Chapter 3

Gradient Methods Using Momentum and Memory

The steepest descent method described in Chapter 2 always steps in the negative gradient direction, which is orthogonal to the boundary of the level set for f at the current iterate. This direction can change sharply from one iteration to the next. For example, when the contours of f are narrow and elongated, this strategy can produce wild swings in direction, and small steps with only slow progress toward a solution.

The steepest descent method is "greedy" in a sense. It uses a direction that is always the most productive direction at this particular iterate, making no explicit use of knowledge gained about the function f at earlier iterations. In this chapter, we examine methods that encode knowledge of the function in various ways, and exploit this knowledge in their choice of search directions and step lengths. One such class of techniques makes use of *momentum*, in which the search direction tends to be similar to that one used on the previous step, with a small tweak in the direction of a negative gradient evaluated at the current point or a nearby point. Each search direction is thus a combination of all gradients encountered so far during the search — a compact encoding of the history of the search. Momentum methods in common use include the heavy-ball method, the conjugate gradient method, and Nesterov's accelerated gradient methods. We will also consider model-based methods, which construct an explicit model of the function f from information gathered at previous iterations, and uses this model to determine the search direction.

3.1 Motivation from Differential Equations

An intuition for momentum methods comes from looking at an optimization algorithm as a dynamical system. By taking the continuous limit of an algorithm, it often traces out the solution path of a differential equation. For instance, the gradient method is akin to moving down a potential well where the potential is defined by the gradient of f:

$$\frac{dx}{dt} = -\nabla f(x)$$

This differential equation has fixed points precisely when $\nabla f(x) = 0$, which are minimizers of a convex smooth function f.

There are other differential equations whose fixed points occur precisely at points for which $\nabla f(x) = 0$. Consider the second-order differential equation that governs a particle with mass moving in a potential defined by the gradient of f:

$$\mu \frac{d^2x}{dt^2} = -\nabla f(x) - b\frac{dx}{dt}. \tag{3.1}$$

As before, points x for which $\nabla f(x) = 0$ are fixed points of this ODE. In this setting μ governs the *mass* of the particle and b governs the friction dissipated during the evolution of the system. If $b = 0$, then the system may always gain acceleration. Trajectories tend to continue to move in the direction they were moving before, and heavier objects move down hill faster than light objects in the presence of friction. Note that in the limit as the mass μ goes to zero, we recover the ODE that we had derived for the gradient method.

If we approximate this ODE using simple finite differences, we have

$$\mu \frac{x(t+\Delta t) - 2x(t) + x(t-\Delta t)}{\Delta t^2} \approx -\nabla f(x(t)) - b\frac{x(t+\Delta t) - x(t)}{\Delta t}$$

Rearranging the terms in this expression and redefining parameters gives the finite difference equation:

$$x(t+\Delta t) = x(t) - \alpha \nabla f(x(t)) + \beta(x(t) - x(t-\Delta t)). \tag{3.2}$$

The algorithm defined by (3.2) is exactly *Heavy-Ball Method* of Polyak, for certain choices of the parameters α and β. With a minor modification, this becomes *Nesterov's optimal method.* When f is a convex quadratic, is known also as *Chebyshev's iterative method.*

By rewriting (3.2) in terms of discrete iterates, we obtain

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}), \tag{3.3}$$

Defining

$$p^k = x^{k+1} - x^k = -\alpha \nabla f(x^k) + \beta(x^k - x^{k-1}) = -\alpha \nabla f(x^k) + \beta p^{k-1}.$$

With this identification, we can rewrite the iteration in terms of two sequences:

$$\begin{aligned} x^{k+1} &= x^k + p^k \\ p^k &= -\alpha \nabla f(x^k) + \beta p^{k-1} \end{aligned}$$

Nesterov's optimal method (also known as *Nesterov's accelerated method*) is defined by the formula

$$x^{k+1} = x^k - \alpha \nabla f(x^k + \beta(x^k - x^{k-1})) + \beta(x^k - x^{k-1}). \tag{3.4}$$

This method only differs in the way that the underlying ODE (3.1) is discretized. In the two-state version, Nesterov's method becomes

$$\begin{aligned} x^{k+1} &= x^k + p^k \\ p^k &= -\alpha \nabla f(x^k + \beta p^{k-1}) + \beta p^{k-1} \end{aligned}$$

By a change of variable, Nesterov's optimal method can also be written in the following equivalent form:

$$\begin{aligned} y^k &= x^k + \beta^k(x^k - x^{k-1}) \\ x^{k+1} &= y^k - \alpha_k \nabla f(y^k). \end{aligned}$$

3.2 Heavy-Ball Method

In this section, we analyze the convergence behavior of the heavy-ball method (3.3), and derive suitable values for its parameters α and β. Our development is done in terms of an inhomogeneous convex quadratic objective

$$f(x) = \frac{1}{2}x^T Qx - b^T x + c, \tag{3.5}$$

with positive definite Hessian Q and eigenvalues

$$0 < m = \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 = L.$$

Note that $Qx^* = b$ at the minimizer x^* of f.

We'll show a powerful linear convergence result for the heavy-ball method on the quadratic function (3.5). But we need to be careful! This result does not imply similar convergence behavior on all strongly sonvex quadratics! A simple one-dimensional example from the Exercises serves to demonstrate the distinction.

By substituting $\nabla f(x^k) = Qx^k - b = Q(x^k - x^*)$ in (3.3), and sprinkling x^* liberally throughout this expression, we obtain

$$x^{k+1} - x^* = (x^k - x^*) - \alpha Q(x^k - x^*) + \beta\left((x^k - x^*) - (x^{k-1} - x^*)\right). \tag{3.6}$$

By concatenating the error vector $x^k - x^*$ over two successive steps, we obtain

$$\begin{bmatrix} x^{k+1} - x^* \\ x^k - x^* \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \alpha Q & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x^k - x^* \\ x^{k-1} - x^* \end{bmatrix} \tag{3.7}$$

By defining

$$w^k := \begin{bmatrix} x^{k+1} - x^* \\ x^k - x^* \end{bmatrix}, \quad T := \begin{bmatrix} (1+\beta)I - \alpha Q & -\beta I \\ I & 0 \end{bmatrix} \tag{3.8}$$

we can write the iteration (3.7) as

$$w^k = Tw^{k-1}, \quad k = 1, 2, \ldots. \tag{3.9}$$

The properties of T are key to analyzing the convergence of the sequence $\{w^k\}$ to zero. The following theorem shows how the eigenvalues of T depend on α, β, and the eigenvalues λ_i of Q.

Theorem 3.1. *If we choose β such that*

$$1 > \beta > \max\left(|1 - \sqrt{\alpha m}|, |1 - \sqrt{\alpha L}|\right)^2, \tag{3.10}$$

the matrix T has all complex eigenvalues, which are as follows:

$$\bar{\lambda}_{i,1} = \frac{1}{2}\left[(1 + \beta - \alpha\lambda_i) + i\sqrt{4\beta - (1 + \beta - \alpha\lambda_i)^2}\right],$$
$$\bar{\lambda}_{i,2} = \frac{1}{2}\left[(1 + \beta - \alpha\lambda_i) - i\sqrt{4\beta - (1 + \beta - \alpha\lambda_i)^2}\right], \quad i = 1, 2, \ldots, n.$$

We have $\bar{\lambda}_{i,1} \neq \bar{\lambda}_{i,2}$ for all i, and all eigenvalues have magnitude β.

Proof. We write the eigenvalue decomposition of Q as $Q = U\Lambda U^T$, where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. By defininining the permutation matrix Π as follows:

$$\Pi_{ij} = \begin{cases} 1 & i \text{ odd}, \ j = (i+1)/2 \\ 1 & i \text{ even}, \ j = n + (i/2) \\ 0 & \text{otherwise.} \end{cases}$$

we have by applying a similarity transformation to the matrix T that

$$\begin{aligned} \Pi \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}^T \begin{bmatrix} (1+\beta)I - \alpha Q & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \Pi^T &= \Pi \begin{bmatrix} (1+\beta)I - \alpha\Lambda & -\beta I \\ I & 0 \end{bmatrix} \Pi^T \\ &= \begin{bmatrix} T_1 & 0 & \ldots & 0 \\ 0 & T_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & T_n \end{bmatrix}, \end{aligned}$$

where

$$T_i = \begin{bmatrix} 1+\beta-\alpha\lambda_i & -\beta \\ 1 & 0 \end{bmatrix}, \quad i = 1, 2, \ldots, n.$$

The eigenvalues of T are the eigenvalues of T_i, for $i = 1, 2, \ldots, n$. The eigenvalues of T_i are the roots of the following quadratic:

$$u^2 - (1 + \beta - \alpha\lambda_i)u + \beta = 0,$$

which are given by the familiar formula:

$$u = \frac{1}{2}\left[(1+\beta-\alpha\lambda_i) \pm \sqrt{(1+\beta-\alpha\lambda_i)^2 - 4\beta}\right].$$

The two roots are distinct complex numbers when $(1+\beta-\alpha\lambda_i)^2 - 4\beta < 0$, which happens when

$$\beta \in \left((1-\sqrt{\alpha\lambda_i})^2, (1+\sqrt{\alpha\lambda_i})^2\right). \tag{3.11}$$

When β is in the range (3.10), it satisfies (3.11) for all i. Hence, the eigenvalues of each T_i are two distinct complex numbers, for all i, and have the values $\bar{\lambda}_{i,1}$, $\bar{\lambda}_{i,2}$ defined above.

It is easy to check that all $\bar{\lambda}_{i,1}$, $\bar{\lambda}_{i,2}$, $i = 1, 2, \ldots, n$, have magnitude β. □

Because $\bar{\lambda}_{i,1} \neq \bar{\lambda}_{i,2}$ for the chosen value of β, there exists a (complex) eigenvalue decomposition of each T_i, of the form

$$U_i^{-1} T_i U_i = S_i, \quad S_i = \begin{bmatrix} \bar{\lambda}_{i,1} & 0 \\ 0 & \bar{\lambda}_{i,2} \end{bmatrix}.$$

In fact, by composing U_i, $= 1, 2, \ldots, n$ with the orthgonal matrices U and Π defined in the proof above, we can define a $(2n) \times (2n)$ nonsingular matrix V such that

$$V^{-1} T V = S, \quad \text{where } S := \begin{bmatrix} S_1 & 0 & \ldots & 0 \\ 0 & S_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & S_n \end{bmatrix}.$$

Therefore, by defining

$$z^k := V^{-1} w^k, \tag{3.12}$$

we can express the relation (3.9) as

$$z^k = S z^{k-1} = S^2 z^{k-2} = \cdots = S^k z_0. \tag{3.13}$$

Since S is a diagonal matrix whose diagonal elements all have magnitude β, we have that

$$\|z^k\| = \beta^k \|z_0\|, \tag{3.14}$$

where $\|z\| = \sqrt{z^H z}$.

When we choose

$$\alpha := \frac{4}{(\sqrt{L} + \sqrt{m})^2}, \tag{3.15}$$

we find that

$$|1 - \sqrt{\alpha m}|^2 = |1 - \sqrt{\alpha L}|^2 = \left(\frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}} \right)^2, \tag{3.16}$$

so a choice of β that satisfies (3.10) would be

$$\beta := \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}}. \tag{3.17}$$

(Note that this value of β lies strictly between the lower bound (3.16) and the upper bound of 1, as required by (3.10).) When we use $\kappa = L/m$ to denote the condition number of Q, we have

$$\beta = 1 - \frac{2}{\sqrt{\kappa} + 1}. \tag{3.18}$$

We conclude with the following result concerning R-linear convergence of x^k to x^*.

Theorem 3.2. *For the choices (3.15) of α and (3.17) β, the heavy-ball iteration (3.3) produces a sequence $\{x^k\}$ that converges R-linearly to x^* with factor $\beta = 1 - 2/(\sqrt{\kappa} + 1)$.*

Proof. We have from (3.8), (3.12), and (3.14) that

$$\|x^k - x^*\| \leq \|w^k\| \leq \|V\| \|z^k\| = \beta^k \|V\| \|z_0\|,$$

proving the result. □

We have monotonic decrease in the norm of the transformed vector z^k (from (3.13)) but not the original error vector w^k. In fact, experiments can show sharp increases in w^k on early iterations before the R-linear convergence promised by Theorem 3.2 takes effect.

Let us compare the linear convergence of heavy ball against steepest descent, on nonconvex quadratics. Recall from (2.18) that the steepest-descent method with constant step $\alpha = 1/L$ requires $O(((L/m) \log \epsilon)$ iterations to obtain a reduction of factor ϵ in the function error $f(x^k) - f^*$. The rate defined by β in Theorem 3.2 suggests a complexity of $O(\sqrt{L/m} \log \epsilon)$ to obtain a reduction of factor ϵ in $\|w^k\|$ (a different quantity). For problems in which the condition number $\kappa = L/m$ is moderate to large, the heavy-ball method has a significant advantage. For example, if $\kappa = 1000$, the improved rate translates into a factor-of-30 reduction in number of iterations required, with virtually identical workload per iteration (one gradient evaluation and a few vector operations).