AGD recap. Recall that we have shown the following convergence guarantee for AGD in class:

$$f(\mathbf{y}_k) - f(\mathbf{x}^*) \le \frac{2L\,\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{(k+2)(k+3)}. \tag{1}$$

Restarting AGD. We can achieve the optimal convergence rate for smooth strongly convex functions by restarting the optimal algorithm (**AGD**) for smooth convex functions.

Assume that f is L-smooth and m-strongly convex. By strong convexity of f, we have

$$f(\mathbf{y}_k) \ge f(\mathbf{x}^*) + \underbrace{\langle \nabla f(\mathbf{x}^*), \mathbf{y}_k - \mathbf{x}^* \rangle}_{=0} + \frac{m}{2}\,\|\mathbf{y}_k - \mathbf{x}^*\|_2^2.$$

Combining this with Eq. (1) (where we use $(k+2)(k+3) \ge k^2$) and rearranging, we get

$$\begin{aligned}\|\mathbf{y}_k - \mathbf{x}^*\|_2^2 &\le \frac{4L\,\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{mk^2}\\ &\le \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2} \quad \text{when } k \ge \sqrt{\frac{8L}{m}}.\end{aligned}$$

This implies that the squared distance to optimality $\|\mathbf{y}_k - \mathbf{x}^*\|_2^2$ halves when we run **AGD** for $k = \left\lceil \sqrt{\frac{8L}{m}} \right\rceil$ iterations. Note that this requires us to know L and m.

Now consider a new algorithm $\mathcal{A}$ that restarts **AGD** every time $\|\mathbf{y}_k - \mathbf{x}^*\|_2^2$ halves.

Algorithm 1 Restarting AGD for smooth and convex functions ($\mathcal{A}$)

Input: $\mathbf{x}_0^{\text{out}} = \mathbf{x}_0$
1: **for** $k = 1$ to K **do**
2: $\quad \mathbf{x}_k^{\text{out}} = \mathbf{AGD}(\mathbf{x}_{k-1}^{\text{out}}, L, \left\lceil \sqrt{\frac{8L}{m}} \right\rceil)$ ▷ Restart of **AGD**
3: **end for**
4: **return** $\mathbf{x}_K^{\text{out}}$

Because we restart **AGD** every time it halves the squared distance to $\mathbf{x}^*$, we have:

$$\begin{aligned}\left\|\mathbf{x}_k^{\text{out}} - \mathbf{x}^*\right\|_2^2 &\le \frac{1}{2}\left\|\mathbf{x}_{k-1}^{\text{out}} - \mathbf{x}^*\right\|_2^2\\ &\le \left(\frac{1}{2}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2.\end{aligned}$$

Hence, to achieve ϵ-distance to the optimum $\mathbf{x}^*$, i.e., $\|\mathbf{x}_k^{\text{out}} - \mathbf{x}^*\| \le \epsilon$, the number of restarts in $\mathcal{A}$ needed is at most

$$K = \log_2\left(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{\epsilon^2}\right) = 2\log_2\left(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2}{\epsilon}\right).$$

And, in order to achieve $f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*) \le \tilde{\epsilon}$, the number of restarts in $\mathcal{A}$ needed is

$$K = O\Big(\log_2\left(\frac{L\,\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{\tilde{\epsilon}}\right)\Big),$$

by setting $\epsilon^2 = \frac{2}{L}\tilde{\epsilon}$ and using the property that f is L-smooth (which leads to $f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*) \le \frac{L}{2}\,\|\mathbf{x}_k^{\text{out}} - \mathbf{x}^*\|^2$).

Therefore, the total number of iterations ($\mathcal{A}$ + **AGD**) needed to achieve $f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*) \le \tilde{\epsilon}$ is

$$O\Big(\sqrt{\frac{L}{m}}\log_2\left(\frac{L\,\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{\tilde{\epsilon}}\right)\Big).$$

Notice that $\mathcal{A}$ achieves faster convergence than the basic descent methods for smooth strongly convex functions that we analyzed in previous lectures.