# Langevin Multiplicative Weights Update with Applications in Polynomial Portfolio Management

**Yi Feng** [*1], **Xiao Wang** [*† 1], **Tian Xie** [*2]

[1]School of Information Management & Engineering, Shanghai University of Finance and Economics
[2]College of Business, Shanghai University of Finance and Economics
fengyi95524@gmail.com, wangxiao@sufe.edu.cn, xietian@shufe.edu.cn

**Abstract**

We consider nonconvex optimization problem over simplex, and more generally, a product of simplices. We provide an algorithm, Langevin Multiplicative Weights Update (LMWU) for solving global optimization problems by adding a noise scaling with the non-Euclidean geometry in the simplex. Non-convex optimization has been extensively studied by machine learning community due to its application in various scenarios such as neural network approximation and finding Nash equilibrium. Despite recent progresses on provable guarantee of escaping and avoiding saddle point (convergence to local minima) and global convergence of Langevin gradient based method without constraints, the global optimization with constraints is less studied. We show that LMWU algorithm is provably convergent to interior global minima with a non-asymptotic convergence analysis. We verify the efficiency of the proposed algorithm in real data set from polynomial portfolio management, where optimization of a highly non-linear objective function plays a crucial role.

## 1 Introduction

In this paper we consider nonconvex 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 $f : \Delta_1 \times ... \times \Delta_N \to \mathbb{R}$ is a sufficiently smooth function and

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

Problem (1) appears naturally in potential game (Monderer and Shapley 1996), i.e., the incentive of all players to change their strategy can be expressed using a single global function (the potential function). A natural approach is to use projected gradient descent, but computing the projection at every iteration might not be an easy task to accomplish. An alternate effective algorithm in solving problem (1) is so called Multiplicative Weights Update (MWU) (Arora, lEad Hazan, and Kale 2012), which is a special case of FTRL that is commonly used in min-max optimization and multi-agent systems (Lei, Panageas, and Wang 2021; Feng et al. 2024; Feng, Piliouras, and Wang 2024). Result of (Panageas, Piliouras, and Wang 2019b) indicates that MWU almost always converges to second-order stationary points with random initialization. Besides MWU, many first-order methods have been proven escaping saddle points or avoiding saddle points asymptotically (Ge et al. 2015; Jin et al. 2017; Jin, Netrapalli, and Jordan 2018; Lee et al. 2016, 2019; Panageas and Piliouras 2016; Criscitiello and Boumal 2019; Sun, Flammarion, and Fazel 2019; Panageas, Piliouras, and Wang 2019a; Sun et al. 2019).

However, in the nonconvex world, finding local minima can be far away from achieving global minima. The classic MWU together with its accelerated variant (Feng, Panageas, and Wang 2022) can only converge to second-order stationary points or interior local minima, and this leaves finding global optima a challenging direction. One approach in designing first-order algorithm converging to global minima is to introduce a random noise into gradient descent, so that the algorithm has a chance to escape local minima. In recent years, progress has been made on this via the Langevin algorithm, an algorithm originally invented to sample from a target distribution proportional to $e^{-f(\mathbf{x})}$ where $f(\mathbf{x})$ is objective function defined on the whole Euclidean space $\mathbb{R}^n$. Successfully, global convergence of Langevin gradient descent with non-asymptotic convergence rate are obtained in (Raginsky, Rakhlin, and Telgarsky 2017; Xu et al. 2018). More recently, projected Langevin algorithm has been investigated in (Lamperski 2021) from the perspective of constrained sampling and optimization.

Despite aforementioned progresses in local and global convergence of gradient based algorithms, it is less understood whether there exists an algorithm that naturally fits distributed optimization framework from game theory and multi-agent systems. It is indicated in (Bailey and Piliouras 2019) that projected gradient descent can spend a few steps at each corner if the constraint has lower dimension. This feature makes projected gradient descent in multi-agent systems less effective, and in contrast, MWU and its variants have proven prominent in learning of games, their behaviors have been extensively studied in literatures, e.g., (Palaiopanos, Panageas, and Piliouras 2017; Bailey and Piliouras 2018; Cheung 2018; Cheung and Piliouras

[*]These authors contributed equally.
[†]Correspondence to Xiao Wang.

2019, 2020). Nevertheless, finding global minima of potential games with MWU or any of its variant seems missing in literature.

Motivated by global convergence analysis of Langevin gradient descent algorithm (Raginsky, Rakhlin, and Telgarsky 2017), we propose a scheme of adding noise that is scalable with a natural geometry of simplex, so that the Langevin Multiplicative Weights Update algorithm (LMWU) enjoys both the efficiency of projecting onto the constraint and the power of escaping saddle points and spurious local minima. LMWU is derived from the geometric Brownian motion on Riemannian manifold, where the natural geometry of simplex, i.e., Shahshahani geometry, plays a crucial role. The main result is stated as follows, and our contributions compared to the most relevant results in literature are illustrated in above table.

**Theorem 1.1** (Informal)**.** *Suppose the global optima of Problem* (1) *is in the interior of the constraints. The Langevin Multiplicative Weights Update converges to the biased global optima in expectation.*

**Other related works.** There have been considerably amount of works in convergence to local and global optima with first-order methods. Apart from the references listed in Table 1, we give a relatively complete review on the literatures about local and global convergence with gradient descent and Langevin algorithms. Local convergence guarantee with non-asymptotic convergence rate are investigated in (Ge et al. 2015; Jin et al. 2017; Jin, Netrapalli, and Jordan 2018; Sun, Flammarion, and Fazel 2019; He et al. 2024).Asymptotic convergence to local optima is studied with techniques from dynamical systems, typical references include (Lee et al. 2019, 2016; Antonakopoulos et al. 2022). On the other hand, convergence of Langevin algorithm in sampling has attracted many attentions. When the target distribution is log-concave, Euler discretization converges rapidly (Roberts and Tweedie 1996; Dalalyan 2017). Later on the convergence rate was improved in (Durmus and Moulines 2017). More recently, rapid convergence of Langevin algorithm for distributions satisfying log-Sobolev inequality has been established in (Vempala and Wibisono 2019; Li and Erdogdu 2020; Wang, Lei, and Panageas 2020; Gatmiry and Vempala 2022). An improved rate analysis for Langevin SGD with variance reduction is provided in (Kinoshita and Suzuki 2022) For sampling in a constrained set, Mirror Langevin diffusion has been studied in (Zhang et al. 2020; Hsieh et al. 2018; Aha and Chewi 2021; Jiang 2021; Li et al. 2021). A Reflected Langevin algorithm is proposed and analyzed in (Sato et al. 2022), we need to mention that the reflected operation has an projection operation embedded, which makes it difficulty to apply the algorithm in simplicial constraint.

## 2 Preliminaries

This section reviews the main background on Riemannian geometry and probability distributions on manifolds.

### 2.1 Riemannian Geometry

**Riemannian metric and exponential map.** A Riemannian manifold $(M, g)$ is real, smooth 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 $\gamma(1)$.

**Riemannian gradient.** 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.

$$\mathrm{grad} f(\mathbf{x}) = \left( \sum_j g^{1j}(\mathbf{x}) \frac{\partial f}{\partial x_j}, ..., \sum_j g^{dj}(\mathbf{x}) \frac{\partial f}{\partial x_j} \right) \quad (2)$$

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

**Retraction.** A retraction on a manifold $M$ is a smooth mapping $\mathrm{Retr}$ 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 $\alpha$ is given as

$$\mathbf{x}_{t+1} = \mathrm{Retr}_{\mathbf{x}_t}(-\epsilon \mathrm{grad} f(\mathbf{x}_t)). \quad (3)$$

### 2.2 Distributions on manifold

**KL divergence.** Let $\rho$ and $\nu$ be probability distributions on $M$ that is absolutely continuous with respect to the Riemannian volume measure on $M$ (denoted as $d\mathbf{x}$). The *Kullback-Leibler* (KL) divergence of $\rho$ with respect to $\nu$ is

$$H(\rho|\nu) = \int_M \rho(\mathbf{x}) \log \frac{\rho(\mathbf{x})}{\nu(\mathbf{x})} d\mathbf{x}$$

KL-divergence measures the "distance" between two probability distributions. Note that KL-divergene is nonnegative: $H(\rho|\nu) \geq 0$, and it is minimized at the target distribution, i.e., $H(\rho|\nu) = 0$ if and only if $\rho = \nu$. Furthermore, $\nu$ is the only stationary point of $H(\cdot|\nu)$, and thus sampling from $\nu$ can be reduced to minimizing $H(\cdot|\nu)$. Note that if $\nu = e^{-\beta f}$, the KL-divergence can be decomposed into

$$H(\rho|\nu) = \mathbb{E}_\rho f + \mathcal{H}(\rho),$$

where $\mathbb{E}_\rho f = \int_M \rho f d\mathrm{Vol}$ is the expected value of $f$ and $\mathcal{H}(\rho) = -\int_M \rho \log \rho d\mathrm{Vol}$ is the differential entropy of $\rho$.

**Wasserstein distance.** The Wasserstein distance between $\mu$ and $\nu$ is defined to be

$$\inf\{\sqrt{\mathbb{E}[d(X, Y)^2]} : \mathrm{law}(X) = \mu, \mathrm{law}(Y) = \nu\}$$

| | Global Convergence | Constraints | Simple Projection | Distributed Constraints |
|---|---|---|---|---|
| MWU (Panageas, Piliouras, and Wang 2019b) | ✗ | ✓ | ✓ | ✓ |
| Langevin GD (Raginsky, Rakhlin, and Telgarsky 2017) | ✓ | ✗ | ✗ | ✗ |
| Projected Langevin (Lamperski 2021) | ✓ | ✓ | ✗ | ✗ |
| Perturbed RGD (Criscitiello and Boumal 2019) | ✗ | ✓ | ✓ | ✗ |
| Accelerated MWU (Feng, Panageas, and Wang 2022) | ✗ | ✓ | ✓ | ✓ |
| Langevin MWU (this work) | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison to related results

**Log-Sobolev inequality.** A probability measure $\mu$ on $M$ is called to satisfy the logarithmic Sobolev inequality if there exists a constant $\alpha > 0$ such that

$$\int_M g^2 \log g^2 d\nu - \left(\int_M g^2 d\nu\right) \log\left(\int_M g^2 d\nu\right) \leq \frac{2}{\alpha} \int_M \|\mathrm{grad} g\|^2 d\nu \tag{4}$$

for all smooth functions $g : M \to \mathbb{R}$ with $\int_M g^2 \leq \infty$. The relative Fisher information of $\rho$ with respect to $\nu$ is $I_\nu(\rho) = \int_M \rho(\mathbf{x}) \left\|\mathrm{grad} \log \frac{\rho(\mathbf{x})}{\nu(\mathbf{x})}\right\| d\mathrm{Vol}$. Log-Sobolev inequality (LSI) is equivalent to the relation between KL-divergence and Fisher information: $H(\rho|\nu) \leq \frac{1}{2\alpha} I_\nu(\rho)$.

## 3 Main Results

In this section, we review classic Multiplicative Weights Update and its linear variant, and then some well known facts about Shahshahani geometry will be discussed. Based on the geometric setting of the simplex, we give a sketched framework how the Langevin Multiplicative Weights Update is derived.

### 3.1 From MWU to Langevin MWU

The classic Multiplicative Weights Update is widely used in constrained optimization, multi-agent system and game theory. It often refers to two forms,

$$x_{ij}(k+1) = \frac{x_{ij}(k) e^{-\epsilon \frac{\partial f}{\partial x_{ij}}}}{\sum_s x_{is}(k) e^{-\epsilon \frac{\partial f}{\partial x_{is}}}}$$

and its linear variant. If not specified, This paper refers MWU to the linear variant. 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 $\mathbf{x}(k)$ is the $k$-th iterate of MWU, the algorithm is written as follows:

$$x_{ij}(k+1) = x_{ij}(k) \frac{1 - \epsilon \frac{\partial f}{\partial x_{ij}}}{1 - \epsilon \sum_s x_{is}(k) \frac{\partial f}{\partial x_{is}}}, \tag{5}$$

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

It is well known that Langevin dynamics corresponds to the gradient flow of relative entropy respect to Wasserstein metric. In the space of measures with the Wasserstein metric, the gradient flow of relative entropy is the following partial differential equation, called Entropy Regularized Wasserstein Gradient Flow:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla f) + \beta^{-1} \Delta \rho$$

The key step in deriving Langevin Multiplicative Weights Update is to implement or approximate the noise scaled with the Shahshahani geometry in the simplex, which is a discretization of geometric Brownian motion in Shahshahani manifold. The geometric Brownian motion inside of the simplex $\Delta_+^{d-1} \subset \mathbb{R}_+^d$ can be obtained from the orthogonal projection of the geometric Brownian motion in $\mathbb{R}_+^d$, where the orthogonal projection is with respect to the Shahshahani metric in $\mathbb{R}_+^d$. Recall the standard Brownian motion in $\mathbb{R}^d$ is a random process $\{X_t\}_{t\geq 0}$ whose density function $\rho(\mathbf{x}, t)$ evolves according to the diffusion equation

$$\frac{\partial \rho(\mathbf{x}, t)}{\partial t} = \beta^{-1} \Delta \rho(\mathbf{x}, t).$$

The Brownian motion in Shahshahani manifold $\mathbb{R}_+^d$ is a random process $\{W_t\}_{t\geq 0}$ whose density function evolves according to the diffusion equation with respect to the Laplace-Beltrami operator, i.e.,

$$\frac{\rho(\mathbf{x}, t)}{\partial t} = \beta^{-1} \Delta_M \rho(\mathbf{x}, t).$$

Since $\mathbb{R}_+^d$ serves as its own local coordinate system as a Riemannian manifold, the geometric Brownian motion in $\mathbb{R}_+^d$ is described by the following stochastic differential equation

$$dX_t = -\beta^{-1} g^{ij} \Gamma_{ij}^k dt + \sqrt{2\beta^{-1} g^{-1}} dB_t$$

Algorithm 1: Langevin-MWU (single-agent)

Input : error threshold $\delta > 0$, large enough $\beta > 0$,
Compute step size $\epsilon < \frac{\delta^2 \alpha}{8C(\frac{M}{2}\sigma + B)}$,
Initialize $\mathbf{x}_0 \sim \rho_0$,
**repeat**
  Compute $S_{\mathbf{x}} = \sum_{j=1}^n \frac{1}{x_j}$, and $z_0^i \sim \mathcal{N}(0, 1)$.
  Compute
  $V_0^i = \frac{\epsilon}{2\beta}(n + 1 - (1 + x_i) S_{\mathbf{x}}) + \sqrt{2\epsilon\beta^{-1} x_i} z_0^i$
  Set $x_i \leftarrow \frac{x_i - \epsilon x_i \frac{\partial f}{\partial x_i} + V_0^i}{1 - \epsilon \sum_{j=1}^n x_j \frac{\partial f}{\partial x_j} + \sum_{j=1}^n V_0^j}$
**until** $k$ large enough, e.g., $k > \frac{16}{3\epsilon}\left(\frac{16(\frac{M}{2}\sigma + B)^2}{\delta_2 \alpha}\right)$

where $dB_t$ is the standard Brownian motion in Euclidean space, $g^{ij}$ is the $(ij)$-entry of the inverse matrix of Shahshahani metric matrix $g_{ij}$, and $\Gamma_{ij}^k$ is the Christoffel symbol of Shahshahani metric that can be calculated explicitly. After establishing the noise discretized from Shahshahani geometric Brownian, we combine the noise and the Riemannian gradient in $\mathbb{R}_+^d$ to finalize the incremental vector in the update rule. We leave the details in Appendix.

### 3.2 Main Theorem

In this section, we firstly state our main theorem that asserts the convergence in expectation of the L-MWU algorithm. Secondly, we will sketch the proof strategies, i.e., decomposition of error $\mathbb{E}f(\mathbf{x}_k) - f^*$ into

$$\mathbb{E}f(\mathbf{x}_k) - \mathbb{E}_\nu f + \mathbb{E}_\nu f - f^* \tag{6}$$

where the expectation $\mathbb{E}_\nu f = \int_M f(\mathbf{x})\nu(\mathbf{x})d\mathrm{Vol}$ and $f^*$ is the global minimum of $f(\mathbf{x})$ over $M$ and $\nu(\mathbf{x})$ is the probability density function that is proportional to $e^{-\beta f}$. We start presenting the main theorem by some a brief discussion on assumptions used in theoretical analysis.

Our analysis relies heavily on the theory of global convergence for Langevin algorithm in Euclidean space (Raginsky, Rakhlin, and Telgarsky 2017; Xu et al. 2018) and the results of rapid convergence results for log-Sobolev distributions such as (Gatmiry and Vempala 2022). Our strategy of giving theoretical analysis is to relate the assumptions in (Raginsky, Rakhlin, and Telgarsky 2017) to the case of Shahshahani manifold, and then generalize the arguments in Euclidean space to Riemannian manifold with special structure. The reason that one can generalize the results in Euclidean space to Shahshahani manifold is the possibility of geometrizing the analytic assumption on $f$ by identifying $\mathbf{0}$ in $\mathbb{R}^n$ with $\frac{1}{n}(1,...,1)$ in $\Delta^{n-1}$, and the interior of $\Delta^{n-1}$ is diffeomorphic to $\mathbb{R}^{n-1}$. We start by giving assumptions function $f$ satisfies.

**Assumption 1.** The function $f$ takes nonnegative real values, and there exist constants $A, B \geq 0$, such that

$$|f(\mathbf{1})| \leq A \text{ and } \|\mathrm{grad} f(\mathbf{1})\| \leq B.$$

This assumption comes from assumption (A.1) in (Raginsky, Rakhlin, and Telgarsky 2017) by relating $\mathbf{0}$ to $\mathbf{1} = \frac{1}{n}(1,...,1)$.

**Assumption 2.** Function $f$ is $M$-smooth for some $M > 0$, i.e.,

$$\|\mathrm{grad} f(\mathbf{y}) - \Gamma_{\mathbf{x}}^{\mathbf{y}} f(\mathbf{x})\| \leq M d(\mathbf{x},\mathbf{y}) \text{ for all } \mathbf{x},\mathbf{y} \in \mathcal{M},$$

where $\Gamma_{\mathbf{x}}^{\mathbf{y}}$ denotes the parallel transport from $\mathbf{x}$ to $\mathbf{y}$. The gradient satisfies

$$\|\mathrm{grad} f(\mathbf{x})\|_{\mathbf{x}} \leq \frac{M}{2} d(\mathbf{1},\mathbf{x}) + B$$

for some constants $M > 0$ and $B > 0$.

$M$-smoothness in Euclidean setting reads as $\|\nabla f(\mathbf{x}) - f(\mathbf{y})\| \leq M\|\mathbf{x}-\mathbf{y}\|$, which is commonly assumed in many theoretical analysis.

**Assumption 3.** There exist positive numbers $m$ and $b$ such that

$$\langle \mathrm{grad} f(\mathbf{x}), d(\mathbf{1},\mathbf{x})\mathbf{v}\rangle_{\mathbf{x}} \geq m d(\mathbf{1},\mathbf{x})^2 - b.$$

where $\mathbf{v}$ is the velocity vector of the geodesic connecting $\mathbf{1}$ and $\mathbf{x}$.

By a constant speed geodesic we mean the velocity has unit length everywhere, so the term $d(\mathbf{1},\mathbf{x})\mathbf{v}$ can be reduced to $\mathbf{x}$ in Euclidean space, where $\mathbf{x}$ means the geodesic of length $\|\mathbf{x}\|$ (straight line) connecting $\mathbf{0}$ and $\mathbf{x}$.

**Assumption 4.** The differential entropy of the distribution $e^{-\beta f}$ is bounded by a constant $K$.

In the case of Euclidean space, the differential entropy has an upper bound by estimating the second moment of Gibbs distribution (Raginsky, Rakhlin, and Telgarsky 2017). The differential entropy of a probability density with a finite second moment is upper-bounded by that of a Gaussian density with the same second moment, $h(\nu) \leq \frac{d}{2}\log\left(\frac{2\pi e(b+d/\beta)}{md}\right)$. Thus there exists an upper-bound for $\beta$ large enough.

**Assumption 5.** $e^{-\beta f}$ satisfies log-Sobolev inequality.

This condition is necessary in bounding the sampling algorithm converges rapidly.

**Theorem 3.1.** *Suppose $f$ satisfies Assumptions 1-5. Then there exists constant $C$, such that*

$$\begin{aligned}|\mathbb{E}f(\mathbf{x}_k) - f^*| \leq &(\frac{M}{2}\sigma + B)\sqrt{\frac{2}{\alpha}}\left(e^{-\frac{3}{16}\alpha\epsilon k} + \frac{C\epsilon}{\alpha}\right)^{\frac{1}{2}} \\ &+ \frac{K}{\beta} + \frac{1}{\beta}\log\left(poly(\beta^{-1})^{-1}\right). \end{aligned} \tag{7}$$

From the theorem we can conclude that for any given $\delta > 0$, there exists $\beta > 0$, $\epsilon < \frac{\delta^2\alpha}{8C(\frac{M}{2}\sigma+B)^2}$, and $k > \frac{16}{3\epsilon}\log\left(\frac{16(\frac{M}{2}\sigma+B)^2}{\delta^2\alpha}\right)$, such that $|\mathbb{E}f(\mathbf{x}_k) - f^*| \leq \delta$.

### 3.3 Outline of Proof

Suppose that the $k$'th iteration $\mathbf{x}_k$, which is a random variable on Shahshahani manifold $M$, has probability density function $\rho_k(\mathbf{x})$. Then the expectation $\mathbb{E}f(\mathbf{x}_k)$ can be written as

$$\int_M f(\mathbf{x})\rho_k(\mathbf{x})d\mathrm{Vol}$$

where $d\mathrm{Vol}$ is the Riemannian volume element induced by Shahshahani metric on $M$. Since the error $\mathbb{E}f(\mathbf{x}_k) - f^*$ has been decomposed into the sum of (6), we need to bound $|\mathbb{E}f(\mathbf{x}_k) - \mathbb{E}_\nu f|$ and $|\mathbb{E}_\nu f - f^*|$ respectively. By the integral on manifold we have the following:

$$\begin{aligned}|\mathbb{E}f(\mathbf{x}_k) - \mathbb{E}_\nu f| &= \left|\int_M f(\mathbf{x})\rho_k(\mathbf{x}) - \int_M f(\mathbf{x})\nu(\mathbf{x})\right| \\ &= \left|\int_M f(\mathbf{x})(\rho_k(\mathbf{x}) - \nu(\mathbf{x}))d\mathrm{Vol}\right| \end{aligned} \tag{8}$$

In Euclidean space, the difference is bounded by the Wasserstein distance between $\rho_k$ and $\nu$ according to Lemma 6 of (Raginsky, Rakhlin, and Telgarsky 2017), where the authors

prove that $\left|\int_{\mathbb{R}^d} g d\mu - \int_{\mathbb{R}^d} g d\nu\right| \leq \text{const} \cdot W_2(\mu,\nu)$, if $g$, $\mu$ and $\nu$ satisfy some assumptions. Therefore, our strategy of bounding $\left|\int_M f\rho_k d\text{Vol} - \int_M f\nu d\text{Vol}\right|$ relies on a generalized version of Lemma 6 of (Raginsky, Rakhlin, and Telgarsky 2017) in the case of Shahshahani manifold, if not for all Riemannian manifolds. Following this idea, we provide the following lemma.

**Lemma 3.1.** *Let $\mu$ and $\nu$ be two density function of probability measures on Shahshahani manifold $M$. Suppose $f : M \to \mathbb{R}$ satisfies*

$$\|\text{grad} f(\mathbf{x})\| \leq \frac{M}{2} d(\mathbf{1},\mathbf{x}) + B$$

*for some constants $\frac{M}{2} > 0$ and $B > 0$. Then*

$$\left|\int_M f\mu d\text{Vol} - \int_M f\nu d\text{Vol}\right| \leq (\frac{M}{2}\sigma + B) W_2(\mu,\nu)$$

*where $\sigma^2 = \int_M d(\mathbf{1},\mathbf{x})^2 \mu(\mathbf{x}) d\text{Vol} \vee \int_M d(\mathbf{1},\mathbf{y})^2 \nu(\mathbf{y}) d\text{Vol}$.*

Letting $\mu = \rho_k$, we can immediately obtain the expected result, i.e.,

$$\begin{aligned} |\mathbb{E} f(\mathbf{x}_k) - \mathbb{E}_\nu f| &= \left|\int_M f\rho_k d\text{Vol} - \int_M f\nu d\text{Vol}\right| \\ &\leq (\frac{M}{2}\sigma + B) W_2(\rho_k,\nu) \end{aligned} \tag{9}$$

Talagrand inequality is a well known connection between Wasserstein distance and KL-divergence. We say that a probability measure $\nu$ satisfies a Talagrand inequality with constant $\alpha > 0$ if for all probability measure $\rho$, absolutely continuous with respect to $\nu$, with finite moments of order 2, it holds that $W_2(\rho,\nu)^2 \leq \frac{2}{\alpha} H(\rho|\nu)$. Therefore, bounding $W_2(\rho_k,\nu)$ boils down to bounding the KL-divergence $H(\rho_k|\nu)$. It has been shown in (Gatmiry and Vempala 2022) that for general Hessian Manifold, Langevin algorithm for sampling from a log-Sobolev distribution converges rapidly to a distribution with bias $\epsilon$. Since simplex with Shahshahani metric is a Hessian manifold, applying Theorem 2 of (Gatmiry and Vempala 2022), we can immediately conclude that there exists a constant $C$ and log-Sobolev constant $\alpha$ such that $H(\rho_k|\nu) \leq e^{-\frac{3}{16}\alpha\epsilon k} + \frac{C\epsilon}{\alpha}$, and therefore

$$\begin{aligned} |\mathbb{E} f(\mathbf{x}_k) - \mathbb{E}_\nu f| &\leq (\frac{M}{2}\sigma + B)\sqrt{\frac{2}{\alpha}} H(\rho_k|\nu)^{\frac{1}{2}} \\ &\leq (\frac{M}{2}\sigma + B)\sqrt{\frac{2}{\alpha}} \left(e^{-\frac{3}{16}\alpha\epsilon k} + \frac{C\epsilon}{\alpha}\right)^{\frac{1}{2}}. \end{aligned} \tag{10}$$

To see that $\phi = \sum_{i=1}^n x_i \ln x_i$ induces a Hessian metric on simplex, let $x_n = 1 - \sum_{i=1}^{n-1} x_i$, then

$$\phi = \left(1 - \sum_{i=1}^{n-1} x_i\right) \ln\left(1 - \sum_{i=1}^{n-1} x_i\right).$$

The Hessian $\nabla^2\phi$ has the form of the following:

$$\begin{bmatrix} \frac{1}{x_1} + \frac{1}{1-\sum_{i=1}^{n-1} x_i} & \cdots & \frac{1}{1-\sum_{i=1}^{n-1} x_i} \\ \vdots & \ddots & \vdots \\ \frac{1}{1-\sum_{i=1}^{n-1} x_i} & \cdots & \frac{1}{x_{n-1}} + \frac{1}{1-\sum_{i=1}^{n-1} x_i} \end{bmatrix}. \tag{11}$$

On the other hand, the mapping $\varphi : (x_1,...,x_{n-1}) \to (x_1,...,x_{n-1}, 1-\sum_{i=1}^{n-1} x_i)$ from $\mathbb{R}^{n-1}$ to $\mathbb{R}^n$ induces a Riemannian metric in the projection of simplex, and this metric matrix $\langle d\varphi(\cdot), d\varphi(\cdot)\rangle$ is exactly the same as (11).

Running Langevin dynamics is equivalent to optimization in the space of probability densities in the underlying space (Wibisono 2018), and thus equivalent to sampling from the stationary distribution of the Wasserstein gradient flow asymptotically. To minimize $\int_M f(\mathbf{x})\rho(\mathbf{x})d\mathbf{x}$ with respect to $\rho(\mathbf{x})$, we introduce the entropy regularized functional of $\rho$ defined by $\mathcal{L}(\rho) = \mathcal{F}(\rho) + \beta^{-1}\mathcal{H}(\rho)$ where $\mathcal{F}(\rho) = \int_M f(\mathbf{x})\rho(\mathbf{x})d\mathbf{x}$, and $\mathcal{H}(\rho) = -\int_M \rho(\mathbf{x})\log\rho(\mathbf{x})d\mathbf{x}$. The Wasserstein space $\mathcal{P}_2(\mathcal{M})$ of probability measures on $\mathcal{M}$ is an infinite dimensional smooth Riemannian manifold. A tangent vector $R \in T_\rho\mathcal{M}$ is of the form $R = -\text{div}(\rho\text{grad}\phi)$ for some function $\phi : \mathcal{M} \to \mathbb{R}$. The gradient of a functional $\mathcal{L} : \mathcal{P} \to \mathbb{R}$ is $\text{grad}_\rho\mathcal{L} = -\text{div}\left(\rho\text{grad}\frac{\delta\mathcal{L}}{\delta\rho}\right)$, where $\frac{\delta\mathcal{L}}{\delta\rho}(\mathbf{x})$ is the first variation of $\mathcal{L}$ with respect to $\rho$. It is well known that the Wasserstein gradient flow of $\mathcal{L}$ is the Fokker-Planck equation

$$\begin{aligned} \frac{\partial\rho(\mathbf{x},t)}{\partial t} &= \text{div}(\rho(\mathbf{x},t)\text{grad} f(\mathbf{x}) + \beta^{-1}\text{grad}\rho(\mathbf{x},t)) \\ &= \text{div}(\rho(\mathbf{x},t)\text{grad} f(\mathbf{x})) + \beta^{-1}\Delta_M\rho(\mathbf{x},t), \end{aligned} \tag{12}$$

where grad, div and $\Delta_M$ are gradient, divergence and Laplace-Beltrami on manifolds. The stationary solution of equation (12) is the density proportional to $e^{-\beta f}$ that minimizes $\mathcal{L}$.

**Lemma 3.2.** *Suppose the entropy of distribution $\nu(\mathbf{x})$ is uniformly bounded for all $\beta$, i.e., $h(\nu) \leq K < \infty$. Then*

$$|\mathbb{E}_\nu f - f^*| \leq \frac{K}{\beta} + \frac{1}{\beta}\log\left(\textit{poly}\left(\frac{1}{\beta}\right)^{-1}\right).$$

Let $p(\mathbf{x}) = \frac{e^{-\beta f(\mathbf{x})}}{\Lambda}$ denote the density of the Gibbs measure with respect to the measure induced by the Shahshahani metric in simplex, where $\Lambda := \int_M e^{-\beta f(\mathbf{x})} d\mathbf{x}$ is the normalization constant known as the partition function. Note that the differential entropy of $p$ has the following expression,

$$h(p) = \frac{1}{\Lambda}\int_M \beta f(\mathbf{x}) e^{-\beta f(\mathbf{x})} d\text{Vol} + \log\Lambda$$

thus we have that

$$\int_M f(\mathbf{x})p(\mathbf{x})d\text{Vol} = \frac{1}{\beta}(h(p) - \log\Lambda).$$

Let $\mathbf{x}^*$ be any point that minimizes $f(\mathbf{x})$. Then $\text{grad} f(\mathbf{x}^*) = 0$. Since $f$ is assumed to be geodesically smooth, we have

$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{M}{2} d(\mathbf{x}, \mathbf{x}^*)^2$, the lower bound of $\log \Lambda$ can be obtained by following calculation,

$$\begin{aligned}\log \Lambda &= \log \int_M e^{-\beta f(\mathbf{x})} d\text{Vol} \\ &= -\beta f(\mathbf{x}^*) + \log \int_M e^{-\beta (f(\mathbf{x}^*) - f(\mathbf{x}))} d\text{Vol} \\ &\geq -\beta f(\mathbf{x}^*) + \log \int_M e^{-\beta d(\mathbf{x}, \mathbf{x}^*)^2/2} d\text{Vol} \qquad (13)\end{aligned}$$

Without loss of generality, we can assume that the global minima $\mathbf{x}^*$ is at the center of simplex, i.e., $\mathbf{x}^* = \mathbf{1} = \left(\frac{1}{n}, ..., \frac{1}{n}\right)$. In appendix, we show that the integral $\int_M e^{-cd(\mathbf{1}, \mathbf{x})^2} d\text{Vol}$ is bounded. By letting $c = \beta \frac{M}{2}$, we furthermore end up with a concrete expression of $\int_M e^{-cd(\mathbf{1}, \mathbf{x})^2} d\text{Vol}$ in terms of a polynomial of $\beta^{-1}$, which is denoted briefly as follows,

$$\log \Lambda \geq -\beta f(\mathbf{x}^*) + \log \left(\text{poly}\left(\beta^{-1}\right)\right),$$

and then we have

$$-f(\mathbf{x}^*) \leq \frac{\log \Lambda}{\beta} + \frac{1}{\beta} \log \left(\text{poly}\left(\beta^{-1}\right)^{-1}\right).$$

Combining with $\mathbb{E}_\nu f = \frac{h(\nu)}{\beta} - \frac{\log \Lambda}{\beta}$, we have the following bound:

$$\mathbb{E}_\nu f - f(\mathbf{x}^*) \leq \frac{K}{\beta} + \frac{1}{\beta} \log \left(\text{poly}\left(\beta^{-1}\right)^{-1}\right).$$

## 4 Application in Portfolio Management

Portfolio management is a critical aspect of finance as it facilitates the efficient and effective management of investments to achieve specific financial goals and objectives. It involves the careful selection, diversification, and alignment of various financial instruments such as stocks, bonds, and other assets, to balance risk and returns according to an individual or institution's risk tolerance, time horizon, and investment objectives. The strategic allocation of assets in a portfolio can enhance returns, mitigate potential losses, and provide a smoother investment journey. Moreover, portfolio management offers a structured approach to monitor, review, and adjust investments in response to changing market conditions, personal circumstances, or shifts in financial goals, making it an indispensable tool for successful financial planning and wealth management.

The polynomial portfolio optimization problem can be formally represented as

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathcal{W}}{\mathbf{argmin}}\, \mathbb{E}[f(\mathbf{w}, \mathbf{r})],$$

where $\mathbb{E}$ denotes the expectation operator, $f(\mathbf{w}, \mathbf{r})$ refers to a polynomial loss function, and $\mathbf{r} = [r_1, r_2, ..., r_n]^\top$ symbolizes the vector of $n$ individual returns within the portfolio. Additionally, $\mathbf{w} = [w_1, w_2, ..., w_n]^\top$ signifies the weights designated to each constituent element of the portfolio. It's important to note that $\mathbf{w}$ is restricted to the feasible set $\mathcal{W}$, $\mathcal{W} \equiv \{\mathbf{w} \in \mathbb{R}^N : \sum_{i=1}^N w_i = 1\}$, which constrains the summation of the weights to be one. This constraint implies that no leveraging or borrowing is permitted in the portfolio construction.

We propose a specific formulation for the loss function $f(\mathbf{w}, \mathbf{r})$ as follows:

$$\begin{aligned}f(\mathbf{w}, \mathbf{r}) = -\lambda_1 m_1(\mathbf{w}, \mathbf{r}) + \lambda_2 m_2(\mathbf{w}, \mathbf{r}) + ... \\ + (-1)^d \lambda_d m_d(\mathbf{w}, \mathbf{r}), \qquad (14)\end{aligned}$$

where $m_1(\mathbf{w}, \mathbf{r}) = \mathbf{w}^\top \mathbf{r}$ represents the sample portfolio return, and

$$m_i(\mathbf{w}, \mathbf{r}) = \Big(m_1(\mathbf{w}, \mathbf{r}) - \mathbb{E}\big(m_1(\mathbf{w}, \mathbf{r})\big)\Big)^i, \quad i = 2, ..., d$$

encapsulates the $i^{th}$ central moment of $m_1(\mathbf{w}, \mathbf{r})$, with $\mathbb{E}\big(m_i(\mathbf{w}, \mathbf{r})\big)$ being the expected value. The parameter vector $\lambda = [\lambda_1, ..., \lambda_d]^\top$ contains the risk preference parameters, each satisfying $\lambda_i \geq 0$, and their summation amounts to one, i.e., $\sum_{i=1}^d \lambda_i = 1$. It's worth noting that the mean-variance (MV), mean-variance-skewness (MVS), and mean-variance-skewness-kurtosis (MVSK) losses can be considered specific instances of this general polynomial portfolio optimization framework.

Our dataset comprises daily entries for $n = 10$ notable NASDAQ stocks, covering the period from January 3, 2011, to December 31, 2021, and thereby accumulating $T = 2517$ periods. We initiate a rolling-window out-of-sample forecasting exercise from the beginning of this data sample. The window length is set at $L = 1000$, approximately corresponding to four years of training data. To calculate the optimal portfolio weights, $\hat{\mathbf{w}}$, we implement four estimation strategies: the traditional Multiplicative Weight Update (MWU) approach, the accelerated MWU algorithm purposed in (Feng, Panageas, and Wang 2022), the projected langevin gradient descent algorithm purposed in (Lamperski 2021) and our newly proposed Langevin Multiplicative Weights Update (LMWU) method.

Following this, we apply the estimated portfolio weights to the returns in the succeeding period and assess the performance of the constructed portfolio using the loss function defined in Equation (14). It's crucial to note that our loss function relies on a predetermined parameter, $\boldsymbol{\lambda}$, which represents different risk preferences. We take into account the following potential values for $\boldsymbol{\lambda}$:

1. Increasing preference: $\frac{1}{15}, \frac{2}{15}, ..., \frac{5}{15}$
2. Degenerate preference: $\frac{5}{15}, \frac{4}{15}, ..., \frac{1}{15}$
3. Mean-Variance (MV) preference: $\frac{1}{2}, \frac{1}{2}, 0, 0, 0$
4. Mean-Variance-Skewness (MVS) preference: $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0$
5. Mean-Variance-Skewness-Kurtosis (MVSK) preference: $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0$
6. Equal preference: $\frac{1}{5}, \frac{1}{5}, ..., \frac{1}{5}$

For each period $t$, we record the loss score, denoted as $\widehat{\text{Loss}}_t$, and compute the average loss score using the formula: $\widehat{\text{Score}} = \frac{1}{T-L} \sum_{t=T-L+1}^{T} \widehat{\text{Loss}}_t$. The outcomes are summarized in Table 2. The table's first column delineates the methods employed in the exercise, while columns two

| Method | Degenerate | Increasing | MV | MVS | MVSK | Equal |
|---|---|---|---|---|---|---|
| MWU | 74.7203 | 17.4527 | 0.8561 | 0.8391 | 16.1104 | 44.5159 |
| AMWU (Feng, Panageas, and Wang 2022) | 76.3657 | 17.5554 | 0.8561 | 0.8579 | 15.6559 | 43.4190 |
| Projected Langevin (Lamperski 2021) | 73.9596 | 17.8720 | 0.8674 | 0.8846 | 16.7519 | 43.2847 |
| LMWU (this work) | 70.8930 | 16.7684 | 0.8464 | 0.8314 | 14.9585 | 42.8307 |

Table 2: Out-of-sample Evaluation Results for Polynomial Portfolio Optimization

through seven display the results of these methods under various risk preferences, as indicated in the header row.

The data clearly demonstrates that the LMWU method outperforms the MWU method and several of its other variants across all risk preferences. For example, under the Degenerate preference, the LMWU method registers a score of 70.8930, a better result (considering the goal is minimization) than the MWU's score of 74.7203. This superiority is consistent across other risk preferences as well. Specifically, for Mean-Variance (MV) and Mean-Variance-Skewness (MVS) preferences, which are likely more commonplace in portfolio management, the LMWU method achieves superior scores (0.8464 and 0.8314, respectively) compared to the MWU method (0.8561 and 0.8391, respectively). Similar better performance can also be observed with Langevin MWU compared to other variants of MWU algorithms. These observations are consistent with our theoretical analysis of LMWU: it has the ability to escape local minima and converge towards global minima.

In summary, these results underscore the efficacy of our proposed LMWU method in the realm of polynomial portfolio optimization.

## 5 Additional Experiments

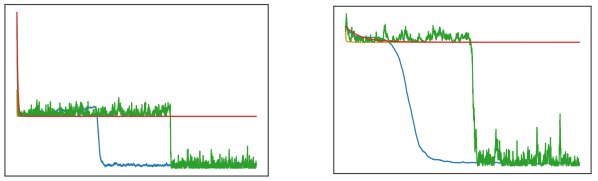
In this section, we present experiments comparing Langevin-MWU with algorithms presented in Table 1. We use Langevin-MWU and other algorithms for comparison to optimize several non-convex functions with many local minima. The experimental results show Langevin-MWU escapes such bad local minima and finds minima with smaller function values, while other algorithms either get stuck at local minima or are more unstable than Langevin-MWU. The experimental results are presented in Figure 1 and Figure 2. Future experiments, especially the examples demonstrating how the trajectories of Langevin-MWU avoid local minima and converge to global minima, are presented in the Appendix.

**Test functions.** We construct non-convex functions to verify the efficiency of LMWU in finding global minima. The functions are given as follows:

$$f_1(x,y,z) = -\ln\left(e^{-10(x-0.3)^2-20(y-0.5)^2-30(z-0.2)^2} + e^{-30(x-0.4)^2-20(y-0.2)^2-36(z-0.4)^2}\right) + y + 10.$$

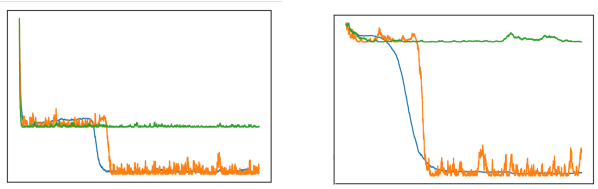
and

$$f_2(x,y,z) = -(x-0.6)^2(x-0.2)^2 + (y-0.3)(y-0.4)^3 + (z-0.2)^3(z-0.8) - xy - 0.4z.$$



(a) Test function : $f_1$ (b) Test function : $f_2$

Figure 1: Comparison of LMWU with Accelerated MWU (Feng, Panageas, and Wang 2022), Projected Lagevin (Lamperski 2021), and PRGD (Criscitiello and Boumal 2019).



(a) Test function : $f_1$ (b) Test function : $f_2$

Figure 2: Future comparison of LMWU with Projected Langevin (Lamperski 2021).

As shown in Figure 1 and 2, LMWU and Projected Langevin can converge to global optima, but Perturbed RGD and Accelerated MWU only converge to local optima, which agree with the claims of (Feng, Panageas, and Wang 2022) and (Criscitiello and Boumal 2019).

## 6 Conclusion

In this paper we focus on a constrained non-convex optimization problem that widely exists in multi-agent learning. We propose a novel algorithm called Langevin Multiplicative Weights Update (LMWU) which is a stochastic version of classic MWU algorithm. Our theoretical analysis shows that LMWU converges to interior global optima of the objective function. Another important setting that is missing in current work is the time-varying environment, e.g., (Feng et al. 2023) in min-max optimization. We leave the time-varying portfolio management for future investigation.

## Acknowledgements

Xiao Wang acknowledges Grant 202110458 from Shanghai University of Finance and Economics and support from the Shanghai Research Center for Data Science and Decision Technology. Xie's research is supported by the Natural Science Foundation of China (72173075) and the Shanghai Research Center for Data Science and Decision Technology.

## References

Aha, K.; and Chewi, S. 2021. Efficient constrained sampling via the mirror-Langevin algorithm. In *NeurIPS*.

Antonakopoulos, K.; Mertikopoulos, P.; Piliouras, G.; and Wang, X. 2022. AdaGrad Avoids Saddle Points. In *ICML*.

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

Bailey, J.; and Piliouras, G. 2018. Multiplicative weights update in zero-sum games. In *EC*.

Bailey, J. P.; and Piliouras, G. 2019. Fast and Furious Learning in Zero-sum Games: Vanishing Regret with Non-vanishing Step Sizes. In *NeurIPS*.

Cheung, Y. K. 2018. Multiplicative Weights Update with Constant Step-size in Graphical Constant-sum Games. In *NeurIPS*.

Cheung, Y. K.; and Piliouras, G. 2019. Vortices instead of equilibria in minmax optimization:chaos and butterfly effects of online learning in zero-sum games. In *COLT*.

Cheung, Y. K.; and Piliouras, G. 2020. Chaos, Extremism and Optimism: Volume Analysis of Learning in Games. In *NeurIPS*.

Criscitiello, C.; and Boumal, N. 2019. Efficiently escaping saddle points on manifolds. In *NeurIPS*.

Dalalyan, A. 2017. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society:Series B*.

Durmus, A.; and Moulines, E. 2017. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Annals of Applied Probability*.

Feng, Y.; Fu, H.; Hu, Q.; Li, P.; Panageas, I.; and Wang, X. 2023. On the last-iterate convergence in time-varying zero-sum games: Extra gradient succeeds where optimism fails. In *NeurIPS*.

Feng, Y.; Li, P.; Panageas, I.; and Wang, X. 2024. Last-iterate Convergence Separation between Extra-gradient and Optimism in Constrained Periodic Game. In *UAI*.

Feng, Y.; Panageas, I.; and Wang, X. 2022. Accelerated Multiplicative Weights Update Avoid Saddle Points Almost Always. In *IJCAI*.

Feng, Y.; Piliouras, G.; and Wang, X. 2024. Prediction Accuracy of Learning in Games: Follow-the-Regularized-Leader meets Heisenberg. In *ICML*.

Gatmiry, K.; and Vempala, S. 2022. Convergence of the Riemannian Langevin Algorithm. In *arXiv:2204.10818*.

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

Gray, A. 1974. The volume of a small geodesic ball of a Riemannian manifold. *Michigan math. J.*

He, C.; Pan, Z.; Wang, X.; and Jiang, B. 2024. Riemannian Accelerated Zeroth-order Algorithm: Improved Robustness and Lower Query Complexity. In *ICML*.

Hofbauer, J.; and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Hsieh, Y.-P.; Kavis, A.; Rolland, P.; and Cevher, V. 2018. Mirrored Langevin dynamics. In *NeurIPS*.

Hsu, E. P. 2002. *Stochastic Analysis on Manifolds*. American Mathematical Society.

Jiang, Q. 2021. Mirror Langevin Monte Carlo:the case under isoperimetry. In *NeurISP*.

Jin, C.; Ge, R.; Netrapalli, P.; Kakade, S. M.; and Jordan, M. I. 2017. 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*, 1724–1732.

Jin, C.; Netrapalli, P.; and Jordan, M. I. 2018. Accelerated gradient descent escapes saddle points faster than gradient descent. In *COLT*.

Kinoshita, Y.; and Suzuki, T. 2022. Improved Convergence Rate of Stochastic Gradient Langevin Dynamics with Variance Reduction and its Application to Optimization. In *NeurIPS*.

Lamperski, A. 2021. Projected Stochastic Gradient Langevin Algorithms for Constrained Sampling Nonconvex Learning. In *COLT*.

Lee, J. 2018. *Introduction to Riemannian Manifolds*, volume 176 GTM. Springer.

Lee, J.; Simchowitz, M.; Jordan, M.; and Recht, B. 2016. Gradient descent only conveges to minimizers. In *COLT*.

Lee, J. D.; Panageas, I.; Piliouras, G.; Simchowitz, M.; Jordan, M. I.; and Recht, B. 2019. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1-2): 311–337.

Lei, Q.; Panageas, S. G. N. I.; and Wang, X. 2021. Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscape. In *AISTATS*.

Li, M. B.; and Erdogdu, M. A. 2020. Riemannian Langevin algorithm for solving semidefinite programs. In *arXiv:2021.11176*.

Li, R.; Tao, M.; Vempala, S. S.; and Wibisono, A. 2021. The mirror Langevin Algorithm Converges with vanishing bias. In *arXiv:2109.12077*.

Monderer, D.; and Shapley, L. 1996. Potential Games. *Games and Economic Behavior*.

Palaiopanos, G.; Panageas, I.; and Piliouras, G. 2017. Multiplicative Weights Update with Constant Step-size in Congestion Games: Convergence, Limit Cycles and Chaos. In *NeurIPS*.

Panageas, I.; and Piliouras, G. 2016. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*.