Lecture 9–10: Accelerated Gradient Descent

Yudong Chen

In previous lectures, we showed that gradient descent achieves a $\frac{1}{k}$ convergence rate for smooth convex functions and a $(1-\frac{m}{L})^k$ geometric rate for L-smooth and m-strongly convex functions. Gradient descent is very greedy: it only uses the gradient $\nabla f(x_k)$ at the current point to choose the next iterate and discards information from past iterates.

It turns out we can do better than gradient descent, achieving a $\frac{1}{k^2}$ rate and a $\left(1-\sqrt{\frac{m}{L}}\right)^k$ rate in the two cases above. Both rates are optimal in a precise sense. The algorithms the attain these rates are known as *Nesterov's accelerated gradient descent (AGD)* or *Nesterov's optimal methods*.

1 Warm-up: the heavy-ball method

The high level idea of acceleration is adding momentum to the GD update. For example, consider the update

$$\begin{aligned} y_k &= x_k + \beta\left(x_k - x_{k-1}\right), && \text{momentum step} \\ x_{k+1} &= y_k - \alpha \nabla f(x_k), && \text{gradient step} \end{aligned}$$

where we first take a step in the direction $(x_k - x_{k-1})$, which is the momentum carried over from the previous update, and then take a standard gradient descent step. One can write these two steps in a single update:

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

This method is known as Polyak's *heavy-ball method*. The update above can be viewed as a discretization of the second order ODE $\ddot{x} = -a\nabla f(x) - b\dot{x}$, which models the motion of a body in a potential field given by f with friction given by b (hence the name heavy-ball).

It can be shown that for a strongly convex *quadratic* function f, the heavy-ball method achieves the accelerated rate $\left(1-\sqrt{\frac{m}{L}}\right)^k$.[1] For non-quadratic functions (e.g., those that are not twice differentiable), theoretical guarantees for heavy-ball method are less clear; in fact, heavy-ball may not even converge for such functions.

Rather than using the gradient at x_k, Nesterov's AGD uses the gradient at the point y_k *after* the momentum update:

$$\begin{aligned} y_k &= x_k + \beta\left(x_k - x_{k-1}\right), && \text{momentum step} \\ x_{k+1} &= y_k - \alpha \nabla f(y_k); && \text{"lookahead" gradient step} \end{aligned}$$

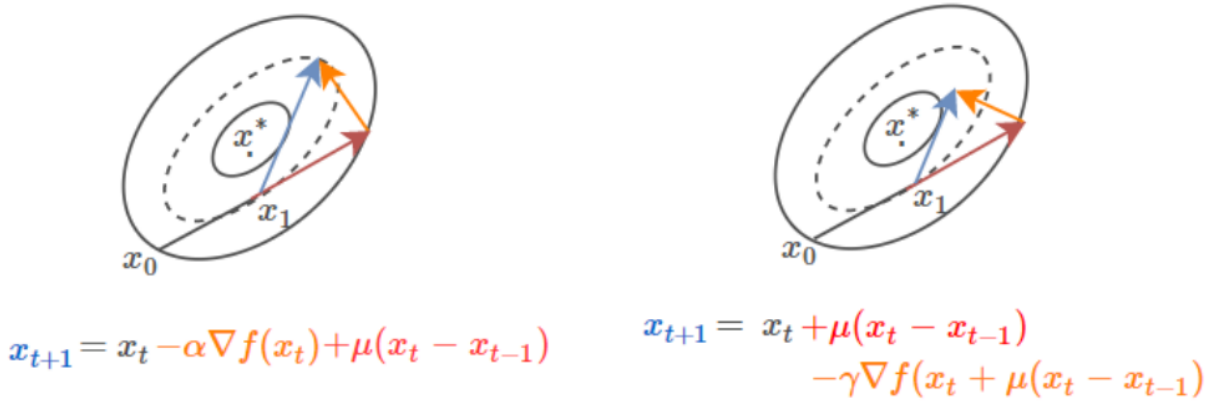
equivalently,

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

[1] This rate can be proved using direct eigenvalue analysis similar to that in Wright-Recht Chapter 4.2.

As we see below, Nesterov's AGD enjoys convergence guarantees for (strongly) convex functions beyond quadratic functions.

Below is an illustration of the updates of the heavy ball method and Nesterov's AGD:[2]


x^*
x_1
x_0
$x_{t+1} = x_t - \alpha\nabla f(x_t) + \mu(x_t - x_{t-1})$
x^*
x_1
x_0
$x_{t+1} = x_t + \mu(x_t - x_{t-1})$
$-\gamma\nabla f(x_t + \mu(x_t - x_{t-1})$

2 AGD for smooth and strongly convex f

Suppose f is m-strongly convex and L-smooth. Nesterov's AGD for minimizing f is given in Algorithm 1.

Algorithm 1 Nesterov's AGD, smooth and strongly convex

```
input: initial $x_0$, strong convexity and smoothness parameters $m, L$, number of iterations $K$
initialize: $x_{-1} = x_0$, $\alpha = \frac{1}{L}$, $\beta = \frac{\sqrt{L/m}-1}{\sqrt{L/m}+1}$.
for $k = 0, 1, \ldots K$
    $y_k = x_k + \beta\left(x_k - x_{k-1}\right)$
    $x_{k+1} = y_k - \alpha\nabla f(y_k)$
return $x_K$
```

Let x^* be the unique minimizer of f and set $f^* := f(x^*)$. By translation of coordinate, we may assume $x^* = 0$ without loss of generality (hence $x_k = x_k - x^*$ and $y_k = y_k - x^*$). Define $\kappa := \frac{L}{m}$ (condition number), $\rho^2 := 1 - \frac{1}{\sqrt{\kappa}}$ (contraction factor), $u_k := \frac{1}{L}\nabla f(y^k)$, and

$$V_k := f(x_k) - f^* + \frac{L}{2}\left\|x_k - \rho^2 x_{k-1}\right\|_2^2.$$

The quantity V_k, viewed a function of (x_k, x_{k-1}), is called a Lyapunov/potential function. We will show $V_{k+1} \le \rho^2 V_k$, hence geometric convergence.

By smoothness and strong convexity:

$$f(z) + \langle\nabla f(z), w - z\rangle + \frac{m}{2}\|w - z\|_2^2 \le f(w) \tag{1}$$

$$\le f(z) + \langle\nabla f(z), w - z\rangle + \frac{L}{2}\|w - z\|_2^2, \qquad \forall w, z \tag{2}$$

[2] Credit: Ioannis Mitliagkas' notes

It follows that

$$\begin{aligned}
V_{k+1} &= f(x_{k+1}) - f^* + \frac{L}{2}\left\|x_{k+1} - \rho^2 x_k\right\|_2^2 && \text{by defnition}\\
&\le f(y_k) - f^* + \langle Lu_k, x_{k+1} - y_k\rangle + \frac{L}{2}\|x_{k+1} - y_k\|_2^2 + \frac{L}{2}\left\|x_{k+1} - \rho^2 x_k\right\|_2^2 && \text{upper bound (2)}\\
&= f(y_k) - f^* - \frac{L}{2}\|u_k\|_2^2 + \frac{L}{2}\left\|x_{k+1} - \rho^2 x_k\right\|_2^2 && x_{k+1} - y_k = -u_k\\
&= \rho^2\Big[f(y_k) - f^* + L\langle u_k, x_k - y_k\rangle\Big] - \rho^2 L\langle u_k, x_k - y_k\rangle && \text{adding and subtracting terms}\\
&\quad + (1-\rho^2)\Big[f(y_k) - f^* - L\langle u_k, y_k\rangle\Big] + (1-\rho^2)L\langle u_k, y_k\rangle\\
&\quad - \frac{L}{2}\|u_k\|_2^2 + \frac{L}{2}\left\|x_{k+1} - \rho^2 x_k\right\|_2^2.
\end{aligned}$$

But

$$f(y_k) \le f(x_k) - L\langle u_k, x_k - y_k\rangle - \frac{m}{2}\|x_k - y_k\|_2^2 \qquad \text{lower bound (1) with } w = x_k, z = y_k$$

and

$$f(x^*) \ge f(y_k) - L\langle u_k, y_k\rangle + \frac{m}{2}\|y_k\|_2^2 \qquad \text{lower bound (1) with } w = x^* = 0, z = y_k.$$

Combining last three equations gives

$$\begin{aligned}
V_{k+1} &\le \rho^2\left[f(x_k) - f^* - \frac{m}{2}\|x_k - y_k\|_2^2\right] - \rho^2 L\langle u_k, x_k - y_k\rangle\\
&\quad - (1-\rho^2)\frac{m}{2}\|y_k\|_2^2 + (1-\rho^2)L\langle u_k, y_k\rangle\\
&\quad - \frac{L}{2}\|u_k\|_2^2 + \frac{L}{2}\left\|x_{k+1} - \rho^2 x_k\right\|_2^2\\
&= \rho^2\underbrace{\left[f(x_k) - f^* + \frac{L}{2}\left\|x_k - \rho^2 x_{k-1}\right\|_2^2\right]}_{V_k} + R_k,
\end{aligned}$$

where

$$\begin{aligned}
R_k := &-\rho^2\frac{m}{2}\|x_k - y_k\|_2^2 - (1-\rho^2)\frac{m}{2}\|y_k\|_2^2\\
&+ L\left\langle u_k, y_k - \rho^2 x_k\right\rangle - \frac{L}{2}\|u_k\|_2^2\\
&+ \frac{L}{2}\left\|x_{k+1} - \rho^2 x_k\right\|_2^2 - \frac{\rho^2 L}{2}\left\|x_k - \rho^2 x_{k-1}\right\|_2^2.
\end{aligned}$$

Claim 1. Under the choice of α, β and ρ above, we have

$$R_k = -\frac{1}{2}L\rho^2\left(\frac{1}{\kappa} + \frac{1}{\sqrt{\kappa}}\right)\|x_k - y_k\|_2^2 \le 0.$$

Proof. Substitute the definitions of $\alpha, \beta, \rho, x_{k+1}, y_k$ into the definition of R_k. (Verify it yourself!) □

It follows hat $V_{k+1} \le \rho^2 V_k, \forall k$, hence

$$\begin{aligned}
f(x_k) - f^* \le V_k &\le \rho^{2k} V_0 \\
&= \rho^{2k}\left(f(x_0) - f^* + \frac{L}{2}\left\|x_0 - \rho^2 x_0\right\|_2^2\right) && x_{-1} = x_0 \\
&= \rho^{2k}\left(f(x_0) - f^* + \frac{m}{2}\left\|x_0\right\|_2^2\right) && (1-\rho^2)^2 = \frac{1}{\kappa} = \frac{m}{L} \\
&= \rho^{2k}\left(f(x_0) - f^* + \frac{m}{2}\left\|x_0 - x^*\right\|_2^2\right) && x^* = 0 \qquad (3) \\
&\le \rho^{2k}\left(\frac{L}{2}\left\|x_0 - x^*\right\|^2 + \frac{m}{2}\left\|x_0 - x^*\right\|_2^2\right) && \text{upper bound (2)}, \nabla f(x^*) = 0 \\
&= \left(1 - \sqrt{\frac{m}{L}}\right)^k \cdot \frac{L+m}{2}\left\|x_0 - x^*\right\|_2^2. && \rho^2 = 1 - \sqrt{\frac{m}{L}}
\end{aligned}$$

We have established the following.

Theorem 1. *For Nesterov's AGD Algorithm 1 applied to m-strongly convex L-smooth f, we have*

$$f(x_k) - f^* \le \left(1 - \sqrt{\frac{m}{L}}\right)^k \cdot \frac{(L+m)\left\|x_0 - x^*\right\|_2^2}{2}, \qquad k = 0, 1, \ldots$$

(Iteration complexity bound) Equivalently, we have $f(x_k) - f^* \le \epsilon$ after at most

$$O\left(\sqrt{\frac{L}{m}} \log \frac{L\left\|x_0 - x^*\right\|_2^2}{\epsilon}\right) \textit{ iterations.}$$

Compare the above results with GD, which satisfies $f(x_k) - f^* = O\left(\left(1 - \frac{m}{L}\right)^k\right)$ and $k = O\left(\frac{L}{m}\log\frac{1}{\epsilon}\right)$. AGD improves by a factor of $\sqrt{\kappa} = \sqrt{\frac{L}{m}}$, which is significant for ill-conditioned problems with a large κ.

Example 1 (Ill-conditioned problems in statistical learning)**.** In statistical learning, we often need to minimize a function of the form

$$f(x) = g(x) + \frac{m}{2}\left\|x\right\|_2^2,$$

where g is a convex function corresponding to the empirical risk/training loss (e.g., the logistic regression loss) of a statistical model with parameter x, and $\frac{m}{2}\left\|x\right\|_2^2$ is called a regularizer. Often, g is *not* strongly convex, so the strong convexity of f comes from the regularizer. In many settings, the smoothness parameter of f is $O(1)$, and the regularization parameter is taken to be $m \propto \frac{1}{n}$, where n is the number of data points. The condition number $\kappa = \frac{L}{m} \propto n$ can be large in this case. We explore this setting in HW3.

3 AGD for smooth convex f

Suppose f is L-smooth, with a minimizer x^* and minimum value $f^* = f(x^*)$. (We dropped the strong convexity assumption.) Nesterov's AGD for such an f is given in Algorithm 2. Note that we allow the momentum parameter β_k to vary with k, and the parameter $\lambda_{k+1} \ge 0$ is chosen to satisfy $\lambda_{k+1}^2 - \lambda_{k+1} = \lambda_k^2$.

Algorithm 2 Nesterov's AGD, smooth convex

```
input: initial $x_0$, smoothness parameter $L$, number of iterations $K$
initialize: $x_{-1} = x_0$, $\alpha = \frac{1}{L}$, $\lambda_0 = 0$, $\beta_0 = 0$.
for $k = 0, 1, \ldots, K$
    $y_k = x_k + \beta_k (x_k - x_{k-1})$
    $x_{k+1} = y_k - \alpha \nabla f(y_k)$
    $\lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}$, $\beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}}$
return $x_K$
```

The Lyapunov function approach in the previous section can be adapted to analyze Algorithm 2; see Wright-Recht Chapter 4.4. Here we present a somewhat different proof.

Recall that a gradient step satisfies the descent property (Descent Lemma, Lecture 6 Lemma 1)

$$f(x_{k+1}) \le f(y_k) - \frac{1}{2L} \|\nabla f(y_k)\|_2^2 \le f(y_k). \tag{4}$$

Therefore, we have

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(x_{k+1}) - f(y_k) + f(y_k) - f(x_k) \\ &\le -\frac{1}{2L} \|\nabla f(y_k)\|_2^2 + \langle \nabla f(y_k), y_k - x_k \rangle && \text{descent property (4), convexity} \\ &= -\frac{L}{2} \|y_k - x_{k+1}\|_2^2 + L \langle y_k - x_{k+1}, y_k - x_k \rangle . && \nabla f(y_k) = L(y_k - x_{k+1}) \end{aligned} \tag{5}$$

Similarly:

$$\begin{aligned} f(x_{k+1}) - f(x^*) &= f(x_{k+1}) - f(y_k) + f(y_k) - f(x^*) \\ &\le -\frac{1}{2L} \|\nabla f(y_k)\|_2^2 + \langle \nabla f(y_k), y_k - x^* \rangle \\ &= -\frac{L}{2} \|y_k - x_{k+1}\|_2^2 + L \langle y_k - x_{k+1}, y_k - x^* \rangle . \end{aligned} \tag{6}$$

Define the optimality gap $\Delta_k := f(x_k) - f(x^*)$. Taking eq.(5)$\times\lambda_k(\lambda_k - 1)$+eq.(6)$\times\lambda_k$, we get

$$\lambda_k(\lambda_k - 1)(\Delta_{k+1} - \Delta_k) + \lambda_k \Delta_{k+1} \le L \langle y_k - x_{k+1}, \lambda_k(\lambda_k - 1)(y_k - x_k) + \lambda_k(y_k - x^*) \rangle - \frac{L}{2} \lambda_k^2 \|y_k - x_{k+1}\|_2^2 .$$

Rearranging terms gives the key inequality:

$$\lambda_k^2 \Delta_{k+1} - (\lambda_k^2 - \lambda_k)\Delta_k \le \frac{L}{2} \cdot \left[2 \langle \lambda_k(y_k - x_{k+1}), \lambda_k y_k - (\lambda_k - 1)x_k - x^* \rangle - \|\lambda_k(y_k - x_{k+1})\|_2^2 \right] . \tag{7}$$

As we show below, the parameters λ_k and β_k are chosen to make the LHS and RHS above telescope.

In particular, substituting $\lambda_k^2 - \lambda_k = \lambda_{k-1}^2$ into LHS of (7) and using the identity $2\langle a, b \rangle - \|a\|_2^2 = \|b\|_2^2 - \|b - a\|_2^2$ on RHS of (7), we obtain

$$\lambda_k^2 \Delta_{k+1} - \lambda_{k-1}^2 \Delta_k \le \frac{L}{2} \cdot \left[\|\lambda_k y_k - (\lambda_k - 1)x_k - x^*\|_2^2 - \|\lambda_k x_{k+1} - (\lambda_k - 1)x_k - x^*\|_2^2 \right] .$$

For the RHS, by definition and our choice of β_{k+1}, we have

$$\begin{aligned}
y_{k+1} = x_{k+1} + \beta_{k+1}\left(x_{k+1} - x_k\right) &= x_{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}}\left(x_{k+1} - x_k\right) \\
\Longleftrightarrow \lambda_{k+1} y_{k+1} - (\lambda_{k+1} - 1) x_{k+1} &= \lambda_k x_{k+1} - (\lambda_k - 1) x_k.
\end{aligned}$$

Combining the last two equations give

$$\lambda_k^2 \Delta_{k+1} - \lambda_{k-1}^2 \Delta_k \leq \frac{L}{2} \cdot \left[\left\|\lambda_k y_k - (\lambda_k - 1) x_k - x^*\right\|_2^2 - \left\|\lambda_{k+1} y_{k+1} - (\lambda_{k+1} - 1) x_{k+1} - x^*\right\|_2^2\right].$$

We sum the above inequalities over k. Note that both sides telescope and $\lambda_0 = 0, \lambda_1 = 1, \beta_1 = -1, y_1 = x_0$, hence

$$\begin{aligned}
\lambda_k^2 \Delta_{k+1} - \lambda_0^2 \Delta_1 &\leq \frac{L}{2}\left\|\lambda_1 y_1 - (\lambda_1 - 1) x_1 - x^*\right\|_2^2 \\
\Longrightarrow \lambda_k^2 \Delta_{k+1} &\leq \frac{L}{2}\left\|x_0 - x^*\right\|_2^2.
\end{aligned}$$

Finally, note that

$$\lambda_k \geq \frac{1 + \sqrt{4\lambda_{k-1}^2}}{2} = \lambda_{k-1} + \frac{1}{2},$$

which, together with $\lambda_1 = 1$, imply $\lambda_k \geq \frac{k+1}{2}, \forall k$. It follows that

$$f(x_{k+1}) - f(x^*) = \Delta_{k+1} \leq \frac{2L\left\|x_0 - x^*\right\|_2^2}{(k+1)^2}.$$

We have established the following.

Theorem 2. *For Nesterov's AGD Algorithm 2 applied to L-smooth convex f, we have*

$$f(x_k) - f(x^*) \leq \frac{2L\left\|x_0 - x^*\right\|_2^2}{k^2}, \qquad k = 0, 1, \ldots$$

(Iteration complexity bound) Equivalently, we have $f(x_k) - f^* \leq \epsilon$ after at most

$$O\left(\sqrt{\frac{L\left\|x_0 - x^*\right\|_2^2}{\epsilon}}\right) \text{ iterations.}$$

Compare with GD, which achieves $f(x_k) - f^* = O\left(\frac{1}{k}\right)$ and $k = O\left(\frac{L}{\epsilon}\right)$. Significant improvement by AGD.

4 Bibliographical notes (optional)

AGD was originally developed in Nesterov (1983). See Nesterov (2004) for a textbook convergence analysis of AGD using bounding functions.

The last decade has witnessed a surge of papers that provide alternative derivation, interpretation or analysis of AGD:

- The Lyapunov function approach in Section 2 follows Lessard et al (2016). In a related direction, Su, Boyd and Candes (2015) connect AGD to a certain second-order ODE. Also related in spirit is a paper by Flammarion and Bach (2015).
- The proof in Section 3 follows Beck and Teboulle (2009).
- Allen-Zhu and Orrechia (2014) view AGD as a linear coupling of GD and mirror descent.
- This blog post by Hardt (2013) relates AGD to Chebyshev polynomials.
- Bubeck et al (2015) provides a geometric perspective and a short proof.
- Diakonikolas and Orecchia (2019) develops the approximate duality gap technique, which can be applied to analyze AGD.

See d'Aspremont et al 2021 for a recent survey on acceleration methods including AGD and beyond.

Momentum/acceleration seems to be quite popular and effective in training neural networks, despite nonconvexity. Standard libraries like PyTorch typically implement (stochastic) gradient descent with the options of momentum and Nesterov acceleration.