

Ter23 K

Up to iteration k ,
query points:

$$x_0, x_1, \dots, x_k$$

gradient answers

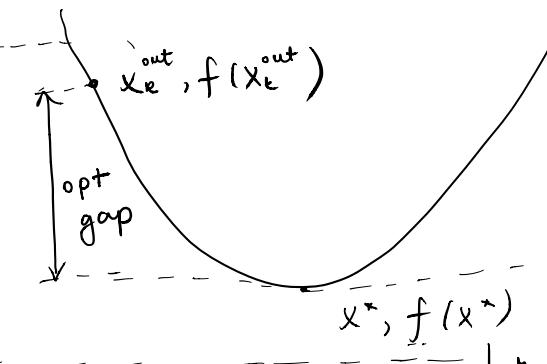
$$\nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_k)$$

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



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

$$L_k \leq f(x^*) ; \quad G_k \geq f(x_k^{out})$$

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

Goal: show that for some **positive** & **strictly increasing**

function A_k , we have:

$$A_k G_k \leq A_{k-1} G_{k-1} + L_k$$

$$\Rightarrow A_k G_k \leq A_0 G_0$$

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

* Lower bound construction:

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

Attempt 1:

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

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

Attempt 2:

$$\begin{aligned}
 \underline{f(x^*)} &\geq \frac{1}{A_k} \sum_{i=0}^k \alpha_i \left(f(x_i) + \langle \nabla f(x_i), x^* - x_i \rangle \right) \\
 &\quad + \underbrace{\left(\frac{L}{2A_k} \right) \|\underline{x^* - x_0}\|_2^2}_{\text{---}} - \underbrace{\frac{L}{2A_k} \|x^* - x_0\|_2^2}_{\text{---}} \\
 &\geq \frac{1}{A_k} \sum_{i=0}^k \alpha_i f(x_i) - \frac{L}{2A_k} \|x^* - x_0\|_2^2 \\
 &\quad + \frac{1}{A_k} \min_{x \in \mathbb{R}^d} \left\{ \sum_{i=0}^k \alpha_i \langle \nabla f(x_i), x - x_i \rangle + \frac{L}{2} \|x - x_0\|_2^2 \right\} \\
 &=: L_k
 \end{aligned}$$

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

$$\text{Let } v_k = \arg \min_{x \in \mathbb{R}^d} m_k(x) = x_0 - \frac{1}{L} \sum_{i=0}^k \alpha_i \nabla f(x_i)$$

$$A_k L_k - A_{k-1} L_{k-1} = \alpha_k f(x_k) + \underbrace{m_k(v_k) - m_{k-1}(v_{k-1})}_{\text{---}}$$

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

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

* Upper bound:

u_k

$$\begin{aligned}
 u_k &= f(x_k^{\text{out}}) \\
 &= f(y_k)
 \end{aligned}$$

Denote $y_k = x_k^{\text{out}}$

$$y_k = x_k^{\text{out}}$$

$$\begin{aligned}
 A_k u_k - A_{k-1} u_{k-1} &= \underbrace{A_k f(y_k)}_{\text{---}} - \underbrace{A_{k-1} f(y_{k-1})}_{\text{---}} - \underbrace{\alpha_k f(x_k)}_{\text{---}} + \underbrace{\alpha_k f(x_k)}_{\text{---}} \\
 &= \underbrace{A_k (f(y_k) - f(x_k))}_{\text{---}} + \underbrace{A_{k-1} (f(x_k) - f(y_{k-1}))}_{\text{---}} + \underbrace{\alpha_k f(x_k)}_{\text{---}}
 \end{aligned}$$

$$G_{ik} = \mathcal{M}_{ik} - b_k$$

* The approximate gap:

$$\underline{A_k G_{ik}} - \underline{A_{k-1} G_{ik-1}} = \underline{A_k (f(y_k) - f(x_k))} + \underline{A_{k-1} (f(x_k) - f(y_{k-1}))}$$

$$- \underbrace{\alpha_k \langle \nabla f(x_k), v_k - x_k \rangle}_{\frac{\alpha_k^2}{2L} \| \nabla f(x_k) \|_2^2}$$

$$\begin{aligned} \langle \nabla f(x_k), v_k - x_k \rangle &= \underbrace{\langle \nabla f(x_k), v_k - v_{k-1} \rangle}_{-\frac{1}{L} \alpha_k \nabla f(x_k)} + \langle \nabla f(x_k), v_{k-1} - x_k \rangle \\ &= -\frac{\alpha_k}{L} \| \nabla f(x_k) \|_2^2 + \underbrace{\langle \nabla f(x_k), v_{k-1} - x_k \rangle}_{\text{---}} \end{aligned}$$

$$\begin{aligned} \underline{A_k G_{ik}} - \underline{A_{k-1} G_{ik-1}} &= \underline{A_k (f(y_k) - f(x_k))} + \underline{A_{k-1} (f(x_k) - f(y_{k-1}))} \\ &- \alpha_k \langle \nabla f(x_k), v_{k-1} - x_k \rangle + \underbrace{\left(\frac{\alpha_k^2}{2L} \| \nabla f(x_k) \|_2^2 \right)}_{\text{---}}. \end{aligned}$$

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

Want: $\underbrace{A_{k-1} (f(x_k) - f(y_{k-1}))}_{\text{convexity of } f} - \alpha_k \langle \nabla f(x_k), v_{k-1} - x_k \rangle \leq 0.$

$$\begin{aligned} &\leq A_{k-1} \langle \nabla f(x_k), x_k - y_{k-1} \rangle - \alpha_k \langle \nabla f(x_k), v_{k-1} - x_k \rangle \\ &= \langle \nabla f(x_k), \underbrace{A_k x_k - A_{k-1} y_{k-1} - \alpha_k v_{k-1}}_{= 0} \rangle \\ &\boxed{x_k = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{\alpha_k}{A_k} v_{k-1}} \end{aligned}$$

$$\underline{A_k G_{ik}} - \underline{A_{k-1} G_{ik-1}} \leq \| \nabla f(x_k) \|_2^2 \left(-\frac{A_k}{2L} + \frac{\alpha_k^2}{2L} \right) \leq 0.$$

$$\boxed{\frac{\alpha_k^2}{A_k} \leq 1}$$

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

$$\begin{aligned} x_k &= \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_k}{A_k} v_{k-1} \\ v_k &= v_{k-1} - \frac{1}{L} a_k \nabla f(x_k) \\ y_k &= x_k - \frac{1}{L} \nabla f(x_k) \end{aligned} \quad \left. \begin{array}{l} \text{where } \frac{a_k^2}{A_k} \leq 1, \quad A_k = \sum_{i=0}^k a_i \\ \text{Nesterov's accelerated gradient descent (AGD)} \end{array} \right\}$$

By construction, we have already proved:

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

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

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

2) Choose "the best" \underline{A}_k .

$$a_0 = A_0 = 1.$$

$$\frac{\underline{a}_k^2}{A_k} \leq 1 \Rightarrow \underline{A}_k = \Theta(k^2)$$

$\underline{a}_k \approx k^p, \underline{A}_k \approx k^{p+1}, p > 0$

An alternative choice:

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

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

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

Putting everything together:

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

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

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

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

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

$$\begin{aligned} y &= y_k, \quad x = x^* \\ \underbrace{C \cdot \frac{L \|x_0 - x^*\|_2^2}{k^2}}_{\text{absolute constant}} &\geq f(y_k) - f(x^*) \geq \langle \nabla f(x^*), y_k - x^* \rangle + \frac{m}{2} \|y_k - x^*\|_2^2 \end{aligned}$$

$\overset{0}{\underset{\text{fixed}}{\textcircled{O}}}$

$$\|y_k - x^*\|_2^2 \leq \underbrace{2 \cdot C \cdot \frac{L \|x_0 - x^*\|_2^2}{m k^2}}$$

In particular, $\|y_k - x^*\|_2^2 \leq \frac{1}{2} \|x_0 - x^*\|_2^2$
for $k \geq 2\sqrt{C} \cdot \sqrt{\frac{L}{m}} = O\left(\sqrt{\frac{L}{m}}\right)$.

Consider the following algorithm:

$$\begin{aligned} (A) \quad x_0^{\text{out}} &= x_0 \\ \text{for } i = 1 : k \\ x_i^{\text{out}} &= \underbrace{\text{AGD}(x_{i-1}^{\text{out}}, L, \underbrace{2\sqrt{C}\sqrt{\frac{L}{m}}}_{\text{fixed}})} \end{aligned}$$

$$\begin{aligned} \|x_{i+1}^{\text{out}} - x^*\|_2^2 &\leq \frac{1}{2} \|x_i^{\text{out}} - x^*\|_2^2 \\ &\leq \left(\frac{1}{2}\right)^i \|x_0^{\text{out}} - x^*\|_2^2 \\ &\leq \left(\frac{1}{2}\right)^{i+1} \|x_0 - x^*\|_2^2 \leq \epsilon^2 \end{aligned}$$

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

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

$$\text{total \# steps / iterations: } \underline{\underline{O\left(\sqrt{L_m} \log\left(\frac{\|x_0 - x^*\|_2}{\epsilon}\right)\right)}}$$