Lec14-new

Monday, February 28, 2011 5:01 PM

Geometric convergence

Monday, February 6, 2023 11:34 AM

See Bubeck p278

- Sébastien Bubeck, "Convex Optimization: Algorithms and Complexity, in Foundations and Trends in Machine Learning, Vol. 8, No. 3-4 (2015)

- With smoothness, we can ensure convergence, but the speed of convergence may be slow

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

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

- Adding strong convexity enables us to prove geometric descent.
  - $\|\nabla f(x)-\nabla f(y)\|_2 \geq \alpha \|x-y\|_2$
  - Intuitively, this ensures that when $x(t)$ is far away from x^*, the improvement of $\|x(t+1)-x^*\|_2^2$ is directly related to $\|x(t)-x^*\|_2^2$.

$$\|x(t+1)-x^*\|^2 \leq \left(1-\left(\frac{2\gamma}{L}-\gamma^2\right)\alpha\right)\|x(t)-x^*\|^2 \quad (*)$$

- The result below is stronger (i.e., faster descent).

Skip

- Lemma: Let f be a L-smooth & α-strongly convex function on R^n. Then for all $x, y \in R^n$,

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

Proof: Since f is α-strongly-convex, we can show that

$$\varphi(x) = f(x) - \frac{\alpha}{2}\|x\|^2$$

is still convex. Further, we can show that $\varphi(x)$ is $(L-\alpha)$-smooth. Thus, using the earlier lemma for smooth functions, we have

$$[\nabla\varphi(x) - \nabla\varphi(y)]^T(x-y) \geq \frac{1}{L-\alpha}\|\nabla\varphi(x) - \nabla f(y)\|^2$$

$$\Leftrightarrow \quad [\nabla f(x) - \nabla f(y)]^T(x-y) - \alpha(x-y)^T(x-y)$$

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

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

$$- \frac{2\alpha}{L-\alpha}(\nabla f(x) - \nabla f(y))^T(x-y)$$

$$\Leftrightarrow \quad \frac{L+\alpha}{L-\alpha}[\nabla f(x) - \nabla f(y)]^T(x-y)$$

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

$$+ \frac{L\alpha}{L-\alpha}\|x-y\|^2$$

The result of the Lemma then follows. #

Theorem: Let f be a L-smooth & α-strongly convex function on R^n. Then, by setting a stepsize

$$\gamma < \frac{2}{\alpha+L},$$

gradient descent satisfies

$$\|x(t) - x^*\|^2 \leq \left(1 - \frac{2\gamma\alpha L}{\alpha+L}\right)^t \|x(0) - x^*\|^2$$

Proof: Starting with the norm approach again.

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

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

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

Note that

$$\left(x(t+1)-x(t)\right)^T\left(x(t)-x^*\right)$$

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

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

Hence,

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

$$\leq \left(1-\frac{2\gamma\alpha L}{\alpha+L}\right)\|x(t)-x^*\|^2 + \underbrace{\left(\gamma^2-\frac{2\gamma}{\alpha+L}\right)}_{\left(\leq 0 \text{ if } \gamma<\frac{2}{\alpha+L}\right)}\|\nabla f(x)-\nabla f(x^*)\|^2$$

$$\leq \left(1-\frac{2\gamma\alpha L}{\alpha+L}\right)\|x(t)-x^*\|^2$$

The result of the Theorem then follows. #

Note:

- For a faster convergence speed, would want

$$1-\frac{2\gamma\alpha L}{\alpha+L}$$

to be smaller, or γ to be larger.

$\Rightarrow$ set $\gamma = \frac{2}{\alpha + L}$.

- Compare with $(*)$ at $\gamma = \frac{2}{\alpha+L}$

$$1 - \gamma\left(\frac{2}{L} - \gamma\right)\alpha$$
$$= 1 - \frac{2\gamma\alpha^2}{L(\alpha+L)}$$

- When $\frac{\alpha}{L^2}$ is small, this is much closer to 1.

- Using $f(x) - f(x^*) \leq \frac{L}{2} \|x - x^*\|^2$

we have

$$f(x(t)) - f(x^*) \leq \frac{L}{2}\left(1 - \frac{2\gamma\alpha L}{\alpha+L}\right)^t \|x(0) - x^*\|^2.$$

- This is usually referred to as "linear convergence"
  - in contrast to "quadratic convergence" for second-order (e.g. Newton's) algorithms.

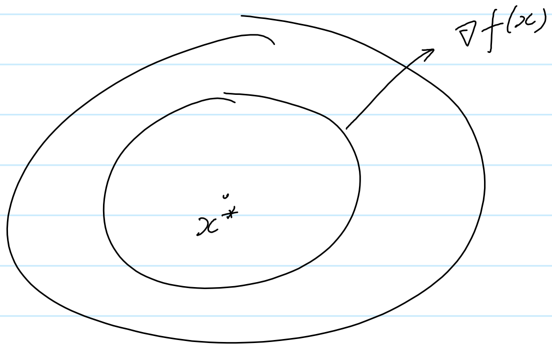
Scaled gradient descent algorithm

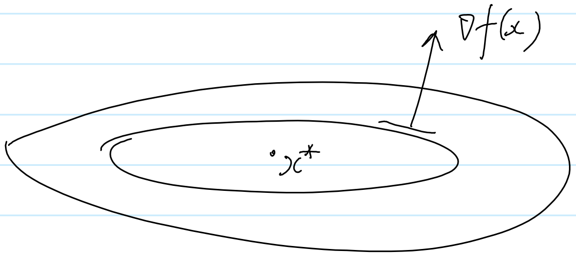
Saturday, February 21, 2009 5:50 PM

- Standard gradient algorithm

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

If $\nabla f(x)$ is Lipschitz & $\gamma < \frac{2}{L}$, then the algorithm converges.

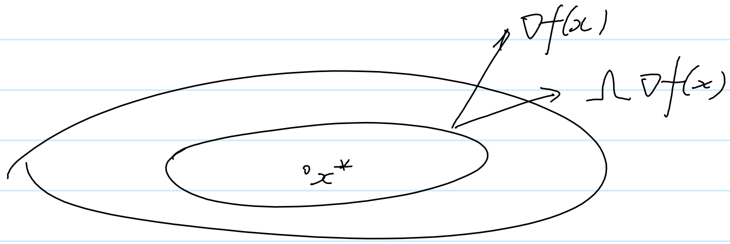




- Scaled gradient algorithm

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

where Ω is a positive-definite matrix



Some choice of Λ can make the algorithm converge faster.

- For example, $\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix}$

  - each processor chooses different step-size.

- Regardless the use of Λ, the scaled gradient algorithm will converge under a similar Lipschitz condition.

$$\gamma < \frac{2}{L \lambda_{max}}$$

  - where λ_{max} is the largest eigenvalue of Λ, i.e.

$$x^T \Lambda x \leq \lambda_{max} \|x\|^2 \quad \text{for all } x$$

<u>Sketch of Proof:</u>

- choose a different norm.

$$(x(t+1) - x^*) \Lambda^{-1} (x(t+1) - x^*)$$

$$= (x(t) - x^*) \Lambda^{-1} (x(t) - x^*)$$

$$+ 2 (x(t+1) - x(t)) \Lambda^{-1} (x(t) - x^*)$$

$$+ (x(t+1) - x(t)) \Lambda^{-1} (x(t+1) - x(t))$$

Note that

$$(x(t+1) - x(t)) \Lambda^{-1} (x(t) - x^*)$$

$$= -\gamma \nabla f(x(t)) (x(t) - x^*)$$

$$= -\gamma [\nabla f(x(t)) - \nabla f(x^*)] [x(t) - x^*]$$

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

Work out the rest in hw.

75

Constrained optimization

Saturday, February 21, 2009 6:06 PM

- Consider the problem

$$\min f(x) \quad \text{sub to} \quad x \in X$$

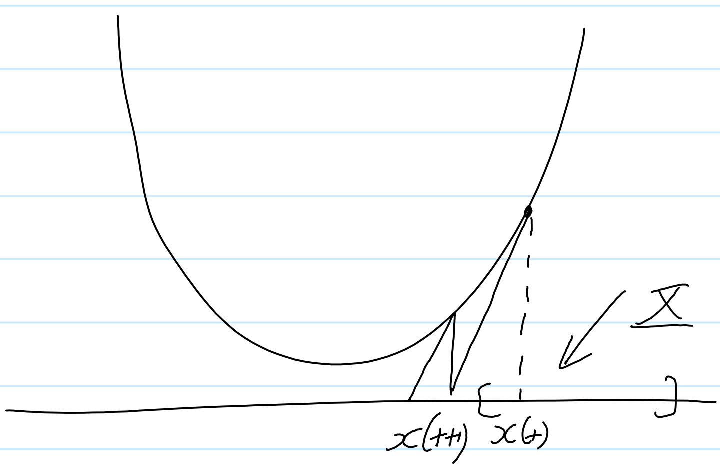
Optimality condition

$$f'(x; y-x) \geq 0 \quad \text{for all } y \in X$$

If f is differentiable, then

$$(\nabla f(x))^T (y-x) \geq 0 \quad \text{for all } y \in X.$$

- However, the normal gradient projection does not work any more because the new $x(t+1)$ can go outside X.



Three solutions

Three solutions

① Penalty function method:

choose $g(x)$ such that

$$g(x) = 0 \quad \text{if } x \in X$$

$$g(x) \geq 0 \quad \text{if } x \notin X$$

- we can now minimize $f(x) + \beta g(x)$

  Let the solution be $x^*(\beta)$

- As $\beta \uparrow +\infty$, the penalty becomes larger & larger, the solution $x^*(\beta)$ will approach x^* (the original constrained problem).

- We will discuss the engineering implication of this approach when we discuss TCP as an example.

---

② Interior point method (Barrier Method).

- Choose $g(x)$ such that $g(x) \to +\infty$ as x approaches the boundary of X from inside.

Example: $\mathbb{X} = \{x \geq 0\}$, $g(x) = -\log x$

$\mathbb{X} = \{x \leq a\}$. $g(x) = -\log(a-x)$

- We then minimize $f(x) + \beta g(x)$.

- Due to the barrier $g(x)$, the optimal solution $x^*(\beta)$ must be in the interior of $\mathbb{X}$

- As $\beta \downarrow 0$, $x^*(\beta) \to x^*$ (the original constrained problem).

In both the penalty-function method or the Interior-point method, the problem is converted to a unconstrained problem. Hence, we can use gradient algorithm to solve $x^*(\beta)$.

- However, it may be difficult to ensure the Lipschitz condition of the gradient.

---

(3) Projection method.

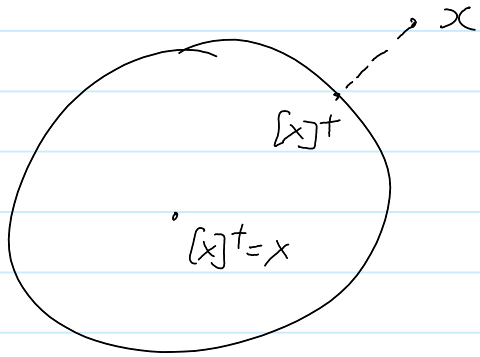
Projection

Saturday, February 21, 2009 6:19 PM

Projection $[x]^+$: (for L_2-norm)

The projection of x onto the convex and closed set $\mathcal{X}$ is the point z in $\mathcal{X}$ that is closest to x, i.e.,

$$[x]^+ = \underset{z \in \mathcal{X}}{\arg\min} \; \|z - x\|_2$$



Example:

① $\mathcal{X} = \bigotimes_{i=1}^{m} [a_i, b_i]$

$$[x]^+ = \{ [x_i]^+ \}$$

where
$$[x_i]^+ = \begin{cases} a_i & x_i \le a_i \\ b_i & x_i \ge b_i \\ x_i & \text{o/w} \end{cases}$$

② Projection to a Polyhedra Boyd p390

$$\min \; \|x - x_0\|_2^2$$
$$\text{sub to} \quad Ax \le b$$

- A quadratic program.

Solution:

- On a hyperplane $a^T x = b$

$$P_C(x_0) = x_0 + (b - a^T x_0) \cdot a \,/\, \|a\|_2^2$$

- On a half-space $a^T x \le b$

$$P_C(x_0) = \begin{cases} x_0 & a^T x_0 \le b \\ x_0 + (b - a^T x_0) \cdot a \,/\, \|a\|_2^2 & \text{if } a^T x_0 > b. \end{cases}$$

② Can we derive these by optimality conditions?

③ In general

$$X = \{x \mid f_i(x) \le 0,\ h_i(x) = 0\}$$

$$[x]^+ = \arg\min_{z} \|z - x\|_2$$
$$\text{sub to} \quad f_i(z) \le 0$$
$$h_i(z) = 0$$

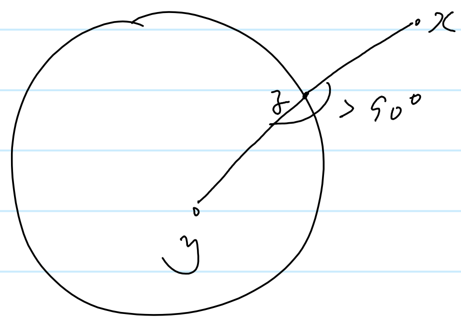
- Not a trivial operation! (Reason to use dual algorithm.)

Projection Theorem (Bertsekas & Tsitsiklis p211)

(a) For every $x \in R^n$, there exists a unique $z \in X$ that minimizes $\|z - x\|_2$ over all $z \in X$, and will be denoted as $[x]^+$

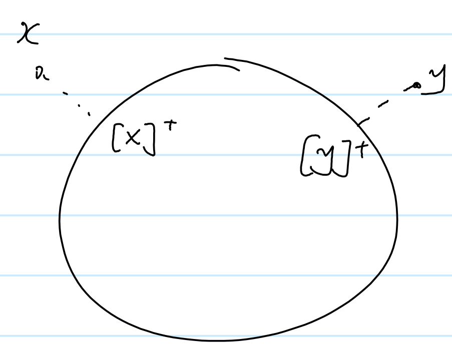
(b) Given some $x \in R^n$, a vector $z \in X$ is equal to $[x]^+$ if & only if

$$(y - z)'(x - z) \leq 0 \quad \text{for all } y \in X$$



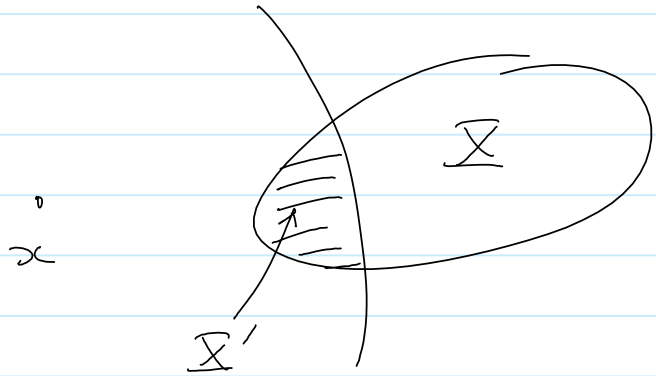
(c) The mapping $f(x) = [x]^+$ is continuous and non-expansive, that is

$$\|[x]^+ - [y]^+\|_2 \leq \|x - y\|_2 \quad \text{for all } x, y \in R$$



Proof:

(a) Can intersect X with a closed & bounded set



- min $\|z-x\|^2$ over X' (closed, bounded)

$\Rightarrow$ exists a minimum point $[x]^+$

- It is unique because $\|z-x\|^2$ is strictly convex.

(b) Consider min $f(z) = \|z-x\|^2$

sub to $z \in X$

- $\nabla f(z) = 2(z-x)$

- By optimality condition

z optimal

$\Leftrightarrow \quad 2(z-x)^T(y-z) \geq 0$ for all $y \in X$

$\Leftrightarrow \quad (y-z)^T(x-z) \leq 0$

(c) From part (b)

$$([y]^+ - [x]^+)^T (x - [x]^+) \leq 0$$

Similarly

$$([x]^+ - [y]^+)^T (y - [y]^+) \le 0$$

$$\Rightarrow \quad ([y]^+ - [x]^+)^T ([y]^+ - [x]^+ - (y - x)) \le 0$$

$$\| [y]^+ - [x]^+ \|^2 \le ([y]^+ - [x]^+) (y - x)$$

$$\le \| [y]^+ - [x]^+ \| \, \| y - x \| .$$

The result then follows.

25