到第 k

Up to iteration k,

query points:

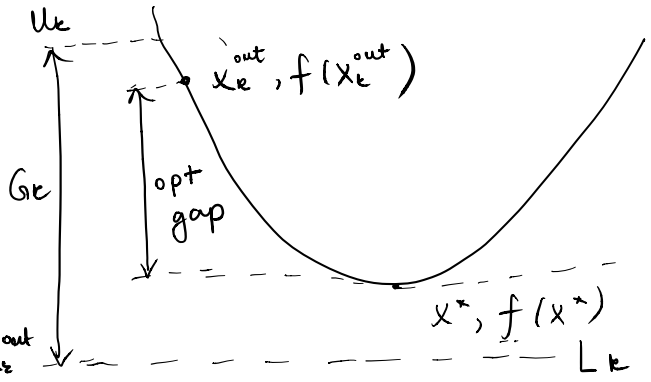
$x_0, x_1, \dots, x_k$

gradient answers

$\nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_k)$

Based on $(x_i, \nabla f(x_i))_{i=0}^{k}$,

we choose an output point x_k^{out}



* Gradient descent (GD):
  - greedy, does not use history.

$x \to$ [1st order oracle] $\to \nabla f(x), f(x)$

* To do better than GD, we need to use convexity somehow

- $L_k \le f(x^*)$; $U_k \ge f(x_k^{out})$

$f(x_k^{out}) - f(x^*) \le G_k = U_k - L_k$

Goal: show that for some positive & strictly increasing function A_k, we have:

$$\boxed{A_k G_k \le A_{k-1} G_{k-1}, \ \forall k}.$$

$$\Rightarrow A_k G_k \le A_0 G_0$$

$$f(x_k^{out}) - f(x^*) \le G_k \le \frac{A_0 G_0}{A_k}.$$

* Lower bound construction:

$$f(x^*) \ge \frac{1}{A_k} \sum_{i=0}^{k} a_i \left(f(x_i) + \langle \nabla f(x_i), x^* - x_i \rangle \right), \quad a_i > 0, \ \forall i$$

$$A_k = \sum_{i=0}^{k} a_i$$

Attempt 1:

$$f(x^*) \ge \frac{1}{A_k} \min_{x \in \mathbb{R}^d} \left\{ \sum_{i=0}^{k} a_i \left(f(x_i) + \langle \nabla f(x_i), x - x_i \rangle \right) \right\}$$

unless $\sum_{i=0}^{k} \nabla f(x_i) = 0$, rhs $= -\infty$.

Attempt 2:

$$f(x^*) \geq \frac{1}{A_k}\sum_{i=0}^{k} a_i\left(f(x_i) + \langle \nabla f(x_i), x^* - x_i\rangle\right)$$
$$+\frac{L}{2A_k}\|x^* - x_0\|_2^2 - \frac{L}{2A_k}\|x^* - x_0\|_2^2 .$$
$$\geq \frac{1}{A_k}\sum_{i=0}^{k} a_i f(x_i) - \frac{L}{2A_k}\|x^* - x_0\|_2^2$$
$$+ \frac{1}{A_k}\min_{x\in\mathbb{R}^d}\underbrace{\left\{\sum_{i=0}^{k} a_i\langle \nabla f(x_i), x - x_i\rangle + \frac{L}{2}\|x - x_0\|_2^2\right\}}_{m_k(x)}$$
$$=: L_k$$

$A_k L_k - A_{k-1}L_{k-1}$, independent of x^*.

Let $v_k = \operatorname{argmin}_{x\in\mathbb{R}^d} m_k(x) = x_0 - \frac{1}{L}\sum_{i=0}^{k} a_i \nabla f(x_i)$

$$A_k L_k - A_{k-1}L_{k-1} = a_k f(x_k) + m_k(v_k) - m_{k-1}(v_{k-1})$$

$$m_k(v_k) = a_k\langle \nabla f(x_k), v_k - x_k\rangle + m_{k-1}(v_k)$$
$$= a_k\langle \nabla f(x_k), v_k - x_k\rangle + \frac{L}{2}\|\underbrace{v_k - v_{k-1}}_{-\frac{1}{L}a_k\nabla f(x_k)}\|_2^2 + m_{k-1}(v_{k-1}).$$

$$A_k L_k - A_{k-1}L_{k-1} = a_k f(x_k) + a_k\langle \nabla f(x_k), v_k - x_k\rangle + \frac{a_k^2}{2L}\|\nabla f(x_k)\|_2^2 .$$

* Upper bound:

u_k $\quad u_k = f(x_k^{out})$ $\quad$ Denote $y_k = x_k^{out}$ $\quad y_k = x_k^{out}$

$$= f(y_k)$$
$$A_k u_k - A_{k-1}u_{k-1} = A_k f(y_k) - A_{k-1} f(y_{k-1}) - \underbrace{a_k}_{A_k - A_{k-1}} f(x_k) + a_k f(x_k)$$
$$= A_k\left(f(y_k) - f(x_k)\right) + A_{k-1}\left(f(x_k) - f(y_{k-1})\right) + a_k f(x_k).$$

$G_k = U_k - L_k$ $\quad A_k(U_k - L_k) - A_{k-1}(U_{k-1} - L_{k-1})$

$= (A_k U_k - A_{k-1} U_{k-1}) - (A_k L_k - A_{k-1} L_{k-1})$

* The approximate gap:

$$A_k G_k - A_{k-1} G_{k-1} = A_k\big(f(y_k) - f(x_k)\big) + A_{k-1}\big(f(x_k) - f(y_{k-1})\big)$$
$$- a_k \langle \nabla f(x_k), v_k - x_k \rangle - \frac{a_k^2}{2L} \|\nabla f(x_k)\|_2^2$$

$$\langle \nabla f(x_k), v_k - x_k \rangle = \langle \nabla f(x_k), \underbrace{v_k - v_{k-1}}_{-\frac{1}{L} a_k \nabla f(x_k)} \rangle + \langle \nabla f(x_k), v_{k-1} - x_k \rangle$$

$$= -\frac{a_k}{L} \|\nabla f(x_k)\|_2^2 + \langle \nabla f(x_k), v_{k-1} - x_k \rangle$$

$$A_k G_k - A_{k-1} G_{k-1} = A_k\big(f(y_k) - f(x_k)\big) + A_{k-1}\big(f(x_k) - f(y_{k-1})\big)$$
$$- a_k \langle \nabla f(x_k), v_{k-1} - x_k \rangle + \frac{a_k^2}{2L} \|\nabla f(x_k)\|_2^2.$$

Take: $\boxed{y_k = x_k - \frac{1}{L} \nabla f(x_k)}$ $\Rightarrow f(y_k) - f(x_k) \le -\frac{1}{2L} \|\nabla f(x_k)\|_2^2.$

Want: $A_{k-1}\big(f(x_k) - f(y_{k-1})\big) - a_k \langle \nabla f(x_k), v_{k-1} - x_k \rangle \le 0.$

$$\overset{\text{convexity of } f}{\le} A_{k-1} \langle \nabla f(x_k), x_k - y_{k-1} \rangle - a_k \langle \nabla f(x_k), v_{k-1} - x_k \rangle$$

$$= \langle \nabla f(x_k), \underbrace{A_k x_k - A_{k-1} y_{k-1} - a_k v_{k-1}}_{=0} \rangle$$

$$\boxed{x_k = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_k}{A_k} v_{k-1}}$$

$$A_k G_k - A_{k-1} G_{k-1} \le \|\nabla f(x_k)\|_2^2 \underbrace{\left(-\frac{A_k}{2L} + \frac{a_k^2}{2L}\right)}_{\le 0}.$$

$$\boxed{\frac{a_k^2}{A_k} \le 1}$$

* Algorithm: Start with $x_0 \in \mathbb{R}^d$; for $k \geq 1$:

$$x_k = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_k}{A_k} v_{k-1}$$

$$v_k = v_{k-1} - \frac{1}{L} a_k \nabla f(x_k)$$

$$y_k = x_k - \frac{1}{L} \nabla f(x_k)$$

where $\frac{a_k^2}{A_k} \leq 1$, $A_k = \sum_{i=0}^{k} a_i$.

} Nesterov's accelerated gradient descent (AGD)

By construction, we have already proved:

$$f(y_k) - f(x^*) \leq G_k \leq \frac{A_0 G_0}{A_k}.$$

1) Bound $A_0 G_0$. $y_0 = x_0 - \frac{1}{L} \nabla f(x_0)$

$A_0 G_0 \leq \frac{A_0 L}{2} \|x^* - x_0\|_2^2$. (verify as an exercise)

2) Choose "the best" A_k.

$a_0 = A_0 = 1$.

$\frac{a_k^2}{A_k} = 1 \Rightarrow A_k = \Theta(k^2)$

$a_k \approx k^p$, $A_k \sim k^{p+1}$, $p > 0$

An alternative choice:

$\bar{a}_i = \frac{i+1}{2}$

$\bar{A}_k = \sum_{i=0}^{k} \bar{a}_i = \frac{k(k+1)}{4} = \Theta(k^2)$

$$\boxed{\frac{\bar{a}_i^2}{\bar{A}_i} \leq 1}$$

Putting everything together:

$$f(y_k) - f(x^*) = O\left(\frac{L \|x^* - x_0\|_2^2}{k^2}\right).$$

$f(y_k) - f(x^*) \le \varepsilon$ after $k = O\left(\sqrt{\frac{L}{\varepsilon}}\, \|x^* - x_0\|_2\right)$ iterations

$\boxed{y_k = AGD(x_0, L, k)}$

* Suppose f was m-strongly convex (in addition to being L-smooth)

$\forall x, y: \; f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|_2^2$

$y = y_k, \; x = x^*$

$$\underbrace{C}_{\text{absolute constant}} \cdot \frac{L\|x_0 - x^*\|_2^2}{k^2} \ge f(y_k) - f(x^*) \ge \underbrace{\langle \nabla f(x^*), y_k - x^* \rangle}_{0} + \frac{m}{2}\|y_k - x^*\|_2^2$$

$$\|y_k - x^*\|_2^2 \le \underbrace{2 \cdot C \cdot \frac{L\|x_0 - x^*\|_2^2}{m\,k^2}}_{\text{fixed}}$$

In particular, $\|y_k - x^*\|_2^2 \le \frac{1}{2}\|x_0 - x^*\|_2^2$

for $k \ge 2\sqrt{C} \cdot \sqrt{\frac{L}{m}} = O\left(\sqrt{\frac{L}{m}}\right)$.

Consider the following algorithm:

(A) $x_0^{out} = x_0$

for $i = 1 : k$

$x_i^{out} = AGD\left(x_{i-1}^{out}, L, 2\sqrt{C}\sqrt{\frac{L}{m}}\right)$

$$\|x_{i+1}^{out} - x^*\|_2^2 \le \frac{1}{2}\|x_i^{out} - x^*\|_2^2$$
$$\le \left(\frac{1}{2}\right)^2 \|x_{i-1}^{out} - x^*\|_2^2$$
$$\vdots$$
$$\le \left(\frac{1}{2}\right)^{i+1} \|x_0 - x^*\|_2^2 \le \varepsilon^2$$

$$\| x_{i+1}^{out} - x^* \|_2^2 \leq \varepsilon^2$$

for $i+1 = \log_2 \left(\frac{\| x_0 - x^* \|_2^2}{\varepsilon^2} \right)$

total # steps / iterations: $O\left(\sqrt{\frac{L}{m}} \log \left(\frac{\| x_0 - x^* \|_2}{\varepsilon} \right) \right)$