* Other basic descent methods:

There are other descent methods for which you can guarantee:

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

* Examples:

1) Preconditioned methods:

$x_{k+1} = x_k - \alpha S_k \nabla f(x_k)$, where S_k is a PD matrix w/ eigenvalues in $[\gamma_1, \gamma_2]$, $0 < \gamma_1 < \gamma_2 < \infty$.

$x_{k+1} - x_k = -\alpha S_k \nabla \sim$

From Lemma 2.2:

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

$$= f(x_k) - \alpha \underbrace{\langle S_k \nabla f(x_k), \nabla f(x_k) \rangle}_{\geq \gamma_1 \|\nabla f(x_k)\|_2^2} + \frac{L}{2} \alpha^2 \underbrace{\|S_k \nabla f(x_k)\|_2^2}_{\leq \gamma_2^2 \|\nabla f(x_k)\|_2^2}$$

$$\leq f(x_k) - \underbrace{\left(\alpha \gamma_1 - \frac{L}{2} \gamma_2^2 \alpha^2\right)}_{>0 \text{ for suff. small } \alpha} \|\nabla f(x_k)\|_2^2.$$

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)

$x_{k+1} = x_k - \alpha \underbrace{\nabla_{i_k} f(x_k)}_{-p_k} e_{i_k}$, $e_{i_k} = [0, 0, \ldots, \underset{i_k \text{ position}}{1}, \ldots, 0]$

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

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

3) Randomized coordinate descent (HW #2)

4) Stochastic gradient descent, where

$p_k = -g(x_k, \xi_k)$, $\mathbb{E}_{\xi_k}[g(x_k, \xi_k)] = \nabla f(x_k)$

(ξ_k ↓ i.i.d. r.v.)

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

under certain assumptions.

* Convergence of basic descent methods: 收敛性

* Assume:

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

① nonconvex case: 非凸情形

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

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

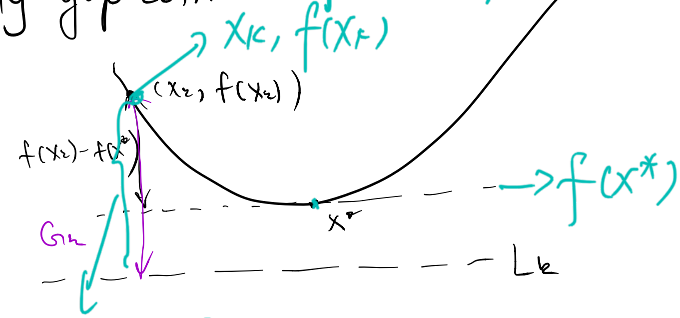
② Convex f.

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

$f(x_{k+1}) - f(x^*) \leq G_k$ optimality gap estimate 最佳gap估计

$f(x^*) \geq L_k$ lower bound 下界

Let A_k be a strictly increasing positive numbers. 严格增正序列



$G_k = f(x^k) - L_k$

使 $A_k G_k$ 收敛

Goal:

$$A_k G_k - A_{k-1} G_{k-1} \le E_k \quad \text{"error"}$$

Sum →

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

$$f(x_{k+1}) - f(x^*) \le G_k \le \frac{A_0 G_0}{A_k} + \frac{\sum_{i=1}^{k} E_i}{A_k}$$

($A_0 G_0$: bounded; $\sum E_i$: want it to grow slowly compared to A_k; $\frac{\sum E_i}{A_k} \to 0$)

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

$\forall\, i$:

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

a_i 是正数序列，A_k 是 a_i 的和

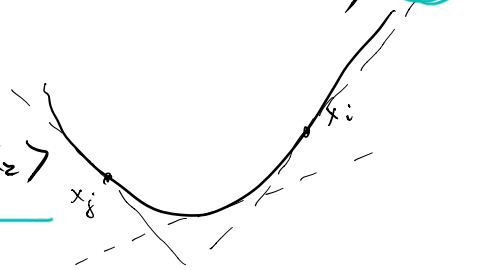
Let $\{a_i\}_{i \ge 0}$ be a sequence of positive numbers

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

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

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

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



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

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

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

$A_k = A_{k-1} + a_k$

$$= A_k \left(f(x_{k+1}) - f(x_k) \right) - a_k \langle \nabla f(x_k), x^* - x_k \rangle$$

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

"Descent Lemma"

$$\le -\frac{A_k \alpha}{2}\|\nabla f(x_k)\|_2^2 - a_k \langle \nabla f(x_k), x^* - x_k \rangle$$

C.S.

$$\le -\frac{A_k \alpha}{2}\|\nabla f(x_k)\|_2^2 + a_k \|\nabla f(x_k)\|_2 \cdot \|x^* - x_k\|_2$$

Cauchy Schwarz ineq

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

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

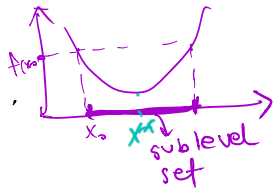
$$A_k G_k - A_{k-1} G_{k-1} \le \frac{a_k^2}{2\alpha A_k}\|x^* - x_k\|_2^2 = E_k$$

Ex. $A_0 G_0 \le \frac{a_0^2}{2\alpha A_0}\|x^* - x_0\|_2^2$

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

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

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



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

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

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

(for integer p, Faulhaber's formula)

One particular choice

$$a_i = \frac{i+1}{2}; \quad A_i = \frac{(i+1)(i+2)}{4}$$

$$\frac{a_i^2}{A_i} \le 1$$

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

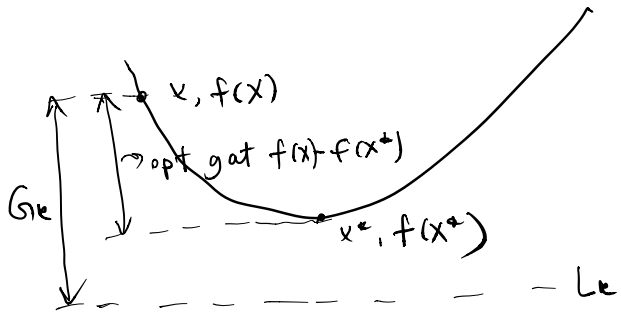
* Can we do better?
  No, unless we say something more about the method we are using.
  E.g., for gradient descent, we can replace R by $\|x_0 - x^*\|_2$.

* Strongly convex case:

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

$L_k \le f(x^*)$

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



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".

Goal: $A_k G_k - A_{k-1} G_{k-1} \le 0$. *

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

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

$\forall i$:

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

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

$$\frac{1}{\partial_k} \cdot \sum a_i \left(f(x_i) + \langle \quad \rangle + \tfrac{m}{2} \| \quad \| \right) = L_k$$

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

$$A_0 G_0 = A_0 f(x_1) - a_0 \left(f(x_0) + \langle \nabla f(x_0), x^* - x_0 \rangle + \frac{m}{2} \|x^* - x_0\|_2^2 \right)$$

$(a_0 = A_0)$

$$= A_0 \left(f(x_1) - f(x_0) - \langle \nabla f(x_0), x^* - x_0 \rangle - \frac{m}{2} \|x^* - x_0\|_2^2 \right)$$

"Descent Lemma"

$$\leq A_0 \left(-\frac{\alpha}{2} \|\nabla f(x_0)\|_2^2 - \langle \nabla f(x_0), x^* - x_0 \rangle - \frac{m}{2} \|x^* - x_0\|_2^2 \right)$$

C.S. $\leq \|\nabla f(x_0)\|_2 \|x^* - x_0\|$

$\forall p, q$: $-\frac{p^2}{2} + pq \leq \frac{q^2}{2}$

$\frac{1}{2\alpha} \|x^* - x_0\|_2^2$ $+$

$$\leq A_0 \left(\frac{1}{\alpha} - m \right) \frac{\|x^* - x_0\|_2^2}{2}.$$

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

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

$A_k = A_{k-1} + a_k$

$$= A_k \left(f(x_{k+1}) - f(x_k) \right) - a_k \langle \nabla f(x_k), x^* - x_k \rangle - \frac{a_k m}{2} \|x^* - x_k\|_2^2$$

C.S.

$$\leq -\frac{A_k \alpha}{2} \|\nabla f(x_k)\|_2^2 + a_k \|\nabla f(x_k)\|_2 \|x^* - x_k\|_2 - \frac{a_k m}{2} \|x^* - x_k\|_2^2$$

$\frac{a_k^2}{2 A_k \alpha} \|x^* - x_k\|_2^2$

$$\leq \left(\frac{a_k^2}{\alpha A_k} - a_k m \right) \frac{\|x^* - x_k\|_2^2}{2}.$$

$$\leq 0 \quad \text{if} \quad \boxed{\frac{a_k}{A_k} \leq \alpha \cdot m}.$$

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

$A_k G_k \leq A_0 G_0$

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

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

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

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

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

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

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

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

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

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

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

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

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

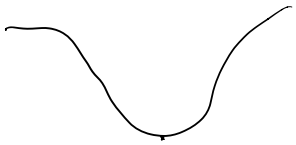
* Did we really need strong convexity to get this fast convergence?

No. We can use Polyak-Łojasiewicz (PL) condition:

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

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 required that $\alpha \leq \frac{1}{L}$.

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:

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

if yes ✓ accept the step

if no, $L_k \leftarrow 2 \cdot L_k$, repeat from (☆).

Q. How many times in total would I need to

double my estimate of L in the worst case?

$< \log\left(\frac{2L}{L_0}\right)$.