Lec13-new

Saturday, February 21, 2009 8:58 PM

HW4 is on the web

Bring convergence/projection handout.

Optimization algorithms

Sunday, February 15, 2009 5:46 PM

- We have now discussed
    - convex problems
    - optimality conditions

- Sometimes we can explicitly solve for the optimal solution from the optimality condition

e.g. $\min \|Ax-b\|_2^2$

$$\Rightarrow \quad A^T A x = A^T b$$
$$x^* = (A^T A)^{-1} A^T b$$

- At other times a closed-form solution may not be possible

- We may then use numerical algorithms.

Numerical Algorithms

- A numerical algorithm starts from some initial estimate x_0, and iteratively generate new estimates by

$$x_k = T(x_{k-1}) \qquad k = 1, 2, \cdots$$

- Hopefully, as $k \to +\infty$, $X_k \to X^*$ the optimal solution.

Q) When will such a sequence converge to the optimal solution?

A) often by showing that the quality of x_k improves in each iteration

(a) The distance between x_k and x^* improves in each iteration. e.g

$$\|x_k - x^*\| \le \alpha \|x_{k-1} - x^*\| \quad , \alpha < 1$$

or $$\|x_k - x^*\| \le \|x_{k-1} - x^*\| - \beta \quad , \beta > 0$$

(b) The function value improves in each iteration. e.g

$$f_0(x_k) \le f_0(x_{k-1}) - \gamma. \qquad \gamma > 0.$$

A trivial example

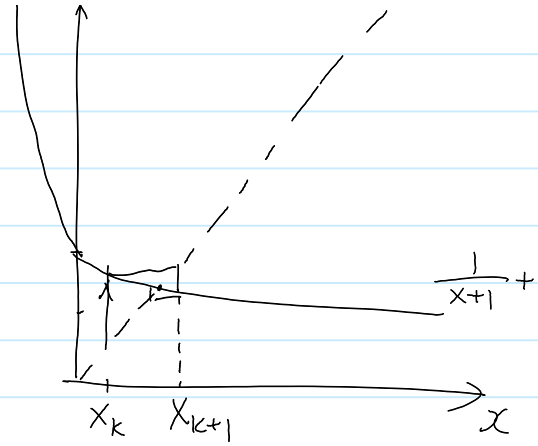
Skip

- How to compute $\sqrt{2}$ using only $+, -, \times, /$
- $x = \sqrt{2} \iff x^2 - 1 = 1$
  $\iff x = \frac{1}{x+1} + 1$
- The iteration
$$x_k = \frac{1}{x_{k-1}+1} + 1$$
- For any two points $\| \ \|^*$ /

- For any two points x, y, let

$$Tx = \frac{1}{x+1} + 1$$

$$Ty = \frac{1}{y+1} + 1$$



$$Tx - Ty = \frac{1}{x+1} - \frac{1}{y+1}$$

$$= -\frac{x-y}{(x+1)(y+1)}$$

$$\|Tx - Ty\| \leq \frac{\|x-y\|}{4}$$ assuming that $x, y \geq 1$. $(*)$

- Note that $\sqrt{2}$ is a fixed point of the mapping

$$T(\sqrt{2}) = \sqrt{2}$$

- Hence, the distance $\|x_k - \sqrt{2}\|$ is cut by $\frac{1}{4}$ in each iteration

$$\Rightarrow \quad x_k \to \sqrt{2}.$$

- The inequality $(*)$ describes a "contraction mapping":
  - The distance between Tx & Ty is less than that of x & y

$$\|Tx - Ty\| \leq c\|x - y\|, \quad 0 \leq c \leq 1$$

  - When $c < 1$ (the contraction mapping is strict), we immediately have

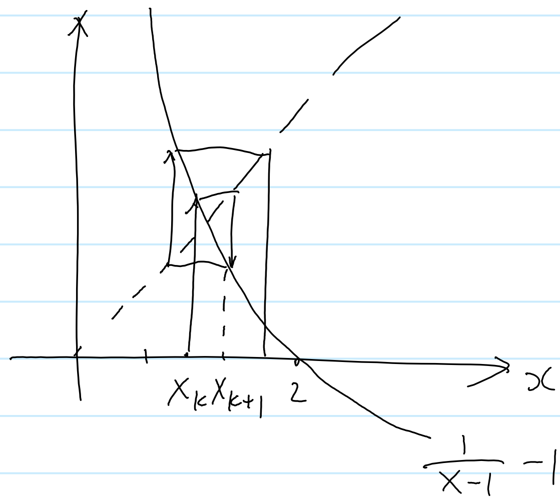
    ① There is a unique fixed point x^* with $Tx^* = x^*$

    ② $x_{k+1} = T(x_k)$ converges "geometrically" to x^*

② $x_{k+1} = T(x_k)$ converges "geometrically" to x^*.

However, not all iterative algorithms converge.

$$x = \sqrt{2} \iff x = \frac{1}{x-1} - 1$$



(10)

Unconstrained optimization

Sunday, February 15, 2009 6:04 PM

- $\min f(x)$ f is convex
- Optimality condition
  - If f is differentiable

    $$\nabla f(x^*) = 0$$

  - If f is not differentiable

    $$f'(x; x - x^*) \geq 0 \quad \forall x$$

    i.e, the directional derivative is positive in all directions.

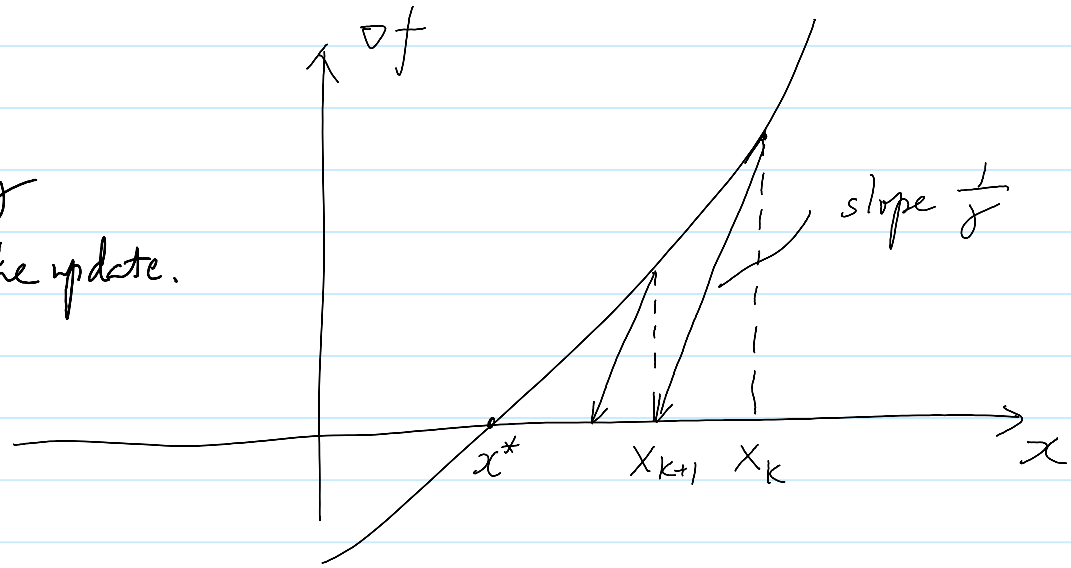
- Assume that $f(\cdot)$ is differentiable.
- Consider the iteration of the type

$$x(t+1) = x(t) - \gamma \underbrace{\nabla f}_{\text{gradient}}(x(t)) \triangleq T(x(t))$$

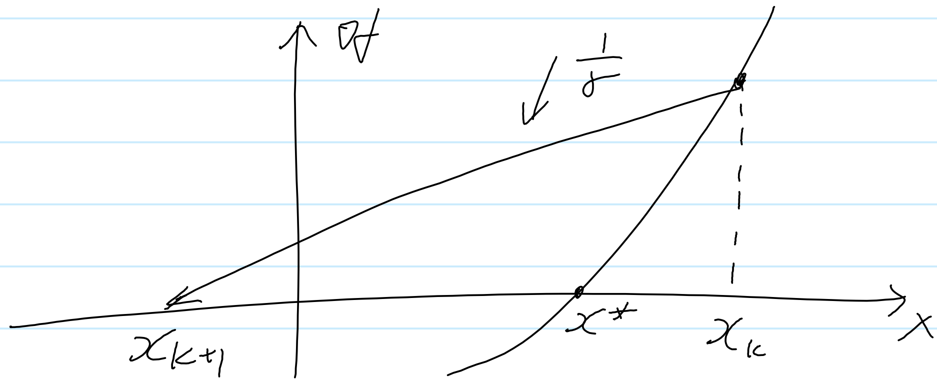
- Note that x^* is a fixed point of the iteration / mapping

- If $x(t) = x^*$, then $x(t+1) = x^*$.

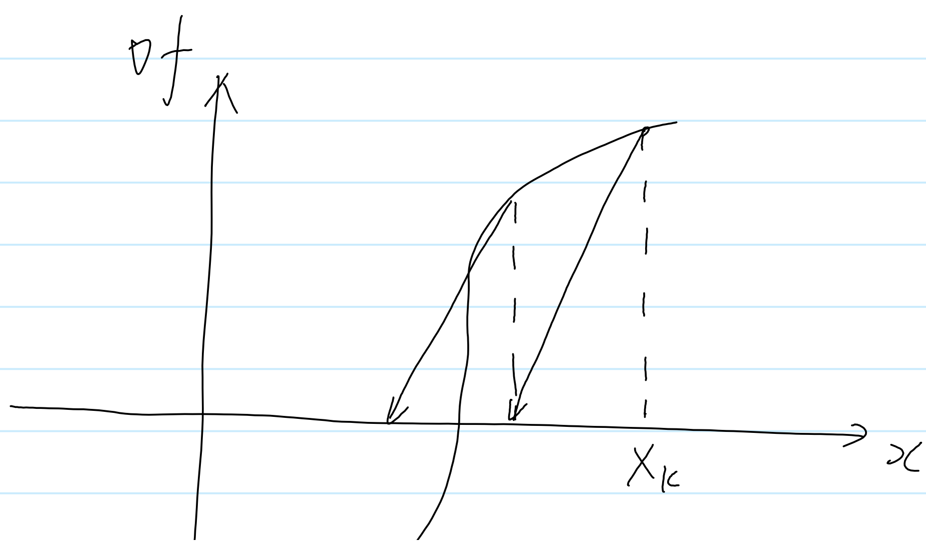
The smaller γ
the smaller the update.

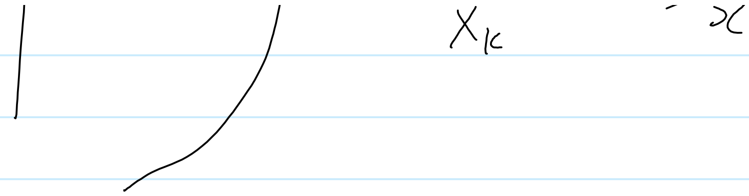


$x_k \to x^*$ if γ is small.



May not converge if γ is too large.





May be more difficult when ∇f is not "smooth"

- when f has sharp corners.

20

Conditions related to convergence

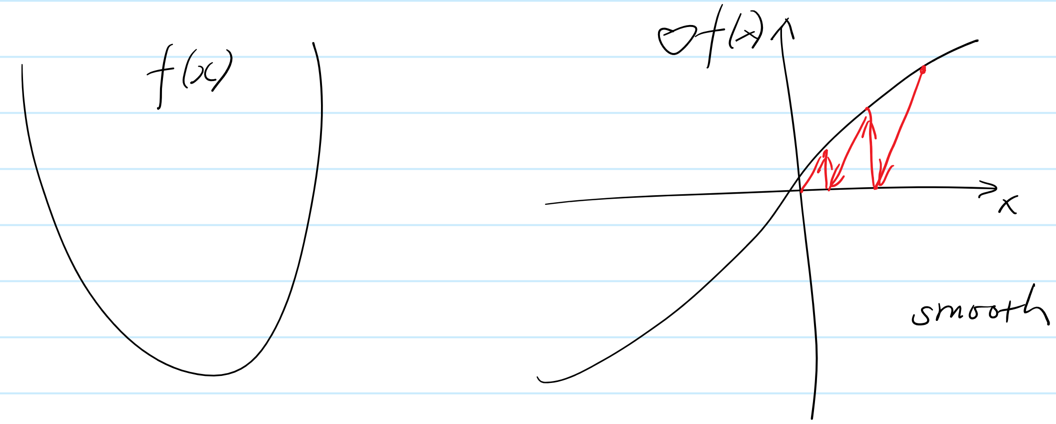
Monday, February 16, 2009 8:24 PM

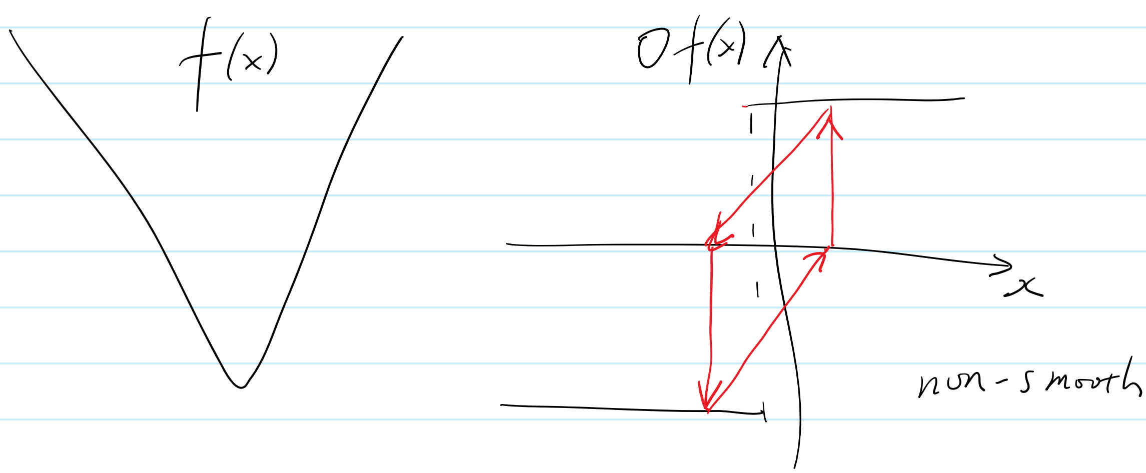
- How can we study the convergence of the gradient algorithm?
- We will usually encounter two types of conditions on the gradient $\nabla f(x)$

① Smoothness

- $\nabla f(x)$ does not abruptly change as x change

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L \, \|x-y\|_2$$





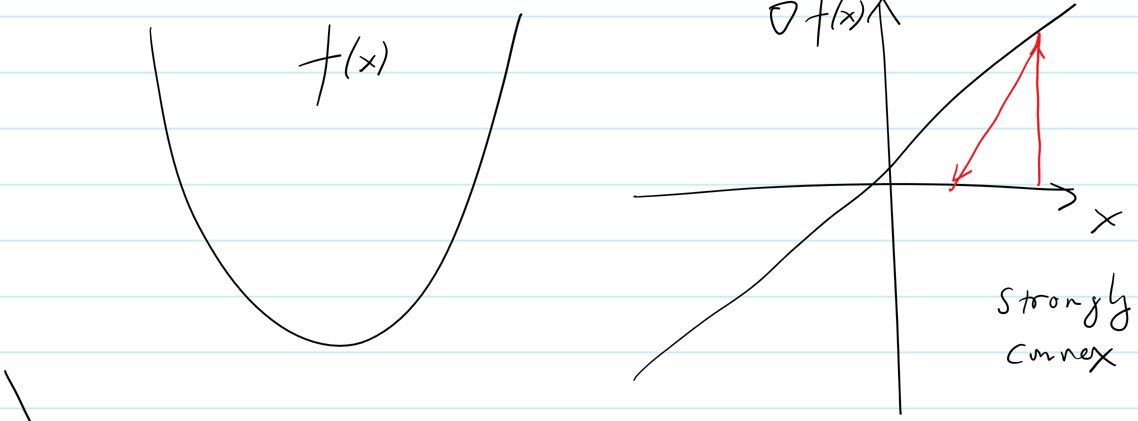
- For non-smooth func, even when you are already in a neighborhood of x^*, it would be difficult for the gradient algorithm to converge exactly to x^*

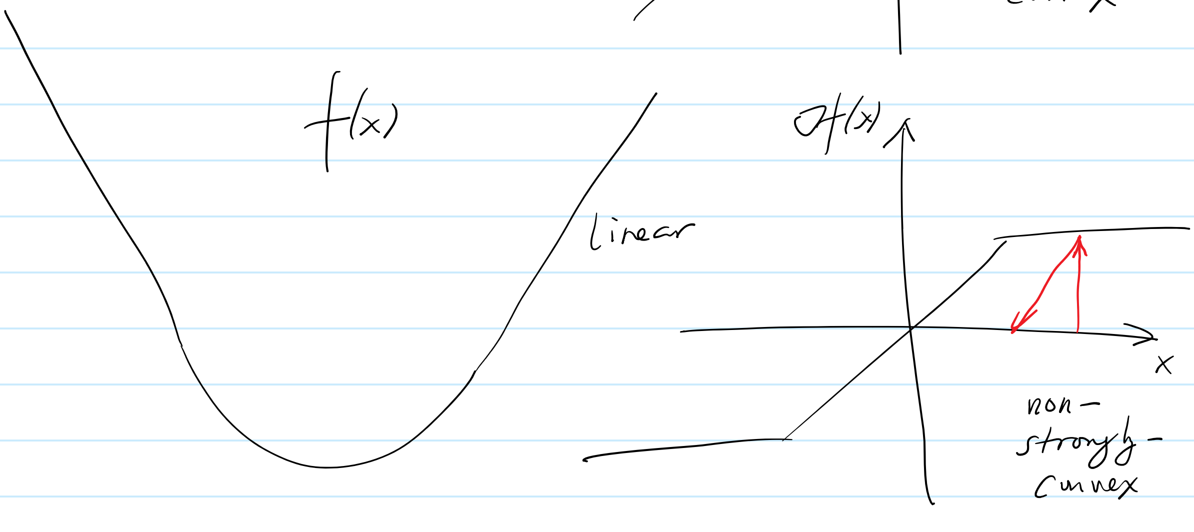
- Instead, smoothness implies convergence, provided that the stepsize is sufficiently small.

② Strong convexity

- $\nabla f(x)$ increases quickly as x lies away from x^*

$$\|\nabla f(x) - \nabla f(y)\|_2 \geq \alpha \|x-y\|_2$$





- If the function is not strongly convex, then starting from $x(0)$ that is far away from x^*, the improvement of the gradient algorithm will be slow

- Instead, strong-convexity & smoothness combined implies geometric descent

$$\|x(t+1) - x^*\|_2 \le \rho \|x(t) - x^*\|_2$$
$$\rho < 1$$

- This is usually referred to as "linear convergence" in the optimization literature

  - In contrast to "quadratic convergence" by Newton's algorithm.

Skip

Two proof techniques:

① Show decrease of $f(x(t))$: More intuitive but less powerful results.

② Show decrease of $\|x(t) - x^*\|$

First approach: Decrease of f - skip

Monday, February 16, 2009 8:24 PM

First Approach:

Lemma: Assume that function f is continuously differentiable & there exists a constant L such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x-y\|_2$$

Smoothness

Then

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2^2$$

Note: The Lipshitz condition bounds how smooth $\nabla f(\cdot)$ is.

$g(t) = f(x + t(y-x))$

$g'(t) = \nabla f(x+t(y-x)) \cdot (y-x)$

Proof:

$$f(y) = \underbrace{f(x + (y-x))}_{= g(\cdot)}$$

$$= \underbrace{f(x)}_{= g(\cdot)} + \int_0^1 \nabla f(x + t(y-x))^T \cdot (y-x)\, dt$$

$$= f(x) + \nabla f(x)^T \cdot (y-x) + \int_0^1 [\nabla f(x+t(y-x)) - \nabla f(x)]^T (y-x)\, dt$$

$$\le f(x) + \nabla f(x)^T (y-x) + \int_0^1 Lt\,\|y-x\|^2\, dt$$

$$= f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2.$$

Theorem: (Bertsekas & Tsitsiklis p203)

Assume that function f is continuously differentiable & there exists a constant L such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x-y\|_2 \quad \forall x, y$$

Suppose f is bounded from below by f^*.

Assume further that $0 < \gamma < \frac{2}{L}$. If the sequence of points $x(t)$ generated by

$$x(t+1) = x(t) - \gamma \nabla f(x(t))$$

has a limit point x^*, then

$$\nabla f(x^*) = 0$$

Start with the basic Taylor-series expansion:

$$f(x(t+1)) \simeq f(x(t)) + \nabla f(x(t))^T \cdot (x(t+1) - x(t))$$
$$= f(x(t)) + \underbrace{\nabla f(x(t))^T \cdot [-\gamma \nabla f(x(t))]}_{\le 0}$$

- However, there may be other higher-order terms, which may create problems when $\nabla f(x(t))$ is already small

- The smoothness condition controls these higher-order terms.

Proof: Use the above Lemma. Let $x = x(t)$,
$y = x(t) - \gamma \nabla f(x(t)) = x(t+1)$

$$\Rightarrow f(x(t+1)) \le f(x)) - \gamma \|\nabla f(x(t))\|^2$$
$$+ \frac{L}{2} \cdot \gamma^2 \|\nabla f(x(t))\|^2 \qquad (*)$$

Since $\gamma < \frac{2}{L}$, we have $\gamma - \frac{L}{2}\gamma^2 > 0$.

Therefore, $f(x(t))$ is non-increasing. But it has a lower bound $f(x^*)$. Hence, the decrement must go to zero.

More precisely, Summing $(*)$ over t, we have

$$f^* \le f(x(t+1)) \le f(x(0)) - \sum_{k=0}^{t} \left(\gamma - \frac{L}{2}\gamma^2\right) \|\nabla f(x(k))\|^2$$

$$\therefore \quad \sum_{k=0}^{t} \|\nabla f(x(k))\|^2 < f(x(0)) - f^*, \ \forall t$$

$$\Rightarrow \nabla f(x(t)) \to 0 \quad \text{as} \quad t \to +\infty$$

Since x^* is a limit point of $x(t)$, and ∇f is continuous, we must have

$$\nabla f(x^*) = 0$$

40

- Does $x(t) \to x^*$?

Second approach: Decrease of norm

Monday, February 16, 2009 8:39 PM

Lemma: If f is convex, then

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq 0 \qquad (*)$$

Note: This holds even if $\nabla f(\cdot)$ is a sub-gradient. A mapping $\nabla f(\cdot)$ that satisfy $(*)$ is called a monotone mapping.

- For 1-dim, $(f'(x) - f'(y))(x-y) \geq 0$

$\Leftrightarrow$ $f'(x) \geq f'(y)$ when $x > y$ (monotonicity)

$\Leftrightarrow$ $f''(x) \geq 0$.

Proof: From first-order condition of convexity,

$$f(y) \geq f(x) + \nabla f(x)(y-x)$$

$$f(x) \geq f(y) + \nabla f(y)(x-y)$$

Summing them,

$$\Rightarrow \quad [\nabla f(y) - \nabla f(x)]^T (y-x) \geq 0 \qquad \#$$

- Later on, we will apply this to

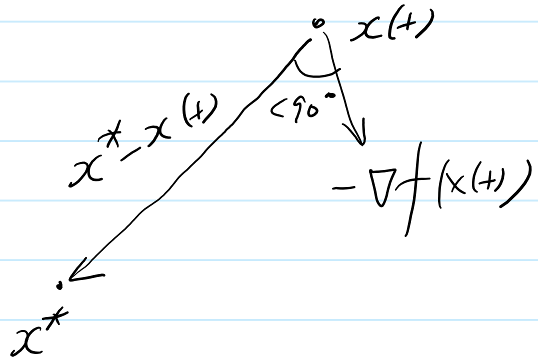
$$y = x(t)$$

$$x = x^*$$

$$\Rightarrow \quad \nabla f(x(t))^T \cdot (x(t) - x^*) \geq 0$$

$$\Rightarrow \quad \nabla f(x(t))^T \cdot (x(t) - x^*) \geq 0$$

$$[- \nabla f(x(t)))^T (x^* - x(t)) \geq 0$$



If the derivative is Lipschitz-continuous, then a stronger version can be shown.

Lemma: Let f be a convex & differentiable function such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 \quad \forall x, y \in R^n$$

Then

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

$$\forall x, y \in R^n \quad (**)$$

Note: If a mapping $\nabla f(\cdot)$ that satisfies $(**)$, the inverse mapping $(\nabla f(x) \to x)$ is called strongly monotone.

We will prove this Lemma later.

Theorem: Assume that

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

$$\forall x, y \in \mathbb{R}^n$$

Further, assume that $0 < \gamma < \frac{2}{L}$ and there exists at least one point x^* with $\nabla f(x^*) = 0$.

Then the sequence of $x(t)$ generated by

$$x(t+1) = x(t) - \gamma \nabla f(x(t))$$

converges, and the limit x_0 satisfies $\nabla f(x_0) = 0$.

Proof: Let x^* be one optimal solution, i.e., $\nabla f(x^*) = 0$.

$$\|x(t+1) - x^*\|^2$$

$$= \|x(t) - x^*\|^2 + 2(x(t+1) - x(t))^T (x(t) - x^*) + \|x(t+1) - x(t)\|^2$$

Note that

$$(x(t+1) - x(t))^T (x(t) - x^*)$$

$$= -\gamma (\nabla f(x(t)) - \underbrace{\nabla f(x^*)}_{=0}) (x(t) - x^*)$$

$$\leq -\frac{\gamma}{L} \|\nabla f(x(t))\|^2$$

$$\therefore \|x(t+1) - x^*\|^2$$

$$\le \|x(t)-x^*\|^2 - \frac{2\gamma}{L}\|\nabla f(x(t))\|^2 + \gamma^2\|\nabla f(x(t))\|^2$$

$$= \|x(t)-x^*\|^2 - \left(\frac{2\gamma}{L}-\gamma^2\right)\|\nabla f(x(t))\|^2$$

If $\gamma < \frac{2}{L}$, then $\alpha \triangleq \frac{2\gamma}{L} - \gamma^2 > 0$

Summing over t

$$\underset{\overset{\vee\!/}{0}}{\|x(t+1)-x^*\|^2} \le \|x(0)-x^*\|^2 - \alpha \sum_{s=0}^{t} \|\nabla f(x(s))\|^2$$

$$\Rightarrow \quad \sum_{s=0}^{t} \|\nabla f(x(s))\|^2 < +\infty$$

$$\Rightarrow \quad \nabla f(x(t)) \to 0$$

Q Why does $x(t)$ converge?

A (Key technique) Assume that a subsequence converges, i.e.,

$$x(t_h) \to x_0, \quad \text{as } h \to +\infty$$

where

$$t_h \to +\infty \quad \text{as } h \to +\infty.$$

We must then have $\nabla f(x_0) = 0$ due to the continuity of $\nabla f(\cdot)$.

Now, replace x^* by x_0,

$$\lim_{t\to+\infty} \|x(t)-x_0\|^2 \le \lim_{h\to+\infty} \|x(t_h)-x_0\|^2 = 0$$

$\Rightarrow \quad x(t) \to x_0 \quad \text{as} \quad t \to +\infty.$

60

Proof of strong monotonicity lemma - skip

Saturday, February 21, 2009 5:38 PM

- Assume that ∇f is Lipschitz

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\| \quad \forall x, y$$

We want to show that

$$(\nabla f(x) - \nabla f(y))^T(x-y) \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

- Note that by the growth Lemma

(see the earlier section "First approach: Decrease of f")

$$f(x) \le f(y) + \nabla f(y)(x-y) + \frac{L}{2}\|x-y\|^2 \quad (*)$$

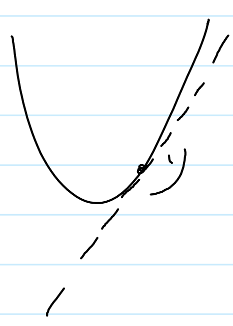
for any convex function and ∇f is Lipschitz.

We will show that

$$f(x) \ge f(y) + \nabla f(y)(x-y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$$

- Fix y.

Let $\tilde{f}(x) = f(x) - f(y) - \nabla f(y)^T(x-y)$



We have

$$\tilde{f}(y) = 0$$

$$\nabla \tilde{f}(x) = \nabla f(x) - \nabla f(y)$$

$$\nabla \tilde{f}(y) = 0$$

Hence, y is the minimum point of $\tilde{f}(\cdot)$.

- Let $z = x - \frac{1}{L}\nabla \tilde{f}(x)$. Using $(*)$ on $\tilde{f}(\cdot)$ have

$$0 \le \tilde{f}(z) \le \tilde{f}(x) + \nabla \tilde{f}(x) \cdot \left(-\frac{1}{L} \nabla \tilde{f}(x)\right)$$

$$+ \frac{L}{2} \cdot \frac{1}{L^2} \cdot \|\nabla \tilde{f}(x)\|^2$$

– Using the definition of $\tilde{f}(x)$, we have

$$f(x) - f(y) - \nabla f(y)^T (x-y) - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \ge 0$$

Interchange the role of x and y, we can similarly show that

$$f(y) - f(x) - \nabla f(x)^T (y-x) - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \ge 0$$

Combine these two inequalities, we have

$$(\nabla f(y) - \nabla f(x))^T (y-x) \ge \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2.$$

#