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

Razvan-Andrei Lascu
RIKEN AIP
razvan-andrei.lascu@riken.jp

Mateusz B. Majka
Heriot-Watt University
m.majka@hw.ac.uk

Łukasz Szpruch
University of Edinburgh
l.szpruch@ed.ac.uk

Abstract

We study two variants of the mirror descent-ascent (MDA) algorithm for solving min-max problems on the space of measures: simultaneous and alternating. 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 establish non-asymptotic convergence rates to mixed Nash equilibria, measured in the Nikaidô-Isoda error, proving an $\mathcal{O}(N^{-1/2})$ rate for simultaneous MDA and an improved $\mathcal{O}(N^{-2/3})$ rate for alternating MDA. The main technical contribution is an infinite-dimensional dual space analysis that relates Bregman divergences on measures to dual Bregman divergences on spaces of bounded continuous functions, allowing us to control asymmetric commutator terms created by alternating updates. The results substantially generalize prior analyses restricted to bilinear objectives and also apply to nonlinear convex-concave problems on measure spaces, thereby providing a unified theoretical foundation for MDA in mean-field min-max optimization.

1 Introduction

Numerous tasks in machine learning can be framed as optimization problems for functions defined on the space of probability measures. For instance, in supervised learning, [19, 38, 48] 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 network parameters. This observation has become a central tool for analysing the convergence of training algorithms for mean-field NNs; see, e.g., [26, 18, 44, 52]. The same mean-field optimization paradigm has also been extended to min-max settings, where the objective is to compute mixed Nash equilibria (MNEs) of games over spaces of probability measures; see, e.g., [25, 21, 54, 37, 23, 31, 13].

In this work, we study the convergence of mirror descent-ascent (MDA) algorithms for convex-concave min-max games over probability measures. The central question we address is how the order of play affects convergence in the infinite-dimensional mean-field setting. In games, learning dynamics depend crucially on whether players update simultaneously, with both players moving from the same current state, or alternately, with one player updating after observing the other player's move. To our knowledge, existing discrete-time convergence guarantees for mean-field min-max games focus on simultaneous updates; see, e.g., [25, 54]. We provide a rigorous comparison between simultaneous and alternating MDA and prove that alternating play can achieve a strictly faster convergence rate.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The mean-field perspective offers a tractable route to analyzing particle algorithms whose finite-particle implementations evolve in high-dimensional spaces. Rather than tracking particles directly, one studies the infinite-particle limit and derives guarantees at the level of probability measures. This viewpoint makes it possible to isolate the effect of algorithmic choices, such as simultaneous versus alternating updates, on the evolution of the players' mixed strategies. Our analysis shows that this distinction is not merely algorithmic. In the alternating scheme, the second player uses additional information from the first player's updated strategy, and this changes the structure of the error terms. Controlling these terms requires a dual-space argument for Bregman divergences on measures and leads to an improved rate for alternating MDA.

Our results provide a theoretical justification for alternating updates in mean-field min-max optimization. In Section 1.4, beyond GANs, we discuss applications to learning MNEs in preference models related to Reinforcement Learning from Human Feedback (RLHF). We also show that the analysis is not restricted to bilinear games. It applies to nonlinear convex-concave min-max problems over measures, including adversarial training of mean-field neural networks.

1.1 Notation and setup

For any $\mathcal{X} \subseteq \mathbb{R}^d$, let $\mathcal{P}(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$. We endow $\mathcal{P}(\mathcal{X})$ with the narrow convergence topology, i.e., convergence against continuous bounded functions in $C_b(\mathcal{X})$. 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 nonempty and convex. We consider a convex-concave (cf. Assumption 2.2) payoff function $F : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ and the associated min-max game

$$\min_{\nu \in \mathcal{C}} \max_{\mu \in \mathcal{D}} F(\nu, \mu). \tag{1}$$

We are interested in finding *mixed Nash equilibria* (MNEs) for game (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

$$F(\nu^*, \mu) \leq F(\nu^*, \mu^*) \leq F(\nu, \mu^*). \tag{2}$$

Throughout, we assume that there exists at least one MNE for game (1).[1] In min-max games, the distance between a pair of strategies (ν, μ) and an MNE is typically measured using the Nikaidô-Isoda (NI) error [43], 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 (2) it follows that $\mathrm{NI}(\nu, \mu) = 0$ if and only if (ν, μ) is an MNE. Our notions of MNE and NI error agree with the classical definitions in the bilinear setting, where $F(\nu, \mu) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y) \nu(\mathrm{d}x) \mu(\mathrm{d}y)$, for some $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We therefore adopt the same terminology for the more general class of convex-concave games over measures.

Following [15, Definition 5.43] and [49, Definition 7.12], we introduce the notion of differentiability on the space of measures that we utilize throughout the paper.

Definition 1.1. *Let $\mathcal{X} \subseteq \mathbb{R}^d$, and let $\mathcal{K} \subseteq \mathcal{P}(\mathcal{X})$ be convex. A functional $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is in $\mathfrak{C}^1(\mathcal{K})$ if there exists a continuous bounded function $\frac{\delta F}{\delta \nu} : \mathcal{K} \times \mathcal{X} \to \mathbb{R}$ satisfying*

$$\lim_{\varepsilon \to 0} \frac{F(\nu + \varepsilon(\nu' - \nu)) - F(\nu)}{\varepsilon} = \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, x)(\nu' - \nu)(\mathrm{d}x) \quad \text{for all } \nu, \nu' \in \mathcal{K}.$$

We call $\frac{\delta F}{\delta \nu}$ the flat derivative (or first variation) of F, normalized such that $\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, x) \nu(\mathrm{d}x) = 0$.

We now introduce the notion of Bregman divergence and its commutator on the space of probability measures using the flat derivative (Definition 1.1), following [6, Definition 3], who defined it via directional derivatives. We assume flat differentiability so that, for each fixed ν, the first-order term is linear in the measure increment $\nu' - \nu$, a property not guaranteed by directional differentiability alone. Let $\mathcal{E} \subseteq \mathcal{P}(\mathcal{X})$ be nonempty and convex.

[1] If F is continuous and convex-concave, and both $\mathcal{C}$ and $\mathcal{D}$ are narrowly compact, then an MNE of (1) exists by Sion's minimax theorem [51]. For the integral payoff $F(\nu, \mu) = \iint_{\mathcal{X} \times \mathcal{X}} f(x, y) \nu(\mathrm{d}x) \mu(\mathrm{d}y)$, an MNE exists if $\mathcal{C}$ and $\mathcal{D}$ are narrowly compact and $f \in C_b(\mathcal{X} \times \mathcal{X})$. If $\mathcal{X}$ is compact, it suffices to assume that f is continuous.

Definition 1.2 (Bregman divergence and commutator)**.** *Assume that $h \in \mathfrak{C}^1(\mathcal{E})$. The h-Bregman divergence and its commutator are the maps $D_h, \mathrm{Com}_h : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$, respectively, given by*

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

Bregman divergence plays a key role in the definition of mirror descent/ascent algorithms (see Section 1.2). Its use as the update penalty allows the algorithm to adapt naturally to the geometry of the underlying space. Since D_h is generally asymmetric, the role of Com_h is to measure this asymmetry.

1.2 Simultaneous and alternating MDA

With the definition of the Bregman divergence in place, we now present the simultaneous and alternating MDA algorithms. 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 *alternating* MDA algorithms are respectively defined in Algorithm 1 and Algorithm 2. The distinction between the two algorithms is important for the

Algorithm 1: SIMULTANEOUS MDA

Input: Objective function F, initial measures (ν^0, μ^0), stepsize $\tau > 0$
for $n = 0, 1, \ldots, N-1$ **do**
$\nu^{n+1} = \operatorname{argmin}_{\nu \in \mathcal{C}} \{\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu, \nu^n)\}$,
$\mu^{n+1} = \operatorname{argmax}_{\mu \in \mathcal{D}} \{\int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu, \mu^n)\}$
end
Output: $\left(\frac{1}{N}\sum_{n=0}^{N-1} \nu^n, \frac{1}{N}\sum_{n=0}^{N-1} \mu^n\right)$

Algorithm 2: ALTERNATING MDA

Input: Objective function F, initial measures (ν^0, μ^0), stepsize $\tau > 0$
for $n = 0, 1, \ldots, N-1$ **do**
$\nu^{n+1} = \operatorname{argmin}_{\nu \in \mathcal{C}} \{\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu, \nu^n)\}$,
$\mu^{n+1} = \operatorname{argmax}_{\mu \in \mathcal{D}} \{\int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu, \mu^n)\}$
end
Output: $\left(\frac{1}{N}\sum_{n=0}^{N-1} \nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1} \mu^n\right)$

convergence rate. In Algorithm 1, both players evaluate their updates at the same old pair (ν^n, μ^n), so the analysis leaves a first-order interaction error that leads to the $\mathcal{O}(N^{-1/2})$ rate. In Algorithm 2, the second player observes the first player's updated move ν^{n+1}. This extra information creates an asymmetric term, but once that term is controlled, the remaining error is of higher order. As we will discuss in the sequel, the proof of Theorem 3.5 shows that this asymmetric term can be rewritten as a Bregman commutator, transported to the dual function space, and bounded cubically by τ^3 using Assumption 3.4. This is the mechanism behind the improved $\mathcal{O}(N^{-2/3})$ rate of the alternating MDA, compared to the simultaneous MDA.

Remark 1.3 (Entropy MDA and the Fisher–Rao gradient flow)**.** *When h is the relative entropy, we can view Algorithms 1 and 2 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 [32] (cf. also [35] for single-agent convex optimization).*

Remark 1.4 (Connection to the continuous-time Bregman gradient flow)**.** *In Appendix H, we provide an implicit MDA algorithm that achieves a convergence rate of $\mathcal{O}(N^{-1})$, which matches the rate $\mathcal{O}(t^{-1})$ of the continuous-time gradient flow obtained by letting $\tau \to 0$. However, this scheme is not*

implementable in practice, as it requires each player to know the next move of their opponent at every iteration. Consequently, while it provides a useful theoretical benchmark, practical algorithms rely on explicit (simultaneous or alternating) mirror schemes.

1.3 Our contribution

We provide a theoretical contribution to the convergence analysis of MDA for mean-field min-max problems. We establish non-asymptotic rates for simultaneous and alternating MDA on spaces of probability measures under convexity-concavity and relative smoothness assumptions. Our contribution can be summarized as follows:

- From the perspective of measure space MD, our work extends the single-player results of [6] to two-player min-max games. This extension is nontrivial because, unlike in convex minimization, the objective is not monotonically decreasing along the iterates. Consequently, the analysis must be carried out in terms of the Nikaidô-Isoda error and requires new proof techniques based on convex conjugates and Bregman divergences on dual function spaces.
- From the perspective of mean-field min-max optimization, our work generalizes [25] in several directions. We go beyond bilinear GAN objectives by covering general convex-concave payoffs, including RLHF and nonlinear adversarial training of mean-field NNs. Our analysis also applies to mirror geometries beyond relative entropy. Finally, whereas [25] establishes an explicit $\mathcal{O}(N^{-1/2})$ rate only for simultaneous MDA, we prove that alternating MDA achieves the faster rate $\mathcal{O}(N^{-2/3})$.
- At the technical level, the main novelty is a dual space commutator argument for alternating MDA on spaces of measures. Alternating updates generate asymmetric Bregman terms. In finite dimensions, such terms can be controlled using the inner product identification between iterates and the mirror map since both live in the same finite-dimensional space, but this identification is unavailable in the measure setting. Iterates are measures, while the mirror map naturally lives in a function space. We therefore pass to the primal-dual pair $(\mathcal{M}(\mathcal{X}), C_b(\mathcal{X}))$, define the convex conjugate h^* on $C_b(\mathcal{X})$, and prove a transport identity showing that primal Bregman commutators on measures coincide with dual Bregman commutators on functions. This allows us to apply Fréchet smoothness of the mirror map to control the update error and obtain the improved $\mathcal{O}(N^{-2/3})$ rate. To our knowledge, the use of this dual formulation of MDA on a function space is novel and may be useful beyond the present setting.

1.4 Motivating examples

Training of GANs. Let $\hat{\xi} \in \mathcal{P}(\mathcal{Y})$ be the empirical measure of the i.i.d. sampled particles $\{x_i\}_{i=1}^M \subseteq \mathcal{Y}$, and let $\xi \in \mathcal{P}(\mathcal{Z})$ be a source measure. Consider the measurable parametrized transport map $T_\theta : \mathcal{Z} \to \mathcal{Y}$ (which typically can be viewed as a neural network with parameters $\theta \in \Theta \subseteq \mathbb{R}^d$). The *pushforward* of the measure ξ on $\mathcal{Z}$ via T_θ is the measure $T_\theta \# \xi$ on $\mathcal{Y}$ characterized by $\int_{\mathcal{Y}} \varphi \mathrm{d}\,(T_\theta \# \xi) = \int_{\mathcal{Z}} (\varphi \circ T_\theta)\,\mathrm{d}\xi$, for any measurable function $\varphi : \mathcal{Y} \to \mathbb{R}$. The aim of a GAN is to search for the optimal set of parameters $\theta^* \in \Theta$ that minimizes the distance between the generated measure $T_{\theta^*} \# \xi$ and the empirical measure $\hat{\xi}$. In order to evaluate this distance, we define the function $D_w : \mathcal{Y} \to \mathbb{R}$ (which can also be viewed as a neural network with parameters $w \in \mathcal{W} \subseteq \mathbb{R}^d$), and solve the min-max problem

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \int_{\mathcal{Y}} D_w(y)(T_\theta \# \xi - \hat{\xi})(\mathrm{d}y).$$

For example, if the family of functions $\{D_w\}_{w \in \mathcal{W}}$ is either 1-Lipschitz continuous or uniformly bounded, the resulting GAN corresponds to the Wasserstein GAN or the Total Variation GAN, respectively [5]. On the other hand, if the family of functions $\{D_w\}_{w \in \mathcal{W}}$ belongs to the norm unit ball of a reproducing kernel Hilbert space (RKHS), we recover the Maximum Mean Discrepancy (MMD) GAN [34]. Solving this problem on the finite-dimensional subspaces $\theta, w \subseteq \mathbb{R}^d$ 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 $f(\theta, w) := \int_{\mathcal{Y}} D_w(y)(T_\theta \# \xi - \hat{\xi})(\mathrm{d}y)$, we solve the mean-field

min-max game

$$\min_{\nu\in\mathcal{P}(\Theta)}\max_{\mu\in\mathcal{P}(\mathcal{W})}\int_{\mathcal{W}}\int_{\Theta} f(\theta,w)\nu(\mathrm{d}\theta)\mu(\mathrm{d}w). \tag{3}$$

Note that the lifted problem is bilinear in ν and μ, so an MNE for (3) exists under mild conditions (see footnote 1). We illustrate that alternating updates speed up GANs training in simple numerical experiments (see Appendix D).

Nash Learning from Human Feedback. Reinforcement Learning from Human Feedback (RLHF) has become essential to LLMs training since it provides a principled way to incorporate human preferences into the learning objective and thereby improve alignment beyond pretraining. This example illustrates that our framework applies to the RLHF (see, e.g., [39, 14] and the references therein for more context). In particular, we show how it can be cast as a convex-concave two-player game. Let $\mathcal{S}$ denote a measurable space of prompts/contexts and $\rho \in \mathcal{P}(\mathcal{S})$ their distribution. Let $\mathcal{A}$ denote a measurable space of responses. A preference model is a measurable map $P(\cdot \succ \cdot|x) : \mathcal{A}\times\mathcal{A} \to [0,1]$ given by $(y,y') \mapsto P(y \succ y'|x)$, and interpreted as the probability that a random human prefers response y over y' given context x. Assume that the preference model is anti-symmetric, i.e., $P(y \succ y'|x) = 1 - P(y' \succ y|x)$. Given two policies $\mu,\nu : \mathcal{S} \to \mathcal{P}(\mathcal{A}|\mathcal{S})$, define the preference between them by

$$P(\mu \succ \nu) := \int_{\mathcal{S}}\int_{\mathcal{A}}\int_{\mathcal{A}} P(y \succ y'|x)\nu(\mathrm{d}y')\mu(\mathrm{d}y)\rho(\mathrm{d}x).$$

Due to anti-symmetry, we have $P(\mu \succ \nu) = 1 - P(\nu \succ \mu)$ for any policies $\mu,\nu \in \mathcal{P}(\mathcal{A}|\mathcal{S})$. Define the convex feasible set of policies $\mathcal{C} = \mathcal{D} := \{\pi :\ \pi(\cdot|x) \in \mathcal{P}(\mathcal{A}),\ \forall x \in \mathcal{S}\}$. The learning objective is to find a policy μ that is preferred over every alternative policy ν and thus can be formulated as

$$\max_{\mu\in\mathcal{C}}\min_{\nu\in\mathcal{C}} P(\mu \succ \nu).^{2}$$

Adversarial training of mean-field NNs. Let $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d-1}$ be compact and convex with $\hat{\mu} \in \mathcal{P}(\mathcal{Y}\times\mathcal{Z})$ representing the training data $(y,z) \in \mathcal{Y}\times\mathcal{Z}$. Let $x = (w,b) \in \mathbb{R}^{d-1}\times\mathbb{R}$ be the parameters of the neural network, let $\varphi : \mathbb{R} \to \mathbb{R}$ be a bounded, continuous, non-constant activation function and define the function $\hat{\varphi}(x,z) \coloneqq \varphi(w\cdot z + b)$. The training of the two-layer mean-field neural network aims to find the optimal distribution of parameters $\nu \in \mathcal{P}(\mathbb{R}^d)$ which minimizes the lifted L^2-loss function (see, e.g., [26, Section 3] and the references therein)

$$\min_{\nu\in\mathcal{P}(\mathbb{R}^d)} F_0(\nu) := \int_{\mathcal{Y}\times\mathcal{Z}} \frac{1}{2}\left|y - \mathbb{E}^{X\sim\nu}[\hat{\varphi}(X,z)]\right|^2 \hat{\mu}(\mathrm{d}y,\mathrm{d}z).$$

To account for potential attacks by an adversary aiming to manipulate the training data $\hat{\mu}$, we minimize over the parameter distribution ν, considering the "worst-case" perturbation of $\hat{\mu}$. This leads to the two-player game

$$\min_{\nu\in\mathcal{P}(\mathbb{R}^d)}\max_{\mu\in\mathcal{P}(\mathcal{Y}\times\mathcal{Z})} \left(F_0(\nu,\mu) - \mathrm{TV}_\varepsilon^2(\mu,\hat{\mu})\right), \tag{4}$$

where, for any $\varepsilon > 0$, TV_ε denotes the smoothed total variation distance, which represents the cost incurred by the adversary to alter the original training data $\hat{\mu}$, and it is defined by

$$\mathrm{TV}_\varepsilon(\mu,\hat{\mu}) := \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} \sqrt{(m(r)-\hat{m}(r))^2 + \varepsilon^2}\,\mathrm{d}r,$$

for μ and $\hat{\mu}$ absolutely continuous with respect to Lebesgue with density m and $\hat{m}$, respectively. The resulting objective $F_\varepsilon(\nu,\mu) \coloneqq F_0(\nu,\mu) - \mathrm{TV}_\varepsilon^2(\mu,\hat{\mu})$ is a non-linear function covered by our general framework (see Proposition C.6). The choice of the incurred cost in (4) is, to an extent, arbitrary, and we focus here on $\mathrm{TV}_\varepsilon^2$ due to its convenience for verifying our assumptions. From a technical point of view, we consider the smoothed TV distance because, unlike the classical TV distance, it is flat differentiable everywhere. Alternative cost functions include the Wasserstein distance [7, 23] and the KL divergence [50].

[2]Because $P(\mu \succ \nu)$ is bilinear in (μ,ν), it admits a Nash equilibrium (ν^*,μ^*) which satisfies $\max_{\mu\in\mathcal{C}}\min_{\nu\in\mathcal{C}} P(\mu \succ \nu) = \min_{\nu\in\mathcal{C}}\max_{\mu\in\mathcal{C}} P(\mu \succ \nu) = P(\mu^* \succ \nu^*)$ [47, Theorem 1].

1.5 Related works

Mirror descent (MD) was originally proposed in [41] for solving convex optimization problems and has been extensively studied on finite-dimensional vector spaces (see, e.g., [9, 12, 36]). In the finite-dimensional min-max setting, [55] analyzed alternating MDA and showed that alternating updates can improve the convergence rate for constrained min-max games. In finite dimensions, the proof relies on Euclidean vector space geometry, which identifies the primal and dual spaces through the Euclidean inner product. In particular, the Bregman potential and the mirror map belong to the same finite-dimensional vector space.

In the space of probability measures, there is no canonical inner product structure. We therefore first construct the appropriate primal space geometry through flat derivatives and Bregman divergences on measures. We then pass to the duality pair between signed measures and bounded continuous functions in order to define the convex conjugate and the corresponding dual mirror map. This dual space construction is essential for the analysis of the alternating scheme.

The MD algorithm has also been extended to infinite-dimensional settings for studying optimization problems over spaces of measures, with applications in machine learning, including Sinkhorn's and Expectation-Maximization algorithms [6], as well as policy optimization for reinforcement learning [53, 29]. By leveraging results from finite-dimensional optimization [8, 36], [6] extends the convergence proof of [36] to infinite-dimensional MD. In particular, they show that convexity and relative smoothness of the objective are sufficient to obtain an $\mathcal{O}(N^{-1})$ convergence rate for single-agent measure optimization. In contrast, our setting is a two-player min-max problem, where monotonic decrease of the objective along the iterates is no longer available and where the alternating scheme creates additional asymmetric terms that must be controlled through the dual space construction.

Other works such as [25, 22] studied infinite-dimensional Mirror-Prox algorithms for finding MNEs of two-player zero-sum games. Mirror-Prox methods can achieve $\mathcal{O}(N^{-1})$ rates under Lipschitz-type conditions on the gradients of the payoff function (see, e.g., [12, Section 5.2.3]). Existing infinite-dimensional analyses (e.g., entropic Mirror-Prox for convex-concave payoffs [22]) rely on a TV-Lipschitz assumption on the Fréchet derivatives of the payoff (see [22, Assumption 4.2]), which is substantially stronger than our relative smoothness assumption (see Assumption 3.3 and Appendix I). In particular, for bilinear games $F(\nu, \mu) = \iint f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, relative smoothness automatically holds with $L_\nu = L_\mu = 0$ for any f, whereas the TV-Lipschitz requirement on the derivatives forces f to be bounded and Lipschitz, excluding many relevant settings such as RLHF and adversarial NNs due to lack of Lipschitz continuity. Moreover, Mirror-Prox requires two mirror steps per player per iteration, thus doubling gradient evaluations, and in the stochastic setting (such as in particle implementations) its rate matches $\mathcal{O}(N^{-1/2})$.

Finally, a convergence theory for an alternating Mirror-Prox scheme (the natural counterpart to our alternating MDA) appears technically challenging and, to our knowledge, is still an open question. We discuss the technical difficulties involved in analyzing Mirror-Prox in greater detail in Appendix I. The appendix also contains a more extensive discussion of related work on optimization over spaces of measures.

2 Convergence of the simultaneous MDA Algorithm 1

In what follows, we state our standing assumptions and the necessary definitions for introducing the convergence results.

Assumption 2.1 (Differentiability and convexity of h)**.** *Assume that $h \in \mathfrak{C}^1(\mathcal{E})$ (cf. Definition 1.1). Moreover, assume that $h : \mathcal{E} \to \mathbb{R}$ is α-strongly convex relative to the total variation (TV) distance, i.e., there exists $\alpha > 0$ such that for all $\varepsilon \in [0, 1]$ and $\nu', \nu \in \mathcal{E}$, with $\nu_\varepsilon := (1 - \varepsilon)\nu + \varepsilon\nu'$,*

$$h(\nu_\varepsilon) \leq (1 - \varepsilon)h(\nu) + \varepsilon h(\nu') - \frac{\alpha}{2}\varepsilon(1 - \varepsilon)\,\mathrm{TV}^2(\nu', \nu).$$

If Assumption 2.1 holds, then we show in Lemma B.3 that for any $\nu', \nu \in \mathcal{E}$, we have

$$h(\nu') - h(\nu) \geq \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu, x)(\nu' - \nu)(\mathrm{d}x) + \frac{\alpha}{2}\,\mathrm{TV}^2(\nu', \nu). \tag{5}$$

Assumption 2.1 implies (5) due to Lemma B.3 and hence, straight from Definition 1.2, we obtain $D_h(\nu', \nu) \geq \frac{\alpha}{2} \mathrm{TV}^2(\nu', \nu)$, for all $\nu', \nu \in \mathcal{E}$, and $D_h(\nu', \nu) = 0$ if and only if $\nu' = \nu$. In Appendix B.1, we provide examples of the corresponding sets $\mathcal{E}$ for the relative entropy and χ^2-divergence cases that satisfy Assumption 2.1. For other examples of functions h that verify the inequality $D_h(\nu', \nu) \geq \frac{\alpha}{2} \mathrm{TV}^2(\nu', \nu)$ and hence Assumption 2.1, see [17, Lemma 3.2].

Let $F : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ be such that $F(\cdot, \mu) \in \mathfrak{C}^1(\mathcal{C})$ and $F(\nu, \cdot) \in \mathfrak{C}^1(\mathcal{D})$ (cf. Definition 1.1). Under Definition 1.2, the Bregman divergences of $F(\cdot, \mu)$ and $F(\nu, \cdot)$ are respectively given by

$$D_{F(\cdot,\mu)}(\nu', \nu) := F(\nu', \mu) - F(\nu, \mu) - \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, \mu, x)(\nu' - \nu)(\mathrm{d}x),$$
$$D_{F(\nu,\cdot)}(\mu', \mu) := F(\nu, \mu') - F(\nu, \mu) - \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu, \mu, y)(\mu' - \mu)(\mathrm{d}y).$$

Assumption 2.2 (Convexity-concavity of F)**.** *Assume that F is convex in ν and concave in μ, i.e., for any $\nu, \nu' \in \mathcal{C}$ and $\mu, \mu' \in \mathcal{D}$, we have*

$$D_{F(\cdot,\mu)}(\nu', \nu) \geq 0, \quad D_{F(\nu,\cdot)}(\mu', \mu) \leq 0.$$

Assumption 2.3 (Uniform boundedness of the flat derivatives of F)**.** *Suppose that there exists $C_1 > 0$ and $C_2 > 0$ such that for all $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, and all $x, y \in \mathcal{X}$, we have*

$$\left| \frac{\delta F}{\delta \nu}(\nu, \mu, x) \right| \leq C_1, \quad \left| \frac{\delta F}{\delta \mu}(\nu, \mu, y) \right| \leq C_2.$$

Assumptions 2.2 and 2.3 are standard in the mean-field optimization literature, see, e.g., [16, 33]. In Proposition C.1 and C.6, we verify that Assumptions 2.2 and 2.3 are satisfied by the examples in Section 1.4. Another important tool utilized in the proofs of the main results is the three-point inequality that is proved in Appendix B following [6][Lemma 3].

Lemma 2.4 (Three-point inequality)**.** *Let Assumption 2.1 hold. Let $G : \mathcal{E} \to \mathbb{R}$ be convex and $G \in \mathfrak{C}^1(\mathcal{E})$. 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 main result on the convergence of the simultaneous MDA algorithm.

Theorem 2.5 (Convergence of the simultaneous MDA Algorithm 1)**.** *Let (ν^0, μ^0) be such that $\Delta := \sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) < \infty$. Suppose that Assumption 2.1, 2.2, 2.3 hold. If $\tau = \Theta\left(N^{-1/2}\right)$ [3], then there exists a constant $K_{sim} \equiv K_{sim}(\Delta, C_1, C_2, \alpha) > 0$ such that*

$$\mathrm{NI}\left(\frac{1}{N} \sum_{n=0}^{N-1} \nu^n, \frac{1}{N} \sum_{n=0}^{N-1} \mu^n \right) \leq \frac{K_{sim}}{\sqrt{N}}.$$

Remark 2.6 (Initialization condition)**.** *The initialization requirement in Theorem 2.5, namely, $\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) < \infty$ must be verified case by case, depending on the choice of h and the admissible classes $\mathcal{C}, \mathcal{D}$. Such verifications for Examples B.1 and B.2 are carried out in Lemmas B.7 and B.8, respectively.*

Remark 2.7 (About the proof of Theorem 2.5)**.** *In their proof of convergence of the infinite-dimensional MD algorithm for convex F, [6] 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 $\mathcal{O}(N^{-1})$. In the case of Algorithm 1, monotonicity no longer holds, and we therefore prove convergence only for the time-averaged iterates. The corresponding convergence rate is expected to be the slower rate $\mathcal{O}\left(N^{-1/2}\right)$, as suggested by the finite-dimensional [12, Theorem 5.1].*

Furthermore, in contrast to the proof strategy in [6], our argument does not require relative smoothness of F, but only uniform boundedness of its first-order flat derivatives. This is made possible by requiring h to be strongly convex relative to the TV^2 distance, which is stronger than the strict convexity of h used in [6], yet still verifiable for typical choices of divergences, see Examples B.1 and

[3] We say $f(n) = \Theta(g(n))$ if there exists $c_1, c_2, n_0 > 0$ such that $c_1 g(n) \leq f(n) \leq c_2 g(n)$, for all $n \geq n_0$.

B.2. The choice of the TV^2 *distance in Assumption 2.1 is motivated by the fact that several functions* h*, as illustrated in Examples B.1 and B.2 and noted in [17, Lemma 3.2], satisfy the inequality* $D_h(\nu', \nu) \geq \frac{\alpha}{2} \mathrm{TV}^2(\nu', \nu)$*.*

By contrast, replacing TV^2 *with, for example, the squared* L^2*-Wasserstein distance* $\mathcal{W}_2^2(\nu', \nu)$ *would reduce the generality of our analysis. Apart from the relative entropy, we are not aware of any divergences satisfying* $D_h(\nu', \nu) \geq \frac{\alpha}{2}\mathcal{W}_2^2(\nu', \nu)$*, and even this would require a considerably stronger (and difficult to verify) condition that the iterates produced by Algorithm 1 satisfy the Talagrand inequality. Thus, Theorem 2.5 should be viewed as the baseline explicit mean-field MDA guarantee. It requires only first-order boundedness of the flat derivatives but does not exploit the additional information structure of alternating play. The improved rate in Theorem 3.5 comes from analyzing precisely this additional structure.*

2.1 Sketch of proof of Theorem 2.5

Here, we present the proof sketch of Theorem 2.5. The full proof is provided in Section A. Applying the Bregman proximal inequality (Lemma 2.4) to each of the simultaneous mirror steps yields two inequalities: one for the ν-update and one for the μ-update. Each inequality relates the corresponding linearized terms in (1) to three Bregman terms: $D_h(\nu, \nu^n)$, $D_h(\nu, \nu^{n+1})$ and $D_h(\nu^{n+1}, \nu^n)$, and likewise for the μ-update. Combining the linearized terms in (1) involving $\nu - \nu^n$ and $\mu - \mu^n$ with the convexity-concavity of F via Assumption 2.2, we obtain upper bounds on the differences $F(\nu^n, \mu^n) - F(\nu, \mu^n)$ and $F(\nu^n, \mu) - F(\nu, \mu^n)$, respectively. Assumptions 2.1 and 2.3 then convert the remaining linearized terms involving $\nu^{n+1} - \nu^n$, $\mu^{n+1} - \mu^n$ and the Bregman divergences $D_h(\nu^{n+1}, \nu^n)$, $D_h(\mu^{n+1}, \mu^n)$ into quadratic bounds in TV. More precisely, for the ν-part one obtains the expression

$$F(\nu^n, \mu^n) - F(\nu, \mu^n) \leq -\frac{\alpha}{2\tau}\left(\mathrm{TV}(\nu^{n+1}, \nu^n) - \frac{2\tau C_1}{\alpha}\right)^2 + \frac{2\tau C_1^2}{\alpha} + \frac{1}{\tau} S_n^\nu,$$

$S_n^\nu := D_h(\nu, \nu^n) - D_h(\nu, \nu^{n+1})$, and similarly for the μ-part. Combining the two inequalities gives the unified bound

$$F(\nu^n, \mu) - F(\nu, \mu^n) \leq \frac{2\tau}{\alpha}\left(C_1^2 + C_2^2\right) + \frac{1}{\tau}(S_n^\nu + S_n^\mu).$$

Summing over $n = 0, ..., N-1$, telescoping $S_n^\nu + S_n^\mu$, dividing by N and using Jensen's inequality, then yields

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^n, \frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right) \leq \frac{2\tau}{\alpha}\left(C_1^2 + C_2^2\right) + \frac{1}{N\tau}\Delta.$$

Setting $\tau = \Theta\left(N^{-1/2}\right)$ leads to the final bound, establishing the claim.

3 Convergence of the alternating MDA Algorithm 2

The analysis of alternating MDA requires a dual-space viewpoint. In the simultaneous scheme, the error terms can be controlled directly using strong convexity of h. Alternating updates, however, produce asymmetric Bregman commutators. In finite-dimensional Euclidean spaces such terms can be handled using self duality of the inner product and derivatives of the dual mirror map on the same space. In the measure setting, the primal variables are signed measures, while the natural dual objects are bounded continuous functions. Therefore, before analyzing the alternating scheme, we introduce the duality pair between signed measures and bounded continuous functions.

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 $\|\mu\|_{\mathrm{TV}} := |\mu|(\mathcal{X})$. Let $(C_b(\mathcal{X}), \|\cdot\|_\infty)$ be the Banach space of bounded continuous functions from $\mathcal{X} \subseteq \mathbb{R}^d$ to $(\mathbb{R}, |\cdot|)$, where $|\cdot|$ is the Euclidean norm. For any $(f, m) \in C_b(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$, we define the duality pairing $\langle\cdot,\cdot\rangle : C_b(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ by

$$\langle f, m\rangle := \int_{\mathcal{X}} f(x) m(\mathrm{d}x). \tag{6}$$

Next, we define the notion of convex conjugate of $h : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ relative to the duality pairing (6).

Definition 3.1 (Convex conjugate)**.** *Let $h : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ be a function. Then the map $h^* : C_b(\mathcal{X}) \to \mathbb{R}$ given by $h^*(f) := \sup_{m\in\mathcal{M}(\mathcal{X})} \{\langle f, m\rangle - h(m)\}$ is called the convex conjugate of h.*

Regardless of the convexity of h, it follows from [11, Theorem 2.112] that h^* is convex on $C_b(\mathcal{X})$, i.e., for all $\lambda \in [0, 1]$ and all $f', f \in C_b(\mathcal{X})$, we have that $h^*((1-\lambda)f + \lambda f') \leq (1-\lambda)h^*(f)+\lambda h^*(f')$. In Examples F.2 and F.4, we provide in Examples B.1 and B.2 the explicit form of h^* when h is the relative entropy and the chi-squared divergence, respectively. Analogous to the characterization of the convexity of h on $\mathcal{P}(\mathcal{X})$ via flat derivatives, we can characterize the convexity of h^* on $C_b(\mathcal{X})$ using its Fréchet derivative (cf. Definition F.1).

We say $h^* : C_b(\mathcal{X}) \to \mathbb{R}$ is *Fréchet convex* if for any $f, f' \in C_b(\mathcal{X})$, we have $h^*(f') - h^*(f) \geq \nabla_{\mathcal{F}} h^*(f)[f' - f]$. As shown in Examples F.3 and F.5, when h is chosen as the relative entropy or the chi-squared divergence, its convex conjugate h^* admits the Fréchet derivatives $\nabla_{\mathcal{F}} h^*(f)$.

Furthermore, using the Fréchet characterization of convexity, we can define the Bregman divergence between f and f' and its commutator on the convex subset $\mathcal{U}$ of the dual space $C_b(\mathcal{X})$.

Definition 3.2 (Dual Bregman divergence and commutator)**.** *Let $h^* : C_b(\mathcal{X}) \to \mathbb{R}$ be the convex conjugate of h. Assume h^* is Fréchet differentiable on $\mathcal{U}$. The dual h^*-Bregman divergence and its commutator are the maps $D_{h^*}, \mathrm{Com}^*_{h^*} : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$, respectively, given by*

$$D_{h^*}(f', f) := h^*(f') - h^*(f) - \nabla_{\mathcal{F}} h^*(f)[f' - f],$$
$$\mathrm{Com}^*_{h^*}(f', f) := D_{h^*}(f', f) - D_{h^*}(f, f').$$

Before stating the next assumption on the dual space, we introduce two additional assumptions on F. These conditions are required only for the analysis of the alternating scheme, as they allow us to control the extra asymmetric terms arising from the updates in (2).

Assumption 3.3 (Relative smoothness of F)**.** *Assume the function F is (L_ν, L_μ)-smooth relative to h, i.e., there exist $L_\nu, L_\mu > 0$ such that, for any $\nu, \nu' \in \mathcal{C}$ and $\mu, \mu' \in \mathcal{D}$, we have*

$$D_{F(\cdot,\mu)}(\nu', \nu) \leq L_\nu D_h(\nu', \nu),\ D_{F(\nu,\cdot)}(\mu', \mu) \geq -L_\mu D_h(\mu', \mu).$$

In Proposition C.1, C.2 and C.6, we verify that Assumptions 3.3 is satisfied by the examples in Section 1.4.

The following assumption on the dual mirror map will turn out to be crucial for showing the improvement in the convergence rate of Algorithm 2 compared to the simultaneous algorithm.

Assumption 3.4 (Third-order smoothness of the dual mirror map)**.** *Suppose that $(C_b(\mathcal{X}) \times C_b(\mathcal{X}) \times C_b(\mathcal{X})) \ni (g, g, g) \mapsto \nabla^3_{\mathcal{F}} h^*(f)[g, g, g] \in \mathbb{R}$ is uniformly bounded, i.e, there exists $L_{h^*} > 0$ such that for all $g \in C_b(\mathcal{X})$,*

$$\left|\nabla^3_{\mathcal{F}} h^*(f)[g, g, g]\right| \leq L_{h^*} \|g\|^3_\infty.$$

Assumption 3.4 is used only in the analysis of the alternating scheme. Its role is to quantify how far the dual mirror geometry is from being locally symmetric. Indeed, alternating updates create Bregman commutators of the form $D_{h^*}(g', g) - D_{h^*}(g, g')$. If the geometry were locally symmetric, this difference would vanish as for χ^2-divergence (see Proposition F.16). Assumption 3.4 ensures that the remaining asymmetry is cubic in $\|g - g'\|_\infty$. Concretely, the Bregman commutators admit the cubic bound $|D_{h^*}(g', g) - D_{h^*}(g, g')| \leq cL_{h^*}\|g' - g\|^3_\infty$, for some constant $c > 0$. Since the mirror update gives $g - g' = \mathcal{O}(\tau)$, the bound becomes $\mathcal{O}(\tau^3)$ along the iterates.

It is also worth comparing Assumption 3.4 with the notion of self-concordance in convex analysis (see, e.g., [42, Chapter 2]). Self-concordance provides a local control of third-order derivatives in terms of the Hessian, namely, it requires that for all $g \in C_b(\mathcal{X})$,

$$\left|\nabla^3_{\mathcal{F}} h^*(f)[g, g, g]\right| \leq 2\left(\nabla^2_{\mathcal{F}} h^*(f)[g, g]\right)^{3/2}.$$

Importantly, Assumption 3.4 is not an artificial technical condition since it holds for two standard mirror geometries, namely relative entropy and the χ^2-divergence. In Propositions F.14 and F.16, we verify Assumption 3.4 in the case where h is the relative entropy and the χ^2-divergence, respectively. On the other hand, self-concordance holds only for the χ^2 mirror map as Proposition F.16 immediately shows. In Example F.15, we explain why self-concordance may fail for the entropy mirror map. In this case, it is possible to prove Theorem 3.5 for the χ^2 mirror map only by assuming self-concordance.

Now, we are ready to state the second main result of the paper.

Theorem 3.5 (Convergence of the alternating MDA Algorithm 2)**.** *Let (ν^0, μ^0) be such that $\Delta := \sup_{\nu\in\mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu\in\mathcal{D}} D_h(\mu, \mu^0) < \infty$ (cf. Remark 2.6). Let Assumptions 2.1, 2.2, 2.3, 3.3 and 3.4 hold. Suppose that $\tau L \leq \frac{1}{2}$, with $L := \max\{L_\nu, L_\mu\}$. If $\tau = \Theta\left(N^{-1/3}\right)$, then there exists a constant $K_{alt} \equiv K_{alt}(\Delta, C_1, C_2, \alpha, L, L_h^*, |F(\nu^0, \mu^0)|) > 0$ such that*

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right) \leq \frac{K_{alt}}{N^{2/3}}. \tag{7}$$

Remark 3.6 (Bilinear games)**.** *In particular, if $F(\nu, \mu) = \iint f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, for a bounded function f, then Assumptions 2.2, 2.3, 3.3 are satisfied and in Assumption 3.3 we have $L_\nu = L_\mu = 0$. Therefore, $L = 0$ in* (7)*, and hence Theorem 3.5 is consistent with the already known rate $\mathcal{O}\left(N^{-2/3}\right)$ of the MDA algorithm in finite-dimensional constrained bilinear min-max games; see [55, 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 3.5 substantially generalizes the results of [55]. Moreover, the proof also differs substantially from the Euclidean argument because the asymmetric alternating terms must be transported from the primal measure space to the dual function space before Assumption 3.4 can be applied.*

3.1 Sketch of proof of Theorem 3.5

The proof of Theorem 3.5 is where the main infinite-dimensional difficulty appears. The alternating update creates a non-symmetric error term that has no analogue in the simultaneous analysis. In finite-dimensional Euclidean space, the corresponding term can be controlled using the self-duality of the inner product and the third derivative of the mirror map. In the measure space setting, one must first identify the correct dual object and prove that the asymmetric terms can be represented as dual Bregman commutators on $C_b(\mathcal{X})$. This commutator transport identity in Lemma G.5 is the key technical step that makes Assumption 3.4 applicable.

Using Assumption 3.3, we combine the asymmetric term $F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n)$ with $\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^n - \nu^{n+1})(\mathrm{d}x)$ and $\int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y)$, which yields the Bregman commutator $\mathrm{Com}_{\nu^n} := D_h(\nu^n, \nu^{n+1}) - D_h(\nu^{n+1}, \nu^n)$, with Com_{μ^n} being defined analogously. Proceeding as in the simultaneous case leads to

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

where the $\mathcal{O}\left(\frac{1}{N}\right)$ term uses the uniform boundedness of F proved in Lemma A.2. Let $f^n := \frac{\delta h}{\delta \nu}(\nu^n, \cdot)$, $g^n := \frac{\delta h}{\delta \mu}(\mu^n, \cdot)$ and define the dual Bregman commutator $\mathrm{Com}^*_{\nu^n} := D_{h^*}\left(f^{n+1}, f^n\right) - D_{h^*}\left(f^n, f^{n+1}\right)$, with $\mathrm{Com}^*_{\mu^n}$ being defined analogously. We transport the commutators from the primal space to the dual space via Lemma G.5 and obtain

$$\mathrm{Com}_{\nu^n} = \mathrm{Com}^*_{\nu^n},\ \mathrm{Com}_{\mu^n} = \mathrm{Com}^*_{\mu^n}.$$

Using the third-order Fréchet derivative of h^* (Definition F.10) and its uniform boundedness (Assumption 3.4), we obtain $\mathrm{Com}^*_{\nu^n} \leq \frac{L_{h^*}}{4}\left\|f^{n+1} - f^n\right\|_\infty^3$. By Proposition B.4, the first-order optimality condition for the ν-update yields (up to an additive constant) $f^{n+1} - f^n = -\tau\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot)$. Hence, by Assumption 2.3,

$$\mathrm{Com}^*_{\nu^n} \leq \frac{L_{h^*}}{4}\tau^3 C_1^3, \quad \mathrm{Com}^*_{\mu^n} \leq \frac{L_{h^*}}{4}\tau^3 C_2^3,$$

where the second inequality follows from applying the same argument to the μ-commutator. Therefore,

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

Choosing $\tau = \Theta\left(N^{-1/3}\right)$ and noting that $N^{-1} \leq N^{-2/3}$, for all $N \geq 1$, yields the conclusion.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their insightful comments, which helped improve the paper. R-AL acknowledges the support of 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. This work was partly carried out while R-AL was a PhD student at Heriot-Watt University. The paper was finalized during R-AL's postdoctoral appointment at the RIKEN Center for Advanced Intelligence Project (RIKEN AIP). R-AL also acknowledges support from JSPS KAKENHI Grant Number JP26K25544 for travel to NeurIPS 2026 to present this work. ŁS 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] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *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.

[6] P.-C. Aubin-Frankowski, A. Korba, and F. Léger. Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. In *Advances in Neural Information Processing Systems*, volume 35, pages 17263–17275. Curran Associates, Inc., 2022.

[7] X. Bai, G. He, Y. Jiang, and J. Obloj. Wasserstein distributional robustness of neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[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] C. Bonet, T. Uscidda, A. David, P.-C. Aubin-Frankowski, and A. Korba. Mirror and preconditioned gradient descent in Wasserstein space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

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

[13] Y. Cai, S. Mitra, X. Wang, and A. Wibisono. On the convergence of min-max Langevin dynamics and algorithm. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 677–754. PMLR, 30 Jun–04 Jul 2025.

[14] D. Calandriello, Z. D. Guo, R. Munos, M. Rowland, Y. Tang, B. A. Pires, P. H. Richemond, C. Le Lan, M. Valko, T. Liu, R. Joshi, Z. Zheng, and B. Piot. Human alignment of large language models through online preference optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[15] 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.

[16] F. Chen, Z. Ren, and S. Wang. Entropic fictitious play for mean field optimization problem. *Journal of Machine Learning Research*, 24(211):1–36, 2023.

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

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

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

[20] P. G. Ciarlet. *Linear and Nonlinear Functional Analysis with Applications*. Society for Industrial and Applied Mathematics, 2013.

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

[22] 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.

[23] C. A. García Trillos and N. García Trillos. On adversarial robustness and the use of Wasserstein ascent-descent dynamics to enforce it. *Information and Inference: A Journal of the IMA*, 13(3), 2024.

[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017.

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

[26] 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.

[27] B. Jourdain and A. Tse. Central limit theorem over non-linear functionals of empirical measures with applications to the mean-field fluctuation of interacting diffusions. *Electronic Journal of Probability*, 2020.

[28] J. E. Katona, X. Wang, and A. Wibisono. A symplectic analysis of alternating mirror descent. *Journal of Machine Learning Research*, 27(44):1–61, 2026.

[29] B. Kerimkulov, J.-M. Leahy, D. Siska, Ł. Szpruch, and Y. Zhang. A fisher–rao gradient flow for entropy-regularised markov decision processes in polish spaces. *Foundations of Computational Mathematics*, 25(5):1959–2002, 2025.

[30] B. Kerimkulov, D. Šiška, Łukasz Szpruch, and Y. Zhang. Mirror descent for stochastic control problems with measure-valued controls. *Stochastic Processes and their Applications*, 190:104765, 2025.

[31] 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.

[32] R.-A. Lascu, M. Majka, and Ł. Szpruch. A Fisher-Rao gradient flow for entropic mean-field min-max games. *Transactions on Machine Learning Research*, 2024.

[33] R.-A. Lascu, M. B. Majka, and Ł. Szpruch. Entropic mean-field min–max problems via Best Response flow. *Appl. Math. Optim.*, 91(2), Mar. 2025.

[34] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] R. Munos, M. Valko, D. Calandriello, M. Gheshlaghi Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, C. Fiegel, A. Michi, M. Selvi, S. Girgin, N. Momchev, O. Bachem, D. J. Mankowitz, D. Precup, and B. Piot. Nash learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2024.

[40] T. Nan, S. D. Gupta, G. Iyengar, and C. Kroer. On the $O(1/T)$ convergence of alternating gradient descent–ascent in bilinear games. In *The Fourteenth International Conference on Learning Representations*, 2026.

[41] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

[42] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

[43] H. Nikaidô and K. Isoda. Note on non-cooperative convex game. *Pacific Journal of Mathematics*, 5:807–815, 1955.

[44] 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.

[45] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1970.

[46] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

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

[48] 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.

[49] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015.

[50] N. Si, F. Zhang, Z. Zhou, and J. Blanchet. Distributionally robust batch contextual bandits. *Management Science*, 69(10):5772–5793, 2023.

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

[52] 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.

[53] A. Tomar, L. Shani, Y. Efroni, and M. Ghavamzadeh. Mirror Descent policy optimization. In *International Conference on Learning Representations*, 2022.

[54] G. Wang and L. Chizat. An Exponentially Converging Particle Method for the Mixed Nash Equilibrium of Continuous Games. *Open Journal of Mathematical Optimization*, 6:1–66, 2025.

[55] A. Wibisono, M. Tao, and G. Piliouras. Alternating Mirror Descent for constrained min-max games. In *Advances in Neural Information Processing Systems*, volume 35, pages 35201–35212. Curran Associates, Inc., 2022.

[56] C. Zalinescu. *Convex Analysis In General Vector Spaces*. World Scientific Publishing Company, 2002.

Appendix

The appendix is organized into several sections (A–I), each providing supporting technical material, extended proofs and further context for the main text. Section A contains the complete proofs of the paper's two main results, namely Theorem 2.5 and Theorem 3.5. Section B contains proofs for supplementary lemmas and propositions referenced earlier in the paper. In Section C, we check that Assumptions 2.2, 2.3 and 3.3 hold in for the examples in Section 1.4. In Section D we describe all implementation details, experimental settings and parameter choices. Section E collects standard definitions concerning differentiability with respect to probability measures that are used throughout the paper. Section F complements the previous section by treating differentiability in the dual space and supporting the duality arguments used in the proof of Theorem 3.5. Section G gathers duality identities used particularly in the proof of Theorem 3.5. Section H shows that an implicit MDA scheme achieves the same sublinear convergence rate as the continuous-time dynamics under convex-concave assumptions on F. Finally, Section I collects references and related literature not included in the main text.

A Proofs of Theorem 2.5 and Theorem 3.5

This section is dedicated to the proofs of the main results, namely Theorem 2.5 and Theorem 3.5. Before proving Theorem 2.5, we summarize the assumptions on h and F, outlining how each applies to the simultaneous and alternating schemes.

A.1 Summary of assumptions and applicability

The assumptions distinguish between properties of the mirror geometry and properties of the payoff. Both the simultaneous and alternating schemes require a differentiable mirror map that is strongly convex in total variation (Assumption 2.1), together with a convex-concave payoff with bounded flat derivatives (Assumptions 2.2 and 2.3). The alternating scheme additionally requires relative smoothness of the payoff with respect to the mirror map (Assumption 3.3) and a bounded third Fréchet derivative of the dual mirror map (Assumption 3.4). Both analyses also require finite initial Bregman divergences. Table 1 summarizes these conditions and their verification in the settings considered in this paper.

Component	**Condition**	**Scheme**	**Verified settings**
Mirror	Assumption 2.1	Both	Relative entropy and χ^2
	Assumption 3.4	Alternating	Relative entropy and χ^2
Payoff	Assumptions 2.2 and 2.3	Both	Propositions C.1, C.2, C.6
	Assumption 3.3	Alternating	Propositions C.1, C.2, C.6
Initial data	Finite initial Bregman divergences	Both	Lemmas B.7 and B.8

Table 1: Summary of the assumptions and their applicability. The listed settings satisfy the corresponding conditions on the admissible domains and under the hypotheses specified in the respective results.

Mirror geometries. The relative entropy and χ^2 mirror geometries considered here both satisfy the TV-strong convexity condition in Assumption 2.1 and the dual third-order condition in Assumption 3.4. For the χ^2 geometry, the third derivative of h^* vanishes on the dual domain.

Bounded bilinear payoffs. Consider $F(\nu, \mu) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, where f is a bounded measurable kernel. The payoff is affine in each argument and therefore convex-concave. Its flat derivatives are uniformly bounded by $\|f\|_\infty$. Moreover, affinity in each argument makes the corresponding first-order expansion exact. Thus the separate-variable relative smoothness conditions hold with $L_\nu = L_\mu = 0$. Hence, for bilinear objectives, the payoff assumptions reduce to boundedness of the kernel. This applies to the lifted GAN objective in Proposition C.1 and the NLHF objective in Proposition C.2.

Nonlinear payoffs. The payoff assumptions do not require bilinearity. Proposition C.6 verifies them for the nonlinear adversarial training problem involving a mean-field neural network. The conditions stated there include a compact data domain, a bounded activation, clipping of the output layer, and a smoothed total variation penalty.

Initialization. Finiteness of the initial Bregman divergences is a separate requirement on the initial measures and the chosen mirror geometry. Lemmas B.7 and B.8 verify this condition for the relative entropy and χ^2 geometries, respectively.

A.2 Proof of Theorem 2.5

Proof of Theorem 2.5. 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 2.4 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,

$$-\tau \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) - D_h(\nu, \nu^n) \leq -\tau \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) - D_h(\nu^{n+1}, \nu^n) - D_h(\nu, \nu^{n+1}). \quad (8)$$

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 2.4 with $\bar{\nu} = \mu^{n+1}$ and $\mu = \mu^n$ implies that, for any $\mu \in \mathcal{D}$, we have

$$\tau \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - D_h(\mu, \mu^n) \leq \tau \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - D_h(\mu^{n+1}, \mu^n) - D_h(\mu, \mu^{n+1}). \quad (9)$$

Using the convexity of $\nu \mapsto F(\nu, \mu)$ in (8), with $\nu = \nu^n$ and $\mu = \mu^n$, we have that

$$F(\nu^n, \mu^n) - F(\nu, \mu^n) - \frac{1}{\tau} D_h(\nu, \nu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^n - \nu^{n+1})(\mathrm{d}x) - \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}). \quad (10)$$

From Assumptions 2.1 and 2.3, it follows from (10) that

$$\begin{aligned} &F(\nu^n, \mu^n) - F(\nu, \mu^n) - \frac{1}{\tau} D_h(\nu, \nu^n) \\ &\quad \leq 2C_1 \, \mathrm{TV}(\nu^{n+1}, \nu^n) - \frac{\alpha}{2\tau} \, \mathrm{TV}^2(\nu^{n+1}, \nu^n) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}) \\ &\quad = -\frac{\alpha}{2\tau} \left(\mathrm{TV}(\nu^{n+1}, \nu^n) - \frac{2\tau C_1}{\alpha} \right)^2 + \frac{2\tau C_1^2}{\alpha} - \frac{1}{\tau} D_h(\nu, \nu^{n+1}) \\ &\quad \leq \frac{2\tau C_1^2}{\alpha} - \frac{1}{\tau} D_h(\nu, \nu^{n+1}), \end{aligned} \quad (11)$$

where the equality follows from the standard identity $-(a-b)^2 + b^2 = -a^2 + 2ab$.

Similarly, using concavity of $\mu \mapsto F(\nu, \mu)$ in (9), with $\nu = \nu^n$ and $\mu = \mu^n$, we have that

$$F(\nu^n, \mu) - F(\nu^n, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu^{n+1}, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^{n+1}). \quad (12)$$

From Assumptions 2.1 and 2.3, it follows from (12) that

$$\begin{aligned} &F(\nu^n,\mu)-F(\nu^n,\mu^n)-\frac{1}{\tau}D_h(\mu,\mu^n)\\ &\quad\le 2C_2\,\mathrm{TV}(\mu^{n+1},\mu^n)-\frac{\alpha}{2\tau}\,\mathrm{TV}^2(\mu^{n+1},\mu^n)-\frac{1}{\tau}D_h(\mu,\mu^{n+1})\\ &\quad=-\frac{\alpha}{2\tau}\left(\mathrm{TV}(\mu^{n+1},\mu^n)-\frac{2\tau C_2}{\alpha}\right)^2+\frac{2\tau C_2^2}{\alpha}-\frac{1}{\tau}D_h(\mu,\mu^{n+1})\\ &\quad\le\frac{2\tau C_2^2}{\alpha}-\frac{1}{\tau}D_h(\mu,\mu^{n+1}). \end{aligned} \tag{13}$$

Adding inequalities (11) and (13) implies that for any $(\nu,\mu)\in\mathcal{C}\times\mathcal{D}$ we have

$$\begin{aligned} &F(\nu^n,\mu)-F(\nu,\mu^n)\\ &\le\frac{2\tau}{\alpha}\left(C_1^2+C_2^2\right)+\frac{1}{\tau}D_h(\nu,\nu^n)+\frac{1}{\tau}D_h(\mu,\mu^n)-\frac{1}{\tau}D_h(\nu,\nu^{n+1})-\frac{1}{\tau}D_h(\mu,\mu^{n+1}). \end{aligned} \tag{14}$$

Summing the previous inequality over $n=0,1,...,N-1$, using $D_h(\nu,\nu^N)+D_h(\mu,\mu^N)\ge 0$, for any $(\nu,\mu)\in\mathcal{C}\times\mathcal{D}$, bounding the right-hand from above by its supremum over $(\nu,\mu)\in\mathcal{C}\times\mathcal{D}$, and dividing by N gives

$$\frac{1}{N}\sum_{n=0}^{N-1}\left(F(\nu^n,\mu)-F(\nu,\mu^n)\right)\le\frac{2\tau}{\alpha}\left(C_1^2+C_2^2\right)+\frac{1}{N\tau}\left(\sup_{\nu\in\mathcal{C}}D_h(\nu,\nu^0)+\sup_{\mu\in\mathcal{D}}D_h(\mu,\mu^0)\right). \tag{15}$$

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

$$\begin{aligned} \frac{1}{N}\sum_{n=0}^{N-1}\left(F(\nu^n,\mu)-F(\nu,\mu^n)\right)&=\frac{1}{N}\sum_{n=0}^{N-1}F(\nu^n,\mu)-\frac{1}{N}\sum_{n=0}^{N-1}F(\nu,\mu^n)\\ &\ge F\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^n,\mu\right)-F\left(\nu,\frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right). \end{aligned} \tag{16}$$

Combining (15) with (16) and taking maximum over (ν,μ) gives

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^n,\frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right)\le\frac{2\tau}{\alpha}\left(C_1^2+C_2^2\right)+\frac{1}{N\tau}\left(\sup_{\nu\in\mathcal{C}}D_h(\nu,\nu^0)+\sup_{\mu\in\mathcal{D}}D_h(\mu,\mu^0)\right).$$

Minimizing the right-hand side over τ amounts to taking

$$\tau=\sqrt{\frac{\alpha\left(\sup_{\nu\in\mathcal{C}}D_h(\nu,\nu^0)+\sup_{\mu\in\mathcal{D}}D_h(\mu,\mu^0)\right)}{2N\left(C_1^2+C_2^2\right)}},$$

and hence we obtain

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^n,\frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right)\le 2\sqrt{\frac{2\left(C_1^2+C_2^2\right)\left(\sup_{\nu\in\mathcal{C}}D_h(\nu,\nu^0)+\sup_{\mu\in\mathcal{D}}D_h(\mu,\mu^0)\right)}{\alpha N}}.$$

□

A.3 Proof of Theorem 3.5

Before we proceed with the proof of Theorem 3.5, we will need two auxiliary results, which will turn out to be essential. The proofs of Lemmas A.1 and A.2 are given in Appendix B.

Lemma A.1. *Let Assumptions 2.1, 2.3 and 3.3 hold. Suppose that $\tau L\le\frac{1}{2}$, with $L:=\max\{L_\nu,L_\mu\}$. Then, for both Algorithms 1 and 2, it holds, for all $n\ge 0$, that*

$$D_h(\nu^{n+1},\nu^n)\le\frac{32\tau^2C_1^2}{\alpha}\quad\text{and}\quad D_h(\mu^{n+1},\mu^n)\le\frac{32\tau^2C_2^2}{\alpha}.$$

Lemma A.2 (Lipschitz continuity and uniform boundedness of F)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{C}, \mathcal{D} \subseteq \mathcal{P}(\mathcal{X})$ be nonempty and convex. Assume that $F \in \mathfrak{C}^1(\mathcal{C} \times \mathcal{D})$ and that Assumption 2.3 holds. Then for all $\nu, \nu' \in \mathcal{C}$ and $\mu, \mu' \in \mathcal{D}$, F is* TV*-Lipschitz, i.e.,*

$$\left|F(\nu', \mu') - F(\nu, \mu)\right| \leq 2\max\{C_1, C_2\}\left(\mathrm{TV}(\nu', \nu) + \mathrm{TV}(\mu', \mu)\right).$$

In particular, for $(\nu^0, \mu^0) \in \mathcal{C} \times \mathcal{D}$ such that $F(\nu^0, \mu^0) < \infty$,

$$\sup_{(\nu,\mu)\in\mathcal{C}\times\mathcal{D}} |F(\nu, \mu)| \leq |F(\nu^0, \mu^0)| + 2(C_1 + C_2),$$

so F is uniformly bounded on $\mathcal{C} \times \mathcal{D}$.

Proof of Theorem 3.5. We start the proof by following the same calculations from Theorem 2.5. For (2), after applying Lemma 2.4 and using convexity-concavity of F, (10) remains unchanged, i.e.,

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

while (12) becomes

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

Adding the previous two inequalities, summing the resulting inequality over $n = 0, 1, ..., N-1$, dividing by N, using (16) and taking maximum over (ν, μ) we arrive at

$$\begin{aligned} \mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right) \leq \frac{1}{N}\sum_{n=0}^{N-1}\left(\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^n - \nu^{n+1})(\mathrm{d}x)\right. \\ \left. + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y)\right) + \frac{1}{N\tau}\left(\sup_{\nu\in\mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu\in\mathcal{D}} D_h(\mu, \mu^0)\right) \\ + \frac{1}{N}\sum_{n=0}^{N-1}\left(F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n)\right) - \frac{1}{N\tau}\sum_{n=0}^{N-1}\left(D_h(\nu^{n+1}, \nu^n) + D_h(\mu^{n+1}, \mu^n)\right), \quad (17) \end{aligned}$$

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 2.5 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 (2). 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 $\mathcal{O}(\tau^3)$.

By Proposition B.4, the first-order conditions for (2) read

$$\begin{cases} \frac{\delta h}{\delta \nu}(\nu^{n+1}, x) - \frac{\delta h}{\delta \nu}(\nu^n, x) = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) + C_{n,1}, \\ \frac{\delta h}{\delta \mu}(\mu^{n+1}, y) - \frac{\delta h}{\delta \mu}(\mu^n, y) = \tau \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y) + C_{n,2}, \end{cases} \quad (18)$$

for all $x \in \mathcal{X}$ ν^{n+1}-a.e. and $y \in \mathcal{X}$ μ^{n+1}-a.e., where $C_{n,1}, C_{n,2} \in \mathbb{R}$. It can be shown directly from Definition 1.2 that

$$\int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu', x) - \frac{\delta h}{\delta \nu}(\nu, x)\right)(\nu' - \nu)(\mathrm{d}x) = D_h(\nu', \nu) + D_h(\nu, \nu'), \quad (19)$$

for all $\nu, \nu' \in \mathcal{C}$, and analogously for $D_h(\mu', \mu) + D_h(\mu, \mu')$. Then, using (18) and (19) we obtain that

$$\begin{aligned} -\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) = \frac{1}{\tau}\int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu^{n+1}, x) - \frac{\delta h}{\delta \nu}(\nu^n, x)\right)(\nu^{n+1} - \nu^n)(\mathrm{d}x) \\ = \frac{1}{\tau}\left(D_h(\nu^{n+1}, \nu^n) + D_h(\nu^n, \nu^{n+1})\right), \quad (20) \end{aligned}$$

and similarly

$$\int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) = \frac{1}{\tau} \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \mu}(\mu^{n+1}, y) - \frac{\delta h}{\delta \mu}(\mu^n, y) \right) (\mu^{n+1} - \mu^n)(\mathrm{d}y)$$
$$= \frac{1}{\tau} \left(D_h(\mu^{n+1}, \mu^n) + D_h(\mu^n, \mu^{n+1}) \right). \quad (21)$$

Therefore, using (20) and (21) in (17), we obtain that

$$\begin{aligned} \mathrm{NI}\left(\frac{1}{N} \sum_{n=0}^{N-1} \nu^{n+1}, \frac{1}{N} \sum_{n=0}^{N-1} \mu^n \right) &\leq \frac{1}{N\tau} \left(\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) \right) \\ &+ \frac{1}{N\tau} \sum_{n=0}^{N-1} \left(D_h(\nu^{n+1}, \nu^n) + D_h(\nu^n, \nu^{n+1}) + D_h(\mu^{n+1}, \mu^n) + D_h(\mu^n, \mu^{n+1}) \right) \\ &+ \frac{1}{N} \sum_{n=0}^{N-1} \left(F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n) \right) - \frac{1}{N\tau} \sum_{n=0}^{N-1} \left(D_h(\nu^{n+1}, \nu^n) + D_h(\mu^{n+1}, \mu^n) \right). \quad (22) \end{aligned}$$

Then, we observe that

$$D_h(\nu^n, \nu^{n+1}) = \frac{1}{2} \left(D_h(\nu^n, \nu^{n+1}) - D_h(\nu^{n+1}, \nu^n) \right) + \frac{1}{2} \left(D_h(\nu^n, \nu^{n+1}) + D_h(\nu^{n+1}, \nu^n) \right), \quad (23)$$

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

$$\begin{aligned} F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n) &= \frac{1}{2} \left(F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n) \right) \\ &+ \frac{1}{2} \left(F(\nu^{n+1}, \mu^n) - F(\nu^{n+1}, \mu^{n+1}) + F(\nu^{n+1}, \mu^{n+1}) - F(\nu^n, \mu^n) \right). \quad (24) \end{aligned}$$

Therefore, putting (23) and (24) into (22) gives

$$\begin{aligned} \mathrm{NI}\left(\frac{1}{N} \sum_{n=0}^{N-1} \nu^{n+1}, \frac{1}{N} \sum_{n=0}^{N-1} \mu^n \right) &\leq \frac{1}{N\tau} \left(\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) \right) \\ &+ \frac{1}{2N\tau} \sum_{n=0}^{N-1} \left(D_h(\nu^n, \nu^{n+1}) - D_h(\nu^{n+1}, \nu^n) + D_h(\mu^n, \mu^{n+1}) - D_h(\mu^{n+1}, \mu^n) \right) \\ &+ \frac{1}{2N} \sum_{n=0}^{N-1} \left(\frac{1}{\tau} (D_h(\nu^n, \nu^{n+1}) + D_h(\nu^{n+1}, \nu^n)) + F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n) \right) \\ &+ \frac{1}{2N} \sum_{n=0}^{N-1} \left(\frac{1}{\tau} (D_h(\mu^n, \mu^{n+1}) + D_h(\mu^{n+1}, \mu^n)) + F(\nu^{n+1}, \mu^n) - F(\nu^{n+1}, \mu^{n+1}) \right. \\ &\left. + F(\nu^{n+1}, \mu^{n+1}) - F(\nu^n, \mu^n) \right). \quad (25) \end{aligned}$$

Combining the fact that $\nu \mapsto F(\nu, \mu)$ is L_ν-smooth relative to h with the first-order condition (18), we have that

$$\begin{aligned} F(\nu^{n+1}, \mu^n) - F(\nu^n, \mu^n) &\leq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) + L_\nu D_h(\nu^{n+1}, \nu^n) \\ &= -\frac{1}{\tau} \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu^{n+1}, x) - \frac{\delta h}{\delta \nu}(\nu^n, x) \right) (\nu^{n+1} - \nu^n)(\mathrm{d}x) + L_\nu D_h(\nu^{n+1}, \nu^n) \\ &= -\frac{1}{\tau} \left(D_h(\nu^{n+1}, \nu^n) + D_h(\nu^n, \nu^{n+1}) \right) + L_\nu D_h(\nu^{n+1}, \nu^n), \quad (26) \end{aligned}$$

where the last equality follows from (19).

Similarly, using L_μ-smoothness of $\mu \mapsto F(\nu, \mu)$ relative to h together with (18), we can show that

$$F(\nu^{n+1}, \mu^n) - F(\nu^{n+1}, \mu^{n+1}) + \frac{1}{\tau}\left(D_h(\mu^n, \mu^{n+1}) + D_h(\mu^{n+1}, \mu^n)\right) \leq L_\mu D_h(\mu^{n+1}, \mu^n). \quad (27)$$

Therefore, using (26) and (27) in (25), and recalling that $L = \max\{L_\nu, L_\mu\}$ gives

$$\begin{aligned}
\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1}\mu^n\right) &\leq \frac{1}{N\tau}\left(\sup_{\nu\in\mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu\in\mathcal{D}} D_h(\mu, \mu^0)\right) \\
&+ \frac{1}{2N\tau}\sum_{n=0}^{N-1}\left(D_h(\nu^n, \nu^{n+1}) - D_h(\nu^{n+1}, \nu^n) + D_h(\mu^n, \mu^{n+1}) - D_h(\mu^{n+1}, \mu^n)\right) \\
&+ \frac{L}{2N}\sum_{n=0}^{N-1}\left(D_h(\nu^{n+1}, \nu^n) + D_h(\mu^{n+1}, \mu^n)\right) + \frac{1}{2N}\left(F\left(\nu^N, \mu^N\right) - F(\nu^0, \mu^0)\right). \quad (28)
\end{aligned}$$

Since, by Lemma A.1,

$$D_h(\nu^{n+1}, \nu^n) + D_h(\mu^{n+1}, \mu^n) \leq \frac{32\tau^2\kappa_2}{\alpha}, \quad (29)$$

where $\kappa_2 \coloneqq C_1^2 + C_2^2$, 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 $\tau = \mathcal{O}\left(N^{-1/3}\right)$, and since by Lemma A.2, $\left|F\left(\nu^N, \mu^N\right)\right| \leq |F(\nu^0, \mu^0)| + 2(C_1 + C_2)$, 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 G.5 together with Assumptions 2.3 and 3.4.

If we denote $f^n \coloneqq \frac{\delta h}{\delta \nu}(\nu^n, \cdot)$, for any $n \geq 0$, then by Lemma G.5, we have that $D_h(\nu^n, \nu^{n+1}) = D_{h^*}(f^{n+1}, f^n)$. For any $\varepsilon \in [0, 1]$ denote $f^{\varepsilon,n} = \varepsilon f^{n+1} + (1-\varepsilon)f^n$ and $\phi(\varepsilon) = h^*(f^{\varepsilon,n})$. Note that $\phi(0) = h^*(f^n)$ and $\phi(1) = h^*(f^{n+1})$. We have

$$\phi'(\varepsilon) = \nabla_{\mathcal{F}} h^*(f^{\varepsilon,n})[f^{n+1} - f^n], \quad \phi''(\varepsilon) = \nabla^2_{\mathcal{F}} h^*(f^{\varepsilon,n})[f^{n+1} - f^n][f^{n+1} - f^n].$$

Note that $\phi'(0) = \nabla_{\mathcal{F}} h^*(f^n)[f^{n+1} - f^n]$. By the fundamental theorem of calculus and integration by parts, we have

$$\begin{aligned}
\phi(1) - \phi(0) = \int_0^1 \phi'(\varepsilon)\mathrm{d}\varepsilon &= [(\varepsilon - 1)\phi'(\varepsilon)]|_{\varepsilon=0}^{\varepsilon=1} - \int_0^1 (\varepsilon - 1)\phi''(\varepsilon)\mathrm{d}\varepsilon \\
&= \phi'(0) + \int_0^1 (1-\varepsilon)\phi''(\varepsilon)\mathrm{d}\varepsilon.
\end{aligned}$$

Hence,

$$h^*(f^{n+1}) - h^*(f^n) - \nabla_{\mathcal{F}} h^*(f^n)[f^{n+1} - f^n] = \int_0^1 (1-\varepsilon)\nabla^2_{\mathcal{F}} h^*(f^{\varepsilon,n})[f^{n+1} - f^n][f^{n+1} - f^n]\mathrm{d}\varepsilon$$

By Definition 3.2, we have that

$$D_{h^*}(f^{n+1}, f^n) = \int_0^1 (1-\varepsilon)\nabla^2_{\mathcal{F}} h^*(f^{\varepsilon,n})[f^{n+1} - f^n][f^{n+1} - f^n]\mathrm{d}\varepsilon$$

Similarly, by Lemma G.5, we have that $D_h(\nu^{n+1}, \nu^n) = D_{h^*}(f^n, f^{n+1})$, and hence

$$D_{h^*}(f^n, f^{n+1}) = \int_0^1 (1-\varepsilon)\nabla^2_{\mathcal{F}} h^*(f^{1-\varepsilon,n})[f^{n+1} - f^n][f^{n+1} - f^n]\mathrm{d}\varepsilon.$$

Therefore, we obtain that

$$\begin{aligned}
&D_{h^*}(f^{n+1}, f^n) - D_{h^*}(f^n, f^{n+1}) \\
&= \int_0^1 (1-\varepsilon) \left(\nabla^2_{\mathcal{F}} h^*(f^{\varepsilon,n}) - \nabla^2_{\mathcal{F}} h^*(f^{1-\varepsilon,n})\right)[f^{n+1} - f^n][f^{n+1} - f^n] \mathrm{d}\varepsilon.
\end{aligned}$$

Note that $f^{\varepsilon,n} - f^{1-\varepsilon,n} = (2\varepsilon - 1)(f^{n+1} - f^n)$. If we denote $g^{\varepsilon,\gamma,n} = f^{1-\varepsilon,n} + \gamma(f^{\varepsilon,n} - f^{1-\varepsilon,n})$, then applying the fundamental theorem of calculus again gives

$$\begin{aligned}
&D_{h^*}(f^{n+1}, f^n) - D_{h^*}(f^n, f^{n+1}) \\
&= \int_0^1 (1-\varepsilon)(2\varepsilon - 1) \int_0^1 \nabla^3_{\mathcal{F}} h^*(g^{\varepsilon,\gamma,n})[f^{n+1} - f^n][f^{n+1} - f^n][f^{n+1} - f^n] \mathrm{d}\gamma \mathrm{d}\varepsilon.
\end{aligned}$$

Using Assumption 3.4, we further obtain

$$\begin{aligned}
D_{h^*}(f^{n+1}, f^n) - D_{h^*}(f^n, f^{n+1}) &\leq L_{h^*} \|f^{n+1} - f^n\|^3_\infty \int_0^1 |(1-\varepsilon)(2\varepsilon - 1)| \mathrm{d}\varepsilon \\
&= \frac{L_{h^*}}{4} \|f^{n+1} - f^n\|^3_\infty,
\end{aligned}$$

The first-order condition for the minimizing player in (18) can be rewritten as

$$f^{n+1}(x) - f^n(x) = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x), \tag{30}$$

for all $x \in \mathcal{X}$ ν^{n+1}-a.e. By Assumption 2.3, there exists $C_1 > 0$ such that $\left\| \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) \right\|_\infty \leq C_1$, for any $n \geq 0$. Hence, we obtain that

$$\begin{aligned}
D_{h^*}(f^{n+1}, f^n) - D_{h^*}(f^n, f^{n+1}) &\leq \frac{L_{h^*}}{4} \|f^{n+1} - f^n\|^3_\infty \\
&= \frac{L_{h^*}}{4} \tau^3 \left\| \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) \right\|^3_\infty \leq \frac{L_{h^*}}{4} \tau^3 C_1^3,
\end{aligned}$$

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

$$\begin{aligned}
D_{h^*}(g^{n+1}, g^n) - D_{h^*}(g^n, g^{n+1}) &\leq \frac{L_{h^*}}{4} \|g^{n+1} - g^n\|^3_\infty \\
&= \frac{L_{h^*}}{4} \tau^3 \left\| \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, \cdot) \right\|^3_\infty \leq \frac{L_{h^*}}{4} \tau^3 C_2^3,
\end{aligned}$$

where $C_2 > 0$ exists due to Assumption 2.3.

Set $\kappa_1 \coloneqq \frac{1}{4}\left(C_1^3 + C_2^3\right) > 0$. Then,

$$D_h(\nu^n, \nu^{n+1}) - D_h(\nu^{n+1}, \nu^n) + D_h(\mu^n, \mu^{n+1}) - D_h(\mu^{n+1}, \mu^n) \leq \kappa_1 L_{h^*} \tau^3. \tag{31}$$

Hence, using (29), (31) and Lemma A.2, estimate (28) becomes

$$\begin{aligned}
&\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1} \nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1} \mu^n\right) \leq \frac{1}{N\tau}\left(\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0)\right) \\
&+ \frac{1}{2N\tau} \sum_{n=0}^{N-1} \left(\left(D_h(\nu^n, \nu^{n+1}) - D_h(\nu^{n+1}, \nu^n)\right) + \left(D_h(\mu^n, \mu^{n+1}) - D_h(\mu^{n+1}, \mu^n)\right) \right) \\
&+ \frac{L}{2N} \sum_{n=0}^{N-1} \left(D_h(\nu^{n+1}, \nu^n) + D_h(\mu^{n+1}, \mu^n)\right) + \frac{1}{2N}\left(F\left(\nu^N, \mu^N\right) - F(\nu^0, \mu^0)\right) \\
&= \frac{1}{N\tau}\left(\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0)\right) + \left(\frac{\kappa_1 L_{h^*}}{2} + \frac{16\kappa_2 L}{\alpha}\right)\tau^2 + \frac{|F(\nu^0, \mu^0)| + C_1 + C_2}{N}.
\end{aligned}$$

Minimizing the right-hand side over τ amounts to taking

$$\tau = \left(\frac{\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0)}{2N}\right)^{1/3} \left(\frac{\kappa_1 L_{h^*}}{2} + \frac{16\kappa_2 L}{\alpha}\right)^{-1/3},$$

and since $\frac{1}{N} \leq \frac{1}{N^{2/3}}$, for any $N \geq 1$, it follows that

$$\mathrm{NI}\left(\frac{1}{N}\sum_{n=0}^{N-1}\nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1}\mu^{n}\right) \leq \frac{1}{(2N)^{2/3}}\left(3\left(\sup_{\nu\in\mathcal{C}} D_h(\nu,\nu^0) + \sup_{\mu\in\mathcal{D}} D_h(\mu,\mu^0)\right)^{2/3} \times \right.$$
$$\left. \times \left(\frac{\kappa_1 L_{h^*}}{2} + \frac{16\kappa_2 L}{\alpha}\right)^{1/3} + 4^{1/3}\left(|F(\nu^0,\mu^0)| + C_1 + C_2\right)\right).$$

□

B Proofs of additional results

In this section, we present the proofs of the additional results of the paper. We start with examples of functions h and the corresponding sets $\mathcal{E}$ such that Assumption 2.1 is satisfied. We continue with the proofs of Lemmas A.1, A.2 and 2.4, which play a key role in proving the main results. Finally, we prove some auxiliary results.

B.1 Examples satisfying Assumption 2.1

Example B.1 (Relative entropy)**.** *Suppose that h is the relative entropy, i.e., $h(m) := \int_{\mathcal{X}} \frac{m(x)}{\pi(x)} \log \frac{m(x)}{\pi(x)} \pi(x)\mathrm{d}x$, where $m, \pi \in \mathcal{P}_\lambda(\mathcal{X})$, i.e., they are absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and π is a fixed reference probability measure on $\mathcal{P}_\lambda(\mathcal{X})$. Fix $\bar{\beta} > 0$ and define $\mathcal{E}_{\bar{\beta}} := \left\{ m \in \mathcal{P}_\pi(\mathcal{X}) : \left\| \log \frac{m(\cdot)}{\pi(\cdot)} \right\|_{L^\infty(\mathcal{X})} < \bar{\beta} \right\}$. Let $\mathcal{C} = \mathcal{D} := \mathcal{E}_{\bar{\beta}}$. Then $\mathcal{E}_{\bar{\beta}}$ is convex. Moreover, it is proved in [30, Proposition 2.17] that h admits the flat derivative*

$$\frac{\delta h}{\delta m}(m,x) = \log \frac{m(x)}{\pi(x)} - h(m), \tag{32}$$

on $\mathcal{E}_{\bar{\beta}}$, and for all $m, m' \in \mathcal{E}_{\bar{\beta}}$, the Bregman divergence $D_h(m', m)$ is in fact the Kullback-Leibler divergence (or relative entropy) $\mathrm{KL}(m', m)$. Therefore, by Pinsker's inequality, that is, $\mathrm{TV}^2(m', m) \leq \frac{1}{2}\mathrm{KL}(m', m)$, Assumption 2.1 holds with $\alpha = 4$. A related observation appears in [6, Remark 5]. A sufficient condition for the relative entropy (or the entropy) to admit a flat derivative in L^∞ at m is that there exists $\kappa_0, \kappa_1 > 0$ such that $\kappa_0 \leq \frac{m(\cdot)}{\pi(\cdot)} \leq \kappa_1$ a.e. on $\mathcal{X}$, which is precisely a similar condition to the one defining the class $\mathcal{E}_{\bar{\beta}}$.

Example B.2 (χ^2-divergence)**.** *Suppose that h is the χ^2-divergence, i.e., $h(m) := \frac{1}{2}\int_{\mathcal{X}} \left(\frac{m(x)}{\pi(x)} - 1\right)^2 \pi(x)\mathrm{d}x$, where $m, \pi \in \mathcal{P}_\lambda(\mathcal{X})$. Let $L^2_\pi(\mathcal{X})$ be the set of square integrable functions on $\mathcal{X}$ with respect to π. Fix $\bar{\eta} > 0$ and define $\mathcal{F}_{\bar{\eta}} := \left\{ m \in \mathcal{P}_\pi(\mathcal{X}) : \frac{m(\cdot)}{\pi(\cdot)} > 0 \ \pi\text{-a.e.}, \left\| \frac{m(\cdot)}{\pi(\cdot)} \right\|_{L^2_\pi(\mathcal{X})} \leq \bar{\eta} \right\}$. Let $\mathcal{C} = \mathcal{D} := \mathcal{F}_{\bar{\eta}}$. Note that $\mathcal{F}_{\bar{\eta}}$ is convex. Moreover, it is proved in [30, Proposition 2.20] that h admits the flat derivative*

$$\frac{\delta h}{\delta m}(m,x) = \frac{m(x)}{\pi(x)} - \int_{\mathbb{R}^d} \frac{m(z)}{\pi(z)} m(z)\mathrm{d}z,$$

on $\mathcal{F}_{\bar{\eta}}$, and for all $m, m' \in \mathcal{F}_{\bar{\eta}}$, the Bregman divergence $D_h(m', m)$ is in fact the L^2-distance $\frac{1}{2}\left\| \frac{m'(\cdot)}{\pi(\cdot)} - \frac{m(\cdot)}{\pi(\cdot)} \right\|^2_{L^2_\pi(\mathcal{X})}$. Since $\pi \in \mathcal{P}_\lambda(\mathcal{X})$, the Cauchy-Schwarz inequality implies $\frac{1}{2}\mathrm{TV}^2(m', m) \leq D_h(m', m)$. Thus, Assumption 2.1 holds with $\alpha = 1$.

B.2 Proof of Lemma A.1

Proof of Lemma A.1. We will only prove the lemma for Algorithm 1 since the argument for Algorithm 2 is almost identical. From L_ν-relative smoothness and the definition of ν^{n+1} in (1), for any $\nu \in \mathcal{C}$,

it follows that

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

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

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

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

$$
\begin{aligned}
\frac{1}{2\tau} D_h(\nu^{n+1}, \nu^n) &\leq F(\nu^n, \mu^n) - F(\nu^{n+1}, \mu^n) \\
&= \int_0^1 \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^{n+1} + \varepsilon(\nu^n - \nu^{n+1}), \mu^n, x)(\nu^n - \nu^{n+1})(\mathrm{d}x)\mathrm{d}\varepsilon \\
&\leq 2C_1 \operatorname{TV}(\nu^{n+1}, \nu^n) \\
&\leq 2C_1 \sqrt{\frac{2}{\alpha} D_h(\nu^{n+1}, \nu^n)},
\end{aligned}
$$

where the penultimate inequality follows from Assumption 2.3 and 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

$$
D_h(\nu^{n+1}, \nu^n) \leq \frac{32\tau^2 C_1^2}{\alpha}.
$$

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

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

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

$$
\frac{1}{2\tau} D_h(\mu^{n+1}, \mu^n) \leq F(\nu^n, \mu^{n+1}) - F(\nu^n, \mu^n) \leq 2C_2 \sqrt{\frac{2}{\alpha} D_h(\mu^{n+1}, \mu^n)},
$$

where the last inequality follows from Assumptions 2.3 and 2.1. Hence, since $D_h(\mu^{n+1}, \mu^n) \geq 0$, for all $n \geq 0$, we obtain that

$$
D_h(\mu^{n+1}, \mu^n) \leq \frac{32\tau^2 C_2^2}{\alpha}.
$$

□

B.3 Proof of Lemma A.2

Proof of Lemma A.2. Fix $\mu \in \mathcal{D}$ and $\nu, \nu' \in \mathcal{C}$, and define $\nu_\lambda := (1 - \lambda)\nu + \lambda\nu'$, $\lambda \in [0, 1]$, which stays in $\mathcal{C}$ by convexity. By Remark G.2 applied to $\lambda \mapsto F(\nu_\lambda, \mu)$,

$$
F(\nu', \mu) - F(\nu, \mu) = \int_0^1 \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu_\lambda, \mu, x)(\nu' - \nu)(\mathrm{d}x)\mathrm{d}\lambda.
$$

Using the dual characterization of total variation (i.e., $|\int \phi \,\mathrm{d}(\nu' - \nu)| \leq 2\|\phi\|_\infty \,\mathrm{TV}(\nu', \nu)$ for bounded ϕ), and Assumption 2.3, we obtain

$$|F(\nu', \mu) - F(\nu, \mu)| \leq 2C_1 \,\mathrm{TV}(\nu', \nu).$$

Similarly, fixing $\nu' \in \mathcal{C}$ and defining $\mu_\lambda := (1-\lambda)\mu + \lambda\mu' \in \mathcal{D}$, we get $|F(\nu', \mu') - F(\nu', \mu)| \leq 2C_2 \,\mathrm{TV}(\mu', \mu)$. Combining the two bounds with the triangle inequality yields

$$\begin{aligned} |F(\nu', \mu') - F(\nu, \mu)| &\leq |F(\nu', \mu') - F(\nu', \mu)| + |F(\nu', \mu) - F(\nu, \mu)| \\ &\leq 2C_2 \,\mathrm{TV}(\mu', \mu) + 2C_1 \,\mathrm{TV}(\nu', \nu), \end{aligned}$$

proving that F is TV-Lipschitz. The uniform bound on F follows by applying this estimate with $(\nu', \mu') = (\nu, \mu)$ and $(\nu, \mu) = (\nu^0, \mu^0)$, noting that $\mathrm{TV}(\cdot, \cdot) \leq 1$ on probability measures and taking the supremum over $\mathcal{C} \times \mathcal{D}$. □

B.4 Proof of Lemma 2.4

Proof of Lemma 2.4. Fix $\nu \in \mathcal{E}$. Define $\mathcal{G} : \mathcal{E} \to \mathbb{R}$ by $\mathcal{G}(\nu) = G(\nu) + D_h(\nu, \mu)$ for all $\mu \in \mathcal{E}$. Since G and h are convex and flat differentiable, $\mathcal{G}$ is flat differentiable. As $\bar{\nu}$ minimises $\mathcal{G}$ over $\mathcal{E}$,

$$\int_{\mathcal{X}} \frac{\delta \mathcal{G}}{\delta \nu}(\bar{\nu}, x)(\nu - \bar{\nu})(\mathrm{d}x) = \lim_{\varepsilon \searrow 0} \frac{\mathcal{G}(\bar{\nu} + \varepsilon(\nu - \bar{\nu})) - \mathcal{G}(\bar{\nu})}{\varepsilon} \geq 0,$$

which implies that

$$D_{\mathcal{G}}(\nu, \bar{\nu}) = \mathcal{G}(\nu) - \mathcal{G}(\bar{\nu}) - \int \frac{\delta \mathcal{G}}{\delta m}(\bar{\nu}, x)(\nu - \bar{\nu})(\mathrm{d}x) \leq \mathcal{G}(\nu) - \mathcal{G}(\bar{\nu}).$$

Adding $\mathcal{G}(\bar{\nu})$ to both sides of the inequality and using the definition of $\mathcal{G}$ gives

$$\begin{aligned} G(\nu) + D_h(\nu, \mu) &\geq D_{\mathcal{G}}(\nu, \bar{\nu}) + G(\bar{\nu}) + D_h(\bar{\nu}, \mu) \\ &= D_G(\nu, \bar{\nu}) + D_{D_h(\cdot, \mu)}(\nu, \bar{\nu}) + G(\bar{\nu}) + D_h(\bar{\nu}, \mu) \\ &\geq D_h(\nu, \bar{\nu}) + G(\bar{\nu}) + D_h(\bar{\nu}, \mu), \end{aligned}$$

where the equality follows from the linearity of the Bregman divergence and the last inequality follows since $D_G(\nu, \bar{\nu}) \geq 0$ by convexity of G and the identity $D_{D_h(\cdot, \mu)}(\nu, \bar{\nu}) = D_h(\nu, \bar{\nu})$. □

B.5 Proofs of auxiliary results

In this subsection, we start by proving the convexity characterization h via its flat derivative.

Lemma B.3 (Strong convexity of h)**.** *Let Assumption 2.1 hold. Then there exists $\alpha > 0$ such that for any $\nu, \nu' \in \mathcal{E}$,*

$$h(\nu') - h(\nu) \geq \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu, x)(\nu' - \nu)(\mathrm{d}x) + \frac{\alpha}{2} \,\mathrm{TV}^2(\nu', \nu).$$

Proof. Let $\varepsilon \in [0, 1]$ and $\nu^\varepsilon = \nu + \varepsilon(\nu' - \nu)$. Since h is strongly convex, and by Definition 1.1,

$$\begin{aligned} \varepsilon \left(h(\nu') - h(\nu)\right) - \frac{\alpha}{2}\varepsilon(1-\varepsilon) \,\mathrm{TV}^2(\nu', \nu) &\geq h(\nu^\varepsilon) - h(\nu) \\ &= \varepsilon \int_0^1 \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu^{s\varepsilon}, x)(\nu' - \nu)(\mathrm{d}x)\mathrm{d}s. \end{aligned}$$

Dividing by ε and passing to the limit $\varepsilon \downarrow 0$ via dominated convergence theorem gives the conclusion since $\frac{\delta h}{\delta \nu}$ is bounded and continuous. □

Following [6, Lemma 2], we prove that the mirror updates in Algorithms 1 and 2 satisfy iterative schemes on the dual space. Before presenting the proof, we define the relative subdifferential of h over $\mathcal{K} \subseteq \mathcal{P}(\mathcal{X})$ at $\rho \in \mathcal{P}(\mathcal{X})$ by

$$\partial_{\mathcal{K}} h(\rho) := \left\{ q : \int_{\mathcal{X}} q(x)(\eta - \rho)(\mathrm{d}x) \leq \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\rho, x)(\eta - \rho)(\mathrm{d}x), \forall\, \eta \in \mathcal{K} \right\}.$$

Proposition B.4 (MDA dual iteration)**.** *For each $n \geq 0$, let $F(\cdot, \mu^n) \in \mathfrak{C}^1(\mathcal{C})$, $F(\nu^n, \cdot) \in \mathfrak{C}^1(\mathcal{D})$. Moreover, let Assumption 2.1 hold. Then the minimizer-maximizer pair $(\nu^{n+1}, \mu^{n+1}) \in \mathcal{C} \times \mathcal{D}$ of each one of Algorithms 1 and 2 satisfies the corresponding dual iterative update*

$$\begin{cases} \frac{\delta h}{\delta \nu}(\nu^n, \cdot) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) \in \partial_{\mathcal{C}} h(\nu^{n+1}), \\ \frac{\delta h}{\delta \mu}(\mu^n, \cdot) + \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, \cdot) \in \partial_{\mathcal{D}} h(\mu^{n+1}), \end{cases} \tag{33}$$

$$\begin{cases} \frac{\delta h}{\delta \nu}(\nu^n, \cdot) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) \in \partial_{\mathcal{C}} h(\nu^{n+1}), \\ \frac{\delta h}{\delta \mu}(\mu^n, \cdot) + \tau \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, \cdot) \in \partial_{\mathcal{D}} h(\mu^{n+1}). \end{cases} \tag{34}$$

Thus, if $\partial_{\mathcal{C}} h(\nu^{n+1}) = \left\{ \frac{\delta h}{\delta \nu}(\nu^{n+1}, \cdot) + c_{\nu,n} : c_{\nu,n} \in \mathbb{R} \right\}$ *and* $\partial_{\mathcal{D}} h(\mu^{n+1}) = \left\{ \frac{\delta h}{\delta \mu}(\mu^{n+1}, \cdot) + c_{\mu,n} : c_{\mu,n} \in \mathbb{R} \right\}$, (33) *and* (34)*, respectively, correspond to*

$$\begin{cases} \frac{\delta h}{\delta \nu}(\nu^{n+1}, \cdot) - \frac{\delta h}{\delta \nu}(\nu^n, \cdot) = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) + c_{n,1}, \quad \nu^{n+1} - \textit{a.e.}, \\ \frac{\delta h}{\delta \mu}(\mu^{n+1}, \cdot) - \frac{\delta h}{\delta \mu}(\mu^n, \cdot) = \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, \cdot) + c_{n,2}, \quad \mu^{n+1} - \textit{a.e.}, \end{cases}$$

$$\begin{cases} \frac{\delta h}{\delta \nu}(\nu^{n+1}, \cdot) - \frac{\delta h}{\delta \nu}(\nu^n, \cdot) = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) + c_{n,3}, \quad \nu^{n+1} - \textit{a.e.}, \\ \frac{\delta h}{\delta \mu}(\mu^{n+1}, \cdot) - \frac{\delta h}{\delta \mu}(\mu^n, \cdot) = \tau \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, \cdot) + c_{n,4}, \quad \mu^{n+1} - \textit{a.e.}, \end{cases}$$

where $c_{n,1}, c_{n,2}, c_{n,3}, c_{n,4} \in \mathbb{R}$.

Proof. We present the proof only for Algorithm 1, as the proof for Algorithm 2 is identical. For convenience, define

$$G(\nu) \coloneqq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x), \quad \nu \in \mathcal{C}.$$

Since ν^{n+1} is the minimizer in Algorithm 1, we have

$$G(\nu^{n+1}) + \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n) \leq G(\nu) + \frac{1}{\tau} D_h(\nu, \nu^n), \quad \forall \nu \in \mathcal{C}.$$

Because $\mathcal{C}$ is convex, fix any $\tilde{\nu} \in \mathcal{C}$ and take $\nu = \nu^{n+1} + \varepsilon(\tilde{\nu} - \nu^{n+1}) \in \mathcal{C}$. Then

$$G(\nu^{n+1}) + \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n) \leq G(\nu^{n+1} + \varepsilon(\tilde{\nu} - \nu^{n+1})) + \frac{1}{\tau} D_h(\nu^{n+1} + \varepsilon(\tilde{\nu} - \nu^{n+1}), \nu^n),$$

which, using linearity of G and rearranging, becomes

$$\varepsilon(G(\tilde{\nu}) - G(\nu^{n+1})) + \frac{1}{\tau} \left(D_h(\nu^{n+1} + \varepsilon(\tilde{\nu} - \nu^{n+1}), \nu^n) - D_h(\nu^{n+1}, \nu^n) \right) \geq 0.$$

Dividing by ε and letting $\varepsilon \searrow 0$, Definition 1.1 yields

$$G(\tilde{\nu}) - G(\nu^{n+1}) + \frac{1}{\tau} \int_{\mathcal{X}} \frac{\delta D_h(\cdot, \nu^n)}{\delta \nu}(\nu^{n+1}, x)(\tilde{\nu} - \nu^{n+1})(\mathrm{d}x) \geq 0.$$

By the definition of Bregman divergence and flat derivative,

$$\frac{\delta D_h(\cdot, \nu^n)}{\delta \nu}(\nu^{n+1}, x) = \frac{\delta h}{\delta \nu}(\nu^{n+1}, x) - \frac{\delta h}{\delta \nu}(\nu^n, x).$$

Hence,

$$\int_{\mathcal{X}} \left(\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) + \frac{1}{\tau} \left(\frac{\delta h}{\delta \nu}(\nu^{n+1}, x) - \frac{\delta h}{\delta \nu}(\nu^n, x) \right) \right) (\tilde{\nu} - \nu^{n+1})(\mathrm{d}x) \geq 0, \ \forall \tilde{\nu} \in \mathcal{C}.$$

Rearranging this inequality gives

$$\frac{\delta h}{\delta \nu}(\nu^n, \cdot) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) \in \partial_{\mathcal{C}} h(\nu^{n+1})$$

An analogous argument gives the optimality condition for the maximizer μ^{n+1}. □

We show that the relative subdifferential of h is a singleton for both the relative entropy on $\mathcal{E}_{\bar{\beta}}$ (Example B.1) and the χ^2-divergence on $\mathcal{F}_{\bar{\eta}}$ (Example B.2).

Lemma B.5 (Singleton subdifferential for relative entropy)**.** *Let $\pi \in \mathcal{P}(\mathcal{X})$ be a reference probability measure and define $h(\nu) := \mathrm{KL}(\nu|\pi)$. For $\bar{\beta} > 0$, consider the convex feasible set $\mathcal{E}_{\bar{\beta}} := \left\{ m \in \mathcal{P}_\pi(\mathcal{X}) : \left\| \log \frac{\mathrm{d}m}{\mathrm{d}\pi} \right\|_{L^\infty(\pi)} < \bar{\beta} \right\}$. Then*

$$\partial_{\mathcal{E}_{\bar{\beta}}} h(\nu) = \left\{ \log \frac{\mathrm{d}\nu}{\mathrm{d}\pi}(\cdot) + c : c \in \mathbb{R} \right\}. \tag{35}$$

Proof. By construction of the feasible set, $\nu \in \mathcal{E}_{\bar{\beta}}$ satisfies the strict slackness condition

$$\left\| \log \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \right\|_{L^\infty(\pi)} \leq \bar{\beta} - \delta \tag{36}$$

for some $\delta > 0$. Let $p_\nu := \log \frac{\mathrm{d}\nu}{\mathrm{d}\pi}$ be the unnormalized flat derivative of h at ν. By definition,

$$\partial_{\mathcal{E}_{\bar{\beta}}} h(\nu) := \left\{ q \in L^\infty(\pi) : \int_{\mathcal{X}} q \, \mathrm{d}(\eta - \nu) \leq \int_{\mathcal{X}} p_\nu \, \mathrm{d}(\eta - \nu), \forall \eta \in \mathcal{E}_{\bar{\beta}} \right\}.$$

Fix $q \in \partial_{\mathcal{E}_{\bar{\beta}}} h(\nu)$. Then

$$\int_{\mathcal{X}} (q - p_\nu) \, \mathrm{d}(\eta - \nu) \leq 0, \quad \forall \eta \in \mathcal{E}_{\bar{\beta}}. \tag{37}$$

Let $u \in L^\infty(\pi)$ be arbitrary with $\int_{\mathcal{X}} u \mathrm{d}\pi = 0$. Condition (36) implies

$$e^{-(\bar{\beta}-\delta)} \leq \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \leq e^{\bar{\beta}-\delta}, \quad \pi\text{-a.e.}$$

Therefore, for sufficiently small $t > 0$,

$$e^{-\bar{\beta}} \leq \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \pm tu \leq e^{\bar{\beta}}, \quad \pi\text{-a.e.}$$

For example, one could take $t\|u\|_{L^\infty(\pi)} < \min\{e^{-(\bar{\beta}-\delta)} - e^{-\bar{\beta}}, e^{\bar{\beta}} - e^{\bar{\beta}-\delta}\}$. Since $\int u \mathrm{d}\pi = 0$, the measures $\frac{\mathrm{d}\eta_\pm}{\mathrm{d}\pi} := \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \pm tu$ belong to $\mathcal{E}_{\bar{\beta}}$. Applying (37) first to η_+ and then to η_- gives

$$t \int_{\mathcal{X}} (q - p_\nu) u \, \mathrm{d}\pi \leq 0, \quad -t \int_{\mathcal{X}} (q - p_\nu) u \, \mathrm{d}\pi \leq 0.$$

Hence

$$\int_{\mathcal{X}} (q - p_\nu) u \, \mathrm{d}\pi = 0, \tag{38}$$

for every $u \in L^\infty(\pi)$ such that $\int_{\mathcal{X}} u \, \mathrm{d}\pi = 0$. Since $q - p_\nu \in L^\infty(\pi)$, we may take $u = q - p_\nu - \int_{\mathcal{X}} (q - p_\nu) \, \mathrm{d}\pi$ in (38). This gives

$$0 = \int_{\mathcal{X}} \left((q - p_\nu) - \int_{\mathcal{X}} (q - p_\nu) \mathrm{d}\pi \right)^2 \mathrm{d}\pi.$$

Hence $q - p_\nu$ is constant π-a.e., and therefore $q = p_\nu + c$ for some $c \in \mathbb{R}$. This proves one inclusion in (35).

Conversely, if $q = p_\nu + c$ for some $c \in \mathbb{R}$, then for every $\eta \in \mathcal{E}_{\bar{\beta}}$,

$$\int_{\mathcal{X}} q \, \mathrm{d}(\eta - \nu) = \int_{\mathcal{X}} p_\nu \, \mathrm{d}(\eta - \nu).$$

Thus $q \in \partial_{\mathcal{E}_{\bar{\beta}}} h(\nu)$, proving (35). □

Lemma B.6 (Singleton subdifferential for χ^2)**.** *Let $\pi \in \mathcal{P}(\mathcal{X})$ be a reference probability measure and define $h(\nu) := \frac{1}{2} \int_{\mathcal{X}} \left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi} - 1 \right)^2 \mathrm{d}\pi$. For $\bar{\eta} > 0$, consider the convex feasible set $\mathcal{F}_{\bar{\eta}} := \left\{ \nu \in \mathcal{P}_\pi(\mathcal{X}) : \frac{\mathrm{d}\nu}{\mathrm{d}\pi} > 0 \ \pi\text{-a.e.}, \| \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \|_{L^2(\pi)} \leq \bar{\eta} \right\}$. Then*

$$\partial_{\mathcal{F}_{\bar{\eta}}} h(\nu) = \left\{ \frac{\mathrm{d}\nu}{\mathrm{d}\pi} + c : c \in \mathbb{R} \right\}, \tag{39}$$

Proof. For $\nu \in \mathcal{F}_{\bar{\eta}}$, let $p_\nu := \frac{\mathrm{d}\nu}{\mathrm{d}\pi}$, which is an unnormalized representative of the flat derivative of h at ν. By definition of relative subdifferential,

$$\partial_{\mathcal{F}_{\bar{\eta}}} h(\nu) := \left\{ q \in L^2(\pi) : \int_{\mathcal{X}} q \, \mathrm{d}(\eta - \nu) \leq \int_{\mathcal{X}} p_\nu \, \mathrm{d}(\eta - \nu), \forall \eta \in \mathcal{F}_{\bar{\eta}} \right\}.$$

Fix $q \in \partial_{\mathcal{F}_{\bar{\eta}}} h(\nu)$. The relative subdifferential inequality then becomes

$$\int_{\mathcal{X}} (q - p_\nu) \mathrm{d}(\eta - \nu) \leq 0, \quad \forall \nu \in \mathcal{F}_{\bar{\eta}}. \tag{40}$$

Let $v \in L^\infty(\pi)$ satisfy $\int_{\mathcal{X}} v \, \mathrm{d}\nu = \int_{\mathcal{X}} \frac{\mathrm{d}\nu}{\mathrm{d}\pi} v \, \mathrm{d}\pi = 0$. For $t > 0$, define $\frac{\mathrm{d}\eta_\pm}{\mathrm{d}\pi} := \frac{\mathrm{d}\nu}{\mathrm{d}\pi}(1 \pm tv)$. If $t > 0$ is chosen such that $t\|v\|_{L^\infty(\pi)} < 1$, then

$$\frac{\mathrm{d}\eta_\pm}{\mathrm{d}\pi} > 0, \quad \pi\text{-a.e.}$$

Moreover,

$$\int_{\mathcal{X}} \frac{\mathrm{d}\eta_\pm}{\mathrm{d}\pi} \, \mathrm{d}\pi = 1 \pm t \int_{\mathcal{X}} \frac{\mathrm{d}\nu}{\mathrm{d}\pi} v \, \mathrm{d}\pi = 1.$$

Finally,

$$\left\| \frac{\mathrm{d}\eta_\pm}{\mathrm{d}\pi} \right\|_{L^2(\pi)} \leq \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \right\|_{L^2(\pi)} + t \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\pi} v \right\|_{L^2(\pi)}.$$

By the L^2-radius condition, the right-hand side is smaller than $\bar{\eta}$ for all sufficiently small $t > 0$. For example, it suffices to take

$$0 < t < \min \left\{ \frac{1}{\|v\|_{L^\infty(\pi)}}, \frac{\bar{\eta} - \|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\|_{L^2(\pi)}}{\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi} v\|_{L^2(\pi)}} \right\}.$$

Hence $\frac{\mathrm{d}\eta_\pm}{\mathrm{d}\pi} \in \mathcal{F}_{\bar{\eta}}$. Applying (40) first with η_+ and then with η_- gives

$$t \int_{\mathcal{X}} (q - p_\nu) v \, \mathrm{d}\nu \leq 0, \quad -t \int_{\mathcal{X}} (q - p_\nu) v \, \mathrm{d}\nu \leq 0.$$

Therefore,

$$\int_{\mathcal{X}} (q - p_\nu) v \mathrm{d}\nu = 0$$

for every bounded v satisfying $\int v \mathrm{d}\nu = 0$. Since $q - p_\nu \in L^2(\pi)$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\pi} \in L^2(\pi)$, the Cauchy-Schwarz inequality gives $q - p_\nu \in L^1(\nu)$. Set $c := \int_{\mathcal{X}} (q - p_\nu) \mathrm{d}\nu$. For any measurable set $A \subseteq \mathcal{X}$, take $v_A := \mathbb{1}_A - \nu(A)$. Then $v_A \in L^\infty(\pi)$ and $\int v_A \mathrm{d}\nu = 0$, and hence

$$0 = \int_{\mathcal{X}} (q - p_\nu) v_A \mathrm{d}\nu = \int_A (q - p_\nu) \mathrm{d}\nu - \nu(A) \int_{\mathcal{X}} (q - p_\nu) \mathrm{d}\nu = \int_A ((q - p_\nu) - c) \mathrm{d}\nu.$$

Since this holds for every measurable A, it follows that $q - p_\nu = c$ ν-a.e. By the strict positivity condition $\frac{\mathrm{d}\nu}{\mathrm{d}\pi} > 0$ π-a.e., the measures ν and π are equivalent, and therefore $q - p_\nu = c$ π-a.e., which proves one inclusion in (39).

Conversely, let $q = p_\nu + c$ for some $c \in \mathbb{R}$. Then it is immediate that $q \in \partial_{\mathcal{F}_{\bar{\eta}}} h(\nu)$, proving (39). □

We now check that the condition $\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) < \infty$ required in Theorems 2.5 and 3.5 is satisfied in the specific cases of Examples B.1 and B.2. In addition, we establish that single-valuedness of the relative subdifferential holds for all iterates along the trajectory.

Lemma B.7. *Let Assumption 2.3 hold and let h denote the relative entropy from Example B.1. Assume $\nu^0, \mu^0 \in \mathcal{E}_{\beta_0}$ for some $\beta_0 > 0$. Fix a horizon $N \geq 1$ and a stepsize $\tau > 0$. Define $\mathcal{C} = \mathcal{D} = \mathcal{E}_{\bar{\beta}}$, with $\bar{\beta} := \beta_0 + 2N\tau \max\{C_1, C_2\} + 1$. Let $(\nu^n, \mu^n)_{n=0}^N$ be the iterates produced by Algorithms 1 and 2 with feasible sets $\mathcal{C}, \mathcal{D}$. Then*

$$(\nu^n, \mu^n)_{n=0}^N \subseteq \mathcal{E}_{\bar{\beta}} \times \mathcal{E}_{\bar{\beta}}.$$

Hence,

$$\partial_{\mathcal{C}} h(\nu^n) = \left\{ \log \frac{\mathrm{d}\nu^n}{\mathrm{d}\pi}(\cdot) + c_{\nu,n} : c_{\nu,n} \in \mathbb{R} \right\},$$
$$\partial_{\mathcal{D}} h(\mu^n) = \left\{ \log \frac{\mathrm{d}\mu^n}{\mathrm{d}\pi}(\cdot) + c_{\mu,n} : c_{\mu,n} \in \mathbb{R} \right\}, \quad n \in \{0, 1, \ldots, N\}.$$

Moreover,

$$\sup_{\nu \in \mathcal{C}} \mathrm{KL}(\nu, \nu^0) \le \bar{\beta} + \beta_0, \quad \sup_{\mu \in \mathcal{D}} \mathrm{KL}(\mu, \mu^0) \le \bar{\beta} + \beta_0,$$

and in particular

$$\sup_{\nu \in \mathcal{C}} \mathrm{KL}(\nu, \nu^0) \; + \; \sup_{\mu \in \mathcal{D}} \mathrm{KL}(\mu, \mu^0) \; < \infty.$$

Proof. We provide the proof only for Algorithm 1, as the argument for the other algorithm is essentially the same. Assume $\nu^0, \mu^0 \in \mathcal{E}_{\beta_0}$ for some $\beta_0 > 0$. Fix a horizon $N \ge 1$ and let $\bar{\beta} := \beta_0 + 2N\tau \max\{C_1, C_2\} + 1$. We take $\mathcal{C} = \mathcal{D} = \mathcal{E}_{\bar{\beta}}$.

Using the flat derivative formula (32), the first-order optimality condition in Proposition B.4 gives

$$\begin{cases} \log \frac{\nu^{n+1}(x)}{\pi(x)} - \log \frac{\nu^n(x)}{\pi(x)} = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) - \log \int_{\mathcal{X}} e^{-\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)} \frac{\nu^n(x)}{\pi(x)} \pi(x) \mathrm{d}x, \\ \log \frac{\mu^{n+1}(y)}{\pi(y)} - \log \frac{\mu^n(y)}{\pi(y)} = \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y) - \log \int_{\mathcal{X}} e^{\tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)} \frac{\mu^n(y)}{\pi(y)} \pi(y) \mathrm{d}y, \end{cases}$$

for all $x, y \in \mathcal{X}$ a.e. Taking the sup-norm on both sides over x, y and using the assumptions gives

$$\left\| \log \frac{\nu^{n+1}(\cdot)}{\pi(\cdot)} \right\|_{L^\infty(\mathcal{X})} \le \left\| \log \frac{\nu^n(\cdot)}{\pi(\cdot)} \right\|_{L^\infty(\mathcal{X})} + 2\tau C_1,$$
$$\left\| \log \frac{\mu^{n+1}(\cdot)}{\pi(\cdot)} \right\|_{L^\infty(\mathcal{X})} \le \left\| \log \frac{\mu^n(\cdot)}{\pi(\cdot)} \right\|_{L^\infty(\mathcal{X})} + 2\tau C_2,$$

Iterating from $n = 0$ and using $\nu^0, \mu^0 \in \mathcal{E}_{\beta_0}$ gives, for all $n = 0, \ldots, N$,

$$\left\| \log \frac{\nu^n(\cdot)}{\pi(\cdot)} \right\|_{L^\infty(\mathcal{X})} \le \beta_0 + 2n\tau C_1 < \bar{\beta}, \quad \left\| \log \frac{\mu^n(\cdot)}{\pi(\cdot)} \right\|_{L^\infty(\mathcal{X})} \le \beta_0 + 2n\tau C_2 < \bar{\beta}.$$

Hence, $\nu^n, \mu^n \in \mathcal{E}_{\bar{\beta}} = \mathcal{C} = \mathcal{D}$ for all $n = 0, \ldots, N$, i.e.,

$$(\nu^n, \mu^n)_{n=0}^N \subseteq \mathcal{E}_{\bar{\beta}} \times \mathcal{E}_{\bar{\beta}}.$$

Hence, choosing, $\mathcal{C} = \mathcal{D} = \mathcal{E}_{\bar{\beta}}$, $\bar{\beta} := \beta_0 + 2N\tau \max\{C_1, C_2\} + 1$, ensures strict slackness at every iterate $n \le N$. Hence, by Lemma B.5,

$$\partial_{\mathcal{C}} h(\nu^{n+1}) = \left\{ \log \frac{\mathrm{d}\nu^{n+1}}{\mathrm{d}\pi}(\cdot) + c_{\nu,n} : c_{\nu,n} \in \mathbb{R} \right\},$$
$$\partial_{\mathcal{D}} h(\mu^{n+1}) = \left\{ \log \frac{\mathrm{d}\mu^{n+1}}{\mathrm{d}\pi}(\cdot) + c_{\mu,n} : c_{\mu,n} \in \mathbb{R} \right\}.$$

Combining this with Proposition B.4 gives, for some constant $c_{\nu,n} \in \mathbb{R}$,

$$\log \frac{\mathrm{d}\nu^{n+1}}{\mathrm{d}\pi} - \log \frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) + c_{\nu,n}.$$

The analogous ascent identity holds for the μ update.

Now fix any $\nu \in \mathcal{E}_{\bar{\beta}}$. Since $\nu^0 \in \mathcal{E}_{\beta_0}$, we have a.e.

$$-\bar{\beta} \le \log \frac{\nu(\cdot)}{\pi(\cdot)} \le \bar{\beta}, \quad -\beta_0 \le \log \frac{\nu^0(\cdot)}{\pi(\cdot)} \le \beta_0,$$

and therefore a.e.

$$\log \frac{\nu(\cdot)}{\nu^0(\cdot)} = \log \frac{\nu(\cdot)}{\pi(\cdot)} - \log \frac{\nu^0(\cdot)}{\pi(\cdot)} \le \bar{\beta} + \beta_0.$$

Hence, since ν is a probability measure,

$$\mathrm{KL}(\nu, \nu^0) \le \int_{\mathcal{X}} (\bar{\beta} + \beta_0)\, \nu(\mathrm{d}x) = \bar{\beta} + \beta_0.$$

Thus, $\sup_{\nu \in \mathcal{C}} \mathrm{KL}(\nu, \nu^0) \le \bar{\beta} + \beta_0 < \infty$. The same argument with $\mu \in \mathcal{E}_{\bar{\beta}}$ and $\mu^0 \in \mathcal{E}_{\beta_0}$ gives $\sup_{\mu \in \mathcal{D}} \mathrm{KL}(\mu, \mu^0) \le \bar{\beta} + \beta_0 < \infty$. Combining the two bounds proves

$$\sup_{\nu \in \mathcal{C}} \mathrm{KL}(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} \mathrm{KL}(\mu, \mu^0) < \infty.$$

□

Lemma B.8. *Let Assumption 2.3 hold and let h denote the χ^2-divergence from Example B.2. Assume $\nu^0, \mu^0 \in \mathcal{F}_{\eta_0}$ for some $\eta_0 > 0$. Fix a horizon $N \ge 1$ and a stepsize $\tau > 0$. Define $\mathcal{C} = \mathcal{D} = \mathcal{F}_{\bar{\eta}}$, with $\bar{\eta} := \eta_0 + N\tau \max\{C_1, C_2\}$. Suppose, moreover, that*

$$\underline{\kappa}_\nu := \operatorname*{ess\,inf}_\pi \frac{\mathrm{d}\nu^0}{\mathrm{d}\pi} > 2N\tau C_1, \quad \underline{\kappa}_\mu := \operatorname*{ess\,inf}_\pi \frac{\mathrm{d}\mu^0}{\mathrm{d}\pi} > 2N\tau C_2. \tag{41}$$

Let $(\nu^n, \mu^n)_{n=0}^N$ be the iterates produced by Algorithms 1 and 2 with feasible sets $\mathcal{C}, \mathcal{D}$. Then, for every $n = 0, ..., N$,

$$\left\| \frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} \right\|_{L^2_\pi} \le \bar{\eta}, \quad \left\| \frac{\mathrm{d}\mu^n}{\mathrm{d}\pi} \right\|_{L^2_\pi} \le \bar{\eta}, \tag{42}$$

and

$$\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} \ge \underline{\kappa}_\nu - 2n\tau C_1 > 0, \quad \frac{\mathrm{d}\mu^n}{\mathrm{d}\pi} \ge \underline{\kappa}_\mu - 2n\tau C_2 > 0, \quad \pi\text{-a.e.} \tag{43}$$

In particular,

$$(\nu^n, \mu^n)_{n=0}^N \subseteq \mathcal{F}_{\bar{\eta}} \times \mathcal{F}_{\bar{\eta}},$$

and all iterates are equivalent to π. Hence,

$$\partial_{\mathcal{C}} h(\nu^n) = \left\{ \frac{\mathrm{d}\nu^n}{\mathrm{d}\pi}(\cdot) + c_{\nu,n} : c_{\nu,n} \in \mathbb{R} \right\},$$
$$\partial_{\mathcal{D}} h(\mu^n) = \left\{ \frac{\mathrm{d}\mu^n}{\mathrm{d}\pi}(\cdot) + c_{\mu,n} : c_{\mu,n} \in \mathbb{R} \right\}, \quad n \in \{0, 1, \ldots, N\}.$$

Moreover,

$$\frac{1}{2} \sup_{\nu \in \mathcal{C}} \left\| \frac{\nu(\cdot)}{\pi(\cdot)} - \frac{\nu^0(\cdot)}{\pi(\cdot)} \right\|^2_{L^2_\pi(\mathcal{X})} \le \bar{\eta}^2 + \eta_0^2, \quad \frac{1}{2} \sup_{\mu \in \mathcal{D}} \left\| \frac{\mu(\cdot)}{\pi(\cdot)} - \frac{\mu^0(\cdot)}{\pi(\cdot)} \right\|^2_{L^2_\pi(\mathcal{X})} \le \bar{\eta}^2 + \eta_0^2,$$

and in particular

$$\frac{1}{2} \sup_{\nu \in \mathcal{C}} \left\| \frac{\nu(\cdot)}{\pi(\cdot)} - \frac{\nu^0(\cdot)}{\pi(\cdot)} \right\|^2_{L^2_\pi(\mathcal{X})} + \frac{1}{2} \sup_{\mu \in \mathcal{D}} \left\| \frac{\mu(\cdot)}{\pi(\cdot)} - \frac{\mu^0(\cdot)}{\pi(\cdot)} \right\|^2_{L^2_\pi(\mathcal{X})} < \infty.$$

Proof. We provide the proof only for Algorithm 1, as the argument for the other algorithm is essentially the same. Assume $\nu^0, \mu^0 \in \mathcal{F}_{\eta_0}$ for some $\eta_0 > 0$. Fix a horizon $N \ge 1$ and let $\bar{\eta} := \eta_0 + N\tau \max\{C_1, C_2\}$. We take $\mathcal{C} = \mathcal{D} = \mathcal{F}_{\bar{\eta}}$. The first-order condition (see e.g., [11, Section 5.1.1]) shows that

$$\left\langle \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) + \frac{1}{\tau} \left(\frac{\mathrm{d}\nu^{n+1}}{\mathrm{d}\pi} - \frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} \right), \phi - \frac{\mathrm{d}\nu^{n+1}}{\mathrm{d}\pi} \right\rangle_{L^2_\pi} \ge 0, \quad \forall \phi \in \mathfrak{F},$$

$$\left\langle \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, \cdot) - \frac{1}{\tau} \left(\frac{\mathrm{d}\mu^{n+1}}{\mathrm{d}\pi} - \frac{\mathrm{d}\mu^n}{\mathrm{d}\pi} \right), \phi - \frac{\mathrm{d}\mu^{n+1}}{\mathrm{d}\pi} \right\rangle_{L^2_\pi} \ge 0, \quad \forall \phi \in \mathfrak{F},$$

where $\langle \cdot, \cdot \rangle_{L^2_\pi}$ is the inner product on $L^2_\pi(\mathcal{X})$, and $\mathfrak{F}$ is the nonempty closed convex set defined by

$$\mathfrak{F} = \left\{ \phi \in L^2_\pi(\mathcal{X}) \middle| \phi \ge 0 \ \pi\text{-a.e. on } \mathcal{X} \text{ and } \int \phi(x) \pi(\mathrm{d}x) = 1 \right\}.$$

Define the projection map $\Pi_{\mathfrak{F}} : L^2_\pi(\mathcal{X}) \mapsto \mathfrak{F}$ such that $\Pi_{\mathfrak{F}}(\varphi) = \operatorname{argmin}_{\phi\in\mathfrak{F}} \|\phi - \varphi\|_{L^2_\pi(\mathcal{X})}$ for all $\varphi \in L^2_\pi(\mathcal{X})$, which satisfies

$$\langle \Pi(\varphi) - \varphi, \phi - \Pi(\varphi)\rangle_{L^2_\pi} \geq 0, \quad \forall \phi \in \mathfrak{F}.$$

Then

$$\frac{\mathrm{d}\nu^{n+1}}{\mathrm{d}\pi} = \Pi_{\mathfrak{F}}\left(\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} - \tau\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot)\right),$$
$$\frac{\mathrm{d}\mu^{n+1}}{\mathrm{d}\pi} = \Pi_{\mathfrak{F}}\left(\frac{\mathrm{d}\mu^n}{\mathrm{d}\pi} + \tau\frac{\delta F}{\delta \mu}(\nu^n, \mu^n, \cdot)\right).$$

Note $\|\Pi_{\mathfrak{F}}(\varphi_1) - \Pi_{\mathfrak{F}}(\varphi_2)\|_{L^2_\pi(\mathcal{X})} \leq \|\varphi_1 - \varphi_2\|_{L^2_\pi(\mathcal{X})}$ for all $\varphi_1, \varphi_2 \in L^2_\pi(\mathcal{X})$ (see e.g., [20, Theorem 4.3-1]). Moreover, since $\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} = \Pi_{\mathfrak{F}}\left(\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi}\right)$, $\frac{\mathrm{d}\mu^n}{\mathrm{d}\pi} = \Pi_{\mathfrak{F}}\left(\frac{\mathrm{d}\mu^n}{\mathrm{d}\pi}\right)$,

$$\left\|\frac{\mathrm{d}\nu^{n+1}}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} \leq \left\|\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} + \tau\left\|\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot)\right\|_{L^2_\pi(\mathcal{X})} \leq \left\|\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} + \tau C_1,$$

$$\left\|\frac{\mathrm{d}\mu^{n+1}}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} \leq \left\|\frac{\mathrm{d}\mu^n}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} + \tau\left\|\frac{\delta F}{\delta \mu}(\nu^n, \mu^n, \cdot)\right\|_{L^2_\pi(\mathcal{X})} \leq \left\|\frac{\mathrm{d}\mu^n}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} + \tau C_2.$$

Iterating from $n = 0$ and using $\nu^0, \mu^0 \in \mathcal{F}_{\eta_0}$ gives, for all $n = 0, \ldots, N$,

$$\left\|\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} \leq \eta_0 + n\tau C_1 \leq \bar{\eta}, \quad \left\|\frac{\mathrm{d}\mu^n}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} \leq \eta_0 + n\tau C_2 \leq \bar{\eta}.$$

Consider the normalized affine candidate

$$\frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi} := \frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} - \tau\left(\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) - \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)\pi(\mathrm{d}x)\right).$$

It belongs to $L^2_\pi(\mathcal{X})$ and has unit π-mass. Moreover, if (43) holds at iteration n,

$$\begin{aligned}\frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi} &\geq \frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} - \tau\left|\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) - \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)\pi(\mathrm{d}x)\right| \\ &\geq \underline{\kappa}_\nu - 2n\tau C_1 - 2\tau C_1 \\ &= \underline{\kappa}_\nu - 2(n+1)\tau C_1 > 0,\end{aligned}$$

where the last inequality follows from Assumption 2.3 and $n + 1 \leq N$. Thus, $\frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi} \in \mathfrak{F}$.

For every $\phi \in \mathfrak{F}$, we have

$$\begin{aligned}&\left\langle \frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi} - \left(\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} - \tau\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot)\right), \phi - \frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi}\right\rangle_{L^2_\pi} \\ &= \tau\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)\pi(\mathrm{d}x)\int_{\mathcal{X}}\left(\phi(x) - \frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi}(x)\right)\pi(\mathrm{d}x) = 0.\end{aligned}$$

Consequently,

$$\frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi} = \Pi_{\mathfrak{F}}\left(\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi} - \tau\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot)\right).$$

In addition,

$$\begin{aligned}\left\|\frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi}\right\|_{L^2_\pi} &\leq \left\|\frac{\mathrm{d}\nu^n}{\mathrm{d}\pi}\right\|_{L^2_\pi} + \tau\left\|\frac{\delta F}{\delta \nu}(\nu^n, \mu^n, \cdot) - \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)\pi(\mathrm{d}x)\right\|_{L^2_\pi} \\ &\leq \eta_0 + (n+1)\tau C_1 \leq \bar{\eta},\end{aligned}$$

for every $0 \leq n \leq N - 1$. Therefore, $\frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi}$ belongs to $\mathcal{F}_{\bar{\eta}}$. Since it minimizes the update objective over the larger set $\mathfrak{F}$, it also minimizes it over $\mathcal{F}_{\bar{\eta}}$ because $\frac{\mathrm{d}\widehat{\nu}^{n+1}}{\mathrm{d}\pi} \in \mathcal{F}_{\bar{\eta}}$. Hence, by Lemma B.6,

$$\partial_{\mathcal{C}} h(\nu^{n+1}) = \left\{\frac{\mathrm{d}\nu^{n+1}}{\mathrm{d}\pi}(\cdot) + c_{\nu,n} : c_{\nu,n} \in \mathbb{R}\right\},$$
$$\partial_{\mathcal{D}} h(\mu^{n+1}) = \left\{\frac{\mathrm{d}\mu^{n+1}}{\mathrm{d}\pi}(\cdot) + c_{\mu,n} : c_{\mu,n} \in \mathbb{R}\right\}.$$

Combining this with Proposition B.4 gives

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

The same argument for the maximizing player gives

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

because this update is the projection of $\frac{\mathrm{d}\mu^n}{\mathrm{d}\pi} + \tau\frac{\delta F}{\delta \mu}(\nu^n, \mu^n, \cdot)$ onto $\mathfrak{F}$. In particular,

$$\frac{\mathrm{d}\mu^{n+1}}{\mathrm{d}\pi} \geq \underline{\kappa}_\mu - 2(n+1)\tau C_2 > 0$$

and

$$\left\|\frac{\mathrm{d}\mu^{n+1}}{\mathrm{d}\pi}\right\|_{L^2_\pi} \leq \eta_0 + (n+1)\tau C_2 \leq \bar{\eta}.$$

Starting from $n = 0$, induction proves (42) and (43).

Hence, $\nu^n, \mu^n \in \mathcal{F}_{\bar{\eta}} = \mathcal{C} = \mathcal{D}$ for all $n = 0, \ldots, N$, i.e.,

$$(\nu^n, \mu^n)_{n=0}^N \subseteq \mathcal{F}_{\bar{\eta}} \times \mathcal{F}_{\bar{\eta}}.$$

Fix any $\nu \in \mathcal{F}_{\bar{\eta}}$. Since $\nu^0 \in \mathcal{F}_{\eta_0}$, we have

$$\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} \leq \bar{\eta}, \quad \left\|\frac{\mathrm{d}\nu^0}{\mathrm{d}\pi}\right\|_{L^2_\pi(\mathcal{X})} \leq \eta_0,$$

and therefore

$$\frac{1}{2}\left\|\frac{\nu(\cdot)}{\pi(\cdot)} - \frac{\nu^0(\cdot)}{\pi(\cdot)}\right\|^2_{L^2_\pi(\mathcal{X})} \leq \bar{\eta}^2 + \eta_0^2.$$

Thus, $\sup_{\nu\in\mathcal{C}} \frac{1}{2}\left\|\frac{\nu(\cdot)}{\pi(\cdot)} - \frac{\nu^0(\cdot)}{\pi(\cdot)}\right\|^2_{L^2_\pi(\mathcal{X})} \leq \bar{\eta}^2 + \eta_0^2 < \infty$. The same argument with $\mu \in \mathcal{F}_{\bar{\eta}}$ and $\mu^0 \in \mathcal{F}_{\eta_0}$ gives $\sup_{\mu\in\mathcal{D}} \frac{1}{2}\left\|\frac{\mu(\cdot)}{\pi(\cdot)} - \frac{\mu^0(\cdot)}{\pi(\cdot)}\right\|^2_{L^2_\pi(\mathcal{X})} \leq \bar{\eta}^2 + \eta_0^2 < \infty$. Combining the two bounds proves

$$\frac{1}{2}\sup_{\nu\in\mathcal{C}}\left\|\frac{\nu(\cdot)}{\pi(\cdot)} - \frac{\nu^0(\cdot)}{\pi(\cdot)}\right\|^2_{L^2_\pi(\mathcal{X})} + \frac{1}{2}\sup_{\mu\in\mathcal{D}}\left\|\frac{\mu(\cdot)}{\pi(\cdot)} - \frac{\mu^0(\cdot)}{\pi(\cdot)}\right\|^2_{L^2_\pi(\mathcal{X})} < \infty.$$

□

When F is strongly-convex-strongly-concave relative to h and Assumption 2.1 holds, it can be shown that (ν^*, μ^*) is the unique MNE of (1) (see the proof of [33, Lemma 6]). Moreover, based on relative convexity-concavity of F, we prove in Lemma B.10 that the NI error satisfies a type of "quadratic growth" inequality relative to h.

Assumption B.9 (Relative convexity-concavity)**.** *Assume F is (ℓ_ν, ℓ_μ)-strongly convex-concave relative to h, i.e., there exist $\ell_\nu, \ell_\mu > 0$ such that for any $\nu, \nu' \in \mathcal{C}$ and $\mu, \mu' \in \mathcal{D}$, we have*

$$D_{F(\cdot,\mu)}(\nu', \nu) = F(\nu', \mu) - F(\nu, \mu) - \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu, \mu, x)(\nu' - \nu)(\mathrm{d}x) \geq \ell_\nu D_h(\nu', \nu), \tag{44}$$

$$D_{F(\nu,\cdot)}(\mu', \mu) = F(\nu, \mu') - F(\nu, \mu) - \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu, \mu, y)(\mu' - \mu)(\mathrm{d}y) \leq -\ell_\mu D_h(\mu', \mu). \tag{45}$$

Lemma B.10 ("Quadratic growth" of NI error relative to h)**.** *Assume that $\mathcal{C}, \mathcal{D}$ are convex and that $(\nu^*, \mu^*) \in \mathcal{C} \times \mathcal{D}$ is an MNE of F. Suppose that Assumption 2.1 and B.9 hold. Then, for all $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, it holds that*

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

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

Remark B.11. *We refer to Lemma B.10 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. Fix $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$. Since (ν^*, μ^*) is a saddle point, we have

$$F(\nu^*, \tilde{\mu}) \leq F(\nu^*, \mu^*) \leq F(\tilde{\nu}, \mu^*), \quad (\tilde{\nu}, \tilde{\mu}) \in \mathcal{C} \times \mathcal{D}.$$

By convexity of $\mathcal{C}$ and $\mathcal{D}$, the measures $\nu_t = (1-t)\nu^* + t\nu$, $\mu_t = (1-t)\mu^* + t\mu$ are feasible for every $t \in [0,1]$. Applying the saddle point inequalities to these measures, dividing by $t > 0$, and letting $t \downarrow 0$ gives

$$\int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^*, \mu^*, x)(\nu - \nu^*)(\mathrm{d}x) \geq 0, \tag{46}$$

$$\int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^*, \mu^*, y)(\mu - \mu^*)(\mathrm{d}y) \leq 0. \tag{47}$$

Applying (44) with base point ν^*, comparison point ν, and fixed second argument μ^*, and using (46), we obtain

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

Similarly, applying (45) with base point μ^*, comparison point μ, and fixed first argument ν^*, and using (47), gives

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

Hence, by the definition of the Nikaidô–Isoda error,

$$\begin{aligned}
\mathrm{NI}(\nu, \mu) &= \sup_{\tilde{\mu} \in \mathcal{D}} F(\nu, \tilde{\mu}) - \inf_{\tilde{\nu} \in \mathcal{C}} F(\tilde{\nu}, \mu) \geq F(\nu, \mu^*) - F(\nu^*, \mu) \\
&= [F(\nu, \mu^*) - F(\nu^*, \mu^*)] + [F(\nu^*, \mu^*) - F(\nu^*, \mu)] \\
&\geq \ell_\nu D_h(\nu, \nu^*) + \ell_\mu D_h(\mu, \mu^*) \geq \ell \left(D_h(\nu, \nu^*) + D_h(\mu, \mu^*) \right),
\end{aligned}$$

where the last inequality follows from the nonnegativity of the Bregman divergences and $\ell = \min\{\ell_\nu, \ell_\mu\}$. □

By Lemma B.10, 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) with the rates proved in Theorem 2.5 and Theorem 3.5, respectively.

C Verification of Assumption 2.2, 2.3 and 3.3 for the examples in Section 1.4

In this section we verify that Assumptions 2.2, 2.3 and 3.3 are satisfied by the objective function F in Section 1.4.

Proposition C.1 (Verification of assumptions for the Wasserstein GAN)**.** *Let $\mathcal{Y}, \mathcal{Z} \subseteq \mathbb{R}^d$, with $\hat{\xi} \in \mathcal{P}(\mathcal{Y})$ and $\xi \in \mathcal{P}(\mathcal{Z})$. Suppose $T_\theta : \mathcal{Z} \to \mathcal{Y}$ is measurable with $\theta \in \Theta \subseteq \mathbb{R}^d$, and for each $w \in \mathcal{W} \subseteq \mathbb{R}^d$, the critic $D_w : \mathcal{Y} \to \mathbb{R}$ is measurable and 1-Lipschitz, i.e., $|D_w(y) - D_w(y')| \leq |y - y'|$, for all $y, y' \in \mathcal{Y}$. Assume moreover that $\mathcal{Y}$ has finite diameter, i.e., $\mathrm{diam}(\mathcal{Y}) := \sup_{y,y' \in \mathcal{Y}} \|y - y'\| < \infty$, (e.g., if $\mathcal{Y}$ is compact), and fix $y_0 \in \mathcal{Y}$. Then Assumptions 2.2, 2.3 and 3.3 are satisfied by the objective*

$$F(\nu, \mu) := \int_{\mathcal{W}} \int_{\Theta} f(\theta, w)\, \nu(\mathrm{d}\theta)\, \mu(\mathrm{d}w)$$

from Example 1.4.

Proof. By Definition 1.1,

$$\frac{\delta F}{\delta \nu}(\nu, \mu, \theta) = \int_{\mathcal{W}} f(\theta, w)\, \mu(\mathrm{d}w), \qquad \frac{\delta F}{\delta \mu}(\nu, \mu, w) = \int_{\Theta} f(\theta, w)\, \nu(\mathrm{d}\theta).$$

Therefore, Assumption 2.2 holds with equality and Assumption 3.3 holds with equality with $L_\nu = L_\mu = 0$.

It remains to verify uniform boundedness of f and hence of F and its flat derivatives. Recall

$$f(\theta, w) = \int_{\mathcal{Y}} D_w(y) \left(T_\theta \# \xi - \hat{\xi}\right)(\mathrm{d}y) = \int_{\mathcal{Y}} D_w(y) \, (T_\theta \# \xi)(\mathrm{d}y) - \int_{\mathcal{Y}} D_w(y) \, \hat{\xi}(\mathrm{d}y).$$

Since adding a constant to D_w does not change $f(\theta, w)$ because $(T_\theta \# \xi - \hat{\xi})$ has total mass 0, we may replace D_w by $\widetilde{D}_w(y) \coloneqq D_w(y) - D_w(y_0)$, so that $\widetilde{D}_w(y_0) = 0$ and $\widetilde{D}_w$ is still 1-Lipschitz. Then for all $y \in \mathcal{Y}$,

$$|\widetilde{D}_w(y)| = |D_w(y) - D_w(y_0)| \leq |y - y_0| \leq \mathrm{diam}(\mathcal{Y}).$$

Hence, using $\hat{\xi} \in \mathcal{P}(\mathcal{Y})$ and $T_\theta \# \xi \in \mathcal{P}(\mathcal{Y})$,

$$\begin{aligned} |f(\theta, w)| &= \left| \int_{\mathcal{Y}} \widetilde{D}_w(y) \left(T_\theta \# \xi - \hat{\xi}\right)(\mathrm{d}y) \right| \\ &\leq \int_{\mathcal{Y}} |\widetilde{D}_w(y)| \, (T_\theta \# \xi)(\mathrm{d}y) + \int_{\mathcal{Y}} |\widetilde{D}_w(y)| \, \hat{\xi}(\mathrm{d}y) \leq 2\mathrm{diam}(\mathcal{Y}). \end{aligned}$$

Consequently, for all $\nu, \mu \in \mathcal{P}(\mathcal{X})$ and all $\theta \in \Theta$, $w \in \mathcal{W}$,

$$\left| \frac{\delta F}{\delta \nu}(\nu, \mu, \theta) \right| \leq 2 \, \mathrm{diam}(\mathcal{Y}), \quad \left| \frac{\delta F}{\delta \mu}(\nu, \mu, w) \right| \leq 2 \, \mathrm{diam}(\mathcal{Y}).$$

Thus Assumption 2.3 holds with

$$C_1 = C_2 = 2 \, \mathrm{diam}(\mathcal{Y}).$$

□

Proposition C.2 (Verification of assumptions for RLHF)**.** *Assume $\mathcal{S}, \mathcal{A}$ are measurable and $P(\cdot \succ \cdot|x) \in [0, 1]$ for all $x \sim \rho \in \mathcal{P}(\mathcal{S})$. Then the game with $F = P(\mu \succ \nu)$ satisfies Assumptions 2.2, 2.3 and 3.3.*

Proof. The map $\mu \mapsto P(\mu \succ \nu)$ is linear, hence concave, and $\nu \mapsto P(\mu \succ \nu)$ is also linear, hence convex. Therefore F is convex in ν and concave in μ, as required by Assumption 2.2. Since F is bilinear, it admits the flat derivatives

$$\begin{aligned} \frac{\delta F}{\delta \nu}(\mu \succ \nu, y') &= \mathbb{E}_{x \sim \rho, y \sim \mu(\cdot|x)} \left[P(y \succ y'|x) \right], \\ \frac{\delta F}{\delta \mu}(\mu \succ \nu, y) &= \mathbb{E}_{x \sim \rho, y' \sim \nu(\cdot|x)} \left[P(y \succ y'|x) \right]. \end{aligned}$$

Because $P(\cdot \succ \cdot|x) \in [0, 1]$, both derivatives are bounded by 1, yielding Assumption 2.3. Since F is bilinear it satisfies Assumption 3.3 with equality and $L_\nu = L_\mu = 0$. □

For the adversarial NN example, first, we prove that $F_\varepsilon(\cdot, \mu)$ is L-smooth in ν, so it satisfies the first condition of Assumption 3.3.

Proposition C.3 (L-smoothness of $F_\varepsilon(\cdot, \mu)$)**.** *Let Assumption 2.1 hold. Let μ be a probability measure on $\mathcal{Y} \times \mathcal{Z}$, let ν, ν' be probability measures on $\mathbb{R}^d$ and $\varepsilon \geq 0$. Assume $\hat{\varphi}$ is bounded by $M_\varphi > 0$. Then*

$$F_\varepsilon(\nu', \mu) - F_\varepsilon(\nu, \mu) - \int_{\mathbb{R}^d} \frac{\delta F_\varepsilon}{\delta \nu}(\nu, \mu, x) \, (\nu' - \nu)(\mathrm{d}x) \leq \frac{4 M_\varphi}{\alpha} D_h(\nu', \nu).$$

In particular, one may take $L_\nu = \frac{4 M_\varphi}{\alpha}$ in Assumption 3.3.

Proof. For $z \in \mathcal{Z}$, define $a_\nu(z) \coloneqq \int_{\mathbb{R}^d} \hat{\varphi}(x, z) \, \nu(\mathrm{d}x)$ and $\Delta a(z) \coloneqq a_{\nu'}(z) - a_\nu(z)$. Note that

$$\frac{\delta F_\varepsilon}{\delta \nu}(\nu, \mu, x) = \int_{\mathcal{Y} \times \mathcal{Z}} \Big(a_\nu(z) - y \Big) \hat{\varphi}(x, z) \, \mu(\mathrm{d}y, \mathrm{d}z).$$

Fix $(y,z) \in \mathcal{Y} \times \mathcal{Z}$. Then $y - a_{\nu'}(z) = y - a_\nu(z) - \Delta a(z)$ and thus

$$\begin{aligned}\frac{1}{2}\big(y - a_{\nu'}(z)\big)^2 - \frac{1}{2}\big(y - a_\nu(z)\big)^2 &= \frac{1}{2}\Big((y - a_\nu(z) - \Delta a(z))^2 - (y - a_\nu(z))^2\Big) \\ &= (a_\nu(z) - y)\Delta a(z) + \frac{1}{2}(\Delta a(z))^2.\end{aligned}$$

Integrating against $\mu(\mathrm{d}y, \mathrm{d}z)$ gives

$$\begin{aligned}&\frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} \big(y - a_{\nu'}(z)\big)^2 \,\mu(\mathrm{d}y,\mathrm{d}z) - \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} \big(y - a_\nu(z)\big)^2 \,\mu(\mathrm{d}y,\mathrm{d}z) \\ &\quad = \int_{\mathcal{Y}\times\mathcal{Z}} (a_\nu(z) - y)\Delta a(z)\,\mu(\mathrm{d}y,\mathrm{d}z) + \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} (\Delta a(z))^2\,\mu(\mathrm{d}y,\mathrm{d}z). \qquad (48)\end{aligned}$$

Next, by Fubini,

$$\begin{aligned}\int_{\mathcal{Y}\times\mathcal{Z}} (a_\nu(z) - y)\Delta a(z)\,\mu(\mathrm{d}y,\mathrm{d}z) &= \int_{\mathcal{Y}\times\mathcal{Z}} (a_\nu(z) - y)\left(\int_{\mathbb{R}^d} \hat{\varphi}(x,z)\,(\nu' - \nu)(\mathrm{d}x)\right)\mu(\mathrm{d}y,\mathrm{d}z) \\ &= \int_{\mathbb{R}^d}\left[\int_{\mathcal{Y}\times\mathcal{Z}} (a_\nu(z) - y)\hat{\varphi}(x,z)\,\mu(\mathrm{d}y,\mathrm{d}z)\right](\nu' - \nu)(\mathrm{d}x) \\ &= \int_{\mathbb{R}^d} \frac{\delta F_\varepsilon}{\delta \nu}(\nu,\mu,x)\,(\nu' - \nu)(\mathrm{d}x).\end{aligned}$$

Since the term $-\,\mathrm{TV}_\varepsilon^2(\mu,\hat{\mu})$ does not depend on ν, it cancels in the difference $F_\varepsilon(\nu',\mu) - F_\varepsilon(\nu,\mu)$, and (48) implies the identity

$$F_\varepsilon(\nu',\mu) - F_\varepsilon(\nu,\mu) - \int_{\mathbb{R}^d} \frac{\delta F_\varepsilon}{\delta \nu}(\nu,\mu,x)\,(\nu' - \nu)(\mathrm{d}x) = \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} (\Delta a(z))^2\,\mu(\mathrm{d}y,\mathrm{d}z). \qquad (49)$$

In particular, the left-hand side is nonnegative, so the absolute value equals the same quantity.

For each $z \in \mathcal{Z}$,

$$|\Delta a(z)| = \left|\int_{\mathbb{R}^d} \hat{\varphi}(x,z)\,(\nu' - \nu)(dx)\right| \le 2M_\varphi\,\mathrm{TV}(\nu',\nu),$$

where we used the dual characterization of total variation. Squaring and inserting into (49) gives

$$\begin{aligned}&F_\varepsilon(\nu',\mu) - F_\varepsilon(\nu,\mu) - \int_{\mathbb{R}^d} \frac{\delta F_\varepsilon}{\delta \nu}(\nu,\mu,x)\,(\nu' - \nu)(\mathrm{d}x) \\ &= \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} (\Delta a(z))^2\,\mu(\mathrm{d}y,\mathrm{d}z) \le 2M_\varphi\,\mathrm{TV}^2(\nu',\nu).\end{aligned}$$

By Assumption 2.1,

$$F_\varepsilon(\nu',\mu) - F_\varepsilon(\nu,\mu) - \int_{\mathbb{R}^d} \frac{\delta F_\varepsilon}{\delta \nu}(\nu,\mu,x)\,(\nu' - \nu)(\mathrm{d}x) \le \frac{4M_\varphi}{\alpha}\,D_h(\nu',\nu),$$

which proves the claim. □

Now, we prove that $\mathrm{TV}_\varepsilon^2$ is flat convex.

Lemma C.4 (Convexity of the squared smoothed TV)**.** *Let $\hat{\mu}$ be a fixed reference probability measure on $\mathcal{Y}\times\mathcal{Z}$ with Lebesgue density $\hat{m}$. Fix $\varepsilon > 0$ and define $\rho_\varepsilon(u) := \sqrt{u^2 + \varepsilon^2}$, so that*

$$\mathrm{TV}_\varepsilon(\mu,\hat{\mu}) := \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} \rho_\varepsilon\big(m(r) - \hat{m}(r)\big)\mathrm{d}r,$$

for any μ absolutely continuous with respect to Lebesgue with density m. Then the function

$$\mu \mapsto \mathrm{TV}_\varepsilon^2(\mu,\hat{\mu})$$

is flat convex on the set of absolutely continuous probability measures.

Proof. We first show that $\mu \mapsto \mathrm{TV}_\varepsilon(\mu, \hat{\mu})$ is convex. Since ρ_ε is twice continuously differentiable with

$$\rho_\varepsilon''(u) = \frac{\varepsilon^2}{(u^2+\varepsilon^2)^{3/2}} \geq 0,$$

it is convex on $\mathbb{R}$. Let μ_0, μ_1 be absolutely continuous measures with densities m_0, m_1 and let $\lambda \in [0,1]$. The convex combination $\mu_\lambda := \lambda\mu_0 + (1-\lambda)\mu_1$ is absolutely continuous with density $m_\lambda = \lambda m_0 + (1-\lambda)m_1$. Using convexity of ρ_ε pointwise, for a.e. $r \in \mathcal{Y} \times \mathcal{Z}$ we have

$$\begin{aligned}\rho_\varepsilon\big(m_\lambda(r) - \hat{m}(r)\big) &= \rho_\varepsilon\big(\lambda(m_0(r) - \hat{m}(r)) + (1-\lambda)(m_1(r) - \hat{m}(r))\big) \\ &\leq \lambda\rho_\varepsilon\big(m_0(r) - \hat{m}(r)\big) + (1-\lambda)\rho_\varepsilon\big(m_1(r) - \hat{m}(r)\big).\end{aligned}$$

Integrating and multiplying by $\frac{1}{2}$ yields

$$\mathrm{TV}_\varepsilon(\mu_\lambda, \hat{\mu}) \leq \lambda\ \mathrm{TV}_\varepsilon(\mu_0, \hat{\mu}) + (1-\lambda)\ \mathrm{TV}_\varepsilon(\mu_1, \hat{\mu}),$$

so $\mathrm{TV}_\varepsilon(\cdot, \hat{\mu})$ is convex. Moreover $\mathrm{TV}_\varepsilon(\mu, \hat{\mu}) \geq 0$ for all μ.

Now define $f(\mu) := \mathrm{TV}_\varepsilon(\mu, \hat{\mu})$. Since $f \geq 0$ and the map $\phi(t) = t^2$ is nondecreasing on $[0, \infty)$, from convexity of f we obtain

$$f(\mu_\lambda)^2 \leq \big(\lambda f(\mu_0) + (1-\lambda) f(\mu_1)\big)^2.$$

Finally, using convexity of $\phi(t) = t^2$ on $\mathbb{R}$ gives

$$f(\mu_\lambda)^2 \leq \lambda f(\mu_0)^2 + (1-\lambda) f(\mu_1)^2,$$

i.e.,

$$\mathrm{TV}_\varepsilon(\mu_\lambda, \hat{\mu})^2 \leq \lambda\, \mathrm{TV}_\varepsilon(\mu_0, \hat{\mu})^2 + (1-\lambda)\, \mathrm{TV}_\varepsilon(\mu_1, \hat{\mu})^2.$$

□

Next, we prove an upper bound on TV_ε.

Lemma C.5 (Upper bound on TV_ε)**.** *Let $\mathcal{Y} \times \mathcal{Z} \subseteq \mathbb{R}^d$ be compact and convex and let $\mu, \hat{\mu}$ be absolutely continuous with densities $m, \hat{m}$, respectively. Then*

$$\mathrm{TV}_\varepsilon(\mu, \hat{\mu}) \ \leq\ 1 + \frac{\varepsilon}{2}\,|\mathcal{Y} \times \mathcal{Z}|. \tag{50}$$

Proof. We use the elementary inequality, valid for all $u \in \mathbb{R}$,

$$\sqrt{u^2 + \varepsilon^2} \leq |u| + \varepsilon. \tag{51}$$

Indeed, squaring both sides gives $u^2 + \varepsilon^2 \leq u^2 + 2\varepsilon|u| + \varepsilon^2$.

Applying (51) with $u = m - \hat{m}$ and integrating,

$$\begin{aligned}\mathrm{TV}_\varepsilon(\mu, \hat{\mu}) &= \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} \sqrt{(m-\hat{m})^2 + \varepsilon^2}\mathrm{d}r \\ &\leq \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} |m - \hat{m}|\mathrm{d}r + \frac{\varepsilon}{2}|\mathcal{Y}\times\mathcal{Z}| = \mathrm{TV}(\mu, \hat{\mu}) + \frac{\varepsilon}{2}\,|\mathcal{Y}\times\mathcal{Z}|.\end{aligned}$$

Since $\mu, \hat{\mu}$ are probability measures, then $\mathrm{TV}(\mu, \hat{\mu}) \leq 1$, yielding (50). □

Proposition C.6 (Verification of assumptions for adversarial NN)**.** *Let $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d-1}$ be compact and convex with $\hat{\mu} \in \mathcal{P}(\mathcal{Y} \times \mathcal{Z})$. Let h be the χ^2-divergence in Example B.2. For $x := (w, b) \in \mathbb{R}^d$ and $z \in \mathbb{R}^{d-1}$, let $\hat{\varphi}(x, z) := \ell(b)\varphi(w \cdot z)$, where $\ell : \mathbb{R} \to [-K, K]$ is a clipping function with clipping threshold $K > 0$ and $\varphi : \mathbb{R} \to \mathbb{R}$ is a bounded, continuous, non-constant function. Let $\varepsilon > 0$. Then Assumptions 2.2, 2.3 and 3.3 are satisfied by the objective*

$$F_\varepsilon(\nu, \mu) = \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} \big|y - \mathbb{E}^{X\sim\nu}[\hat{\varphi}(X, z)]\big|^2\, \mu(\mathrm{d}y, \mathrm{d}z) - \mathrm{TV}_\varepsilon^2(\mu, \hat{\mu}).$$

Proof. Observe that by linearity of the expectation in ν and convexity of $|\cdot|^2$, the function

$$F^0(\nu,\mu) = \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} \left|y - \mathbb{E}^{X\sim\nu}[\hat{\varphi}(X,z)]\right|^2 \mu(\mathrm{d}y,\mathrm{d}z)$$

satisfies the flat-convexity condition

$$F^0((1-\lambda)\nu + \lambda\nu', \mu) \le (1-\lambda)F^0(\nu,\mu) + \lambda F^0(\nu',\mu),$$

for any $\nu,\nu' \in \mathcal{P}(\mathbb{R}^d)$, $\mu \in \mathcal{P}(\mathcal{Y}\times\mathcal{Z})$ and any $\lambda \in [0,1]$. Since $F^0(\cdot,\mu) \in \mathfrak{C}^1(\mathbb{R}^d)$, by [26, Lemma 4.1], $\nu \mapsto F_\varepsilon(\nu,\mu)$ satisfies $D_{F_\varepsilon(\cdot,\mu)}(\nu',\nu) \ge 0$.

By Lemma C.4, it holds that $\mathrm{TV}_\varepsilon^2$ is convex, that is,

$$\mathrm{TV}_\varepsilon^2((1-\lambda)\mu + \lambda\mu', \hat{\mu}) \le (1-\lambda)\,\mathrm{TV}^2(\mu,\hat{\mu}) + \lambda\,\mathrm{TV}^2(\mu',\hat{\mu}),$$

for any $\mu,\mu' \in \mathcal{P}(\mathcal{Y}\times\mathcal{Z})$ and any $\lambda\in[0,1]$. We claim that $\mathrm{TV}_\varepsilon^2(\cdot,\hat{\mu}) \in \mathfrak{C}^1(\mathcal{Y}\times\mathcal{Z})$.

Assume $\mu,\mu',\hat{\mu}$ are absolutely continuous with respect to Lebesgue on $\mathcal{Y}\times\mathcal{Z}$ with densities $m, m', \hat{m}$, respectively. Fix $\varepsilon > 0$ and define $\rho_\varepsilon(u) := \sqrt{u^2+\varepsilon^2}$, $\psi_\varepsilon(u) := \rho_\varepsilon'(u) = \frac{u}{\sqrt{u^2+\varepsilon^2}}$. For a perturbed measure $\mu + \delta(\mu'-\mu)$, with $\delta > 0$,

$$\lim_{\delta\to 0}\frac{1}{\delta}\left(\mathrm{TV}_\varepsilon(\mu+\delta(\mu'-\mu),\hat{\mu}) - \mathrm{TV}_\varepsilon(\mu,\hat{\mu})\right) = \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} \psi_\varepsilon\big(m(r)-\hat{m}(r)\big)(\mu'-\mu)(\mathrm{d}r),$$

so the flat derivative is

$$\frac{\delta\,\mathrm{TV}_\varepsilon(\cdot,\hat{\mu})}{\delta\mu}(\mu,r) = \frac{1}{2}\,\psi_\varepsilon\big(m(r)-\hat{m}(r)\big).$$

By the chain rule for $-\mathrm{TV}_\varepsilon^2$,

$$-\frac{\delta\,\mathrm{TV}_\varepsilon^2(\cdot,\hat{\mu})}{\delta\mu}(\mu,r) = -2\,\mathrm{TV}_\varepsilon(\mu,\hat{\mu})\frac{1}{2}\,\psi_\varepsilon\big(m(r)-\hat{m}(r)\big) = -\,\mathrm{TV}_\varepsilon(\mu,\hat{\mu})\,\psi_\varepsilon\big(m(r)-\hat{m}(r)\big).$$

Also, by linearity of F^0 in μ, it follows from [26, Lemma 4.1] that $\mu \mapsto F_\varepsilon(\nu,\mu)$ satisfies $D_{F_\varepsilon(\nu,\cdot)}(\mu',\mu) \le 0$. Therefore, F_ε satisfies Assumption 2.2.

To verify Assumption 3.3, we have by Proposition C.3 that $F_\varepsilon(\cdot,\mu)$ satisfies the first condition of Assumption 3.3. For the second condition of Assumption 3.3, recall $\rho_\varepsilon(u) := \sqrt{u^2+\varepsilon^2}$, $\psi_\varepsilon(u) := \rho_\varepsilon'(u) = \frac{u}{\sqrt{u^2+\varepsilon^2}}$. Then $|\psi_\varepsilon(u)| \le 1$ for all $u\in\mathbb{R}$ and moreover

$$\psi_\varepsilon'(u) = \frac{\varepsilon^2}{(u^2+\varepsilon^2)^{3/2}},$$

implies $0 \le \psi_\varepsilon'(u) \le \frac{1}{\varepsilon}$, hence $|\psi_\varepsilon(a) - \psi_\varepsilon(b)| \le \frac{1}{\varepsilon}|a-b|$, for all $a,b\in\mathbb{R}$.

Define the loss function

$$\ell_\nu(y,z) := \left|y - \mathbb{E}^{X\sim\nu}[\hat{\varphi}(X,z)]\right|^2.$$

Since the loss term $\frac{1}{2}\int \ell_\nu\, d\mu$ is linear in μ, its flat derivative of F^0 is $\frac{1}{2}\,\ell_\nu(r)$. Therefore

$$\frac{\delta F_\varepsilon}{\delta\mu}(\nu,\mu,r) = \frac{1}{2}\,\ell_\nu(r) \;-\; \mathrm{TV}_\varepsilon(\mu,\hat{\mu})\,\psi_\varepsilon\big(m(r)-\hat{m}(r)\big). \tag{52}$$

Subtracting (52) for μ' and μ cancels the loss term, yielding

$$\left|\frac{\delta F_\varepsilon}{\delta\mu}(\nu,\mu',r) - \frac{\delta F_\varepsilon}{\delta\mu}(\nu,\mu,r)\right| = \left|\,\mathrm{TV}_\varepsilon(\mu',\hat{\mu})\psi_\varepsilon(m'(r)-\hat{m}(r)) - \mathrm{TV}_\varepsilon(\mu,\hat{\mu})\psi_\varepsilon(m(r)-\hat{m}(r))\right|.$$

Using the triangle inequality,

$$\begin{aligned}
&\left|\,\mathrm{TV}_\varepsilon(\mu',\hat{\mu})\psi_\varepsilon(m'(r)-\hat{m}(r)) - \mathrm{TV}_\varepsilon(\mu,\hat{\mu})\psi_\varepsilon(m(r)-\hat{m}(r))\right| \\
&\le |\,\mathrm{TV}_\varepsilon(\mu',\hat{\mu}) - \mathrm{TV}_\varepsilon(\mu,\hat{\mu})||\psi_\varepsilon(m'(r)-\hat{m}(r))| \\
&+ \mathrm{TV}_\varepsilon(\mu,\hat{\mu})|\psi_\varepsilon(m'(r)-\hat{m}(r)) - \psi_\varepsilon(m(r)-\hat{m}(r))|.
\end{aligned}$$

We bound each term. Since ρ_ε is 1-Lipschitz, $|\rho_\varepsilon(a) - \rho_\varepsilon(b)| \le |a-b|$, hence

$$\begin{aligned}|\operatorname{TV}_\varepsilon(\mu', \hat{\mu}) - \operatorname{TV}_\varepsilon(\mu, \hat{\mu})| &= \frac{1}{2}\left|\int_{\mathcal{Y}\times\mathcal{Z}} \rho_\varepsilon(m' - \hat{m}) - \rho_\varepsilon(m - \hat{m})\mathrm{d}r\right| \\ &\le \frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Z}} |m' - m|\mathrm{d}r = \operatorname{TV}(\mu', \mu).\end{aligned}$$

Also, $|\psi_\varepsilon(m'(r) - \hat{m}(r))| \le 1$. By the Lipschitz bound on ψ_ε,

$$\left|\psi_\varepsilon\big(m'(r) - \hat{m}(r)\big) - \psi_\varepsilon\big(m(r) - \hat{m}(r)\big)\right| \le \frac{1}{\varepsilon}|m'(r) - m(r)|.$$

Combining,

$$\left|\frac{\delta F_\varepsilon}{\delta \mu}(\nu, \mu', r) - \frac{\delta F_\varepsilon}{\delta \mu}(\nu, \mu, r)\right| \le \operatorname{TV}(\mu', \mu) + \frac{\operatorname{TV}_\varepsilon(\mu, \hat{\mu})}{\varepsilon}\,|m'(r) - m(r)|,$$

for a.e. $r \in \mathcal{Y}\times\mathcal{Z}$.

Letting $\mu_\theta = \mu + \theta(\mu' - \mu)$, for $\theta \in [0,1]$, gives

$$\begin{aligned}\left|\frac{\delta F_\varepsilon}{\delta \mu}(\nu, \mu_\theta, r) - \frac{\delta F_\varepsilon}{\delta \mu}(\nu, \mu, r)\right| &\le \operatorname{TV}(\mu_\theta, \mu) + \frac{\operatorname{TV}_\varepsilon(\mu_\theta, \hat{\mu})}{\varepsilon}\,|m_\theta(r) - m(r)| \\ &\le \theta \operatorname{TV}(\mu', \mu) + \theta\left(\frac{1}{\varepsilon} + \frac{1}{2}|\mathcal{Y}\times\mathcal{Z}|\right)|m'(r) - m(r)|,\end{aligned}$$

where the last inequality follows from the convexity of TV and Lemma C.5. Lastly, by Definition 1.1,

$$\begin{aligned}&\left|F_\varepsilon(\nu, \mu') - F_\varepsilon(\nu, \mu) - \int_{\mathcal{Y}\times\mathcal{Z}} \frac{\delta F_\varepsilon}{\delta \mu}(\nu, \mu, r)(\mu' - \mu)(\mathrm{d}r)\right| \\ &\le \int_0^1 \theta \int_{\mathcal{Y}\times\mathcal{Z}} \left(\operatorname{TV}(\mu', \mu) + \left(\frac{1}{\varepsilon} + \frac{1}{2}|\mathcal{Y}\times\mathcal{Z}|\right)|m'(r) - m(r)|\right)|m'(r) - m(r)|\,\mathrm{d}r\mathrm{d}\theta \\ &= \frac{1}{2}\operatorname{TV}^2(\mu', \mu) + \frac{1}{2}\left(\frac{1}{\varepsilon} + \frac{1}{2}|\mathcal{Y}\times\mathcal{Z}|\right)\|m' - m\|^2_{L^2_{\mathcal{Y}\times\mathcal{Z}}} \\ &\le D_h(\mu', \mu) + \left(\frac{1}{\varepsilon} + \frac{1}{2}|\mathcal{Y}\times\mathcal{Z}|\right) D_h(\mu', \mu) \\ &= \left(1 + \frac{1}{\varepsilon} + \frac{1}{2}|\mathcal{Y}\times\mathcal{Z}|\right) D_h(\mu', \mu),\end{aligned}$$

where the last inequality follows since h is the χ^2-divergence in Example B.2 and since $\mathcal{Y}\times\mathcal{Z}$ is compact, we may take π to be the Lebesgue measure in that example. In particular, we take $L_\mu = 1 + \frac{1}{\varepsilon} + \frac{1}{2}|\mathcal{Y}\times\mathcal{Z}|$ in Assumption 3.3.

To verify Assumptions 2.3, note that

$$\begin{aligned}\left|\frac{\delta F_\varepsilon}{\delta \nu}(\nu, \mu, x)\right| &\le \int_{\mathcal{Y}\times\mathcal{Z}} \left|y - \mathbb{E}^{X\sim\nu}[\hat{\varphi}(X, z)]\right|\,|\hat{\varphi}(x, z)|\,\mu(\mathrm{d}y, \mathrm{d}z) \\ &\le M_\varphi\,(\mu_{\mathcal{Y}} + M_\varphi) \coloneqq C_1,\end{aligned}$$

where

$$\mu_{\mathcal{Y}} \coloneqq \int_{\mathcal{Y}\times\mathcal{Z}} |y|\mu(\mathrm{d}y, \mathrm{d}z) < \infty$$

since $\mathcal{Y}\times\mathcal{Z}$ is compact.

Similarly,

$$\begin{aligned}\left|\frac{\delta F_\varepsilon}{\delta \mu}(\nu, \mu, r)\right| &= \left|\frac{1}{2}\left|y - \mathbb{E}^{X\sim\nu}[\hat{\varphi}(X, z)]\right|^2 - \frac{1}{2}\,\psi_\varepsilon\big(m(r) - \hat{m}(r)\big)\right| \\ &\le \frac{1}{2}\,(1 + |\mathcal{Y}| + M_\varphi)^2 \coloneqq C_2,\end{aligned}$$

since $\mathcal{Y}$ is compact and $|\psi_\varepsilon| \le 1$. □

D Numerical experiments

In this section, we outline how to implement the infinite-dimensional algorithms 1 and 2 in the case where h is the relative entropy. For brevity, we present the derivations only for Algorithm 1, as the arguments for Algorithm 2 are entirely analogous. The complete algorithms for both the simultaneous and alternating MDA schemes can be found in Algorithm 3 and Algorithm 4 in Section D.4.

D.1 Scope and interpretation of the GAN experiments

The GAN experiments in this appendix provide a qualitative comparison of the practical stability and optimization behavior of simultaneous and alternating MDA. They are not designed to estimate or directly verify the theoretical convergence rates $\mathcal{O}(N^{-1/2})$ and $\mathcal{O}(N^{-2/3})$, respectively.

Theoretical guarantees versus empirical diagnostics. The main theorems control the NI error of averaged iterates generated by exact infinite-dimensional mirror steps, with stepsizes chosen as a function of the iteration horizon N. In contrast, the experiments report the W_1 distance at the current generated distribution using a finite-particle, minibatch-based Langevin approximation with neural-network parametrizations. Thus the theoretical and empirical quantities differ in the error criterion, the use of averaged versus current iterates, and the implementation of the updates. Therefore, the reported W_1 trajectories should not be identified with the averaged-iterate NI error appearing in the convergence bounds.

Averaging and last-iterate behavior. In particular, the simultaneous convergence theorem does not assert last-iterate convergence of the error. Averaging is essential to the analysis because simultaneous MDA can exhibit oscillations even in elementary bilinear games. Thus an NI error bound for averaged iterates is compatible with oscillatory behavior of the current iterates. Accordingly, the noisy behavior of the current generator in Figure 5 does not, by itself, contradict the simultaneous convergence theorem. Nor does that theorem directly control the last-iterate W_1 diagnostic of the approximate implementation.

Connection with the error analysis. Figure 5 primarily illustrates a qualitative difference in practical stability and optimization behavior, and should not be viewed as an asymptotic rate comparison. This distinction is related to the structure of the proof. The simultaneous update uses the opponent's previous strategy and leaves a first-order contribution in the stepsize to the relevant error estimate, whereas the alternating update uses the newly updated first-player strategy in the second-player step and reduces the corresponding contribution to higher order under its additional assumptions. This improved error control provides a theoretical distinction between the schemes, but it does not prove that the observed difference in W_1 is entirely attributable to their asymptotic NI convergence rates. Therefore, the experiments should be interpreted as a qualitative comparison of the implemented methods, not as a rate estimation experiment.

D.2 Simulation of infinite-dimensional MDA

As shown in Example B.1, by taking h to be the entropy, the corresponding h-Bregman divergence is exactly the KL divergence. Moreover, using the flat derivative formula (32), the first-order optimality condition in Proposition B.4 gives

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

for every $n \geq 0$ and, for all $x, y \in \mathcal{X}$ a.e., where $C, C' \in \mathbb{R}$. Exponentiating both sides, we obtain

$$\begin{cases} \nu^{n+1}(x) \propto \nu^n(x) e^{-\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)}, \\ \mu^{n+1}(y) \propto \mu^n(y) e^{\tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)}, \end{cases}$$

where the constants C, C' are absorbed into the normalizations. Iterating these multiplicative updates gives

$$\begin{cases} \nu^n(x) \propto \nu^0(x) \exp\left(-\tau \sum_{k=0}^{n-1} \frac{\delta F}{\delta \nu}(\nu^k, \mu^k, x)\right), \\ \mu^n(y) \propto \mu^0(y) \exp\left(\tau \sum_{k=0}^{n-1} \frac{\delta F}{\delta \mu}(\nu^k, \mu^k, y)\right). \end{cases}$$

Hence,

$$\begin{cases} \nabla \log \nu^n(x) = \nabla \log \nu^0(x) - \tau \sum_{k=0}^{n-1} \nabla \frac{\delta F}{\delta \nu}(\nu^k, \mu^k, x), \\ \nabla \log \mu^n(y) = \nabla \log \mu^0(y) - \tau \sum_{k=0}^{n-1} \nabla \frac{\delta F}{\delta \mu}(\nu^k, \mu^k, y), \end{cases}$$

since the normalization constant disappears after taking the gradient. Thus, in the Langevin implementation one does not need to compute $\nabla \log \nu^1$ and $\nabla \log \mu^1$ explicitly. One only needs $\nabla \log \nu^0$ and $\nabla \log \mu^0$, which is easy to evaluate for standard initializations such as Gaussian or uniform measures, together with the gradients of the flat derivatives along the iterates, which are computed using empirical averages on minibatches.

Suppose the initial samples $(X_j, Y_j)_{j=1}^J$ are drawn from (ν^0, μ^0). We set $(X_{j,0}, Y_{j,0})_{j=1}^J = (X_j, Y_j)_{j=1}^J$ and sample from (ν^1, μ^1) via Langevin dynamics:

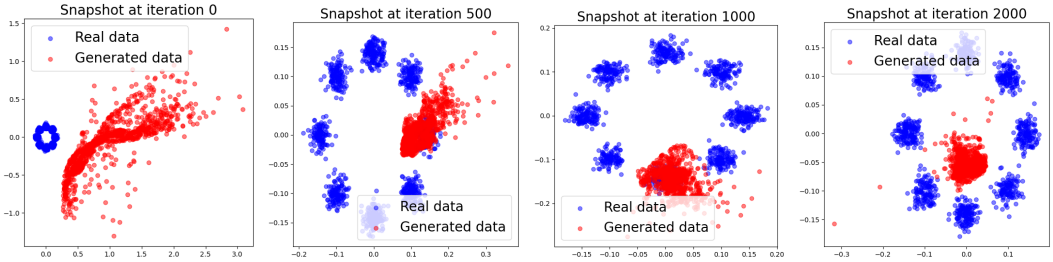
$$X_{j,t+1} = X_{j,t} + \gamma \left(\nabla \log \nu^0(X_{j,t}) - \tau \nabla \frac{\delta F}{\delta \nu}(\nu^0, \mu^0, X_{j,t}) \right) + \sqrt{\frac{2\gamma}{\tau}} \mathcal{N}_{j,t},$$

$$Y_{j,t+1} = Y_{j,t} + \gamma \left(\nabla \log \mu^0(Y_{j,t}) + \tau \nabla \frac{\delta F}{\delta \mu}(\nu^0, \mu^0, Y_{j,t}) \right) + \sqrt{\frac{2\gamma}{\tau}} \mathcal{N}_{j,t},$$

for $1 \le j \le J$ and $0 \le t \le T-1$, where $\gamma > 0$ is the step size and $\mathcal{N}_{j,t}$ are i.i.d standard Gaussian variables. For sufficiently large J and T, the terminal particles $(X_{j,T}, Y_{j,T})_{j=1}^J$ approximate samples from (ν^1, μ^1). Repeating this procedure recursively then yields samples from $(\nu^2, \mu^2), \ldots, (\nu^n, \mu^n)$. For simplicity, we may take ν_0, μ_0 be uniform distributions, so that $\nabla \log \nu^0 = \nabla \log \mu^0 = 0$.

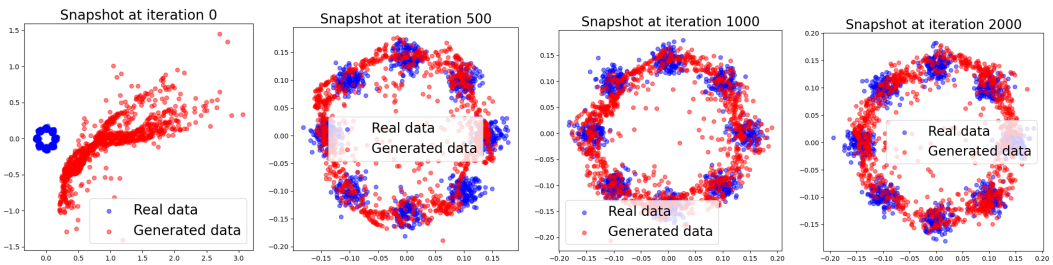
D.3 Training GANs by MDA

We train the mean-field GAN from Example 1.4 using simultaneous and alternating MDA-GAN (Algorithms 5 and 6) on the 8-Gaussian mixture and Swiss Roll datasets [24]. Full algorithmic details, including hyperparameters and network architectures, are in Section D.4. Both methods are run for 2000 iterations, with performance assessed by visualizing generated samples at 400, 1000, and 2000 iterations.



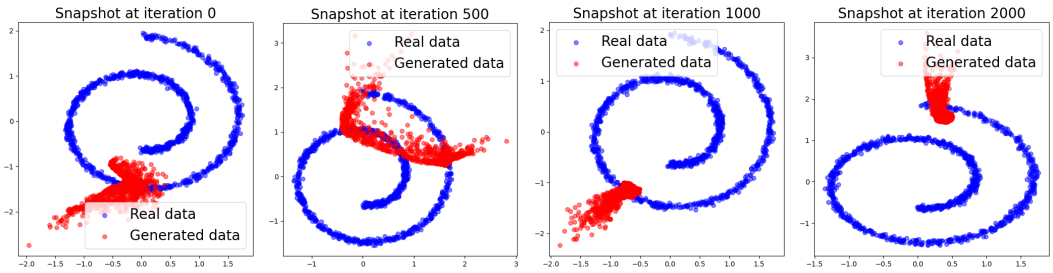
(a) Iteration $n = 0$ (b) Iteration $n = 500$ (c) Iteration $n = 1000$ (d) Iteration $n = 2000$

Figure 1: Simultaneous MDA-GAN (Algorithm 5) learning an 8-Gaussian mixture



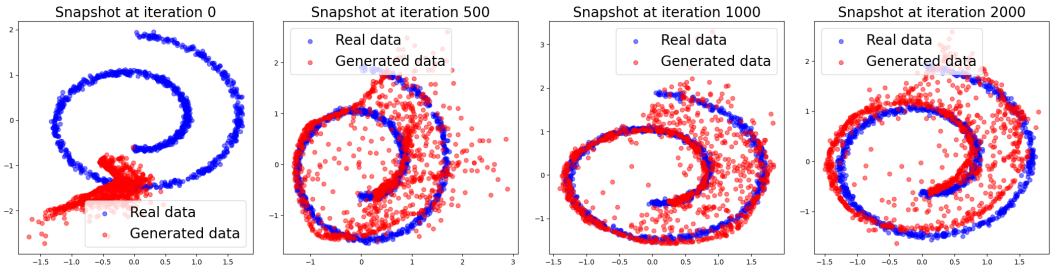
(a) Iteration $n = 0$ (b) Iteration $n = 500$ (c) Iteration $n = 1000$ (d) Iteration $n = 2000$

Figure 2: Alternating MDA-GAN (Algorithm 6) learning an 8-Gaussian mixture



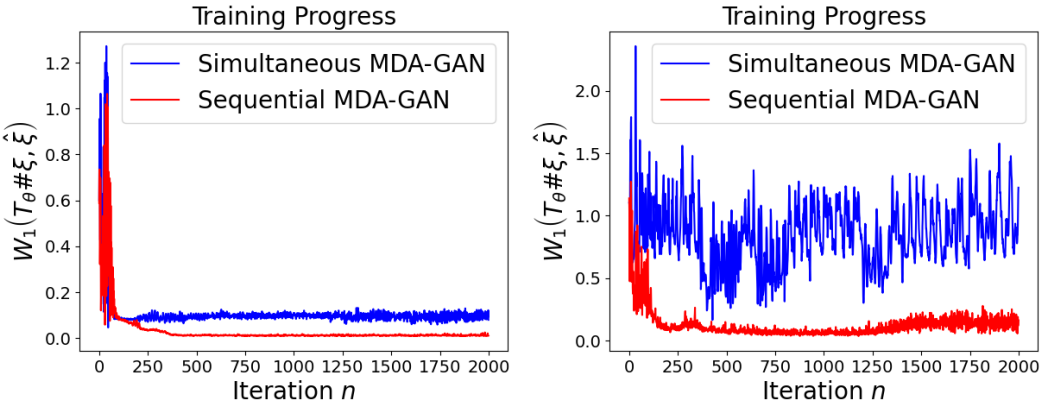
(a) Iteration $n = 0$ (b) Iteration $n = 500$ (c) Iteration $n = 1000$ (d) Iteration $n = 2000$

Figure 3: Simultaneous MDA-GAN (Algorithm 5) learning the Swiss Roll



(a) Iteration $n = 0$ (b) Iteration $n = 500$ (c) Iteration $n = 1000$ (d) Iteration $n = 2000$

Figure 4: Alternating MDA-GAN (Algorithm 6) learning the Swiss Roll



(a) 8-Gaussian mixture (b) Swiss Roll

Figure 5: L^1-Wasserstein distance between generated and real data for the 8-Gaussian mixture and Swiss Roll

Figures 1 and 2 show the training dynamics of simultaneous and alternating MDA-GANs on the 8-Gaussian mixture, with analogous results on the Swiss Roll in Figures 3 and 4. In both settings, generated samples start far from the data but the alternating variant captures the multi-modal structure and the spiral geometry of the Swiss Roll more clearly and at earlier iterations. In Figure 5, we plot the L^1-Wasserstein distance $W_1\big(T_{\theta^n}\#\xi, \hat{\xi}\big)$ for both tasks over iterations n, illustrating the more favorable empirical behavior of alternating MDA-GAN.

D.4 Details on numerical experiments

In this section, we present the additional details of the numerical experiments. We begin by summarizing the implementable versions of the simultaneous and alternating MDA algorithms introduced in Section D. We now turn to Algorithms 3 and 4 in the setting where F corresponds to the GAN

Algorithm 3: IMPLEMENTABLE SIMULTANEOUS MDA

Input: objective function F, initial measures (ν^0, μ^0), stepsize $\tau, \gamma > 0$, time horizons K, N and number of particles J

Generate i.i.d $\big(X_j^0, Y_j^0\big)_{j=1}^J \sim (\nu^0, \mu^0)$

Set $\big(X_{j,0}^0, Y_{j,0}^0\big)_{j=1}^J = \big(X_j^0, Y_j^0\big)_{j=1}^J$

for $n = 0, 1, \ldots, N-1$ **do**

  for $k = 0, 1, \ldots, K-1$ **do**

    Generate independent Gaussian random variables $\mathcal{N}_{j,k}^n$

    for $j = 1, 2, \ldots, J$ **do**

$$X_{j,k+1}^n = X_{j,k}^n + \gamma\left(\nabla \log \nu^n(X_{j,k}^n) - \tau\nabla\tfrac{\delta F}{\delta \nu}(\nu^n, \mu^n, X_{j,k}^n)\right) + \sqrt{\tfrac{2\gamma}{\tau}}\mathcal{N}_{j,k}^n$$

$$Y_{j,k+1}^n = Y_{j,k}^n + \gamma\left(\nabla \log \mu^n(Y_{j,k}^n) + \tau\nabla\tfrac{\delta F}{\delta \mu}(\nu^n, \mu^n, Y_{j,k}^n)\right) + \sqrt{\tfrac{2\gamma}{\tau}}\mathcal{N}_{j,k}^n$$

    end

  end

  for $j = 1, 2, \ldots, J$ **do**

$$X_{j,0}^{n+1} = X_{j,K}^n, \quad Y_{j,0}^{n+1} = Y_{j,K}^n$$

  end

$$\nu^n = \tfrac{1}{J}\textstyle\sum_{j=1}^J \delta_{X_{j,0}^n}, \quad \mu^n = \tfrac{1}{J}\sum_{j=1}^J \delta_{Y_{j,0}^n}$$

end

Generate $\hat{n} \sim \mathrm{Unif}\{0, \ldots, N-1\}$

Output: $(\nu^{\hat{n}}, \mu^{\hat{n}})$

objective introduced in Example 1.4. Recall that F takes the form

$$\begin{aligned} F(\nu, \mu) &= \int_{\mathcal{W}}\int_{\Theta}\int_{\mathcal{Y}} D_w(y)\left(T_\theta\#\xi - \hat{\xi}\right)(\mathrm{d}y)\nu(\mathrm{d}\theta)\mu(\mathrm{d}w) \\ &= \int_{\mathcal{W}}\int_{\Theta}\int_{\mathcal{Y}} D_w(y)\left(T_\theta\#\xi\right)(\mathrm{d}y)\nu(\mathrm{d}\theta)\mu(\mathrm{d}w) - \int_{\mathcal{W}}\int_{\mathcal{Y}} D_w(y)\hat{\xi}(\mathrm{d}y)\mu(\mathrm{d}w) \end{aligned}$$

By Definition 1.1, we have

$$\frac{\delta F}{\delta \nu}(\nu, \mu, \theta) = \int_{\mathcal{W}}\int_{\mathcal{Y}} D_w(y)\left(T_\theta\#\xi\right)(\mathrm{d}y)\mu(\mathrm{d}w),$$

$$\frac{\delta F}{\delta \mu}(\nu, \mu, w) = \int_{\Theta}\int_{\mathcal{Y}} D_w(y)\left(T_\theta\#\xi\right)(\mathrm{d}y)\nu(\mathrm{d}\theta) - \int_{\mathcal{Y}} D_w(y)\hat{\xi}(\mathrm{d}y).$$

The flat derivatives can be approximated using empirical averages. For a batch of real data $\{\xi_1^{\text{real}}, \ldots, \xi_M^{\text{real}}\} \sim \hat{\xi}$, we have

$$\int_{\mathcal{Y}} D_w(y)\hat{\xi}(\mathrm{d}y) \approx \frac{1}{M}\sum_{i=1}^M D_w(\xi_i^{\text{real}}).$$

Algorithm 4: IMPLEMENTABLE ALTERNATING MDA

Input: objective function F, initial measures (ν^0, μ^0), stepsize $\tau, \gamma > 0$, time horizons K, N and number of particles J

Generate i.i.d $\left(X_j^0, Y_j^0\right)_{j=1}^J \sim (\nu^0, \mu^0)$

Set $\left(X_{j,0}^0, Y_{j,0}^0\right)_{j=1}^J = \left(X_j^0, Y_j^0\right)_{j=1}^J$

for $n = 0, 1, \ldots, N-1$ **do**

  for $k = 0, 1, \ldots, K-1$ **do**

    Generate independent Gaussian random variables $\mathcal{N}_{j,t}^n$

    for $j = 1, 2, \ldots, J$ **do**

      $X_{j,k+1}^n = X_{j,k}^n + \left(\nabla \log \nu^n(X_{j,k}^n) - \gamma \nabla \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, X_{j,k}^n)\right) + \sqrt{\frac{2\gamma}{\tau}}\mathcal{N}_{j,k}^n$

    end

  end

  for $j = 1, 2, \ldots, J$ **do**

    $X_{j,0}^{n+1} = X_{j,K}^n$

  end

  $\nu^{n+1} = \frac{1}{J}\sum_{j=1}^J \delta_{X_{j,0}^{n+1}}$

  for $k = 0, 1, \ldots, K-1$ **do**

    Generate independent Gaussian random variables $\mathcal{N}_{j,k}^n$

    for $j = 1, 2, \ldots, J$ **do**

      $Y_{j,k+1}^n = Y_{j,k}^n + \gamma\left(\nabla \log \mu^n(Y_{j,k}^n) + \tau \nabla \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, Y_{j,k}^n)\right) + \sqrt{\frac{2\gamma}{\tau}}\mathcal{N}_{j,k}^n$

    end

  end

  for $j = 1, 2, \ldots, J$ **do**

    $Y_{j,0}^{n+1} = Y_{j,K}^n$

  end

  $\mu^n = \frac{1}{J}\sum_{j=1}^J \delta_{Y_{j,0}^n}$

end

Generate $\hat{n} \sim \mathrm{Unif}\{0, \ldots, N-1\}$

Output: $(\nu^{\hat{n}}, \mu^{\hat{n}})$

For the term in $\frac{\delta F}{\delta \mu}(\nu, \mu, w)$ that involves integration with respect to both ν and the generated data $T_\theta \# \xi$, we approximate via sampling as follows. We sample

$$\{\theta^{(1)}, \theta^{(2)}, ..., \theta^{(J)}\} \sim \nu, \quad \left\{Z_i^{(j)}\right\}_{i=1}^M \sim T_{\theta^{(j)}} \# \xi,$$

leading to the estimator

$$\int_\Theta \int_{\mathcal{Y}} D_w(y) \left(T_\theta \# \xi\right)(\mathrm{d}y) \nu(\mathrm{d}\theta) \approx \frac{1}{JM} \sum_{i=1}^M \sum_{j=1}^J D_w\left(X_i^{(j)}\right).$$

Analogously, for $\frac{\delta F}{\delta \nu}(\nu, \mu, \theta)$ we sample

$$\{w^{(1)}, w^{(2)}, ..., w^{(J)}\} \sim \mu, \quad \{Z_i\}_{i=1}^M \sim T_\theta \# \xi,$$

and approximate

$$\int_{\mathcal{W}} \int_{\mathcal{Y}} D_w(y) \left(T_\theta \# \xi\right)(\mathrm{d}y) \mu(\mathrm{d}w) \approx \frac{1}{JM} \sum_{i=1}^M \sum_{j=1}^J D_{w^{(j)}}\left(Z_i\right).$$

To mitigate the computational cost of Algorithms 3 and 4, we follow the approach of [25] and employ Langevin dynamics with exponential damping (see also their Algorithm 3). Algorithm 3 is essentially a realization of the Mirror-GAN algorithm of [25], with different architectures and hyperparameter

Algorithm 5: SIMULTANEOUS MDA-GAN

Input: Initial parameters w^0, θ^0, step sizes $\{\gamma^n\}_{n=0}^{N-1}$, $\{\tau^n\}_{n=0}^{N-1}$, time horizon $\{K^n\}_{n=0}^{N-1}$, averaging parameter $\beta \in [0,1]$, source probability measure ξ

for $n = 0, 1, \ldots, N-1$ **do**

Set $\bar{w}^n, w_0^n = w^n$ and $\bar{\theta}^n, \theta_0^n = \theta^n$;

for $k = 0, 1, \ldots, K_n - 1$ **do**

$A = \{Z_1, \ldots, Z_M\} \sim T_{\theta_k^n} \# \xi$;

$$\theta_{k+1}^n = \theta_k^n - \frac{\gamma^n}{M} \nabla_\theta \sum_{Z_i \in A} D_{w^n}(Z_i) + \sqrt{\frac{2\gamma^n}{\tau^n}} \mathcal{N}_k^n;$$

$B = \{\xi_1^{\text{real}}, \ldots, \xi_M^{\text{real}}\} \sim \hat{\xi}$;
$B' = \{Z_1', \ldots, Z_M'\} \sim T_{\theta^n} \# \xi$;

$$w_{k+1}^n = w_k^n + \frac{\gamma^n}{M} \nabla_w \sum_{Z_i' \in B'} D_{w_k^n}(Z_i') - \frac{\gamma^n}{M} \nabla_w \sum_{\xi_i^{\text{real}} \in B} D_{w_k^n}(\xi_i^{\text{real}}) + \sqrt{\frac{2\gamma^n}{\tau^n}} \mathcal{N}_k^n;$$

$\bar{w}^n = (1-\beta)\bar{w}^n + \beta w_{k+1}^n, \quad \bar{\theta}^n = (1-\beta)\bar{\theta}^n + \beta \theta_{k+1}^n$;

end

$w^{n+1} = (1-\beta)w^n + \beta \bar{w}^n, \quad \theta^{n+1} = (1-\beta)\theta^n + \beta \bar{\theta}^n$;

end

Output: w^N, θ^N

Algorithm 6: ALTERNATING MDA-GAN

Input: Initial parameters w^0, θ^0, step sizes $\{\gamma^n\}_{n=0}^{N-1}$, $\{\tau^n\}_{n=0}^{N-1}$, time horizon $\{K^n\}_{n=0}^{N-1}$, averaging parameter $\beta \in [0,1]$, source probability measure ξ

for $n = 0, 1, \ldots, N-1$ **do**

Set $\bar{w}^n, w_0^n = w^n$ and $\bar{\theta}^n, \theta_0^n = \theta^n$;

for $k = 0, 1, \ldots, K_n - 1$ **do**

$A = \{Z_1, \ldots, Z_M\} \sim T_{\theta_k^n} \# \xi$;

$$\theta_{k+1}^n = \theta_k^n - \frac{\gamma^n}{M} \nabla_\theta \sum_{Z_i \in A} D_{w^n}(Z_i) + \sqrt{\frac{2\gamma^n}{\tau^n}} \mathcal{N}_k^n;$$

$\bar{\theta}^n = (1-\beta)\bar{\theta}^n + \beta \theta_{k+1}^n$;

end

$\theta^{n+1} = (1-\beta)\theta^n + \beta \bar{\theta}^n$;

for $k = 0, 1, \ldots, K_n - 1$ **do**

$B = \{\xi_1^{\text{real}}, \ldots, \xi_M^{\text{real}}\} \sim \hat{\xi}$;
$B' = \{Z_1', \ldots, Z_M'\} \sim T_{\theta^{n+1}} \# \xi$;

$$w_{k+1}^n = w_k^n + \frac{\gamma^n}{M} \nabla_w \sum_{Z_i' \in B'} D_{w_k^n}(Z_i') - \frac{\gamma^n}{M} \nabla_w \sum_{\xi_i^{\text{real}} \in B} D_{w_k^n}(\xi_i^{\text{real}}) + \sqrt{\frac{2\gamma^n}{\tau^n}} \mathcal{N}_k^n;$$

$\bar{w}^n = (1-\beta)\bar{w}^n + \beta w_{k+1}^n$;

end

$w^{n+1} = (1-\beta)w^n + \beta \bar{w}^n$;

end

Output: w^N, θ^N

choices. The alternating variant, Algorithm 4, which is not studied in [25], is obtained by adapting

that simultaneous algorithm to the alternating update setting. Below, we present the simultaneous and alternating algorithms used in our experiments.

In all experiments, we closely follow the specifications from [25]. We adopt the gradient-penalized discriminator of [24] as a soft-constraint alternative to the original Wasserstein GAN formulation to increase stability. The gradient penalty parameter is set to $\lambda = 0.1$. For our Simultaneous and alternating MDA-GANs, we fix the damping factor to $\beta = 0.8$. The scheduling of the parameters K^n, γ^n, and τ^n is $K^n = \lfloor (1 + 10^{-5})^n \rfloor$, $\gamma^n = \gamma(1 - 10^{-5})^n$, with $\gamma = 0.01$, and $\tau^n = \tau(1 - 5 \times 10^{-5})^{-n}$, with $\tau = 100$. The number of samples per batch is $M = 1024$. For both the 8-Gaussian mixture and Swiss Roll datasets, we use fully connected networks for the generator and discriminator, each consisting of two-hidden-layers with $J = 512$ neurons on each layer. The generator and discriminator networks use ReLU activations, except for the output layer of the discriminator, which employs a tanh activation. All network parameters are initialized from a normal distribution $\mathcal{N}(0, 0.01)$.

Remark D.1 (Implementation of the measure space mirror steps)**.** *Algorithms 1 and 2 are formulated as idealized infinite-dimensional mirror algorithms, and our convergence results concern their exact measure space updates. For a general mirror potential h, the corresponding proximal step is itself an optimization problem over probability measures and may require an expensive inner solver. Moreover, even when the proximal problem has a tractable variational characterization, as we saw in the case of relative entropy, the flat derivatives entering the update generally involve expectations or integrals that must be approximated from samples.*

Thus, relative entropy gives a closed-form density update at the variational level, but its numerical realization can still be nontrivial since the normalizing constant may be unavailable in closed form, the flat derivatives may require Monte Carlo estimation, and sampling from the exponentially tilted measures may require an additional sampling procedure. Therefore, in line with particle algorithms that are standard in the mean-field literature, we implement an inexact particle approximation using finitely many particles, minibatch estimates, and Langevin dynamics. The experiments should accordingly be interpreted as an illustrative approximate realization of the ideal measure space algorithms, rather than as claiming generic efficient implementability. For other mirror geometries, such as the χ^2 divergence, the proximal update may additionally require an infinite-dimensional projection and may activate feasibility constraints. A joint analysis of finite-particle approximation, stochastic estimation, and Langevin sampling errors is beyond the scope of the present convergence theory.

E Differentiability on the primal space

Remark E.1. *One can show that if $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ admits a flat derivative $\frac{\delta F}{\delta \mu}$, then for all $\mu, \mu' \in \mathcal{P}(\mathbb{R}^d)$, the function $[0,1] \ni \varepsilon \mapsto F(\mu^\varepsilon)$ is continuous on $[0,1]$ and differentiable on $(0,1)$ with derivative $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} F(\mu^\varepsilon) = \int_{\mathbb{R}^d} \frac{\delta F}{\delta \mu}(\mu^\varepsilon, x)(\mu' - \mu)(\mathrm{d}x)$ (see [27, Theorem 2.3]). Hence, by the fundamental theorem of calculus, $F(\mu') - F(\mu) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta \mu}(\mu^\varepsilon, x)(\mu' - \mu)(\mathrm{d}x)\mathrm{d}\varepsilon$, provided that $\varepsilon \mapsto \int \frac{\delta F}{\delta \mu}(\mu^\varepsilon, x)(\mu' - \mu)(\mathrm{d}x)$ is integrable.*

F Differentiability on the dual space

In this section, we start by recalling the notions of Fréchet and Gâteaux derivative for functions $H : C_b(\mathcal{X}) \to \mathbb{R}$, where $(C_b(\mathcal{X}), \|\cdot\|_\infty)$ is the Banach space of real-valued bounded continuous functions on $\mathcal{X} \subseteq \mathbb{R}^d$; see e.g. Chapters $7, 1, 3$ in [2, 3, 45], respectively. Based on these notions of differentiablity, we will introduce the notion of first variation for functions H.

F.1 Preliminaries on Fréchet and Gâteaux derivatives

For $\mathcal{X} \subseteq \mathbb{R}^d$, let $\mathcal{L}(C_b(\mathcal{X}), \mathbb{R})$ and $\mathcal{L}(C_b(\mathcal{X}))$ denote the space of continuous linear maps from $C_b(\mathcal{X})$ to $\mathbb{R}$, and from $C_b(\mathcal{X})$ to itself, respectively.

Definition F.1 (Fréchet differentiability)**.** *Let $\mathcal{U} \subseteq C_b(\mathcal{X})$ be open. Given $f \in \mathcal{U}$, the function $H : \mathcal{U} \to \mathbb{R}$ is Fréchet differentiable at f if there exists $T \in \mathcal{L}(C_b(\mathcal{X}), \mathbb{R})$ such that, for all*

$g \in C_b(\mathcal{X})$,

$$\lim_{\|g\|_\infty \to 0} \frac{|H(f+g) - H(f) - T[g]|}{\|g\|_\infty} = 0.$$

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 F.2 (Convex conjugate of the relative entropy)**.** *If h is the relative entropy in Example B.1, then a straightforward calculation directly from Definition 3.1 shows that its dual h^* is given by*

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

Example F.3 (Fréchet derivative of the relative entropy)**.** *A straightforward calculation directly from Definition F.1 shows that h^* is Fréchet differentiable on $C_b(\mathcal{X})$ with Fréchet derivative given by*

$$\nabla_{\mathcal{F}} h^*(f)[g] = \int_{\mathcal{X}} g(z) \frac{e^{f(z)}}{\int_{\mathcal{X}} e^{f(y)} \pi(\mathrm{d}y)} \pi(\mathrm{d}z),$$

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

Example F.4 (Convex conjugate of the χ^2-divergence)**.** *If h is the χ^2-divergence in Example B.2, then [46, Example 7.4] shows that its dual h^* is given by*

$$h^*(f) = \frac{1}{2}\int_{\mathcal{X}} f(z)\pi(\mathrm{d}z) + \frac{1}{8}\int_{\mathcal{X}} f^2(z)\pi(\mathrm{d}z).$$

Example F.5 (Fréchet derivative of the relative χ^2-divergence)**.** *A straightforward calculation directly from Definition F.1 shows that h^* is Fréchet differentiable on $C_b(\mathcal{X})$ with Fréchet derivative given by*

$$\nabla_{\mathcal{F}} h^*(f)[g] = \frac{1}{2}\int_{\mathcal{X}} g(z)\pi(\mathrm{d}z) + \frac{1}{4}\int_{\mathcal{X}} g(z) f(z)\pi(\mathrm{d}z),$$

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

Definition F.6 (Gâteaux differentiability)**.** *Let $\mathcal{U} \subseteq C_b(\mathcal{X})$ be open. Given $f \in \mathcal{U}$, the function $H : \mathcal{U} \to \mathbb{R}$ is Gâteaux differentiable at f if there exists $T \in \mathcal{L}(C_b(\mathcal{X}), \mathbb{R})$ such that for any direction $f' \in C_b(\mathcal{X})$,*

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

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, 45], 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, [45, 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 [45, Proposition 3.1.4].

Following the discussions in [2, 3, 45], it is possible to extend Definition F.1 to higher-order Fréchet derivatives.

Definition F.7 (Second-order Fréchet differentiability)**.** *Let $\mathcal{U} \subseteq C_b(\mathcal{X})$ be open and let $f \in \mathcal{U}$. Suppose that $H : \mathcal{U} \to \mathbb{R}$ is Fréchet differentiable (cf. Definition F.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 $T \in \mathcal{L}(C_b(\mathcal{X}), \mathcal{L}(C_b(\mathcal{X}), \mathbb{R}))$ such that for all $f', f'' \in C_b(\mathcal{X})$,*

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

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 F.8 (Second-order Fréchet derivative of the relative entropy)**.** *If h is the relative entropy in Example B.1, using Example F.3, we can show that $\nabla_{\mathcal{F}} h^*(f)$ is Fréchet differentiable on $C_b(\mathcal{X})$ with Fréchet derivative given by*

$$\begin{aligned}\nabla^2_{\mathcal{F}} h^*(f)[g'][g] &= \int_{\mathcal{X}} g(x)g'(x)\varphi(f)(\mathrm{d}x) - \left(\int_{\mathcal{X}} g'(z)\varphi(f)(\mathrm{d}z)\right)\left(\int_{\mathcal{X}} g(z)\varphi(f)(\mathrm{d}z)\right) \\ &= \mathrm{Cov}_{\varphi(f)}(g', g),\end{aligned}$$

for all $g, g' \in C_b(\mathcal{X})$, where

$$\varphi(f)(\mathrm{d}x) \coloneqq \frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(y)}\pi(\mathrm{d}y)}\pi(\mathrm{d}x)$$

If $g' = g$, then

$$\nabla^2_{\mathcal{F}} h^*(f)[g, g] = \mathrm{Var}_{\varphi(f)}(g).$$

Example F.9 (Second-order Fréchet derivative of the χ^2-divergence)**.** *If h is the χ^2-divergence in Example B.2, using Example F.5, we can show that $\nabla_{\mathcal{F}} h^*(f)$ is Fréchet differentiable on $C_b(\mathcal{X})$ with Fréchet derivative given by*

$$\nabla^2_{\mathcal{F}} h^*(f)[g'][g] = \frac{1}{4}\int_{\mathcal{X}} g(z)g'(z)\pi(\mathrm{d}z),$$

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

Definition F.10 (Third-order Fréchet differentiability)**.** *Let $\mathcal{U} \subseteq C_b(\mathcal{X})$ be open and let $f \in \mathcal{U}$. Suppose that $H : \mathcal{U} \to \mathbb{R}$ is twice Fréchet differentiable (cf. Definition F.7) at f, and admits second-order Fréchet derivative $\nabla^2_{\mathcal{F}} H(f)$. Then $\nabla^2_{\mathcal{F}} H(f)$ is Fréchet differentiable at f, if there exists $T \in \mathcal{L}(C_b(\mathcal{X}), \mathcal{L}(C_b(\mathcal{X}), \mathcal{L}(C_b(\mathcal{X}), \mathbb{R})))$ such that for all $g, g', g'' \in C_b(\mathcal{X})$,*

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

If it exists, the map T is unique, we write $T = \nabla^3_{\mathcal{F}} H(f)$, and call $\nabla^3_{\mathcal{F}} H(f)$ the third Fréchet derivative of H at f.

Example F.11 (Third-order Fréchet derivative of the relative entropy)**.** *If h is the relative entropy in Example B.1, using Example F.8, we can show that $\nabla^2_{\mathcal{F}} h^*(f)$ is Fréchet differentiable on $C_b(\mathcal{X})$ with Fréchet derivative $\nabla^3_{\mathcal{F}} h^*(f)$. We differentiate the variance $\mathrm{Var}_{\varphi(f)}(g)$ with respect to f in a direction g'. Using the identity*

$$\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0}\left(\int_{\mathcal{X}} G(x)\varphi(f + \varepsilon g')(\mathrm{d}x)\right) = \mathrm{Cov}_{\varphi(f)}(G, g'),$$

we obtain

$$\begin{aligned}\nabla^3_{\mathcal{F}} h^*(f)[g, g][g'] &= \left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0}\left(\int_{\mathcal{X}} g(x)^2\varphi(f + \varepsilon g')(\mathrm{d}x) - \left(\int_{\mathcal{X}} g(x)\varphi(f + \varepsilon g')(\mathrm{d}x)\right)^2\right) \\ &= \mathrm{Cov}_{\varphi(f)}(g^2, g') - 2\,\mathrm{Cov}_{\varphi(f)}(g, g')\int_{\mathcal{X}} g(x)\varphi(f)(\mathrm{d}x),\end{aligned}$$

for all $g, g' \in C_b(\mathcal{X})$. If $g' = g$, then

$$\nabla^3_{\mathcal{F}} h^*(f)[g, g, g] = \int_{\mathcal{X}}\left(g(x) - \int_{\mathcal{X}} g(y)\varphi(f)(\mathrm{d}y)\right)^3\varphi(f)(\mathrm{d}x).$$

Example F.12 (Third-order Fréchet derivative of the χ^2-divergence)**.** *If h is the χ^2-divergence in Example B.2, using Example F.9, we observe that since $\nabla^2_{\mathcal{F}} h^*(f)$ is independent of f, the third-order Fréchet derivative of h^* is given by*

$$\nabla^3_{\mathcal{F}} h^*(f)[g, g, g] = 0,$$

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

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 $\mathcal{L}\left(C_b(\mathcal{X}), \mathcal{L}\left(C_b(\mathcal{X}), \mathcal{L}\left(C_b(\mathcal{X}), \mathbb{R}\right)\right)\right)$ is isometrically isomorphic to $\mathcal{L}_3\left(C_b(\mathcal{X}), \mathbb{R}\right)$, i.e., the space of continuous trilinear maps from $C_b(\mathcal{X}) \times C_b(\mathcal{X}) \times C_b(\mathcal{X})$ to $\mathbb{R}$, and therefore, we could naturally view the third-order Fréchet derivative of H, if it exists, as a continuous trilinear map. Furthermore, due to [3, Theorem 3.5], we have that the third-order Fréchet derivative is always symmetric. On the contrary, the second-order Gâteaux derivative is not necessarily symmetric as noted on page 78 in [45].

Remark F.13. *If we replace $C_b(\mathcal{X})$ with $\mathbb{R}^d$, then the first and second-order Fréchet derivatives are precisely the gradient and Hessian matrix of H at f.*

Proposition F.14 (Verification of Assumption 3.4 for the relative entropy in Example B.1)**.** *For h being the relative entropy in Example B.1, its third Frechet derivative in Example F.11 satisfies Assumption 3.4.*

Proof. Let $g \in C_b(\mathcal{X})$. Recall that

$$\nabla^3_{\mathcal{F}} h^*(f)[g,g,g] = \int_{\mathcal{X}} \left(g(x) - \int_{\mathcal{X}} g(y)\varphi(f)(\mathrm{d}y) \right)^3 \varphi(f)(\mathrm{d}x).$$

Note that since $\varphi(f) \in \mathcal{P}(\mathcal{X})$ we have

$$\left| g(x) - \int_{\mathcal{X}} g(y)\varphi(f)(\mathrm{d}y) \right| \leq 2\|g\|_\infty,$$

and hence

$$|\nabla^3_{\mathcal{F}} h^*(f)[g,g,g]| \leq 8\|g\|^3_\infty.$$

Using the fact that

$$\left\|\nabla^3_{\mathcal{F}} h^*(f)\right\|_{\mathcal{L}_3(C_b(\mathcal{X}),\mathbb{R})} \coloneqq \sup_{\|g\|_\infty = 1} \frac{|\nabla^3_{\mathcal{F}} h^*(f)[g,g,g]|}{\|g\|^3_\infty},$$

we conclude that

$$\left\|\nabla^3_{\mathcal{F}} h^*(f)\right\|_{\mathcal{L}_3(C_b(\mathcal{X}),\mathbb{R})} \leq 8,$$

for all $f \in C_b(\mathcal{X})$. □

Example F.15 (Why self-concordance may not hold for the entropy mirror map)**.** *When h is the relative entropy in Example B.1, $\nabla^2_{\mathcal{F}} h^*(f)[g,g] = \mathrm{Var}_{\varphi(f)}(g)$ and $\nabla^3_{\mathcal{F}} h^*(f)[g,g,g]$ is the centered third moment of g under $\varphi(f)$, i.e.,*

$$\nabla^3_{\mathcal{F}} h^*(f)[g,g,g] = \int_{\mathcal{X}} \left(g(x) - \int_{\mathcal{X}} g(y)\varphi(f)(\mathrm{d}y) \right)^3 \varphi(f)(\mathrm{d}x).$$

In this case, self-concordance can fail globally because the normalized skewness

$$\frac{\int_{\mathcal{X}} \left(g(x) - \int_{\mathcal{X}} g(y)\varphi(f)(\mathrm{d}y) \right)^3 \varphi(f)(\mathrm{d}x)}{\left(\mathrm{Var}_{\varphi(f)}(g)\right)^{3/2}}$$

can be arbitrarily large as $\varphi(f)$ becomes highly concentrated.

Proposition F.16 (Verification of Assumption 3.4 for the relative χ^2-divergence in Example B.2)**.** *For h being the χ^2-divergence in Example B.2, its third Frechet derivative in Example F.12 satisfies Assumption 3.4.*

Proof. Recall that

$$\nabla^3_{\mathcal{F}} h^*(f)[g,g,g] = 0,$$

hence the conclusion is immediate. □

F.2 First variation

Following Chapter 2 from [1], we introduce the notion of first variation for Fréchet differentiable functions H, relative to the duality pairing (6).

Definition F.17 (First variation of H)**.** *Let $H : C_b(\mathcal{X}) \to \mathbb{R}$ be Fréchet differentiable at $f \in C_b(\mathcal{X})$. If it exists, the first variation of H at f is the element $\frac{\delta H}{\delta f}(f) \in \mathcal{M}(\mathcal{X})$ such that, for all $g \in C_b(\mathcal{X})$,*

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

Example F.18 (First variation of the dual of relative entropy)**.** *From Example F.2, we observe that the first variation $\frac{\delta h^*}{\delta f}(f) \in \mathcal{P}(\mathcal{X}) \subseteq \mathcal{M}(\mathcal{X})$ of h^* at f is given by*

$$\frac{\delta h^*}{\delta f}(f)(\mathrm{d}z) = \varphi(f)(\mathrm{d}z).$$

Example F.19 (First variation of the dual of χ^2-divergence)**.** *From Example F.4, we observe that the first variation $\frac{\delta h^*}{\delta f} \in \mathcal{M}(\mathcal{X})$ of h^* at f is given by*

$$\frac{\delta h^*}{\delta f}(f)(\mathrm{d}z) = \left(1 + \frac{1}{2} f(z)\right) \pi(\mathrm{d}z).$$

Assuming that $H : C_b(\mathcal{X}) \to \mathbb{R}$ is Fréchet differentiable at $f \in C_b(\mathcal{X})$ with Fréchet derivative $\nabla_{\mathcal{F}} H(f)$, then it is Gâteaux differentiable (cf. Definition F.6) with the same derivative, and therefore the first variation of H at f can be characterized as

$$\left\langle g, \frac{\delta H}{\delta f}(f) \right\rangle = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(H\left(f + \varepsilon g\right) - H\left(f\right) \right), \tag{53}$$

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

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

Lemma F.20 (Necessary first-order condition on $C_b(\mathcal{X})$)**.** *Suppose $H : C_b(\mathcal{X}) \to \mathbb{R}$ 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 F.21 (Sufficient first-order condition on $C_b(\mathcal{X})$)**.** *Let $\mathcal{U} \subseteq C_b(\mathcal{X})$ 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 F.22. *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 Lemma B.3, it can be showed 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. □

G 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 3.

Proposition G.1. *Let Assumption 2.1 hold. Let $h^* : C_b(\mathcal{X}) \to \mathbb{R}$ be the convex conjugate of h. Then, the following are equivalent:*

1. *The supremum of $\mathcal{M}(\mathcal{X}) \ni m \mapsto \langle g^*, m\rangle - h(m) \in \mathbb{R}$ is attained at $m = m^*$.*
2. *We have the first-order condition $g^*(x) - \frac{\delta h}{\delta m}(m^*, x) = 0$, for all $x \in \mathcal{X}$, m^*-a.e.*
3. *The supremum of $C_b(\mathcal{X}) \ni g \mapsto \langle g, m^*\rangle - h^*(g) \in \mathbb{R}$ is attained at $g = g^*$.*
4. *It holds that $m^* = \frac{\delta h^*}{\delta g}(g^*)$.*

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 $m^* = \operatorname{argmax}_{m\in\mathcal{M}(\mathcal{X})} \{\langle g^*, m\rangle - h(m)\}$. Hence,

$$\langle g^*, m^* - m\rangle - (h(m^*) - h(m)) \geq 0,$$

for all $m \in \mathcal{M}(\mathcal{X})$. Let $\tilde{m} \in \mathcal{M}(\mathcal{X})$ and set $m \coloneqq m^* + t(\tilde{m} - m^*)$, for $t \in [0, 1]$. Then

$$-t\langle g^*, \tilde{m} - m^*\rangle + (h(m^* + t(\tilde{m} - m^*)) - h(m^*)) \geq 0.$$

Dividing by t and letting $t \searrow 0$ gives

$$-\langle g^*, \tilde{m} - m^*\rangle + \int_{\mathcal{X}} \frac{\delta h}{\delta m}(m^*, x)(\tilde{m} - m^*)(\mathrm{d}x) \geq 0,$$

or equivalently

$$\left\langle -g^* + \frac{\delta h}{\delta m}(m^*, \cdot), \tilde{m} - m^* \right\rangle \geq 0.$$

Since $\tilde{m}$ is arbitrary, m^* satisfies the first-order condition

$$g^*(x) - \frac{\delta h}{\delta m}(m^*, x) = 0,$$

for all $x \in \mathcal{X}$, m^*-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 $m \mapsto \langle g^*, m\rangle$. Therefore, m^* is the maximizer of the map $\mathcal{M}(\mathcal{X}) \ni m \mapsto \langle g^*, m\rangle - h(m) \in \mathbb{R}$, 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 $g^* \in \operatorname{argmax}_{g\in C_b(\mathcal{X})} \{\langle g, m^*\rangle - h^*(g)\}$. Hence, by Lemma F.20, 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 $C_b(\mathcal{X})$ is convex 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 F.21, it follows that g^* is a maximizer of the map $C_b(\mathcal{X}) \ni g \mapsto \langle g, m^*\rangle - h^*(g) \in \mathbb{R}$, and so (3) holds.

$(1) \implies (3)$: Suppose that (1) holds. Then, by Definition 3.1, we have that $h^*(g) = \langle g, m^*\rangle - h(m^*)$, and equivalently $h(m^*) = \langle g, m^*\rangle - h^*(g)$. Clearly, $\mathcal{M}(\mathcal{X})$ is convex and $(\mathcal{M}(\mathcal{X}), \mathrm{TV})$ is Hausdorff since it is a metric space, hence we can apply the Fenchel-Moreau theorem [56, Theorem 2.3.3] to conclude that $h^{**} = h$, i.e., $h(m^*) = \sup_{g\in C_b(\mathcal{X})}\{\langle g, m^*\rangle - h^*(g)\}$. 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 [56, Theorem 2.3.3], $h^{**}(m) = h(m)$, for all $m \in \mathcal{M}(\mathcal{X})$, and hence $h^*(g^*) = \langle g^*, m^*\rangle - h(m^*)$. Hence, by Definition 3.1, the supremum of $m \mapsto \langle g^*, m\rangle - h(m)$ is realized at $m = m^*$. □

Corollary G.2. *Let $h^* : C_b(\mathcal{X}) \to \mathbb{R}$ be the convex conjugate of h. If Assumption 2.1 holds and h^* admits the first variation $\frac{\delta h^*}{\delta f}(f)$ (cf.* (53)*) on $C_b(\mathcal{X})$, then*

$$\frac{\delta h^*}{\delta f}(f) = \operatorname*{argmax}_{m \in \mathcal{M}(\mathcal{X})} \{\langle f, m\rangle - h(m)\}. \tag{54}$$

Remark G.3 (Bregman divergence via first variation)**.** *Definition 3.2 can be relaxed as follows. Provided that h^* admits a first variation (see Examples F.18 and F.19), Corollary G.2 shows that if Assumption 2.1 holds, then the first variation $\frac{\delta h^*}{\delta f}(f)$ of h^* at f is the unique maximizer of $m \mapsto \langle f, m\rangle - h(m)$. Consequently, from Definition F.17, since $f, f' \in C_b(\mathcal{X})$ and $\frac{\delta h^*}{\delta f}(f) \in \mathcal{M}(\mathcal{X})$, it follows that $\nabla_{\mathcal{F}} h^*(f)[f' - f] = \langle f' - f, \frac{\delta h^*}{\delta f}(f)\rangle$. Moreover, because h^* is Fréchet-convex, $D_{h^*}(f', f) \geq 0$, for all $f', f \in C_b(\mathcal{X})$.*

Lemma G.4. *Let Assumption 2.1 hold. Let $h^* : C_b(\mathcal{X}) \to \mathbb{R}$ be the convex conjugate of h. Fix $f, g \in C_b(\mathcal{X})$ 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 $z \in \mathcal{X}$, μ-a.e. and μ'-a.e., respectively, then*

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

Proof. By Definition 3.2, we have that

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

where the second and third equalities follow from Lemma G.1 and Corollary G.2, while the last equality follows from the definition of the Bregman divergence. □

Lemma G.5. *Consider Algorithms 1 and 2. Let Assumption 2.1 hold. Let $h^* : C_b(\mathcal{X}) \to \mathbb{R}$ be the convex conjugate of h. For each $n \geq 0$, fix $f^n, g^n \in C_b(\mathcal{X})$, $\nu^n \in \mathcal{C}$ 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 2.1, the pairs (ν^{n+1}, μ^{n+1}) in (1) and (2) are unique. We will only present the proof for (1) since the argument for (2) is identical. The updates in (1) can be equivalently written as

$$\begin{aligned}
\nu^{n+1} &= \operatorname*{argmin}_{\nu \in \mathcal{C}} \left\{ \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu, \nu^n) \right\} \\
&= \operatorname*{argmin}_{\nu \in \mathcal{C}} \left\{ \int_{\mathcal{X}} \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)(\nu - \nu^n)(\mathrm{d}x) + h(\nu) - h(\nu^n) - \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu^n, x)(\nu - \nu^n)(\mathrm{d}x) \right\} \\
&= \operatorname*{argmin}_{\nu \in \mathcal{C}} \left\{ \int_{\mathcal{X}} \left(\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) - \frac{\delta h}{\delta \nu}(\nu^n, x)\right) (\nu - \nu^n)(\mathrm{d}x) + h(\nu) \right\} \qquad (55) \\
&= \operatorname*{argmax}_{\nu \in \mathcal{C}} \left\{ \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu^n, x) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)\right) (\nu - \nu^n)(\mathrm{d}x) - h(\nu) \right\} \\
&= \operatorname*{argmax}_{\nu \in \mathcal{C}} \left\{ \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \nu}(\nu^n, x) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x)\right) \nu(\mathrm{d}x) - h(\nu) \right\},
\end{aligned}$$

and

$$\begin{aligned}
\mu^{n+1} &= \operatorname*{argmax}_{\mu\in\mathcal{D}} \left\{ \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu, \mu^n) \right\} \\
&= \operatorname*{argmax}_{\mu\in\mathcal{D}} \left\{ \int_{\mathcal{X}} \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - h(\mu) + h(\mu^n) + \int_{\mathcal{X}} \frac{\delta h}{\delta \mu}(\mu^n, y)(\mu - \mu^n)(\mathrm{d}y) \right\} \\
&= \operatorname*{argmax}_{\mu\in\mathcal{D}} \left\{ \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \mu}(\mu^n, y) + \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y) \right) (\mu - \mu^n)(\mathrm{d}y) - h(\mu) \right\} \\
&= \operatorname*{argmax}_{\mu\in\mathcal{D}} \left\{ \int_{\mathcal{X}} \left(\frac{\delta h}{\delta \mu}(\mu^n, y) + \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y) \right) \mu(\mathrm{d}y) - h(\mu) \right\}.
\end{aligned} \tag{56}$$

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) in Proposition B.4 can be equivalently written as

$$f^{n+1}(x) - f^n(x) = -\tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x), \tag{57}$$

$$g^{n+1}(y) - g^n(y) = \tau \frac{\delta F}{\delta \mu}(\nu^n, \mu^n, y), \tag{58}$$

for all $(x, y) \in \mathcal{X} \times \mathcal{X}$, ν^{n+1}-a.e. and μ^{n+1}-a.e., respectively. Then, using (54), (55) becomes

$$\begin{aligned}
\nu^{n+1} &= \operatorname*{argmax}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} \left(f^n(x) - \tau \frac{\delta F}{\delta \nu}(\nu^n, \mu^n, x) \right) \nu(\mathrm{d}x) - h(\nu) \right\} \\
&= \operatorname*{argmax}_{\nu\in\mathcal{C}} \left\{ \int_{\mathcal{X}} f^{n+1}(x)\nu(\mathrm{d}x) - h(\nu) \right\} = \frac{\delta h^*}{\delta f}(f^{n+1}),
\end{aligned} \tag{59}$$

for all $n \geq 0$. Similarly, from (56), we have that

$$\mu^{n+1} = \frac{\delta h^*}{\delta f}(g^{n+1}), \tag{60}$$

for all $n \geq 0$. The conclusion follows directly from Lemma G.4. □

H Convergence of the continuous-time dynamics and the MDA implicit algorithm

In this section, we provide a formal calculation showing that the continuous-time gradient flow obtained by taking the limit $\tau \to 0$ in the dual iterative MDA schemes of Proposition B.4 converges at rate $\mathcal{O}(1/t)$ in NI for the time-averaged flows.

Moreover, we show that an implicit Euler discretization of this gradient flow achieves a linear convergence rate $\mathcal{O}(1/N)$, matching the continuous-time rate under the same convexity–concavity assumptions on F. However, this implicit scheme is not practically implementable, unlike the explicit Algorithms 1 and 2.

Formally letting $\tau \to 0$ in the updates of Proposition B.4 yields the continuous-time flow

$$\partial_t \frac{\delta h}{\delta \nu}(\nu_t, x) = -\frac{\delta F}{\delta \nu}(\nu_t, \mu_t, x), \quad \partial_t \frac{\delta h}{\delta \mu}(\mu_t, y) = \frac{\delta F}{\delta \mu}(\nu_t, \mu_t, y), \quad t > 0, \tag{61}$$

with initial condition $(\nu_0, \mu_0) \in \mathcal{C} \times \mathcal{D}$. For convenience, we assume this flow is well-posed, i.e., it admits a unique solution $(\nu_t, \mu_t)_{t\geq 0}$.

For any $(\nu, \mu) \in \mathcal{C} \times \mathcal{D}$, and assuming the interchange of derivatives and integrals is valid, a direct calculation gives

$$\begin{aligned}
\partial_t D_h(\nu, \nu_t) &= \partial_t \left(h(\nu) - h(\nu_t) - \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu_t, x)(\nu - \nu_t)(\mathrm{d}x) \right) \\
&= -\partial_t h(\nu_t) - \partial_t \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu_t, x)(\nu - \nu_t)(\mathrm{d}x) \\
&= -\int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu_t, x)\partial_t \nu_t(\mathrm{d}x) - \int_{\mathcal{X}} \partial_t \frac{\delta h}{\delta \nu}(\nu_t, x)(\nu - \nu_t)(\mathrm{d}x) - \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu_t, x)\partial_t(\nu - \nu_t)(\mathrm{d}x) \\
&= -\int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu_t, x)\partial_t \nu_t(\mathrm{d}x) - \int_{\mathcal{X}} \partial_t \frac{\delta h}{\delta \nu}(\nu_t, x)(\nu - \nu_t)(\mathrm{d}x) + \int_{\mathcal{X}} \frac{\delta h}{\delta \nu}(\nu_t, x)\partial_t \nu_t(\mathrm{d}x) \\
&= \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu_t, \mu_t, x)(\nu - \nu_t)(\mathrm{d}x).
\end{aligned}$$

Following the same calculation for $D_h(\mu, \mu_t)$ we obtain

$$\partial_t D_h(\mu, \mu_t) = -\int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu_t, \mu_t, y)(\mu - \mu_t)(\mathrm{d}y).$$

Adding these and applying the convexity–concavity of F (Assumption 2.2) yields

$$\partial_t \left(D_h(\nu, \nu_t) + D_h(\mu, \mu_t) \right) \leq F(\nu, \mu_t) - F(\nu_t, \mu_t) + F(\nu_t, \mu_t) - F(\nu_t, \mu).$$

Integrating, dividing by t and applying Jensen's inequality to F gives

$$F\left(\frac{1}{t}\int_0^t \nu_s \mathrm{d}s, \mu \right) - F\left(\nu, \frac{1}{t}\int_0^t \mu_s \mathrm{d}s \right) \leq \frac{1}{t}\left(\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu_0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu_0) \right).$$

Hence, maximizing over (ν, μ), we conclude that

$$\mathrm{NI}\left(\frac{1}{t}\int_0^t \nu_s \mathrm{d}s, \frac{1}{t}\int_0^t \mu_s \mathrm{d}s \right) \leq \frac{1}{t}\left(\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu_0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu_0) \right),$$

establishing the $\mathcal{O}(1/t)$ rate.

We now turn to the implicit MDA scheme. 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 algorithm is defined by

Algorithm 7: IMPLICIT MDA

Input: Objective function F, initial measures (ν_0, μ_0), stepsize $\tau > 0$

for $n = 0, 1, \ldots, N-1$ **do**

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

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

end

Output: $\left(\frac{1}{N}\sum_{n=0}^{N-1} \nu^{n+1}, \frac{1}{N}\sum_{n=0}^{N-1} \mu^n \right)$

Theorem H.1 (Convergence of the implicit MDA Algorithm 7)**.** *Let (ν^0, μ^0) be such that $\sup_{\nu \in \mathcal{C}} D_h(\nu, \nu^0) + \sup_{\mu \in \mathcal{D}} D_h(\mu, \mu^0) < \infty$. Let Assumption 2.1, 2.2 and 3.3 hold. Suppose that $\tau L \leq 1$, where $L \coloneqq \max\{L_\nu, L_\mu\}$. Then, we have*

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

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 2.4 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,

$$-\tau \int \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu - \nu^n)(\mathrm{d}x) - D_h(\nu, \nu^n) \leq -\tau \int \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) - D_h(\nu^{n+1}, \nu^n) - D_h(\nu, \nu^{n+1}). \quad (62)$$

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 2.4 with $\bar{\nu} = \mu^{n+1}$ and $\mu = \mu^n$ implies that, for any $\mu \in \mathcal{D}$, we have

$$\tau \int \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu - \mu^n)(\mathrm{d}y) - D_h(\mu, \mu^n) \leq \tau \int \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - D_h(\mu^{n+1}, \mu^n) - D_h(\mu, \mu^{n+1}). \quad (63)$$

Using the convexity of $\nu \mapsto F(\nu, \mu)$ in (62), with $\nu = \nu^n$ and $\mu = \mu^{n+1}$, we have that

$$F(\nu^n, \mu^{n+1}) - F(\nu, \mu^{n+1}) - \frac{1}{\tau} D_h(\nu, \nu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^n - \nu^{n+1})(\mathrm{d}x) - \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}). \quad (64)$$

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

$$\begin{aligned} F(\nu^{n+1}, \mu^{n+1}) &\leq F(\nu^n, \mu^{n+1}) + \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) + L_\nu D_h(\nu^{n+1}, \nu^n) \\ &\leq F(\nu^n, \mu^{n+1}) + \int_{\mathcal{X}} \frac{\delta F}{\delta \nu}(\nu^n, \mu^{n+1}, x)(\nu^{n+1} - \nu^n)(\mathrm{d}x) + \frac{1}{\tau} D_h(\nu^{n+1}, \nu^n). \quad (65) \end{aligned}$$

Hence, combining (64) with (65), we obtain that

$$F(\nu^n, \mu^{n+1}) - F(\nu, \mu^{n+1}) - \frac{1}{\tau} D_h(\nu, \nu^n) \leq F(\nu^n, \mu^{n+1}) - F(\nu^{n+1}, \mu^{n+1}) - \frac{1}{\tau} D_h(\nu, \nu^{n+1}). \quad (66)$$

Similarly, using concavity of $\mu \mapsto F(\nu, \mu)$ in (63), with $\nu = \nu^{n+1}$ and $\mu = \mu^n$, we have that

$$F(\nu^{n+1}, \mu) - F(\nu^{n+1}, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^n) \leq \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu^{n+1}, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^{n+1}). \quad (67)$$

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

$$\begin{aligned} F(\nu^{n+1}, \mu^{n+1}) &\geq F(\nu^{n+1}, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - L_\mu D_h(\mu^{n+1}, \mu^n) \\ &\geq F(\nu^{n+1}, \mu^n) + \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\nu^{n+1}, \mu^n, y)(\mu^{n+1} - \mu^n)(\mathrm{d}y) - \frac{1}{\tau} D_h(\mu^{n+1}, \mu^n). \quad (68) \end{aligned}$$

Hence, combining (67) with (68), we obtain that

$$F(\nu^{n+1}, \mu) - F(\nu^{n+1}, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^n) \leq F(\nu^{n+1}, \mu^{n+1}) - F(\nu^{n+1}, \mu^n) - \frac{1}{\tau} D_h(\mu, \mu^{n+1}). \quad (69)$$

Adding inequalities (66) and (69) implies that

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

Summing the previous inequality over $n = 0, 1, ..., N-1$, bounding the right-hand side from above by its supremum over (ν, μ), dividing by N, applying Jensen's inequality and taking maximum over (ν, μ) in the left-hand side leads to

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

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

I Further related works

Comparison with Wasserstein MD and other mean-field min-max methods. Our work is also related to recent developments on optimization over spaces of probability measures. A first point of comparison is [10], who study mirror descent on $\mathcal{P}_2(\mathbb{R}^d)$ using the Wasserstein geometry. By contrast, we work with the flat geometry and study min-max problems over $\mathcal{P}(\mathbb{R}^d)$. These two geometries are complementary and are suited to different goals.

In the flat geometry, convexity-concavity and relative smoothness are formulated along linear interpolations of measures $[0,1] \ni \varepsilon \mapsto (1-\varepsilon)\mu_0 + \varepsilon\mu_1$. This formulation is aligned with our MDA analysis, since it interacts naturally with the duality pairing between measures and functions. It also leads to assumptions that are relatively easy to verify in several motivating examples, including GANs and RLHF. In the Wasserstein geometry, by contrast, convexity and smoothness are typically formulated along transport curves, in particular along geodesics of the form $[0,1] \ni \varepsilon \mapsto \big((1-\varepsilon)\,\mathrm{Id} + \varepsilon\,\mathrm{T}_{\mu_0}^{\mu_1}\big)_{\#}\mu_0$, where $\mathrm{T}_{\mu_0}^{\mu_1}$ denotes an optimal transport map from μ_0 to μ_1. This structure is natural for transport-based algorithms, but it imposes substantially different requirements. Such maps need not exist in general. For example, a standard sufficient condition is that μ_0 is absolutely continuous with respect to Lebesgue measure. Moreover, for bilinear objective, i.e., $F(\nu,\mu) = \iint f(x,y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$, Wasserstein geodesic convexity-concavity would impose Euclidean convexity-concavity conditions on f. This would exclude several of our main motivating examples, such as GANs, RLHF, and adversarial training of mean-field neural networks, unless additional restrictive assumptions were imposed.

More generally, Wasserstein regularity is stronger than flat regularity. The Wasserstein gradient is expressed in terms of $\nabla\frac{\delta F}{\delta\mu}$, and therefore requires differentiability of the flat derivative itself. This is often delicate and can be difficult to verify in applications such as RLHF. Furthermore, certain important functionals, including relative entropy, are not Wasserstein differentiable in general and require subgradient formulations, as discussed in [10]. For these reasons, the flat geometry is better suited to the assumptions used in our MDA analysis. Relatedly, Remark 2.7 explains why replacing the TV^2 strong convexity assumption on the mirror potential h by a W_2^2 condition would substantially reduce the generality of our results and would require stronger assumptions, such as Talagrand inequalities along the iterates.

The limitation of the flat geometry is that it does not include the transport structure intrinsic to Wasserstein space. In the Wasserstein setting, an update may be represented by a transport map T^{n+1} and written as $\mu^{n+1} = (\mathrm{T}^{n+1})_{\#}\mu^n$, so the algorithm directly moves particles from μ^n to μ^{n+1}. This perspective is particularly natural for particle-based algorithms and for objectives involving Sinkhorn divergences, sliced-Wasserstein distances, or interaction energies, as in [10]. In contrast, the flat derivative $\frac{\delta F}{\delta\mu}$ describes infinitesimal signed perturbations of the measure. This is the appropriate object for our analysis, but it does not capture the geodesic transport structure exploited by Wasserstein methods.

Another related contribution is [13], which studies an entropy-regularized bilinear game over distributions using min-max Langevin dynamics and finite-particle approximations. Their setting assumes a four-times differentiable, strongly convex-strongly concave payoff with Lipschitz gradient. In contrast, we study simultaneous and alternating MDA schemes for unregularized convex-concave min-max problems on spaces of measures, under relative smoothness and mirror geometry. In particular, in the bilinear case we do not assume convexity-concavity of f, nor do we require f to be differentiable or to have Lipschitz gradient. A sufficient condition for Assumption 2.3 is simply boundedness of f. As discussed in Remark 3.6, relative smoothness is automatic for bilinear games in our framework, without additional regularity assumptions on f. Thus, the regularity requirements in our setting are substantially weaker.

The guarantees are also different in nature. Our results concern simultaneous and alternating MDA, with performance measured by the Nikaido–Isoda error. A main contribution of our paper is to show that, in our measure space setting, alternating MDA improves the rate from $\mathcal{O}(N^{-1/2})$ for simultaneous MDA to $\mathcal{O}(N^{-2/3})$. In this sense, the results of [13] neither imply nor are implied by ours since they address a different algorithmic and analytical regime.

Finally, [31] study an entropy-regularized distributional minimax problem and analyze Wasserstein descent-ascent dynamics with historical averaging. Their setting differs from ours in several respects.

First, they consider a regularized Wasserstein dynamics, whereas we study unregularized convex-concave games on spaces of measures under mirror geometry. Second, their Theorem 3.4 gives continuous-time rates, that is, the unweighted averaging scheme achieves an $\mathcal{O}(\log T/T)$ rate, while the weighted averaging scheme achieves an $\mathcal{O}(1/T)$ rate. These continuous-time guarantees are not directly comparable to our discrete-time MDA rates, especially because their method is a history-averaged Wasserstein dynamics rather than an MDA scheme. As discussed further in Appendix H, continuous-time rates should be interpreted as idealized limits and do not automatically translate into achievable rates for implementable discrete-time algorithms.

Moreover, the discrete-time result of [31], stated in their Theorem 3.7 is a propagation-of-chaos result controlling time and particle approximations of the continuous-time dynamics. Our paper does not address propagation-of-chaos guarantees. Instead, our contribution is to establish convergence rates for explicit simultaneous and alternating MDA schemes on measure spaces, and to prove that alternating updates yield a faster rate than simultaneous updates in this framework. Thus, while [31] provides important results for regularized Wasserstein dynamics, it does not directly imply that the rates obtained here are suboptimal.

Comparison with infinite-dimensional Mirror Prox. Besides the MDA algorithm, [25] considers the entropic Mirror Prox algorithm, which is computationally more expensive than simultaneous MDA since it requires two mirror descent steps per player per iteration, effectively doubling the number of gradient evaluations. Although [25] proves it achieves a $\mathcal{O}(N^{-1})$ convergence rate with deterministic gradients, in the stochastic case it attains only the same $\mathcal{O}(N^{-1/2})$ rate as simultaneous MDA. Since the stochastic setting is precisely the one arising in the particle-based algorithmic implementation, even if the theoretical infinite-dimensional Mirror Prox analysis yielded a $\mathcal{O}(N^{-1})$ rate, the corresponding particle system would still converge at best at $\mathcal{O}(N^{-1/2})$.

Another approach based on reproducing kernel Hilbert spaces (RKHS) is developed in [22] 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. Finally, we note that the convergence analysis of an alternating Mirror Prox scheme remains, to our knowledge, open. Designing how to alternate the four mirror descent steps in a way that improves convergence rates appears to be technically delicate. We view this as a potential direction for future work.

Extending our analysis to the Mirror Prox scheme is not straightforward due to the following technical issue. To replicate the assumptions on the objective F used in [12, Section 5.2.3] or [22, Assumption 4.2], one would need to impose the following TV-Lipschitz condition on the flat derivatives of F. Specifically, there exist $C_{11}, C_{12}, C_{21}, C_{22} > 0$ such that

$$\left|\frac{\delta F}{\delta \nu}(\nu', \mu, x') - \frac{\delta F}{\delta \nu}(\nu, \mu, x)\right| \leq C_{11}(\mathrm{TV}(\nu', \nu) + |x' - x|),$$

$$\left|\frac{\delta F}{\delta \nu}(\nu, \mu', x') - \frac{\delta F}{\delta \nu}(\nu, \mu, x)\right| \leq C_{12}(\mathrm{TV}(\mu', \mu) + |x' - x|),$$

$$\left|\frac{\delta F}{\delta \mu}(\nu', \mu, y') - \frac{\delta F}{\delta \mu}(\nu, \mu, y)\right| \leq C_{21}(\mathrm{TV}(\nu', \nu) + |y' - y|).$$

$$\left|\frac{\delta F}{\delta \mu}(\nu, \mu', y') - \frac{\delta F}{\delta \mu}(\nu, \mu, y)\right| \leq C_{22}(\mathrm{TV}(\mu', \mu) + |y' - y|).$$

This TV-Lipschitz assumption is substantially stronger than the relative smoothness condition (Assumption 3.3), which is needed only for the alternating scheme but not for the simultaneous one. For instance, the bilinear game

$$F(\nu, \mu) = \int_{\mathcal{X}} \int_{\mathcal{X}} f(x, y)\nu(\mathrm{d}x)\mu(\mathrm{d}y)$$

satisfies Assumption 3.3 for $L_\nu = L_\mu = 0$, yet it satisfies the TV-Lipschitz conditions above only when f is both bounded and Lipschitz. Thus, showing that Mirror Prox achieves an $\mathcal{O}(N^{-1})$ convergence rate under the weaker relative smoothness assumption remains a challenging open question.

Comparison with Euclidean alternating updates. A closely related line of work studies alternating methods for finite-dimensional bilinear zero-sum games. In particular, [28] analyze alternating mirror descent (AMD) from a symplectic/Hamiltonian perspective. Their $\mathcal{O}(N^{-4/5})$ bound on the average-iterate duality gap is obtained by viewing AMD as a symplectic Euler discretization, constructing a modified Hamiltonian, and exploiting high-order smoothness properties of the convex conjugate h^* associated with the Bregman potential.

The techniques in [28], however, appear to rely in an essential way on finite-dimensional Euclidean structure. The symplectic Euler interpretation uses a finite-dimensional representation in which AMD is related to a symplectic scheme through an invertible linear transformation. Moreover, the modified Hamiltonian is constructed using Poisson brackets, which are defined for sufficiently smooth functions on Euclidean spaces. Their argument also involves an infinite-series representation of the modified Hamiltonian and requires delicate control of this series, established only in certain regimes. Consequently, extending this approach to our setting would require infinite-dimensional analogues of several nontrivial objects, including a suitable notion of symplectic integrator on spaces of measures, an associated Hamiltonian formalism, and an appropriate version of Poisson brackets. Developing such a framework is an interesting direction, but it is not a direct consequence of the methods used here. Thus, the sharper rate in [28] exploits finite-dimensional bilinear structure that is not available in our general mean-field convex-concave setting.

Another related result is due to [40], who study alternating projected gradient descent-ascent (Alt-GDA) for bilinear games over compact convex sets. They prove a global $\mathcal{O}(1/T)$ rate under the assumption that an interior Nash equilibrium exists, and a local $\mathcal{O}(1/T)$ rate near a maximum-support equilibrium when this assumption is removed. This result is specific to Euclidean projected GDA rather than mirror descent. In particular, it corresponds to the quadratic Bregman potential $\frac{1}{2}\|\cdot\|^2$ and does not immediately extend to general Bregman geometries, even in finite dimensions. In our setting, the updates are defined on spaces of measures, where there is no canonical Euclidean gradient step and the analysis must instead be formulated through the duality pairing between measures and functions. Hence the connection between Euclidean projected GDA and mirror descent-type dynamics on measure spaces is far from immediate.

Furthermore, the analysis of [40] uses structural properties of finite-dimensional simplices. Concepts such as interior Nash equilibria and maximum support equilibria play a central role in both their assumptions and their proofs, and their admissible stepsize condition also depends on this finite-dimensional simplex structure. Therefore, as in the comparison with [28], the sharper guarantees in [40] are obtained in a substantially more specialized setting using tools that do not transfer to infinite-dimensional mean-field games over measures.

Overall, the results of [28] and [40] show that sharper convergence rates for alternating methods are possible in special finite-dimensional bilinear regimes. Our setting is different in that it concerns general convex-concave games over spaces of measures. Extending such improved rates to this level of generality would require new geometric and analytic tools, and remains a highly nontrivial open direction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Sections 2, 3 and A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Section 2, 3, A and the rest of the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details of the algorithm and the numerical experiments are provided in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The experiments conducted are limited to toy simulations.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are provided in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experiments are toy illustrative examples only meant to demonstrate the qualitative behavior of the proposed approach, so I do not report error bars, confidence intervals, or statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments were run on an Apple M4 laptop, with each experiment taking on the order of tens of minutes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have verified that the research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is primarily theoretical, and no immediate societal impact is anticipated.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper is primarily theoretical.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.