Two-Scale Gradient Descent Ascent Dynamics Finds Mixed Nash Equilibria of Continuous Games: A Mean-Field Perspective

Yulong Lu [1]

Abstract

Finding the mixed Nash equilibria (MNE) of a two-player zero sum continuous game is an important and challenging problem in machine learning. A canonical algorithm to finding the MNE is the noisy gradient descent ascent method which in the infinite particle limit gives rise to the *Mean-Field Gradient Descent Ascent* (GDA) dynamics on the space of probability measures. In this paper, we first study the convergence of a two-scale Mean-Field GDA dynamics for finding the MNE of the entropy-regularized objective. More precisely we show that for each finite temperature (or regularization parameter), the two-scale Mean-Field GDA with a suitable *finite* scale ratio converges exponentially to the unique MNE without assuming the convexity or concavity of the interaction potential. The key ingredient of our proof lies in the construction of new Lyapunov functions that dissipate exponentially along the Mean-Field GDA. We further study the simulated annealing of the Mean-Field GDA dynamics. We show that with a temperature schedule that decays logarithmically in time the annealed Mean-Field GDA converges to the MNE of the original unregularized objective.

1. Introduction

Minmax learning underpins numerous machine learning methods including Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), adversarial training (Madry et al., 2018) and reinforcement learning (Busoniu et al., 2008). Minmax learning is often formulated as a zero-sum game between min and max players and hence can be formalized as a minmax optimization problem of the form

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

where the function K is the game objective, x and y represent the player strategies. When K is nonconvex in x or nonconcave in y, finding the pure Nash (Nash, 1951) equilibria of K is difficult and sometimes impossible since the pure Nash equilibrium may not even exist. On the other hand, the mixed Nash equilibria (MNE) (Glicksberg, 1952) where the pure strategies are replaced by mixed strategies modeled by a probability distribution over the set of strategies, do exist for more general objective functions. Formally the mixed Nash equilibria consist of pairs of probability distributions μ and ν that solve

$$\begin{aligned}&\min_{\mu\in\mathcal{P}(\mathcal{X})}\max_{\nu\in\mathcal{P}(\mathcal{Y})} E_0(\mu,\nu), \text{ where}\\ &E_0(\mu,\nu):=\int_{\mathcal{X}}\int_{\mathcal{Y}} K(x,y)\mu(dx)\nu(dy).\end{aligned} \tag{1}$$

Here $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ denote the space of probability measures (or the set of all mixed strategies) on the state spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Thanks to Glicksberg's theorem (Glicksberg, 1952), a MNE exists if $\mathcal{X}$ and $\mathcal{Y}$ are finite or if $\mathcal{X}$ and $\mathcal{Y}$ are compact and K is continuous. While MNE exist in much generality, it is still difficult to find them. Several progress have been made recently for finding MNE of high dimensional game problems with applications in GANs. For instance, the work (Hsieh et al., 2019) proposed a mirror-descent algorithm for finding MNE of (1) and applied the algorithms for Wasserstein GANs with provable convergence guarantees. The recent work (Domingo-Enrich et al., 2020; Ma & Ying, 2021) proposed and analyzed an entropy-regularized version of (1):

$$\begin{aligned}&\min_{\mu\in\mathcal{P}(\mathcal{X})}\max_{\nu\in\mathcal{P}(\mathcal{Y})} E_\tau(\mu,\nu), \text{ where}\\ &E_\tau(\mu,\nu):=\int_{\mathcal{X}}\int_{\mathcal{Y}} K(x,y)d\nu(y)d\mu(x)-\tau\mathcal{H}(\mu)+\tau\mathcal{H}(\nu).\end{aligned} \tag{2}$$

Here $\mathcal{H}(\mu)=-\int\log\frac{d\mu}{dx}d\mu$ is the entropy functional of the probability measure μ, and the parameter $\tau>0$ is the entropy regularization parameter(or temperature). Observe that the objective functional E_τ is strongly convex in μ

[1]Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts, USA . Correspondence to: Yulong Lu <yulonglu@umass.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

and strongly concave in ν. As a result of the famous Von Neumann's minmax theorem (Von. Neumann, 1928; Sion, 1958; Nikaido & Isoda, 1955; Nikaido, 1954), one has that

$$\min_{\mu\in\mathcal{P}(\mathcal{X})}\max_{\nu\in\mathcal{P}(\mathcal{Y})} E_\tau(\mu,\nu) = \max_{\nu\in\mathcal{P}(\mathcal{Y})}\min_{\mu\in\mathcal{P}(\mathcal{X})} E_\tau(\mu,\nu). \quad (3)$$

Under very mild assumptions on K, Problem (2) has a unique MNE (μ^*,ν^*) (see a proof in (Domingo-Enrich et al., 2020)) in the sense that for any $\mu\in\mathcal{P}(\mathcal{X}),\nu\in\mathcal{P}(\mathcal{Y})$,

$$E_\tau(\mu^*,\nu) \le E_\tau(\mu^*,\nu^*) \le E_\tau(\mu,\nu^*).$$

Furthermore, the MNE (μ^*,ν^*) is given by the unique solution of the fixed point equations

$$\begin{aligned}\mu^*(dx) &\propto \exp\Big(-\int_{\mathcal{Y}}\tau^{-1}K(x,y)d\nu^*(y)\Big),\\ \nu^*(dy) &\propto \exp\Big(\int_{\mathcal{X}}\tau^{-1}K(x,y)d\mu^*(x))\Big).\end{aligned} \quad (4)$$

In the setting where both players play finite mixtures of N strategies, i.e. $\mu=\sum_{j=1}^N\frac{1}{N}\delta_{X^j}$ and $\nu=\sum_{j=1}^N\frac{1}{N}\delta_{Y^j}$, the following noisy gradient descent-ascent dynamics is perhaps one of the most natural algorithms for finding MNE of (1) and (2):

$$\begin{aligned}dX_t^i &= -\frac{1}{N}\sum_{j=1}^N\nabla_x K(X_t^i,Y_t^j)dt+\sqrt{2\tau}dW_t^i,\\ dY_t^i &= \frac{\eta}{N}\sum_{j=1}^N\nabla_y K(X_t^j,Y_t^i)dt+\sqrt{2\tau\eta}dB_t^i,\end{aligned} \quad (5)$$

where $W_t^i,B_t^i,i=1,\cdots N$ are independent Brownian motions. The parameter $\eta>0$ in (5) represents the ratio of timescales at which the gradient descent and ascent dynamics is undergoing. When $\eta=1$, there is no timescale separation. However, GDA with different time scales is commonly used in minmax optimization (Jin et al., 2020; Lin et al., 2020) and sometimes leads to better convergence property (Heusel et al., 2017).

In the limit of large number of strategies ($N\to\infty$), the empirical measures $\mu_t^N=\frac{1}{N}\sum_{i=1}^N\delta_{X_t^i}$ and $\nu_t^N=\frac{1}{N}\sum_{i=1}^N\delta_{Y_t^i}$ of the interacting particle, with $\{X_0^i\}$ and $\{Y_0^i\}$ identically independent sampled from the initial distributions μ_0 and ν_0, converge weakly within finite time to the solutions μ_t and ν_t of the *Mean-Field GDA* (also named the Interacting Wasserstein Gradient Flow in (Domingo-Enrich et al., 2020)) dynamics:

$$\begin{aligned}\partial_t\mu_t &= \nabla\cdot(\mu_t\int_{\mathcal{Y}}\nabla_x K(x,y)d\nu_t(y))+\tau\Delta\mu_t,\\ \partial_t\nu_t &= \eta\Big(-\nabla\cdot(\nu_t\int_{\mathcal{X}}\nabla_y K(x,y)d\mu_t(x))+\tau\Delta\nu_t\Big).\end{aligned} \quad (6)$$

The dynamics (6) can be viewed as the minmax analogue of the Mean-Field Langevin (Nitanda et al., 2022; Chizat, 2022; Hu et al., 2021) dynamics that arises naturally from the mean field analysis of optimization of two-layer neural networks.

Despite its simplicity and popularity, the long-time convergence of the Mean-Field GDA (6) is still not well understood, except in the extreme quasi-static (Ma & Ying, 2021) regime where the ascent dynamics is infinitely faster or slower than the descent dynamics ($\eta=0$ or $+\infty$). This motivates us to study the first question, which to the best of our knowledge, remains open:

***Question 1**: Does the Mean-Field GDA (6) with a finite scale ratio $\eta>0$ converge to the unique MNE? If so, what is the rate of convergence?*

We provide an affirmative answer to **Question 1** and establish a quantitative exponential convergence of (6) to the MNE for any fixed $\tau>0$. Furthermore, motivated by the simulated annealing results for Langevin dynamics (Holley et al., 1989; Tang & Zhou, 2021; Raginsky et al., 2017) in the context of global optimization and the recent results for Mean-Field Langevin dynamics (Chizat, 2022; Nitanda et al., 2022) in the setting of training two-layer neural networks, it is natural to ask what would happen to the dynamics (6) in the annealed regime that $\tau\to0$ as $t\to\infty$. In particular, it is natural to ask

***Question 2:** Does the annealed Mean-Field GDA (6) with a decreasing temperature $\tau=\tau_t$ converge to a MNE of the unregularized objective E_0 defined in (1)?*

Summary of contributions

In this work we first address **Question 1** by providing the first convergence analysis of the two-scale continuous-time Mean-Field GDA dynamics (6) with a **finite** scale ratio. This improves substantially the earlier convergence results by (Ma & Ying, 2021; Domingo-Enrich & Bruna, 2022) on Mean-Field GDA in the quasistatic setting where the scale ratio either vanishes or explodes. The key ingredient of the proof is the construction of a novel Lyapunov function which decreases exponentially along the dynamics (6). We then further address **Question 2** by proving that the annealed Mean-Field GDA converges to a global MNE of the original unregularized objective E_0. The latter result, to the best our knowledge, is the first rigorous justification of the convergence of GDA to MNE in the mean field regime.

We highlight the major contributions as follows.

- For any fixed temperature $\tau>0$, we show that in the fast ascent/descent regime (or the scale ratio η is either larger or smaller than certain threshold), the Mean-Field GDA dynamics (6) converges exponentially to

the MNE of the entropy-regularized objective E_τ with respect to certain Lyapunov functions; see Theorem 2.4.

- The convergence in Lyapunov functions further implies the global exponential convergence of the dynamics with respect to the relative entropy (see Theorem 2.7). Our (non-asymptotic) convergence results hold for any positive temperature (regularization parameter) and do not assume convexity or concavity of the interaction potential K. The convergence rate is characterized explicitly in terms of τ, bounds on K and diameters of the state spaces.
- We also study the simulated annealing of the Mean-Field GDA dynamics (6). We prove that with the cooling schedule τ_t decaying with a logarithimc rate, the Mean-Field GDA dynamics (6) converges to the global mixed Nash equilibrium of E_0 with respect to the Nikaidò-Isoda error (defined in (7)). See Theorem 2.8 for the precise statements.

1.1. Related work

MINMAX OPTIMIZATION

Most previous works about minmax optimization focused on the analysis of discrete-time optimization schemes in finite dimensional Euclidean spaces under various assumptions on the objective function. In the convex-concave setting, global convergence guarantees were obtained for various optimization schemes, such as GDA (Daskalakis & Panageas, 2019), mirror descent (Mertikopoulos et al., 2018) and Hamiltonian gradient descent (Abernethy et al., 2019). In the non-convex and/or non-concave setting, local convergence to several notions of stationary points were studied in (Daskalakis & Panageas, 2018; Lin et al., 2020). In the special case where the objective satisfies a two-sided Polyak-Łojasiewicz (PL) condition, convergence to global Nash equilibria were recently obtained in (Yang et al., 2020; 2022; Doan, 2022) for two-scale GDA.

MEAN-FIELD ANALYSIS

Our study of the Mean-Field GDA dynamics for minmax optimization is strongly motivated by a growing amount of work on mean-field analysis of gradient descent algorithms for minimizing neural networks. The study of the latter originates from the work on two-layer networks (Mei et al., 2018; Sirignano & Spiliopoulos, 2020; Rotskoff & Vanden-Eijnden, 2022; Chizat & Bach, 2018), to residual networks (Lu et al., 2020; Wojtowytsch et al., 2020) and general deep neural networks (Araújo et al., 2019; Sirignano & Spiliopoulos, 2022; Nguyen, 2019). These results show that the mean field dynamics can be realized as Wasserstein gradient flows of certain nonlinear energy functionals on the space of probability measures. Global convergence of the Wasserstein gradient flows were obtained for the mean-field Langevin dynamics in (Mei et al., 2018; Chizat, 2022; Nitanda et al., 2022; Hu et al., 2021) and for the noiseless mean-field dynamics in (Chizat & Bach, 2018). We note that mean-field Langevin dynamics can be viewed as a special Mckean-Vlasov dynamics whose long-time convergence are often established by assuming either the small interaction strength or large noise level; see e.g. (Eberle et al., 2019; Hu et al., 2021).

Among the existing work on mean field Langevin dynamics, (Chizat, 2022) is closest to ours where the author proved the exponential convergence of Mean-Field Langevin dynamics under a certain uniform log-Sobolev inequality assumption, and also proved the convergence of the annealed Mean-Field Langevin to the global minimizers of the objective function. Our results are parallel to the results of (Chizat, 2022), but require new proof strategies. A key ingredient of our proof is constructing new Lyapunov functions that decay exponentially along the Mean-Field GDA (6).

In the context of minmax optimization, (Domingo-Enrich et al., 2020) first proposed and analyzed the noisy gradient descent ascent flow (5) without scale separation ($\eta = 1$) for finding the MNE. Specifically, the authors therein proved the convergence of (5) to the mean field GDA (6) in the limit of large agent size, and characterized the equilibrium of (6) as the unique solution of the fixed point equation (4). However, (Domingo-Enrich et al., 2020) did not provide a proof of the long-time convergence of (6). In (Ma & Ying, 2021), the authors obtained convergence guarantees for the mean field GDA (6) in the quasi-static regime where the ascent dynamics is infinitely faster or slower than the descent dynamics (i.e. $\eta = +\infty$ or 0). One of the major contributions of the present work goes beyond the quasi-static regime and provides the first proof for exponential convergence of (6) with a finite scale ratio.

WASSERSTEIN-FISHER-RAO GRADIENT FLOWS

As an alternative to Wasserstein gradient flows, gradient flow with respect to the Wasserstein-Fisher-Rao (WFR) metric (Chizat et al., 2018; Kondratyev et al., 2016; Laschos & Mielke, 2019) has recently sparked a large amount of research on PDEs (Brenier & Vorotnikov, 2020; Kondratyev & Vorotnikov, 2019), neural network training (Rotskoff et al., 2019) and statistical sampling (Lu et al., 2019; 2022). Especially, WFR gradient flows give rise to birth-death interacting particle dynamics that enables "teleportation" of mass and locations of particles and hence lead to better convergence properties than the Langevin dynamics in certain scenarios (Rotskoff et al., 2019; Lu et al., 2019; 2022). More recently, (Domingo-Enrich et al., 2020) adopts a WFR-gradient flow for two-player zero-mean continuous games

and proves that the time-averaging of the WFR-dynamics converges to the MNE in the regime where Fisher-Rao dynamics dominates the Wasserstein counterpart. In (Wang & Chizat, 2022), a new particle algorithm motivated by the implicit time-discretization of WFR-gradient flow was proposed for finding the MNE in the space of atomic measures. The authors proved a local exponential convergence of the proposed algorithm under certain non-degeneracy assumptions on the game objective and the assumption that the MNE is unique.

1.2. Notations

For a measurable space $\mathcal{X}$, we use $\mathcal{P}(\mathcal{X})$ to denote the space of probability measures on $\mathcal{X}$. Given a canonical Borel measure dx on $\mathcal{X}$, such as the uniform measure considered in the paper, if μ is absolutely continuous with respect to dx, we denote the Radon-Nikodym derivative by $\frac{d\mu}{dx}$. We define the shannon entropy of μ by $\mathcal{H}(\mu) = -\int_{\mathcal{X}} \log(\frac{d\mu}{dx}) d\mu(x)$. The relative entropy (or Kullback-Leibler divergence) $\mathcal{H}(\mu_1|\mu_2)$ and the relative Fisher information $\mathcal{I}(\mu_1|\mu_2)$ between two probability measures μ_1 and μ_2 are defined by

$$\mathcal{H}(\mu_1|\mu_2) = \int \log\Big(\frac{d\mu_1}{d\mu_2}\Big) d\mu_1,$$
$$\mathcal{I}(\mu_1|\mu_2) = \int \Big|\nabla \log\Big(\frac{d\mu_1}{d\mu_2}\Big)\Big|^2 d\mu_1$$

if μ_1 is absolutely continuous with respect to μ_2 and $+\infty$ otherwise. Given an approximate mixed Nash equilibrium (μ, ν), we define the Nikaidò-Isoda (Nikaido & Isoda, 1955) error $\mathrm{NI}(\mu, \nu)$ by

$$\mathrm{NI}(\mu, \nu) := \sup_{\nu' \in \mathcal{P}(\mathcal{Y})} E_0(\mu, \nu') - \inf_{\mu' \in \mathcal{P}(\mathcal{X})} E_0(\mu', \nu). \quad (7)$$

Note that by definition $\mathrm{NI}(\mu, \nu) \geq 0$ for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ and $\mathrm{NI}(\mu, \nu) = 0$ if and only if (μ, ν) is a mixed Nash equilibrium.

1.3. Organization

The remaining of the paper is organized as follows. In Section 2 we state the key assumptions and present the main results of the paper. In Section 3 we discuss applications of our results in training GANs and adversarial learning of PDEs. All the proofs are provided in the Appendix. We finish with several conclusion remarks and future directions in Section 4.

2. Assumptions and Main Results

Let us first recall the minmax problem of the entropy-regularized continuous game:

$$\min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{\nu \in \mathcal{P}(\mathcal{Y})} E_\tau(\mu, \nu), \text{ where}$$
$$E_\tau(\mu, \nu) := \int_{\mathcal{X}} \int_{\mathcal{Y}} K(x, y) d\nu(y) d\mu(x) - \tau \mathcal{H}(\mu) + \tau \mathcal{H}(\nu). \quad (8)$$

When the regularization parameter (or temperature) $\tau = 0$, we write

$$E_0(\mu, \nu) = \int_{\mathcal{X}} \int_{\mathcal{Y}} K(x, y) d\nu(y) d\mu(x).$$

We are interested in the global convergence of the Mean-Field GDA dynamics to the MNE of (8):

$$\partial_t \mu_t = \nabla \cdot (\mu_t \int_{\mathcal{Y}} \nabla_x K(x, y) d\nu_t(y)) + \tau \Delta \mu_t,$$
$$\partial_t \nu_t = \eta\Big(- \nabla \cdot (\nu_t \int_{\mathcal{X}} \nabla_y K(x, y) d\mu_t(x)) + \tau \Delta \nu_t \Big).$$

In what follows, we may suppress the dependence of E_τ on τ and write $E = E_\tau$ to simplify notations when the $\tau > 0$ is fixed; we will indicate such dependence in later discussions on the annealed dynamics where $\tau = \tau_t$ shrinks to zero in time.

2.1. Assumptions

Following (Domingo-Enrich & Bruna, 2022) and (Wang & Chizat, 2022), we make the following assumptions on the state spaces $\mathcal{X}$ and $\mathcal{Y}$.

Assumption 2.1. The state spaces $\mathcal{X}$ and $\mathcal{Y}$ are (i) either smooth compact Riemannian manifolds without boundary such that the eigenvalues of the Ricci curvature are strictly positive everywhere or (ii) Euclidean tori (with possibly unequal dimensions).

We also make the following smoothness assumption on the potential K.

Assumption 2.2. The function $K \in C^2(\mathcal{X} \times \mathcal{Y})$ and there exists $K_{xy} > 0$ such that

$$\|\nabla^2_{xy} K\|_\infty \leq K_{xy}.$$

It will be very useful to introduce operators $\mathcal{K}^+ : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ and $\mathcal{K}^- : \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X})$ where

$$\mathcal{K}^+ \mu(dy) = \frac{1}{Z^+(\mu)} \exp\Big(\int_{\mathcal{X}} \tau^{-1} K(x, y) d\mu(x) \Big),$$
$$\mathcal{K}^- \nu(dx) = \frac{1}{Z^-(\nu)} \exp\Big(- \int_{\mathcal{Y}} \tau^{-1} K(x, y) d\nu(y) \Big). \quad (9)$$

Here $Z^-(\nu)$ and $Z^+(\mu)$ are the corresponding partition functions defined by

$$\begin{aligned} Z^-(\nu) &= \int_{\mathcal{X}} \exp\Big(-\int_{\mathcal{Y}} \tau^{-1}K(x,y)d\nu(y)\Big)dx, \\ Z^+(\mu) &= \int_{\mathcal{Y}} \exp\Big(\int_{\mathcal{X}} \tau^{-1}K(x,y)d\mu(x)\Big)dy. \end{aligned} \tag{10}$$

It is easy to verify by the definition of $E(\mu,\nu)$ that

$$\arg\max_{\nu\in\mathcal{P}(\mathcal{Y})} E(\mu,\nu) = \mathcal{K}^+(\mu), \quad \arg\min_{\mu\in\mathcal{P}(\mathcal{X})} E(\mu,\nu) = \mathcal{K}^-(\nu). \tag{11}$$

With the notations above, the MNE (μ^*,ν^*) of (8) is characterized as the unique solution of the following fixed point equations (see Theorem 1 of (Domingo-Enrich et al., 2020)):

$$\mu^* = \mathcal{K}^-(\nu^*), \quad \nu^* = \mathcal{K}^+(\mu^*). \tag{12}$$

The lemma below shows that under the assumptions above the measures $\mathcal{K}^+(\mu)$ and $\mathcal{K}^-(\nu)$ satisfy the logarithmic Sobolev inequalities (LSI) uniformly for any $\mu\in\mathcal{P}(\mathcal{X})$ and $\nu\in\mathcal{P}(\mathcal{Y})$.

Lemma 2.3. *There exists a constant $\lambda_{LS}>0$ such that for any $\mu\in\mathcal{P}(\mathcal{X})$ and $\nu\in\mathcal{P}(\mathcal{Y})$,*

$$\lambda_{LS}\mathcal{H}(\mu|\mathcal{K}^-\nu) \le \mathcal{I}(\mu|\mathcal{K}^-\nu), \lambda_{LS}\mathcal{H}(\nu|\mathcal{K}^+\mu) \le \mathcal{I}(\nu|\mathcal{K}^+\mu). \tag{13}$$

Lemma 2.3 follows from the classical Holley-Stroock's perturbation argument and the fact that the uniform measures on $\mathcal{X}$ and $\mathcal{Y}$ satisfy the LSI thanks to Assumption 2.1; see e.g. Proposition 5.2 of (Chizat, 2022) and Proposition 6 of (Domingo-Enrich & Bruna, 2022)). It is worthwhile to note that in the regime where $\tau^{-1}\|K\|_\infty \gg 1$, the log-Sobolev constant $\lambda_{LS}=\lambda_{LS}(\tau)$ satisfies that

$$\lambda_{LS}(\tau) = \mathcal{O}(\exp(-\tau^{-1}\|K\|_\infty)). \tag{14}$$

2.2. Exponential convergence of Mean-Field GDA with a fixed temperature

The goal of this section is to present the global convergence of the GDA dynamics (6) to the MNE (μ^*,ν^*). To this end, it will be useful to define the following functionals

$$\begin{aligned} \mathcal{L}_1(\mu) &:= \max_{\nu\in\mathcal{P}(\mathcal{Y})} E(\mu,\nu) - \min_{\mu\in\mathcal{P}(\mathcal{X})}\max_{\nu\in\mathcal{P}(\mathcal{Y})} E(\mu,\nu), \\ \mathcal{L}_2(\mu,\nu) &:= \max_{\nu\in\mathcal{P}(\mathcal{Y})} E(\mu,\nu) - E(\mu,\nu), \\ \mathcal{L}_3(\nu) &:= \max_{\nu\in\mathcal{P}(\mathcal{Y})}\min_{\mu\in\mathcal{P}(\mathcal{X})} E(\mu,\nu) - \min_{\mu\in\mathcal{P}(\mathcal{X})} E(\mu,\nu), \\ \mathcal{L}_4(\mu,\nu) &:= E(\mu,\nu) - \min_{\mu\in\mathcal{P}(\mathcal{X})} E(\mu,\nu). \end{aligned} \tag{15}$$

Note that by definition the functionals $\mathcal{L}_i, i=1,\cdots,4$ defined above are non-negative. For a fixed constant $\gamma>0$, we also define the Lyapunov functions

$$\begin{aligned} \mathcal{L}(\mu,\nu) &= \mathcal{L}_1(\mu) + \gamma\mathcal{L}_2(\mu,\nu), \\ \widetilde{\mathcal{L}}(\mu,\nu) &= \mathcal{L}_3(\mu) + \gamma\mathcal{L}_4(\mu,\nu). \end{aligned} \tag{16}$$

Let κ be the *effective* condition number defined by $\kappa = \frac{K_{xy}}{\tau\lambda_{LS}}$. Our first main theorem characterizes the exponential convergence of (6) in terms of the Lyapunov functions $\mathcal{L}$ and $\widetilde{\mathcal{L}}$.

Theorem 2.4. *Let Assumption 2.1 and Assumption 2.2 hold. For a fixed $\gamma<1$, let (μ_t,ν_t) be the solution to the GDA dynamics (6) with an initial condition (μ_0,ν_0) satisfying $\mathcal{L}(\mu_0,\nu_0)<\infty$ and $\widetilde{\mathcal{L}}(\mu_0,\nu_0)<\infty$, and with a finite time-scale ratio $\eta>0$.*

*(i) **Fast ascent regime:** set $\eta = \frac{2\lambda_{LS}\kappa^2\mathrm{Diam}(\mathcal{Y})^2}{\gamma}$. Then for all $t>0$,*

$$\mathcal{L}(\mu_t,\nu_t) \le e^{-\alpha_1 t}\mathcal{L}(\mu_0,\nu_0)$$

with

$$\alpha_1 = \tau\lambda_{LS}\Big(\frac{1-\gamma}{2} \wedge \frac{\kappa^2\mathrm{Diam}(\mathcal{Y})^2(1+3\gamma)}{\gamma}\Big);$$

*(ii) **Fast descent regime:** set $\eta = \frac{\gamma}{2\lambda_{LS}\kappa^2\mathrm{Diam}(\mathcal{X})^2(1+3\gamma)}$. Then for all $t>0$,*

$$\widetilde{\mathcal{L}}(\mu_t,\nu_t) \le e^{-\alpha_2 t}\widetilde{\mathcal{L}}(\mu_0,\nu_0)$$

with

$$\alpha_2 = \frac{\tau\lambda_{LS}}{2}\Big(1\wedge\eta(1-\gamma)\Big).$$

Proof. See Appendix B.2. □

Remark 2.5. Theorem 2.4 states that the two-scale GDA dynamics (6) with a suitable finite scale ratio converges exponentially to the equilibrium. This improves substantially the earlier result by (Ma & Ying, 2021) on GDA in the quasistatic setting where the scale ratio $\eta=0$ or $\eta=\infty$. We emphasize that we chose the specific scale ratios merely for the purpose of simplifying the expression of the convergence rate α. In fact, by tracking the proof of Theorem 2.4, one can obtain that the convergence rate $\alpha_1 = C_1\tau\lambda_{LS}$ if $\eta > c_1\lambda_{LS}\kappa^2\mathrm{Diam}(\mathcal{Y})^2$ and the rate $\alpha_2 = C_2\tau\lambda_{LS}^2$ if $\eta < \frac{c_2\gamma}{\lambda_{LS}\kappa^2\mathrm{Diam}(\mathcal{Y})^2(1+3\gamma)}$, where constants $C_i, c_i > 0, i=1,2$ are independent of quantities of interest. Moreover, the dependence of the convergence rates on the diameters of the state spaces via $K_{xy}\mathrm{Diam}(\mathcal{Y})$ and $K_{xy}\mathrm{Diam}(\mathcal{X})$ can be avoided, and especially the latter two quantities can be replaced by $\|\nabla K\|_\infty$; see discussions around (37).

Observe also that in the low temperature regime ($\tau\ll 1$), the log-Sobolev constant $\lambda_{LS} = \mathcal{O}(\exp(-\tau^{-1}\|K\|_\infty))$. Hence Theorem 2.4 requires $\eta = \Omega((\tau^{-1}K_{xy})^2/\lambda_{LS})$ (and

$\eta = o(\lambda_{LS}\tau^2/K_{xy}^2))$ to guarantee an exponential convergence rate $\alpha = \mathcal{O}(\tau\lambda_{LS})$ (and $\alpha = \mathcal{O}(\lambda_{LS}^2\tau^3)$) in the fast ascent (descent) regime. Notice that the quadratic dependence of the convergence rate α on λ_{LS} in the fast descent regime, in contrast to the linear dependence on λ_{LS} in the fast ascent regime, is due to the fact that the ascent dynamics itself is running in the slow time-scale $\eta = o(\lambda_{LS}\tau^2/K_{xy}^2)$; one would recover the same rate of convergence as in the fast ascent regime if one used the time scales $1/\eta$ for the descent dynamics and 1 the ascent dynamics.

Remark 2.6. Our construction of Lyapunov functions (16) is strongly motivated by the recent study (Yang et al., 2020; Doan, 2022) of two-scale GDA for minmax optimization on Euclidean spaces. In the finite dimensional setting, a two-sided PL condition is sufficient to guarantee the exponential convergence of two-scale GDA in both continuous-time (Doan, 2022) and discrete-time (Yang et al., 2020; 2022) cases. In our infinite dimensional setting of the Mean-Field GDA, the uniform log-Sobolev inequality plays the same role as the two-sided PL condition, which however, needs to be combined with the entropy sandwich inequalities in Lemma A.1 to obtain the exponential dissipation of Lyapunov functions.

The exponential convergence of (6) in the Lyapunov functions further implies exponential convergence of (6) with respect to the relative entropy as shown in the next theorem.

Theorem 2.7. *Let the assumptions of Theorem 2.4 hold. Let the rates $\alpha_i, i = 1, 2$ be defined as in Theorem 2.4.*

(i) ***Fast ascent regime:*** *set $\eta = \frac{2\lambda_{LS}\kappa^2 \mathrm{Diam}(\mathcal{Y})^2}{\gamma}$. Then for all $t > 0$,*

$$\begin{aligned} \tau\mathcal{H}(\mu_t|\mu^*) &\leq \frac{e^{-\alpha_1 t}}{\gamma}\mathcal{L}(\mu_0,\nu_0), \\ \tau\mathcal{H}(\nu_t|\nu^*) &\leq \Big(\frac{2}{\gamma} + \frac{4\|K\|_\infty^2}{\tau^2}\Big)e^{-\alpha_1 t}\mathcal{L}(\mu_0,\nu_0). \end{aligned} \tag{17}$$

(ii) ***Fast descent regime:*** *set $\eta = \frac{2\lambda_{LS}\kappa^2 \mathrm{Diam}(\mathcal{Y})^2}{\gamma}$. Then for all $t > 0$,*

$$\begin{aligned} \tau\mathcal{H}(\mu_t|\mu^*) &\leq \frac{e^{-\alpha_2 t}}{\gamma}\mathcal{L}(\mu_0,\nu_0), \\ \tau\mathcal{H}(\nu_t|\nu^*) &\leq \Big(\frac{2}{\gamma} + \frac{4\|K\|_\infty^2}{\tau^2}\Big)e^{-\alpha_2 t}\mathcal{L}(\mu_0,\nu_0). \end{aligned} \tag{18}$$

Proof. See Appendix B.3. □

2.3. Convergence of annealed Mean-Field GDA

In this section, we proceed to presenting the convergence of the "annealed" Mean-Field GDA dynamics

$$\begin{aligned} \partial_t\mu_t &= \nabla\cdot\Big(\mu_t\int_{\mathcal{Y}}\nabla_x K(x,y)d\nu_t(y)\Big) + \tau_t\Delta\mu_t, \\ \partial_t\nu_t &= \eta_t\Big(-\nabla\cdot\Big(\nu_t\int_{\mathcal{X}}\nabla_y K(x,y)d\mu_t(x)\Big) + \tau_t\Delta\nu_t\Big), \end{aligned} \tag{19}$$

where $\tau_t > 0$ is now a time-dependent temperature which shrinks in time. Given any initial condition $(\mu_0,\nu_0) \in \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})$, the existence and uniqueness of the global solution to (19) follow directly from the classical well-posedness of the theory of nonlinear Mckean-Vlasov-type PDEs (Funaki, 1984; Sznitman, 1991; Wang, 2018; Huang et al., 2021). Our goal is to show that by carefully choosing the cooling schedule τ_t and the time-scale ratio η_t the solution (μ_t,ν_t) to the annealed dynamics (19) converges to the MNE of E_0.

Let (μ_τ^*,ν_τ^*) be the solution of (12) corresponding to temperature τ. Recall the Nikaidò-Isoda defined by (7).

Theorem 2.8. *Let Assumption 2.2 and Assumption 2.1 hold. Let (μ_t,ν_t) be the solution to (19) with an initial condition $(\mu_0,\nu_0) \in \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})$. Assume that the log-Sobolev constant $\lambda_{LS} = \lambda_{LS}(\tau) \geq C_0 e^{-\xi^*/\tau}$ for some $\xi^*, C_0 > 0$.*

(i) Fast ascent regime: *Assume further that τ_t is smooth, decreasing in t and for some $\xi > \xi^*$, $\tau_t = \xi/\log t$ for large values of t. Set $\eta_t = \frac{M}{(\log t)^2}t^{\xi^*/\xi}$ for some large $M > 0$. Then for every $0 < \epsilon < 1 - \xi^*/\xi$, there exists $C, C' > 0$ such that for t sufficiently large,*

$$\mathcal{H}(\mu_t|\mu_{\tau_t}^*) \leq Ct^{-(1-\xi^*/\xi-\epsilon)}, \mathcal{H}(\nu_t|\nu_{\tau_t}^*) \leq Ct^{-(1-\xi^*/\xi-\epsilon)}, \tag{20}$$

and that

$$0 \leq NI(\mu_t,\nu_t) \leq \frac{C'\log\log t}{\log t}. \tag{21}$$

(ii) Fast descent regime: *Assume further that τ_t is smooth, decreasing in t and for some $\xi > 2\xi^*$, $\tau_t = \xi/\log t$ for large values of t. Set $\eta_t = \frac{\log t}{Mt}$ for some large M. Then for every $0 < \epsilon < 1 - 2\xi^*/\xi$, there exists $C, C' > 0$ such that for t sufficiently large,*

$$\mathcal{H}(\mu_t|\mu_{\tau_t}^*) \leq Ct^{-(1-2\xi^*/\xi-\epsilon)}, \mathcal{H}(\nu_t|\nu_{\tau_t}^*) \leq Ct^{-(1-2\xi^*/\xi-\epsilon)}, \tag{22}$$

and that

$$0 \leq NI(\mu_t,\nu_t) \leq \frac{C'\log\log t}{\log t}. \tag{23}$$

Proof. See Appendix C. □

3. Applications

In this section, we discuss briefly applications of our theoretical results in training of GANs and adversarial learning of PDEs.

3.1. Training of GANs

Let μ_m be the empirical measure associated to the i.i.d. samples $\{x_i\}_{i=1}^m \in \mathcal{X}$ from a target measure $\mu \in \mathcal{P}(\mathcal{X})$. Let $D_\mathcal{F}$ be an IPM on the space of probability measures $\mathcal{P}(\mathcal{X})$ parameterized by a set $\mathcal{F}$ of discriminators, i.e.

$$D_\mathcal{F}(\mu,\nu) := \sup_{f\in\mathcal{F}} \int f d\mu - \int f d\nu.$$

IPM-based GANs learn an optimal μ that minimizes $D_\mathcal{F}(\mu,\mu_m)$ over $\mathcal{P}(\mathcal{X})$:

$$\inf_{\mu\in\mathcal{P}(\mathcal{X})} D_\mathcal{F}(\mu,\mu_m) = \inf_{\mu\in\mathcal{P}(\mathcal{X})} \sup_{f\in\mathcal{F}} \int f d\mu - \int f d\mu_m. \tag{24}$$

Consider the witness function class $\mathcal{F}$ given by the unit ball of Barron space $\mathcal{B}$ which consists of functions admitting the representation

$$f(x) = \int_\mathcal{Y} a\sigma(b\cdot x + c) d\nu(y),$$

where $y = (a,b,c) \in \mathcal{Y}$ and ν is a probability measure on the parameter space $\mathcal{Y}$. Observe that Barron functions arise as natural infinite-width limit of two-layer neural networks with a dimension-free rate (Bach, 2017; Ma et al., 2022; Barron, 1993). When the activation function satisfies that $\sup_x |a\sigma(b\cdot x + c)| \le \phi(y)$ for some nonnegative function ϕ and for all $y \in \mathcal{Y}$, the Barron norm $\|f\|_\mathcal{B}$ is defined by

$$\begin{aligned}\|f\|_\mathcal{B} := \inf_\nu \Big\{ \int_\mathcal{Y} \phi(y)\nu(dy)\Big| \\ f(x) = \int_\mathcal{Y} a\sigma(b\cdot x + c) d\nu(y)\Big\}.\end{aligned} \tag{25}$$

Setting $\mathcal{F} = \{f \in \mathcal{B} \mid \|f\|_\mathcal{B} \le 1\}$ in (24) leads to

$$\begin{aligned}&\inf_{\mu\in\mathcal{P}(\mathcal{X})} \sup_{\nu\in\mathcal{P}(\mathcal{Y})} \int_\mathcal{X}\int_\mathcal{Y} K(x,y)\mu(dx)\nu(dy), \text{ where} \\ &K(x,y) = \Sigma(x,y) - \int_\mathcal{X} \Sigma(x,y)\mu_m(dx) - \phi(y).\end{aligned} \tag{26}$$

Here we adopted the short-notation $\Sigma(x,y) = a\sigma(b\cdot x + c)$ with $y = (a,b,c) \in \mathcal{Y}$ in the above. Assume that the activation function $\sigma \in C^2(\mathbb{R})$, and the input space $\mathcal{X}$ and the parameter space satisfy the Assumption 2.1. Then it is straightforward to see that $K \in C^2(\mathcal{X}\times\mathcal{Y})$ and there exists $C_\sigma < \infty$ such that for any multi-indices $\mathbf{i}$ and $\mathbf{j}$ with $0 \le |\mathbf{i}| + |\mathbf{j}| \le 2$,

$$\|\nabla_x^\mathbf{i}\nabla_y^\mathbf{j} K(x,y)\|_\infty \le 2\|\nabla_x^\mathbf{i}\nabla_y^\mathbf{j}\Sigma\|_\infty + \|\nabla_y^\mathbf{j}\phi\| \le C_\sigma.$$

Therefore the convergence results in Theorem 2.4-2.7 for the Mean-Field GDA hold for the entropy-regularization of the GAN objective defined in (26). Moreover, Theorem 2.8 implies that the annealed GDA dynamics finds the MNE of the unregularized GAN objective.

3.2. Adversarial learning of PDEs

We provide another usage of our results in adversarial learning of PDEs. To demonstrate the idea, we focus on a simple linear elliptic PDE on a bounded Lipschitz domain $\mathcal{Z} \subset \mathbb{R}^d$ equipped with the Neumann boundary condition

$$\begin{aligned}-\Delta u(z) + Vu(z) = f(z), z \in \mathcal{Z}, \\ \partial_\nu u(z) = 0, z \in \partial\mathcal{Z}.\end{aligned}$$

Assume that $0 < V_{\min} \le V \le V_{\max} < \infty$ and $f \in (H^1(\mathcal{Z}))^*$. The weak solution $u \in H^1(\mathcal{Z})$ satisfies that

$$\int_\mathcal{Z} \nabla u\cdot\nabla v + Vuvdz = \int_\mathcal{Z} fvdz, \forall v \in H^1(\mathcal{Z}). \tag{27}$$

We seek an approximate solution to (27) in the framework of Petrov-Galerkin (Petrov, 1940; Mitchell & Griffiths, 1980) where we choose the spaces of trial functions and test functions as two different Barron functions. More precisely, consider a trial function $u(z) \in \mathcal{U} := \{u \in \mathcal{B}_1 \mid \|u\|_{\mathcal{B}_1} \le 1\}$ and a test function $v \in \mathcal{V} := \{v \in \mathcal{B}_2 \mid \|v\|_{\mathcal{B}_2} \le 1\}$, where $\mathcal{B}_i, i = 1,2$ are Barron spaces defined in Section 3.1 with activation function σ_i and Barron norm $\|\cdot\|_{\mathcal{B}_i}$ defined in (25) with ϕ replaced by nonnegative weight functions ϕ_i. We look for a solution $u \in \mathcal{U}$ parameterized by some probability measure $\mu \in \mathcal{P}(\mathcal{X})$,

$$u(z) = \int_\mathcal{X} a_1\sigma_1(b_1\cdot z + c_1)\mu(dx)$$

with $x = (a_1,b_1,c_1) \in \mathcal{X}$ satisfying equation (27) for any $v \in \mathcal{V}$ with $\|v\|_{\mathcal{B}_2} \le 1$ parameterized by $\nu \in \mathcal{P}(\mathcal{Y})$ such that

$$v(z) = \int_\mathcal{Y} a_2\sigma_2(b_2\cdot z + c_2)\nu(dy).$$

Notice that the state spaces $\mathcal{X}$ and $\mathcal{Y}$ denote the parameter spaces associated to the two Barron functions u and v respectively. Putting these into the weak formulation (27) leads to

$$\inf_{\mu\in\mathcal{P}(\mathcal{X})} \sup_{\nu\in\mathcal{P}(\mathcal{Y})} \int_\mathcal{X}\int_\mathcal{Y} K(x,y)\mu(dx)\nu(dy),$$

where the potential $K(x,y)$ is given for $x = (a_1,b_1,c_1), y = (a_2,b_2,c_2)$ by

$$\begin{aligned}K(x,y) = \int_\mathcal{Z} \Big(&a_1a_2b_1\cdot b_2\sigma_1'(b_1\cdot z + c_1)\sigma_2'(b_2\cdot z + c_2) \\ &+ V(z)a_1a_2\sigma_1(b_1\cdot z + c_1)\sigma_2(b_2\cdot z + c_2) \\ &- f(z)a_2\sigma_2(b_2\cdot z + c_2)\Big)dz - \phi_1(x) + \phi_2(y).\end{aligned}$$

Assume that the activation functions $\sigma_i \in C^2(\mathbb{R})$ and the parameter spaces $\mathcal{X}$ and $\mathcal{Y}$ satisfy Assumption 2.1. Assume also that $\phi_1 \in C^2(\mathcal{X})$ and $\phi_2 \in C^2(\mathcal{Y})$. Then it is easy to verify that $K \in C^2(\mathcal{X} \times \mathcal{Y})$ and that $\|K\|_{C^2} \leq C$ for some constant $C > 0$ depending on $\sigma_i, \phi_i, \mathcal{X}, \mathcal{Y}, V$ and f. Hence the convergence results on Mean-Field GDA and its annealed version established in Section 2 apply to this problem.

4. Conclusion and future work

In this paper, we proved the global exponential convergence of two-scale Mean-Field GDA dynamics for finding the unique MNE of the entropy-regularized game objective on the space of probability measures. The obtained convergence rate depends on the uniform log-Sobolev constant of proximal Gibbs distributions. We also proved that the convergence of the annealed GDA dynamics to the MNE of the unregularized game objective with respect to the Nikaidò-Isoda error. The key ingredient of our proofs are new Lyapunov functions which are used to capture the dissipation of the two-scale GDA dynamics in different scaling regimes.

We would like to mention several open problems and future research directions. First, although we have proved the global convergence of the Mean-Field GDA in the fast ascent/descent regime (with a finite but large/small time-scale ratio), it remains an open question to show the convergence or nonconvergence of the Mean-Field GDA in the intermediate time-scale regime, including particularly the case where no time-scale separation occurs (i.e. $\eta = 1$ in (6)). Also, currently our results only hold on bounded state spaces and the convergence rate depends on the uniform bounds of the potential function K on the state spaces. It remains an important question to extend the results to minmax optimization on unbounded state spaces. In practice, the Mean-Field GDA needs to be implemented by a certain interacting particle algorithm such as (5) with a large number of particles. Existing results on the mean field limit of (5) (e.g. Theorem 3 of (Domingo-Enrich et al., 2020)) only holds in finite time interval and the error bound can potentially grow exponentially in time. To obtain a quantitative error analysis of the GDA for minmax optimization, it is an interesting open question to derive a uniform-in-time quantitative error bound between the particle system and the mean field dynamics. Finally, we anticipate our results can be exploited to prove theoretical convergence guarantees for a variety of minmax learning problems, including especially the training of GANs (Goodfellow et al., 2020), adversarial learning problems (Madry et al., 2018), dual training of energy based models (Dai et al., 2019; Domingo-Enrich et al., 2021) and weak adversarial networks for PDEs (Zang et al., 2020).

Acknowledgements

The author thanks Lexing Ying for suggesting the question and thanks Jianfeng Lu, Yiping Lu and Chao Ma for helpful discussions at the early stage of the problem setup. The author also thank Fei Cao and Lénaïc Chizat for the valuable feedback on an early version of the paper. This work is supported by the National Science Foundation through the award DMS-2107934.

References

Abernethy, J., Lai, K. A., and Wibisono, A. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*, 2019.

Araújo, D., Oliveira, R. I., and Yukimura, D. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Brenier, Y. and Vorotnikov, D. On optimal transport of matrix-valued measures. *SIAM Journal on Mathematical Analysis*, 52(3):2849–2873, 2020.

Busoniu, L., Babuska, R., and De Schutter, B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Chizat, L. Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 2022.

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

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

Dai, B., Liu, Z., Dai, H., He, N., Gretton, A., Song, L., and Schuurmans, D. Exponential family estimation via adversarial dynamics embedding. *Advances in Neural Information Processing Systems*, 32, 2019.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in neural information processing systems*, 31, 2018.

Daskalakis, C. and Panageas, I. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *10th Innovations in Theoretical Computer Science*, 2019.

Doan, T. Convergence rates of two-time-scale gradient descent-ascent dynamics for solving nonconvex min-max problems. In *Learning for Dynamics and Control Conference*, pp. 192–206. PMLR, 2022.

Domingo-Enrich, C. and Bruna, J. Simultaneous transport evolution for minimax equilibria on measures. *arXiv preprint arXiv:2202.06460*, 2022.

Domingo-Enrich, C., Jelassi, S., Mensch, A., Rotskoff, G., and Bruna, J. A mean-field analysis of two-player zero-sum games. *Advances in neural information processing systems*, 33:20215–20226, 2020.

Domingo-Enrich, C., Bietti, A., Gabrié, M., Bruna, J., and Vanden-Eijnden, E. Dual training of energy-based models with overparametrized shallow neural networks. *arXiv preprint arXiv:2107.05134*, 2021.

Eberle, A., Guillin, A., and Zimmer, R. Quantitative harris-type theorems for diffusions and mckean–vlasov processes. *Transactions of the American Mathematical Society*, 371(10):7135–7173, 2019.

Funaki, T. A certain class of diffusion processes associated with nonlinear parabolic equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67 (3):331–348, 1984.

Glicksberg, I. L. 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.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Holley, R. A., Kusuoka, S., and Stroock, D. W. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *Journal of functional analysis*, 83 (2):333–347, 1989.

Hsieh, Y.-P., Liu, C., and Cevher, V. Finding mixed nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pp. 2810–2819. PMLR, 2019.

Hu, K., Ren, Z., Šiška, D., and Szpruch, Ł. Mean-field langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 57, pp. 2043–2065. Institut Henri Poincaré, 2021.

Huang, X., Ren, P., and Wang, F.-Y. Distribution dependent stochastic differential equations. *Frontiers of Mathematics in China*, 16:257–301, 2021.

Jin, C., Netrapalli, P., and Jordan, M. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pp. 4880–4889. PMLR, 2020.

Kondratyev, S. and Vorotnikov, D. Spherical hellinger–kantorovich gradient flows. *SIAM Journal on Mathematical Analysis*, 51(3):2053–2084, 2019.

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

Laschos, V. and Mielke, A. Geometric properties of cones with applications on the hellinger–kantorovich space, and a new distance on the space of probability measures. *Journal of Functional Analysis*, 276(11):3529–3576, 2019.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Lu, Y., Lu, J., and Nolen, J. Accelerating langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*, 2019.

Lu, Y., Ma, C., Lu, Y., Lu, J., and Ying, L. A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pp. 6426–6436. PMLR, 2020.

Lu, Y., Slepčev, D., and Wang, L. Birth-death dynamics for sampling: Global convergence, approximations and their asymptotics. *arXiv preprint arXiv:2211.00450*, 2022.

Ma, C. and Ying, L. Provably convergent quasistatic dynamics for mean-field two-player zero-sum games. In *International Conference on Learning Representations*, 2021.

Ma, C., Wu, L., et al. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

Mitchell, A. R. and Griffiths, D. F. The finite difference method in partial differential equations. *A Wiley-Interscience Publication*, 1980.

Nash, J. Non-cooperative games. *Annals of Mathematics*, 54:286–295, 1951.

Nguyen, P.-M. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*, 2019.

Nikaido, H. On von neumann's minimax theorem. *Pacific J. Math*, 4(1):65–72, 1954.

Nikaido, H. and Isoda, K. Note on non-cooperative convex games. *Pacific Journal of Mathematics*, 5(S1):807–815, 1955.

Nitanda, A., Wu, D., and Suzuki, T. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 9741–9757. PMLR, 2022.

Petrov, G. I. Application of the method of galerkin to a problem involving the stationary flow of a viscous fluid. *Prikl. Matem. Mekh.*, 4(3):3–12, 1940.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703. PMLR, 2017.

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

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

Sion, M. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

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

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

Sznitman, A.-S. Topics in propagation of chaos. In *Ecole d'Eté de probabilits de Saint-Flour XIX-1989*, pp. 165–251. Springer, 1991.

Tang, W. and Zhou, X. Y. Simulated annealing from continuum to discretization: a convergence analysis via the eyring–kramers law. *arXiv preprint arXiv:2102.02339*, 2021.

Von. Neumann, J. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Wang, F.-Y. Distribution dependent sdes for landau type equations. *Stochastic Processes and their Applications*, 128(2):595–621, 2018.

Wang, G. and Chizat, L. An exponentially converging particle method for the mixed nash equilibrium of continuous games. *arXiv preprint arXiv:2211.01280*, 2022.

Wojtowytsch, S. et al. On the banach spaces associated with multi-layer relu networks: Function representation, approximation theory and gradient descent dynamics. *arXiv preprint arXiv:2007.15623*, 2020.

Yang, J., Kiyavash, N., and He, N. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165, 2020.

Yang, J., Orvieto, A., Lucchi, A., and He, N. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pp. 5485–5517. PMLR, 2022.

Zang, Y., Bao, G., Ye, X., and Zhou, H. Weak adversarial networks for high-dimensional partial differential equations. *Journal of Computational Physics*, 411:109409, 2020.

A. Proofs of preliminary lemmas

We first state a lemma below summarizing some important properties on the functionals $\mathcal{L}_i, i = 1, \cdots, 4.$ defined in (15).

Lemma A.1. *For any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the following hold*

$$\mathcal{L}_2(\mu, \nu) = \tau \mathcal{H}(\nu | \mathcal{K}^+(\mu)), \tag{28}$$

$$\mathcal{L}_4(\mu, \nu) = \tau \mathcal{H}(\mu | \mathcal{K}^-(\nu)), \tag{29}$$

$$\tau \mathcal{H}(\mu | \mu^*) \le \mathcal{L}_1(\mu) \le \tau \mathcal{H}(\mu | \mathcal{K}^-(\mathcal{K}^+(\mu))), \tag{30}$$

$$\tau \mathcal{H}(\nu | \nu^*) \le \mathcal{L}_3(\nu) \le \tau \mathcal{H}(\nu | \mathcal{K}^+(\mathcal{K}^-(\nu))). \tag{31}$$

Remark A.2. The sandwich inequalities (30) and (31) play an essential role in controlling terms in the entropy production of the Mean-Field GDA dynamics via the Lyapunov functions in order to close the Grönwall argument to obtain the dissipation of the latter along the dynamics. A similar sandwich inequality appeared in Lemma 3.4 of (Chizat, 2022; Nitanda et al., 2022) in the proof of convergence for the Mean-Field Langevin dynamics.

Proof of Lemma A.1. First we have from the definition of E and (11) that

$$\begin{aligned}
\mathcal{L}_2(\mu, \nu) &= \max_{\nu \in \mathcal{P}(\mathcal{Y})} E(\mu, \nu) - E(\mu, \nu) \\
&= \tau \log Z^+(\mu) - \int_{\mathcal{X}} \int_{\mathcal{Y}} K(x, y) \mu(dx) \nu(dy) - \tau \mathcal{H}(\nu) \\
&= \tau \mathcal{H}(\nu | \mathcal{K}^+(\mu))
\end{aligned}$$

which proves (28). To see (29), notice again from (11) that

$$\begin{aligned}
\mathcal{L}_4(\mu, \nu) &= E(\mu, \nu) - \min_{\mu \in \mathcal{P}(\mathcal{X})} E(\mu, \nu) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} K(x, y) \mu(dx) \nu(dy) - \tau \mathcal{H}(\mu) + \tau \log Z^-(\nu) \\
&= \tau \mathcal{H}(\mu | \mathcal{K}^-(\nu)).
\end{aligned}$$

Next, we prove (30). In fact, replacing ν by $\mathcal{K}^-(\mu)$ in the last display leads to

$$\begin{aligned}
\tau \mathcal{H}(\mu | \mathcal{K}^-(\mathcal{K}^+(\mu))) &= \mathcal{L}_4(\mu, \mathcal{K}^+(\mu)) \\
&= E(\mu, \mathcal{K}^+(\mu)) - \min_{\mu \in \mathcal{P}(\mathcal{X})} E(\mu, \mathcal{K}^+(\mu)) \\
&= \max_{\nu \in \mathcal{P}(\mathcal{Y})} E(\mu, \nu) - \min_{\mu \in \mathcal{P}(\mathcal{X})} E(\mu, \mathcal{K}^+(\mu)) \\
&\ge \max_{\nu \in \mathcal{P}(\mathcal{Y})} E(\mu, \nu) - \min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{\nu \in \mathcal{P}(\mathcal{Y})} E(\mu, \nu) \\
&= \mathcal{L}_1(\mu),
\end{aligned}$$

which proves the lower bound part of (30). To prove the upper bound part, we first compute the first-variation of the log partition function $\log Z^+(\mu)$ as

$$\frac{\partial}{\partial \mu} \log Z^+(\mu) = \tau^{-1} \int_{\mathcal{Y}} K(x, y) d\mathcal{K}^+(\mu)(y) = -\log(\mathcal{K}^-(\mathcal{K}^+(\mu))) - \log Z^-(\mathcal{K}^+(\mu)). \tag{32}$$

Then an application of the convexity of $\mu \to \log Z^+(\mu)$ shown in Lemma A.3 yields

$$\begin{aligned}
\log Z^+(\mu) - \log Z^+(\mu^*) &\ge \int_{\mathcal{X}} \frac{\partial \log Z^+(\mu)}{\partial \mu}\Big|_{\mu = \mu^*} (d\mu - d\mu^*) \\
&= -\int_{\mathcal{X}} \log(\mathcal{K}^-(\mathcal{K}^+(\mu^*)))(d\mu - d\mu^*) \\
&= -\int_{\mathcal{X}} \log(\mu^*)(d\mu - d\mu^*).
\end{aligned}$$

Consequently, one obtains that

$$\begin{aligned}\mathcal{L}_1(\mu) &= \mathcal{E}_1(\mu) - \mathcal{E}_1(\mu^*) \\ &= -\tau(\mathcal{H}(\mu) - \mathcal{H}(\mu^*)) + \tau(\log Z^+(\mu) - \log Z^+(\mu^*)) \\ &\geq \tau \int \log \mu d\mu - \int \log \mu^* d\mu^* - \int_{\mathcal{X}} \log(\mu^*)(d\mu - d\mu^*) \\ &= \tau \mathcal{H}(\mu|\mu^*).\end{aligned}$$

Finally, the inequality (31) follows from the same reasoning above by exploiting the convexity of $\nu \to \log Z^-(\nu)$.

□

The following lemma establishes the convexity of the log partition functions $\log Z^+(\mu)$ and $\log Z^-(\nu)$. It was proved in Proposition 3 of (Ma & Ying, 2021), but for completeness we include its proof here.

Lemma A.3. *Both the functional $\mu \to \log Z^+(\mu)$ and the functional $\nu \to \log Z^-(\nu)$ are convex.*

Proof. We only present the proof for the convexity of the map $\mu \to \log Z^+(\mu)$ since the other case can be proved in the same manner. For any $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$ and $\alpha \in (0,1)$,

$$\begin{aligned}&Z^+(\alpha\mu_1 + (1-\alpha)\mu_2) \\ &= \log\Big(\int_{\mathcal{Y}} \exp\Big(\int_{\mathcal{X}} \tau^{-1} K(x,y)(\alpha\mu_1(dx) + (1-\alpha)\mu_2(dx))\Big)dy\Big) \\ &= \log\Big(\int_{\mathcal{Y}} \exp\Big(\alpha\int_{\mathcal{X}} \tau^{-1} K(x,y)\mu_1(dx)\Big) \cdot \exp\Big((1-\alpha)\int_{\mathcal{X}} \tau^{-1} K(x,y)\mu_2(dx)\Big)dy\Big) \\ &\leq \log\Big(\Big(\int_{\mathcal{Y}} \exp\Big(\int_{\mathcal{X}} \tau^{-1} K(x,y)\mu_1(dx)\Big)dy\Big)^{\alpha} \cdot \Big(\int_{\mathcal{Y}} \exp\Big(\int_{\mathcal{X}} \tau^{-1} K(x,y)\mu_2(dx)\Big)dy\Big)^{1-\alpha}\Big) \\ &= \alpha Z^+(\mu_1) + (1-\alpha)Z^+(\mu_2).\end{aligned}$$

□

We restate Theorem 5 of (Domingo-Enrich et al., 2020) in the following lemma which is used to control $\mathrm{NI}(\mu_\tau^*, \nu_\tau^*)$.

Lemma A.4. *(Theorem 5 of (Domingo-Enrich et al., 2020)) Let $\epsilon > 0, \delta := \epsilon/(2Lip(K))$ and let V_δ be a lower bound on the volume of a ball of radius of δ in $\mathcal{X}$ and $\mathcal{Y}$. Let (μ_τ^*, ν_τ^*) be the solution of the solution of (12). Then (μ_τ^*, ν_τ^*) is an ϵ-Nash equilibrium of E_0 if*

$$\tau \leq \frac{\epsilon}{4\log\left(\frac{2(1-V_\delta)}{V_\delta}(4\|K\|_\infty/\epsilon - 1)\right)}.$$

In particular, when $\mathcal{X}$ and $\mathcal{Y}$ are Riemannian manifolds or Euclidean tori of dimensions $d_\mathcal{X}$ and $d_\mathcal{Y}$ respectively so that $Vol(B_x^\delta) \geq C\delta^{d_\mathcal{X}}$ and $Vol(B_y^\delta) \geq C\delta^{d_\mathcal{Y}}$, the bounds above become

$$\tau \leq \frac{C\epsilon}{\log \epsilon^{-1}}$$

for some constant $C > 0$ depending only on $K, d_\mathcal{X}, d_\mathcal{Y}$. Alternatively, if τ is sufficiently small, then (μ_τ^*, ν_τ^*) is an ϵ-Nash equilibrium with

$$\epsilon = \beta\tau(\log(1/\tau)) \text{ for } \beta > d_\mathcal{X} \vee d_\mathcal{Y} + 1. \tag{33}$$

Proof of Lemma A.4. The proof of the lemma can be found in Appendix C.4 of (Domingo-Enrich et al., 2020). We would like to elaborate on the proof of (33). In fact, by tracking the proof of Theorem 5 in (Domingo-Enrich et al., 2020), one sees that (μ_τ^*, ν_τ^*) is an ϵ-Nash equilibrium provided that

$$e^{\frac{\varepsilon}{2\tau}}\Big(\frac{\mathrm{Vol}(B_x^\delta)}{1-\mathrm{Vol}(B_x^\delta)} \vee \frac{\mathrm{Vol}(B_y^\delta)}{1-\mathrm{Vol}(B_y^\delta)}\Big) \geq 2(4\|K\|_\infty/\epsilon - 1),$$

where we recall that $\delta = \epsilon/(2\mathrm{Lip}(K))$. Thanks to the lower bound on the volume of small balls in $\mathcal{X}$ and $\mathcal{Y}$, the inequality above holds if

$$e^{\frac{\epsilon}{2\tau}} \geq C_1 \epsilon^{-(d_{\mathcal{X}} \vee d_{\mathcal{Y}}+1)}$$

for some $C_1 > 0$ depending only on $K, d_{\mathcal{X}}, d_{\mathcal{Y}}$. Now with the choice (33) for ϵ with $\beta > d_{\mathcal{X}} \vee d_{\mathcal{Y}} + 1$, one has for τ sufficiently small that

$$e^{\frac{\varepsilon}{2\tau}} \geq \frac{1}{\tau^{\beta}} > \frac{C_1}{\beta^{d_{\mathcal{X}} \vee d_{\mathcal{Y}}+1}} \Big(\frac{1}{\tau \log(1/\tau)} \Big)^{d_{\mathcal{X}} \vee d_{\mathcal{Y}}+1} = C_1 \epsilon^{-(d_{\mathcal{X}} \vee d_{\mathcal{Y}}+1)}.$$

□

B. Proofs of convergence for Mean-Field GDA with a fixed temperature

B.1. Calculations of the time-derivatives of $\mathcal{L}_i$ along (6)

In the next two propositions we keep track of the time-derivatives of functionals $\mathcal{L}_i$ along the Mean-Field GDA dynamics (6). The next proposition concerns $\mathcal{L}_1(\mu_t)$ and $\mathcal{L}_2(\mu_t, \nu_t)$.

Proposition B.1. *Let (μ_t, ν_t) be the solution to the DGA dynamics (6). Then*

$$\frac{d}{dt}\mathcal{L}_1(\mu_t) \leq -\frac{\tau^2}{2}\mathcal{I}(\mu_t | \mathcal{K}^-(\mathcal{K}^+(\mu_t))) + K_{xy}^2 \mathrm{Diam}(\mathcal{Y})^2 \cdot \mathcal{H}(\nu_t | \mathcal{K}^+(\mu_t)) \tag{34}$$

and

$$\frac{d}{dt}\mathcal{L}_2(\mu_t, \nu_t) \leq -\tau^2 \eta \mathcal{I}(\nu_t | \mathcal{K}^+(\mu_t)) + \frac{\tau^2}{2}\mathcal{I}(\mu_t | \mathcal{K}^-(\mathcal{K}^+(\mu_t))) + 3K_{xy}^2 \mathrm{Diam}(\mathcal{Y})^2 \cdot \mathcal{H}(\nu_t | \mathcal{K}^+(\mu_t)). \tag{35}$$

Proof. First let us compute the functional gradient $\frac{\partial \mathcal{L}_1}{\partial u}(\mu_t)$. In fact, thanks to the fact that $\mathcal{L}_1(\mu) = \mathcal{E}_1(\mu) - \mathcal{E}_1(\mu^*)$ and (50), one has that

$$\begin{aligned}
\frac{\partial \mathcal{L}_1}{\partial u}\Big|_{\mu=\mu_t} &= -\tau \frac{\partial}{\partial u}\Big(\mathcal{H}(\mu) - \log Z^+(\mu) \Big)\Big|_{\mu=\mu_t} \\
&= \tau \Big(\log \mu + \tau^{-1} \int_{\mathcal{Y}} K(x,y) d\mathcal{K}^+(\mu)(y) \Big)\Big|_{\mu=\mu_t} \\
&= \tau \Big(\log \Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))} \Big) - \log Z^-(\mathcal{K}^+(\mu_t)) \Big),
\end{aligned}$$

where in the second identity we have used (32). Therefore it follows from (6) and the Cauchy-Schwarz inequality that

$$\begin{aligned}
\frac{d}{dt}\mathcal{L}_1(\mu_t) &= \int_{\mathcal{X}} \frac{\partial \mathcal{L}_1}{\partial u}(\mu_t) d(\partial_t \mu_t) \\
&= \tau^2 \int_{\mathcal{X}} \Big(\log \Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))} \Big) - \log Z^-(\mathcal{K}^-(\mathcal{K}^+(\mu_t))) \Big) \nabla_x \cdot \Big(\mu_t \nabla_x \log \Big(\frac{d\mu_t}{d\mathcal{K}^-(\nu_t)} \Big) \Big) dx \\
&= -\tau^2 \int_{\mathcal{X}} \nabla_x \log \Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))} \Big) \cdot \nabla_x \log \Big(\frac{d\mu_t}{d\mathcal{K}^-(\nu_t)} \Big) d\mu_t(x) \\
&= -\tau^2 \int_{\mathcal{X}} \Big| \nabla_x \log \Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))} \Big) \Big|^2 d\mu_t(x) \\
&\quad - \tau^2 \int_{\mathcal{X}} \nabla_x \log \Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)} \Big) \cdot \nabla_x \log \Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))} \Big) d\mu_t(x) \\
&\leq -\frac{\tau^2}{2}\mathcal{I}(\mu_t | \mathcal{K}^-(\mathcal{K}^+(\mu_t))) + \frac{\tau^2}{2} \int_{\mathcal{X}} \Big| \nabla_x \log \Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)} \Big) \Big|^2 d\mu_t(x).
\end{aligned}$$

Furthermore, using the fact that

$$\mathcal{K}^-(\mathcal{K}^+(\mu_t)) \propto \exp\Big(-\int_{\mathcal{Y}} \tau^{-1} K(x,y) d\mathcal{K}^+(\mu_t)(y) \Big), \quad \mathcal{K}^-(\nu_t) \propto \exp\Big(-\int_{\mathcal{Y}} \tau^{-1} K(x,y) d\nu_t(y) \Big),$$

one derives that

$$\frac{1}{2} \int_{\mathcal{X}} \Big| \nabla_x \log \Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)} \Big) \Big|^2 d\mu_t(x) = \frac{1}{2\tau^2} \int_{\mathcal{X}} \Big| \int_{\mathcal{Y}} \nabla_x K(x,y) (d\mathcal{K}^+(\mu_t)(y) - d\nu_t(y)) \Big|^2 d\mu_t(x). \tag{36}$$

Moreover, using the mean value theorem, one has for any $y_0 \in \mathcal{Y}$ that

$$\nabla_x K(x,y) - \nabla_x K(x,y_0) = \Big[\int_0^1 \nabla^2_{xy} K(x, y_0 + s(y-y_0)) s\, ds\Big](y-y_0).$$

Inserting the last identity into (36) leads to

$$\begin{aligned}\frac{1}{2}\int_{\mathcal{X}} \Big|\nabla_x \log\Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)}\Big)\Big|^2 d\mu_t(x) &\le \frac{K_{xy}^2 \mathrm{Diam}(\mathcal{Y})^2}{2\tau^2}\mathrm{TV}^2(\mathcal{K}^+(\mu_t),\nu_t)\\ &\le \frac{K_{xy}^2\mathrm{Diam}(\mathcal{Y})^2}{\tau^2}\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)),\end{aligned} \tag{37}$$

where we have used the Pinsker's inequality in the last inequality above. Combining the last two estimates proves (34).

Next we proceed with the proof of (35). Recall from (28) that

$$\mathcal{L}_2(\mu,\nu) = \tau\mathcal{H}(\nu|\mathcal{K}^+(\mu)) = \tau\int_{\mathcal{Y}} \log\Big(\frac{\nu}{\mathcal{K}^+(\mu)}\Big)d\nu.$$

Hence

$$\frac{d}{dt}\mathcal{L}_2(\mu_t,\nu_t) = \tau\Big(\int_{\mathcal{Y}} \log\Big(\frac{\nu_t}{\mathcal{K}^+(\mu_t)}\Big)d(\partial_t\nu_t) - \int_{\mathcal{Y}}\frac{d}{dt}\log(\mathcal{K}^+(\mu_t))d\nu_t\Big). \tag{38}$$

Using the second equation of (6) that ν_t solves, one sees that the first term on the right side above becomes

$$\tau\Big(\int_{\mathcal{Y}} \log\Big(\frac{\nu_t}{\mathcal{K}^+(\mu_t)}\Big)d(\partial_t\nu_t) = -\eta\tau^2\mathcal{I}(\nu_t|\mathcal{K}^+(\mu_t)). \tag{39}$$

To compute the second term on the right side of (38), observe that

$$\log\mathcal{K}^+(\mu_t) = \tau^{-1}\int_{\mathcal{X}} K(x,y)d\mu_t(x) - \log(Z^+(\mu_t)).$$

An a result of above and (32), one obtains that

$$\begin{aligned}
&-\tau\int_{\mathcal{Y}}\frac{d}{dt}\log(\mathcal{K}^+(\mu_t))d\nu_t\\
&= -\tau\int_{\mathcal{Y}}\int_{\mathcal{X}}\Big(\tau^{-1}K(x,y) - \frac{\partial\log(Z^+(\mu))}{\partial\mu}(\mu_t)\Big)d(\partial_t\mu_t(x))d\nu_t(y)\\
&= -\tau\int_{\mathcal{X}}\Big(\int_{\mathcal{Y}}\tau^{-1}K(x,y)d\nu_t(y) - \int_{\mathcal{Y}}\tau^{-1}K(x,y)d\mathcal{K}^+(\mu_t)(y)\Big)d(\partial_t\mu_t(x))\\
&= -\tau\int_{\mathcal{X}}\Big(\log(\mathcal{Z}^-(\mathcal{K}^+(\mu_t))\mathcal{K}^-(\mathcal{K}^+(\mu_t))) - \log(\mathcal{Z}^-(\nu_t)\mathcal{K}^-(\nu_t))\Big)d(\partial_t\mu_t(x))\\
&= \tau^2\int_{\mathcal{X}}\nabla_x\log\Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)}\Big)\cdot\nabla_x\log\Big(\frac{d\mu_t}{d\mathcal{K}^-(\nu_t)}\Big)d\mu_t(x)\\
&= \tau^2\Big(\int_{\mathcal{X}}\Big|\nabla_x\log\Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)}\Big)\Big|^2 d\mu_t(x)\\
&\qquad + \int_{\mathcal{X}}\nabla_x\log\Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)}\Big)\cdot\nabla_x\log\Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}\Big)d\mu_t(x)\Big)\\
&\le \frac{\tau^2}{2}\mathcal{I}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))) + \frac{3\tau^2}{2}\int_{\mathcal{X}}\Big|\nabla_x\log\Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)}\Big)\Big|^2 d\mu_t(x)\\
&\le \frac{\tau^2}{2}\mathcal{I}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))) + 3K_{xy}^2\mathrm{Diam}(\mathcal{Y})^2\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)),
\end{aligned}$$

where we have used (37) in the last inequality above. The estimate (35) then follows from above, (39) and (38). This completes the proof of the proposition. □

The proposition below bounds the time-derivative of $\mathcal{L}_2(\nu_t)$ and $\mathcal{L}_4(\mu_t, \nu_t)$.

Proposition B.2. *Let (μ_t, ν_t) be the solution to the DGA dynamics (6). Then*

$$\frac{d}{dt}\mathcal{L}_3(\nu_t) \le -\frac{\eta\tau^2}{2}\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) + \eta K_{xy}^2 \mathrm{Diam}(\mathcal{X})^2 \cdot \mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)) \tag{40}$$

and

$$\frac{d}{dt}\mathcal{L}_4(\mu_t, \nu_t) \le -\tau^2\mathcal{I}(\mu_t|\mathcal{K}^-(\nu_t)) + \frac{\eta\tau^2}{2}\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) + 3\eta K_{xy}^2 \mathrm{Diam}(\mathcal{X})^2 \cdot \mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)). \tag{41}$$

Proof. The proof follows essentially the same reasoning as the proof of Proposition B.1. In fact, it can be shown by a straightforward calculation that

$$\begin{aligned}\frac{d}{dt}\mathcal{L}_2(\nu_t) &= -\eta\tau^2 \int_{\mathcal{Y}} \nabla_y \log\Big(\frac{d\nu_t}{d\mathcal{K}^+(\mathcal{K}^-(\nu_t))}\Big) \cdot \nabla_y \log\Big(\frac{d\nu_t}{d\mathcal{K}^+(\mu_t)}\Big) d\nu_t(y) \\ &\le -\frac{\eta\tau^2}{2}\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) + \frac{\eta\tau^2}{2}\int_{\mathcal{Y}} \Big|\nabla_y \log\Big(\frac{d\mathcal{K}^+(\mathcal{K}^-(\nu_t))}{d\mathcal{K}^+(\mu_t)}\Big)\Big|^2 d\nu_t(y).\end{aligned}$$

Similar to (37), one can obtain that

$$\frac{\eta\tau^2}{2}\int_{\mathcal{Y}} \Big|\nabla_y \log\Big(\frac{d\mathcal{K}^+(\mathcal{K}^-(\nu_t))}{d\mathcal{K}^+(\mu_t)}\Big)\Big|^2 d\nu_t(y) \le \eta K_{xy}^2 \mathrm{Diam}(\mathcal{X})^2 \cdot \mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)).$$

Combining the last two inequalities yield (40).

As for time-derivative of $\mathcal{L}_4(\mu_t, \nu_t)$, one has that

$$\frac{d}{dt}\mathcal{L}_4(\mu_t, \nu_t) = -\tau^2\mathcal{I}(\mu_t|\mathcal{K}^-(\nu_t)) - \eta\tau^2 \int_{\mathcal{Y}} \nabla_y \log\Big(\frac{d\mathcal{K}^+(\mu_t)}{d\mathcal{K}^+(\mathcal{K}^-(\nu_t))}\Big) \cdot \nabla_y \log\Big(\frac{d\nu_t}{d\mathcal{K}^+(\mu_t)}\Big) d\nu_t(y).$$

The second term on the right side above is bounded by

$$\begin{aligned}&\eta\tau^2\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) + \frac{3\eta\tau^2}{2}\int_{\mathcal{Y}} \Big|\nabla_y \log\Big(\frac{d\mathcal{K}^+(\mathcal{K}^-(\nu_t))}{d\mathcal{K}^+(\mu_t)}\Big)\Big|^2 d\nu_t(y) \\ &\le \frac{\eta\tau^2}{2}\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) + 3\eta K_{xy}^2 \mathrm{Diam}(\mathcal{X})^2 \cdot \mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)).\end{aligned}$$

The estimate (41) follows directly from the last two inequalities. □

B.2. Proof of Theorem 2.4

With Proposition B.1 and Proposition B.2, we are ready to present the proof of Theorem 2.4.

(i) Fast ascent regime. Thanks to Proposition B.1 and identity (28), we have

$$\begin{aligned}\frac{d}{dt}\mathcal{L}(\mu_t, \nu_t) \le &-\eta\tau^2\gamma\mathcal{I}(\nu_t|\mathcal{K}^+(\mu_t)) - \frac{\tau^2}{2}(1-\gamma)\mathcal{I}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))) \\ &+ \frac{K_{xy}^2 \mathrm{Diam}(\mathcal{Y})^2(1+3\gamma)}{\tau}\mathcal{L}_2(\mu_t, \nu_t).\end{aligned}$$

Observe also from Lemma 2.3 and sandwich inequality (30) that

$$\begin{aligned}\tau\mathcal{I}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))) &\ge \tau\lambda_{LS}\mathcal{H}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))) \\ &\ge \lambda_{LS}\mathcal{L}_1(\mu_t).\end{aligned}$$

Combining the last two displays leads to

$$\begin{aligned}\frac{d}{dt}\mathcal{L}(\mu_t, \nu_t) &\le -\eta\tau^2\gamma\mathcal{I}(\nu_t|\mathcal{K}^+(\mu_t)) - \frac{\tau}{2}(1-\gamma)\lambda_{LS}\mathcal{L}_1(\mu_t) + \frac{K_{xy}^2 \mathrm{Diam}(\mathcal{Y})^2(1+3\gamma)}{\tau}\mathcal{L}_2(\mu_t, \nu_t) \\ &\le -\eta\tau^2\gamma\lambda_{LS}\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) - \frac{\tau}{2}(1-\gamma)\lambda_{LS}\mathcal{L}_1(\mu_t) + \frac{K_{xy}^2 \mathrm{Diam}(\mathcal{Y})^2(1+3\gamma)}{\tau}\mathcal{L}_2(\mu_t, \nu_t) \\ &= -\frac{\tau}{2}(1-\gamma)\lambda_{LS}\mathcal{L}_1(\mu_t) - \tau\Big(\eta\gamma\lambda_{LS} - \frac{K_{xy}^2 \mathrm{Diam}(\mathcal{Y})^2(1+3\gamma)}{\tau^2}\Big)\mathcal{L}_2(\mu_t, \nu_t).\end{aligned}$$

Now for any fixed $\gamma < 1$, we set

$$\eta = \frac{2K_{xy}^2 \mathrm{Diam}(\mathcal{Y})^2(1+3\gamma)}{\tau^2\gamma\lambda_{LS}} = \frac{2\lambda_{LS}\kappa^2\mathrm{Diam}(\mathcal{Y})^2}{\gamma}.$$

Then it follows from the last inequality that

$$\frac{d}{dt}\mathcal{L}(\mu_t,\nu_t) \le -\alpha\mathcal{L}(\mu_t,\nu_t)$$

with

$$\alpha = \tau\lambda_{LS}\Big(\frac{1-\gamma}{2} \wedge \frac{\kappa^2\mathrm{Diam}(\mathcal{Y})^2(1+3\gamma)}{\gamma}\Big).$$

(ii) Fast descent regime. It follows from Proposition B.2 that

$$\begin{aligned}\frac{d}{dt}\widetilde{\mathcal{L}}(\mu_t,\nu_t) \le -\frac{\eta\tau^2}{2}(1-\gamma)\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) - \gamma\tau^2\mathcal{I}(\mu_t|\mathcal{K}^-(\nu_t))\\ + \eta K_{xy}^2\mathrm{Diam}(\mathcal{X})^2(1+3\gamma)\cdot\mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)).\end{aligned}$$

Thanks to Lemma 2.3 and (31),

$$\begin{aligned}\tau\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) &\ge \tau\lambda_{LS}\mathcal{H}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t)))\\ &\ge \lambda_{LS}\mathcal{L}_3(\nu_t).\end{aligned}$$

As a result of above and (29),

$$\frac{d}{dt}\widetilde{\mathcal{L}}(\mu_t,\nu_t) \le -\frac{\eta\tau}{2}(1-\gamma)\lambda_{LS}\mathcal{L}_3(\nu_t) - \tau\Big(\gamma\lambda_{LS} - \eta K_{xy}^2\mathrm{Diam}(\mathcal{X})^2(1+3\gamma)\Big)\mathcal{L}_4(\mu_t,\nu_t).$$

Therefore setting

$$\eta = \frac{\gamma\tau^2\lambda_{LS}}{2K_{xy}^2\mathrm{Diam}(\mathcal{X})^2(1+3\gamma)}$$

and applying the Grönwall's inequality to above leads to

$$\widetilde{\mathcal{L}}(\mu_t,\nu_t) \le e^{-\alpha t}\widetilde{\mathcal{L}}(\mu_0,\nu_0)$$

with

$$\alpha = \frac{\tau\lambda_{LS}}{2}\Big(1 \wedge \eta(1-\gamma)\Big).$$

B.3. Proof of Theorem 2.7

(i) Fast ascent regime. First, thanks to Theorem 2.4, $\mathcal{L}(\mu_t,\nu_t) \le e^{-\alpha t}\mathcal{L}(\mu_0,\nu_0)$. In particular, for any $t > 0$,

$$\mathcal{L}_1(\mu_t) \le e^{-\alpha t}\mathcal{L}(\mu_0,\nu_0) \tag{42}$$

and

$$\tau\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) = \mathcal{L}_2(\mu_t,\nu_t) \le \frac{e^{-\alpha t}}{\gamma}\mathcal{L}(\mu_0,\nu_0). \tag{43}$$

As a result of (42) and (30), one obtains that

$$\tau\mathcal{H}(\mu_t|\mu^*) \le e^{-\alpha t}\mathcal{L}(\mu_0,\nu_0). \tag{44}$$

Next, to obtain the exponential decay of $\mathcal{H}(\nu_t|\nu^*)$, notice first that

$$\begin{aligned}\tau\mathcal{H}(\nu_t|\nu^*) &= \tau\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) + \tau\int_{\mathcal{Y}}(\log(\mathcal{K}^+(\mu_t)) - \log(\nu^*))d\nu_t\\ &= \tau\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) + \tau\int_{\mathcal{Y}}(\log(\mathcal{K}^+(\mu_t)) - \log(\nu^*))d(\nu_t-\nu^*) - \tau\mathcal{H}(\nu^*|\mathcal{K}^+(\nu_t))\\ &\le \tau\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) + \tau\int_{\mathcal{Y}}(\log(\mathcal{K}^+(\mu_t)) - \log(\nu^*))d(\nu_t-\nu^*).\end{aligned}$$

Since $\nu^* = \mathcal{K}^+(\mu^*)$, we have

$$(\log(\mathcal{K}^+(\mu_t)) - \log(\nu^*))(y) = \tau^{-1}\int_{\mathcal{X}} K(x,y)(\mu_t(dx) - \mu^*(dx)) - (\log Z^+(\mu_t) - \log Z^+(\mu^*)).$$

It follows from the last two displays that

$$\tau\mathcal{H}(\nu_t|\nu^*) \le \tau\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) + \int_{\mathcal{Y}}\int_{\mathcal{X}} K(x,y)(\mu_t(dx) - \mu^*(dx))(\nu_t(dy) - \nu^*(dy)). \tag{45}$$

The last term on the right side above can be upper bounded by

$$\begin{aligned}\int_{\mathcal{Y}}\int_{\mathcal{X}} K(x,y)(\mu_t(dx) - \mu^*(dx))(\nu_t(dy) - \nu^*(dy)) &\le \|K\|_\infty \cdot \mathrm{TV}(\mu_t,\mu^*)\cdot \mathrm{TV}(\nu_t,\nu^*)\\ &\le 2\|K\|_\infty\sqrt{\mathcal{H}(\mu_t|\mu^*)}\cdot\sqrt{\mathcal{H}(\nu_t|\nu^*)}\\ &\le \frac{\tau}{2}\mathcal{H}(\nu_t|\nu^*) + \frac{2\|K\|_\infty^2}{\tau}\mathcal{H}(\mu_t|\mu^*).\end{aligned} \tag{46}$$

where we have used the Pinsker's inequality and Young's inequality in the last two lines above. Finally combining the last two estimates leads to

$$\begin{aligned}\tau\mathcal{H}(\nu_t|\nu^*) &\le 2\tau\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) + \frac{4\|K\|_\infty^2}{\tau}\mathcal{H}(\mu_t|\mu^*)\\ &\le \Big(\frac{2}{\gamma} + \frac{4\|K\|_\infty^2}{\tau^2}\Big)e^{-\alpha t}\mathcal{L}(\mu_0,\nu_0),\end{aligned}$$

where we have used inequalities (43) and (44) in the last inequality above.

(ii) Fast descent regime. Thanks to Part (ii) of Theorem 2.4 and the bound (31), we have that

$$\tau\mathcal{H}(\nu_t|\nu^*) \le \widetilde{L}(\mu_t,\nu_t) \le e^{-\alpha_1 t}\widetilde{L}(\mu_0,\nu_0) \tag{47}$$

and that

$$\tau\mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)) = \mathcal{L}_4(\mu_t,\nu_t) \le \frac{e^{-\alpha_1 t}}{\gamma}\widetilde{L}(\mu_0,\nu_0). \tag{48}$$

Next to obtain the exponential decay of $\tau\mathcal{H}(\mu_t|\mu^*)$, observe that

$$\begin{aligned}\tau\mathcal{H}(\mu_t|\mu^*) &= \tau\mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)) + \tau\int_{\mathcal{X}}(\log\mathcal{K}^-(\nu_t) - \log\mu^*)d\mu_t\\ &= \tau\mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)) + \tau\int_{\mathcal{X}}(\log\mathcal{K}^-(\nu_t) - \log\mu^*)d(\mu_t-\mu^*) - \tau\mathcal{H}(\mu^*|\mathcal{K}^-(\nu_t))\\ &\le \tau\mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)) + \tau\int_{\mathcal{X}}(\log\mathcal{K}^-(\nu_t) - \log\mu^*)d(\mu_t-\mu^*).\end{aligned} \tag{49}$$

Since $\mu^* = \mathcal{K}^-(\nu^*)$, we can write

$$\log\mathcal{K}^-(\nu_t) - \log\mu^* = -\tau^{-1}\int_{\mathcal{Y}} K(x,y)(\nu_t-\nu^*)(dy) - (\log Z^-(\nu_t) - \log Z^-(\nu^*)).$$

Hence the second term on the RHS of the last line of (49) can be bounded by in a similar manner as (46). Namely,

$$\begin{aligned}\tau\int_{\mathcal{X}}(\log\mathcal{K}^-(\nu_t) - \log\mu^*)d(\mu_t-\mu^*) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} K(x,y)(\nu_t-\nu^*)(dy)(\mu_t-\mu^*)(dx)\\ &\le \frac{\tau}{2}\mathcal{H}(\mu_t|\mu^*) + \frac{2\|K\|_\infty^2}{\tau}\mathcal{H}(\nu_t|\nu^*).\end{aligned}$$

Combining the last inequality with (49) leads to

$$\begin{aligned}\tau\mathcal{H}(\mu_t|\mu^*) &\le 2\tau\mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)) + \frac{4\|K\|_\infty^2}{\tau}\mathcal{H}(\nu_t|\nu^*)\\ &\le \Big(\frac{2}{\gamma} + \frac{4\|K\|_\infty^2}{\tau^2}\Big)e^{-\alpha_2 t}\widetilde{L}(\mu_0,\nu_0),\end{aligned}$$

where we have used (47) and (48) in the last inequality.

C. Proof of convergence for the annealed dynamics

Recall that the MNE of entropy-regularized objective E_τ is given by (μ_τ^*, ν_τ^*) characterized by (1). Note that we emphasized the dependence of the objective and the optimizer on the temperature τ since the later will be time-dependent throughout this section.

It will be useful to define energies $\mathcal{E}_i, i = 1, 2$ as follows.

$$\begin{aligned}\mathcal{E}_1(\mu) &:= \max_{\nu\in\mathcal{P}(\mathcal{Y})} E_\tau(\mu,\nu) = -\tau\mathcal{H}(\mu) + \tau\log Z^+(\mu),\\ \mathcal{E}_2(\nu) &:= \min_{\mu\in\mathcal{P}(\mathcal{X})} E_\tau(\mu,\nu) = \tau\mathcal{H}(\nu) - \tau\log Z^-(\nu).\end{aligned} \tag{50}$$

It is clear that $\mathcal{L}_1(\mu) = \mathcal{E}_1(\mu) - \mathcal{E}_1(\mu^*)$.

(i) Proof of Theorem 2.8 in the fast ascent regime. The proof of the entropy decays in (20) follows largely the proof of Theorem 2.4 and is also inspired by the proof of Theorem 4.1 in (Chizat, 2022).

Step 1. Bounding $\mathcal{H}(\mu_t|\mu_{\tau_t}^*)$ and $\mathcal{H}(\nu_t|\nu_{\tau_t}^*)$.

First, let us compute the time-derivative $\frac{d}{dt}\mathcal{E}_1(\mu_{\tau_t}^*)$. Note that since

$$\mathcal{E}_1(\mu) = -\tau\mathcal{H}(\mu) + \tau\log Z^+(\mu),$$

we have for $\tau > 0$,

$$\begin{aligned}\frac{d}{d\tau}\mathcal{E}_1(\mu_\tau^*) &= -\mathcal{H}(\mu_\tau^*) + \log Z^+(\mu_\tau^*)\\ &\quad + \tau\Big(\langle\log\mu_\tau^*, \partial_\tau\mu_\tau^*\rangle + \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_\tau^*)(y)K(x,y)\frac{d}{d\tau}(\tau^{-1}\mu_\tau^*(x))dydx\Big).\end{aligned}$$

Moreover,

$$\begin{aligned}\tau\int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_\tau^*)(y)K(x,y)\frac{d}{d\tau}(\tau^{-1}\mu_\tau^*(x))dydx &= -\int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_\tau^*)(y)\tau^{-1}K(x,y)dy\mu_\tau^*(x)dx\\ &\quad + \tau\int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_\tau^*)(y)\tau^{-1}K(x,y)dy\partial_\tau\mu_\tau^*(dx)\\ &= -\int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_\tau^*)(y)\tau^{-1}K(x,y)dy\mu_\tau^*(x)dx - \tau\langle\log\mathcal{K}^-(\mathcal{K}^+(\mu_\tau^*)), \partial_\tau\mu_\tau^*\rangle.\end{aligned}$$

Combining the last two identities leads to

$$\begin{aligned}\frac{d}{d\tau}\mathcal{E}_1(\mu_\tau^*) &= -\mathcal{H}(\mu_\tau^*) + \log Z^+(\mu_\tau^*) - \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_\tau^*)(y)\tau^{-1}K(x,y)dy\mu_\tau^*(x)dx\\ &\quad + \underbrace{\tau\langle\log\mu_\tau^* - \log\mathcal{K}^-(\mathcal{K}^+(\mu_\tau^*)), \partial_\tau\mu_\tau^*\rangle}_{=\langle\partial_u\mathcal{E}_1^\tau(u)|_{u=\mu_\tau^*},\, \partial_\tau\mu_\tau^*\rangle=0}\\ &= \tau^{-1}\mathcal{E}_1(\mu_\tau^*) - \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_\tau^*)(y)\tau^{-1}K(x,y)dy\mu_\tau^*(x)dx.\end{aligned}$$

Consequently, we have

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_1(\mu_{\tau_t}^*) &= \tau_t'\frac{d}{d\tau}\mathcal{E}_1(\mu_\tau^*)\Big|_{\tau=\tau_t}\\ &= \tau_t'\Big(\tau_t^{-1}\mathcal{E}_1(\mu_{\tau_t}^*) - \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_{\tau_t}^*)(y)\tau_t^{-1}K(x,y)dy\mu_{\tau_t}^*(x)dx\Big).\end{aligned} \tag{51}$$

Next, a similar calculation as above applied to $\mathcal{E}_1(\mu_t)$ gives

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_1(\mu_t) &= \tau_t'\Big(\tau_t^{-1}\mathcal{E}_1(\mu_t) - \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_t)(y)\tau_t^{-1}K(x,y)dy\mu_t(x)dx\Big)\\ &\quad + \tau_t\int_{\mathcal{X}} \log\Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}\Big)\cdot\partial_t\mu_t(dx)\\ &= \tau_t'\Big(\tau_t^{-1}\mathcal{E}_1(\mu_t) - \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathcal{K}^+(\mu_t)(y)\tau_t^{-1}K(x,y)dy\mu_t(x)dx\Big)\\ &\quad - \tau_t^2\int_{\mathcal{X}} \nabla_x\log\Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}\Big)\cdot\nabla_x\log\Big(\frac{d\mu_t}{d\mathcal{K}^-(\nu_t)}\Big)\mu_t(dx).\end{aligned} \tag{52}$$

As a result of (52) and (51),

$$\begin{aligned}&\frac{d}{dt}\mathcal{L}_1(\mu_t) = \frac{d}{dt}(\mathcal{E}_1(\mu_t) - \mathcal{E}_1(\mu_{\tau_t}^*))\\ &= \frac{\tau_t'}{\tau_t}\Big[\mathcal{L}_1(\mu_t) - \Big(\int_{\mathcal{X}}\int_{\mathcal{Y}}\mathcal{K}^+(\mu_t)(y)K(x,y)dy\mu_t(x)dx - \int_{\mathcal{X}}\int_{\mathcal{Y}}\mathcal{K}^+(\mu_{\tau_t}^*)(y)K(x,y)dy\mu_{\tau_t}^*(x)dx\Big)\Big]\\ &\quad - \tau_t^2\int_{\mathcal{X}}\nabla_x\log\Big(\frac{d\mu_t}{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}\Big)\cdot\nabla_x\log\Big(\frac{d\mu_t}{d\mathcal{K}^-(\nu_t)}\Big)\mu_t(dx)\\ &\le \frac{\tau_t'}{\tau_t}\Big(\mathcal{L}_1(\mu_t) + 2\|K\|_\infty\Big) - \frac{\tau_t^2}{2}\mathcal{I}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))) + K_{xy}^2\mathrm{Diam}(\mathcal{Y})^2\cdot\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)).\end{aligned} \tag{53}$$

Next, we have from the ascent dynamics for ν_t,

$$\begin{aligned}\frac{d}{dt}\mathcal{L}_2(\mu_t,\nu_t) &= \tau_t'\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) + \tau_t\int_{\mathcal{Y}}\log\Big(\frac{d\nu_t}{d\mathcal{K}^+(\mu_t)}\Big)\partial_t\nu_t(dy) - \tau_t\int_{\mathcal{Y}}\frac{d}{dt}(\log\mathcal{K}^+(\mu_t))\nu_t(dy)\\ &\le \tau_t'\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) - \eta_t\tau_t^2\mathcal{I}(\nu_t|\mathcal{K}^+(\mu_t)) - \tau_t\int_{\mathcal{Y}}\frac{d}{dt}(\log\mathcal{K}^+(\mu_t))\nu_t(dy).\end{aligned} \tag{54}$$

The third term on the RHS above is

$$\begin{aligned}&-\tau_t\Big(\int_{\mathcal{Y}}\int_{\mathcal{X}}\frac{d}{dt}(\tau_t^{-1}\mu_t(x))K(x,y)dx\nu_t(y)dy - \frac{d}{dt}\log(Z^+(\mu_t))\Big)\\ &= -\tau_t\Big(\int_{\mathcal{Y}}\int_{\mathcal{X}}\frac{d}{dt}(\tau_t^{-1}\mu_t(x))K(x,y)dx(\nu_t(y) - \mathcal{K}^+(\mu_t)(y))dy\Big)\\ &= -\frac{\tau_t'}{\tau_t}\int_{\mathcal{Y}}\int_{\mathcal{X}}\mu_t(x)K(x,y)dx(\nu_t(y) - \mathcal{K}^+(\mu_t)(y))dy\\ &\qquad - \int_{\mathcal{Y}}\int_{\mathcal{X}}\partial_t\mu_t(x)K(x,y)dx(\nu_t(y) - \mathcal{K}^+(\mu_t)(y))dy\\ &= -\frac{\tau_t'}{\tau_t}\int_{\mathcal{Y}}\int_{\mathcal{X}}\mu_t(x)K(x,y)dx(\nu_t(y) - \mathcal{K}^+(\mu_t)(y))dy\\ &\quad + \tau_t^2\int_{\mathcal{X}}\nabla_x\log\Big(\frac{d\mathcal{K}^-(\mathcal{K}^+(\mu_t))}{d\mathcal{K}^-(\nu_t)}\Big)\cdot\nabla_x\log\Big(\frac{d\mu_t}{d\mathcal{K}^-(\nu_t)}\Big)d\mu_t(x)\\ &\le \frac{2\|K\|_\infty|\tau_t'|}{\tau_t} + \frac{\tau_t^2}{2}\mathcal{I}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))) + 3K_{xy}^2\mathrm{Diam}(\mathcal{Y})^2\cdot\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)).\end{aligned}$$

Combining the last two displays yields

$$\begin{aligned}\frac{d}{dt}\mathcal{L}_2(\mu_t,\nu_t) &\le \Big(\tau_t' + 3K_{xy}^2\mathrm{Diam}(\mathcal{Y})^2\Big)\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t)) - \eta_t\tau_t^2\mathcal{I}(\nu_t|\mathcal{K}^+(\mu_t))\\ &\quad + \frac{2\|K\|_\infty|\tau_t'|}{\tau_t} + \frac{\tau_t^2}{2}\mathcal{I}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))).\end{aligned} \tag{55}$$

It then follows from (53) (55) that

$$\begin{aligned}
&\frac{d}{dt}\mathcal{L}(\mu_t,\nu_t) \le -\frac{\tau_t^2(1-\gamma)\lambda_{LS}(\tau_t)}{2}\mathcal{H}(\mu_t|\mathcal{K}^-(\mathcal{K}^+(\mu_t))) + \Big(\gamma\tau_t' + (1+3\gamma)K_{xy}^2\mathrm{Diam}(\mathcal{Y})^2\Big)\mathcal{H}(\nu_t|\mathcal{K}^+(\mu_t))\\
&+ \frac{2(1+\gamma)\|K\|_\infty|\tau_t'|}{\tau_t} + \frac{\tau_t'}{\tau_t}\mathcal{L}_1(\mu_t) - \gamma\eta_t\tau_t^2\mathcal{I}(\nu_t|\mathcal{K}^+(\mu_t))\\
&\le -\Big(\frac{(1-\gamma)\tau_t\lambda_{LS}(\tau_t)}{2} - \frac{\tau_t'}{\tau_t}\Big)\mathcal{L}_1(\mu_t) - \Big(\gamma\eta_t\tau_t\lambda_{LS}(\tau_t) - \frac{\gamma\tau_t'}{\tau_t} - \frac{(1+3\gamma)K_{xy}^2\mathrm{Diam}(\mathcal{Y})^2}{\tau_t}\Big)\mathcal{L}_2(\mu_t,\nu_t)\\
&+ \frac{2(1+\gamma)\|K\|_\infty|\tau_t'|}{\tau_t}\\
&\le -\alpha_t\mathcal{L}(\mu_t,\nu_t) + \frac{2(1+\gamma)\|K\|_\infty}{t\log t},
\end{aligned}$$

where the time-dependent rate satisfies for large t that

$$\begin{aligned}
\alpha_t &= \Big(\Big(\frac{(1-\gamma)\tau_t\lambda_{LS}(\tau_t)}{2} - \frac{\tau_t'}{\tau_t}\Big) \wedge \Big(\eta_t\tau_t\lambda_{LS}(\tau_t) - \frac{\tau_t'}{\tau_t} - \frac{(1+3\gamma)K_{xy}^2\mathrm{Diam}(\mathcal{Y})^2}{\gamma\tau_t}\Big)\Big)\\
&\ge \frac{\xi}{\log t}\Big(C_1 t^{-\xi_*/\xi} \wedge \Big(\eta_t t^{-\xi^*/\xi} + \frac{C_2}{t\log t} - \frac{C_3}{(\log t)^2}\Big)\Big).
\end{aligned}$$

Note that in the above we have used the decreasing of τ_t for large t. By choosing $\eta_t \ge M\frac{t^{\xi^*/\xi}}{(\log t)^2}$ for some large $M>0$, we have

$$\alpha_t \ge \frac{Ct^{-\xi^*/\xi-\epsilon}}{\log t}.$$

As a result, we have obtained that for any $\epsilon' < 1-\xi^*/\xi$ and for t large enough,

$$\frac{d}{dt}\mathcal{L}(\mu_t,\nu_t) \le -Ct^{-\xi^*/\xi-\epsilon'}\mathcal{L}(\mu_t,\nu_t) + \frac{C'}{t\log t}.$$

Define $Q(t) = \mathcal{L}(\mu_t,\nu_t) - \frac{C'}{C}t^{-1+\xi^*/\xi+\epsilon'}$. Then it is straightforward to show that for $t > t_*$ large enough,

$$\frac{d}{dt}Q(t) \le -Ct^{-\xi^*/\xi-\epsilon'}Q(t),$$

which implies that

$$\mathcal{L}(\mu_t,\nu_t) \le Q(t_*)e^{-\frac{C}{1-\xi^*/\xi-\epsilon}(t^{1-\xi^*/\xi-\epsilon'}-t_*^{1-\xi^*/\xi-\epsilon'})} + \frac{C'}{C}t^{-1+\xi^*/\xi+\epsilon'} \le C''t^{-1+\xi^*/\xi+\epsilon'} \tag{56}$$

since $\xi^* < \xi$ and $Q(t_*)$ is finite. By the definition of $\mathcal{L}$ and the fact that $\log t \ll t^\epsilon$ for any $\epsilon > 0$, the last estimate further implies that for any $0 < \epsilon < 1-\xi^*/\xi$,

$$\mathcal{H}(\mu_t|\mu_{\tau_t}^*) \le C''t^{-1+\xi^*/\xi+\epsilon}, \text{ for } t > t_*. \tag{57}$$

In addition, using the same arguments used in the proof of the second bound in (17) from Theorem 2.7, one can obtain that

$$\mathcal{H}(\nu_t|\nu_{\tau_t}^*) \le C''t^{-1+\xi^*/\xi+\epsilon}, \text{ for } t > t_*.$$

Step 2: Bounding $\mathrm{NI}(\mu_t,\nu_t)$. Let us first claim that the difference $\mathrm{NI}(\mu_t,\nu_t) - \mathrm{NI}(\mu_{\tau_t}^*,\nu_{\tau_t}^*)$ satisfies

$$\mathrm{NI}(\mu_t,\nu_t) - \mathrm{NI}(\mu_{\tau_t}^*,\nu_{\tau_t}^*) \le Ct^{-\frac{1-\xi^*/\xi-\epsilon}{2}}. \tag{58}$$

In fact, by definition,

$$\begin{aligned}
&\mathrm{NI}(\mu_t,\nu_t) - \mathrm{NI}(\mu_{\tau_t}^*,\nu_{\tau_t}^*) = \max_{\nu\in\mathcal{P}(\mathcal{X})}E_0(\mu_t,\nu) - \max_{\nu\in\mathcal{P}(\mathcal{X})}E_0(\mu_{\tau_t}^*,\nu) + \min_{\mu\in\mathcal{P}(\mathcal{X})}E_0(\mu,\nu_{\tau_t}^*) - \min_{\mu\in\mathcal{P}(\mathcal{X})}E_0(\mu,\nu_t)\\
&= \max_{y\in\mathcal{Y}}\int_{\mathcal{X}}K(x,y)\mu_t(dx) - \max_{y\in\mathcal{Y}}\int_{\mathcal{X}}K(x,y)\mu_{\tau_t}^*(dx)\\
&\quad + \min_{x\in\mathcal{X}}\int_{\mathcal{Y}}K(x,y)\nu_{\tau_t}^*(dy) - \min_{x\in\mathcal{X}}\int_{\mathcal{Y}}K(x,y)\nu_t(dy)\\
&=: J_1 + J_2.
\end{aligned}$$

To bound J_1, let us define $y_t \in \arg\max_{y\in\mathcal{Y}} \int_{\mathcal{X}} K(x,y)\mu_t(dx)$ and $y_t^* \in \arg\max_{y\in\mathcal{Y}} \int_{\mathcal{X}} K(x,y)\mu^*_{\tau_t}(dx)$. Then by the optimality of y_t^* and the boundedness of K,

$$\begin{aligned} J_1 &= \int_{\mathcal{X}} K(x,y_t)\mu_t(dx) - \int_{\mathcal{X}} K(x,y_t^*)\mu^*_{\tau_t}(dx) \\ &\le \int_{\mathcal{X}} K(x,y_t)\mu_t(dx) - \int_{\mathcal{X}} K(x,y_t)\mu^*_{\tau_t}(dx) \\ &\le \|K\|_\infty \mathrm{TV}(\mu_t, \mu^*_{\tau_t}) \\ &\le \sqrt{2}\|K\|_\infty \sqrt{\mathcal{H}(\mu_t|\mu^*_{\tau_t})} \\ &\le Ct^{-\frac{1-\xi^*/\xi-\epsilon}{2}}. \end{aligned}$$

The same bound holds for J_2 after applying a similar argument as above, which completes the proof of (58). Finally, applying Lemma A.4 with $\tau = \tau_t = \xi/\log t$, we have for t large enough,

$$\mathrm{NI}(\mu^*_{\tau_t}, \nu^*_{\tau_t}) \le \frac{C\log(\log t)}{\log t}. \tag{59}$$

Hence the estimate (21) follows from (58) and (59).

(ii) Proof of Theorem 2.8 in the fast descent regime. By a direct calculation, one has

$$\frac{d}{dt}\mathcal{L}_3(\nu_t) \le \frac{\tau_t'}{\tau_t}(\mathcal{L}_3(\nu_t) + 2\|K\|_\infty) - \frac{\eta_t\tau_t^2}{2}\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) + \eta_t K_{xy}^2 \mathrm{Diam}(\mathcal{X})^2 \mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)) \tag{60}$$

and

$$\begin{aligned} \frac{d}{dt}\mathcal{L}_4(\mu_t,\nu_t) \le \frac{\tau_t'}{\tau_t}(\mathcal{L}_4(\nu_t) + 2\|K\|_\infty) - \tau_t^2\mathcal{I}(\mu_t|\mathcal{K}^-(\nu_t)) + \frac{\tau_t^2\eta_t}{2}\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) \\ + 3\eta_t K_{xy}^2 \mathrm{Diam}(\mathcal{X})^2 \mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)). \end{aligned} \tag{61}$$

Combining the last two displays, one can obtain the time-derivative of $\widetilde{\mathcal{L}}(\mu_t,\nu_t)$ as follows

$$\begin{aligned} \frac{d}{dt}\widetilde{\mathcal{L}}(\mu_t,\nu_t) &\le \frac{\tau_t'}{\tau_t}\widetilde{\mathcal{L}}(\mu_t,\nu_t) + \frac{\tau_t'}{\tau_t}2(1+\gamma)\|K\|_\infty - \gamma\tau_t^2\mathcal{I}(\mu_t|\mathcal{K}^-(\nu_t)) \\ &\quad - \frac{(1-\gamma)\eta_t\tau_t^2}{2}\mathcal{I}(\nu_t|\mathcal{K}^+(\mathcal{K}^-(\nu_t))) + \eta_t(1+3\gamma)K_{xy}^2\mathrm{Diam}(\mathcal{X})^2\mathcal{H}(\mu_t|\mathcal{K}^-(\nu_t)) \\ &\le \frac{\tau_t'}{\tau_t}\widetilde{\mathcal{L}}(\mu_t,\nu_t) - \frac{(1-\gamma)\eta_t\tau_t\lambda_{LS}(\tau_t)}{2}\mathcal{L}_3(\nu_t) - \tau_t\Big(\gamma\lambda_{LS}(\tau_t) \\ &\quad - \frac{\eta_t(1+3\gamma)K_{xy}^2\mathrm{Diam}(\mathcal{X})^2}{\tau_t^2}\Big)\mathcal{L}_4(\mu_t,\nu_t) + \frac{2\tau_t'}{\tau_t}(1+\gamma)\|K\|_\infty \\ &\le -\alpha_2(t)\widetilde{\mathcal{L}}(\mu_t,\nu_t) + \frac{2\tau_t'}{\tau_t}(1+\gamma)\|K\|_\infty, \end{aligned}$$

where

$$\begin{aligned} \alpha_2(t) &= \frac{(1-\gamma)\eta_t\tau_t\lambda_{LS}(\tau_t)}{2} \wedge \tau_t\Big(\lambda_{LS}(\tau_t) - \frac{\eta_t(1+3\gamma)K_{xy}^2\mathrm{Diam}(\mathcal{X})^2}{\gamma\tau_t^2} - \frac{\tau_t'}{\tau_t}\Big) \\ &\ge \frac{\xi}{\log t}t^{-(\xi^*/\xi)}\Big(C_1\eta_t \wedge \Big(t^{-(\xi^*/\xi)} + \frac{C_2}{t\log t} - \frac{C_3\eta_t}{\log(t)^2}\Big)\Big). \end{aligned}$$

Setting $\eta_t = c\log t/t$ for some $c < C_2/C_3$ in the above yields that for every $0 < \epsilon < 1 - \xi^*/\xi$ and t large enough

$$\alpha_2(t) \ge Ct^{-2\xi^*/\xi-\epsilon}.$$

Therefore we have obtained that

$$\frac{d}{dt}\widetilde{\mathcal{L}}(\mu_t,\nu_t) \le -Ct^{-2\xi^*/\xi-\epsilon}\widetilde{\mathcal{L}}(\mu_t,\nu_t) + \frac{C'}{t\log t}.$$

Similar to the proof of (56), one can obtain from above that for t large enough

$$\widetilde{\mathcal{L}}(\mu_t, \nu_t) \leq C'' t^{-1+2\xi^*/\xi+\epsilon}.$$

This directly implies the entropy decay bounds in (22). Finally the estimate (23) follows from (22) and the arguments used in Step 2 of the proof of Theorem 2.8- Part(i).