb{R}$ admits first-order flat derivative on $\mathcal{P}_p(\mathcal{M})$, if there exists a function $\frac{\delta F}{\delta \nu} : \mathcal{P}_p(\mathcal{M}) \times \mathcal{M} \to \mathbb{R}$, such that*

1. *the map $\mathcal{P}_p(\mathcal{M}) \times \mathcal{M} \ni (m, x) \mapsto \frac{\delta F}{\delta m}(m, x)$ is jointly continuous with respect to the product topology, where $\mathcal{P}_p(\mathcal{M})$ is endowed with the weak topology,*
2. *For any $m \in \mathcal{P}_p(\mathcal{M})$, there exists $C > 0$ such that, for all $x \in \mathcal{M}$, we have*

$$\left|\frac{\delta F}{\delta m}(m, x)\right| \leq C\left(1 + |x|^p\right),$$

3. *For all $m, m' \in \mathcal{P}_p(\mathcal{M})$, it holds that*

$$F(m') - F(m) = \int_0^1 \int_{\mathcal{M}} \frac{\delta F}{\delta m}(m + \varepsilon(m' - m), x)\,(m' - m)\,(\mathrm{d}x)\mathrm{d}\varepsilon. \quad (34)$$

The functional $\frac{\delta F}{\delta m}$ *is then called the flat derivative of* F *on* $\mathcal{P}_p(\mathcal{M})$*. We note that* $\frac{\delta F}{\delta m}$ *exists up to an additive constant, and thus we make the normalizing convention* $\int_{\mathcal{M}} \frac{\delta F}{\delta m}(m, x) m(\mathrm{d}x) = 0$.

If, for fixed $x \in \mathcal{M}$, the map $m \mapsto \frac{\delta F}{\delta m}(m, x)$ satisfies Definition B.1, we say that F admits a second-order flat derivative denoted by $\frac{\delta^2 F}{\delta m^2}$. Consequently, by Definition B.1, there exists a functional $\frac{\delta^2 F}{\delta m^2} : \mathcal{P}_p(\mathcal{M}) \times \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ such that

$$\frac{\delta F}{\delta m}(m', x) - \frac{\delta F}{\delta m}(m, x) = \int_0^1 \int_{\mathcal{M}} \frac{\delta^2 F}{\delta m^2}(\nu + \varepsilon(m' - m), x, x') \, (m' - m) \, (\mathrm{d}x') \mathrm{d}\varepsilon. \tag{35}$$

Definition B.2 (TV distance between probability measures; (Tsybakov, 2008), Definition 2.4)**.** *Let* $(\mathcal{M}, \mathcal{A})$ *be a measurable space and let* P *and* Q *be probability measures on* $(\mathcal{M}, \mathcal{A})$*. Assume that* μ *is a* σ*-finite measure on* $(\mathcal{M}, \mathcal{A})$ *such that* P *and* Q *are absolutely continuous with respect to* μ *and let* p *and* q *denote their probability density functions, respectively. The total variation distance between* P *and* Q *is defined as:*

$$\mathrm{TV}(P, Q) \coloneqq \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \sup_{A \in \mathcal{A}} \left| \int_A (p - q) \mathrm{d}\mu \right|.$$