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. ← Backend of GD.

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

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

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. This is known as Polyak's *heavy-ball method*. The update above is equivalent to 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).

$f(x) = \frac{1}{2}x^TAx + bx + c$

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:

At $k+1$'s step, use info from x_k, x_{k-1}

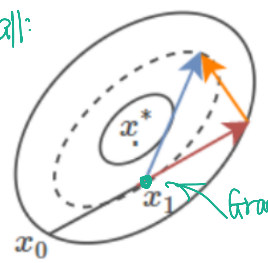
$$
\begin{aligned}
y_k &= x_k + \beta(x_k - x_{k-1}), && \text{momentum step}\\
x_{k+1} &= y_k - \alpha \nabla f(y_k). && \text{"lookahead" gradient step}
\end{aligned}
$$

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

"探头", toward $x_k - x_{k-1}$'s direction.

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

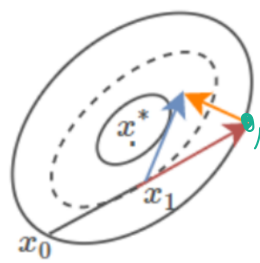
Heavy Ball:



Grad on this pt.

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

AGD:



Grad on this pt.

$$
\begin{aligned}
x_{t+1} = x_t &+ \mu(x_t - x_{t-1})\\
&- \gamma\nabla f(x_t + \mu(x_t - x_{t-1}))
\end{aligned}
$$

2. AGD for smooth, strongly cvx 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 $k = 0, 1, \dots K$

  $y_k = x_k + \beta(x_k - x_{k-1})$

  $x_{k+1} = y_k - \alpha \nabla f(y_k)$

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

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

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.

Smoothness & Strongly-convexity together ensures a bound for $f(w)$

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

$$V_{k+1} = f(x_{k+1}) - f^* + \frac{L}{2}\|x_{k+1} - \rho^2 x_k\|_2^2$$

$$\le f(y_k) - f^* + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \frac{L}{2}\|x_{k+1} - y_k\|_2^2 + \frac{L}{2}\|x_{k+1} - \rho^2 x_k\|_2^2$$

$x_{k+1} - y_k = -\frac{1}{L}\nabla f(y_k) = -u_k$

$$= f(y_k) - f^* - \langle Lu_k, u_k\rangle + \frac{L}{2}\|u_k\|^2 + \frac{L}{2}\|x_{k+1} - \rho^2 x_k\|_2^2$$

$$= f(y_k) - f^* - \frac{L}{2}\|u_k\|_2^2 + \frac{L}{2}\|x_{k+1} - \rho^2 x_k\|_2^2$$

$$= \rho^2\left(f(y_k) - f^* + L\langle u_k, x_k - y_k\rangle\right) - \rho^2 L\langle u_k, x_k - y_k\rangle$$

$$+ (1-\rho^2)\left(f(y_k) - f^* - L\langle u_k, y_k\rangle\right) + (1-\rho^2)L\langle u_k, y_k\rangle$$

$$- \frac{L}{2}\|u_k\|_2^2 + \frac{L}{2}\|x_{k+1} - \rho^2 x_k\|_2^2.$$

Plug $\Big\{$ $f(x_k) \ge f(y_k) + \langle Lu_k, x_k - y_k\rangle + \frac{m}{2}\|x_k - y_k\|_2^2 \Rightarrow f(y_k) \le f(x_k) - \langle Lu_k, x_k - y_k\rangle - \frac{m}{2}\|x_k - y_k\|_2^2.$

back. $f(x^*) \geq f(y_k) + \langle L u_k, x^* - y_k \rangle + \frac{m}{2}\|x^* - y_k\|_2^2 \Rightarrow f(y_k) - f(x^*) - L\langle u_k, y_k \rangle \leq -\frac{m}{2}\|y_k\|_2^2$

$$\leq \rho^2\left(f(x_k) - f^* - \frac{m}{2}\|x_k - y_k\|_2^2\right) - \rho^2 L\langle u_k, x_k - y_k \rangle$$
$$- \frac{m}{2}(1-\rho^2)\|y_k\|_2^2 + (1-\rho^2) L\langle u_k, y_k \rangle$$
$$- \frac{L}{2}\|u_k\|_2^2 + \frac{L}{2}\|x_{k+1} - \rho^2 x_k\|_2^2$$

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

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

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

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

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

pf: $x_{k+1} = y_k - u_k.$

$$\begin{cases} x_k - y_k = -\beta(x_k - x_{k-1}) \Rightarrow \|x_k - y_k\|^2 = \beta^2\|x_k - x_{k-1}\|^2 \\ x_{k+1} - \rho^2 x_k = y_k - u_k - \rho^2 x_k \end{cases}$$

$$R_k = -\rho^2\frac{m}{2}\beta^2\left(\|x_k\|^2 + \|x_{k-1}\|^2 + 2\langle x_k, x_{k-1}\rangle\right) + (\rho^2 - 1)\frac{m}{2}\|y_k\|^2 + L\langle u_k, y_k \rangle - L\rho^2\langle u_k, x_k \rangle$$
$$-\frac{L}{2}\|u_k\|^2 + \frac{L}{2}\left(\|y_k\|^2 + \|u_k\|^2 + \rho^4\|x_k\|^2 - 2\langle u_k, y_k \rangle - 2\rho^2\langle x_k, y_k \rangle + 2\rho^2\langle u_k, x_k \rangle\right)$$
$$-\frac{\rho^2 L}{2}\left(\|x_k\|^2 + \rho^4\|x_{k-1}\|^2 - 2\rho^2\langle x_k, x_{k-1}\rangle\right)$$

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

$\Rightarrow V_{k+1} \leq \rho^2 V_k.$

$\therefore f(x_k) - f^* \leq V_k \leq \rho^{2k} V_0$

$$= \rho^{2k}\left(f(x_0) - f^* + \frac{L}{2}\|x_0 - \rho^2 x_0\|_2^2\right)$$

$x_{-1} = x_0$

$$= \rho^{2k}\left(f(x_0) - f^* + \frac{m}{2}\|x_0\|_2^2\right)$$

$(1-\rho^2)^2 = \frac{1}{\kappa} = \frac{m}{L}$

$$= \rho^{2k}\left(f(x_0) - f(x^*) + \frac{m}{2}\|x_0 - x^*\|^2\right)$$

$x^* = 0.$

$f(x_0) - f(x^*) \le \langle \underbrace{\nabla f(x^*)}_{=0}, x_0 - x^* \rangle + \frac{L}{2}\|x_0 - x^*\|^2$

$$\le \rho^{2k} \cdot \frac{L+m}{2}\|x_0 - x^*\|^2$$

$$= \left(1 - \sqrt{\frac{m}{L}}\right)^k \cdot \frac{L+m}{2}\|x_0 - x^*\|_2^2$$

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

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

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

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

Recall 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.

$\kappa = \frac{L}{m}.$

3 AGD for smooth convex f

Suppose f is L-smooth, with a minimizer x^* and minimum value $f^* = f(x^*)$. 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 $\lambda_{k+1} \ge 0$ is chosen to satisfy $\lambda_{k+1}^2 - \lambda_{k+1} = \lambda_k^2$. $\Rightarrow \lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}$

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

Const l.r.
Adjusted momentum.

Analysis of Nesterov's AGD:

By descent lemma, $f(x_{k+1}) \le f(y_k) - \frac{\alpha}{2}\|\nabla f(y_k)\|_2^2 = f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|^2 \le f(y_k)$

$$\Rightarrow f(x_{k+1}) - f(x_k) = f(x_{k+1}) - f(y_k) + f(y_k) - f(x_k)$$

$$\leq -\frac{1}{2L}\|\nabla f(y_k)\|_2^2 + \langle \nabla f(y_k), y_k - x_k \rangle$$

$$= -\frac{L}{2}\|y_k - x_{k+1}\|_2^2 + L\langle y_k - x_{k+1}, y_k - x_k \rangle \quad ①$$

Cvx:

$$f(x_k) \geq f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle$$

$$\nabla f(y_k) = L(y_k - x_{k+1})$$

Similarly,

$$f(x_{k+1}) - f(x^*) = f(x_{k+1}) - f(y_k) + f(y_k) - f(x^*)$$

$$\leq -\frac{1}{2L}\|\nabla f(y_k)\|_2^2 + \langle \nabla f(y_k), y_k - x^* \rangle$$

$$= -\frac{L}{2}\|y_k - x_{k+1}\|_2^2 + L\langle y_k - x_{k+1}, y_k - x^* \rangle \quad ②$$

Let $\Delta_k := f(x_k) - f(x^*)$. Take ① $\times \lambda_k(\lambda_k - 1)$ + ② $\times \lambda_k$.

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

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

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

We show that the above λ_k, β_k are well chosen to make this inequality meaningful. Apply

$$\begin{cases} \lambda_k^2 - \lambda_k = \lambda_{k-1}^2 \\ 2\langle a, b\rangle - \|a\|^2 = \|b\|^2 - \|b - a\|^2 \end{cases}$$

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

Follow identity of β_k:

$$y_{k+1} = x_{k+1} + \beta_{k+1}(x_{k+1} - x_k) = x_{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}}(x_{k+1} - x_k)$$

$$\lambda_{k+1} y_{k+1} = \lambda_{k+1} x_{k+1} + (\lambda_k - 1)(x_{k+1} - x_k)$$

$$\lambda_{k+1} y_{k+1} - (\lambda_{k+1} - 1) x_{k+1} = \lambda_k x_{k+1} - (\lambda_k - 1) x_k$$

Plug in.

$$\lambda_k^2 \Delta_{k+1} - \lambda_{k-1}^2 \Delta_k \le \frac{L}{2}\Big[\underbrace{\|\lambda_k y_k - (\lambda_k - 1) x_k - x^*\|_2^2}_{a_k} - \underbrace{\|\lambda_{k+1} y_{k+1} - (\lambda_{k+1} - 1) x_{k+1} - x^*\|_2^2}_{a_{k+1}}\Big]$$

Do telescope sum. $\sum_{i=1}^{k}$ $(a_i - a_{i+1})$ $(a_1 - a_2)$

$$\lambda_k^2 \Delta_{k+1} - \lambda_0^2 \Delta_1 \le \frac{L}{2}\Big[\|\lambda_1 y_1 - (\lambda_1 - 1) x_1\|^2 - \|\lambda_{k+1} y_{k+1} - (\lambda_k - 1) x_{k+1} - x^*\|_2^2\Big]$$

$$\le \frac{L}{2}\Big[\|\lambda_1 y_1 - (\lambda_1 - 1) x_1\|_2^2\Big]$$

$\lambda_0 = 0$, $\lambda_1 = 1$, $\beta_1 = -1$, $y_1 = x_0$.

$$\lambda_k^2 \Delta_{k+1} \le \frac{L}{2}\|x_0\|^2 = \frac{L}{2}\|x_0 - x^*\|^2.$$

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

$$\Rightarrow \lambda_k \ge \frac{k}{2}$$

$$\Delta_{k+1} \le \frac{4}{k^2} \cdot \frac{L}{2}\|x_0 - x^*\|^2 = \frac{2L\|x_0 - x^*\|^2}{k^2}$$

$O(\frac{1}{k^2})$

Then Nesterov's AGD attains $f(x_{k+1}) - f(x^*) \le \varepsilon$

after $O\left(\sqrt{\frac{L\|x_0 - x^*\|^2}{\varepsilon}}\right)$ # iterations. $O(\frac{1}{\sqrt{\varepsilon}})$ iteration complexity.

Compare GD: $O(\frac{1}{k})$, $O(\frac{L}{\varepsilon})$.

Additional Proofs: Course note.