Heavy-ball Algorithms Always Escape Saddle Points

Tao Sun[1] , **Dongsheng Li**[1]* , **Zhe Quan**[2] , **Hao Jiang**[1] , **Shengguo Li**[1] and **Yong Dou**[1]

[1] Computer of College, National University of Defense Technology
[2] Hunan University

nudtsuntao@163.com, dsli@nudt.edu.cn, quanzhe@hnu.edu.cn,{haojiang, nudtlsg, yongdou}@nudt.edu.cn

Abstract

Nonconvex optimization algorithms with random initialization have attracted increasing attention recently. It has been showed that many first-order methods always avoid saddle points with random starting points. In this paper, we answer a question: can the nonconvex heavy-ball algorithms with random initialization avoid saddle points? The answer is yes! Direct using the existing proof technique for the heavy-ball algorithms is hard due to that each iteration of the heavy-ball algorithm consists of current and last points. It is impossible to formulate the algorithms as iteration like $x^{k+1} = g(x^k)$ under some mapping g. To this end, we design a new mapping on a new space. With some transfers, the heavy-ball algorithm can be interpreted as iterations after this mapping. Theoretically, we prove that heavy-ball gradient descent enjoys larger stepsize than the gradient descent to escape saddle points to escape the saddle point. And the heavy-ball proximal point algorithm is also considered; we also proved that the algorithm can always escape the saddle point.

1 Introduction

We consider the smooth nonconvex optimization problem

$$\text{minimize}_{x \in \mathbb{R}^N} f(x) \qquad (1)$$

where f is a differentiable closed function whose gradient is L-Lipschitz continuous (may be nonconvex). And more we assume f is $\mathcal{C}^2$ over $\mathbb{R}^N$, that is, f is twice-differentiable and $\nabla^2 f(x)$ is continuous. This paper is devoted to the study of two heavy-ball algorithms for minimizing f with random initializations. Specifically, we focus on whether these two algorithms escape the saddle points.

1.1 Heavy-ball Algorithms

We consider the Heavy-Ball Gradient Descent (HBGD) with random initializations x^0, x^1 presented as

$$x^{k+1} = x^k - \gamma \nabla f(x^k) + \beta(x^k - x^{k-1}), k \geq 1, \qquad (2)$$

---
*Corresponding Author.

where γ is the stepsize and β is the inertial parameter. If $\beta = 0$, the algorithm then reduces to the basic gradient descent. The heavy-ball method is also named as momentum method and has been widely used in different cases to accelerate the gradient descent iteration. Theoretically, it has been proved that HBGD enjoys better convergence factor than both the gradient and Nesterov's accelerated gradient method with linear convergence rates under the condition that the objective function is twice continuously differentiable, strongly convex and has Lipschitz continuous gradient. With the convex and smooth assumption on the objective function, the ergodic $O(1/k)$ rate in terms of the objective value, i.e., $f\left(\frac{\sum_{i=1}^k x^i}{k}\right) - \min f = O\left(\frac{1}{k}\right)$, is proved by [Ghadimi *et al.*, 2015]. HBGD was proved to converge linearly in the strongly convex case by [Ghadimi *et al.*, 2015]. But the authors used a somehow restrictive assumption on the parameter β, which leads to a small range values of choose for β when the strongly convex constant is tiny. Due to the inertial term $\beta(x^k - x^{k-1})$, HBGD breaks Fejér monotonicity that gradient descent obeys. Consequently, it is difficult to prove its non-ergodic convergence rates. To this end, [Sun *et al.*, 2019] designed a novel Lyapunov analysis and derived better convergence results like non-ergodic sublinear convergence rate $\left(f(x^k) - \min f = O\left(\frac{1}{k}\right)\right)$ and larger stepsize for linear convergence. [Combettes and Glaudin, 2017] developed and studied heavy-ball methods for operator research. In the nonconvex case, if the objective function is nonconvex but Lipschitz differentiable, [Zavriev and Kostyuk, 1993] proved that the sequence generated by the heavy-ball gradient descent is convergent to a critical point, but without specifying related convergence rates. With semi-algebraic assumption on the objective function, [Ochs *et al.*, 2014] proved the sequence generated by HBGD is convergent to some critical point. Actually, [Ochs *et al.*, 2014] also extended HBGD to a nonsmooth composite optimization problem. [Xu and Yin, 2013] developed and analyzed the Heavy-ball algorithm for tensor minimization problems. [Loizou and Richtárik, 2017a; Loizou and Richtárik, 2017b] introduced the stochastic versions of heavy-ball algorithms and its relations with existing methods.

On the other hand, [Alvarez, 2000] studied a second-order ODE whose implicit discretization yields the Heavy-Ball Proximal Point Algorithm (HBPPA) mathematically de-

scribed as

$$x^{k+1} = \mathbf{Prox}_{\gamma f}(x^k + \beta(x^k - x^{k-1})), \qquad (3)$$

which admits the acceleration of the proximal point algorithm. In fact, the setting $\beta = 0$ will push the scheme back to be the basic proximal point algorithm. The convergence of HBPPA is studied by [Alvarez, 2000] with several assumptions on the parameters γ and β when f is convex. Noticing the fact that ∇f is actually the maximal monotone operator in the convex case, [Alvarez and Attouch, 2001] extended HBPPA to solve the inclusion problem. The inexact HBPPA was proposed and studied by [Moudafi and Elizabeth, 2003]. Existing works on HBPPA are all based on the convex assumption on the f. In this paper, we will present the convergence of nonconvex HBPPA including the gradient convergence under smooth assumption and sequence convergence under semi-algebraic assumption.

The performances of HBGD and HBPPA with convex settings exclude the interest of this paper. What we care is that can these two algorithms escape the saddle points when f is nonconvex? This paper answers this problem.

1.2 First-order Algorithms Escape Saddle Points

In the community of nonconvex continuous optimization, saddle points, which are usually unavoidable and outnumber the cardinality of local minima, have long been regarded as a major obstacle. In many applications like tensor decomposition [Ge *et al.*, 2015], matrix completion [Ge *et al.*, 2016] and dictionary learning [Sun *et al.*, 2017], all local optima are also globally optimal. In some networks, under several conditions, local optimas are close to the global optima [Chaudhari *et al.*, 2016]. Although these conditions are barely satisfied by many networks, the conclusion coincides with existing experimental observations. Thus if all local minimizers are relatively good, we just need to push the iterations to escape the saddle points. And how to escape the saddle points has been a hot topic recently for first-order optimization algorithms. [Pemantle, 1990] established the convergence of the Robbins-Monro stochastic approximation to local minimizers by introducing sufficiently large unbiased noise. For tensor problem, [Ge *et al.*, 2015] provided the quantitative rates on the convergence of noisy gradient descent. For general smooth nonconvex functions, the analysis of noisy gradient descent is established by [Jin *et al.*, 2017a]. In a latter paper, [Jin *et al.*, 2017b] considered the noisy accelerated gradient descent and established related convergence theory. Another technique to escape the saddle points is using random initializations. It has been shown that, even without any random perturbations, a good initialization can make gradient descent converge to the global minimum for various non-convex problems like matrix factorization [Keshavan *et al.*, 2010] , phase retrieval [Cai *et al.*, 2016], dictionary learning [Arora *et al.*, 2015]. [Lee *et al.*, 2016; Lee *et al.*, 2017] proved that for plenty of first-order nonconvex optimization algorithms, with random initializations, they can converge to local minima with probability 1. In the perspective of mathematics, the core techniques consist of two parts: 1). they reformulate the algorithms as mapping iterations; 2). the stable manifold theorem [Shub, 2013] is

employed. In this paper, we establish the theoretical guarantees for heavy-ball algorithms escaping the saddle points.

1.3 Difficulty in Analyzing Heavy-ball Iterations

The difficulty of study on the performance of heavy-ball with random initializations lies in how to reformulate the algorithm as iterations under a mapping. Taking the gradient descent for example, one can use $g := \mathbb{I} - \gamma\nabla f$ with γ being the stepsize. And then, the gradient descent can be represented as $x^{k+1} = g(x^k)$. Constructing the mapping for gradient descent is obvious and easy. But for heavy-ball ones, the construction tasks are quite different due to the inertial term $\beta(x^k - x^{k-1})$. Point x^{k+1} is generated by both the current variable x^k and the last one x^{k-1}. Thus, it is impossible to find a mapping g to rewrite the heavy-ball as $x^{k+1} = g(x^k)$.

1.4 Contribution

Due to that the analysis of nonconvex HBPPA is still blank, we establish related theory in this paper. And with the recently developed techniques in variational analysis, the sequence convergence of nonconvex HBPPA is also studied. But our main interests are still the analysis of HBGD and HBPPA with random initializations. We prove guarantees for both algorithms to escape saddle points under mild assumptions. Direct using the existing method is unapplicable for heavy-ball algorithms. Thus, we propose a novel construction technique. It can be proved that heavy-ball gradient descent enjoys larger stepsize than gradient descent to escape the saddle points with random starting points, which somehow indicates the advantages of heavy-ball methods.

2 Preliminaries

In this part, we introduce several basic definitions and tool lemmas. For a closed function J (not necessary to be convex), the proximal map is defined as

$$\mathbf{Prox}_J(\cdot) := \arg\min_x \left\{ J(x) + \frac{\|x - \cdot\|^2}{2} \right\}. \qquad (4)$$

Under the condition that J is differentiable, if $y = \mathbf{Prox}_J(z)$, the K.K.T. condition can yields that $\nabla J(y) + y - z = \mathbf{0}$. Given another point arbitrary $x \in \mathbb{R}^N$, it holds that

$$J(y) + \frac{\|y - z\|^2}{2} \le J(x) + \frac{\|x - z\|^2}{2}. \qquad (5)$$

For matrices A and B, $A \sim B$ denotes that there exist perturbation matrices P, Q such that $A = PBQ$. We use $\mathbb{I}_N$ and $\mathbf{0}_N$ to denote the $N \times N$ unit and zero matrix, respectively. $\det(A)$ is determinant of matrix A. It is easy to see if $A \sim B$, $\det(A) = 0 \Leftrightarrow \det(B) = 0$.

Definition 1 (Strict Saddle) *We present the following two definitions,*

1. *A point x^* is a critical point of f if $\nabla f(x^*) = 0$. x^* is isolated if there is a neighborhood U around x^*, and x^* is the only critical point in U.*

2. *A point x^* is a strict saddle point of f if x^* is a critical point and $\lambda_{\min}(\nabla^2 f(x^*)) < 0$. Let χ_f^* denote the set of strict saddle points of function f.*

Let g be a mapping; and its domain is defined as $\text{dom}(g) := \{x \mid g(x) \neq \emptyset\}$. And g^k denotes the k-fold composition of g. Our interest is the sequence generated by

$$x^k = g(x^{k-1}) = g^k(x_0). \qquad (6)$$

Definition 2 (Global Stable Set) *The global stable set of the strict saddles is the set of initial conditions where iteration of the mapping g converges to a strict saddle, i.e., $W_g := \{x_0 : \lim_k g^k(x_0) \in \chi^*\}$. For an iterative algorithm $\mathcal{M}$, $W_{\mathcal{M}} := \{x_0 : \lim_k x^k \in \chi^*\}$ and $\{x^1, \ldots\}$ is generated by $\mathcal{M}$ with starting point x^0.*

Definition 3 (Unstable fixed point) *The differential of the mapping g, denoted as $Dg(x)$, is a linear operator from $\mathcal{T}(x) \rightarrow \mathcal{T}(g(x))$, where $\mathcal{T}(x)$ is the tangent space of $\mathcal{X}$ at point x. Given a curve γ in $\mathcal{X}$ with $\gamma(0) = x$ and $\frac{d\gamma}{dt}(0) := v \in \mathcal{T}(x)$, the linear operator is defined as $Dg(x)v = \frac{d(g \circ \gamma)}{dt}(0) \in \mathcal{T}(g(x))$. Let*

$$\mathcal{A}_g^* = \{x : g(x) = x, \exists \lambda(Dg(x)) \text{ such that } |\lambda(Dg(x))| > 1\}$$

be the set of fixed points where the differential has at least a single real eigenvalue with magnitude greater than one. These are the unstable fixed points.

With previous definitions, the authors in [Lee *et al.*, 2017] presented the following result. The core technique is employing the stable manifold theorem given in [Shub, 2013].

Lemma 1 *Let g be a $\mathcal{C}^1$ mapping and $\det(Dg(x)) \neq \mathbf{0}$ for all $x \in \text{dom}(g)$. Assume the sequence $(x^k)_{k \geq 0}$ is generated by scheme (6). Then the set of initial points that converge to an unstable fixed point has measure zero, $\mu(\{x^0 : \lim x^k \in \mathcal{A}_g^*\}) = 0$. Further if $\mathcal{X}^* \subset \mathcal{A}_g^*$, $\mu(W_g) = 0$.*

3 Main Results

In this section, we present the theoretical guarantees to escape the saddle points for the heavy-ball algorithms presented before. We need a smooth assumption on the objective function.

Assumption 1 *The objective function f is twice-differentiable and $\|\nabla^2 f(x)\|_2 \leq L$, where $\|\cdot\|_2$ denotes the spectral norm of a matrix. Otherwise the objective function satisfies that $\lim_{x \in \mathbb{R}^N} f(x) > -\infty$.*

The smooth assumption is used to derive the differential of the designed mapping. And the lower bounded assumption is used to provide a lower bound for the descent variables.

Lemma 2 *Let $(x^k)_{k \geq 0}$ be generated by the HBGD (2) with $0 < \beta < 1$. If the stepsize is selected as*

$$0 < \gamma < \frac{2(1-\beta)}{L},$$

we have $\mu\{(x^0, x^1) : \lim_k x^k \in \chi_f^*\} = 0$ and $\mu\{W_{HBGD}\} = 0$.

It is easy to see that if $\beta < \frac{1}{2}$ the upper bound of the stepsize is therefore larger than $\frac{1}{L}$ for HBGD with random initializations. And the upper bound is closed to $\frac{2}{L}$ provided that β is sufficiently small. [Lee *et al.*, 2016] proved that the stepsize for the GD with random initializations is required to be smaller than $\frac{1}{L}$. Lemma 2 means a larger stepsize can be used under the heavy-ball scheme, which somehow demonstrates the advantage of heavy-ball method.

The difficulty of the analysis for HBGD has been presented in Sec. 1: it is impossible to find to a mapping such that $x^{k+1} = g(x^k)$ directly. Noticing the kth iteration involves with x^k and x^{k-1}, we turn to consider combining the these two points together, i.e., (x^k, x^{k-1}). And then, the task then turns to find a mapping such that $(x^{k+1}, x^k) = \hat{g}(x^k, x^{k-1})$. The mapping $\hat{g}$ is from $\mathbb{R}^{2N}$ to $\mathbb{R}^{2N}$. After building the relations between the unstable fixed point of $\hat{g}$ and the strict saddle points, the theorem then can be proved. Different from existing results in [Lemma 1,[Sun *et al.*, 2019]], $\beta \neq 0$ here. This is to make sure that the differential of the constructed mapping in the proofs is invertible. And the differential is non-symmetric.

Lemma 2 actually means $\mathbf{Prob}(x^0 \in W_{HBGD}) = 0$, that is, HBGD always escapes saddle points provided the stepsize is set properly. Lemma 2 just means that $(x^k)_{k \geq 0}$ will not fall into a saddle point. But whether the sequence converges or not is out of scope of this lemma. In the nonconvex optimization community, such a problem is answered by posing the semi-algebracity assumption on the objective function [Łojasiewicz, 1993; Kurdyka, 1998] or the isolated assumption on the critical point [Lange, 2013]. Thus, we can derive the following result.

Theorem 1 *Assume that conditions of Lemma 2 hold and f is coercive[1]. Suppose the starting points x^0 and x^1 are totally random and one of the following conditions is satisfied.*

1. *f is semi-algebraic.*

2. *All critical points of f are isolated points.*

Then, for HBGD, $\lim_k x^k$ exists and that limit x^* is a local minimizer of f with probability 1.

Now, we study the HBPPA. First, we provide the convergence in the nonconvex settings.

Lemma 3 *Let $(x^k)_{k \geq 0}$ be generated by scheme (3) with $0 < \beta < \frac{1}{2}$ and $\gamma > 0$. Then, we have $\lim_k \|\nabla f(x^k)\| = 0$. Further more, if the function f is semi-algebraic and coercive, $(x^k)_{k \geq 0}$ is convergent to some critical point x^*.*

To obtain the sequence convergence of HBPPA, besides the semi-algebraicity, the coercivity is also needed, which is used to guarantee the boundedness of the generated sequence. In Lemma 3, f is set to be differentiable. Actually, the smooth setting can be removed. Alternatively, we need to use the notion about subdifferentials of closed functions [Rockafellar and Wets, 2009] which are frequently used in the variational analysis. Specifically, in the smooth case, subdifferential is

---

[1]We say function coercive if $f(x) \rightarrow +\infty$ as $\|x\| \rightarrow +\infty$.

actually the gradient. With the same parameters, the sequence of HBPPA is still provably to converge to a critical point of f.

We are prepared to present the guarantee for HBPPA to escape the saddle point.

Lemma 4 *Let $(x^k)_{k \geq 0}$ be generated by scheme (3) with $0 < \beta < \frac{1}{2}$, if the stepsize is selected as*

$$0 < \gamma < \frac{1}{L},$$

we have $\mu\{(x^0, x^1) : \lim_k x^k \in \chi_f^*\} = 0$ and $\mu\{W_{HBPPA}\} = 0$.

For HBPPA, the stepsize requirement coincides with existing result [Lee *et al.*, 2017]. Main idea of proof of Lemma 4 is similar to Lemma 2, but more complicated in details.

Theorem 2 *Assume that conditions of Lemma 4 hold and f is coercive. Suppose the starting points x^0 and x^1 are totally random and one of the following conditions is satisfied.*

1. f is semi-algebraic.

2. All critical points of f are isolated points.

Then, for HBPPA, $\lim_k x^k$ exists and that limit x^* is a local minimizer of f with probability 1.

4 Proofs

This section collects the proofs for the this paper.

4.1 Proof of Lemma 2

To guarantee the convergence of HBGD, from [Lemma 1,[Sun *et al.*, 2019]], we need $0 \leq \beta < 1$ and $0 < \gamma < \frac{2(1-\beta)}{L}$. Let $y, z \in \mathbb{R}^N$ and $w := (y, z)$, denote a map as

$$F(w) = F(y, z) := \left(\begin{array}{c} y - \gamma \nabla f(y) + \beta(y - z) \\ y \end{array} \right).$$

It is easy to check that if $F(w) = \mathbf{0} \Longrightarrow y = z$. We further denote $w^k := (x^k, x^{k-1}) \in \mathbb{R}^{2N}$. It can be easily verified that the Heavy-ball methods can be reformulated as

$$w^{k+1} = F(w^k). \tag{7}$$

To use Lemma 1, we turn to the proofs of two facts:

1. $\det(DF(w)) \neq \mathbf{0}$ for all $w \in \mathbb{R}^{2N}$.

2. $\chi_f^* \times \chi_f^* \bigcap \{(y, z) \in \mathbb{R}^{2N} \mid y = z\} \subseteq \mathcal{A}_F^*$.

Proof of fact 1: Direct calculations give us

$$DF(w) = \left(\begin{array}{cc} (1 + \beta)\mathbb{I}_N - \gamma \nabla^2 f(y) & -\beta \mathbb{I}_N \\ \mathbb{I}_N & \mathbf{0}_N \end{array} \right).$$

We use the short-hand notation $A(y) := (1 + \beta)\mathbb{I}_N - \gamma \nabla^2 f(y)$. We can obtain the following relation

$$DF(w) = \left(\begin{array}{cc} A(y) & -\beta \mathbb{I}_N \\ \mathbb{I}_N & \mathbf{0}_N \end{array} \right) \sim \left(\begin{array}{cc} \mathbf{0}_N & -\beta \mathbb{I}_N \\ \mathbb{I}_N & \mathbf{0}_N \end{array} \right).$$

Therefore, we can derive $\det(DF(w)) \neq 0$ for any $w \in \mathbb{R}^{2N}$.

Proof of fact 2: For any x^* being a strict saddle, with the symmetry of $\nabla^2 f(x^*)$, then

$$\lambda_{\max}(A(x^*)) > 1 + \beta.$$

For $\lambda \in \mathbb{R}$ and $\lambda \neq 0$ [2], denoting $w^* = (x^*, x^*)$, we consider

$$\lambda \mathbb{I}_{2N} - DF(w^*) = \left(\begin{array}{cc} \lambda \mathbb{I}_N - A(x^*) & \beta \mathbb{I}_N \\ -\mathbb{I}_N & \lambda \mathbb{I}_N \end{array} \right).$$

After simplifications, we can get

$$\left(\begin{array}{cc} \lambda \mathbb{I}_N - A(x^*) & \beta \mathbb{I}_N \\ -\mathbb{I}_N & \lambda \mathbb{I}_N \end{array} \right)$$

$$\sim \left(\begin{array}{cc} \lambda \mathbb{I}_N + \frac{\beta}{\lambda} \mathbb{I}_N - A(x^*) & \\ & \lambda \mathbb{I}_N \end{array} \right).$$

That indicates $\det(\lambda \mathbb{I}_{2N} - DF(w^*)) = 0 \iff \det(\lambda \mathbb{I}_N + \frac{\beta}{\lambda} \mathbb{I}_N - A(x^*)) = 0$. Denote function $s(t) := t + \frac{\beta}{t}$, $t > 0$. All the eigenvalues of $DF(w^*)$ are the roots of the equation above. We exploit the two roots of $\lambda + \frac{\beta}{\lambda} = \lambda_{\max}(A(x^*))$. It is easy to check that $\lambda_{\max}(A(x^*))^2 - 4\beta > (1 - \beta)^2 \geq 0$, which means this equation has two real roots denoted as $0 < \underline{\lambda} < \overline{\lambda}$. Noting $\lambda_{\max}(A(x^*)) > 1 + \beta$, that is $s(\overline{\lambda}) > s(1)$. We can see that $s(t)$ is increasing on $[\sqrt{\beta}, +\infty)$. With the fact $0 \leq \beta < 1$, $s(t)$ is increasing on $[1, +\infty)$. Thus, there exists some real number $\lambda_0 > 1$ such that $\overline{\lambda} > \lambda_0$. Thus, we get $\lambda_{\max}(DF(w^*)) \geq \overline{\lambda} \geq \lambda_0 > 1$, i.e., $w^* \in \mathcal{A}_F^*$. That is also

$$\{(x^0, x^1) : \lim_k x^k \in \chi_f^*\} \subseteq \{w^0 : \lim_k w^k \in \mathcal{A}_F^*\}.$$

Using Lemma 1, $\mu\{w^0 : \lim_k w^k \in \mathcal{A}_F^*\} = 0$ and then

$$\mu\{(x^0, x^1) : \lim_k x^k \in \chi_f^*\} \leq \mu\{w^0 : \lim_k w^k \in \mathcal{A}_F^*\} = 0.$$

On the other hand, due to the nonnegativity of the measurement,

$$\mu\{(x^0, x^1) : \lim_k x^k \in \chi_f^*\} = 0.$$

4.2 Proof of Theorem 1

The first case has been proved in [Ochs *et al.*, 2014], in which $(x^k)_{k \geq 0}$ is convergent to some critical point x^*. For the second case, [Proposition 12.4.1, [Lange, 2013]] and equation (10) means that the stationary set of $(x^k)_{k \geq 0}$ is finite. Otherwise, the stationary set is connected which contradicts with the fact all critical points are isolated. Once using [Proposition 12.4.1, [Lange, 2013]], the stationary set has only one element x^*; and then, $(x^k)_{k \geq 0} \to x^*$.

Therefore, in both cases, it can be proved that $(x^k)_{k \geq 0} \to x^* \in \mathbb{R}^N$. With Lemma 2, x^* is a local minimizer with probability 1 if the starting points x^0, x^1 are random. The theorem is then proved.

4.3 Proof of Lemma 3

This subsection contains two parts. Part 1 is to prove the gradient convergence, while Part 2 contains the sequence convergence proof.

---

[2] We have proved that $DF(w)$ is nonsingular.

Part 1. In (5), by substituting $y \leftarrow x^{k+1}$, $z \leftarrow x^k + \beta(x^k - x^{k-1})$, $x \leftarrow x^k$, $J \leftarrow \gamma f$,

$$\gamma f(x^{k+1}) + \frac{\|x^k + \beta(x^k - x^{k-1}) - x^{k+1}\|^2}{2}$$

$$\leq \gamma f(x^k) + \frac{\|\beta(x^k - x^{k-1})\|^2}{2}.$$

After simplifications, we then get

$$f(x^{k+1}) + \frac{\|x^{k+1} - x^k\|^2}{2\gamma}$$

$$\leq f(x^k) + \frac{\beta}{\gamma}\langle x^{k+1} - x^k, x^k - x^{k-1}\rangle$$

$$\leq f(x^k) + \frac{\beta}{2\gamma}\|x^{k+1} - x^k\|^2 + \frac{\beta}{2\gamma}\|x^k - x^{k-1}\|^2. \quad (8)$$

Denote a novel function as

$$H(w) = H(x, y) := f(x) + \frac{\beta}{2\gamma}\|x - y\|^2$$

and sequence $w^k := (x^k, x^{k-1})$. Thus, inequality (8) offers the descent as

$$H(w^{k+1}) + \frac{1 - 2\beta}{2\gamma}\|x^{k+1} - x^k\|^2 \leq H(w^k), \ k \geq 1. \quad (9)$$

With the fact $\inf_{x \in \mathbb{R}^N} f(x) > -\infty$, $\inf_k H(w^k) > -\infty$. The setting $0 < \beta < \frac{1}{2}$ gives

$$\lim_k \|x^{k+1} - x^k\| = 0. \quad (10)$$

On the other hand, from the property of the proximal map,

$$\gamma \nabla f(x^k) + x^{k+1} - x^k - \beta(x^k - x^{k-1}) = \mathbf{0}, \quad (11)$$

which can derive

$$\lim_k \|\nabla f(x^k)\|$$

$$\leq \frac{\lim_k \|x^{k+1} - x^k\| + \beta \lim_k \|x^k - x^{k-1}\|}{\gamma} = 0 \quad (12)$$

Part 2. We prove that $H(x, y)$ is coercive provided f is coercive. If this claim fails to hold, there exists some $C > 0$ such that

$$f(x) \leq H(x, y) \leq C \ \text{as} \ \|(x, y)\| \to +\infty.$$

With the coercivity of f, there exists some $C_1 > 0$ such that $\|x\| \leq C_1$. On the other hand, $\|x - y\|$ is bounded due to the lower boundedness of f; thus, $\|y\|$ is bounded, which contradicts the fact $\|(x, y)\| \to +\infty$.

From (9), we see that $(w^k)_{k \geq 1}$ is bounded; and then, $(x^k)_{k \geq 0}$ is bounded. The descent property of $(H(w^k))_{k \geq 1}$ and the continuity of function H directly give that

$$\lim_k H(w^k) = H(w^*). \quad (13)$$

That means the sequence has a stationary point x^*; (12) means x^* admits the critical point of f. It is easy to see $w^* := (x^*, x^*)$ is the critical point of H. Without loss of generality, we assume $w^k \neq w^*$ as $k \geq 1$.

Noting polynomial functions are semi-algebraic and the fact that sum of semi-algebraic functions is still semi-algebraic, $H(w)$ is then semi-algebraic. Denote the set of all stationary points of $(w^k)_{k \geq 0}$ as Ω. From [Lemma 3.6, [Bolte *et al.*, 2014]], there exist $\eta, \varepsilon \in (0, +\infty]$ and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}^+$ such that

1. $\varphi(0) = 0$; φ is $\mathcal{C}^1$ on $(0, \eta)$; for all $s \in (0, \eta)$, $\varphi'(s) > 0$.
2. for $w \in \{w | H(w^*) < H(w) < H(w^*) + \eta\} \bigcap \{w | \text{dist}(w, \Omega) < \varepsilon\}$, it holds that

$$\varphi'(H(w) - H(w^*)) \cdot \|\nabla H(w)\| \geq 1. \quad (14)$$

For the η, ε given above, as k is large enough, w^k will fall into the set $\{w | H(w^*) < H(w) < H(w^*) + \eta\} \bigcap \{w | \text{dist}(w, \Omega) < \varepsilon\}$. The concavity of φ gives

$$\varphi(H(w^k) - H(w^*)) - \varphi(H(w^{k+1}) - H(w^*))$$

$$\geq \varphi'(H(w^k) - H(w^*)) \cdot (H(w^k) - H(w^{k+1}))$$

$$\overset{a)}{\geq} \nu \cdot \varphi'(H(w^k) - H(w^*)) \cdot \|x^{k+1} - x^k\|^2$$

$$\overset{b)}{\geq} \frac{\nu \|x^{k+1} - x^k\|^2}{\|\nabla H(w^k)\|}, \quad (15)$$

where $\nu = \frac{1 - 2\beta}{2\gamma}$, and $a)$ uses 9, and $b)$ comes from (14). The gradient of H satisfies

$$\|\nabla H(w^k)\| \leq \|\nabla f(x^k)\| + \frac{\beta}{\gamma}\|x^k - x^{k-1}\|$$

$$\overset{c)}{\leq} \frac{1}{\gamma}(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|), \quad (16)$$

where $c)$ depends on (11) and, we used the fact $0 < \beta < \frac{1}{2}$. Combining (15) and (16),

$$2\|x^{k+1} - x^k\|$$

$$\leq 2 \times \frac{1}{2}\sqrt{\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|}$$

$$\times 2\sqrt{\frac{1}{\nu\gamma}}\sqrt{\varphi(H(w^k) - H(w^*)) - \varphi(H(w^{k+1}) - H(w^*))}$$

$$\overset{d)}{\leq} \frac{4}{\nu\gamma}[\varphi(H(w^k) - H(w^*)) - \varphi(H(w^{k+1}) - H(w^*))]$$

$$+ \frac{\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|}{4}, \quad (17)$$

where $d)$ employs the Schwarz inequality $2(xy)^{\frac{1}{2}} \leq x + y$ with $x = 2\sqrt{\frac{\varsigma}{\nu}}\sqrt{\varphi(H(w^k) - H(w^*)) - \varphi(H(w^{k+1}) - H(w^*))}$, and $y = \frac{1}{2}\sqrt{\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|}$. Summing both sides of (17) from sufficiently large k to K, we have

$$\frac{3}{2}\sum_{i=k}^{K-1}\|x^{i+1} - x^i\| + \frac{7}{4}\|x^{K+1} - x^K\| \leq \frac{1}{4}\|x^k - x^{k-1}\|$$

$$+ \frac{4}{\nu\gamma}[\varphi(H(w^k) - H(w^*)) - \varphi(H(w^{K+1}) - H(w^*))]. \quad (18)$$

Letting $K \to +\infty$, from (10), (13), $\varphi(0) = 0$ and the continuity of φ, (18) then yields $\sum_{i=k}^{+\infty} \|x^{i+1} - x^i\| \leq \frac{1}{6}\|x^k - x^{k-1}\| + \frac{4}{\nu\gamma}\varphi(H(w^k) - H(w^*)) < +\infty$. We are then led to $\sum_{i=k}^{+\infty} \|x^{i+1} - x^i\| < +\infty \Rightarrow \sum_{i=0}^{+\infty} \|x^{i+1} - x^i\| < +\infty$. That means $(x^k)_{k\geq 0}$ is convergent to some point. With the fact, x^* is a stationary point, $(x^k)_{k\geq 0} \to x^*$.

4.4 Proof of Lemma 4

In this proof, we denote the mapping as

$$F(w) = F(y, z) := \begin{pmatrix} \mathbf{Prox}_{\gamma f}(y + \beta(y - z)) \\ y \end{pmatrix}.$$

To calculate the differential of F, we denote

$$g(y, z) := \mathbf{Prox}_{\gamma f}(y + \beta(y - z)).$$

It is easy to see that

$$DF(w) = \begin{pmatrix} \frac{\partial g(y,z)}{\partial y} & \frac{\partial g(y,z)}{\partial z} \\ \mathbb{I}_N & \mathbf{0}_N \end{pmatrix}. \tag{19}$$

The definition of the proximal map then gives

$$\mathbf{0} = \gamma\nabla f[g(y, z)] + [g(y, z) - (y + \beta(y - z))]. \tag{20}$$

By implicit differentiation on variable y, we can get

$$\mathbf{0} = \gamma\nabla^2 f[g(y, z)]\frac{\partial g(y, z)}{\partial y} + \frac{\partial g(y, z)}{\partial y} - (1 + \beta)\mathbb{I}_N.$$

Due to that $0 < \gamma < \frac{1}{L}$, $\gamma\nabla^2 f[g(y, z)] + \mathbb{I}_N$ is invertible. We use a shorthand notation

$$A(y, z) := (\gamma\nabla^2 f[g(y, z)] + \mathbb{I}_N)^{-1}.$$

For z, by the same procedure, then we can derive

$$\frac{\partial g(y, z)}{\partial y} = (1 + \beta)A(y, z), \ \frac{\partial g(y, z)}{\partial z} = -\beta A(y, z). \tag{21}$$

With (19) and (21), we are then led to

$$DF(w) = \begin{pmatrix} (1 + \beta)A(y, z) & -\beta A(y, z) \\ \mathbb{I}_N & \mathbf{0}_N \end{pmatrix}$$

$$\sim \begin{pmatrix} \mathbf{0}_N & -\beta A(y, z) \\ \mathbb{I}_N & \mathbf{0}_N \end{pmatrix}. \tag{22}$$

That means for any $w \in \mathbb{R}^{2N}$, $\det(DF(w)) \neq 0$. On the other hand, if x^* is a strict saddle, $g(x^*, x^*) = x^*$, $\lambda_{\max}(A(x^*, x^*)) > 1$. For $\lambda \in \mathbb{R}$, denoting $w^* = (x^*, x^*)$, we consider

$$\lambda\mathbb{I}_{2N} - DF(w^*)$$
$$= \begin{pmatrix} \lambda\mathbb{I}_N - (1 + \beta)A(x^*, x^*) & \beta A(x^*, x^*) \\ -\mathbb{I}_N & \lambda\mathbb{I}_N \end{pmatrix}.$$

After simplifications, we can derive

$$\lambda\mathbb{I}_{2N} - DF(w^*$$
$$\sim \begin{pmatrix} \lambda\mathbb{I}_N + [\frac{\beta}{\lambda} - (1 + \beta)]A(x^*, x^*) & \\ & \lambda\mathbb{I}_N \end{pmatrix}.$$

With direct computations, we turn to

$$\det(\lambda^2\mathbb{I}_N + (\beta - (1 + \beta)\lambda)A(x^*, x^*)) = 0.$$

We consider the following equation

$$H(\lambda) := \lambda^2 + (\beta - (1 + \beta)\lambda)\lambda_{\max}(A(x^*, x^*)) = 0.$$

Direct calculations give

$$(1 + \beta)^2\lambda_{\max}^2(A(x^*, x^*)) - 4\beta\lambda_{\max}(A(x^*, x^*))$$
$$= 4\beta\lambda_{\max}(A(x^*, x^*))(\lambda_{\max}(A(x^*, x^*)) - 1)$$
$$+ (1 - \beta)^2\lambda_{\max}^2(A(x^*, x^*)) \geq 0.$$

Thus, the equation enjoys two real roots denoted by $\underline{\lambda} < \overline{\lambda}$. It is easy to see that $\underline{\lambda} > 0$. Noticing that $H(1) < 0$ and $\lim_{\lambda\to+\infty} H(\lambda) = +\infty$, we have $\overline{\lambda} > 1$. That means $\lambda_{\max}(DF(w^*)) \geq \overline{\lambda} > 1$, thus, $w^* \in \mathcal{A}_F^*$. Consequently, the proof is proved by Lemma 1.

4.5 Proof of Theorem 2

This proof is similar to the one of Theorem 1 and will not be reproduced.

5 Conclusion

In this paper, we proved that HBGD and HBPPA always escape the saddle points with random initializations. This paper also established the convergence of nonconvex HBPPA. The core part in the proofs is bundling current and the last point as one point in an enlarged space. The heavy algorithms then can be represented as an iteration after a mapping. An interesting finding is that the HBGD can enjoy larger stepsize than the gradient descent to escape saddle points.

References

[Alvarez and Attouch, 2001] Felipe Alvarez and Hedy Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1-2):3–11, 2001.

[Alvarez, 2000] Felipe Alvarez. On the minimizing property of a second order dissipative system in hilbert spaces. *SIAM Journal on Control and Optimization*, 38(4):1102–1119, 2000.

[Arora et al., 2015] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. 2015.

[Bolte *et al.*, 2014] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[Cai *et al.*, 2016] T Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

[Chaudhari *et al.*, 2016] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

[Combettes and Glaudin, 2017] Patrick L. Combettes and Lilian E. Glaudin. Quasinonexpansive iterations on the affine hull of orbits: From mann's mean value algorithm to inertial methods. *Siam Journal on Optimization*, 27(4), 2017.

[Ge *et al.*, 2015] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[Ge *et al.*, 2016] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[Ghadimi *et al.*, 2015] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *Control Conference (ECC), 2015 European*, pages 310–315. IEEE, 2015.

[Jin *et al.*, 2017a] 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-Volume 70*, pages 1724–1732. JMLR. org, 2017.

[Jin *et al.*, 2017b] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017.

[Keshavan *et al.*, 2010] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[Kurdyka, 1998] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–784. Chartres: L'Institut, 1950-, 1998.

[Lange, 2013] Kenneth Lange. Elementary optimization. In *Optimization*, pages 1–21. Springer, 2013.

[Lee *et al.*, 2016] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

[Lee *et al.*, 2017] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *To be appeared in Mathematical Progamming*, 2017.

[Loizou and Richtárik, 2017a] Nicolas Loizou and Peter Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *arXiv preprint arXiv:1710.10737*, 2017.

[Loizou and Richtárik, 2017b] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.

[Łojasiewicz, 1993] Stanislas Łojasiewicz. Sur la géométrie semi-et sous-analytique. *Ann. Inst. Fourier*, 43(5):1575–1595, 1993.

[Moudafi and Elizabeth, 2003] Abdellatif Moudafi and E Elizabeth. Approximate inertial proximal methods using the enlargement of maximal monotone operators. *International Journal of Pure and Applied Mathematics*, 5(3):283–299, 2003.

[Ochs *et al.*, 2014] Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[Pemantle, 1990] Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, 18(2):698–712, 1990.

[Rockafellar and Wets, 2009] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[Shub, 2013] Michael Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

[Sun *et al.*, 2017] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[Sun *et al.*, 2019] Tao Sun, Penghang Yin, Dongsheng Li, Chun Huang, Lei Guan, and Hao Jiang. Non-ergodic convergence analysis of heavy-ball algorithms. *AAAI*, 2019.

[Xu and Yin, 2013] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[Zavriev and Kostyuk, 1993] SK Zavriev and FV Kostyuk. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 4(4):336–341, 1993.