Mathematical Introduction to Machine Learning

Lecture 5: Gradient Descent and Momentum Accelerations

November 14, 2024

Lecturer: Lei Wu *Scribe: Lei Wu*

1 Problem setup

Let $f : \Omega \mapsto \mathbb{R}$ be an objective function, where $\Omega \subset \mathbb{R}^d$ is the input domain. Our task is to solve the optimization problem:

$$\inf_{x\in\Omega} f(x). \tag{1}$$

Here we use $\inf$ instead of $\min$ since the minimum might be not attainable.

When Ω is a constrained domain, this is called constraint optimization. Otherwise, it is often called unconstrained optimization. In this lecture, we focus on the unconstrained case for simplicity. Handling constraints can be quite complex, and, whenever possible, it is generally advisable to reformulate your problem as an unconstrained optimization problem.

For most problems, it is impossible to solve the optimization problem (1) analytically. An optimization method [1] solves (1) by certain *iterative methods*, e.g., the gradient descent (GD):

$$x_{t+1} = x_t - \eta\nabla f(x_t),$$

where x_t is the solution at the t-step. The key question in optimization is understanding when and how these iterations converge.

Criteria for measuring convergence. Depending on the properties of the objective function, commonly used criteria include the following:

- **Point convergence.** If $x^* = \operatorname{argmin}_x f(x)$ exists, we can measure the convergence by $\|x_t - x^*\|$.
- **Loss convergence.** We can also measure convergence using $f(x_t) - \inf_x f(x)$, which works even if x^* is not attainable.
- **Gradient convergence.** For non-convex problems, x_t may only converge to a critical point where $\nabla f(x) = 0$. In such cases, we can use $\|\nabla f(x_t)\|$ to measure the convergence.

It is important to emphasize that in machine learning (ML), the most relevant criterion is loss convergence, as it directly influences model performance. Point convergence is less applicable in modern ML, where models are often over-parameterized or degenerate; in such cases, $\operatorname{argmin}_x f(x)$ may consist of multiple minima, making point convergence either impractical or not computable. Gradient convergence is frequently used when analyzing convergence in non-convex landscapes. However, it should be noted that a small gradient norm does not necessarily indicate good model performance, as the loss can still remain high.

[1] In machine learning, it is often referred to as an optimizer.

2 Gradient Descent

Gradient descent (GD) iterates as follows

$$x_{t+1} = x_t - \eta_t \nabla f(x_t),$$

where η_t is the learning rate (also called step size) of the t-th step. The intuition behind this is that GD iterates along the steepest descent direction, which is exactly $-\nabla f(x)$ if the ℓ_2 metric is considered.

Popular schedules of tuning learning rates include the following three ones.

- Line search: $\eta_t = \operatorname{argmin}_{\eta \geq 0} f(x_t - \eta \nabla f(x_t))$. This sophisticated approach is popular in classical numerical optimization but not in machine learning.
- Constant learning rate: $\eta_t = \eta$. This is most commonly-used one in machine learning.
- Decay learning rate, e.g., $\eta_t = \eta_0/(1+t)$. This type of schedules are often used when $f(\cdot)$ is non-smooth.

Before proceeding to the convergence analysis, we provide a concrete example to illustrate why the last one is needed.

Example 2.1. *Consider $f(x) = |x|$. GD becomes $x_{t+1} = x_t - \eta_t \operatorname{sign}(x_t)$. In such a case, if we want x_t to converge we must decay the learning rate towards zero; otherwise $\{x_t\}$ may oscillate around the minimum.*

In the following, we focus on the case of $\eta_t = \eta$. In such a case, when $\eta \to 0$, the gradient descent becomes the gradient flow:

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

Discrete-time GD can be viewed as the forward-Euler discretization of the continuous-time GF. The analysis of the continuous-time GF is often much simpler than discrete-time GD.

2.1 One-step loss descent

The convergence analysis relies an estimate of the one-step loss descent. For the GF, it is easy to verify

$$\frac{\mathrm{d} f(x_t)}{\mathrm{d} t} = \langle \nabla f(x_t), \dot{x}_t \rangle = -\|\nabla f(x_t)\|^2. \tag{2}$$

This provides an intution behind GF/GD convergence. The rate of loss descent depends on the gradient's norm. For the discrete-time GD, we need to further make assumption on the objective function's smoothness.

Definition 2.2 (Smoothness)**.** $f \in C^1(\mathbb{R}^d)$ is said to be L-smooth, if $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$ holds for any $x, y \in \mathbb{R}^d$.

If $f \in C^2(\mathbb{R}^d)$, the above condition is equivanlent to $\sup_x \|\nabla^2 f(x)\|_2 \leq L$. The following lemma shows that if smooth functions growth at most quadratically.

Lemma 2.3. *If f is L-smooth, we have $f(y) \leq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{L}{2}\|y - x\|^2$.*

Proof. Omitted! □

Lemma 2.4. *Assume f is L-smooth and $\eta \leq 1/L$. Then, $f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2$.*

Proof. Using the L-smoothness and Lemma (2.3), we have

$$\begin{aligned} f(x_{t+1}) &= f(x_t - \eta \nabla f(x_t) \\ &\leq \langle x_{t+1} - x_t, \nabla f(x_t) \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ &= (-\eta + \frac{L\eta^2}{2}) \|\nabla f(x_t)\|^2 \leq -\frac{\eta}{2} \|\nabla f(x_t)\|^2, \end{aligned}$$

where the last step uses $\eta \leq 1/L$.

□

The effect of finite learning rate. The choice of learning rate depends heavily on the smoothness of the objective function, as an excessively large learning rate can lead to instability and increases in the objective value. To better understand the influence of a finite learning rate, consider the one-dimensional quadratic function $f(x) = \frac{a}{2}x^2$, where $a > 0$ represents the curvature of the loss landscape near the global minimum of interest. In this case, the GD updates becomes

$$x_t = x_{t-1} - \eta a x_{t-1} = (1 - \eta a) x_{t-1} = (1 - \eta a)^t x_0.$$

We observe the following behaviors:

- When $\eta \leq 1/a$, x_t decays exponentially and monotonically to zero. This dynamical behavior is similar to GF.
- When $1/a < \eta < 2/a$, $|x_t|$ decays exponentially to zero but x_t will oscillate across the valley. The dynamical behavior is quite different from GF.
- When $\eta = 2/a$, GD converges to the periodic orbit $(x_0, -x_0)$.
- When $\eta > 2/a$, GD blows up exponentially.

These observations suggest that the dynamical behavior of gradient descent is governed by the curvature of the objective function. A similar pattern also applies to more general objective functions.

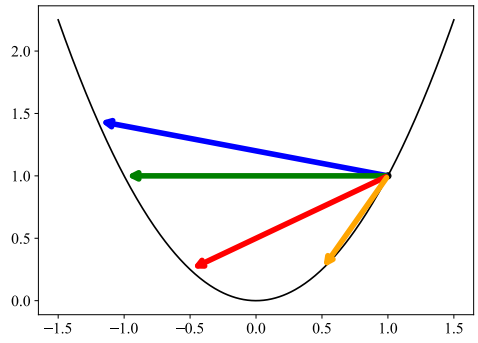

Figure 1: The effect of finite learning rate. orange: $\eta \leq 1/a$; red: $1/a < \eta < 2/a$; green: $\eta = 2/a$; blue: $\eta > 2/a$.

2.2 Non-convex analysis

Theorem 2.5. *Let $f \in C^1(\mathbb{R}^d)$. Then we have $\inf_{s\in[0,t]} \|\nabla f(x_s)\| = O(1/\sqrt{t})$.*

This theorem shows that the gradient norm decreases to zero in a $O(1/t)$ rate.

Proof. By Eq. (2), we have

$$f(x_t) - f(x_0) = \int_0^t \frac{\mathrm{d}f(x_t)}{\mathrm{d}t}\,\mathrm{d}t = -\int_0^t \|\nabla f(x_t)\|^2\,\mathrm{d}t.$$

Therefore,

$$\inf_{s\in[0,t]} \|\nabla f(x_s)\| \le \sqrt{\frac{f(x_0) - f(x_t)}{t}} \le \sqrt{\frac{f(x_0) - \inf_x f(x)}{t}}$$

□

Theorem 2.6. *Let f be L-smooth and $\{x_t\}$ be the GD solutions. Suppose $\eta \le 1/L$. Then,*

$$\min_{s=0,1\ldots,t-1} \|\nabla f(x_s)\| \le \sqrt{\frac{f(x_0) - \inf_x f(x)}{2\eta t}}.$$

Proof. By Lemma (2.4), we have

$$f(x_{t+1}) - f(x_t) \le -\frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

Summing over t and noticing that the left side is a telescoping sum, we obtain

$$\inf_x f(x) - f(x_0) \le f(x_t) - f(x_0) \le -\frac{\eta}{2}\sum_{s=0}^{t-1} \|f(x_t)\|^2.$$

This implies that

$$\min_{s=0,1\ldots,t-1} \|f(x_s)\| \le \sqrt{\frac{\sum_{s=0}^{t-1} \|f(x_t)\|^2}{t}} \le \sqrt{\frac{f(x_0) - \inf_x f(x)}{2\eta t}}.$$

□

2.3 Convex analysis

Here, we only consider the continuous-time case for simplicity. Let $S_f = \operatorname{argmin}_x f(x)$ and $d(x, A) = \inf_{x'\in A} \|x - x'\|$ for $x \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$. Note that when $f(\cdot)$ is not strongly convex, S_f may contain many points and is even a manifold. For instance, for $f(x_1, x_2) := (x_1 - 1)^2$,

$$S_f = \{x \in \mathbb{R}^2 \,:\, x_1 = 1\}.$$

Theorem 2.7. *Suppose that f is convex. Then, we have*

$$f(x_t) - \inf_x f(x) \le \frac{\mathrm{dist}^2(x_0, S_f)}{2t}$$

Proof. For any $\bar{x} \in \mathbb{R}^d$, consider the Lyapnov function

$$J(t) = t(f(x_t) - f(\bar{x})) + \frac{1}{2}\|x_t - \bar{x}\|^2. \tag{3}$$

Then, by the convexity, we have

$$\dot{J}(t) = f(x_t) - f(\bar{x}) - t\|\nabla f(x_t)\|^2 + \langle \bar{x} - x_t, \nabla f(x_t)\rangle \leq -t\|\nabla f(x_t)\|^2 \leq 0.$$

Then, we have $J(t) \leq J(0)$, which implies

$$t(f(x_t) - f(\bar{x})) + \frac{1}{2}\|x_t - \bar{x}\|^2 \leq \frac{1}{2}\|x_0 - \bar{x}\|^2. \tag{4}$$

Thus for any $\bar{x} \in S$, we have

$$f(x_t) - f(\bar{x}) \leq \frac{\|x_0 - \bar{x}\|^2}{2t}.$$

This leads to the conclusion. □

Remark 2.8 (Implicit bias). According to (4), we have for any $\bar{x} \in S$ that $\|x_t - \bar{x}\| \leq \|x_0 - \bar{x}\|$. Taking $x_0 = 0$ gives rise to

$$\|x_t\| \leq 2 \inf_{x \in S} \|\bar{x}\|, \quad \forall t \geq 0.$$

This implies that up to a constant factor, GD with zero initialization converges to minima with the roughly smallest norm.

Optimality The decay rate $O(1/t)$ is optimal for convex objective functions with minimizers in a compact domain, i.e., $S_f \neq \emptyset$.

Example 2.9. *Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|^n$. This function is convex for $n \geq 1$ since $f''(x) = n(n-1)|x|^{n-2}$. By the energy dissipation identity , we have*

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x_t) = -f'(x_t)^2 = -n^2 x_t^{2n-2} = -n^2 f^{2-\frac{2}{n}}(x_t).$$

We denote $z_t = f(x_t)$ for brevity and solve

$$\dot{z} = -n^2 z^{2-\frac{2}{n}} \quad \Rightarrow \quad \frac{d}{dt} z^{\frac{2}{n}-1} = z^{\frac{2}{n}-2}\dot{z} = -\frac{n^2}{\frac{2}{n} - 2}$$

so $z_t = \left(z_0 + \frac{n}{n-2} t\right)^{\frac{-n}{n-2}}$. Since $\frac{n}{n-2} \to 1$ as $n \to \infty$, there is no $\alpha > 1$ such that we could guarantee that $f(x_t) - \inf_x f(x) \leq \frac{C}{t^\alpha}$ for any $C > 0$ without making additional assumptions on f.

Remark 2.10. For convex functions whose minimizers lie at infinity (i.e., classification problem with cross entropy loss), the convergence rate of gradient descent can be substantially slower than $1/t$. An illustrative example is provided below.

Example 2.11. *Consider $f_\alpha : (0,\infty) \to \mathbb{R}$, $f_\alpha(x) = x^{-\alpha}$ for $\alpha > 0$. Since $f'_\alpha(x) = -\alpha\, x^{-\alpha-1}$ and $f''_\alpha(x) = -\alpha(-\alpha-1)x^{\alpha-2}$, the function f_α is convex. We can solve the gradient flow equation*

$$\dot{x} = -f'_\alpha(x) = \alpha\, x^{-\alpha-1}$$

with initial condition $x_0 = 1$ explicitly since

$$\frac{d}{dt}x^{2+\alpha} = (2+\alpha)\, x^{1+\alpha}\dot{x} = C_\alpha \quad \Rightarrow \quad x_t = (1 + C_\alpha t)^{-\frac{1}{2+\alpha}},$$

which satisfies

$$f(x(t)) = (1 + C_\alpha t)^{-\frac{\alpha}{2+\alpha}} \sim t^{-\frac{\alpha}{2+\alpha}}.$$

If α is close to zero, the objective function decays very slowly. Intuitively, the reason is that the objective function is very flat, so the gradient is too small to induce significant changes in x over a short time, and small changes in x do not decrease f by a noticeable amount.

Next, we show that the same convergence rate also hold for the discrete-time GD.

Theorem 2.12. *Assume that $f \in C^1(\mathbb{R})$ is L-smooth and convex. If the learning rate $\eta \le 1/L$, then the GD solution satisfies*

$$f(\bar{x}_T) - \inf f \le \frac{d^2(x_0, S_f)}{2T\eta},$$

where $\bar{x}_T = \sum_{t=0}^{T} x_t/T$ be the average GD solution.

Proof. By Lemma 2.4, it holds that

$$f(x_{t+1}) \le f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

Since f is convex, we have $f(x^*) \ge f(x_t) + \langle \nabla f(x_t), x^* - x_t\rangle$ for any $x^* \in S_f$. Thus, we have

$$\begin{aligned} f(x_{t+1}) - f(x^*) &\le \langle \nabla f(x_t), x_t - x^*\rangle - \frac{\eta_t}{2}\|\nabla f(x_t)\|^2 \\ &\le -\frac{1}{2\eta}\left(\|x_t - \eta\nabla f(x_t) - x^*\|^2 - \|x_t - x^*\|^2\right) \\ &= -\frac{1}{2\eta}\left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right). \end{aligned} \tag{5}$$

Hence,

$$\begin{aligned} f(\bar{x}_T) - f(x^*) \le \frac{1}{T}\sum_{t=0}^{T}(f(x_t) - f(x^*)) &\le \frac{1}{T}\sum_{t=1}^{T} -\frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t-1} - x^*\|^2\right) \\ &\le \frac{1}{2\eta T}(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2) \le \frac{\|x_0 - x^*\|^2}{2\eta T}, \end{aligned}$$

where the first step follows from the convexity of f. □

Remark 2.13. A refined analysis can prove the same result for the last-iterate solution x_T, showing that $f(x_T) = O(1/t)$. However, this analysis is complex and does not provide additional insights, so we omit it here.

2.4 KL analysis

Let us look at again the energy dissipation of GF: $\frac{\mathrm{d}f(x_t)}{\mathrm{d}t} = -\|\nabla f(x_t)\|^2$. Obviously, the following condition ensures that the rate of energy dissipation does not degenerate.

Definition 2.14. $f \in C^1(\mathbb{R}^d)$ is said to satisfy the Kurtyak-Lojasiewicz (KL) inequality if there exist $\mu > 0$ such that

$$\|\nabla f(x)\|^2 \geq \mu \left(f(x) - \inf_x f(x)\right)^\alpha \qquad \forall x \in \mathbb{R}^d,$$

where α is often called the Lojasiewicz exponent

An immediate consequence of this inequality is stated as follows.

Lemma 2.15. *If f satisfies the KL condition, then all stationary points are global minima.*

The KL condition captures how "sharp" or "flat" the landscape of f is around its critical points. It suggests that if $f(x)$ is close to $f(x^*)$, the gradient $\nabla f(x)$ must be also small, which provides a kind of control to the descent behavior of gradient-based methods.

The particular case of $\alpha = 1$ is referred to as the Polyak-Lojasiewicz (PL) condition. This case is important as it yields an exponential convergence:

Theorem 2.16. *If f satisfies the PL inequality, we have*

$$f(x_t) - \inf_x f(x) \leq e^{-\mu t}(f(x_0) - \inf_x f(x)).$$

Proof. Suppose $\inf_x f(x) = 0$. Then, $\frac{\mathrm{d}f(x_t)}{\mathrm{d}t} = -\|\nabla f(x_t)\|^2 \leq -\mu f(x_t)$, which implies the conclusion. □

If $f \in C^1(\mathbb{R}^d)$ is strongly convex, f satisfies the PL condition (see Lemma 2.19). This implies that GD converges exponentially fast for strongly convex function. However, PL is much weaker and includes situations like:

- Let $g : \mathbb{R}^k \mapsto \mathbb{R}$ be μ_0-strongly convex and $A \in \mathbb{R}^{k\times d}$. Suppose that $\sigma_k(A)$ be the smallest singular value of A. Then $f(x) := g(Ax)$ satisfy PL.

  $$\begin{aligned} \nabla f(x) &= A^\top \nabla g(Ax) \Rightarrow \\ \|\nabla f(x)\|^2 = \|A^\top \nabla g(Az)\|^2 &\geq \sigma_k^2(A)\|\nabla g(Az)\|^2 \geq \sigma_k^2(A)\mu_0 g(Ax) = \sigma_k^2(A)\mu_0 f(x). \end{aligned}$$

  When $k < d$, f cannot be strongly convex but still PL. This example includes the popular case of over-parameterized linear regression.

  $$F(\beta) = \frac{1}{n}\sum_{i=1}^n (\Phi(x_i)^T\beta - y_i)^2 = \frac{1}{n}\|\Phi(X)\beta - y\|^2,$$

  where $\Phi(X) = (\Phi(x_1), \ldots, \Phi(x_n))^\top \in \mathbb{R}^{n\times d}, \beta \in \mathbb{R}^d$. In the over-parameterized case, i.e., $d > n$, F is PL but not strongly convex. Then, minimizing F with GD will see an exponential convergence even if F is not strongly convex.

- $f(x) = x^2 + 3\sin^2(x)$ is non-convex but PL.

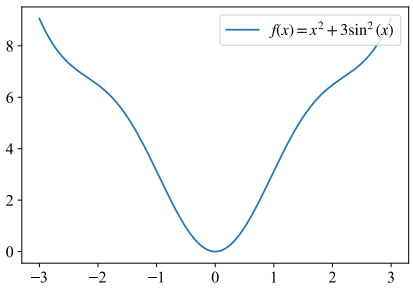

Figure 2: An illustration of $f(x) := x^2 + 3\sin^2(x)$. This function is non-convex but PL.

Theorem 2.17. *Assume f satisfies the KL condition Then,*

- *If $\alpha > 1$, then $f(x_t) - \inf_x f(x) \sim t^{-1/(\alpha-1)}$.*
- *If $\alpha = 1$, then $f(x_t) - \inf_x f(x) \sim e^{-t}$.*
- *If $\alpha < 1$, then*

$$f(x_t) - \inf_x f(x) \leq (f(x_0) - \lambda(1-\alpha)t)^{1/(1-\alpha)} \qquad \forall t < \frac{f(x_0)^{1-\alpha}}{\lambda(1-\alpha)}.$$

This means that x_t stops at finite time.

Proof. The proof is left to homework. □

So depending on the exponent α in KL equalities, three types of behaviors may occur: convergence at an algebraic rate (see also Example 2.11), convergence at an exponential rate, and in finite time. Note that the convergence in finite time cannot be recovered in practice, as the condition also prevents the objective function from being smooth close to a minimum. This requires choosing a decaying learning rate for GD, which pushes the time of convergence to infinity.

2.5 Strongly convex analysis

Definition 2.18. $f \in C^1(\mathbb{R}^d)$ is said to be strongly convex if there exist a $\mu > 0$ such that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{6}$$

Note that if $f \in C^2(\mathbb{R}^d)$, the condition (6) is equivalent to $\inf_x \lambda_{\min}(\nabla^2 f(x)) \geq \mu$.

Lemma 2.19. *If f is strongly convex with constant μ, then f satisfies the PL condition:*

$$\|\nabla f(x)\|^2 \geq 2\mu \left(f(x) - f(x^*)\right).$$

Proof. Note that the minimum of the right hand side of (6) is attained in $\tilde{y} = x - \frac{1}{\mu}\nabla f(x)$. Thus,

$$f(y) \geq f(x) + \langle \nabla f(x), \tilde{y} - x \rangle + \frac{\mu}{2}\|\tilde{y} - x\|^2$$

$$\geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

Taking $y = x^*$ completes the proof. □

Remark 2.20. It is more intuitive to check this property with $f(x) = \frac{1}{2}x^\top Ax$.

As noted above, in the case of strong convexity, the function value converges at an exponential rate. The following lemma further establishes that the points also converge exponentially.

Lemma 2.21. *When $f \in C^1(\mathbb{R})$ is strongly convex, $\|x_t - x^*\| \leq \frac{2}{\mu}e^{-2\mu t}$.*

Proof. By Theorem 2.16, we have $f(x_t) - f(x^*) \leq e^{-2\mu t}$. The strong convexity implies

$$\|x_t - x^*\|^2 \leq \frac{2}{\mu}(f(x_t) - f(x^*)) \leq \frac{2}{\mu}e^{-2\mu t}.$$

□

However, this does not imply that minimizing a quadratic function is always easy, as we have not yet considered the impact of a finite learning rate. To explore this, let us examine the following toy example.

Example 2.22. *Consider the objective function $f(x, y) = \frac{1}{2}x^2 + \frac{1}{2\varepsilon}y^2$. For GD,*

$$\begin{cases} x_{t+1} = x_t - \eta x_t \\ y_{t+1} = y_t - \frac{\eta}{\varepsilon}y_t \end{cases} \Rightarrow f(x_t, y_t) = (1-\eta)^t x_0^2 + \frac{1}{2\varepsilon}(1 - \frac{\eta}{\varepsilon})^2 y_0^2.$$

For convergence, we can only take $\eta < 2\varepsilon$. This small learning rate results in a $O((1-\varepsilon)^2)$ convergence rate.

The above calculation also holds for a general quadratic objective function $f(x) = x^\top Hx/2$ where $x \in \mathbb{R}^d$, for which GD iterates by

$$x_{t+1} = x_t - \eta Hx_t = (I - \eta H)x_t.$$

Let $H = \sum_{j=1}^d \lambda_j u_j u_j^\top = U\Sigma U^\top$ be the eigen decomposition of H. WLOG, assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Then, we can decompose the dynamics of x_t into eigen spaces:

$$x_t = \sum_{i=1}^d \tilde{x}_j(t)u_j,$$

where $\tilde{x}_j(t) = x_t^\top u_j$. Then, it is easy to obtain

$$\tilde{x}_j(t+1) = (1 - \eta\lambda_j)\tilde{x}_j(t) = (1 - \eta\lambda_j)^t \tilde{x}_j(0).$$

One can see that each eigen components updates independently. GD converges with different rates in different eigen components. To ensure stability, the learning rate must satisfy

$$\max_j |1 - \eta\lambda_j| < 1. \tag{7}$$

Then, the stability condition becomes

$$\eta \leq \frac{2}{\lambda_1}. \tag{8}$$

Consequently, we are facing a trade-off between the directions u_1 and u_d. The sharpest direction u_1 limits the maximum allowable learning rate. Consequently, with the decay in the flattest direction u_d is too slow: $(1 - \lambda_1/\lambda_d)^t$. In the literature, $\kappa := \lambda_d/\lambda_1$ is often called the **condition number**.

$$f(x_t) = \sum_{j=1}^{d} \lambda_j \tilde{x}_j^2(t) = \sum_{j=1}^{d} \lambda_j (1 - \eta\lambda_j)^t \tilde{x}_j^2(0). \tag{9}$$

If assuming that $\tilde{x}_j^2(0) \neq 0$ for all $j \in [d]$, then optimizing the choice of learning rate leads to the following convergence rate

$$f(x_t) \approx \left(\frac{\kappa - 1}{\kappa + 1}\right)^t C_0,$$

where C_0 is a constant depending on the initialization and $\{\lambda_j\}$. A straightforward calculation shows that to achieve the target precision of ϵ, we require approximately

$$\kappa \log(1/\epsilon)$$

steps. Unfortunately, for many practical problems, the condition number κ is often quite large, which makes achieving convergence challenging. Intuitively speaking, this type of slow convergence is reflected by the zig-zag oscillation in GD trajectory. See Figure 3 for an illustration.

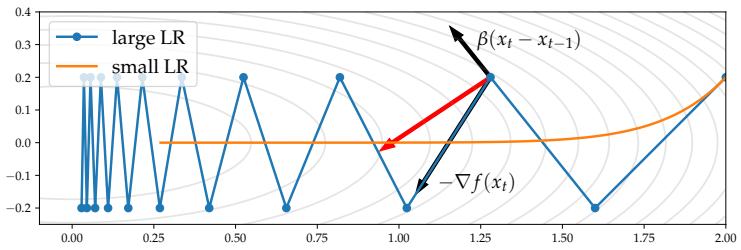

Figure 3: The GD trajectories for $f(x, y) := \frac{1}{2}(x^2 + 10y^2)$. Large-LR GD wastes time in oscillating around the valley. Small-LR GD converges well but move too little in each step. The red arrow denotes the direction proposed by heavy-ball momentum (HBM) method, which is better than the negative gradient direction.

3 Heavy-ball Momentum

To alleivate the zig-zag phenomenon of large-LR GD and accelerate GD, one idea is to use past informations to construct a better update direction. By looking at Figure 3, this seems doable and in particular visually it seems that $-\nabla f(x_t) + \beta(x_t - x_{t-1})$ can yield a better direction if choosing β appropriately. It turns out that this is exactly the Heavy-ball momentum (HBM) method introduced by [Polyak, 1964]:

$$x_{t+1} = x_t - \eta\nabla f(x_t) + \beta(x_t - x_{t-1}). \tag{10}$$

By introduce the momentum $v_t = (x_t - x_{t-1})/\eta$, the above update can be also written as

$$\begin{aligned} v_{t+1} &= \beta v_t - \nabla f(x_t) \\ x_{t+1} &= x_t + \eta v_{t+1}. \end{aligned} \tag{11}$$

In this regard, β is called the momentum factor.

Other variants. In the literature, there are also two other formulations of heavy-ball momentum.

- By let $v_t = x_t - x_{t-1}$, (10) can be rewritten as

$$\begin{aligned} v_{t+1} &= \beta v_t - \eta \nabla f(x_t) \\ x_{t+1} &= x_t + v_{t+1}, \end{aligned} \tag{12}$$

- Let $\eta = (1-\beta)\bar{\eta}$ and $v_t = (x_t - x_{t-1})/\bar{\eta}$. Then, (10) can be written as

$$\begin{aligned} v_{t+1} &= \beta v_t + (1-\beta)(-\nabla f(x_t)) \\ x_{t+1} &= x_t + \bar{\eta} v_{t+1}, \end{aligned} \tag{13}$$

  In this variant, the momentum is updated as a convex combination of the previous momentum and the current negative gradient.

These formulations are essentially equivalent but hyper-parameters may have different meanings. In all formulations, β's are the same but the learning rates may scale differently with β. Different machine learning packages may use different formulations to implement HBM. For instance, PyTorch uses (11) but TensorFlow use (12). Therefore, one should be careful about the choice of learning rate.

3.1 Preliminary analyses

Considering the update (11), we have

$$v_t = -\sum_{s=0}^{t} \beta^{t-s-1} \nabla f(x_s) + \beta^t v_0,$$

which implies that the momentum is just a *moving average* of past gradients. In particular, it tell us that we must set $\beta < 1$; otherwise, $\|v_t\|$ will blow up.

Continuous-time limit. Let $\beta = 1 - \alpha$. Then, (10) gives

$$x_{t+1} - 2x_t + x_{t-1} = -\alpha(x_t - x_{t-1}) - \eta \nabla f(x_t).$$

Deviding both sides with η gives

$$\frac{x_{t+1} - 2x_t + x_{t-1}}{(\sqrt{\eta})^2} = \frac{\alpha}{\sqrt{\eta}} \frac{x_t - x_{t-1}}{\sqrt{\eta}} - \nabla f(x_t).$$

Consider the following limiting scaling

$$\alpha, \eta \to 0, \qquad \frac{\alpha}{\sqrt{\eta}} \to \gamma.$$

Let $X(t\tau) = x_t$. Then, the continuous-time limit is given by

$$\ddot{X} = -\gamma \dot{X} - \nabla f(X). \tag{14}$$

This ODE is exactly the Newton's Law for the motion of a ball of mass 1, where $f(\cdot)$ is the potential energy and $-\gamma \dot{x}_t$ is the friction force. This explains why this method is called heavy-ball momentum method. We refer to Figure 4 for an illustration.

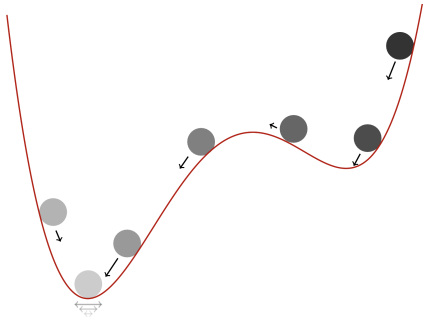

Figure 4: An illustration of the behavior of HBM dynamics.

Lemma 3.1. *For this physical system, the total energy is* $J(x, v) = f(x) + \frac{v^2}{2}$. *The energy dissipation of* (14) *is*

$$\frac{\mathrm{d}J(X_t, V_t)}{\mathrm{d}t} = -\gamma \|V_t\|^2$$

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}t} J(X_t, V_t) = \langle \nabla f(X_t), V_t \rangle + \langle V_t, -\gamma \dot{X}_t - \nabla f(X_t) \rangle = -\gamma \|V_t\|^2.$$

□

This tells us that the energy dissipation speed depends on the friction $\gamma \approx \frac{1-\beta}{\sqrt{\eta}}$. This offer an explanation of how the choice (η, β) affects the dynamical behavior of HBM.

The behavior of HBM dynamics. The continuous-time limit HBM-ODE along with its physical interpretation is very helpful for understanding the convergence behavior of HBM. In particular, when γ is not too big,

- Momentum can help escape saddle, local minima, and plateau.
- HBM may converge non-monotonically.

From the continuous-time analysis above indicates that, to realize these benefits, the momentum factor β to be close to 1, which aligns well with practical choice.

3.2 Acceleration for strongly convex problem

Figure 5 shows the comparison of GD and HBM. One can see clearly that HBM converges to the minimum with a better trajectory and the zig-zag phenomenon is greatly alleviated. This property can lead substantial acceleration for optimizing strongly convex problem, improving the iteration complexity from $\kappa \log(1/\epsilon)$ to $\sqrt{\kappa}\log(1/\epsilon)$. For simplicity, we consider a quadratic objective $f(x) = \frac{1}{2}x^\top Hx$, whose minimizer is $x^* = 0$, to illustrate the underlying mechanism.

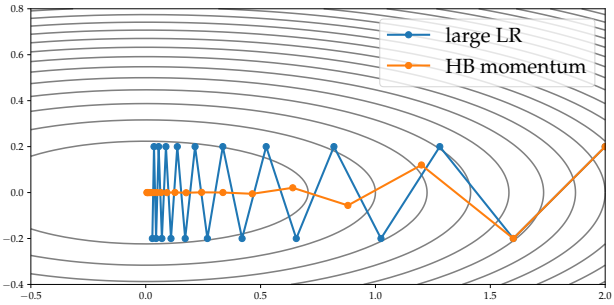

Figure 5: HBM can provide a better convergence direction without needing to reduce the learning rate.

First, analogous to the GD, the dynamics for HBM along different eigen directions is also decoupled. Still consider the decomposition $x_t = \sum_{j=1}^d x_j(t)u_j$. Then, multiplying both sides of Eq. (10) with u_j gives

$$\begin{aligned} x_j(t+1) &= x_j(t) - \eta\lambda_j x_j(t) + \beta(x_j(t) - x_j(t-1)) \\ &= (1+\beta-\eta\lambda_j)x_j(t) - \beta x_j(t-1). \end{aligned}$$

Thus, the eigen component satisfies a second-order linear recurrence, whose solution is determined by the characteristic equation

$$\mu^2 - (1+\beta-\eta\lambda_j)\mu + \beta = 0.$$

The two roots are given by

$$\mu_{j,\pm} = \frac{(1+\beta-\eta\lambda_j) \pm \sqrt{\Delta_j}}{2}, \quad \Delta_j = (1+\beta-\eta\lambda_j)^2 - 4\beta.$$

WLOG, assuming $\mu_{j,+} \neq \mu_{j,-}$. Then, the eigen component's dynamics follows

$$x_j(t) = C_+\mu_{j,+}^t + C_-\mu_{j,-}^t,$$

where $C_\pm$ are constants. To ensure convergence, (η, β) must satisfy

$$\sup_{j\in[d],\zeta\in\{+,-\}} |\mu_{j,\zeta}| < 1.$$

Optimal convergence. Note that when $\Delta_j < 0$, the two roots are conjugates of each other and satisfy

$$|\mu_{j,+}| = |\mu_{j,-}| = \sqrt{\mu_{j,+}\mu_{j,-}} = \sqrt{\beta}.$$

Under this condition, we have $|x_j(t)| \leq C\sqrt{\beta}^t$. Suprisingly, the convergence rate is independent of the curvature λ_j. Consequently, if we can choose (η, β) such that $\max_{j\in[d]} \Delta_j < 0$, then

$$\|x_t - x^*\| = O(\beta^{t/2}). \tag{15}$$

The remaining task is to determine the smallest feasible value of β.

Note that

$$\begin{aligned}\Delta_j = (1+\beta-\eta\lambda_j)^2 - 4\beta < 0 &\iff -1 \leq \frac{1+\beta-\eta\lambda_j}{2\sqrt{\beta}} \leq 1 \\ &\iff (1-\sqrt{\beta})^2 \leq \eta\lambda_j \leq (1+\sqrt{\beta})^2. \end{aligned} \tag{16}$$

To satisfy this condition for all $j \in [d]$, it is required that

$$\eta\lambda_d \geq (1-\sqrt{\beta})^2, \quad \eta\lambda_1 \leq (1+\sqrt{\beta})^2.$$

Assuming equality in both cases, we obtain

$$\sqrt{\beta} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \qquad \eta = \frac{2(1+\beta)}{\lambda_1+\lambda_d}.$$

Substituting this into (15) yields

$$\|x_t\| \leq C\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t. \tag{17}$$

Compared with GD, the dependence on the condition number is improved from κ to $\sqrt{\kappa}$.

Remark 3.2. We refer to https://distill.pub/2017/momentum/ for a better visual explanation of how momentum works.

4 Nesterov momentum

Note that the improvement of HBM applies only when the objective function is quadratic (or strongly convex). For general convex problems, both HBM and GD share the same rate of $O(1/t)$. This raise a natural question:

Can the rate $O(1/t)$ be improved by using momentum?

Nesterov's seminal work [Nesterov, 1983] fully addressed the above question. However, Nesterov's original motivation was to explore a broader question: Given the iteration methods that use only gradient information,

$$x_{t+1} = G(x_0, \nabla f(x_t), \nabla f(x_{t-1}), \cdots, \nabla f(x_0)), \tag{18}$$

where G is a general map, he asked

What the fastest rate a general gradient-based method can achieve when f is convex?

Nesterov's findings provided the following insights:

- The optimal rate is $O(1/t^2)$ in the sense that it cannot be improved by using only gradient information.
- He introduced the following accelerated gradient method:

$$\begin{aligned} x_{t+1} &= y_t - \eta \nabla f(y_t) \\ y_{t+1} &= x_{t+1} + \beta_t (x_{t+1} - x_t), \end{aligned} \tag{19}$$

  where $\beta_t = \frac{t-1}{t+2}$. Moreover, he showed (19) achieves the optimal rate; see Theorem 4.1.

Theorem 4.1 (Nesterov's Theorem)**.** *Let f be a L-smooth convex function. If x_t is generated by the NAG scheme with learning rate $\eta \leq \frac{1}{L}$, then*

$$f(x_t) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{\eta(t+1)^2}$$

In the optimization literature, (19) is often referred to as the Nesterov Accelerated Gradient (NAG) method. Comparing it with (10) reveals its distinction from HBM. The term "accelerated gradient' is used due to the improvement guaranteed by the theorem.

The proof is very complicated and highly relies on the delicate choice of β_t.

The variant in non-convex optimization NAG method becomes popular in training neural network-like models because of the work [Sutskever et al., 2013]. It provides a large number of experimental results showing the better performance of NAG compared with HBM and vanilla GD. In particular, [Sutskever et al., 2013] rewrote NAG in a way that emphasizes its similarity to HBM:

$$\begin{aligned} v_{t+1} &= \beta v_t - \eta \nabla f(x_t + \beta v_t) \\ x_{t+1} &= x_t + v_{t+1} \end{aligned}$$

Figure 6 provides a visual comparison between Nesterov momentum and heavy ball momentum.

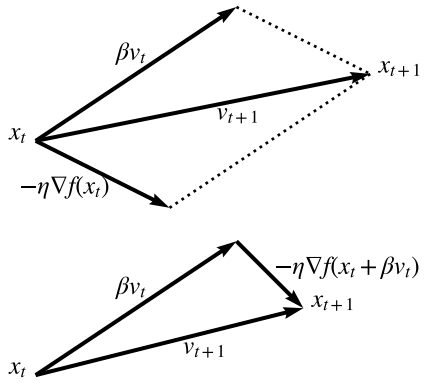

Figure 6: The comparison between two types of momentum.

In deep learning, we tend to use a constant value for the momentum factor β, e.g., 0.99. The delicate $\beta_t = (t-1)/(t+2)$ is chosen to achieve optimal rates for convex problem. However, in deep learning, objective functions are always non-convex and this particular choice is not necessary.

4.1 A continuous-time analysis

Note that

$$\begin{aligned} x_{t+1} - x_t &= y_t - \eta \nabla f(y_t) - x_t \\ &= \frac{t-1}{t+2}(x_t - x_{t-1}) - \eta \nabla f(y_t) \\ &= x_t - x_{t-1} - \frac{3}{t+2}(x_t - x_{t-1}) - \eta \nabla f(y_t), \end{aligned}$$

which can be rephrased as

$$\frac{x_{t+1} - 2x_t + x_{t-1}}{\eta} = -\frac{3}{t+1}\frac{x_t - x_{t-1}}{\eta} - \nabla f(y_t). \tag{20}$$

Let $\tau = t\sqrt{\eta}$ and $X(t\sqrt{\eta}) = x_t$. Then, the above

$$\frac{\ddot{X}(\tau)(\sqrt{\eta})^2}{\eta} + o(\sqrt{\eta}) = -\frac{3}{(t+1)\sqrt{\eta}}\dot{X}(\tau) - \nabla f(X(\tau)) + o(\sqrt{\eta}).$$

Taking $\eta \to 0$ and considering the leading term, we obtain the limiting ODE as follows

$$\ddot{X} = -\frac{3}{\tau}\dot{X} - \nabla f(X)$$

The above ODE is analogous to the HBM ODE (14). The difference is that in HBM, the friction factor is a constant, while in NAG the friction factor $3/\tau$ decays to zero as $\tau \to \infty$.

For brevity, we will still use t to denote the continuous time.

Theorem 4.2. *Suppose $\dot{X}_t = 0$. Then,*

$$f(X_t) - f^* \le \frac{2\|X_0 - x^*\|^2}{t^2}$$

Proof. Consider the energy functional defined as

$$\mathcal{E}(t) := t^2 \left(f(X_t) - f^\star\right) + 2\left\|X_t + \frac{t}{2}\dot{X}_t - x^\star\right\|^2$$

whose time derivative is

$$\dot{\mathcal{E}} = 2t\left(f(X) - f^\star\right) + t^2\langle \nabla f, \dot{X}\rangle + 4\left\langle X + \frac{t}{2}\dot{X} - x^\star, \frac{3}{2}\dot{X} + \frac{t}{2}\ddot{X}\right\rangle$$

Substituting $3\dot{X}/2 + t\ddot{X}/2$ with $-t\nabla f(X)/2$, (3.3) gives

$$\dot{\mathcal{E}} = 2t\left(f(X) - f^\star\right) + 4\left\langle X - x^\star, -\frac{t}{2}\nabla f(X)\right\rangle = 2t\left(f(X) - f^\star\right) - 2t\left\langle X - x^\star, \nabla f(X)\right\rangle \le 0,$$

where the inequality follows from the convexity of f. Hence by monotonicity of $\mathcal{E}$ and non-negativity of $2\left\|X + t\dot{X}/2 - x^\star\right\|^2$, the gap obeys $f(X_t) - f^\star \le \mathcal{E}(t)/t^2 \le \mathcal{E}(0)/t^2 = 2\left\|x_0 - x^\star\right\|^2/t^2$. □

The above elegant continuous-time analysis is from [Su et al., 2016].

References

[Nesterov, 1983] Nesterov, Y. (1983). A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543.

[Polyak, 1964] Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17.

[Su et al., 2016] Su, W., Boyd, S., and Candès, E. J. (2016). A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43.

[Sutskever et al., 2013] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147. PMLR.