Mirror Descent-Ascent for mean-field min-max problems

MIRROR DESCENT-ASCENT FOR MEAN-FIELD MIN-MAX PROBLEMS

RAZVAN-ANDREI LASCU, MATEUSZ B. MAJKA, AND ŁUKASZ SZPRUCH

Abstract. We study two variants of the mirror descent-ascent algorithm for solving min-max problems on the space of measures: simultaneous and sequential. We work under assumptions of convexity-concavity and relative smoothness of the payoff function with respect to a suitable Bregman divergence, defined on the space of measures via flat derivatives. We show that the convergence rates to mixed Nash equilibria, measured in the Nikaidò-Isoda error, are of order $\mathcal{O}\left(N^{-1/2}\right)$ and $\mathcal{O}\left(N^{-2/3}\right)$ for the simultaneous and sequential schemes, respectively, which is in line with the state-of-the-art results for related finite-dimensional algorithms.

1. Introduction

Numerous tasks in machine learning can be framed as the optimization of a function over the space of probability measures. For instance, in supervised learning, pioneering works [15, 31, 36] showed that training a shallow neural network (NN) in the mean-field regime (i.e., an infinite-width one-hidden-layer NN) can be viewed as minimizing a convex function over the space of probability distributions of the parameters of the network. This key insight proved to be a fruitful approach in analyzing the convergence of training algorithms for infinite-width one-hidden-layer NNs (see, e.g., [21, 14, 34, 38]).

The paradigm of mean-field optimization has been extended in several works, e.g., [20, 16, 41, 30, 40, 24], which formulate the training of Generative Adversarial Networks (GANs) and adversarial robustness as a problem of finding mixed Nash equilibria (MNEs) of min-max games over the space of probability measures.

In this work, we study the convergence of an infinite-dimensional mirror descent-ascent algorithm (MDA) to mixed Nash equilibria of a min-max game with convex-concave payoff function. In games, the design of learning algorithms heavily depends on the playing conventions the players can adopt: simultaneous (players move at the same time) or sequential (each player moves upon observing the opponents' moves). To the best of our knowledge, the works concerned with studying the convergence of discrete-time algorithms for mean-field min-max games only analyze the case of simultaneous playing (see, e.g., [20, 41]). In contrast, we make a rigorous comparison between the simultaneous and sequential algorithms, and prove that sequential playing leads to faster convergence rate. This result theoretically underpins the common practice of training GANs in an alternating fashion.

1.1. Notation and setup.

For any $\mathcal{X} \subset \mathbb{R}^d$, let $\mathcal{P}(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$. In game theory, if $\mathcal{X}$ is the set of *(pure) strategies* available to the players, then $\mathcal{P}(\mathcal{X})$ is known as the set of *mixed strategies*. Let $\mathcal{C}, \mathcal{D} \subseteq \mathcal{P}(\mathcal{X})$ be convex. We consider a convex-concave (cf. Assumption 1.1) payoff function $F : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$, and the associated min-max game

$$\min_{\nu \in \mathcal{C}} \max_{\mu \in \mathcal{D}} F(\nu, \mu). \tag{1.1}$$

We are interested in finding *mixed Nash equilibria* (MNEs) for game (1.1), i.e., pairs of strategies $(\nu^*, \mu^*) \in \mathcal{C} \times \mathcal{D}$ such that, for any $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, we have

$$F(\nu^*, \mu) \leq F(\nu^*, \mu^*) \leq F(\nu, \mu^*). \tag{1.2}$$

We observe that in the case in which F is bilinear, i.e., $F(\nu, \mu) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, for some $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, measures characterized by (1.2) are MNEs in the classical sense of two-player zero-sum games. Throughout, we assume that there exists at least one MNE for game (1.1). [1]

In min-max games, the distance between a pair of strategies (ν, μ) and an MNE is typically measured using the Nikaidò-Isoda (NI) error [33], which, for all $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, is defined by

$$\mathrm{NI}(\nu, \mu) \coloneqq \max_{\mu' \in \mathcal{D}} F(\nu, \mu') - \min_{\nu' \in \mathcal{C}} F(\nu', \mu).$$

Straight from the definition, we see that $\mathrm{NI}(\nu, \mu) \geq 0$ for all $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, and from (1.2) it follows that $\mathrm{NI}(\nu, \mu) = 0$ if and only if (ν, μ) is an MNE.

1.2. **Motivating example: Training of GANs.** Let $\hat{\xi} \in \mathcal{P}(\mathcal{Y})$ be the empirical measure of the i.i.d. sampled particles $\{x_i\}_{i=1}^M \subset \mathcal{Y}$, and let $\xi \in \mathcal{P}(\mathcal{Z})$ be a source measure. Consider the measurable parametrized transport map $T_{\theta_g} : \mathcal{Z} \to \mathcal{Y}$ (which typically can be viewed as a neural network with parameters $\theta_g \in \Theta_g \subset \mathbb{R}^d$). The *pushforward* of the measure ξ on $\mathcal{Z}$ is the measure $T_{\theta_g}\#\xi$ on $\mathcal{Y}$ characterized by

$$\int_{\mathcal{Y}} \varphi \mathrm{d}\left(T_{\theta_g}\#\xi\right) = \int_{\mathcal{Z}} \left(\varphi \circ T_{\theta_g}\right) \mathrm{d}\xi,$$

for every measurable function $\varphi : \mathcal{Y} \to \mathbb{R}$.

The aim of the GAN is to search for the optimal set of parameters $\theta_g^* \in \Theta_g$ that minimizes the distance between the generated measure $T_{\theta_g^*}\#\xi$ and the empirical measure $\hat{\xi}$. In order to evaluate this distance, we define the function $D_{\theta_d} : \mathcal{Y} \to \mathbb{R}$ (which can also be viewed as a neural network with parameters $\theta_d \in \Theta_d \subset \mathbb{R}^d$), and solve the min-max problem

$$\min_{\theta_g \in \Theta_g} \max_{\theta_d \in \Theta_d} \left\{ \int_{\mathcal{Y}} D_{\theta_d}(y) \left(T_{\theta_g}\#\xi - \hat{\xi}\right)(\mathrm{d}y) \right\}.$$

For instance, observe that if the function $y \mapsto D_{\theta_d}(y)$ is continuous and 1-Lipschitz, we obtain the Wasserstein GAN [6]. On the other hand, if the function $y \mapsto D_{\theta_d}(y)$ belongs to the norm unit ball of a reproducing kernel Hilbert space (RKHS), we recover the Maximum Mean Discrepancy (MMD) GAN [27]. Solving this problem on the finite-dimensional subspaces $\Theta_g, \Theta_d \subset \mathbb{R}^d$ may pose serious challenges such as the lack of existence of pure Nash equilibria. Instead, we lift the problem to the space of probability measures and search for MNEs, i.e., optimal distributions over the set of parameters. That is, by setting $f(\theta_g, \theta_d) \coloneqq \int_{\mathcal{Y}} D_{\theta_d}(y) \left(T_{\theta_g}\#\xi - \hat{\xi}\right)(\mathrm{d}y)$, we solve the mean-field min-max game

$$\min_{\nu \in \mathcal{P}(\Theta_g)} \max_{\mu \in \mathcal{P}(\Theta_d)} \left\{ \int_{\Theta_d} \int_{\Theta_g} f(\theta_g, \theta_d)\nu(\mathrm{d}\theta_g)\mu(\mathrm{d}\theta_d) \right\}. \tag{1.3}$$

Observe that the lifted problem is bilinear in ν and μ, hence an MNE for (1.3) exists under mild assumptions (see footnote 1). We would like to emphasize that since we

[1] If F is continuous and $\mathcal{C}, \mathcal{D}$ are compact, then the existence of an MNE of (1.1) follows from Sion's minimax theorem [37]. For the particular case when $F(\nu, \mu) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, an MNE exists due to Glicksberg's minimax theorem [19] if f is continuous and $\mathcal{C}, \mathcal{D}$ are compact.

work with a convex-concave objective function F, the framework we propose is more general and includes (1.3) as a particular case. Moreover, we demonstrate theoretically by comparing Theorem 2.6 and Theorem 2.16 that sequential updates speed up GANs training significantly.

1.3. **Simultaneous and sequential MDA.** In what follows, we state our standing assumptions and the necessary definitions for introducing the simultaneous and sequential MDA schemes.

Assumption 1.1 (Convexity-concavity of F)**.** Let $F : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ be a function such that $\nu \mapsto F(\nu, \mu)$ and $\mu \mapsto F(\nu, \mu)$ admit first-order flat derivatives (cf. Definition B.1) on $\mathcal{C}$ and $\mathcal{D}$, respectively. Assume that F is convex in ν and concave in μ, i.e., for any $\nu, \nu' \in \mathcal{C}$ and any $\mu, \mu' \in \mathcal{D}$, we have

$$F(\nu', \mu) - F(\nu, \mu) \geq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, \mu, x)(\nu' - \nu)(\mathrm{d}x), \tag{1.4}$$

$$F(\nu, \mu') - F(\nu, \mu) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu, \mu, y)(\mu' - \mu)(\mathrm{d}y). \tag{1.5}$$

Let $h : \mathcal{P}(\mathcal{X}) \to \mathbb{R} \cup \{\pm\infty\}$ be a function and set $\mathcal{E} \coloneqq \mathcal{C} \cup \mathcal{D} \subseteq \mathcal{P}(\mathcal{X})$.

Assumption 1.2 (Differentiability and convexity of h)**.** Assume that $|h(\nu)| < \infty$, for all $\nu \in \mathcal{E}$, h is lower semi-continuous on $\mathcal{E}$, and admits first-order flat derivative (cf. Definition B.1) on $\mathcal{E}$. Moreover, assume that $h : \mathcal{E} \to \mathbb{R}$ is convex on $\mathcal{E}$, i.e., for all $\lambda \in [0, 1]$ and all $\nu', \nu \in \mathcal{E}$, we have $h\left((1-\lambda)\nu + \lambda\nu'\right) \leq (1-\lambda)h(\nu) + \lambda h(\nu')$.

If Assumption 1.2 holds, then it can be shown, via [21, Lemma 4.1], that h is convex on $\mathcal{E}$ in the sense of Assumption 1.1, i.e., for any $\nu', \nu \in \mathcal{E}$, we have

$$h(\nu') - h(\nu) \geq \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu, x)(\nu' - \nu)(\mathrm{d}x).$$

Following [7, Definition 3], we define h-Bregman divergence (or simply Bregman divergence) on the space of probability measures.

Definition 1.3 (Bregman divergence)**.** Suppose that Assumption 1.2 holds. The h-Bregman divergence is the map $D_h : \mathcal{E} \times \mathcal{E} \to [0, \infty)$ given by

$$D_h(\nu', \nu) \coloneqq h(\nu') - h(\nu) - \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu, x)(\nu' - \nu)(\mathrm{d}x).$$

We observe that, by Assumption 1.2, $\int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu, x)(\nu' - \nu)(\mathrm{d}x)$ is well-defined, and that $D_h(\nu', \nu) \geq 0$, for all $\nu', \nu \in \mathcal{E}$.

We now give an example of a function h and a corresponding set $\mathcal{E}$ such that Assumption 1.2 is satisfied.

Example 1.4 (Relative entropy)**.** Suppose that h is the relative entropy, i.e., $h(\nu) \coloneqq \int_{\mathcal{X}} \log \frac{\nu(x)}{\pi(x)} \nu(x)\mathrm{d}x$, where $\nu, \pi \in \mathcal{P}(\mathcal{X})$ are absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and π is a fixed reference probability measure on $\mathcal{P}(\mathcal{X})$. From [17, Lemma 1.4.3], we know that the relative entropy is convex and lower semi-continuous on $\mathcal{P}(\mathcal{X})$. Define $\mathcal{E} \coloneqq \left\{\nu \in \mathcal{P}(\mathcal{X}) : \left\|\log \frac{\nu(\cdot)}{\pi(\cdot)}\right\|_{\infty} < \infty, \text{ with } \nu, \pi \ll \mathrm{Leb}\right\}$. Clearly, we see that h is finite on $\mathcal{E}$. Moreover, it is proved in [23, Proposition 2.16] that h admits flat derivative on $\mathcal{E}$, and for all $\nu, \nu' \in \mathcal{E}$, the Bregman divergence $D_h(\nu', \nu)$ is in fact the

Kullback-Leibler divergence (or relative entropy) $\mathrm{KL}(\nu', \nu)$. For other examples of regularizers h that frequently appear in machine learning applications, and their corresponding domains $\mathcal{E}$, see [23, Proposition 2.18, 2.20].

For a given stepsize $\tau > 0$, and fixed initial pair of strategies $(\nu_0, \mu_0) \in \mathcal{C} \times \mathcal{D}$, for $n \geq 0$, the *simultaneous* and *sequential* MDA iterative schemes are respectively defined by

$$\begin{cases} \nu^{n+1} \in \underset{\nu \in \mathcal{C}}{\operatorname{argmin}} \{ \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu, \nu^n) \}, \\ \mu^{n+1} \in \underset{\mu \in \mathcal{D}}{\operatorname{argmax}} \{ \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu, \mu^n) \}, \end{cases} \tag{1.6}$$

$$\begin{cases} \nu^{n+1} \in \underset{\nu \in \mathcal{C}}{\operatorname{argmin}} \{ \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu, \nu^n) \}, \\ \mu^{n+1} \in \underset{\mu \in \mathcal{D}}{\operatorname{argmax}} \{ \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu, \mu^n) \}. \end{cases} \tag{1.7}$$

Although we abuse the notation by denoting both (1.6) and (1.7) by $(\nu^n, \mu^n)_{n \geq 0}$, we will make it clear from the context which scheme we consider.

Scheme (1.6) is referred to as *simultaneous* because both players update their strategy from step n to $n+1$ at the same time, whereas scheme (1.7) is called *sequential* because the minimizing player is first updating their move from step n to $n+1$, and then the maximizing player is acting upon observing the minimizing player's $(n+1)$-th action. Note that due to the symmetry of the players, the analysis of scheme (1.7) also covers the case when the maximizing player moves first followed by the minimizing player.

The motivation behind the use of the terms involving $\frac{\delta F}{\delta \nu}$ and $\frac{\delta F}{\delta \mu}$ in schemes (1.6) and (1.7), is that instead of minimizing and maximizing directly on F (which could be a potentially intractable problem), we minimize and maximize over ν and μ in the first-order linear approximations $F(\nu^n, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x)$ and $F(\nu^n, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y)$. In order to make sure that these approximations around (ν^n, μ^n) are precise enough, we penalize the distance between (ν^{n+1}, μ^{n+1}) and (ν^n, μ^n) by introducing the Bregman regularization terms $\frac{1}{\tau} D_h(\nu, \nu^n)$ and $\frac{1}{\tau} D_h(\mu, \mu^n)$.

We observe that by varying the choices of h in Definition 1.3 we obtain a collection of different update rules in the MDA algorithms (1.6) and (1.7). When h is the relative entropy, we can view (1.6) and (1.7) as Euler discretizations of a Fisher-Rao gradient flow, whose continuous-time convergence with explicit rates for mean-field min-max games was proved in [26] (cf. also [28] for single-player convex optimization).

1.4. **Related works.** Mirror descent (MD) was originally proposed in [32] for solving convex optimization problems and has been extensively studied on finite-dimensional vector spaces, see e.g. [9, 11, 29]. One of its main advantages over traditional gradient descent is that, by utilizing Bregman divergence as a regularization term instead of the usual squared Euclidean norm, the MD method captures the geometry of the ambient space better than the gradient descent scheme (see [9] for a detailed discussion).

Recently, the MD algorithm has been extended to infinite-dimensional settings, in order to study optimization problems on spaces of measures, thus making it a suitable method for training algorithms with applications in machine learning (ML) (Sinkhorn's and Expectation Maximization algorithms), see [7], but also for policy optimization in reinforcement learning, see [39, 22].

By leveraging results from optimization on $\mathbb{R}^d$ (see [8, 29]), the work of [7] extends the convergence proof from [29] to the case of the infinite-dimensional MD method by

showing that in order for the MD procedure to converge with rate $\mathcal{O}\left(N^{-1}\right)$, it suffices to require convexity of F in the sense of Assumption 1.1 and relative smoothness of F with respect to h. Below, we state the notion of relative smoothness in the min-max games setting.

Assumption 1.5 (Relative smoothness)**.** Let $F : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ be a function such that $\nu \mapsto F(\nu, \mu)$ and $\mu \mapsto F(\nu, \mu)$ admit first-order flat derivatives (cf. Definition B.1) on $\mathcal{C}$ and $\mathcal{D}$, respectively. Assume that, given $L_\nu, L_\mu > 0$, the function F is L_ν-smooth in ν and L_μ-smooth in μ relative to h, i.e., for any $\nu, \nu' \in \mathcal{C}$ and any $\mu, \mu' \in \mathcal{D}$, we have

$$F(\nu', \mu) - F(\nu, \mu) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, \mu, x)(\nu' - \nu)(\mathrm{d}x) + L_\nu D_h(\nu', \nu),$$

$$F(\nu, \mu') - F(\nu, \mu) \geq \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu, \mu, y)(\mu' - \mu)(\mathrm{d}y) - L_\mu D_h(\mu', \mu).$$

In Lemma A.2, we show that if $\nu \mapsto F(\nu, \mu), \mu \mapsto F(\nu, \mu)$, and h admit second-order flat derivative (cf. (B.2)) on $\mathcal{C}, \mathcal{D}$ and $\mathcal{E}$, respectively, and F is convex-concave and smooth relative to h, then the second-order flat derivatives $\frac{\delta^2 F}{\delta \nu^2}, -\frac{\delta^2 F}{\delta \mu^2}$ are non-negative and bounded above by $\frac{\delta^2 h}{\delta \nu^2}, \frac{\delta^2 h}{\delta \mu^2}$ multiplied by the respective smoothness constants. This result corresponds to the intuition we have from optimization on $\mathbb{R}^d$, where convexity and relative smoothness are equivalent, respectively, to the Hessian of F being non-negative, and upper bounded by the Hessian of h weighted by the smoothness constant.

Other works such as [20, 18] studied infinite-dimensional MDA and Mirror Prox algorithms for finding MNEs of two-player zero-sum games. The most closely related work to ours is [20], which focuses on min-max games for bilinear objective functions and utilizes a particular case of the MDA algorithm with relative entropy regularization. Our paper generalizes the setting of [20] by considering convex-concave objective function and MDA algorithm with general Bregman divergence. While [20] proves an explicit convergence rate $\mathcal{O}\left(N^{-1/2}\right)$ only for the simultaneous MDA algorithm, we also prove a faster convergence rate $\mathcal{O}\left(N^{-2/3}\right)$ for the sequential scheme. Besides the vanilla MDA algorithm, [20] considers the entropic Mirror Prox algorithm, which requires the computation of an extra gradient at an intermediate point and two projections onto the dual space. Although it is proved in [20] that the Mirror Prox algorithm achieves $\mathcal{O}\left(N^{-1}\right)$ converges rate for deterministic gradients, it is also outlined that for stochastic gradients (which one has typically access to in practice) Mirror Prox and simultaneous MDA achieve the same rate $\mathcal{O}\left(N^{-1/2}\right)$. Another approach based on reproducing kernel Hilbert spaces (RKHS) is developed in [18] and achieves the same convergence rates $\mathcal{O}\left(N^{-1}\right)$ and $\mathcal{O}\left(N^{-1/2}\right)$ for the deterministic and stochastic Mirror Prox algorithm, respectively. To our knowledge, the analysis of a sequential version of the Mirror Prox algorithm has not appeared in the literature.

1.5. **Our contribution.** In the present paper, we extend [7] to the min-max games setting. This is not straightforward due to the following two main reasons:

(1) In optimization, the monotonic decrease of the objective function along the iterates plays a central role in proving convergence of the MD algorithm as demonstrated in both the infinite-dimensional (see [7]) and the finite-dimensional (see [29]) setting. For min-max games this monotonic decrease does no longer hold due to the competing behaviour of the players. We deal with this issue by studying convergence at the level of time-averaged iterates (for a similar result in continuous-time, see [26, Theorem 2.3]).

(2) Unlike in optimization, in games players can act *simultaneously* or *sequentially*. Indeed, both players can update their moves simultaneously from iteration n to iteration $n+1$ or they can move in turns, i.e., only one of the player moves from iteration n to $n+1$ and the other player updates their strategy after observing the first player's $(n+1)$-th move. As we will highlight in this paper, different playing patterns (simultaneous or sequential) lead to different convergence rates.

In this work, we first prove that sufficient regularity of F will allow us to circumvent the absence of monotonic decrease along the NI error, and still obtain convergence rates of the MDA scheme to an MNE of F. Secondly, by exploiting the connection between MDA and its dual formulation, we prove that playing game (1.1) sequentially via the MDA algorithm yields a faster convergence rate than playing simultaneously.

To our knowledge, the technique of using the dual formulation of MDA on the space of measures is new, and we believe it may be of independent interest. In particular, we utilize the concept of duality between the Bregman divergence on the space of measures, and the dual Bregman divergence on the space of bounded measurable functions (see Section 2.2 for details).

2. Main results

In this section, we state the main results of the paper. We start by introducing the necessary assumptions.

2.1. Convergence of the simultaneous MDA scheme (1.6).

Proving that the simultaneous MDA procedure (1.6) converges relies on the following key assumption.

Assumption 2.1. Suppose that F is Lipschitz relative to h, i.e., there exists $L_F > 0$ such that, for any $\nu, \nu' \in \mathcal{C}$ and any $\mu, \mu' \in \mathcal{D}$,

$$|F(\nu', \mu') - F(\nu, \mu)|^2 \leq L_F \left(D_h(\nu', \nu) + D_h(\mu', \mu) \right).$$

Remark 2.2. We show in Lemma A.1 that Assumption 2.1 is satisfied via Pinsker's inequality, that is, $\mathrm{TV}^2(\nu', \nu) \leq \frac{1}{2}\mathrm{KL}(\nu', \nu)$, when F has bounded first-order flat derivatives in ν and μ, and h is the relative entropy, i.e., $h(\nu) := \int_{\mathcal{X}} \log \frac{\nu(x)}{\pi(x)} \nu(x) \mathrm{d}x$, where $\nu, \pi \in \mathcal{P}(\mathcal{X})$ are absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and π is a fixed reference probability measure on $\mathcal{P}(\mathcal{X})$. For other examples of functions h which satisfy the inequality $\mathrm{TV}^2(\nu', \nu) \leq \frac{1}{2} D_h(\nu', \nu)$, and hence Assumption 2.1, see [13, Lemma 3.2].

Remark 2.3. A similar notion to the Lipschitz property from Assumption 2.1, which goes under the name of *Bregman continuity*, was introduced in [5] as a generalization of the standard Lipschitz continuity.

Assumption 2.1 allows us to prove the following lemma, which will turn out to be essential for proving the main results of the paper. The proof of Lemma 2.4 is given in Appendix A.

Lemma 2.4. *Let Assumption 1.2, 1.5 and 2.1 hold. Suppose that* $\tau L \leq \frac{1}{2}$, *with* $L := \max\{L_\nu, L_\mu\}$. *Then, for both schemes* (1.6) *and* (1.7), *it holds, for all* $n \geq 0$, *that*

$$D_h(\nu^{n+1}, \nu^n) \leq 4 L_F \tau^2 \quad \text{and} \quad D_h(\mu^{n+1}, \mu^n) \leq 4 L_F \tau^2.$$

Besides the previous lemma, another important tool utilized in the proofs of the main results is the three-point inequality that is proved in Appendix A following [7, Lemma 3].

Lemma 2.5 (Three-point inequality)**.** *Let Assumption 1.2 hold. Let $G : \mathcal{E} \to \mathbb{R}$ be convex and admit flat derivative on $\mathcal{E}$. For all $\mu \in \mathcal{E}$, suppose that there exists $\bar{\nu} \in \mathcal{E}$ such that*

$$\bar{\nu} \in \operatorname*{argmin}_{\nu \in \mathcal{E}} \{G(\nu) + D_h(\nu, \mu)\}.$$

Then, for any $\nu \in \mathcal{E}$, we have

$$G(\nu) + D_h(\nu, \mu) \geq G(\bar{\nu}) + D_h(\bar{\nu}, \mu) + D_h(\nu, \bar{\nu}).$$

We are ready to state the first main result of the paper. If F is convex-concave, we prove that time-averaged iterates $\left(\frac{1}{N}\sum_{n=0}^{N-1} \nu^n, \frac{1}{N}\sum_{n=0}^{N-1} \mu^n\right)_{N\geq 1}$ generated by the simultaneous MDA algorithm (1.6), where N is the total number of iterations, converge to a local MNE of game (1.1) with rate $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ along the NI error.

Theorem 2.6 (Convergence of the simultaneous MDA scheme (1.6))**.** *Let (ν^*, μ^*) be an MNE of (1.1) and (ν^0, μ^0) be such that $\max_{\nu\in\mathcal{C}} D_h(\nu, \nu^0) + \max_{\mu\in\mathcal{D}} D_h(\mu, \mu^0) < \infty$. Suppose that Assumption 1.1, 1.2, 1.5 and 2.1 hold. Suppose that $\tau L \leq \frac{1}{2}$, with $L \coloneqq \max\{L_\nu, L_\mu\}$. Then, we have*

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1} \nu^n, \frac{1}{N}\sum_{n=0}^{N-1} \mu^n\right) \leq 4\sqrt{\frac{L_F\left(\max_{\nu\in\mathcal{C}} D_h(\nu, \nu^0) + \max_{\mu\in\mathcal{D}} D_h(\mu, \mu^0)\right)}{N}}.$$

Remark 2.7. Theorem 2.6 is consistent with the already known convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ of the MDA algorithm for min-max games with strategies in compact convex subsets of $\mathbb{R}^d$; see e.g. [11, Theorem 5.1].

Proof outline for Theorem 2.6. In their proof of convergence of the infinite-dimensional MD procedure for convex F, the authors of [7] show that relative smoothness is sufficient to prove that F is monotonically decreasing along the sequence $(\nu^n)_{n\geq 0}$ generated by MD, i.e., $F(\nu^{n+1}) \leq F(\nu^n)$, for all $n \geq 0$. The monotonicity property is key to establishing that the MD scheme converges to a minimizer of F with rate $\mathcal{O}(\frac{1}{N})$. In the case of (1.6), Assumption 1.5 and the fact that $\tau L \leq \frac{1}{2}$ imply that $F(\nu^{n+1}, \mu^n) \leq F(\nu^n, \mu^n) \leq F(\nu^n, \mu^{n+1})$, for all $n \geq 0$. Thus, in our setup, the relative smoothness assumption only illustrates the competitive behaviour of the players. We show that combining Assumption 1.5 with Assumption 2.1 allows us to control the Bregman divergence between consecutive iterates, i.e., $D_h(\nu^{n+1}, \nu^n)$ and $D_h(\mu^{n+1}, \mu^n)$, by $\mathcal{O}(\tau^2)$ (see Lemma 2.4). This condition will turn out to be sufficient to bypass the lack of monotonicity of F and also will guarantee the convergence in NI of the simultaneous MDA algorithm (1.6) with rate $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. For the proof, see Section 3.

2.2. **Convergence of the sequential MDA scheme** (1.7)**.** Before we state the main result concerning the convergence of the sequential MDA scheme (1.7), we introduce the necessary notions on the dual space of the space of probability measures.

Let $(\mathcal{M}(\mathcal{X}), \|\cdot\|_{\mathrm{TV}})$ be the Banach space of finite signed measures μ on $\mathcal{X}$ equipped with the total variation norm $\|\mu\|_{\mathrm{TV}} \coloneqq |\mu|(\mathcal{X})$. Let $(B_b(\mathcal{X}), \|\cdot\|_\infty)$ be the Banach space of bounded measurable functions from $\mathcal{X} \subset \mathbb{R}^d$ to $(\mathbb{R}, |\cdot|)$, where $|\cdot|$ is the Euclidean norm. For any $(f, m) \in B_b(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$, we define the duality pairing $\langle\cdot,\cdot\rangle : B_b(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ by

$$\langle f, m\rangle \coloneqq \int_{\mathcal{X}} f(x) m(\mathrm{d}x). \tag{2.1}$$

Next, we define the notion of convex conjugate of $h : \mathcal{P}(\mathcal{X}) \to \mathbb{R} \cup \{\pm\infty\}$ relative to the duality pairing (2.1).

Definition 2.8 (Convex conjugate)**.** Let $h : \mathcal{P}(\mathcal{X}) \to \mathbb{R} \cup \{\pm\infty\}$ be a function. Then the map $h^* : B_b(\mathcal{X}) \to \mathbb{R} \cup \{\pm\infty\}$ given by

$$h^*(f) \coloneqq \sup_{m \in \mathcal{P}(\mathcal{X})} \{\langle f, m\rangle - h(m)\}$$

is called the *convex conjugate* of h.

Assumption 2.9. Let $V \subseteq B_b(\mathcal{X})$ be convex. Assume that $|h^*(f)| < \infty$, for all $f \in V$, and h^* admits the first variation $\frac{\delta h^*}{\delta f}(f)$ (cf. (C.2)) on V.

Following [10, Theorem 2.112], it can be proved that h^* is convex on V, i.e., for all $\lambda \in [0,1]$ and all $f', f \in V$, we have that $h^*\left((1-\lambda)f + \lambda f'\right) \leq (1-\lambda)h^*(f) + \lambda h^*(f')$.

Assumption 2.10 (Strict convexity of h)**.** Suppose that $h : \mathcal{E} \to \mathbb{R}$ is strictly convex on $\mathcal{E}$.

Under Assumption 2.10, the following corollary shows that the first variation of h^* is the unique maximizer of $m \mapsto \langle f, m\rangle - h(m)$. This result is expected since on $\mathbb{R}^d$ the "gradient" of the convex conjugate (of a strictly convex function) is the maximizer of the Legendre–Fenchel transformation.

Corollary 2.11. *Suppose that Assumption 1.2, 2.9 and 2.10 hold. Let $h^* : V \to \mathbb{R}$ be the convex conjugate of h. Then, it follows that*

$$\frac{\delta h^*}{\delta f}(f) = \operatorname*{argmax}_{m \in \mathcal{E}} \{\langle f, m\rangle - h(m)\}. \tag{2.2}$$

If Assumption 1.2, 2.9 and 2.10 hold, then, via [21, Lemma 4.1], we can characterize the convexity of h^* with respect to its first variation, i.e., for any $f, f' \in V$,

$$h^*(f') - h^*(f) \geq \int_{\mathcal{X}} (f'(x) - f(x)) \frac{\delta h^*}{\delta f}(f)(\mathrm{d}x).$$

Definition 2.12 (Dual Bregman divergence)**.** Suppose that Assumption 1.2, 2.9 and 2.10 hold. Let $h^* : V \to \mathbb{R}$ be the convex conjugate of h. The *dual h^*-Bregman divergence* is the map $D_{h^*} : V \times V \to [0, \infty)$ given by

$$D_{h^*}(f', f) \coloneqq h^*(f') - h^*(f) - \int_{\mathcal{X}} (f'(x) - f(x)) \frac{\delta h^*}{\delta f}(f)(\mathrm{d}x).$$

Since f, g are bounded and $\frac{\delta h^*}{\delta g}(g)$ is a probability measure (cf. Definition C.7), it follows that $\int_{\mathcal{X}} (f(x) - g(x)) \frac{\delta h^*}{\delta g}(g)(\mathrm{d}x)$ is well-defined. Moreover, since h^* is convex, $D_{h^*}(f', f) \geq 0$, for all $f', f \in V$.

The following Lipschitzness assumption on the second variation $\frac{\delta^2 h^*}{\delta f^2}$ (cf. Definition C.11) will turn out to be crucial for showing the improvement in the convergence rate of the sequential algorithm (1.7) compared to the simultaneous algorithm.

Assumption 2.13. Suppose that $(B_b(\mathcal{X}) \times B_b(\mathcal{X})) \ni (f, g) \mapsto \frac{\delta^2 h^*}{\delta f^2}(f)(g) \in \mathcal{M}(\mathcal{X} \times \mathcal{X})$ is *two-sided L_{h^*}-Lipschitz* in the sense that there exists $L_{h^*} > 0$ such that, for all $f, g, f', g', \psi \in V$, it holds that

$$\left| \int_{\mathcal{X}\times\mathcal{X}} \psi(x)\psi(y) \left(\frac{\delta^2 h^*}{\delta f^2}(f')(g') - \frac{\delta^2 h^*}{\delta f^2}(f)(g) \right) (\mathrm{d}y \otimes \mathrm{d}x) \right|$$
$$\leq L_{h^*} \left(\|f' - f\|_\infty + \|g' - g\|_\infty \right) \int_{\mathcal{X}} \int_{\mathcal{X}} |\psi(x)||\psi(y)| \mathrm{d}y \mathrm{d}x,$$

where $\frac{\delta^2 h^*}{\delta f^2}(f)(g)(\mathrm{d}y \otimes \mathrm{d}x) \coloneqq \frac{\delta^2 h^*}{\delta f^2}(f)(\mathrm{d}x)(g)(\mathrm{d}y)$ and the norm defined on $B_b(\mathcal{X}) \times B_b(\mathcal{X})$ is given by $\|(f,g)\|_{B_b(\mathcal{X})\times B_b(\mathcal{X})} \coloneqq \|f\|_\infty + \|g\|_\infty$, for all $f, g \in B_b(\mathcal{X})$.

For a discussion on how to verify Assumption 2.13 in the case where h is the entropy, see Example C.12. The following two assumptions ensure that F and its flat derivatives $\frac{\delta F}{\delta \nu}, \frac{\delta F}{\delta \mu}$ are uniformly bounded along the sequence of iterates generated by (1.7).

Assumption 2.14 (Uniform boundedness of F)**.** Let $(\nu^n, \mu^n)_{n\geq 0}$ be defined by (1.7). Suppose that there exists $M > 0$ such that, for all $n \geq 0$,

$$|F(\nu^n, \mu^n)| \leq M.$$

Assumption 2.15 (Uniform boundedness of the flat derivatives of F)**.** Let $(\nu^n, \mu^n)_{n\geq 0}$ be defined by (1.7). Suppose that there exist $C_\nu > 0$ and $C_\mu > 0$ such that, for all $n \geq 0$, and all $(x, y) \in \mathcal{X} \times \mathcal{X}$,

$$\left|\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)\right| \leq C_\nu \quad \text{and} \quad \left|\frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)\right| \leq C_\mu.$$

Now, we are ready to state the second main result of the paper. If F is convex-concave, we prove that the time-averaged iterates $\left(\frac{1}{N}\sum_{n=0}^{N-1} \nu^n, \frac{1}{N}\sum_{n=0}^{N-1} \mu^n\right)_{N\geq 1}$ generated by the sequential MDA algorithm (1.7), where N is the total number of iterations, converge to a local MNE of F with an enhanced rate $\mathcal{O}\left(\frac{1}{N^{2/3}}\right)$ along the NI error.

Theorem 2.16 (Convergence of the sequential MDA scheme (1.7))**.** *Let (ν^*, μ^*) be an MNE of (1.1) and (ν^0, μ^0) be such that $\max_{\nu\in\mathcal{C}} D_h(\nu, \nu^0) + \max_{\mu\in\mathcal{D}} D_h(\mu, \mu^0) < \infty$. Let Assumption 1.1, 1.2, 1.5, 2.1, 2.9, 2.10, 2.13, 2.14 and 2.15 hold. Suppose that $\operatorname{diam}(\mathcal{X}) < \infty$ and $\tau L \leq \frac{1}{2}$, with $L \coloneqq \max\{L_\nu, L_\mu\}$. Then, we have*

$$\text{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1} \nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1} \mu^n\right) \leq \frac{1}{2N^{2/3}}\Bigg(3\left(\max_{\nu\in\mathcal{C}} D_h(\nu, \nu^0) + \max_{\mu\in\mathcal{D}} D_h(\mu, \mu^0)\right)^{2/3} \times \\ \times (\kappa L_{h^*} + 2L_F L))^{1/3} + 2M\Bigg), \tag{2.3}$$

where $\kappa \coloneqq \operatorname{diam}(\mathcal{X})^2\left(C_\nu^3 + C_\mu^3\right)$.

Remark 2.17. In particular, if $F(\nu, \mu) = \int_{\mathcal{X}}\int_{\mathcal{X}} f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, for a bounded function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then Assumption 2.1 is satisfied and in Definition 1.5 we have $L_\nu = L_\mu = 0$. Therefore, $L = 0$ in (2.3), and hence Theorem 2.16 is consistent with the already known convergence rate $\mathcal{O}\left(\frac{1}{N^{2/3}}\right)$ of the MDA algorithm for min-max games with strategies in compact convex subsets of $\mathbb{R}^d$ and bilinear payoff function; see [42, Theorem 3.2 and Corollary 3.3]. Since we work in an infinite-dimensional setting with a non-linear convex-concave objective function F, Theorem 2.16 substantially generalizes the results of [42].

Proof outline for Theorem 2.16. In contrast to the simultaneous MDA scheme, where the primal formulation (1.6) is sufficient to prove that $D_h(\nu^{n+1}, \nu^n)$ and $D_h(\mu^{n+1}, \mu^n)$ are of order $\mathcal{O}(\tau^2)$, and still obtain the convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, the situation for the sequential scheme (1.7) is more delicate. Besides terms such as $D_h(\nu^{n+1}, \nu^n)$ and $D_h(\mu^{n+1}, \mu^n)$, which are of order $\mathcal{O}(\tau^2)$ (see Lemma 2.4), our estimates for the sequential game will also contain the Bregman commutators $D_h(\nu^{n+1}, \nu^n) - D_h(\nu^n, \nu^{n+1})$ and $D_h(\mu^{n+1}, \mu^n) - D_h(\mu^n, \mu^{n+1})$. In order to deal with these terms, we move from the space

of probability measures to the dual space of bounded measurable functions and use Assumption 2.13. As a consequence, we prove that the Bregman commutators are of order $\mathcal{O}(\tau^3)$, which leads to the faster convergence rate $\mathcal{O}\left(\frac{1}{N^{2/3}}\right)$. For the proof, see Section 3.

3. Proofs of Theorem 2.6 and Theorem 2.16

This section is dedicated to the proofs of the main results, namely Theorem 2.6 and Theorem 2.16.

3.1. **Proof of Theorem 2.6.**

Proof of Theorem 2.6. Since $\nu \mapsto \tau \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x)$ is convex, applying Lemma 2.5 with $\bar{\nu} = \nu^{n+1}$ and $\mu = \nu^n$ implies that, for any $\nu \in \mathcal{C}$, we have

$$\tau \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + D_h(\nu, \nu^n) \geq \tau \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) + D_h(\nu^{n+1}, \nu^n) + D_h(\nu, \nu^{n+1}),$$

or, equivalently,

$$-\tau \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) - D_h(\nu, \nu^n) \leq -\tau \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) - D_h(\nu^{n+1}, \nu^n) - D_h(\nu, \nu^{n+1}). \tag{3.1}$$

Similarly, since $\mu \mapsto -\tau \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y)$ is convex, applying Lemma 2.5 with $\bar{\nu} = \mu^{n+1}$ and $\mu = \mu^n$ implies that, for any $\mu \in \mathcal{D}$, we have

$$\tau \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - D_h(\mu, \mu^n) \leq \tau \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - D_h(\mu^{n+1}, \mu^n) - D_h(\mu, \mu^{n+1}). \tag{3.2}$$

Using the convexity of $\nu \mapsto F(\nu, \mu)$ in (3.1), i.e., (1.4) with $\nu = \nu^n$ and $\mu = \mu^n$, we have that

$$F(\nu^n, \mu^n) - F(\nu, \mu^n) - \frac{1}{\tau} D_h(\nu, \nu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^n - \nu^{n+1})(\mathrm{d}x) - \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}). \tag{3.3}$$

From L_ν-relative smoothness and the fact that $\tau L \leq \frac{1}{2} < 1$, it follows that

$$\begin{aligned} F(\nu^{n+1}, \mu^n) &\leq F(\nu^n, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) + L_\nu D_h(\nu^{n+1}, \nu^n) \\ &\leq F(\nu^n, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n). \end{aligned} \tag{3.4}$$

Hence, combining (3.3) with (3.4), we obtain for any $\nu \in \mathcal{C}$ that

$$F(\nu^n, \mu^n) - F(\nu, \mu^n) - \frac{1}{\tau} D_h(\nu, \nu^n) \leq F(\nu^n, \mu^n) - F(\nu^{n+1}, \mu^n) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}). \tag{3.5}$$

Similarly, using concavity of $\mu \mapsto F(\nu, \mu)$ in (3.2), i.e., (1.5) with $\nu = \nu^n$ and $\mu = \mu^n$, we have that

$$F(\nu^n, \mu) - F(\nu^n, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu^{n+1}, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^{n+1}). \tag{3.6}$$

From L_μ-relative smoothness and the fact that $\tau L \leq \frac{1}{2} < 1$, it follows that

$$\begin{aligned} F(\nu^n, \mu^{n+1}) &\geq F(\nu^n, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - L_\mu D_h(\mu^{n+1}, \mu^n) \\ &\geq F(\nu^n, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu^{n+1}, \mu^n). \end{aligned} \tag{3.7}$$

Hence, combining (3.6) with (3.7), we obtain for any $\mu \in \mathcal{D}$ that

$$F(\nu^n, \mu) - F(\nu^n, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^n) \leq F(\nu^n, \mu^{n+1}) - F(\nu^n, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^{n+1}). \tag{3.8}$$

Adding inequalities (3.5) and (3.8) implies that for any $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$ we have

$$\begin{aligned} F(\nu^n, \mu) - F(\nu, \mu^n) &\leq F(\nu^n, \mu^n) - F(\nu^{n+1}, \mu^n) + F(\nu^n, \mu^{n+1}) - F(\nu^n, \mu^n) \\ &\quad + \frac{1}{\tau} D_h(\nu, \nu^n) + \frac{1}{\tau} D_h(\mu, \mu^n) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}) - \frac{1}{\tau} D_h(\mu, \mu^{n+1}). \end{aligned} \tag{3.9}$$

By Assumption 2.1, we have that

$$|F(\nu^n, \mu^n) - F(\nu^{n+1}, \mu^n)|^2 = |F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n)|^2 \leq L_F D_h(\nu^{n+1}, \nu^n) \leq 4 L_F^2 \tau^2,$$

and

$$|F(\nu^n, \mu^{n+1}) - F(\nu^n, \mu^n)|^2 \leq L_F D_h(\mu^{n+1}, \mu^n) \leq 4 L_F^2 \tau^2,$$

where the last inequalities follows from Lemma 2.4. Therefore, from (3.9), we obtain

$$F(\nu^n, \mu) - F(\nu, \mu^n) \leq 4 L_F \tau + \frac{1}{\tau} D_h(\nu, \nu^n) + \frac{1}{\tau} D_h(\mu, \mu^n) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}) - \frac{1}{\tau} D_h(\mu, \mu^{n+1}),$$

for any $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$. Summing the previous inequality over $n = 0, 1, ..., N-1$ and dividing by N gives

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} (F(\nu^n, \mu) - F(\nu, \mu^n)) &\leq 4 L_F \tau + \frac{1}{N\tau} \left(D_h(\nu, \nu^0) + D_h(\mu, \mu^0) - D_h(\nu, \nu^N) - D_h(\mu, \mu^N) \right) \\ &\leq 4 L_F \tau + \frac{1}{N\tau} \left(D_h(\nu, \nu^0) + D_h(\mu, \mu^0) \right), \end{aligned} \tag{3.10}$$

since $D_h(\nu, \nu^N) + D_h(\mu, \mu^N) \geq 0$, for any $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, by definition.

Since $\nu \mapsto F(\nu, \mu)$ and $\mu \mapsto -F(\nu, \mu)$ are convex, it follows by Jensen's inequality that

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} (F(\nu^n, \mu) - F(\nu, \mu^n)) &= \frac{1}{N} \sum_{n=0}^{N-1} F(\nu^n, \mu) - \frac{1}{N} \sum_{n=0}^{N-1} F(\nu, \mu^n) \\ &\geq F\left(\frac{1}{N} \sum_{n=0}^{N-1} \nu^n, \mu \right) - F\left(\nu, \frac{1}{N} \sum_{n=0}^{N-1} \mu^n \right). \end{aligned} \tag{3.11}$$

Combining (3.10) with (3.11) and taking maximum over (ν, μ) gives

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^n, \frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right) \leq 4L_F\tau + \frac{1}{N\tau}\left(\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0) + \max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right).$$

Minimizing the right-hand side over τ amounts to taking $\tau = \frac{1}{2}\sqrt{\frac{\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0)+\max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)}{L_F N}}$, and hence we obtain

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^n, \frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right) \leq 4\sqrt{\frac{L_F\left(\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0) + \max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right)}{N}}.$$

□

3.2. **Proof of Theorem 2.16.**

Proof of Theorem 2.16. We start the proof by following the same calculations from Theorem 2.6. For (1.7), after applying Lemma 2.5 and using convexity-concavity of F, (3.3) remains unchanged, i.e.,

$$F(\nu^n,\mu^n) - F(\nu,\mu^n) - \frac{1}{\tau}D_h(\nu,\nu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n,\mu^n,x)(\nu^n-\nu^{n+1})(\mathrm{d}x) - \frac{1}{\tau}D_h(\nu^{n+1},\nu^n) - \frac{1}{\tau}D_h(\nu,\nu^{n+1}),$$

while (3.6) becomes

$$F(\nu^{n+1},\mu) - F(\nu^{n+1},\mu^n) - \frac{1}{\tau}D_h(\mu,\mu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1},\mu^n,y)(\mu^{n+1}-\mu^n)(\mathrm{d}y) - \frac{1}{\tau}D_h(\mu^{n+1},\mu^n) - \frac{1}{\tau}D_h(\mu,\mu^{n+1}).$$

Adding the previous two inequalities, summing the resulting inequality over $n = 0, 1, ..., N-1$, dividing by N, using (3.11) and taking maximum over (ν, μ) we arrive at

$$\begin{aligned}(3.12)\quad \mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right) &\leq \frac{1}{N}\sum_{n=0}^{N-1}\left(\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n,\mu^n,x)(\nu^n-\nu^{n+1})(\mathrm{d}x) + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1},\mu^n,y)(\mu^{n+1}-\mu^n)(\mathrm{d}y)\right) + \frac{1}{N\tau}\left(\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0) + \max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right) \\ &+ \frac{1}{N}\sum_{n=0}^{N-1}\left(F(\nu^{n+1},\mu^n) - F(\nu^n,\mu^n)\right) - \frac{1}{N\tau}\sum_{n=0}^{N-1}\left(D_h(\nu^{n+1},\nu^n) + D_h(\mu^{n+1},\mu^n)\right),\end{aligned}$$

where we used the fact that $D_h(\nu,\nu^N) + D_h(\mu,\mu^N) \geq 0$, for any $(\nu,\mu) \in \mathcal{C}\times\mathcal{D}$.

Note that the key difference to the estimates from Theorem 2.6 is the appearance of the term $F(\nu^{n+1},\mu^n) - F(\nu^n,\mu^n)$ due to the non-symmetry of the flat derivatives of F in (1.7). The idea is to combine $F(\nu^{n+1},\mu^n) - F(\nu^n,\mu^n)$ with both $\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n,\mu^n,x)(\nu^n - \nu^{n+1})(\mathrm{d}x)$ and $\int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1},\mu^n,y)(\mu^{n+1}-\mu^n)(\mathrm{d}y)$ via relative smoothness in order to obtain $D_h(\nu^n,\nu^{n+1}) - D_h(\nu^{n+1},\nu^n)$ and $D_h(\mu^n,\mu^{n+1}) - D_h(\mu^{n+1},\mu^n)$, which will prove to be of order $\mathcal{O}(\tau^3)$.

Since the flat derivative of $\mathcal{E} \ni m \mapsto D_h(m, m') \in [0, \infty)$ is given by $\frac{\delta}{\delta m} D_h(\cdot, m') = \frac{\delta h}{\delta m}(m, x) - \frac{\delta h}{\delta m}(m', x)$, it follows that the first-order conditions for (1.7) read

$$\begin{cases} \frac{\delta h}{\delta \nu}(\nu^{n+1}, x) - \frac{\delta h}{\delta \nu}(\nu^n, x) = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x), \\ \frac{\delta h}{\delta \mu}(\mu^{n+1}, y) - \frac{\delta h}{\delta \mu}(\mu^n, y) = \tau \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y), \end{cases} \tag{3.13}$$

for all $(x, y) \in \mathcal{X} \times \mathcal{X}$ Lebesgue a.e. It can be shown directly from Definition 1.3 that

$$\int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu', x) - \frac{\delta h}{\delta \nu}(\nu, x) \right) (\nu' - \nu)(\mathrm{d}x) = D_h(\nu', \nu) + D_h(\nu, \nu'), \tag{3.14}$$

for all $\nu, \nu' \in \mathcal{C}$, and analogously for $D_h(\mu', \mu) + D_h(\mu, \mu')$. Then, using (3.13) and (3.14) we obtain that

$$\begin{aligned} -\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) &= \frac{1}{\tau} \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu^{n+1}, x) - \frac{\delta h}{\delta \nu}(\nu^n, x) \right) (\nu^{n+1} - \nu^n)(\mathrm{d}x) \\ &= \frac{1}{\tau} \left(D_h(\nu^{n+1}, \nu^n) + D_h(\nu^n, \nu^{n+1}) \right), \end{aligned} \tag{3.15}$$

and similarly

$$\begin{aligned} \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) &= \frac{1}{\tau} \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \mu}(\mu^{n+1}, y) - \frac{\delta h}{\delta \mu}(\mu^n, y) \right) (\mu^{n+1} - \mu^n)(\mathrm{d}y) \\ &= \frac{1}{\tau} \left(D_h(\mu^{n+1}, \mu^n) + D_h(\mu^n, \mu^{n+1}) \right). \end{aligned} \tag{3.16}$$

Therefore, using (3.15) and (3.16) in (3.12), we obtain that

$$\begin{aligned} \mathrm{NI} \left(\frac{1}{N} \sum_{n=0}^{N-1} \nu^{n+1}, \frac{1}{N} \sum_{n=0}^{N-1} \mu^n \right) &\leq \frac{1}{N\tau} \left(\max_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \max_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) \right) \\ &+ \frac{1}{N\tau} \sum_{n=0}^{N-1} \left(D_h(\nu^{n+1}, \nu^n) + D_h(\nu^n, \nu^{n+1}) + D_h(\mu^{n+1}, \mu^n) + D_h(\mu^n, \mu^{n+1}) \right) \\ &+ \frac{1}{N} \sum_{n=0}^{N-1} \left(F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n) \right) - \frac{1}{N\tau} \sum_{n=0}^{N-1} \left(D_h(\nu^{n+1}, \nu^n) + D_h(\mu^{n+1}, \mu^n) \right). \end{aligned} \tag{3.17}$$

Then, we observe that

$$D_h(\nu^n, \nu^{n+1}) = \frac{1}{2} \left(D_h(\nu^n, \nu^{n+1}) - D_h(\nu^{n+1}, \nu^n) \right) + \frac{1}{2} \left(D_h(\nu^n, \nu^{n+1}) + D_h(\nu^{n+1}, \nu^n) \right), \tag{3.18}$$

and a similar representation holds for $D_h(\mu^n, \mu^{n+1})$. Similarly, we can write

$$\begin{aligned} F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n) &= \frac{1}{2} \left(F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n) \right) \\ &+ \frac{1}{2} \left(F(\nu^{n+1}, \mu^n) - F(\nu^{n+1}, \mu^{n+1}) + F(\nu^{n+1}, \mu^{n+1}) - F(\nu^n, \mu^n) \right). \end{aligned} \tag{3.19}$$

Therefore, putting (3.18) and (3.19) into (3.17) gives

$$\begin{aligned}(3.20)\quad \mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1},\frac{1}{N}\sum_{n=0}^{N-1}\mu^{n}\right) &\leq \frac{1}{N\tau}\left(\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0)+\max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right)\\ &+\frac{1}{2N\tau}\sum_{n=0}^{N-1}\left(D_h(\nu^n,\nu^{n+1})-D_h(\nu^{n+1},\nu^n)+D_h(\mu^n,\mu^{n+1})-D_h(\mu^{n+1},\mu^n)\right)\\ &+\frac{1}{2N}\sum_{n=0}^{N-1}\left(\frac{1}{\tau}(D_h(\nu^n,\nu^{n+1})+D_h(\nu^{n+1},\nu^n))+F(\nu^{n+1},\mu^n)-F(\nu^n,\mu^n)\right)\\ &+\frac{1}{2N}\sum_{n=0}^{N-1}\left(\frac{1}{\tau}(D_h(\mu^n,\mu^{n+1})+D_h(\mu^{n+1},\mu^n))+F(\nu^{n+1},\mu^n)-F(\nu^{n+1},\mu^{n+1})\right.\\ &\qquad\left.+F(\nu^{n+1},\mu^{n+1})-F(\nu^n,\mu^n)\right).\end{aligned}$$

Combining the fact that $\nu\mapsto F(\nu,\mu)$ is L_ν-smooth relative to h with the first-order condition (3.13), we have that

$$\begin{aligned}(3.21)\quad F(\nu^{n+1},\mu^n)-F(\nu^n,\mu^n) &\leq \int_{\mathcal{X}}\frac{\delta F}{\delta\nu}(\nu^n,\mu^n,x)(\nu^{n+1}-\nu^n)(\mathrm{d}x)+L_\nu D_h(\nu^{n+1},\nu^n)\\ &=-\frac{1}{\tau}\int_{\mathcal{X}}\left(\frac{\delta h}{\delta\nu}(\nu^{n+1},x)-\frac{\delta h}{\delta\nu}(\nu^n,x)\right)(\nu^{n+1}-\nu^n)(\mathrm{d}x)+L_\nu D_h(\nu^{n+1},\nu^n)\\ &=-\frac{1}{\tau}\left(D_h(\nu^{n+1},\nu^n)+D_h(\nu^n,\nu^{n+1})\right)+L_\nu D_h(\nu^{n+1},\nu^n),\end{aligned}$$

where the last equality follows from (3.14).

Similarly, using L_μ-smoothness of $\mu\mapsto F(\nu,\mu)$ relative to h together with (3.13), we can show that

(3.22)

$$F(\nu^{n+1},\mu^n)-F(\nu^{n+1},\mu^{n+1})+\frac{1}{\tau}\left(D_h(\mu^n,\mu^{n+1})+D_h(\mu^{n+1},\mu^n)\right)\leq L_\mu D_h(\mu^{n+1},\mu^n).$$

Therefore, using (3.21) and (3.22) in (3.20), and recalling that $L=\max\{L_\nu,L_\mu\}$ gives

$$\begin{aligned}(3.23)\quad \mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1},\frac{1}{N}\sum_{n=0}^{N-1}\mu^{n}\right) &\leq \frac{1}{N\tau}\left(\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0)+\max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right)\\ &+\frac{1}{2N\tau}\sum_{n=0}^{N-1}\left(D_h(\nu^n,\nu^{n+1})-D_h(\nu^{n+1},\nu^n)+D_h(\mu^n,\mu^{n+1})-D_h(\mu^{n+1},\mu^n)\right)\\ &+\frac{L}{2N}\sum_{n=0}^{N-1}\left(D_h(\nu^{n+1},\nu^n)+D_h(\mu^{n+1},\mu^n)\right)+\frac{1}{2N}\left(F\left(\nu^N,\mu^N\right)-F(\nu^0,\mu^0)\right).\end{aligned}$$

Since, by Lemma 2.4, $D_h(\nu^{n+1},\nu^n)\leq 4L_F\tau^2$ and $D_h(\mu^{n+1},\mu^n)\leq 4L_F\tau^2$, it suffices to show that $D_h(\nu^n,\nu^{n+1})-D_h(\nu^{n+1},\nu^n)+D_h(\mu^n,\mu^{n+1})-D_h(\mu^{n+1},\mu^n)$ is of order $\mathcal{O}(\tau^3)$. Indeed, we could then choose $\tau=\mathcal{O}\left(\frac{1}{N^{1/3}}\right)$, and since by Assumption 2.14, $\left|F\left(\nu^N,\mu^N\right)\right|\leq M$, we would obtain that

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1},\frac{1}{N}\sum_{n=0}^{N-1}\mu^{n}\right)\leq\mathcal{O}\left(\frac{1}{N^{2/3}}\right)+\mathcal{O}\left(\frac{1}{N}\right)=\mathcal{O}\left(\frac{1}{N^{2/3}}\right),$$

because $\frac{1}{N} \leq \frac{1}{N^{2/3}}$, for all $N \geq 1$.

In order to show that $D_h(\nu^n, \nu^{n+1}) - D_h(\nu^{n+1}, \nu^n) + D_h(\mu^n, \mu^{n+1}) - D_h(\mu^{n+1}, \mu^n)$ is $\mathcal{O}(\tau^3)$, we will leverage the connection between Bregman divergence and dual Bregman divergence given by Lemma D.3 together with Assumption 2.13, 2.15.

If we denote $f^n := \frac{\delta h}{\delta \nu}(\nu^n, \cdot)$, for any $n \geq 0$, then by Lemma D.3, we have that $D_h(\nu^n, \nu^{n+1}) = D_{h^*}(f^{n+1}, f^n)$. By Definition 2.12, we have that

$$
\begin{aligned}
D_{h^*}(f^{n+1}, f^n) &= h^*(f^{n+1}) - h^*(f^n) - \int_{\mathcal{X}} \left(f^{n+1}(x) - f^n(x)\right) \frac{\delta h^*}{\delta f}(f^n)(\mathrm{d}x) \\
&= \int_0^1 \left\langle f^{n+1} - f^n, \frac{\delta h^*}{\delta f}(\lambda f^{n+1} + (1-\lambda) f^n) \right\rangle \mathrm{d}\lambda - \left\langle f^{n+1} - f^n, \frac{\delta h^*}{\delta f}(f^n) \right\rangle \\
&= \int_0^1 \left\langle f^{n+1} - f^n, \frac{\delta h^*}{\delta f}(\lambda f^{n+1} + (1-\lambda) f^n) - \frac{\delta h^*}{\delta f}(f^n) \right\rangle \mathrm{d}\lambda \\
&= \int_0^1 \int_0^1 \int_{\mathcal{X}\times\mathcal{X}} \lambda \left(f^{n+1}(x) - f^n(x)\right) \left(f^{n+1}(x') - f^n(x')\right) \times \\
&\quad \times \frac{\delta^2 h^*}{\delta f^2}(\eta\lambda f^{n+1} + (1-\eta\lambda) f^n)(\eta\lambda f^{n+1} + (1-\eta\lambda) f^n)(\mathrm{d}x' \otimes \mathrm{d}x)\mathrm{d}\eta\mathrm{d}\lambda,
\end{aligned}
$$

where the second and last equalities follow from (C.3) and (C.6), respectively. Similarly, by Lemma D.3, we have that $D_h(\nu^{n+1}, \nu^n) = D_{h^*}(f^n, f^{n+1})$, and hence

$$
\begin{aligned}
D_{h^*}(f^n, f^{n+1}) &= \int_0^1 \int_0^1 \int_{\mathcal{X}\times\mathcal{X}} \lambda \left(f^n(x) - f^{n+1}(x)\right) \left(f^n(x') - f^{n+1}(x')\right) \times \\
&\quad \times \frac{\delta^2 h^*}{\delta f^2}(\eta\lambda f^n + (1-\eta\lambda) f^{n+1})(\eta\lambda f^n + (1-\eta\lambda) f^{n+1})(\mathrm{d}x' \otimes \mathrm{d}x)\mathrm{d}\eta\mathrm{d}\lambda.
\end{aligned}
$$

Therefore, we obtain that

$$
\begin{aligned}
D_{h^*}(f^{n+1}, f^n) - D_{h^*}(f^n, f^{n+1}) &= \int_0^1 \int_0^1 \int_{\mathcal{X}\times\mathcal{X}} \lambda (f^{n+1}(x) - f^n(x))(f^{n+1}(x') - f^n(x')) \times \\
&\quad \times \left(\frac{\delta^2 h^*}{\delta f^2}(\eta\lambda f^{n+1} + (1-\eta\lambda) f^n)(\eta\lambda f^{n+1} + (1-\eta\lambda) f^n) \right. \\
&\quad \left. - \frac{\delta^2 h^*}{\delta f^2}(\eta\lambda f^n + (1-\eta\lambda) f^{n+1})(\eta\lambda f^n + (1-\eta\lambda) f^{n+1}) \right) (\mathrm{d}x' \otimes \mathrm{d}x)\mathrm{d}\eta\mathrm{d}\lambda.
\end{aligned}
$$

Using Assumption 2.13, we further obtain

$$
\begin{aligned}
D_{h^*}(f^{n+1},f^n)-D_{h^*}(f^n,f^{n+1}) \le \int_0^1\int_0^1 \Bigg| \int_{\mathcal{X}\times\mathcal{X}} \lambda(f^{n+1}(x)-f^n(x))(f^{n+1}(x')-f^n(x'))\times \\
\times \Bigg(\frac{\delta^2 h^*}{\delta f^2}(\eta\lambda f^{n+1}+(1-\eta\lambda)f^n)(\eta\lambda f^{n+1}+(1-\eta\lambda)f^n) \\
- \frac{\delta^2 h^*}{\delta f^2}(\eta\lambda f^{n}+(1-\eta\lambda)f^{n+1})(\eta\lambda f^{n}+(1-\eta\lambda)f^{n+1}) \Bigg)(\mathrm{d}x'\otimes\mathrm{d}x)\Bigg|\mathrm{d}\eta\mathrm{d}\lambda \\
\le 2L_{h^*}\int_0^1\int_0^1\int_{\mathcal{X}}\int_{\mathcal{X}} \lambda|f^{n+1}(x)-f^n(x)||f^{n+1}(x')-f^n(x')|\times \\
\times|1-2\eta\lambda|\|f^{n+1}-f^n\|_\infty \mathrm{d}x'\mathrm{d}x\mathrm{d}\eta\mathrm{d}\lambda \\
\le 2L_{h^*}\int_0^1\int_0^1\int_{\mathcal{X}}\int_{\mathcal{X}} \lambda\|f^{n+1}-f^n\|_\infty\|f^{n+1}-f^n\|_\infty\|f^{n+1}-f^n\|_\infty \mathrm{d}x'\mathrm{d}x\mathrm{d}\eta\mathrm{d}\lambda \\
\le L_{h^*}\operatorname{diam}(\mathcal{X})^2\|f^{n+1}-f^n\|_\infty^3,
\end{aligned}
$$

where the third inequality follows since $|1-2\eta\lambda|\le 1$, for all $\eta,\lambda\in[0,1]$. The first-order condition for the minimizing player in (3.13) can be rewritten as

$$
f^{n+1}(x)-f^n(x) = -\tau\frac{\delta F}{\delta\nu}(\nu^n,\mu^n,x), \tag{3.24}
$$

for all $x\in\mathcal{X}$ Lebesgue a.e. By Assumption 2.15, there exists $C_\nu>0$ such that $\left\|\frac{\delta F}{\delta\nu}(\nu^n,\mu^n,\cdot)\right\|_\infty \le C_\nu$, for any $n\ge 0$. Hence, we obtain that

$$
\begin{aligned}
D_{h^*}(f^{n+1},f^n)-D_{h^*}(f^n,f^{n+1}) &\le L_{h^*}\operatorname{diam}(\mathcal{X})^2\|f^{n+1}-f^n\|_\infty^3 \\
&= L_{h^*}\operatorname{diam}(\mathcal{X})^2\tau^3\left\|\frac{\delta F}{\delta\nu}(\nu^n,\mu^n,\cdot)\right\|_\infty^3 \le L_{h^*}\operatorname{diam}(\mathcal{X})^2\tau^3 C_\nu^3,
\end{aligned}
$$

Similarly, denoting $g^n := \frac{\delta h}{\delta\mu}(\mu^n,\cdot)$, for any $n\ge 0$, and repeating the steps above, we can prove that

$$
\begin{aligned}
D_{h^*}(g^{n+1},g^n)-D_{h^*}(g^n,g^{n+1}) &\le L_{h^*}\operatorname{diam}(\mathcal{X})^2\|g^{n+1}-g^n\|_\infty^3 \\
&= L_{h^*}\operatorname{diam}(\mathcal{X})^2\tau^3\left\|\frac{\delta F}{\delta\mu}(\nu^{n+1},\mu^n,\cdot)\right\|_\infty^3 \le L_{h^*}\operatorname{diam}(\mathcal{X})^2\tau^3 C_\mu^3,
\end{aligned}
$$

where $C_\mu>0$ exists due Assumption 2.15.

Set $\kappa := \operatorname{diam}(\mathcal{X})^2 C_\nu^3 + \operatorname{diam}(\mathcal{X})^2 C_\mu^3 > 0$. Then,

$$
D_h(\nu^n,\nu^{n+1})-D_h(\nu^{n+1},\nu^n)+D_h(\mu^n,\mu^{n+1})-D_h(\mu^{n+1},\mu^n) \le \kappa L_{h^*}\tau^3. \tag{3.25}
$$

Hence, using Lemma 2.4, (3.25) and Assumption 2.14, estimate (3.23) becomes

$$\begin{aligned}\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1},\frac{1}{N}\sum_{n=0}^{N-1}\mu^{n}\right) &\leq \frac{1}{N\tau}\left(\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0)+\max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right)\\ &+\frac{1}{2N\tau}\sum_{n=0}^{N-1}\left(\left(D_h(\nu^n,\nu^{n+1})-D_h(\nu^{n+1},\nu^n)\right)+\left(D_h(\mu^n,\mu^{n+1})-D_h(\mu^{n+1},\mu^n)\right)\right)\\ &+\frac{L}{2N}\sum_{n=0}^{N-1}\left(D_h(\nu^{n+1},\nu^n)+D_h(\mu^{n+1},\mu^n)\right)+\frac{1}{2N}\left(F\left(\nu^N,\mu^N\right)-F(\nu^0,\mu^0)\right)\\ &=\frac{1}{N\tau}\left(\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0)+\max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right)+\left(\frac{1}{2}\kappa L_{h^*}+4L_F L\right)\tau^2+\frac{M}{N}.\end{aligned}$$

Minimizing the right-hand side over τ amounts to taking $\tau=\left(\frac{\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0)+\max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)}{N(\kappa L_{h^*}+2L_F L)}\right)^{1/3}$, and since $\frac{1}{N}\leq\frac{1}{N^{2/3}}$, for any $N\geq 1$, it follows that

$$\begin{aligned}\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1},\frac{1}{N}\sum_{n=0}^{N-1}\mu^{n}\right)\leq\frac{1}{2N^{2/3}}\Bigg(3\left(\max_{\nu\in\mathcal{C}} D_h(\nu,\nu^0)+\max_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right)^{2/3}\times\\ \times\left(\kappa L_{h^*}+2L_F L\right)\Big)^{1/3}+2M\Bigg).\end{aligned}$$

□

Acknowledgements

R-AL was supported by the EPSRC Centre for Doctoral Training in Mathematical Modelling, Analysis and Computation (MAC-MIGS) funded by the UK Engineering and Physical Sciences Research Council (grant EP/S023291/1), Heriot-Watt University and the University of Edinburgh. LS acknowledges the support of the UKRI Prosperity Partnership Scheme (FAIR) under EPSRC Grant EP/V056883/1 and the Alan Turing Institute.

References

[1] R. Abraham, J. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications*. Applied Mathematical Sciences. Springer New York, 2012.

[2] C. Aliprantis and K. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2007.

[3] A. Ambrosetti and G. Prodi. *A Primer of Nonlinear Analysis*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

[4] M. Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

[5] K. Antonakopoulos, V. Belmega, and P. Mertikopoulos. An adaptive mirror-prox method for variational inequalities with singular operators. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[6] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[7] P.-C. Aubin-Frankowski, A. Korba, and F. Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17263–17275. Curran Associates, Inc., 2022.

[8] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42:330–348, 2017.

[9] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[10] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research, 2000.

[11] S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3–4):231–357, 2015.

[12] R. A. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*. Springer International Publishing, 2018.

[13] L. Chizat. Convergence Rates of Gradient Methods for Convex Optimization in the Space of Measures. *Open Journal of Mathematical Optimization*, 3:1–19, 2022.

[14] L. Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

[15] L. Chizat and F. R. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NeurIPS*, 2018.

[16] C. Domingo-Enrich, S. Jelassi, A. Mensch, G. Rotskoff, and J. Bruna. A mean-field analysis of two-player zero-sum games. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20215–20226. Curran Associates, Inc., 2020.

[17] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, New York, NY, 1997.

[18] P. Dvurechensky and J.-J. Zhu. Analysis of kernel mirror prox for measure optimization. In *27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

[19] 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.

[20] Y.-P. Hsieh, C. Liu, and V. Cevher. Finding mixed Nash equilibria of generative adversarial networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2810–2819. PMLR, 09–15 Jun 2019.

[21] K. Hu, Z. Ren, D. Šiška, and Łukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(4):2043 – 2065, 2021.

[22] B. Kerimkulov, J.-M. Leahy, D. Šiška, Ł. Szpruch, and Y. Zhang. A Fisher-Rao gradient flow for entropy-regularised Markov decision processes in Polish spaces, 2023. arXiv:2310.02951.

[23] B. Kerimkulov, D. Šiška, Ł. Szpruch, and Y. Zhang. Mirror Descent for Stochastic Control Problems with Measure-valued Controls, 2024. arXiv:2401.01198.

[24] J. Kim, K. Yamamoto, K. Oko, Z. Yang, and T. Suzuki. Symmetric mean-field langevin dynamics for distributional minimax problems. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] R.-A. Lascu, M. B. Majka, and Ł. Szpruch. Entropic mean-field min-max problems via Best Response flow, 2023. arXiv:2306.03033.

[26] R.-A. Lascu, M. B. Majka, and Łukasz Szpruch. A fisher-rao gradient flow for entropic mean-field min-max games, 2024. arXiv:2405.15834.

[27] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. Mmd gan: Towards deeper understanding of moment matching network. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[28] L. Liu, M. Majka, and Ł. Szpruch. Polyak–łojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes. *Applied Mathematics and Optimization*, 87(3), 2023.

[29] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

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

[31] 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 of the United States of America*, 115:E7665 – E7671, 2018.
[32] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.
[33] H. Nikaidô and K. Isoda. Note on non-cooperative convex game. *Pacific Journal of Mathematics*, 5:807–815, 1955.
[34] A. Nitanda, D. Wu, and T. Suzuki. Convex analysis of the mean field langevin dynamics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.
[35] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1970.
[36] G. M. Rotskoff and E. Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error, 2018. arXiv:1805.00915.
[37] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958.
[38] T. Suzuki, D. Wu, and A. Nitanda. Mean-field langevin dynamics: Time-space discretization, stochastic gradient, and variance reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
[39] M. Tomar, L. Shani, Y. Efroni, and M. Ghavamzadeh. Mirror Descent Policy Optimization, 2021. arXiv:2005.09814.
[40] C. G. Trillos and N. G. Trillos. On adversarial robustness and the use of wasserstein ascent-descent dynamics to enforce it, 2023. arXiv:2301.03662.
[41] G. Wang and L. Chizat. An exponentially converging particle method for the mixed nash equilibrium of continuous games, 2023. arXiv:2211.01280.
[42] A. Wibisono, M. Tao, and G. Piliouras. Alternating mirror descent for constrained min-max games. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35201–35212. Curran Associates, Inc., 2022.
[43] C. Zalinescu. *Convex Analysis In General Vector Spaces*. World Scientific Publishing Company, 2002.

Appendix A. Proofs of additional results

In this section, we present the proofs of the additional results of the paper. We start with the proofs of Lemma 2.4 and Lemma 2.5, which play a key role in proving the main results. Then we continue with the proofs of some auxiliary results.

A.1. **Proof of Lemma 2.4.**

Proof of Lemma 2.4. We will only prove the lemma for scheme (1.6) since the argument for (1.7) is almost identical. From L_ν-relative smoothness and the definition of ν^{n+1} in (1.6), for any $\nu \in \mathcal{C}$, it follows that

$$F(\nu^{n+1},\mu^n) \le F(\nu^n,\mu^n)+\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n,\mu^n,x)(\nu^{n+1}-\nu^n)(\mathrm{d}x)+\left(\frac{1}{\tau}+L_\nu-\frac{1}{\tau}\right)D_h(\nu^{n+1},\nu^n)$$
$$\le F(\nu^n,\mu^n)+\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n,\mu^n,x)(\nu-\nu^n)(\mathrm{d}x)+\frac{1}{\tau}D_h(\nu,\nu^n)+\left(L_\nu-\frac{1}{\tau}\right)D_h(\nu^{n+1},\nu^n).$$

Setting $\nu=\nu^n$, we obtain that

$$F(\nu^{n+1},\mu^n) \le F(\nu^n,\mu^n)+\left(L_\nu-\frac{1}{\tau}\right)D_h(\nu^{n+1},\nu^n).$$

Recall $L \coloneqq \max\{L_\nu, L_\mu\} > 0$. By assumption, $\tau L \le \frac{1}{2}$, and so we get

$$\frac{1}{2\tau}D_h(\nu^{n+1},\nu^n) \le F(\nu^n,\mu^n)-F(\nu^{n+1},\mu^n) \le \sqrt{L_F}\sqrt{D_h(\nu^{n+1},\nu^n)},$$

where the last inequality follows from Assumption 2.1. Hence, since $D_h(\nu^{n+1}, \nu^n) \geq 0$, for all $n \geq 0$, we obtain that

$$D_h(\nu^{n+1}, \nu^n) \leq 4L_F\tau^2.$$

From L_μ-relative smoothness and the definition of μ^{n+1} in (1.6), for any $\mu \in \mathcal{D}$, it follows that

$$F(\nu^n, \mu^{n+1}) \geq F(\nu^n, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - \left(\frac{1}{\tau} + L_\mu - \frac{1}{\tau}\right) D_h(\mu^{n+1}, \mu^n)$$
$$\geq F(\nu^n, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu, \mu^n) - \left(L_\mu - \frac{1}{\tau}\right) D_h(\mu^{n+1}, \mu^n).$$

Setting $\mu = \mu^n$, we obtain that

$$F(\nu^n, \mu^{n+1}) \geq F(\nu^n, \mu^n) - \left(L_\mu - \frac{1}{\tau}\right) D_h(\mu^{n+1}, \mu^n).$$

Using again the assumption $\tau L \leq \frac{1}{2}$, we get

$$\frac{1}{2\tau} D_h(\mu^{n+1}, \mu^n) \leq F(\nu^n, \mu^{n+1}) - F(\nu^n, \mu^n) \leq \sqrt{L_F}\sqrt{D_h(\mu^{n+1}, \mu^n)},$$

where the last inequality follows from Assumption 2.1. Hence, since $D_h(\mu^{n+1}, \mu^n) \geq 0$, for all $n \geq 0$, we obtain that

$$D_h(\mu^{n+1}, \mu^n) \leq 4L_F\tau^2.$$

□

A.2. **Proof of Lemma 2.5.**

Proof of Lemma 2.5. From Definition 1.3, we have

$$D_h(\nu, \mu) = h(\nu) - h(\mu) - \int_{\mathcal{X}} \frac{\delta h}{\delta \mu}(\mu, y)(\nu - \mu)(\mathrm{d}y),$$

and hence, for any $\mu \in \mathcal{K}$, and all $y \in \mathcal{X}$ Lebesgue a.e., we have

$$\left(\frac{\delta D_h}{\delta \nu}(\nu, \mu, y)\right)\Big|_{\nu=\bar{\nu}} = \frac{\delta h}{\delta \nu}(\bar{\nu}, y) - \frac{\delta h}{\delta \mu}(\mu, y).$$

Therefore, for any $\mu \in \mathcal{K}$, we have that

$$\begin{aligned}
D_{D_h(\cdot,\mu)}(\nu, \bar{\nu}) &= D_h(\nu, \mu) - D_h(\bar{\nu}, \mu) - \int_{\mathcal{X}} \left(\frac{\delta D_h}{\delta \nu}(\nu, \mu, y)\right)\Big|_{\nu=\bar{\nu}} (\nu - \bar{\nu})(\mathrm{d}y) \\
&= D_h(\nu, \mu) - D_h(\bar{\nu}, \mu) - \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\bar{\nu}, y)(\nu - \bar{\nu})(\mathrm{d}y) + \int_{\mathcal{X}} \frac{\delta h}{\delta \mu}(\mu, y)(\nu - \bar{\nu})(\mathrm{d}y) \\
&= h(\nu) - h(\bar{\nu}) - \int_{\mathcal{X}} \frac{\delta h}{\delta \mu}(\mu, y)(\nu - \mu)(\mathrm{d}y) + \int_{\mathcal{X}} \frac{\delta h}{\delta \mu}(\mu, y)(\bar{\nu} - \mu)(\mathrm{d}y) \\
&\quad - \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\bar{\nu}, y)(\nu - \bar{\nu})(\mathrm{d}y) + \int_{\mathcal{X}} \frac{\delta h}{\delta \mu}(\mu, y)(\nu - \bar{\nu})(\mathrm{d}y) \\
&= h(\nu) - h(\bar{\nu}) - \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\bar{\nu}, y)(\nu - \bar{\nu})(\mathrm{d}y) \\
&= D_h(\nu, \bar{\nu}).
\end{aligned}$$

Given $\mu \in \mathcal{K}$, if we denote $g(\nu) \coloneqq G(\nu) + D_h(\nu, \mu)$, then by linearity of flat derivative, we further obtain that

$$D_g(\nu, \bar{\nu}) = D_{G(\cdot)+D_h(\cdot,\mu)}(\nu, \bar{\nu}) = D_G(\nu, \bar{\nu}) + D_{D_h(\cdot,\mu)}(\nu, \bar{\nu}) = D_G(\nu, \bar{\nu}) + D_h(\nu, \bar{\nu}) \geq D_h(\nu, \bar{\nu}),$$

since $D_G(\nu, \bar{\nu}) \geq 0$ by convexity of G. By optimality of $\bar{\nu}$, the first-order condition $\frac{\delta g}{\delta \nu}(\bar{\nu}, y) = \text{constant}$ holds for all $y \in \mathcal{X}$ Lebesgue a.e., and hence

$$g(\nu) - g(\bar{\nu}) - D_g(\nu, \bar{\nu}) = 0.$$

Therefore, we obtain that

$$g(\nu) = g(\bar{\nu}) + D_g(\nu, \bar{\nu}) \geq g(\bar{\nu}) + D_h(\nu, \bar{\nu}),$$

which is the desired inequality. □

A.3. **Proofs of auxiliary results.**

Lemma A.1. *Suppose that there exists $C_{F,\nu} > 0$ and $C_{F,\mu} > 0$ such that, for all $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, and all $(x, y) \in \mathcal{X} \times \mathcal{X}$, it holds that*

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

Take h to be the relative entropy, i.e., $h(\nu) \coloneqq \int_{\mathcal{X}} \log \frac{\nu(x)}{\pi(x)} \nu(\mathrm{d}x)$, where $\nu, \pi \in \mathcal{P}(\mathcal{X})$ are absolutely continuous with respect to Lebesgue measure on $\mathcal{X}$ and π is fixed reference probability measures on $\mathcal{P}(\mathcal{X})$. Then Assumption 2.1 is satisfied.

Proof. Since h is the relative entropy, it follows from Example 1.4 that Bregman divergence is in fact the Kullback-Leibler divergence. Then, from Definition B.1, we have

$$\begin{aligned}
|F(\nu', \mu') - F(\nu, \mu)| &= |F(\nu', \mu') - F(\nu, \mu') + F(\nu, \mu') - F(\nu, \mu)| \\
&\leq |F(\nu', \mu') - F(\nu, \mu')| + |F(\nu, \mu') - F(\nu, \mu)| \\
&= \left| \int_0^1 \int_{\mathcal{X}} \frac{\delta F}{\delta \nu} (\nu + \varepsilon(\nu' - \nu), \mu', x) (\nu' - \nu)(\mathrm{d}x)\mathrm{d}\varepsilon \right| \\
&+ \left| \int_0^1 \int_{\mathcal{X}} \frac{\delta F}{\delta \mu} (\nu, \mu + \varepsilon(\mu' - \mu), y) (\mu' - \mu)(\mathrm{d}y)\mathrm{d}\varepsilon \right| \\
&\leq C_{F,\nu} \operatorname{TV}(\nu', \nu) + C_{F,\mu} \operatorname{TV}(\mu', \mu),
\end{aligned}$$

where the last inequality follows since F is assumed to have bounded first-order flat derivatives by $C_{F,\nu}, C_{F,\mu} > 0$, respectively. The conclusion follows by squaring both sides and applying Pinsker's inequality, that is, $\operatorname{TV}^2(\nu', \nu) \leq \frac{1}{2} \operatorname{KL}(\nu', \nu)$. □

Lemma A.2. *Let Assumption 1.1, 1.2 and 1.5 hold. Suppose that $\nu \mapsto F(\nu, \mu)$, $\mu \mapsto F(\nu, \mu)$, and h admit second-order flat derivative (cf.* (B.2)*) on $\mathcal{C}, \mathcal{D}$ and $\mathcal{E}$, respectively. Then, we have*

$$\begin{aligned}
0 &\leq \int_0^1 \int_{\mathcal{X}} \int_0^{\varepsilon} \int_{\mathcal{X}} \frac{\delta^2 F}{\delta \nu^2} (\nu + \eta(\nu' - \nu), \mu, x, x') (\nu' - \nu)(\mathrm{d}x')\mathrm{d}\eta(\nu' - \nu)(\mathrm{d}x)\mathrm{d}\varepsilon \\
&\leq L_\nu \int_0^1 \int_{\mathcal{X}} \int_0^{\varepsilon} \int_{\mathcal{X}} \frac{\delta^2 h}{\delta \nu^2} (\nu + \eta(\nu' - \nu), x, x') (\nu' - \nu)(\mathrm{d}x')\mathrm{d}\eta(\nu' - \nu)(\mathrm{d}x)\mathrm{d}\varepsilon,
\end{aligned}$$

$$0 \le -\int_0^1 \int_{\mathcal{X}} \int_0^{\varepsilon} \int_{\mathcal{X}} \frac{\delta^2 F}{\delta \mu^2}\left(\nu, \mu + \eta(\mu' - \mu), y, y'\right)(\mu' - \mu)(\mathrm{d}y')\mathrm{d}\eta(\mu' - \mu)(\mathrm{d}y)\mathrm{d}\varepsilon$$
$$\le L_\mu \int_0^1 \int_{\mathcal{X}} \int_0^{\varepsilon} \int_{\mathcal{X}} \frac{\delta^2 h}{\delta \mu^2}\left(\mu + \eta(\mu' - \mu), y, y'\right)(\mu' - \mu)(\mathrm{d}y')\mathrm{d}\eta(\mu' - \mu)(\mathrm{d}y)\mathrm{d}\varepsilon.$$

Proof. We observe that combining relative smoothness and convexity for $\nu \mapsto F(\nu, \mu)$ gives that for some $L_\nu > 0$, any $\nu, \nu' \in \mathcal{C}$ and any $\mu, \mu' \in \mathcal{D}$, we have

$$0 \le F(\nu', \mu) - F(\nu, \mu) - \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, \mu, x)(\nu' - \nu)(\mathrm{d}x) \le L_\nu D_h(\nu', \nu). \tag{A.1}$$

Since $\nu \mapsto F(\nu, \mu)$, $\mu \mapsto F(\nu, \mu)$, and h admit second-order flat derivative (cf. (B.2)) on $\mathcal{C}, \mathcal{D}$ and $\mathcal{E}$, respectively, from (A.1), we obtain

$$0 \le \int_0^1 \int_{\mathcal{X}} \int_0^{\varepsilon} \int_{\mathcal{X}} \frac{\delta^2 F}{\delta \nu^2}\left(\nu + \eta(\nu' - \nu), \mu, x, x'\right)(\nu' - \nu)(\mathrm{d}x')\mathrm{d}\eta(\nu' - \nu)(\mathrm{d}x)\mathrm{d}\varepsilon$$
$$\le L_\nu \int_0^1 \int_{\mathcal{X}} \int_0^{\varepsilon} \int_{\mathcal{X}} \frac{\delta^2 h}{\delta \nu^2}\left(\nu + \eta(\nu' - \nu), x, x'\right)(\nu' - \nu)(\mathrm{d}x')\mathrm{d}\eta(\nu' - \nu)(\mathrm{d}x)\mathrm{d}\varepsilon.$$

The analogous inequalities are similarly obtained for relative smoothness and relative concavity. □

When F is strongly-convex-strongly-concave relative to h and Assumption 2.10 holds, it can be shown that (ν^*, μ^*) is the unique MNE of (1.1) (see the proof of [25, Lemma A.5]). Moreover, based on relative convexity-concavity of F, we prove in Lemma A.4 that the NI error satisfies a type of "quadratic growth" inequality relative to h.

Assumption A.3 (Relative convexity-concavity)**.** Let $F : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ be a function such that $\nu \mapsto F(\nu, \mu)$ and $\mu \mapsto F(\nu, \mu)$ admit first-order flat derivatives (cf. Definition B.1) on $\mathcal{C}$ and $\mathcal{D}$, respectively. Assume that, given $\ell_\nu, \ell_\mu > 0$, the function F is ℓ_ν-strongly convex in ν and ℓ_μ-strongly concave in μ relative to h, i.e., for any $\nu, \nu' \in \mathcal{C}$ and any $\mu, \mu' \in \mathcal{D}$, we have

$$F(\nu', \mu) - F(\nu, \mu) \ge \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, \mu, x)(\nu' - \nu)(\mathrm{d}x) + \ell_\nu D_h(\nu', \nu), \tag{A.2}$$

$$F(\nu, \mu') - F(\nu, \mu) \le \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu, \mu, y)(\mu' - \mu)(\mathrm{d}y) - \ell_\mu D_h(\mu', \mu). \tag{A.3}$$

Lemma A.4 ("Quadratic growth" of NI error relative to h)**.** *Suppose that Assumption 1.2 and A.3 hold. Then, for all $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, it holds that*

$$\mathrm{NI}(\nu, \mu) \ge \ell\left(D_h(\nu, \nu^*) + D_h(\mu, \mu^*)\right),$$

where $\ell \coloneqq \min\{\ell_\nu, \ell_\mu\}$.

Remark A.5. We refer to Lemma A.4 as "quadratic growth" of NI error relative to h due to the similar notion of quadratic growth of a convex function relative to the squared Euclidean norm on $\mathbb{R}^d$ (see e.g. [4]).

Proof. Let $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$. Since F is ℓ_ν-strongly convex in ν and ℓ_μ-strongly concave in μ, it follows that

$$F(\nu, \mu^*) - F(\nu^*, \mu^*) \ge \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^*, \mu^*, x)(\nu - \nu^*)(\mathrm{d}x) + \ell_\nu D_h(\nu, \nu^*),$$

$$F(\nu^*, \mu) - F(\nu^*, \mu^*) \le \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^*, \mu^*, y)(\mu - \mu^*)(\mathrm{d}y) - \ell_\mu D_h(\mu, \mu^*).$$

Since (ν^*, μ^*) is the MNE of F, we have

$$\frac{\delta F}{\delta \nu}(\nu^*, \mu^*, x) = \text{constant}, \quad \frac{\delta F}{\delta \mu}(\nu^*, \mu^*, y) = \text{constant},$$

for all $(x, y) \in \mathcal{X} \times \mathcal{X}$ Lebesgue a.e. Hence, adding the inequalities above and using the definition of NI error, we get

$$\mathrm{NI}(\nu, \mu) \geq \ell \left(D_h(\nu, \nu^*) + D_h(\mu, \mu^*) \right).$$

□

By Lemma A.4, the time-averaged iterates $\left(\frac{1}{N} \sum_{n=0}^{N-1} \nu^n, \frac{1}{N} \sum_{n=0}^{N-1} \mu^n \right)$ converge in Bregman divergence to the unique MNE (ν^*, μ^*) of (1.1) with the rates proved in Theorem 2.6 and Theorem 2.16, respectively.

Appendix B. Differentiability on the primal space

In this section, following [12, Definition 5.43], we introduce the notion of differentiability on the space of probability measure that we utilize throughout the paper.

Definition B.1. For any $\mathcal{X} \subset \mathbb{R}^d$, let $\mathcal{K} \subseteq \mathcal{P}(\mathcal{X})$ be convex. A function $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ admits first-order flat derivative on $\mathcal{K}$, if there exists a function $\frac{\delta F}{\delta \nu} : \mathcal{K} \times \mathcal{X} \to \mathbb{R}$, such that

(1) the map $\mathcal{K} \times \mathcal{X} \ni (\nu, x) \mapsto \frac{\delta F}{\delta \nu}(\nu, x)$ is continuous with respect to the product topology, where $\mathcal{P}(\mathcal{X})$ is equipped with the Total Variation (TV) distance,
(2) For any $\nu \in \mathcal{K}$, there exists $C > 0$ such that, for all $x \in \mathcal{X}$, we have

$$\left| \frac{\delta F}{\delta \nu}(\nu, x) \right| \leq C,$$

(3) For all $\nu, \nu' \in \mathcal{K}$, it holds that

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

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

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

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

Appendix C. Differentiability on the dual space

In this section, we start by recalling the notions of Fréchet and Gâteaux derivative for functions $H : B_b(\mathcal{X}) \to \mathfrak{X}$, where $(B_b(\mathcal{X}), \|\cdot\|_\infty)$ is the Banach space of real-valued bounded measurable functions on $\mathcal{X} \subset \mathbb{R}^d$ and $(\mathfrak{X}, \|\cdot\|_{\mathfrak{X}})$ is a normed vector space; see e.g. Chapters 7, 1, 3 in [2, 3, 35], respectively. Based on these notions of differentiablity, we will introduce the notions of first and second variation for functions H.

C.1. **Preliminaries on Fréchet and Gâteaux derivatives.** For $\mathcal{X} \subset \mathbb{R}^d$, let $\mathcal{L}\left(B_b\left(\mathcal{X}\right), \mathfrak{X}\right)$ and $\mathcal{L}\left(B_b\left(\mathcal{X}\right)\right)$ denote the space of continuous linear maps from $B_b\left(\mathcal{X}\right)$ to $\mathfrak{X}$, and from $B_b\left(\mathcal{X}\right)$ to itself, respectively.

Definition C.1 (Fréchet differentiability)**.** Let $\mathcal{U} \subset B_b\left(\mathcal{X}\right)$ be open. Given $f \in \mathcal{U}$, the function $H : \mathcal{U} \to \mathfrak{X}$ is *Fréchet differentiable* at f if there exists $T \in \mathcal{L}(B_b(\mathcal{X}), \mathfrak{X})$ such that, for all $g \in B_b(\mathcal{X})$,

$$\lim_{\|g\|_\infty \to 0} \frac{\|H\left(f+g\right) - H(f) - T\left(g\right)\|_{\mathfrak{X}}}{\|g\|_\infty} = 0.$$

If it exists, the map T is unique, we write $T = \nabla_{\mathcal{F}} H(f)$, and call $\nabla_{\mathcal{F}} H(f)$ the *Fréchet derivative* of H at f. If H is Fréchet differentiable at every $f \in \mathcal{U}$, then we say that H is Fréchet differentiable on $\mathcal{U}$.

Example C.2. If h is the entropy, then a straightforward calculation directly from Definition 2.8 shows that its dual h^* is given by

$$h^*(f) = \log\left(\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z\right).$$

Consequently, from Definition C.1 and following the argument from [22, Proposition 3.9], we can show that that h^* is Fréchet differentiable on $B_b(\mathcal{X})$ with Fréchet derivative given by

$$\nabla_{\mathcal{F}} h^*(f)(g) = \int_{\mathcal{X}} g(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z, \tag{C.1}$$

for all $g \in B_b(\mathcal{X})$.

Definition C.3 (Gâteaux differentiability)**.** Let $\mathcal{U} \subset B_b\left(\mathcal{X}\right)$ be open. Given $f \in \mathcal{U}$, the function $H : \mathcal{U} \to \mathfrak{X}$ is *Gâteaux differentiable* at f if there exists $T \in \mathcal{L}(B_b(\mathcal{X}), \mathfrak{X})$ such that for any direction $f' \in B_b(\mathcal{X})$,

$$\lim_{\varepsilon \downarrow 0} \frac{H\left(f + \varepsilon f'\right) - H\left(f\right)}{\varepsilon} = T\left(f'\right).$$

If it exists, the map T is unique, we write $T = \nabla_{\mathcal{G}} H(f)$, and call $\nabla_{\mathcal{G}} H(f)$ the *Gâteaux derivative* of H at f. If H is Gâteaux differentiable at every $f \in \mathcal{U}$, then we say that H is Gâteaux differentiable on $\mathcal{U}$.

As observed in Chapter 1, 3 in [3, 35], if H is Fréchet differentiable, then it is automatically Gâteaux differentiable and the two derivatives coincide, i.e., $\nabla_{\mathcal{F}} H = \nabla_{\mathcal{G}} H$. Moreover, [35, Proposition 3.1.6] proves that Fréchet differentiability of H at $f \in \mathcal{U}$ implies that H is continuous at f, whereas in the case of Gâteaux differentiability, this does not necessarily hold; see [35, Proposition 3.1.4].

Following the discussions in [2, 3, 35], it is possible to extend Definition C.1 to higher-order Fréchet derivatives.

Definition C.4 (Second-order Fréchet differentiability)**.** Let $\mathcal{U} \subset B_b\left(\mathcal{X}\right)$ be open and let $f \in \mathcal{U}$. Suppose that $H : \mathcal{U} \to \mathfrak{X}$ is *Fréchet differentiable* (cf. Definition C.1) at f, and admits Fréchet derivative $\nabla_{\mathcal{F}} H(f)$. Then $\nabla_{\mathcal{F}} H(f)$ is *Fréchet differentiable* at f, if there exists $T \in \mathcal{L}\left(B_b(\mathcal{X}), \mathcal{L}\left(B_b(\mathcal{X}), \mathfrak{X}\right)\right)$ such that for all $f', f'' \in B_b(\mathcal{X})$,

$$\lim_{\|f''\|_\infty \to 0} \frac{\|\nabla_{\mathcal{F}} H\left(f + f''\right)\left(f'\right) - \nabla_{\mathcal{F}} H(f)(f') - T\left(f''\right)\left(f'\right)\|_{\mathfrak{X}}}{\|f''\|_\infty} = 0.$$

If it exists, the map T is unique, we write $T = \nabla^2_{\mathcal{F}} H(f)$, and call $\nabla^2_{\mathcal{F}} H(f)$ the *second Fréchet derivative* of H at f.

Example C.5. If h is the entropy, using (C.1) and following the argument from [22, Proposition 3.6], we can show that $\nabla_{\mathcal{F}} h^*(f)$ is Fréchet differentiable on $B_b(\mathcal{X})$ with Fréchet derivative given by

$$\begin{aligned}
\nabla^2_{\mathcal{F}} h^*(f)(f')(g) &= \int_{\mathcal{X}} g(x) \left(f'(x) - \int_{\mathcal{X}} f'(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z \right) \frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}x \\
&= \int_{\mathcal{X}} \left(g(x) - \int_{\mathcal{X}} g(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z + \int_{\mathcal{X}} g(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z \right) \times \\
&\qquad \times \left(f'(x) - \int_{\mathcal{X}} f'(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z \right) \frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}x \\
&= \int_{\mathcal{X}} \left(g(x) - \int_{\mathcal{X}} g(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z \right) \left(f'(x) - \int_{\mathcal{X}} f'(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z \right) \frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}x,
\end{aligned}$$

for all $g \in B_b(\mathcal{X})$, where the last line used the fact that

$$\int_{\mathcal{X}} \int_{\mathcal{X}} g(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \left(f'(x) - \int_{\mathcal{X}} f'(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z \right) \frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z \mathrm{d}x = 0.$$

The motivation behind working with Fréchet instead of Gâteaux differentiability is that the higher-order derivatives in the case of the former could be identified with continuous symmetric multilinear maps. As proved in Section 3 of Chapter 1 from [3], the space $\mathcal{L}\left(B_b(\mathcal{X}), \mathcal{L}\left(B_b(\mathcal{X}), \mathfrak{X}\right)\right)$ is isometrically isomorphic to $\mathcal{L}_2\left(B_b(\mathcal{X}), \mathfrak{X}\right)$, i.e., the space of continuous bilinear maps from $B_b(\mathcal{X}) \times B_b(\mathcal{X})$ to $\mathfrak{X}$, and therefore, we could naturally view the second-order Fréchet derivative of H, if it exists, as a continuous bilinear map.

Furthermore, due to [3, Theorem 3.5], we have that the second-order Fréchet derivative is always symmetric. On the contrary, the second-order Gâteaux derivative is not necessarily symmetric as noted on page 78 in [35].

Remark C.6. If we replace $B_b(\mathcal{X})$ with $\mathbb{R}^d$ and $\mathfrak{X}$ with $\mathbb{R}$, then the first and second-order Fréchet derivatives are precisely the gradient and Hessian matrix of H at f.

C.2. **First and second variations.** Following Chapter 2 from [1], we introduce the notions of first and second variation for Fréchet differentiable functions H, relative to the duality pairing (2.1).

Definition C.7 (First variation of H)**.** Let $H : B_b(\mathcal{X}) \to \mathfrak{X}$ be Fréchet differentiable at $f \in B_b(\mathcal{X})$. If it exists, the *first variation* of H at f is the unique continuous map $B_b(\mathcal{X}) \ni f \mapsto \frac{\delta H}{\delta f}(f) \in \mathcal{P}(\mathcal{X})$ such that, for all $g \in B_b(\mathcal{X})$,

$$\left\langle g, \frac{\delta H}{\delta f}(f) \right\rangle \coloneqq \nabla_{\mathcal{F}} H(f)(g).$$

From Example C.2, we observe that the first variation $\frac{\delta h^*}{\delta f} : B_b(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ of h^* is given by

$$\frac{\delta h^*}{\delta f}(f)(\mathrm{d}z) = \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)} \mathrm{d}z} \mathrm{d}z.$$

Assuming that $H : B_b(\mathcal{X}) \to \mathbb{R}$ is Fréchet differentiable at $f \in B_b(\mathcal{X})$ with Fréchet derivative $\nabla_{\mathcal{F}} H(f)$, then it is Gâteaux differentiable (cf. Definition C.3) with the same

derivative, and therefore the first variation of H at f can be characterized as

$$\left\langle g, \frac{\delta H}{\delta f}(f) \right\rangle = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(H\left(f + \varepsilon g\right) - H\left(f\right) \right), \tag{C.2}$$

for all $g \in B_b(\mathcal{X})$.

Let $f, g \in B_b(\mathcal{X})$. For any $\lambda \in [0,1]$, set $f^\lambda \coloneqq f + \lambda g$. Then since $f^\lambda \in B_b(\mathcal{X})$, for all $\lambda \in [0,1]$, it follows by (C.2) that

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(H\left(f^\lambda + \varepsilon g\right) - H\left(f^\lambda\right) \right) = \left\langle g, \frac{\delta H}{\delta f}\left(f^\lambda\right) \right\rangle.$$

Since $f^\lambda + \varepsilon g = f^{\lambda + \varepsilon}$, it follows by the fundamental theorem of calculus that

$$H(f+g) - H(f) = \int_0^1 \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(H\left(f^{\lambda+\varepsilon}\right) - H\left(f^\lambda\right) \right) \mathrm{d}\lambda = \int_0^1 \left\langle g, \frac{\delta H}{\delta f}\left(f^\lambda\right) \right\rangle \mathrm{d}\lambda. \tag{C.3}$$

With the definition of first variation at hand, we can introduce necessary and sufficient conditions for H to have an extremum at $f \in B_b(\mathcal{X})$.

Lemma C.8 (Necessary first-order condition on $B_b(\mathcal{X})$)**.** *Let $\mathfrak{X} = \mathbb{R}$. Suppose that $H : B_b(\mathcal{X}) \to \mathbb{R}$ admits first variation at f. If H has an extremum at f, then it holds that*

$$\frac{\delta H}{\delta f}(f) = 0.$$

Proof. For a proof, see [1, Proposition 2.4.22]. □

Lemma C.9 (Sufficient first-order condition on $B_b(\mathcal{X})$)**.** *Let $\mathcal{U} \subset B_b(\mathcal{X})$ be non-empty and convex. Suppose that $H : \mathcal{U} \to \mathbb{R}$ admits first variation on $\mathcal{U}$ and is convex in the sense that, for all $\lambda \in [0,1]$, and all $f, g \in \mathcal{U}$, it holds that $H\left((1-\lambda)f + \lambda g\right) \leq (1-\lambda)H(f) + \lambda H(g)$. If $\frac{\delta H}{\delta f}(f^*) = 0$, for some $f^* \in \mathcal{U}$, then f^* is a global minimum of H.*

Remark C.10. An analogous result can be identically proved for concave functions and global maxima, so we will give the proof only for the convex case.

Proof. Since H is convex and admits first variation, following the argument in [21, Lemma 4.1], it can be shown that for any $f, g \in \mathcal{U}$

$$H(g) \geq H(f) + \left\langle g - f, \frac{\delta H}{\delta f}(f) \right\rangle.$$

For $f = f^*$ and using the assumption that $\frac{\delta H}{\delta f}(f^*) = 0$, we get

$$H(g) \geq H(f^*),$$

for all $g \in \mathcal{U}$, i.e. f^* is a global minimum. □

Definition C.11 (Second variation of H)**.** Let $H : B_b(\mathcal{X}) \to \mathfrak{X}$ be twice Fréchet differentiable at $f \in B_b(\mathcal{X})$. If it exists, the *second variation* of H at f is the unique element $\frac{\delta^2 H}{\delta f^2}(f) \in \mathcal{L}_2\left(B_b(\mathcal{X}), \mathfrak{X}\right)$ such that, for all $g, g' \in B_b(\mathcal{X})$,

$$\int_{\mathcal{X}\times\mathcal{X}} g(x) \frac{\delta^2 H}{\delta f^2}(f)(f)(\mathrm{d}y \otimes \mathrm{d}x) g'(y) \coloneqq \nabla^2_{\mathcal{F}} H(f)(g)(g'),$$

where $\frac{\delta^2 H}{\delta f^2}(f)(f)(\mathrm{d}y \otimes \mathrm{d}x) \coloneqq \frac{\delta^2 H}{\delta f^2}(f)(\mathrm{d}x)(f)(\mathrm{d}y)$.

From Example C.5, we observe that the second variation $\frac{\delta^2 h^*}{\delta f^2} : B_b(\mathcal{X}) \to \mathcal{L}\left(B_b(\mathcal{X}), \mathcal{M}(\mathcal{X})\right)$ of h^* is given by

$$\frac{\delta^2 h^*}{\delta f^2}(f)(g)(\mathrm{d}x) = \left(g(x) - \int_{\mathcal{X}} g(z)\frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(z)}\mathrm{d}z}\mathrm{d}z\right)\frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(z)}\mathrm{d}z}\mathrm{d}x. \tag{C.4}$$

Assume that $H : B_b(\mathcal{X}) \to \mathfrak{X}$ is twice Fréchet differentiable at $f \in B_b(\mathcal{X})$. Then its first-order Fréchet derivative $\nabla_{\mathcal{F}} H(f)$ is Fréchet differentiable at f, and thus it is Gâteaux differentiable (cf. Definition C.3) with the same second-order derivative. Hence, using Definition C.7, the second variation of H at f can be characterized in terms of the first variation as

$$\int_{\mathcal{X}\times\mathcal{X}} g(x)\frac{\delta^2 H}{\delta f^2}(f)(f)(\mathrm{d}y \otimes \mathrm{d}x)g'(y) = \lim_{\varepsilon \downarrow 0}\frac{1}{\varepsilon}\left\langle g, \left(\frac{\delta H}{\delta f}(f + \varepsilon g') - \frac{\delta H}{\delta f}(f)\right)\right\rangle, \tag{C.5}$$

for all $g, g' \in B_b(\mathcal{X})$.

Let $f, g, g' \in B_b(\mathcal{X})$. For any $\lambda \in [0,1]$, set $f^\lambda \coloneqq f + \lambda g'$. Then since $f^\lambda \in B_b(\mathcal{X})$, for all $\lambda \in [0,1]$, it follows by (C.5) that

$$\lim_{\varepsilon \downarrow 0}\frac{1}{\varepsilon}\left\langle g, \left(\frac{\delta H}{\delta f}\left(f^\lambda + \varepsilon g'\right) - \frac{\delta H}{\delta f}\left(f^\lambda\right)\right)\right\rangle = \int_{\mathcal{X}\times\mathcal{X}} g(x)\frac{\delta^2 H}{\delta f^2}\left(f^\lambda\right)\left(f^\lambda\right)(\mathrm{d}y \otimes \mathrm{d}x)g'(y).$$

Since $f^\lambda + \varepsilon g' = f^{\lambda+\varepsilon}$, it follows that

$$\begin{aligned}
\left\langle g, \left(\frac{\delta H}{\delta f}(f + g') - \frac{\delta H}{\delta f}(f)\right)\right\rangle &= \left\langle g, \int_0^1 \lim_{\varepsilon \downarrow 0}\frac{1}{\varepsilon}\left(\frac{\delta H}{\delta f}\left(f^{\lambda+\varepsilon}\right) - \frac{\delta H}{\delta f}\left(f^\lambda\right)\right)\mathrm{d}\lambda\right\rangle \\
&= \int_0^1 \lim_{\varepsilon \downarrow 0}\frac{1}{\varepsilon}\left\langle g, \left(\frac{\delta H}{\delta f}\left(f^{\lambda+\varepsilon}\right) - \frac{\delta H}{\delta f}\left(f^\lambda\right)\right)\right\rangle \mathrm{d}\lambda \\
&= \int_0^1 \int_{\mathcal{X}\times\mathcal{X}} g(x)\frac{\delta^2 H}{\delta f^2}\left(f^\lambda\right)\left(f^\lambda\right)(\mathrm{d}y \otimes \mathrm{d}x)g'(y)\mathrm{d}\lambda,
\end{aligned} \tag{C.6}$$

where the first equality follows from the fundamental theorem of calculus and the second equality from Fubini's theorem and the dominated convergence theorem.

Example C.12 (Verification of Assumption 2.13 for entropy)**.** Let $f, g, \psi \in V$. Then, for $f \in V$, denote $\varphi(f)(\mathrm{d}x) \coloneqq \frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(z)}\mathrm{d}z}\mathrm{d}x \in \mathcal{P}(\mathcal{X})$. Thus, for h being the entropy, its second variation (C.4) can be written as

$$\frac{\delta^2 h^*}{\delta f^2}(f)(g)(\mathrm{d}x) = \left(g(x) - \int_{\mathcal{X}} g(z)\varphi(f)(\mathrm{d}z)\right)\varphi(f)(\mathrm{d}x),$$

and since $\varphi(f)$ is absolutely continuous with respect to Lebesgue measure on $\mathcal{X}$, it follows that

$$\begin{aligned}
&\int_{\mathcal{X}} |\psi(x)|\left|\frac{\delta^2 h^*}{\delta f^2}(f')(g') - \frac{\delta^2 h^*}{\delta f^2}(f)(g)\right|(\mathrm{d}x) \\
&= \int_{\mathcal{X}} |\psi(x)|\left|\left(g'(x) - \int_{\mathcal{X}} g'(z)\varphi(f')(z)\mathrm{d}z\right)\varphi(f')(x) - \left(g(x) - \int_{\mathcal{X}} g(z)\varphi(f)(z)\mathrm{d}z\right)\varphi(f)(x)\right|\mathrm{d}x \\
&= \int_{\mathcal{X}} |\psi(x)|\Bigg|g'(x)\varphi(f')(x) - \varphi(f')(x)\int_{\mathcal{X}} g'(z)\varphi(f')(z)\mathrm{d}z \\
&\qquad - g(x)\varphi(f)(x) + \varphi(f)(x)\int_{\mathcal{X}} g(z)\varphi(f)(z)\mathrm{d}z\Bigg|\mathrm{d}x \leq I_1 + I_2,
\end{aligned}$$

where

$$I_1 = \int_{\mathcal{X}} |\psi(x)| \, |g'(x)\varphi(f')(x) - g(x)\varphi(f)(x)| \, \mathrm{d}x,$$

$$I_2 = \int_{\mathcal{X}} |\psi(x)| \left| \varphi(f)(x) \int_{\mathcal{X}} g(z)\varphi(f)(z)\mathrm{d}z - \varphi(f')(x) \int_{\mathcal{X}} g'(z)\varphi(f')(z)\mathrm{d}z \right| \mathrm{d}x,$$

and the last inequality follows from triangle inequality. For I_1, we observe that

$$\begin{aligned} I_1 &= \int_{\mathcal{X}} |\psi(x)| \, |g'(x)\varphi(f')(x) - g'(x)\varphi(f)(x) + g'(x)\varphi(f)(x) - g(x)\varphi(f)(x)| \, \mathrm{d}x \\ &\leq \int_{\mathcal{X}} |\psi(x)| \, |g'(x)| \, |\varphi(f')(x) - \varphi(f)(x)| \, \mathrm{d}x + \int_{\mathcal{X}} |\psi(x)| \, |\varphi(f)(x)| \, |g'(x) - g(x)| \, \mathrm{d}x. \end{aligned}$$

Since $f, g' \in V \subseteq B_b(\mathcal{X})$, there exist $C_{g'}, C_f > 0$ such that $|g'(x)| \leq C_{g'}$ and $|\varphi(f)(x)| \leq C_f$, for all $x \in \mathcal{X}$. Since f, f' are bounded on $\mathcal{X}$, following the argument in [25, Lemma A.2], we deduce that $f \mapsto \varphi(f)$ is Lipschitz, i.e., there exists $L_\varphi > 0$ such that, for all $x \in \mathcal{X}$,

$$|\varphi(f')(x) - \varphi(f)(x)| \leq L_\varphi \, |f'(x) - f(x)| \, .$$

Hence, I_1 becomes

$$\begin{aligned} I_1 &\leq C_{g'} L_\varphi \int_{\mathcal{X}} |\psi(x)| \, |f'(x) - f(x)| \, \mathrm{d}x + C_f \int_{\mathcal{X}} |\psi(x)| \, |g'(x) - g(x)| \, \mathrm{d}x \\ &\leq \max\{C_{g'} L_\varphi, C_f\} \int_{\mathcal{X}} |\psi(x)| \, (|f'(x) - f(x)| + |g'(x) - g(x)|) \, \mathrm{d}x \\ &\leq \max\{C_{g'} L_\varphi, C_f\} \, (\|f' - f\|_\infty + \|g' - g\|_\infty) \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x. \end{aligned}$$

Similarly, for I_2, we have that

$$\begin{aligned} I_2 &= \int_{\mathcal{X}} |\psi(x)| \Bigg| \varphi(f)(x) \int_{\mathcal{X}} g(z)\varphi(f)(z)\mathrm{d}z - \varphi(f)(x) \int_{\mathcal{X}} g'(z)\varphi(f')(z)\mathrm{d}z \\ &\qquad + \varphi(f)(x) \int_{\mathcal{X}} g'(z)\varphi(f')(z)\mathrm{d}z - \varphi(f')(x) \int_{\mathcal{X}} g'(z)\varphi(f')(z)\mathrm{d}z \Bigg| \mathrm{d}x \\ &\leq \int_{\mathcal{X}} |\psi(x)||\varphi(f)(x)| \int_{\mathcal{X}} |g(z)\varphi(f)(z) - g'(z)\varphi(f')(z)| \mathrm{d}z\mathrm{d}x \\ &\qquad + \int_{\mathcal{X}} |\psi(x)||\varphi(f)(x) - \varphi(f')(x)| \int_{\mathcal{X}} |g'(z)||\varphi(f')(z)| \mathrm{d}z\mathrm{d}x \\ &\leq C_f \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x \int_{\mathcal{X}} |g(z)\varphi(f)(z) - g'(z)\varphi(f')(z)| \mathrm{d}z + C_{g'} C_{f'} L_\varphi \int_{\mathcal{X}} |\psi(x)||f(x) - f'(x)| \mathrm{d}x \\ &\leq C_f \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x \int_{\mathcal{X}} |g(z)\varphi(f)(z) - g'(z)\varphi(f')(z)| \mathrm{d}z + C_{g'} C_{f'} L_\varphi \|f' - f\|_\infty \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x. \end{aligned}$$

We observe that

$$
\begin{aligned}
&C_f \int_{\mathcal{X}} |\psi(x)| \int_{\mathcal{X}} |g(z)\varphi(f)(z) - g'(z)\varphi(f')(z)| \mathrm{d}z \mathrm{d}x \\
&\quad = C_f \int_{\mathcal{X}} |\psi(x)| \int_{\mathcal{X}} |g(z)\varphi(f)(z) - g'(z)\varphi(f)(z) + g'(z)\varphi(f)(z) - g'(z)\varphi(f')(z)| \mathrm{d}z \mathrm{d}x \\
&\leq C_f \int_{\mathcal{X}} |\psi(x)| \int_{\mathcal{X}} |g(z) - g'(z)||\varphi(f)(z)| \mathrm{d}z \mathrm{d}x + C_f \int_{\mathcal{X}} |\psi(x)| \int_{\mathcal{X}} |g'(z)|\varphi(f)(z) - \varphi(f')(z)| \mathrm{d}z \mathrm{d}x \\
&\qquad \leq C_f^2 \|g' - g\|_\infty \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x + C_f C_{g'} L_\varphi \|f' - f\|_\infty \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x.
\end{aligned}
$$

Hence, we get that

$$
\begin{aligned}
&I_1 + I_2 \leq \max\{C_{g'} L_\varphi, C_f\} \left(\|f' - f\|_\infty + \|g' - g\|_\infty\right) \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x \\
&+ C_f^2 \|g' - g\|_\infty \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x + C_f C_{g'} L_\varphi \|f' - f\|_\infty \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x + C_{g'} C_{f'} L_\varphi \|f' - f\|_\infty \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x \\
&\qquad = \max\{C_{g'} L_\varphi, C_f\} \left(\|f' - f\|_\infty + \|g' - g\|_\infty\right) \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x \\
&\qquad + C_f^2 \|g' - g\|_\infty \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x + \left(C_f + C_{f'}\right) C_{g'} L_\varphi \|f' - f\|_\infty \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x \\
&\leq \left(\max\{C_{g'} L_\varphi, C_f\} + \max\{C_f^2, \left(C_f + C_{f'}\right) C_{g'} L_\varphi\}\right) \left(\|f' - f\|_\infty + \|g' - g\|_\infty\right) \int_{\mathcal{X}} |\psi(x)| \mathrm{d}x.
\end{aligned}
$$

Setting $L_{h^*} \coloneqq \max\{C_{g'} L_\varphi, C_f\} + \max\{C_f^2, \left(C_f + C_{f'}\right) C_{g'} L_\varphi\}$ finishes the verification.

Appendix D. Technical results on duality

In this section, we state and prove some technical results which are central to the proof technique via dual Bregman divergence that we developed in Subsection 2.2.

Proposition D.1. *Let Assumption 1.2, 2.9 and 2.10 hold. Let $h^* : V \to \mathbb{R}$ be the convex conjugate of h. Then, the following are equivalent:*

1. *The supremum of $\mathcal{E} \ni m \mapsto \langle g^*, m\rangle - h(m) \in \mathbb{R}$ is attained at $m = m^*$,*
2. *$g^*(x) - \frac{\delta h}{\delta m}(m^*, x) =$ constant, for all $x \in \mathcal{X}$ Lebesgue a.e.,*
3. *The supremum of $V \ni g \mapsto \langle g, m^*\rangle - h^*(g) \in \mathbb{R}$ is attained at $g = g^*$,*
4. *$m^* = \frac{\delta h^*}{\delta g}(g^*)$.*

Proof. (1) $\implies$ (2): Suppose that (1) holds. Then the supremum of $m \mapsto \langle g^*, m\rangle - h(m)$ is attained at the maximizer $m^* = \operatorname{argmax}_{m \in \mathcal{E}} \{\langle g^*, m\rangle - h(m)\}$. Hence, by [21, Proposition 2.5], m^* satisfies the first-order condition

$$
g^*(z) - \frac{\delta h}{\delta m}(m^*, z) = \text{constant},
$$

for all $z \in \mathcal{X}$ Lebesgue a.e.

(2) $\implies$ (1): Suppose that (2) holds. Observe that the map $m \mapsto \langle g^*, m\rangle - h(m)$ is strictly concave due to the strict convexity of h and the linearity of $m \mapsto \langle g^*, m\rangle$. Then by the converse of [21, Proposition 2.5], it follows that m^* is the maximizer of the map $\mathcal{E} \ni m \mapsto \langle g^*, m\rangle - h(m) \in \mathbb{R}$, and so (1) holds.

(3) $\implies$ (4): Suppose that (3) holds. Then the supremum in $g \mapsto \langle g, m^*\rangle - h^*(g)$ is attained at a maximizer $g^* \in \operatorname{argmax}_{g \in V} \{\langle g, m^*\rangle - h^*(g)\}$. Hence, by Lemma C.8, it

follows that g^* satisfies the first-order condition

$$m^* = \frac{\delta h}{\delta g}(g^*).$$

(4) $\implies$ (3): Suppose that (4) holds. Observe that V is a convex non-empty subset of $B_b(\mathcal{X})$ and the map $g \mapsto \langle g, m^* \rangle - h^*(g)$ is concave due to the convexity of h^* and the linearity of $g \mapsto \langle g, m^* \rangle$. Hence, by Lemma C.9, it follows that g^* is a maximizer of the map $V \ni g \mapsto \langle g, m^* \rangle - h^*(g) \in \mathbb{R}$, and so (3) holds.

(1) $\implies$ (3): Suppose that (1) holds. Then, by Definition 2.8, we have that $h^*(g) = \langle g, m^* \rangle - h(m^*)$, and equivalently $h(m^*) = \langle g, m^* \rangle - h^*(g)$. Clearly, $\mathcal{P}(\mathcal{X})$ is convex and $(\mathcal{P}(\mathcal{X}), \mathrm{TV})$, where TV is the total variation distance, is Hausdorff since it is a metric space, hence we can apply the Fenchel-Moreau theorem [43, Theorem 2.3.3] to conclude that $h^{**} = h$, i.e., $h(m^*) = \sup_{g \in V} \{\langle g, m^* \rangle - h^*(g)\}$. Therefore, $h(m^*)$ is the supremum of $g \mapsto \langle g, m^* \rangle - h^*(g)$ attained at $g = g^*$.

(3) $\implies$ (1): Suppose (3) holds. Then $h^{**}(m^*) = \langle g^*, m^* \rangle - h^*(g^*)$, or equivalently $h^*(g^*) = \langle g^*, m^* \rangle - h^{**}(m^*)$. Again, by the Fenchel-Moreau theorem [43, Theorem 2.3.3], $h^{**}(m) = h(m)$, for all $m \in \mathcal{E}$, and hence $h^*(g^*) = \langle g^*, m^* \rangle - h(m^*)$. Hence, by Definition 2.8, the supremum of $m \mapsto \langle g^*, m \rangle - h(m)$ is realized at $m = m^*$. □

Lemma D.2. *Let Assumption 1.2, 2.9 and 2.10 hold. Let $h^* : V \to \mathbb{R}$ be the convex conjugate of h. Fix $f, g \in V$ and $\mu, \mu' \in \mathcal{E}$. If $f(z) = \frac{\delta h}{\delta m}(\mu, z)$ and $g(z) = \frac{\delta h}{\delta m}(\mu', z)$, for all $z \in \mathcal{X}$ Lebesgue a.e., up to an additive constant, then*

$$D_{h^*}(f, g) = D_h(\mu', \mu).$$

Proof. By Definition 2.12, we have that

$$\begin{aligned} D_{h^*}(f,g) &= h^*(f) - h^*(g) - \int_{\mathcal{X}} (f(z) - g(z)) \frac{\delta h^*}{\delta g}(g)(\mathrm{d}z) \\ &= \langle f, \mu \rangle - h(\mu) - \langle g, \mu' \rangle + h(\mu') - \int_{\mathcal{X}} (f(z) - g(z)) \frac{\delta h^*}{\delta g}(g)(\mathrm{d}z) \\ &= h(\mu') - h(\mu) + \int_{\mathcal{X}} \frac{\delta h}{\delta m}(\mu, z)\mu(\mathrm{d}z) - \int_{\mathcal{X}} \frac{\delta h}{\delta m}(\mu', z)\mu'(\mathrm{d}z) - \int_{\mathcal{X}} \left(\frac{\delta h}{\delta m}(\mu, z) - \frac{\delta h}{\delta m}(\mu', z) \right) \mu'(\mathrm{d}z) \\ &= h(\mu') - h(\mu) - \int_{\mathcal{X}} \frac{\delta h}{\delta m}(\mu, z)(\mu' - \mu)(\mathrm{d}z) = D_h(\mu', \mu), \end{aligned}$$

where the second and third equalities follow from Lemma D.1 and Corollary 2.11, while the last equality follows from the definition of the Bregman divergence. □

Lemma D.3. *Consider* (1.6) *and* (1.7). *Let Assumption 1.2, 2.9 and 2.10 hold. Let $h^* : V \to \mathbb{R}$ be the convex conjugate of h. For each $n \geq 0$, fix $f^n, g^n \in V$, $\nu^n \in \mathcal{C}$ and $\mu^n \in \mathcal{D}$. If $f^n = \frac{\delta h}{\delta \nu}(\nu^n, \cdot)$ and $g^n = \frac{\delta h}{\delta \mu}(\mu^n, \cdot)$, then, for any $n \geq 0$, we have that*

$$\begin{aligned} D_h(\nu^{n+1}, \nu^n) = D_{h^*}(f^n, f^{n+1}), \quad D_h(\nu^n, \nu^{n+1}) = D_{h^*}(f^{n+1}, f^n), \\ D_h(\mu^{n+1}, \mu^n) = D_{h^*}(g^n, g^{n+1}), \quad D_h(\mu^n, \mu^{n+1}) = D_{h^*}(g^{n+1}, g^n). \end{aligned}$$

Proof. First, observe that due to Assumption 2.10, the pairs (ν^{n+1}, μ^{n+1}) in (1.6) and (1.7) are unique. We will only present the proof for (1.6) since the argument for (1.7) is

identical. The updates in (1.6) can be equivalently written as

$$\text{(D.1)}\quad \nu^{n+1} = \operatorname*{argmin}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu, \nu^n) \right\}$$

$$= \operatorname*{argmin}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + h(\nu) - h(\nu^n) - \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu^n, x)(\nu - \nu^n)(\mathrm{d}x) \right\}$$

$$= \operatorname*{argmin}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} \left(\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) - \frac{\delta h}{\delta \nu}(\nu^n, x) \right) (\nu - \nu^n)(\mathrm{d}x) + h(\nu) \right\}$$

$$= \operatorname*{argmax}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu^n, x) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) \right) (\nu - \nu^n)(\mathrm{d}x) - h(\nu) \right\}$$

$$= \operatorname*{argmax}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu^n, x) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) \right) \nu(\mathrm{d}x) - h(\nu) \right\},$$

and

$$\text{(D.2)}\quad \mu^{n+1} = \operatorname*{argmax}_{\mu\in\mathcal{D}} \left\{ \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu, \mu^n) \right\}$$

$$= \operatorname*{argmax}_{\mu\in\mathcal{D}} \left\{ \int_{\mathcal{X}} \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - h(\mu) + h(\mu^n) + \int_{\mathcal{X}} \frac{\delta h}{\delta \mu}(\mu^n, y)(\mu - \mu^n)(\mathrm{d}y) \right\}$$

$$= \operatorname*{argmax}_{\mu\in\mathcal{D}} \left\{ \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \mu}(\mu^n, y) + \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y) \right) (\mu - \mu^n)(\mathrm{d}y) - h(\mu) \right\}$$

$$= \operatorname*{argmax}_{\mu\in\mathcal{D}} \left\{ \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \mu}(\mu^n, y) + \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y) \right) \mu(\mathrm{d}y) - h(\mu) \right\}.$$

Using the notation $f^n = \frac{\delta h}{\delta \nu}(\nu^n, \cdot)$ and $g^n = \frac{\delta h}{\delta \mu}(\mu^n, \cdot)$, for each $n \geq 0$, the first-order conditions for (1.6) can be equivalently written as

$$\text{(D.3)}\qquad f^{n+1}(x) - f^n(x) = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x),$$

$$\text{(D.4)}\qquad g^{n+1}(y) - g^n(y) = \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y),$$

for all $(x, y) \in \mathcal{X} \times \mathcal{X}$ Lebesgue a.e. Then using (2.2), (D.1) becomes

$$\text{(D.5)}\quad \nu^{n+1} = \operatorname*{argmax}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} \left(f^n(x) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) \right) \nu(\mathrm{d}x) - h(\nu) \right\}$$

$$= \operatorname*{argmax}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} f^{n+1}(x)\nu(\mathrm{d}x) - h(\nu) \right\} = \frac{\delta h^*}{\delta f}(f^{n+1}),$$

for all $n \geq 0$. Similarly, from (D.2), we have that

$$\text{(D.6)}\qquad \mu^{n+1} = \frac{\delta h^*}{\delta f}(g^{n+1}),$$

for all $n \geq 0$. The conclusion follows directly from Lemma D.2. □

Appendix E. Proof of convergence for the MDA implicit scheme

In this section, we prove that an implicit Euler discretization of the Fisher-Rao flows studied in [26] yields a linear convergence rate $\mathcal{O}(1/N)$, which matches the result in continuous-time under the same assumption of convexity-concavity of F (see [26, Theorem

2.3]). However, a major weakness of this implicit game is that it is not implementable in practice as opposed to (1.6) and (1.7).

For a given stepsize $\tau > 0$, and fixed initial pair of strategies $(\nu_0, \mu_0) \in \mathcal{C} \times \mathcal{D}$, for $n \geq 0$, the *implicit* MDA iterative scheme is defined by

$$\text{(E.1)} \qquad \begin{cases} \nu^{n+1} \in \operatorname{argmin}_{\nu \in \mathcal{C}} \{ \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu, \nu^n) \}, \\ \mu^{n+1} \in \operatorname{argmax}_{\mu \in \mathcal{D}} \{ \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu, \mu^n) \}, \end{cases}$$

Theorem E.1 (Convergence of the implicit MDA scheme (E.1))**.** *Let (ν^*, μ^*) be an MNE of (1.1) and (ν^0, μ^0) be such that $\max_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \max_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) < \infty$. Let Assumption 1.1, 1.2 and 1.5 hold. Suppose that $\tau L \leq 1$, where $L \coloneqq \max\{L_\nu, L_\mu\}$. Then, we have*

$$\mathrm{NI}\left(\frac{1}{N} \sum_{n=0}^{N-1} \nu^{n+1}, \frac{1}{N} \sum_{n=0}^{N-1} \mu^{n+1} \right) \leq \frac{1}{N\tau} \left(\max_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \max_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) \right).$$

Proof. Since $\nu \mapsto \tau \int \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu - \nu^n)(\mathrm{d}x)$ is convex, applying Lemma 2.5 with $\bar{\nu} = \nu^{n+1}$ and $\mu = \nu^n$ implies that, for any $\nu \in \mathcal{C}$, we have

$$\begin{aligned} \tau \int \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu - \nu^n)(\mathrm{d}x) + D_h(\nu, \nu^n) \geq \tau \int \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) \\ + D_h(\nu^{n+1}, \nu^n) + D_h(\nu, \nu^{n+1}), \end{aligned}$$

or, equivalently,

$$\text{(E.2)} \quad \begin{aligned} -\tau \int \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu - \nu^n)(\mathrm{d}x) - D_h(\nu, \nu^n) \leq -\tau \int \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) \\ - D_h(\nu^{n+1}, \nu^n) - D_h(\nu, \nu^{n+1}). \end{aligned}$$

Similarly, since $\mu \mapsto -\tau \int \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y)$ is convex, applying Lemma 2.5 with $\bar{\nu} = \mu^{n+1}$ and $\mu = \mu^n$ implies that, for any $\mu \in \mathcal{D}$, we have

$$\text{(E.3)} \quad \begin{aligned} \tau \int \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - D_h(\mu, \mu^n) \leq \tau \int \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) \\ - D_h(\mu^{n+1}, \mu^n) - D_h(\mu, \mu^{n+1}). \end{aligned}$$

Using the convexity of $\nu \mapsto F(\nu, \mu)$ in (E.2), i.e., (1.4) with $\nu = \nu^n$ and $\mu = \mu^{n+1}$, we have that

$$\text{(E.4)} \quad \begin{aligned} F(\nu^n, \mu^{n+1}) - F(\nu, \mu^{n+1}) - \frac{1}{\tau} D_h(\nu, \nu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^n - \nu^{n+1})(\mathrm{d}x) \\ - \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}). \end{aligned}$$

From L_ν-relative smoothness and the fact that $\tau L \leq 1$, it follows that

$$\text{(E.5)} \quad \begin{aligned} F(\nu^{n+1}, \mu^{n+1}) &\leq F(\nu^n, \mu^{n+1}) + \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) + L_\nu D_h(\nu^{n+1}, \nu^n) \\ &\leq F(\nu^n, \mu^{n+1}) + \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n). \end{aligned}$$

Hence, combining (E.4) with (E.5), we obtain that

(E.6)
$$F(\nu^n,\mu^{n+1})-F(\nu,\mu^{n+1})-\frac{1}{\tau}D_h(\nu,\nu^n)\le F(\nu^n,\mu^{n+1})-F(\nu^{n+1},\mu^{n+1})-\frac{1}{\tau}D_h(\nu,\nu^{n+1}).$$

Similarly, using concavity of $\mu\mapsto F(\nu,\mu)$ in (E.3), i.e., (1.5) with $\nu=\nu^{n+1}$ and $\mu=\mu^n$, we have that

$$F(\nu^{n+1},\mu)-F(\nu^{n+1},\mu^n)-\frac{1}{\tau}D_h(\mu,\mu^n)\le\int_{\mathcal{X}}\frac{\delta F}{\delta\mu}(\nu^{n+1},\mu^n,y)(\mu^{n+1}-\mu^n)(\mathrm{d}y)-\frac{1}{\tau}D_h(\mu^{n+1},\mu^n)-\frac{1}{\tau}D_h(\mu,\mu^{n+1}). \tag{E.7}$$

From L_μ-relative smoothness and the fact that $\tau L\le 1$, it follows that

(E.8)
$$\begin{aligned}F(\nu^{n+1},\mu^{n+1})&\ge F(\nu^{n+1},\mu^n)+\int_{\mathcal{X}}\frac{\delta F}{\delta\mu}(\nu^{n+1},\mu^n,y)(\mu^{n+1}-\mu^n)(\mathrm{d}y)-L_\mu D_h(\mu^{n+1},\mu^n)\\&\ge F(\nu^{n+1},\mu^n)+\int_{\mathcal{X}}\frac{\delta F}{\delta\mu}(\nu^{n+1},\mu^n,y)(\mu^{n+1}-\mu^n)(\mathrm{d}y)-\frac{1}{\tau}D_h(\mu^{n+1},\mu^n).\end{aligned}$$

Hence, combining (E.7) with (E.8), we obtain that

(E.9)
$$F(\nu^{n+1},\mu)-F(\nu^{n+1},\mu^n)-\frac{1}{\tau}D_h(\mu,\mu^n)\le F(\nu^{n+1},\mu^{n+1})-F(\nu^{n+1},\mu^n)-\frac{1}{\tau}D_h(\mu,\mu^{n+1}).$$

Adding inequalities (E.6) and (E.9) implies that

$$\begin{aligned}F(\nu^{n+1},\mu)-F(\nu,\mu^{n+1})&\le F(\nu^{n+1},\mu^{n+1})-F(\nu^{n+1},\mu^{n+1})\\&\quad+\frac{1}{\tau}D_h(\nu,\nu^n)+\frac{1}{\tau}D_h(\mu,\mu^n)-\frac{1}{\tau}D_h(\nu,\nu^{n+1})-\frac{1}{\tau}D_h(\mu,\mu^{n+1}).\end{aligned}$$

Summing the previous inequality over $n=0,1,...,N-1$, dividing by N, applying Jensen's inequality and taking maximum over (ν,μ) leads to

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1},\frac{1}{N}\sum_{n=0}^{N-1}\mu^{n+1}\right)\le\frac{1}{N\tau}\left(\max_{\nu\in\mathcal{C}}D_h(\nu,\nu^0)+\max_{\mu\in\mathcal{D}}D_h(\mu,\mu^0)\right),$$

where the last inequality follows since $D_h(\nu,\nu^N)+D_h(\mu,\mu^N)\ge 0$, for all $(\nu,\mu)\in\mathcal{C}\times\mathcal{D}$. □

School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK, and Maxwell Institute for Mathematical Sciences, Edinburgh, UK
Email address: rl2029@hw.ac.uk

School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK, and Maxwell Institute for Mathematical Sciences, Edinburgh, UK
Email address: m.majka@hw.ac.uk

School of Mathematics, University of Edinburgh, UK, and The Alan Turing Institute, UK and Simtopia, UK
Email address: l.szpruch@ed.ac.uk