Nesterov's Accelerated Gradient Method

Momentum, Lyapunov Functions, and Convergence

Contents

1 Why introduce momentum? 2
- 1.1 Gradient descent . . . 2
- 1.2 Heavy-ball momentum . . . 2
- 1.3 Nesterov momentum . . . 3

2 Lyapunov method 3
- 2.1 Why a decreasing scalar quantity is useful . . . 4
- 2.2 Lyapunov functions in optimization . . . 5
- 2.3 Example: gradient flow . . . 6
- 2.4 Optional motivation: the accelerated ODE . . . 7
- 2.5 The discrete-time analogue . . . 8
- 2.6 Example: gradient descent . . . 9

3 Nesterov acceleration: convex and smooth case 11
- 3.1 The three-sequence formulation . . . 11
- 3.2 The Lyapunov function . . . 12
- 3.3 One-step Lyapunov estimate . . . 13
- 3.4 Monotonicity of the energy . . . 14
- 3.5 Quadratic growth of the weight . . . 15
- 3.6 Accelerated convergence theorem . . . 16

4 Recovering the standard two-sequence Nesterov method 17
- 4.1 An important parameter identity . . . 17
- 4.2 The usual t_k-notation . . . 18
- 4.3 Relation to the explicit weights $A_k = k(k+1)/(4L)$ . . . 18

5 Nesterov acceleration: strongly convex case 19
- 5.1 Three-sequence formulation . . . 19
- 5.2 The strongly convex Lyapunov function . . . 20
- 5.3 A clean contraction proof . . . 20

6 Standard two-sequence form in the strongly convex case 23

7 Complexity comparison 24

8 How to explain the Lyapunov strategy 25

9 Common sources of confusion 26

1 Why introduce momentum?

1.1 Gradient descent

The gradient descent method with step size $1/L$ is

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k).$$

For convex and L-smooth functions, the usual rate is

$$f(x_k) - f^\star = O\left(\frac{1}{k}\right).$$

For μ-strongly convex and L-smooth functions, gradient descent converges linearly:

$$f(x_k) - f^\star \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_0) - f^\star\right).$$

Since

$$1 - \frac{\mu}{L} = 1 - \frac{1}{\kappa},$$

gradient descent requires on the order of

$$O\left(\kappa \log \frac{1}{\varepsilon}\right)$$

iterations to obtain error at most ε.

Nesterov acceleration improves these rates to

$$O\left(\frac{1}{k^2}\right)$$

in the convex case and

$$O\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$$

in the strongly convex case.

1.2 Heavy-ball momentum

Polyak's heavy-ball method takes the form

$$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1}).$$

The vector

$$x_k - x_{k-1}$$

represents the previous direction of motion. The parameter $\beta \geq 0$ introduces inertia or momentum.

The method first computes the gradient at x_k, and then adds a momentum correction:

gradient at the current point + momentum.

Heavy-ball momentum can be extremely effective for quadratic functions. However, parameter choices that are optimal for quadratics do not automatically guarantee convergence for all smooth strongly convex functions.

1.3 Nesterov momentum

Nesterov's method changes the order of the operations. In one common notation,

$$y_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{1}$$
$$x_{k+1} = y_{k+1} + \beta_k (y_{k+1} - y_k). \tag{2}$$

The point x_k is an extrapolated point. The gradient is evaluated at x_k, rather than at y_k.

Equivalently, if

$$x_k = y_k + \beta_{k-1}(y_k - y_{k-1}),$$

then

$$y_{k+1} = y_k + \beta_{k-1}(y_k - y_{k-1}) - \alpha_k \nabla f\left(y_k + \beta_{k-1}(y_k - y_{k-1})\right).$$

Thus Nesterov's method has the schematic form

extrapolate first $\longrightarrow$ evaluate the gradient at the extrapolated point.

This is different from heavy-ball momentum, where the gradient is evaluated before the momentum correction is added.

Teaching point 1.1. It is useful to distinguish three types of points:

- y_k: the point produced by the previous gradient step;
- x_k: the extrapolated point where the next gradient is evaluated;
- z_k: an auxiliary momentum state used in the convergence proof.

The sequence z_k may initially seem artificial, but it is exactly the variable that makes the Lyapunov cancellations transparent.

2 Lyapunov method

Consider the autonomous differential equation

$$\dot{z}(t) = g(z(t)),$$

where $g : \mathbb{R}^m \to \mathbb{R}^m$, and let $z(t)$ be a sufficiently regular solution. A point $z^\star$ is an equilibrium if

$$g(z^\star) = 0.$$

For simplicity, suppose that the equilibrium of interest is $z^\star = 0$.

The state $z(t)$ may be high-dimensional, and it may be difficult to understand its trajectory directly. The Lyapunov idea is to replace this vector-valued evolution by a carefully chosen scalar quantity

$$V(z(t)).$$

This scalar quantity is designed to summarize how far the system is from the desired equilibrium and how that distance evolves under the dynamics.

A useful Lyapunov function must therefore perform two different tasks:

(i) it must measure the behavior we want to control;

(ii) its evolution must be compatible with the differential equation.

For stability of the equilibrium 0, the first task is usually expressed by requiring

$$V(0) = 0 \qquad \text{and} \qquad V(z) > 0 \quad \text{for } z \neq 0.$$

Thus V assigns zero energy to the equilibrium and positive energy to every other state.

The second task is to show that the dynamics cannot increase this energy. Along a trajectory $z(t)$, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} V(z(t)) = \langle \nabla V(z(t)), \dot{z}(t) \rangle .$$

Substituting the equation of motion $\dot{z}(t) = g(z(t))$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} V(z(t)) = \langle \nabla V(z(t)), g(z(t)) \rangle .$$

The expression

$$\dot{V}(z) := \langle \nabla V(z), g(z) \rangle$$

is the derivative of V in the direction of the vector field g.

If

$$\dot{V}(z) \leq 0,$$

then the vector field always points in a direction in which V does not increase. Consequently, for $t_2 \geq t_1$,

$$V(z(t_2)) \leq V(z(t_1)).$$

Indeed,

$$V(z(t_2)) - V(z(t_1)) = \int_{t_1}^{t_2} \dot{V}(z(s)) \, \mathrm{d}s \leq 0.$$

This integration step is elementary. It is not the main content of the Lyapunov method. The main problem is to discover a function V for which the sign of

$$\langle \nabla V(z), g(z) \rangle$$

can be controlled.

2.1 Why a decreasing scalar quantity is useful

Suppose V is positive away from the equilibrium and nonincreasing along trajectories. Then

$$V(z(t)) \leq V(z(0)) \qquad \text{for every } t \geq 0.$$

The trajectory is therefore confined to the sublevel set

$$\{ z \in \mathbb{R}^m : V(z) \leq V(z(0)) \} .$$

If small sublevel sets of V are small neighborhoods of the origin, then a trajectory that starts close to the origin cannot move far away from it. This is the basic mechanism by which a Lyapunov function proves stability.

It is important to distinguish stability from convergence. The inequality

$$\dot{V}(z) \leq 0$$

only says that the energy cannot increase. It does not, by itself, imply that

$$z(t) \longrightarrow 0.$$

The energy may remain constant along a nontrivial trajectory. To prove convergence to the equilibrium, one needs additional information, such as strict decay

$$\dot{V}(z) < 0 \qquad \text{for } z \neq 0,$$

or a more refined invariance argument.

Thus a Lyapunov function translates questions about a complicated trajectory into questions about a scalar quantity:

Does V measure the desired notion of error?
Does the dynamics force V to decrease?

Both questions are essential.

2.2 Lyapunov functions in optimization

In optimization, the objective is not merely to prove that a trajectory remains near an equilibrium. We want to control an optimization error, such as

$$f(x(t)) - f^\star, \qquad f^\star = \min_x f(x),$$

and ideally obtain a quantitative convergence rate.

For this purpose, it is often useful to allow the Lyapunov function to depend explicitly on time. We then consider an energy

$$\mathcal{E}(t) = \mathcal{V}(t, x(t)).$$

Its derivative along a trajectory is

$$\mathcal{E}'(t) = \partial_t \mathcal{V}(t, x(t)) + \langle \nabla_x \mathcal{V}(t, x(t)), \dot{x}(t) \rangle .$$

Once again, the differential equation is substituted for $\dot{x}(t)$. The energy is designed so that the resulting terms cancel or combine into nonpositive quantities.

A common form is

$$\mathcal{E}(t) = A(t)\big(f(x(t)) - f^\star\big) + R(t),$$

where

$$A(t) > 0 \qquad \text{and} \qquad R(t) \geq 0.$$

Here $R(t)$ may measure distance to a minimizer, velocity, momentum, or a combination of these quantities.

If the dynamics and the assumptions on f imply

$$\mathcal{E}'(t) \leq 0,$$

then

$$\mathcal{E}(t) \leq \mathcal{E}(0).$$

Since $R(t) \geq 0$, we also have

$$A(t)\big(f(x(t)) - f^\star\big) \leq \mathcal{E}(t).$$

Combining the two inequalities gives

$$f(x(t)) - f^\star \leq \frac{\mathcal{E}(0)}{A(t)}.$$

The growth of $A(t)$ therefore determines the convergence rate. For example,

$$A(t) \asymp t \quad \Longrightarrow \quad f(x(t)) - f^\star = O\left(\frac{1}{t}\right),$$

whereas

$$A(t) \asymp t^2 \quad \Longrightarrow \quad f(x(t)) - f^\star = O\left(\frac{1}{t^2}\right).$$

This is the central Lyapunov strategy in optimization:

1. Decide which error and which convergence rate should be proved.
2. Construct an energy that contains the appropriately weighted error.
3. Add auxiliary terms that make the derivative manageable.
4. Differentiate the complete energy.
5. Substitute the dynamics.
6. Use properties such as convexity or strong convexity to show that the resulting expression is nonpositive.
7. Read the convergence estimate from the energy.

The creative step is the construction of the energy. The auxiliary terms are not arbitrary: they are chosen so that their derivatives cancel unfavorable terms arising from the weighted objective gap.

2.3 Example: gradient flow

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and convex, and suppose that $x^\star$ is a minimizer of f. Consider the gradient-flow equation

$$\dot{x}(t) = -\nabla f(x(t)).$$

The objective value itself is nonincreasing, because

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) = \langle \nabla f(x(t)), \dot{x}(t) \rangle = -\left\|\nabla f(x(t))\right\|^2 \leq 0.$$

This is useful, but it only shows monotonicity of the objective value. It does not yet provide a rate at which $f(x(t))$ approaches $f^\star$.

To obtain a rate, consider instead the time-dependent energy

$$\mathcal{E}(t) = t\big(f(x(t)) - f^\star\big) + \frac{1}{2}\left\|x(t) - x^\star\right\|^2.$$

This energy has been designed with two components:

- The term
$$t\big(f(x(t)) - f^\star\big)$$
encodes the desired $O(1/t)$ rate.
- The term
$$\frac{1}{2}\left\|x(t) - x^\star\right\|^2$$
is included because its derivative will interact favorably with convexity.

We now differentiate the complete energy:

$$\mathcal{E}'(t) = f(x(t)) - f^\star + t\,\langle \nabla f(x(t)), \dot{x}(t)\rangle + \langle x(t) - x^\star, \dot{x}(t)\rangle\,.$$

Substituting the gradient-flow equation

$$\dot{x}(t) = -\nabla f(x(t))$$

gives

$$\mathcal{E}'(t) = f(x(t)) - f^\star - t\,\|\nabla f(x(t))\|^2 - \langle \nabla f(x(t)), x(t) - x^\star\rangle\,.$$

Convexity implies

$$f(x^\star) \geq f(x(t)) + \langle \nabla f(x(t)), x^\star - x(t)\rangle\,.$$

Equivalently,

$$f(x(t)) - f^\star \leq \langle \nabla f(x(t)), x(t) - x^\star\rangle\,.$$

Therefore,

$$\mathcal{E}'(t) \leq -t\,\|\nabla f(x(t))\|^2 \leq 0.$$

The important point is the cancellation

$$f(x(t)) - f^\star - \langle \nabla f(x(t)), x(t) - x^\star\rangle \leq 0.$$

The distance term was added precisely to produce the inner product needed for this convexity inequality.

It follows that

$$\mathcal{E}(t) \leq \mathcal{E}(0).$$

Since

$$\mathcal{E}(0) = \frac{1}{2}\,\|x(0) - x^\star\|^2$$

and the distance term in $\mathcal{E}(t)$ is nonnegative, we obtain, for every $t > 0$,

$$t\big(f(x(t)) - f^\star\big) \leq \mathcal{E}(t) \leq \frac{1}{2}\,\|x(0) - x^\star\|^2\,.$$

Consequently,

$$f(x(t)) - f^\star \leq \frac{\|x(0) - x^\star\|^2}{2t}.$$

This example illustrates the Lyapunov method more fully than the simple observation that $f(x(t))$ decreases. The rate follows from a specifically designed combination of the objective gap and the distance to a minimizer.

2.4 Optional motivation: the accelerated ODE

A continuous-time model often associated with Nesterov acceleration is

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \qquad t > 0. \tag{3}$$

For a convex function f, consider the energy

$$\mathcal{E}(t) = t^2\big(f(x(t)) - f^\star\big) + 2\left\|x(t) + \frac{t}{2}\dot{x}(t) - x^\star\right\|^2.$$

Set

$$u(t) := x(t) + \frac{t}{2}\dot{x}(t) - x^\star.$$

Using the ODE,

$$\begin{aligned}\dot{u}(t) &= \dot{x}(t) + \frac{1}{2}\dot{x}(t) + \frac{t}{2}\ddot{x}(t) \\ &= \frac{3}{2}\dot{x}(t) + \frac{t}{2}\left(-\frac{3}{t}\dot{x}(t) - \nabla f(x(t))\right) \\ &= -\frac{t}{2}\nabla f(x(t)).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(t) &= 2t(f(x(t)) - f^\star) + t^2\left\langle \nabla f(x(t)), \dot{x}(t)\right\rangle + 4\left\langle u(t), \dot{u}(t)\right\rangle \\ &= 2t(f(x(t)) - f^\star) + t^2\left\langle \nabla f(x(t)), \dot{x}(t)\right\rangle \\ &\quad - 2t\left\langle x(t) + \frac{t}{2}\dot{x}(t) - x^\star, \nabla f(x(t))\right\rangle \\ &= 2t\left(f(x(t)) - f^\star - \langle \nabla f(x(t)), x(t) - x^\star\rangle\right) \\ &\leq 0,\end{aligned}$$

where the last inequality follows from convexity.

Consequently,

$$f(x(t)) - f^\star \leq \frac{\mathcal{E}(0)}{t^2}.$$

This calculation gives the continuous-time analogue of accelerated $O(t^{-2})$ convergence.

2.5 The discrete-time analogue

For an iterative method, the trajectory $x(t)$ is replaced by a sequence $x_0, x_1, \ldots$. A discrete Lyapunov function is a sequence of nonnegative energies

$$\mathcal{E}_0, \mathcal{E}_1, \ldots$$

constructed so that

$$\mathcal{E}_{k+1} - \mathcal{E}_k \leq 0.$$

Then

$$\mathcal{E}_k \leq \mathcal{E}_0.$$

As in continuous time, this final implication is elementary. The main work is to design $\mathcal{E}_k$ so that:

(i) it controls the optimization error;

(ii) the update rule makes the energy difference nonpositive.

For example, if

$$\mathcal{E}_k = A_k\big(f(x_k) - f^\star\big) + R_k, \qquad R_k \geq 0,$$

and one proves

$$\mathcal{E}_{k+1} \leq \mathcal{E}_k,$$

then

$$f(x_k) - f^\star \leq \frac{\mathcal{E}_0}{A_k}.$$

The continuous- and discrete-time viewpoints are therefore parallel:

$$\mathcal{E}'(t) \leq 0 \qquad \longleftrightarrow \qquad \mathcal{E}_{k+1} - \mathcal{E}_k \leq 0.$$

In both cases, the Lyapunov idea is to construct a scalar certificate whose evolution reflects the desired behavior of the original system.

A Lyapunov proof is often best understood backwards:

1. Decide which convergence rate you want.
2. Choose a weight A_k with the corresponding growth.
3. Add an auxiliary nonnegative term.
4. Design the algorithmic parameters so that the increment $\mathcal{E}_{k+1} - \mathcal{E}_k$ is nonpositive.

2.6 Example: gradient descent

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, and let $x^\star$ be a minimizer of f. Consider gradient descent with step size

$$0 < \alpha \leq \frac{1}{L} :$$
$$x_{k+1} = x_k - \alpha \nabla f(x_k).$$

We will construct a discrete Lyapunov function that proves

$$f(x_k) - f^\star = O\left(\frac{1}{k}\right), \qquad f^\star = f(x^\star).$$

Define

$$\boxed{\mathcal{E}_k = k\big(f(x_k) - f^\star\big) + \frac{1}{2\alpha}\left\|x_k - x^\star\right\|^2.}$$

The factor k is chosen to encode the desired $O(1/k)$ rate. The distance term is included because its change under the gradient descent update produces the terms needed to control the objective gap.

We begin by deriving the key one-step inequality.

By L-smoothness,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\left\|y - x\right\|^2.$$

Applying this inequality with $x = x_k$ and $y = x_{k+1}$, and using

$$x_{k+1} - x_k = -\alpha \nabla f(x_k),$$

gives

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \alpha \left\|\nabla f(x_k)\right\|^2 + \frac{L\alpha^2}{2}\left\|\nabla f(x_k)\right\|^2 \\ &= f(x_k) - \alpha\left(1 - \frac{L\alpha}{2}\right)\left\|\nabla f(x_k)\right\|^2. \end{aligned}$$

Because $\alpha \leq 1/L$,

$$1 - \frac{L\alpha}{2} \geq \frac{1}{2}.$$

Therefore,

$$\boxed{f(x_{k+1}) \leq f(x_k) - \frac{\alpha}{2}\left\|\nabla f(x_k)\right\|^2.}$$

In particular,

$$f(x_{k+1}) \leq f(x_k),$$

so the objective values are nonincreasing.

Convexity gives

$$f(x_k) - f^\star \le \langle \nabla f(x_k), x_k - x^\star \rangle .$$

Combining this inequality with the descent estimate yields

$$\begin{aligned} f(x_{k+1}) - f^\star &= f(x_{k+1}) - f(x_k) + f(x_k) - f^\star \\ &\le \langle \nabla f(x_k), x_k - x^\star \rangle - \frac{\alpha}{2} \|\nabla f(x_k)\|^2 . \end{aligned}$$

We now examine the change in the squared distance to a minimizer. From the gradient descent update,

$$\begin{aligned} \|x_{k+1} - x^\star\|^2 &= \|x_k - x^\star - \alpha \nabla f(x_k)\|^2 \\ &= \|x_k - x^\star\|^2 - 2\alpha \langle \nabla f(x_k), x_k - x^\star \rangle + \alpha^2 \|\nabla f(x_k)\|^2 . \end{aligned}$$

Rearranging gives

$$\frac{1}{2\alpha} \left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2 \right) = \langle \nabla f(x_k), x_k - x^\star \rangle - \frac{\alpha}{2} \|\nabla f(x_k)\|^2 .$$

Consequently,

$$\boxed{f(x_{k+1}) - f^\star \le \frac{1}{2\alpha} \left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2 \right).}$$

This is the key one-step estimate. It says that the decrease in the distance term is large enough to pay for one copy of the objective gap at the next iterate.

We can now verify that the Lyapunov function decreases. Its one-step change is

$$\begin{aligned} \mathcal{E}_{k+1} - \mathcal{E}_k &= (k+1)\big(f(x_{k+1}) - f^\star\big) - k\big(f(x_k) - f^\star\big) \\ &\quad + \frac{1}{2\alpha} \left(\|x_{k+1} - x^\star\|^2 - \|x_k - x^\star\|^2 \right). \end{aligned}$$

The key one-step estimate implies

$$\frac{1}{2\alpha} \left(\|x_{k+1} - x^\star\|^2 - \|x_k - x^\star\|^2 \right) \le -\big(f(x_{k+1}) - f^\star\big).$$

It follows that

$$\begin{aligned} \mathcal{E}_{k+1} - \mathcal{E}_k &\le k\big(f(x_{k+1}) - f(x_k)\big) \\ &\le 0, \end{aligned}$$

where the last inequality follows from the monotonicity of the objective values. Thus

$$\mathcal{E}_{k+1} \le \mathcal{E}_k,$$

and hence

$$\mathcal{E}_k \le \mathcal{E}_0.$$

At $k = 0$, the weighted objective gap vanishes, so

$$\mathcal{E}_0 = \frac{1}{2\alpha} \|x_0 - x^\star\|^2 .$$

Moreover, the distance term in $\mathcal{E}_k$ is nonnegative. Therefore, for every $k \ge 1$,

$$k\big(f(x_k) - f^\star\big) \le \mathcal{E}_k \le \mathcal{E}_0 = \frac{1}{2\alpha} \|x_0 - x^\star\|^2 .$$

Dividing by k gives

$$\boxed{f(x_k) - f^\star \le \frac{\|x_0 - x^\star\|^2}{2\alpha k}.}$$

Thus gradient descent satisfies an $O(1/k)$ convergence rate for every step size

$$0 < \alpha \le \frac{1}{L}.$$

A smaller step size gives the same order of convergence, but the bound has a larger constant proportional to $1/\alpha$.

The Lyapunov mechanism can be seen particularly clearly by writing

$$\delta_k = f(x_k) - f^\star, \qquad R_k = \frac{1}{2\alpha}\|x_k - x^\star\|^2.$$

The key one-step estimate says

$$R_{k+1} - R_k \le -\delta_{k+1}.$$

On the other hand,

$$\mathcal{E}_k = k\delta_k + R_k,$$

so increasing the weight from k to $k+1$ creates one additional copy of δ_{k+1}:

$$\begin{aligned}\mathcal{E}_{k+1} - \mathcal{E}_k &= (k+1)\delta_{k+1} - k\delta_k + R_{k+1} - R_k \\ &\le k(\delta_{k+1} - \delta_k) \le 0.\end{aligned}$$

The decrease in the distance term pays for the additional objective-gap term. This cancellation is the central idea behind the discrete Lyapunov construction.

3 Nesterov acceleration: convex and smooth case

Assume throughout this section that:

- $f : \mathbb{R}^d \to \mathbb{R}$ is convex;
- f is L-smooth;
- $x^\star \in \arg\min f$.

3.1 The three-sequence formulation

Let

$$A_0 = 0.$$

For each $k \ge 0$, choose $a_{k+1} > 0$ and define

$$A_{k+1} = A_k + a_{k+1}.$$

We impose the key relation

$$La_{k+1}^2 = A_{k+1}. \tag{4}$$

Given A_k, this relation determines a_{k+1}. Indeed,

$$La_{k+1}^2 = A_k + a_{k+1},$$

and hence

$$a_{k+1} = \frac{1 + \sqrt{1 + 4LA_k}}{2L}. \tag{5}$$

Define

$$\tau_k := \frac{a_{k+1}}{A_{k+1}}.$$

Starting from

$$y_0 = z_0 = x^{\text{init}},$$

consider the iteration

$$\begin{aligned} x_k &= \frac{A_k}{A_{k+1}} y_k + \frac{a_{k+1}}{A_{k+1}} z_k \\ &= (1-\tau_k) y_k + \tau_k z_k, \qquad (6) \\ y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k), \qquad (7) \\ z_{k+1} &= z_k - a_{k+1}\nabla f(x_k). \qquad (8) \end{aligned}$$

The roles of the three sequences are:

- x_k is the point where the gradient is evaluated;
- y_{k+1} is obtained by a usual gradient step from x_k;
- z_k stores accumulated gradient or momentum information.

The point x_k is a convex combination of y_k and z_k.

Remark 3.1. At $k = 0$, relation (4) gives

$$a_1 = A_1 = \frac{1}{L}, \qquad \tau_0 = 1.$$

Thus $x_0 = z_0$.

For $k \geq 1$, one has $0 < \tau_k < 1$.

3.2 The Lyapunov function

Define

$$\mathcal{E}_k := \frac{1}{2}\|z_k - x^\star\|^2 + A_k\big(f(y_k) - f^\star\big). \qquad (9)$$

The two terms have different roles.

- The optimization error
$$A_k(f(y_k) - f^\star)$$
is the quantity from which we will read off the rate.
- The auxiliary term
$$\frac{1}{2}\|z_k - x^\star\|^2$$
creates inner products involving $z_k - x^\star$. Those inner products will cancel the terms obtained from convexity.

Since both terms are nonnegative,

$$A_k(f(y_k) - f^\star) \leq \mathcal{E}_k.$$

Thus, if we can prove $\mathcal{E}_k \leq \mathcal{E}_0$, then

$$f(y_k) - f^\star \leq \frac{\mathcal{E}_0}{A_k}.$$

The entire proof therefore reduces to two tasks:

1. prove $\mathcal{E}_{k+1} \leq \mathcal{E}_k$;
2. prove A_k grows quadratically.

3.3 One-step Lyapunov estimate

For conciseness, define

$$g_k := \nabla f(x_k).$$

Lemma 3.2 (Convex Lyapunov increment)**.** *For the iteration (6)–(8),*

$$\mathcal{E}_{k+1} - \mathcal{E}_k \le \frac{1}{2}\left(a_{k+1}^2 - \frac{A_{k+1}}{L}\right)\|g_k\|^2 .$$

Proof. We divide the calculation into three parts.

Step 1: change of the distance term. Since

$$z_{k+1} = z_k - a_{k+1}g_k,$$

we have

$$\begin{aligned}\frac{1}{2}\|z_{k+1} - x^\star\|^2 - \frac{1}{2}\|z_k - x^\star\|^2 &= \frac{1}{2}\|z_k - x^\star - a_{k+1}g_k\|^2 - \frac{1}{2}\|z_k - x^\star\|^2 \\ &= -a_{k+1}\langle g_k, z_k - x^\star\rangle + \frac{a_{k+1}^2}{2}\|g_k\|^2 . \end{aligned} \tag{10}$$

Step 2: change of the weighted objective error. Using $A_{k+1} = A_k + a_{k+1}$, we write

$$\begin{aligned}&A_{k+1}(f(y_{k+1}) - f^\star) - A_k(f(y_k) - f^\star) \\ &\quad = A_{k+1}(f(y_{k+1}) - f(x_k)) + a_{k+1}(f(x_k) - f^\star) + A_k(f(x_k) - f(y_k)).\end{aligned} \tag{11}$$

Combining (10) and (11) gives

$$\mathcal{E}_{k+1} - \mathcal{E}_k = \frac{a_{k+1}^2}{2}\|g_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x_k)) + R_k, \tag{12}$$

where

$$R_k := a_{k+1}(f(x_k) - f^\star) + A_k(f(x_k) - f(y_k)) - a_{k+1}\langle g_k, z_k - x^\star\rangle .$$

Step 3: use convexity to show $R_k \le 0$. By convexity,

$$f(x_k) - f^\star \le \langle g_k, x_k - x^\star\rangle ,$$

and

$$f(x_k) - f(y_k) \le \langle g_k, x_k - y_k\rangle .$$

Therefore,

$$\begin{aligned}R_k &\le a_{k+1}\langle g_k, x_k - x^\star\rangle + A_k\langle g_k, x_k - y_k\rangle - a_{k+1}\langle g_k, z_k - x^\star\rangle \\ &= \langle g_k, a_{k+1}(x_k - z_k) + A_k(x_k - y_k)\rangle .\end{aligned}$$

But the definition of x_k gives

$$A_{k+1}x_k = A_k y_k + a_{k+1}z_k.$$

Since $A_{k+1} = A_k + a_{k+1}$,

$$\begin{aligned}a_{k+1}(x_k - z_k) + A_k(x_k - y_k) &= (A_k + a_{k+1})x_k - A_k y_k - a_{k+1}z_k \\ &= 0.\end{aligned}$$

Hence

$$R_k \leq 0.$$

It follows from (12) that

$$\mathcal{E}_{k+1} - \mathcal{E}_k \leq \frac{a_{k+1}^2}{2} \|g_k\|^2 + A_{k+1}(f(y_{k+1}) - f(x_k)).$$

Finally, since

$$y_{k+1} = x_k - \frac{1}{L} g_k,$$

smoothness gives

$$f(y_{k+1}) - f(x_k) \leq -\frac{1}{2L} \|g_k\|^2 .$$

Therefore,

$$\mathcal{E}_{k+1} - \mathcal{E}_k \leq \frac{1}{2} \left(a_{k+1}^2 - \frac{A_{k+1}}{L} \right) \|g_k\|^2 .$$

□

Teaching point 3.3. The proof contains two independent cancellations.

1. The interpolation rule
$$A_{k+1} x_k = A_k y_k + a_{k+1} z_k$$
cancels all first-order convexity terms.

2. The weight relation
$$L a_{k+1}^2 = A_{k+1}$$
cancels the remaining squared-gradient term.

This illustrates that the algorithm and the Lyapunov function are designed together.

3.4 Monotonicity of the energy

By construction,

$$L a_{k+1}^2 = A_{k+1}.$$

Consequently,

$$a_{k+1}^2 - \frac{A_{k+1}}{L} = 0.$$

The previous lemma therefore gives

$$\mathcal{E}_{k+1} \leq \mathcal{E}_k.$$

By induction,

$$\mathcal{E}_k \leq \mathcal{E}_0.$$

Because $A_0 = 0$ and $z_0 = x^{\text{init}}$,

$$\mathcal{E}_0 = \frac{1}{2} \left\| x^{\text{init}} - x^\star \right\|^2 .$$

Thus, for $k \geq 1$,

$$f(y_k) - f^\star \leq \frac{\left\| x^{\text{init}} - x^\star \right\|^2}{2A_k}. \tag{13}$$

It remains to estimate the growth of A_k.

3.5 Quadratic growth of the weight

Define

$$t_k := \sqrt{LA_k}.$$

Since

$$La_{k+1}^2 = A_{k+1},$$

we have

$$a_{k+1} = \frac{t_{k+1}}{L}.$$

The identity $A_{k+1} = A_k + a_{k+1}$ becomes

$$\frac{t_{k+1}^2}{L} = \frac{t_k^2}{L} + \frac{t_{k+1}}{L}.$$

Therefore,

$$t_{k+1}^2 - t_{k+1} = t_k^2. \tag{14}$$

Equivalently,

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

Since $A_0 = 0$, we have $t_0 = 0$, and

$$t_1 = 1.$$

Lemma 3.4. *For every $k \geq 1$,*

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

Consequently,

$$A_k = \frac{t_k^2}{L} \geq \frac{(k+1)^2}{4L}.$$

Proof. We use induction.

For $k = 1$,

$$t_1 = 1 = \frac{1+1}{2}.$$

Suppose

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

Then

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

This proves the result. □

3.6 Accelerated convergence theorem

Theorem 3.5 (Nesterov acceleration for convex smooth functions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, and suppose $x^\star \in \arg\min f$.*

Let the sequences $A_k, a_{k+1}, x_k, y_k, z_k$ be generated by (4) and (6)–(8), with

$$y_0 = z_0 = x^{\text{init}}.$$

Then, for every $k \geq 1$,

$$f(y_k) - f^\star \leq \frac{2L \left\|x^{\text{init}} - x^\star\right\|^2}{(k+1)^2}.$$

Proof. The Lyapunov estimate gives

$$\mathcal{E}_k \leq \mathcal{E}_0 = \frac{1}{2} \left\|x^{\text{init}} - x^\star\right\|^2.$$

Since

$$A_k(f(y_k) - f^\star) \leq \mathcal{E}_k,$$

we obtain

$$f(y_k) - f^\star \leq \frac{\left\|x^{\text{init}} - x^\star\right\|^2}{2A_k}.$$

Using

$$A_k \geq \frac{(k+1)^2}{4L}$$

gives

$$f(y_k) - f^\star \leq \frac{2L \left\|x^{\text{init}} - x^\star\right\|^2}{(k+1)^2}.$$

□

Corollary 3.6 (Iteration complexity)**.** *To guarantee*

$$f(y_k) - f^\star \leq \varepsilon,$$

it is sufficient that

$$k + 1 \geq \sqrt{\frac{2L \left\|x^{\text{init}} - x^\star\right\|^2}{\varepsilon}}.$$

Thus the iteration complexity is

$$O\left(\sqrt{\frac{L}{\varepsilon}}\right),$$

up to the dependence on the initial distance.

Beginning of Lecture 2

4 Recovering the standard two-sequence Nesterov method

The three-sequence formulation is ideal for analysis, but the method is usually implemented using only two sequences.

We now show that the preceding algorithm is exactly equivalent to the standard Nesterov iteration.

4.1 An important parameter identity

Recall that

$$\tau_k = \frac{a_{k+1}}{A_{k+1}}.$$

The key weight relation implies

$$\tau_k a_{k+1} = \frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{L}.$$

Equivalently,

$$a_{k+1} = \frac{1}{L\tau_k}. \tag{15}$$

This identity makes it possible to eliminate z_k.

From

$$x_k = (1-\tau_k)y_k + \tau_k z_k,$$

we obtain

$$z_k = y_k + \frac{1}{\tau_k}(x_k - y_k).$$

Furthermore,

$$y_{k+1} - x_k = -\frac{1}{L}\nabla f(x_k).$$

Using (15),

$$\begin{aligned} z_{k+1} &= z_k - a_{k+1}\nabla f(x_k) \\ &= z_k + \frac{1}{\tau_k}(y_{k+1} - x_k) \\ &= y_k + \frac{1}{\tau_k}(x_k - y_k) + \frac{1}{\tau_k}(y_{k+1} - x_k) \\ &= y_k + \frac{1}{\tau_k}(y_{k+1} - y_k). \end{aligned}$$

Now,

$$\begin{aligned} x_{k+1} &= (1-\tau_{k+1})y_{k+1} + \tau_{k+1}z_{k+1} \\ &= y_{k+1} + \tau_{k+1}(z_{k+1} - y_{k+1}) \\ &= y_{k+1} + \tau_{k+1}\left(\frac{1}{\tau_k} - 1\right)(y_{k+1} - y_k). \end{aligned}$$

Therefore,

$$x_{k+1} = y_{k+1} + \beta_k(y_{k+1} - y_k), \qquad \beta_k = \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}. \tag{16}$$

4.2 The usual t_k-notation

Recall

$$t_k = \sqrt{LA_k}.$$

Since

$$a_{k+1} = \frac{t_{k+1}}{L}, \qquad A_{k+1} = \frac{t_{k+1}^2}{L},$$

we obtain

$$\tau_k = \frac{1}{t_{k+1}}.$$

Define a shifted sequence

$$s_k := t_{k+1}.$$

Then

$$s_0 = 1$$

and

$$s_{k+1} = \frac{1 + \sqrt{1 + 4s_k^2}}{2}.$$

Moreover,

$$\beta_k = \frac{s_k - 1}{s_{k+1}}.$$

We have therefore recovered the familiar implementation.

Nesterov accelerated gradient method: convex case
Choose $y_0 = x_0 \in \mathbb{R}^d$ and set $s_0 = 1$. For $k \geq 0$, compute

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k), \\ s_{k+1} &= \frac{1 + \sqrt{1 + 4s_k^2}}{2}, \\ x_{k+1} &= y_{k+1} + \frac{s_k - 1}{s_{k+1}}(y_{k+1} - y_k). \end{aligned}$$

Remark 4.1 (Why the momentum increases). The sequence s_k grows approximately like $k/2$. Hence

$$\beta_k = \frac{s_k - 1}{s_{k+1}} \longrightarrow 1.$$

Thus the momentum becomes stronger over time.

This does not mean that the method ignores the gradient. At every step the gradient is evaluated at the extrapolated point x_k, so the gradient continues to correct the direction of motion.

4.3 Relation to the explicit weights $A_k = k(k+1)/(4L)$

A simpler explicit choice is

$$A_k = \frac{k(k+1)}{4L}.$$

Then

$$a_{k+1} := A_{k+1} - A_k = \frac{k+1}{2L},$$

and

$$a_{k+1}^2 = \frac{(k+1)^2}{4L^2} \leq \frac{(k+1)(k+2)}{4L^2} = \frac{A_{k+1}}{L}.$$

This inequality is enough to prove

$$\mathcal{E}_{k+1} \leq \mathcal{E}_k$$

and therefore gives the same $O(k^{-2})$ rate.

For this explicit choice,

$$\tau_k = \frac{a_{k+1}}{A_{k+1}} = \frac{2}{k+2}.$$

The corresponding formal momentum parameter is

$$\beta_k = \frac{\tau_{k+1}(1-\tau_k)}{\tau_k} = \frac{k}{k+3}.$$

However, in this explicit-weight version,

$$a_{k+1}\tau_k = \frac{k+1}{L(k+2)} \neq \frac{1}{L}.$$

Therefore, the elimination of z_k does not give the standard two-sequence Nesterov method exactly.

This is why the equality choice

$$La_{k+1}^2 = A_{k+1}$$

is especially convenient pedagogically: it proves the accelerated rate and gives exact equivalence with the standard implementation.

5 Nesterov acceleration: strongly convex case

We now assume:

- $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex;
- f is L-smooth;
- $0 < \mu < L$.

The minimizer $x^\star$ is unique.

Define

$$\tau := \sqrt{\frac{\mu}{L}} \in (0, 1).$$

5.1 Three-sequence formulation

Consider

$$x_k = \frac{\tau}{1+\tau} z_k + \frac{1}{1+\tau} y_k, \tag{17}$$

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \tag{18}$$

$$z_{k+1} = z_k + \tau(x_k - z_k) - \frac{\tau}{\mu}\nabla f(x_k). \tag{19}$$

The interpolation identity (17) can be rewritten as

$$y_k - x_k = \tau(x_k - z_k). \tag{20}$$

This identity will be used repeatedly in the proof.

5.2 The strongly convex Lyapunov function

Define

$$\mathcal{E}_k := f(y_k) - f^\star + \frac{\mu}{2}\|z_k - x^\star\|^2. \tag{21}$$

Unlike the convex case, there is no growing weight A_k. Instead, we will prove a direct contraction:

$$\mathcal{E}_{k+1} \le (1-\tau)\mathcal{E}_k.$$

Iterating this inequality yields

$$\mathcal{E}_k \le (1-\tau)^k \mathcal{E}_0.$$

The convergence rate therefore comes from the contraction factor rather than from a growing coefficient.

5.3 A clean contraction proof

Theorem 5.1 (Strongly convex accelerated convergence)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth, with $0 < \mu < L$, and let*

$$\tau = \sqrt{\frac{\mu}{L}}.$$

Suppose (x_k, y_k, z_k) is generated by (17)–(19). Then

$$\mathcal{E}_{k+1} \le (1-\tau)\mathcal{E}_k,$$

where

$$\mathcal{E}_k = f(y_k) - f^\star + \frac{\mu}{2}\|z_k - x^\star\|^2.$$

Consequently,

$$\mathcal{E}_k \le \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathcal{E}_0.$$

Proof. Fix k and introduce the shorter notation

$$x = x_k, \qquad y = y_k, \qquad z = z_k, \qquad g = \nabla f(x),$$

and

$$y^+ = y_{k+1}, \qquad z^+ = z_{k+1}.$$

Define

$$h := y - x.$$

By (20),

$$h = \tau(x - z). \tag{22}$$

The update of z becomes

$$\begin{aligned} z^+ - z &= \tau(x - z) - \frac{\tau}{\mu} g \\ &= h - \frac{\tau}{\mu} g. \end{aligned} \tag{23}$$

We estimate

$$\mathcal{E}_{k+1} - (1-\tau)\mathcal{E}_k.$$

Step 1: use smoothness for the gradient step. Since

$$y^+ = x - \frac{1}{L} g,$$

smoothness gives

$$f(y^+) \le f(x) - \frac{1}{2L} \|g\|^2 . \tag{24}$$

Therefore,

$$\begin{aligned} \mathcal{E}_{k+1} - (1-\tau)\mathcal{E}_k \le f(x) - f^\star - (1-\tau)(f(y) - f^\star) - \frac{1}{2L}\|g\|^2 \\ + \frac{\mu}{2}\left(\|z^+ - x^\star\|^2 - (1-\tau)\|z - x^\star\|^2\right). \end{aligned} \tag{25}$$

Step 2: estimate the function-value expression. We rewrite

$$f(x) - f^\star - (1-\tau)(f(y) - f^\star) = \tau(f(x) - f^\star) + (1-\tau)(f(x) - f(y)).$$

Strong convexity, applied with $x^\star$, gives

$$f(x) - f^\star \le \langle g, x - x^\star \rangle - \frac{\mu}{2}\|x - x^\star\|^2 .$$

Strong convexity, applied with y, gives

$$f(x) - f(y) \le -\langle g, y - x \rangle - \frac{\mu}{2}\|y - x\|^2 .$$

Consequently,

$$\begin{aligned} & f(x) - f^\star - (1-\tau)(f(y) - f^\star) \\ & \quad \le \tau \langle g, x - x^\star \rangle - \frac{\tau\mu}{2}\|x - x^\star\|^2 \\ & \qquad - (1-\tau)\langle g, h \rangle - \frac{(1-\tau)\mu}{2}\|h\|^2 . \end{aligned} \tag{26}$$

Step 3: expand the distance expression. Let

$$r := x - x^\star.$$

From $h = \tau(x - z)$, we have

$$z - x^\star = x - x^\star - \frac{h}{\tau} = r - \frac{h}{\tau}.$$

Also, by (23),

$$z^+ - z = h - \frac{\tau}{\mu} g.$$

Write

$$d := z^+ - z = h - \frac{\tau}{\mu} g.$$

Then

$$\begin{aligned} & \|z^+ - x^\star\|^2 - (1-\tau)\|z - x^\star\|^2 \\ & \quad = \tau\|z - x^\star\|^2 + 2\langle d, z - x^\star \rangle + \|d\|^2 . \end{aligned} \tag{27}$$

Substitute

$$z - x^\star = r - \frac{h}{\tau}, \qquad d = h - \frac{\tau}{\mu} g.$$

A direct expansion yields

$$\begin{aligned}
&\tau \|z - x^\star\|^2 + 2\langle d, z - x^\star\rangle + \|d\|^2 \\
&\quad = \tau \|r\|^2 + \left(1 - \frac{1}{\tau}\right)\|h\|^2 - \frac{2\tau}{\mu}\langle g, r\rangle \\
&\qquad + \frac{2(1-\tau)}{\mu}\langle g, h\rangle + \frac{\tau^2}{\mu^2}\|g\|^2 .
\end{aligned} \tag{28}$$

Multiplying by $\mu/2$, we obtain

$$\begin{aligned}
&\frac{\mu}{2}\left(\left\|z^+ - x^\star\right\|^2 - (1-\tau)\|z - x^\star\|^2\right) \\
&\quad = \frac{\tau\mu}{2}\|r\|^2 + \frac{\mu}{2}\left(1 - \frac{1}{\tau}\right)\|h\|^2 - \tau\langle g, r\rangle \\
&\qquad + (1-\tau)\langle g, h\rangle + \frac{\tau^2}{2\mu}\|g\|^2 .
\end{aligned} \tag{29}$$

Step 4: combine the estimates. Insert (26) and (29) into (25).

The following terms cancel:

$$\tau\langle g, r\rangle \quad \text{with} \quad -\tau\langle g, r\rangle ,$$

$$-\frac{\tau\mu}{2}\|r\|^2 \quad \text{with} \quad \frac{\tau\mu}{2}\|r\|^2 ,$$

and

$$-(1-\tau)\langle g, h\rangle \quad \text{with} \quad (1-\tau)\langle g, h\rangle .$$

We are left with

$$\begin{aligned}
\mathcal{E}_{k+1} - (1-\tau)\mathcal{E}_k \le \frac{\mu}{2}\left(\tau - \frac{1}{\tau}\right)\|h\|^2 \\
+ \frac{1}{2}\left(\frac{\tau^2}{\mu} - \frac{1}{L}\right)\|g\|^2 .
\end{aligned} \tag{30}$$

Because

$$\tau = \sqrt{\frac{\mu}{L}},$$

we have

$$\frac{\tau^2}{\mu} = \frac{1}{L}.$$

Thus the squared-gradient term vanishes.

Moreover, $0 < \tau < 1$, so

$$\tau - \frac{1}{\tau} < 0.$$

It follows that

$$\mathcal{E}_{k+1} - (1-\tau)\mathcal{E}_k \le 0.$$

Therefore,

$$\mathcal{E}_{k+1} \le (1-\tau)\mathcal{E}_k.$$

Iterating gives

$$\mathcal{E}_k \le (1-\tau)^k \mathcal{E}_0 = \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathcal{E}_0.$$

□

Teaching point 5.2. The parameter

$$\tau = \sqrt{\frac{\mu}{L}}$$

does not appear by magic. It is chosen to cancel the coefficient

$$\frac{\tau^2}{\mu} - \frac{1}{L}.$$

This is a typical Lyapunov-design calculation:

compute the energy increment first, then choose parameters to make the remaining terms nonpositive.

6 Standard two-sequence form in the strongly convex case

We now eliminate z_k from the strongly convex three-sequence system.

From

$$x_k = \frac{\tau}{1+\tau} z_k + \frac{1}{1+\tau} y_k,$$

we obtain

$$z_k = \frac{1+\tau}{\tau} x_k - \frac{1}{\tau} y_k.$$

Recall

$$z_{k+1} = z_k + \tau(x_k - z_k) - \frac{\tau}{\mu} \nabla f(x_k).$$

Since

$$y_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$$

and

$$\frac{\tau^2}{\mu} = \frac{1}{L},$$

one can simplify the z-update to

$$z_{k+1} = y_k + \frac{1}{\tau}(y_{k+1} - y_k).$$

Hence

$$\begin{aligned} x_{k+1} &= \frac{\tau}{1+\tau} z_{k+1} + \frac{1}{1+\tau} y_{k+1} \\ &= y_{k+1} + \frac{1-\tau}{1+\tau}(y_{k+1} - y_k). \end{aligned}$$

Therefore the standard method is

Nesterov accelerated gradient method: strongly convex case

Set

$$\tau = \sqrt{\frac{\mu}{L}}, \qquad \beta = \frac{1-\tau}{1+\tau}.$$

Choose $y_0 = x_0$. For $k \geq 0$, compute

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L} \nabla f(x_k), \\ x_{k+1} &= y_{k+1} + \beta(y_{k+1} - y_k). \end{aligned}$$

Since

$$\tau = \sqrt{\frac{\mu}{L}},$$

the momentum parameter is

$$\begin{aligned}\beta &= \frac{1 - \sqrt{\mu/L}}{1 + \sqrt{\mu/L}} \\ &= \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}.\end{aligned}$$

In terms of the condition number $\kappa = L/\mu$,

$$\tau = \frac{1}{\sqrt{\kappa}}, \qquad \beta = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}.$$

7 Complexity comparison

The strongly convex Lyapunov estimate gives

$$\mathcal{E}_k \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \mathcal{E}_0.$$

Using

$$1 - s \leq e^{-s}, \qquad s \geq 0,$$

we obtain

$$\mathcal{E}_k \leq \exp\left(-\frac{k}{\sqrt{\kappa}}\right) \mathcal{E}_0.$$

To guarantee

$$\mathcal{E}_k \leq \varepsilon,$$

it is sufficient that

$$\exp\left(-\frac{k}{\sqrt{\kappa}}\right) \mathcal{E}_0 \leq \varepsilon.$$

Equivalently,

$$k \geq \sqrt{\kappa} \log\left(\frac{\mathcal{E}_0}{\varepsilon}\right).$$

Thus Nesterov acceleration has complexity

$$O\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right).$$

By comparison, gradient descent with step size $1/L$ satisfies a rate involving

$$1 - \frac{\mu}{L} = 1 - \frac{1}{\kappa},$$

and therefore has complexity

$$O\left(\kappa \log \frac{1}{\varepsilon}\right).$$

The improvement is

$$\boxed{\kappa \quad \longrightarrow \quad \sqrt{\kappa}.}$$

For an ill-conditioned problem with, for example,

$$\kappa = 10^4,$$

this corresponds heuristically to replacing a factor of 10^4 by a factor of 10^2.

8 How to explain the Lyapunov strategy

A useful classroom explanation is the following.

Step 1: the objective value need not decrease

For accelerated methods, neither the iterates nor the objective values are necessarily monotone:

$$f(y_{k+1}) \leq f(y_k)$$

need not hold.

This does not mean the method is unstable. It means that the objective value alone is not the right quantity to monitor.

Step 2: search for a combined energy

We combine the optimization error with a geometric or momentum term:

$$\mathcal{E}_k = \text{objective error} + \text{distance/momentum energy}.$$

The distance term absorbs oscillations that may appear in the objective values.

Step 3: compute one step of the energy

The essential calculation is always

$$\mathcal{E}_{k+1} - \mathcal{E}_k$$

or, in the strongly convex case,

$$\mathcal{E}_{k+1} - (1-\tau)\mathcal{E}_k.$$

The proof uses:

- smoothness to control the gradient step;
- convexity or strong convexity to control first-order terms;
- the interpolation identities to cancel cross terms;
- parameter choices to cancel the remaining squared norms.

Step 4: read off the rate

In the convex case,

$$\mathcal{E}_k = \frac{1}{2}\left\|z_k - x^\star\right\|^2 + A_k(f(y_k) - f^\star),$$

and

$$\mathcal{E}_k \leq \mathcal{E}_0.$$

Therefore,

$$f(y_k) - f^\star \leq \frac{\mathcal{E}_0}{A_k}.$$

Since $A_k \asymp k^2/L$, the rate is $O(L/k^2)$.

In the strongly convex case,

$$\mathcal{E}_{k+1} \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)\mathcal{E}_k.$$

Therefore,

$$\mathcal{E}_k \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathcal{E}_0.$$

9 Common sources of confusion

1. **Is the continuous-time Lyapunov theorem formally used?**

   No. It provides the conceptual template. The actual discrete convergence proof directly verifies a discrete energy inequality.

2. **Why do we introduce z_k?**

   Because the update of
   $$\frac{1}{2}\|z_k - x^\star\|^2$$
   produces the inner product
   $$-a_{k+1}\langle \nabla f(x_k), z_k - x^\star\rangle .$$
   This is exactly the type of term needed to cancel the convexity bounds for the objective error.

3. **Does the objective decrease at every iteration?**

   Not necessarily. The Lyapunov energy decreases, but the individual terms in the energy may oscillate.

4. **Where does the k^2 rate come from?**

   It comes from the growth of A_k, not directly from a one-step decrease of $f(y_k)$. The proof gives
   $$f(y_k) - f^\star \le \frac{\mathcal{E}_0}{A_k},$$
   and the weight recurrence gives $A_k \asymp k^2/L$.

5. **Where does the square root $\sqrt{\mu/L}$ come from?**

   It is the parameter that cancels the squared-gradient coefficient in the Lyapunov increment:
   $$\frac{\tau^2}{\mu} - \frac{1}{L} = 0.$$

6. **Is Nesterov momentum the same as heavy-ball momentum?**

   No. Heavy ball evaluates the gradient at the current point and then adds momentum. Nesterov first forms an extrapolated point and evaluates the gradient there.

7. **Why are there several equivalent versions of Nesterov's method?**

   Different index conventions and variable names lead to formulas that look different. The two common representations are:

   $$\text{two-sequence implementation} \quad \leftrightarrow \quad \text{three-sequence Lyapunov representation.}$$

   The three-sequence representation is usually easier to analyze. The two-sequence representation is usually easier to implement.

References

[1] S. Weissmann, *Introduction to Stochastic Gradient Methods*, lecture notes, arXiv:2606.03953. Available at: https://arxiv.org/abs/2606.03953.