Multiplicative Weights Update as a Distributed Constrained Optimization Algorithm: Convergence to Second-order Stationary Points Almost Always

Ioannis Panageas [1] **Georgios Piliouras** [1] **Xiao Wang** [1]

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 saddle 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 $\max_{\mathbf{x}\in D} P(\mathbf{x})$, 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, by combining 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), (see also (Daskalakis & Panageas, 2018) for the analogue in min-max optimization). Moreover by adding noise, it is guaranteed that GD converges to a local maximum in polynomial time (see (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 (Allen-Zhu, 2017a) (Allen-Zhu, 2017b)(Allen-Zhu, 2018a)(Allen-Zhu, 2018b) 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 not to 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), (Daskalakis & Panageas, 2019)), 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 A.1 in the supplementary material).

In this paper we focus on solving problems of the form

$$\max_{\mathbf{x}\in D} P(\mathbf{x}), \tag{1}$$

where P is a non-concave, twice continuously differentiable function and D is some compact set, which will be a prod-

[1]Singapore University of Technology and Design, Singapore. Correspondence to: Ioannis Panageas <ioannis@sutd.edu.sg>, Georgios Piliouras <georgios@sutd.edu.sg>, Xiao Wang <xiao_wang@sutd.edu.sg>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

uct 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 $\mathbf{x}$ can be also interpreted as a collection of N probability distributions (having N players), where each distribution $\mathbf{x}_i$ has support of size M (strategies). For this particular problem (1), one natural algorithm that is commonly used is the Baum-Eagon dynamics (2) (see the seminal paper by Baum and Eagon (Baum & Eagon, 1967)) with many applications to inference problems, Hidden Markov Models (HMM) in particular (see also discussion in Section 4).

$$x_{ij}^{t+1} = x_{ij}^{t} \frac{\frac{\partial P}{\partial x_{ij}}\big|_{\mathbf{x}^t}}{\sum_s x_{is}^t \frac{\partial P}{\partial x_{is}}\big|_{\mathbf{x}^t}}, \tag{2}$$

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

Despite its power, Baum-Eagon dynamics has its limitations. First and foremost, the Baum-Eagon dynamics is not always well-defined; the denominator term $\sum_s x_{is}^t \frac{\partial P}{\partial x_{is}}\big|_{\mathbf{x}^t}$ must be non-zero at all times and moreover the fraction in equations (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 settings (see section 2.3) the Baum-Eagon dynamics 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 $\mathbf{x}^t$ is the t-th iterate of MWU, the equations of which can be described as follows:

$$x_{ij}^{t+1} = x_{ij}^{t} \frac{1 + \epsilon_i \frac{\partial P}{\partial x_{ij}}\big|_{\mathbf{x}^t}}{1 + \epsilon_i \sum_s x_{is}^t \frac{\partial P}{\partial x_{is}}\big|_{\mathbf{x}^t}}, \tag{3}$$

where ϵ_i the *stepsize* (learning rate) of the dynamics. Intuitively (in game theory terms), for strategy profile (vector) $\tilde{\mathbf{x}} := (\tilde{\mathbf{x}}_1, ...\tilde{\mathbf{x}}_N)$, each player i that chooses strategy j has utility to be $\frac{\partial P}{\partial x_{ij}}\big|_{\mathbf{x}=\tilde{\mathbf{x}}}$. We call a strategy profile $\mathbf{y} \in D$ a fixed point if it is invariant under the update rule dynamics (3). It is also clear that the set D is invariant under the dynamics in the sense that if $\mathbf{x}^t \in D$ then $\mathbf{x}^{t+1} \in D$ 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.

[1] A function is called a diffeomorphism if it is differentiable and a bijection and its inverse is differentiable as well.

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

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). Formally, it holds

$$\begin{aligned} &\mathbf{x}^* \in D \\ &x_{ij}^* > 0 \Rightarrow \tfrac{\partial P}{\partial x_{ij}}(\mathbf{x}^*) = \textstyle\sum_{j'} x_{ij'}^* \tfrac{\partial P}{\partial x_{ij'}}(\mathbf{x}^*) \\ &x_{ij}^* = 0 \Rightarrow \tfrac{\partial P}{\partial x_{ij}}(\mathbf{x}^*) \leq \textstyle\sum_{j'} x_{ij'}^* \tfrac{\partial P}{\partial x_{ij'}}(\mathbf{x}^*). \end{aligned} \tag{4}$$

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:

$$\mathbf{y}^\top \nabla^2 P(\mathbf{x}^*) \mathbf{y} \leq 0. \tag{5}$$

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 ϵ_i such that the set of initial conditions $\mathbf{x}^0$ of which the MWU dynamics (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 μ 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, with isolated[2] stationary points. It follows that with probability one (randomness induced by μ), MWU dynamics converges to second order stationary points.*

[2] A stationary point is isolated if there exists a neighborhood around it so that there is no other stationary point in that neighborhood.

Remark 1.5. It is obvious that when the learning rates $\epsilon_i = 0$, MWU (3) is trivially the identity map. On the other hand, whenever the dynamics 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).

We conclude our results by showing that it is unlikely that MWU dynamics converges fast to second (or even first) order stationary points when MWU is applied to solve problem (1). The problem of finding first (resp. second) order stationary points are inherently connected with the problem of finding mixed (resp. pure) Nash equilibria in congestion games. Currently, no polynomial time algorithms are known for computing mixed Nash in congestion games (the problem lies between P and CLS[3]) (Daskalakis & Papadimitriou, 2011), whereas computing pure Nash Nash equilibria even in linear congestion games, is known to be PLS-complete (Fabrikant et al., 2004; Ackermann et al., 2008). The reductions between the problems is based on the fact that congestion games are potential games and hence (3) captures the behavior of self-interested learning agents playing a congestion game.

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 dynamics is well defined). The proof of this statement leverages recently discovered connections between MWU and the Baum-Eagon dynamics (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).

[3]CLS is a computational complexity class that captures continuous local search. It lies on the intersection of the mores well studied classes of PLS and PPAD.

Notation Throughout this article, D is the product of N simplices of size M each, $D = \{(x_{ij}) | x_{ij} \geq 0, \sum_{j=1}^{M} x_{ij} = 1 \text{ for all } 1 \leq i \leq N\}$, where we interpret i as the index for the N agents and j the index of strategies M. We also use boldfaces to denote vectors, i.e., $\mathbf{x}$ and $[N]$ denotes $\{1, ..., N\}$.

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 (6) 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}, $i = 1, ..., k, j = 1, ..., n_i$. Let $n = \sum_{i=1}^{k} n_i$. Let D be the product of simplexes. Define $\mathbf{x}' := T(\mathbf{x})$ as the vector in D with component ij given by

$$x'_{ij} = T(\mathbf{x})_{ij} := \frac{x_{ij} \frac{\partial P}{\partial x_{ij}}}{\sum_{h=1}^{n_i} x_{ih} \frac{\partial P}{\partial x_{ih}}}. \tag{6}$$

Theorem 2.1 ((Baum & Eagon, 1967)). *Let $P(\{x_{ij}\})$ be a polynomial with non-negative coefficients homogeneous of degree d in its variables $\{x_{ij}\}$. Let $\mathbf{x} = \{x_{ij}\}$ be any point of the domain $D = \{x_{ij} \geq 0, \sum_{j=1}^{n_i} x_{ij} = 1, i = 1, 2, ..., k, j = 1, 2, ..., n_i\}$. For $\mathbf{x} = \{x_{ij}\} \in D$, let $T(\mathbf{x}) = T(\{x_{ij}\})$ be the point of D whose i, j coordinate is*

$$T(\mathbf{x})_{ij} = \frac{x_{ij} \frac{\partial P}{\partial x_{ij}}}{\sum_{h=1}^{n_i} x_{ih} \frac{\partial P}{\partial x_{ih}}}. \tag{7}$$

Then $P(T(\mathbf{x})) > P(\mathbf{x})$ unless $T(\mathbf{x}) = \mathbf{x}$.

2.2. Optimization for rational functions

According to (Gopalakrishnan et al., 1991), one can define a Baum-Eagon dynamics for rational functions $R(\mathbf{x}) = \frac{S_1(\mathbf{x})}{S_2(\mathbf{x})}$ with positive denominator so that the update rule of the Baum-Eagon dynamics increases the value of the rational

function R for any given vector $\mathbf{y}$ unless $\mathbf{y}$ is a fixed point. This can be done by starting with the Baum-Eagon map of the following polynomial: Let $\mathbf{y} \in D$ be an arbitrary point.

$$Q_{\mathbf{y}}(\mathbf{x}) = P_{\mathbf{y}}(\mathbf{x}) + C_{\mathbf{y}}(\mathbf{x}),$$

where $P_{\mathbf{y}}(\mathbf{x}) = S_1(\mathbf{x}) - R(\mathbf{y}) \cdot S_2(\mathbf{x}), C_{\mathbf{y}}(\mathbf{x}) = N_{\mathbf{y}}(\sum_{ij} x_{ij} + 1)^d$, where d is the degree of $P_{\mathbf{y}}(\mathbf{x})$ and $N_{\mathbf{y}}$ is a constant such that $P_{\mathbf{y}}(\mathbf{x}) + C_{\mathbf{y}}(\mathbf{x})$ only has nonnegative coefficients.

It is proved in (Gopalakrishnan et al., 1991) that $R(T(\mathbf{y})) > R(\mathbf{y})$ along the Baum-Eagon dynamics (update rule T) induced by polynomial $Q_{\mathbf{y}}(\mathbf{x})$.

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[4] of D onto itself if and only if the polynomial P can be expressed as a sum that *contains* monomials of the form $c_{i,j}x_{i,j}^{w_{i,j}}$ for all $i = 1, ..., k, j = 1, ..., n_i$ where $c_{i,j} > 0$ and $w_{i,j}$ is an integer greater than zero (this means that P might also contain other terms, i.e, products of different variables). But this condition is incorrect and we give a counter example below. We note that our example 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 τ defined on a single simplex (with n variables)

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

and τ can be written as

$$x_i' = \tau(\mathbf{x})_i := \frac{x_i \frac{\partial P}{\partial x_i}}{\sum x_i \frac{\partial P}{\partial x_i}} \tag{8}$$

The map defined in equation (8) can be expressed as a composition of τ_1 and τ_2 defined in the following way:

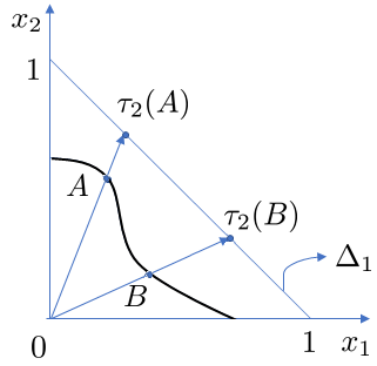
$$\tau_1 : (x_1, ..., x_n) \mapsto (x_1 \frac{\partial P}{\partial x_1}, ..., x_n \frac{\partial P}{\partial x_n}) \tag{9}$$

$$\tau_2 : (x_1 \frac{\partial P}{\partial x_1}, ..., x_n \frac{\partial P}{\partial x_n}) \mapsto \tag{10}$$

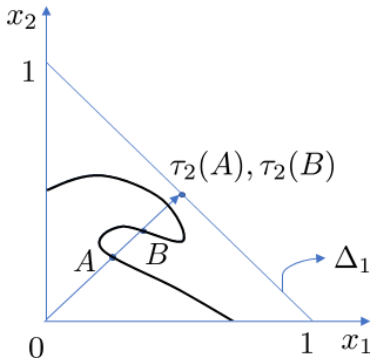
$$\frac{1}{\sum_{i=1}^{n} x_i \frac{\partial P}{\partial x_i}} (x_1 \frac{\partial P}{\partial x_1}, ..., x_n \frac{\partial P}{\partial x_n}) \tag{11}$$

Consider 1-dimensional simplex as an example (i.e, $n = 2$), τ_1 maps the simplex Δ_1 to a curve and τ_2 maps points on the curve back to Δ_1 by scaling. From Figure 1a, we notice

[4] 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.



(a) $\tau = \tau_2 \circ \tau_1$



(b) $\tau = \tau_2 \circ \tau_1$

.

Figure 1. Illustration

that a necessary condition for τ to be a homeomorphism is that the curve $\tau_1(\Delta_1)$ (image of Δ_1 under τ_1, see thick, black curve in Figure 1a, 1b) does not cross twice (or more times) any line that passes through the origin and has slope non-negative (see also Figure 1b). A necessary condition for τ to be a homeomorphism is that τ must be one to one. In 1-dimensional case, the ratio

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

must be monotone with respect to x_1. The following example is a polynomial that satisfies Baum's condition, however it holds that function k is not monotone with respect to x_1.

Example 2.2. Suppose $P = x_1 + x_1^7 x_2 + x_2^7$, then

$$x_1 \frac{\partial P}{\partial x_1} = x_1 + 7x_1^7 x_2$$
$$x_2 \frac{\partial P}{\partial x_2} = x_1^7 x_2 + 7x_2^7$$

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.

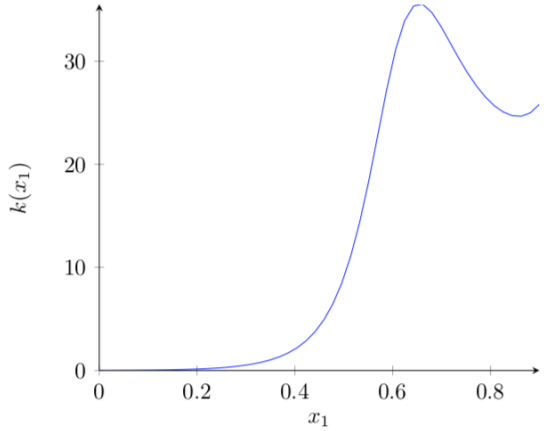


Figure 2. Non-monotonicity of $k(x_1)$

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 D. The update rule of the Baum-Eagon dynamics for the function $Q := \sum_{i,j} x_{ij} + \epsilon P$ (as defined in (2)) is a diffeomorphism for ϵ sufficiently small (we note that Baum-Eagon dynamics for Q coincides with the MWU dynamics for P, see Equations (3)). This is what next theorem captures.

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

Proof. Firstly, we prove that the Baum-Eagon map of Q is a local diffeomorphism. For a fixed i, denote

$$T(\mathbf{x})_{ij} = \frac{x_{ij} + \epsilon x_{ij} \frac{\partial P}{\partial x_{ij}}}{\sum_j x_{ij} + \epsilon \sum_j x_{ij} \frac{\partial P}{\partial x_{ij}}}.$$

Since the roots of the characteristic polynomial of a matrix vary continuously as a function of coefficients (see Theorem VI.1.2 in (Bhatia, 1992)), let J_ϵ be the Jacobian of the function $T(\mathbf{x})$, i.e., of the update rule of the Baum-Eagon dynamics induced by function $Q = \sum_{ij} x_{ij} + \epsilon P$ (note that T coincides with the MWU dynamics for function P with same stepsize ϵ (i.e., same learning rates)). The determinant $|J_\epsilon|$ is continuous with respect to ϵ. When $\epsilon \to 0$, it holds that $|J_\epsilon| \to 1$ at each point $p \in D$ where the Jacobian is computed, thus for each point $p \in D$ there exists ϵ_p, such that for all $\epsilon < \epsilon_p$, we get that $|J_\epsilon(p)| > 1/2$.

Since the determinant is also continuous with respect to points in D, for ϵ_p, there is a neighborhood of p, denoted as $U(p, \epsilon_p)$, such that for all $\mathbf{x} \in U(p, \epsilon_p)$, $|J_{\epsilon_p}(\mathbf{x})| > 1/2$. 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 δ in the lemma.

To prove that the Baum-Eagon 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 (we suggest the reader to see the supplementary material for all the missing definitions). □

Remark 2.4. The above theorem essentially can be generalized for different stepsizes (learning rates) ϵ 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., 2017), we will also make use of the Center-stable manifold theorem (Theorem A.1). The challenging part technically in this paper is to prove that every stationary point $\mathbf{x}$ that is not a local maximum has the property that the Jacobian of the MWU dynamics computed at $\mathbf{x}$ 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 $\mathbf{x}^t$ is the t-th iterate of MWU. Recall that:

$$x_{ij}^{t+1} = x_{ij}^t \frac{1 + \epsilon_i \frac{\partial P}{\partial x_{ij}}|_{\mathbf{x}=\mathbf{x}^t}}{1 + \epsilon_i \sum_s x_{is}^t \frac{\partial P}{\partial x_{is}}|_{\mathbf{x}=\mathbf{x}^t}} \tag{12}$$

where ϵ_i the stepsize of the dynamics. Let $T : D \to D$ be the update rule of the MWU dynamics (12). Fix indexes $i, i' \in [N]$ for players and $j, s \in [M]$ for strategies. Set $S_i = 1 + \epsilon_i \sum_{j'} x_{ij'} \frac{\partial P}{\partial x_{ij'}}$. The equations of the Jacobian look as follows:

$$\frac{\partial T_{ij}}{\partial x_{ij}} = \frac{1 + \epsilon_i \frac{\partial P}{\partial x_{ij}}}{S_i} + \frac{x_{ij}}{S_i^2}\Big(\epsilon_i \frac{\partial^2 P}{\partial x_{ij}^2} \cdot S_i - \epsilon_i(1 + \epsilon \frac{\partial P}{\partial x_{ij}}) \cdot (\frac{\partial P}{\partial x_{ij}} + x_{ij} \frac{\partial^2 P}{\partial x_{ij}^2} + \sum_{s \neq j} x_{is} \frac{\partial P}{\partial x_{is} \partial x_{ij}})\Big),$$

$$\frac{\partial T_{ij}}{\partial x_{is}} = \frac{x_{is}}{S_i^2}\Big(\epsilon_i \frac{\partial^2 P}{\partial x_{ij} \partial x_{is}} \cdot S_i - \epsilon_i(1 + \epsilon_i \frac{\partial P}{\partial x_{ij}} \cdot (\frac{\partial P}{\partial x_{is}} + x_{is} \frac{\partial^2 P}{\partial x_{is}^2} + \sum_{j' \neq s} x_{ij'} \frac{\partial^2 P}{\partial x_{ij'} \partial x_{is}}))\Big)$$

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

$$\frac{\partial T_{ij}}{\partial x_{i's}} = \frac{x_{ij}}{S_i^2}\Big(\epsilon_i \frac{\partial^2 P}{\partial x_{i's}\partial x_{ij}} \cdot S_i - \epsilon_i(1+\epsilon_i \frac{\partial P}{\partial x_{ij}}) \cdot (\sum_{j'} x_{ij'} \frac{\partial^2 P}{\partial x_{ij'}\partial x_{i's}})\Big) \text{ for } i \neq i' \in [N], j, s \in [M].$$

Let $\mathbf{y}$ to be a fixed point of MWU dynamics. We define the projected MWU mapping to be the function $T_{\mathbf{y}}$ 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 $D_{\mathbf{y}}$ to be the projection of D in the same way. Now the mapping is $T_{\mathbf{y}} : \mathcal{S} \to \mathcal{S}$ for $\mathcal{S} \subset \mathbb{R}^{NM-M}$ is still a diffeomorphism where $\mathcal{S}$ is an open set that contains $D_{\mathbf{y}}$. We define the corresponding Jacobian (called *projected 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 $\mathbf{y}$ 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 $\mathbf{y}$ (formally now is the projected point $\mathbf{y} \in D_{\mathbf{y}}$) has an eigenvalue with absolute value greater than one.*

Proof. Since $\mathbf{y}$ is not a stationary point, there exist i, j and so that $y_{ij} = 0$ but $\frac{\partial P}{\partial x_{ij}}\big|_{\mathbf{x}=\mathbf{y}} > \sum_{j'} y_{ij'} \frac{\partial P}{\partial x_{ij'}}\big|_{\mathbf{x}=\mathbf{y}}$. The projected Jacobian computed at $\mathbf{y}$ 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. □

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 $\mathbf{x}^*$ 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 $\mathbf{x}^*$ (formally now is the projected point $\mathbf{x}^* \in D_{\mathbf{x}^*}$) has an eigenvalue with absolute value greater than one.*

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

$$D_{xs} = \begin{pmatrix} \frac{\epsilon_1 x^*_{11}}{S_1} & & 0 & & & \\ & \ddots & & & 0 & \\ 0 & & \frac{\epsilon_1 x^*_{1M}}{S_1} & & & \\ & & & \ddots & & \\ & & & & \frac{\epsilon_N x^*_{N1}}{S_N} & \\ & 0 & & & & \ddots \\ & & & & 0 & & \frac{\epsilon_N x^*_{NM}}{S_N} \end{pmatrix}$$

and

$$D_{xx} = \begin{pmatrix} x^*_{11} & \cdots & x^*_{1M} & & & \\ \vdots & & \vdots & & 0 & \\ x^*_{11} & \cdots & x^*_{1M} & & & \\ & & & \ddots & & \\ & & & & x^*_{N1} & \cdots & x^*_{NM} \\ & 0 & & & \vdots & & \vdots \\ & & & & x^*_{N1} & \cdots & x^*_{NM} \end{pmatrix}$$

where D_{xs} is a diagonal matrix with positive diagonal entries and D_{xx} has rank N. The Jacobian (not projected) of MWU dynamics computed at $\mathbf{x}^*$ can be expressed in a compact form (see Section 3.1 for the equations of the Jacobian) as

$$I + D_{xs}(I - D_{xx})\nabla^2 P(\mathbf{x}^*) = I + D_{xs}\nabla^2 P(\mathbf{x}^*) - D_{xs}D_{xx}\nabla^2 P(\mathbf{x}^*), \tag{13}$$

where I denotes the identity matrix (in particular of size NM in the aforementioned expression). Observe that if $\mathbf{x}^*$ violates the second order KKT conditions it means that the symmetric matrix $\nabla^2 P(\mathbf{x}^*)$ has an eigenvector $\mathbf{z}$ orthogonal to all ones vector (for each player) with positive eigenvalue λ. Moreover by law of inertia of Sylvester, the same holds for the matrix

$$D_{xs}^{1/2}\nabla^2 P(\mathbf{x}^*)D_{xs}^{1/2}.$$

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

$$D_{xs}^{1/2}D_{xs}^{1/2}\nabla^2 P(\mathbf{x}^*)D_{xs}^{1/2}D_{xs}^{-1/2} = D_{xs}\nabla^2 P(\mathbf{x}^*),$$

therefore matrix $D_{xs}\nabla^2 P(\mathbf{x}^*)$ has a positive eigenvalue with an eigenvector $\mathbf{z}'$. To finish the proof, observe that

since $\mathbf{z}$ is orthogonal to all ones vector (the vector with all entries equal to 1), it holds that the null space of $D_{xs}D_{xx}$ and $D_{xs}D_{xx}\nabla^2 P(\mathbf{x}^*)$ span the whole space, hence $\mathbf{z}'$ should lie in the null space of $(D_{xs}D_{xx}\nabla^2 P(\mathbf{x}^*))^\top$. Therefore $\mathbf{z}'$ is an eigenvector of $D_{xs}(I - D_{xx})\nabla^2 P(\mathbf{x}^*)$ with positive eigenvalue, hence $\mathbf{z}'$ is an eigenvector of $I + D_{xs}(I - D_{xx})\nabla^2 P(\mathbf{x}^*)$ (i.e., of the Jacobian) with eigenvalue greater than one. It is easy to see that this is also an eigenvalue of the projected Jacobian and the claim follows. □

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., 2015), (Lee et al., 2017) 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 $\mathbf{y}$ be an unstable fixed point of the MWU (as a dynamical system) with update rule a function $T_{\mathbf{y}} : \mathcal{S} \to \mathcal{S}$. For such unstable fixed point $\mathbf{y}$, there is an associated open neighborhood $B_{\mathbf{y}} \subset \mathcal{S}$ promised by the Stable Manifold Theorem A.1.

Define $W_{\mathbf{y}} = \{\mathbf{x}^0 \in D_{\mathbf{y}} : \lim_{t\to\infty} \mathbf{x}^t = \mathbf{y}\}$. Fix a point $\mathbf{x}^0 \in W_{\mathbf{y}}$. Since $\mathbf{x}^k \to \mathbf{y}$, then for some non-negative integer K and all $t \geq K, T^t_{\mathbf{y}}(\mathbf{x}^0) \in B_{\mathbf{y}}$ ($T^t_{\mathbf{y}}$ denotes composition of $T_{\mathbf{y}}$ t times). We mentioned above that $T_{\mathbf{y}}$ is a diffeomorphism in $\mathcal{S}$. By Theorem A.1, $Q_{\mathbf{y}} := \cap_{k=0}^{\infty} T_{\mathbf{y}}^{-k}(B_{\mathbf{y}})$ is a subset of the local center-stable manifold which has co-dimension at least one, and $Q_{\mathbf{y}}$ is thus measure zero.

Finally, $T^K_{\mathbf{y}}(\mathbf{x}^0) \in Q_{\mathbf{y}}$ implies that $\mathbf{x}^0 \in T^{-K}_{\mathbf{y}}(Q_{\mathbf{y}})$. Since K is unknown we union over all non-negative integers, to obtain $\mathbf{x}^0 \in \cup_{j=0}^{\infty} T^{-j}_{\mathbf{y}}(Q_{\mathbf{y}})$. Since $\mathbf{x}^0$ was arbitrary, we have shown that $W_{\mathbf{y}} \subset \cup_{j=0}^{\infty} T^{-j}_{\mathbf{y}}(Q_{\mathbf{y}})$. Using Lemma 1 of page 5 in (Lee et al., 2017) and that countable union of measure zero sets is measure zero, $W_{\mathbf{y}}$ has measure zero. The claim follows since by mapping $W_{\mathbf{y}}$ 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 $\mathbf{y}$ and is of measure zero in D. □

3.3. On the speed of convergence

In this section we argue about the limitations of any algorithm that aims at solving maximization problem subject to simplex constraints (even for polynomial objectives), i.e., problem (1). We conclude that it is unlikely that MWU dynamics (or any other algorithm) converges in polynomial time to a local maximum (for problem (1)). In fact, as we will show providing a polynomial time algorithm for finding even first order stationary points for an arbitrary polynomial function is at least as hard as computing Nash equilibria for general congestion games, a problem for which no polynomial time algorithm is known and whose time complexity lies in CLS (Daskalakis & Papadimitriou, 2011). Computing second order stationary points even for general bilinear functions, specifically even for function of the form $f(\mathbf{x}) = \sum_{i,i',i\neq i'} \sum_{j,j'} a_{ii'jj'} x_{ij} x_{i'j'} + \sum_i \sum_j b_{ij} x_{ij}$ is strongly connected with the problem finding pure Nash equilibria even in linear congestion games that is known to be PLS-complete (Fabrikant et al., 2004; Ackermann et al., 2008).

Specifically, it suffices to focus on a special class of congestion games which are called threshold games. These are congestion games in which the set of resources R is divided into two disjoint subsets R_{in} and R_{out}. The set R_{out} contains a resource r_i for every player $i \in N$. This resource has a fixed delay T_i called the threshold of player i. Each player i has exactly two strategies: a strategy $S_i^{out} = \{r_i\}$ with $r_i \in R_i^{out}$, and a strategy $S_i^{in} \subseteq R_{in}$. Agent i prefers strategy S_i^{in} to strategy S_i^{out} if the total cost of playing S_i^{in} is smaller than the threshold cost T_i. Quadratic threshold games are a subclass of threshold games in which the set R_{in} contains exactly one resource $r_{ii'}$ for every unordered pair of players $\{i, i'\} \subset N$. For every player $i \in N$ of a quadratic threshold game, his strategy set $S_{in} = \{r_{ii'} | i' \in N, ji' \neq i\}$. Without loss of generality let any resource $r_{ii'}$ have a linear delay function of the form $c_{ii'}(k) = a_{ii'}k$ with $a_{ii'} > 0$. Furthermore, all thresholds can be assumed to be positive. (Ackermann et al., 2008) proves that computing a Nash equilibrium of a quadratic threshold game with nondecreasing delay functions is PLS-complete.

Theorem 3.3. *Finding a first-order stationary point for a general polynomial function f is at least as hard as computing a Nash equilibrium for general congestion games. Let $f(\mathbf{x}) = \sum_{i,i',i\neq i'} \sum_{j,j'} a_{ii'jj'} x_{ij} x_{i'j'} + \sum_i \sum_j b_{ij} x_{ij}$, where for all i, $\sum x_{ij} = 1$. Finding a second-order stationary point of $f(x)$ is at least as hard as computing a pure Nash equilibrium in a generic quadratic threshold game.*

Proof. Firstly, any first order stationary point of the expected value of the potential is a Nash equilibrium, since the gradient of the potential corresponds to the vector of deviating payoffs for all agents and all strategies. Thus. first order stationarity implies that only strategies that give maximal payoff are played with positive probability, i.e. the strategy is a Nash equilibrium. The expected value of the potential function of a quadratic threshold congestion games is a bilinear function. This is trivially true since each resource can only be used by at most two agents. Specifically the expected value of the potential function of the game when each agent i is using mixed strategy $(x_{iS_i^{in}}, x_{iS_i^{out}})$ is equal to $\sum_{i\in N} x_{iS_i^{in}} T_i +$

$\sum_{i\in N} x_{iS_i^{out}} \sum_{i'\neq i} a_{ii'} + \sum_{i,i',i\neq i'} x_{iS_i^{out}} x_{i'S_{i'}^{out}} a_{ii'}$. By the genericity assumption we can assume that the number of fixed points of MWU are finite and isolated, e.g. (Kleinberg et al., 2009). If this Nash equilibrium is pure then we are done. Suppose not, in which case there exist some agents that play mixed strategies with support equal to 2. Since the potential is a bilinear function it can be computed without error using its gradient and Hessian via Taylor expansion. Second order stationarity now implies that for any coordinated set of deviations of two of the randomizing agents the potential can still not improve. Consider the continuum of strategy profiles (ζ_i, x_{-i}) where i was a randomizing agent that now deviates and plays strategy S_i^{in} with arbitrary probability $\zeta_i \in [0,1]$. Since the original strategy profile x is a NE, agent i is still indifferent between his two actions. As we have argued any profile that exactly two randomizing agents deviate does not affect the value of the expected potential for so the value of the potential does not change. So, even if agent i' was to deviate to strategy $\zeta_i' \in [0,1]$, the value of the potential at $(\zeta_i, \zeta_i', x_{-i,i'})$ cannot be higher that its value at (ζ_i, x_{-i}) and x. So, none of the randomizing agents at any strategy profile (ζ_i, x_{-i}) can profit by deviating. Each point on the line segment (ζ_i, x_{-i}) with $\zeta_i \in [0,1]$ is a stationary point of MWU, and we reach a contradiction to our genericity assumption. Thus, the second order stationary point of the potential is a pure Nash. □

4. Applications

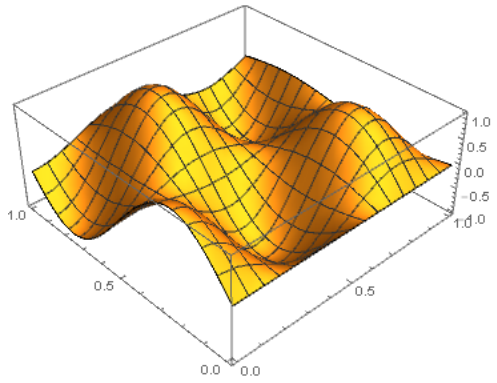


Figure 3. Landscape of non-concave function $cos(8x)sin(6y)$.

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

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

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:

$$\begin{aligned} x^{t+1} &= \frac{x^t(1+\epsilon(-8\sin(8x)\sin(6y)))}{1+\epsilon x\cdot(-8\sin(8x)\sin(6y))+\epsilon(1-x)\cdot(8\sin(8x)\sin(6y))} \\ y^{t+1} &= \frac{y^t(1+\epsilon(6\cos(8x)\cos(6y)))}{1+\epsilon y\cdot(6\cos(8x)\cos(6y))+\epsilon(1-y)\cdot(-6\cos(8x)\cos(6y))} \end{aligned} \tag{14}$$

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 (14)). The three dots indicate the local maxima of P and the rest of the points do not satisfy the second order KKT conditions. We see that MWU dynamics avoids those points that do not satisfy the second order KKT conditions (avoids those that are not local maxima).

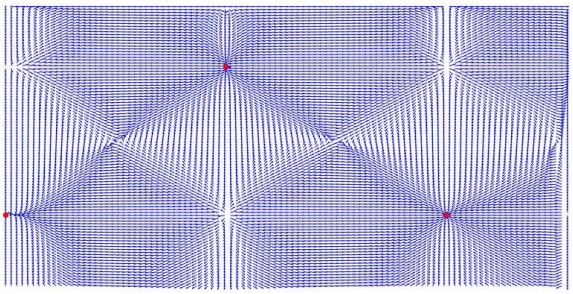


Figure 4. Vector field of MWU dynamics in the case of non-concave function $cos(8x)sin(6y)$. Only local maxima (red dots) have positive regions of attraction.

Acknowledgements

Ioannis Panageas acknowledges SRG ISTD 2018 136. Georgios Piliouras and Xiao Wang acknowledge MOE AcRF Tier 2 Grant 2016-T2-1-170, grant PIE-SGP-AI-2018-01 and NRF 2018 Fellowship NRF-NRFF2018-07.

References

Ackermann, H., Röglin, H., and Vöcking, B. On the impact of combinatorial structure on congestion games. *J. ACM*, 55(6):25:1–25:22, December 2008. ISSN 0004-5411. doi: 10.1145/1455248.1455249. URL http://doi.acm.org/10.1145/1455248.1455249.

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

Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *arXiv:1702.00763v5*, 2017a.

Allen-Zhu, Z. Natasha 2: Faster non-convex optimization than sgd ? how to swing by saddle points. In *arXiv:1708.08694v4*, 2017b.

Allen-Zhu, Z. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization. In *arXiv:1802.03866v1*, 2018a.

Allen-Zhu, Z. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. In *arXiv:1801.02982v2*, 2018b.

Arora, S., Hazan, E., and Kale, S. The multiplicative weights update method: a meta algorithm and applications. In *Theory of Computing*, 2012.

Baum, L. E. and Eagon, J. A. An inequality with applications to statistical prediction for functions of markov processes and to a model of ecology. In *Bulletin of the American Mathematical Society*, 1967.

Bhatia, R. *Matrix Analysis*. Springer, 1992.

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

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

Daskalakis, C. and Papadimitriou, C. Continuous local search. In *Proceedings of the Twenty-second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, pp. 790–804, Philadelphia, PA, USA, 2011. Society for Industrial and Applied Mathematics. URL http://dl.acm.org/citation.cfm?id=2133036.2133098.

Fabrikant, A., Papadimitriou, C., and Talwar, K. The complexity of pure nash equilibria. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pp. 604–612, New York, NY, USA, 2004. ACM. ISBN 1-58113-852-0. doi: 10.1145/1007352.1007445. URL http://doi.acm.org/10.1145/1007352.1007445.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *COLT*, 2015.

Gopalakrishnan, P., Kanevsky, D., Nadas, A., and Nahamoo, D. An inequality for rational functions with applications to some statistical estimation problems. In *IEEE Transactions on Information Theory*, 1991.

Ho, C. A note on proper maps. In *Proceedings of the American Mathematical Society*, 1975.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *ICML*, 2017.

Kleinberg, B., Piliouras, G., and Tardos, E. Multiplicative updates outperform generic no-regret learning in congestion games. In *ACM Symposium on Theory of Computing*, 2009.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid saddle points. In *arXiv preprint*, 2017.

Mehta, R., Panageas, I., and Piliouras, G. Natural selection as an inhibitor of genetic diversity: Multiplicative weights updates algorithm and a conjecture of haploid genetics. In *ITCS*, 2015.

Palaiopanos, G., Panageas, I., and Piliouras, G. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *NIPS*, 2017.

Stebe, P. Invariant functions of an iterative process for maximization of a polynomial. In *Pacific Journal of Mathematics*, 1972.