Mathematics of Data Science
UU Math 5750/6880

Numerical Methods for Optimization: Momentum Descent[1]

Braxton Osting

Fall 2025

[1] last updated: October 28, 2025

Gradient descent Convergence

Consider the *unconstrained optimization problem*,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}).$$

Gradient descent. $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x}_k)$

Theorem. Let $f \in C_L^{1,1}(\mathbb{R}^n)$ and assume that f is bounded below on $\mathbb{R}^n$, i.e., there exists $m \in \mathbb{R}$ such that $f(\boldsymbol{x}) > m$, $\forall \boldsymbol{x} \in \mathbb{R}^n$. Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient descent method with stepsize $\alpha \in \left(0, \frac{2}{L}\right)$. Then

(a) For any k,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2.$$

In particular, $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ unless $\nabla f(\boldsymbol{x}_k) = \mathbf{0}$.

(b) $\nabla f(\boldsymbol{x}_k) \to \mathbf{0}$ as $k \to \infty$.

(c) Let $f^\star = \lim_{k\to\infty} f(\boldsymbol{x}_k)$. Then

$$\min_{k=0,\ldots,t} \|\nabla f(\boldsymbol{x}_k)\|^2 \leq 2L\frac{f(\boldsymbol{x}_0) - f^\star}{t+1}.$$

Theorem. Assume $f \in C_L^{1,1}(\mathbb{R}^n)$ is convex and admits a minimizer. Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient descent method with stepsize $\alpha \in \left(0, \frac{2}{L}\right)$. Then for any $t \in \mathbb{N}$,

$$f(\boldsymbol{x}_t) - f^\star \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}{2\alpha t}$$

where $\boldsymbol{x}^\star$ is any minimizer of f.

Theorem. Assume $f \in C_L^{1,1}(\mathbb{R}^n)$ satisfies the (PŁ) inequality. Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient descent method with stepsize $\alpha \in \left(0, \frac{1}{L}\right)$. Then for any $t \in \mathbb{N}$,

$$f(\boldsymbol{x}_t) - f^\star \leq (1-\alpha\mu)^t\left(f(\boldsymbol{x}_0) - f^\star\right).$$

Assumptions

Definition (L-smoothness)

Let $L \geq 0$. A function $f \in C^1(\mathbb{R}^n)$ is said to be *L-smooth*, written $f \in C_L^{1,1}(\mathbb{R}^n)$, if it satisfies

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

The constant L is called the *smoothness parameter*.

Definition (μ-strong convexity)

A function $f \in C^1(\mathbb{R}^n)$ is μ-strongly convex if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Definition (PŁinequality)

We say that $f \in C^1(\mathbb{R}^n)$ satisfies the Polyak-Łojasiewicz (PŁ) inequality if

$$\frac{1}{2}\|\nabla f(\boldsymbol{x})\|^2 \geq \mu\left(f(\boldsymbol{x}) - f^\star\right), \qquad \boldsymbol{x} \in \mathbb{R}^n. \tag{PŁ}$$

Overview of accelerated methods

In the μ-strongly convex setting (or, more generally, under the (PL) inequality) gradient descent converges linearly. This is still slow! Since the gradient ∇f vanishes at the minimizer, it is small nearby, and thus the gradient descent iterations change very little.

There are (many!) ideas to accelerate the convergence of gradient descent.

Heavy ball method. Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$ and set $\boldsymbol{x}_1 = \boldsymbol{x}_0$,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k) + \beta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}), \qquad k \in \mathbb{N},$$

where α is the step length and $\beta \in [0,1]$ is the momentum parameter.

Nesterov Accelerated Descent. Choose $\boldsymbol{x}_1 \in \mathbb{R}^n$ and set $\boldsymbol{y}_1 = \boldsymbol{x}_1$. Compute the numbers $\lambda_k = \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2}$, $k \in \mathbb{N}$ where $\lambda_0 = 0$.

$$\begin{aligned} \boldsymbol{y}_{k+1} &= \boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k) \\ \boldsymbol{x}_{k+1} &= \boldsymbol{y}_{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}} (\boldsymbol{y}_{k+1} - \boldsymbol{y}_k) \end{aligned}$$

$x_{n+1} = x_n - \alpha \nabla f(x_n)$

Stochastic Gradient Descent (SGD). In ML, we typically optimize a function of the form

$$f(\boldsymbol{x}) = \frac{1}{n} \sum_{i \in [n]} f_i(\boldsymbol{x}).$$

Idea: approximate $\nabla f(\boldsymbol{x}) \approx \frac{1}{|I|} \sum_{i \in I} \nabla f_i(\boldsymbol{x})$ where $I \subseteq [n]$.

Adaptive methods. Idea: scales the learning rate for each parameter based on historical gradients. Many variants: Root Mean Square Propagation (RMSprop), Adaptive Gradient (ADAgrad), Adam (Adaptive Moment Estimation)

Newton's Method. Use second order derivative (Hessian) information:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - [\nabla^2 f(\boldsymbol{x}_n)]^{-1} \nabla f(\boldsymbol{x}_n)$$

quasi-Newton, Gauss-Newton variants

Table of Contents

- Momentum descent / Heavy ball method
- Continuum-time analysis and momentum flow
- Convergence analysis for momentum descent

Momentum descent / Heavy ball method

Momentum methods are loosely based on the idea of rolling a ball with some positive mass down the energy landscape, in the presence of friction forces to slow down the ball.

Momentum can build up speed over time, provided the descent directions are similar over many steps, leading to faster convergence near the minimizer.

When descent directions change rapidly over each step (like in the zig-zag effect for a quadratic form with large condition number), momentum acts to average out the descent directions over time and reduces the amount of zig and zag, leading to faster convergence.

There are several variations on this idea. We'll discuss the heavy ball method of Polyak.

Heavy ball method. Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$ and set $\boldsymbol{x}_1 = \boldsymbol{x}_0$,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k) + \beta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}), \qquad k \in \mathbb{N},$$

where α is the step length and $\beta \in [0,1]$ is the momentum parameter.

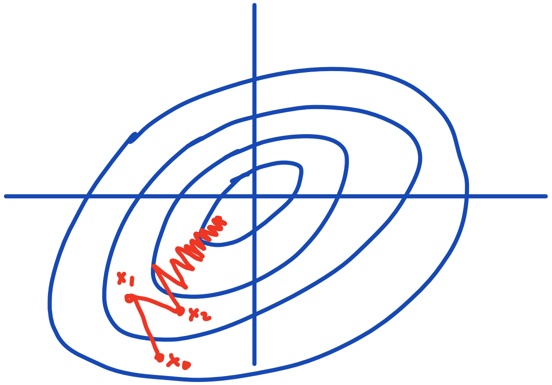

Table of Contents

- Momentum descent / Heavy ball method
- Continuum-time analysis and momentum flow
- Convergence analysis for momentum descent

Continuum-time limit and gradient descent

To understand accelerated methods, it is useful to take a *continuum-time* limit and obtain an ODE.

First consider gradient descent using a continuum-time limit.

$$x(t) = x(0) + (t-0)\,\dot{x}(0) + \cdots$$

We rewrite the gradient descent iteration as

$$\frac{\boldsymbol{x}_{k+1} - \boldsymbol{x}_k}{\alpha} = -\nabla f(\boldsymbol{x}_k).$$

$$x(\alpha k) = x(\alpha(k-1)) + \alpha\,\dot{x}(\alpha(k-1))$$

$$x_k = x_{k-1} + \alpha\,\dot{x}_{k-1}$$

By assuming that $\boldsymbol{x}_k = \boldsymbol{x}(\alpha k)$ for a smooth curve $\boldsymbol{x}(t)$, we find that the LHS is merely a forward differences approximation for $\dot{\boldsymbol{x}}(\alpha k) \approx \frac{\boldsymbol{x}(\alpha(k+1)) - \boldsymbol{x}(\alpha k)}{\alpha}$. Thus, gradient descent corresponds to the *gradient flow ODE*

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= -\nabla f(\boldsymbol{x}(t)), \qquad\qquad t \geq 0 \qquad\qquad \text{(gf)} \\ \boldsymbol{x}(0) &= \boldsymbol{x}_0. \end{aligned}$$

Convergence analysis. Suppose f is μ-strongly convex and we consider (gf) with initial condition $\boldsymbol{x}(0) = \boldsymbol{x}_0$. Let $\boldsymbol{x}^\star \in \mathbb{R}^n$ denote the unique minimizer of f.

Note that (gf) can be written

$$\frac{d}{dt}\left(\boldsymbol{x}(t) - \boldsymbol{x}^\star\right) = -\left(\nabla f(\boldsymbol{x}(t) - \nabla f(\boldsymbol{x}^\star)\right).$$

Take the inner product of both sides with $\boldsymbol{x}(t) - \boldsymbol{x}^\star$ to obtain

$$\frac{1}{2}\frac{d}{dt}\|\boldsymbol{x}(t) - \boldsymbol{x}^\star\|^2 = -\left(\nabla f(\boldsymbol{x}(t) - \nabla f(\boldsymbol{x}^\star)\right)^\top \left(\boldsymbol{x}(t) - \boldsymbol{x}^\star\right) \leq -\mu\|\boldsymbol{x}(t) - \boldsymbol{x}^\star\|^2,$$

where we used the μ-strong convexity of f. Using Grönwall's inequality, we have that

$$\|\boldsymbol{x}(t) - \boldsymbol{x}^\star\|^2 \leq e^{-2\mu t}\|\boldsymbol{x}(\overset{0}{\cancel{t}}) - \boldsymbol{x}^\star\|^2,$$

which is the continuous-time equivalent to the linear convergence rate for (discrete) gradient descent.

Suppose $e(t) \geq 0$ satisfies

$$\dot{e}(t) \leq a\, e(t)$$

for $a \in \mathbb{R}$. Then $e(t) \leq e(0)\, e^{at}$

pf. $\frac{d}{dt} \log e(t) = \frac{\dot{e}}{e} \leq a.$

$$\int_0^t \frac{d}{ds} \log e(s)\, ds \leq \int_0^t a\, ds = at$$

$$\shortparallel$$

$$\log e(t) - \log e(0)$$

$$\boxed{e(t) \leq e(0)\, e^{at}}$$

Continuum-time limit and momentum descent

We can rearrange the heavy ball method iteration to obtain

$$\frac{\boldsymbol{x}_{k+1} - 2\boldsymbol{x}_k + \boldsymbol{x}_{k-1}}{\boxed{\alpha}} + \frac{1-\beta}{\alpha}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) = -\nabla f(\boldsymbol{x}_k).$$

For a smooth curve $\boldsymbol{x}(t)$, we use the approximation

$$(x_{k+1} - x_k) - (x_k - x_{k-1})$$

$$\frac{\boldsymbol{x}(h(k+1)) - 2\boldsymbol{x}(hk) + \boldsymbol{x}(h(k-1))}{\boxed{h^2}} = \ddot{\boldsymbol{x}}(hk) + O(h^2).$$

Interpreting $\alpha = h^2$ for a time step h, we have $\boldsymbol{x}_k = \boldsymbol{x}(hk)$ and

$$\frac{\boldsymbol{x}_{k+1} - 2\boldsymbol{x}_k + \boldsymbol{x}_{k-1}}{\alpha} \approx \ddot{\boldsymbol{x}}(\sqrt{\alpha}k).$$

Additionally,

$$\frac{1-\beta}{\alpha}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) = \frac{1-\beta}{\sqrt{\alpha}} \frac{\left(\boldsymbol{x}(\sqrt{\alpha}k) - \boldsymbol{x}(\sqrt{\alpha}(k-1))\right)}{\sqrt{\alpha}} \approx a\dot{\boldsymbol{x}}(\sqrt{\alpha}k),$$

where $a = \frac{1-\beta}{\sqrt{\alpha}}$. Thus, the heavy ball method iteration has the continuum limit

$$\begin{aligned} &\ddot{\boldsymbol{x}}(t) + a\dot{\boldsymbol{x}}(t) = -\nabla f(\boldsymbol{x}(t)), \qquad t \geq 0 \\ &\boldsymbol{x}(0) = \boldsymbol{x}_0 \\ &\dot{\boldsymbol{x}}(0) = \mathbf{0}. \end{aligned} \tag{mf}$$

This ODE corresponds to Newton's law of motion for a body under the forcing of $-\nabla f$ and friction coefficient a. We can see that $\beta \in [0,1]$ is required for positivity of the friction coefficient, which ensures the system will dissipate energy and slow down over time. To keep the amount of friction fixed, we fix a and choose $\beta = 1 - a\sqrt{\alpha}$.

We refer to continuum-time momentum descent as *momentum flow*.

Convergence analysis for Momentum flow

Momentum flow.

$$\begin{aligned} &\ddot{\boldsymbol{x}}(t) + a\dot{\boldsymbol{x}}(t) = -\nabla f(\boldsymbol{x}(t)), \qquad && t \geq 0 \\ &\boldsymbol{x}(0) = \boldsymbol{x}_0 && \\ &\dot{\boldsymbol{x}}(0) = \mathbf{0}. && \end{aligned} \tag{mf}$$

Theorem

Assume f is L-smooth and μ-strongly convex and that $\boldsymbol{x}(t)$ satisfies (mf). Let $\boldsymbol{x}^\star$ denote the unique minimizer of f. Then we have

$$\|\boldsymbol{x}(t) - \boldsymbol{x}^\star\|^2 \leq \frac{3L + 2a^2}{3\mu} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 \exp\left(-\frac{2\mu a t}{3L + 2a^2}\right).$$

Remark. (gf) and (mf) converge at the exponential rate e^{-ct}, with different constants $c > 0$. It isn't clear from this analysis which converges more quickly. The difference in the two methods only appears upon discretizing the methods with the backward Euler scheme. The heavy ball method involves a second derivative in time, which allows for a larger time step, α, and gives a faster convergence to the minimizer.

Proof

Let $\boldsymbol{y}(t) = \boldsymbol{x}(t) - \boldsymbol{x}^\star$ and note that $\boldsymbol{y}(t)$ satisfies

$$\ddot{\boldsymbol{y}}(t) + a\dot{\boldsymbol{y}}(t) = -\nabla f(\boldsymbol{x}(t)).$$

$f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|$

Define the energy

$$e(t) = \tfrac{3}{2}\|\dot{\boldsymbol{y}}(t)\|^2 + 3\left(f(\boldsymbol{x}(t)) - f^\star\right) + \tfrac{a^2}{2}\|\boldsymbol{y}(t)\|^2 + a\boldsymbol{y}(t)^\top\dot{\boldsymbol{y}}(t),$$

where $f^\star = f(\boldsymbol{x}^\star)$. Using the μ-strong convexity of f, we have

$$\begin{aligned} e(t) &\geq \tfrac{3}{2}\|\dot{\boldsymbol{y}}(t)\|^2 + \tfrac{3\mu}{2}\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2 + \tfrac{a^2}{2}\|\boldsymbol{y}(t)\|^2 + a\,\boldsymbol{y}(t)^\top\dot{\boldsymbol{y}}(t) \\ &= \tfrac{3}{2}\|\dot{\boldsymbol{y}}(t)\|^2 + \tfrac{3\mu}{2}\|\boldsymbol{y}(t)\|^2 + \tfrac{1}{2}\left(\|a\boldsymbol{y}(t) + \dot{\boldsymbol{y}}(t)\|^2 - \|\dot{\boldsymbol{y}}(t)\|^2\right) \\ &= \|\dot{\boldsymbol{y}}(t)\|^2 + \tfrac{3\mu}{2}\|\boldsymbol{y}(t)\|^2 + \tfrac{1}{2}\|a\boldsymbol{y}(t) + \dot{\boldsymbol{y}}(t)\|^2 \\ &\geq \tfrac{3\mu}{2}\|\boldsymbol{y}(t)\|^2. \end{aligned}$$

Therefore $e(t) \geq 0$ and, in particular,

$$\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2 = \|\boldsymbol{y}(t)\|^2 \leq \frac{2}{3\mu}e(t).$$

The rest of the proof will focus on showing that $e(t)$ decays to zero exponentially fast.

We differentiate $e(t)$ and use (mf) and the identities $\dot{\boldsymbol{x}} = \dot{\boldsymbol{y}}$ and $\ddot{\boldsymbol{x}} = \ddot{\boldsymbol{y}}$ to compute

$$\begin{aligned} \dot{e}(t) &= 3\dot{\boldsymbol{y}}(t)^\top\ddot{\boldsymbol{y}}(t) + 3\nabla f(\boldsymbol{x}(t))^\top\dot{\boldsymbol{x}}(t) + a^2\dot{\boldsymbol{y}}(t)^\top\cancel{\dot{\boldsymbol{y}}}(t) + a\|\dot{\boldsymbol{y}}(t)\|^2 + a\boldsymbol{y}(t)^\top\ddot{\boldsymbol{y}}(t) \\ &= 3\dot{\boldsymbol{y}}(t)^T\left(\ddot{\boldsymbol{y}}(t) + \nabla f(\boldsymbol{x}(t))\right) + a\boldsymbol{y}(t)^T\left(\ddot{\boldsymbol{y}}(t) + a\dot{\boldsymbol{y}}(t)\right) + a\|\dot{\boldsymbol{y}}(t)\|^2 \\ &= -3a\|\dot{\boldsymbol{y}}(t)\|^2 - a\nabla f(\boldsymbol{x}(t))^T\boldsymbol{y}(t) + a\|\dot{\boldsymbol{y}}(t)\|^2 \\ &= -2a\|\dot{\boldsymbol{y}}(t)\|^2 - a\left(\nabla f(\boldsymbol{x}(t)) - \nabla f(\boldsymbol{x}^*)\right)^T\left(\boldsymbol{x}(t) - \boldsymbol{x}^*\right) \\ &\leq -2a\|\dot{\boldsymbol{y}}(t)\|^2 - a\mu\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2 \\ &= -a\left(\mu\|\boldsymbol{y}(t)\|^2 + 2\|\dot{\boldsymbol{y}}(t)\|^2\right), \end{aligned}$$

where we have used strong convexity of f (and $\nabla f(\boldsymbol{x}^*) = 0$).

Proof II

By the Cauchy–Schwarz inequality and $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ we have

$$a\mathbf{y}(t)^T\dot{\mathbf{y}}(t) \leq \|a\mathbf{y}(t)\|\|\dot{\mathbf{y}}(t)\| \leq \tfrac{a^2}{2}\|\mathbf{y}(t)\|^2 + \tfrac{1}{2}\|\dot{\mathbf{y}}(t)\|^2.$$

Since f is L-smooth, we have $f(\mathbf{x}(t)) - f^* \leq \frac{L}{2}\|\mathbf{x}(t) - \mathbf{x}^*\|^2 = \frac{L}{2}\|\mathbf{y}(t)\|^2$, and so

$$\begin{aligned}
e(t) &= \tfrac{3}{2}\|\dot{\mathbf{y}}(t)\|^2 + 3\left(f(\mathbf{x}(t)) - f^\star\right) + \tfrac{a^2}{2}\|\mathbf{y}(t)\|^2 + a\mathbf{y}(t)^\top\dot{\mathbf{y}}(t) \\
&\leq \tfrac{3}{2}\|\dot{\mathbf{y}}(t)\|^2 + \tfrac{3L}{2}\|\mathbf{y}(t)\|^2 + \tfrac{a^2}{2}\|\mathbf{y}(t)\|^2 + \tfrac{a^2}{2}\|\mathbf{y}(t)\|^2 + \tfrac{1}{2}\|\dot{\mathbf{y}}(t)\|^2 \\
&= \left(\tfrac{3L}{2} + a^2\right)\|\mathbf{y}(t)\|^2 + 2\|\dot{\mathbf{y}}(t)\|^2 \\
&= \left(\tfrac{3L+2a^2}{2\mu}\right)\mu\|\mathbf{y}(t)\|^2 + 2\|\dot{\mathbf{y}}(t)\|^2 \\
&\leq \left(\tfrac{3L+2a^2}{2\mu}\right)\left(\mu\|\mathbf{y}(t)\|^2 + 2\|\dot{\mathbf{y}}(t)\|^2\right),
\end{aligned}$$

where the last inequality follows from $L \geq \mu$ and so $\frac{3L+2a^2}{2\mu} \geq \frac{3L}{2\mu} \geq \frac{3}{2} \geq 1$.

Therefore we have

$$\mu\|\mathbf{y}(t)\|^2 + 2\|\dot{\mathbf{y}}(t)\|^2 \geq \frac{2\mu}{3L + 2a^2}\, e(t).$$

Substituting this into the upper bound on $\dot{e}(t)$ from above, we obtain

$$\dot{e}(t) \leq -\frac{2\mu a}{3L + 2a^2}\, e(t).$$

Since $e(t) \geq 0$, it follows from Grönwall's inequality that

$$e(t) \leq e(0)\exp\!\left(-\frac{2\mu a t}{3L + 2a^2}\right).$$

Proof III

Combining this with $\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2 \leq \dfrac{2}{3\mu}\, e(t)$ (derived above), yields

$$\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2 \leq \frac{2}{3\mu}\, e(0)\, \exp\Big(-\frac{2\mu a t}{3L + 2a^2}\Big).$$

Finally, we use $\dot{\boldsymbol{y}}(0) = 0$ to estimate

$$\begin{aligned} e(0) &\leq \big(\tfrac{3L+2a^2}{2\mu}\big)\big(\mu\|\boldsymbol{y}(0)\|^2 + 2\|\dot{\boldsymbol{y}}(0)\|^2\big) \\ &= \tfrac{1}{2}\big(3L + 2a^2\big)\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \end{aligned}$$

We now combine these inequalities to obtain

$$\|\boldsymbol{x}(t) - \boldsymbol{x}^*\|^2 \leq \frac{3L + 2a^2}{3\mu}\, \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2\, \exp\Big(-\frac{2\mu a t}{3L + 2a^2}\Big),$$

as desired. □

Table of Contents

- Momentum descent / Heavy ball method

- Continuum-time analysis and momentum flow

- Convergence analysis for momentum descent

Minimizing a quadratic form

The analysis of momentum descent is more challenging.

To simplify the setting, we will consider the special case of minimizing the quadratic form

$$f(\boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x} - \boldsymbol{x}^\top \boldsymbol{b}.$$

where $A \succ 0$.

Let A have spectral decomposition $A = V\Lambda V^t$, where $V = [\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_n]$ and $\Lambda = \mathrm{diag}(\lambda)$, with $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$.

We know that

- $\nabla f(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$
- $\nabla^2 f(\boldsymbol{x}) = A$
- f is μ-strongly convex for $\mu = \lambda_{\min} = \lambda_1$.
- f is L-smooth for $L = \lambda_{\max} = \lambda_n$
- The minimizer of f is given by

$$\begin{aligned} \boldsymbol{x}^\star &= A^{-1}\boldsymbol{b} \\ &= \xcancel{V\Lambda^{-1}V^t b} \\ &= \xcancel{\sum_{i\in[n]} \lambda_i^{-1}(\boldsymbol{v}_i^\top \boldsymbol{b})\boldsymbol{v}_i} \end{aligned}$$

Minimizing a quadratic form: convergence analysis for gradient descent

Theorem

Suppose $\boldsymbol{x}_k$ satisfies

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha(A\boldsymbol{x}_k - \boldsymbol{b}), \qquad k \geq 1$$

with $\alpha \in \left(0, \frac{1}{L}\right)$. Then we have

$$(1-\alpha L)^k \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|}{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|} \leq (1-\alpha\mu)^k$$

$\mu < L$

$0 < \alpha\mu < 1$

Proof. We compute

$$\begin{aligned}
\boldsymbol{x}_k &= (I-\alpha A)\boldsymbol{x}_{k-1} + \alpha\boldsymbol{b} \\
&= (I-\alpha A)^k \boldsymbol{x}_0 + \alpha\left[\sum_{j=0}^{n-1}(I-\alpha A)^j\right]\boldsymbol{b} \\
&= (I-\alpha A)^k \boldsymbol{x}_0 + \left[I - (I-\alpha A)^k\right] A^{-1}\boldsymbol{b} \\
&= (I-\alpha A)^k \left(\boldsymbol{x}_0 - A^{-1}\boldsymbol{b}\right) + A^{-1}\boldsymbol{b},
\end{aligned}$$

$$\sum_{j=0}^{k-1} r^j = \frac{1-r^k}{1-r}$$

where we have used the equation recursively and summed the finite geometric series

$$\sum_{j=0}^{k-1} B^j = (I-B^k)(I-B)^{-1}.$$

Thus,

$$\boldsymbol{x}_k - \boldsymbol{x}^\star = (I-\alpha A)^k(\boldsymbol{x}_0 - \boldsymbol{x}^\star)$$

Since $(1-\alpha L)^k I \preceq (I-\alpha A)^k \preceq (1-\alpha\mu)^k I$, we have that

$$(1-\alpha L)^k\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\| \leq \|\boldsymbol{x}_k - \boldsymbol{x}^\star\| \leq (1-\alpha\mu)^k\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|,$$

as desired.

Minimizing a quadratic form: convergence analysis for momentum descent

Theorem

Suppose $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\boldsymbol{x}_1 = \boldsymbol{x}_0$, *and*

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha(A\boldsymbol{x}_k - b) + \beta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}), \qquad k \geq 1.$$

Let $\alpha \in \left(0, \frac{1}{L}\right)$ *and assume*

$$(1 - \sqrt{\alpha\mu})^2 \leq \beta \leq 1.$$

Then for all $k \geq 1$, *we have*

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2 + \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|^2 \leq 2\beta^k\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2.$$

In particular, for $\beta = (1 - \sqrt{\alpha\mu})^2$ *(and dropping the second term), we have*

$$\|\boldsymbol{x}_k - \boldsymbol{x}^\star\| \leq \sqrt{2}(1 - \sqrt{\alpha\mu})^k\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|.$$

Remark. If we choose $\alpha = \frac{1}{L}$, then the linear convergence rate for momentum descent is $1 - \sqrt{\kappa^{-1}}$, compared to $1 - \kappa^{-1}$ for gradient descent, where $\kappa = \frac{L}{\mu} = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1$ is the condition number of the matrix A. Thus, for a poorly conditioned matrix, $\kappa \gg 1$ and momentum descent gives a substantial improvement over gradient descent.

This Theorem is stated for a quadratic form, but holds for more general functions.

Proof I

We can write the iterations as

$$\begin{pmatrix} \boldsymbol{x}_{k+1} \\ \boldsymbol{x}_k \end{pmatrix} = B \begin{pmatrix} \boldsymbol{x}_k \\ \boldsymbol{x}_{k-1} \end{pmatrix} + \begin{pmatrix} \alpha \boldsymbol{b} \\ 0 \end{pmatrix}, \qquad B = \begin{pmatrix} I + \beta I - \alpha A & -\beta I \\ I & 0 \end{pmatrix}$$

Using this equation recursively, we have

$$\begin{pmatrix} \boldsymbol{x}_{k+1} \\ \boldsymbol{x}_k \end{pmatrix} = B^k \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_0 \end{pmatrix} + \alpha \sum_{j=0}^{k-1} B^j \begin{pmatrix} \boldsymbol{b} \\ 0 \end{pmatrix}$$

To compute the proof, we need to find the eigenvalues of B. We notice the following identity

$$\begin{pmatrix} I + \beta I - \alpha A & -\beta I \\ I & 0 \end{pmatrix} = \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} I + \beta I - \alpha \Lambda & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix}^\top .$$

Writing $\tilde{B} = \begin{pmatrix} I + \beta I - \alpha \Lambda & -\beta I \\ I & 0 \end{pmatrix}$, this shows that

$$\begin{pmatrix} V^\top \boldsymbol{x}_{k+1} \\ V^\top \boldsymbol{x}_k \end{pmatrix} = \tilde{B}^k \begin{pmatrix} V^\top \boldsymbol{x}_1 \\ V^\top \boldsymbol{x}_0 \end{pmatrix} + \alpha \sum_{j=0}^{k-1} \tilde{B}^j \begin{pmatrix} V^\top \boldsymbol{b} \\ 0 \end{pmatrix}$$

Thus, the system decomposes into 2×2 sub-blocks, with characteristic polynomials

$$p_i(\mu) = \mu^2 - (1 + \beta - \alpha\lambda_i)\mu + \beta$$

The eigenvalues of B are then

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

Proof II

The discriminant $(1+\beta-\alpha\lambda_i)^2 - 4\beta = (\beta-(1-s)^2)(\beta-(1+s)^2)$ where $s = \sqrt{\alpha\lambda_i}$ is non-positive iff

$$\beta \geq (1-\sqrt{\alpha\lambda_i})^2,$$

which is guaranteed by the assumptions $\beta \geq (1-\sqrt{\alpha\mu})^2$ and $\alpha \leq \frac{1}{L}$. In this case, both roots of p_i are complex valued and have square magnitude

$$|\mu|^2 \;=\; \frac{1+\beta-\alpha\lambda_i^2}{4} - \frac{(1+\beta-\alpha\lambda_i)^2-4\beta}{4} \;=\; \beta.$$

In particular $\|B\| \leq \sqrt{\beta} < 1$ and $\lambda = 1$ is not an eigenvalue of B and so $I - B$ is invertible and we can sum the finite geometric series to obtain

$$\begin{aligned}\begin{pmatrix}\boldsymbol{x}_{k+1}\\ \boldsymbol{x}_k\end{pmatrix} &= B^k\begin{pmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_0\end{pmatrix} + \alpha\sum_{j=0}^{k-1} B^j\begin{pmatrix}\boldsymbol{b}\\ 0\end{pmatrix}\\ &= B^k\begin{pmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_0\end{pmatrix} + \alpha(I-B^k)(I-B)^{-1}\begin{pmatrix}\boldsymbol{b}\\ 0\end{pmatrix}\end{aligned}$$

Furthermore, $(I-B)^{-1}\begin{pmatrix}\boldsymbol{b}\\ 0\end{pmatrix} = \begin{pmatrix}A^{-1}\boldsymbol{b}\\ A^{-1}\boldsymbol{b}\end{pmatrix} = \begin{pmatrix}\boldsymbol{x}^\star\\ \boldsymbol{x}^\star\end{pmatrix}$ so that

$$\begin{pmatrix}\boldsymbol{x}_{k+1}-\boldsymbol{x}^\star\\ \boldsymbol{x}_k-\boldsymbol{x}^\star\end{pmatrix} = B^k\begin{pmatrix}\boldsymbol{x}_1-\boldsymbol{x}^\star\\ \boldsymbol{x}_0-\boldsymbol{x}^\star\end{pmatrix}$$

Taking the squared norm of both sides, gives

$$\|\boldsymbol{x}_k-\boldsymbol{x}_\star\|^2 + \|\boldsymbol{x}_{k+1}-\boldsymbol{x}_\star\|^2 \;\leq\; \|B^k\|\left(\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2+\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2\right) \;\leq\; 2\beta^k\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2,$$

as desired.

References

The material in these slides is mostly based on the lecture notes by Jeff Calder on Intro to the Mathematics of Image and Data Analysis.[1] See also [1, 2, 3], [4, Ch.12], and UU Math 5770/6640.

[1] A. Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with Python and MATLAB, Second Edition*. Society for Industrial and Applied Mathematics, 2023. DOI: 10.1137/1.9781611977622.

[2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2017.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. DOI: 10.1017/cbo9780511804441.

[4] E. K. Ryu and W. Yin. *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge University Press, Nov. 2022. DOI: 10.1017/9781009160865.

[1] https://www-users.cse.umn.edu/~jwcalder/5467/index.html