MULTIPLICATIVE WEIGHTS UPDATES AS A DISTRIBUTED CONSTRAINED OPTIMIZATION ALGORITHM: CONVERGENCE TO SECOND-ORDER STATIONARY POINTS ALMOST ALWAYS

IOANNIS PANAGEAS, GEORGIOS PILIOURAS, AND XIAO WANG
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Abstract. Non-concave maximization has been the subject of much recent study in the optimization and machine learning communities, specifically in deep learning. Recent papers ((Ge et al. 2015, Lee et al. 2017) and references therein) indicate that first order methods work well and avoid saddles points. Results as in (Lee et al. 2017), however, are limited to the unconstrained case or for cases where the critical points are in the interior of the feasibility set, which fail to capture some of the most interesting applications. In this paper we focus on constrained non-concave maximization. We analyze a variant of a well-established algorithm in machine learning called Multiplicative Weights Update (MWU) for the maximization problem $\max_{x \in D} P(x)$, where $P$ is non-concave, twice continuously differentiable and $D$ is a product of simplices. We show that MWU converges almost always for small enough stepsizes to critical points that satisfy the second order KKT conditions. We combine techniques from dynamical systems as well as taking advantage of a recent connection between Baum Eagon inequality and MWU (Palaiopanos et al. 2017).

1. Introduction

The interplay between the structure of saddle points and the performance of first order algorithms is a critical aspect of non-concave maximization. In the unconstrained setting, there have been many recent results indicating that gradient descent (GD) avoids strict saddle points with random initialization (Lee et al. 2017). Moreover by adding noise, it is guaranteed that GD converges to a local maximum in polynomial time (Ge et al. 2015, Jin et al. 2017 and references therein). By adding a non-smooth function in the objective (e.g., the indicator function of a convex set) it can be shown that there are stochastic first order methods that converge to a local minimum point in the constrained case (see Zhu 2017, Zhu 2018) under the assumption of oracle access to the stochastic (sub)gradients. What is less understood is the problem of convergence to second order stationary points in constrained optimization (under the weaker assumption that we do not have access to the subgradient of the indicator of the feasibility set; in other words when projection to the feasibility set is not a trivial task). In the case of constrained optimization, we also note that the techniques of (Lee et al. 2017) are not applicable in a straightforward way.

Non-concave maximization problems with saddle points/local optima on the boundary are very common. For example in game theory, it is typical for a Nash equilibrium to not have full support (and thus to lie on the boundary of the simplex). In such cases, one natural approach is to use projected gradient descent, but computing the projection at every iteration might not be an easy task to accomplish. Several distributed, concurrent optimization techniques have been studied in such settings (Kleinberg et al. 2009, Ackermann et al. 2009), however they are known to work only for very specific type of optimization problems, i.e., multilinear potential functions. Moreover, having saddle points/local optima on the boundary of a closed set that has (Lebesgue) measure zero compared to the full domain (e.g., simplex with $n$ variables has measure zero in $\mathbb{R}^n$) makes impossible to use as a black box the result in (Lee et al. 2017) in which they make use of well-known center-stable manifold theorem from the dynamical systems literature (see Theorem 5.1).
In this paper we focus on solving problems of the form
\[
\max_{x \in D} P(x),
\]
(1.1)
where \( P \) is a non-concave, twice continuously differentiable function and \( D \) is some compact set, which will be a product of simplices for our purposes, i.e., \( D = \{(x_{ij})|x_{ij} \geq 0, \sum_{j=1}^{M} x_{ij} = 1 \text{ for all } 1 \leq i \leq N\} \), where \( N, M \) are natural numbers. As a result, vector \( x \) can be also interpreted as a collection of \( N \) probability distributions (having \( N \) players), where each distribution \( x_i \) has support of size \( M \) (strategies). For this particular problem (1.1), one natural algorithm that is commonly used is the Baum-Eagon dynamics (1.2) (see the seminal paper by Baum and Eagon (Baum and Eagon 1967) with many applications to inference problems, Hidden Markov Models (HMM) in particular (see discussion in Section 4).

\[
x^{t+1}_{ij} = x^{t}_{ij} \frac{\partial P}{\partial x_{ij}} \bigg|_{x^t},
\]
(1.2)

The denominator of the above fraction is for renormalization purposes. It is clear that as long as \( x^t \in D \) then \( x^{t+1} \in D \).

Despite its power, Baum-Eagon dynamics has its limitations. First and foremost, the Baum-Eagon dynamics is not always well-defined; the denominator term \( \sum_s x^{t}_{is} \frac{\partial P}{\partial x_{is}} \bigg|_{x^t} \) must be non-zero at all times and moreover the fraction in equations (1.2) should always be non-negative. This provides a restriction to the class of functions \( P \) to which the Baum-Eagon dynamics can be applied. Moreover, it turns out that the update rule of the Baum-Eagon dynamics is not always a diffeomorphism\(^1\). In fact, as we show even in simple setting (see section 2.3) the Baum-Eagon dynamic may not be even a homeomorphism or one-to-one. This counterexample disproves a conjecture by Stebe (Stebe 1972). Since the map is not even a local diffeomorphism one cannot hope to leverage the power of center-stable manifold theorem to argue convergence towards local maxima.

To counter this, in this paper we focus on multiplicative weights update algorithm (MWU) (Arora et al. 2012) which can be interpreted as an instance of Baum-Eagon dynamics in the presence of learning rates. Introducing learning rates gives us a lot of flexibility and will allow us to formally prove strong convergence properties which would be impossible without this adaptation. Assume that \( x^t \) is the \( t \)-th iterate of MWU, the equations of which can be described as follows:

\[
x^{t+1}_{ij} = x^{t}_{ij} \frac{1 + \epsilon_i \frac{\partial P}{\partial x_{ij}} \bigg|_{x^t}}{1 + \epsilon_i \sum_s x^{t}_{is} \frac{\partial P}{\partial x_{is}} \bigg|_{x^t}},
\]
(1.3)
where \( \epsilon_i \) the stepsizes (learning rate) of the dynamics. Intuitively (in game theory terms), for strategy profile (vector) \( \tilde{x} := (\tilde{x}_1, \ldots, \tilde{x}_N) \), each player \( i \) that chooses strategy \( j \) has utility to be \( \frac{\partial P}{\partial x_{ij}} \bigg|_{x=\tilde{x}} \). We call a strategy profile \( y \in D \) a fixed point if it is invariant under the update rule dynamics (1.3). It is also clear that the set \( D \) is invariant under the dynamics in the sense that if \( x^t \in D \) then \( x^{t+1} \in D \) for \( t \in \mathbb{N} \). This last observation indicates that MWU has the projection step for free (compared to projected gradient descent). We would also like to note that MWU can be computed in a distributed manner and this makes the algorithm more important for Machine Learning applications.

Statement of our results. We will need the following two definitions (well-known in optimization literature, as applied to simplex constraints):

\(\footnote{A function is called a diffeomorphism if it is differentiable and a bijection and its inverse is differentiable as well.}
Definition 1.1 (Stationary point). $\mathbf{x}^*$ is called a stationary point as long as it satisfies the first order KKT conditions for the problem (1.1). Formally, it holds

$$
\begin{align*}
\mathbf{x}^* &\in D \\
x_{ij}^* > 0 &\Rightarrow \frac{\partial P}{\partial x_{ij}}(\mathbf{x}^*) = \sum_{j'} x_{ij'}^* \frac{\partial P}{\partial x_{ij'}}(\mathbf{x}^*) \\
x_{ij}^* = 0 &\Rightarrow \frac{\partial P}{\partial x_{ij}}(\mathbf{x}^*) \leq \sum_{j'} x_{ij'}^* \frac{\partial P}{\partial x_{ij'}}(\mathbf{x}^*). 
\end{align*}
$$

(1.4)

The stationary point is called strict if the last inequalities hold strictly.

Definition 1.2 (Second order stationary point). $\mathbf{x}^*$ is called a second order stationary point as long as it is a stationary point and moreover it holds that:

$$
y^\top \nabla^2 P(\mathbf{x}^*) y \leq 0.
$$

(1.5)

for all $\mathbf{y}$ such that $\sum_{j=1}^M y_{ij} = 0$ (for all $1 \leq i \leq N$) and $y_{ij} = 0$ whenever $x_{ij}^* = 0$, i.e., it satisfies the second order KKT conditions.

Our main result are stated below:

Theorem 1.3 (Avoid non-stationary). Assume that $P$ is twice continuously differentiable in a set containing $D$. There exists small enough fixed stepsizes $\epsilon_i$ such that the set of initial conditions $\mathbf{x}^0$ of which the MWU dynamics (1.3) converges to fixed points that violate second order KKT conditions is of (Lebesgue) measure zero.

The following corollary is immediate from Theorem 1.3 and the Baum-Eagon inequality for rational functions (see Section 2).

Corollary 1.4. Assume $\mu$ is a measure that is absolutely continuous with respect to the Lebesgue measure and $P$ is a rational function (fraction of polynomials) that is twice continuously differentiable in a set containing $D$. It follows that with probability one (randomness induced by $\mu$), MWU dynamics converges to second order stationary points.

Remark 1.5. It is obvious that when the learning rates $\epsilon_i = 0$, MWU (1.3) is trivially the identity map. On the other hand, whenever the dynamic is well defined in the limit $\epsilon \to \infty$ (i.e. when $P$ is sufficiently well behaved, e.g. a polynomial with positive coefficients) this corresponds to the well known class of Baum-Eagon maps (Stebe 1972).

Our techniques. The first step of the proof given in Section 3 is to prove that MWU converges to fixed points for all rational functions and any possible set of learning rates (as long as the dynamic is well defined). The proof of this statement leverages recently discovered connections between MWU and the Baum-Eagon dynamic (Palaiopanos et al. 2017). However, this does not even allow us to exclude very suboptimal fixed points (i.e. saddle points or even local minima) from having a positive region of attraction.

The other two steps of the proof work on weeding out the "bad" stationary points and showing that the set of initial conditions that converge to them is of measure zero. The key tool for proving that type of statements is the center-stable manifold theorem (Lee et al. 2017). However, in order to leverage the power of the theorem we first show in Theorem 2.3 that for small enough learning rates MWU is a diffeomorphism. The second and third step of the proof respectively is to show that fixed points that do not satisfy the first (resp. second) order stationary point conditions are unstable under MWU.

Even for the first step of the proof (lemma 3.1), we have to use ad-hoc techniques to deal with problems due to the constraints. Specifically, we start by projecting the domain $D$ to a subspace that is full dimensional (for example simplex of size $n$ is mapped to the Euclidean subspace of dimension $n - 1$). Next, we show that non-first order stationary points result to fixed points where the Jacobian of MWU has eigenvalue larger than 1. Proving a similar statement for the fixed points...
that correspond to non-second order stationary fixed points (lemma 3.2) is the most technical part of the proof as we have to deal with the asymmetry of the resulting Jacobian. Nevertheless we manage to do so by using Sylvester’s law of inertia and exploiting newly discovered decompositions for this class of matrices. Putting everything together results in our main theorem (Theorem 1.3).

2. Optimization with Baum-Eagon algorithm

In this section, we state the important result of Baum and Eagon providing a method to increase the value of a polynomial with nonnegative coefficients and (later generalized for) rational functions with nonzero denominators. The update rule defined by (2.1) increases the value of the polynomial \( P \) if the initial point is not a fixed point of Baum-Eagon dynamics.

2.1. Baum-Eagon map. Let \( P \) be a polynomial with real positive coefficients and variables \( x_{ij}, i = 1, \ldots, k, j = 1, \ldots, n_i \). Let \( n = \sum_i n_i \). Let \( D \) be the product of simplexes. Define \( T(x) \) as the vector in \( D \) with component \( x_{ij}' \) given by

\[
x_{ij}' = \frac{x_{ij} \frac{\partial P}{\partial x_{ij}}}{\sum_{h=1}^{n_i} x_{ih} \frac{\partial P}{\partial x_{ih}}}. \tag{2.1}
\]

**Theorem 2.1** (Baum-Eagon). Let \( P(\{x_{ij}\}) \) be a polynomial with non-negative coefficients homogeneous of degree \( d \) in its variables \( \{x_{ij}\} \). Let \( x = \{x_{ij}\} \) be any point of the domain \( D = \{x_{ij} \geq 0, \sum_{j=1}^{n_i} x_{ij} = 1, i = 1, 2, \ldots, k, j = 1, 2, \ldots, n_i \} \). For \( x = \{x_{ij}\} \in D \) let \( T(x) = T(\{x_{ij}\}) \) be the point of \( D \) whose \( i, j \) coordinate is

\[
x_{ij}' = \frac{x_{ij} \frac{\partial P}{\partial x_{ij}}}{\sum_{h=1}^{n_i} x_{ih} \frac{\partial P}{\partial x_{ih}}}. \tag{2.2}
\]

Then \( P(T(x)) > P(x) \) unless \( T(x) = x \).

2.2. Optimization for rational functions. According to (Gopalakrishnan et al. 1991), one can define a Baum-Eagon dynamics for rational functions \( R(x) = \frac{S_1(x)}{S_2(x)} \) with positive denominator so that the update rule of the Baum-Eagon dynamics increases the value of the rational function unless the initial point is a fixed point. This can be done by starting with the Baum-Eagon map of the following polynomial:

\[
Q_x(X) = P_x(X) + C_x(X),
\]

where

\[
P_x(X) = S_1(X) - R(x)S_2(X)
\]

and

\[
C_x(X) = N_x(\sum_{ij} x_{ij} + 1)^d,
\]

where \( d \) is the degree of \( P_x(X) \) and \( N_x \) is a constant such that \( P_x(X) + C_x(X) \) only has nonnegative coefficients.

It is proved in (Gopalakrishnan et al. 1991) that the value of \( R(X) \) increases along the Baum-Eagon dynamics induced by polynomial \( Q_x \).
2.3. Bad example on Baum-Eagon dynamics. L. Baum has an unpublished result (Stebe 1972) claiming that the Baum-Eagon map $T$ is a homeomorphism of $D$ onto itself if and only if the expression for $P$ as a sum of distinct monomials with positive coefficients monomials $c_{i,j}x_iw_{i,j}$ for all $i = 1, ..., k, j = 1, ..., n_i$ where $c_{i,j} > 0$ and $w_{i,j}$ is an integer greater than zero. But this condition is incorrect and we give a counter example. This indicates that the Baum-Eagon dynamics does not satisfy the nice property of being a diffeomorphism. For a special case, we focus on the map $\tau$ defined on a single simplex

$$\Delta_{n-1} = \{(x_1, ..., x_n) | \sum_{i=1}^{n} x_i = 1\},$$

and $\tau$ can be written as

$$x_i' = \frac{x_i \frac{\partial P}{\partial x_i}}{\sum x_i \frac{\partial P}{\partial x_i}} \quad (2.3)$$

The map defined by equation 2.3 can be considered a composition of $\tau_1$ and $\tau_2$ defined in the following way:

$$\tau_1 : (x_1, ..., x_2) \mapsto (x_1 \frac{\partial P}{\partial x_1}, ..., x_2 \frac{\partial P}{\partial x_2}) \quad (2.4)$$

$$\tau_2 : (x_1 \frac{\partial P}{\partial x_1}, ..., x_2 \frac{\partial P}{\partial x_2}) \mapsto \frac{1}{\sum x_i \frac{\partial P}{\partial x_i}} (x_1 \frac{\partial P}{\partial x_1}, ..., x_2 \frac{\partial P}{\partial x_2}) \quad (2.5)$$

Consider 1-dimensional simplex as an example, $\tau_1$ maps the simplex $\Delta_1$ to a curve and $\tau_2$ maps points on the curve back to $\Delta_1$ by scaling. From figure 1a we notice that a necessary condition for $\tau$ to be a homeomorphism is that boundaries of the curve given by $\tau_1$ are not equal to $(0, 0)$. A necessary condition for $\tau$ to be a homeomorphism is that $\tau$ must be one to one. In 1-dimensional case, the ratio

$$k = \frac{x_1 \frac{\partial P}{\partial x_1}}{x_2 \frac{\partial P}{\partial x_2}}$$

must be monotone with respect to $x_1$. While the following example gives a polynomial satisfying Baum’s condition but $k$ is not monotone with respect to $x_1$.  

\footnote{A function is called a homeomorphism if it is continuous and a bijection and its inverse is continuous as well. Thus if a function is not a homeomorphism, then it is not a diffeomorphism.}

Figure 1. Illustration
Example 2.2. Suppose $P = x_1 + x_1^7 x_2 + x_2^7$, then
\[
\begin{align*}
x_1 \frac{\partial P}{\partial x_1} &= x_1 + 7x_1^7 x_2 \\
x_2 \frac{\partial P}{\partial x_2} &= x_1^7 x_2 + 7x_2^7
\end{align*}
\]
As it is shown in Figure 2, the ratio $k = x_1 \frac{\partial P}{\partial x_1} / x_2 \frac{\partial P}{\partial x_2}$ is not monotone with respect to $x_1$. So the Baum-Eagon map is not one to one implying that it is not a homeomorphism.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure2.png}
\caption{Non-monotonicity of $k(x_1)$}
\end{figure}

2.4. Baum-Eagon map of $\sum_{i,j} x_{ij} + \epsilon P$. Let $P$ be a twice continuously differentiable function on the product of simplexes. The update rule of the dynamics for $P$ is a diffeomorphism for $\epsilon$ sufficiently small. This is what next theorem captures.

**Theorem 2.3.** For any twice continuously differentiable function $P$, there exists a positive number $\delta(P)$ depending on $P$, such that for any $\epsilon < \delta(P)$, the Baum-Eagon map applied to $Q = \sum_{ij} x_{ij} + \epsilon P$ is a diffeomorphism.

**Proof.** Firstly, we prove that B-E map of $Q$ is a local diffeomorphism. For a fixed $i$, denote
\[
g_{ij} = \frac{x_{ij} + \epsilon x_{ij} \frac{\partial P}{\partial x_{ij}}}{\sum_j x_{ij} + \epsilon \sum_j x_{ij} \frac{\partial P}{\partial x_{ij}}}.
\]

Since the roots of characteristic polynomial of Jacobian vary continuously as a function of coefficients (see Theorem VI.1.2 in (Bhatia 1997)), let $J_i$ be the Jacobian of the Baum-Eagon dynamics induced by function $Q = \sum_{ij} x_{ij} + \epsilon P$ (this coincides with the MWU dynamics for function $P$ with same stepsize $\epsilon$ (i.e., same learning rates)). The determinant $|J_i|$ at each point is continuous with respect to $\epsilon$. Since as $\epsilon \to 0$, $|J_i| \to 1$ at each point $p$, for each point $p \in D$, there exists $\epsilon_p$ such that for all $\epsilon < \epsilon_p$, $|J_i(p)| > 1/2$. Since determinant is continuous with respect to points in $D$, for $\epsilon_p$, there is a neighborhood of $p$, denoted as $U(p, \epsilon_p)$, such that for all $x \in U(p, \epsilon_p)$, $|J_i(x)| > 1/2$. Thus we have obtained an open cover of $D$, which is $\bigcup_{p \in D} U(p, \epsilon_p)$. Since $D$ is compact, there is a finite subcover of $\bigcup_{p \in D} U(p, \epsilon_p)$, denoted as $\bigcup_{i=1}^n U(p_i, \epsilon_{p_i})$. Then the minimum of $\{\epsilon_{p_i}\}$ gives the $\delta(P)$ in the lemma.

To prove that the B-E map $T$ of $Q$ is a global diffeomorphism, one needs Theorem 2 in (Ho 1975). Since $T$ is proper (preimage of compact set is compact) and $D$ is simply connected and path connected, we conclude that $T$ is a homeomorphism on $D$. 
\[\Box\]
Remark 2.4. The above theorem essentially can be generalized for different stepsizes (learning rates) $\epsilon$ for each player. The idea is that we should apply the same techniques on the function $\sum_{i=1}^{N} \frac{1}{\epsilon_i} \sum_{j=1}^{M} x_{ij} + P$.

3. Convergence Analysis of MWU for arbitrary functions

In this section we provide the proof of Theorem 1.3. As has already been proven in previous section (Theorem 2.3), the update rule of the MWU dynamics is a diffeomorphism for appropriately small enough learning rates. Following the general framework of (Lee et al. 17), we will also make use of the Center stable manifold theorem (Theorem 5.1). The challenging part technically in this paper is to prove that every stationary point $x$ that is not a local maximum has the property that the Jacobian of the MWU dynamics computed at $x$ has a repelling direction (eigenvector).

3.1. Equations of the Jacobian at a fixed point and projection. We focus on multiplicative weights updates algorithm. Assume that $x^t$ is the $t$-th iterate of MWU. Recall the equations:

$$x_{ij}^{t+1} = x_{ij}^t + \frac{1 + \epsilon_i \frac{\partial P}{\partial x_{ij}^t}}{1 + \epsilon_i \sum_{s} x_{is}^t \frac{\partial P}{\partial x_{is}^t}} x_{ij}^t$$

where $\epsilon_i$ the stepsize of the dynamics.

Let $T: D \rightarrow D$ be the update rule of the MWU dynamics (3.1). Fix indexes $i, i'$ for players and $j, s$ for strategies. Set $S_i = 1 + \epsilon_i \sum_{j' \neq s} x_{ij'}^t \frac{\partial P}{\partial x_{ij'}^t}$. The equations of the Jacobian look as follows:

$$\frac{\partial T_{ij}}{\partial x_{ij}} = \frac{1 + \epsilon_i \frac{\partial P}{\partial x_{ij}^t}}{1 + \epsilon_i \sum_{s} x_{is}^t \frac{\partial P}{\partial x_{is}^t}}$$

$$+ \frac{x_{ij}^t}{S_i^2} \left( \epsilon_i \frac{\partial^2 P}{\partial x_{ij} \partial x_{is}^t} - \epsilon_i (1 + \epsilon) \frac{\partial P}{\partial x_{is}^t} \right)$$

$$\cdot \left( \frac{\partial P}{\partial x_{is}^t} + x_{is}^t \frac{\partial^2 P}{\partial x_{is}^2} + x_{ij}^t \frac{\partial P}{\partial x_{ij}^t} \right)$$

for all $i \in [N], j \in [M]$.

$$\frac{\partial T_{ij}}{\partial x_{is}^t} = \frac{x_{is}^t}{S_i^2} \left( \epsilon_i \frac{\partial^2 P}{\partial x_{ij} \partial x_{is}^t} \cdot S_i 

- \epsilon_i (1 + \epsilon) \frac{\partial P}{\partial x_{is}^t} \right)$$

$$\cdot \left( \frac{\partial P}{\partial x_{is}^t} + x_{is}^t \frac{\partial^2 P}{\partial x_{is}^2} + \sum_{j' \neq s} x_{ij}^t \frac{\partial^2 P}{\partial x_{ij}^t \partial x_{is}^t} \right)$$

for all $i \in [N], j, s \in [M], j \neq s$.

$$\frac{\partial T_{ij}}{\partial x_{i's}^t} = \frac{x_{ij}^t}{S_i^2} \left( \epsilon_i \frac{\partial^2 P}{\partial x_{ij} \partial x_{i's}^t} \cdot S_i 

- \epsilon_i (1 + \epsilon) \frac{\partial P}{\partial x_{ij}^t} \cdot \left( \sum_{j' \neq s} x_{ij}^t \frac{\partial^2 P}{\partial x_{ij}^t \partial x_{i's}^t} \right) \right)$$

for all $i, i' \in [N], j, s \in [M], \text{with } i \neq i'$.

Let $y$ to be a fixed point of MWU dynamics. We define the projected MWU mapping to be the function $T_y$ by removing one variable $j \in [M]$ for each player $i \in [N]$ (i.e, $x_{ij}$) such that $y_{ij} > 0$. 

We also define $D_y$ to be the projection of $D$ in the same way. Now the mapping is $T_y: S \to S$ for $S \subset \mathbb{R}^{NM-M}$ is still a diffeomorphism where $S$ is an open set that contains $D_y$. We define the corresponding Jacobian to be the submatrix by removing rows and columns that correspond to variables $x_{ij}$ that were removed.

### 3.2. Stability and proof of Theorem 1.3

We prove the following important lemma that characterizes (partially) the unstable fixed points (meaning the spectral radius of the Jacobian computed at the fixed point is greater than one) of the MWU dynamics and relates them to the stationary points.

**Lemma 3.1** (Non first order stationary points are unstable). Let $y$ be a fixed point of MWU dynamics that violates the first order KKT conditions (is not a first order stationary point). It holds that the projected Jacobian computed at $y$ (formally now is the projected point $y \in D_y$) has an eigenvalue with absolute value greater than one.

**Proof.** Since $y$ is not a stationary point, there exist $i,j$ and so that $y_{ij} = 0$ but $\frac{\partial P}{\partial x_{ij}}\bigg|_{x=y} > \sum_{j'} y_{ij'} \frac{\partial P}{\partial x_{ij'}}\bigg|_{x=y}$. The projected Jacobian computed at $y$ has the property that for variable $x_{ij}$, the corresponding row has entries zeros, apart from the corresponding diagonal entry that is $\frac{1+\epsilon_i \frac{\partial P}{\partial x_{ij}}}{1+\epsilon_i \sum_{j'} x_{ij'} \frac{\partial P}{\partial x_{ij'}}} > 1$ (from the definition of stationary point). Since the projected Jacobian has as eigenvalue $\frac{1+\epsilon_i \frac{\partial P}{\partial x_{ij}}}{1+\epsilon_i \sum_{j'} x_{ij'} \frac{\partial P}{\partial x_{ij'}}}$ the claim follows. \(\square\)

The following technical lemma gives a full characterization among the unstable fixed points of MWU dynamics and the second order stationary points (local maxima). This lemma is more challenging than the stability analysis in (Lee et al. 2017) due to the fact that we have constraints on simplex.

**Lemma 3.2** (Non second order stationary points are unstable). Let $x^*$ be a fixed point of MWU dynamics that is a stationary point (satisfies first order KKT conditions) and violates the second order KKT conditions (is not a second order stationary point). It holds that the projected Jacobian computed at $x^*$ (formally now is the projected point $x^* \in D_{x^*}$) has an eigenvalue with absolute value greater than one.

**Proof.** Because of Lemma 3.1 we may assume that $x^*$ in the interior of $D$ (all coordinates are positive). Set $S_i = 1 + \epsilon_i \sum_{j=1}^{M} x_{ij'} \frac{\partial P}{\partial x_{ij'}}$ for $i \in [N]$. Set

$$
D_{xs} = \begin{pmatrix}
\frac{\epsilon_1 x^*_{11}}{S_1} & 0 & \cdots & 0 \\
0 & \frac{\epsilon_1 x^*_{1M}}{S_1} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \frac{\epsilon_N x^*_{NM}}{S_N}
\end{pmatrix}
$$
Moreover, where \( \lambda \) is the null space of the null space of \( \nabla^2 P(x^*) \). Moreover by law of inertia of Sylvester, the same holds for the matrix \( I + D_{xs}(I - D_{xx})\nabla^2 P(x^*) = I + D_{xs}\nabla^2 P(x^*) - D_{xs}D_{xx}\nabla^2 P(x^*). \) (3.2)

Observe that if \( x^* \) violates the second order KKT conditions it means that the symmetric matrix \( \nabla^2 P(x^*) \) has an eigenvector \( z \) orthogonal to all ones vector (for each player) with positive eigenvalue \( \lambda \). Moreover by law of inertia of Sylvester, the same holds for the matrix \( D_{zs}^{1/2}\nabla^2 P(x^*)D_{zs}^{1/2}. \) Moreover \( D_{zs}^{1/2}\nabla^2 P(x^*)D_{zs}^{1/2} \) has the same eigenvalues with matrix

\[
\begin{pmatrix}
 x_{11}^* & \cdots & x_{1M}^* \\
 \vdots & \ddots & \vdots \\
 x_{N1}^* & \cdots & x_{NM}^* \\
 0 & \cdots & 0 \\
\end{pmatrix}
\]

where \( D_{zs} \) is a diagonal matrix with positive diagonal entries and \( D_{xx} \) has rank \( N \). The Jacobian (not projected) of MWU dynamics computed at \( x^* \) can be expressed in a compact form as

\[
I + D_{zs}(I - D_{xx})\nabla^2 P(x^*) = I + D_{zs}\nabla^2 P(x^*) - D_{zs}D_{xx}\nabla^2 P(x^*). \] (3.2)

Moreover \( D_{zs}^{1/2}\nabla^2 P(x^*)D_{zs}^{1/2} \) has the same eigenvalues with matrix

\[
\begin{pmatrix}
 x_{11}^* & \cdots & x_{1M}^* \\
 \vdots & \ddots & \vdots \\
 x_{N1}^* & \cdots & x_{NM}^* \\
 0 & \cdots & 0 \\
\end{pmatrix}
\]

We can now prove our second main Theorem \[1,3\].

**Proof of Theorem \[1,3\].** As long as we establish the idea of projecting the Jacobian, then the proof follows the lines of work of (Mehta et al. 15, Lee et al. 17) and is rather generic. We shall show that the set of initial conditions so that MWU dynamics converges to unstable fixed points (meaning that the spectral radius of the Jacobian computed at the fixed point is greater than one) is of measure zero and then by Lemma \[3,1\] the proof follows. Let \( y \) be an unstable fixed point of the MWU (as a dynamical system) with update rule a function \( T_y : S \to S \). For such unstable fixed point \( y \), there is an associated open neighborhood \( B_y \subset S \) promised by the Stable Manifold Theorem \[5,1\].

Define \( W_y = \{ x^0 \in D_y : \lim_{t \to \infty} x^t = y \} \). Fix a point \( x^0 \in W_y \). Since \( x^k \to y \), then for some non-negative integer \( K \) and all \( t \geq K \), \( T_y^t(x^0) \in B_y \) (\( T_y^t \) denotes composition of \( T_y \) \( t \) times). We mentioned above that \( T_y \) is a diffeomorphism in \( S \). By Theorem \[5,1\] \( Q_y := \cap_{k=0}^{\infty} T_y^{-k}(B_y) \) is a subset of the local center stable manifold which has co-dimension at least one, and \( Q_y \) is thus measure zero.

Finally, \( T_y^K(x^0) \in Q_y \) implies that \( x^0 \in T_y^{-K}(Q_y) \). Since \( K \) is unknown we union over all non-negative integers, to obtain \( x^0 \in \cup_{j=0}^{\infty} T_y^{-j}(Q_y) \). Since \( x^0 \) was arbitrary, we have shown that \( W_y \subset \cup_{j=0}^{\infty} T_y^{-j}(Q_y) \). Using Lemma 1 of page 5 in (Lee et al. 17) and that countable union of measure zero sets is measure zero, \( W_y \) has measure zero. The claim follows since by mapping \( W_y \) to the set \( W \) (which is defined by padding the removed variables), then \( W \) is the set of initial conditions that MWU dynamics converges to \( y \) and is of measure zero in \( D \). □
4. Applications

One application of Baum-Eagon algorithm is parameter estimation via maximum likelihood. Suppose that \(X_1, ..., X_n\) are samples from a population with probability density function \(f(x|\theta_1, ..., \theta_k)\), the likelihood function is defined by

\[
L(\theta|x) = L(\theta_1, ..., \theta_k|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, ..., \theta_k).
\]

Maximum likelihood estimator has many applications in machine learning and statistics (e.g., regression) and when is consistent, the problem of estimation boils down to maximizing the likelihood function. This can be achieved via the E-M algorithm based on the Baum-Eagon inequality. For example, the estimation of the parameters of hidden Markov models (motivated by real world problems, see (Gopalakrishnan et al. 1991) for an example on speech recognition) result in the maximization of rational functions over a domain of probability values. The rational functions are conditional likelihood functions of parameters \(\theta = (\theta_1, ..., \theta_k)\). The Baum-Eagon dynamics is used to estimate the parameters of hidden Markov models. Our main result indicates that MWU dynamics should be used for the optimization part as MWU has some nice properties (well-defined, update rule is a diffeomorphism, avoids non-stationary points) in which Baum-Eagon dynamics might not have.

Below we provide a pictorial illustration of MWU dynamics applied to a non-concave function (not rational). The function we consider is \(P(x,y) = \cos(8x)\sin(6y)\) and we want to optimize it over \(R = \{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}\) (see Figure 3 for the landscape). The aforementioned instance is captured by our model for \(N = M = 2\), in which we have essentially projected the space by using one variable for each player (for player one, the second variable is \(1 - x\) and for player two is \(1 - y\)). The equations of MWU dynamics boil down to the following:

\[
\begin{align*}
x^{t+1} &= x^t + \frac{1+\epsilon(-8\sin(8x)\sin(6y))}{1+\epsilon(6\cos(8x)\cos(6y))} \\
y^{t+1} &= y^t + \frac{1+\epsilon(8\sin(8x)\sin(6y))}{1+\epsilon(-6\cos(8x)\cos(6y))}
\end{align*}
\]

We demonstrate in Figure 4 the “vector field” of MWU dynamics (because it is a discrete time system it is not precisely vector field, at point \((x, y)\) we plot a vector with direction \(T(x, y) - (x, y)\), where \(T\) is the update rule of dynamics (4.1)). The three dots indicate the local maxima of \(P\) and the rest of the points do not satisfy the second order KKT conditions. We observe that the MWU dynamics avoids those points that do not satisfy the second order KKT conditions (avoids those that are not local maxima).
MULTIPLICATIVE WEIGHTS UPDATES AS A DISTRIBUTED CONSTRAINED OPTIMIZATION ALGORITHM: CONVERGENCE TO SECOND-ORDER STATIONARY POINTS ALMOST ALWAYS

Figure 4. Vector field of MWU dynamics in the case of non-concave function $\cos(8x)\sin(6y)$. Only local maxima (red dots) have positive regions of attraction.

5. SUPPLEMENTARY MATERIAL

Theorem 5.1 (Center-stable manifold theorem, III.7 (Shub 1987)). Let $x^*$ be a fixed point for the $C^r$ local diffeomorphism $g : X \to X$. Suppose that $E = E_s \oplus E_u$, where $E_s$ is the span of the eigenvectors corresponding to eigenvalues of magnitude less than or equal to one of $Dg(x^*)$, and $E_u$ is the span of the eigenvectors corresponding to eigenvalues of magnitude greater than one of $Dg(x^*)$. Then there exists a $C^r$ embedded disk $W_{loc}^{cs}$ of dimension $\dim(E_s)$ that is tangent to $E_s$ at $x^*$ called the local stable center manifold. Moreover, there exists a neighborhood $B$ of $x^*$, such that $g(W_{loc}^{cs}) \cap B \subset W_{loc}^{cs}$, and $\cap_{k=0}^{\infty} g^{-k}(B) \subset W_{loc}^{cs}$.

REFERENCES

H. Ackermann, P. Berenbrink, S. Fischer, and M. Hoefer. Concurrent imitation dynamics in congestion games. In Proceedings of the 28th ACM symposium on Principles of distributed computing, pp. 63-72. ACM, 2009.

S. Arora, E. Hazan, S. Kale. The Multiplicative Weights Update Method: a Meta Algorithm and Applications. Theory of Computing 2012.

L. E. Baum, and J. A. Eagon, An inequality with applications to statistical prediction for functions of Markov processes and to a model of ecology, Bull. Amer. Math. Soc., Vol. 73, 1967.

R. Bhatia, Matrix Analysis. Springer, 1992.

P.S. Gopalakrishnan, Dimitri Kanevsky, Arthur Nadas and David Nahamoo (1991). An inequality for rational functions with applications to some statistical estimation problems, IEEE Transactions on Information Theory, Vol. 37, No. 1, 1991.

C. Ho. A note on proper maps. Proceedings of the American Mathematical Society, Vol. 51, No.1, 1975.

C. Jin, R. Ge, P. Netrapali, S. M. Kakade, M. I. Jordan. How to Escape Saddle Points Efficiently. ICML 2017

B. Kleinberg, Georgios Piliouras and Eva Tardos, Multiplicative updates outperform generic no-regret learning in congestion games, ACM Symposium on Theory of Computing (STOC), 2009.

R. Mehta, I. Panageas, G. Piliouras. Natural Selection as an Inhibitor of Genetic Diversity: Multiplicative Weights Updates Algorithm and a Conjecture of Haploid Genetics. ITCS 2015

J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan and B. Recht. First-order Methods Almost Always Avoid Saddle Points. Arxiv preprint 2017.

R. Ge, F. Huang, C. Jin, Y. Yuan. Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition. COLT 2015

G. Paliapopoulos, I. Panageas and G. Piliouras. Multiplicative Weights Update with Constant step-size in Congestion Games: Convergence, Limit Cycles and Chaos. NIPS 2017

M. Shub. Global Stability of Dynamical Systems. Springer Science & Business Media. 1987

P. Stebe (1972). Invariant functions of an iterative process for maximization of a polynomial. Pacific Journal of Mathematics, Vol. 43, No. 3.

Z. Allen-Zhu (2017). Natasha: Faster Non-Convex Stochastic Optimization Via Strongly Non-Convex Parameter.

Z. Allen-Zhu (2017). Natasha 2: Faster Non-Convex Optimization Than SGD ? How to Swing By Saddle Points.

3Jacobian of function $g$. 

Z. Allen-Zhu (2018). Katyusha X: Practical Momentum Method for Stochastic Sum-of-Nonconvex Optimization. 
*arXiv:1802.03866v1*

Z. Allen-Zhu (2018). How To Make the Gradients Small Stochastically: Even Faster Convex and Nonconvex SGD. 
*arXiv:1801.02982v2*