arXiv:1807.04252v2 [math.OC] 17 Sep 2018

Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization

Constantinos Daskalakis
MIT
costis@csail.mit.edu

Ioannis Panageas
MIT
ioannis@csail.mit.edu

Abstract

Motivated by applications in Game Theory, Optimization, and Generative Adversarial Networks, recent work of Daskalakis et al [8] and follow-up work of Liang and Stokes [11] have established that a variant of the widely used Gradient Descent/Ascent procedure, called "Optimistic Gradient Descent/Ascent (OGDA)", exhibits last-iterate convergence to saddle points in *unconstrained* convex-concave min-max optimization problems. We show that the same holds true in the more general problem of *constrained* min-max optimization under a variant of the no-regret Multiplicative-Weights-Update method called "Optimistic Multiplicative-Weights Update (OMWU)". This answers an open question of Syrgkanis et al [19].

The proof of our result requires fundamentally different techniques from those that exist in no-regret learning literature and the aforementioned papers. We show that OMWU monotonically improves the Kullback-Leibler divergence of the current iterate to the (appropriately normalized) min-max solution until it enters a neighborhood of the solution. Inside that neighborhood we show that OMWU becomes a contracting map converging to the exact solution. We believe that our techniques will be useful in the analysis of the last iterate of other learning algorithms.

1 Introduction

A central problem in Game Theory and Optimization is computing a pair of probability vectors $(\mathbf{x}, \mathbf{y})$, solving

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

where $\Delta_n \subset \mathbb{R}^n$ and $\Delta_m \subset \mathbb{R}^m$ are probability simplices, and A is a $n \times m$ matrix. Von Neumann's celebrated minimax theorem informs us that

$$\min_{\mathbf{y}\in\Delta_m} \max_{\mathbf{x}\in\Delta_n} \mathbf{x}^\top A\mathbf{y} = \max_{\mathbf{x}\in\Delta_n} \min_{\mathbf{y}\in\Delta_m} \mathbf{x}^\top A\mathbf{y}, \tag{2}$$

and that all solutions to the LHS are solutions to the RHS, and vice versa. This result was a founding stone in the development of Game Theory. Indeed, interpreting $\mathbf{x}^\top A\mathbf{y}$ as the payment of the "min player" to the "max player" when the former selects a distribution $\mathbf{y}$ over columns and the latter selects a distribution $\mathbf{x}$ over rows of matrix A, a solution to (1) constitutes an equilibrium of the game defined by matrix A, called a "minimax equilibrium", a pair of randomized strategies such that neither player can improve their payoff by unilaterally changing their distribution.

Besides their fundamental value for Game Theory, it is known that (1) and (2) are also intimately related to Linear Programming. It was shown by von Neumann that (2) follows from strong linear programming duality. Moreover, it was suggested by Dantzig [7] and recently proven by Adler [1]

that any linear program can be solved by solving some min-max problem of the form (1). In particular, min-max problems of form (1) are exactly as expressive as min-max problems of the following form, which capture any linear program (by Lagrangifying the constraints):

$$\min_{\mathbf{y}\geq 0}\max_{\mathbf{x}\geq 0}\left(\mathbf{x}^\top A\mathbf{y}+b^\top\mathbf{x}+c^\top\mathbf{y}\right). \tag{3}$$

Soon after the minimax theorem was proven and its connection to linear programming was forged, researchers proposed dynamics for solving min-max optimization problems by having the min and max players of (1) run a simple learning procedure in tandem. An early method, proposed by Brown [4] and analyzed by Robinson [18], was fictitious play. Soon after, Blackwell's approachability theorem [3] propelled the field of online learning, which lead to the discovery of several learning algorithms converging to minimax equilibrium at faster rates, while also being robust to adversarial environments, situations where one of the players of the game deviates from the prescribed dynamics; see e.g. [5]. These learning methods, called "no-regret", include the celebrated multiplicative-weights-update method, follow-the-regularized-leader, and follow-the-perturbed-leader. Compared to centralized linear programming procedures the advantage of these methods is the simplicity of executing their steps, and their robustness to adversarial environments, as we just discussed.

Last vs Average Iterate Convergence. Despite the extensive literature on no-regret learning, an unsatisfactory feature of known results is that min-max equilibrium is shown to be attained only in an average sense. To be precise, if $(\mathbf{x}^t,\mathbf{y}^t)$ is the trajectory of a no-regret learning method, it is usually shown that the average $\frac{1}{t}\sum_{\tau\leq t}\mathbf{x}^{\tau\top}A\mathbf{y}^\tau$ converges to the optimal value of (1), as $t\to\infty$. Moreover, if the solution to (1) is unique, then $\frac{1}{t}\sum_{\tau\leq t}(\mathbf{x}^\tau,\mathbf{y}^\tau)$ converges to the optimal solution. Unfortunately that does not mean that the last iterate $(\mathbf{x}^t,\mathbf{y}^t)$ converges to an optimal solution, and indeed it commonly diverges or enters a limit cycle. Furthermore, in the optimization literature, Nesterov [15] provides a method that can give pointwise convergence (i.e., convergence of the last iterate) to problem (1)[1], however his algorithm is not a no-regret learning algorithm. Recent work by Daskalakis et al [8] and Liang and Stokes [11] studies whether last iterate convergence can be established for no-regret learning methods in the simple unconstrained min-max problem of the form:

$$\min_{\mathbf{y}\in\mathbb{R}^m}\max_{\mathbf{x}\in\mathbb{R}^n}\left(\mathbf{x}^\top A\mathbf{y}+b^\top\mathbf{x}+c^\top\mathbf{y}\right). \tag{4}$$

For this problem, it is known that Gradient Descent/Ascent (GDA) is a no-regret learning procedure, corresponding to follow-the-regularized leader (FTRL) with ℓ_2^2-regularization. As such, the average trajectory traveled by GDA converges to a min-max solution, in the afore-described sense. On the other hand, it is also known that GDA may diverge from the min-max solution, even in trivial cases such as $A=\mathrm{I}, n=m=1, b=c=0$. Interestingly, [8, 11] show that a variant of GDA, called "Optimistic Gradient Descent/Ascent (OGDA),[2] exhibits last iterate convergence. Inspired by their theoretical result for the performance of OGDA in (4), Daskalakis et al. [8] even propose the use of OGDA for training Generative Adversarial Networks (GANs) [10]. Moreover, Syrgkanis et al. [19] provide numerical experiments which indicate that the trajectories of Optimistic Hedge (variant of Hedge in the same way OGDA is a variant of GDA) stabilize (i.e., converge pointwise) as opposed to (classic) Hedge and they *posed the question whether Optimistic Hedge actually converges pointwise.*

[1] Nesterov showed that by optimizing $f_\mu(\mathbf{x}):=\mu\ln(\frac{1}{m}\sum_{j=1}^m e^{-\frac{1}{\mu}(A\mathbf{x})_j})$, $g_\nu(\mathbf{x}):=\nu\ln(\frac{1}{n}\sum_{j=1}^n e^{\frac{1}{\nu}(A^\top\mathbf{y})_j})$ for $\mu=\Theta(\frac{\epsilon}{\log m})$, $\nu=\Theta(\frac{\epsilon}{\log n})$ yields an $O(\epsilon)$ approximation to the problem problem (1).

[2] OGDA is tantamount to Optimistic FTRL with ℓ_2^2-regularization, in the same way that GDA is tantamount to FTRL with ℓ_2^2-regularization; see e.g. [17]. OGDA essentially boils down to GDA with negative momentum.

Motivated by the afore-described lines of work, and the importance of last iterate convergence for Game Theory and the modern applications of GDA-style methods in Optimization, our goal in this work is to generalize the results of [8, 11] to the general min-max problem (3), or equivalently (1); indeed, we will focus on the latter, but our algorithms are readily applicable to the former as the two problems are equivalent [1]. With the constraint that $(\mathbf{x}, \mathbf{y})$ should remain in $\Delta_n \times \Delta_m$, GDA and OGDA are not applicable. Indeed, the natural GDA-style method for min-max problems in this case is the celebrated Multiplicative-Weights-Update (MWU) method, which is tantamount to FTRL with entropy-regularization. Unsurprisingly, in the same way that GDA suffers in the unconstrained problem (4), MWU exhibits cycling in the constrained problem (1) (a recent work is [2] and was also shown empirically in [19]). So it is natural for us to study instead its optimistic variant, "Optimistic Multiplicative-Weights-Update (OMWU)," (called Optimistic Hedge in [19]) which corresponds to Optimistic FTRL with entropy-regularization, the equations of which are given in Section 2.2. Our main result is the following (restated as Theorem 2.7 after Section 2.2) and answers an open question asked in [19] as applicable to two player zero sum games:

Theorem 1.1 (**Last-Iterate Convergence of OMWU**). *Whenever* (1) *has a unique optimal solution* $(\mathbf{x}^*, \mathbf{y}^*)$*, OMWU with a small enough learning rate and initialized at the pair of uniform distributions* $(\frac{1}{n}\mathbf{1}, \frac{1}{m}\mathbf{1})$ *exhibits last-iterate convergence to the optimal solution. That is, if* $(\mathbf{x}^t, \mathbf{y}^t)$ *are the vectors maintained by OMWU at step* t*, then* $\lim_{t\to\infty}(\mathbf{x}^t, \mathbf{y}^t) = (\mathbf{x}^*, \mathbf{y}^*)$.

Remark 1.2. *We note that the assumption about uniqueness of the optimal solution for problem* (1) *is generic in the following sense: Within the set of all zero-sum games, the set of zero-sum games with non-unique equilibrium has Lebesgue measure zero [2, 6]. This implies that if* A*'s entries are sampled independently from some continuous distribution, then with probability one the min-max problem* (1) *will have a unique solution.*

Our paper provides *two important messages*:

- It strengthens the intuition that optimism helps the trajectories of learning dynamics stabilize (e.g., Optimistic MWU vs MWU or Optimistic GDA vs GDA; as the papers of Syrgkanis et al [19] and Daskalakis et al [8] also do).
- The techniques we use (typically appear in dynamical systems literature) to prove convergence for the last iterate, are fundamentally different from those commonly used to prove convergence of the time average of a learning algorithm.

Notation: Vectors in Δ_n, Δ_m are denoted in boldface $\mathbf{x}, \mathbf{y}$. Time indices are denoted by superscripts. Thus, a time indexed vector $\mathbf{x}$ at time t is denoted as $\mathbf{x}^t$. We use the letter J to denote the Jacobian of a function (with appropriate subscript), $\mathbf{I}, \mathbf{0}, \mathbf{1}$ to denote the identity, zero matrix and all ones vector respectively with appropriate subscripts to indicate the size. Moreover, $(A\mathbf{y})_i$ captures $\sum_j A_{ij} y_j$. The support of $\mathbf{x}$ is denoted by $\mathrm{Supp}(\mathbf{x})$. Finally we use $(\mathbf{x}^*, \mathbf{y}^*)$ to denote the optimal solution for the min-max problem (1) and $[n]$ to denote $\{1, ..., n\}$.

2 Preliminaries

2.1 Definitions and facts

Dynamical Systems. 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 $\mathcal{S} = \Delta_n \times \Delta_m \times \Delta_n \times \Delta_m$ for our purposes. The point $\mathbf{z}$ is called a *fixed point* or *equilibrium* of w if $w(\mathbf{z}) = \mathbf{z}$. We will be interested in the following well known fact that will be used in our proofs.

Proposition 2.1 (e.g. [9]). *If the Jacobian of the update rule w[3] at a fixed point $\mathbf{z}$ has spectral radius less than one, then there exists a neighborhood U around $\mathbf{z}$ such that for all $\mathbf{x} \in U$, the dynamics converges to $\mathbf{z}$, i.e., $\lim_{n\to\infty} w^n(\mathbf{x}) = \mathbf{z}$. We call w a contraction mapping in U.*

2.2 OMWU Method

Our main contribution is that the last iterate of OMWU converges to the optimal solution. The OMWU dynamics is defined as follows ($t \geq 1$):

$$\begin{aligned} x_i^{t+1} &= x_i^t \frac{e^{2\eta(A\mathbf{y}^t)_i - \eta(A\mathbf{y}^{t-1})_i}}{\sum_{j=1}^n x_j^t e^{2\eta(A\mathbf{y}^t)_j - \eta(A\mathbf{y}^{t-1})_j}} \text{ for all } i \in [n], \\ y_i^{t+1} &= y_i^t \frac{e^{-2\eta(A^\top\mathbf{x}^t)_i + \eta(A^\top\mathbf{x}^{t-1})_i}}{\sum_{j=1}^m y_j^t e^{-2\eta(A^\top\mathbf{x}^t)_j + \eta(A^\top\mathbf{x}^{t-1})_j}} \text{ for all } i \in [m]. \end{aligned} \tag{5}$$

Points $(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^0, \mathbf{y}^0)$ are the initial conditions and are given as input. We call $0 < \eta < 1$ the *stepsize* of the dynamics. It is more convenient to interpret OMWU dynamics as mapping a quadruple to quadruple ($(\mathbf{x}^t, \mathbf{y}^t, \mathbf{x}^{t-1}, \mathbf{y}^{t-1}) \to (\mathbf{x}^{t+1}, \mathbf{y}^{t+1}, \mathbf{x}^t, \mathbf{y}^t)$, see Section 3.2 for the construction of the dynamical system).

Remark 2.2. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be the optimal solution. We see that $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$ is a fixed point of the mapping. Furthermore, $\Delta_n \times \Delta_m \times \Delta_n \times \Delta_m$ is invariant under OMWU dynamics. For $t \geq 1$, if $x_i^t = 0$ then x_i remains zero for all times greater than t, and if it is positive, it remains positive (both numerator and denominator are positive)* [4]*. In words, at all times the OMWU satisfies the non-negativity constraints and the renormalization factor (denominator) makes both $\mathbf{x}, \mathbf{y}$'s coordinates sum up to one. A last observation is that every fixed point of OMWU dynamics (mapping a quadruple to quadruple) has the form $(\mathbf{x}, \mathbf{y}, \mathbf{x}, \mathbf{y})$ (two same copies). Equation (8) shows how to express OMWU dynamics as a dynamical system.*

2.3 Linear Variant of OMWU

We provide the linear variant of OMWU dynamics (5) because we use it in some intermediate lemmas (appear in appendix).

$$\begin{aligned} x_i^{t+1} &= x_i^t \frac{1 + 2\eta(A\mathbf{y}^t)_i - \eta(A\mathbf{y}^{t-1})_i}{\sum_{j=1}^n x_j^t (1 + 2\eta(A\mathbf{y}^t)_j - \eta(A\mathbf{y}^{t-1})_j)} \text{ for all } i \in [n], \\ y_i^{t+1} &= y_i^t \frac{1 - 2\eta(A^\top\mathbf{x}^t)_i + \eta(A^\top\mathbf{x}^{t-1})_i}{\sum_{j=1}^m y_j^t (1 - 2\eta(A^\top\mathbf{x}^t)_j + \eta(A^\top\mathbf{x}^{t-1})_j)} \text{ for all } i \in [m]. \end{aligned} \tag{6}$$

This dynamics is derived by considering the first order approximation of the exponential function. Stepsize η in this case should be chosen sufficiently small so that both numerator and denominator are positive.

2.4 More definitions and statement of our result

Definition 2.3 ([12]). *Assume $\alpha > 0$. We call a point $(\mathbf{x}, \mathbf{y}) \in \Delta_n \times \Delta_m$ α-close if for each i we have that $x_i \leq \alpha$ or $|\mathbf{x}^\top A\mathbf{y} - (A\mathbf{y})_i| \leq \alpha$ and for each j it holds $y_j \leq \alpha$ or $|\mathbf{x}^\top A\mathbf{y} - (A^\top\mathbf{x})_j| \leq \alpha$.*

Remark 2.4. *Think of α-close points as α-approximate optimal solutions for min-max problems that are induced by* submatrices *of A (α-approximate stationary points). Moreover, if $(\mathbf{x}, \mathbf{y})$ is 0-close point does not necessarily imply $(\mathbf{x}, \mathbf{y})$ is the optimal solution of problem (1)!*

[3] We assume w is a continuously differential function.

[4] Same holds for vector $\mathbf{y}$.

Definition 2.5 (Approximate solution)**.** *Assume $\epsilon > 0$. We call a point $(\mathbf{x}, \mathbf{y}) \in \Delta_n \times \Delta_m$ ϵ-approximate (or ϵ-approximate Nash equilibrium) if for all $\tilde{\mathbf{x}} \in \Delta_n$ we get that $\tilde{\mathbf{x}}^\top A\mathbf{y} \leq \mathbf{x}^\top A\mathbf{y} + \epsilon$ (max player deviates) and for all $\tilde{\mathbf{y}} \in \Delta_m$ we get that $\mathbf{x}^\top A\tilde{\mathbf{y}} \geq \mathbf{x}^\top A\mathbf{y} - \epsilon$ (min player deviates).*

Remark 2.6. *Think of ϵ-approximate points as approximate optimal solutions to the min-max problem (1). Moreover, if $(\mathbf{x}, \mathbf{y})$ is 0-approximate then $(\mathbf{x}, \mathbf{y})$ is the optimal solution of problem (1).*

Statement of our results. We finish the preliminary section by stating formally the main result.

Theorem 2.7 (OMWU converges)**.** *Let A be a $n \times m$ matrix and assume that*

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

has a unique solution $(\mathbf{x}^*, \mathbf{y}^*)$. It holds that for η sufficiently small (depends on n, m, A), starting from the uniform distribution, i.e., $(\mathbf{x}^1, \mathbf{y}^1) = (\mathbf{x}^0, \mathbf{y}^0) = (\frac{1}{n}\mathbf{1}, \frac{1}{m}\mathbf{1})$, it holds

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

under OMWU dynamics. The stepsize η is constant, i.e., does not scale with time[5]*.*

We need to note that it is not clear from our theorem how small η is and its dependence on the size of A. Nevertheless, our convergence result holds for constant stepsizes as opposed to the classic no-regret learning literature where η scales like $\frac{1}{\sqrt{T}}$ after T iterations. Another result we know of this flavor is about MWU algorithm on congestion games [16].

3 Last iterate convergence of OMWU

In this section we show our main result (Theorem 2.7), by breaking the proof into three key theorems. The first theorem says that KL divergence from the t-th iterate $(\mathbf{x}^t, \mathbf{y}^t)$ to the optimal solution $(\mathbf{x}^*, \mathbf{y}^*)$, i.e., (sum of KL divergences to be exact)

$$\sum_i x_i^* \ln(x_i^*/x_i^t) + \sum_i y_i^* \ln(y_i^*/y_i^t),$$

decreases with time $t \geq 2$ by at least a factor of η^3 per iteration, unless the iterate $(\mathbf{x}^t, \mathbf{y}^t)$ is $O(\eta^{1/3})$-close (see Definition 2.3). Moreover, provided that the stepsize η is small enough, we can show the structural result that $(\mathbf{x}^t, \mathbf{y}^t)$ lies in a neighborhood of $(\mathbf{x}^*, \mathbf{y}^*)$ that becomes smaller and smaller as $\eta \to 0$. Finally, as long as OMWU dynamics has reached a small neighborhood around $(\mathbf{x}^*, \mathbf{y}^*)$, we show that the update rule of the dynamical system induced by OMWU is contracting, and the last iterate convergence result follows. Formally we show:

Theorem 3.1 (KL decreasing)**.** *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be the unique optimal solution of problem (1) and η sufficiently small. Then*

$$D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^t, \mathbf{y}^t))$$

is decreasing with time t by (at least) $\Omega(\eta^3)$ unless $(\mathbf{x}^t, \mathbf{y}^t)$ is $O(\eta^{1/3})$-close.

[5] Our proof also works if the starting points $(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^0, \mathbf{y}^0)$ are both in the interior of $\Delta_n \times \Delta_m$ and not necessarily uniform, however the choice of η depends on the initial distributions as well and not only on n, m, A.

Theorem 3.2 ($\eta^{1/3}$-close implies close to optimum in ℓ_1)**.** *Assume that $(\mathbf{x}^*, \mathbf{y}^*)$ is unique optimal solution of the problem (1). Let T (depends on η) be the first time KL divergence does not decrease by $\Omega(\eta^3)$. It follows that as $\eta \to 0$, the $\eta^{1/3}$-close point $(\mathbf{x}^T, \mathbf{y}^T)$ has distance from $(\mathbf{x}^*, \mathbf{y}^*)$ that goes to zero, i.e., $\lim_{\eta\to 0} \left\|(\mathbf{x}^*, \mathbf{y}^*) - (\mathbf{x}^T, \mathbf{y}^T)\right\|_1 = 0$.*

Theorem 3.3 (OMWU is a contraction)**.** *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be the unique optimal solution to the min-max problem (1). There exists a neighborhood $U \subset \Delta_n \times \Delta_m \times \Delta_n \times \Delta_m$ of $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$[6] so that for all $(\mathbf{x}^1, \mathbf{y}^1, \mathbf{x}^0, \mathbf{y}^0) \in U$ we have that $\lim_{t\to\infty}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{x}^{t-1}, \mathbf{y}^{t-1}) = (\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$ under OMWU dynamics as defined in (5) and (8) (Section 3.2).*

Assuming these three theorems, our main result is straightforward.

Proof of Theorem 2.7. Let η be sufficiently small. If $(\mathbf{x}^1, \mathbf{y}^1) = (\frac{1}{n}\mathbf{1}, \frac{1}{m}\mathbf{1})$ (starting point is uniform) then an easy upper bound (by removing negative terms) on KL divergence from $(\mathbf{x}^1, \mathbf{y}^1)$ to $(\mathbf{x}^*, \mathbf{y}^*)$ is $-\sum_{i=1}^n x_i^* \log x_i^1 + \sum_{i=1}^m y_i^* \log y_i^1 = \log(nm)$. Therefore using Theorem 3.1 we have that after at most T that is $O(\frac{\log(nm)}{\eta^3})$ steps, OMWU reaches a $O(\eta^{1/3})$-close point (T is the first time so that KL divergence from current iterate to optimal solution $(\mathbf{x}^*, \mathbf{y}^*)$ has not decreased by at least a factor of η^3) or the KL divergence between the optimal solution and $(\mathbf{x}^T, \mathbf{y}^T)$ is $O(\eta^3)$ (KL divergence was decreasing by at least a factor of η^3 for all iterations until the iterate reached a ℓ_1 distance $O(\eta^3)$). In the latter case it follows $\left\|(\mathbf{x}^*, \mathbf{y}^*) - (\mathbf{x}^T, \mathbf{y}^T)\right\|_1^2$ is $O(\eta^3)$ and hence $(\mathbf{x}^T, \mathbf{y}^T)$ is $O(\eta^{3/2})$ in ℓ_1 distance from the optimal solution, therefore for small η, $(\mathbf{x}^{T+1}, \mathbf{y}^{T+1}, \mathbf{x}^T, \mathbf{y}^T)$ is in the neighborhood U that is needed for contraction (Theorem 3.3). In the former case, by Theorem 3.2 (for η sufficiently small) it follows that $(\mathbf{x}^{T+1}, \mathbf{y}^{T+1}, \mathbf{x}^T, \mathbf{y}^T)$ is also in the neighborhood U that is needed for contraction (Theorem 3.3)[7]. The proof follows by Theorem 3.3. □

In the next subsections we will provide the proofs to all three key theorems.

3.1 KL decreases and OMWU reaches neighborhood

In this subsection we argue about the proofs of Theorems 3.1 and 3.2. The inequality we managed to prove (see in the appendix the proof of Theorem 3.1) is the following:

$$\begin{aligned}
&D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^{t+1}, \mathbf{y}^{t+1})) - D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^t, \mathbf{y}^t)) \leq \\
&-\textstyle\sum_{i=1}^n x_i^t \left((\frac{1}{2} - O(\eta))\eta^2 \left(2(A\mathbf{y}^t)_i - 2\mathbf{x}^{t\,\top} A\mathbf{y}^t - (A\mathbf{y}^{t-1})_i + \mathbf{x}^{t\,\top} A\mathbf{y}^{t-1}\right)^2\right) \\
&-\textstyle\sum_{i=1}^m y_i^t \left((\frac{1}{2} - O(\eta))\eta^2 \left(2(A^\top\mathbf{x}^t)_i - 2\mathbf{x}^{t\,\top} A\mathbf{y}^t - (A^\top\mathbf{x}^{t-1})_i + \mathbf{x}^{t-1\,\top} A\mathbf{y}^t\right)^2\right) + O(\eta^3).
\end{aligned} \tag{7}$$

The proof of the inequality is quite long, we choose to provide intuition and skip the details. We refer to the appendix for a proof. The inequality says that OMWU dynamics has a good progress (KL divergence decreases by at least a factor of η^3) as long as the current and previous iterate $(\mathbf{x}^t, \mathbf{y}^t), (\mathbf{x}^{t-1}, \mathbf{y}^{t-1})$ are not α-close for α chosen to be $O(\eta^{1/3})$. This situation appears a lot in gradient methods when the dynamics is close to a stationary point, the gradient of f is small and the progress is small as opposed to the case where the gradient of f is big and there is satisfying progress. The RHS of inequality (7) captures the "distance" from stationarity. Thus, as long as we are not close to a stationary point (i.e., $O(\eta^{1/3})$-close) in a time window between 1,2,...,k, KL divergence from current iterate (k-th) to the optimum has decreased by (at least) $\Omega(k\eta^3)$ compared to KL divergence from first iterate to the optimum.

[6] Since $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$ might be on the boundary of $\Delta_n \times \Delta_m \times \Delta_n \times \Delta_m$, U is the intersection of an open ball around $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$ with $\Delta_n \times \Delta_m \times \Delta_n \times \Delta_m$.

[7] In both cases we used that iterate $(\mathbf{x}^T, \mathbf{y}^T)$ and $(\mathbf{x}^{T+1}, \mathbf{y}^{T+1})$ have ℓ_1 distance $O(\eta)$, this is Lemma B.1.

Moreover, suppose that at some point of OMWU dynamics, KL divergence from current iterate to the optimum did not decrease by at least a factor of η^3 and let T be the iteration this happened. As we have already argued, $(\mathbf{x}^T, \mathbf{y}^T)$ is a $O(\eta^{1/3})$-close point. We can show that as long as η is sufficiently small, then for all i, j in the support of $(\mathbf{x}^*, \mathbf{y}^*)$, x_i^T, y_j^T are (at least) $\Omega(\eta^{1/3})$ i.e., coordinates in the support of the optimum will have non negligible probability in $(\mathbf{x}^T, \mathbf{y}^T)$. Formally:

Lemma 3.4. *Let $i \in Supp(\mathbf{x}^*)$ and $j \in Supp(\mathbf{y}^*)$. It holds that $x_i^T \geq \frac{1}{2}\eta^{1/3}$ and $y_i^T \geq \frac{1}{2}\eta^{1/3}$ as long as*

$$\eta^{1/3} \leq \min_{s \in Supp(\mathbf{x}^*)} \frac{1}{(nm)^{1/x_s^*}}, \min_{s \in Supp(\mathbf{y}^*)} \frac{1}{(nm)^{1/y_s^*}}.$$

Proof. By definition of T, the KL divergence is decreasing for $2 \leq t \leq T-1$, thus

$$D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^{T-1}, \mathbf{y}^{T-1})) < D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^1, \mathbf{y}^1)).$$

Therefore $x_i^* \log \frac{1}{x_i^{T-1}} < \sum_i x_i^* \log \frac{1}{x_i^1} + \sum_i y_j^* \log \frac{1}{y_j^1} = \log(mn)$. It follows that $x_i^T > 1/(mn)^{\frac{1}{x_i^*}} \geq \eta^{1/3}$ for $x_i^* > 0$ ($i \in \text{Supp}(\mathbf{x}^*)$). Since $|x_i^T - x_i^{T-1}|$ is $O(\eta)$ (Lemma B.1) the result follows. Similarly, the argument works for y_j^T. □

Lemma 3.4 indicates that the stepsize η might have to be exponentially small in the dimension (OMWU dynamics is slow when η is very small). We can now prove Theorem 3.2.

Proof of Theorem 3.2. From Lemma 3.4 and definition of T, we get that $|(A\mathbf{y}^T)_i - \mathbf{x}^{T\ \top} A\mathbf{y}^T|$ is $O(\eta^{1/3})$ for all i in the support of $\mathbf{x}^*$ and $|(A^\top \mathbf{x}^T)_j - \mathbf{x}^{T\ \top} A\mathbf{y}^T|$ is $O(\eta^{1/3})$ for all j in the support of $\mathbf{y}^*$. We consider $(\mathbf{w}^T, \mathbf{z}^T)$ to be the projection of the point $(\mathbf{x}^T, \mathbf{y}^T)$ by removing all the coordinates that have probability mass less than $\frac{1}{2}\eta^{1/3}$ and rescale so that the coordinates sum up to one.

We restrict ourselves to the corresponding subproblem (submatrix). It is clear that $(\mathbf{w}^T, \mathbf{z}^T)$ is a $O(\eta^{1/3})$-approximate solution [8] for the subproblem. Let $v = \mathbf{x}^* A\mathbf{y}^*$ be the optimal value. By uniqueness of the optimal solution, we get that $(A\mathbf{y}^*)_i = v$ for all $i \in \text{Supp}(\mathbf{x}^*)$ and $(A\mathbf{y}^*)_i < v$ otherwise (check Lemma C.3 in paper [13] for a proof, where they use Farkas' lemma to show it, we use this fact later in Section 3.2). Similarly $(A^\top \mathbf{x}^*)_j = v$ for the min player $\mathbf{y}$ if j lies in the support of $\mathbf{y}^*$ and $(A^\top \mathbf{x}^*)_j > v$ otherwise. We choose η so small that every $O(\eta^{1/3})$-approximate solution $(\mathbf{x}, \mathbf{y})$ has the property that $(A\mathbf{y})_i \leq v - \eta^{1/4}$, $(A^\top \mathbf{x})_j \geq v + \eta^{1/4}$ for all $i \notin \text{Supp}(\mathbf{x}^*)$ and $j \notin \text{Supp}(\mathbf{y}^*)$ respectively (this is possible by continuity of the bilinear function and Claim 3.5 below). Hence we conclude that if η is small enough, the coordinates in the vector $(\mathbf{w}^T, \mathbf{z}^T)$ that are not in the support of the optimal solution (since $\eta^{1/4} \gg \eta^{1/3}$), should have probability mass $O(\eta^{1/3})$ at time T.

Claim 3.5. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be the unique optimal solution to the problem (1). For every $\epsilon > 0$, there exists an $\delta(\epsilon) > 0$ so that for every δ-approximate solution $(\mathbf{x}, \mathbf{y})$ we get that $|x_i - x_i^*| < \epsilon$ for all $i \in [n]$. Analogously holds for player $\mathbf{y}$.*

Proof. We will prove this by contradiction. Assume there is an ϵ that violates this statement. We choose a sequence δ_k so that $\lim_{k\to\infty} \delta_k = 0$ and also there is a sequence $(\mathbf{x}_k, \mathbf{y}_k)$ of δ_k-approximate Nash equilibrium with $|x_{k,i} - x_i^*| \geq \epsilon$ for some strategy i. Since $\Delta_n \times \Delta_m$ is compact and the sequence above is bounded, there is a convergent subsequence. The limit of the convergent subsequence is a Nash equilibrium by definition of δ-approximate (Definition 2.5). By uniqueness it follows that the i-th coordinate of the convergent sequence must converge to x_i^*, hence we reached a contradiction. □

[8] By ϵ-approximate optimal solution we mean the ϵ-approximate Nash equilibrium notion (additive), see Definition 2.5.

Therefore, if we restrict to the subproblem induced by the strategies in the support of $(\mathbf{x}^*, \mathbf{y}^*)$, the projected vector $(\mathbf{w}^T, \mathbf{z}^T)$ is a $O(\eta^{1/3})$-approximate solution of the subgame.

From Claim 3.5, as $\eta \to 0$ it follows that the ℓ_1 distance (any distance suffices) between the projected $(\mathbf{w}^T, \mathbf{z}^T)$ and the optimal solution (Nash equilibrium) of the subgame goes to zero. Since the optimal solution of the subgame is exactly the same as the optimal solution of the original game we get that as $\eta \to 0$, $(\mathbf{x}^T, \mathbf{y}^T)$ reaches $(\mathbf{x}^*, \mathbf{y}^*)$. In particular, since $\left\|(\mathbf{x}^{T+1}, \mathbf{y}^{T+1}) - (\mathbf{x}^T, \mathbf{y}^T)\right\|_1$ is $O(\eta)$ (see Lemma B.1) there exists a η small so that $(\mathbf{x}^{T+1}, \mathbf{y}^{T+1}, \mathbf{x}^T, \mathbf{y}^T)$ is inside the necessary neighborhood U of $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$ that gives contraction (Theorem 3.3). □

3.2 Proving contraction

The purpose of this section is to prove Theorem 3.3. To show contraction of OMWU dynamics in a neighborhood of the optimal solution $(\mathbf{x}^*, \mathbf{y}^*)$, we first construct a dynamical system that captures OMWU. Moreover, we prove that the Jacobian of the update rule of that particular dynamical system computed at the optimal solution, has spectral radius less than one. This suffices to prove contraction (see Proposition 2.1). As a result, as long as OMWU reaches a small neighborhood of $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$, it converges pointwise (last iterate convergence) to it[9]. Below we provide the update rule g of the dynamical system, which consists of 4 components:

$$\begin{aligned}
&g(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}) := (g_1(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}), g_2(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}), g_3(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}), g_4(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w})),\\
&g_{1,i}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}) := (g_1(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}))_i := x_i \frac{e^{2\eta(A\mathbf{y})_i - \eta(A\mathbf{w})_i}}{\sum_t x_t e^{2\eta(A\mathbf{y})_t - \eta(A\mathbf{w})_t}} \text{ for all } i \in [n],\\
&g_{2,i}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}) := (g_2(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}))_i := y_i \frac{e^{-2\eta(A^\top\mathbf{x})_i + \eta(A^\top\mathbf{z})_i}}{\sum_t y_t e^{-2\eta(A^\top\mathbf{x})_t + \eta(A^\top\mathbf{z})_t}} \text{ for all } i \in [m],\\
&g_3(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}) := \mathbf{I}_{n\times n}\mathbf{x},\\
&g_4(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}) := \mathbf{I}_{m\times m}\mathbf{y}.
\end{aligned} \tag{8}$$

It is not hard to check that

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

so g captures exactly the dynamics of OMWU (5). The equations of the Jacobian of g can be found in the appendix (see Section A).

Spectral analysis the Jacobian of OMWU at the optimal solution. The rest of the section constitutes the proof of Theorem 3.3. Assume $v = \mathbf{x}^{*\top} A\mathbf{y}^*$, i.e., v is the value of the bilinear function $\mathbf{x}^\top A\mathbf{y}$ at the optimal solution. We will analyze the Jacobian computed at $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$[10].

Assume $i \notin \mathrm{Supp}(\mathbf{x}^*)$, then

$$\frac{\partial g_{1,i}}{\partial x_i} = \frac{e^{\eta(A\mathbf{y}^*)_i}}{\sum x_t^* e^{\eta(A\mathbf{y}^*)_t}} = \frac{e^{\eta(A\mathbf{y}^*)_i}}{e^{\eta v}}$$

and all other partial derivatives of $g_{1,i}$ are zero, thus $\frac{e^{\eta(A\mathbf{y}^*)_i}}{e^{\eta v}}$ is an eigenvalue of the Jacobian computed at $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$. Moreover because of uniqueness of the optimal solution, it holds that $\frac{e^{\eta(A\mathbf{y}^*)_i}}{e^{\eta v}} < 1$ because $(A\mathbf{y}^*)_i - v < 0$ (check Lemma C.3 in [13] for a proof, where they use Farkas' Lemma to show it). Similarly, it holds for $j \notin \mathrm{Supp}(\mathbf{y}^*)$ that $\frac{\partial g_{2,j}}{\partial y_j} = \frac{e^{-\eta(A^\top\mathbf{x}^*)_j}}{e^{-\eta v}} < 1$ (again by C.3 in

[9] Since the dynamical system is from a quadruple to a quadruple, it is a neighborhood of $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$.

[10] See also Equations (14) of the Jacobian computed at $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$.

[13] it holds that $(A\mathbf{x}^*)_j - v > 0$) and all other partial derivatives of $g_{2,j}$ are zero, hence $\frac{e^{-\eta(A^\top \mathbf{x}^*)_j}}{e^{-\eta v}}$ is an eigenvalue of the Jacobian computed at the optimal solution.

Let D_x be the diagonal matrix of size $|\mathrm{Supp}(\mathbf{x}^*)| \times |\mathrm{Supp}(\mathbf{x}^*)|$ that has on the diagonal the nonzero entries of $\mathbf{x}^*$ and similarly we define D_y of size $|\mathrm{Supp}(\mathbf{y}^*)| \times |\mathrm{Supp}(\mathbf{y}^*)|$. We set $k_1 = |\mathrm{Supp}(\mathbf{x}^*)|, k_2 = |\mathrm{Supp}(\mathbf{y}^*)|$ and $k = k_1 + k_2$. Let $\mathbf{x}', \mathbf{y}'$ be the optimal solution to the min-max problem with payoff matrix the corresponding submatrix of payoff matrix A (denoted by B) after removing the rows/columns which correspond to the coordinates that are not in the support of the unique optimal solution $(\mathbf{x}^*, \mathbf{y}^*)$[11]. We consider the submatrix $\tilde{J}$ of the Jacobian matrix that is created by removing rows and columns of the corresponding coordinates that are not in the support of optimum (for the variables $\mathbf{x}$ and $\mathbf{y}$, these are exactly $n+m-k$). It is clear from above, that the Jacobian of OMWU has eigenvalues with absolute value less than one iff $\tilde{J}$ has as well. After also removing the rows (and the corresponding columns) that have only zero entries (these are exactly $n+m-k$, result zero eigenvalues and correspond to variables $\mathbf{z}$ and $\mathbf{w}$) the resulting submatrix (denote it by J) boils down to the following:

$$J = \begin{pmatrix} \mathbf{I}_{k_1\times k_1} - D_x \mathbf{1}_{k_1}\mathbf{1}_{k_1}^\top & 2\eta D_x(B - v\mathbf{1}_{k_1}\mathbf{1}_{k_2}^\top) & \mathbf{0}_{k_1\times k_1} & -\eta D_x(B - v\mathbf{1}_{k_1}\mathbf{1}_{k_2}^\top) \\ 2\eta D_y(v\mathbf{1}_{k_2}\mathbf{1}_{k_1}^\top - B^\top) & \mathbf{I}_{k_2\times k_2} - D_y\mathbf{1}_{k_2}\mathbf{1}_{k_2}^\top & -\eta D_y(v\mathbf{1}_{k_2}\mathbf{1}_{k_1}^\top - B^\top) & \mathbf{0}_{k_2\times k_2} \\ \mathbf{I}_{k_1\times k_1} & \mathbf{0}_{k_1\times k_2} & \mathbf{0}_{k_1\times k_1} & \mathbf{0}_{k_1\times k_2} \\ \mathbf{0}_{k_2\times k_1} & \mathbf{I}_{k_2\times k_2} & \mathbf{0}_{k_2\times k_1} & \mathbf{0}_{k_2\times k_2} \end{pmatrix}. \tag{9}$$

It is clear that $(\mathbf{1}_{k_1}, \mathbf{0}_{k_2}, \mathbf{0}_{k_1}, \mathbf{0}_{k_2})$, $(\mathbf{0}_{k_1}, \mathbf{1}_{k_2}, \mathbf{0}_{k_1}, \mathbf{0}_{k_2})$ are left eigenvectors with eigenvalues zero and thus any right eigenvector $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}})$ with nonzero eigenvalue has the property that $\tilde{\mathbf{x}}^\top \mathbf{1}_{k_1} = 0$ and $\tilde{\mathbf{y}}^\top \mathbf{1}_{k_2} = 0$. Hence every nonzero eigenvalue of the matrix above is an eigenvalue of the matrix below:

$$J_{\text{new}} = \begin{pmatrix} \mathbf{I}_{k_1\times k_1} & 2\eta D_x B & \mathbf{0}_{k_1\times k_1} & -\eta D_x B \\ -2\eta D_y B^\top & \mathbf{I}_{k_2\times k_2} & \eta D_y B^\top & \mathbf{0}_{k_2\times k_2} \\ \mathbf{I}_{k_1\times k_1} & \mathbf{0}_{k_1\times k_2} & \mathbf{0}_{k_1\times k_1} & \mathbf{0}_{k_1\times k_2} \\ \mathbf{0}_{k_2\times k_1} & \mathbf{I}_{k_2\times k_2} & \mathbf{0}_{k_2\times k_1} & \mathbf{0}_{k_2\times k_2} \end{pmatrix}. \tag{10}$$

Let $p(\lambda)$ be the characteristic polynomial of the matrix (10). After row/column operations it boils down to

$$(-1)^k \det \begin{pmatrix} \lambda(1-\lambda)\mathbf{I}_{k_1\times k_1} & (2\lambda-1)\eta D_x B \\ -\eta(2\lambda-1)D_y B^\top & \lambda(1-\lambda)\mathbf{I}_{k_2\times k_2} \end{pmatrix} = (1-2\lambda)^k q\left(\frac{\lambda(\lambda-1)}{2\lambda-1}\right), \tag{11}$$

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

$$J_{\text{small}} = \begin{pmatrix} \mathbf{0}_{k_1\times k_1} & \eta D_x B \\ -\eta D_y B^\top & \mathbf{0}_{k_2\times k_2} \end{pmatrix}. \tag{12}$$

Observe that

$$J_{\text{small}} \cdot \begin{pmatrix} D_x & \mathbf{0}_{k_1\times k_2} \\ \mathbf{0}_{k_2\times k_1} & D_y \end{pmatrix} \text{ is real skew symmetric,}$$

and hence by Lemma B.6, J_{small} has eigenvalues of the form[12] $\pm i\eta\tau$ with $\tau \in \mathbb{R}$ (i.e., imaginary eigenvalues; we include η in the expression to conclude that $\sigma := \eta\tau$ can be sufficiently small in absolute value). We conclude that any nonzero eigenvalue λ of the matrix J should satisfy the equation $\frac{\lambda(\lambda-1)}{2\lambda-1} = i\sigma$ for some small in absolute value $\sigma \in \mathbb{R}$. Finally we get that

$$\lambda = \frac{1 + 2i\sigma \pm \sqrt{1-4\sigma^2}}{2}.$$

[11] Note that $(\mathbf{x}', \mathbf{y}')$ should be the *unique* optimal solution to the min-max problem with payoff matrix B.

[12] We denote $i = \sqrt{-1}$.

We compute the square of the magnitude of λ and we get $|\lambda|^2 = \frac{2-4\sigma^2 \pm 2\sqrt{1-4\sigma^2}+4\sigma^2}{4} = \frac{1\pm\sqrt{1-4\sigma^2}}{2} < 1$ unless $\sigma = 0$ (i.e., $\tau = 0$). If $\sigma = 0$, it means that J_{new} has an eigenvalue which is equal to one. Assume that $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ is the corresponding right eigenvector, it holds that $B\tilde{\mathbf{y}} = \mathbf{0}$ and $B^\top \tilde{\mathbf{x}} = \mathbf{0}$. Assume also that there exists an eigenvalue that is equal to one in the original matrix J. It follows that $\mathbf{1}_{k_2}^\top \tilde{\mathbf{y}} = 0$ and $\mathbf{1}_{k_1}^\top \tilde{\mathbf{x}} = 0$. It holds that $\tilde{\mathbf{x}} = \mathbf{0}_{k_1}$ and $\tilde{\mathbf{y}} = \mathbf{0}_{k_2}$ otherwise $(\mathbf{x}', \mathbf{y}') + t(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ would be another optimal solution (for the min-max problem with payoff matrix B; by padding zeros to the vector, we could create another optimal solution for the original min-max problem with payoff matrix A) for small enough t. We reached contradiction because we have assumed uniqueness. Hence all the eigenvalues of J are less than 1, i.e., the mapping is a contraction mapping and the proof is complete.

4 Experiments

The purpose of our experiments is primarily to understand how the speed of convergence of OMWU dynamics (5) scales with the size of matrix A. Moreover, for A of fixed size, we are interested in how the speed of convergence scales with the error of the output of OMWU dynamics. By error we mean the ℓ_1 distance between the last iterate of OMWU and the optimal solution.

For the former case, we fix the error to be 0.1 and we run OMWU for $n = 25, 50, ..., 250$ where the input matrix A has size $n \times n$ with entries i.i.d random variables sampled from uniform $[-1, 1]$. We output the number of iterations OMWU needs starting from uniform $(\frac{1}{n}, ..., \frac{1}{n})$ to reach a solution that is at most 0.1 away from optimal in ℓ_1 distance. We note that we computed the optimal solutions using LP-solvers.

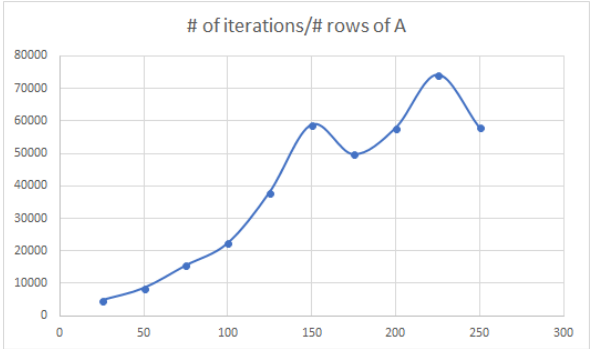
For the latter case, we fix $n = 50$ and we consider the error ϵ to be $\{0.5, 0.25, 0.0625, 0.015625, 0.007812\}$. Starting from uniform distribution, we count the number of iterations to reach error ϵ. The stepsize η is fixed at 0.01 at all times. The results can be found in the figure below (Figure 4). If we had to guess, it seems that the relation between dimension and iterations is between linear and quadratic (i.e., OMWU dynamics has roughly cubic-quartic running time in n if we count the cost of each iteration as quadratic) and the dependence between error ϵ and iterations t seems like t is inverse polynomial in ϵ.

We note the importance of stepsize η. η must be sufficiently small for our proofs to work. If η is chosen to be big, then OMWU might not converge (might cycle, we observed such behavior in experiments). On the other hand, the smaller η is chosen, the smaller the progress of OMWU dynamics (see the inequality claim for KL divergence) and hence the slower the dynamics.

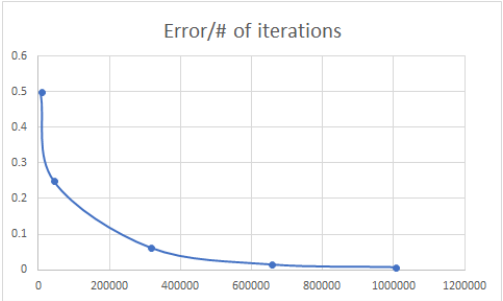
5 Conclusion

In this paper we showed that a no-regret algorithm called Optimistic Multiplicative Weights Update (OMWU) converges pointwise to a Nash equilibrium in two player zero sum games (See also a concurrent work to ours [14], in which the authors provide a pointwise result about other dynamics, using different techniques). Our analysis is novel and does not follow the standard approaches of the literature of no-regret learning. We believe that our techniques can be useful in the analysis of other learning algorithms with no provable guarantees of pointwise convergence.

One interesting open question is to show that OMWU algorithm converges in polynomial time in n, m (for proper choice of stepsize η) and find exact rates of convergence. Another possible future direction is to generalize our results about OMWU beyond the bilinear setting.



(a) In the x axis we have the number of rows of a square matrix A and on y axis the number of iterations of OMWU. This figure captures how the number of iterations depends on the dimensionality of the min-max problem.



(b) In the x axis we have the number of iterations of OMWU and on y axis the ℓ_1 distance from the optimal solution. This figure captures how the number of iterations scales with the error.

References

[1] Ilan Adler. The equivalence of linear programs and zero-sum games. In *International Journal of Game Theory*, pages 165–177, 2013.

[2] 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.

[3] David Blackwell. An analog of the minimax theorem for vector payoffs. In *Pacific J. Math.*, pages 1–8, 1956.

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

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

[6] Eric Van Damme. *Stability and perfection of Nash equilibria*. Springer, 1991.

[7] George B Dantzig. A proof of the equivalence of the programming problem and the game problem. *Activity analysis of production and allocation*, (13):330–338, 1951.

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

[9] Oded Galor. *Discrete Dynamical Systems*. Springer, 2007.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

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

[12] Ruta Mehta, Ioannis Panageas, Georgios Piliouras, Prasad Tetali, and Vijay V. Vazirani. Mutation, sexual reproduction and survival in dynamic environments. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 16:1–16:29, 2017.

[13] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 2703–2717, 2018.

[14] 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.

[15] Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

[16] 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.

[17] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, pages 993–1019, 2013.

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

[19] 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.

A Equations of the Jacobian of OMWU dynamics

A.1 Equations computed at point $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w})$

Set $S_x = \sum_{t=1}^{n} x_t e^{2\eta(A\mathbf{y})_t - \eta(A\mathbf{w})_t}$, $S_y = \sum_{t=1}^{m} y_t e^{-2\eta(A^\top \mathbf{x})_t + \eta(A^\top \mathbf{z})_t}$ and let i, j be arbitrary indexes ($g_{1,i}$ captures the i-th coordinate of function g_1 etc),

$$
\begin{aligned}
\frac{\partial g_{1,i}}{\partial x_i} &= \frac{e^{2\eta(A\mathbf{y})_i - \eta(A\mathbf{w})_i}}{S_x} - x_i \frac{\left(e^{2\eta(A\mathbf{y})_i - \eta(A\mathbf{w})_i}\right)^2}{S_x^2} \text{ for all } i \in [n],\\
\frac{\partial g_{1,i}}{\partial x_j} &= -x_i e^{2\eta(A\mathbf{y})_j - \eta(A\mathbf{w})_j} \cdot \frac{e^{2\eta(A\mathbf{y})_i - \eta(A\mathbf{w})_i}}{S_x^2} \text{ for all } i \in [n], j \in [n] \text{ and } j \neq i,\\
\frac{\partial g_{1,i}}{\partial y_j} &= x_i e^{2\eta(A\mathbf{y})_i - \eta(A\mathbf{w})_i} \cdot \frac{2\eta A_{ij} S_x - 2\eta \sum_t A_{tj} x_t e^{2\eta(A\mathbf{y})_t - \eta(A\mathbf{w})_t}}{S_x^2} \text{ for all } i \in [n], j \in [m],\\
\frac{\partial g_{1,i}}{\partial z_j} &= 0 \text{ for all } i, j \in [n],\\
\frac{\partial g_{1,i}}{\partial w_j} &= x_i e^{2\eta(A\mathbf{y})_i - \eta(A\mathbf{w})_i} \cdot \frac{-\eta A_{ij} S_x + \eta \sum_t A_{tj} x_t e^{2\eta(A\mathbf{y})_t - \eta(A\mathbf{w})_t}}{S_x^2} \text{ for all } i \in [n], j \in [m],\\
\frac{\partial g_{2,i}}{\partial y_i} &= \frac{e^{-2\eta(A^\top \mathbf{x})_i + \eta(A^\top \mathbf{z})_i}}{S_y} - y_i \frac{\left(e^{-2\eta(A^\top \mathbf{x})_i + \eta(A^\top \mathbf{z})_i}\right)^2}{S_y^2} \text{ for all } i \in [m],\\
\frac{\partial g_{2,i}}{\partial y_j} &= -y_i e^{-2\eta(A^\top \mathbf{x})_j + \eta(A^\top \mathbf{z})_j} \cdot \frac{e^{-2\eta(A^\top \mathbf{x})_i + \eta(A^\top \mathbf{z})_i}}{S_y^2} \text{ for all } i \in [m], j \in [m] \text{ and } j \neq i,\\
\frac{\partial g_{2,i}}{\partial x_j} &= y_i e^{-2\eta(A^\top \mathbf{x})_i + \eta(A^\top \mathbf{z})_i} \cdot \frac{-2\eta A_{ij}^\top S_y + 2\eta \sum_t A_{tj}^\top y_t e^{-2\eta(A^\top \mathbf{x})_t + \eta(A^\top \mathbf{z})_t}}{S_y^2} \text{ for all } i \in [m], j \in [n],\\
\frac{\partial g_{2,i}}{\partial z_j} &= y_i e^{-2\eta(A^\top \mathbf{x})_i + \eta(A^\top \mathbf{z})_i} \cdot \frac{\eta A_{ij}^\top S_y - \eta \sum_t A_{tj}^\top x_t e^{-2\eta(A^\top \mathbf{x})_t + \eta(A^\top \mathbf{z})_t}}{S_y^2} \text{ for all } i \in [m], j \in [n],\\
\frac{\partial g_{2,i}}{\partial w_j} &= 0 \text{ for any } i, j \in [m],\\
\frac{\partial g_{3,i}}{\partial x_i} &= 1 \text{ for all } i \in [n] \text{ and zero 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 all the other partial derivatives of } g_{4,i}.
\end{aligned}
\tag{13}
$$

A.2 Equations computed at point $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$

Set $S_x = \sum_{t=1}^{n} x_t^* e^{\eta(A\mathbf{y}^*)_t}$, $S_y = \sum_{t=1}^{m} y_t^* e^{-\eta(A^\top \mathbf{x}^*)_t}$ and let i, j be arbitrary indexes ($g_{1,i}$ captures the i-th coordinate of function g_1 etc). Assume $v = \mathbf{x}^{*\top} A\mathbf{y}^*$, it is not hard to see that $(A^\top \mathbf{x}^*)_i = (A\mathbf{y}^*)_j = v$ for all $i \in \text{Supp}(\mathbf{x}^*), j \in \text{Supp}(\mathbf{y}^*)$ and $S_x = e^{\eta v}, S_y = e^{-\eta v}$. We get that:

$$
\begin{aligned}
&\tfrac{\partial g_{1,i}}{\partial x_i} = 1 - x_i^* \text{ for all } i \in \text{Supp}(\mathbf{x}^*),\\
&\tfrac{\partial g_{1,i}}{\partial x_i} = \tfrac{e^{\eta(A\mathbf{y}^*)_i}}{e^{\eta v}} \text{ for all } i \notin \text{Supp}(\mathbf{x}^*),\\
&\tfrac{\partial g_{1,i}}{\partial x_j} = -x_i^* \text{ for all } i, j \in \text{Supp}(\mathbf{x}^*) \text{ and } j \neq i,\\
&\tfrac{\partial g_{1,i}}{\partial x_j} = 0 \text{ for all } i \notin \text{Supp}(\mathbf{x}^*), j \in [n] \text{ and } j \neq i,\\
&\tfrac{\partial g_{1,i}}{\partial y_j} = x_i^*(2\eta A_{ij} - 2\eta v) \text{ for all } i \in \text{Supp}(\mathbf{x}^*), j \in \text{Supp}(\mathbf{y}^*),\\
&\tfrac{\partial g_{1,i}}{\partial y_j} = 0 \text{ for all } i \notin \text{Supp}(\mathbf{x}^*), j \in [m],\\
&\tfrac{\partial g_{1,i}}{\partial z_j} = 0 \text{ for all } i, j \in [n],\\
&\tfrac{\partial g_{1,i}}{\partial w_j} = x_i^*(-\eta A_{ij} + \eta v) \text{ for all } i \in \text{Supp}(\mathbf{x}^*), j \in \text{Supp}(\mathbf{y}^*),\\
&\tfrac{\partial g_{1,i}}{\partial w_j} = 0 \text{ for all } i \notin \text{Supp}(\mathbf{x}^*), j \in [m],\\
&\tfrac{\partial g_{2,i}}{\partial y_i} = 1 - y_i^* \text{ for all } i \in \text{Supp}(\mathbf{y}^*),\\
&\tfrac{\partial g_{2,i}}{\partial y_i} = \tfrac{e^{-\eta(A\mathbf{x}^*)_i}}{e^{-\eta v}} \text{ for all } i \notin \text{Supp}(\mathbf{y}^*),\\
&\tfrac{\partial g_{2,i}}{\partial y_j} = -y_i^* \text{ for all } i, j \in \text{Supp}(\mathbf{y}^*) \text{ and } j \neq i,\\
&\tfrac{\partial g_{2,i}}{\partial y_j} = 0 \text{ for all } i \notin \text{Supp}(\mathbf{y}^*), j \in [m] \text{ and } j \neq i,\\
&\tfrac{\partial g_{2,i}}{\partial x_j} = y_i^*(-2\eta A_{ij}^\top + 2\eta v) \text{ for all } i \in \text{Supp}(\mathbf{y}^*), j \in \text{Supp}(\mathbf{x}^*),\\
&\tfrac{\partial g_{2,i}}{\partial x_j} = 0 \text{ for all } i \notin \text{Supp}(\mathbf{y}^*), j \in [n],\\
&\tfrac{\partial g_{2,i}}{\partial z_j} = y_i^*(\eta A_{ij}^\top - \eta v) \text{ for all } i \in \text{Supp}(\mathbf{y}^*), j \in \text{Supp}(\mathbf{x}^*),\\
&\tfrac{\partial g_{2,i}}{\partial z_j} = 0 \text{ for all } i \notin \text{Supp}(\mathbf{y}^*), j \in [n],\\
&\tfrac{\partial g_{2,i}}{\partial w_j} = 0 \text{ for any } i, j \in [m],\\
&\tfrac{\partial g_{3,i}}{\partial x_i} = 1 \text{ for all } i \in [n] \text{ and zero all the other partial derivatives of } g_{3,i},\\
&\tfrac{\partial g_{4,i}}{\partial y_i} = 1 \text{ for all } i \in [m] \text{ and zero all the other partial derivatives of } g_{4,i}.
\end{aligned}
\tag{14}
$$

B Missing claims and proofs

Lemma B.1 shows that the change between next and current iterate in both OMWU algorithms (classic and linear variant) is of order $O(\eta)$ and that the difference between the next iterate of both algorithms is $O(\eta^2)$.

Lemma B.1. *Let $\mathbf{x} \in \Delta_n$ be the vector of the max player, $\mathbf{w}, \mathbf{z} \in \Delta_m$ and suppose $\mathbf{x}', \mathbf{x}''$ are the next iterates of OMWU and its linear variant with current vector $\mathbf{x}$ and vectors $\mathbf{w}, \mathbf{z}$ of the min player. It holds that*

$$\left\|\mathbf{x}' - \mathbf{x}''\right\|_1 \text{ is } O(\eta^2), \text{and } \left\|\mathbf{x}' - \mathbf{x}\right\|_1, \left\|\mathbf{x}'' - \mathbf{x}\right\|_1 \text{ are } O(\eta).$$

Analogously, it holds for vector $\mathbf{y} \in \Delta_m$ of the min player and its next iterates.

Proof. Let η be sufficiently small (smaller than maximum in absolute value entry of A).

$$
\begin{aligned}
|x_i' - x_i''| &= x_i \left| \frac{e^{2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i}}{\sum_j x_j e^{2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j}} - \frac{1 + 2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i}{\sum_j x_j(1 + 2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j)} \right| \\
&= x_i \left| \frac{1 + 2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i \pm O(\eta^2)}{\sum_j x_j(1 + 2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j) \pm O(\eta^2)} - \frac{1 + 2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j}{\sum_j x_j(1 + 2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j)} \right| \\
&\quad \text{which is } O(\eta^2)x_i
\end{aligned}
$$

and hence $\|\mathbf{x}' - \mathbf{x}''\|_1$ is $O(\eta^2)$. Moreover we have that

$$
\begin{aligned}
|x_i - x_i''| &= x_i \left| 1 - \frac{1 + 2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i}{\sum_j x_j(1 + 2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i)} \right| \\
&= x_i \left| \frac{\sum_j x_j(1 + 2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j) - (1 + 2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i)}{\sum_j x_j(1 + 2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i)} \right| \\
&= x_i \left| \frac{\sum_j x_j(2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j) - 2\eta(A\mathbf{w})_i + \eta(A\mathbf{z})_i}{\sum_j x_j(1 + 2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i)} \right| \text{ which is } O(\eta)x_i.
\end{aligned}
$$

By triangle inequality and the two above proofs we get the third part of the lemma. □

Lemmas B.2, B.3 and B.5 will be used in the proof of Theorem 3.1.

Lemma B.2. *Let $\mathbf{x} \in \Delta_n$, $\mathbf{w}, \mathbf{z} \in \Delta_m$ and suppose $\mathbf{x}', \mathbf{x}''$ are the next iterates of OMWU and its linear variant with current vector $\mathbf{x}$ and inputs $\mathbf{w}, \mathbf{z}$, i.e., $\mathbf{x}'$ has coordinates $x_i' = x_i \frac{e^{2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i}}{\sum_j x_j e^{2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j}}$ and $\mathbf{x}''$ has coordinates $x_i'' = x_i \frac{1 + 2\eta(A\mathbf{w})_i - \eta(A\mathbf{z})_i}{\sum_j x_j(1 + 2\eta(A\mathbf{w})_j - \eta(A\mathbf{z})_j)}$. It holds that (for η sufficiently small)*

$$
\begin{aligned}
&\eta \mathbf{x}'^\top A(2\mathbf{w} - \mathbf{z}) - \eta \mathbf{x}^\top A(2\mathbf{w} - \mathbf{z}) = \\
&= \left(\eta \mathbf{x}''^\top A(2\mathbf{w} - \mathbf{z}) - \eta \mathbf{x}^\top A(2\mathbf{w} - \mathbf{z})\right) - O(\eta^3) = \\
&= (1 - O(\eta))\eta^2 \sum_i x_i(2\mathbf{x}^\top A\mathbf{w} - \mathbf{x}^\top A\mathbf{z} - 2(A\mathbf{w})_i + (A\mathbf{z})_i)^2 - O(\eta^3) \\
&= (1 - O(\eta))\eta^2 \sum_i x_i'(2\mathbf{x}'^\top A\mathbf{w} - \mathbf{x}'^\top A\mathbf{z} - 2(A\mathbf{w})_i + (A\mathbf{z})_i)^2 - O(\eta^3).
\end{aligned}
$$

Proof. It suffices to prove the second equality. The rest follow from Lemma B.1. Set $B = (\mathbf{1}_n \mathbf{1}_m^\top + \eta A)$. We have that $x_i'' = x_i \frac{(B(2\mathbf{w} - \mathbf{z}))_i}{\mathbf{x}^\top B(2\mathbf{w} - \mathbf{z})}$ (from definition of linear variant of OMWU dynamics). It follows that

$$
\begin{aligned}
&(\mathbf{x}''^\top B(2\mathbf{w} - \mathbf{z})) \cdot (x^\top B(2\mathbf{w} - \mathbf{z})) = \sum_{ij} B_{ij} x_i''(2\mathbf{w} - \mathbf{z})_j \cdot \left(\mathbf{x}^\top B(2\mathbf{w} - \mathbf{z})\right) \\
&= \sum_{ij} B_{ij} \left(x_i \frac{(B(2\mathbf{w} - \mathbf{z}))_i}{\mathbf{x}^\top B(2\mathbf{w} - \mathbf{z})} \right) (2\mathbf{w} - \mathbf{z})_j \cdot \left(x^\top B(2\mathbf{w} - \mathbf{z})\right) \\
&= \sum_{ij} B_{ij} \left(x_i (B(2\mathbf{w} - \mathbf{z})_i) \right) (2\mathbf{w} - \mathbf{z})_j \\
&= \sum_i x_i (B(2\mathbf{w} - \mathbf{z}))_i^2 \\
&= (\mathbf{x}^\top B(2\mathbf{w} - \mathbf{z}))^2 + \sum_i x_i (\mathbf{x}^\top B(2\mathbf{w} - \mathbf{z}) - (B(2\mathbf{w} - \mathbf{z}))_i)^2.
\end{aligned}
$$

where the last inequality comes from the fact that for a random variable ξ we have $\mathbb{E}[\xi^2] = \mathbb{E}^2[\xi] + \mathbb{V}[\xi]$. Therefore by diving LHS and RHS by $\mathbf{x}^\top B(2\mathbf{w} - \mathbf{z})$ which is $1 \pm O(\eta)$ we get

$$\begin{aligned}(\mathbf{x}''^\top B(2\mathbf{w} - \mathbf{z})) &= (\mathbf{x}^\top B(2\mathbf{w} - \mathbf{z})) + (1 - O(\eta)) \sum_i x_i (\mathbf{x}^\top B(2\mathbf{w} - \mathbf{z}) - (B(2\mathbf{w} - \mathbf{z}))_i)^2 \\ &= (\mathbf{x}^\top (\mathbf{1}_n \mathbf{1}_m^\top + \eta A)(2\mathbf{w} - \mathbf{z})) + \\ &\quad + (1 - O(\eta))\eta^2 \sum_i x_i (\mathbf{x}^\top A(2\mathbf{w} - \mathbf{z}) - (A(2\mathbf{w} - \mathbf{z}))_i)^2.\end{aligned}$$

The proof is complete by Lemma B.1. □

Using same arguments as in proof of Lemma B.2 we have the following lemma:

Lemma B.3. *Let $\mathbf{y} \in \Delta_m$, $\mathbf{w}, \mathbf{z} \in \Delta_n$ and suppose $\mathbf{y}'$ is the next iterate of OMWU with current vector $\mathbf{y}$ and inputs $\mathbf{w}, \mathbf{z}$, i.e., $\mathbf{y}'$ has coordinates $y'_i = y_i \frac{e^{-2\eta(A^\top \mathbf{w})_i + \eta(A^\top \mathbf{z})_i}}{\sum_j y_j e^{-2\eta(A^\top \mathbf{w})_j + \eta(A^\top \mathbf{z})_j}}$. It holds that (for η sufficiently small)*

$$\begin{aligned}&\eta \mathbf{y}'^\top A^\top (\mathbf{z} - 2\mathbf{w}) - \eta \mathbf{y}^\top A^\top (\mathbf{z} - 2\mathbf{w}) \\ &= (1 - O(\eta))\eta^2 \sum_i y'_i (\mathbf{y}'^\top A^\top \mathbf{z} - 2\mathbf{y}'^\top A^\top \mathbf{w} - (A^\top \mathbf{z})_i + 2(A^\top \mathbf{w})_i)^2 - O(\eta^3).\end{aligned}$$

Lemma B.4. *Let $(\mathbf{x}^t, \mathbf{y}^t)$ be the t-th iterate of OMWU dynamics (5). For each time step $t \geq 2$ it holds that*

$$\begin{aligned}&\eta \mathbf{x}^{t-1\ \top} A \mathbf{y}^t - \eta \mathbf{x}^{t\ \top} A \mathbf{y}^{t-1} \leq \\ &\leq -(1 - O(\eta))\eta^2 \sum_i x_i^t (2\mathbf{x}^{t\top} A \mathbf{y}^t - \mathbf{x}^{t\top} A \mathbf{y}^{t-1} - 2(A\mathbf{y}^t)_i + (A\mathbf{y}^{t-1})_i)^2 - \\ &- (1 - O(\eta))\eta^2 \sum_i y_i^t (\mathbf{y}^{t\top} A^\top \mathbf{x}^{t-1} - 2\mathbf{y}^{t\top} A^\top \mathbf{x}^t - (A^\top \mathbf{x}^{t-1})_i + 2(A^\top \mathbf{x}^t)_i)^2 + O(\eta^3).\end{aligned}$$

Proof.

$$\begin{aligned}\eta \mathbf{x}^{t-1\ \top} A \mathbf{y}^t - \eta \mathbf{x}^{t\ \top} A \mathbf{y}^{t-1} &= \eta \mathbf{x}^{t-1\ \top} A \mathbf{y}^t - \frac{1}{2}\eta \mathbf{x}^{t-1\ \top} A \mathbf{y}^{t-1} + \frac{1}{2}\eta \mathbf{x}^{t-1\ \top} A \mathbf{y}^{t-1} - \eta \mathbf{x}^{t\ \top} A \mathbf{y}^{t-1} \\ &= \eta \mathbf{x}^{t\ \top} A \mathbf{y}^t - \frac{1}{2}\eta \mathbf{x}^{t\ \top} A \mathbf{y}^{t-1} + \frac{1}{2}\eta \mathbf{x}^{t-1\ \top} A \mathbf{y}^t - \eta \mathbf{x}^{t\ \top} A \mathbf{y}^t - \\ &- (\frac{1}{2} - O(\eta))\eta^2 \sum_i x_i^t (2\mathbf{x}^{t\top} A \mathbf{y}^t - \mathbf{x}^{t\top} A \mathbf{y}^{t-1} - 2(A\mathbf{y}^t)_i + (A\mathbf{y}^{t-1})_i)^2 - \\ &- (\frac{1}{2} - O(\eta))\eta^2 \sum_i y_i^t (\mathbf{y}^{t\top} A^\top \mathbf{x}^{t-1} - 2\mathbf{y}^{t\top} A^\top \mathbf{x}^t - (A^\top \mathbf{x}^{t-1})_i + \\ &+ 2(A^\top \mathbf{x}^t)_i)^2 + O(\eta^3).\end{aligned}$$

The second equality comes from Lemmas B.2 and B.3. By canceling out the common terms and bring to the LHS the appropriate remaining terms, the claim follows. □

Lemma B.5. *Let $(\mathbf{x}^t, \mathbf{y}^t)$ denote the t-th iterate of OMWU dynamics. It holds for $t \geq 2$ that*

$$\mathbf{x}^{*\ \top} A(2\mathbf{y}^t - \mathbf{y}^{t-1}) \geq \mathbf{x}^{*\ \top} A \mathbf{y}^* \text{ and } (2\mathbf{x}^{t\ \top} - \mathbf{x}^{t-1\ \top}) A \mathbf{y}^* \leq \mathbf{x}^{*\ \top} A \mathbf{y}^*,$$

where $(\mathbf{x}^*, \mathbf{y}^*)$ is the optimal solution of the min-max problem.

Proof. It is true that $x_i^t \geq (1 - O(\eta))x_i^{t-1}$, hence $x_i^t \geq \frac{1}{2}x_i^{t-1}$ for η sufficiently small. Therefore $2\mathbf{x}^t - \mathbf{x}^{t-1}$ lies in the simplex Δ_n. Hence since $(\mathbf{x}^*, \mathbf{y}^*)$ is the optimum (Nash equilibrium) we get that $(2\mathbf{x}^{t\ \top} - \mathbf{x}^{t-1\ \top})A\mathbf{y}^* \leq \mathbf{x}^{*\ \top}A\mathbf{y}^*$ ($\mathbf{x}$ is the max player). Similarly the second inequality can be proved. □

Proof of Theorem 3.1. We compute the difference between
$D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}))$ and $D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^t, \mathbf{y}^t))$

$$D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^{t+1}, \mathbf{y}^{t+1})) - D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^t, \mathbf{y}^t)) = -\left(\sum_i \mathbf{x}_i^* \ln \frac{x_i^{t+1}}{x_i^t} + \sum_i y_i^* \ln \frac{y_i^{t+1}}{y_i^t}\right)$$

$$= -\left(\sum_i \mathbf{x}_i^* \ln e^{2\eta(A\mathbf{y}^t)_i - \eta(A\mathbf{y}^{t-1})_i} + \sum_i y_i^* \ln e^{-2\eta(A^\top \mathbf{x}^t)_i + \eta(A^\top \mathbf{x}^{t-1})_i}\right) +$$

$$+ \ln\left(\sum_i x_i^t e^{2\eta(A\mathbf{y}^t)_i - \eta(A\mathbf{y}^{t-1})_i}\right) + \ln\left(\sum_i y_i^t e^{-2\eta(A^\top \mathbf{x}^t)_i + \eta(A^\top \mathbf{x}^{t-1})_i}\right)$$

$$= -2\eta\mathbf{x}^{*\ \top}A\mathbf{y}^t + \eta\mathbf{x}^{*\ \top}A\mathbf{y}^{t-1} + 2\eta\mathbf{x}^{t\ \top}A\mathbf{y}^* - \eta\mathbf{x}^{t-1\ \top}A\mathbf{y}^* +$$

$$+ \ln\left(\sum_i x_i^t e^{2\eta(A\mathbf{y}^t)_i - \eta(A\mathbf{y}^{t-1})_i}\right) + \ln\left(\sum_i y_i^t e^{-2\eta(A^\top \mathbf{x}^t)_i + \eta(A^\top \mathbf{x}^{t-1})_i}\right).$$

We use Lemma B.5 and we get that $-2\eta\mathbf{x}^{*\ \top}A\mathbf{y}^t + \eta\mathbf{x}^{*\ \top}A\mathbf{y}^{t-1} + 2\eta\mathbf{x}^{t\ \top}A\mathbf{y}^* - \eta\mathbf{x}^{t-1\ \top}A\mathbf{y}^* \leq 0$, therefore the LHS (difference in the KL divergence) is at most

$$\leq \overbrace{-2\eta\mathbf{x}^{t\ \top}A\mathbf{y}^t + \eta\mathbf{x}^{t\ \top}A\mathbf{y}^{t-1} + 2\eta\mathbf{x}^{t\ \top}A\mathbf{y}^t - \eta\mathbf{x}^{t-1\ \top}A\mathbf{y}^t - \eta\mathbf{x}^{t\ \top}A\mathbf{y}^{t-1} + \eta\mathbf{x}^{t-1\ \top}A\mathbf{y}^t}^{=0} +$$

$$+ \ln\left(\sum_i x_i^t e^{2\eta(A\mathbf{y}^t)_i - \eta(A\mathbf{y}^{t-1})_i}\right) + \ln\left(\sum_i y_i^t e^{-2\eta(A^\top \mathbf{x}^t)_i + \eta(A^\top \mathbf{x}^{t-1})_i}\right)$$

$$= \ln\left(\sum_i x_i^t e^{2\eta((A\mathbf{y}^t)_i - \mathbf{x}^{t\ \top}A\mathbf{y}^t) - \eta((A\mathbf{y}^{t-1})_i - \mathbf{x}^{t\ \top}A\mathbf{y}^{t-1})}\right) +$$

$$+ \ln\left(\sum_i y_i^t e^{-2\eta((A^\top \mathbf{x}^t)_i - \mathbf{x}^{t\ \top}A\mathbf{y}^t) + \eta((A^\top \mathbf{x}^{t-1})_i - \mathbf{x}^{t-1\ \top}A\mathbf{y}^t)}\right) - \eta\mathbf{x}^{t\ \top}A\mathbf{y}^{t-1} + \eta\mathbf{x}^{t-1\ \top}A\mathbf{y}^t$$

We furthermore use second order Taylor approximation (η is sufficiently small) to the function e^x and we get that previous expression is at most

$$\leq \ln(\sum_i x_i^t \left(1 + 2\eta((A\mathbf{y}^t)_i - \mathbf{x}^{t\ \top}A\mathbf{y}^t) - \eta((A\mathbf{y}^{t-1})_i - \mathbf{x}^{t\ \top}A\mathbf{y}^{t-1})\right) +$$

$$+ \sum_i x_i^t \left((\frac{1}{2} + O(\eta))\eta^2 \left(2(A\mathbf{y}^t)_i - 2\mathbf{x}^{t\ \top}A\mathbf{y}^t - (A\mathbf{y}^{t-1})_i + \mathbf{x}^{t\ \top}A\mathbf{y}^{t-1}\right)^2\right))+$$

$$+ \ln(\sum_i y_i^t \left(1 - 2\eta((A^\top \mathbf{x}^t)_i - \mathbf{x}^{t\ \top}A\mathbf{y}^t) + \eta((A^\top \mathbf{x}^{t-1})_i - \mathbf{x}^{t-1\ \top}A\mathbf{y}^t)\right) +$$

$$+ \sum_i y_i^t \left((\frac{1}{2} + O(\eta))\eta^2 \left(2(A^\top \mathbf{x}^t)_i - 2\mathbf{x}^{t\ \top}A\mathbf{y}^t - (A^\top \mathbf{x}^{t-1})_i + \mathbf{x}^{t-1\ \top}A\mathbf{y}^t\right)^2\right))-$$

$$- \eta\mathbf{x}^{t\ \top}A\mathbf{y}^{t-1} + \eta\mathbf{x}^{t-1\ \top}A\mathbf{y}^t$$

Finally, using Taylor approximation on $\log(1+x)$ and Lemma B.4 (last equality) we get the following system:

$$
\begin{aligned}
&\leq \eta \mathbf{x}^{t-1\ \top} A\mathbf{y}^t - \eta \mathbf{x}^{t\ \top} A\mathbf{y}^{t-1} + \\
&+ \sum_i x_i^t \left((\frac{1}{2} + O(\eta))\eta^2 \left(2(A\mathbf{y}^t)_i - 2\mathbf{x}^{t\ \top} A\mathbf{y}^t - (A\mathbf{y}^{t-1})_i + \mathbf{x}^{t\ \top} A\mathbf{y}^{t-1} \right)^2 \right) + \\
&+ \sum_i y_i^t \left((\frac{1}{2} + O(\eta))\eta^2 \left(2(A^\top \mathbf{x}^t)_i - 2\mathbf{x}^{t\ \top} A\mathbf{y}^t - (A^\top \mathbf{x}^{t-1})_i + \mathbf{x}^{t-1\ \top} A\mathbf{y}^t \right)^2 \right) \\
&= - \sum_i x_i^t \left((\frac{1}{2} - O(\eta))\eta^2 \left(2(A\mathbf{y}^t)_i - 2\mathbf{x}^{t\ \top} A\mathbf{y}^t - (A\mathbf{y}^{t-1})_i + \mathbf{x}^{t\ \top} A\mathbf{y}^{t-1} \right)^2 \right) \\
&- \sum_i y_i^t \left((\frac{1}{2} - O(\eta))\eta^2 \left(2(A^\top \mathbf{x}^t)_i - 2\mathbf{x}^{t\ \top} A\mathbf{y}^t - (A^\top \mathbf{x}^{t-1})_i + \mathbf{x}^{t-1\ \top} A\mathbf{y}^t \right)^2 \right) + O(\eta^3).
\end{aligned}
$$

It is clear that as long as $(\mathbf{x}^t, \mathbf{y}^t)$ (and thus $(\mathbf{x}^{t-1}, \mathbf{y}^{t-1})$ by Lemma B.1) is not $O(\eta^{1/3})$-close, from above inequalities/equalities we get

$$D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^{t+1}, \mathbf{y}^{t+1})) - D_{KL}((\mathbf{x}^*, \mathbf{y}^*)||(\mathbf{x}^t, \mathbf{y}^t)) \leq -\Omega(\eta^3),$$

meaning that KL divergence decreases by at least a factor of η^3 and the claim follows. □

Lemma B.6. *Let D be a real diagonal matrix with positive diagonal entries and S be a real skew-symmetric matrix ($S^\top = -S$). It holds that SD has eigenvalues with real part zero (i.e., it has only imaginary eigenvalues).*

Proof. Let $\mathbf{z}^*$ be the conjugate transpose of $\mathbf{z}$ and $\mathbf{z}^*$ be a left eigenvector of SD with complex eigenvalue λ. It holds that

$$
\begin{aligned}
\lambda \mathbf{z}^* D^{-1} \mathbf{z} &= \mathbf{z}^* S D D^{-1} \mathbf{z} \\
&= \mathbf{z}^* S \mathbf{z} \\
&= -(\mathbf{z}^* S \mathbf{z})^* \text{ (since } S \text{ is skew symmetric)} \\
&= -(\lambda \mathbf{z}^* D^{-1} \mathbf{z})^* \text{ (using first and second equalities above)} \\
&= -\overline{\lambda} \mathbf{z}^* D^{-1} \mathbf{z}.
\end{aligned}
$$

Since D has positive diagonal entries, we conclude that $\mathbf{z}^* D^{-1} \mathbf{z} \neq 0$ (since $\mathbf{z} \neq \mathbf{0}$), thus $\lambda = -\overline{\lambda}$ and the claim follows. □