A mean-field analysis of two-player zero-sum games

Carles Domingo-Enrich[a], Samy Jelassi[a,d], Arthur Mensch[e],
Grant M. Rotskoff[a], and Joan Bruna[a, b, c]

[a]Courant Institute of Mathematical Sciences, New York University, New York
[b]Center for Data Science, New York University, New York
[c]Institute for Advanced Study, Princeton
[d]Princeton University, Princeton
[e]ENS Paris, France

May 7, 2021

Abstract

Finding Nash equilibria in two-player zero-sum continuous games is a central problem in machine learning, e.g. for training both GANs and robust models. The existence of pure Nash equilibria requires strong conditions which are not typically met in practice. Mixed Nash equilibria exist in greater generality and may be found using mirror descent. Yet this approach does not scale to high dimensions. To address this limitation, we parametrize mixed strategies as mixtures of particles, whose positions and weights are updated using gradient descent-ascent. We study this dynamics as an interacting gradient flow over measure spaces endowed with the Wasserstein-Fisher-Rao metric. We establish global convergence to an approximate equilibrium for the related Langevin gradient-ascent dynamic. We prove a law of large numbers that relates particle dynamics to mean-field dynamics. Our method identifies mixed equilibria in high dimensions and is demonstrably effective for training mixtures of GANs.

1 Introduction

Multi-objective optimization problems arise in many fields, from economics to civil engineering. Tasks that require optimizing multiple objectives have also become a routine part of many agent-based machine learning algorithms including generative adversarial networks [Goodfellow et al., 2014], imaginative agents [Racanière et al., 2017], hierarchical reinforcement learning [Wayne and Abbott, 2014] and multi-agent reinforcement learning [Bu et al., 2008]. It not only remains difficult to carry out the necessary optimization, but also to assess the optimality of a given solution.

Multi-agent optimization is generally cast as finding equilibria in the space of strategies. The classic notion of equilibrium is due to Nash [Nash, 1951]: a Nash equilibrium is a set of agent strategies for which no agent can unilaterally improve its loss value. Pure Nash equilibria, in which each agent adopts a single strategy, provide a limited notion of optimality because they exist only under restrictive conditions. On the other hand, mixed Nash equilibria (MNE), where agents adopt a strategy from a probability distribution over the set of all strategies, exist in much greater generality [Glicksberg, 1952]. Importantly, MNE exist for games with infinite-dimensional compact strategy spaces, in which each player observes a loss function that is continuous in its strategy. We encounter this setting in different game formulations of machine learning problems, like GANs [Goodfellow et al., 2014].

Although MNE are guaranteed to exist, it is difficult to identify them. Indeed, worst-case complexity analyses have shown that without additional assumptions on the losses there is no efficient algorithm for finding a MNE, even in the case of two-player finite games [Daskalakis et al., 2009]. Some recent progress has been made; Hsieh et al. [2019] proposed a mirror-descent algorithm with convergence guarantees, which is approximately realizable in high-dimension.

Contributions. Following Hsieh et al. [2019], we formulate continuous two-player zero-sum games as a multi-agent optimization problem over the space of probability measures on strategies. We describe two gradient descent-ascent dynamics in this space, both involving a transport term.

- We show that the stationary points of a gradient ascent-descent flow with Langevin diffusion over the space of mixed strategies are approximate MNE.
- We analyse a gradient ascent-descent dynamics that jointly updates the positions and weights of two mixed strategies to converge to an *exact* MNE. This dynamics corresponds to a gradient descent-ascent flow over the space of measures endowed with a Wasserstein-Fisher-Rao (WFR) metric [Chizat et al., 2018].
- We discretize both dynamics in space and time to obtain implementable training algorithms. We provide mean-field type consistency results on the discretization. We demonstrate numerically how both dynamics overcome the curse of dimensionality for finding MNE on synthetic games. On real data, we use WFR flows to train mixtures of GANs, that explicitly discover data clusters while maintaining good performance.

2 Related work

Equilibria in continuous games. Most of the works that study convergence to equilibria in continuous games or GANs do not frame the problem in the infinite-dimensional space of measures, but on finite-dimensional spaces. That is because they either (i) restrict their attention to games with convexity-concavity assumptions in which pure equilibria exist [Mertikopoulos et al., 2019, Lin et al., 2018, Nouiehed et al., 2019], or (ii) provide algorithms with convergence guarantees to local notions of equilibrium such as stable fixed points, local Nash equilibria and local minimax points [Heusel et al., 2017, Adolphs et al., 2018, Mazumdar et al., 2019, Jin et al., 2019, Fiez et al., 2019, Balduzzi et al., 2018]. Both approaches differ from ours, which is to give global convergence guarantees without convexity assumptions. Some works have studied approximate MNE in infinite-dimensional measure spaces. Arora et al. [2017] proved the existence of approximate MNE and studied the generalization properties of this approximate solution; their analysis, however, does not provide a constructive method to identify such a solution. In a more explicit setting, Grnarova et al. [2017] designed an online-learning algorithm for finding a MNE in GANs under the assumption that the discriminator is a single hidden layer neural network. Balandat et al. [2016] apply the dual averaging algorithm to the minimax problem and show that it recovers a MNE, but they do not provide any convergence rate nor a practical algorithm for learning mixed NE. Our framework holds without making any assumption on the architectures of the discriminator and generator and provides explicit algorithms with some convergence guarantees.

Mean-field view of nonlinear gradient descent. Our approach is closely related to the mean-field perspective on wide neural networks [Mei et al., 2018, Rotskoff and Vanden-Eijnden, 2018, Chizat and Bach, 2018, Sirignano and Spiliopoulos, 2019, Rotskoff et al., 2019]. These methods view training algorithms as approximations of Wasserstein gradient flows, which are dynamics on measures over the space of neurons. In our setting, a mixed strategy corresponds to a measure over the space of strategies.

Particle approaches for two-player games. Our theoretical work sheds a new light on the results of Hsieh et al. [2019], and rigorously justifies important algorithmic modifications the authors introduced. Specifically, they give rates of convergence for infinite-dimensional mirror descent on measures (i.e. updating strategy weights but not their positions). The straightforward implementation of this algorithm performs poorly unless the dimension is low (Figure 1), which is why they proposed an 'implementable' two-timescale version, in which the inner loop is a transport-based sampling procedure closely related to our Algorithm 1. This implementable version is not studied theoretically, as the two-timescale structure hinders a thorough analysis. Our analysis includes transport on equal footing with mirror descent updates.

3 Problem setup and mean-field dynamics

Notation. For a topological space $\mathcal{X}$ we denote by $\mathcal{P}(\mathcal{X})$ the space of Borel probability measures on $\mathcal{X}$, and $\mathcal{M}_+(\mathcal{X})$ the space of Borel (positive) measures. For a given measure $\mu \in \mathcal{P}(\mathcal{X})$ that is absolutely continuous with respect to the canonical Borel measure dx of $\mathcal{X}$ and has Radon-Nikodym derivative $\frac{d\mu}{dx} \in \mathcal{C}(\mathcal{X})$, we define its differential entropy $H(\mu) = -\int \log(\frac{d\mu}{dx}) d\mu$. For measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $\mathcal{W}_2$ is the 2-Wasserstein distance.

3.1 Lifting differentiable games to spaces of strategy distributions

Differentiable two-player zero-sum games. We recall the definition of a differentiable zero-sum game, and show how finding a mixed Nash equilibrium to such a game is equivalent to solving a bi-linear game in the infinite dimensional space of distributions on strategies. We will use gradient flow approaches for solving the lifted problem.

Definition 1. *A two-player zero-sum game consists of a set of two players with parameters $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where players observe a loss functions $\ell_1 \colon \mathcal{Z} \to \mathbb{R}$ and $\ell_2 \colon \mathcal{Z} \to \mathbb{R}$ that satisfy for all $(x, y) \in \mathcal{Z}$, $\ell_1(x, y) + \ell_2(x, y) = 0$. $\ell \triangleq \ell_1 = -\ell_2$ is the loss of the game.*

The compact finite-dimensional spaces of strategies $\mathcal{X}$ and $\mathcal{Y}$ are endowed with a certain distance function d (which we assume Euclidean in what follows—Subsec. G.5 derives our results on arbitrary strategy manifolds). This allows to define differentiable games, amenable to first-order optimization. We make the following mild assumption over the regularity of losses and constraints [Glicksberg, 1952].

Assumption 1. *The parameter spaces $\mathcal{X}$ and $\mathcal{Y}$ are compact Riemannian manifolds without boundary of dimensions d_x, d_y embedded in $\mathbb{R}^{D_x}, \mathbb{R}^{D_y}$ respectively. The loss ℓ is continuously differentiable and L-smooth with respect to each parameter. That is, for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$, $\|\nabla_x \ell(x, y) - \nabla_x \ell(x', y')\|_2 \leqslant L(d(x, x') + d(y, y'))$, $\|\nabla_y \ell(x, y) - \nabla_y \ell(x', y')\|_2 \leqslant L(d(x, x') + d(y, y'))$.*

From pure to mixed Nash equilibria. Assuming that both players play simultaneously, a pure Nash equilibrium point is a pair of strategies $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ such that, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(x^\star, y) \leqslant \ell(x^\star, y^\star) \leqslant \ell(x, y^\star)$. Such points do not always exist in continuous games. In contrast, mixed Nash equilibria (MNE) are guaranteed to exist [Glicksberg, 1952] under Assumption 1. Those distributions $(\mu_x^\star, \mu_y^\star) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ are global saddle points of the expected loss $\mathcal{L}(\mu_x, \mu_y) \triangleq \iint \ell(x, y) d\mu_x(x) d\mu_y(y)$. Formally, for all $\mu_x, \mu_y \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$,

$$\mathcal{L}(\mu_x^*, \mu_y) \leqslant \mathcal{L}(\mu_x^*, \mu_y^*) \leqslant \mathcal{L}(\mu_x, \mu_y^*). \tag{1}$$

We quantify the accuracy of an estimation $(\hat{\mu}_x, \hat{\mu}_y)$ of a MNE using the Nikaidô and Isoda [1955] error

$$\mathrm{NI}(\hat{\mu}_x, \hat{\mu}_y) = \sup_{\mu_y \in \mathcal{P}(\mathcal{Y})} \mathcal{L}(\hat{\mu}_x, \mu_y) - \inf_{\hat{\mu}_x \in \mathcal{P}(\mathcal{X})} \mathcal{L}(\mu_x, \hat{\mu}_y). \tag{2}$$

We track the evolution of this metric in our theoretical results (Subsec. 4.2) and in our experiments. We obtain guarantees on finding ε-MNE $(\mu_x^\varepsilon, \mu_y^\varepsilon)$, i.e. distribution pairs such that $\mathrm{NI}(\mu_x^\varepsilon, \mu_y^\varepsilon) \leqslant \varepsilon$.

Algorithm 1 Langevin Descent-Ascent (L-DA).

Input: IID samples $x_0^1, \dots, x_0^n$ from $\mu_{x,0} \in \mathcal{P}(\mathcal{X})$, IID samples $y_0^1, \dots, y_0^n \in \mathcal{Y}$ from $\mu_{y,0} \in \mathcal{P}(\mathcal{Y})$
for $t = 0, \dots, T$ **do**
  for $i = 1, \dots, n$ **do**
    Sample $\Delta W_t^i \sim \mathcal{N}(0, I)$, $x_{t+1}^i = x_t^i - \frac{\eta}{n} \sum_{j=1}^n \nabla_x \ell(x_t^i, y_t^j) + \sqrt{2\eta\beta^{-1}} \Delta W_t^i$
    Sample $\Delta \bar{W}_t^i \sim \mathcal{N}(0, I)$, $y_{t+1}^i = y_t^i + \frac{\eta}{n} \sum_{j=1}^n \nabla_y \ell(x_t^j, y_t^i) + \sqrt{2\eta\beta^{-1}} \Delta \bar{W}_t^i$
Return $\mu_{x,T}^n = \frac{1}{n} \sum_{i=1}^n \delta_{x_T^i}, \quad \mu_{y,T}^n = \frac{1}{n} \sum_{i=1}^n \delta_{y_T^i}$

3.2 Training dynamics on discrete mixtures of strategies

We study three different dynamics for solving (1). Let us first assume that the two players play *finite* mixtures of n strategies $\mu_x = \sum_{i=1}^n w_x^i \delta_{x^i} \in \mathcal{P}(\mathcal{X})$, $\mu_y = \sum_{i=1}^n w_y^i \delta_{y^i} \in \mathcal{P}(\mathcal{Y})$, where $\{x^i, y^i\}_{i \in [1:n]}$ are the positions of the strategies and $w_x^i, w_y^i \geqslant 0$ are their weights. In the simplest setting, those mixtures are assumed uniform, i.e. $w_x^i = w_y^i = 1/n$. Finding the best $2n$ strategies involve finding a saddle point of $\mathcal{L}(\mu_x, \mu_y) = \frac{1}{n^2} \sum_i \sum_j \ell(x_i, y_j)$. Starting from random independent initial strategies $x_0^i = \xi_i \sim \mu_{x,0}, y_0^i = \bar{\xi}_i \sim \mu_{y,0}$, we may hope that the gradient descent-ascent dynamics

$$\frac{dx_t^i}{dt} = -\frac{1}{n} \sum_{j=1}^n \nabla_x \ell(x_t^i, y_t^j), \quad \frac{dy_t^i}{dt} = \frac{1}{n} \sum_{j=1}^n \nabla_y \ell(x_t^j, y_t^i), \quad \forall i \in [1:n] \tag{3}$$

finds such a saddle point. Yet this may fail in simple nonconvex-nonconcave games, as illustrated in Subsec. G.2—the particle distributions collapse to a stationary point that is not a MNE.

To mitigate this convergence problem, we analyse a perturbed dynamics analogous to Langevin gradient descent. Using the same initialization as in (3), we add a small amount of noise in the gradient dynamics and obtain the stochastic differential equations

$$dX_t^i = -\frac{1}{n} \sum_{j=1}^n \nabla_x \ell(X_t^i, Y_t^j) dt + \sqrt{\frac{2}{\beta}} dW_t^i, \; dY_t^i = \frac{1}{n} \sum_{j=1}^n \nabla_y \ell(X_t^j, Y_t^i) dt + \sqrt{\frac{2}{\beta}} d\bar{W}_t^i, \tag{4}$$

where $W_t^i, \bar{W}_t^i$ are independent Brownian motions. The discretization of (4) results in Alg. 1; it is similar to Alg. 4 in Hsieh et al. [2019].

We propose a second alternative dynamics to (3), that updates both the positions and the weights of the particles, using relative updates for weights. We will show that it enjoys better convergence properties in the mean-field limit.

$$\frac{dx_t^i}{dt} = -\gamma \sum_{j=1}^n w_{y,t}^j \nabla_x \ell(x_t^i, y_t^j), \quad \frac{dw_{x,t}^i}{dt} = \alpha \left(-\sum_{j=1}^n w_{y,t}^j \ell(x_t^i, y_t^j) + K(t) \right) w_{x,t}^i \tag{5}$$

and similarly for all y_t^i (flipping the sign of ℓ). $K(t) \triangleq \sum_{k=1}^n \sum_{j=1}^n w_{y,t}^j w_{x,t}^k \ell(x_t^i, y_t^j)$ keeps $w_{x,t}$ in the simplex. We use uniform weights for initialization. When $\gamma = 0$ and $\alpha = 1$, only the weights are updated: this results in the continuous-time version of the infinite-dimensional mirror descent studied by Hsieh et al. [2019]. The Euler discretization of (5) results in Alg. 2.

3.3 Training dynamics as gradient flows on measures

The three dynamics that we have introduced at the level of particles induces dynamics on the associated empirical probability measures. If $\{x_t^i, y_t^i\}_{i \in [1,n]}$ is a solution of (3), then $\mu_x(t) = \frac{1}{n} \sum_{i=1}^n \delta_{x_t^i}$ and $\mu_y(t) =$

Algorithm 2 Wasserstein-Fisher-Rao Descent-Ascent (WFR-DA).

Input: IID samples $x_0^{(1)}, \dots, x_0^{(n)}$ from $\nu_{x,0} \in \mathcal{P}(\mathcal{X})$, IID samples $y_0^{(1)}, \dots, y_0^{(n)}$ from $\nu_{y,0} \in \mathcal{P}(\mathcal{Y})$. Initial weights: For all $i \in [1:n]$, $w_x^{(i)} = 1,\ w_y^{(i)} = 1$.
for $t = 0, \dots, T$ **do**
  $[x_{t+1}^{(i)}]_{i=1}^n = [x_t^{(i)} - \eta \sum_{j=1}^n w_{y,t}^{(i)} \nabla_x \ell(x_t^{(i)}, y_t^{(j)})]_{i=1}^n$
  $[\hat{w}_{x,t+1}^{(i)}]_{i=1}^n = \Big[w_{x,t}^{(i)} \exp\big(-\eta' \sum_{j=1}^n w_{y,t}^{(j)} \ell(x_t^{(i)}, y_t^{(j)})\big)\Big]_{i=1}^n, \quad [w_{x,t+1}^{(i)}]_{i=1}^n = [\hat{w}_{x,t+1}^{(i)}]_{i=1}^n / \sum_{j=1}^n \hat{w}_{x,t+1}^{(j)}$
  $[y_{t+1}^{(i)}]_{i=1}^n = [y_t^{(i)} + \eta \sum_{j=1}^n w_{x,t}^{(j)} \nabla_y \ell(x_t^{(j)}, y_t^{(i)})]_{i=1}^n,$
  $[\hat{w}_{y,t+1}^{(i)}]_{i=1}^n = \Big[w_{y,t}^{(i)} \exp\big(\eta' \sum_{j=1}^n w_{x,t}^{(j)} \ell(x_t^{(j)}, y_t^{(i)})\big)\Big]_{i=1}^n, \quad [w_{y,t+1}^{(i)}]_{i=1}^n = [\hat{w}_{y,t+1}^{(i)}]_{i=1}^n / \sum_{j=1}^n \hat{w}_{y,t+1}^{(j)}$
Return $\bar{\nu}_{x,T}^n = \frac{1}{T+1} \sum_{t=0}^T \sum_{i=1}^n w_{x,T}^{(i)} \delta_{x_T^{(i)}}, \quad \bar{\nu}_{y,T}^n = \frac{1}{T+1} \sum_{t=0}^T \sum_{i=1}^n w_{y,T}^{(i)} \delta_{y_T^{(i)}}$

$\frac{1}{n} \sum_{i=1}^n \delta_{y_t^i}$ are solutions of the *Interacting Wasserstein Gradient Flow* (IWGF) of $\mathcal{L}$:

$$\begin{cases} \partial_t \mu_x = \nabla \cdot (\mu_x \nabla_x V_x(\mu_y, x)), & \mu_x(0) = \frac{1}{n} \sum_{i=1}^n \delta_{x_0^i}, \\ \partial_t \mu_y = -\nabla \cdot (\mu_y \nabla_y V_y(\mu_x, y)), & \mu_y(0) = \frac{1}{n} \sum_{i=1}^n \delta_{y_0^i}. \end{cases} \tag{6}$$

The derivation of (6) is provided in Subsec. G.3. We use the notation $V_x(\mu_y, x) \triangleq \frac{\delta \mathcal{L}}{\delta \mu_x}(\mu_x, \mu_y)(x) = \int \ell(x, y) d\mu_y(y)$ for the first variations of the functional $\mathcal{L}(\mu_x, \mu_y)$. Holding μ_y fixed, the evolution of μ_x is a Wasserstein gradient flow on $\mathcal{L}(\cdot, \mu_y)$ [Ambrosio et al., 2005]. We interpret these PDEs in the weak sense, i.e. equality holds when integrating measures against bounded continuous functions.

The distributions $\mu_x(t) = \frac{1}{n} \sum_{i=1}^n \delta_{X_t^i}$ and $\mu_y(t) = \frac{1}{n} \sum_{i=1}^n \delta_{Y_t^i}$, where $\{X^i, Y^i\}_{i \in [1:n]}$ are solutions of (4) follows a *Entropy-Regularized Interacting Wasserstein Gradient Flow (ERIWGF)*:

$$\begin{cases} \partial_t \mu_x = \nabla_x \cdot (\mu_x \nabla_x V_x(\mu_y, x)) + \beta^{-1} \Delta_x \mu_x, & \mu_x(0) = \frac{1}{n} \sum_{i=1}^n \delta_{x_0^i} \\ \partial_t \mu_y = -\nabla_y \cdot (\mu_y \nabla_y V_y(\mu_x, y)) + \beta^{-1} \Delta_y \mu_y, & \mu_y(0) = \frac{1}{n} \sum_{i=1}^n \delta_{y_0^i} \end{cases} \tag{7}$$

The derivation of (7) is provided in Lemma 10. It is a system of coupled nonlinear Fokker-Planck equations, that are the Kolmogorov forward equations of the SDE (4). They correspond to the IWGF of the entropy-regularized loss $\mathcal{L}_\beta(\mu_x, \mu_y) \triangleq \mathcal{L}(\mu_x, \mu_y) + \beta^{-1}(H(\mu_y) - H(\mu_x)$.

Finally, if $\{x^i, y^i, w_x^i, w_y^i\}_{i \in [1:n]}$ solve (5), then $\mu_x(t) = \sum_{i=1}^n w_{x,t}^i \delta_{x_t^i}$, $\mu_y(t) = \sum_{i=1}^n w_{y,t}^i \delta_{y_t^i}$ solve the *Interacting Wasserstein-Fisher-Rao Gradient Flow (IWFRGF)* of $\mathcal{L}$:

$$\begin{cases} \partial_t \mu_x &= \gamma \nabla_x \cdot (\mu_x \nabla_x V_x(\mu_y, x)) - \alpha \mu_x (V_x(\mu_y, x) - \mathcal{L}(\mu_x, \mu_y)), \ \mu_x(0) = \sum_{i=1}^n w_{x,0}^i \delta_{x_0^i}, \\ \partial_t \mu_y &= -\gamma \nabla_y \cdot (\mu_y \nabla_y V_y(\mu_x, y)) + \alpha \mu_y (V_y(\mu_x, y) - \mathcal{L}(\mu_x, \mu_y)), \ \mu_y(0) = \sum_{i=1}^n w_{y,0}^i \delta_{y_0^i}. \end{cases} \tag{8}$$

The derivation of (8) is provided in App. A and Lemma 11. The Wasserstein-Fisher-Rao or Hellinger-Kantorovich metric [Chizat et al., 2015, Kondratyev et al., 2016, Gallouët and Monsaingeon, 2016] is a metric on the probability space $\mathcal{M}_+(\mathcal{X})$ induced by a lifting to the space $\mathcal{P}(\mathcal{X} \times \mathbb{R}^+)$ of the form $\nu \mapsto \mu = \int_{\mathbb{R}^+} w \, d\nu(\cdot, w)$. If we keep ν_y fixed, the first equation in (8) is a Wasserstein-Fisher-Rao gradient flow (slightly modified by the term $\alpha \mu_x \mathcal{L}(\mu_x, \mu_y)$ to constrain μ_x in $\mathcal{P}(\mathcal{X})$). The term $-\alpha \mu_x (V_x(\mu_y, x) - \mathcal{L}(\mu_x, \mu_y))$, which also arises in entropic mirror descent, allow mass to 'teleport' from bad strategies to better ones with finite cost by moving along the weight coordinate. Wasserstein-Fisher-Rao gradient flows have been used by Chizat [2019], Rotskoff et al. [2019], Liero et al. [2018] in the context of optimization.

Initialization of (6), (7) and (8) may be done with the measures $\mu_{x,0}$ and $\mu_{y,0}$ from which $\{x_0^i\}, \{y_0^i\}$ are sampled, in which case the measures $\mu_x(t)$ and $\mu_y(t)$ are not discrete and follow the *mean-field* dynamics. In Subsec. 4.3 we link the dynamics starting from discrete realizations to the mean-field dynamics.

4 Convergence analysis

We establish convergence results for the entropy-regularized dynamics and the WFR dynamics.

4.1 Convergence of the entropy-regularized Wasserstein dynamics

The following theorem characterizes the stationary points of the entropy-regularized dynamics.

Theorem 1. *Suppose that Assumption 1 holds, that $\ell \in C^2(\mathcal{X} \times \mathcal{Y})$ and that the initial measures $\mu_{x,0}, \mu_{y,0}$ have densities in $L^1(\mathcal{X}), L^1(\mathcal{Y})$. If a solution $(\mu_x(t), \mu_y(t))$ of the ERIWGF (7) converges in time, it must converge to the point $(\hat{\mu}_x, \hat{\mu}_y)$ which is the unique fixed point of the problem*

$$\rho_x(x) = \frac{1}{Z_x} e^{-\beta \int \ell(x,y)\ d\mu_y(y)}, \quad \rho_y(y) = \frac{1}{Z_y} e^{\beta \int \ell(x,y)\ d\mu_x(x)}. \tag{9}$$

$(\hat{\mu}_x, \hat{\mu}_y)$ is an ε-Nash equilibrium of the game given by $\mathcal{L}$ when $\beta \geqslant \frac{4}{\varepsilon} \log\left(2\frac{1-V_\delta}{V_\delta}(2K_\ell/\varepsilon - 1)\right)$, where $K_\ell := \max_{x,y} \ell(x,y) - \min_{x,y} \ell(x,y)$ is the length of the range of ℓ, $\delta := \varepsilon/(2Lip(\ell))$ and V_δ is a lower bound on the volume of a ball of radius δ in $\mathcal{X}, \mathcal{Y}$.

The proof is in App. C. Theorem 1 characterizes the stationary points of the ERIWGF but does not provide a guarantee of convergence in time. It implies that if the dynamics (7) converges in time, the limit will be an ε-Nash equilibrium of $\mathcal{L}$, with $\varepsilon = \tilde{O}(1/\beta)$ (disregarding log factors). The dynamics (7) correspond to a McKean-Vlasov process on the joint probability measure $\mu_x \times \mu_y$. While convergence to stationary solutions of such processes have been studied in the Euclidean case [Eberle et al., 2019]l, their results would only guarantee convergence for temperatures $\beta^{-1} \gtrsim Lip(\ell)$ in our setup, which is not strong enough to certify convergence to arbitrary ε-NE.

There is a trade-off between setting a low temperature β^{-1}, which yields an ε-Nash equilibrium with small ε but possibly slow or no convergence, and setting a high temperature, which has the opposite effect. Linear potential Fokker-Planck equations (that we recover when both players are decoupled) indeed converge exponentially with rate $e^{-\lambda_\beta t}$ for all β, with λ_β decreasing exponentially with β for nonconvex potentials [Markowich and Villani, 1999, sec. 5]. Entropic regularization also biases the dynamics towards measures with full support and hence precludes convergence to sparse equilibria even if they exist. This problem does not arise in the WFR dynamics.

4.2 Analysis of the Wasserstein-Fisher-Rao dynamics

Theorem 2 states that, at a certain time t_0, the time averaged measures of the solution (ν_x, ν_y) of (8) are an ε-MNE, where ε can be made arbitrarily small by adjusting the constants γ, α of the dynamics. We define $\bar{\nu}_x(t) = \frac{1}{t}\int_0^t \nu_x(s)\ ds$ and $\bar{\nu}_y(t) = \frac{1}{t}\int_0^t \nu_y(s)\ ds$, where ν_x and ν_y are solutions of (8).

Theorem 2. *Let $\varepsilon > 0$ arbitrary. Suppose that $\nu_{x,0}, \nu_{y,0}$ are such that their Radon-Nikodym derivatives with respect to the Borel measures of $\mathcal{X}, \mathcal{Y}$ are lower-bounded by $e^{-K'_x}, e^{-K'_y}$ respectively. For any $\delta \in (0, 1/2)$, there exists a constant $C_{\delta,\mathcal{X},\mathcal{Y},K'_x,K'_y} > 0$ depending on the dimensions of $\mathcal{X}, \mathcal{Y}$, their curvatures and K'_x, K'_y, such that if $\gamma/\alpha < 1$, $\frac{\gamma}{\alpha} \leqslant \left(\varepsilon / C_{\delta,\mathcal{X},\mathcal{Y},K'_x,K'_y}\right)^{\frac{2}{1-\delta}}$*

$$NI(\bar{\nu}_x(t_0), \bar{\nu}_y(t_0)) \leqslant \varepsilon \quad \textit{where} \quad t_0 = (\alpha\gamma)^{-1/2}.$$

The proof (App. D) builds on the convergence properties of continuous-time mirror descent and closely follows the proof of Theorem 3.8 from Chizat [2019]. We explicit the dependency of $C_{\delta,\mathcal{X},\mathcal{Y},K'_x,K'_y}$ on the

dimensions of the manifolds and the properties of the loss ℓ. Notice that Theorem 2 ensures convergence towards an ε-Nash equilibrium of the non-regularized game. Following Chizat [2019], it is possible to replace the regularity assumption on the initial measures $\nu_{x,0}, \nu_{y,0}$ by a singular initialisation, at the expense of using $O(\exp(d))$ particles. This result is not a convergence result for the measures, but rather on the value of the NI error. Notice that it involves time-averaging and a finite horizon. Similar results are common for mirror descent in convex games [Juditsky et al., 2011], albeit in the discrete-time setting.

Theorem 2 does not capture the benefits of transport, as it regards it as a perturbation of mirror descent (which corresponds to $\gamma = 0$). When targetting a small error ε, we need to set $\gamma \ll \alpha$ because of the bound on γ/α. In this case, mirror descent is the main driver of the dynamics. However, it is seen empirically that taking much higher ratios γ/α (i.e. increasing the importance of the transport term) results in better performance. A satisfying explanation of this phenomenon is still sought after in the simpler optimization setting [Chizat, 2019].

4.3 Convergence to mean-field

The following theorem (proof in App. F) links the empirical measures of the systems (4), (5) to the solutions of the mean field dynamics (7) and (8) respectively. It can be seen as a law of large numbers. It shows that by Theorem 3, Alg. 1 and Alg. 2 approximate the mean-field dynamics studied in Subsec. 4.1 and Subsec. 4.2.

Theorem 3. *(i) Let $\mu_x^n = \frac{1}{n}\sum_{i=1}^n \delta_{X^{(i)}} \in \mathcal{C}([0,T], \mathcal{P}(\mathcal{X})), \mu_y^n = \frac{1}{n}\sum_{i=1}^n \delta_{Y^{(i)}} \in \mathcal{C}([0,T], \mathcal{P}(\mathcal{Y}))$ be the empirical measures of a solution of (4) up to an arbitrary time T. Let $\mu_x \in \mathcal{C}([0,T], \mathcal{P}(\mathcal{X})), \mu_y \in \mathcal{C}([0,T], \mathcal{P}(\mathcal{Y}))$ be a solution of the ERIWGF (7) with mean-field initial conditions $\mu_x(0) = \mu_{x,0}, \mu_y(0) = \mu_{y,0}$. Then,*

$$\mathbb{E}[\mathcal{W}_2^2(\mu_{x,t}^n, \mu_{x,t}) + \mathcal{W}_2^2(\mu_{y,t}^n, \mu_{y,t})] \xrightarrow{n\to\infty} 0, \quad \mathbb{E}[|NI(\mu_{x,t}^n, \mu_{y,t}^n) - NI(\mu_{x,t}, \mu_{y,t})|] \xrightarrow{n\to\infty} 0,$$

uniformly over $t \in [0,T]$. NI is the Nikaido-Isoda error defined in (2).

(ii) Let $\nu_x^n = \sum_{i=1}^n w_{x,t}^i \delta_{X^{(i)}} \in \mathcal{C}([0,T], \mathcal{P}(\mathcal{X}))$, $\mu_y^n = \sum_{i=1}^n w_{y,t}^i \delta_{Y^{(i)}} \in \mathcal{C}([0,T], \mathcal{P}(\mathcal{Y}))$ be the (projected) empirical measures of a solution of (5) up to an arbitrary time T. Let $\nu_x \in \mathcal{C}([0,T], \mathcal{P}(\mathcal{X})), \nu_y \in \mathcal{C}([0,T], \mathcal{P}(\mathcal{Y}))$ be a solution of (8) with mean-field initial conditions $\mu_x(0) = \mu_{x,0}, \mu_y(0) = \mu_{y,0}$. Then,

$$\mathbb{E}[\mathcal{W}_2^2(\nu_{x,t}^n, \nu_{x,t}) + \mathcal{W}_2^2(\nu_{y,t}^n, \nu_{y,t})] \xrightarrow{n\to\infty} 0, \quad \mathbb{E}[|NI(\bar{\nu}_{x,t}^n, \bar{\nu}_{y,t}^n) - NI(\bar{\nu}_{x,t}, \bar{\nu}_{y,t})|] \xrightarrow{n\to\infty} 0,$$

uniformly over $t \in [0,T]$. $\bar{\nu}_{x,t}, \bar{\nu}_{y,t}, \bar{\nu}_{x,t}^n, \bar{\nu}_{y,t}^n$ are the time-averaged measures, as in Theorem 2.

5 Numerical Experiments

We show that WFR and Langevin dynamics outperform mirror descent in high dimension, on synthetic games. We then show the interests of using WFR-DA for training GANs. Code has been made available for reproducibility.

5.1 Polynomial games on spheres

We study two different games with losses $\ell_a, \ell_b : \mathcal{S}^{d-1} \times \mathcal{S}^{d-1} \to \mathbb{R}$ of the form

$$\ell_a(x,y) = x^\top A_0 x + x^\top A_1 y + y^\top A_2 y + y^\top A_3(x^2) + a_0^\top x + a_1^\top y$$
$$\ell_b(x,y) = x^\top A_0^\top A_0 x + x^\top A_1 y + y^\top A_2^\top A_2 y + a_0^\top x + a_1^\top y,$$

where $A_0, A_1, A_2, A_3, a_0, a_1$ are matrices and vectors with components sampled from a normal distribution $\mathcal{N}(0,1)$, and x^2 is the vector given by component-wise multiplication of x. ℓ_b is a convex loss on the sphere,

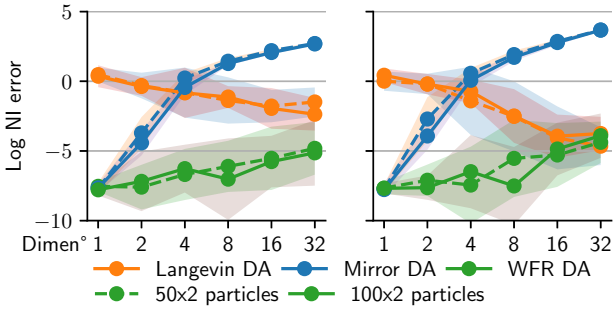

Figure 1: Nikaido-Isoida errors for L-DA, WFR-DA and mirror descent, as a function of the problem dimension, for a nonconvex loss ℓ_a (left) and convex loss ℓ_b (right). L-DA and WFR-DA outperforms mirror descent for large dimensions. Values averaged over 20 runs after 30000 iterations. Error bars show standard deviation across runs.

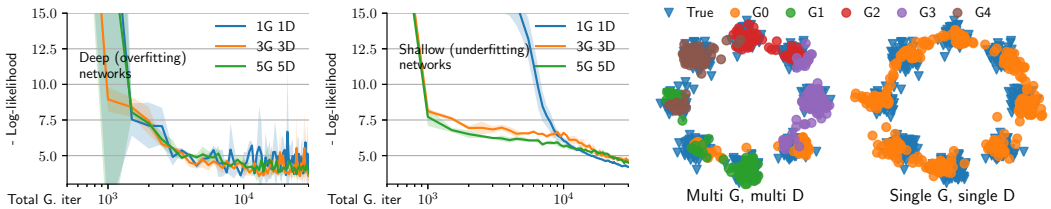

Figure 2: Training mixtures of GANs over a synthetic mixture of Gaussians in 2D. WFR-DA converges faster with models with low number of parameters, and similar performance with over-parametrized models. Mixtures naturally perform a form of clustering of the data. Errors bars show variance across 5 runs.

while ℓ_a is not. We run Langevin Descent-Ascent (updates of positions) and WFR Descent-Ascent (updates of weights and positions), and compare it with mirror descent (updates of weights).We note that the computation of the NI error (2) entails solving two optimization problems on measures, or equivalently in parameter space. We solve each of them by performing 2000 gradient acsent runs with random uniform initialization and selecting the highed minimum final value. This gives a lower bound on the NI error which is precise enough for our purposes. We perform time averaging on the weights of mirror descent and WFR-DA, but not on the positions of WFR-DA because that would incur an $O(t)$ overhead on memory.

Results. Wirror descent performs like WFR-DA in low dimensions, but suffers strongly from the curse of dimensionality (Figure 1). On the other hand, algorithms that incorporate a transport term keep performing well in high dimensions. In particular, WFR-DA is consistently the algorithm with lowest NI error. Notice that the errors in the $n = 50$ and $n = 100$ plots do not differ much, confirming that we reach a mean-field regime.

5.2 Training GAN mixtures

We now use WFR-DA to train mixtures of generator networks. We consider the Wasserstein-GAN [Arjovsky et al., 2017] setting. We seek to approximate a distribution $\mathcal{P}_{\text{data}}$ with a distribution $\mathcal{G}_x$, defined as the push-forward of a noise distribution $\mathcal{N}(0, I)$ by a neural-network g_x. The discrepancy between $\mathcal{P}_{\text{data}}$ and $\mathcal{G}_x$ is estimated by a neural-network discriminator f_y, leading to the problem

$$\min_x \max_y \ell(x, y) \triangleq \mathbb{E}_{a\sim p_{\text{data}}}[f_y(a)] - \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}[f_y(g_x(\varepsilon))].$$

We lift this problem in the space of distributions over the parameters x and y (see Subsec. G.4), that we represent through weighted discrete distributions of $\sum_{i=1}^p w_x^{(i)}\delta_{x^{(i)}}$ and $\sum_{j=1}^q w_y^{(j)}\delta_{y^{(j)}}$. We solve

$$\min_{x^{(i)}, w_x\in\triangle_p} \max_{y^{(j)}, w_y\in\triangle^q} \sum_{i=1}^p \sum_{j=1}^q w_x^{(i)} w_y^{(j)} \ell(x^{(i)}, y^{(j)}) \ ,$$

Figure 3: Training mixtures of GANs over CIFAR10. We compare the algorithm that updates the mixture weights and parameters (WFR-DA flow) with the algorithm that only updates parameters (W-DA flow). Using several discriminators and a WFR-DA flow brings more stable convergence. Each generator tends to specialize in a type of images. Errors bars show variance across 5 runs.

using Alg. 2, where $\triangle^q$ is the q-dimensional simplex. The optimal generation strategy corresponding to an equilibrium point $(x^{(i)})_i, w_x, (y^{(j)})_j, w_y$ is then to randomly select a generator g_{x_I} with I sampled among $[n]$ with probability $w_x^{(i)}$, and use it to generate $g_{x_I}(\varepsilon)$, with $\varepsilon \sim \mathcal{N}(0, I)$. Training mixtures of generators has been proposed by Ghosh et al. [2018], with a tweaked discriminator loss. Our formulation only involves a lifting in the space of measures, and uses a new training algorithm.

Results on 2D GMMs. We first set $\mathcal{P}_{\text{data}}$ to be an 8-mode mixture of Gaussians in two dimensions. We use the original W-GAN loss, with weight cropping for the discriminators $(f_{y^{(j)}})_j$. We measure the interest of using mixtures when a single generator $g_{x^{(i)}}$ cannot fit $\mathcal{P}_{\text{data}}$ (single-layer MLP), and when it can (4-layer MLP). We report results in Figure 2, measuring the log likelihood of $\mathcal{G}_x$ for the GMM during training. The WFR dynamic is stable even with few particles. When training under-parametrized generators, using mixtures permits faster convergence (in terms of generator updates). In the over-parametrized setting, training a single generator or a mixture of generators perform similarly. WFR-DA is thus useful to train mixtures of simple generators. In this setting, each simple generator identifies modes in the training data, doing data clustering at no cost (Figure 2 right).

Results on real data. We train a mixture of ResNet generators on CIFAR10 and MNIST. We replace the position updates in Alg. 2 by extrapolated Adam steps [Gidel et al., 2019] to achieve faster convergence, and perform grid search over generator and discriminators learning rates. Convergence curves for the best learning rates are displayed in Figure 3 right, measuring test FID [Heusel et al., 2017]. With a sufficient number of generators and discriminators ($G > 5, D > 2$), the model trains as fast as a normal GAN. WFR-DA is thus stable and efficient even with a reasonable number of particles. Using the discretized WFR versus the Wasserstein flow provides a slight improvement over updating parameters only. As with GMMs, each generator trained with WFR-DA becomes specialised in generating a fraction of the target data, thereby identifying clusters. Those could be used for unsupervised conditional generation of images.

6 Conclusions and future work

We have explored non-convex-non-concave, high-dimensional games from the perspective of optimal transport. As with non-convex optimization, framing the problem in terms of measures provides geometric benefits, at the expense of moving into non-Euclidean metric spaces over measures. Our theoretical results establish approximate mean-field convergence for two setups: Langevin Descent-Ascent and WFR D-A, and directly applies to GANs, for mixtures of generators and discriminators.

Despite the positive convergence guarantees our results are qualitative in nature, i.e. without rates. In the entropic case, the unfavorable tradeoff between temperature and convergence of the associated McKean-Vlasov scheme deserves further study, maybe through log-Sobolev-type inequalities [Markowich and Villani, 1999]. In

the WFR case, we lack a local convergence analysis explaining the benefits of transport observed empirically, perhaps leveraging sharpness Polyak-Łojasiewicz results such as those in [Chizat, 2019] or [Sanjabi et al., 2018]. Finally, in our GAN formulation, each generator is associated to a single particle in a high-dimensional product space of all network parameters, which is not scalable to large population sizes that would approximate their mean-field limit. A natural question is to understand to what extent our framework could be combined with specific choices of architecture, as recently studied in [Lei et al., 2019].

Broader impact

We study algorithms designed to find equilibria in games, provide theoretical guarantees of convergence and test their performance empirically. Among other applications, our results give insight into training algorithms for generative adversarial networks (GANs), which are useful for many relevant tasks such as image generation, image-to-image or text-to-image translation and video prediction. As always, we note that machine learning improvements like ours come in the form of "building machines to do X better". For a sufficiently malicious or ill-informed choice of X, such as surveillance or recidivism prediction, almost any progress in machine learning might indirectly lead to a negative outcome, and our work is not excluded from that.

Funding disclosure

C. Domingo-Enrich thanks J. De Dios Pont for conversations on the subject. This work is partially supported by the Alfred P. Sloan Foundation, NSF RI-1816753, NSF CAREER CIF 1845360, NSF CHS-1901091, Samsung Electronics, and the Institute for Advanced Study. The work of C. Domingo-Enrich is partially supported by the La Caixa Fellowship. The work of A. Mensch is supported by the European Research Council (ERC project NORIA). The work of G. Rotskoff is supported by the James S. McDonnell Foundation.

References

L. Adolphs, H. Daneshmand, A. Lucchi, and T. Hofmann. Local saddle point optimization: A curvature exploitation approach. *arXiv preprint arXiv:1805.05751*, 2018.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Lectures in mathematics ETH Zürich. Birkhäuser, 2005.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 224–232. JMLR. org, 2017.

M. Balandat, W. Krichene, C. Tomlin, and A. Bayen. Minimizing regret on reflexive banach spaces and nash equilibria in continuous zero-sum games. In *Advances in Neural Information Processing Systems*, pages 154–162, 2016.

D. Balduzzi, S. Racanière, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The mechanics of n-player differentiable games. In *Proceedings of the International Conference on Machine Learning*, 2018.

V. I. Bogachev, N. V. Krylov, and M. Röckner. On regularity of transition probabilities and invariant measures of singular diffusions under minimal conditions. *Communications in Partial Differential Equations*, 26 (11-12):2037–2080, 2001.

L. Bu, R. Babu, B. De Schutter, et al. A comprehensive survey of multi-agent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

G. Chirikjian. *Stochastic models, information theory, and Lie groups*. Number v. 1 in Applied and numerical harmonic analysis. Birkhäuser, 2009.

L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300v1*, 2019.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. 18(1):1–44, 2018.

C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

A. Eberle, A. Guillin, and R. Zimmer. Quantitative Harris-type theorems for diffusions and McKean-Vlasov processes. *Trans. Amer. Math. Soc.*, 371:7135–7173, 2019.

T. Fiez, B. Chasnov, and L. J. Ratliff. Convergence of learning dynamics in stackelberg games, 2019.

T. O. Gallouët and L. Monsaingeon. A jko splitting scheme for kantorovich-fisher-rao gradient flows. *SIAM J. Math. Analysis*, 49:1100–1130, 2016.

A. Ghosh, V. Kulharia, V. Namboodiri, P. H. S. Torr, and P. K. Dokania. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

I. L. Glicksberg. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

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

P. Grnarova, K. Y. Levy, A. Lucchi, T. Hofmann, and A. Krause. An online learning approach to generative adversarial networks. *arXiv preprint arXiv:1706.03269*, 2017.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Y.-P. Hsieh, C. Liu, and V. Cevher. Finding mixed Nash equilibria of generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2810–2819, Long Beach, California, USA, 06 2019. PMLR.

W. Huang, M. Ji, Z. Liu, and Y. Yi. Steady states of fokker–planck equations: I. existence. *Journal of Dynamics and Differential Equations*, 27:721–742, 12 2015.

C. Jin, P. Netrapalli, and M. I. Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

O. Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer New York, 2002. ISBN 9780387953137.

S. Kondratyev, L. Monsaingeon, D. Vorotnikov, et al. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12):1117–1164, 2016.

D. Lacker. Mean field games and interacting particle systems. *Preprint*, 2018.

Q. Lei, J. D. Lee, A. G. Dimakis, and C. Daskalakis. SGD learns one-layer networks in WGANs. *arXiv preprint arXiv:1910.07030*, 2019.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 03 2018.

Q. Lin, M. Liu, H. Rafique, and T. Yang. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 2018.

P. A. Markowich and C. Villani. On the trend to equilibrium for the fokker-planck equation: An interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*, pages 1–29, 1999.

E. V. Mazumdar, M. I. Jordan, and S. S. Sastry. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *arXiv:1901.00838*, 2019.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):7665–7671, 2018.

P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2019.

J. Nash. Non-cooperative games. *Annals of Mathematics*, pages 286–295, 1951.

H. Nikaidô and K. Isoda. Note on non-cooperative convex games. *Pacific Journal of Mathematics*, 5(Suppl. 1):807–815, 1955.

M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14905–14916, 2019.

E. C. Posner. Random coding strategies for minimum entropy. *IEEE Transations on Information Theory*, 21 (4):388–391, 1975.

S. Racanière, T. Weber, D. Reichert, L. Buesing, A. Guez, D. J. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li, et al. Imagination-augmented agents for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5690–5701, 2017.

J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33 (3):520–534, 1965.

G. Rotskoff and E. Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

G. Rotskoff, S. Jelassi, J. Bruna, and E. Vanden-Eijnden. Global convergence of neuron birth-death dynamics. *arXiv preprint arXiv:1902.01843*, 2019.

M. Sanjabi, M. Razaviyayn, and J. D. Lee. Solving non-convex non-concave min-max games under polyaklojasiewicz condition. *arXiv preprint arXiv:1812.02878*, 2018.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 2019.

A.-S. Sznitman. Topics in propagation of chaos. In P.-L. Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

G. Wayne and L. Abbott. Hierarchical control using networks trained with higher-level forward models. *Neural Computation*, 26(10):2163–2193, 2014.

A Lifted dynamics for the Interacting Wasserstein-Fisher-Rao Gradient Flow

Recall the IWFRGF in (8), which we reproduce here for convenience.

$$\begin{cases} \partial_t \mu_x &= \gamma \nabla_x \cdot (\mu_x \nabla_x V_x(\mu_y, x)) - \alpha \mu_x (V_x(\mu_y, x) - \mathcal{L}(\mu_x, \mu_y)), \quad \mu_x(0) = \mu_{x,0} \\ \partial_t \mu_y &= -\gamma \nabla_y \cdot (\mu_y \nabla_y V_y(\mu_x, y)) + \alpha \mu_y (V_y(\mu_x, y) - \mathcal{L}(\mu_x, \mu_y)), \quad \mu_y(0) = \mu_{y,0} \end{cases}$$

Given $\nu_x \in \mathcal{P}(\mathcal{X} \times \mathbb{R}^+)$ define $\mu_x = \int_{\mathcal{X}} w_x \, d\nu_x(\cdot, w_x) \in \mathcal{P}(\mathcal{X})$, that is

$$\int_{\mathcal{X}} \varphi(x) \, d\mu_x(x) = \int_{\mathcal{X} \times \mathbb{R}^+} w_x \varphi(x) \, d\nu_x(x, w_x),$$

for all $\varphi \in C(\mathcal{X})$. Given $\nu_y \in \mathcal{P}(\mathcal{Y} \times \mathbb{R}^+)$, define $\mu_y = \int_{\mathcal{X}} w_y \, d\nu_y(\cdot, w_y) \in \mathcal{P}(\mathcal{Y})$ analogously. We say that ν_x, ν_y are "lifted" measures of μ_x, μ_y, and reciprocally μ_x, μ_y are "projected" measures of ν_x, ν_y.

By Lemma 1 below, we can view a solution of (8) as the projection of a solution of the following dynamics on the lifted domains $\mathcal{X} \times \mathbb{R}^+$ and $\mathcal{Y} \times \mathbb{R}^+$:

$$\begin{cases} \partial_t \nu_x = \nabla_{w_x, x} \cdot (\nu_x g_{\mu_y}(x, w_x)), \quad \nu_x(0) = \mu_{x,0} \times \delta_{w_x = 1} \\ \partial_t \nu_y = -\nabla_{w_y, y} \cdot (\nu_y g_{\mu_x}(y, w_y)), \quad \nu_y(0) = \mu_{y,0} \times \delta_{w_y = 1} \end{cases} \tag{10}$$

where

$$\begin{aligned} g_{\mu_y}(x, w_x) &= (\alpha w_x (V_x(\mu_y, x) - \mathcal{L}(\mu_x, \mu_y)), \gamma \nabla_x V_x(\mu_y, x))), \\ g_{\mu_x}(y, w_y) &= (\alpha w_y (V_y(\mu_x, x) - \mathcal{L}(\mu_x, \mu_y)), \gamma \nabla_y V_y(\mu_x, y))). \end{aligned}$$

Lemma 1. *For a solution $\nu_x : [0, T] \to \mathcal{P}(\mathcal{X} \times \mathbb{R}^+), \nu_y : [0, T] \to \mathcal{P}(\mathcal{Y} \times \mathbb{R}^+)$ of (10), the projections μ_x, μ_y are solutions of (8).*

That is, given any $\varphi_x \in \mathcal{C}^1(\mathcal{X}), \varphi_y \in \mathcal{C}^1(\mathcal{Y})$, we have

$$\begin{aligned} \frac{d}{dt} \int_{\mathcal{X}} \varphi_x(x) \, d\mu_x &= -\gamma \int_{\mathcal{X}} \nabla_x \varphi_x(x) \cdot \nabla_x V_x(\mu_y, x) \, d\mu_x - \alpha \int_{\mathcal{X}} \varphi_x(x) (V_x(\mu_y, x) - \mathcal{L}(\mu_x, \mu_y)) \, d\mu_x, \\ \frac{d}{dt} \int_{\mathcal{Y}} \varphi_y(y) \, d\mu_y &= \gamma \int_{\mathcal{Y}} \nabla_y \varphi_y(y) \cdot \nabla_y V_y(\mu_x, y)) \, d\mu_y + \alpha \int_{\mathcal{Y}} \varphi_y(y) (V_y(\mu_x, y) - \mathcal{L}(\mu_x, \mu_y)) \, d\mu_y, \\ \mu_x(0) &= \mu_{x,0}, \quad \mu_y(0) = \mu_{y,0} \end{aligned} \tag{11}$$

From (10) in the weak form, we obtain that given any $\psi_x \in \mathcal{C}^1(\mathcal{X} \times \mathbb{R}^+), \psi_y \in \mathcal{C}^1(\mathcal{Y} \times \mathbb{R}^+)$,

$$\begin{aligned} \frac{d}{dt} \int_{\mathcal{X} \times \mathbb{R}^+} \psi_x(x, w_x) \, d\nu_x(x, w_x) &= \int_{\mathcal{X} \times \mathbb{R}^+} -\gamma \nabla_x \psi_x(x, w_x) \cdot \nabla_x V_x(\mu_y, x) \\ &\quad - \alpha w_x \frac{d\psi_x}{dw_x}(x, w_x)(V_x(\mu_y, x) - \mathcal{L}(\mu_x, \mu_y)) \, d\mu_x, \\ \frac{d}{dt} \int_{\mathcal{Y} \times \mathbb{R}^+} \psi_y(y, w_y) \, d\nu_y(y, w_y) &= \int_{\mathcal{Y} \times \mathbb{R}^+} \gamma \nabla_y \psi_y(y, w_y) \cdot \nabla_y V_y(\mu_x, y) \\ &\quad + \alpha w_y \frac{d\psi_y}{dw_y}(y, w_y)(V_y(\mu_x, y) - \mathcal{L}(\mu_x, \mu_y)) \, d\mu_y, \\ \nu_x(0) = \mu_{x,0} \times \delta_{w_x = 1}, &\quad \nu_y(0) = \mu_{y,0} \times \delta_{w_y = 1}. \end{aligned} \tag{12}$$

Taking $\psi_x(x, w_x) = w_x\varphi_x(x), \psi_y(y, w_y) = w_y\varphi_y(y)$ yields

$$\begin{aligned}\frac{d}{dt}\int_{\mathcal{X}\times\mathbb{R}^+} w_x\varphi_x(x)\ d\nu_x(x, w_x) &= \int_{\mathcal{X}\times\mathbb{R}^+} -\gamma w_x\nabla_x\varphi_x(x)\cdot\nabla_x V_x(\mu_y, x)\\ &\quad - \alpha w_x\varphi_x(x)(V_x(\mu_y, x) - \mathcal{L}(\mu_x, \mu_y))\ d\mu_x,\\ \frac{d}{dt}\int_{\mathcal{Y}\times\mathbb{R}^+} w_y\psi_y(y, w_y)\ d\nu_y(y, w_y) &= \int_{\mathcal{Y}\times\mathbb{R}^+} \gamma w_y\nabla_y\varphi_y(y)\cdot\nabla_y V_y(\mu_x, y)\\ &\quad + \alpha w_y\varphi_y(y)(V_y(\mu_x, y) - \mathcal{L}(\mu_x, \mu_y))\ d\mu_y.\end{aligned} \tag{13}$$

Notice that (13) is indeed (11).

B Continuity and convergence properties of the Nikaido-Isoda error

Lemma 2. *The Nikaido-Isoda error $NI : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ defined in (2) is continuous when we endow $\mathcal{P}(\mathcal{X}), \mathcal{P}(\mathcal{Y})$ with the topology of weak convergence. Specifically, it is Lip(ℓ)-Lipschitz when we use the distance $\mathcal{W}_1(\mu_x, \mu'_x) + \mathcal{W}_1(\mu_y, \mu'_y)$ between (μ_x, μ_y) and (μ'_x, μ'_y) in $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$.*

Proof. For any μ_y, the function $V_x(\mu_y, \cdot) : \mathcal{X} \to \mathbb{R}$ defined as $x \mapsto \int \ell(x, y)\ d\mu_y$ is continuous and it has the same Lipschitz constant Lip(ℓ) as ℓ. Hence, for any $\mu_x, \mu'_x \in \mathcal{P}(\mathcal{X})$,

$$\begin{aligned}&\sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x, \mu_y) - \sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu'_x, \mu_y) = \sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \int V_x(\mu_y, x)d\mu_x - \sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \int V_x(\mu_y, x)d\mu'_x\\ &\leqslant \sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \int V_x(\mu_y, x)d\mu'_x + \sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \int V_x(\mu_y, x)d(\mu_x - \mu'_x) - \sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \int V_x(\mu_y, x)d\mu'_x\\ &= \sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \int V_x(\mu_y, x)d(\mu_x - \mu'_x) \leqslant \mathrm{Lip}(\ell)\mathcal{W}_1(\mu_x, \mu'_x)\end{aligned}$$

The same inequality interchanging the roles of μ_x, μ'_x shows that $|\sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x, \mu_y) - \sup_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu'_x, \mu_y)| \leqslant \mathrm{Lip}(\ell)\mathcal{W}_1(\mu_x, \mu'_x)$ holds. An analogous reasoning for $\ell(\mu_x, \cdot) : \mathcal{Y} \to \mathbb{R}$ and the triangle inequality complete the proof. □

Lemma 3. *Suppose that $(\mu_x^n)_{n\in\mathbb{N}}$ is a sequence of random elements valued in $\mathcal{P}(\mathcal{X})$ such that*

$$\mathbb{E}[\mathcal{W}_2^2(\mu_x^n, \mu_x)] \xrightarrow{n\to\infty} 0,$$

where $\mu_x \in \mathcal{P}(X)$. Analogously, suppose that $(\mu_y^n)_{n\in\mathbb{N}}$ is a sequence of random elements valued in $\mathcal{P}(\mathcal{Y})$ such that

$$\mathbb{E}[\mathcal{W}_2^2(\mu_y^n, \mu_y)] \xrightarrow{n\to\infty} 0,$$

where $\mu_y \in \mathcal{P}(Y)$.

Then,

$$\mathbb{E}[|NI(\mu_x^n, \mu_y^n) - NI(\mu_x, \mu_y)|] \xrightarrow{n\to\infty} 0$$

Proof. First,

$$\mathbb{E}[\mathcal{W}_1(\mu_x^n, \mu_x)] \leqslant \mathbb{E}[\mathcal{W}_2(\mu_x^n, \mu_x)] \leqslant \left(\mathbb{E}[\mathcal{W}_2^2(\mu_x^n, \mu_x)]\right)^{1/2}, \tag{14}$$

which results from two applications of the Cauchy-Schwarz inequality on the appropriate scalar products. An analogous inequality holds for $\mathbb{E}[\mathcal{W}_1(\mu_y^n, \mu_y)]$. Hence, by Lemma 2,

$$\begin{aligned}\mathbb{E}[|\mathrm{NI}(\mu_x^n, \mu_y^n) - \mathrm{NI}(\mu_x, \mu_y)|] &\leqslant \mathrm{Lip}(\ell)\mathbb{E}[\mathcal{W}_1(\mu_x^n, \mu_x) + \mathcal{W}_1(\mu_y^n, \mu_y)] \\ &\leqslant \mathrm{Lip}(\ell)\left(\left(\mathbb{E}[\mathcal{W}_2^2(\mu_x^n, \mu_x)]\right)^{1/2} + \left(\mathbb{E}[\mathcal{W}_2^2(\mu_x^n, \mu_x)]\right)^{1/2}\right) \\ &\leqslant \mathrm{Lip}(\ell)\sqrt{2}\left(\mathbb{E}[\mathcal{W}_2^2(\mu_x^n, \mu_x)] + \mathbb{E}[\mathcal{W}_2^2(\mu_x^n, \mu_x)]\right)^{1/2},\end{aligned}$$

where the second inequality uses (14) and the third inequality is another application of the Cauchy-Schwarz inequality. Since the right hand side converges to 0 by assumption, this concludes the proof. □

C Proof of Theorem 1

We restate Theorem 1 for convenience.

Theorem 1. *Suppose that Assumption 1 holds, that $\ell \in C^{2,\alpha}(\mathcal{X} \times \mathcal{Y})$ for some $\alpha \in (0,1)$ and that the initial measures $\mu_{x,0}, \mu_{y,0}$ have densities in $L^1(\mathcal{X}), L^1(\mathcal{Y})$. If a solution $(\mu_x(t), \mu_y(t))$ of the ERIWGF (7) converges in time, it must converge to the point $(\hat{\mu}_x, \hat{\mu}_y)$ which is the unique fixed point of the problem*

$$\rho_x(x) = \frac{1}{Z_x} e^{-\beta \int \ell(x,y)\ d\mu_y(y)}, \quad \rho_y(y) = \frac{1}{Z_y} e^{\beta \int \ell(x,y)\ d\mu_x(x)}.$$

$(\hat{\mu}_x, \hat{\mu}_y)$ is an ε-Nash equilibrium of the game given by $\mathcal{L}$ when $\beta \geqslant \frac{4}{\varepsilon}\log\left(2\frac{1-V_\delta}{V_\delta}(2K_\ell/\varepsilon - 1)\right)$, where $K_\ell := \max_{x,y}\ell(x,y) - \min_{x,y}\ell(x,y)$ is the length of the range of ℓ, $\delta := \varepsilon/(2Lip(\ell))$ and V_δ is a lower bound on the volume of a ball of radius δ in $\mathcal{X}, \mathcal{Y}$.

Theorem 1 is a consequence of the following three results, which we prove separately.

Theorem 4. *Assume $\mathcal{X}, \mathcal{Y}$ are compact Polish metric spaces equipped with canonical Borel measures, and that ℓ is a continuous function on $\mathcal{X} \times \mathcal{Y}$. Let us consider the fixed point problem*

$$\begin{cases}\rho_x(x) &= \frac{1}{Z_x} e^{-\beta \int \ell(x,y)\ d\mu_y(y)}, \\ \rho_y(y) &= \frac{1}{Z_y} e^{\beta \int \ell(x,y)\ d\mu_x(x)},\end{cases}$$

where Z_x and Z_y are normalization constants and ρ_x, ρ_y are the densities of μ_x, μ_y. This fixed point problem has a unique solution $(\hat{\mu}_x, \hat{\mu}_y)$ that is also the unique Nash equilibrium of the game given by $\mathcal{L}_\beta(\mu_x, \mu_y) \triangleq \mathcal{L}(\mu_x, \mu_y) + \beta^{-1}(H(\mu_y) - H(\mu_x))$.

Theorem 5. *Let $K_\ell := \max_{x,y}\ell(x,y) - \min_{x,y}\ell(x,y)$ be the length of the range of ℓ. Let $\varepsilon > 0$, $\delta := \varepsilon/(2Lip(\ell))$ and V_δ be a lower bound on the volume of a ball of radius δ in $\mathcal{X}, \mathcal{Y}$. Then the solution $(\hat{\mu}_x, \hat{\mu}_y)$ of (9) is an ε-Nash equilibrium of the game given by $\mathcal{L}$ when*

$$\beta \geqslant \frac{4}{\varepsilon}\log\left(2\frac{1-V_\delta}{V_\delta}(2K_\ell/\varepsilon - 1)\right).$$

Theorem 6. *Suppose that Assumption 1 holds and $\ell \in C^{2,\alpha}(\mathcal{X} \times \mathcal{Y})$ for some $\alpha \in (0,1)$, i.e. the second derivatives of ℓ are α-Hölder. Then, there exists only one stationary solution of the ERIWGF (7) and it is the solution of the fixed point problem (9).*

C.1 Proof of Theorem 4: Preliminaries

Definition 2 (Upper hemicontinuity). *A set-valued function $\varphi : X \to 2^Y$ is upper hemicontinuous if for every open set $W \subset Y$, the set $\{x|\varphi(x) \subset W\}$ is open.*

Alternatively, set-valued functions can be seen as correspondences $\Gamma : X \to Y$. The graph of Γ is $\mathrm{Gr}(\Gamma) = \{(a, b) \in X \times Y | b \in \Gamma(a)\}$. If Γ is upper hemicontinuous, then $\mathrm{Gr}(\Gamma)$ is closed. If Y is compact, the converse is also true.

Definition 3 (Kakutani map). *Let X and Y be topological vector spaces and $\varphi : X \to 2^Y$ be a set-valued function. If Y is convex, then φ is termed a Kakutani map if it is upper hemicontinuous and $\varphi(x)$ is non-empty, compact and convex for all $x \in X$.*

Theorem 7 (Kakutani-Glicksberg-Fan). *Let S be a non-empty, compact and convex subset of a Hausdorff locally convex topological vector space. Let $\varphi : S \to 2^S$ be a Kakutani map. Then φ has a fixed point.*

Definition 4 (Lower semi-continuity). *Suppose X is a topological space, x_0 is a point in X and $f : X \to \mathbb{R} \cup \{-\infty, \infty\}$ is an extended real-valued function. We say that f is lower semi-continuous (l.s.c.) at x_0 if for every $\varepsilon > 0$ there exists a neighborhood U of x_0 such that $f(x) \geqslant f(x_0) - \varepsilon$ for all x in U when $f(x_0) < +\infty$, and $f(x)$ tends to $+\infty$ as x tends towards x_0 when $f(x_0) = +\infty$.*

We can also characterize lower-semicontinuity in terms of level sets. A function is lower semi-continuous if and only if all of its lower level sets $\{x \in X : \ f(x) \leqslant \alpha\}$ are closed. This property will be useful.

Theorem 8 (Weierstrass theorem for l.s.c. functions). *Let $f : T \to (-\infty, +\infty]$ be a l.s.c. function on a compact Hausdorff topological space T. Then f attains its infimum over T, i.e. there exists a minimum of f in T.*

Proof. Proof. Let $\alpha_0 = \inf f(T)$. If $\alpha_0 = +\infty$, then f is infinite and the assertion trivially holds. Let $\alpha_0 < +\infty$. Then, for each real $\alpha > \alpha_0$, the set $\{f \leqslant \alpha\}$ is closed and nonempty. Any finite collection of such sets has a nonempty intersection. By compactness, also the set $\bigcap_{\alpha > \alpha_0} \{f \leqslant \alpha\} = \{f \leqslant \alpha_0\} = f^{-1}(\alpha_0)$ is nonempty. (In particular, this implies that α_0 is finite.) □

Remark 1. *By Prokhorov's theorem, since $\mathcal{X}$ and $\mathcal{Y}$ are compact separable metric spaces, $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ are compact in the topology of weak convergence.*

C.2 Proof of Theorem 4: Existence

Lemma 4 and 5 are intermediate results, and Lemma 6 shows existence of the solution.

Lemma 4. *For any $\mu_y \in \mathcal{P}(\mathcal{Y})$, $\mathcal{L}_\beta(\cdot, \mu_y) : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is lower semicontinuous, and it achieves a unique minimum in $\mathcal{P}(\mathcal{X})$. Moreover, the minimum $m_x(\mu_y)$ is absolutely continuous with respect to the Borel measure, it has full support and its density takes the form*

$$\frac{dm_x(\mu_y)}{dx}(x) = \frac{1}{Z_{\mu_y}} e^{-\beta \int L(x,y) d\mu_y}, \tag{15}$$

where Z_{μ_y} is a normalization constant.

Analogously, for any $\mu_x \in \mathcal{P}(\mathcal{X})$, $-\mathcal{L}_\beta(\mu_x, \cdot) : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ is lower semicontinuous, and it achieves a unique minimum in $\mathcal{P}(\mathcal{Y})$. The minimum $m_y(\mu_x)$ is absolutely continuous with respect to the Borel measure, it has full support and its density takes the form

$$\frac{dm_y(\mu_x)}{dy}(y) = \frac{1}{Z_{\mu_x}} e^{\beta \int L(x,y) d\mu_x},$$

where Z_{μ_x} is a normalization constant.

Proof. We will prove the result for $\mathcal{L}_\beta(\cdot, \mu_y)$, as the other one is analogous. Let dx denote the canonical Borel measure on $\mathcal{X}$, and let $\tilde{p}$ be the probability measure proportional to the canonical Borel measure, i.e. $\frac{d\tilde{p}}{dx} = \frac{1}{\text{vol}(\mathcal{X})}$. Notice that $\text{vol}(\mathcal{X})$ is by definition the value of the canonical Borel measure on the whole $\mathcal{X}$. We rewrite

$$\begin{aligned}
\mathcal{L}_\beta(\mu_x, \mu_y) &= \iint \ell(x,y) d\mu_y d\mu_x + \beta^{-1} \int \log\left(\frac{d\mu_x}{dx}\right) d\mu_x + \beta^{-1} H(\mu_y) \\
&= \iint \ell(x,y) d\mu_y d\mu_x + \beta^{-1} \int \log\left(\frac{d\mu_x}{d\tilde{p}} \frac{d\tilde{p}}{dx}\right) d\mu_x + \beta^{-1} H(\mu_y) \\
&= \iint \left(\ell(x,y) - \beta^{-1} \log\left(\text{vol}(\mathcal{X})\right)\right) d\mu_y d\mu_x + \beta^{-1} \int \log\left(\frac{d\mu_x}{d\tilde{p}}\right) d\mu_x + \beta^{-1} H(\mu_y)
\end{aligned}$$

Notice that the first term in the right hand side is a lower semi-continuous (in weak convergence topology) functional in μ_x when μ_y is fixed. That is because it is a linear functional in μ_x with a continuous integrand, which implies that it is continuous in the weak convergence topology. The second to last term can be seen as the relative entropy (or Kullback-Leibler divergence) between μ_x and $\tilde{p}$:

$$H_{\tilde{p}}(\mu_x) := \int \log\left(\frac{d\mu_x}{d\tilde{p}}\right) d\mu_x$$

The relative entropy $H_{\tilde{p}}(\mu_x)$ is a lower semi-continuous functional with respect to μ_x (see Theorem 1 of Posner [1975], which proves a stronger statement: joint semi-continuity with respect to both measures).

Therefore, we conclude that $\mathcal{L}_\beta(\cdot, \mu_y)$ (with $\mu_y \in \mathcal{P}(\mathcal{Y})$ fixed) is a l.s.c. functional on $\mathcal{P}(\mathcal{X})$. By Theorem 8 and using the compactness of $\mathcal{P}(\mathcal{X})$, there exists a minimum of $\mathcal{L}_\beta(\cdot, \mu_y)$ in $\mathcal{P}(\mathcal{X})$.

Denote a minimum of $\mathcal{L}_\beta(\cdot, \mu_y)$ by $\hat{\mu}_x$. $\hat{\mu}_x$ must be absolutely continuous, because otherwise $-\beta^{-1} H(\hat{\mu}_x)$ would take an infinite value. By the Euler-Lagrange equations for functionals on probability measures, a necessary condition for $\hat{\mu}_x$ to be a minimum of $\mathcal{L}_\beta(\cdot, \mu_y)$ is that the first variation $\frac{\delta \mathcal{L}_\beta(\cdot, \mu_y)}{\delta \mu_x}(\hat{\mu}_x)(x)$ must take a constant value for all $x \in \text{supp}(\hat{\mu}_x)$ and values larger or equal outside of $\text{supp}(\hat{\mu}_x)$. The intuition behind this is that otherwise a zero-mean signed measure with positive mass on the minimizers of $\frac{\delta \mathcal{L}_\beta(\cdot, \mu_y)}{\delta \mu_x}(\hat{\mu}_x)$ and negative mass on the maximizers would provide a direction of decrease of the functional. We compute the first variation at $\hat{\mu}_x$:

$$\begin{aligned}
\frac{\delta \mathcal{L}_\beta(\cdot, \mu_y)}{\delta \mu_x}(\hat{\mu}_x)(x) &= \frac{\delta}{\delta \mu_x}\left(\int L(x,y) d\mu_y d\mu_x - \beta^{-1} H(\hat{\mu}_x) + \beta^{-1} H(\mu_y)\right) \\
&= \int L(x,y) d\mu_y + \beta^{-1} \log\left(\frac{d\hat{\mu}_x}{dx}(x)\right),
\end{aligned}$$

We equate $\int \ell(x,y) d\mu_y + \beta^{-1} \log(\frac{d\hat{\mu}_x}{dx}(x)) = K,\ \forall x \in \text{supp}(\hat{\mu}_x)$, where K is a constant. The first variation must take values larger or equal than K outside of $\text{supp}(\hat{\mu}_x)$, but since $\log(\frac{d\hat{\mu}_x}{dx}(x)) = -\infty$ outside of $\text{supp}(\hat{\mu}_x)$, we obtain that $\text{supp}(\hat{\mu}_x) = \mathcal{X}$. Then, for all $x \in \mathcal{X}$,

$$\frac{d\hat{\mu}_x}{dx}(x) = e^{-\beta \int L(x,y) d\mu_y + \beta K} = \frac{1}{Z_{\mu_y}} e^{-\beta \int L(x,y) d\mu_y}$$

where Z_{μ_y} is a normalization constant obtained from imposing $\int \frac{d\hat{\mu}_x}{dx}(x)\, dx = \int 1\, d\hat{\mu}_x = 1$. Since the necessary condition for optimality specifies a unique measure and the minimum exists, we obtain that $m_x(\mu_y) = \hat{\mu}_x$ is the unique minimum. An analogous argument holds for $m_y(\hat{\mu}_x)$ □

Lemma 5. *Suppose that the measures $(\mu_{y,n})_{n\in\mathbb{N}}$ and μ_y are in $\mathcal{P}(\mathcal{Y})$. Recall the definition of $m_x : \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X})$ in equation (15). If $(\mu_{y,n})_{n\in\mathbb{N}}$ converges weakly to μ_y, then $(m_x(\mu_{y,n}))_{n\in\mathbb{N}}$ converges weakly to $m_x(\mu_y)$, i.e. m_x is a continuous mapping when we endow $\mathcal{P}(\mathcal{Y})$ and $\mathcal{P}(\mathcal{X})$ with their weak convergence topologies.*

The same thing holds for m_y and measures $(\mu_{x,n})_{n\in\mathbb{N}}$ and μ_x on $\mathcal{X}$.

Proof. Given $x \in \mathcal{X}$, we have $\int \ell(x,y)d\mu_{y,n} \to \int \ell(x,y)d\mu_y$, because $\ell(x,\cdot)$ is a continuous bounded function on $\mathcal{Y}$. By continuity of the exponential function, we have that for all $x \in \mathcal{X}$, $e^{-\beta \int \ell(x,y)d\mu_{y,n}} \to e^{-\beta \int \ell(x,y)d\mu_y}$. Using the dominated convergence theorem,

$$\int_{\mathcal{X}} e^{-\beta \int \ell(x,y)d\mu_{y,n}} dx \to \int_{\mathcal{X}} e^{-\beta \int \ell(x,y)d\mu_y} dx$$

We need to find a dominating function. It is easy, because $\forall n \in \mathbb{N}$, $\forall x \in \mathcal{X}$, $e^{-\beta \int \ell(x,y)d\mu_{y,n}} \leqslant e^{-\beta \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \ell(x,y)}$. And $\int_{\mathcal{X}} e^{-\beta \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \ell(x,y)} dx = e^{-\beta \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \ell(x,y)} \mathrm{vol}(\mathcal{X}) < \infty$. By the Portmanteau theorem, we just need to prove that for all continuity sets B of $m_x(\mu_y)$, we have $m_x(\mu_{y,n})(B) \to m_x(\mu_y)(B)$. This translates to

$$\frac{\int_B e^{-\beta \int \ell(x,y)d\mu_{y,n}} dx}{\int_{\mathcal{X}} e^{-\beta \int \ell(x,y)d\mu_{y,n}} dx} \to \frac{\int_B e^{-\beta \int \ell(x,y)d\mu_y} dx}{\int_{\mathcal{X}} e^{-\beta \int \ell(x,y)d\mu_y} dx}$$

We have proved that the denominators converge appropriately, and the numerator converges as well using the same reasoning with dominated convergence. And both the numerators and the denominators are positive and the numerator is always smaller denominator, the quotient must converge. □

Lemma 6. *There exists a solution of (9), which is the Nash equilibrium of the game given by $\mathcal{L}_\beta$.*

Proof. We use Theorem 7 on the set $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, with the map $m : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ given by $m(\mu_x, \mu_y) = (m_x(\mu_y), m_y(\mu_x))$. The only condition to check is upper hemicontinuity of m. By Lemma 5 we know that m_x, m_y are continuous, and since continuous functions are upper hemicontinuous as set valued functions, this concludes the argument. Indeed, we could have used Tychonoff's theorem, which is similar to Theorem 7 but for single-valued functions. □

C.3 Proof of Theorem 4: Uniqueness

Lemma 7. *The solution of (9) is unique.*

Proof. The argument is analogous to the proof of Theorem 2 of Rosen [1965]. Suppose $(\mu_{x,1}, \mu_{y,1})$ and $(\mu_{x,2}, \mu_{y,2})$ are two different solutions of (9). We use the notation $F_1(\mu_x, \mu_y) = \mathcal{L}_\beta(\mu_x, \mu_y), F_2(\mu_x, \mu_y) = -\mathcal{L}_\beta(\mu_x, \mu_y)$. Hence, there exist constants $K_{x,1}, K_{y,1}, K_{x,2}, K_{y,2}$ such that

$$\frac{\delta F_1}{\delta \mu_x}(\mu_{x,1}, \mu_{y,1})(x) + K_{x,1} = 0, \frac{\delta F_2}{\delta \mu_y}(\mu_{x,1}, \mu_{y,1})(y) + K_{y,1} = 0,$$
$$\frac{\delta F_1}{\delta \mu_x}(\mu_{x,2}, \mu_{y,2})(x) + K_{x,2} = 0, \frac{\delta F_2}{\delta \mu_y}(\mu_{x,2}, \mu_{y,2})(y) + K_{y,2} = 0$$

On the one hand, we know that

$$\begin{aligned} &\int \frac{\delta F_1}{\delta \mu_x}(\mu_{x,1}, \mu_{y,1})(x)\ d(\mu_{x,2} - \mu_{x,1}) + \int \frac{\delta F_2}{\delta \mu_y}(\mu_{x,1}, \mu_{y,1})(y)\ d(\mu_{y,2} - \mu_{y,1}) \\ &+ \int \frac{\delta F_1}{\delta \mu_x}(\mu_{x,2}, \mu_{y,2})(x)\ d(\mu_{x,1} - \mu_{x,2}) + \int \frac{\delta F_2}{\delta \mu_y}(\mu_{x,2}, \mu_{y,2})(y)\ d(\mu_{y,1} - \mu_{y,2}) \\ =& - \int K_{x,1}\ d(\mu_{x,2} - \mu_{x,1}) - \int K_{y,1}\ d(\mu_{y,2} - \mu_{y,1}) \\ &- \int K_{x,2}\ d(\mu_{x,1} - \mu_{x,2}) - \int K_{y,2}\ d(\mu_{y,1} - \mu_{y,2}) = 0 \end{aligned} \tag{16}$$

We will now prove that the left hand side of (16) must be strictly larger than 0, reaching a contradiction. We can write

$$\begin{aligned}\frac{\delta F_1}{\delta \mu_x}(\mu_{x,2},\mu_{y,2})(x) - \frac{\delta F_1}{\delta \mu_x}(\mu_{x,1},\mu_{y,1})(x) &= \int L(x,y)\ d(\mu_{y,2}-\mu_{y,1}) \\ &\quad + \beta^{-1}(\log(\mu_{x,2}(x)) - \log(\mu_{x,1}(x))), \\ \frac{\delta F_2}{\delta \mu_y}(\mu_{x,2},\mu_{y,2})(x) - \frac{\delta F_2}{\delta \mu_y}(\mu_{x,1},\mu_{y,1})(x) &= -\int L(x,y)\ d(\mu_{x,2}-\mu_{x,1}) \\ &\quad + \beta^{-1}(\log(\mu_{y,2}(x)) - \log(\mu_{y,1}(x)))\end{aligned}$$

Hence, we rewrite the left hand side of (16) as

$$\begin{aligned}&\iint L(x,y)\ d(\mu_{y,2}-\mu_{y,1})d(\mu_{x,2}-\mu_{x,1}) + \beta^{-1}\int (\log(\mu_{x,2}(x)) - \log(\mu_{x,1}(x)))\ d(\mu_{x,2}-\mu_{x,1}) \\ -&\iint L(x,y)\ d(\mu_{x,2}-\mu_{x,1})d(\mu_{y,2}-\mu_{y,1}) + \beta^{-1}\int (\log(\mu_{y,2}(x)) - \log(\mu_{y,1}(x)))\ d(\mu_{y,2}-\mu_{y,1}) \\ &= \beta^{-1}(H_{\mu_{x,1}}(\mu_{x,2}) + H_{\mu_{x,2}}(\mu_{x,1}) + H_{\mu_{y,1}}(\mu_{y,2}) + H_{\mu_{y,1}}(\mu_{y,2})).\end{aligned}$$

Since the relative entropy is always non-negative and zero only if the two measures are equal, we have reached the desired contradiction. □

C.4 Proof of Theorem 5

We will use the shorthand $V_x(x) = V_x(\hat{\mu}_y)(x) = \int \mathcal{L}(x,y)d\hat{\mu}_y$, $V_y(y) = V_y(\hat{\mu}_x)(y) = \int \mathcal{L}(x,y)d\hat{\mu}_x$. Since $\ell : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ is a continuous function on a compact metric space, it is uniformly continuous. Hence,

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ st. } \sqrt{d(x,x')^2 + d(y,y')^2} < \delta \implies |\ell(x,y) - \ell(x',y')| < \varepsilon$$

Which means that

$$d(x,x') < \delta \implies |V_x(x) - V_x(x')| = \left|\int (\ell(x,y) - \ell(x',y))dy\right| < \varepsilon$$

This proves that V_x is uniformly continuous on $\mathcal{X}$ (and V_y is uniformly continuous on $\mathcal{Y}$ using the same argument).

We can write the Nikaido-Isoda function of the game with loss $\mathcal{L}$ (equation (2)) evaluated at $(\hat{\mu}_x, \hat{\mu}_y)$ as

$$\begin{aligned}\mathrm{NI}(\hat{\mu}_x,\hat{\mu}_y) &:= \mathcal{L}(\hat{\mu}_x,\hat{\mu}_y) - \min_{\mu'_x}\{\mathcal{L}(\mu'_x,\hat{\mu}_y)\} + (-\mathcal{L}(\hat{\mu}_x,\hat{\mu}_y) + \max_{\mu'_y}\{\mathcal{L}(\hat{\mu}_x,\mu'_y)\}) \\ &= \frac{\int V_x(x)e^{-\beta V_x(x)}dx}{\int e^{-\beta V_x(x)}dx} - \min_{x\in\mathcal{C}_1} V_x(x) + \frac{-\int V_y(y)e^{\beta V_y(y)}dy}{\int e^{\beta V_y(y)}dy} + \max_{y\in\mathcal{C}_2} V_y(y)\end{aligned} \tag{17}$$

The second equality follows from the definitions of $\mathcal{L}, V_x, V_y$. We observe that in the right-most expression the first two terms and the last two terms are analogous. Let us show the first two terms can be made smaller than an arbitrary $\varepsilon > 0$ by taking β large enough; the last two will be dealt with in an analogous manner. Let us define $\tilde{V}_x(x) = V_x(x) - \min_{x'\in\mathcal{C}_1} V_x(x')$.

$$\begin{aligned}&\frac{\int V_x(x)e^{-\beta V_x(x)}dx}{\int e^{-\beta V_x(x)}dx} - \min_{x\in\mathcal{C}_1} V_x(x) = \frac{\int (V_x(x) - \min_{x'\in\mathcal{C}_1} V_x(x'))e^{-\beta V_x(x)}dx}{\int e^{-\beta V_x(x)}dx} \\ &= \frac{\int \tilde{V}_x(x)e^{-\beta V_x(x)}\left(\mathbb{1}_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + \mathbb{1}_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}} + \mathbb{1}_{\{\varepsilon<\tilde{V}_x(x)\}}\right)dx}{\int e^{-\beta V_x(x)}\mathbb{1}_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}}dx + \int e^{-\beta V_x(x)}\mathbb{1}_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}}dx + \int e^{-\beta V_x(x)}\mathbb{1}_{\{\varepsilon<\tilde{V}_x(x)\}}dx}\end{aligned} \tag{18}$$

Let us define

$$q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} = \int e^{-\beta V_x(x)} \mathbb{1}_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} dx,$$

and $q_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}}$ and $q_{\{\varepsilon<\tilde{V}_x(x)\}}$ analogously.

Similarly, let

$$r_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} = \int \tilde{V}_x(x) e^{-\beta V_x(x)} \mathbb{1}_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} dx,$$

and $r_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}}$ and $r_{\{\varepsilon<\tilde{V}_x(x)\}}$ analogously.

Let

$$\tilde{p} = \frac{q_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}}}{q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + q_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}} + q_{\{\varepsilon<\tilde{V}_x(x)\}}}$$

Then, we can rewrite the right-most expression of (18) as

$$\begin{aligned} &\frac{r_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + r_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}} + r_{\{\varepsilon<\tilde{V}_x(x)\}}}{q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + q_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}} + q_{\{\varepsilon<\tilde{V}_x(x)\}}} \\ &= \tilde{p}\frac{r_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}}}{q_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}}} + (1-\tilde{p})\frac{r_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + r_{\{\varepsilon<\tilde{V}_x(x)\}}}{q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + q_{\{\varepsilon<\tilde{V}_x(x)\}}} \end{aligned} \tag{19}$$

Since $\tilde{V}(x) \leqslant \varepsilon$ in the set $\{x|\varepsilon/2 < \tilde{V}_x(x) \leqslant \varepsilon\}$, $r_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}}/q_{\{\varepsilon/2<\tilde{V}_x(x)\leqslant\varepsilon\}} \leqslant \varepsilon$.

Let $x_{\min}$ be such that $V(x_{\min}) = \min_{x\in C_1} V(x)$ (possibly not unique). By uniform continuity of V_x, we know there exists $\delta > 0$ (dependent only on ε) such that $B(x_{\min}, \delta) \subseteq \{x|\tilde{V}_x(x) \leqslant \varepsilon/2\}$. The following inequalities hold:

$$\begin{aligned} r_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} &\leqslant \frac{\varepsilon}{2} q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}}, \\ r_{\{\varepsilon<\tilde{V}_x(x)\}} &\leqslant (\max_{x\in\mathcal{C}_1} V_x(x) - \min_{x\in\mathcal{C}_1} V_x(x)) q_{\{\varepsilon<\tilde{V}_x(x)\}} \leqslant (\max_{x,y} L(x,y) - \min_{x,y} L(x,y)) q_{\{\varepsilon<\tilde{V}_x(x)\}} \\ &= K_L q_{\{\varepsilon<\tilde{V}_x(x)\}}. \end{aligned} \tag{20}$$

where we define $K_\ell = \max_{x,y} \ell(x,y) - \min_{x,y} \ell(x,y)$. Using (20), we obtain

$$\frac{r_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + r_{\{\varepsilon<\tilde{V}_x(x)\}}}{q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + q_{\{\varepsilon<\tilde{V}_x(x)\}}} \leqslant \frac{\frac{\varepsilon}{2} q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + K_L q_{\{\varepsilon<\tilde{V}_x(x)\}}}{q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} + q_{\{\varepsilon<\tilde{V}_x(x)\}}}.$$

If the right-hand side is smaller or equal than ε, then equation (19) would be smaller than ε and the proof would be concluded. For that to happen, we need $(K_\ell - \varepsilon) q_{\{\varepsilon<\tilde{V}_x(x)\}} \leqslant \frac{\varepsilon}{2} q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} \iff q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}}/q_{\{\varepsilon<\tilde{V}_x(x)\}} \geqslant 2(K_\ell/\varepsilon - 1)$. The following bounds hold:

$$\begin{aligned} q_{\{\tilde{V}_x(x)\leqslant\varepsilon/2\}} &\geqslant \mathrm{Vol}(B(x_{\min},\delta)) e^{-\beta(\min_{x\in\mathcal{C}_1} V_x(x) + \varepsilon/2)}, \\ q_{\{\varepsilon<\tilde{V}_x(x)\}} &\leqslant (1 - \mathrm{Vol}(B(x_{\min},\delta))) e^{-\beta(\min_{x\in\mathcal{C}_1} V_x(x) + \varepsilon)}. \end{aligned}$$

Thus, the following condition is sufficient:

$$\frac{\mathrm{Vol}(B(x_{\min},\delta))}{1 - \mathrm{Vol}(B(x_{\min},\delta))} e^{\beta\varepsilon/2} \geqslant 2(K_L/\varepsilon - 1).$$

Hence, if we take

$$\beta \geqslant \frac{2}{\varepsilon} \log\left(2 \frac{1 - \mathrm{Vol}(B(x_{\min},\delta))}{\mathrm{Vol}(B(x_{\min},\delta))} (K_L/\varepsilon - 1) \right) \tag{21}$$

then $(\hat{\mu}_x, \hat{\mu}_y)$ is an ε-Nash equilibrium. Since we have only bound the first two terms in the right hand side of (17) and the other two are bounded in the same manner, the statement of the theorem results from setting $\varepsilon = \varepsilon/2$ in (21).

C.5 Proof of Theorem 6

First, we show that any pair $\hat{\mu}_x, \hat{\mu}_y$ such that

$$\frac{d\hat{\mu}_x}{dx}(x) = \frac{1}{Z_x} e^{-\beta \int \ell(x,y)\ d\hat{\mu}_y(y)}, \quad \frac{d\hat{\mu}_y}{dy}(y) = \frac{1}{Z_y} e^{\beta \int \ell(x,y)\ d\hat{\mu}_x(x)}$$

is a stationary solution of (7). Denoting the Radon-Nikodym derivatives $\frac{d\hat{\mu}_x}{dx}, \frac{d\hat{\mu}_y}{dy}$ by $\hat{\rho}_x, \hat{\rho}_y$, it is sufficient to see that

$$\begin{cases} 0 = \nabla_x \cdot (\hat{\rho}_x \nabla_x V_x(\mu_y, x)) + \beta^{-1} \Delta_x \hat{\rho}_x \\ 0 = -\nabla_y \cdot (\hat{\rho}_y \nabla_y V_y(\mu_x, y)) + \beta^{-1} \Delta_y \hat{\rho}_y \end{cases} \tag{22}$$

holds weakly. And

$$\begin{aligned} \nabla_x \hat{\rho}_x &= \frac{1}{Z_x} e^{-\beta \int \ell(x,y)\ d\hat{\mu}_y(y)} \left(-\beta \nabla_x \int \ell(x,y)\ d\hat{\mu}_y(y) \right) = -\hat{\rho}_x \nabla_x V_x(\hat{\mu}_y, x), \\ \nabla_y \hat{\rho}_y &= \frac{1}{Z_y} e^{\beta \int \ell(x,y)\ d\hat{\mu}_x(x)} \left(\beta \nabla_y \int \ell(x,y)\ d\hat{\mu}_x(x) \right) = \hat{\rho}_y \nabla_y V_y(\hat{\mu}_x, y), \end{aligned}$$

implies that (22) holds.

Now we will prove the converse. Suppose that $\hat{\mu}_x, \hat{\mu}_y$ are (weak) stationary solutions of (7). That is, if $\varphi_x \in C^2(\mathcal{X}), \varphi_y \in C^2(\mathcal{Y})$ are arbitrary twice continuously differentiable functions, the following holds

$$\begin{aligned} 0 &= \int_{\mathcal{X}} \left(-\int_{\mathcal{Y}} \nabla_x \varphi_x(x) \cdot \nabla_x \ell(x,y)\ d\hat{\mu}_y + \beta^{-1} \Delta_x \varphi_x(x) \right) d\hat{\mu}_x \\ 0 &= \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} -\nabla_y \varphi_y(y) \cdot \nabla_y \ell(x,y)\ d\hat{\mu}_x - \beta^{-1} \Delta_y \varphi_y(x,y) \right) d\hat{\mu}_y \end{aligned} \tag{23}$$

(23) can be seen as two measure-valued stationary Fokker-Planck equations. We want to see that they have densities and that the densities satisfy the corresponding classical stationary Fokker-Planck equations (22). Works in the theory of PDEs have studied sufficient conditions for measure-valued stationary Fokker-Planck equations to correspond to weak stationary Fokker-Planck equations, and further to classical stationary Fokker-Planck equations. See page 3 of Huang et al. [2015] for a more detailed explanation on the two steps. That measure-valued stationary correspond to weak stationary solutions is shown in Theorem 2.2 of Bogachev et al. [2001]. That weak stationary solutions are classical stationary solutions requires that the drift term is in $C^{1,\alpha}_{\text{loc}}$ (locally α-Hölder continuous with exponent 1), meaning that it is in C^1 and that its derivatives are α-Hölder in compact sets. The result follows from the theory of Schauder estimates. Differentiating under the integral sign, the drift terms $-\int_{\mathcal{Y}} \nabla_x \ell(x,y)\ d\hat{\mu}_y, \int_{\mathcal{X}} \nabla_y \ell(x,y)\ d\hat{\mu}_x$ fulfill the condition if $\ell \in C^{2,\alpha}$.

D Proof of Theorem 2

Recall the expression of an *Interacting Wasserstein-Fisher-Rao Gradient Flow (IWFRGF)* in (8):

$$\begin{cases} \partial_t \mu_x &= \gamma \nabla \cdot (\mu_x \nabla_x V_x(\mu_y, x)) \\ & \quad -\alpha \mu_x (V_x(\mu_y, x) - \mathcal{L}(\mu_x, \mu_y)), \quad \mu_x(0) = \mu_{x,0} \\ \partial_t \mu_y &= -\gamma \nabla \cdot (\mu_y \nabla_y V_y(\mu_x, y)) \\ & \quad +\alpha \mu_y (V_y(\mu_x, y) - \mathcal{L}(\mu_x, \mu_y)), \quad \mu_y(0) = \mu_{y,0} \end{cases}$$

The aim is to obtain a global convergence result like the one in Theorem 3.8 of Chizat [2019]. First, we will rewrite Lemma 3.10 of Chizat [2019] in our case.

Lemma 8. *Let μ_x, μ_y be the solution of the IWFRGF in (8). Let $\mu_x^\star, \mu_y^\star$ be arbitrary measures on $\mathcal{X}, \mathcal{Y}$. Let $\bar{\mu}_x(t) = \frac{1}{t}\int_0^t \mu_x(s)\ ds$ and $\bar{\mu}_y(t) = \frac{1}{t}\int_0^t \mu_y(s)\ ds$. Let $\|\cdot\|_{BL}$ be the bounded Lipschitz norm, i.e. $\|f\|_{BL} = \|f\|_\infty + Lip(f)$. Let*

$$\mathcal{Q}_{\mu^\star,\mu_0}(\tau) = \inf_{\mu\in\mathcal{P}(\Theta)} \|\mu^\star - \mu\|_{BL}^* + \frac{1}{\tau}\mathcal{H}(\mu,\mu_0) \tag{24}$$

with $\Theta = \mathcal{X}$ or $\mathcal{Y}$. Let

$$B = \frac{1}{2}\left(\max_{x\in\mathcal{X},y\in\mathcal{Y}} \ell(x,y) - \min_{x\in\mathcal{X},y\in\mathcal{Y}} \ell(x,y)\right) + Lip(\ell) \tag{25}$$

Then,

$$\mathcal{L}(\bar{\mu}_x(t),\mu_y^\star) - \mathcal{L}(\mu_x^\star,\bar{\mu}_y(t)) \leqslant B\mathcal{Q}_{\mu_x^\star,\mu_{x,0}}(\alpha B t) + B\mathcal{Q}_{\mu_y^\star,\mu_{y,0}}(\alpha B t) + \gamma B^2 t \tag{26}$$

Proof. The proof is as in Lemma 3.10 of Chizat [2019], but in this case we have to do everything twice. Namely, we define the dynamics

$$\begin{aligned}\frac{d\mu_x^\varepsilon}{dt} &= \gamma\nabla\cdot(\mu_x^\varepsilon\nabla V_x(\mu_y,x))\\ \frac{d\mu_y^\varepsilon}{dt} &= -\gamma\nabla\cdot(\mu_y^\varepsilon\nabla V_y(\mu_x,y))\end{aligned}$$

initialized at $\mu_x^\varepsilon(0) = \mu_{x,0}^\varepsilon, \mu_y^\varepsilon(0) = \mu_{y,0}^\varepsilon$ arbitrary such that $\mu_{x,0}^\varepsilon$ and $\mu_{y,0}^\varepsilon$ are absolutely continuous with respect to $\mu_{x,0}$ and $\mu_{y,0}$ respectively.

Let us show that

$$\frac{1}{\alpha}\frac{d}{dt}\mathcal{H}(\mu_x^\varepsilon,\mu_x) = \int \frac{\delta\mathcal{L}}{\delta\mu_x}(\mu_x,\mu_y)(x)\ d(\mu_x^\varepsilon - \mu_x) \tag{27}$$

where $\mathcal{H}(\mu_x^\varepsilon,\mu_x)$ is the relative entropy, i.e.

$$\frac{d}{dt}\mathcal{H}(\mu_x^\varepsilon,\mu_x) = \frac{d}{dt}\int \log\left(\rho_x^\varepsilon\right)\ d\mu_x^\varepsilon,$$

ρ_x^ε being the Radon-Nikodym derivative $d\mu_x^\varepsilon/d\mu_x$.

Assume to begin with that μ_x^ε remains absolutely continuous with respect to μ_x through time. We can write

$$\frac{d}{dt}\int \varphi_x(x)\rho_x^\varepsilon(x)d\mu_x(x) = \frac{d}{dt}\int \varphi(x)d\mu_x^\varepsilon(x)$$

We can develop the left hand side into

$$\begin{aligned}\frac{d}{dt}\int \varphi_x(x)\rho_x^\varepsilon(x)d\mu_x(x) &= \int -\gamma\nabla(\varphi_x(x)\rho_x^\varepsilon(x))\cdot\nabla V_x(\mu_y,x)d\mu_x(x)\\ &+ \int -\alpha\varphi_x(x)\rho_x^\varepsilon(x)(V_x(\mu_y,x) - \mathcal{L}(\mu_x,\mu_y))d\mu_x(x)\\ &+ \int \varphi_x(x)\frac{\partial\rho_x^\varepsilon}{\partial t}(x)d\mu_x(x)\\ &= \int -\gamma(\nabla\varphi_x(x)\rho_x^\varepsilon(x) + \varphi_x(x)\nabla\rho_x^\varepsilon(x))\cdot\nabla V_x(\mu_y,x)\ d\mu_x(x)\\ &+ \int -\alpha\varphi_x(x)\rho_x^\varepsilon(x)(V_x(\mu_y,x) - \mathcal{L}(\mu_x,\mu_y))d\mu_x(x)\\ &+ \int \varphi_x(x)\frac{\partial\rho_x^\varepsilon}{\partial t}(x)d\mu_x(x)\end{aligned}$$

and the right hand side into

$$\frac{d}{dt}\int \varphi(x)d\mu_x^\varepsilon(x) = \int -\gamma\nabla\varphi_x(x)\cdot\nabla V_x(\mu_y,x)d\mu_x^\varepsilon(x)$$

Note that comparing terms, we obtain

$$\begin{aligned}&\int -\gamma\varphi_x(x)\nabla\rho_x^\varepsilon(x)\cdot\nabla V_x(\mu_y,x)\ d\mu_x(x)\\ &= \int \alpha\varphi_x(x)\rho_x^\varepsilon(x)(V_x(\mu_y,x)-\mathcal{L}(\mu_x,\mu_y)) - \varphi_x(x)\frac{\partial\rho_x^\varepsilon}{\partial t}(x)\ d\mu_x(x)\end{aligned}$$

Since φ_x is arbitrary, it must be that

$$-\gamma\nabla\rho_x^\varepsilon(x)\cdot\nabla V_x(\mu_y,x) = \alpha\rho_x^\varepsilon(x)(V_x(\mu_y,x)-\mathcal{L}(\mu_x,\mu_y)) - \frac{\partial}{\partial t}\rho_x^\varepsilon(x) \tag{28}$$

holds μ_x-almost everywhere. Now,

$$\begin{aligned}\frac{d}{dt}\int \log\left(\rho_x^\varepsilon\right)\ d\mu_x^\varepsilon &= -\gamma\int \nabla\left(\log\left(\rho_x^\varepsilon(x)\right)\right)\cdot\nabla V_x(\mu_y,x)\ d\mu_x^\varepsilon(x)\\ &= -\gamma\int \frac{1}{\rho_x^\varepsilon(x)}\nabla\left(\rho_x^\varepsilon(x)\right)\cdot\nabla V_x(\mu_y,x)\ d\mu_x^\varepsilon(x)\\ &= \alpha\int (V_x(\mu_y,x)-\mathcal{L}(\mu_x,\mu_y))\ d\mu_x^\varepsilon(x) - \int\frac{1}{\rho_x^\varepsilon(x)}\frac{\partial}{\partial t}\rho_x^\varepsilon(x)d\mu_x^\varepsilon(x)\end{aligned}$$

Here,

$$\int \frac{1}{\rho_x^\varepsilon(x)}\frac{\partial}{\partial t}\rho_x^\varepsilon(x)d\mu_x^\varepsilon(x) = \int\frac{\partial}{\partial t}\rho_x^\varepsilon(x)d\mu_x(x) = 0$$

And since

$$\mathcal{L}(\mu_x,\mu_y) = \int \frac{\delta\mathcal{L}}{\delta\mu_x}(\mu_x,\mu_y)(x)\ d\mu_x,$$

the first term yields (27). We assumed that ρ_x^ε existed and was regular enough. To make the argument precise, we can define the density of μ_x^ε with respect to μ_x to be a solution ρ_x^ε of (28), and thus specify μ_x^ε.

Now, recall that $\mu_x^\star$ is an arbitrary measure in $\mathcal{P}(\mathcal{X})$. By linearity of $\mathcal{L}$ with respect to μ_x,

$$\begin{aligned}\int \frac{\delta\mathcal{L}}{\delta\mu_x}(\mu_x,\mu_y)(x)\ d(\mu_x^\varepsilon-\mu_x) &= \int \frac{\delta\mathcal{L}}{\delta\mu_x}(\mu_x,\mu_y)(x)\ d(\mu_x^\star-\mu_x) + \int\frac{\delta\mathcal{L}}{\delta\mu_x}(\mu_x,\mu_y)(x)\ d(\mu_x^\varepsilon-\mu_x^\star)\\ &\leqslant -(\mathcal{L}(\mu_x,\mu_y)-\mathcal{L}(\mu_x^\star,\mu_y)) + \|\frac{\delta\mathcal{L}}{\delta\mu_x}(\mu_x,\mu_y)\|_{\mathrm{BL}}\|\mu_x^\varepsilon-\mu_x^\star\|_{\mathrm{BL}}^*\end{aligned} \tag{29}$$

Notice that we can take $\|\frac{\delta\mathcal{L}}{\delta\mu_x}(\mu_x,\mu_y)\|_{\mathrm{BL}}$ to be smaller than B (defined in (25)). If we integrate (27) and (29) from 0 to t and divide by t, we obtain

$$\begin{aligned}&\frac{1}{t}\int_0^t \mathcal{L}(\mu_x(s),\mu_y(s))\ ds - \frac{1}{t}\int_0^t \mathcal{L}(\mu_x^\star,\mu_y(s))\ ds\\ &\leqslant \frac{1}{\alpha t}(\mathcal{H}(\mu_{x,0}^\varepsilon,\mu_{x,0})-\mathcal{H}(\mu_x^\varepsilon(t),\mu_x(t))) + \frac{B}{t}\int_0^t \|\mu_x^\varepsilon-\mu_x^\star\|_{\mathrm{BL}}^*\ ds\end{aligned} \tag{30}$$

We bound the last term on the RHS:

$$\frac{B}{t}\int_0^t \|\mu_x^\varepsilon-\mu_x^\star\|_{\mathrm{BL}}^*\ ds \leqslant B\|\mu_{x,0}^\varepsilon-\mu_x^\star\|_{\mathrm{BL}}^* + \frac{B}{t}\int_0^t \|\mu_{x,0}^\varepsilon-\mu_x^\varepsilon\|_{\mathrm{BL}}^*\ ds \tag{31}$$

And

$$\begin{aligned}\|\mu_x^\varepsilon(t)-\mu_{x,0}^\varepsilon\|_{\mathrm{BL}}^* &= \sup_{\|f\|_{\mathrm{BL}}\leqslant 1, f\in C^2(\mathcal{X})} \int f\ d(\mu_x^\varepsilon(t)-\mu_{x,0}^\varepsilon) = \sup_{\|f\|_{\mathrm{BL}}\leqslant 1, f\in C^2(\mathcal{X})} \int_0^t \frac{d}{ds}\int f\ d\mu_x^\varepsilon(s)\ ds \\ &= \sup_{\|f\|_{\mathrm{BL}}\leqslant 1, f\in C^2(\mathcal{X})} -\int_0^t \int \gamma\nabla f(x)\cdot\nabla\frac{\delta\mathcal{L}}{\delta\mu_x}(\mu_x^\varepsilon,\mu_y)(x)\ d\mu_x^\varepsilon(s)\ ds \leqslant \int_0^t\int \gamma B\ d\mu_x^\varepsilon(s)\ ds = \gamma B t\end{aligned} \tag{32}$$

Also, by linearity of $\mathcal{L}$ with respect to μ_y,

$$-\frac{1}{t}\int_0^t \mathcal{L}(\mu_x^\star,\mu_y(s))\ ds = -\mathcal{L}(\mu_x^\star,\bar{\mu}_y(t)) \tag{33}$$

If we use (31), (32) and (33) and the non-negativeness of the relative entropy on (30), we obtain:

$$\frac{1}{t}\int_0^t \mathcal{L}(\mu_x(s),\mu_y(s))\ ds - \mathcal{L}(\mu_x^\star,\bar{\mu}_y(t)) \leqslant \frac{\mathcal{H}(\mu_{x,0}^\varepsilon,\mu_{x,0})}{4\alpha t} + B\|\mu_{x,0}^\varepsilon-\mu_x^\star\|_{\mathrm{BL}}^* + \frac{B^2\gamma}{2}t \tag{34}$$

$$-\frac{1}{t}\int_0^t \mathcal{L}(\mu_x(s),\mu_y(s))\ ds + \mathcal{L}(\bar{\mu}_x(t),\mu_y^\star) \leqslant \frac{\mathcal{H}(\mu_{y,0}^\varepsilon,\mu_{y,0})}{4\alpha t} + B\|\mu_{y,0}^\varepsilon-\mu_y^\star\|_{\mathrm{BL}}^* + \frac{B^2\gamma}{2}t \tag{35}$$

Equation (35) is obtained by performing the same argument switching the roles of x and y, and $\mathcal{L}$ by $-\mathcal{L}$. By adding equations (34) and (35) and considering the definition of $\mathcal{Q}$ in (24), we obtain the inequality (26).

□

Notice that by taking the supremum wrt $\mu_x^\star,\mu_y^\star$ on (26) we obtain a bound on the Nikaido-Isoda error of $(\bar{\mu}_x(t),\bar{\mu}_y(t))$ (see (2)).

Next, we will obtain a result like Lemma E.1 from Chizat [2019] in which we bound $\mathcal{Q}$. The proof is a variation of the argument in Lemma E.1 from Chizat [2019], as in our case no measures are necessarily sparse.

Lemma 9. *Let Θ be a Riemannian manifold of dimension d. Assume that $Vol(B_{\theta,\varepsilon}) \geqslant e^{-K}\varepsilon^d$ for all $\theta\in\Theta$, where the volume is defined of course in terms of the Borel measure[1] of Θ. If $\rho := \frac{d\mu_0}{d\theta}$ is the Radon-Nikodym derivative of μ_0 with respect to the Borel measure of Θ, assume that $\rho(\theta)\geqslant e^{-K'}$ for all $\theta\in\Theta$. The function $\mathcal{Q}_{\mu^\star,\mu_0}(\tau)$ defined in (24) can be bounded by*

$$\mathcal{Q}_{\mu^\star,\mu_0}(\tau) \leqslant \frac{d}{\tau}(1-\log d+\log\tau)+\frac{1}{\tau}(K+K')$$

Proof. We will choose μ^ε in order to bound the infimum. For $\theta\in\Theta,\varepsilon>0$, let $\xi_{\theta,\varepsilon}$ be a probability measure on Θ with support on the ball $B_{\theta,\varepsilon}$ of radius ε centered at θ and proportional to the Borel measure for all subsets of the ball. Let us define the measure

$$\mu^\varepsilon(A) = \int_\Theta \xi_{\theta,\varepsilon}(A)\ d\mu^\star(\theta)$$

for all Borel sets A of $\mathcal{X}$. Now, we can bound $\|\mu^\varepsilon-\mu^\star\|_{\mathrm{BL}}^* \leqslant W_1(\mu^\varepsilon,\mu^\star)$. Let us consider the coupling γ between μ^ε and $\mu^\star$ defined as:

$$\gamma(A\times B) = \int_A \xi_{\theta,\varepsilon}(B)\ d\mu^\star(\theta)$$

for A,B arbitrary Borel sets of Θ. Notice that γ is indeed a coupling between μ^ε and $\mu^\star$, because $\gamma(A\times\Theta) = \mu^\star(A)$ and $\gamma(\Theta\times B)=\mu^\varepsilon(B)$. Hence,

$$W_1(\mu^\varepsilon,\mu^\star) \leqslant \int_{\Theta\times\Theta} d_\Theta(\theta,\theta')\ d\gamma(\theta,\theta') = \int_\Theta \frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})}\int_{B_{\theta',\varepsilon}} d_\Theta(\theta,\theta')\ d\theta\ d\mu^\star(\theta') \tag{36}$$

[1] The metric of the manifold gives a natural choice of a Borel (volume) measure, the one given by integrating the canonical volume form.

where the inner integral is with respect to the Borel measure on Θ. Since $d_\Theta(\theta, \theta') \leqslant \varepsilon$ for all $\theta \in B_{\theta',\varepsilon}$, we conclude from that (36) that $W_1(\mu^\varepsilon, \mu^\star) \leqslant \varepsilon$.

Next, let us bound the relative entropy term. Define ρ_ε as the Radon-Nikodym derivative of μ^ε with respect to the Borel measure of Θ, i.e.

$$\rho_\varepsilon(\theta) := \frac{d\mu^\varepsilon}{d\theta}(\theta) = \int_\Theta \frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \mathbb{1}_{B_{\theta',\varepsilon}}(\theta)\ d\mu^\star(\theta').$$

Also, recall that $\rho := \frac{d\mu_0}{d\theta}$. Then, we write

$$\mathcal{H}(\mu^\varepsilon, \mu_0) = \int_\Theta \log \frac{\rho_\varepsilon}{\rho} d\mu^\varepsilon = \int_\Theta \log(\rho_\varepsilon)\rho_\varepsilon d\theta - \int_\Theta \log(\rho)\rho_\varepsilon d\theta. \tag{37}$$

On the one hand, we use the convexity of the function $x \to x \log x$:

$$\begin{aligned}
\rho_\varepsilon(\theta) \log \rho_\varepsilon(\theta) &= \left(\int_\Theta \frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \mathbb{1}_{B_{\theta',\varepsilon}}(\theta)\ d\mu^\star(\theta') \right) \log \left(\int_\Theta \frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \mathbb{1}_{B_{\theta',\varepsilon}}(\theta)\ d\mu^\star(\theta') \right) \\
&\leqslant \int_\Theta \left(\frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \mathbb{1}_{B_{\theta',\varepsilon}}(\theta) \right) \log \left(\frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \mathbb{1}_{B_{\theta',\varepsilon}}(\theta) \right) d\mu^\star(\theta').
\end{aligned}$$

We use Fubini's theorem:

$$\begin{aligned}
&\int_\Theta \rho_\varepsilon(\theta) \log \rho_\varepsilon(\theta)\ d\theta \leqslant \int_\Theta \int_\Theta \left(\frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \mathbb{1}_{B_{\theta',\varepsilon}}(\theta) \right) \log \left(\frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \mathbb{1}_{B_{\theta',\varepsilon}}(\theta) \right)\ d\theta\ d\mu^\star(\theta') \\
&= \int_\Theta \frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \int_{B_{\theta',\varepsilon}} -\log\left(\mathrm{Vol}(B_{\theta',\varepsilon})\right)\ d\theta\ d\mu^\star(\theta') = -\int_\Theta \log\left(\mathrm{Vol}(B_{\theta',\varepsilon})\right) d\mu^\star(\theta') \\
&\leqslant -d \log \varepsilon + K
\end{aligned} \tag{38}$$

where d is the dimension of Θ and K is a constant such that $\mathrm{Vol}(B_{\theta',\varepsilon}) \geqslant e^{-K}\varepsilon^d$ for all $\theta' \in \Theta$.

On the other hand,

$$\begin{aligned}
-\int_\Theta \log(\rho(\theta))\rho_\varepsilon(\theta)\ d\theta &= \int_\Theta \frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \int_{\mathrm{Vol}(B_{\theta',\varepsilon})} -\log(\rho(\theta))\ d\theta\ d\mu^\star(\theta') \\
&\leqslant \int_\Theta \frac{1}{\mathrm{Vol}(B_{\theta',\varepsilon})} \int_{\mathrm{Vol}(B_{\theta',\varepsilon})} K'\ d\theta\ d\mu^\star(\theta') = K'
\end{aligned} \tag{39}$$

where K' is defined such that $\rho(\theta) \geqslant e^{-K'}$ for all $\theta \in \Theta$.

By plugging (38) and (39) into (37) we obtain:

$$\|\mu^\star - \mu^\varepsilon\|_{\mathrm{BL}}^* + \frac{1}{\tau}\mathcal{H}(\mu^\varepsilon, \mu_0) \leqslant \varepsilon + \frac{1}{\tau}(-d \log \varepsilon + K + K').$$

If we optimize the bound with respect to ε we obtain the final result. □

Theorem 2. *Let $\varepsilon > 0$ arbitrary. Suppose that $\mu_{x,0}, \mu_{y,0}$ are such that their Radon-Nikodym derivatives with respect to the Borel measures of $\mathcal{X}, \mathcal{Y}$ are lower-bounded by $e^{-K'_x}, e^{-K'_y}$ respectively. For any $\delta \in (0, 1/2)$, there exists a constant $C_{\delta,\mathcal{X},\mathcal{Y},K'_x,K'_y} > 0$ depending on the dimensions of $\mathcal{X}, \mathcal{Y}$, their curvatures and K'_x, K'_y, such that if $\gamma/\alpha < 1$ and*

$$\frac{\gamma}{\alpha} \leqslant \left(\frac{\varepsilon}{C_{\delta,\mathcal{X},\mathcal{Y},K'_x,K'_y}} \right)^{\frac{2}{1-\delta}} \tag{40}$$

Then, at $t_0 = (\alpha\gamma)^{-1/2}$ we have

$$NI(\bar{\mu}_x(t_0), \bar{\mu}_y(t_0)) := \sup_{\mu_x^\star, \mu_y^\star} \mathcal{L}(\bar{\mu}_x(t_0), \mu_y^\star) - \mathcal{L}(\mu_x^\star, \bar{\mu}_y(t_0)) \leqslant \varepsilon$$

Proof. We plug the bound of Theorem 9 into the result of Theorem 8, obtaining

$$\begin{aligned}\mathcal{L}(\bar{\mu}_x(t), \mu_y^\star) - \mathcal{L}(\mu_x^\star, \bar{\mu}_y(t)) \leqslant & \frac{d_x}{\alpha t}(1 - \log d_x + \log(\alpha B t)) \\ & + \frac{d_y}{\alpha t}(1 - \log d_y + \log(\alpha B t)) \\ & + \frac{1}{\alpha t}(K_x + K_x' + K_y + K_y') + \gamma B^2 t\end{aligned}$$

Now, we set $t = (\alpha\gamma)^{-1/2}$, and thus the right hand side becomes

$$\sqrt{\frac{\gamma}{\alpha}}\left(d_x\left(1 - \log\frac{d_x}{B} + \log\sqrt{\frac{\alpha}{\gamma}}\right) + d_y\left(1 - \log\frac{d_y}{B} + \log\sqrt{\frac{\alpha}{\gamma}}\right) + K_x + K_x' + K_y + K_y' + B^2\right) \tag{41}$$

Let $\varepsilon > 0$ arbitrary. We want (41) to be lower or equal than ε. For any δ such that $0 < \delta < 1/2$, there exists C_δ such that $\log(x) \leqslant C_\delta x^\delta$. This yields

$$\begin{aligned}&\sqrt{\frac{\gamma}{\alpha}}\left(d_x\left(1 - \log\frac{d_x}{B} + C_\delta\left(\frac{\alpha}{\gamma}\right)^{-\delta/2}\right) + d_y\left(1 - \log\frac{d_y}{B} + C_\delta\left(\frac{\alpha}{\gamma}\right)^{-\delta/2}\right)\right) \\ &+ \sqrt{\frac{\gamma}{\alpha}}\left(K_x + K_x' + K_y + K_y' + B^2\right)\end{aligned} \tag{42}$$

If we set $\gamma < \alpha$, $(\gamma/\alpha)^{-\delta/2} > 1$ then (42) is upper-bounded by

$$\left(\frac{\gamma}{\alpha}\right)^{\frac{1-\delta}{2}}\left(d_x(1 - \log\frac{d_x}{B} + C_\delta) + d_y(1 - \log\frac{d_y}{B} + C_\delta) + K_x + K_x' + K_y + K_y' + B^2\right)$$

If we bound this by ε, we obtain the bound in (40). □

Corollary 1. *Let $(\mathcal{X}_{d_x}, \mathcal{Y}_{d_y}, l_{d_x,d_y})_{d_x\in\mathbb{N}, d_y\in\mathbb{N}}$ be a family indexed by $\mathbb{N}^2$. Assume that $\mu_{x,0}, \mu_{y,0}$ are set to be the Borel measures in $\mathcal{X}_{d_x}, \mathcal{Y}_{d_y}$, that $\mathcal{X}_{d_x}, \mathcal{Y}_{d_y}$ are locally isometric to the d_x, d_y-dimensional Euclidean spaces, and that the volumes of $\mathcal{X}_{d_x}$, $\mathcal{Y}_{d_y}$ grow no faster than exponentially on the dimensions d_x, d_y. Assume that l_{d_x,d_y} are such that B is constant. Then, we can rewrite (40) as*

$$\frac{\gamma}{\alpha} \leqslant O\left(\left(\frac{\varepsilon}{(d_x + d_y)\log(B) + d_x\log(d_x) + d_y\log(d_y) + B^2}\right)^{\frac{2}{1-\delta}}\right)$$

Proof. The volume of n-dimensional ball of radius r in n-dimensional Euclidean space is

$$V_n(r) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)}R^n,$$

and hence, if $\mathcal{X}, \mathcal{Y}$ are locally isometric to the d_x and d_y-dimensional Euclidean spaces we can take

$$\begin{aligned}K_x &= \log\Gamma\left(\frac{d_x}{2} + 1\right) - \frac{d_x}{2}\log(\pi) \leqslant \left(\frac{d_x}{2} + 1\right)\log\left(\frac{d_x}{2} + 1\right) - \frac{d_x}{2}\log(\pi) \leqslant O(d_x\log d_x) \\ K_y &= \log\Gamma(\frac{d_y}{2} + 1) - \frac{n}{2}\log(\pi) \leqslant O(d_x\log d_x)\end{aligned}$$

If the volumes of $\mathcal{X}, \mathcal{Y}$ grow no faster than an exponential of the dimensions d_x, d_y and we take $\mu_{x,0}, \mu_{y,0}$ to be the Borel measures, we can take $K_x' = \log(\mathrm{Vol}(\mathcal{X})), K_y' = \log(\mathrm{Vol}(\mathcal{Y}))$ to be constant with respect to the dimensions d_x, d_y. □

E Proof of Theorem 3(i)

E.1 Preliminaries

Throughout the section we will use the techniques shown in Subsec. G.5 to deal with SDEs on manifolds. Effectively, this means that for SDEs we have additional drift terms $\hat{\mathbf{h}}_x$ or $\hat{\mathbf{h}}_x$ induced by the geometry of the manifold, and that we must project the variations of the Brownian motion onto the tangent space.

Define the processes $\mathbf{X}^n = (X^1, \dots, X^n)$ and $\mathbf{Y}^n = (Y^1, \dots, Y^n)$ such that for all $i \in \{1, \dots, n\}$,

$$\begin{aligned} dX_t^i &= \left(-\frac{1}{n} \sum_{j=1}^n \nabla_x \ell(X_t^i, Y_t^j) + \hat{\mathbf{h}}_x(X_t^i) \right) dt + \sqrt{2\beta^{-1}} \operatorname{Proj}_{T_{X_t^i}\mathcal{X}}(dW_t^i), \quad X_0^{n,i} = \xi^i \sim \mu_{x,0} \\ dY_t^i &= \left(\frac{1}{n} \sum_{j=1}^n \nabla_y \ell(X_t^j, Y_t^i) + \hat{\mathbf{h}}_y(Y_t^i) \right) dt + \sqrt{2\beta^{-1}} \operatorname{Proj}_{T_{Y_t^i}\mathcal{Y}}(d\bar{W}_t^i), \quad Y_0^{n,i} = \bar{\xi}^i \sim \mu_{y,0} \end{aligned} \tag{43}$$

where $\mathbf{W}_t = (W_t^1, \dots, W_t^n)$, and $\bar{\mathbf{W}}_t = (\bar{W}_t^1, \dots, \bar{W}_t^n)$ are Brownian motions on $\mathbb{R}^{nD_x}$ and $\mathbb{R}^{nD_y}$ respectively. Notice that $\mathbf{X}_t$ is valued in $\mathcal{X}^n \subseteq \mathbb{R}^{nD_x}$ and $\mathbf{Y}_t$ is valued in $\mathcal{Y}^n \subseteq \mathbb{R}^{nD_y}$. (43) can be seen as a system of $2n$ interacting particles in which each particle of one player interacts with all the particles of the other one. It also corresponds to noisy continuous-time mirror descent on parameter spaces for an augmented game in which there are n replicas of each player, choosing $\frac{1}{2}\|\cdot\|_2^2$ for the mirror map.

Now, define $\tilde{\mathbf{X}} = (\tilde{X}^1, \dots, \tilde{X}^n)$ and $\tilde{\mathbf{Y}} = (\tilde{Y}^1, \dots, \tilde{Y}^n)$ for all $i \in \{1, \dots, n\}$ let

$$\begin{aligned} d\tilde{X}_t^i &= \left(-\int_{\mathcal{Y}} \nabla_x \ell(\tilde{X}_t^i, y)\, d\mu_{y,t} + \hat{\mathbf{h}}_x(\tilde{X}_t^i) \right) dt + \sqrt{2\beta^{-1}} \operatorname{Proj}_{T_{\tilde{X}_t^i}\mathcal{X}}(dW_t^i), \\ d\tilde{Y}_t^i &= \left(\int_{\mathcal{X}} \nabla_y \ell(x, \tilde{Y}_t^i)\, d\mu_{x,t} + \hat{\mathbf{h}}_y(\tilde{Y}_t^i) \right) dt + \sqrt{2\beta^{-1}} \operatorname{Proj}_{T_{\tilde{Y}_t^i}\mathcal{Y}}(d\bar{W}_t^i), \\ \tilde{X}_0^i &= \xi^i \sim \mu_{x,0}, \quad \mu_{y,t} = \operatorname{Law}(\tilde{Y}_t^i), \quad \tilde{Y}_0^i = \bar{\xi}^i \sim \mu_{y,0}, \quad \mu_{x,t} = \operatorname{Law}(\tilde{X}_t^i) \end{aligned} \tag{44}$$

Lemma 10 (Forward Kolmogorov equation)**.** *The laws $(\mu_x)_{t\in[0,T]}, (\mu_y)_{t\in[0,T]}$ of a solution $\tilde{X}, \tilde{Y}$ of (44) with $n = 1$ (seen as elements of $\mathcal{C}([0,T], \mathcal{P}(\mathcal{X})), \mathcal{C}([0,T], \mathcal{P}(\mathcal{Y})))$ are a solution of (45).*

$$\begin{cases} \partial_t \mu_x = \nabla_x \cdot (\mu_x \nabla_x V_x(\mu_y, x)) + \beta^{-1} \Delta_x \mu_x, & \mu_x(0) = \mu_{x,0} \\ \partial_t \mu_y = -\nabla_y \cdot (\mu_y \nabla_y V_y(\mu_x, y)) + \beta^{-1} \Delta_y \mu_y, & \mu_y(0) = \mu_{y,0} \end{cases} \tag{45}$$

Proof. We sketch the derivation for the forward Kolmogorov equation on manifolds. First, we define the semigroups

$$P_t^x \varphi_x(x) = \mathbb{E}[\varphi_x(\tilde{X}_t) | \tilde{X}_0 = x], \quad P_t^y \varphi_y(y) = \mathbb{E}[\varphi_y(\tilde{Y}_t) | \tilde{Y}_0 = y],$$

where $\tilde{X}, \tilde{Y}$ are solutions of (44) with $n = 1$. We obtain that if $\mathcal{L}_t^x, \mathcal{L}_t^y$ are the infinitesimal generators (i.e., $\mathcal{L}_t^x \varphi_x(x) = \lim_{t\to 0^+} \frac{1}{t}(P_t^x \varphi_x(x) - \varphi_x(x))$), the backward Kolmogorov equations $\frac{d}{dt} P_t^x \varphi_x(x) = \mathcal{L}_t^x P_t^x \varphi_x(x)$, $\frac{d}{dt} P_t^y \varphi_y(y) = \mathcal{L}_t^y P_t^y \varphi_y(y)$ hold for φ_x, φ_y in the domains of the generators. Since $\mathcal{L}_t^x$ and P_t^x commute for these choices of φ_x, we have $\frac{d}{dt} P_t^x \varphi_x(x) = P_t^x \mathcal{L}_t^x \varphi_x(x)$, $\frac{d}{dt} P_t^y \varphi_y(y) = P_t^y \mathcal{L}_t^y \varphi_y(y)$. By integrating these two equations over the initial measures $\mu_{x,0}, \mu_{y,0}$, we get

$$\frac{d}{dt} \int \varphi_x(x)\, d\mu_{x,t} = \int \mathcal{L}_t^x \varphi_x(x)\, d\mu_{x,t}, \quad \frac{d}{dt} \int \varphi_y(y)\, d\mu_{y,t} = \int \mathcal{L}_t^y \varphi_y(y)\, d\mu_{y,t}.$$

We can write an explicit form for $\mathcal{L}_t^x P_t^x \varphi_x(x)$ by using Itô's lemma on (44):

$$\mathcal{L}_t^x \varphi_x(x) = \left(\int_{\mathcal{Y}} \nabla_x \ell(x,y)\ d\mu_{y,s}\ ds - \hat{\mathbf{h}}_x(x) \right) \nabla_x \varphi_x(x) + \beta^{-1} \mathrm{Tr}\left(\left(\mathrm{Proj}_{T_x\mathcal{X}}\right)^\top H\varphi_x(x)\ \mathrm{Proj}_{T_x\mathcal{X}} \right),$$

where we use $\mathrm{Proj}_{T_{\tilde{X}_t^i}\mathcal{X}}$ to denote its matrix in the canonical basis.

Let $\{\xi_k\}$ be a partition of unity for $\mathcal{X}$ (i.e. a set of functions such that $\sum_k \xi_k(x) = 1$) in which each ξ_k is regular enough and supported on a patch of $\mathcal{X}$. We can write

$$\begin{aligned}\frac{d}{dt}\int_{\mathcal{X}} \varphi_x(x)\ d\mu_{x,t}(x) &= \frac{d}{dt}\int_{\mathcal{X}} \varphi_x(x)\ d\mu_{x,t}(x) = \sum_k \frac{d}{dt}\int_{\mathcal{X}} \xi_k(x)\varphi_x(x)\ d\mu_{x,t}(x) \\ &= \sum_k \int \mathcal{L}_t^x(\xi_k(x)\varphi_x(x))\ d\mu_{x,t}\end{aligned}$$

Now, let $\tilde{\varphi}_x^k(x) = \xi_k(x)\varphi_x(x)$.

$$\begin{aligned}&\int_{\mathcal{X}} \mathcal{L}_t^x \tilde{\varphi}_x^k(x)\ d\mu_{x,t} \\ &= \int_{\mathcal{X}} \left(\nabla_x V_x(\mu_{y,s},x) - \hat{\mathbf{h}}_x(x) \right) \nabla_x \tilde{\varphi}_x^k(x) + \beta^{-1}\mathrm{Tr}\left(\left(\mathrm{Proj}_{T_x\mathcal{X}}\right)^\top H\tilde{\varphi}_x^k(x)\ \mathrm{Proj}_{T_x\mathcal{X}} \right)\ d\mu_{x,t}\end{aligned}$$

Notice that this equation is analogous to (66). We reverse the argument made in Subsec. G.5. Using the fact that the support of $\tilde{\varphi}_x^k(x)$ is contained on some patch of $\mathcal{X}$ given by the mapping $\psi_k : U_{\mathbb{R}^d} \subseteq \mathbb{R}^d \to U \subseteq \mathcal{X} \subseteq \mathbb{R}^D$, the corresponding Fokker-Planck on $U_{\mathbb{R}^d}$ is

$$\begin{aligned}&\frac{d}{dt}\int_{U_{\mathbb{R}^d}} \tilde{\varphi}_x^k(\psi_k(q))\ d(\psi_k^{-1})_*\mu_{x,t}(q) \\ &= \int_{U_{\mathbb{R}^d}} \nabla V_x(\mu_{y,s},\psi_k(q)) \cdot \nabla\tilde{\varphi}_x^k(\psi_k(q)) + \beta^{-1}\Delta\tilde{\varphi}_x^k(\psi_k(q))\ d(\psi_k^{-1})_*\mu_{x,t}(q),\end{aligned}$$

where the gradients and the Laplacian are in the metric inherited from the embedding (as in Subsec. G.5). The pushforward definition implies

$$\frac{d}{dt}\int_{\mathcal{X}} \tilde{\varphi}_x^k(x)\ d\mu_{x,t}(x) = \int_{U_{\mathbb{R}^d}} \nabla V_x(\mu_{y,s},x) \cdot \nabla\tilde{\varphi}_x^k(x) + \beta^{-1}\Delta\tilde{\varphi}_x^k(x)\ d\mu_{x,t}(x),$$

By substituting $\tilde{\varphi}_x^k(x) = \xi_k(x)\varphi_x(x)$, summing for all k and using $\sum_k \xi_k(x) = 1$, we obtain:

$$\frac{d}{dt}\int_{\mathcal{X}} \varphi_x(x)\ d\mu_{x,t}(x) = \int_{\mathcal{X}} \nabla_x V_x(\mu_{y,s},x) \cdot \nabla_x\varphi_x(x) + \beta^{-1}\Delta_x\varphi_x(x)\ d\mu_{x,t}(x)$$

which is the same as the first equation in (7). The second equation is obtained analogously. □

Let $\mu_x^n = \frac{1}{n}\sum_{i=1}^n \delta_{X^i}$ be a $\mathcal{P}(\mathcal{C}([0,T],\mathcal{X}))$-valued random element that corresponds to the empirical measure of a solution $\mathbf{X}^n$ of (43). Analogously, let $\mu_y^n = \frac{1}{n}\sum_{i=1}^n \delta_{Y^i}$ be a $\mathcal{P}(\mathcal{C}([0,T],\mathcal{Y}))$-valued random element corresponding to the empirical measure of $\mathbf{Y}^n$.

Define the 2-Wasserstein distance on $\mathcal{P}(\mathcal{C}([0,T],\mathcal{X}))$ as

$$\mathcal{W}_2^2(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{C([0,T],\mathcal{X})^2} d(x,y)^2\ d\pi(x,y) \tag{46}$$

where $d(x,y) = \sup_{t\in[0,T]} d_{\mathcal{X}}(x(t),y(t))$. Define it analogously on $\mathcal{P}(\mathcal{C}([0,T],\mathcal{Y}))$.

We state a stronger version of the law of large numbers in the first statement of Theorem 3(i).

Theorem 9. *There exists a solution of the coupled McKean-Vlasov SDEs (44). Pathwise uniqueness and uniqueness in law hold. Let $\mu_x \in \mathcal{P}(\mathcal{C}([0,T],\mathcal{X})), \mu_y \in \mathcal{P}(\mathcal{C}([0,T],\mathcal{Y}))$ be the unique laws of the solutions for $n = 1$ (all pairs have the same solutions). Then,*

$$\mathbb{E}[\mathcal{W}_2^2(\mu_x^n, \mu_x) + \mathcal{W}_2^2(\mu_y^n, \mu_y)] \xrightarrow{n\to\infty} 0$$

Let us comment on why Theorem 9 implies the first statement in Theorem 3(i). We make use of the mapping $\mathcal{P}(\mathcal{C}([0,T],\mathcal{X})) \ni \mu \mapsto (\mu_t)_{t\in[0,T]} \in \mathcal{C}([0,T],\mathcal{P}(\mathcal{X}))$ into the time marginals. By the definition (46), $\sup_{t\in[0,t]} \mathcal{W}_2^2(\mu_{x,t}^n, \mu_{x,t}) \leqslant \mathcal{W}_2^2(\mu_x^n, \mu_x)$ and the same holds for μ_y^n, μ_y. At this point, Lemma 10 states that $(\mu_x)_{t\in[0,T]}, (\mu_y)_{t\in[0,T]}$ is a solution of the mean-field ERIWGF (45) and concludes the argument. The proof of Theorem 9 uses a propagation of chaos argument, originally due to Sznitman [1991] in the context of interacting particle systems. Our argument follows Theorem 3.3 of Lacker [2018].

E.2 Existence and uniqueness

We prove existence and uniqueness of the system given by

$$\begin{aligned}
\tilde{X}_t &= \int_0^t \left(-\int_{\mathcal{Y}} \nabla_x \ell(\tilde{X}_s, y)\ d\mu_{y,s}\ ds + \hat{\mathbf{h}}_x(\tilde{X}_s) \right)\ ds + \sqrt{2\beta^{-1}}\ \int_0^t \mathrm{Proj}_{T_{\tilde{X}_s}\mathcal{X}}(dW_s), \\
\tilde{Y}_t &= \int_0^t \left(\int_{\mathcal{X}} \nabla_y \ell(x, \tilde{Y}_s)\ d\mu_{x,s} + \hat{\mathbf{h}}_y(Y_s^{n,i}) \right)\ ds + \sqrt{2\beta^{-1}}\ \int_0^t \mathrm{Proj}_{T_{\tilde{Y}_s}\mathcal{Y}}(d\bar{W}_s), \\
\mu_{x,t} &= \mathrm{Law}(\tilde{X}_t^n), \quad \mu_{y,t} = \mathrm{Law}(\tilde{Y}_t^n), \quad \tilde{X}_0 = \xi \sim \mu_{x,0}, \quad \tilde{Y}_0 = \bar{\xi} \sim \mu_{y,0}.
\end{aligned} \tag{47}$$

Path-wise uniqueness means that given $W, \bar{W}, \xi, \bar{\xi}$, two solutions are equal almost surely. Uniqueness in law means that regardless of the Brownian motion and the initialization random variables chosen (as long as they are $\mu_{x,0}$ and $\mu_{y,0}$-distributed), the law of the solution is unique. We prove that both hold for (47).

We have that for all $x, x' \in \mathcal{X}, \mu, \nu \in \mathcal{P}(\mathcal{Y})$,

$$\left| \int \nabla_x \ell(x, y)\ d\mu - \int \nabla_x \ell(x', y)\ d\nu \right| \leqslant L(d(x, x') + \mathcal{W}_2(\mu, \nu)) \tag{48}$$

This is obtained by adding and subtracting the term $\int \nabla_x \ell(x'y)\ d\mu$, by using the triangle inequality and the inequality $\mathcal{W}_1(\mu,\nu)) \leqslant \mathcal{W}_2(\mu,\nu))$ (which is proven using the Cauchy-Schwarz inequality). Hence,

$$\left| \int \nabla_x \ell(x, y)\ d\mu - \int \nabla_x \ell(x', y)\ d\nu \right|^2 \leqslant 2L^2(d(x, x')^2 + \mathcal{W}_2^2(\mu, \nu)) \tag{49}$$

On the other hand, using the regularity of the manifold, there exists $\mathcal{L}_{\mathcal{X}}$ such that

$$\begin{aligned}
|\hat{\mathbf{h}}_x(x) - \hat{\mathbf{h}}_x(x')| &\leqslant L_{\mathcal{X}} d(x, x'), \\
|\mathrm{Proj}_{T_x\mathcal{X}} - \mathrm{Proj}_{T_{x'}\mathcal{X}}| &\leqslant L_{\mathcal{X}} d(x, x')
\end{aligned}$$

where $\mathrm{Proj}_{T_x\mathcal{X}}$ denotes its matrix in the canonical basis and the norm in the second line is the Frobenius norm. Also, let $\|x - x'\|$ be the Euclidean norm of $\mathcal{X}$ in $\mathbb{R}^{D_x}$ (the Euclidean space where $\mathcal{X}$ is embedded) and let $K_{\mathcal{X}} > 1$ be such that $d(x, x') \leqslant K_{\mathcal{X}} \|x - x'\|$.

Let $\mu_y, \nu_y \in \mathcal{P}(\mathcal{C}([0,T],\mathcal{X}))$ and let X^{μ_y}, X^{ν_y} be the solutions of the first equation of (47) when we plug μ_y (ν_y resp.) as the measure for the other player. X^{μ_y} and X^{ν_y} exist and are unique by the classical theory of SDEs (see Chapter 18 of Kallenberg [2002]). Following the procedure in Theorem 3.3 of Lacker [2018], we

obtain

$$\begin{aligned}\mathbb{E}[\|X^{\mu_y} - X^{\nu_y}\|_t^2] &\leqslant 3t\mathbb{E}\bigg[\int_0^t \bigg|\int \nabla_x \ell(X^{\mu_y}, y)\ d\mu_{y,r} - \int \nabla_x \ell(X^{\nu_y}, y)\ d\nu_{y,r}\bigg|^2\ dr\bigg] \\ &\quad + 3t\mathbb{E}\bigg[\int_0^t |\hat{\mathbf{h}}_x(X^{\mu_y}) - \hat{\mathbf{h}}_x(X^{\nu_y})|^2\ dr\bigg] \\ &\quad + 12\mathbb{E}\bigg[\int_0^t |\mathrm{Proj}_{T_x\mathcal{X}} - \mathrm{Proj}_{T_{x'}\mathcal{X}}|^2\ dr\bigg] \\ &\leqslant 3(3t+4)\tilde{L}^2\mathbb{E}\bigg[\int_0^t (\|X^{\mu_y} - X^{\nu_y}\|_r^2 + \mathcal{W}_2^2(\mu_{y,r}, \nu_{y,r}))\ dr\bigg],\end{aligned} \tag{50}$$

where $\tilde{L}^2 = (L^2 + L_{\mathcal{X}}^2)K_{\mathcal{X}}^2$. Using Fubini's theorem and Gronwall's inequality, we obtain

$$\mathbb{E}[\|X^{\mu_y} - X^{\nu_y}\|_t^2] \leqslant 3(3T+4)\tilde{L}^2 \exp(3(3T+4)\tilde{L}^2) \int_0^t \mathcal{W}_2^2(\mu_{y,r}, \nu_{y,r}))\ dr \tag{51}$$

Let $C_T := 3(3T+4)\tilde{L}^2 \exp(3(3T+4)\tilde{L}^2)$. For $\mu, \nu \in \mathcal{P}(C([0,T], \mathcal{X}))$, define

$$\mathcal{W}_{2,t}^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{C([0,T],\mathcal{X})^2} \sup_{r\in[0,t]} d(x(r), y(r))\ \pi(dx, dy)$$

Hence, (51) and the bound $\mathcal{W}_2^2(\mu_{y,r}, \nu_{y,r}) \leqslant \mathcal{W}_{2,r}^2(\mu_y, \nu_y)$ yield

$$\mathbb{E}[\|X^{\mu_y} - X^{\nu_y}\|_t^2] \leqslant C_T \int_0^t \mathcal{W}_{2,r}^2(\mu_y, \nu_y)\ dr$$

Reasoning analogously for the other player, we obtain

$$\mathbb{E}[\|X^{\mu_y} - X^{\nu_y}\|_t^2 + \|Y^{\mu_x} - Y^{\nu_x}\|_t^2] \leqslant C_T \int_0^t \mathcal{W}_{2,r}^2(\mu_y, \nu_y)\ dr + C_T \int_0^t \mathcal{W}_{2,r}^2(\mu_x, \nu_x)\ dr$$

Given $\mu_y \in \mathcal{P}(C([0,T], \mathcal{Y}))$, define $\Phi_x(\mu_y) = \mathrm{Law}(X^{\mu_y}) \in \mathcal{P}(C([0,T], \mathcal{X}))$, and define Φ_y analogously. Notice that $\mathcal{W}_{2,t}^2(\Phi_x(\mu_y), \Phi_x(\nu_y)) \leqslant \mathbb{E}[\|X^{\mu_y} - X^{\nu_y}\|_t^2], \mathcal{W}_{2,t}^2(\Phi_y(\mu_x), \Phi_y(\nu_x)) \leqslant \mathbb{E}[\|X^{\mu_x} - X^{\nu_x}\|_t^2]$. Hence, we obtain

$$\mathcal{W}_{2,t}^2(\Phi_x(\mu_y), \Phi_x(\nu_y)) + \mathcal{W}_{2,t}^2(\Phi_y(\mu_x), \Phi_y(\nu_x)) \leqslant C_T \int_0^t \mathcal{W}_{2,r}^2(\mu_y, \nu_y) + \mathcal{W}_{2,r}^2(\mu_x, \nu_x)\ dr$$

Observe that $\mathcal{W}_{2,t}^2(\mu_x, \nu_x) + \mathcal{W}_{2,t}^2(\mu_y, \nu_y)$ is the square of a distance between (μ_x, μ_y) and (ν_x, ν_y) on $\mathcal{P}(C([0,T], \mathcal{X})) \times \mathcal{P}(C([0,T], \mathcal{Y}))$. Hence, we can apply the Piccard iteration argument to obtain the existence result and another application of Gronwall's inequality yields pathwise uniqueness.

Uniqueness in law (i.e., regardless of the specific Brownian motions and initialization random variables) follows from the typical uniqueness in law result for SDEs (see Chapter 18 of Kallenberg [2002] for example). The idea is that when we solve the SDEs with $W', \bar{W}', \xi', \bar{\xi}'$ plugging in the drift the laws of a solution for $W, \bar{W}, \xi, \bar{\xi}$, the solution has the same laws by uniqueness in law of SDEs. Hence, that new solution solves the coupled McKean-Vlasov for $W', \bar{W}', \xi', \bar{\xi}'$.

E.3 Propagation of chaos

Let $\mu_x^n = \frac{1}{n}\sum_{i=1}^n \delta_{X^i}, \mu_y^n = \frac{1}{n}\sum_{i=1}^n \delta_{Y^i}$. Using the argument from existence and uniqueness on the i-th components of $\mathbf{X}, \tilde{\mathbf{X}}$,

$$\mathbb{E}[\|X^i - \tilde{X}^i\|_t^2] \leqslant 3(3T+4)\tilde{L}^2\mathbb{E}\bigg[\int_0^t (\|X^i - \tilde{X}^i\|_r^2 + \mathcal{W}_2^2(\mu_{y,r}^n, \mu_{y,r}))\ dr\bigg]$$

Arguing as before, we obtain

$$\mathbb{E}[\|X^i - \tilde{X}^i\|_t^2] \leqslant C_T \mathbb{E}\bigg[\int_0^t \mathcal{W}_{2,r}^2(\mu_y^n, \mu_y)\ dr\bigg]$$

Let $\nu_x^n = \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{X}^i}$ be the empirical measure of the mean field processes in (44). Notice that $\frac{1}{n}\sum_{i=1}^n \delta_{(X^i, \tilde{X}^i)}$ is a coupling between ν_x^n and μ_x^n, and so

$$\mathcal{W}_{2,t}^2(\mu_x^n, \nu_x^n) \leqslant \frac{1}{n}\sum_{i=1}^n \|X^i - \tilde{X}^i\|_t^2$$

Thus, we obtain

$$\mathbb{E}[\mathcal{W}_{2,t}^2(\mu_x^n, \nu_x^n)] \leqslant C_T \mathbb{E}\bigg[\int_0^t \mathcal{W}_{2,r}^2(\mu_y^n, \mu_y)\ dr\bigg]$$

We use the triangle inequality

$$\begin{aligned}\mathbb{E}[\mathcal{W}_{2,t}^2(\mu_x^n, \mu_x)] &\leqslant 2\mathbb{E}[\mathcal{W}_{2,t}^2(\mu_x^n, \nu_x^n)] + 2\mathbb{E}[\mathcal{W}_{2,t}^2(\nu_x^n, \mu_x)] \\ &\leqslant 2C_T \mathbb{E}\bigg[\int_0^t \mathcal{W}_{2,r}^2(\mu_y^n, \mu_y)\ dr\bigg] + 2\mathbb{E}[\mathcal{W}_{2,t}^2(\nu_x^n, \mu_x)]\end{aligned}$$

At this point we follow an analogous procedure for the other player and we end up with

$$\begin{aligned}\mathbb{E}[\mathcal{W}_{2,t}^2(\mu_x^n, \mu_x) + \mathcal{W}_{2,t}^2(\mu_y^n, \mu_y)] \leqslant 2C_T \mathbb{E}\bigg[\int_0^t \mathcal{W}_{2,r}^2(\mu_y^n, \mu_y) + \mathcal{W}_{2,r}^2(\mu_x^n, \mu_x)\ dr\bigg] \\ + 2\mathbb{E}[\mathcal{W}_{2,t}^2(\nu_x^n, \mu_x) + \mathcal{W}_{2,t}^2(\nu_y^n, \mu_y)]\end{aligned}$$

We use Fubini's theorem and Gronwall's inequality again.

$$\mathbb{E}[\mathcal{W}_{2,t}^2(\mu_x^n, \mu_x) + \mathcal{W}_{2,t}^2(\mu_y^n, \mu_y)] \leqslant 2\exp(2C_T T)\mathbb{E}[\mathcal{W}_{2,t}^2(\nu_x^n, \mu_x) + \mathcal{W}_{2,t}^2(\nu_y^n, \mu_y)]$$

If we set $t = T$ we get

$$\mathbb{E}[\mathcal{W}_2^2(\mu_x^n, \mu_x) + \mathcal{W}_2^2(\mu_y^n, \mu_y)] \leqslant 2\exp(2C_T T)\mathbb{E}[\mathcal{W}_2^2(\nu_x^n, \mu_x) + \mathcal{W}_2^2(\nu_y^n, \mu_y)]$$

and the factor $\mathbb{E}[\mathcal{W}_2^2(\nu_x^n, \mu_x) + \mathcal{W}_2^2(\nu_y^n, \mu_y)]$ goes to 0 as $n \to \infty$ by the law of large numbers (see Corollary 2.14 of [Lacker, 2018]).

E.4 Convergence of the Nikaido-Isoda error

Corollary 2. *For $t \in [0, T]$, if $\mu_{x,t}^n, \mu_{x,t}, \mu_{y,t}^n, \mu_{y,t}$ are the marginals of $\mu_x^n, \mu_x, \mu_y^n, \mu_y$ at time t, we have*

$$\mathbb{E}[|NI(\mu_{x,t}^n, \mu_{y,t}^n) - NI(\mu_{x,t}, \mu_{y,t})|] \xrightarrow{n\to\infty} 0$$

Proof. See Lemma 3. □

F Proof of Theorem 3(ii)

F.1 Preliminaries

Define the processes $\mathbf{X} = (X^1, \dots, X^n), \mathbf{w}_x = (w_x^1, \dots, w_x^n)$ and $\mathbf{Y} = (Y^1, \dots, Y^n), \mathbf{w}_y = (w_y^1, \dots, w_y^n)$ such that for all $i \in \{1, \dots, n\}$

$$\begin{aligned}
\frac{dX_t^i}{dt} &= -\gamma \frac{1}{n} \sum_{j=1}^n w_{y,t}^j \nabla_x \ell(X_t^i, Y_t^j), \quad X_0^i = \xi^i \sim \mu_{x,0} \\
\frac{dw_{x,t}^i}{dt} &= \alpha \left(-\frac{1}{n} \sum_{j=1}^n w_{y,t}^j \ell(X_t^i, Y_t^j) + \frac{1}{n^2} \sum_{k=1}^n \sum_{j=1}^n w_{y,t}^j w_{x,t}^k \ell(X_t^i, Y_t^j) \right) w_{x,t}^i, \quad w_{x,0}^i = 1 \\
\frac{dY_t^i}{dt} &= \gamma \frac{1}{n} \sum_{j=1}^n w_{x,t}^j \nabla_y \ell(X_t^j, Y_t^i), \quad Y_0^i = \bar{\xi}^i \sim \mu_{y,0} \\
\frac{dw_{y,t}^i}{dt} &= \alpha \left(\frac{1}{n} \sum_{j=1}^n w_{x,t}^j \ell(X_t^i, Y_t^j) - \frac{1}{n^2} \sum_{k=1}^n \sum_{j=1}^n w_{y,t}^j w_{x,t}^k \ell(X_t^i, Y_t^j) \right) w_{x,t}^i, \quad w_{y,0}^i = 1
\end{aligned} \tag{52}$$

Let $\nu_{x,t}^n = \frac{1}{n} \sum_{i=1}^n \delta_{(X_t^i, w_{x,t}^i)} \in \mathbb{P}(\mathcal{X} \times \mathbb{R}^+), \nu_{y,t}^n = \frac{1}{n} \sum_{i=1}^n \delta_{(Y_t^i, r_{y,t}^{n,i})} \in \mathbb{P}(\mathcal{Y} \times \mathbb{R}^+)$. Let $\mu_{x,t}^n = \frac{1}{n} \sum_{i=1}^n w_{x,t}^i \delta_{X_t^i} \in \mathbb{P}(\mathcal{X}), \mu_{y,t}^n = \frac{1}{n} \sum_{i=1}^n w_{y,t}^i \delta_{Y_t^i} \in \mathbb{P}(\mathcal{Y})$ be the projections of $\nu_{x,t}^n, \nu_{y,t}^n$. Notice that we have changed the notation with respect to the main text, multiplying w_x^i by n: now $w_{x,0}^i = 1$ and $\sum_i w_{x,t}^i = n, \forall t \geqslant 0$ instead of $w_{x,0}^i = 1/n$ and $\sum_i w_{x,t}^i = 1, \forall t \geqslant 0$.

Let h_x, h_y be the projection operators, i.e. $h_x \nu_x = \int_{\mathcal{R}^+} w_x \nu_x(\cdot, w_x)$. We also define the mean field processes $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \tilde{\mathbf{w}}_x, \tilde{\mathbf{w}}_y$ given component-wise by

$$\begin{aligned}
\frac{d\tilde{X}_t^i}{dt} &= -\gamma \nabla_x \int \ell(\tilde{X}_t^i, y) d\mu_{y,t}, \quad \tilde{X}_0^i = \xi^i \sim \mu_{x,0} \\
\frac{d\tilde{w}_{x,t}^i}{dt} &= \alpha \left(-\int \ell(\tilde{X}_t^i, y) d\mu_{y,t} + \mathcal{L}(\mu_{x,t}, \mu_{y,t}) \right) \tilde{w}_{x,t}^i, \quad \tilde{w}_{x,0}^i = 1 \\
\frac{d\tilde{Y}_t^i}{dt} &= \gamma \nabla_y \int \ell(x, \tilde{Y}_t^i) d\mu_{x,t}, \quad \tilde{Y}_0^i = \bar{\xi}^i \sim \mu_{y,0} \\
\frac{d\tilde{w}_{y,t}^i}{dt} &= \alpha \left(\int \ell(x, \tilde{Y}_t^i) d\mu_{x,t} - \mathcal{L}(\mu_{x,t}, \mu_{y,t}) \right) \tilde{w}_{x,t}^i, \quad \tilde{w}_{y,0}^i = 1 \\
\mu_{x,t} &= h_x \mathrm{Law}(\tilde{X}_t^i, \tilde{w}_{x,t}^i), \quad \mu_{y,t} = h_y \mathrm{Law}(\tilde{Y}_t^i, \tilde{w}_{y,t}^i)
\end{aligned} \tag{53}$$

for i between 1 and n.

Lemma 11 (Forward Kolmogorov equation)**.** *If $\tilde{X}, \tilde{w}_x, \tilde{Y}, \tilde{w}_y$ is a solution of (53) with $n = 1$, then its laws ν_x, ν_y fulfill (10).*

Proof. Let $\psi_x : \mathcal{X} \times \mathbb{R}^+ \to \mathbb{R}$. Plug the laws ν_x, ν_y of the solution $(\tilde{X}, \tilde{w}_x), (\tilde{Y}, \tilde{w}_y)$ into the ODE (53). Let $\Phi_{x,t} = (X_{x,t}^\Phi, w_{x,t}^\Phi) : (\mathcal{X} \times \mathbb{R}^+) \to (\mathcal{X} \times \mathbb{R}^+)$ denote the flow that maps an initial condition of the ODE (53) to the corresponding solution at time t. Then, we can write $\nu_{x,t} = (\Phi_{x,t})_* \nu_{x,0}$, where $(\Phi_{x,t})_*$ is the pushforward.

Hence,

$$\begin{aligned}
&\frac{d}{dt}\int_{\mathcal{X}\times\mathbb{R}^+} \psi_x(x,w_x)\ d\nu_{x,t}(x,w_x)\\
&= \frac{d}{dt}\int_{\mathcal{X}\times\mathbb{R}^+} \psi_x(\Phi_{x,t}(x,w_x))\ d\nu_{x,0}(x,w_x)\\
&= \int_{\mathcal{X}\times\mathbb{R}^+} \left(\nabla_x\psi_x(\Phi_{x,t}(x,w_x)), \frac{d\psi_x}{dw_x}(\Phi_{x,t}(x,w_x))\right)\cdot \frac{d}{dt}\Phi_{x,t}(x,w_x)\ d\nu_{x,0}(x,w_x)\\
&= \int_{\mathcal{X}\times\mathbb{R}^+} \nabla_x\psi_x(\Phi_{x,t}(x,w_x))\cdot(-\gamma\nabla_x V_x(h_y\nu_{y,t}, X^{\Phi}_{x,t}))\\
&+ \frac{d\psi_x}{dw_x}(\Phi_{x,t}(x,w_x))\alpha(-V_x(h_y\nu_{y,t}, X^{\Phi}_{x,t}) + \mathcal{L}(h_x\nu_{x,t}, h_y\mu_{y,t}))\ d\nu_{x,0}(x,w_x)
\end{aligned}$$

And we can identify the right hand side as the weak form of (10), shown in (12). The argument for ν_y is analogous. □

We state a stronger version of the law of large numbers in the first statement of Theorem 3(ii).

Theorem 10. *There exists a solution of the coupled SDEs (53). Pathwise uniqueness and uniqueness in law hold. Let $\nu_x \in \mathcal{P}(\mathcal{C}([0,T],\mathcal{X}\times\mathbb{R}^+)), \nu_y \in \mathcal{P}(\mathcal{C}([0,T],\mathcal{Y}\times\mathbb{R}^+))$ be the unique laws of the solutions for $n=1$ (all pairs have the same solutions). Then,*

$$\mathbb{E}[\mathcal{W}_2^2(\nu_x^n,\nu_x) + \mathcal{W}_2^2(\nu_y^n,\nu_y)] \xrightarrow{n\to\infty} 0$$

Theorem 10 is the law of large numbers for the WFR dynamics, and its proof follows the same argument of Theorem 9. The reason Theorem 10 implies Theorem 3(ii) is analogous to the reason for which Theorem 9 implies Theorem 3(i), with the additional step that $\mathcal{W}_2^2(\mu_{x,t}^n,\mu_{x,t}) = \mathcal{W}_2^2(h_x\nu_{x,t}^n, h_x\nu_{x,t}) \leqslant e^{4MT}\mathcal{W}_2^2(\nu_{x,t}^n,\nu_{x,t})$, and this inequality is shown in (55).

F.2 Existence and uniqueness

We choose to do an argument close to Sznitman [1991] (see Lacker [2018]), which yields convergence of the expectation of the square of the 2-Wasserstein distances between the empirical and the mean field measures.

First, to prove existence and uniqueness of the solution $(\mu_{x,t},\mu_{y,t})$ in the time interval $[0,T]$ for arbitrary T, we can use the same argument as in the App. E. Now, instead of (47) we have

$$\begin{aligned}
\tilde{X}_t &= \xi - \gamma\int_0^t\int_{\mathcal{Y}} \nabla_x\ell(\tilde{X}_s,y)\ d\mu_{y,s}\ ds,\\
\tilde{w}_{x,t} &= 1 + \alpha\int_0^t\left(-\int \ell(\tilde{X}_t,y)d\mu_{y,t} + \mathcal{L}(\mu_{x,t},\mu_{y,t})\right)\tilde{w}_{x,s}\ ds,\\
\tilde{Y}_t &= \bar{\xi} + \gamma\int_0^t\int_{\mathcal{X}} \nabla_y\ell(x,\tilde{Y}_s)\ d\mu_{x,s}\ ds,\\
\tilde{w}_{y,t} &= 1 + \alpha\int_0^t\left(\int \ell(x,\tilde{Y}_t)d\mu_{x,t} - \mathcal{L}(\mu_{x,t},\mu_{y,t})\right)\tilde{w}_{y,s}\ ds,\\
\mu_{x,t} &= h_x\mathrm{Law}(\tilde{X}_t,\tilde{w}_{x,t}), \quad \mu_{y,t} = h_y\mathrm{Law}(\tilde{Y}_t,\tilde{w}_{y,t}),
\end{aligned}$$

where ξ and $\bar{\xi}$ are arbitrary random variables with laws $\mu_{x,0},\mu_{y,0}$ respectively. For $x,x'\in\mathcal{X}$, $r,r'\in\mathbb{R}^+$,

$\mu_x, \mu_x' \in \mathcal{P}(\mathcal{X})$, $\mu_y, \mu_y' \in \mathcal{P}(\mathcal{Y})$, notice that using an argument similar to (48) the following bound holds

$$\begin{aligned}
&\left| \left(-\int \ell(x,y) d\mu_y + \mathcal{L}(\mu_x, \mu_y) \right) w - \left(-\int \ell(x',y) d\mu_y' + \mathcal{L}(\mu_x', \mu_y') \right) w' \right| \\
&\leqslant 2M|w-w'| + |w'|\tilde{L}(|x-x'| + 3\mathcal{W}_1(\nu,\mu)) \leqslant 2M|w-w'| + |w'|\tilde{L}(|x-x'| + 3\mathcal{W}_2(\mu_y,\mu_y')) \\
&\implies \left| \left(-\int \ell(x,y) d\mu_y + \mathcal{L}(\mu_x, \mu_y) \right) r - \left(-\int \ell(x',y) d\mu_y' + \mathcal{L}(\mu_x', \mu_y') \right) r' \right|^2 \\
&\quad \leqslant 12M^2|w-w'|^2 + 3|w'|^2\tilde{L}^2(|x-x'|^2 + 9\mathcal{W}_2^2(\mu_y,\mu_y'))
\end{aligned}$$

Recall that M is a bound on the absolute value of ℓ and $\tilde{L}$ is the Lipschitz constant of the loss ℓ. A simple application of Gronwall's inequality shows $|\tilde{w}_{x,t}|$ is bounded by e^{2MT} for all $t \in [0,T]$. Hence, we can write

$$\begin{aligned}
&\mathbb{E}[\|X^{\mu_y} - X^{\mu_y'}\|_t^2 + \|w_x^{\mu_y} - w_x^{\mu_y'}\|_t^2] \leqslant \gamma^2 t \mathbb{E}\left[\int_0^t \left| \nabla_x \int \ell(X_s^{\mu_y}, y) d\mu_{y,s} - \nabla_x \int \ell(X_s^{\mu_y'}, y) d\mu_{y,s}' \right|^2 ds \right] \\
&+\alpha^2 t \mathbb{E}\left[\int_0^t \left| \left(-\int \ell(X_s^{\mu_y}, y) d\mu_y + \mathcal{L}(\mu_x, \mu_y) \right) w_x^{\mu_y} - \left(-\int \ell(X_s^{\mu_y'}, y) d\mu_y' + \mathcal{L}(\mu_x', \mu_y') \right) w_x^{\mu_y'} \right|^2 ds \right] \\
&\qquad \leqslant Kt\mathbb{E}\left[\int_0^t \|X^{\mu_y} - X^{\mu_y'}\|_s^2 + \|w^{\mu_y} - w^{\mu_y'}\|_s^2 \, ds \right] + K' t \mathbb{E}\left[\int_0^t \mathcal{W}_2^2(\mu_{y,s}, \mu_{y,s}') \, ds \right],
\end{aligned}$$

where $K = \max\{12\alpha^2M^2, 2L^2\gamma^2 + 3\tilde{L}^2e^{4MT}\alpha^2\}, K' = 2L^2\gamma^2 + 27\tilde{L}^2e^{4MT}\alpha^2$. Notice that we have used (49) as well. This equation is analogous to equation (50), and upon application of Fubini's theorem and Gronwall's inequality it yields

$$\mathbb{E}[\|X^{\mu_y} - X^{\mu_y'}\|_t^2 + \|w_x^{\mu_y} - w_x^{\mu_y'}\|_t^2] \leqslant TK' \exp(TK) \mathbb{E}\left[\int_0^t \mathcal{W}_2^2(\mu_{y,s}, \mu_{y,s}') \, ds \right] \tag{54}$$

Now we will prove that

$$\mathcal{W}_2^2(h_x\nu_x, h_x\nu_x') \leqslant e^{4MT}\mathcal{W}_2^2(\nu_x, \nu_x'), \tag{55}$$

where $\nu_x, \nu_x' \in \mathcal{P}(\mathcal{X} \times [0, e^{2MT}])$. Define the homogeneous projection operator $\tilde{h} : \mathcal{P}((\mathcal{X} \times \mathbb{R}^+)^2) \to \mathcal{P}(\mathcal{X}^2)$ as $\forall f \in C(\mathcal{X}^2)$,

$$\int_{\mathcal{X}^2} f(x,y) \, d(\tilde{h}\pi)(x,y) = \int_{(\mathcal{X}\times[0,e^{2MT}])^2} w_x w_y f(x,y) \, d\pi(x,w_x,y,w_y), \ \forall \pi \in \mathcal{P}((\mathcal{X} \times \mathbb{R}^+)^2).$$

Let π be a coupling between $h_x\nu_x, h_x\nu_x'$. Then $\tilde{h}\pi$ is a coupling between $h_x\nu_x, h_x\nu_x'$ and

$$\begin{aligned}
\int_{\mathcal{X}^2} \|x-y\|^2 \, d(\tilde{h}\pi)(x,y) &= \int_{(\mathcal{X}\times[0,e^{2MT}])^2} w_x w_y \|x-y\|^2 \, d\pi(x,w_x,y,w_y) \\
&\leqslant e^{4MT} \int_{(\mathcal{X}\times[0,e^{2MT}])^2} \|x-y\|^2 \, d\pi(x,w_x,y,w_y) \\
&\leqslant e^{4MT} \int_{(\mathcal{X}\times[0,e^{2MT}])^2} \|x-y\|^2 + |w_x - w_y|^2 \, d\pi'(x,w_x,y,w_y)
\end{aligned}$$

Taking the infimum with respect to π on both sides we obtain the desired inequality.

Let $\nu_{x,t} = \text{Law}(X_t^{\mu_y}, w_{x,t}^{\mu_y}), \nu_{x,t}' = \text{Law}(X_t^{\mu_y'}, w_{x,t}^{\mu_y'})$ and recall that $\mu_{x,t} = h_x\nu_{x,t}, \mu_{x,t}' = h_x\nu_{x,t}'$. Given $\nu_y \in \mathcal{P}(C([0,T], \mathcal{Y} \times \mathbb{R}^+))$, define $\Phi_x(\nu_y) = \text{Law}(X^{\nu_y}, w_x^{\nu_y}) \in \mathcal{P}(C([0,T], \mathcal{X}))$ where we abuse the notation and use $(X^{\nu_y}, w_x^{\nu_y})$ to refer to $(X^{\mu_y}, w_x^{\mu_y})$. Notice also that

$$\begin{aligned}
\mathcal{W}_{2,t}^2(\Phi_x(\nu_y), \Phi_x(\nu_y')) &\leqslant \mathbb{E}\left[\sup_{s\in[0,t]} \|X_s^{\mu_y} - X_s^{\mu_y'}\|^2 + \|w_{x,s}^{\mu_y} - w_{x,s}^{\mu_y'}\|^2 \right] \\
&\leqslant \mathbb{E}[\|X^{\mu_y} - X^{\mu_y'}\|_t^2 + \|w_x^{\mu_y} - w_x^{\mu_y'}\|_t^2]
\end{aligned} \tag{56}$$

We use (55) and (56) on (54) to conclude

$$\mathcal{W}_{2,t}^2(\Phi_x(\nu_y),\Phi_x(\nu_y')) \leqslant TK'\exp(TK)\mathbb{E}\bigg[\int_0^t \mathcal{W}_{2,s}^2(\nu_y,\nu_y')\ ds\bigg]$$

The rest of the argument is sketched in App. E.

F.3 Propagation of chaos

Following the reasoning in the existence and uniqueness proof, we can write

$$\begin{aligned}&\mathbb{E}[\|X^i-\tilde{X}^i\|_t^2+\|w_x^i-\tilde{w}_x^i\|_t^2]\\&\leqslant Kt\mathbb{E}\bigg[\int_0^t \|X^i-\tilde{X}^i\|_s^2+\|w_x^i-\tilde{w}_x^i\|_s^2\ ds\bigg]+K't\mathbb{E}\bigg[\int_0^t \mathcal{W}_2^2(\mu_{y,s}^n,\mu_{y,s})\ ds\bigg],\end{aligned}$$

Hence, we obtain

$$\mathbb{E}[\|X^i-\tilde{X}^i\|_t^2+\|w_x^i-\tilde{w}_x^i\|_t^2] \leqslant TK'\exp(TK)\mathbb{E}\bigg[\int_0^t \mathcal{W}_2^2(\mu_{y,s}^n,\mu_{y,s})\ ds\bigg]$$

Let $\tilde{\nu}_{x,t}^n=\frac{1}{n}\sum_{i=1}^n \delta_{(\tilde{X}_t^i,\tilde{w}_t^i)}\in\mathbb{P}(\mathcal{X}\times\mathbb{R}^+)$ be the marginal at time t of the empirical measure of (52). As in App. E,

$$\mathcal{W}_{2,t}^2(\nu_x^n,\tilde{\nu}_x^n)\leqslant\frac{1}{n}\sum_{i=1}^n\sup_{s\in[0,t]}\|X_s^i-\tilde{X}_s^i\|^2+|w_{x,s}^i-\tilde{w}_{x,s}^i|^2\leqslant\frac{1}{n}\sum_{i=1}^n\|X^i-\tilde{X}^i\|_t^2+\|w_x^i-\tilde{w}_x^i\|_t^2$$

which yields

$$\begin{aligned}\mathbb{E}[\mathcal{W}_{2,t}^2(\nu_x^n,\tilde{\nu}_x^n)]&\leqslant TK'\exp(TK)\mathbb{E}\bigg[\int_0^t \mathcal{W}_2^2(\mu_{y,s}^n,\mu_{y,s})\ ds\bigg]\\&\leqslant TK'\exp((K+4M)T)\mathbb{E}\bigg[\int_0^t \mathcal{W}_{2,s}^2(\nu_y^n,\nu_y)\ ds\bigg]\end{aligned}$$

The second inequality above follows from inequality (55) $\mathcal{W}_2^2(\nu_{y,s}^n,\nu_{y,s})\leqslant\mathcal{W}_{2,s}^2(\nu_y^n,\nu_y)$. Now we use the triangle inequality as in App. E:

$$\begin{aligned}\mathbb{E}[\mathcal{W}_{2,t}^2(\nu_x^n,\nu_x)]&\leqslant 2\mathbb{E}[\mathcal{W}_{2,t}^2(\nu_x^n,\tilde{\nu}_x^n)]+2\mathbb{E}[\mathcal{W}_{2,t}^2(\tilde{\nu}_x^n,\nu_x)]\\&\leqslant 2TK'\exp((K+4M)T)\mathbb{E}\bigg[\int_0^t \mathcal{W}_{2,s}^2(\nu_y^n,\nu_y)\ ds\bigg]+2\mathbb{E}[\mathcal{W}_{2,t}^2(\tilde{\nu}_x^n,\nu_x)]\end{aligned}$$

If we denote $C:=2TK'\exp((K+4M)T)$ and we make the same developments for the other player, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{W}_{2,t}^2(\nu_x^n,\nu_x)+\mathcal{W}_{2,t}^2(\nu_y^n,\nu_y)]\leqslant C\mathbb{E}\bigg[\int_0^t &\mathcal{W}_{2,s}^2(\nu_y^n,\nu_y)+\mathcal{W}_{2,s}^2(\nu_x^n,\nu_x)\ ds\bigg]\\&+2\mathbb{E}[\mathcal{W}_{2,t}^2(\tilde{\nu}_x^n,\nu_x)+\mathcal{W}_{2,t}^2(\tilde{\nu}_y^n,\nu_y)]\end{aligned}$$

From this point on, the proof works as in App. E.

F.4 Convergence of the Nikaido-Isoda error

Corollary 3. *For $t\in[0,T]$, let $\bar{\mu}_{x,t}^n=\frac{1}{t}\int_0^t h_x\nu_{x,r}^n\ dr$, $\bar{\mu}_{x,t}=\frac{1}{t}\int_0^t h_x\nu_{x,r}\ dr$ and define $\bar{\mu}_{y,t}^n,\bar{\mu}_{y,t}$ analogously. Then,*

$$\mathbb{E}[|NI(\bar{\mu}_{x,t}^n,\bar{\mu}_{y,t}^n)-NI(\bar{\mu}_{x,t},\bar{\mu}_{y,t})|]\xrightarrow{n\to\infty}0$$

Proof. Notice that since the integral over time and the homogeneous projection commute, we have $\bar{\mu}^n_{x,t} = h_x(\frac{1}{t}\int_0^t \nu^n_{x,r}\ dr), \bar{\mu}_{x,t} = h_x(\frac{1}{t}\int_0^t \nu_{x,r}\ dr)$. Since $\frac{1}{t}\int_0^t \nu^n_{x,r}\ dr$ and $\frac{1}{t}\int_0^t \nu_{x,r}\ dr$ belong to $\mathcal{P}(\mathcal{X}\times[0,e^{2MT}])$, (55) implies

$$\mathcal{W}_2^2\left(h_x\left(\frac{1}{t}\int_0^t \nu^n_{x,r}\ dr\right), h_x\left(\frac{1}{t}\int_0^t \nu_{x,r}\ dr\right)\right) \leqslant e^{4MT}\mathcal{W}_2^2\left(\frac{1}{t}\int_0^t \nu^n_{x,r}\ dr, \frac{1}{t}\int_0^t \nu_{x,r}\ dr\right)$$

Notice that $\mathcal{W}_2^2(\frac{1}{t}\int_0^t \nu^n_{x,r}\ dr, \frac{1}{t}\int_0^t \nu_{x,r}\ dr) \leqslant \frac{1}{t}\int_0^t \mathcal{W}_2^2(\nu^n_{x,r},\nu_{x,r})\ dr$. Indeed,

$$\begin{aligned}
\mathcal{W}_2^2\left(\frac{1}{t}\int_0^t \nu^n_{x,r}\ dr, \frac{1}{t}\int_0^t \nu_{x,r}\ dr\right) &= \max_{\varphi\in\Psi_c(\mathcal{X})} \frac{1}{t}\int_0^t \int \varphi\ d\nu^n_{x,r}\ dr + \frac{1}{t}\int_0^t \int \varphi^c\ d\nu^n_{x,r}\ dr \\
&\leqslant \frac{1}{t}\int_0^t \left(\max_{\varphi\in\Psi_c(\mathcal{X})} \int \varphi\ d\nu^n_{x,r} + \int \varphi^c\ d\nu^n_{x,r}\right) dr \\
&= \frac{1}{t}\int_0^t \mathcal{W}_2^2(\nu^n_{x,r},\nu_{x,r})\ dr
\end{aligned}$$

Hence, using the inequality $\mathcal{W}_2^2(\nu^n_{x,r},\nu_{x,r}) \leqslant \mathcal{W}_2^2(\nu^n_x,\nu_x)$:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{W}_2^2\left(h_x\left(\frac{1}{t}\int_0^t \nu^n_{x,r}\ dr\right), h_x\left(\frac{1}{t}\int_0^t \nu_{x,r}\ dr\right)\right)\right] &\leqslant e^{4MT}\mathbb{E}\left[\frac{1}{t}\int_0^t \mathcal{W}_2^2(\nu^n_{x,r},\nu_{x,r})\ dr\right] \\
&\leqslant e^{4MT}\mathbb{E}[\mathcal{W}_2^2(\nu^n_x,\nu_x)]
\end{aligned}$$

Since the right hand side goes to zero as $n\to\infty$ by Theorem 10, we conclude by applying Lemma 3. □

F.5 Hint of the infinitesimal generator approach

Let $\varphi_x : \mathcal{X}\to\mathbb{R}, \varphi_y : \mathcal{Y}\to\mathbb{R}$ be arbitrary continuously differentiable functions, i.e. $\varphi_x \in C^1(\mathcal{X},\mathbb{R}), \varphi_y \in C^1(\mathcal{Y},\mathbb{R})$. Let us define the operators $\mathcal{L}^{(n)}_{x,t} : C^1(\mathcal{X},\mathbb{R})\to C^0(\mathcal{X},\mathbb{R}), \mathcal{L}^{(n)}_{y,t} : C^1(\mathcal{Y},\mathbb{R})\to C^0(\mathcal{Y},\mathbb{R})$ as

$$\begin{aligned}
\mathcal{L}^{(n)}_{x,t}\varphi_x(x) &= -\gamma\nabla_x \int \ell(x,y)d\mu^n_{y,t}\cdot\nabla_x\varphi_x(x) + \alpha\left(-\int \ell(x,y)d\mu^n_{y,t} + \mathcal{L}(\mu^n_{x,t},\mu^n_{y,t})\right) \\
\mathcal{L}^{(n)}_{y,t}\varphi_y(y) &= \gamma\nabla_y \int \ell(x,y)d\mu^n_{x,t}\cdot\nabla_y\varphi_y(x) + \alpha\left(\int \ell(x,y)d\mu^n_{x,t} - \mathcal{L}(\mu^n_{x,t},\mu^n_{y,t})\right)
\end{aligned} \tag{57}$$

Notice that from (52) and (57), we have

$$\begin{aligned}
\frac{d}{dt}\int_{\mathcal{X}} \varphi_x(x)\ d\mu^n_{x,t}(x) &= \frac{d}{dt}\int_{\mathcal{X}\times\mathbb{R}^+} w_x\varphi_x(x)\ d\nu^n_{x,t}(x,w_x) = \frac{d}{dt}\sum_{i=1}^n w^i_{x,t}\varphi_x(X^i_t) \\
&= \sum_{i=1}^n \frac{dw^i_{x,t}}{dt}\varphi_x(X^i_t) + \sum_{i=1}^n w^i_{x,t}\nabla_x\varphi_x(X^i_t)\cdot\frac{dX^i_t}{dt} \\
&= \int_{\mathcal{X}\times\mathbb{R}^+} w_x\mathcal{L}^{(n)}_{x,t}\varphi_x(x)\ d\nu^n_{x,t}(x,w_x) = \int_{\mathcal{X}} \mathcal{L}^{(n)}_{x,t}\varphi_x(x)\ d\mu^n_{x,t}(x)
\end{aligned} \tag{58}$$

The analogous equation holds for $\mu^n_{x,t}$:

$$\frac{d}{dt}\int_{\mathcal{Y}} \varphi_y(y)\ d\mu^n_{y,t}(y) = \int_{\mathcal{Y}} \mathcal{L}^{(n)}_{y,t}\varphi_y(y)\ d\mu^n_{y,t}(y) \tag{59}$$

Formally taking the limit $n \to \infty$ on (58) and (59) yields

$$\frac{d}{dt}\int_{\mathcal{X}} \varphi_x(x)\ d\mu_{x,t}(x) = \int_{\mathcal{X}} \mathcal{L}_{x,t}\varphi_x(x)\ d\mu_{x,t}(x)$$
$$\frac{d}{dt}\int_{\mathcal{Y}} \varphi_y(y)\ d\mu_{y,t}(y) = \int_{\mathcal{Y}} \mathcal{L}_{y,t}\varphi_y(y)\ d\mu_{y,t}(y),$$

where

$$\mathcal{L}_{x,t}\varphi_x(x) = -\gamma\nabla_x \int \ell(x,y)d\mu_{y,t} \cdot \nabla_x\varphi_x(x) + \alpha\left(-\int \ell(x,y)d\mu_{y,t} + \mathcal{L}(\mu_{x,t},\mu_{y,t})\right)$$
$$\mathcal{L}_{y,t}\varphi_y(y) = \gamma\nabla_y \int \ell(x,y)d\mu_{x,t} \cdot \nabla_y\varphi_y(x) + \alpha\left(\int \ell(x,y)d\mu_{x,t} - \mathcal{L}(\mu_{x,t},\mu_{y,t})\right)$$

and $\mu_{x,0}, \mu_{y,0}$ are set as in (52).

To make the limit $n \to \infty$ rigorous, an argument analogous to Theorem 2.6 of Chizat and Bach [2018] would result in almost sure convergence of the 2-Wasserstein distances between the empirical and the mean field measures. In our case almost sure convergence of the squared distance implies convergence of the expectation of the squared distance through dominated convergence, and hence the almost sure convergence result is stronger. Nonetheless, such an argument would require proving uniqueness of the mean field measure PDE through some notion of geodesic convexity, which is not clear in our case.

G Auxiliary material

G.1 ε-Nash equilibria and the Nikaido-Isoda error

Recall that an ε-NE (μ_x, μ_y) satisfies $\forall \mu_x^* \in \mathcal{P}(\mathcal{X}),\ \mathcal{L}(\mu_x, \mu_y) \leqslant \mathcal{L}(\mu_x^*, \mu_y) + \varepsilon$ and $\forall \mu_y^* \in \mathcal{P}(\mathcal{Y}),\ \mathcal{L}(\mu_x, \mu_y) \geqslant \mathcal{L}(\mu_x, \mu_y^*) - \varepsilon$. That is, each player can improve its value by at most ε by deviating from the equilibrium strategy, supposing that the other player is kept fixed.

Recall the Nikaido-Isoda error defined in (2). This equation can be rewritten as:

$$\mathrm{NI}(\mu_x, \mu_y) = \sup_{\mu_y^* \in \mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x, \mu_y^*) - \mathcal{L}(\mu_x, \mu_y) + \mathcal{L}(\mu_x, \mu_y) - \inf_{\mu_x^* \in \mathcal{P}(\mathcal{X})} \mathcal{L}(\mu_x^*, \mu_y)\ .$$

The terms $\sup_{\mu_y^* \in \mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x, \mu_y^*) - \mathcal{L}(\mu_x, \mu_y) > 0$ measure how much player y can improve its value by deviating from μ_y while μ_x stays fixed. Analogously, the terms $\mathcal{L}(\mu_x, \mu_y) - \inf_{\mu_x^* \in \mathcal{P}(\mathcal{X})} \mathcal{L}(\mu_x^*, \mu_y) > 0$ measure how much player x can improve its value by deviating from μ_x while μ_y stays fixed.

Notice that

$$\forall \mu_x^* \in \mathcal{P}(\mathcal{X}),\ \mathcal{L}(\mu_x, \mu_y) \leqslant \mathcal{L}(\mu_x^*, \mu_y) + \varepsilon \iff \mathcal{L}(\mu_x, \mu_y) - \inf_{\mu_x^* \in \mathcal{P}(\mathcal{X})} \mathcal{L}(\mu_x^*, \mu_y) \leqslant \varepsilon$$
$$\forall \mu_y^* \in \mathcal{P}(\mathcal{Y}),\ \mathcal{L}(\mu_x, \mu_y) \geqslant \mathcal{L}(\mu_x, \mu_y^*) - \varepsilon \iff \sup_{\mu_y^* \in \mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x, \mu_y^*) - \mathcal{L}(\mu_x, \mu_y) \leqslant \varepsilon$$

Thus, an ε-Nash equilibrium (μ_x, μ_y) fulfills $\mathrm{NI}(\mu_x, \mu_y) \leqslant 2\varepsilon$, and any pair (μ_x, μ_y) such that $\mathrm{NI}(\mu_x, \mu_y) \leqslant \varepsilon$ is an ε-Nash equilibrium.

G.2 Example: failure of the Interacting Wasserstein Gradient Flow

Let us consider the polynomial $f(x) = 5x^4 + 10x^2 - 2x$, which is an asymmetric double well as shown in Figure 4.

Let us define the loss $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as $\ell(x, y) = f(x) - f(y)$. That is, the two players are non-interacting and hence we obtain $V_x(x, \mu_y) = f(x) + K$, $V_y(y, \mu_x) = -f(y) + K'$. This means that the IWGF in equation (6) becomes two independent Wasserstein Gradient Flows

$$\begin{aligned} \partial_t \mu_x &= \nabla \cdot (\mu_x f'(x)), \quad \mu_x(0) = \mu_{x,0}, \\ \partial_t \mu_y &= -\nabla \cdot (\mu_y f'(y)), \quad \mu_y(0) = \mu_{y,0}. \end{aligned}$$

The particle flows in (3) become

$$\frac{dx_i}{dt} = -f'(x_i), \quad \frac{dy_i}{dt} = f'(y_i).$$

That is, the particles of player x follow the gradient flow of f and the particles of player y follow the gradient flow of $-f$. It is clear from Figure 4 that if the initializations $x_{0,i}, y_{0,i}$ are on the left of the barrier, they will not end up in the global minimum f (resp., the global maximum of $-f$). And in this case, the pair of measures supported on the global minimum of f is the only (pure) Nash equilibrium.

The game given by ℓ does not fall exactly in the framework that we describe in this work because ℓ is not defined on compact spaces. However, it is easy to construct very similar continuously differentiable functions on compact spaces that display the same behavior.

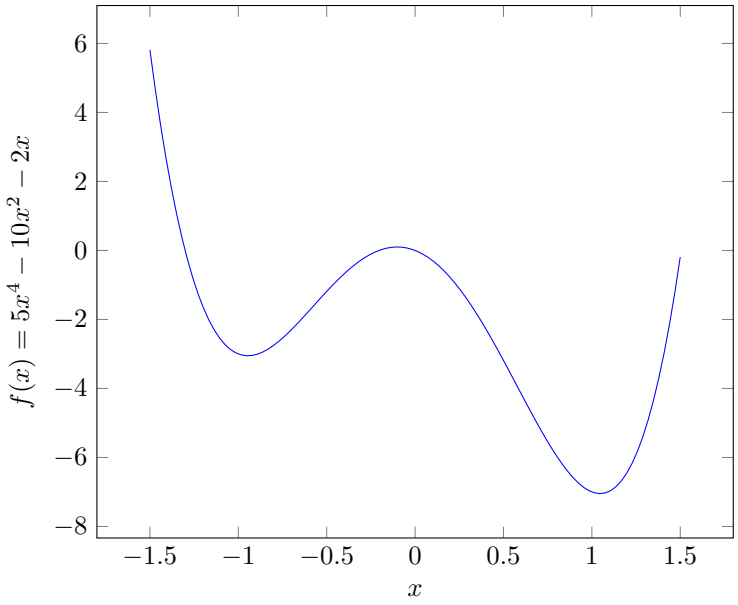

Figure 4: Plot of the function $f(x) = 5x^4 + 10x^2 - 2x$.

G.3 Link between Interacting Wasserstein Gradient Flow and interacting particle gradient flows

Recall (3):

$$\frac{dx_i}{dt} = -\frac{1}{n}\sum_{j=1}^{n} \nabla_x \ell(x_i, y_j), \quad \frac{dy_i}{dt} = \frac{1}{n}\sum_{j=1}^{n} \nabla_x \ell(x_j, y_i).$$

Let $\Phi_t = (\Phi_{x,t}, \Phi_{y,t}) : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{X}^n \times \mathcal{Y}^n$ be the flow mapping initial conditions $\mathbf{X}_0 = (x_{i,0})_{i\in[1:n]}, \mathbf{Y}_0 = (y_{i,0})_{i\in[1:n]}$ to the solution of (3). Let $\mu^n_{x,t} = \frac{1}{n}\sum_{i=1}^n \delta_{\Phi^{(i)}_{x,t}(\mathbf{X}_0,\mathbf{Y}_0)}, \mu^n_{y,t} = \frac{1}{n}\sum_{i=1}^n \delta_{\Phi^{(i)}_{y,t}(\mathbf{X}_0,\mathbf{Y}_0)}$. For all $\psi_x \in \mathcal{C}(\mathcal{X})$,

$$\begin{aligned}
\frac{d}{dt}\int_{\mathcal{X}} \psi_x(x)\, d\mu^n_{x,t}(x) &= \frac{1}{n}\sum_{i=1}^n \frac{d}{dt}\psi_x(\Phi^{(i)}_{x,t}(\mathbf{X}_0,\mathbf{Y}_0)) \\
&= \frac{1}{n}\sum_{i=1}^n \nabla_x \psi_x(\Phi^{(i)}_{x,t}(\mathbf{X}_0,\mathbf{Y}_0)) \cdot \left(-\frac{1}{n}\sum_{j=1}^n \nabla_x \ell(\Phi^{(i)}_{x,t}(\mathbf{X}_0,\mathbf{Y}_0), \Phi^{(j)}_{y,t}(\mathbf{X}_0,\mathbf{Y}_0))\right) \\
&= -\frac{1}{n}\sum_{i=1}^n \nabla_x \psi_x(\Phi^{(i)}_{x,t}(\mathbf{X}_0,\mathbf{Y}_0)) \cdot \nabla_x V_x(\mu^n_{y,t}, \Phi^{(i)}_{x,t}(\mathbf{X}_0,\mathbf{Y}_0)) \\
&= -\int_{\mathcal{X}} \nabla_x \psi_x(x) \cdot \nabla_x V_x(\mu^n_{y,t}, x)\, d\mu^n_{x,t}(x),
\end{aligned}$$

which is the first line of (6). The second line follows analogously.

G.4 Minimax problems and Stackelberg equilibria

Several machine learning problems, including GANs, are framed as a minimax problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \ell(x, y).$$

A minimax point (also known as a Stackelberg equilibrium or sequential equilibrium) is a pair $(\tilde{x}, \tilde{y})$ at which the minimum and maximum of the problem are attained, i.e.

$$\begin{cases} \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \ell(x,y) = \max_{y\in\mathcal{Y}} \ell(\tilde{x},y) \\ \max_{y\in\mathcal{Y}} \ell(\tilde{x},y) = \ell(\tilde{x},\tilde{y}) \end{cases}.$$

We consider the lifted version of the minimax problem (G.4) in the space of probability measures.

$$\min_{\mu_x\in\mathcal{P}(\mathcal{X})} \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x,\mu_y). \tag{60}$$

By the generalized Von Neumann's minimax theorem, a Nash equilibrium of the game given by $\mathcal{L}$ is a solution of the lifted minimax problem (60) (see Lemma 12 in the case $\varepsilon = 0$).

The converse is not true: minimax points (solutions of (60)) are not necessarily mixed Nash equilibria even in the case where the loss function is convex-concave. An example is $\mathcal{L} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ given by $\mathcal{L}(\mu_x,\mu_y) = \iint (x^2 + 2xy)\, d\mu_x\, d\mu_y$. Let $\mathcal{M}$ be the set of measures $\mu \in \mathcal{P}(\mathbb{R})$ such that $\int x\, d\mu = 0$. Notice that any pair (δ_0, μ_y) with $\mu_y \in \mathcal{P}(\mathbb{R})$ is a minimax point. That is because

$$\max_{\mu_y\in\mathcal{P}(\mathbb{R})} \mathcal{L}(\mu_x,\mu_y) = \begin{cases} +\infty & \text{if } \mu_x \notin \mathcal{M} \\ \text{positive} & \text{if } \mu_x \in \mathcal{M}\setminus\{\delta_0\} \\ 0 & \text{if } \mu_x = \delta_0, \end{cases}$$

and hence $\delta_0 = \operatorname{argmin}_{\mu_x\in\mathcal{P}(\mathbb{R})} \max_{\mu_y\in\mathcal{P}(\mathbb{R})} \mathcal{L}(\mu_x,\mu_y)$. But if $\mu_x = \delta_0$, we have $\operatorname{argmax}_{\mu_y\in\mathcal{P}(\mathbb{R})} \mathcal{L}(\mu_x,\mu_y) = \mathcal{P}(\mathbb{R})$, because for all measures $\mu_y \in \mathcal{P}(\mathbb{R})$, $\mathcal{L}(\delta_0,\mu_y) = 0$. However, for $\mu_y \notin \mathcal{M}$, $\mathcal{L}(\mu_x,\mu_y)$ as a function of μ_x does not have a minimum at δ_0, but at $\delta_{-\int y\, d\mu_y}$. Hence, the only mixed Nash equilibria are of the form (δ_0,μ_y), with $\mu_y \in \mathcal{M}$.

The intuition behind the counterexample is that minimax points only require the minimizing player to be non-exploitable, but the maximizing player is only subject to a weaker condition.

We define a ε-minimax point (or ε-Stackelberg equilibrium) of an objective $\mathcal{L}(\mu_x,\mu_y)$ as a couple $(\tilde{\mu}_x,\tilde{\mu}_y)$ such that

$$\begin{cases} \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x,\mu_y) \geqslant \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\tilde{\mu}_x,\mu_y) - \varepsilon \\ \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\tilde{\mu}_x,\mu_y) \leqslant \mathcal{L}(\tilde{\mu}_x,\tilde{\mu}_y) + \varepsilon \end{cases}.$$

Lemma 12. *An ε-Nash equilibrium is a 2ε-minimax point, and it holds that*

$$\min_{\mu_x\in\mathcal{P}(\mathcal{X})} \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x,\mu_y) - \varepsilon \leqslant \mathcal{L}(\hat{\mu}_x,\hat{\mu}_y) \leqslant \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \mathcal{L}(\mu_x,\hat{\mu}_y) + \varepsilon$$

Proof. Let $(\hat{\mu}_x,\hat{\mu}_y)$ be an ε-Nash equilibrium. Notice that $\max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \mathcal{L}(\tilde{\mu}_x,\mu_y) \leqslant \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\tilde{\mu}_x,\mu$
Also,

$$\begin{aligned} \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\mu_x,\mu_y) &\leqslant \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\hat{\mu}_x,\mu_y) \leqslant \mathcal{L}(\hat{\mu}_x,\hat{\mu}_y) + \varepsilon \leqslant \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \mathcal{L}(\mu_x,\hat{\mu}_y) + 2\varepsilon \\ &\leqslant \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \mathcal{L}(\mu_x,\hat{\mu}_y) + 2\varepsilon \end{aligned} \tag{61}$$

and this yields the chain of inequalities in the statement of the theorem. The condition $\max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\tilde{\mu}_x,\mu_y) \leqslant \mathcal{L}(\tilde{\mu}_x,\tilde{\mu}_y) + \varepsilon$ of the definition of ε-minimax point follows directly from the definition of an ε-Nash equilibrium. Using part of (61), we get

$$\max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\hat{\mu}_x,\mu_y) - 2\varepsilon \leqslant \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \mathcal{L}(\mu_x,\hat{\mu}_y) \leqslant \min_{\mu_x\in\mathcal{P}(\mathcal{X})} \max_{\mu_y\in\mathcal{P}(\mathcal{Y})} \mathcal{L}(\tilde{\mu}_x,\mu_y),$$

which is the first condition of a 2ε-minimax. □

Lemma 12 provides the link between approximate Nash equilibria and approximate Stackelberg equilibria, and it allows to translate our convergence results into minimax problems such as GANs.

G.5 Itô SDEs on Riemannian manifolds: a parametric approach

We provide a brief summary on how to deal with SDEs on Riemannian manifolds and their corresponding Fokker-Planck equations (see Chapter 8 of Chirikjian [2009]). While ODEs have a straightforward translation into manifolds, the same is not true for SDEs. Recall that the definitions of the gradient and divergence for Riemannian manifolds are

$$\nabla \cdot X = |g|^{-1/2}\partial_i(|g|^{1/2}X^i), \quad (\nabla f)^i = g^{ij}\partial_j f,$$

where g_{ij} is the metric tensor, $g^{ij} = (g_{ij})^{-1}$ and $|g| = \det(g_{ij})$. We use the Einstein convention for summing repeated indices.

The parametric approach to SDEs in manifolds is to define the SDE for the variables $\mathbf{q} = (q_1, \cdots, q_d)$ of a patch of the manifold:

$$d\mathbf{q} = \mathbf{h}(\mathbf{q}, t)dt + H(\mathbf{q}, t)d\mathbf{w}. \tag{62}$$

The corresponding forward Kolmogorov equation is

$$\frac{\partial f}{\partial t} + |g|^{-1/2}\sum_{i=1}^{d}\frac{\partial}{\partial q_i}\left(|g|^{1/2}h_i f\right) = \frac{1}{2}|g|^{-1/2}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial q_i \partial q_j}\left(|g|^{1/2}\sum_{k=1}^{D}H_{ik}H_{kj}^{\top}f\right), \tag{63}$$

which is to be understood in the weak form.

Assume that the manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$. If $\varphi : \mathcal{U}_{\mathbb{R}^d} \subseteq \mathbb{R}^d \to \mathcal{U} \subseteq \mathcal{M} \subseteq \mathbb{R}^D$ is the mapping corresponding to the patch $\mathcal{U}$ and (62) is defined on $\mathcal{U}_{\mathbb{R}^d}$, let us set $H(\mathbf{q}) = (D\varphi(\mathbf{q}))^{-1}$. In this case, $\sum_k H_{ik}H_{kj}^{\top} = \sum_k (D\varphi)_{ik}^{-1}((D\varphi)_{kj}^{-1})^{\top} = g^{ij}(\mathbf{q})$. Hence, the right hand side of (63) becomes

$$\begin{aligned}
&\frac{1}{2}|g|^{-1/2}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial q_i \partial q_j}\left(|g|^{1/2}g^{ij}f\right)\\
&= |g|^{-1/2}\sum_{i=1}^{d}\frac{\partial}{\partial q_i}\left(|g|^{1/2}\tilde{h}_i f\right) + \frac{1}{2}|g|^{-1/2}\sum_{i,j=1}^{d}\frac{\partial}{\partial q_i}\left(|g|^{1/2}g^{ij}\frac{\partial}{\partial q_j}f\right)\\
&= |g|^{-1/2}\sum_{i=1}^{d}\frac{\partial}{\partial q_i}\left(|g|^{1/2}\tilde{h}_i f\right) + \frac{1}{2}|g|^{-1/2}\sum_{i,j=1}^{d}\frac{\partial}{\partial q_i}\left(|g|^{1/2}g^{ij}\frac{\partial}{\partial q_j}f\right)\\
&= \nabla \cdot (\tilde{\mathbf{h}}f) + \frac{1}{2}\nabla \cdot \nabla f
\end{aligned}$$

where

$$\tilde{h}_i(\mathbf{q}) = \frac{1}{2}\sum_{j=1}^{d}\left(|g(\mathbf{q})|^{-1/2}g^{ij}(\mathbf{q})\frac{\partial |G(\mathbf{q})|^{1/2}}{\partial q_j} + \frac{\partial g^{ij}(\mathbf{q})}{\partial q_j}\right)$$

Hence, we can rewrite (63) as

$$\frac{\partial f}{\partial t} = \nabla \cdot ((-\mathbf{h} + \tilde{\mathbf{h}})f) + \frac{1}{2}\nabla \cdot \nabla f$$

For this equation to be a Fokker-Planck equation with potential E (i.e. with a Gibbs equilibrium solution), we need $-\mathbf{h} + \tilde{\mathbf{h}} = \nabla E$, which implies $\mathbf{h} = -\nabla E + \tilde{\mathbf{h}}$.

We can convert an SDE in parametric form like (62) into an SDE on $\mathbb{R}^D$ by using Ito's lemma on $X = \varphi(\mathbf{q})$:

$$dX_i = d\varphi_i(\mathbf{q}) = \left(D\varphi_i(\mathbf{q})\mathbf{h}(\mathbf{q}) + \frac{1}{2}\mathrm{Tr}(H(\mathbf{q}, t)^{\top}(H\varphi_i)(\mathbf{q})H(\mathbf{q}, t))\right)dt + D\varphi_i(\mathbf{q})H(\mathbf{q}, t)d\mathbf{w} \tag{64}$$

If we set $H(\mathbf{q}) = (D\varphi(\mathbf{q}))^{-1}$ as before, $D\varphi(\mathbf{q})H(\mathbf{q},t)$ is the projection onto the tangent space of the manifold, i.e. $D\varphi(\mathbf{q})H(\mathbf{q},t)v = \mathrm{Proj}_{T_{\varphi(\mathbf{q})}M} v,\ \forall v \in \mathbb{R}^D$. In the case $\mathbf{h} = \nabla E + \tilde{\mathbf{h}}$, $D\varphi_i(\mathbf{q})\mathbf{h}(\mathbf{q}) = D\varphi_i(\mathbf{q})\nabla E(\mathbf{q}) + D\varphi_i(\mathbf{q})\tilde{\mathbf{h}}(\mathbf{q})$. It is very convenient to abuse the notation and denote $D\varphi(\mathbf{q})\nabla E(\mathbf{q})$ by $\nabla E(\varphi(\mathbf{q}))$. We also use $\hat{\mathbf{h}}(\varphi(\mathbf{q})) := D\varphi(\mathbf{q})\tilde{\mathbf{h}}(\mathbf{q}) + \frac{1}{2}\mathrm{Tr}(((D\varphi(\mathbf{q}))^{-1})^\top (H\varphi)(\mathbf{q})(D\varphi(\mathbf{q}))^{-1})$. Both definitions are well-defined because the variables are invariant by changes of coordinates. Hence, under these assumptions (64) becomes

$$dX = (-\nabla E(X) + \hat{\mathbf{h}}(X))\ dt + \mathrm{Proj}_{T_X M}(d\mathbf{w}) \tag{65}$$

In short that means that we can treat SDEs on embedded manifolds as SDEs on the ambient space by projecting the Brownian motions to the tangent space and adding a drift term $\hat{\mathbf{h}}$ that depends on the geometry of the manifold. Notice that for ODEs on manifolds the additional drift term does not appear and (65) reads simply $dX = \nabla E(X)dt$.

Notice that the forward Kolmogorov equation for (65) on $\mathbb{R}^D$ reads

$$\frac{d}{dt}\int f(x)\ d\mu_t(x) = \int (\nabla E(x) - \hat{\mathbf{h}}(x)) \cdot \nabla_x f(x) + \frac{1}{2}\mathrm{Tr}((\mathrm{Proj}_{T_x M})^\top H f(x) \mathrm{Proj}_{T_x M})\ d\mu_t(x), \tag{66}$$

for an arbitrary f.