FULL LENGTH PAPER

Series B

CrossMark

Behavior of accelerated gradient methods near critical points of nonconvex functions

Michael O'Neill[1] · Stephen J. Wright[1]

Abstract

We examine the behavior of accelerated gradient methods in smooth nonconvex unconstrained optimization, focusing in particular on their behavior near strict saddle points. Accelerated methods are iterative methods that typically step along a direction that is a linear combination of the previous step and the gradient of the function evaluated at a point at or near the current iterate. (The previous step encodes gradient information from earlier stages in the iterative process). We show by means of the stable manifold theorem that the heavy-ball method is unlikely to converge to strict saddle points, which are points at which the gradient of the objective is zero but the Hessian has at least one negative eigenvalue. We then examine the behavior of the heavy-ball method and other accelerated gradient methods in the vicinity of a strict saddle point of a nonconvex quadratic function, showing that both methods can diverge from this point more rapidly than the steepest-descent method.

Keywords Accelerated gradient methods · Nonconvex optimization

Mathematics Subject Classification 90C26 · 49M30

✉ Stephen J. Wright
swright@cs.wisc.edu

Michael O'Neill
moneill@cs.wisc.edu

[1] Computer Sciences Department, University of Wisconsin, Madison, WI 53706, USA

1 Introduction

We consider methods for the smooth unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function. We say that x^* is a *critical point* of (1) if $\nabla f(x^*) = 0$. Critical points that are not local minimizers are of little interest in the context of the optimization problem (1), so a desirable property of any algorithm for solving (1) is that it not be attracted to such a point. Specifically, we focus on functions with *strict saddle points*, that is, functions where the Hessian at each saddle point has at least one negative eigenvalue.

Our particular interest here is in methods that use *gradients* and *momentum* to construct steps. In many such methods, each step is a linear combination of two components: the gradient ∇f evaluated at a point at or near the latest iterate, and a momentum term, which is the step between the current iterate and the previous iterate. There are rich convergence theories for these methods in the case in which f is convex or strongly convex, along with extensive numerical experience in some important applications. However, although these methods are applied frequently to nonconvex functions, little is known from a mathematical viewpoint about their behavior in such settings. Early results showed that a certain modified accelerated gradient method achieves the same order of convergence on a nonconvex problem as gradient descent [7,10]—not a faster rate, as in the convex setting.

The heavy-ball method was studied in the nonconvex setting in [17]. From an argument based on a Lyapunov function, this work shows that heavy-ball converges to some set of stationary points when short step sizes are used. Their result also implies that with these shorter stepsizes, heavy-ball converges to these stationary points with a sublinear rate, just as gradient descent does in the nonconvex case. Another work studied the continuous time heavy-ball method [2]. For Morse functions (functions where all critical points have a non-singular Hessian matrix), this paper shows that the set of initial conditions from which heavy-ball converges to a local minimizer is an open dense subset of $\mathbb{R}^n \times \mathbb{R}^n$. We present a similar result for a larger class of functions, using techniques like those of [9], where the authors show that gradient descent, started from a random initial point, converges to a strict saddle point with probability zero. We show that the discrete heavy-ball method essentially shares this property. We also study whether momentum methods can "escape" strict saddle points more rapidly than gradient descent. Experience with nonconvex quadratics indicate that, when started close to the (measure-zero) set of points from which convergence to the saddle point occurs, momentum methods do indeed escape more quickly.

After submission of our paper, [8] described a method that combines accelerated gradient, perturbation at points with small gradients and explicit negative curvature detection to attain a method with worst-case complexity guarantees.

Notation   For compactness, we sometimes use the notation (y, z) to denote the vector $[y^T z^T]^T$, for $y \in \mathbb{R}^n$ and $z \in \mathbb{R}^n$.

2 Heavy-ball is unlikely to converge to strict saddle points

We show in this section that the heavy-ball method is not attracted to strict saddle points, unless initialized in a very particular way, that cannot occur if the starting point is chosen at random and the algorithm is modified slightly. Following [9], our proof is based on the stable manifold theorem.

We make the following assumption throughout this section.

Assumption 1   The function $f : \mathbb{R}^n \to \mathbb{R}$ is $r + 1$ times continuously differentiable, for some integer $r \geq 1$, and ∇f has Lipschitz constant L.

Under this assumption, the eigenvalues of the Hessian $\nabla^2 f(x^*)$ are bounded in magnitude by L.

The heavy-ball method is a prototypical momentum method (see [13]), which proceeds as follows from a starting point x^0:

$$x^{k+1} := x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}), \quad \text{with } x^{-1} = x^0. \tag{2}$$

Following [13], we can write (2) as follows:

$$\begin{bmatrix} x^{k+1} \\ x^k \end{bmatrix} = \begin{bmatrix} x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}) \\ x^k \end{bmatrix}. \tag{3}$$

Convergence for this method is known for the special case in which f is a strongly convex quadratic. Denote by m the positive lower bound on the eigenvalues of the Hessian of this quadratic, and recall that L is the upper bound. For the settings

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{m})^2}, \quad \beta = \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}} \tag{4}$$

a rigorous version of the eigenvalue-based argument in [13, Section 3.2] can be applied to show R-linear convergence with rate constant $\sqrt{\beta}$, which is approximately $1 - \sqrt{m/L}$ when the ratio L/m is large. This suggests a complexity of $O(\sqrt{L/m} \log \epsilon)$ iterations to reduce the error $\|x^k - x^*\|$ by a factor of ϵ (where x^* is the unique solution). Such rates are typical of accelerated gradient methods. They contrast with the $O((L/m) \log \epsilon)$ rates attained by the steepest-descent method on such functions.

We note that the eigenvalue-based argument that is "sketched" by [13] does not extend rigorously beyond strongly convex quadratic functions. A more sophisticated argument based on Lyapunov functions is needed, like the one presented for Nesterov's accelerated gradient method in [14, Chapter 4].

The key to our argument for non-convergence to strict saddle points lies in formulating the heavy-ball method as a mapping whose fixed points are stationary points of

f and to which we can apply the stable manifold theorem. Following (3), we define this mapping to be

$$G(z_1, z_2) = \begin{bmatrix} z_1 - \alpha \nabla f(z_1) + \beta(z_1 - z_2) \\ z_1 \end{bmatrix}, \quad (z_1, z_2) \in \mathbb{R}^n \times \mathbb{R}^n. \quad (5)$$

Note that

$$DG(z_1, z_2) = \begin{bmatrix} (1 + \beta)I - \alpha \nabla^2 f(z_1) & -\beta I \\ I & 0 \end{bmatrix}. \quad (6)$$

We have the following elementary result about the relationship of critical points for (1) to fixed points for the mapping G.

Lemma 1 *If x^* is a critical point of f, then $(z_1^*, z_2^*) = (x^*, x^*)$ is a fixed point for G. Conversely, if (z_1^*, z_2^*) is a fixed point for G, then $x^* = z_1^* = z_2^*$ is a critical point for f.*

Proof The first claim is obvious by substitution into (5). For the second claim, we have that if (z_1^*, z_2^*) is a fixed point for G, then

$$\begin{bmatrix} z_1^* \\ z_2^* \end{bmatrix} = \begin{bmatrix} z_1^* - \alpha \nabla f(z_1^*) + \beta(z_1^* - z_2^*) \\ z_1^* \end{bmatrix},$$

from which we have $z_2^* = z_1^*$ and $\nabla f(z_1^*) = 0$, giving the result. □

We now establish that G is a diffeomorphic mapping, a property needed for application of the stable manifold result.

Lemma 2 *Suppose that Assumption 1 holds. Then the mapping G defined in (5) is a C^r diffeomorphism.*

Proof We need to show that G is injective and surjective, and that G and its inverse are r times continuously differentiable.

To show injectivity of G, suppose that $G(x_1, x_2) = G(y_1, y_2)$. Then, we have

$$\begin{bmatrix} x_1 - \alpha \nabla f(x_1) + \beta(x_1 - x_2) \\ x_1 \end{bmatrix} = \begin{bmatrix} y_1 - \alpha \nabla f(y_1) + \beta(y_1 - y_2) \\ y_1 \end{bmatrix}. \quad (7)$$

Therefore, $x_1 = y_1$, and so

$$x_1 - y_1 + \beta(x_1 - y_1 + y_2 - x_2) = \alpha(\nabla f(x_1) - \nabla f(y_1)) \quad \Rightarrow \quad x_2 = y_2, \quad (8)$$

demonstrating injectivity. To show that G is surjective, we construct its inverse G^{-1} explicitly. Let (y_1, y_2) be such that

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = G(z_1, z_2) = \begin{bmatrix} z_1 - \alpha \nabla f(z_1) + \beta(z_1 - z_2) \\ z_1 \end{bmatrix}, \quad (9)$$

Then $z_1 = y_2$. From the first partition in (9), we obtain $z_2 = (z_1 - y_1 - \alpha \nabla f(z_1))/\beta + z_1$, which after substitution of $z_1 = y_2$ leads to

$$
\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = G^{-1}(y_1, y_2) = \begin{bmatrix} y_2 \\ \frac{1}{\beta}(y_2 - y_1 - \alpha \nabla f(y_2)) + y_2 \end{bmatrix}. \tag{10}
$$

Thus, G is a bijection. Both G and G^{-1} are continuously differentiable one time less than f, so by Assumption 1, G is a C^r-diffeomorphism. □

We are now ready to state the stable manifold theorem, which provides tools to let us characterize the set of escaping points.

Theorem 1 (Theorem III.7 of [15]) *Let 0 be a fixed point for the C^r local diffeomorphism $\phi : U \to E$ where U is a neighborhood of 0 in the Banach space E. Suppose that $E = E_{cs} \oplus E_u$, where E_{cs} is the invariant subspace corresponding to the eigenvalues of $D\phi(0)$ whose magnitude is less than or equal to 1, and E_u is the invariant subspace corresponding to eigenvalues of $D\phi(0)$ whose magnitude is greater than 1. Then there exists a C^r embedded disc W_{loc}^{cs} that is tangent to E_{cs} at 0 called the local stable center manifold. Additionally, there exists a neighborhood B of 0 such that $\phi(W_{loc}^{cs}) \cap B \subset W_{loc}^{cs}$, and that if z is a point such that $\phi^k(z) \in B$ for all $k \geq 0$, then $z \in W_{loc}^{cs}$.*

This is a similar statement of the stable manifold theorem to the one found in [9], except that since we have to deal with complex eigenvalues here, we emphasize that the decomposition is between the eigenvalues whose *magnitude* is less than or equal to 1, and greater than 1, respectively. It guarantees the existence of a stable center manifold of dimension equal to the number of eigenvalues of the Jacobian at the critical point that are less than or equal to 1.

We show now that the Jacobian $DG(x^*, x^*)$ has the properties required for application of this result, for values of α and β similar to the choices (4) (note that the conditions on α and β in this result hold when $\alpha \in (0, 4/L)$ and $\beta \in (-1 + \alpha L/2, 1)$, where L is the Lipschitz constant from Assumption 1). For purposes of this and future results in this section, we assume that at the point x^* we have $\nabla f(x^*) = 0$ and that the eigenvalue decomposition of $\nabla^2 f(x^*)$ can be written as

$$
\nabla^2 f(x^*) = V \Lambda V^T = \sum_{i=1}^{n} \lambda_i v_i (v_i)^T, \tag{11}
$$

where the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ have

$$
\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-p} \geq 0 > \lambda_{n-p+1} \geq \cdots \geq \lambda_n, \tag{12}
$$

for some p with $1 \leq p < n$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, and where v_i, $i = 1, 2, \ldots, n$ are the orthonormal set of eigenvectors that correspond to the eigenvalues in (12). The matrix $V = [v_1 \mid v_2 \mid \cdots \mid v_n]$ is orthogonal.

Theorem 2 *Suppose that Assumption 1 holds. Let x^* be a critical point for f at which $\nabla^2 f(x^*)$ has p negative eigenvalues, where $p \geq 1$. Consider the mapping G defined by (5) where*

$$0 < \alpha < \frac{4}{\lambda_1}, \quad \beta \in \left(\max \left(-1 + \frac{\alpha \lambda_1}{2}, 0 \right), 1 \right),$$

where λ_1 is the largest positive eigenvalue of $\nabla^2 f(x^*)$. Then there are matrices $\tilde{V}_s \in \mathbb{R}^{2n \times (2n-p)}$ and $\tilde{V}_u \in \mathbb{R}^{2n \times p}$ such that (a) the $2n \times 2n$ matrix $\tilde{V} = [\tilde{V}_s \mid \tilde{V}_u]$ is nonsingular; (b) the columns of $\tilde{V}_s$ span an invariant subspace of $DG(x^*, x^*)$ corresponding to eigenvalues of $DG(x^*, x^*)$ whose magnitude is less than or equal to 1; (c) the columns of $\tilde{V}_u$ span an invariant subspace of $DG(x^*, x^*)$ corresponding to eigenvalues of $DG(x^*, x^*)$ whose magnitude is greater than 1.

Proof Since

$$DG(x^*, x^*) = \begin{bmatrix} (1+\beta)I - \alpha \nabla^2 f(x^*) & -\beta I \\ I & 0 \end{bmatrix}, \tag{13}$$

we have from (11) that

$$\begin{bmatrix} V^T & 0 \\ 0 & V^T \end{bmatrix} DG(x^*, x^*) \begin{bmatrix} V & 0 \\ 0 & V \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \alpha \Lambda & -\beta I \\ I & 0 \end{bmatrix}.$$

By performing a symmetric permutation P on this matrix, interleaving rows/columns from the first block with rows/columns from the second block, we obtain a block diagonal matrix with 2×2 blocks of the following form on the diagonals, that is,

$$P^T \begin{bmatrix} V^T & 0 \\ 0 & V^T \end{bmatrix} DG(x^*, x^*) \begin{bmatrix} V & 0 \\ 0 & V \end{bmatrix} P = \begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \ddots & \\ & & & M_n \end{bmatrix}, \tag{14}$$

where

$$M_i := \begin{bmatrix} (1+\beta) - \alpha \lambda_i & -\beta \\ 1 & 0 \end{bmatrix}, \quad i = 1, 2, \ldots, n. \tag{15}$$

The eigenvalues of M_i are obtained from the following quadratic in μ:

$$t(\mu) := ((1+\beta) - \alpha \lambda_i - \mu)(-\mu) + \beta = 0, \tag{16}$$

that is,

$$t(\mu) = \mu^2 - (1 + \beta - \alpha \lambda_i)\mu + \beta = 0, \tag{17}$$

for which the roots are

$$\mu_i^{\text{hi,lo}} = \frac{1}{2}\left[(1 + \beta - \alpha\lambda_i) \pm \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}\right]. \tag{18}$$

We examine first the matrices M_i for which $\lambda_i < 0$. We have

$$(1 + \beta - \alpha\lambda_i)^2 - 4\beta = (1 - \beta)^2 - 2\alpha\lambda_i(1 + \beta) + \alpha^2|\lambda_i|^2 > 0,$$

so both roots in (18) are real. Since $t(\cdot)$ is convex quadratic, with $t(0) = \beta > 0$ and $t(1) = \alpha\lambda_i < 0$, one root is in $(0, 1)$ and the other is in $(1, \infty)$. We can thus write

$$M_i = S_i \Lambda_i S_i^{-1}, \quad \text{where} \tag{19a}$$

$$\Lambda_i = \begin{bmatrix} \mu_i^{\text{hi}} & 0 \\ 0 & \mu_i^{\text{lo}} \end{bmatrix}, \quad S_i = \begin{bmatrix} \mu_i^{\text{hi}} & 1 \\ 1 & \frac{1}{\mu_i^{\text{lo}}} \end{bmatrix}, \quad S_i^{-1} = \left(\frac{\mu_i^{\text{hi}}}{\mu_i^{\text{lo}}} - 1\right)^{-1} \begin{bmatrix} \frac{1}{\mu_i^{\text{lo}}} & -1 \\ -1 & \mu_i^{\text{hi}} \end{bmatrix}. \tag{19b}$$

where μ_i^{hi} is the eigenvalue of M_i in the range $(1, \infty)$ and μ_i^{lo} is the eigenvalue of M_i in the range $(0, 1)$ (This claim can be verified by direct calculation of the product (19a)).

Consider now the matrices M_i for which $\lambda_i = 0$. From (18), we have that the roots are 1 and β, which are distinct, since $\beta \in (0, 1)$. The eigenvalue decompositions of these matrices have the form

$$M_i = S_i \Lambda_i S_i^{-1}, \quad \text{where } \Lambda_i = \text{diag}(1, \beta), \tag{20}$$

and the S_i are 2×2 nonsingular matrices.

When $\lambda_i > 0$, we show that the eigenvalues of M_i both have magnitude less than 1, under the given conditions on α and β. Both roots in (18) are complex exactly when the term under the square root is negative, and in this case the magnitude of both roots is

$$\frac{1}{2}\sqrt{(1 + \beta - \alpha\lambda_i)^2 + \left(4\beta - (1 + \beta - \alpha\lambda_i)^2\right)} = \sqrt{\beta},$$

which is less than 1 by assumption. When both roots are real, we have $(1+\beta-\alpha\lambda_i)^2 - 4\beta \geq 0$, and we require the following to be true to ensure that both are less than 1 in absolute value:

$$-2 < (1 + \beta - \alpha\lambda_i) \pm \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta} < 2. \tag{21}$$

We deal with the right-hand inequality in (21) first. By rearranging, we show that this is implied by the following sequence of equivalent inequalities:

$$(1 + \beta - \alpha\lambda_i) + \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta} < 2$$

$$\Leftrightarrow \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta} < 1 - \beta + \alpha\lambda_i$$

$$\Leftrightarrow (1 + \beta - \alpha\lambda_i)^2 - 4\beta < (1 - \beta + \alpha\lambda_i)^2$$

$$\Leftrightarrow \beta^2 + 2\beta(1 - \alpha\lambda_i) + (1 - \alpha\lambda_i)^2 - 4\beta < \beta^2 - 2\beta(1 + \alpha\lambda_i) + (1 + \alpha\lambda_i)^2$$

$$\Leftrightarrow 2\beta - 2\beta\alpha\lambda_i - 4\beta + (1 - \alpha\lambda_i)^2 < -2\beta - 2\beta\alpha\lambda_i + (1 + \alpha\lambda_i)^2$$

$$\Leftrightarrow (1 - \alpha\lambda_i)^2 < (1 + \alpha\lambda_i)^2$$

$$\Leftrightarrow -2\alpha\lambda_i < 2\alpha\lambda_i,$$

where the last is clearly true, because of $\alpha > 0$ and $\lambda_i > 0$. Thus the right-hand inequality in (21) is satisfied.

For the left-hand inequality in (21), we have

$$-2 < (1 + \beta - \alpha\lambda_i) - \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}$$

$$\Leftrightarrow -3 - \beta + \alpha\lambda_i < -\sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}$$

$$\Leftrightarrow 3 + \beta - \alpha\lambda_i > \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}$$

$$\Leftrightarrow (3 + \beta - \alpha\lambda_i)^2 > (1 + \beta - \alpha\lambda_i)^2 - 4\beta$$

$$\Leftrightarrow \beta^2 + 2\beta(3 - \alpha\lambda_i) + (3 - \alpha\lambda_i)^2 > \beta^2 + 2\beta(1 - \alpha\lambda_i) + (1 - \alpha\lambda_i)^2 - 4\beta$$

$$\Leftrightarrow 6\beta - 2\beta\alpha\lambda_i + 9 - 6\alpha\lambda_i + \alpha^2(\lambda_i)^2 > -2\beta - 2\beta\alpha\lambda_i + 1 - 2\alpha\lambda_i + \alpha^2(\lambda_i)^2$$

$$\Leftrightarrow 8\beta + 8 - 4\alpha\lambda_i > 0$$

$$\Leftrightarrow \beta > -1 + \alpha\lambda_i/2,$$

and the last condition holds because of the assumption that $\beta > -1 + \alpha\lambda_1/2$. This completes our proof of the claim (21). Thus our assumptions on α and β suffice to ensure that both eigenvalues of M_i defined in (15) have magnitude less than 1 when $\lambda_i > 0$.

By defining

$$S := \begin{bmatrix} I & & & & & \\ & \ddots & & & & \\ & & I & & & \\ & & & S_{n-p+1} & & \\ & & & & \ddots & \\ & & & & & S_n \end{bmatrix},$$

where S_i, $i = n - p + 1, \ldots, n$ are the matrices defined in (19), we have from (14) that

$$S^{-1} P^T \begin{bmatrix} V^T & 0 \\ 0 & V^T \end{bmatrix} DG(x^*, x^*) \begin{bmatrix} V & 0 \\ 0 & V \end{bmatrix} PS$$

$$
= \begin{bmatrix} M_1 & & & & & \\ & \ddots & & & & \\ & & M_{n-p} & & & \\ & & & \Lambda_{n-p+1} & & \\ & & & & \ddots & \\ & & & & & \Lambda_n \end{bmatrix}. \tag{22}
$$

We now define another $2n$-dimensional permutation matrix $\tilde{P}$ that sorts the entries of the diagonal matrices Λ_i, $i = n - p + 1, \ldots, n$ into those whose magnitude is greater than one and those whose magnitude is less than or equal to one, to obtain

$$
\tilde{P}^T S^{-1} P^T \begin{bmatrix} V^T & 0 \\ 0 & V^T \end{bmatrix} DG(x^*, x^*) \begin{bmatrix} V & 0 \\ 0 & V \end{bmatrix} P S \tilde{P}
$$

$$
= \begin{bmatrix} M_1 & & & & \\ & \ddots & & & \\ & & M_{n-p} & & \\ & & & \tilde{\Lambda}^{\text{lo}} & \\ & & & & \tilde{\Lambda}^{\text{hi}} \end{bmatrix}, \tag{23}
$$

where

$$
\tilde{\Lambda}^{\text{lo}} = \operatorname{diag}(\mu^{\text{lo}}_{n-p+1}, \mu^{\text{lo}}_{n-p+2}, \ldots, \mu^{\text{lo}}_n), \quad \tilde{\Lambda}^{\text{hi}} = \operatorname{diag}(\mu^{\text{hi}}_{n-p+1}, \mu^{\text{hi}}_{n-p+2}, \ldots, \mu^{\text{hi}}_n).
$$

We now define

$$
\tilde{V} = \begin{bmatrix} V & 0 \\ 0 & V \end{bmatrix} P S \tilde{P},
$$

which is a nonsingular matrix, by nonsingularity of S and orthogonality of V, P, and $\tilde{P}$. As in the statement of the theorem, we define $\tilde{V}_s$ to be the first $2n - p$ columns of $\tilde{V}$ and $\tilde{V}_u$ to be the last p columns. These define invariant spaces. For the stable space, we have

$$
DG(x^*, x^*)\tilde{V}_s = \tilde{V}_s \tilde{\Lambda}_s, \quad \text{where} \quad \tilde{\Lambda}_s := \begin{bmatrix} M_1 & & & \\ & \ddots & & \\ & & M_{n-p} & \\ & & & \tilde{\Lambda}^{\text{lo}} \end{bmatrix},
$$

where all eigenvalues of $\tilde{\Lambda}_s$ have magnitude less than or equal to 1. For the unstable space, we have

$$
DG(x^*, x^*)\tilde{V}_u = \tilde{V}_u \tilde{\Lambda}^{\text{hi}},
$$

where $\tilde{\Lambda}^{\text{hi}}$ is a diagonal matrix with all diagonal elements greater than 1. $\qquad\square$

We find a basis for the eigenspace that corresponds to the eigenvalues of $DG(x^*, x^*)$ that are greater than 1 (that is, the column space of $\tilde{V}_u$) in the following result.

Corollary 1 *Suppose that the assumptions of Theorem 2 hold. Then the eigenvector of $DG(x^*, x^*)$ that corresponds to the unstable eigenvalue $\mu_i^{hi} > 1, i = n - p + 1, \ldots, n$ defined in (18) is*

$$\begin{bmatrix} v_i \\ (1/\mu_i^{hi})v_i \end{bmatrix}, \tag{24}$$

where v_i is an eigenvector of $\nabla^2 f(x^*)$ that corresponds to $\lambda_i < 0$. The set of such vectors forms an orthogonal basis for the subspace of $\mathbb{R}^{2n}$ corresponding to the eigenvalues of $DG(x^*, x^*)$ whose magnitude is greater than 1.

Proof We have from (13) that

$$DG(x^*, x^*) \begin{bmatrix} v_i \\ (1/\mu_i^{hi})v_i \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \alpha\nabla^2 f(x^*) & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} v_i \\ (1/\mu_i^{hi})v_i \end{bmatrix}$$
$$= \begin{bmatrix} ((1 + \beta - \alpha\lambda_i) - \beta/\mu_i^{hi})v_i \\ v_i \end{bmatrix},$$

so the result holds provided that

$$(1 + \beta - \alpha\lambda_i) - \beta/\mu_i^{hi} = \mu_i^{hi}.$$

But this is true because of (17), so (24) is an eigenvector of $DG(x^*, x^*)$ corresponding to the eigenvalue μ_i^{hi}. Since the vectors $\{v_i \mid i = n - p + 1, \ldots, n\}$ form an orthogonal set, so do the vectors (24) for $i = n - p + 1, \ldots, n$, completing the proof. $\square$

Our next result is similar to [9, Theorem 4.1]. It is for a modified version of the heavy-ball method in which the initial value for x^{-1} is perturbed from its usual choice of x^0.

Theorem 3 *Suppose that the assumptions of Theorem 2 hold. Suppose that the heavy-ball method is applied from an initial point of $(x^0, x^{-1}) = (x^0, x^0 + \epsilon y)$, where x^0 and y are random vectors with i.i.d. elements, and $\epsilon > 0$ is small. We then have*

$$\Pr\left(\lim_k x^k = x^*\right) = 0,$$

where the probability is taken over the starting vectors x^0 and y.

Proof Our proof tracks that of [9, Theorem 4.1]. As there, we define the stable set for x^* to be

$$W^s(x^*) := \left\{ (x^0, x^{-1}) : \lim_{k \to \infty} G^k(x^0, x^{-1}) = (x^*, x^*) \right\}. \tag{25}$$

For the neighborhood B of $(x^*, x^*) \in \mathbb{R}^{2n}$ promised by Theorem 1, we have for all $z \in W^s(x^*)$ that there is some $l \geq 0$ such that $G^t(z) \in B$ for all $t \geq l$, and therefore by Theorem 1 we must have $G^l(z) \in W_{loc}^{cs} \cap B$. Thus $W^s(x^*)$ is the set of points z such that $G^l(z) \in W_{loc}^{cs}$ for some finite l. From Theorem 1, W_{loc}^{cs} is tangent to the subspace E_{cs} at (x^*, x^*), and the dimension of E_{cs} is $2n - p$, by Theorem 2 (since E_{cs} is the space spanned by the columns of $\tilde{V}_s$). This subspace has measure zero in $\mathbb{R}^{2n}$, since $p \geq 1$. Since diffeomorphisms map sets of measure zero to sets of measure zero, and countable unions of measure zero sets have measure zero, we conclude that $W^s(x^*)$ has measure zero. Thus the initialization strategy we have outlined produces a starting vector in $W^s(x^*)$ with probability zero. □

Theorem 3 does not guarantee that once the iterates leave the neighborhood of x^*, they never return. It does not exclude the possibility that the sequence $\{(x^{k+1}, x^k)\}$ returns infinitely often to a neighborhood of (x^*, x^*).

We note that the tweak of taking x^{-1} slightly different from x^0 does not affect practical performance of the heavy-ball method, and has in fact been proposed before [17]. It also does not disturb the theory that exists for this method, which for the case of quadratic f discussed in [13] rests on an argument based on the eigendecomposition of the (linear) operator DG, which is not affected by the modified starting point. We note too that the accelerated gradient methods to be considered in the next section can also allow $x^{-1} \neq x^0$ without significantly affecting the convergence theory. A Lyapunov-function-based convergence analysis of this method (see, for example [14, Chapter 4], based on arguments in [16]) requires only trivial modification to accommodate $x^{-1} \neq x^0$.

For the variant of heavy-ball method in which $x^0 = x^{-1}$, we could consider a random choice of x^0 and ask whether there is zero probability of (x^0, x^0) belonging to the measure-zero set $W^s(x^*)$ defined by (25). The problem is of course that (x^0, x^0) lies in the n-dimensional subspace $Y^n := \{(z_1, z_1) \mid z_1 \in \mathbb{R}^n\}$, and we would need to establish that the intersection $W^s(x^*) \cap Y^n$ has measure zero in Y^n. In other words, we need that the set $\{z_1 \mid (z_1, z_1) \in W^s(x^*)\}$ has measure zero in $\mathbb{R}^n$. We have a partial result in this regard, pertaining to the set W_{loc}^{cs}, which is the local counterpart of $W^s(x^*)$. This result also makes use of the subspace E_{cs}, defined as in Theorem 1, which is the invariant subspace corresponding to eigenvalues of $DG(x^*, x^*)$ whose magnitudes are less than or equal to one.

Theorem 4 *Suppose that the assumptions of Theorem 2 hold. Then any vector of the form (w, w) where $w \in \mathbb{R}^n$ lies in the stable subspace E_{cs} only if $w \in \text{span}\{v_1, v_2, \ldots, v_{n-p}\}$ that is, the span of eigenvectors of $\nabla^2 f(x^*)$ that correspond to nonnegative eigenvalues of this matrix.*

Proof We write $w = \sum_{i=1}^n \tau_i v_i$ for some coefficients τ_i, $i = 1, 2, \ldots, n$, and show that $\tau_i = 0$ for $i = n - p + 1, \ldots, n$.

We first show that

$$DG(x^*, x^*)^k \begin{bmatrix} w \\ w \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \sigma_{k,i} v_i \\ \sum_{i=1}^n \eta_{k,i} v_i \end{bmatrix}, \tag{26}$$

where $\sigma_{0,i} = \tau_i$ and $\eta_{0,i} = \tau_i$, $i = 1, 2, \ldots, n$. To derive recurrences for $\sigma_{k,i}$ and $\eta_{k,i}$, we consider the multiplication by $DG(x^*, x^*)$ that takes us from stages k to $k + 1$. We have

$$\begin{bmatrix} \sum_{i=1}^{n} \sigma_{k+1,i} v_i \\ \sum_{i=1}^{n} \eta_{k+1,i} v_i \end{bmatrix} = \begin{bmatrix} (1 + \beta)I - \alpha \nabla^2 f(x^*) & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} \sigma_{k,i} v_i \\ \sum_{i=1}^{n} \eta_{k,i} v_i \end{bmatrix}$$
$$= \begin{bmatrix} \sum_{i=1}^{n} (1 - \alpha \lambda_i) \sigma_{k,i} v_i + \beta \sum_{i=1}^{n} (\sigma_{k,i} - \eta_{k,i}) v_i \\ \sum_{i=1}^{n} \sigma_{k,i} v_i \end{bmatrix}.$$

By matching terms, we have

$$\begin{bmatrix} \sigma_{k+1,i} \\ \eta_{k+1,i} \end{bmatrix} = \begin{bmatrix} (1 + \beta - \alpha \lambda_i) & -\beta \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sigma_{k,i} \\ \eta_{k,i} \end{bmatrix} = M_i \begin{bmatrix} \sigma_{k,i} \\ \eta_{k,i} \end{bmatrix},$$

where M_i is defined in (15). Using the factorization (19), we have

$$\begin{bmatrix} \sigma_{k,i} \\ \eta_{k,i} \end{bmatrix} = M_i^k \begin{bmatrix} \sigma_{0,i} \\ \eta_{0,i} \end{bmatrix} = S_i \Lambda_i^k S_i^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tau_i,$$

By substitution from (19), we obtain

$$\begin{bmatrix} \sigma_{k,i} \\ \eta_{k,i} \end{bmatrix} = \begin{bmatrix} \mu_i^{\text{hi}} & 1 \\ 1 & \frac{1}{\mu_i^{\text{lo}}} \end{bmatrix} \begin{bmatrix} (\mu_i^{\text{hi}})^k & 0 \\ 0 & (\mu_i^{\text{lo}})^k \end{bmatrix} \begin{bmatrix} 1 - \mu_i^{\text{lo}} \\ \mu_i^{\text{lo}}(\mu_i^{\text{hi}} - 1) \end{bmatrix} \frac{\tau_i}{\mu_i^{\text{hi}} - \mu_i^{\text{lo}}}.$$

Because $0 < \mu_i^{\text{lo}} < 1 < \mu_i^{\text{hi}}$, it follows from this formula that

$$\tau_i \neq 0 \quad \Rightarrow \quad \frac{\sigma_{k,i}}{\tau_i} \to_k \infty, \quad \frac{\eta_{k,i}}{\tau_i} \to_k \infty,$$

so if w has any component in the span of v_i, $i = n - p + 1, \ldots, n$ (that is, if $\tau_i \neq 0$), repeated multiplications of $(-w, w)$ by $DG(x^*, x^*)$ will lead to divergence, so $(-w, w)$ cannot be in the subspace E_{cs}. $\square$

A consequence of this theorem is that for a random choice of x^0, there is probability zero that $(x^0 - x^*, x^0 - x^*) \in E_{cs}$, which is tangential to W_{loc}^{cs} at x^*. Thus for x^0 close to x^*, there is probability zero that (x^0, x^0) is in the measure-zero set W_{loc}^{cs}. Successive iterations of (2) are locally similar to repeated multiplications of $(x^0 - x^*, x^0 - x^*)$ by the matrix $DG(x^*, x^*)$, that is, for $(x^{k+1} - x^*, x^k - x^*)$ small, we have

$$\begin{bmatrix} x^{k+1} - x^* \\ x^k - x^* \end{bmatrix} \approx DG(x^*, x^*) \begin{bmatrix} x^k - x^* \\ x^{k-1} - x^* \end{bmatrix} \approx DG(x^*, x^*)^{k+1} \begin{bmatrix} x^0 - x^* \\ x^0 - x^* \end{bmatrix}.$$

Under the probability-one event that $x^0 - x^* \notin E_{cs}$, this suggests divergence of the iteration (2) away from (x^*, x^*).

On the other hand, we can show that if the sequence passes sufficiently close to a point (x^*, x^*) such that x^* satisfies second-order sufficient conditions to be a solution

of (1), it subsequently converges to (x^*, x^*). For this result we need the following variant of the stable manifold theorem.

Theorem 5 (Theorem III.7 of [15]) *Let 0 be a fixed point for the C^r local diffeomorphism $\phi : U \to E$ where U is a neighborhood of 0 in the Banach space E. Suppose that E_s is the invariant subspace corresponding to the eigenvalues of $D\phi(0)$ whose magnitude is strictly less than 1. Then there exists a C^r embedded disc W^s_{loc} that is tangent to E_s at 0, and a neighborhood B of 0 such that $W^s_{loc} \subset B$, and for all $z \in W^s_{loc}$, we have $\phi^k(z) \to 0$ at a linear rate.*

When x^* satisfies second-order conditions for (1), all eigenvalues of $\nabla^2 f(x^*)$ are strictly positive. It follows from the proof of Theorem 2 that under the assumptions of this theorem, all eigenvalues of $DG(x^*, x^*)$ have magnitude strictly less than 1. Thus, the invariant subspace E_s in Theorem 5 is the full space (in our case, $\mathbb{R}^{2n}$), so W^s_{loc} is a neighborhood of (x^*, x^*). It follows that there is some $\epsilon > 0$ such that if $\|(x^{K+1}, x^K) - (x^*, x^*)\| < \epsilon$ for some K, the sequence (x^{k+1}, x^k) for $k \geq K$ converges to (x^*, x^*) at a linear rate.

3 Speed of divergence on a Toy problem

In this section, we investigate the rate of divergence of an accelerated method on a simple nonconvex objective function, the quadratic with $n = 2$ defined by

$$f(x) = \frac{1}{2}(x_1^2 - \delta x_2^2), \quad \text{where } 0 < \delta \ll 1. \tag{27}$$

Obviously, this function is unbounded below with a saddle point at $(0, 0)^T$. Its gradient has Lipschitz constant $L = 1$. Despite being a trivial problem, it captures the behavior of gradient algorithms near strict saddle points for indefinite quadratics of arbitrary dimension, as is apparent from the analysis below.

We have described the heavy-ball method in (2). The steepest-descent method, by contrast, takes steps of the form

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \tag{28}$$

for some $\alpha_k > 0$. When $\nabla f(x)$ has Lipschitz constant L, the choice $\alpha_k \equiv 1/L$ leads to decrease in f at each iteration that is consistent with convergence of $\|\nabla f(x^k)\|$ to zero at a sublinear rate when f is bounded below [12] (The classical theory for gradient descent says little about the case in which f is unbounded below, as in this example).

The gradient descent and heavy-ball methods will converge to the saddle point 0 for (27) only from starting points of the form $x^0 = (x_1^0, 0)$ for any $x_1^0 \in \mathbb{R}$. (In the case of heavy-ball, this claim follows from Theorem 4, using the fact that $(1, 0)^T$ is the eigenvector of $\nabla^2 f$ that corresponds to the positive eigenvalue 1.) From any other starting point, both methods will diverge, with function values going to $-\infty$. When

the starting point x^0 is very close to (but not on) the x_1 axis, the typical behavior is that these algorithms pass close to 0 before diverging along the x_2 axis. We are interested in the question: *Does the heavy-ball method diverge away from* 0 *significantly faster than the steepest-descent method?* The answer is "yes," as we show in this section.

We consider a starting point that is just off the horizontal axis, that is,

$$x^0 = \begin{bmatrix} 1 \\ \epsilon \end{bmatrix}, \quad \text{for some small } \epsilon > 0. \tag{29}$$

For the steepest-descent method with constant steplength, we have

$$\begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix} = \begin{bmatrix} x_1^k \\ x_2^k \end{bmatrix} - \alpha \begin{bmatrix} x_1^k \\ -\delta x_2^k \end{bmatrix},$$

so that

$$\begin{bmatrix} x_1^k \\ x_2^k \end{bmatrix} = \begin{bmatrix} (1-\alpha)^k \\ (1+\delta\alpha)^k \epsilon \end{bmatrix}. \tag{30}$$

One measure of repulsion from the saddle point is the number of iterations required to obtain $|x_2^k| \geq 1$. Here it suffices for k to be large enough that $(1 + \delta\alpha)^k \epsilon \geq 1$, for which (using the usual bound $\log(1 + \gamma) \leq \gamma$) a sufficient condition is that

$$k \geq \frac{|\log \epsilon|}{\delta\alpha}.$$

Making the standard choice of steplength $\alpha = 1/L = 1$, we obtain

$$k \geq \frac{|\log \epsilon|}{\delta}. \tag{31}$$

Consider now the heavy-ball method. Following (2), the iteration has the form:

$$\begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix} = \begin{bmatrix} (1-\alpha)x_1^k \\ (1+\delta\alpha)x_2^k \end{bmatrix} + \beta \begin{bmatrix} x_1^k - x_1^{k-1} \\ x_2^k - x_2^{k-1} \end{bmatrix}. \tag{32}$$

(For this quadratic problem, the operator G defined by (5) is linear, so that DG is constant.) We can partition this recursion into x_1 and x_2 components, and write

$$\begin{bmatrix} x_1^{k+1} \\ x_1^k \end{bmatrix} = M_1 \begin{bmatrix} x_1^k \\ x_1^{k-1} \end{bmatrix}, \quad \begin{bmatrix} x_2^{k+1} \\ x_2^k \end{bmatrix} = M_2 \begin{bmatrix} x_2^k \\ x_2^{k-1} \end{bmatrix}, \tag{33}$$

where

$$M_1 = \begin{bmatrix} 1 - \alpha + \beta & -\beta \\ 1 & 0 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 + \delta\alpha + \beta & -\beta \\ 1 & 0 \end{bmatrix}. \tag{34}$$

The eigenvalues of these two matrices are given by (18), by setting $\lambda_1 = 1$ and $\lambda_2 = -\delta$, respectively. For α and β satisfying the conditions of Theorem 2, which translate here to

$$0 < \alpha < 4, \quad \beta \in (-1 + \alpha/2, 1), \tag{35}$$

both eigenvalues of M_1 are less than 1 in magnitude (as we show in the proof of Theorem 2), so the x_1 components converge to zero. Again referring to the proof of Theorem 2, the eigenvalues of M_2 are both real, with one of them greater than 1, suggesting divergence in the x_2 component.

To understand rigorously the behavior of the x_2 sequence, we make some specific choices of α and β. Consider

$$\alpha \in (0, 3], \quad \beta = 1 - \alpha\delta - \gamma, \tag{36}$$

for some parameter $\gamma \geq 0$. Note that for small δ and γ, these choices are consistent with (35). By substituting into (18), we see that the two eigenvalues of M_2 are

$$\mu_2^{\text{hi,lo}} = \frac{1}{2}\left[(2 - \gamma) \pm \sqrt{\gamma^2 + 4\alpha\delta}\right].$$

For reasonable choices of γ, we have that $\mu_2^{\text{hi}} = 1 + c\sqrt{\delta}$ for a modest positive value of c. For specificity (and simplicity) let us consider $\alpha = 3$ and $\gamma = 0$, for which we have

$$\mu_2^{\text{hi}} = 1 + \sqrt{3\delta}, \quad \mu_2^{\text{lo}} = 1 - \sqrt{3\delta}. \tag{37}$$

The formula (19) yields $M_2 = S_2 \Lambda_2 S_2^{-1}$, where $\Lambda_2 = \text{diag}(1 + \sqrt{3\delta}, 1 - \sqrt{3\delta})$ and

$$S_2 = \begin{bmatrix} 1 + \sqrt{3\delta} & 1 \\ 1 & \frac{1}{1-\sqrt{3\delta}} \end{bmatrix}, \quad S_2^{-1} = \frac{1 - \sqrt{3\delta}}{2\sqrt{3\delta}} \begin{bmatrix} \frac{1}{1-\sqrt{3\delta}} & -1 \\ -1 & 1 + \sqrt{3\delta} \end{bmatrix}.$$

From (33), and setting $x_2^0 = x_2^{-1} = \epsilon$, we have

$$\begin{bmatrix} x_2^k \\ x_2^{k-1} \end{bmatrix} = S_2 \Lambda_2^k S_2^{-1} \begin{bmatrix} \epsilon \\ \epsilon \end{bmatrix}.$$

By substituting for Λ_2 and S_2, we obtain

$$\begin{bmatrix} x_2^k \\ x_2^{k-1} \end{bmatrix} = \epsilon S_2 \Lambda_2^k S_2^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \epsilon \frac{1 - \sqrt{3\delta}}{2\sqrt{3\delta}} S_2 \Lambda_2^k \begin{bmatrix} \frac{\sqrt{3\delta}}{1-\sqrt{3\delta}} \\ \sqrt{3\delta} \end{bmatrix}$$

$$= \epsilon S_2 \Lambda_2^k \begin{bmatrix} 1/2 \\ (1 - \sqrt{3\delta})/2 \end{bmatrix}$$

$$= \epsilon S_2 \begin{bmatrix} (\mu_2^{\text{hi}})^k/2 \\ (\mu_2^{\text{lo}})^k(1 - \sqrt{3\delta})/2 \end{bmatrix}$$

$$\geq \frac{1}{2}\epsilon \begin{bmatrix} (1 + \sqrt{3\delta})(\mu_2^{\text{hi}})^k \\ (\mu_2^{\text{hi}})^k \end{bmatrix},$$

where we simply drop the term involving μ_2^{lo} in the final step and use $1 - \sqrt{3\delta} > 0$. It follows that

$$x_2^k \geq \frac{1}{2}\epsilon(1 + \sqrt{3\delta})(\mu_2^{\text{hi}})^k = \frac{1}{2}\epsilon(1 + \sqrt{3\delta})^{k+1}.$$

It follows from this bound, by a standard argument, that a sufficient condition for $x_2^k \geq 1$ is

$$k + 1 \geq \frac{\log(2/\epsilon)}{\sqrt{3\delta}}.$$

Thus we have confirmed that divergence from the saddle point occurs in $O(|\log \epsilon|/\sqrt{\delta})$ iterations for heavy-ball, versus $O(|\log \epsilon|/\delta)$ iterations for gradient descent.

For larger values of δ, the divergence of steepest-descent and heavy-ball methods are both rapid, For appropriate choices of α and β, the iterates generated by both algorithms leave the vicinity of the saddle point quickly.

Figure 1 illustrates the divergence behavior of steepest descent and heavy-ball on the function (27) with $\delta = .02$. We set $\alpha = .75$ for both steepest descent and heavy-ball. For heavy-ball, we chose $\beta = 1 - \alpha\delta = .985$. Both methods were started from $x^0 = (.25, .01)^T$. We see that the trajectory traced by steepest descent approaches the saddle point quite closely before diverging slowly along the x_2 axis. The heavy-ball method "overshoots" the x_2 axis (because of the momentum term) but quickly returns to diverging along the x_2 direction at a faster rate than for steepest descent.

4 General accelerated gradient methods applied to quadratic functions

Here we analyze the rate at which a general class of accelerated gradient methods escape the saddle point of an n-dimensional quadratic function:

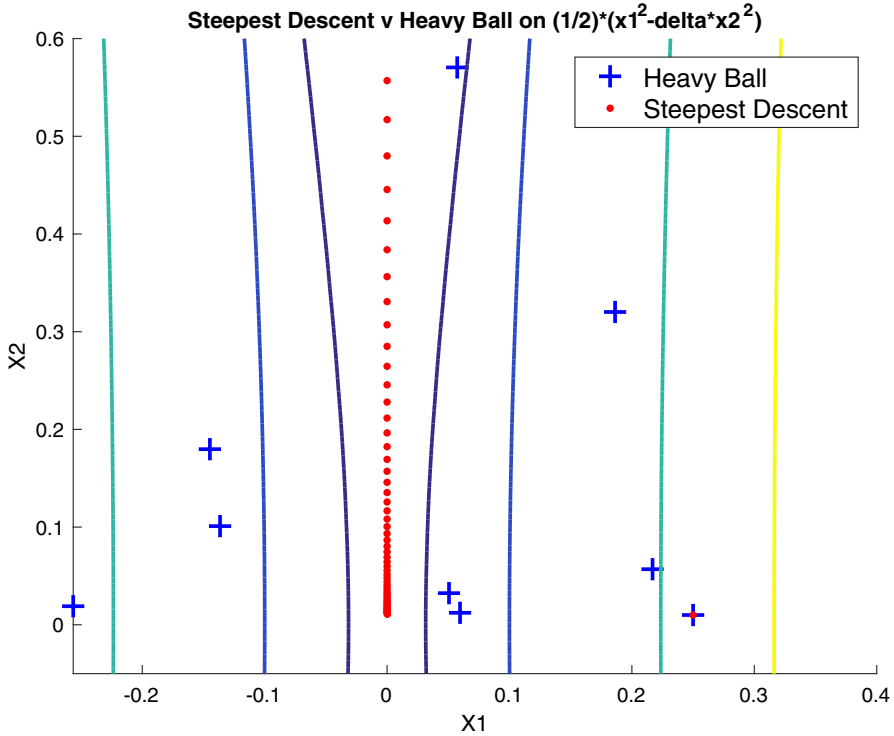$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2}x^T H x \tag{38}$$

Fig. 1 Steepest descent and heavy-ball on (27) with $\delta = 0.02$, from starting point $(0.25, 0.01)^T$, with $\alpha = 0.75$, $\beta = 1 - \alpha\delta = 0.985$. Every 5th iterate is plotted for each method

where H is a symmetric matrix with eigenvalues satisfying (12). We assume without loss of generality that H is in fact diagonal, that is,

$$H = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \tag{39}$$

The Lipschitz constant L for ∇f is $L = \max(\lambda_1, -\lambda_n)$.

Algorithm 1 General Accelerated Gradient Framework

Choose $x^1 \in \mathbb{R}^n$, $\alpha < \frac{1}{L}$;
Set $x^0 = x^1$;
for $k = 1, 2, \ldots$ **do**
  Choose $\gamma_k \in [0, 1]$ and $\beta_k \in [0, 1]$;
  $y^k = x^k + \gamma_k(x^k - x^{k-1})$;
  $x^{k+1} = x^k + \beta_k(x^k - x^{k-1}) - \alpha\nabla f(y^k)$;
end for

As in Sect. 3, gradient descent with $\alpha \in (0, 1/L)$ satisfies

$$x_i^{k+1} = (1 - \alpha\lambda_i)x_i^k = (1 - \alpha\lambda_i)^k x_i^1, \quad i = 1, 2, \ldots, n. \tag{40}$$

It follows that for all $i \geq n - p + 1$, for which $\lambda_i < 0$, gradient descent diverges in that component at a rate of $(1 - \alpha\lambda_i)$.

Algorithm 1 describes a general accelerated gradient framework, including gradient descent when $\gamma_k = \beta_k = 0$, heavy-ball when $\gamma_k = 0$ and $\beta_k > 0$, and accelerated gradient methods when $\gamma_k = \beta_k > 0$. With f defined by (38), the update formula can be written as

$$
\begin{aligned}
x^{k+1} &= x^k + \beta_k(x^k - x^{k-1}) - \alpha H(x^k + \gamma_k(x^k - x^{k-1})) \\
&= ((1 + \beta_k)I - \alpha(1 + \gamma_k)H)x^k - (\beta_k I - \alpha\gamma_k H)x^{k-1},
\end{aligned}
$$

which because of (39) is equivalent to

$$
x_i^{k+1} = ((1 + \beta_k) - (1 + \gamma_k)\alpha\lambda_i)x_i^k - (\beta_k - \gamma_k\alpha\lambda_i)x_i^{k-1}, \quad i = 1, 2, \ldots, n. \tag{41}
$$

The following theorem describes the dynamics of x_i^{k+1} in (41) when $\lambda_i < 0$.

Theorem 6 *For all i such that $\lambda_i < 0$, we have from (41) that*

$$
x_i^{k+1} = x_i^0 \prod_{m=0}^{k} (1 + b_{i,m}) \tag{42}
$$

where

$$
b_{i,k} = \begin{cases} 0, & \text{for } k = 0 \\ (\beta_k + \gamma_k\alpha|\lambda_i|)\left(1 - \frac{1}{1+b_{i,k-1}}\right) + \alpha|\lambda_i|, & \text{otherwise.} \end{cases} \tag{43}
$$

In addition if $\gamma_{k+1} \geq \gamma_k$ and $\beta_{k+1} \geq \beta_k$ for all k then,

$$
b_{i,k+1} \geq b_{i,k}, \quad k = 1, 2, \ldots. \tag{44}
$$

Proof We begin by showing that (43) holds for $k = 0$ and $k = 1$. The case for $k = 0$ is trivial as $x^1 = x^0$. In addition, for $k = 1$, the update formula (41) becomes

$$
x_i^2 = (1 - \alpha\lambda_i)x_i^0.
$$

Thus because $b_{i,0} = 0$, we can make this consistent with (42) by setting $b_{i,1} = \alpha|\lambda_i|$ which is exactly (43) for $k = 1$.

Now assume that (42) holds for all $k \leq K - 1$. From (41), using the inductive hypothesis for $K - 1$ and $K - 2$, we need to show

$$
\begin{aligned}
x_i^0 \prod_{m=0}^{K} (1 + b_{i,m}) &= x_i^0((1 + \beta_K) - (1 + \gamma_K)\alpha\lambda_i) \prod_{m=0}^{K-1} (1 + b_{i,m}) \\
&\quad - x_i^0(\beta_K - \gamma_K\alpha\lambda_i) \prod_{m=0}^{K-2} (1 + b_{i,m})
\end{aligned} \tag{45}
$$

by the given definition of $b_{i,K}$ in (43). Dividing both sides by $x_i^0 \prod_{m=0}^{K-1}(1 + b_{i,m})$, this is equivalent to

$$1 + b_{i,K} = 1 + \beta_K + (1 + \gamma_K)\alpha|\lambda_i| - \frac{\beta_K + \gamma_K\alpha|\lambda_i|}{1 + b_{i,K-1}},$$

which is true because

$$b_{i,K} = (\beta_K + \gamma_K\alpha|\lambda_i|)\left(1 - \frac{1}{1 + b_{i,K-1}}\right) + \alpha|\lambda_i|$$

is (43) with $k = K$, as required.

Now we assume that $\gamma_{K+1} \geq \gamma_K$ and $\beta_{K+1} \geq \beta_K$ holds for all $K \geq 1$ and show by induction that $b_{i,K+1} \geq b_{i,K}$ holds for all $K \geq 0$. This is clearly true for $K = 0$ since $\alpha|\lambda_i| > 0$. Assume now that $b_{i,k+1} \geq b_{i,k}$ holds for all $0 \leq k \leq K - 1$. We have

$$b_{i,K+1} = (\beta_{K+1} + \gamma_{K+1}\alpha|\lambda_i|)\left(1 - \frac{1}{1 + b_{i,K}}\right) + \alpha|\lambda_i|$$

$$\geq (\beta_{K+1} + \gamma_{K+1}\alpha|\lambda_i|)\left(1 - \frac{1}{1 + b_{i,K-1}}\right) + \alpha|\lambda_i|$$

$$\geq (\beta_K + \gamma_K\alpha|\lambda_i|)\left(1 - \frac{1}{1 + b_{i,K-1}}\right) + \alpha|\lambda_i| = b_{i,K}.$$

where the second inequality above follows from $\gamma_{K+1} \geq \gamma_K$, $\beta_{K+1} \geq \beta_K$ and $b_{i,K-1} \geq b_{i,0} = 0$. □

Since $b_{i,k} \geq \alpha|\lambda_i|$ for all $k \geq 1$, Theorem 6 shows that Algorithm 1 diverges at a faster rate than gradient descent when at least one of $\gamma_k > 0$ or $\beta_k > 0$ is true. Now we explore the rate of divergence by finding a limit for the sequence $\{b_{i,k}\}_{k=1,2,\ldots}$.

Theorem 7 *Let $\gamma_{k+1} \geq \gamma_k$ and $\beta_{k+1} \geq \beta_k$ hold for all k and denote $\bar{\gamma} = \lim_{k\to\infty}\gamma_k$ and $\bar{\beta} = \lim_{k\to\infty}\beta_k$. Then, for all i such that $\lambda_i < 0$, we have $\lim_{k\to\infty} b_{i,k} = \bar{b}_i$, where $\bar{b}_i$ is defined by by*

$$\bar{b}_i := \frac{1}{2}\left(\bar{\beta} - 1 + \alpha|\lambda_i|(1 + \bar{\gamma})\right) + \frac{1}{2}\sqrt{(\bar{\beta} - 1 + \alpha|\lambda_i|(1 + \bar{\gamma}))^2 + 4\alpha|\lambda_i|} \quad (46)$$

Proof We can write (41) as follows:

$$x_i^{k+1} = (1 + \alpha|\lambda_i|)x_i^k + (\beta_k + \gamma_k\alpha|\lambda_i|)(x_i^k - x_i^{k-1}).$$

Recall from Theorem 6 that $x_i^k = (1 + b_{i,k-1})x_i^{k-1}$. By substituting into the equation above, we have

$$x_i^{k+1} = \left[(1 + \alpha|\lambda_i|)(1 + b_{i,k-1}) + (\beta_k + \gamma_k\alpha|\lambda_i|)b_{i,k-1}\right]x_i^{k-1}$$

$$= \left[1 + \alpha|\lambda_i| + (1 + \alpha|\lambda_i| + \beta_k + \gamma_k\alpha|\lambda_i|)b_{i,k-1}\right]x_i^{k-1}. \quad (47)$$

Using Theorem 6 again, we have

$$x_i^{k+1} = \left[(1 + b_{i,k})(1 + b_{i,k-1})\right] x_i^{k-1} = \left[1 + b_{i,k} + b_{i,k-1} + b_{i,k-1}b_{i,k}\right] x_i^{k-1}.$$

By matching this expression with (47), we obtain

$$\alpha|\lambda_i| + (1 + \alpha|\lambda_i| + \beta_k + \gamma_k\alpha|\lambda_i|) b_{i,k-1} = b_{i,k} + b_{i,k-1} + b_{i,k-1}b_{i,k}, \quad (48)$$

which after division by $b_{i,k-1}$ yields

$$\frac{\alpha|\lambda_i|}{b_{i,k-1}} + (1 + \alpha|\lambda_i| + \beta_k + \gamma_k\alpha|\lambda_i|) = \frac{b_{i,k}}{b_{i,k-1}} + 1 + b_{i,k}. \quad (49)$$

Now assume for contradiction that the nondecreasing sequence $\{b_{i,k}\}_{k=1,2,\ldots}$ has no finite limit, that is, $b_{i,k} \to \infty$. Recalling that γ_k and β_k have a finite limit (as they are nondecreaseing sequences restricted to the interval $[0, 1]$), we have by taking the limit as $k \to \infty$ in (49) that the left-hand side approaches $\left(1 + \alpha|\lambda_i| + \bar{\beta} + \bar{\gamma}\alpha|\lambda_i|\right)$, while the right-hand side approaches ∞, a contradiction. Thus, the nondecreasing sequence $\{b_{i,k}\}_{k=1,2,\ldots}$ has a finite limit, which we denote by $\bar{b}_i$.

To find the value for $\bar{b}_i$, we take limits as $k \to \infty$ in (48) to obtain

$$\alpha|\lambda_i| + \left(1 + \alpha|\lambda_i| + \bar{\beta} + \bar{\gamma}\alpha|\lambda_i|\right) \bar{b}_i = 2\bar{b}_i + \bar{b}_i^2.$$

By solving this quadratic for $\bar{b}_i$, we obtain

$$\bar{b}_i = \frac{1}{2} \left(\bar{\beta} - 1 + \alpha|\lambda_i|(1 + \bar{\gamma})\right) \pm \frac{1}{2}\sqrt{(\bar{\beta} - 1 + \alpha|\lambda_i|(1 + \bar{\gamma}))^2 + 4\alpha|\lambda_i|}.$$

By Theorem 6, we know that $b_{i,k} \geq 0$ for all k, so that $\bar{b}_i \geq 0$. Therefore, $\bar{b}_i$ satisfies (46), as claimed. □

We apply Theorem 7 to parameter choices that typically appear in accelerated gradient methods.

Corollary 2 *Let the assumptions of Theorem 7 hold, let $\gamma_k = \beta_k$ hold for all k and let $\bar{\gamma} = \bar{\beta} = 1$. Then,*

$$\bar{b}_i = \alpha|\lambda_i| + \sqrt{\alpha|\lambda_i|}\sqrt{1 + \alpha|\lambda_i|}. \quad (50)$$

Proof By direct computation with $\bar{\beta} = \bar{\gamma} = 1$, we have

$$\bar{b}_i = \alpha|\lambda_i| + \frac{1}{2}\sqrt{4(\alpha|\lambda_i|)^2 + 4\alpha|\lambda_i|} = \alpha|\lambda_i| + \sqrt{\alpha|\lambda_i|}\sqrt{1 + \alpha|\lambda_i|}.$$

□

The above corollary gives a rate of divergence for many standard choices of the extrapolation parameters found in the accelerated gradient literature. In particular, it includes the sequence $\beta_k = \gamma_k = \frac{t_{k-1}-1}{t_k}$ where $t_0 = 1$ and

$$t_k = \frac{\sqrt{4t_{k-1}^2 + 1} + 1}{2} \tag{51}$$

which was used in a seminal work by Nesterov [11]. (For completeness, we provide a proof that $t_k \to \infty$, so that the assumptions of Corollary 2 hold for this sequence in the "Appendix"). Another setting used in recent works $\beta_k = \gamma_k = \frac{k-1}{k+\eta+1}$ [1,3,5]. For proper choices of $\eta > 0$, this scheme has a number of impressive properties such as fast convergence of iterates for accelerated proximal gradient as well as achieving a $o(\frac{1}{k^2})$ of convergence in the weakly convex case.

We can also use Theorem 7 to derive a bound for the heavy-ball method. If we target the n-th eigenvalue and set $\gamma_k = 0$ and $\beta_k = 1 - \alpha|\lambda_n|$ for all k, simple manipulation shows that $\bar{b}_n = \sqrt{\alpha|\lambda_n|}$, which gives us an equivalent rate to that derived in (37). Note that for $\bar{b}_n$ defined in (50) we also have $\bar{b}_n \geq \sqrt{\alpha|\lambda_n|}$.

The divergence rates for accelerated gradient and heavy-ball methods are significantly faster than the per-iteration rate of $(1 + \alpha|\lambda_n|)$ obtained for steepest descent.
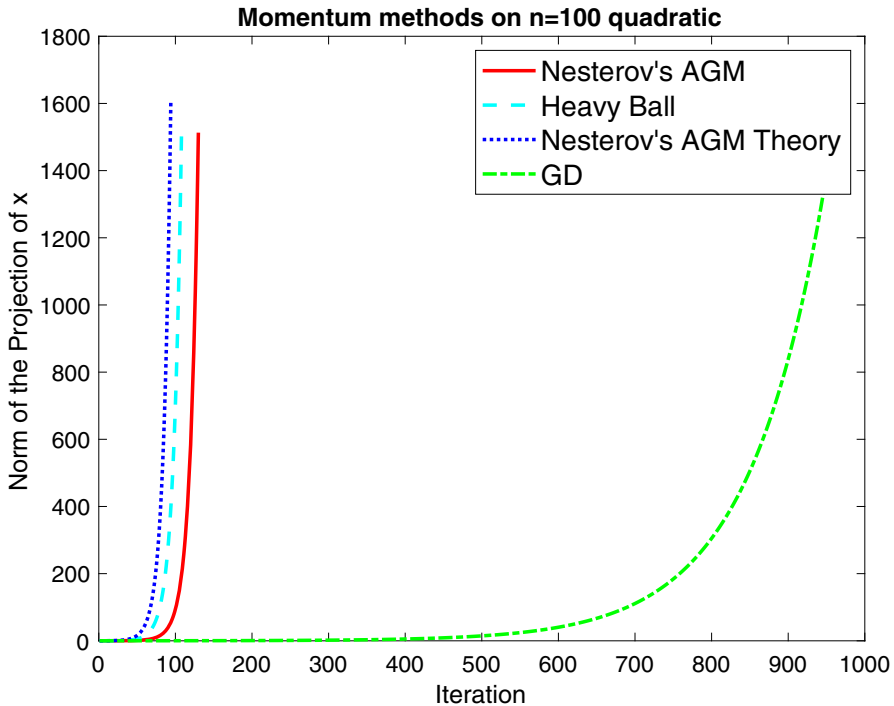
5 Experiments

Some computational experiments verify that accelerated gradient methods escape saddle points on nonconvex quadratics faster than steepest descent.

We apply these methods to a quadratic with diagonal Hessian, with $n = 100$ and a single negative eigenvalue, $\lambda_n = -\delta = -0.01$. The nonnegative eigenvalues are i.i.d. from the uniform distribution on $[0, 1]$, and starting vector x^0 is drawn from a uniform distribution on the unit ball. Figure 2 plots the norm of the component of x^k in the direction of the negative eigenvector $e_n = (0, 0, \ldots, 0, 1)^T$ at each iteration k, for accelerated gradient, heavy-ball, and steepest descent. It also shows the divergence that would be attained if the theoretical limit $\bar{b}_i$ from Theorem 7 applied at every iteration. Steepest descent and heavy-ball were run with $\alpha = 1/L$. Heavy-ball uses (36) to calculate β, yielding $\beta = 0.989$ in the case of $\delta = .01$. Accelerated gradient is run with $\alpha = 0.99/L$ and $\beta_k = \gamma_k = \frac{t_k-1}{t_{k+1}}$ where t_k is defined in (51).

It is clear from Fig. 2 that accelerated gradient and heavy-ball diverge at a significantly faster rate than steepest descent. In addition, there is only a small discrepancy between applying accelerated gradient and its limiting rate that is derived in Corollary 2, suggesting that $b_{i,k}$ approaches $\bar{b}_i$ rapidly as $k \to \infty$.

Next we investigate how these methods behave for various dimensions n and various distributions of the eigenvalues. For two values of n ($n = 100$ and $n = 1000$), we generate 100 random matrices with $n - 5$ eigenvalues uniformly distributed in the interval $[0, 1]$, with the 5 negative eigenvalues uniformly distributed in $[-2\delta, -\delta]$. The starting vector x^0 is uniformly distributed on the unit ball. Algorithmic constants were the same as those used to generate Fig. 2. Each trial was run until the norm of

Fig. 2 Momentum methods and theoretical divergence applied to a quadratic function with $n = 100$ and a single negative eigenvalue. The vertical axis displays the norm of the projection of x^k onto the negative eigenvector

the projection of the current iterate into the negative eigenspace of the Hessian was greater than the dimension n. The results of these trials are shown in Table 1.

As expected, accelerated gradient outperforms gradient descent in all respects. All convergence results are slightly faster for $n = 100$ than for $n = 1000$, because the random choice of x^0 will, in expectation, have a smaller component in the span of the negative eigenvectors in the latter case. The eigenvalue spectrum has a much stronger effect on the divergence rate. For steepest descent, an order of magnitude decrease in the absolute value of the negative eigenvalues corresponds to an order of magnitude increase in iterations, whereas Nesterov's accelerated gradient sees significantly less growth in the iteration count. While the accelerated gradient method diverges at a slightly slower rate than the theoretical limit, the relative difference between the two does not change much as the dimensions change. Thus, Theorem 7 provides a strong indication of the practical behavior of Nesterov's method on these problems.

6 Conclusion

We have derived several results about the behavior of accelerated gradient methods on nonconvex problems, in the vicinity of critical points at which at least one of the

Table 1 Divergence Behavior of Gradient Algorithms

n	δ	Method	Av. Iters	Max. Iters
100	10^{-2}	Steepest Descent	379	518
		Accelerated Gradient	71	87
		$\bar{b}$ Divergence Rate	46	59
100	10^{-3}	Steepest Descent	3855	5603
		Accelerated Gradient	242	299
		$\bar{b}$ Divergence Rate	155	194
1000	10^{-2}	Steepest Descent	582	773
		Accelerated Gradient	99	116
		$\bar{b}$ Divergence Rate	71	85
1000	10^{-3}	Steepest Descent	5775	8240
		Accelerated Gradient	332	399
		$\bar{b}$ Divergence Rate	235	282

eigenvalues of the Hessian $\nabla^2 f(x^*)$ is negative. Section 2 shows that the heavy-ball method does not converge to such a point when started randomly, while Sects. 3 and 4 show that when f is an indefinite quadratic, momentum methods diverge faster than the steepest-descent method.

It would be interesting to extend the results on speed of divergence to non-quadratic smooth functions f. It would also be interesting to know what can be proved about the complexity of convergence to a point satisfying second-order necessary conditions, for unadorned accelerated gradient methods. A recent work [6] shows that gradient descent can take exponential time to escape from a set of saddle points. We believe that a similar result holds for accelerated methods as well. The report [8], which appeared after this paper was submitted, describes an accelerated gradient method that add noise selectively to some iterates, and exploits negative curvature search directions when they are detected in the course of the algorithm. This approach is shown to have the $O(\epsilon^{-7/4})$ rate that characterizes the best known gradient-based algorithms for finding second-order necessary points of smooth nonconvex functions.

A properties of the sequence $\{t_k\}$ defined by (51)

In this "Appendix" we show that the following two properties hold for the sequence defined by (51):

$$\frac{t_{k-1} - 1}{t_k} \quad \text{is an increasing nonnegative sequence} \tag{52}$$

and

$$\lim_{k \to \infty} \frac{t_{k-1} - 1}{t_k} = 1. \tag{53}$$

We begin by noting two well known properties of the sequence t_k (see for example [4, Section 3.7.2]):

$$t_k^2 - t_k = t_{k-1}^2 \tag{54}$$

and

$$t_k \geq \frac{k+1}{2}. \tag{55}$$

To prove that $\frac{t_{k-1}-1}{t_k}$ is monotonically increasing, we need

$$\frac{t_{k-1} - 1}{t_k} = \frac{t_{k-1}}{t_k} - \frac{1}{t_k} \leq \frac{t_k}{t_{k+1}} - \frac{1}{t_{k+1}} = \frac{t_k - 1}{t_{k+1}}, \quad k = 1, 2, \ldots.$$

Since $t_{k+1} \geq t_k$ (which follows immediately from (51)), it is sufficient to prove that

$$\frac{t_{k-1}}{t_k} \leq \frac{t_k}{t_{k+1}}.$$

By manipulating this expression and using (54), we obtain the equivalent expression

$$t_{k-1} \leq \frac{t_k^2}{t_{k+1}} = \frac{t_{k+1}^2 - t_{k+1}}{t_{k+1}} = t_{k+1} - 1. \tag{56}$$

By definition of t_{k+1}, we have

$$t_{k+1} = \frac{\sqrt{4t_k^2 + 1} + 1}{2} \geq t_k + \frac{1}{2} = \frac{\sqrt{4t_{k-1}^2 + 1} + 1}{2} + \frac{1}{2} \geq t_{k-1} + 1.$$

Thus (56) holds, so the claim (52) is proved. The sequence $\{(t_{k-1} - 1)/t_k\}$ is nonnegative, since $(t_0 - 1)/t_1 = 0$.

Now we prove (53). We can lower-bound $(t_{k-1} - 1)/t_k$ as follows:

$$\begin{aligned}
\frac{t_{k-1} - 1}{t_k} &= \frac{2(t_{k-1} - 1)}{\sqrt{4t_{k-1}^2 + 1} + 1} \geq \frac{2(t_{k-1} - 1)}{\sqrt{4t_{k-1}^2 + 2}} \\
&= \frac{2(t_{k-1} - 1)}{2(t_{k-1} + 1)} = 1 - \frac{2}{t_{k-1} + 1}.
\end{aligned} \tag{57}$$

For an upper bound, we have from $t_k \geq t_{k-1}$ that

$$\frac{t_{k-1} - 1}{t_k} \leq \frac{t_{k-1}}{t_k} \leq 1. \tag{58}$$

Since $t_{k-1} \to \infty$ (because of (55)), it follows from (57) and (58) that (53) holds.

References

1. Attouch, H., Cabot, A.: Convergence rates of inertial forward–backward algorithms. SIAM J. Optim. **28**(1), 849–874 (2018)
2. Attouch, H., Goudou, X., Redont, P.: The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. Commun. Contemp. Math. **2**(01), 1–34 (2000)
3. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. SIAM J. Optim. **26**(3), 1824–1834 (2016)
4. Bubeck, S.: Convex optimization: algorithms and complexity. Found. Trends Mach. Learn. **8**(3–4), 231–357 (2015)
5. Chambolle, A., Dossal, Ch.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". J. Optim. Theory Appl. **166**(3), 968–982 (2015)
6. Du, S.S., Jin, C., Lee, J.D., Jordan, M.I., Singh, A., Poczos, B.: Gradient descent can take exponential time to escape saddle points. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 1067–1077. Curran Associates Inc, Red Hook (2017)
7. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. **156**(1–2), 59–99 (2016)
8. Jin, C., Netrapalli, P., Jordan, M.I.: Accelerated gradient descent escapes saddle points faster than gradient descent. arXiv preprint arXiv:1711.10456, (2017)
9. Lee, J.D., Simchowitz, M., Jordan, M.I., Recht, B.: Gradient descent only converges to minimizers. JMLR Workshop Conf. Proc. **49**(1), 1–12 (2016)
10. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28, pp. 379–387. Curran Associates Inc, Red Hook (2015)
11. Nesterov, Y.: A method for unconstrained convex problem with the rate of convergence $O(1/k^2)$. Dokl AN SSSR **269**, 543–547 (1983)
12. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Springer, New York (2004)
13. Polyak, B.T.: Introduction to Optimization. Optimization Software (1987)
14. Recht, B., Wright, S.J.: Nonlinear Optimization for Machine Learning (2017). (Manuscript in preparation)
15. Shub, M.: Global Stability of Dynamical Systems. Springer, Berlin (1987)
16. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. Technical report, Department of Mathematics, University of Washington, (2008)
17. Zavriev, S.K., Kostyuk, F.V.: Heavy-ball method in nonconvex optimization problems. Comput. Math. Model. **4**(4), 336–341 (1993)