Non-Asymptotic Convergence Analysis of the Multiplicative Gradient Algorithm for the Log-Optimal Investment Problems∗

Renbo Zhao†
September 15, 2021

Abstract
We analyze the non-asymptotic convergence rate of the multiplicative gradient (MG) algorithm for the log-optimal investment problems, and show that it exhibits $O(1/t)$ convergence rates, in both ergodic and non-ergodic senses.

1 Introduction
Let us consider the following optimization problem:

$$(P) : \max_{x \in \Delta_n} \left\{ f(x) := \sum_{j=1}^{m} p_j \ln(a_j^\top x) \right\},$$

where $\Delta_n := \{ x \in \mathbb{R}^n : x \geq 0, \sum_{i=1}^{n} x_i = 1 \}$ denotes the probability simplex in $\mathbb{R}^n$, $p_j > 0$ for all $j \in [m]$, $a_j \in \mathbb{R}_+^n$, for all $j \in [m]$ and $\mathbb{R}_+^n := \{ x \in \mathbb{R}^n : x_i \geq 0, \forall i \in [n] \}$, namely the nonnegative orthant in $\mathbb{R}^n$. Without loss of generality, we may assume that $\sum_{j=1}^{m} p_j = 1$. Furthermore, we assume that $a_j \neq 0$ for all $j \in [m]$ so that $\text{dom } f \cap \Delta_n \neq \emptyset$ and hence $(P)$ has an optimal solution.

The problem $(P)$ subsumes many diverse applications, including computing the rate distortion in information theory [1], the positron emission tomography problem in medical imaging [2], maximum likelihood estimation for mixture models in statistics [3] and the log-optimal investment problem [4]. In the following, we will briefly describe the last application due to its simplicity.

In the log-optimal investment problem, we consider $n$ stocks in the market, and let $R_i$ denote the random per-unit capital return on investing in stock $i$, for $i \in [n]$. The random vector $R := (R_1, \ldots, R_n)$ has unknown distribution $P$. An investor allocates her investment capital over these $n$ stocks, and let $x_i$ denote the (nonnegative) proportion of capital invested in stock $i$, whereby $w_i \geq 0$ for all $i \in [n]$ and $\sum_{i=1}^{n} x_i = 1$. Define $x := (x_1, \ldots, x_n)$. The goal of the investor is to maximize her expected log return $\mathbb{E}_{R \sim P}[\ln(x^\top R)]$ subject to $x \in \Delta_n$. From Cover [4], the naturalness of this objective can be justified from several perspectives involving the principle that money compounds multiplicatively rather than additively. Since $P$ is unknown, one can use a (historical) data-driven empirical distribution such as $\hat{P}_m := \sum_{j=1}^{m} p_j \delta_{r_j}$, where $p_j > 0$, $\sum_{j=1}^{m} p_j = 1$, $r_j \in \mathbb{R}^n$ is a realization of $R$ and $\delta_{r_j}$ denotes the unit point mass at $r_j$ for $j \in [m]$. Under this empirical distribution, the investor instead solves the problem:

$$\min_{x \in \Delta_n} - \sum_{j=1}^{m} p_j \ln(r_j^\top x). \tag{1}$$

∗Research supported by AFOSR Grant No. FA9550-19-1-0240.
†MIT Operations Research Center, 77 Massachusetts Avenue, Cambridge, MA 02139 (mailto: renboz@mit.edu).
One of the earliest algorithms for solving (P) is the multiplicative gradient (MG) algorithm: Given \( x^0 \in \text{ri} \Delta_n = \{ x \in \mathbb{R}^n : x > 0, \sum_{i=1}^n x_i = 1 \} \), at each iteration \( t \geq 0 \), we have

\[
x_i^{t+1} = x_i^t \nabla_i f(x^t), \quad \forall i \in [n], \quad \text{where} \quad \nabla_i f(x^t) := \sum_{j=1}^m p_j \frac{a_{ji}}{a_j^x} x^t
\]

(2)
denotes the \( i \)-th entry of \( \nabla_i f(x^t) \), and \( a_{ji} \) denotes the \( i \)-th entry of \( a_j \). As a sanity check, note that if \( x^t \in \Delta_n \), then we always have \( x^{t+1} \in \Delta_n \) since \( \nabla_i f(x^t) \geq 0 \) and hence \( x_i^{t+1} \geq 0 \) for all \( i \in [n] \), and

\[
\sum_{i=1}^n x_i^{t+1} = \sum_{i=1}^n x_i^t \sum_{j=1}^m p_j \frac{a_{ji}}{a_j^x} x^t = \sum_{j=1}^m p_j = 1.
\]

(3)

In addition, note that if the MG algorithm terminates finitely, meaning that \( x^{t+1} = x^t \) for some \( t \geq 0 \), then \( \nabla f(x^t) = e \), where \( e \) denotes the vector with all unit entries of proper dimension. Since the constraint set is the probability simplex \( \Delta_n \), we have \( \langle \nabla f(x^t), x - x^t \rangle = 0 \) for all \( x \in \Delta_n \) and hence \( x^t \) is an optimal solution for (P).

The MG algorithm is one of the most popular methods for solving (P), mainly for two reasons. The first second is simplicity. Unlike many other first-order methods, it does not require selecting step-sizes or solving projection or linear minimization subproblems. The second reason is good empirical performance. Indeed, as noted in some recent references (e.g., [5, 6]), the MG algorithm empirically runs faster compared to several standard first-order methods, including the Frank-Wolfe method [6, 7], the Bregman proximal gradient method [8, 9] and the primal-dual hybrid gradient method [10].

Despite the popularity and empirical efficacy of the MG algorithm, to the best of our knowledge, the non-asymptotic convergence guarantees have not been given in the literature, in contrast to the well-understanding to its asymptotic convergence properties (see e.g., [2, 4, 11, 12]). In this work, we provide a simple proof showing that the MG algorithm has \( O(1/t) \) convergence rate, in both ergodic and non-ergodic senses.

## 2 Convergence Analysis

Our analysis is based on the following simple but important observations about (P). Throughout the analysis, let \( x^* \) be any optimal solution of (P).

**Lemma 1.** There exist \( \nu \in \mathbb{R}_+^n \) such that
\[
\nu_i x_i^* = 0 \quad \text{and} \quad \nabla_i f(x^*) + \nu_i = 1, \quad \forall i \in [n].
\]

In particular, if \( x_i^* > 0 \) for some \( i \in [n] \), then \( \nabla_i f(x^*) = 1 \).

**Proof.** First, let us observe that for any \( x \in \Delta_n \),

\[
\langle \nabla f(x), x \rangle = \sum_{i=1}^n x_i \sum_{j=1}^m p_j \frac{a_{ji}}{a_j^x} x = \sum_{j=1}^m p_j = 1,
\]

(4)

where \( a_{ji} \) denotes the \( i \)-th entry of \( a_j \). (Note that (4) can also been seen from that \( -f \) is \( 1 \)-logarithmically homogeneous; see e.g., [13, Section 2.3.3].) Using KKT conditions, we know there exist \( \lambda \in \mathbb{R} \) and \( \nu \in \mathbb{R}_+^n \) such that \( \nu_i x_i^* = 0 \) for all \( i \in [n] \) and \( \nabla f(x^*) + \lambda e + \nu = 0 \), where recall that \( e \) denotes the vector with all unit entries. Consequently,

\[
1 + \lambda = \langle \nabla f(x^*), x^* \rangle + \lambda \langle e, x^* \rangle + \langle \nu, x^* \rangle = 0 \quad \implies \quad \lambda = -1.
\]
This implies that $\nabla f(x^\ast) + \nu = e$. \hfill \Box

The next lemma lower bounds the improvement of the objective value at each iteration of the MG algorithm. For completeness, we provide a proof of Lemma 2 in Appendix A.

**Lemma 2** (Cover [4, Theorem 1]). Let $h(x) := \sum_{i=1}^n x_i \ln x_i - x_i$ for $x \geq 0$ and let

\[
D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x \rangle \quad (\forall y \geq 0, \forall x > 0)
\]

denote the Bregman divergence induced by $h$. Then for all $t \geq 0$, we have

\[
f(x^{t+1}) - f(x^t) \geq D_h(x^{t+1}, x^t) \geq 0.
\]

Note that for any $y \in \Delta_n$ and $x \in \text{ri} \Delta_n$, we have

\[
D_h(y, x) = \sum_{i=1}^n y_i \ln \left( \frac{y_i}{x_i} \right),
\]

namely the Kullback-Leibler (KL) divergence between $y$ and $x$. Now, for convenience, define the optimality gap at $x \in \Delta_n$ as $\delta(x) := f^\ast - f(x)$, and from Lemma 2, we see that $\{\delta(x^t)\}_{t\geq 0}$ is a monotonically non-increasing sequence.

**Lemma 3.** For any $x \in \Delta_n$, we have

\[
\delta(x) \leq \sum_{i=1}^n x^*_i \ln (\nabla_i f(x)).
\]

**Proof.** Define $I := \{i \in [n] : x^*_i > 0\}$, and from Lemma 1, we see that

\[
\sum_{j=1}^m p_j a_{ji} \frac{1}{x^*_j} = \nabla_i f(x^*) = 1.
\]

As a result,

\[
\sum_{i=1}^n x^*_i \ln (\nabla_i f(x)) = \sum_{i \in I} x^*_i \ln (\nabla_i f(x))
\]

\[
= \sum_{i \in I} x^*_i \ln \left( \sum_{j=1}^m \frac{p_j a_{ji}}{x^*_j} \frac{a^*_j x^*}{a^*_j x} \right)
\]

\[
\geq \sum_{i \in I} x^*_i \sum_{j=1}^m \frac{p_j a_{ji}}{a^*_j x^*} \ln \left( \frac{a^*_j x^*}{a^*_j x} \right)
\]

\[
= \sum_{j=1}^m p_j \ln \left( \frac{a^*_j x^*}{a^*_j x} \right)
\]

\[
= f(x^*) - f(x) = \delta(x),
\]

the inequality follows from concavity of $\ln(\cdot)$ and (7). \hfill \Box
Equipped with the lemmas above, we now prove the $O(1/t)$ convergence rate of the MG algorithm, in both ergodic and non-ergodic senses.

**Theorem 1.** For all $t \geq 0$, we have

1. Ergodic rate: $\delta(\bar{x}^t) \leq \frac{D_h(x^*, x^0)}{t+1}$, where $\bar{x}^t := (t+1)^{-1} \sum_{k=0}^t x^k$.

2. Non-ergodic rate: $\delta_t \leq \frac{D_h(x^*, x^0)}{t+1}$.

**Proof.** We use $D_h(x^*, x^t)$ as the Lyapunov function. Note that from (5) we have

$$D_h(x^*, x^t) - D_h(x^*, x^{t+1}) = \sum_{i=1}^n x_i^t \ln \left( \frac{x_i^{t+1}}{x_i^t} \right) = \sum_{i=1}^n x_i^t \ln (\nabla_i f(x^t)) \geq \delta(x^t),$$

where the last step follows from Lemma 3. Telescoping over $i = 0, \ldots, t$, we have

$$D_h(x^*, x^0) \geq D_h(x^*, x^0) - D_h(x^*, x^t) \geq \sum_{k=0}^t \delta(x^k).$$

Using the convexity of $\delta(\cdot)$, we have $\sum_{k=0}^t \delta(x^k) \geq (t+1)\delta(\bar{x}^t)$, and this proves the first part. Alternatively, using the non-increasing property of $\{\delta(x^t)\}_{t\geq 0}$, we have $\sum_{k=0}^t \delta(x^k) \geq (t+1)\delta(x^t)$, and this proves the second part. 

---

**A Proof of Lemma 2**

We have

$$f(x^{t+1}) - f(x^t) = \sum_{j=1}^m p_j \ln \left( \frac{a_j^T x^{t+1}}{a_j^T x^t} \right)$$

$$= \sum_{j=1}^m p_j \ln \left( \sum_{i=1}^n \frac{a_{ji} x_i^t x^{t+1}}{a_j^T x^t x_i^t} \right)$$

$$\geq \sum_{j=1}^m p_j \sum_{i=1}^n \frac{a_{ji} x_i^t}{a_j^T x^t} \ln \left( \frac{x_i^{t+1}}{x_i^t} \right)$$

$$= \frac{n}{i=1} x_i^t \nabla_i f(x^t) \ln \left( \frac{x_i^{t+1}}{x_i^t} \right)$$

$$= D_h(x^{t+1}, x^t),$$

where in the inequality we use the concavity of $\ln(\cdot)$.

---

**References**

[1] I. Csiszar, “On the computation of rate-distortion functions (corresp.),” *IEEE Trans. Inf. Theory*, vol. 20, no. 1, pp. 122–124, 1974.

[2] Y. Vardi, L. A. Shepp, and L. Kaufman, “A statistical model for positron emission tomography,” *J. Amer. Stat. Assoc.*, vol. 80, no. 389, pp. 8–20, 1985.
[3] Y. Vardi and D. Lee, “From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems,” *J. R. Stat. Soc. Ser. B*, vol. 55, no. 3, pp. 569–598, 1993.

[4] T. Cover, “An algorithm for maximizing expected log investment return,” *IEEE Transactions on Information Theory*, vol. 30, no. 2, pp. 369–373, 1984.

[5] Y.-H. Li, “Online positron emission tomography by online portfolio selection,” in *Proc. ICASSP*, pp. 1110–1114, 2020.

[6] R. Zhao and R. M. Freund, “Analysis of the frank-wolfe method for logarithmically-homogeneous barriers, with an extension.” arXiv:2010.08999, 2020.

[7] G. Odor, Y.-H. Li, A. Yurtsever, Y.-P. Hsieh, Q. Tran-Dinh, M. E. Halabi, and V. Cevher, “Frank-wolfe works for non-lipschitz continuous gradient objectives: Scalable poisson phase retrieval,” in *Proc. ICASSP*, pp. 6230–6234, 2016.

[8] H. H. Bauschke, J. Bolte, and M. Teboulle, “A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications,” *Math. Oper. Res.*, vol. 42, no. 2, pp. 330–348, 2017.

[9] H. Lu, R. M. Freund, and Y. Nesterov, “Relatively smooth convex optimization by first-order methods, and applications,” *SIAM J. Optim.*, vol. 28, no. 1, pp. 333–354, 2018.

[10] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb, “Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications,” *SIAM J. Optim.*, vol. 28, no. 4, pp. 2783–2808, 2018.

[11] I. Csiszar, “Information geometry and alternating minimization procedures,” *Stat. Decis.*, vol. 1, pp. 205–237, 1984.

[12] A. N. Iusem, “A short convergence proof for the em algorithm with independent poisson variable,” *Braz. J. Prob. Stat.*, vol. 6, no. 1, pp. 57–67, 1992.

[13] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*. SIAM, 2001.