Last Iterate Convergence in No-regret Learning: Constrained Min-max Optimization for Convex-concave Landscapes

Qi Lei
EE, Princeton

Sai Ganesh Nagarajan
ESD, SUTD

Ioannis Panageas
CS, UC Irvine

Xiao Wang
SIME, SUFE

Abstract

In a recent series of papers it has been established that variants of Gradient Descent/Ascent and Mirror Descent exhibit last iterate convergence in convex-concave zero-sum games. Specifically, Daskalakis et al. (2018); Liang and Stokes (2018) show last iterate convergence of the so called "Optimistic Gradient Descent/Ascent" for the case of *unconstrained* min-max optimization. Moreover, in Mertikopoulos et al. (2018) the authors show that Mirror Descent with an extra gradient step displays last iterate convergence for convex-concave problems (both constrained and unconstrained), though their algorithm uses *vanishing stepsizes*. In this work, we show that "Optimistic Multiplicative-Weights Update (OMWU)" with *constant stepsize*, exhibits last iterate convergence locally for convex-concave games, generalizing the results of Daskalakis and Panageas (2019) where last iterate convergence of OMWU was shown only for the *bilinear case*. To the best of our knowledge, this is the first result about last-iterate convergence for constrained zero sum games (beyond the bilinear case) in which the dynamics use constant step-sizes.

1 Introduction

In classic (normal form) zero-sum games, one has to compute two probability vectors $\mathbf{x}^* \in \Delta_n, \mathbf{y}^* \in \Delta_m$ [1] that consist an equilibrium of the following problem

$$\min_{\mathbf{x}\in\Delta_n} \max_{\mathbf{y}\in\Delta_m} \mathbf{x}^\top A\mathbf{y}, \tag{1}$$

where A is $n \times m$ real matrix (called payoff matrix). Here $\mathbf{x}^\top A\mathbf{y}$ represents the payment of the $\mathbf{x}$ player to the $\mathbf{y}$ player under choices of strategies by the two players and is a *bilinear* function.

Arguably, one of the most celebrated theorems and a founding stone in Game Theory, is the minimax theorem by Von Neumann Von Neumann (1928). It states

$$\min_{\mathbf{x}\in\Delta_n} \max_{\mathbf{y}\in\Delta_m} f(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\Delta_m} \min_{\mathbf{x}\in\Delta_n} f(\mathbf{x},\mathbf{y}), \tag{2}$$

where $f : \Delta_n \times \Delta_m \to \mathbb{R}$ is convex in $\mathbf{x}$, concave in $\mathbf{y}$. The aforementioned result holds for any convex compact sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$. The min-max theorem reassures us that an equilibrium always exists in the bilinear game (1) or its convex-concave analogue (again $f(\mathbf{x},\mathbf{y})$ is interpreted as the payment of the $\mathbf{x}$ player to the $\mathbf{y}$ player). An equilibrium is a pair of randomized strategies $(\mathbf{x}^*, \mathbf{y}^*)$ such that neither player can improve their payoff by unilaterally changing their distribution.

Soon after the appearance of the minimax theorem, research was focused on whether minmax solutions of (1) can be reached when agents adopt *simple and natural* update rules aligned with their selfish (and possibly myopic behavior). An early method, proposed by Brown Brown (1951) and analyzed by Robinson Robinson (1951), was *fictitious play*. Later on, researchers discover several learning robust algorithms converging to minimax equilibrium at faster rates, see Cesa-Bianchi and Lugosi (2006). This class of learning algorithms, are the so-called "no-regret" and include Multiplicative Weights Update (MWU) method Arora et al. (2012) and Follow the regularized leader (FTRL) Abernethy et al. (2008). Formally, in the online learning framework, at time t, each player chooses a probability distribution ($\mathbf{x}^t, \mathbf{y}^t$ respectively) simultaneously depending *only* on the past choices of both players (i.e., $\mathbf{x}^1, ..., \mathbf{x}^{t-1}, \mathbf{y}^1, ..., \mathbf{y}^{t-1}$) and experiences payoff that depends on choices $\mathbf{x}^t, \mathbf{y}^t$.

[1] Δ_n denotes the simplex of size n.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1.1 Average Iterate Convergence vs Last Iterate

Despite the rich literature on no-regret learning, most of the known results have the feature that min-max equilibrium is shown to be attained only by the time *average*. This means that the trajectory of a no-regret learning method $(\mathbf{x}^t, \mathbf{y}^t)$ has the property that $\frac{1}{t}\sum_{\tau \leq t} \mathbf{x}^{\tau\top} A \mathbf{y}^\tau$ converges to the equilibrium of (1), as $t \to \infty$. Unfortunately that does not mean that the last iterate $(\mathbf{x}^t, \mathbf{y}^t)$ converges to an equilibrium, it commonly diverges or cycles. One such example is the well-known Multiplicative Weights Update Algorithm, the time average of which is known to converge to an equilibrium, but the actual trajectory cycles towards the boundary of the simplex (Bailey and Piliouras (2018)). This is even true for the vanilla Gradient Descent/Ascent, where one can show for even bilinear landscapes (unconstrained case) last iterate fails to converge Daskalakis et al. (2018). It is important to note that for most no-regret dynamics, the chosen step-size is vanishing with time and it is not known whether time-average convergence results persist once the agents use *constant step-sizes* that are independent of the dynamics' time-steps. The reason is that once constant step-size are used, the no-regret property collapses.

Motivated by the training of Generative Adversarial Networks (GANs), the last couple of years researchers have focused on designing and analyzing procedures that exhibit *last iterate* convergence (or pointwise convergence) for zero-sum games. This is crucial for training GANs, the landscapes of which are typically non-convex non-concave and averaging now as before does not give much guarantees (e.g., note that Jensen's inequality is not applicable anymore). In Daskalakis et al. (2018); Liang and Stokes (2018) the authors show that a variant of Gradient Descent/Ascent, called Optimistic Gradient Descent/Ascent has last iterate convergence for the case of bilinear functions $\mathbf{x}^\top A \mathbf{y}$ where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ (this is called the unconstrained case, since there are no restrictions on the vectors). Later on, Daskalakis and Panageas (2019) generalized the above result with simplex constraints, where the online method that the authors analyzed was Optimistic Multiplicative Weights Update. In Mertikopoulos et al. (2018), it is shown that Mirror Descent with extra gradient computation converges pointwise for a class of zero-sum games that includes the convex-concave setting (with arbitrary constraints), though their algorithm does not fit in the online no-regret framework since it uses information twice about the payoffs before it iterates and moreover the chosen step-size is vanishing with time. This was later generalized with one call of the gradient in Hsieh et al. (2019), again using vanishing step-sizes. Last but not least there have appeared other works that show pointwise convergence for other settings (see Palaiopanos et al. (2017); Daskalakis and Panageas (2018) and Abernethy et al. (2019) and references therein) to stationary points (but not local equilibrium solutions).

1.2 Main Results

In this paper, we focus on the constrained min-max optimization problem

$$\min_{\mathbf{x}\in\Delta_n} \max_{\mathbf{y}\in\Delta_m} f(\mathbf{x}, \mathbf{y}), \tag{3}$$

where f is a convex-concave function (convex in $\mathbf{x}$, concave in $\mathbf{y}$). We analyze the no-regret online algorithm Optimistic Multiplicative Weights Update (OMWU). OMWU is an instantiation of the Optimistic Follow the Regularized Leader (OFTRL) method with entropy as a regularizer (for both players, see Preliminaries section for the definition of OMWU).

We prove that OMWU exhibits local last iterate convergence with *constant step-size*, generalizing the result of Daskalakis and Panageas (2019) and proving an open question of Syrgkanis et al. (2015) (for convex-concave games). Formally, our main theorem is stated below:

Theorem 1.1 (Last iterate convergence of OMWU)**.** *Let $f : \Delta_n \times \Delta_m \to \mathbb{R}$ be a twice differentiable function $f(\mathbf{x}, \mathbf{y})$ that is convex in $\mathbf{x}$ and concave in $\mathbf{y}$. Assume that there exists an equilibrium $(\mathbf{x}^*, \mathbf{y}^*)$ that satisfies the KKT conditions with strict inequalities (see (4)). It holds that for sufficiently small constant step-size, there exists a neighborhood $U \subseteq \Delta_n \times \Delta_m$ of $(\mathbf{x}^*, \mathbf{y}^*)$ such that for all for all initial conditions $(\mathbf{x}^0, \mathbf{y}^0), (\mathbf{x}^1, \mathbf{y}^1) \in U$, OMWU exhibits last iterate (pointwise) convergence, i.e.,*

$$\lim_{t\to\infty} (\mathbf{x}^t, \mathbf{y}^t) = (\mathbf{x}^*, \mathbf{y}^*),$$

where $(\mathbf{x}^t, \mathbf{y}^t)$ denotes the t-th iterate of OMWU.

Moreover, we provide experiments that indicate that our results should hold globally (we conjecture global convergence).

1.3 Technical Overview

In this subsection, we present a brief technical overview. The main result of the paper boils down to perform stability analysis of OMWU dynamics near the min-max equilibrium. The stability analysis, the understanding of the local behavior and the local convergence guarantees of OMWU rely on the spectral analysis of the computed Jacobian matrix. One of the

key challenges is understanding the equilibrium min-max solution in the boundary of the simplex. The main reason is that the min-max solution $(\mathbf{x}^*, \mathbf{y}^*)$ *does not necessarily satisfy* $\mathbf{\nabla f(x^*, y^*) = 0}$ (gradient zero). Especially, a min-max solution $(\mathbf{x}^*, \mathbf{y}^*)$ that lies on the boundary (some coordinates of $(\mathbf{x}^*, \mathbf{y}^*)$ are zero) of the simplex most likely will not have $\nabla f(\mathbf{x}^*, \mathbf{y}^*) = 0$. So, works such as Malitsky and Tam (2020), that try to understand convergence to first order stationary points alone is not applicable. Furthermore, the techniques for bilinear games (as in Daskalakis and Panageas (2019)) are no longer valid in convex-concave games. Allow us to explain the differences from Daskalakis and Panageas (2019). In general, one cannot expect a trivial generalization from linear to non-linear scenarios. The properties of bilinear games are fundamentally different from that of convex-concave games, and this makes the analysis much more challenging in the latter. The key result of spectral analysis in Daskalakis and Panageas (2019) is in a lemma (Lemma B.6) which states that a skew symmetric[2] has imaginary eigenvalues. Skew symmetric matrices appear since in bilinear cases there are terms that are linear in $\mathbf{x}$ and linear in $\mathbf{y}$ but no higher order terms in $\mathbf{x}$ or $\mathbf{y}$. However, the skew symmetry has no place in the case of convex-concave landscapes and the Jacobian matrix of OMWU is far more complicated. One key technique to overcome the lack of skew symmetry is the use of Ky Fan inequality Moslehian (2011) which states that the sequence of the eigenvalues of $\frac{1}{2}(W + W^\top)$ majorizes the real part of the sequence of the eigenvalues of W for any square matrix W (see Lemma 3.1).

Notation The boldface $\mathbf{x}$ and $\mathbf{y}$ denote the vectors in Δ_n and Δ_m. $\mathbf{x}^t$ denotes the t-th iterate of the dynamical system. The letter J denote the Jacobian matrix. $\mathbf{I}$, $\mathbf{0}$ and $\mathbf{1}$ are preserved for the identity, zero matrix and the vector with all the entries equal to 1. The support of $\mathbf{x}$ is the set of indices of x_i such that $x_i \neq 0$, denoted by $\text{Supp}(\mathbf{x})$. $(\mathbf{x}^*, \mathbf{y}^*)$ denotes the optimal solution for minimax problem. $[n]$ denote the set of integers $\{1, ..., n\}$.

2 Preliminaries

In this section, we present some background that will be used later.

2.1 Equilibria for Constrained Minimax

From Von Neumann's minimax theorem, one can conclude that the problem $\min_{\mathbf{x}\in\Delta_n} \max_{\mathbf{y}\in\Delta} f(\mathbf{x}, \mathbf{y})$ has always an equilibrium $(\mathbf{x}^*, \mathbf{y}^*)$ with $f(\mathbf{x}^*, \mathbf{y}^*)$ be unique. Moreover from KKT conditions (as long as f is twice differentiable), such an equilibrium must satisfy the following ($\mathbf{x}^*$ is a local minimum for fixed $\mathbf{y} = \mathbf{y}^*$ and $\mathbf{y}^*$ is a local maximum for fixed $\mathbf{x} = \mathbf{x}^*$):

Definition 2.1 (KKT conditions)**.** The facts below are straightforward from KKT conditions because of non-negativity and equality constraints (linear constraints).

$$\begin{aligned}
&\mathbf{x}^* \in \Delta_n, \text{i.e., } \mathbf{x}^* \text{ is in simplex,}\\
&\text{If } x_i^* > 0 \text{ then } \tfrac{\partial f}{\partial x_i}(\mathbf{x}^*, \mathbf{y}^*) = \textstyle\sum_{j=1}^n x_j^* \tfrac{\partial f}{\partial x_j}(\mathbf{x}^*, \mathbf{y}^*),\\
&\text{If } x_i^* = 0 \text{ then } \tfrac{\partial f}{\partial x_i}(\mathbf{x}^*, \mathbf{y}^*) \geq \textstyle\sum_{j=1}^n x_j^* \tfrac{\partial f}{\partial x_j}(\mathbf{x}^*, \mathbf{y}^*).\\
&\text{Last two equations indicate best strategy for } \mathbf{x}.\\
&\\
&\mathbf{y}^* \in \Delta_m \text{i.e., } \mathbf{y}^* \text{ is in simplex,}\\
&y_i^* > 0 \text{ then } \tfrac{\partial f}{\partial y_i}(\mathbf{x}^*, \mathbf{y}^*) = \textstyle\sum_{j=1}^m y_j^* \tfrac{\partial f}{\partial y_j}(\mathbf{x}^*, \mathbf{y}^*),\\
&y_i^* = 0 \text{ then } \tfrac{\partial f}{\partial y_i}(\mathbf{x}^*, \mathbf{y}^*) \leq \textstyle\sum_{j=1}^m y_j^* \tfrac{\partial f}{\partial y_j}(\mathbf{x}^*, \mathbf{y}^*),\\
&\text{Last two equations indicate best strategy for } \mathbf{y}.
\end{aligned} \tag{4}$$

Remark 2.2 (No degeneracies). For the rest of the paper we assume the last inequalities hold strictly for both players (if $x_i^* = 0$ then $\frac{\partial f}{\partial x_i}(\mathbf{x}^*, \mathbf{y}^*) > \sum_{j=1}^n x_j^* \frac{\partial f}{\partial x_j}(\mathbf{x}^*, \mathbf{y}^*)$ and similarly for $\mathbf{y}$ player). Intuitively, this assumption means that any unilateral deviation incurs a strict loss to the player that deviated. We would like to note that this assumption is satisfied for "generic" bilinear zero-sum games Ritzberger (1994) and guarantee uniqueness of minmax solution $(\mathbf{x}^*, \mathbf{y}^*)$, so we adapt the same assumption for convex-concave landscapes. Moreover, we assume that the Hessian of $f(\mathbf{x}, \mathbf{y})$ is invertible at the min-max solution $(\mathbf{x}^*, \mathbf{y}^*)$. Let us note that these two assumptions guarantee uniqueness (no continuums) of the min-max solution for convex-concave landscapes. Finally, it is easy to see that since f is convex concave and twice differentiable, then $\nabla^2_{\mathbf{xx}} f$ (part of the Hessian that involves $\mathbf{x}$ variables) is positive semi-definite and $\nabla^2_{\mathbf{yy}} f$ (part of the Hessian that involves $\mathbf{y}$ variables) is negative semi-definite.

2.2 Optimistic Multiplicative Weights Update

The equations of Optimistic Follow-the-Regularized-Leader (OFTRL) applied to a problem $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ with regularizers (strongly convex functions) $h_1(\mathbf{x}), h_2(\mathbf{y})$ (for player $\mathbf{x}, \mathbf{y}$ respectively) and $\mathcal{X} \subset \mathbb{R}^n, \mathcal{Y} \subset \mathbb{R}^m$ is given below (see

[2] A is skew symmetric if $A^\top = -A$.

Daskalakis et al. (2018)):

$$\mathbf{x}^{t+1} = \underset{\mathbf{x}\in\mathcal{X}}{\mathbf{argmin}}\{\eta\sum_{s=1}^{t}\mathbf{x}^\top\nabla_{\mathbf{x}}f(\mathbf{x}^s,\mathbf{y}^s) + \underbrace{\eta\mathbf{x}^\top\nabla_{\mathbf{x}}f(\mathbf{x}^t,\mathbf{y}^t)}_{\text{optimistic term}} + h_1(\mathbf{x})\}$$

$$\mathbf{y}^{t+1} = \underset{\mathbf{y}\in\mathcal{Y}}{\mathbf{argmax}}\{\eta\sum_{s=1}^{t}\mathbf{y}^\top\nabla_{\mathbf{y}}f(\mathbf{x}^s,\mathbf{y}^s) + \underbrace{\eta\mathbf{y}^\top\nabla_{\mathbf{y}}f(\mathbf{x}^t,\mathbf{y}^t)}_{\text{optimistic term}} - h_2(\mathbf{y})\}.$$

η is called the *step-size* of the online algorithm. If η depends on time (iteration) and goes to zero as $t\to\infty$, then it is vanishing; moreover if η does not depend on time, then it is constant. OFTRL is uniquely defined if f is convex-concave and domains $\mathcal{X}$ and $\mathcal{Y}$ are convex. For simplex constraints and entropy regularizers, i.e., $h_1(\mathbf{x}) = \sum_i x_i\ln x_i, h_2(\mathbf{y}) = \sum_i y_i\ln y_i$, we can solve for the explicit form of OFTRL using KKT conditions, the update rule is the Optimistic Multiplicative Weights Update (OMWU) and is described as follows:

$$x_i^{t+1} = \frac{x_i^t\cdot e^{-2\eta\frac{\partial f}{\partial x_i}(\mathbf{x}^t,\mathbf{y}^t)+\eta\frac{\partial f}{\partial x_i}(\mathbf{x}^{t-1},\mathbf{y}^{t-1})}}{\sum_k x_k^t e^{-2\eta\frac{\partial f}{\partial x_k}(\mathbf{x}^t,\mathbf{y}^t)+\eta\frac{\partial f}{\partial x_k}(\mathbf{x}^{t-1},\mathbf{y}^{t-1})}},$$

$$y_j^{t+1} = \frac{y_j^t\cdot e^{2\eta\frac{\partial f}{\partial y_j}(\mathbf{x}^t,\mathbf{y}^t)-\eta\frac{\partial f}{\partial y_j}(\mathbf{x}^{t-1},\mathbf{y}^{t-1})}}{\sum_k y_k^t e^{2\eta\frac{\partial f}{\partial y_j}(\mathbf{x}^t,\mathbf{y}^t)-\eta\frac{\partial f}{\partial y_k}(\mathbf{x}^{t-1},\mathbf{y}^{t-1})}}$$

for all $i\in[n], j\in[m]$.

2.3 Fundamentals of Dynamical Systems

We conclude Preliminaries section with some basic facts from dynamical systems.

Definition 2.3. A recurrence relation of the form $\mathbf{x}^{t+1} = w(\mathbf{x}^t)$ is a discrete time dynamical system, with update rule $w:\mathcal{S}\to\mathcal{S}$ where $\mathcal{S}$ is a subset of $\mathbb{R}^k$ for some positive integer k. The point $\mathbf{z}\in\mathcal{S}$ is called a *fixed point* if $w(\mathbf{z}) = \mathbf{z}$.

Remark 2.4. Using KKT conditions (4), it is not hard to observe that an equilibrium point $(\mathbf{x}^*,\mathbf{y}^*)$ must be a fixed point of the OMWU algorithm, i.e., if $(\mathbf{x}^t,\mathbf{y}^t) = (\mathbf{x}^{t-1},\mathbf{y}^{t-1}) = (\mathbf{x}^*,\mathbf{y}^*)$ then $(\mathbf{x}^{t+1},\mathbf{y}^{t+1}) = (\mathbf{x}^*,\mathbf{y}^*)$.

Proposition 2.5 (Galor (2007))**.** *Assume that w is a differentiable function and the Jacobian of the update rule w at a fixed point $\mathbf{z}^*$ has spectral radius less than one. It holds that there exists a neighborhood U around $\mathbf{z}^*$ such that for all $\mathbf{z}^0\in U$, the dynamics $\mathbf{z}^{t+1} = w(\mathbf{z}^t)$ converges to $\mathbf{z}^*$, i.e. $\lim_{n\to\infty}w^n(\mathbf{z}^0) = \mathbf{z}^*$ [3]. w is called a contraction mapping in U.*

[3] w^n denotes the composition of w with itself n times.

Note that we will make use of Proposition 2.5 to prove our Theorem 1.1 (by proving that the Jacobian of the update rule of OMWU has spectral radius less than one).

3 Last iterate convergence of OMWU

In this section, we prove that OMWU converges pointwise (exhibits last iterate convergence) if the initializations $(\mathbf{x}^0,\mathbf{y}^0),(\mathbf{x}^1,\mathbf{y}^1)$ belong in a neighborhood U of the equilibrium $(\mathbf{x}^*,\mathbf{y}^*)$.

3.1 Dynamical System of OMWU

We first express OMWU algorithm as a dynamical system so that we can use Proposition 2.5. The idea (similar to Daskalakis and Panageas (2019)) is to lift the space to consist of four components $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}$, in such a way we can include the history (current and previous step, see Section 2.2 for the equations). First, we provide the update rule $g:\Delta_n\times\Delta_m\times\Delta_n\times\Delta_m\to\Delta_n\times\Delta_m\times\Delta_n\times\Delta_m$ of the lifted dynamical system and is given by $g(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}) = (g_1,g_2,g_3,g_4)$, where $g_i = g_i(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w})$ for $i\in[4]$ are defined as follows:

$$g_{1,i}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}) = x_i\frac{e^{-2\eta\frac{\partial f}{\partial x_i}(\mathbf{x},\mathbf{y})+\eta\frac{\partial f}{\partial z_i}(\mathbf{z},\mathbf{w})}}{\sum_k x_k e^{-2\eta\frac{\partial f}{\partial x_k}(\mathbf{x},\mathbf{y})+\eta\frac{\partial f}{\partial z_k}(\mathbf{z},\mathbf{w})}}, i\in[n]$$

$$g_{2,i}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}) = y_i\frac{e^{2\eta\frac{\partial f}{\partial y_i}(\mathbf{x},\mathbf{y})-\eta\frac{\partial f}{\partial w_i}(\mathbf{z},\mathbf{w})}}{\sum_k y_k e^{2\eta\frac{\partial f}{\partial y_k}(\mathbf{x},\mathbf{y})-\eta\frac{\partial f}{\partial w_k}(\mathbf{z},\mathbf{w})}}, i\in[m]$$

$$g_{3,i}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}) = x_i,\ i\in[n]$$

$$g_{4,i}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w}) = y_i,\ i\in[m].$$

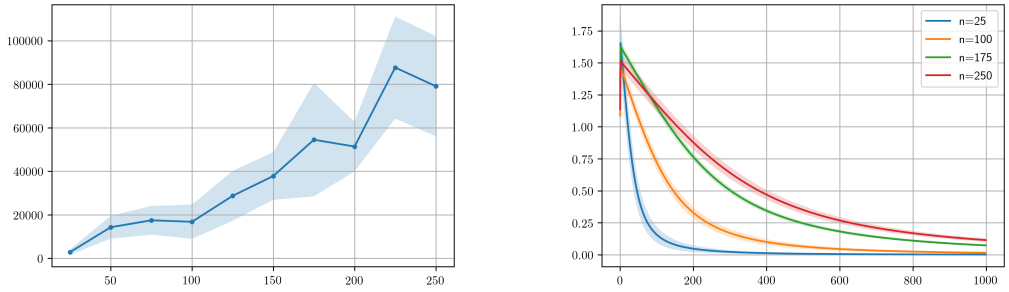
Then the dynamical system of OMWU can be written in compact form as

$$(\mathbf{x}_{t+1},\mathbf{y}_{t+1},\mathbf{x}_t,\mathbf{y}_t) = g(\mathbf{x}_t,\mathbf{y}_t,\mathbf{x}_{t-1},\mathbf{y}_{t-1}).$$

In what follows, we will perform spectral analysis on the Jacobian of the function g, computed at the fixed point $(\mathbf{x}^*,\mathbf{y}^*)$. Since g has been lifted, the fixed point we analyze is $(\mathbf{x}^*,\mathbf{y}^*,\mathbf{x}^*,\mathbf{y}^*)$ (see Remark 2.4). By showing that the spectral radius is less than one, our Theorem 1.1 follows by Proposition 2.5. The computations of the Jacobian of g are deferred to the supplementary material.

3.2 Spectral Analysis

Let $(\mathbf{x}^*,\mathbf{y}^*)$ be the equilibrium of min-max problem (2). Next, we compute the equations of the Jacobian at the fixed point $(\mathbf{x}^*,\mathbf{y}^*,\mathbf{z}^*,\mathbf{w}^*)$. The following equations can be derived from the equations of the Jacobian (see supplementary material) combined with the fact that the fixed point for g is $(\mathbf{x}^*,\mathbf{y}^*,\mathbf{x}^*,\mathbf{y}^*)$.



(a) #iterations vs size of n (b) l_1 error vs #iterations

Figure 1: *Convergence of OMWU vs different sizes of the problem.* For Figure (a), x-axis is n and y-axis is the number of iterations to reach convergence for Eqn. (6). In Figure (b) we choose four cases of n to illustrate how l_1 error of the problem decreases with the number of iterations.

Derivatives of g_1

$$\frac{\partial g_{1,i}}{\partial x_i} = 1 - x_i^* - 2\eta x_i^* \left(\frac{\partial^2 f}{\partial x_i^*} - \sum_k x_k^* \frac{\partial^2 f}{\partial x_i \partial x_k} \right), i \in [n].$$

$$\frac{\partial g_{1,i}}{\partial x_j} = -x_i^* - 2\eta x_i^* \left(\frac{\partial^2 f}{\partial x_i \partial x_j} - \sum_k x_k^* \frac{\partial^2 f}{\partial x_j \partial x_k} \right), j \in [n].$$

$$\frac{\partial g_{1,i}}{\partial y_j} = -2\eta x_i^* \left(\frac{\partial^2 f}{\partial x_i \partial y_j} - \sum_k x_k^* \frac{\partial^2 f}{\partial x_k \partial y_j} \right), j \in [m].$$

$$\frac{\partial g_{1},i}{\partial z_j} = \eta x_i^* \left(\frac{\partial^2 f}{\partial x_i \partial x_j} - \sum_k x_k^* \frac{\partial^2 f}{\partial x_k \partial x_j} \right), j \in [n].$$

$$\frac{\partial g_{1,i}}{\partial w_j} = \eta x_i^* \left(\frac{\partial^2 f}{\partial x_i \partial y_j} - \sum_k x_k^* \frac{\partial^2 f}{\partial x_k \partial y_j} \right), j \in [m].$$

Derivatives of g_2, g_3, g_4

$$\frac{\partial g_{2,i}}{\partial x_j} = 2\eta y_i^* \left(\frac{\partial^2 f}{\partial x_j \partial y_i} - \sum_k y_k^* \frac{\partial^2 f}{\partial x_j \partial y_k} \right), j \in [n]$$

$$\frac{\partial g_{2,i}}{\partial y_i} = 1 - y_i^* + 2\eta \left(\frac{\partial^2 f}{\partial y_i^2} - \sum_k y_k^* \frac{\partial^2 f}{\partial y_i \partial y_k} \right), i \in [m]$$

$$\frac{\partial g_{2,i}}{\partial y_j} = -y_i^* + 2\eta \left(\frac{\partial^2 f}{\partial y_i \partial y_j} - \sum_k y_k^* \frac{\partial^2 f}{\partial y_j \partial y_k} \right), j \in [m]$$

$$\frac{\partial g_{2,i}}{\partial z_j} = \eta y_i^* \left(-\frac{\partial^2 f}{\partial x_j \partial y_i} + \sum_k y_k^* \frac{\partial^2 f}{\partial x_j \partial y_k} \right), j \in [n]$$

$$\frac{\partial g_{2,i}}{\partial w_j} = \eta y_i^* \left(-\frac{\partial^2 f}{\partial y_i \partial y_j} + \sum_k y_k^* \frac{\partial^2 f}{\partial y_k \partial y_j} \right), j \in [m]$$

$$\frac{\partial g_{3,i}}{\partial x_i} = 1 \text{ for all } i \in [n] \text{ and zero for all the other partial derivatives of } g_{3,i}$$

$$\frac{\partial g_{4,i}}{\partial y_i} = 1 \text{ for all } i \in [m] \text{ and zero for all the other partial derivatives of } g_{4,i}.$$

Let i be the i-th coordinate of $\mathbf{x}^*$. If $i \notin \mathbf{Supp}(\mathbf{x}^*)$ (same for support of $\mathbf{y}^*$), then from the KKT conditions, the partial derivatives w.r.t the i-th coordinate have entries that are zero except the diagonal entry which is less than one. For instance, $\frac{\partial g_{1,i}}{\partial x_i}(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*) = \frac{e^{-\eta \frac{\partial f}{\partial x_i}(\mathbf{x}^*, \mathbf{y}^*)}}{\sum_{t=1}^n x_t^* e^{-\eta \frac{\partial f}{\partial x_t}(\mathbf{x}^*, \mathbf{y}^*)}}$ and all other partial derivatives of $g_{1,i}$ are zero, thus $\frac{e^{-\eta \frac{\partial f}{\partial x_i}(\mathbf{x}^*, \mathbf{y}^*)}}{\sum_{t=1}^n x_t^* e^{-\eta \frac{\partial f}{\partial x_t}(\mathbf{x}^*, \mathbf{y}^*)}}$ is an **eigenvalue** of the Jacobian computed at $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$. This is true because the row of the Jacobian that corresponds to $g_{1,i}$ has zeros everywhere but the diagonal entry. Using Remark 2.2 (degeneracy assumption), It also holds $\frac{e^{-\eta \frac{\partial f}{\partial x_i}(\mathbf{x}^*, \mathbf{y}^*)}}{\sum_{t=1}^n x_t^* e^{-\eta \frac{\partial f}{\partial x_t}(\mathbf{x}^*, \mathbf{y}^*)}} < 1$. Similarly, it holds for $j \notin \text{Supp}(\mathbf{y}^*)$ that $\frac{\partial g_{2,j}}{\partial y_j}(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*) = \frac{e^{\eta \frac{\partial f}{\partial y_j}(\mathbf{x}^*, \mathbf{y}^*)}}{\sum_{t=1}^m y_t^* e^{\eta \frac{\partial f}{\partial y_t}(\mathbf{x}^*, \mathbf{y}^*)}} < 1$.

Thus it suffices to restrict our analysis to the minor of the Jacobian of size $\mathbf{Supp}(\mathbf{x}^*) \times \mathbf{Supp}(\mathbf{y}^*)$ (the rows/cols of the variables we keep correspond to the support of $(\mathbf{x}^*, \mathbf{y}^*)$, and show the particular submatrix has spectral radius less than 1.

We focus on the submatrix of the Jacobian of g computed at $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$ that corresponds to the nonzero probabilities of $\mathbf{x}^*$ and $\mathbf{y}^*$. We denote $D_{\mathbf{x}^*}$ to be the diagonal matrix of size $|\text{Supp}(\mathbf{x}^*)| \times |\text{Supp}(\mathbf{x}^*)|$ that has on the diagonal the nonzero entries of $\mathbf{x}^*$ and similarly we define $D_{\mathbf{y}^*}$ of size $|\text{Supp}(\mathbf{y}^*)| \times |\text{Supp}(\mathbf{y}^*)|$. For convenience, let us denote $k_x := |\text{Supp}(\mathbf{x}^*)|$ and $k_y := |\text{Supp}(\mathbf{y}^*)|$. The Jacobian submatrix is the fol-

lowing

$$J = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ \mathbf{I}_{k_x \times k_x} & \mathbf{0}_{k_x \times k_y} & \mathbf{0}_{k_x \times k_x} & \mathbf{0}_{k_x \times k_y} \\ \mathbf{0}_{k_y \times k_x} & \mathbf{I}_{k_y \times k_y} & \mathbf{0}_{k_y \times k_x} & \mathbf{0}_{k_y \times k_y} \end{bmatrix}$$

where

$$\begin{aligned} A_{11} &= \mathbf{I}_{k_x \times k_x} - D_{\mathbf{x}^*} \mathbf{1}_{k_x} \mathbf{1}_{k_x}^\top - 2\eta D_{\mathbf{x}^*} (\mathbf{I}_{k_x \times k_x} \\ &\quad - \mathbf{1}_{k_x} \mathbf{x}^{*\top}) \nabla^2_{\mathbf{xx}} f, \\ A_{12} &= -2\eta D_{\mathbf{x}^*} (\mathbf{I}_{k_x \times k_x} - \mathbf{1}_{k_x} \mathbf{x}^{*\top}) \nabla^2_{\mathbf{xy}} f \\ A_{13} &= \eta D_{\mathbf{x}^*} (\mathbf{I}_{k_x \times k_x} - \mathbf{1}_{k_x} \mathbf{x}^{*\top}) \nabla^2_{\mathbf{xx}} f, \\ A_{14} &= \eta D_{\mathbf{x}^*} (\mathbf{I}_{k_x \times k_x} - \mathbf{1}_{k_x} \mathbf{x}^{*\top}) \nabla^2_{\mathbf{xy}} f \\ A_{21} &= 2\eta D_{\mathbf{y}^*} (\mathbf{I}_{k_y \times k_y} - \mathbf{1}_{k_y} \mathbf{y}^{*\top}) \nabla^2_{\mathbf{yx}} f, \\ A_{22} &= \mathbf{I}_{k_y \times k_y} - D_{\mathbf{y}^*} \mathbf{1}_{k_y} \mathbf{1}_{k_y}^\top \\ &\quad + 2\eta D_{\mathbf{y}^*} (\mathbf{I}_{k_y \times k_y} - \mathbf{1}_{k_y} \mathbf{y}^{*\top}) \nabla^2_{\mathbf{yy}} f \\ A_{23} &= -\eta D_{\mathbf{y}^*} (\mathbf{I}_{k_y \times k_y} - \mathbf{1}_{k_y} \mathbf{y}^{*\top}) \nabla^2_{\mathbf{yx}} f, \\ A_{24} &= -\eta D_{\mathbf{y}^*} (\mathbf{I}_{k_y \times k_y} - \mathbf{1}_{k_y} \mathbf{y}^{*\top}) \nabla^2_{\mathbf{yy}} f. \end{aligned} \tag{5}$$

We note that $\mathbf{I}, \mathbf{0}$ capture the identity matrix and the all zeros matrix respectively (the appropriate size is indicated as a subscript). The vectors $(\mathbf{1}_{k_x}, \mathbf{0}_{k_y}, \mathbf{0}_{k_x}, \mathbf{0}_{k_y})$ and $(\mathbf{0}_{k_x}, \mathbf{1}_{k_y}, \mathbf{0}_{k_x}, \mathbf{0}_{k_y})$ are left eigenvectors with eigenvalue zero for the above matrix. Hence, any right eigenvector $(\mathbf{v_x}, \mathbf{v_y}, \mathbf{v_z}, \mathbf{v_w})$ should satisfy the conditions $\mathbf{1}^\top \mathbf{v_x} = 0$ and $\mathbf{1}^\top \mathbf{v_y} = 0$. Thus, every non-zero eigenvalue of the above matrix is also a non-zero eigenvalue of the matrix below:

$$J_{\text{new}} = \begin{bmatrix} B_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & B_{22} & A_{23} & A_{24} \\ \mathbf{I}_{k_x \times k_x} & \mathbf{0}_{k_x \times k_y} & \mathbf{0}_{k_x \times k_x} & \mathbf{0}_{k_x \times k_y} \\ \mathbf{0}_{k_y \times k_x} & \mathbf{I}_{k_y \times k_y} & \mathbf{0}_{k_y \times k_x} & \mathbf{0}_{k_y \times k_y} \end{bmatrix}$$

where

$$\begin{aligned} B_{11} &= \mathbf{I}_{k_x \times k_x} - 2\eta D_{\mathbf{x}^*} (\mathbf{I}_{k_x \times k_x} - \mathbf{1}_{k_x} \mathbf{x}^{*\top}) \nabla^2_{\mathbf{xx}} f, \\ B_{22} &= \mathbf{I}_{k_y \times k_y} + 2\eta D_{\mathbf{y}^*} (\mathbf{I}_{k_y \times k_y} - \mathbf{1}_{k_y} \mathbf{y}^{*\top}) \nabla^2_{\mathbf{yy}} f. \end{aligned}$$

The characteristic polynomial of J_{new} is obtained by finding $\det(J_{\text{new}} - \lambda \mathbf{I})$. First observe that $\lambda = 1$ is an eigenvalue of J_{new} with left eigenvector the all ones (which correspond to a zero eigenvalue for initial matrix J). Moreover, if the eigenvalue $\lambda = 1$ has multiplicity more than two, it follows that the Hessian is singular (which violates our assumption). So we may assume for the rest of the proof that $\lambda \neq 1$.

One can perform row/column operations on J_{new} to calculate this determinant, which gives us the following relation:

$$\det(J_{\text{new}} - \lambda \mathbf{I}_{2k_x \times 2k_y}) = (1 - 2\lambda)^{(k_x + k_y)} \, q\left(\frac{\lambda(\lambda - 1)}{2\lambda - 1}\right)$$

where $q(\lambda)$ is the characteristic polynomial of the following matrix

$$J_{\text{small}} = \begin{bmatrix} B_{11} - \mathbf{I}_{k_x \times k_x} & A_{12} \\ A_{21} & B_{22} - \mathbf{I}_{k_y \times k_y}, \end{bmatrix}$$

and $B_{11}, B_{12}, A_{12}, A_{21}$ are the aforementioned submatrices. Notice that J_{small} can be written as

$$J_{\text{small}} = 2\eta \begin{bmatrix} -(D_{\mathbf{x}^*} - \mathbf{x}^* \mathbf{x}^{*\top}) & \mathbf{0}_{k_x \times k_y} \\ \mathbf{0}_{k_y \times k_x} & (D_{\mathbf{y}^*} - \mathbf{y}^* \mathbf{y}^{*\top}) \end{bmatrix} H$$

where

$$H = \begin{bmatrix} \nabla^2_{\mathbf{xx}} f & \nabla^2_{\mathbf{xy}} f \\ \nabla^2_{\mathbf{yx}} f & \nabla^2_{\mathbf{yy}} f \end{bmatrix}.$$

Notice here that H is the Hessian matrix evaluated at the fixed point $(\mathbf{x}^*, \mathbf{y}^*)$, and is the appropriate submatrix restricted to the support of $|\text{Supp}(\mathbf{y}^*)|$ and $|\text{Supp}(\mathbf{x}^*)|$. Although, the Hessian matrix is symmetric, we would like to work with the following representation of J_{small}:

$$J_{\text{small}} = 2\eta \begin{bmatrix} (D_{\mathbf{x}^*} - \mathbf{x}^* \mathbf{x}^{*\top}) & \mathbf{0}_{k_x \times k_y} \\ \mathbf{0}_{k_y \times k_x} & (D_{\mathbf{y}^*} - \mathbf{y}^* \mathbf{y}^{*\top}) \end{bmatrix} H^-$$

where

$$H^- = \begin{bmatrix} -\nabla^2_{\mathbf{xx}} f & -\nabla^2_{\mathbf{xy}} f \\ \nabla^2_{\mathbf{yx}} f & \nabla^2_{\mathbf{yy}} f \end{bmatrix}.$$

Let us denote any non-zero eigenvalue of J_{small} by ϵ which may be a complex number. Thus ϵ is where $q(\cdot)$ vanishes and hence the eigenvalue of J_{new} must satisfy the relation $\frac{\lambda(\lambda-1)}{2\lambda-1} = \epsilon$.

We are to now show that the magnitude of any eigenvalue of J_{new} is strictly less than 1, i.e, $|\lambda| < 1$. Trivially, $\lambda = \frac{1}{2}$ satisfies the above condition. Thus we need to show that the magnitude of λ where $q(\cdot)$ vanishes is strictly less than 1. The remainder of the proof proceeds by showing the following two lemmas:

Lemma 3.1 (Real part non-positive)**.** *Let λ be an eigenvalue of matrix J_{small}. It holds that $Re(\lambda) \leq 0$.*

Proof. Assume that $\lambda \neq 0$. All the non-zero eigenvalues of matrix J_{small} coincide with the eigenvalues of the matrix

$$\begin{aligned} R := &\begin{bmatrix} (D_{\mathbf{x}^*} - \mathbf{x}^* \mathbf{x}^{*\top}) & \mathbf{0}_{k_x \times k_y} \\ \mathbf{0}_{k_y \times k_x} & (D_{\mathbf{y}^*} - \mathbf{y}^* \mathbf{y}^{*\top}) \end{bmatrix}^{\frac{1}{2}} \times H^- \\ &\times \begin{bmatrix} (D_{\mathbf{x}^*} - \mathbf{x}^* \mathbf{x}^{*\top}) & \mathbf{0}_{k_x \times k_y} \\ \mathbf{0}_{k_y \times k_x} & (D_{\mathbf{y}^*} - \mathbf{y}^* \mathbf{y}^{*\top}) \end{bmatrix}^{\frac{1}{2}}. \end{aligned}$$

This is well-defined since

$$\begin{bmatrix} (D_{\mathbf{x}^*} - \mathbf{x}^* \mathbf{x}^{*\top}) & \mathbf{0}_{k_x \times k_y} \\ \mathbf{0}_{k_y \times k_x} & (D_{\mathbf{y}^*} - \mathbf{y}^* \mathbf{y}^{*\top}) \end{bmatrix}$$

is positive semi-definite. Moreover, we use KyFan inequalities which state that the sequence (in decreasing order) of the eigenvalues of $\frac{1}{2}(W+W^\top)$ majorizes the real part of the sequence of the eigenvalues of W for any square matrix W (see Moslehian (2011), page 4). We conclude that for any eigenvalue λ of R, it holds that $\mathrm{Re}(\lambda)$ is at most the maximum eigenvalue of $\frac{1}{2}(R+R^\top)$. Observe now that $R+R^\top$ is equal to

$$\begin{bmatrix} (D_{\mathbf{x}^*}-\mathbf{x}^*\mathbf{x}^{*\top}) & \mathbf{0}_{k_x\times k_y} \\ \mathbf{0}_{k_y\times k_x} & (D_{\mathbf{y}^*}-\mathbf{y}^*\mathbf{y}^{*\top}) \end{bmatrix}^{\frac{1}{2}} \times (H^- + H^{-\top}) \times \begin{bmatrix} (D_{\mathbf{x}^*}-\mathbf{x}^*\mathbf{x}^{*\top}) & \mathbf{0}_{k_x\times k_y} \\ \mathbf{0}_{k_y\times k_x} & (D_{\mathbf{y}^*}-\mathbf{y}^*\mathbf{y}^{*\top}) \end{bmatrix}^{\frac{1}{2}}.$$

Since

$$H^- + H^{-\top} = \begin{bmatrix} -\nabla^2_{\mathbf{xx}} f & 0 \\ 0 & \nabla^2_{\mathbf{yy}} f \end{bmatrix}$$

by the convex-concave assumption on f it follows that the matrix above is negative semi-definite (see Remark 2.2) and so is $R+R^\top$. We conclude that the maximum eigenvalue of $R+R^\top$ is non-positive. Therefore any eigenvalue of R has real part non-positive and the same is true for J_{small}. □

Lemma 3.2. *If ϵ is a non-zero eigenvalue of J_{small} then, $Re(\epsilon)\le 0$ and $|\epsilon|\downarrow 0$ as the stepsize $\eta\to 0$.*

Proof. Let $\lambda = x+\sqrt{-1}y$ and $\epsilon = a+\sqrt{-1}b$. The relation $\frac{\lambda(\lambda-1)}{2\lambda-1}=\epsilon$ gives two equations based on the equality of real and imaginary parts, i.e., : $x^2-x-y^2 = 2ax-a-2by$ and $2xy-y = 2bx+2ay-b$. Notice that the above equations can be transformed to the following forms:

$$(x-\frac{2a+1}{2})^2-(y-b)^2 = -a-b^2+\frac{(2a+1)^2}{4}$$
$$(x-\frac{2a+1}{2})(y-b) = ab.$$

For each $\epsilon = a+\sqrt{-1}b$, there exist two pairs of points (x_1,y_1) and (x_2,y_2) that are the intersections of the above two hyperbola. First consider the case when, $a<0$. As $|\epsilon|\to 0$, the hyperbola can be obtained from the translation by $(\frac{2a+1}{2},b)$ of the hyperbola

$$x^2-y^2 = -a-b^2+\frac{(2a+1)^2}{4} \text{ and } xy = ab$$

where the translated symmetric center is close to $(\frac{1}{2},0)$ since (a,b) is close to $(0,0)$. So the two intersections of the above hyperbola, (x_1,y_1) and (x_2,y_2), satisfy the property that $x_1^2+y_1^2$ is small and $x_2>\frac{1}{2}$ since the two intersections are on two sides of the axis $x=\frac{2a+1}{2}$, On the other hand, we have

$$\frac{\lambda(\lambda-1)}{2\lambda-1} = \frac{(x+\sqrt{-1}y)(x-1+\sqrt{-1}y)}{2x-1+\sqrt{-1}2y} = \epsilon = a+\sqrt{-1}b$$

and then the condition $a<0$ gives the inequality

$$\mathrm{Re}(\epsilon) = \frac{(x^2-x+y^2)(2x-1)}{(2x-1)^2+4y^2} < 0$$

that is equivalent to $x>\frac{1}{2}$ and $x^2-x+y^2<0$, where only the case $x>\frac{1}{2}$ is considered since if the intersection whose x-component satisfying $x<\frac{1}{2}$ has the property that x^2+y^2 is small and then less than 1. Thus to prove that $|\lambda|<1$, it suffices to assume $x>\frac{1}{2}$. It is obvious that $x^2-x+y^2 = (x-\frac{1}{2})^2+y^2-\frac{1}{4}<0$ implies that $x^2+y^2<1$.

Now consider the case when $a=0$. Here, the quadratic system becomes $(x-\frac{1}{2})^2-(y-b)^2 = -b^2+\frac{1}{4}$ and $(x-\frac{1}{2})(y-b)=0$. It must hold that $x=\frac{1}{2}$ or $y=b$. Since the step-size η can be arbitrarily small constant, when $|\epsilon|\to 0$, we get $b\to 0$ and then y must be a complex number, contradicting that y is real. So we can only have $y=b$, and from the equation $(x-\frac{1}{2})^2 = -b^2+\frac{1}{4}$, we have $x=\frac{1}{2}\pm\sqrt{-b^2+\frac{1}{4}}$. In this case it also holds $x^2+y^2 = \frac{1}{2}\pm\sqrt{-b^2+\frac{1}{4}}<1$ for $|b|,|\epsilon|\to 0$ (i.e., small enough). □

Remark 3.3. In our proof we showed that the spectral radius $\mathrm{sp}(J)$ of the Jacobian of the update rule of OMWU is upper-bounded by a positive number $\mathrm{sp}(J)<\rho<1$ less than one. From a standard dynamical systems argument, if this is the case for a twice continuously differentiable function f, then $\|\mathbf{x}_t-\mathbf{x}^*\|\le Ce^{-\rho t}$ where ρ depends on the dimension of f, Lipschitzness L, strong-convexity μ etc but not on t and C on the initialization. Thus, to reach ϵ close (in ℓ_2) to the solution, we can do it after $O\left(\log\frac{1}{\epsilon}\right)$ iterations (as long as we start close enough to $\mathbf{x}^*$). Observe also that it must be the case that $\epsilon\ll\eta$ (the claim holds for ϵ much smaller than η). Nevertheless, the constants in $O\left(\log\frac{1}{\epsilon}\right)$ are not clear as we do not know how ρ depends on n,m,L.

4 Experiments

In this section, we primarily target to understand two factors that influence the convergence speed of OMWU: the problem size and the learning rate. We also compare our algorithm with Optimistic Gradient Descent Ascent (OGDA) with projection, and demonstrate our superiority against it. We start with a simple bilinear min-max game:

$$\min_{\mathbf{x}\in\Delta_n}\max_{\mathbf{y}\in\Delta_n}\mathbf{x}^\top A\mathbf{y}. \tag{6}$$

We first vary the value of n to study how the learning speed scales with the size of the problem. The learning rate is fixed at 1.0, and we run OMWU with

$n \in \{25, 50, 75, \cdots, 250\}$ and matrix $A \in \mathbb{R}^{n \times n}$ is generated with i.i.d random Gaussian entries. We output the number of iterations for OMWU to reach convergence, i.e., with l_1 error to the optimal solution to be less or equal to 10^{-5}. The results are averaged from 10 runs with different randomly initializations. As reported in Figure 1, generally a larger problem size requires more iterations to reach convergence. We also provide four specific cases of n to show the convergence in l_1 distance in Figure 1(b). The shaded area demonstrates the standard deviation from the 50 runs.

To understand how learning rate affects the speed of convergence, we conduct similar experiments on Eqn. (6) and plot the l_1 error with different step sizes in Figure 2(a)-(c). For this experiment the matrix size is fixed as $n = 100$. We also include a comparison with the Optimistic Gradient Descent AscentDaskalakis and Panageas (2018). For the setting we considered, we observe a larger learning rate effectively speeds up our learning process, and our algorithm is relatively more stable to the choice of step-size. In comparison, OGDA is quite sensitive to the choice of step-size. As shown in Figure 2(b), a larger step-size makes the algorithm diverge, while a smaller step-size will make very little progress. Furthermore, we also choose to perform our algorithm over a convex-concave but not bilinear function $f(\mathbf{x}, \mathbf{y}) = x_1^2 - y_1^2 + 2x_1y_1$, where $\mathbf{x}, \mathbf{y} \in \Delta_2$ and x_1 and y_1 are the first coefficients of $\mathbf{x}$ and $\mathbf{y}$. With this low dimensional function, we could visually show the convergence procedure as in Figure 2(b), where each arrow indicates an OMWU step. This figure demonstrates that at least in this case, a larger step size usually makes sure a bigger progress towards the optimal solution.

Remark 4.1. Notice that our experiments doesn't show slow convergence of OGDA. Figure 5(b)(c) shows that when we use the learning rate of e^2, OGDA is even faster than OMWU, with linear convergence. However, OGDA is much more sensitive to hyperparameter tuning. When the learning rate is too big (e^4 in (b)), it diverges, and while it's too small (e^{-2}), it converges slowly (but still achieves linear convergence). The update rule for OGDA follows exactly from Daskalakis and Panageas (2019), and we use the same projection steps for our method and OGDA as presented in Chen and Ye (2011).

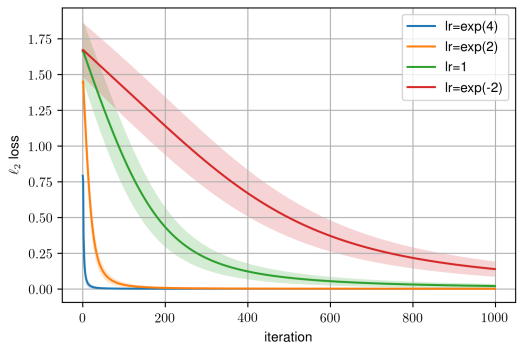
Finally we show how the KL divergence $D_{KL}((\mathbf{x}^*, \mathbf{y}^*) \parallel (\mathbf{x}^t, \mathbf{y}^t))$ decreases under different circumstances. Figure 3 again considers the bilinear problem (Eqn.(6)) with multiple dimensions n and a simple convex-concave function $f(\mathbf{x}, \mathbf{y}) = x_1^2 - y_1^2 + 2x_1y_1$ with different learning rate. We note that in all circumstances we consider, OMWU achieves global convergence.

5 Conclusion

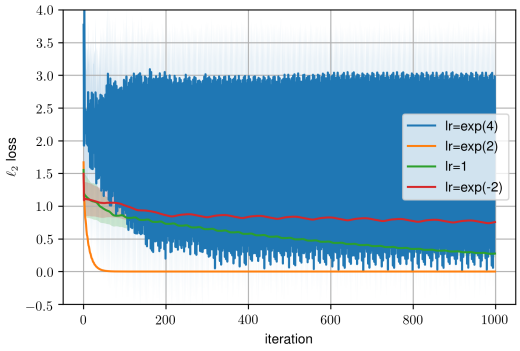
In this paper we analyze the last iterate behavior of a no-regret learning algorithm called Optimistic Multiplicative Weights Update for convex-concave landscapes. We prove that OMWU with constant stepsize exhibits last iterate convergence in a neighborhood of the fixed point of OMWU algorithm, generalizing previous results that showed last iterate convergence for bilinear functions. The provided experiments indicate that OMWU achieves global convergence. One possible open question is to show global last iterate convergence of OMWU for generic convex-concave games.

Acknowledgements

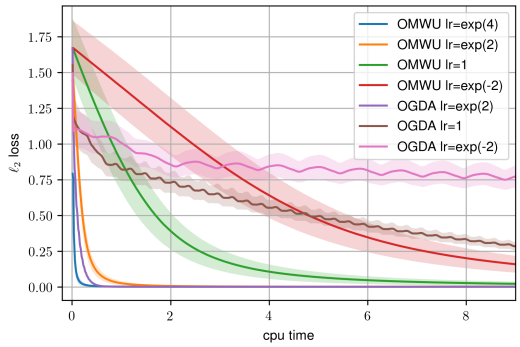
Qi Lei is supported by NSF #2030859 and the Computing Research Association for the CIFellows Project. Sai Ganesh Nagarajan would like to acknowledge SUTD President's Graduate Fellowship (SUTD-PGF). Ioannis Panageas would like to acknowledge UC Irvine start-up grant. Xiao Wang would like to acknowledge NRF for AI Fellowship 2019.



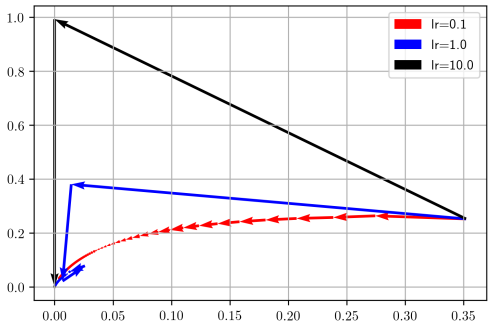
(a) OMWU



(b) OGDA

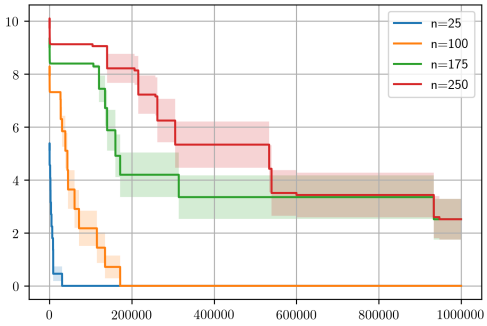


(c) Convergence time comparisons

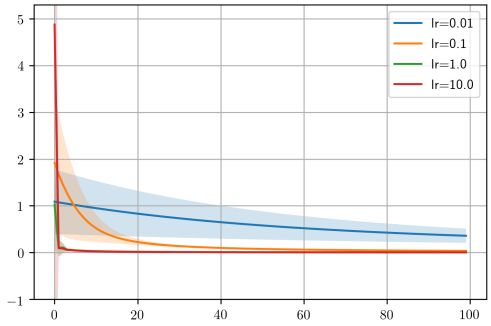


(d) OMWU trajectories with different learning rate

Figure 2: *Time comparisons of OMWU and projected OGDA vs different choices of learning rate.* For Figure (a)(b)(c), x-axis is iterations and y-axis is the l_1 error to the stationary point for Eqn. (6) with $n = 100$. We observe that OMWU (as in (a)) always converges while projected OGDA (as in (b)) will diverge for large learning rate. In figure (c) we remove the divergent case and compare the efficiency of the two algorithm measured in CPU time. In Figure (d) we visually present the trajectories for the min-max game of $\min_{\mathbf{x}\in\Delta_2}\max_{\mathbf{y}\in\Delta_2}\{x_1^2 - y_1^2 + 2x_1y_1\}$ with learning rate $0.1, 1.0$ and 10. Here x-axis is the value of x_1 and y-axis is the value of y_1 respectively. The equilibrium point the algorithm converges to is $\mathbf{x} = [0, 1], \mathbf{y} = [0, 1]$.



(a) KL divergence vs #iterations with different n



(b) KL divergence vs #iterations with different η

Figure 3: *KL divergence decreases with #iterations under different settings.* For both images, x-axis is the number of iterations, and y-axis is KL divergence. Figure (a) is OMWU on bilinear function Eqn.(6) with $n = \{25, 100, 175, 250\}$. Figure (b) is OMWU on the quadratic function $f(\mathbf{x}, \mathbf{y}) = x_1^2 - y_1^2 + 2x_1y_1$ with different learning rate η in $\{0.01, 0.1, 1.0, 10.0\}$. Shaded area indicates standard deviation from 10 runs with random initializations. OMWU with smaller learning rate tends to have higher variance.

References

Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 263–274. Omnipress, 2008.

Jacob D. Abernethy, Kevin A. Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *CoRR*, abs/1906.02027, 2019. URL `http://arxiv.org/abs/1906.02027`.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*, pages 321–338, 2018.

G.W Brown. Iterative solutions of games by fictitious play. In *Activity Analysis of Production and Allocation*, 1951.

Nikolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Yunmei Chen and Xiaojing Ye. Projection onto a simplex. *arXiv preprint arXiv:1101.6081*, 2011.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 9256–9266, 2018.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 27:1–27:18, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. In *Proceedings of ICLR*, 2018.

Oded Galor. *Discrete Dynamical Systems*. Springer, 2007.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 6936–6946, 2019.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *arXiv preprint:1802.06132*, 2018.

Yura Malitsky and Matthew K Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

Panayotis Mertikopoulos, Houssam Zenati, Bruno Lecouat, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *CoRR*, abs/1807.02629, 2018.

Mohammad Sal Moslehian. Ky fan inequalities. *CoRR*, abs/1108.1467, 2011.

Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5874–5884, 2017.

Klaus Ritzberger. The theory of normal form games from the differentiable viewpoint. *International Journal of Game Theory*, pages 207–236, 1994.

J. Robinson. An iterative method of solving a game. In *Annals of Mathematics*, pages 296–301, 1951.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Annual Conference on Neural Information Processing Systems 2015*, pages 2989–2997, 2015.

J Von Neumann. Zur theorie der gesellschaftsspiele. In *Math. Ann.*, pages 295–320, 1928.