

# BEHAVIOR OF ACCELERATED GRADIENT METHODS NEAR CRITICAL POINTS OF NONCONVEX FUNCTIONS\*

MICHAEL O'NEILL<sup>†</sup> AND STEPHEN J. WRIGHT<sup>‡</sup>

**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 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 Nesterov's accelerated gradient method 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.

**Key words.** Accelerated Gradient Methods, Nonconvex Optimization

**AMS subject classifications.** 90C26

**1. Introduction.** We consider methods for the smooth unconstrained optimization problem

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

where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a smooth function. We say that  $x^*$  is a *critical point* of (1) if  $\nabla f(x^*) = 0$ . Critical points that are not 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.

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  $\nabla f$  evaluated at a point 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. The best known results are that a certain modified accelerated gradient method achieves the same order of convergence on a nonconvex problem as gradient descent [3] [6] — not a faster rate, as in the convex setting.

The heavy-ball method was studied in the nonconvex setting in [13]. From an argument based on a Lyapunov function, this work shows that heavy ball converges to

---

\*Version of November 27, 2017.

**Funding:** This work was supported by NSF Awards IIS-1447449, 1628384, 1634597, and 1740707; AFOSR Award FA9550-13-1-0138; and Subcontract 3F-30222 from Argonne National Laboratory. Part of this work was done while the second author was visiting the Simons Institute for the Theory of Computing, and partially supported by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF Award CCF-1740425.

<sup>†</sup>Computer Sciences Department, University of Wisconsin, Madison, WI 53706 ([mon-eill@cs.wisc.edu](mailto:mon-eill@cs.wisc.edu)).

<sup>‡</sup>Computer Sciences Department, University of Wisconsin, Madison, WI 53706 ([swright@cs.wisc.edu](mailto:swright@cs.wisc.edu)).

some set of stationary points when short step sizes are used. Their result also implies that when using these shorter stepsizes, heavy-ball converges to these stationary points with a sublinear rate, just as gradient descent does in the nonconvex case.

A recent paper [5] shows that gradient descent, started from a random initial point, converges to a strict saddle point with probability zero. (Strict saddle point are critical points at which the Hessian has at least one negative eigenvalue.) We show that the 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.

*Notation.* For compactness, we sometimes use the notation  $(y, z)$  to denote the vector  $\begin{bmatrix} y \\ z \end{bmatrix}$ , 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 [5], our proof is based on the stable manifold theorem.

We make the following assumption throughout this section.

ASSUMPTION 1. *The function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $r + 1$  times continuously differentiable, for some integer  $r \geq 1$ , and  $\nabla f$  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 [9]), which proceeds as follows from a starting point  $x^0$ :

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

Following [9], we can write (2) as follows:

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

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

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

a rigorous version of the eigenvalue-based argument in [9, 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  $\epsilon$  (where  $x^*$  is the unique solution). Such rates are typical of accelerated gradient methods, including the Nesterov method to be discussed below. 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 [9] 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 method in [10, 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

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

Note that

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

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.  $\square$

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

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

Therefore,  $x_1 = y_1$ , and so

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

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

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

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

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

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.  $\square$

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

**THEOREM 3** (Theorem III.7 of [11]). *Let  $0$  be a fixed point for the  $C^r$  local diffeomorphism  $\phi : U \rightarrow E$  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 [5], 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  $\alpha$  and  $\beta$  similar to the choices (4). (Note that the conditions on  $\alpha$  and  $\beta$  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

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

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

$$(12) \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{n-p} \geq 0 > \lambda_{n-p+1} \geq \dots \geq \lambda_n,$$

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

**THEOREM 4.** *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  $\lambda_1$  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  $\tilde{V} = [\tilde{V}_s | \tilde{V}_u]$  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

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

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  $2 \times 2$  blocks of the following form on the diagonals, that is,

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

where

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

The eigenvalues of  $M_i$  are obtained from the following quadratic in  $\mu$ :

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

that is,

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

for which the roots are

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

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

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

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

where  $\mu_i^{hi}$  is the eigenvalue of  $M_i$  in the range  $(1, \infty)$  and  $\mu_i^{lo}$  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  $\beta$ , which are distinct, since  $\beta \in (0, 1)$ . The eigenvalue decompositions of these matrices have the form

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

and the  $S_i$  are  $2 \times 2$  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  $\alpha$  and  $\beta$ . 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

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

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:

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

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:

$$\begin{aligned} & (1 + \beta - \alpha\lambda_i) + \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta} < 2 \\ \Leftrightarrow & \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta} < 1 - \beta + \alpha\lambda_i \\ \Leftrightarrow & (1 + \beta - \alpha\lambda_i)^2 - 4\beta < (1 - \beta + \alpha\lambda_i)^2 \\ \Leftrightarrow & \beta^2 + 2\beta(1 - \alpha\lambda_i) + (1 - \alpha\lambda_i)^2 - 4\beta < \beta^2 - 2\beta(1 + \alpha\lambda_i) + (1 + \alpha\lambda_i)^2 \\ \Leftrightarrow & 2\beta - 2\beta\alpha\lambda_i - 4\beta + (1 - \alpha\lambda_i)^2 < -2\beta - 2\beta\alpha\lambda_i + (1 + \alpha\lambda_i)^2 \\ \Leftrightarrow & (1 - \alpha\lambda_i)^2 < (1 + \alpha\lambda_i)^2 \\ \Leftrightarrow & -2\alpha\lambda_i < 2\alpha\lambda_i, \end{aligned}$$

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

$$\begin{aligned} & -2 < (1 + \beta - \alpha\lambda_i) - \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta} \\ \Leftrightarrow & -3 - \beta + \alpha\lambda_i < -\sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta} \\ \Leftrightarrow & 3 + \beta - \alpha\lambda_i > \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta} \\ \Leftrightarrow & (3 + \beta - \alpha\lambda_i)^2 > (1 + \beta - \alpha\lambda_i)^2 - 4\beta \\ \Leftrightarrow & \beta^2 + 2\beta(3 - \alpha\lambda_i) + (3 - \alpha\lambda_i)^2 > \beta^2 + 2\beta(1 - \alpha\lambda_i) + (1 - \alpha\lambda_i)^2 - 4\beta \\ \Leftrightarrow & 6\beta - 2\beta\alpha\lambda_i + 9 - 6\alpha\lambda_i + \alpha^2(\lambda_i)^2 > -2\beta - 2\beta\alpha\lambda_i + 1 - 2\alpha\lambda_i + \alpha^2(\lambda_i)^2 \\ \Leftrightarrow & 8\beta + 8 - 4\alpha\lambda_i > 0 \\ \Leftrightarrow & \beta > -1 + \alpha\lambda_i/2, \end{aligned}$$

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  $\alpha$  and  $\beta$  suffice to ensure that both eigenvalues of  $M_i$  defined in (15) have magnitude less than 1 when  $\lambda_i > 0$ .



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

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

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

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 5.** *Suppose that the assumptions of Theorem 4 hold. Then the eigenvector of  $DG(x^*, x^*)$  that corresponds to the unstable eigenvalue  $\mu_i^{hi} > 1$ ,  $i = n - p + 1, \dots, n$  defined in (18) is*

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

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

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

so the result holds provided that

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

But this is true because of (18), so (24) is an eigenvector of  $DG(x^*, x^*)$  corresponding to the eigenvalue  $\mu_i^{hi}$ . Since the vectors  $\{v_i \mid i = n - p + 1, \dots, n\}$  form an orthogonal set, so do the vectors (24) for  $i = n - p + 1, \dots, n$ , completing the proof.  $\square$

Our next result is similar to [5, 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 6.** *Suppose that the assumptions of Theorem 4 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 [5, Theorem 4.1]. As there, we define the stable set for  $x^*$  to be

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

For the neighborhood  $B$  of  $(x^*, x^*) \in \mathbb{R}^{2n}$  promised by Theorem 3, 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 3 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 3,  $W_{loc}^{cs}$  is tangent to the subspace  $E_{cs}$  at  $(x^*, x^*)$ , and the dimension of  $E_{cs}$  is  $2n - p$ , by Theorem 4 (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.  $\square$

Theorem 6 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 [13]. It also does not disturb the theory that exists for this method, which for the case of quadratic  $f$  discussed in [9] 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 method to be considered in the next section — Nesterov's method — 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 [10, Chapter 4], based on arguments in [12]) 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 under 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 3, is the invariant subspace corresponding to eigenvalues of  $DG(x^*, x^*)$  whose magnitudes are less than or equal to one.

**THEOREM 7.** *Suppose that the assumptions of Theorem 4 hold. Then any vector of the form  $(w, w)$  where  $w \in \mathbb{R}^n$  lies in the stable subspace  $E_{cs}$  only if  $w \in \text{span}\{v_1, v_2, \dots, v_{n-p}\}$  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  $\tau_i$ ,  $i = 1, 2, \dots, n$ , and show that  $\tau_i = 0$  for  $i = n - p + 1, \dots, n$ .

We first show that

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

where  $\sigma_{0,i} = \tau_i$  and  $\eta_{0,i} = \tau_i$ ,  $i = 1, 2, \dots, n$ . 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

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

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

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

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

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

so if  $w$  has any component in the span of  $v_i$ ,  $i = n - p + 1, \dots, n$  (that is, if  $\tau_i \neq 0$ ), repeated multiplications of  $\begin{bmatrix} w \\ w \end{bmatrix}$  by  $DG(x^*, x^*)$  will lead to divergence, so  $\begin{bmatrix} w \\ w \end{bmatrix}$  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 8** (Theorem III.7 of [11]). *Let  $0$  be a fixed point for the  $C^r$  local diffeomorphism  $\phi : U \rightarrow E$  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  $W_{loc}^s$  that is tangent to  $E_s$  at  $0$ , and a neighborhood  $B$  of  $0$  such that  $W_{loc}^s \subset B$ , and for all  $z \in W_{loc}^s$ , we have  $\phi^k(z) \rightarrow 0$  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 4 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 8 is the full space (in our case,  $\mathbb{R}^{2n}$ ), so  $W_{loc}^s$  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

$$(27) \quad f(x) = \frac{1}{2}(x_1^2 - \delta x_2^2), \quad \text{for some small } \delta > 0.$$

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

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

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 [8]. (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 7, 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,

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

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

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

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

$$(31) \quad k \geq \frac{|\log \epsilon|}{\delta}.$$

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

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

(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

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

where

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

The eigenvalues of these two matrices are given by (18), by setting  $\lambda_1 = 1$  and  $\lambda_2 = -\delta$ , respectively. For  $\alpha$  and  $\beta$  satisfying the conditions of Theorem 4, which translate here to

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

both eigenvalues of  $M_1$  are less than 1 in magnitude (as we show in the proof of Theorem 4, so the  $x_1$  components converge to zero. Again referring to the proof of Theorem 4, 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  $\alpha$  and  $\beta$ . Consider

$$(36) \quad \alpha \in (0, 3), \quad \beta = 1 - \alpha\delta - \gamma,$$

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

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

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

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

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  $\Lambda_2$  and  $S_2$ , we obtain

$$\begin{aligned}
\begin{bmatrix} x_2^k \\ x_2^{k-1} \end{bmatrix} &= \epsilon S_2 \Lambda_2^k S_2^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= \epsilon \frac{1 - \sqrt{3\delta}}{2\sqrt{3\delta}} S_2 \Lambda_2 \begin{bmatrix} \frac{\sqrt{3\delta}}{1 - \sqrt{3\delta}} \\ \sqrt{3\delta} \end{bmatrix} \\
&= \epsilon S_2 \Lambda_2^k \begin{bmatrix} 1/2 \\ (1 - \sqrt{3\delta})/2 \end{bmatrix} \\
&= \epsilon S_2 \begin{bmatrix} (\mu_2^{hi})^k / 2 \\ (\mu_2^{lo})^k (1 - \sqrt{3\delta})/2 \end{bmatrix} \\
&\geq \frac{1}{2} \epsilon \begin{bmatrix} (1 + \sqrt{3\delta})(\mu_2^{hi})^k \\ (\mu_2^{hi})^k \end{bmatrix},
\end{aligned}$$

where we simply drop the term involving  $\mu_2^{lo}$  in the final step and use  $1 - \sqrt{3\delta} > 0$ . It follows that

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

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.

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. Nesterov’s Accelerated Gradient Method on General Quadratic Functions.** Here we analyze the rate at which the accelerated gradient method due to Nesterov [7] escapes the saddle point of an  $n$ -dimensional quadratic function:

$$(37) \quad \min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^T H x$$

where  $H$  is a symmetric matrix with  $p \geq 1$  negative eigenvalues  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , satisfying (12). We assume without loss of generality that  $H$  is in fact diagonal, that is,

$$(38) \quad H = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

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

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

$$(39) \quad x_i^{k+1} = (1 - \alpha\lambda_i) x_i^k = (1 - \alpha\lambda_i)^k x_i^1, \quad i = 1, 2, \dots, n.$$

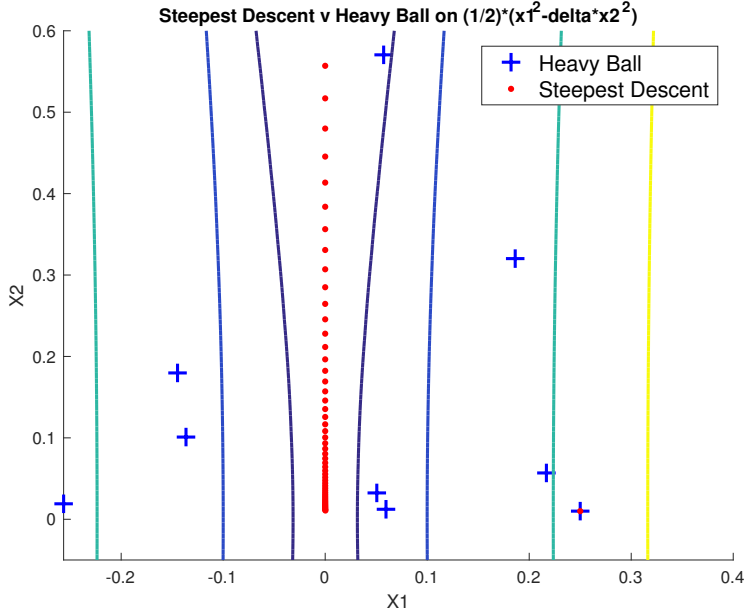


Fig. 1: Steepest descent and heavy-ball on (27) with  $\delta = .02$ , from starting point  $(.25, .01)^T$ , with  $\alpha = .75$ ,  $\beta = 1 - \alpha\delta = .985$ . Every 5th iterate is plotted for each method.

---

**Algorithm 1** Nesterov's Accelerated Gradient Method

---

Choose  $x^1 \in \mathbb{R}^n$ ,  $\alpha < \frac{1}{L}$ ;  
 Set  $x^0 = x^1$ ,  $t_0 = 1$ ;  
**for**  $k = 1, 2, \dots$  **do**  
 $t_k = \frac{\sqrt{4t_{k-1}^2 + 1} + 1}{2}$ ;  
 $y^k = x^k + \frac{t_{k-1} - 1}{t_k}(x^k - x^{k-1})$ ;  
 $x^{k+1} = y^k - \alpha \nabla f(y^k)$ ;  
**end for**

---

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 the method of Nesterov [7]. With  $f$  defined by (37), and defining

$$(40) \quad \gamma_k = t_{k-1} - 1,$$

the update formula can be written as

$$x^{k+1} = \frac{t_k + \gamma_k}{t_k}(I - \alpha H)x^k - \frac{\gamma_k}{t_k}(I - \alpha H)x^{k-1},$$

which because of (38) is equivalent to

$$(41) \quad x_i^{k+1} = \frac{t_k + \gamma_k}{t_k}(1 - \alpha\lambda_i)x_i^k - \frac{\gamma_k}{t_k}(1 - \alpha\lambda_i)x_i^{k-1}, \quad i = 1, 2, \dots, n.$$

It is easy to see from the definitions of  $t_k$  and  $\gamma_k$  that

$$(42a) \quad \{t_k\}_{k=0,1,2,\dots} \text{ is an increasing sequence, with } t_0 = 1;$$

$$(42b) \quad \{\gamma_k\}_{k=1,2,\dots}, \text{ is an increasing sequence, with } \gamma_1 = 0.$$

It can also be shown that

$$(43) \quad \frac{\gamma_k}{t_k} \text{ is an increasing nonnegative sequence}$$

and

$$(44) \quad \lim_{k \rightarrow \infty} \frac{\gamma_k}{t_k} = 1.$$

(For completeness, we include proofs of these two claims in the appendix.)

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

**THEOREM 9.** *For all  $i$  such that  $\lambda_i < 0$ , we have from (41) that*

$$(45) \quad x_i^{k+1} = x_i^1 \prod_{m=1}^k (1 + b_{i,m})$$

where

$$(46) \quad b_{i,k} = \begin{cases} \alpha|\lambda_i|, & \text{for } k = 1 \\ \frac{\gamma_k}{t_k} \left(1 - \frac{1 + \alpha|\lambda_i|}{1 + b_{i,k-1}}\right) + \alpha|\lambda_i| \left(1 + \frac{\gamma_k}{t_k}\right), & \text{otherwise.} \end{cases}$$

We have in addition that

$$(47) \quad b_{i,k+1} \geq b_{i,k}, \quad k = 1, 2, \dots$$

*Proof.* To begin, we prove that (46) holds for  $k = 1$ . Since  $\gamma_1 = 0$ , the update formula (41) is

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

so  $b_{i,1} = \alpha|\lambda_i|$ , as required. Now assume that (45) holds for all  $k \leq K-1$ . From (41), using the inductive hypothesis for  $K-1$  and  $K-2$ , we have

$$(48) \quad x_i^1 \prod_{m=1}^K (1 + b_{i,m}) = x_i^{K+1} = x_i^1 (1 - \alpha\lambda_i) \prod_{m=1}^{K-2} (1 + b_{i,m}) \left( \frac{t_K + \gamma_K}{t_K} (1 + b_{i,K-1}) - \frac{\gamma_K}{t_K} \right).$$

By dividing both sides by  $x_i^1 \prod_{m=1}^{K-1} (1 + b_{i,m})$ , we obtain

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

It follows that

$$\begin{aligned} b_{i,K} &= (1 - \alpha\lambda_i) \frac{\gamma_K}{t_K} \left( 1 - \frac{1}{1 + b_{i,K-1}} \right) - \alpha\lambda_i \\ &= \frac{\gamma_K}{t_K} \left( 1 - \alpha\lambda_i - \frac{1 - \alpha\lambda_i}{1 + b_{i,K-1}} \right) - \alpha\lambda_i \\ &= \frac{\gamma_K}{t_K} \left( 1 - \frac{1 + \alpha|\lambda_i|}{1 + b_{i,K-1}} \right) + \alpha|\lambda_i| \left( 1 + \frac{\gamma_K}{t_K} \right), \end{aligned}$$

proving (46).

Now we show by induction that  $b_{i,K+1} \geq b_{i,K}$  for all  $K \geq 1$ . When  $K = 1$ , we have:

$$b_{i,2} = \frac{\gamma_2}{t_2} \left( 1 - \frac{1 + \alpha|\lambda_i|}{1 + b_{i,1}} \right) + \alpha|\lambda_i| \left( 1 + \frac{\gamma_2}{t_2} \right) = \alpha|\lambda_i| \left( 1 + \frac{\gamma_2}{t_2} \right) \geq \alpha|\lambda_i| = b_{i,1},$$

where the second equality follows by  $b_{i,1} = \alpha|\lambda_i|$  and the inequality follows because  $\frac{\gamma_k}{t_k} \geq 0$  for all  $k$ , by (43).

Assume now that  $b_{i,k+1} \geq b_{i,k}$  holds for all  $k \leq K-1$ . We have

$$\begin{aligned} b_{i,K+1} &= \frac{\gamma_{K+1}}{t_{K+1}} \left( 1 - \frac{1 + \alpha|\lambda_i|}{1 + b_{i,K}} \right) + \alpha|\lambda_i| \left( 1 + \frac{\gamma_{K+1}}{t_{K+1}} \right) \\ &\geq \frac{\gamma_{K+1}}{t_{K+1}} \left( 1 - \frac{1 + \alpha|\lambda_i|}{1 + b_{i,K-1}} \right) + \alpha|\lambda_i| \left( 1 + \frac{\gamma_{K+1}}{t_{K+1}} \right) \quad \text{from (47)} \\ &\geq \frac{\gamma_K}{t_K} \left( 1 - \frac{1 + \alpha|\lambda_i|}{1 + b_{i,K-1}} \right) + \alpha|\lambda_i| \left( 1 + \frac{\gamma_K}{t_K} \right) = b_{i,K}. \end{aligned}$$

where the second inequality above follows from (43) and  $b_{i,K-1} \geq b_{i,1} = \alpha|\lambda_i|$ .  $\square$

Since  $b_{i,k} \geq \alpha|\lambda_i|$  for all  $k$ , Theorem 9 shows that Algorithm 1 diverges at least as quickly as gradient descent. We explore the limiting rate of divergence by finding a limit for the sequence  $b_{i,k}$ ,  $k = 1, 2, \dots$ .

**THEOREM 10.** *For all  $i$  such that  $\lambda_i < 0$ , we have  $\lim_{k \rightarrow \infty} b_{i,k} = \bar{b}_i$ , where  $\bar{b}_i$  is defined by*

$$(49) \quad \bar{b}_i := \alpha|\lambda_i| + \sqrt{\alpha|\lambda_i|} \sqrt{1 + \alpha|\lambda_i|}$$

*Proof.* We can write (41) as follows:

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

Recall from Theorem 9 that  $x_i^k = (1 + b_{i,k-1})x_i^{k-1}$ , where  $b_{i,k-1} \geq \alpha|\lambda_i|$ . By substituting into the equation above, we have

$$\begin{aligned} (50) \quad x_i^{k+1} &= \left[ (1 + \alpha|\lambda_i|)(1 + b_{i,k-1}) + \frac{\gamma_k}{t_k}(1 + \alpha|\lambda_i|)b_{i,k-1} \right] x_i^{k-1} \\ &= \left[ 1 + \alpha|\lambda_i| + (1 + \alpha|\lambda_i|) \left( 1 + \frac{\gamma_k}{t_k} \right) b_{i,k-1} \right] x_i^{k-1}. \end{aligned}$$

Using Theorem 9 again, we have

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

By matching this expression with (50), we obtain

$$(51) \quad \alpha|\lambda_i| + (1 + \alpha|\lambda_i|) \left( 1 + \frac{\gamma_k}{t_k} \right) b_{i,k-1} = b_{i,k} + b_{i,k-1} + b_{i,k-1}b_{i,k},$$

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

$$(52) \quad \frac{\alpha|\lambda_i|}{b_{i,k-1}} + (1 + \alpha|\lambda_i|) \left( 1 + \frac{\gamma_k}{t_k} \right) = \frac{b_{i,k}}{b_{i,k-1}} + 1 + b_{i,k}.$$



Now assume for contradiction that the nondecreasing sequence  $\{b_{i,k}\}_{k=1,2,\dots}$  has no finite limit, that is,  $b_{i,k} \rightarrow \infty$ . Recalling that  $\frac{\gamma^k}{t_k} \rightarrow 1$  by (44), we have by taking the limit as  $k \rightarrow \infty$  in (52) that the left-hand side approaches  $2(1 + \alpha|\lambda_i|)$ , while the right-hand side approaches  $\infty$ , a contradiction. Thus, the nondecreasing sequence  $\{b_{i,k}\}_{k=1,2,\dots}$  has a finite limit, which we denote by  $\bar{b}_i$ .

To find the value for  $\bar{b}_i$ , we take limits as  $k \rightarrow \infty$  in (51) to obtain

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

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

$$\bar{b}_i = \alpha|\lambda_i| \pm \sqrt{\alpha|\lambda_i|}\sqrt{1 + \alpha|\lambda_i|}.$$

By Theorem 9, we know that  $b_{i,k} \geq \alpha|\lambda_i|$  for all  $k$ , so that  $\bar{b}_i \geq \alpha|\lambda_i|$ . Therefore,  $\bar{b}_i$  satisfies (49), as claimed.  $\square$

We have shown that the limiting rate of divergence for Nesterov's method for the negative-eigenvalue components is  $1 + \alpha|\lambda_i| + \sqrt{1 + \alpha|\lambda_i|}\sqrt{\alpha|\lambda_i|}$ , a significant improvement on the rate of  $1 + \alpha|\lambda_i|$  attained by gradient descent. For the toy problem of the previous section, with  $\lambda_1 = 1$ ,  $\lambda_2 = -\delta$ , and  $\alpha = 1$ , we would see similar divergence behavior in the  $x_2$  component for Nesterov's method as for the heavy-ball method —  $1 + c\sqrt{\delta}$  in both cases, for a modest positive value of  $c$ .

**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  $e_n = (0, 0, \dots, 0, 1)^T$  at each iteration  $k$ , for Nesterov's method, heavy ball, and steepest descent. It also shows the divergence that would be attained if the theoretical limit  $\bar{b}_i$  from Theorem 10 applied at every iteration. Steepest descent and heavy-ball were run with  $\alpha = 1/L$ . Heavy ball uses (36) to calculate  $\beta$ , yielding  $\beta = 0.989$  in the case of  $\delta = .01$ . Nesterov's method is run with  $\alpha = 0.99/L$ .

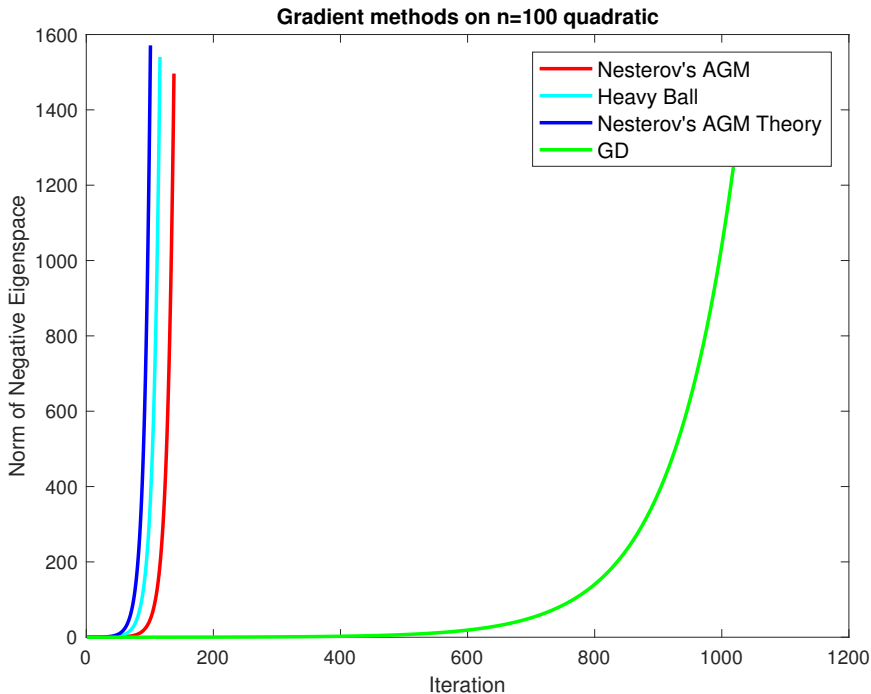
It is clear from Figure 2 that Nesterov's method and heavy-ball diverge at a significantly faster rate than steepest descent. In addition, there is only a small discrepancy between applying Nesterov's method and its limiting rate that is captured in the bound  $\bar{b}_i$ , suggesting that  $b_{i,k}$  approaches  $\bar{b}_i$  rapidly as  $k \rightarrow \infty$ .

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 Figure 2. Each trial was run until the norm of 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 Nesterov's 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 10 provides a strong indication of the practical behavior of Nesterov's method on these problems.

Fig. 2: Gradient methods and theoretical divergence applied to a quadratic function with  $n = 100$  and a single negative eigenvalue.



**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 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 Sections 3 and 4 show that when  $f$  is an indefinite quadratic, the accelerated gradient methods diverge faster than the steepest descent method.

Several extensions are possible. It would be interesting to extend the results on speed of divergence to non-quadratic smooth functions  $f$ . It would also be interesting to strengthen the results of Section 2 to show not just that the methods do not converge to a point of the type that we are considering, but also that they permanently escape the neighborhood of such points.

Another possible angle is to build on the recent work [4], in which the divergence rate of gradient descent from strict saddle points was used to obtain complexity results for a perturbed gradient descent method to second order stationary points. It is

Table 1: Divergence Behavior of Gradient Algorithms

$n$	$\delta$	Method	Av. Iters	Max. Iters	Av. Time
100	$10^{-2}$	Steepest Descent	379	518	0.011
		Nesterov's AGM	71	87	0.003
		$\bar{b}$ Divergence Rate	46	59	
100	$10^{-3}$	Steepest Descent	3855	5603	0.087
		Nesterov's AGM	242	299	0.009
		$\bar{b}$ Divergence Rate	155	194	
1000	$10^{-2}$	Steepest Descent	582	773	0.36
		Nesterov's AGM	99	116	0.17
		$\bar{b}$ Divergence Rate	71	85	
1000	$10^{-3}$	Steepest Descent	5775	8240	3.85
		Nesterov's AGM	332	399	0.61
		$\bar{b}$ Divergence Rate	235	282	

worth investigating whether an accelerated gradient method such as Algorithm 1 could escape saddle points at a faster rate and thus improve upon the complexity bounds of [4]. Another recent work [2] shows that gradient descent can take exponential time to escape from a set of saddle points, unlike the perturbed method in [4]. We believe that a similar result holds for accelerated methods as well.

**Acknowledgments.** We are grateful to Bin Hu for his advice and suggestions on the manuscript.

## REFERENCES

- [1] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.
- [2] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Barnabas Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. *arXiv preprint arXiv:1705.10412*, 2017.
- [3] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.
- [4] C. Jin, R. Ge, P. Netrapalli, S. Kakade, and M. I. Jordan. How to escape saddle points efficiently. Technical Report arXiv:1703.00887, University of California-Berkeley, March 2017.
- [5] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent converges to minimizers. Technical Report arXiv:1602.04915, University of California-Berkeley, 2016.
- [6] H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in neural information processing systems*, pages 379–387, 2015.
- [7] Y. Nesterov. A method for unconstrained convex problem with the rate of convergence  $O(1/k^2)$ . *Doklady AN SSSR*, 269:543–547, 1983.
- [8] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science and Business Media, New York, 2004.
- [9] B. T. Polyak. *Introduction to Optimization*. Optimization Software, 1987.
- [10] B. Recht and S. J. Wright. Nonlinear Optimization for Machine Learning, 2017. (Manuscript in preparation).
- [11] M. Shub. *Global stability of dynamical systems*. Springer, 1987.
- [12] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, Department of Mathematics, University of Washington, May 2008.
- [13] SK Zavriev and FV Kostyuk. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 4(4):336–341, 1993.

**Appendix A. Proof of (43) and (44).** We begin by noting two well known

properties of the sequence  $t_k$  (see for example Section 3.7.2 [1]):

$$(53) \quad t_k^2 - t_k = t_{k-1}^2$$

and

$$(54) \quad t_k \geq \frac{k+1}{2}.$$

To prove that  $\frac{\gamma_k}{t_k}$  is monotonically increasing, we need

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

Since  $t_{k+1} \geq t_k$  (42), 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 (53), we obtain the equivalent expression

$$(55) \quad t_{k-1} \leq \frac{t_k^2}{t_{k+1}} = \frac{t_{k+1}^2 - t_{k+1}}{t_{k+1}} = t_{k+1} - 1.$$

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 (55) holds, so the claim (43) is proved. The sequence  $\{\gamma_k/t_k\}$  is nonnegative, since  $\gamma_1/t_1 = 0$ .

Now we prove (44). We can lower-bound  $\gamma_k/t_k$  as follows:

$$(56) \quad \begin{aligned} \frac{\gamma_k}{t_k} &= \frac{t_{k-1} - 1}{t_k} = \frac{2(t_{k-1} - 1)}{\sqrt{4t_{k-1}^2 + 1} + 1} \\ &\geq \frac{2(t_{k-1} - 1)}{\sqrt{4t_{k-1}^2 + 2}} = \frac{2(t_{k-1} - 1)}{2(t_{k-1} + 1)} = 1 - \frac{2}{t_{k-1} + 1}. \end{aligned}$$

For an upper bound, we have from (42) that

$$(57) \quad \frac{\gamma_k}{t_k} = \frac{t_{k-1} - 1}{t_k} \leq \frac{t_{k-1}}{t_k} \leq 1.$$

Since  $t_{k-1} \rightarrow \infty$  (because of (54)), it follows from (56) and (57) that (44) holds.