

* Other basic descent methods:

There are other descent methods for which you can guarantee:

$$f(x_{k+1}) \leq f(x_k) - \frac{\beta}{2} \|\nabla f(x_k)\|_2^2 \text{ for some } \beta > 0.$$

* Examples:

1) Preconditioned methods:

$$x_{k+1} = x_k - \alpha S_k \nabla f(x_k), \text{ where } S_k \text{ is a PD matrix w/ eigenvalues in } [\mu_1, \mu_2] \quad 0 < \mu_1 < \mu_2 < \infty.$$

From Lemma 2.2:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 \\ &= f(x_k) - \alpha \underbrace{\langle S_k \nabla f(x_k), \nabla f(x_k) \rangle}_{\geq \mu_1 \|\nabla f(x_k)\|_2^2} \\ &\quad + \frac{L}{2} \alpha^2 \underbrace{\|S_k \nabla f(x_k)\|_2^2}_{\leq \mu_2^2 \|\nabla f(x_k)\|_2^2} \\ &\leq f(x_k) - \underbrace{(\alpha \mu_1 - \frac{L}{2} \mu_2^2 \alpha^2)}_{> 0 \text{ for suff. small } \alpha} \|\nabla f(x_k)\|_2^2. \end{aligned}$$

Newton's method uses $S_k = (\nabla^2 f(x_k))^{-1}$; need $\nabla^2 f(x_k)$ to have positive evals for this work

2) Gauss-Southwell (greedy coordinate descent)

$$x_{k+1} = x_k - \alpha \underbrace{\nabla_i f(x_k) e_{ik}}_{-p_k}, \quad e_{ik} = [0, 0, \dots, \underset{i \text{ position}}{1}, \dots, 0]$$

* $i_k = \arg \max_{1 \leq i \leq n} |\nabla_i f(x_k)|$

$$\|p_k\|_2 \geq \frac{1}{\alpha} \|\nabla f(x_k)\|_2$$

3) Randomized coordinate descent (HW #2)

4) Stochastic gradient descent, where

$$p_k = -g(x_k, \bar{z}_k), \quad \mathbb{E}_{\bar{z}_k} [g(x_k, \bar{z}_k)] = \nabla f(x_k)$$

i.i.d. r.v.

$$x_{k+1} = x_k + \alpha p_k,$$

under certain assumptions.

* Convergence of basic descent methods: y2/4.4

* Assume:

$$f(x_{k+1}) \leq f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|_2^2.$$

① Nonconvex case:

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|_2 \leq \sqrt{\frac{2(f(x_0) - f_*)}{\alpha(k+1)}},$$

where $f(x) \geq f_* > -\infty, \forall x.$

② Convex f .

Convexity:

$$\forall x: f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle$$

$$f(x_{k+1}) - f(x^*) \leq \underline{G_k}$$

optimality gap estimate

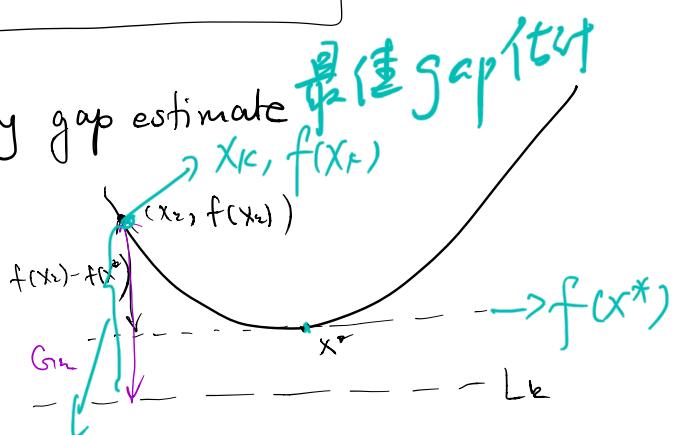
最佳 gap 估計

$$f(x^*) \geq \underline{L_k}$$

lower bound

Let Δ_k be a strictly increasing positive numbers.

严格增正序列



$$\underline{G_k} = f(x^k) - L_k$$

便 $A_k G_k$ は?

Goal:

$$A_k G_k - A_{k-1} G_{k-1} \leq \widehat{E}_k \quad \text{"error"}$$

Sum →

$$A_k G_k \leq A_0 G_0 + \sum_{i=1}^k E_i$$

$$f(x_{k+1}) - f(x^*) \leq G_k$$

$$\frac{A_0 G_0}{A_k} + \frac{\sum_{i=1}^k E_i}{A_k}$$

want if grow slowly compared to A_k

$$\frac{\sum E_i}{A_k} \rightarrow 0$$

Lower bound on $f(x^*)$:

$$f(x^*) \geq f(x_i) + \langle \nabla f(x_i), x^* - x_i \rangle$$

α_i 是正数序列 / A_k 是 α_i 的和

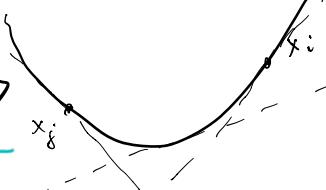
Let $\{\alpha_i\}_{i \geq 0}$ be a sequence of positive numbers

$$\text{s.t. } A_k = \sum_{i=0}^k \alpha_i \text{ so that } \frac{1}{A_k} \sum_{i=0}^k \alpha_i = 1.$$

$$\underbrace{\sum_{i=0}^k \alpha_i f(x^*)}_{A_k} \geq \sum_{i=0}^k \alpha_i (\langle \nabla f(x_i), x^* - x_i \rangle + f(x_i)) \quad L_k$$

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

$$A_k L_k - A_{k-1} L_{k-1} = \underbrace{\alpha_k f(x_k)}_{+ \alpha_k \langle \nabla f(x_k), x^* - x_k \rangle}$$



$$G_k = f(x_{k+1}) - L_k$$

$$A_k G_k - A_{k-1} G_{k-1} = \underbrace{A_k f(x_{k+1}) - A_{k-1} f(x_k)}_{\alpha_k f(x_k) - \alpha_k \langle \nabla f(x_k), x^* - x_k \rangle}$$

$$\begin{aligned} &= A_k (f(x_{k+1}) - f(x_k)) \\ &\quad - \alpha_k \langle \nabla f(x_k), x^* - x_k \rangle \end{aligned}$$

$$f(x_{k+1}) - f(x_k) \leq -\frac{\alpha}{2} \|\nabla f(x_k)\|_2^2$$

"Descent Lemma"

$$\leq -\frac{A_k \alpha}{2} \|\nabla f(x_k)\|_2^2$$

$$-\underbrace{\alpha_k}_{\text{Cauchy Schur neg}} \langle \nabla f(x_k), x^* - x_k \rangle$$

C.S.

$$\leq -\frac{A_k \alpha}{2} \|\nabla f(x_k)\|_2^2 + \underbrace{\alpha_k \|\nabla f(x_k)\|_2}_{\text{C.S.}} \underbrace{\|x^* - x_k\|_2}_{\text{C.S.}}$$

Useful inequality: $-\frac{p^2}{2} + pq \leq \frac{q^2}{2}$

$$p = \sqrt{\alpha A_k} \|\nabla f(x_k)\|_2, q = \frac{\alpha_k}{\sqrt{\alpha A_k}} \|x^* - x_k\|_2$$

$$\underbrace{A_k G_k - A_{k-1} G_{k-1}}_2 \leq \frac{\alpha_k^2}{2\alpha A_k} \|x^* - x_k\|_2^2 = E_k$$

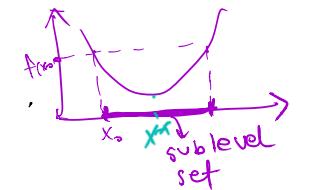
$$\text{Ex. } A_0 G_0 \leq \frac{\alpha_0}{2\alpha A_0} \|x^* - x_0\|_2^2$$

$$\Rightarrow f(x_{k+1}) - f(x^*) \leq G_k \leq \sum_{i=0}^k \frac{\alpha_i^2}{2\alpha A_i} \|x^* - x_k\|_2^2.$$

Define $R = \max \{ \|x^* - x\|_2 : f(x) \leq f(x_0) \}$

sublevel set: $\{ x : f(x) \leq f(x_0) \}$.

$$\Rightarrow f(x_{k+1}) - f(x^*) \leq \frac{R^2}{2\alpha A_k} \sum_{i=0}^k \frac{\alpha_i^2}{A_i}$$



How should we choose α_i ($A_i = \sum_{j=0}^i \alpha_j$) ?

$\alpha_i \propto i^p$, $p > 0$, for k large enough $A_k \propto \frac{k^{p+1}}{p+1}$.

(for integer p , Faulhaber's formula)

One particular choice

$$\alpha_i = \frac{i+1}{2}, \quad A_i = \frac{(i+1)(i+2)}{4}$$

$$\frac{\alpha_i^2}{A_i} \leq 1$$

$$\Rightarrow f(x_{k+1}) - f(x^*) \leq \frac{(2) R^2}{\alpha(k+2)}.$$

* Can we do better?

No, unless we say something more about the method we are using.

E.g., for gradient descent, we can replace

R by $\|x_0 - x^*\|_2$.

* Strongly convex case:

$$G_k = f(x_{k+1}) - L_k$$

$$L_k \leq f(x^*)$$

$$f(x_{k+1}) - f(x^*) \leq G_k$$

Find a fast growing sequence of positive numbers A_k s.t.

$A_k G_k$ is either non-increasing, or it increases

"slowly" compared to A_k .

$$\text{Goal: } \underbrace{A_k G_k - A_{k-1} G_{k-1}}_{\geq 0} \leq 0. *$$

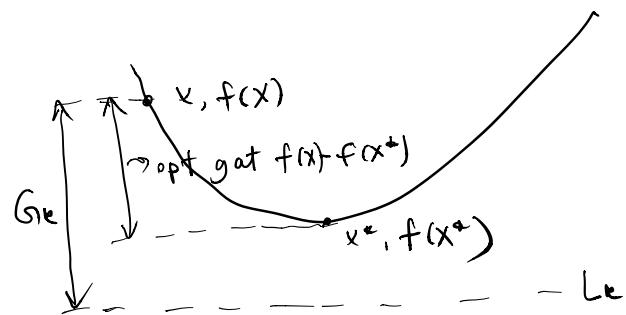
$$\Rightarrow f(x_{k+1}) - f(x^*) \leq \frac{A_0 G_0}{A_k}$$

Function f is m -strongly convex for some $m > 0$.
 $+ i :$

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

$$f(x^*) \geq \frac{1}{A_k} \sum_{i=0}^k a_i \left(f(x_i) + \langle \nabla f(x_i), x^* - x_i \rangle + \frac{m}{2} \|x^* - x_i\|_2^2 \right),$$

$$\frac{1}{A_k} \cdot \sum a_i (f(x_i) + \langle \nabla f(x_i), x^* - x_i \rangle + \frac{m}{2} \|x^* - x_i\|_2^2) = L_k$$



where $a_i > 0$, $\forall i \geq 0$ and $A_k = \sum_{i=0}^k a_i$.

$$\begin{aligned}
 A_0 G_0 &= A_0 f(x_1) - a_0 (f(x_0) + \langle \nabla f(x_0), x^* - x_0 \rangle + \frac{m}{2} \|x^* - x_0\|_2^2) \\
 &\stackrel{(a_0 = A_0)}{=} A_0 (f(x_1) - f(x_0) - \langle \nabla f(x_0), x^* - x_0 \rangle - \frac{m}{2} \|x^* - x_0\|_2^2) \\
 &\stackrel{\text{"Descent Lemma"} }{\leq} A_0 \left(-\frac{\alpha}{2} \|\nabla f(x_0)\|_2^2 - \underbrace{\langle \nabla f(x_0), x^* - x_0 \rangle}_{\text{C.S.}} - \frac{m}{2} \|x^* - x_0\|_2^2 \right) \\
 &\stackrel{P_1, P_2}{\leq} -\frac{P_1^2}{2} + P_2 \leq \frac{Q^2}{2} \\
 &\quad \left(\frac{1}{2\alpha} \|x^* - x_0\|_2^2 \neq \frac{\|\nabla f(x_0)\|_2^2}{2} \right) \\
 &\leq A_0 \left(\frac{1}{\alpha} - m \right) \frac{\|x^* - x_0\|_2^2}{2}.
 \end{aligned}$$

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

$$\begin{aligned}
 A_k G_k - A_{k-1} G_{k-1} &= A_k f(x_{k+1}) - A_{k-1} f(x_k) \\
 &\quad - a_k (f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle + \frac{m}{2} \|x^* - x_k\|_2^2).
 \end{aligned}$$

$$\begin{aligned}
 &\stackrel{A_k = A_{k-1} + a_k}{=} A_k (f(x_{k+1}) - f(x_k)) \\
 &\quad - a_k \underbrace{\langle \nabla f(x_k), x^* - x_k \rangle}_{\text{C.S.}} - \frac{a_k m}{2} \|x^* - x_k\|_2^2 \\
 &\leq -\frac{A_k \alpha}{2} \|\nabla f(x_k)\|_2^2 + a_k \|\nabla f(x_k)\|_2 \|x^* - x_k\|_2 \\
 &\quad - \frac{a_k m}{2} \|x^* - x_k\|_2^2 \quad \frac{a_k^2}{2 A_k \alpha} \|x^* - x_k\|_2^2 \\
 &\leq \left(\frac{a_k^2}{\alpha A_k} - a_k m \right) \frac{\|x^* - x_k\|_2^2}{2} \\
 &\leq 0 \quad \text{if} \quad \boxed{\frac{a_k}{A_k} \leq \alpha \cdot m}.
 \end{aligned}$$

$$A_k G_k - A_{k-1} G_{k-1} \leq 0.$$

$$A_k G_k \leq A_0 G_0$$

$$f(x_{k+1}) - f(x^*) \leq G_k \leq \frac{A_0 G_0}{A_k} = A_0 \underbrace{\left(\frac{1}{\alpha} - m \right)}_{\frac{1}{\alpha}(1-m\alpha)} \frac{\|x^* - x_0\|_2^2}{2A_k}.$$

Remains to bound $\frac{A_0}{A_k}$.

$\forall k : \frac{a_k}{A_k} \leq \alpha m$. Fastest growing choice:

$$\frac{a_k}{A_k} = \alpha m$$

$$a_k = A_k - A_{k-1}; \quad \left(1 - \frac{A_{k-1}}{A_k}\right) = \alpha m$$

$$\frac{A_{k-1}}{A_k} = (1 - \alpha m)$$

$$\frac{A_0}{A_k} = \frac{A_0}{A_1} \cdot \frac{A_1}{A_2} \cdots \frac{A_{k-1}}{A_k} = (1 - \alpha m)^k *$$

$$\boxed{f(x_{k+1}) - f(x^*) \leq (1 - \alpha m)^{k+1} \cdot \frac{\|x_0 - x^*\|_2^2}{2\alpha}}$$

$$\leq e^{-\alpha m(k+1)} \frac{\|x_0 - x^*\|_2^2}{2\alpha} \leq \epsilon \quad \log(\cdot)$$

$$f(x_{k+1}) - f(x^*) \leq \underline{\epsilon} \text{ after at most}$$

$$\underbrace{k = \frac{1}{\alpha m} \log \left(\frac{\|x_0 - x^*\|_2^2}{2\alpha \epsilon} \right) - 1}_{\text{iterations.}}$$

If f is smooth & strongly convex, we can also write:

$$\underbrace{f(x_{k+1}) - f(x^*) \leq \min \left\{ (1 - \alpha m)^{k+1} \frac{\|x_0 - x^*\|_2^2}{2\alpha}, \frac{2R^2}{\alpha(k+1)} \right\}}_{\text{.}}$$

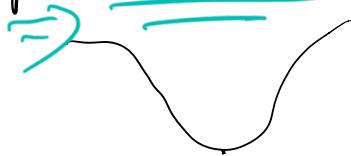
- * Did we really need strong convexity? to get this fast convergence?
No. We can use Polyak-Lojasiewicz (PL) condition:

Why?

$$(\forall x \in \mathbb{R}^d) : \|\nabla f(x)\|_2^2 \geq 2m(f(x) - f(x^*)) , m > 0.$$

Ex. Prove that m -strong convexity implies PL condition.

$$f(x^*) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2$$



- * "Simple line search" for gradient descent.
Setting: you know that f is smooth, but you don't know L .

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

but how do you set α if you don't know L ?
for convergence, we require that $\alpha \leq \frac{1}{L}$. $\text{Upd } \alpha \leq \frac{1}{L}$

Start w/ some guess $L_0 > 0$.

At iteration k , $L_k = L_{k-1}$

Try: $x_{k+1} = x_k - \frac{1}{L_k} \nabla f(x_k)$

check whether Lemma 2.2. hold for this L_k :

$$f(x_{k+1}) \stackrel{?}{\leq} f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_k}{2} \|x_{k+1} - x_k\|_2^2$$

if yes ✓ accept the step

if no, $L_k \leftarrow 2 \cdot L_k$, repeat from \star .

Q. How many times in total would I need to

double my estimate of L in the worst case?

$$\left\lfloor \log \left(\frac{2L}{L_0} \right) \right\rfloor$$