Lecture 2
Gradient Descent and Subgradient Methods

Jie Lu (jielu@kth.se)

Richard Combes
Alexandre Proutiere

Automatic Control, KTH

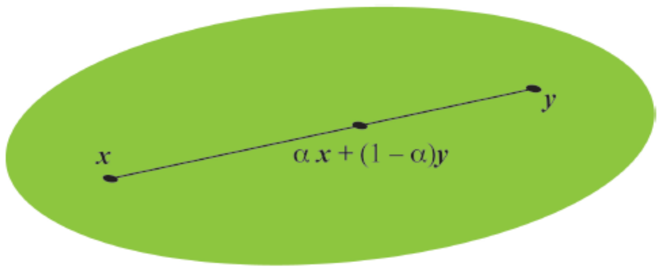
September 12, 2013

Outline

- Convex analysis
- Gradient descent method
- Gradient projection method
- Subgradient method

Convex Set

- A set $X \subset \mathbb{R}^n$ is *convex* iff $\forall x, y \in X$, $\forall \alpha \in [0,1]$, $\alpha x + (1-\alpha) y \in X$.

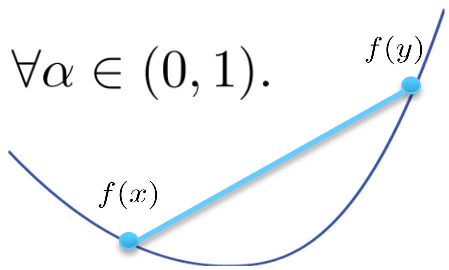


- Examples
  - Hyperplane $\{x \in \mathbb{R}^n : a^T x + b = 0\}$, $a \neq 0$
  - Polyhedral $\{x \in \mathbb{R}^n : Ax + b \preceq 0\}$, $A \in \mathbb{R}^{m \times n}$
  - Ellipsoid $\{x \in \mathbb{R}^n : (x - x_0)^T P^{-1} (x - x_0) \leq 1\}$, $P \in \mathbb{S}^n_+$
- *Convex hull*: the set of all convex combinations of the points in X
  - Convex Combination:

$$\sum_{i=1}^m \alpha_i x_i,\ \alpha_i \in [0,1],\ \sum_{i=1}^m \alpha_i = 1,\ x_i \in X$$

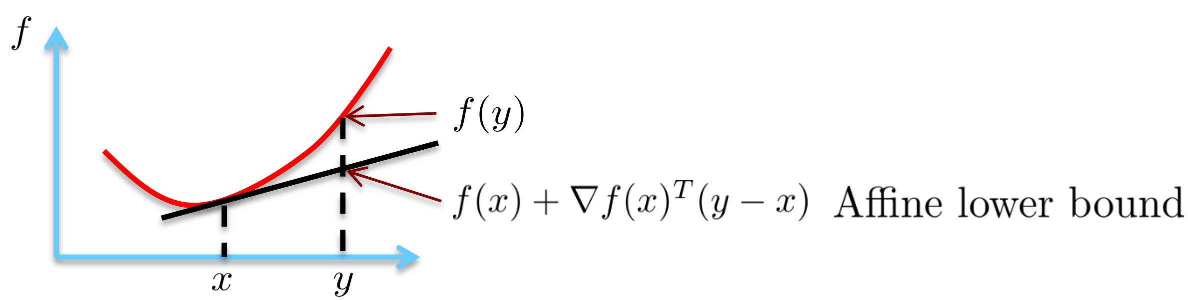
Convex Function

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* iff
  $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \ \forall x, y \in \mathbb{R}^n, \ \forall \alpha \in (0,1).$



  - f is *strictly convex* if the equality holds only when $x = y$.
  - Jensen's Inequality $f(\sum_{i=1}^N \alpha_i x_i) \leq \sum_{i=1}^N \alpha_i f(x_i)$, $\alpha_i \in [0,1]$, $\sum_{i=1}^N \alpha_i = 1$.

- Suppose f is differentiable. Then, f is convex iff
  $f(y) \geq f(x) + \nabla f(x)^T (y - x), \ \forall x, y \in \mathbb{R}^n.$ (1.1)



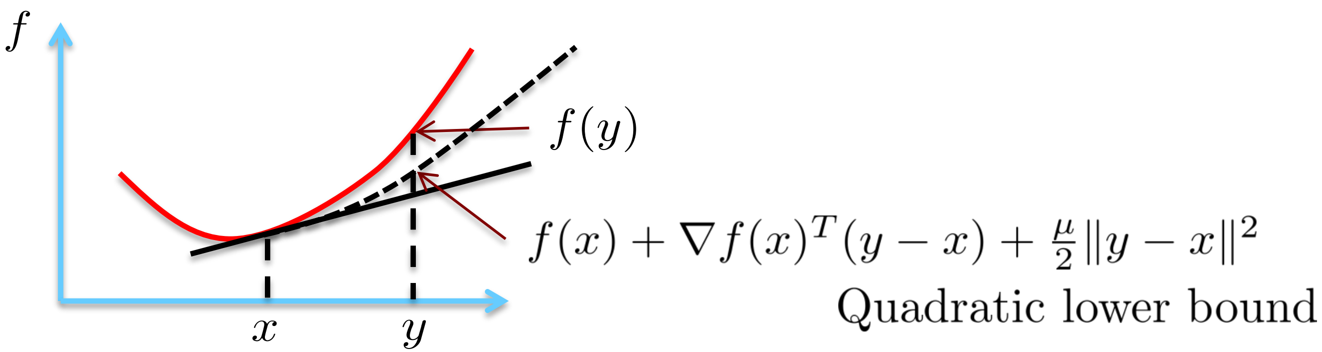
- Eq(1.1) is equivalent to $(\nabla f(y) - \nabla f(x))^T (y - x) \geq 0$.
  If f is twice differentiable, it is equivalent to $\nabla^2 f(x) \geq 0$.

- f is *concave* if $-f$ is convex.

Strong Convexity

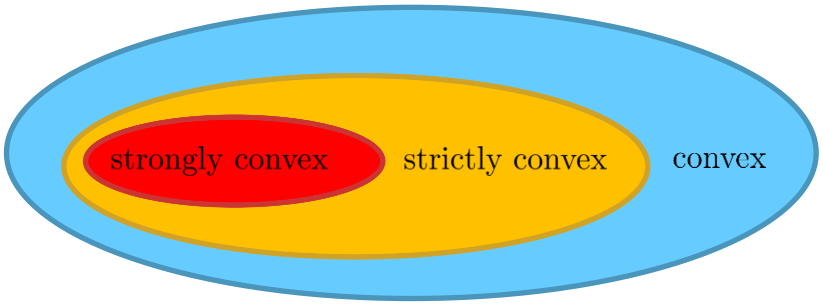
- A differentiable function f is *strongly convex* iff $\exists \mu > 0$ such that one of the following holds:

(i) $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2, \ \forall x, y \in \mathbb{R}^n.$



(ii) $(\nabla f(y) - \nabla f(x))^T(y-x) \geq \mu\|y-x\|^2, \ \forall x, y \in \mathbb{R}^n.$

(iii) $\nabla^2 f(x) \geq \mu I, \ \forall x \in \mathbb{R}^n$, if f is twice differentiable.

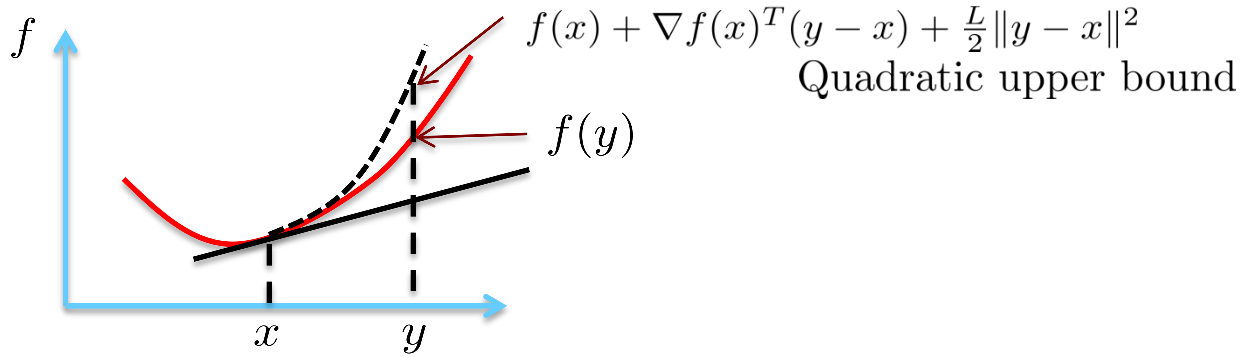


Convex Function with Lipschitz Continuous Gradient

- Let ∇f be Lipschitz continuous, i.e., there exists $L > 0$ such that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^n$.

- f is convex and has Lipschitz continuous gradient iff one of the following holds:

$$0 \leq f(y) - f(x) - \nabla f(x)^T(y - x) \leq \frac{L}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^n.$$



$$f(y) - f(x) - \nabla f(x)^T(y - x) \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2, \forall x, y \in \mathbb{R}^n.$$

$$(\nabla f(y) - \nabla f(x))^T(y - x) \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2, \forall x, y \in \mathbb{R}^n.$$

- If f is strongly convex and has Lipschitz continuous gradient, then

$$(\nabla f(y) - \nabla f(x))^T(y - x) \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(y) - \nabla f(x)\|^2, \forall x, y \in \mathbb{R}^n.$$

Gradient Descent Method

- Unconstrained convex optimization

$$\min_{x\in\mathbb{R}^n} f(x)$$

  - f is convex and continuously differentiable

- Optimality condition: $x^\star \in \arg\min_{x\in\mathbb{R}^n} f(x) \Leftrightarrow \nabla f(x^\star) = 0$
  - Unique if f is strictly convex

- Basic gradient method

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),\ \alpha_k > 0$$

- A descent method (for sufficiently small stepsize α_k)

$$\begin{aligned} f(x_k + \alpha_k d) &= f(x_k) + \alpha_k \nabla f(x_k)^T d + o(\alpha_k) \\ &= f(x_k) + \alpha_k \Big(\nabla f(x_k)^T d + o(\alpha_k)/\alpha_k\Big) \end{aligned}$$

If $\alpha_k > 0$ is small enough so that $o(\alpha_k)/\alpha_k$ is negligible,
$f(x_{k+1}) - f(x_k) \approx -\alpha_k \|\nabla f(x_k)\|^2 \leq 0$

Convergence Analysis

- Choose sufficiently small stepsize α_k so that $f(x_{k+1}) \le f(x_k)\ \forall k \ge 0$
- $f(x_k) - f^\star \le \dfrac{\|x_0 - x^\star\|^2 + \sum_{t=0}^{k-1} \alpha_t^2 \|\nabla f(x_t)\|^2}{2\sum_{t=0}^{k-1} \alpha_t}, \quad \forall k \ge 1$
  - Need further assumptions to guarantee convergence
- Suppose f is Lipschitz continuous with $L_f > 0 \Rightarrow \|\nabla f(x)\| \le L_f,\ \forall x \in \mathbb{R}^n$

$$f(x_k) - f^\star \le \frac{\|x_0 - x^\star\|^2 + L_f^2 \sum_{t=0}^{k-1} \alpha_t^2}{2\sum_{t=0}^{k-1} \alpha_t}, \quad \forall k \ge 1$$

  - For constant stepsize $\alpha_k = \alpha$,

$$\lim_{k\to\infty} f(x_k) \le f^\star + \frac{\alpha L_f^2}{2}$$

  - For diminishing stepsize $\sum_{k=0}^{\infty} \alpha_k = \infty,\ \sum_{k=0}^{\infty} \alpha_k^2 < \infty,$

$$\lim_{k\to\infty} f(x_k) = f^\star$$

  - Accuracy ϵ can be obtained in $(\|x_0 - x^\star\| L_f)^2/\epsilon^2$ iterations

$$\text{With } \alpha_t = \frac{\|x_0 - x^\star\|}{L_f\sqrt{k}},\ t = 0, 1, \ldots, k-1,\ f(x_k) - f^\star \le \frac{\|x_0 - x^\star\| L_f}{\sqrt{k}}$$

Convergence Rate

- Suppose f has Lipschitz continuous gradient with L > 0 and use constant stepsize $\alpha \in (0, \frac{2}{L})$. Then,

$$f(x_k) - f^\star \leq \frac{2(f(x_0) - f^\star)\|x_0 - x^\star\|^2}{2\|x_0 - x^\star\|^2 + (f(x_0) - f^\star)\alpha(2 - L\alpha)k}.$$

  - R.H.S. achieves minimum when $\alpha = \frac{1}{L}$

$$f(x_k) - f^\star \leq \frac{2L\|x_0 - x^\star\|^2}{k + 4}$$

- Further suppose f is strongly convex with $\mu > 0$ and use constant stepsize $\alpha \in (0, \frac{2}{\mu + L}]$. Then,

$$\|x_k - x^\star\|^2 \leq q^k \|x_0 - x^\star\|^2, \text{ where } q = 1 - \frac{2\alpha\mu L}{\mu + L}.$$

  - q achieves minimum $\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^2$ when $\alpha = \frac{2}{\mu + L}$
  - L/μ is condition number

Gradient Projection Method

- Constrained convex optimization

$$\min_{x \in X} f(x)$$

  - f is convex and continuously differentiable
  - X is a nonempty, closed, and convex set

- Optimality condition

$$x^\star \in \arg\min_{x \in X} f(x) \Leftrightarrow \nabla f(x^\star)^T (x - x^\star) \geq 0, \forall x \in X$$

  - Unique if f is strictly convex

- Gradient projection method

$$x_{k+1} = P_X[x_k - \alpha_k \nabla f(x_k)] \text{ with } x_0 \in X$$

  - Projection operator

$$P_X[x] = \arg\min_{y \in X} \|y - x\| \qquad \text{(unique)}$$

  - Similar convergence analysis as unconstrained case, using properties of projection
  - Suppose ∇f is Lipschitz with $L > 0$. If $\alpha \in (0, 2/L)$, $f(x_k) - f^\star \leq O(1/k)$.

Important Facts of Projection

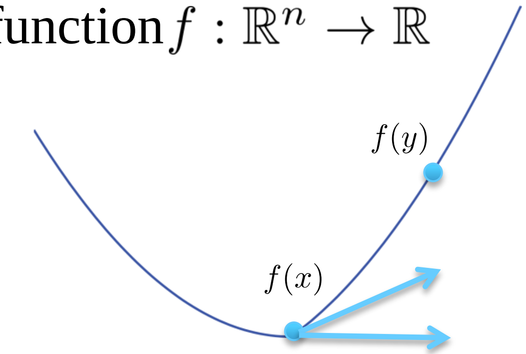
- For any $x \in \mathbb{R}^n$, $(x - P_X[x])^T(z - P_X[x]) \leq 0$, $\forall z \in X$.
- For any $x, y \in \mathbb{R}^n$, $\|P_X[x] - P_X[y]\| \leq \|x - y\|$.
- For any $z \in X$, $z \in \arg\min_{x \in X} f(x) \Leftrightarrow P_X[z - \alpha \nabla f(z)] = z$, $\forall \alpha > 0$.

Subgradient and Subdifferential

- Consider a convex and possibly non-differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$

- A vector $s \in \mathbb{R}^n$ is a *subgradient* of f at x if

$$f(y) \geq f(x) + s^T(y - x), \forall y \in \mathbb{R}^n$$



- *Subdifferential* at x (denoted as $\partial f(x)$): the set of all subgradients at x
  - If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.

- $\partial f(x)$ is nonempty, convex, and compact for all $x \in \mathbb{R}^n$.

- For any compact set $X \subset \mathbb{R}^n$, $\cup_{x \in X} \partial f(x)$ is bounded.

- $f'(x; d) = \max_{s \in \partial f(x)} s^T d$
  - $f'(x; d)$: directional derivative of f at x along direction d

    $f'(x : d) = \lim_{h \rightarrow 0} \frac{f(x+hd) - f(x)}{h}$

Subgradient Method

- Consider the (constrained) nonsmooth convex optimization problem

$$\min_{x \in X} f(x)$$

- Optimality condition

$$x^\star \in \arg\min_{x \in X} f(x) \Leftrightarrow \exists s \in \partial f(x^\star) \text{ such that } s^T(x - x^\star) \geq 0, \forall x \in X$$

  - For unconstrained case ($X = \mathbb{R}^n$), the condition becomes $0 \in \partial f(x^\star)$.

- Subgradient method

$$x_{k+1} = P_X[x_k - \alpha_k s_k] \text{ with } x_0 \in X \text{ and } s_k \in \partial f(x_k)$$

Convergence Analysis

- $$\min_{t\in\{0,1,\ldots,k\}} f(x_t) - f^\star \le \frac{\|x_0 - x^\star\|^2 + \sum_{t=0}^{k} \alpha_t^2 \|s_t\|^2}{2\sum_{t=0}^{k}\alpha_t}, \quad \forall k \ge 1$$
  - Very similar to the convergence analysis of the gradient descent method
- If every $\|s_k\|$ is bounded by $L > 0$, then accuracy ϵ can be obtained in $(\|x_0 - x^\star\| L)^2/\epsilon^2$ iterations.
- Averages behave better

$$\bar{x}_K = \frac{1}{K}\sum_{k=0}^{K-1} x_k$$

Note that $f(\bar{x}_K) \le \frac{1}{K}\sum_{k=0}^{K-1} f(x_k)$.

Choose stepsize $\alpha_k = \frac{\gamma}{\sqrt{K}}$ $\forall k = 0, 1, \ldots, K-1$, where $\gamma > 0$.

$$f(\bar{x}_K) - f^\star \le \frac{\|x_0 - x^\star\|^2 + \gamma^2 L^2}{\gamma\sqrt{K}}$$

Summary

- Convex set
- Convex function
  - Strictly convex, strongly convex, Lipschitz continuous gradient
- Gradient descent method
  - Smooth unconstrained convex optimization
  - Convergence performance
    - Lipschitz continuous function: $O(1/\epsilon^2)$
    - Lipschitz continuous gradient: sublinear convergence $O(1/k)$
    - Strongly convex function with Lipschitz continuous gradient: linear convergence $q^k,\ q \in [0,1)$
- Gradient projection method
  - Smooth constrained convex optimization
  - Facts of projection
  - Similar convergence results as gradient descent method
- Subgradient method
  - Subgradient and subdifferential
  - Nonsmooth convex optimization
  - Convergence complexity $O(1/\epsilon^2)$

References

- Y. Nesterov, *Introductory lectures on Convex Optimization: A Basic Course*. Norwell, MA: Kluwer Academic Publishers, 2004.
- D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.
- D. P. Bertsekas, A. Nedich, and A. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA: Athena Scientific, 2003.
- S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY: Cambridge University Press, 2004.