Mathematical Introduction to Machine Learning

Lecture 5: Gradient Descent and Momentum Accelerations

October 26, 2023

Lecturer: Lei Wu *Scribe: Lei Wu*

1 Problem setup

Let $f : \Omega \mapsto \mathbb{R}$ be an objective function, where $\Omega \subset \mathbb{R}^d$. 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 most problems, it is impossible to solve (1) analytically. An optimization method [1] solves (1) by certain iteration methods, e.g., the gradient descent (GD) $x_{t+1} = x_t - \eta\nabla f(x_t)$, where x_t be the solution at t-step. The most important question in optimization is to understand when and how the iteration converges?

Criteria of measuring convergence.

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

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.

- Constant learning rate: $\eta_t = \eta$. This is most commonly-used one in machine learning.
- Line search: $\eta_t = \operatorname{argmin}_{\eta\geq 0} f(x_t - \eta\nabla f(x_t))$. This sophisticated approach is common in classical numerical optimization but not very popular in machine learning.

[1] In machine learning, it is also called optimizer.

- 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 Non-convex analysis

Theorem 2.2. *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. Note that

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

Therefore,

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

Therefore,

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

□

The discrete-time analysis needs stronger condition.

Definition 2.3 (Smoothness)**.** $f \in C^1(\mathbb{R}^d)$ is said to be L-smooth, if $\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|$ holds for any $x, y \in \mathbb{R}^d$.

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

Lemma 2.4. *If f is L-smooth, we have $f(y) \le f(x) + \langle y - x, \nabla f(x)\rangle + \frac{L}{2}\|y - x\|^2$.*

Proof. Omitted! □

Theorem 2.5. *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. Using the L-smoothness and Lemma (2.4), we have

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

Since $\eta \leq 1/L$, we have

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

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

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

This implies that

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

□

From the proof, one can see that the choice of learning rate depends on the smoothness. A too-large learning rate may cause an increase in objective value. (Instability!)

2.2 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, if $f(x) = (x_1x_2 - 1)^2$, then

$$S_f = \{x \in \mathbb{R}^2 \,:\, x_1x_2 = 1\}.$$

Theorem 2.6. *Suppose that f is convex. Then, we have*

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

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

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

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

$$t(f(x_t) - f(\bar{x})) + \frac{1}{2}\|x_t - \bar{x}\|^2 \leq \frac{1}{2}\|x_0 - \bar{x}\|^2. \tag{3}$$

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 2.7 (Implicit bias). According to (3), 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 almost optimal for convex objective functions with minimizers, i.e., $S_f \neq \emptyset$.

Example 2.8. *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.

For convex functions which do not have minimizers, the value of the objective function may decay significantly more slowly than $1/t$.

Example 2.9. *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.

2.3 KL analysis

Definition 2.10. $f \in C^1(\mathbb{R}^d)$ is said to satisfy the Kurtyak-Lojasiewicz (KL) inequality if there exist $\mu > 0$ such that

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

In particular, when $\alpha = 1$, it is often referred to as the Polyak-Lojasiewicz (PL) inequality.

Note that $\frac{\mathrm{d}f(x_t)}{\mathrm{d}t} = -\|\nabla f(x_t)\|^2$. Therefore, the KL condition ensures that the energy dissipation of GF is sufficiently large.

Lemma 2.11. *If f satisfies the KL condition, then all stationary points are global minima.*

Theorem 2.12. *If f satisfy 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. □

When $f \in C^2(\mathbb{R}^d)$ is strongly convex with $\inf_x \lambda_{\min}(\nabla^2 f(x)) \geq \mu_0 > 0$. Then, f satisfies the PL condition with $\mu = \mu_0$. This implies that GD converges exponentially fast for strongly convex function. However, PL is much weaker than strong convexity:

- 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.

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

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

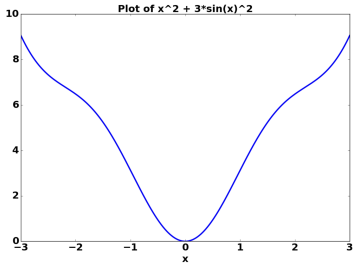
  $$F(\beta) = \frac{1}{n}\sum_{i=1}^n (\Phi(x_i)^T\beta - y_i)^2 = \frac{1}{n}\|\hat{\Phi}(X)\beta - y\|^2,$$

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

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

Theorem 2.13. *Assume f satisfies the KL condition and $\inf_x f(x) = 0$. Then,*

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



- *If $\alpha < 1$, then*

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

  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 in finite time, convergence at an exponential rate, and convergence at an algebraic rate (see also Example 2.9). 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 decaying time-step sizes in the gradient descent scheme, which pushes the time of convergence to infinity.

2.4 Convergence of x_t

Definition 2.14. $f \in C^1(\mathbb{R}^d)$ is said to be strongly convex if there exist a $\mu > 0$ such that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{4}$$

Note that if $f \in C^2(\mathbb{R}^d)$, (4) is equivalent to $\inf_x \lambda_{\min}(\nabla^2 f(x)) \geq \mu$.

Lemma 2.15. *If f is strongly convex with constant μ, then f also satisfy the PL condition, i.e.,*

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

Proof. Note that the minimum of the right hand side of (4) is attained in $\tilde{y} = x - \frac{1}{\mu}\nabla f(x)$. Thus,

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), \tilde{y} - x \rangle + \frac{\mu}{2}\|\tilde{y} - x\|^2 \\ &\geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2. \end{aligned}$$

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

Remark 2.16. It is more intuitive to check this property with $f(x) = \frac{1}{2}x^T A x$.

Lemma 2.17. *When $f \in C^1(\mathbb{R})$ is strongly convex, $\|x_t - x^*\| \leq \frac{2}{\mu}e^{-2\mu t}$.*

Proof. By Theorem 2.12, we have $f(x) - f(x^*) \leq e^{-2\mu t}$. The strong convexity implies

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

□

Other behaviors for point convergence.

- When global minima are at the infinite, x_t diverges even if the objective is convex. For instance, the classification problem with cross-entropy/exponential loss:

$$\min_{\beta,\beta_0} \frac{1}{n} \sum_{i=1}^{n} e^{-y_i(\beta^T x_i + \beta_0)}.$$

- Even if minimizers exist and the objective function decays to its global minimum along the gradient flow, it is entirely possible that gradient flow trajectories do not converge to a limit.

  Consider the function $f(x, y) = x^2 e^{-y^4}$. The set of minimizers of f is the line $\{x = 0\}$, but the decay of the exponential term is so rapid that the objective function is reduced more by increasing $|y|$ than be taking $x \to 0$. We present a trajectory of the gradient flow such that

$$\lim_{t\to\infty} f\big(x(t), y(t)\big) = 0, \qquad \lim_{t\to\infty} x(t) = \infty, \qquad \lim_{t\to\infty} y(t) = 0.$$

  The gradient of f is given by

$$\nabla f(x, y) = e^{-y^4}\big(2x,\ -4x^2y^3\big) = 2x\, e^{-y^2}\big(1, -2xy^3\big),$$

  so the curve $(x, y)(t) = \big(t, 1/(\sqrt{2}t)\big)$ satisfies

$$(\dot{x}, \dot{y})(t) = \left(1, -\frac{1}{\sqrt{2}\, t^2}\right) = \left(1, -2t \cdot \left(\frac{1}{\sqrt{2}\, t}\right)^3\right) = \big(1, 2x(t)\, y^3(t)\big) = \frac{\nabla f(x, y)}{2x\, e^{-y^2}}.$$

  In particular, after reversing and rescaling time appropriately, $(x, y)(t)$ is a solution of the gradient flow.

2.5 The effect of finite learning rate

We have only considered GF in the preceding analysis for simplicity but what is the difference between GF and GD? To understand the effect of the finite learning rate, we here consider the quadratic objective $f(x) = \frac{1}{2}x^T Ax$, for which $x^* = 0$.

We first take a look at one-dimensional case: $f(x) = \frac{a}{2}x^2$. In this case,

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

We have the following observation:

- When $\eta \leq 1/a$, x_t decays exponentially and monotonically to zero. The dynamical behavior is similar to GF.
- When $1/a \leq \eta < 2/a$, x_t decays exponentially to zero but x_t will osccilate 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.

We next take a look at multi-dimensional example.

Example 2.18. *Consider the objective function $f(x,y) = \frac{1}{2}x^2 + \frac{1}{2\varepsilon}y^2$. For gradient descent,*

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

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 general quadratic objective functions, for which GD iterates by

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

Consider the eigen-decomposition $A = \sum_{j=1}^d \lambda_j u_j u_j^T = U \Sigma U^T$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Let $\tilde{x}(t) = U x_t$, we have

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

One can see that each coordinate (in the eigenspace) updates independently. GD converges with different rates in different coordinates. To ensure stability, the learning rate must satisfy

$$\max_j |1 - \eta\lambda_j| < 1. \tag{5}$$

Then, the stability condition becomes

$$\eta \leq \frac{2}{\lambda_1}. \tag{6}$$

Consequently, we are facing a tradeoff between the directions u_1 and u_d. Taking a small η, the decay in the u_1 is slow; but taking a large η, saying close to $2/\lambda_1$, the decay in u_d is at most $(1 - \lambda_1/\lambda_d)^t$. In the literature, $\kappa := \lambda_d/\lambda_1$ is often called the condition number, which characterize how stiff our problem is.

$$f(x_t) = \sum_{j=1}^d \lambda_j \tilde{x}_j^2(t) = \sum_{j=1}^d \lambda_j (1-\eta\lambda_j)^t \tilde{x}_j^2(0). \tag{7}$$

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

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

where C_0 is a constant depending on the initialization.

Intuitively speaking, this type of slow convergence is reflected by the zig-zag oscillation in GD trajectory. See Figure 1 for an illustration.

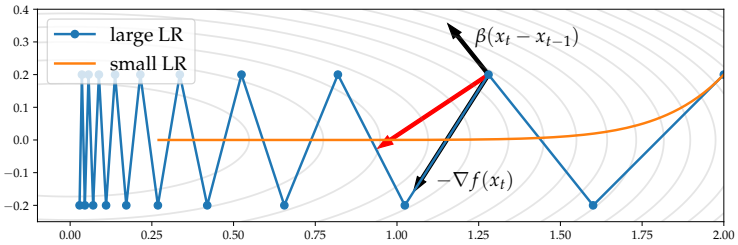


Figure 1: The GD trajectories. Large-LR GD wastes time in oscillating around the valley. Small-LR GD converges well but move too little in each step. The red arrow denotes the direction proposed by heavy-ball momentum (HBM) method, which is better than the negative gradient direction.

3 Heavy-ball Momentum

To alleivate the zig-zag phenomenon of large-LR GD, one idea is to use past informations to construct a better update direction. By looking at Figure 1, 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 in 1964:

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1}). \tag{8}$$

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

$$\begin{aligned} v_{t+1} &= \beta v_t - \nabla f(x_t)) \\ x_{t+1} &= x_t + \eta v_{t+1}. \end{aligned} \tag{9}$$

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 $v_t = x_t - x_{t-1}$, (8) can be rewritten as

$$\begin{aligned} v_{t+1} &= \beta v_t - \eta \nabla f(x_t) \\ x_{t+1} &= x_t + v_{t+1}, \end{aligned} \tag{10}$$

- Let $\eta = (1-\beta)\bar{\eta}$ and $v_t = (x_t - x_{t-1})/\bar{\eta}$. THen, (8) can be written as

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

These formulations are (almost) equivalent but hyper-parameters may have different meanings. But in all formulations, β's are the same but the learning rates may scale differently with β.

Remark 3.1. Different machine learning packages may use different formulations to implement HBM. For instance, PyTorch uses (9) but TensorFlow use (10). Therefore, one should be careful about the choice of learning rate.

The momentum method uses the trajectory history to modify the update direction, which expectedly can accelerate the convergence.

3.1 Preliminary analyses

Considering the update (9), we have

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

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

Continuous-time limit. Let $\beta = 1 - \alpha$. Then, (8) 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

$$\ddot{X} = -\gamma \dot{X} - \nabla f(X) \tag{12}$$

which 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. For this phsycial system, the total energy is

$$J(x, v) = f(x) + \frac{v^2}{2}.$$

The understanding explains why this method is called heavy-ball momentum method.

Lemma 3.2. *The energy dissipation of* (12) *is*

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

This lemma tells us that the dissipation speed depends on the friction γ. This physical interpretation implies that the momentum in HBM can help escape from saddle points and bad local minima.

3.2 Acceleration for strongly convex problem

Figure 2 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 alleivated. To mathematically analyze this acceleration of convergence, we consider a quadratic objective $f(x) = \frac{1}{2} x^T H x$, whose minimizer is $x^* = 0$.

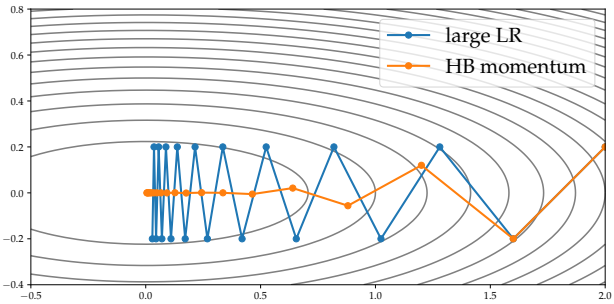


Figure 2: Heavy-ball momentum can provide a better convergence direction without needing to reduce the learning rate.

According to the Polyak formulation (8), we can rewrite HBM as

$$\begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix} = \begin{pmatrix} (1+\beta)I - \eta H & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix} =: J \begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix}$$

It is well-known that the convergence of this update is determined by the eigenvalue of J. Let $H = \sum_{j=1}^{d} \lambda_j u_j u_j^T = U \Sigma U^T$ be the eigen-decomposition of H. Let $y_t = U(x_t, x_{t-1})^T$. Then,

$$y_{t+1} = \begin{pmatrix} J_1 & 0 & \dots 0 & \\ 0 & J_2 & \dots 0 & \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & J_d \end{pmatrix} y_t,$$

where

$$J_j = \begin{pmatrix} 1+\beta-\eta\lambda_j & -\beta \\ 1 & 0 \end{pmatrix}$$

Note that $\det(J_j) = \beta$ and the eigenvalues of J_j are

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

Taking η, β such that

$$(1+\beta-\eta\lambda_j)^2 - 4\beta < 0 \iff -1 \le \frac{1+\beta-\eta\lambda_j}{2\sqrt{\beta}} \le 1, \tag{13}$$

then we have that $\mu_{j,\pm}$ are complex and

$$|\mu_{j,\pm}|^2 = \mu_{j,-}^* \mu_{j,+} = \det(J_j) = \beta.$$

Therefore, in such a case, $|\mu_{j,\pm}| = \sqrt{\beta}$.

Noticing that $\lambda_j \in [\lambda_d, \lambda_1]$, thus we can choose β as small as possible but with the following condition satisfied

$$\frac{1+\beta-\eta\lambda_d}{2\sqrt{\beta}} \le 1 \quad \text{and} \quad \frac{1+\beta-\eta\lambda_1}{2\sqrt{\beta}} \ge -1$$

Suppose that the equalities hold. Then, we obtain

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

Thus,

$$\|x_t\|^2 + \|x_{t-1}\|^2 = \|y_t\|^2 \leq \sqrt{\beta}^t \|y_0\|^2 \lesssim \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t.$$

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

We remark that the above improvement only holds when the objective function is quadratic or the local quadratic approximation is validated. For general convex problems, the improvement is not significant; HBM and GD have the same rate of $O(1/t)$. Then, a natural question is:

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

4 Nesterov momentum

Nesterov's classical work completely answers the above question.

- GD with Nesterov momentum converges in a $O(1/t^2)$ rate for convex problems.
- The rate $O(1/t^2)$ is optimal in the sense that it cannot be improved by using only gradient information.

Remark 4.1. Newton's method has a faster rate but needs second-order information.

Nesterov in 1984 introduced the following update schedule

$$\begin{aligned} x_{t+1} &= y_t - \eta \nabla f(y_t) \\ y_{t+1} &= x_{t+1} + \beta_t (x_{t+1} - x_t), \end{aligned} \tag{14}$$

where $\beta_t = \frac{t-1}{t+2}$. This is often refered as the Nesterov accelerate gradient (NAG) method in the literature of convex optimization. The difference from HBM can be seen from the comparison with (8). It is called "accelerate gradient" due to the following theorem.

Theorem 4.2 (Nesterov's Theorem)**.** *Let f be a convex function such that ∇f is C_L-Lipschitz. If x_t is generated by the NAG scheme with learning rate $\eta \leq \frac{1}{C_L}$, then*

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

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:

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

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.

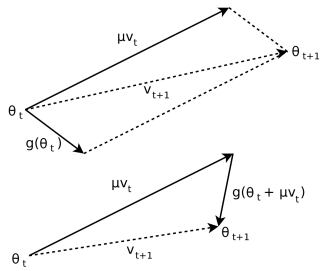


Figure 3: The comparison between two types of momentums.

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

$$\frac{x_{t+1} - 2x_t + x_{t-1}}{\eta} = -\frac{3}{t+1}\frac{x_t - x_{t-1}}{\eta} - \nabla f(y_t). \tag{15}$$

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 ODE (12) that is for HBM. 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 4.3. *Suppose $\dot{X}_t = 0$. Then,*

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

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 nonnegativity 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

[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.