MIT 6.7220/15.084 — Nonlinear Optimization Tue, Mar 7th 2024

Lecture 8
Acceleration and momentum

Instructor: Prof. Gabriele Farina (✉ gfarina@mit.edu)*

In a groundbreaking paper in 1983, Nesterov, Y. [Nes83] showed that a simple variant of gradient descent—called *accelerated* gradient descent and applicable to any L-smooth convex function—produces iterates with optimality gap $f(x_t) - f_\star$ of order $1/t^2$, as opposed to the $1/t$ rate seen in the previous lecture. The intuition behind accelerated gradient descent is notoriously hard to grasp. The original proof, rife with algebraic manipulations, is notoriously elusive and has led several authors to investigate what principles make acceleration possible at a deep level, hoping to generalize the fundamental principles beyond just gradient descent. These efforts include at least the following directions.

- Some authors have explained accelerated gradient descent as a reflection of a discretization of specific ordinary differential equations. This includes the works [SBC16], [KBB15], [WWJ16].
- The work [BLS15] proposed a simple geometric explanation for the possibility of acceleration based on certain properties of balls in Euclidean space.
- The works [WA18] and [CST21] proposed an interpretation leveraging in light of certain advancements in the theory of *online* optimization algorithm.
- The work [AS22] analyzed acceleration as an approximation of another—much more well-understood—method called the "proximal point method".

Finally, Allen-Zhu, Z., & Orecchia, L. [AO17] proposed a different framing, whereby accelerated gradient descent is simply an interpolation ("linear coupling") between two fundamental descent modes: gradient descent—fast for large gradients—and mirror descent—fast for small gradients. This is the point of view we will adopt today.

1 A tale of two descent modes: a second look at the descent lemmas

Consider a convex and L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$, and let $x_\star \in \mathbb{R}^n$ be a minimizer of f. Consider the iterates x_t produced by gradient descent run with step size $\eta > 0$. The *gradient* descent lemma and the Euclidean *mirror* descent lemma we saw in Lecture 7 provide two *conceptually different* mechanisms for measuring progress at each gradient descent step.

- The gradient descent lemma asserts that the progress made *in the function value* in two consecutive iterates x_t and x_{t+1} is at least as big as *the norm of the gradient of f at x_t*: provided $\eta \le \frac{1}{L}$, then
$$(f(x_{t+1}) - f_\star) \le (f(x_t) - f_\star) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2.$$
- The Euclidean mirror descent lemma asserts that

*These notes are class material that has not undergone formal peer review. The TAs and I are grateful for any reports of typos.

$$\|x_{t+1} - x_\star\|_2^2 \leq \|x_t - x_\star\|_2^2 + \eta^2 \|\nabla f(x_t)\|_2^2 - 2\eta \cdot (f(x_t) - f_\star). \tag{1}$$

Hence, the Euclidean mirror descent lemma establishes that the *distance from the optimal solution* decreases fast when the optimality gap $f(x_t) - f_\star$ is large and the gradient norm $\|\nabla f(x_t)\|_2$ is small.

Remark 1.1. When the stepsize η is chosen so that $0 < \eta \leq \frac{1}{L}$, we can apply the gradient descent lemma once in (1) and find that

$$\|x_{t+1} - x_\star\|_2^2 \leq \|x_t - x_\star\|_2^2 - 2\eta \cdot (f(x_{t+1}) - f_\star),$$

which implies a monotonic decrease in the Euclidean distance to optimality.

The two lemmas focus on two different performance metrics. The gradient descent lemma focuses on progress in the function value. The Euclidean mirror descent lemma focuses on progress on the Euclidean distance to optimality.

1.1 A thought experiments

We will consider a thought experiment to build intuition behind the construction of accelerated gradient descent. Imagine running gradient descent on an L-smooth function using stepsize η. We now consider two extreme cases.

■ **Large gradients.** As a first case, we consider the case of "large" gradient norms.

Theorem 1.1. Suppose that all the gradients of the points produced by gradient descent satisfy

$$\|\nabla f(x_t)\|_2^2 \geq \gamma \qquad \text{at all } t = 0, 1, ...$$

for some constant $\gamma > 0$. In this case, using the stepsize $\eta = \frac{1}{L}$, after $T_{\text{half}} := \frac{L}{\gamma}(f(x_0) - f_\star)$ iterations the optimality gap will halve, that is,

$$f(x_{T_{\text{half}}}) - f_\star \leq \frac{f(x_0) - f_\star}{2}.$$

Proof. The gradient descent lemma implies that

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L}\|\nabla f(x_t)\|_2^2 \leq -\frac{\gamma}{2L}.$$

So, after T_{half} iterations the function value is

$$f(x_{T_{\text{half}}}) - f(x_0) \leq -\frac{\gamma}{2L} T_{\text{half}} = -\frac{1}{2}(f(x_0) - f_\star).$$

Rearranging yields the statement. □

■ **Small gradients.** On the other extreme, we consider the case in which all the gradients encountered are "small", in the sense that their square norm is below some threshold value γ.

Theorem 1.2. Consider the case where

$$\|\nabla f(x_t)\|_2^2 < \gamma \qquad \text{at all } t = 0, 1, ...$$

for some constant $\gamma > 0$. In this case, using the stepsize $\eta := (f(x_0) - f_\star)/(2\gamma)$, after

$$T_{\text{half}} := 4\gamma \frac{\|x_0 - x_\star\|_2^2}{(f(x_0) - f_\star)^2}$$

iterations we will find a point $x_{\text{half}} \in \{x_0, x_1, ..., x_{T_{\text{half}}}\}$ with optimality gap

$$f(x_{\text{half}}) - f_\star \leq \frac{f(x_0) - f_\star}{2}.$$

Proof. In this case, *for every* $\eta > 0$, the Euclidean mirror descent lemma guarantees that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} f(x_t) &\leq f_\star + \frac{1}{2\eta T}\|x_\star - x_0\|_2^2 + \frac{\eta}{2T}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|_2^2 \\ &< f_\star + \frac{1}{2\eta T}\|x_\star - x_0\|_2^2 + \frac{\eta\gamma}{2}.\end{aligned}$$

Rearranging terms, we find

$$\frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f_\star) < \frac{1}{2\eta T}\|x_\star - x_0\|_2^2 + \frac{\eta\gamma}{2}$$

Setting $\eta = \frac{f(x_0)-f_\star}{2\gamma}$ and plugging the value of T_{half} given in the statement, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f_\star) &< \frac{\gamma}{T(f(x_0) - f_\star)}\|x_\star - x_0\|_2^2 + \frac{f(x_0) - f_\star}{4} \\ &= \frac{f(x_0) - f_\star}{2}.\end{aligned}$$

Finally, recognizing that the minimum is upper bounded by the average, we conclude that

$$\min_{t=1}^{T} f(x_t) - f_\star < \frac{f(x_0) - f_\star}{2},$$

which implies the statement. □

■ **Balancing the two cases.** In summary, the two cases above reveal that, assuming the stepsize is chosen well:

- when $\|\nabla f(x_t)\|_2^2 \geq \gamma$ at all times t, we can halve the optimality gap within

  $$\frac{L}{\gamma}(f(x_0) - f_\star)$$

  iterations; and
- when $\|\nabla f(x_t)\|_2^2 \leq \gamma$ at all times t, we can halve the optimality gap within

  $$4\gamma \frac{\|x_0 - x_\star\|_2^2}{(f(x_0) - f_\star)^2}$$

  iterations.

In the first case, the number of required iterations decreases as γ increases, while in the second case, it increases. The value of γ that minimizes the maximum halving time across the two cases is therefore attained when

$$\frac{L}{\gamma}(f(x_0) - f_\star) = 4\gamma \frac{\|x_0 - x_\star\|_2^2}{(f(x_0) - f_\star)^2} \qquad \Longrightarrow \qquad \gamma = \frac{\sqrt{L}(f(x_0) - f_\star)^{3/2}}{2\|x_0 - x_\star\|}.$$

For such a value of the threshold γ, both cases require at most

$$T_{\text{half}} := \frac{2\|x_0 - x_\star\|_2 \sqrt{L}}{\sqrt{f(x_0) - f_\star}}$$

iterations to halve the optimality gap.

This is in contrast with running gradient descent with the optimal stepsize $\eta = \frac{1}{L}$, which instead requires a number of iterations at most (Theorem 2.6 of Lecture 7[1])

$$\frac{L\|x_0 - x_\star\|_2^2}{f(x_0) - f_\star},$$

a quadratic slowdown compared to T_{half}.

2 Allen-Zhu and Orecchia's linear coupling

Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex, with minimum value $f_\star$ attained in (at least one) point $x_\star \in \mathbb{R}^n$. The discussion above reveals that *there is hope* to construct an accelerated gradient descent method. However, the approach is not formal since it might be the case that neither $\|\nabla f(x_t)\|_2^2 \geq \gamma$ at all times, nor $\|\nabla f(x_t)\|_2^2 \leq \gamma$ at all times.

To fix the construction, Allen-Zhu and Orecchia propose an algorithm that performs *both*

- a step with learning rate $1/L$, corresponding to the "large gradient" case above ("short step"); and
- a step using the larger stepsize corresponding to the "small gradients" case above ("long step").

The "final" iterate produced by the algorithm is computed as a linear interpolation between these two steps.

In formulas, the algorithm keeps track of three sequences $\{x_t\}, \{y_t\}, \{z_t\}$. The sequence x_t corresponds to the "final" iterate, while the sequences y_t and z_t correspond to the short and long steps, respectively. At the beginning,

$$x_0 = y_0 = z_0.$$

Then, at each iteration t, we let

$$\begin{aligned}
x_{t+1} &:= (1 - \tau) y_t + \tau z_t && \text{(interpolation with coupling rate } \tau\text{)} \\
y_{t+1} &:= x_{t+1} - \frac{1}{L}\nabla f(x_{t+1}) && \text{("short" gradient step)} \\
z_{t+1} &:= z_t - \alpha \nabla f(x_{t+1}) && \text{("long" gradient step, with stepsize } \alpha\text{)}
\end{aligned}$$

[1] Theorem 2.6 of Lecture 7 asserts that the t-th iterate produced by the gradient descent algorithm run with stepsize $\eta = \frac{1}{L}$ satisfies $f(x_t) - f_\star \leq L\frac{\|x_0 - x_\star\|_2^2}{2t}$. So, after $t = L\frac{\|x_0 - x_\star\|_2^2}{f(x_0) - f_\star}$ iterations, we have $f(x_t) - f_\star \leq \frac{f(x_0) - f_\star}{2}$.

Remark 2.1. The quantity $z_t = z_0 - \alpha \sum_{s=1}^{t} \nabla f(x_s)$ keeps track of the sum of past gradients, which is then combined into the definition of x_{t+1}. This term is often called *momentum*.

Intuitively, when the optimization algorithm is unstable, and the gradients go back and forth, the momentum is small, so the learning rate can be decreased to stabilize the algorithm. On the other hand, when the algorithm is making steady progress in the same direction, the momentum term increases the learning rate to accelerate convergence.

The analysis of the convergence rate of this interpolated variant of gradient descent follows the same conceptual steps we saw last time: first, we will establish an interpolated version of the gradient descent lemma, and then we will establish an interpolated version of the Euclidean mirror descent lemma.

2.1 The coupled gradient descent lemma

Since y_{t+1} is obtained from x_{t+1} by taking a step in the direction $-\nabla f(x_{t+1})$ using the theoretically-optimal stepsize of $1/L$, the proof of the gradient descent lemma seen in Lecture 7 applies verbatim, yielding that all times t,

$$f(y_{t+1}) \leq f(x_{t+1}) - \frac{1}{2L} \left\| \nabla f(x_{t+1}) \right\|_2^2.$$

2.2 The coupled Euclidean mirror descent lemma

The derivation of an interpolated version of the Euclidean mirror descent lemma is significantly more laborious but only involves elementary techniques. In particular, we have the following.

Theorem 2.1. At all times t,

$$f(x_{t+1}) - f_\star \leq \frac{1}{2\alpha} \Big(\|x_\star - z_t\|_2^2 - \|x_\star - z_{t+1}\|_2^2 + \|z_t - z_{t+1}\|_2^2 \Big) + \frac{1-\tau}{\tau} (f(y_t) - f(x_{t+1}))$$

Proof (Optional). Like in Lecture 7, we start from the three-point equality to write

$$\langle z_t - z_{t+1}, z_t - x_\star \rangle = \frac{1}{2} \Big(\|x_\star - z_t\|_2^2 - \|x_\star - z_{t+1}\|_2^2 + \|z_t - z_{t+1}\|_2^2 \Big).$$

Using the fact that $z_{t+1} = z_t - \alpha \nabla f(x_{t+1})$, we can therefore write

$$\alpha \langle \nabla f(x_{t+1}), z_t - x_\star \rangle = \frac{1}{2} \Big(\|x_\star - z_t\|_2^2 - \|x_\star - z_{t+1}\|_2^2 + \|z_t - z_{t+1}\|_2^2 \Big).$$

We now use the definition of linear coupling

$$x_{t+1} = (1-\tau) y_t + \tau z_t \qquad \Longrightarrow \qquad (x_{t+1} - x_\star) = (z_t - x_\star) + \frac{1-\tau}{\tau} (y_t - x_{t+1})$$

to write

$$
\begin{aligned}
&\langle \nabla f(x_{t+1}), x_{t+1} - x_\star \rangle \\
&\quad = \langle \nabla f(x_{t+1}), z_t - x_\star \rangle + \frac{1-\tau}{\tau} \langle \nabla f(x_{t+1}), y_t - x_{t+1} \rangle \\
&\quad = \frac{1}{2\alpha} \Big(\|x_\star - z_t\|_2^2 - \|x_\star - z_{t+1}\|_2^2 + \|z_t - z_{t+1}\|_2^2 \Big) + \frac{1-\tau}{\tau} \langle \nabla f(x_{t+1}), y_t - x_{t+1} \rangle \\
&\quad \leq \frac{1}{2\alpha} \Big(\|x_\star - z_t\|_2^2 - \|x_\star - z_{t+1}\|_2^2 + \|z_t - z_{t+1}\|_2^2 \Big) + \frac{1-\tau}{\tau} (f(y_t) - f(x_{t+1})). \quad \text{(Convexity)}
\end{aligned}
$$

Since by convexity

$$
f_\star \geq f(x_{t+1}) + \langle \nabla f(x_{t+1}), x_\star - x_{t+1} \rangle \qquad \Longrightarrow \qquad \langle \nabla f(x_{t+1}), x_{t+1} - x_\star \rangle \geq f(x_{t+1}) - f_\star,
$$

we obtain the statement. □

Just like in the proof in Lecture 7, we can use the fact that $z_{t+1} = z_t - \alpha \nabla f(x_{t+1})$ in the coupled Euclidean mirror descent lemma above to find that

$$
\begin{aligned}
f(x_{t+1}) - f_\star &\leq \frac{1}{2\alpha} \Big(\|x_\star - z_t\|_2^2 - \|x_\star - z_{t+1}\|_2^2 \Big) + \frac{\alpha}{2} \|\nabla f(x_{t+1})\|_2^2 + \frac{1-\tau}{\tau} (f(y_t) - f(x_{t+1})) \\
&\leq \frac{1}{2\alpha} \Big(\|x_\star - z_t\|_2^2 - \|x_\star - z_{t+1}\|_2^2 \Big) + \alpha L (f(x_{t+1}) - f(y_{t+1})) + \frac{1-\tau}{\tau} (f(y_t) - f(x_{t+1})),
\end{aligned}
$$

where we used the coupled gradient descent lemma in the second inequality. Now, if τ is chosen such that

$$
\frac{1-\tau}{\tau} = \alpha L, \qquad \text{that is,} \qquad \tau = \frac{1}{1+\alpha L},
$$

the previous inequality simplifies into the following result.

Theorem 2.2 (Coupling, Lemma 3.2 in [AO17]). Pick τ such that $\frac{1-\tau}{\tau} = \alpha L$. Then, at all times t,

$$
f(x_{t+1}) - f_\star \leq \frac{1}{2\alpha} \Big(\|x_\star - z_t\|_2^2 - \|x_\star - z_{t+1}\|_2^2 \Big) + \alpha L (f(y_t) - f(y_{t+1})).
$$

2.3 Putting the pieces together

We are now ready to perform the telescoping step that we saw in Lecture 7 using the new coupled variant of the Euclidean mirror descent lemma given in Theorem 2.2. Specifically, we can prove the following.

Theorem 2.3. Let α and τ be defined so that

$$
\alpha := \frac{\|x_\star - z_0\|_2}{\sqrt{2(f(x_0) - f_\star)}}, \qquad \tau := \frac{1}{1+\alpha L}.
$$

Then, Allen-Zhu and Orecchia's accelerated gradient descent finds at least one iterate x_t such that

$$
f(x_t) - f_\star \leq \frac{1}{2} (f(x_0) - f_\star)
$$

within

$$T_{\text{half}} := \frac{2\|x_\star - x_0\|_2 \sqrt{2L}}{\sqrt{f(x_0) - f_\star}}$$

iterations.

Proof. Averaging the inequality in Theorem 2.2 over $t = 0, 1, ..., T-1$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f_\star) &\leq \frac{1}{2T\alpha}\left(\|x_\star - z_0\|_2^2 - \|x_\star - z_T\|_2^2\right) + \frac{\alpha L}{T}(f(y_0) - f(y_T)) \\ &\leq \frac{1}{2T\alpha}\|x_\star - z_0\|_2^2 + \frac{\alpha L}{T}(f(y_0) - f_\star).\end{aligned}$$

Plugging in the proposed values of α and T_{half} verifies the statement. □

3 Final remarks

The idea of *momentum*, tracking the sum or average of all past gradients and using it when defining the next point, is extremely useful in machine learning. It extends well past gradient descent.

While the algorithm above is theoretically safe and interpolates the momentum term, in practice, people like to rewrite the update step to use momentum directly as

$$x_{t+1} \approx x_t - \eta g_t, \qquad \text{where} \quad g_t := \mu \sum_{s=1}^{t} (1-\mu)^{t-s} \nabla f(x_t).$$

This is essentially what goes on when you call `optim.SGD(nesterov=True)` from `pytorch`.

Bibliography

[Nes83] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$", *Proceedings of the USSR Academy of Sciences*, 1983.

[SBC16] W. Su, S. Boyd, and E. J. Candès, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights," *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016, [Online]. Available: `https://jmlr.org/papers/v17/15-084.html`

[KBB15] W. Krichene, A. Bayen, and P. L. Bartlett, "Accelerated Mirror Descent in Continuous and Discrete Time," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[WWJ16] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 47, p. E7351–E7358, Nov. 2016, doi: `10.1073/pnas.1614734113`.

[BLS15] S. Bubeck, Y. T. Lee, and M. Singh, "A geometric alternative to Nesterov's accelerated gradient descent."

[WA18] J.-K. Wang and J. D. Abernethy, "Acceleration through Optimistic No-Regret Dynamics," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[CST21] M. B. Cohen, A. Sidford, and K. Tian, "Relative Lipschitzness in Extragradient Methods and a Direct Recipe for Acceleration," in *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, 2021.

[AS22] K. Ahn and S. Sra, "Understanding Nesterov's Acceleration via Proximal Point Method," in *Symposium on Simplicity in Algorithms (SOSA)*, 2022, pp. 117–130.

[AO17] Z. Allen-Zhu and L. Orecchia, "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 2017.