Accelerated Multiplicative Weights Update Avoids Saddle Points Almost Always

Yi Feng[1] , **Ioannis Panageas**[2] and **Xiao Wang**[1*]
[1]Shanghai University of Finance and Economics
[2]University of California, Irvine
2021310186@live.sufe.edu.cn, ipanagea@ics.uci.edu, wangxiao@sufe.edu.cn

Abstract

We consider non-convex optimization problems with constraint that is a product of simplices. A commonly used algorithm in solving this type of problem is the Multiplicative Weights Update (MWU), an algorithm that is widely used in game theory, machine learning and multi-agent systems. Despite it has been known that MWU avoids saddle points (Panageas et al.,2019b), there is a question that remains unaddressed: "Is there an accelerated version of MWU that avoids saddle points provably?" In this paper we provide a positive answer to above question. We provide an accelerated MWU based on Riemannian Accelerated Gradient Descent in (Zhang and Sra, 2018), and prove that the Riemannian Accelerated Gradient Descent, thus the accelerated MWU, almost always avoid saddle points.

1 Introduction

In this paper we consider non-convex optimization problem with constraint that is a product of simplices, i.e.,

$$\min_{\mathbf{x}\in\Delta_1\times\ldots\times\Delta_n} f(\mathbf{x}) \tag{1}$$

where $f:\Delta_1\times\ldots\times\Delta_n\rightarrow\mathbb{R}$ is a sufficiently smooth function and

$$\Delta_i=\left\{(x_{i1},...,x_{id}):\sum_{s=1}^{d}x_{is}=1,x_{is}\geq 0\right\},$$

Especially, we are interested in the Multiplicative Weights Update (MWU) algorithm [Arora *et al.*, 2012], which is commonly used in various fields including game theory, optimization, machine learning and multi-agent systems. We propose an Accelerated Multiplicative Weights Update (A-MWU), and study its saddle point avoidance behavior from a general perspective of Riemannian Accelerated Gradient Descent [Zhang and Sra, 2018].

Escaping saddle points in non-convex optimization has been studied extensively by the Machine Learning community [Ge *et al.*, 2015; Jin *et al.*, 2017; Jin *et al.*, 2018; Lee *et al.*, 2016; Lee *et al.*, 2019; Panageas and Piliouras, 2016; Criscitiello and Boumal, 2019; Sun *et al.*, 2019b; Panageas *et al.*, 2019a; Sun *et al.*, 2019a]. The first-order optimization algorithms can be studied from a dynamical system perspective and results in[Lee *et al.*, 2016; Lee *et al.*, 2019] guarantee that the algorithms asymptotically avoid saddle points in probability 1 with random initialization. The main technique of using Center-Stable Manifold theorem was extended to heavy-ball algorithm on Euclidean space by [Sun *et al.*, 2019a].

Despite the aforementioned progresses, the saddle avoidance of accelerated algorithms in non-Euclidean setting is less studied, especially for the case where the step-size is varying with time. The only result on variable step-size is given by [Panageas *et al.*, 2019a], but the Riemannian Accelerated Gradient Gescent is omitted. It is known that MWU on the simplex is a special case of mirror descent with entropy regularizer, and the mirror descent algorithm has many applications in optimization [Dekel *et al.*, 2011; Juditsky *et al.*, 2011]. Using this connection, an accelerated version of mirror descent is provided by [Krichene *et al.*, 2015], in which the authors discretize a system of ODEs of continuous-time mirror descent to get acceleration. However, [Krichene *et al.*, 2015] focuses on convex optimization only and leaves the saddle avoidance in non-convex optimization. [Panageas *et al.*, 2019b] proved that MWU almost always converges to second-order stationary points with constant step-size, but the result for accelerated MWU with variable step-size is missing in the literatures.

Motivated by the question that if there exists an accelerated version of MWU which also provably avoids saddle point, we investigate the Riemannian Accelerated Gradient Descent (RAGD) proposed by Zhang and Sra in [Zhang and Sra, 2018]. The main results and contributions of this paper are the following:

- We propose an Accelerated Multiplicative Weights Update which is derived and simplified from RAGD of [Zhang and Sra, 2018];
- We prove that the RAGD of [Zhang and Sra, 2018] avoids saddle points and moreover, this provides the first saddle avoidance result of Accelerated Multiplicative Weights Update.

Our contributions compared to the most relevant results in literature are illustrated in Table 1.

*Contact Author

	Acceleration	Saddle Avoidance	Variable Step-size	Types of Manifolds
Accelerated MD [Krichene *et al.*, 2015]	✓	✗	✓	Convex set
Heavy-ball [Sun *et al.*, 2019a]	✓	✓	✗	Euclidean space
MWU [Panageas *et al.*, 2019b]	✗	✓	✗	Simplices
RAGD [Zhang and Sra, 2018]	✓	✗	✓	General
AMWU (this work)	✓	✓	✓	General, Simplices

Table 1: Comparison to related results

2 Preliminaries

Riemannian metric and geodesic. A d-dimension Riemannian manifold (M, g) is real, smooth d-dimension manifold M equipped with a Riemannian metric g. For each $\mathbf{x} \in M$, let $T_{\mathbf{x}}M$ denote the tangent space at $\mathbf{x}$. The metric g induces a inner product $\langle \cdot, \cdot \rangle_{\mathbf{x}} : T_{\mathbf{x}}M \times T_{\mathbf{x}}M \to \mathbb{R}$. We call a curve $\gamma(t) : [0, 1] \to M$ a geodesic if it satisfies

- The curve $\gamma(t)$ is parametrized with constant speed, i.e. $\left\| \frac{d}{dt}\gamma(t) \right\|_{\gamma(t)}$ is constant for $t \in [0, 1]$.
- The curve is locally length minimized between $\gamma(0)$ and $\gamma(1)$.

Exponential and logarithmic map. The exponential map $\mathrm{Exp}_{\mathbf{x}}(\mathbf{v})$ maps $\mathbf{v} \in T_{\mathbf{x}}M$ to $\mathbf{y} \in M$ such that there exists a geodesic γ with $\gamma(0) = \mathbf{x}$, $\gamma(1) = \mathbf{y}$ and $\gamma'(0) = \mathbf{v}$. For $\mathbf{x} \in M$, let $\mathrm{Log}_{\mathbf{x}}$ denote the logarithmic map at $\mathbf{x}$,

$$\mathrm{Log}_{\mathbf{x}}(\mathbf{y}) = \underset{\mathbf{u} \in T_{\mathbf{x}}M}{\mathbf{argmin}} \text{ subject to } \mathrm{Exp}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y},$$

with domain such that this is uniquely defined.

Remark 1. *The existence and uniqueness of geodesic is guaranteed by the Fundametal Theorem of ODE, one cannot expect an explicit expression of the geodesic, usually we need to apply a numerical scheme to approximate the solution of the geodesic equation. However, there are cases for which the geodesic equation can be expressed with a closed form, we give the following examples.*

Geodesically convex. A set $U \subset M$ is geodesically convex if for any $\mathbf{x}, \mathbf{y} \in U$, there is a geodesic γ with $\gamma(0) = \mathbf{x}$, $\gamma(1) = \mathbf{y}$ and $\gamma(t) \in U$ for $t \in [0, 1]$.

Let U be a geodesically convex subset of M. A function $f : M \to \mathbb{R}$ is called geodesically convex on U if for any $\mathbf{x}, \mathbf{y} \in M$ and any geodesic γ such that $\gamma(0) = \mathbf{x}$, $\gamma(1) = \mathbf{y}$ and $\gamma(t) \in U$ for all $t \in [0, 1]$, it holds that

$$f(\gamma(t)) \le (1 - t)f(\mathbf{x}) + tf(\mathbf{y}).$$

A function $f : M \to \mathbb{R}$ is called geodesically μ-convex on U if for any $\mathbf{x}, \mathbf{y} \in U$ and gradient $\mathrm{grad}f(\mathbf{x})$ at $\mathbf{x}$, it holds that

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \mathrm{grad}f(\mathbf{x}), \mathrm{Log}_{\mathbf{x}}(\mathbf{y}) \rangle + \frac{\mu}{2} \left\| \mathrm{Log}_{\mathbf{x}}(\mathbf{y}) \right\|^2, \quad (2)$$

where we assume the logarithmic map is well defined in U.

Riemannian Gradient and Hessian. For differentiable function $f : M \to \mathbb{R}$, $\mathrm{grad}f(\mathbf{x}) \in T_{\mathbf{x}}M$ denotes the Riemannian gradient of f that satisfies $\frac{d}{dt}f(\gamma(t)) = \langle \gamma'(t), \mathrm{grad}f(\mathbf{x}) \rangle$ for any differentiable curve $\gamma(t)$ passing through $\mathbf{x}$. The local coordinate expression of gradient is useful in our analysis.

$$\mathrm{grad}f(\mathbf{x}) = \left(\sum_j g^{1j}(\mathbf{x}) \frac{\partial f}{\partial x_j}, ..., \sum_j g^{dj}(\mathbf{x}) \frac{\partial f}{\partial x_j} \right) \quad (3)$$

where $g^{ij}(\mathbf{x})$ is the ij-th entry of the inverse of the metric matrix $\{g_{ij}(\mathbf{x})\}$ at each point.

The Hessian of f is the covariant derivative of the gradient vector field: $\mathrm{Hess}f(\mathbf{x})[\mathbf{u}] = \nabla_{\mathbf{u}} \mathrm{grad}f(\mathbf{x})$ for any vector field $\mathbf{u}$ on M.

Strict saddle point. A strict saddle point $\mathbf{x}^*$ of $f : M \to \mathbb{R}$ satisfies

$$\|\mathrm{grad}f(\mathbf{x}^*)\| = 0 \text{ and } \lambda_{\min}(\mathrm{Hess}f(\mathbf{x}^*)) < 0.$$

Retraction. A retraction on a manifold M is a smooth mapping Retr from the tangent bundle TM to M satisfying properties 1 and 2 below: Let $\mathrm{Retr}_{\mathbf{x}} : T_{\mathbf{x}}M \to M$ denote the restriction of $Retr$ to $T_{\mathbf{x}}M$.

1. $\mathrm{Retr}_{\mathbf{x}}(0) = \mathbf{x}$, where 0 is the zero vector in $T_{\mathbf{x}}M$.
2. The differential of $\mathrm{Retr}_{\mathbf{x}}$ at 0 is the identity map.

Then the Riemannian gradient descent with stepsize α is given as

$$\mathbf{x}_{t+1} = \mathrm{Retr}_{\mathbf{x}_t}(-\alpha \mathrm{grad}f(\mathbf{x}_t)). \quad (4)$$

Multiplicative Weights Update. The paper focuses on Accelerated Multiplicative Weights Update, for completeness, we recall the linear variant of MWU. Suppose that $\mathbf{x}_i = (x_{i1}, ..., x_{id_i})$ is in the i-th component of $\Delta_1 \times ... \times \Delta_n$. Assume that $\mathbf{x}(t)$ is the t-th iterate of MWU, the algorithm is written as follows:

$$x_{ij}(t+1) = x_{ij}(t) \frac{1 - \alpha \frac{\partial f}{\partial x_{ij}}}{1 - \alpha \sum_s x_{is}(t) \frac{\partial f}{\partial x_{is}}}, \quad (5)$$

where $j \in \{1, ..., d_i\}$.

Algorithm 1 Single-Agent A-MWU

input : $\mathbf{x}_0, \mathbf{v}_0, 0 < c \le \alpha_t < \frac{1}{L}, \beta_t > 0, \delta > 0,$

repeat

Compute $s_t \in (0,1)$ from the equation $s_t^2 = \alpha_t((1-s_t)\gamma_t + s_t\mu)$,

Set $\bar{\gamma}_{t+1} = (1-s_t)\gamma_t + s_t\mu$, $\gamma_{t+1} = \frac{1}{1+\beta_t}\bar{\gamma}_{t+1}$,

$S = \left(\prod_{i=1}^d \frac{x_i(t)}{y_i(t)}\right)^{1/d}$

Set $y_i(t+1) = \frac{x_i(t)\exp\left(\frac{s_t\gamma_t}{\gamma_t+s_t\mu}\ln\left(S\frac{v_i(t)}{x_i(t)}\right)\right)}{\sum_j x_j(t)\exp\left(\frac{s_t\gamma_t}{\gamma_t+s_t\mu}\ln\left(S\frac{v_j(t)}{x_j(t)}\right)\right)}$

Set $x_i(t+1) = y_i(t)\frac{1-\alpha_t\frac{\partial f}{\partial x_i}(\mathbf{y}_t)}{1-\alpha_t\sum_j\frac{\partial f}{\partial x_j}(\mathbf{y}_t)}$

Compute $S' = \left(\prod_{i=1}^d \frac{y_i(t)}{v_i(t)}\right)^{1/d}$,

$$u_i = \frac{(1-s_t)\gamma_t}{\bar{\gamma}_t}\ln\left(S'\frac{v_i(t)}{y_i(t)}\right) + y_i(t)\frac{1-\alpha_t\frac{\partial f}{\partial x_i}(\mathbf{y}_t)}{1-\alpha_t\sum_j y_j(t)\frac{\partial f}{\partial x_j}(\mathbf{y}_t)} - y_i(t)$$

Set $v_i(t+1) = \frac{y_i(t)\exp(u_i)}{\sum_j y_j(t)\exp(u_j)}$

until $\|\mathrm{grad} f(\mathbf{y}_t)\| \le \delta$

3 Algorithm

We give the Single-Agent version of the Accelerated Multiplicative Weights Update (A-MWU) in Algorithm 1, which is derived from the Riemannian Accelerated Gradient Descent based on the geometry of the positive orthant $\mathbb{R}^d_+$. We leave the full multi-agent A-MWU algorithm to Appendix.

Even the derivation of Algorithm 1 is based on the Riemannian Accelerated Gradient Descent, the above implementation is simplified compared to the original algorithm in [Zhang and Sra, 2018]. Since it is impossible to have an explicit form of exponential and logarithmic map in general, all the updates of the algorithm are computed based on the property of Riemannian geometry of the positive orthant $\mathbb{R}^d_+ = \{\mathbf{x} : x_i > 0 \ \text{ for all } \ i \in [d]\}$, [Shahshahani, 1979; Hofbauer and Sigmund, 1998]. We recall the background that is necessary for the interpretation that MWU is indeed a manifold gradient descent. Formally, the positive orthant $\mathbb{R}^d_+$ is endowed with a Riemannian metric (called the **Shahshahani metric**) whose metric matrix $\{g_{ij}(\mathbf{x})\}$ is diagonal with $g_{ii}(\mathbf{x}) = \frac{|\mathbf{x}|}{x_i}$ where $|\mathbf{x}| = \sum_j x_j$, The positive orthant $\mathbb{R}^d_+$ with the Shahshahani metric is call a Shahshahani manifold. The tangent spaces $T_{\mathbf{x}}\mathbb{R}^d_+$ for all $\mathbf{x} \in \mathbb{R}^d_+$ are all identified with $\mathbb{R}^d$. Now consider a differentiable function $f : \mathbb{R}^d_+ \to \mathbb{R}$, one can define the gradient of f at each point with respect to the Shahshahani metric. The following local-coordinate expression of the Shahshahani gradient for each $\mathbf{x} \in \mathbb{R}^d_+$ is straightforward from (3):

$$\mathrm{grad} f(\mathbf{x}) = g^{-1} \cdot \nabla f(\mathbf{x}) = \left(\frac{x_1}{|\mathbf{x}|}\frac{\partial f}{\partial x_1}, ..., \frac{x_d}{|\mathbf{x}|}\frac{\partial f}{\partial x_d}\right).$$

We know from [Hofbauer and Sigmund, 1998] that the exponential map on the positive simplex $\Delta_+ \subset \mathbb{R}^n_+$ is

$$\mathrm{Exp}_{\mathbf{x}}(\mathbf{v}) = \left(\frac{x_1 e^{v_1}}{\sum x_i e^{v_i}}, ..., \frac{x_n e^{v_n}}{\sum x_i e^{v_i}}\right),$$

where $\mathbf{v} = (v_1, ..., v_n)$ is a tangent vector of Δ_+ at point $\mathbf{x}$.

Moreover, the Multiplicative Weights Update given by (5) can be understood as a retraction on Δ_+ and the logarithmic map can also be computed explicitly. These makes the formulation of Algorithm 1 possible, and we leave the details in Appendix.

4 Saddle Avoidance Analysis

We show that the Accelerated Gradient Descent on the product manifold $M_1 \times ... \times M_n$ actually avoids saddle points almost always. The algorithm is extended from the Accelerated Riemannian Gradient Descent of [Zhang and Sra, 2018] with the structure of product manifold.

Suppose that each manifold component M_i is equipped with a Riemannian metric g_i, then the product manifold $M_1 \times ... \times M_n$ has the product metric $g = g_1 \otimes ... \otimes g_n$ whose metric matrix is blocked matrix with g_i's the non-trivial blocks. Then the gradient is also the Cartesian product of the gradients of each component manifold M_i, i.e.

$$\mathrm{grad} f(\mathbf{x}) = (\mathrm{grad}_{\mathbf{x}_1} f(\mathbf{x}), ..., \mathrm{grad}_{\mathbf{x}_n} f(\mathbf{x})).$$

Recall the Accelerated Riemannian Gradient Descent in [Zhang and Sra, 2018] as follows,

$$\mathbf{y}_t = \mathrm{Exp}_{\mathbf{x}_t}\left(\frac{s_t\gamma_t}{\gamma_t + s_t\mu}\mathrm{Log}_{\mathbf{x}_t}(\mathbf{v}_t)\right) \tag{6}$$

$$\mathbf{x}_{t+1} = \mathrm{Exp}_{\mathbf{y}_t}(-\alpha_t \mathrm{grad} f(\mathbf{y}_t)) \tag{7}$$

$$\mathbf{v}_{t+1} = \mathrm{Exp}_{\mathbf{y}_t}\left(\frac{(1-s_t)\gamma_t}{\bar{\gamma}_t}\mathrm{Log}_{\mathbf{y}_t}(\mathbf{v}_t) - \frac{s_t}{\bar{\gamma}_t}\mathrm{grad} f(\mathbf{y}_t)\right) \tag{8}$$

where the $0 < c \le \alpha_t < \frac{1}{L}$, $\beta_t > 0$, and s_t, γ_t, $\bar{\gamma}_t$ are computed according to $s_t \in (0,1)$, such that $s_t^2 = \alpha_t((1-s_t)\gamma_t + s_t\mu)$, and

$$\bar{\gamma}_{t+1} = (1-s_t)\gamma_t + s_t\mu, \quad \gamma_{t+1} = \frac{1}{1+\beta_t}\bar{\gamma}_{t+1}.$$

In order to write the distributed Accelerated Riemannian Gradient Descent, we simply compute the exponential and logarithmic map component-wise. We denote

$$\overrightarrow{\mathbf{x}} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in M_1 \times, ..., \times M_n$$

and the t'th update

$$\overrightarrow{\mathbf{x}}_t = (\mathbf{x}_{1,t}, ..., \mathbf{x}_{n,t}), \quad \overrightarrow{\mathbf{v}}_t = (\mathbf{v}_{1,t}, ..., \mathbf{v}_{n,t})$$

are computed according to (6) in a distributed manner.

Note that the update rule of RAGD can be written as the composition of three maps, i.e., denote

$$\mathbf{y}(\mathbf{x}, \mathbf{v}) = \mathrm{Exp}_{\mathbf{x}}\left(\frac{s_t\gamma_t}{\gamma_t + s_t\mu}\mathrm{Log}_{\mathbf{x}}(\mathbf{v})\right), \tag{9}$$

$$F(\mathbf{y}) = \mathrm{Exp}_{\mathbf{y}}(-\alpha_t \mathrm{grad} f(\mathbf{y})), \tag{10}$$

$$G(\mathbf{y}, \mathbf{v}) = \mathrm{Exp}_{\mathbf{y}}\left(\frac{(1-s_t)\gamma_t}{\bar{\gamma}_t}\mathrm{Log}_{\mathbf{y}}(\mathbf{v}) - \frac{s_t}{\bar{\gamma}_t}\mathrm{grad} f(\mathbf{y})\right). \tag{11}$$

Therefore, the update rule, which can be viewed as a map $\psi(\mathbf{x}, \mathbf{v}) : M \times M \to M \times M$, can be written in the following way:

$$\psi(\mathbf{x}, \mathbf{v}) \stackrel{\text{def}}{=} (F(\mathbf{y}), G(\mathbf{y}, \mathbf{v})) = (F(\mathbf{y}(\mathbf{x}, \mathbf{v})), G(\mathbf{y}(\mathbf{x}, \mathbf{v}), \mathbf{v})) . \tag{12}$$

If $\mathbf{x}^* = \mathbf{v}^*$ where $\mathbf{x}^*$ is a critical point, then the intermediate variable $\mathbf{y}(\mathbf{x}^*, \mathbf{v}^*) = \mathbf{x}^*$, and $\psi(\mathbf{x}^*, \mathbf{v}^*) = (\mathbf{x}^*, \mathbf{v}^*)$. This means that if $\mathbf{x}^*$ is a critical point of $f : M \to \mathbb{R}$, the point $(\mathbf{x}^*, \mathbf{x}^*)$ is a stationary point of the dynamical system defined by iterations of ψ on $M \times M$.

The main technical step is the proof of "Stable-manifold theorem" for the dynamical system defined by iteration of the augmented map (12), which is a non-autonomous discrete-time dynamical system. We state the new stable manifold theorem as follows:

Theorem 2. *For a fixed dimension $d \geq 2$, there exists $s \in [d]$, and positive constant $K_1, K_2 < 1$, such that for all $t \in \mathbb{N}$, the following holds*

$$\sup_{1\leq j\leq s} \{\mathcal{L}_j(t)\} \leq K_1, \qquad \sup_{s+1\leq j\leq d} \{\mathcal{L}_j^{-1}(t)\} \leq K_2. \tag{13}$$

Suppose $\boldsymbol{\eta}(t, \mathbf{x})$ satisfies that $\boldsymbol{\eta}(t, \mathbf{0}) = \mathbf{0}$ and for each $\epsilon > 0$, there exists a neighborhod of $\mathbf{0}$ such that

$$\|\boldsymbol{\eta}(t, \mathbf{x}) - \boldsymbol{\eta}(t, \mathbf{y})\| \leq \epsilon \|\mathbf{x} - \mathbf{y}\| .$$

Then the dynamical system,

$$\mathbf{x}_{t+1} = \mathbf{L}\mathbf{x}_t + \boldsymbol{\eta}(t, \mathbf{x}_t),$$

where $\mathbf{L} = \mathbf{diag}\{\mathcal{L}_j(t)\}$, has a local stable manifold in a small neighborhood of $\mathbf{0}$.

Note that the Center-Stable Manifold Theorem only secures the existence of local stable manifold, which implies that there exists a small neighborhood of an unstable fixed point $\mathbf{x}^*$ such that the all initial conditions taken from this neighborhood and converges to $\mathbf{x}^*$ must belong to a lower dimensional curved space, thus has measure zero. The local result can be extended to global through an argument in Appendix, for the case when the dynamical system is non-autonomous. In particular, if the update rule ψ of the algorithm is not time-dependent such as the cases of constant step, the above result holds trivially by letting $\psi(t, \mathbf{x}) \equiv \psi(\mathbf{x})$. We give the following general version on manifold also for further referrings.

Proposition 3. *Let $\psi(t, \mathbf{x}) : \mathbb{N} \times M \to M$ be a dynamical system on a finite dimensional manifold M. Suppose that $\psi(t, \cdot)$ is a local diffeomorphism on M for each $t \in \mathbb{N}$. Let $\mathcal{A}^*$ be the set of unstable fixed points. Then the global stable set $W^s(\mathcal{A}^*)$ of $\mathcal{A}^*$ has measure of zero with respect to the volume measure induced by the Riemannian metric on M.*

Here the stable and unstable set of a point $\mathbf{x}$ are defined to be

- Stable set of $\mathbf{x}$:
$$W^s(\mathbf{x}) = \{\mathbf{z} \in M : \lim_{t\to\infty} d(\psi(t, \mathbf{z}), \psi(t, \mathbf{x})) \to 0\}$$
- Unstable set of $\mathbf{x}$:
$$W^u(\mathbf{x}) = \{\mathbf{z} \in M : \lim_{t\to-\infty} d(\psi(t, \mathbf{z}), \psi(t, \mathbf{x})) \to 0\}.$$

To prove following theorem 4, we leverage the property of the dynamical system induced by the map ψ constructed in (12). Its orbit is the same as the orbit of RAGD and it also satisfies all other conditions in the center-stable manifold theorem. As mentioned above, the theorem tells us in a local neighborhood of a strict saddle point $\mathbf{x}^*$, all points in this neighborhood that will converge to $\mathbf{x}^*$ under ψ lie in a submanifold, and it's dimension equals to the number of the eigenvalues of the Jacobian of ψ at $\mathbf{x}^*$ that are less than 1. Then we verify $DF(\mathbf{x}^*, \mathbf{x}^*)$ has a eigenvalue greater or equal to 1. So all initial points in the local neighborhood that will converge to $\mathbf{x}^*$ lies on a low dimension space, thus has measure 0. Now we assume critical points are uncountable. Since points converge to a strict saddle point under RAGD will finally fall into a local neighborhood of this strict saddle point, the set consisted by points in the whole space that will converge to a strict saddle point is a countable union of measure 0 sets, which is also measure 0. If critical points are uncountable, the Lindelof's lemma, which says every open cover there is a countable subcover, leads us to the fact that the set of initial conditions that converge to the set of saddle points (uncountable) is a countable union of measure zero sets, thus, a measure zero set. Formally, it is stated as follows.

Theorem 4. *Suppose $f : M \to \mathbb{R}$ is geodesically μ-convex in neighborhood of local minima. Then the set of initial points $(\mathbf{x}_0, \mathbf{v}_0) \in M \times M$ converging to saddle points of f under the Riemannian Accelerated Gradient Descent [Zhang and Sra, 2018] has measure (induced by the product metric) of zero.*

The proof is left in Appendix.

Corollary 5. *Let $\mathcal{M} = M_1 \times ... \times M_n$. Suppose that $f : \mathcal{M} \to \mathbb{R}$ is geodesically μ-convex in neighborhood of local minima. Then the set of initial points $(\overrightarrow{\mathbf{x}}_0, \overrightarrow{\mathbf{v}}_0) \in \mathcal{M} \times \mathcal{M}$ converging to saddle points of f has measure (induced by the product metric) of zero.*

The proof is completed by trivially considering $\mathcal{M}$ as the manifold in Theorem 4, all the structures of the proof can be carried over. The following corollary is immediate from Theorem 4 and Corollary 5.

Corollary 6. *Let $\mathcal{M} = \Delta_1 \times ... \times \Delta_n$, where*

$$\Delta_i = \left\{ \mathbf{x}_i \in \mathbb{R}^{d_i} : \sum_{s=1}^{d_i} x_{is} = 1, x_{is} > 0 \right\}.$$

Suppose $f : \mathcal{M} \to \mathbb{R}$ is C^2 and geodesically convex w.r.t. the product Shahshahani metric. Then the Accelerated Multiplicative Weights Update algorithm avoids interior saddle points almost always, i.e., randomly choose an initial point, the algorithm will avoid interior saddle points with probability one.

Remark 7. *In many applications, the constraint is compact, i.e., $\Delta_i = \left\{ \mathbf{x}_i \in \mathbb{R}^{d_i} : \sum_{s=1}^{d_i} x_{is} = 1, x_{is} \geq 0 \right\}$, and we might be interested in convergence to second-order stationary points [Panageas* et al.*, 2019b; Lu* et al.*, 2019]. With Theorem 2, Lemma 3.1 and Lemma 3.2 in [Panageas* et al.*, 2019b], one can extend the result of Corollary 6 to convergence to second-order stationary points with compact constraints.*

5 Experiments

Algorithms for comparison. The experiments are designed to understand the behavior of A-MWU compared to classic MWU and Accelerated Mirror Descent (A-MD) proposed by [Krichene *et al.*, 2015], especially their behaviors near a saddle point. Since the performance of A-MD is controlled by the parameter r in [Krichene *et al.*, 2015], i.e., the larger r ends up with a smoother curve of convergence, but with a slower rate of convergence. In order to understand exactly how A-MWU performs, we choose different r in Accelerated Mirror Descent while keep the step-size of A-MWU and A-MD the same.

Parameter setting. In our experiments, we set the parameters as follows:

- $\beta = 0.001$ and $\mu = 1$ for Rosenbrock function, Figure 1.
- $\beta = 0.1$ and $\mu = 1$ for Bohachevsky function, Figure 2.
- $\beta = 0.1$ and $\mu = 0.2$ for Test function 1, Figure 3.
- $\beta = 0.001$ and $\mu = 0.001$ for Test function 2, Figure 4.
- $\beta = 0.1$ and $\mu = 0.5$ for two-agent, Figure 5.

In general, by definition of geodesic convexity, the principle of choosing μ is to estimate a lower bound of the actual geodesic convexity parameter μ. Since when $\mu > \mu_0$, then the convexity inequality holds for μ_0 if it holds for μ. To achieve fast local convergence rate, one can choose $\beta \approx \frac{1}{5}\sqrt{\mu/L}$ according to [Zhang and Sra, 2018]. There is a further discussion on parameters in Appendix.

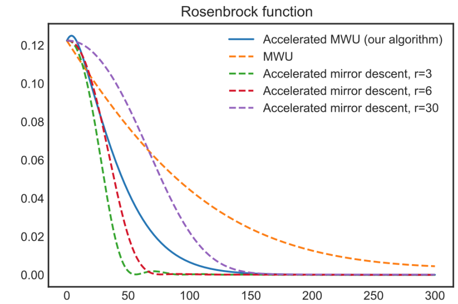
Efficiency in convergence. For the local convergence behavior, we use Rosenbrock function and Bohachevsky function as the test functions.

- The Rosenbrock function: $(0.5-x)^2 + 0.25(y-x^2)^2 + x + y + z - 1$, with initial points $(0.2, 0.4, 0.4)$, and the step sizes are 0.01 for A-MWU, MWU and A-MD, see Figure 1.
- The Bohachevsky function: $x^2 + 2y^2 - 0.3\cos(3\pi x) - 0.4\cos(4\pi y) + x + y + z - 1$, with initial point $(0.35, 0.3, 0.35)$, and step-size 0.001, see Figure 2.

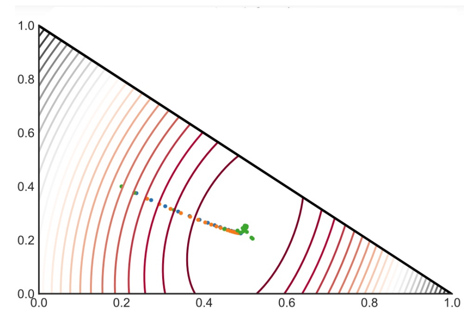
Compared to the usual definitions of 2-dimension Rosenbrock function and Bohachevsky function, our expression contains an additional term $x+y+z-1$ to make the function well defined in $\mathbb{R}^3$. But in simplices, $x+y+z-1=0$ always holds, thus our definition boils down to the original definitions.

Efficiency in escaping saddle points. We compare the curves of convergence and trajectories of A-MWU, A-MD and MWU for the function with many saddle points, in order to illustrate the different efficiency of three algorithms in escaping saddle points.

- Test function 1: $\cos(8.5x)\sin(8.5(y-0.4)) + \sin(8.5z)$, with initial point $(0.42, 0.24, 0.33)$, and the step-size is 0.005.
- Test function 2: $\cos(0.7x)\sin(y)\sin(0.9z) + x^2$, with initial point $(0.6, 0.2, 0.2)$, and the step-size is 0.05

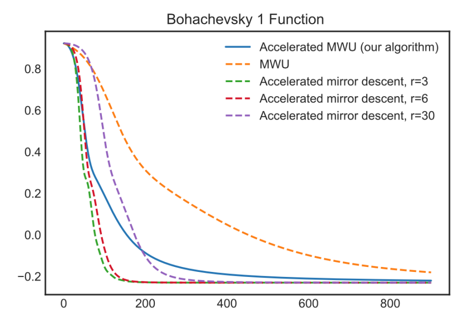


(a) A-MWU, MWU, A-MD, step-size=0.01

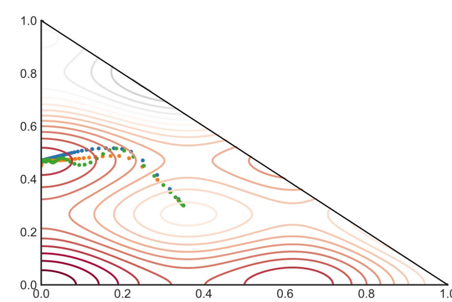


(b) Trajectories

Figure 1: Rosenbrock function.

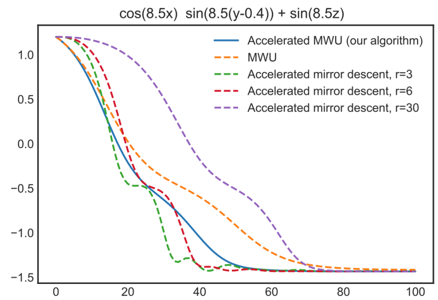


(a) A-MWU, MWU, A-MD, step-size=0.001

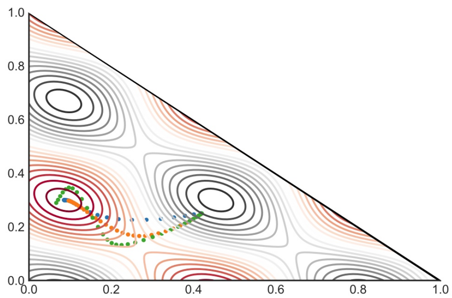


(b) Trajectories

Figure 2: Bohachevsky function

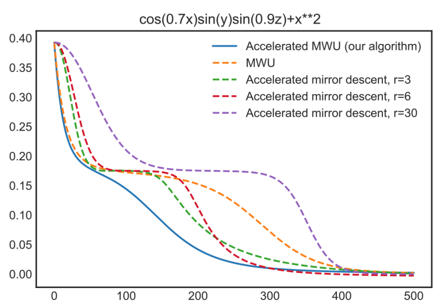


(a) A-MWU, MWU, A-MD, step-size=0.005

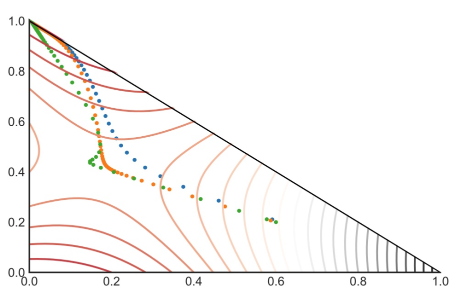


(b) Trajectories

Figure 3: Test function 1:$\cos(8.5x)\sin(8.5(y-0.4)) + \sin(8.5z)$

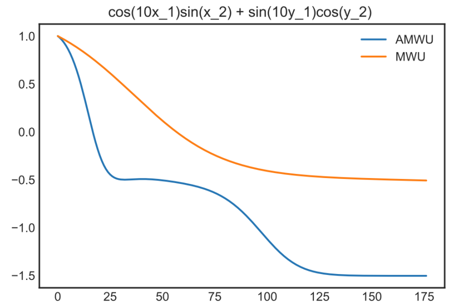


(a) A-MWU, MWU, A-MD, step-size=0.01

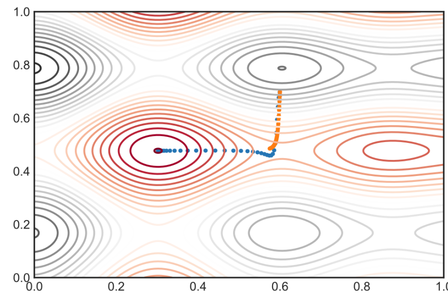


(b) Trajectories

Figure 4: Test function 2:$\cos(0.7x)\sin(y)\sin(0.9z) + x^2$



(a) A-MWU, MWU, step-size = 0.001



(b) Trajectories

Figure 5: Product of 1-simplices.

Multiple simplices. We give a final illustration on the behavior of A-MWU in two-agent case. Suppose $(x_1, x_2) \in [0,1] \times [0,1]$ and satisfies $x_1 + x_2 = 1$, $(y_1, y_2) \in [0,1] \times [0,1]$ and $y_1 + y_2 = 1$. The function is of the form $\cos(10x_1)\sin(x_2) + \sin(10y_1)\cos(y_2)$.

Experimental conclusions. The experimental results indicate the following:

- As expected, all experiments have verified that A-MWU has a better convergence behavior and efficiency in escaping saddle points compared to classic MWU.
- Compared to Accelerated Mirror Descent with entropy regularizer [Krichene *et al.*, 2015], the proposed algorithm A-MWU has the similar acceleration effect as A-MD with small r value, and outperforms A-MD with larger r value. See Figure 1 and 2.
- From Figure 3-(a) and Figure 4-(a), we believe that A-MWU has potential to outperform A-MD significantly in escaping saddle points.

6 Conclusion

In this paper we study an Accelerated Multiplicative Weights Update from the Riemannian geometric viewpoint. We prove that with locally geometric convexity in local minima, RAGD avoids saddle points with random initialization, which implies that A-MWU avoids saddle points. This indicates that A-MWU converges to local minima provided RAGD converges. Our experiments have verified the efficiency of A-MWU in escaping saddle points.

Acknowledgements

Yi Feng acknowledges Grant 202110458 from SUFE. Ioannis Panageas wants to thank a startup grant. Xiao Wang acknowledges Grant 202110458 from SUFE and support from the Shanghai Research Center for Data Science and Decision Technology.

References

[Arora *et al.*, 2012] Sanjeev Arora, lEad Hazan, and Satyen Kale. The multiplicative weights update method: a meta algorithm and applications. In *Theory of Computing*, 2012.

[Criscitiello and Boumal, 2019] Chris Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. In *NeurIPS*, 2019.

[Dekel *et al.*, 2011] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction. In *ICML*, 2011.

[Ge *et al.*, 2015] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *COLT*, 2015.

[Hofbauer and Sigmund, 1998] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[Jin *et al.*, 2017] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1724–1732, 2017.

[Jin *et al.*, 2018] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *COLT*, 2018.

[Juditsky *et al.*, 2011] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prod algorithm. *Stoch. Syst.*, 2011.

[Krichene *et al.*, 2015] Walid Krichene, Alexandre M. Bayen, and Peter L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *NIPS*, 2015.

[Lee *et al.*, 2016] Jason Lee, Max Simchowitz, Michael Jordan, and Benjamin Recht. Gradient descent only conveges to minimizers. In *COLT*, 2016.

[Lee *et al.*, 2019] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1-2):311–337, 2019.

[Lu *et al.*, 2019] Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, and Mingyi Hong. SNAP: finding approximate second-order stationary solutions efficiently for non-convex linearly constrained problems. *CoRR*, abs/1907.04450, 2019.

[Panageas and Piliouras, 2016] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.

[Panageas *et al.*, 2019a] Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In *NeurIPS*, 2019.

[Panageas *et al.*, 2019b] Ioannis Panageas, Georgios Piliouras, and Xiao Wang. Multiplicative weights updates as a distributed constrained optimization algorithm: Convergence to second-order stationary points almost always. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 4961–4969, 2019.

[Shahshahani, 1979] S Shahshahani. *A New Mathematical Framework for the Study of linage and Selection*, volume 17. American Mathematical Society, 1979.

[Sun *et al.*, 2019a] Tao Sun, Dongsheng Li, Zhe Quan, Hao Jiang, Shengguo Li, and Yong Dou. Heavy-ball algorithms always escape saddle points. In *IJCAI*, 2019.

[Sun *et al.*, 2019b] Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. In *NeurIPS*, 2019.

[Zhang and Sra, 2018] Hongyi Zhang and Suvrit Sra. Towards riemannian accelerated gradient methods. In *COLT*, 2018.