Lecture 19:

Convex Non-Smooth Optimization

April 2, 2007

Outline

- Convex non-smooth problems
- Examples
- Subgradients and subdifferentials
- Subgradient properties
- Operations with subgradients and subdifferentials
- Optimality conditions
- Subgradient algorithms

Convex-Constrained Non-smooth Minimization

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in C \end{aligned}$$

- **Characteristics**:
  - The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex and possibly non-differentiable
  - The set $C \subseteq \mathbb{R}^n$ is nonempty and convex
  - The optimal value f^* is finite
- Our focus here is *non-differentiability*

Renewed interest comes from large-scale problems and the need for distributed computations. Main questions:

- Where do such problems arise?
- How do we deal with non-differentiability? How can we solve them?

Where they arise

- Naturally in some **applications** (comm. nets, data fitting, neural-nets):
  - Least-squares problems

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^m \|h(w, x_j) - y_j\|^2 \\ \text{subject to} & w \succeq 0 \end{array}$$

    here (x_j, y_j), $j = 1, \ldots, m$ are the input-output pairs, w are weights (decision variables) to be optimized, h is convex possibly nonsmooth
- In **Lagrangian duality**

$$\begin{array}{ll} \text{minimize} & -q(\mu, \lambda) \\ \text{subject to} & \mu \succeq 0 \end{array}$$

  - A systematic approach for generating primal optimal bounds
  - A part of some primal-dual scheme
- In (sharp) **penalty approaches**

$$\min_{x \in C} \{f(x) + tP(g(x))\}$$

  where $t > 0$ is a penalty parameter and the penalty function is

$$P(u) = \sum_{j=1}^m \max\{u_j, 0\} \quad \text{or} \quad P(u) = \max\{u_1, \ldots, u_m, 0\}$$

Example: Optimization in Network Coding

Linear Cost Model

$$\begin{array}{ll} \text{minimize} & \sum_{(i,j)\in L} a_{ij} \max_{s\in S} x_{ij}^s \\ \text{subject to} & 0 \le \max_{s\in S} x_{ij}^s \le c_{ij} \quad \text{for all } (i,j) \in L \\ & \sum_{\{j|(i,j)\in L\}} x_{ij}^s - \sum_{\{j|(j,i)\}} x_{ji}^s = b_i^s \quad \text{for all } i \in \mathcal{N},\ s \in S \end{array}$$

- $\mathcal{N}$ is the set of nodes in the communication network
- S is the set of sessions [a session is a pair of nodes that communicate]
- L is the set of directed links
- (i,j) denotes a link originating at node i and ending at node j
- c_{ij} is the communication rate capacity of the link (i,j)
- a_{ij} is the cost for the link (i,j)
- x_{ij}^s is the communication rate for session s on link (i,j)
- **Problem**
  - Assign the rates x_{ij} so as to minimize the cost subject to link capacities and the network balance equations
  - The full knowledge of the network is not centralized

Non-differentiability Issue

- Any convex function is *differentiable almost everywhere on its domain* [the points of non-differentiability are countable - zero measure]
- However, the points of nondifferentiability cannot be ignored

$$\begin{aligned} &\text{minimize} && |x| \\ &\text{subject to} && x \in \mathbb{R} \end{aligned}$$

  - Consider a gradient descent method $d_k = -\nabla f(x_k)$ starting with $x^0 = 2$, and a backtracking line search with $\alpha = 1$ and $\sigma = 0.2$
  - Point $x = 1$ will be accepted since $f(2) = 2$, $f(1) = 1$, and $\nabla f(2) = 2$, so that $1 = f(2) - f(1) \geq 0.2\|\nabla f(2)\|^2 = 0.8$
  - Thus, $x_1 = 1$ and $f(1) = 1$, and $x = 0$ will be accepted: $x_2 = 0$. To proceed, we need $\nabla f(0)$, which is not defined!!!
- Is there a direction playing role of a gradient?

Subgradient

- What is a subgradient?
- For a convex and differentiable f, the linearization of f at a vector $\widehat{x} \in \operatorname{dom} f$ underestimates f at all points in $\operatorname{dom} f$

$$f(x) \geq f(\widehat{x}) + \nabla f(\widehat{x})^T (x - \widehat{x}) \qquad \text{for all } x \in \operatorname{dom} f$$

- For a differentiable function this linearization is unique at any given $\widehat{x} \in \operatorname{dom} f$
- A convex non-differentiable f may have multiple linearizations at some points in $\operatorname{dom} f$
- For such functions, a subgradient provides a linearization of f that underestimates f globally (at all points of the domain of f)

Definition of Subgradient and Subdifferential

Def. A vector $s \in \mathbb{R}^n$ is a **subgradient of f at $\widehat{x} \in \operatorname{dom} f$** when

$$f(x) \geq f(\widehat{x}) + s^T(x - \widehat{x}) \qquad \text{for all } x \in \operatorname{dom} f$$

Def. A **subdifferential of f at $\widehat{x} \in \operatorname{dom} f$** is the set of all subgradients s of f at $\widehat{x} \in \operatorname{dom} f$

- The **subdifferential of f at $\widehat{x}$** is denoted by $\partial f(\widehat{x})$
- When f is differentible at $\widehat{x}$, we have $\partial f(\widehat{x}) = \{\nabla f(\widehat{x})\}$ (the subdifferential is a singleton)
- Examples

$$f(x) = |x|, \qquad \partial f(0) = \begin{cases} \operatorname{sign}(x) & \text{for } x \neq 0 \\ [-1, 1] & \text{for } x = 0 \end{cases}$$

$$f(x) = \begin{cases} x^2 + 2|x| - 3 & \text{for } |x| > 1 \\ 0 & \text{for } |x| \leq 1 \end{cases}$$

Subgradients and Epigraph

- Let s be a subgradient of f at $\widehat{x}$:

$$f(x) \geq f(\widehat{x}) + s^T(x - \widehat{x}) \qquad \text{for all } x \in \operatorname{dom} f$$

- The subgradient inequality is equivalent to

$$-s^T\widehat{x} + f(\widehat{x}) \leq -s^T x + f(x) \qquad \text{for all } x \in \operatorname{dom} f$$

- Let $f(x) > -\infty$ for all $x \in \mathbb{R}^n$. Then

$$\operatorname{epi} f = \{(x, w) \mid f(x) \leq w,\ x \in \mathbb{R}^n\}$$

Thus, $-s^T\widehat{x} + f(\widehat{x}) \leq -s^T x + w$ for all $(x, w) \in \operatorname{epi} f$, equivalent to

$$\begin{bmatrix} -s \\ 1 \end{bmatrix}^T \begin{bmatrix} \widehat{x} \\ f(\widehat{x}) \end{bmatrix} \leq \begin{bmatrix} -s \\ 1 \end{bmatrix}^T \begin{bmatrix} x \\ w \end{bmatrix} \qquad \text{for all } (x, w) \in \operatorname{epi} f$$

Therefore, **the hyperplane**

$$H = \left\{(x, \gamma) \in \mathbb{R}^{n+1} \mid (-s, 1)^T(x, \gamma) = (-s, 1)^T(\widehat{x}, f(\widehat{x}))\right\}$$

supports $\operatorname{epi} f$ **at the vector** $(\widehat{x}, f(\widehat{x}))$

Subdifferential Set Properties

• When nonempty, a subdifferential $\partial f(\widehat{x})$ is **convex and closed**

Existence Theorem Let f be convex with $f(x) > -\infty$ for all x and a nonempty $\operatorname{int}(\operatorname{dom} f)$. Then, the subdifferential $\partial f(\widehat{x})$ **is nonempty compact convex set for every $\widehat{x}$ in the interior of $\operatorname{dom} f$.**

Proof: $\partial f(\widehat{x})$ *Nonempty.*

Let $\widehat{x}$ be in the interior of $\operatorname{dom} f$. The vector $(\widehat{x}, f(\widehat{x}))$ does not belong to the interior of $\operatorname{epi} f$. The epigraph $\operatorname{epi} f$ is convex and by the *Supporting Hyperlane Theorem*, there is a vector $(d, \beta) \in \mathbb{R}^{n+1}$, $(d, \beta) \neq 0$ such that

$$d^T\widehat{x} + \beta f(\widehat{x}) \leq d^T x + \beta w \qquad \text{for all } (x, w) \in \operatorname{epi} f$$

By $f(x) > -\infty$ for all x, we have $\operatorname{epi} f = \{(x, w) \mid f(x) \leq w,\ x \in \mathbb{R}^n\}$. Hence,

$$d^T\widehat{x} + \beta f(\widehat{x}) \leq d^T x + \beta w \qquad \text{for all } x \in \mathbb{R}^n \text{ with } f(x) \leq w$$

We must have $\beta \geq 0$. We cannot have $\beta = 0$ (it would imply $d = 0$). Dividing by β, we see that $-d/\beta$ is a subgradient of f at $\widehat{x}$

$\partial f(\widehat{x})$ *Bounded*. By the subgradient inequality, we have

$$f(x) \geq f(\widehat{x}) + s^T(x - \widehat{x}) \qquad \text{for all } x \in \operatorname{dom} f$$

Suppose that the subdifferential $\partial f(\widehat{x})$ is unbounded. Let s_k be a sequence of subgradients in $\partial f(\widehat{x})$ with $\|s_k\| \to \infty$.

Since $\widehat{x}$ lies in the interior of domain, there exists a $\delta > 0$ such that $\widehat{x} + \delta y \in \operatorname{dom} f$ for any $y \in \mathbb{R}^n$. Letting $x = \widehat{x} + \delta \frac{s_k}{\|s_k\|}$ for any k, we have

$$f\left(\widehat{x} + \delta \frac{s_k}{\|s_k\|}\right) \geq f(\widehat{x}) + \delta\|s_k\| \qquad \text{for all } k$$

As $k \to \infty$, we have $f\left(\widehat{x} + \delta \frac{s_k}{\|s_k\|}\right) - f(\widehat{x}) \to \infty$.

However, this relation contradicts the continuity of f at $\widehat{x}$. [Recall, a convex function is continuous over the interior of its domain.]

Example Consider $f(x) = -\sqrt{x}$ with $\operatorname{dom} f = \{x \mid x \geq 0\}$. We have $\partial f(0) = \emptyset$. Note that 0 is not in the interior of the domain of f

NOTE When f is closed convex and $\operatorname{dom} f \neq \emptyset$, then

$$f(x) > -\infty \text{ for all } x \in \mathbb{R}^n$$

Operations with Subdifferential

Let f, f_1, and f_2 be *convex* functions with *nonempty domains*

- **Scaling** For $\lambda > 0$, the function λf is convex and

$$\partial(\lambda f)(x) = \lambda \partial f(x) \qquad \text{for all } x \in \operatorname{int}(\operatorname{dom} f)$$

- **Sum** The function $f_1 + f_2$ is convex, and

$$\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$$

for all $x \in \operatorname{int}(\operatorname{dom} f) = \operatorname{int}(\operatorname{dom} f_1) \cap \operatorname{int}(\operatorname{dom} f_2)$

- **Composition with Affine Mapping** Let $\phi(x) = f(Ax + b)$. Then, the function ϕ is convex and

$$\partial\phi(x) = A^T \partial f(Ax + b) \qquad \text{for all } x \in \operatorname{int}(\operatorname{dom} \phi)$$

where $\operatorname{dom} \phi = \{x \mid Ax + b \in \operatorname{dom} f\}$

Max-Type Functions

Let $f_i(x), i = 1, \ldots, m$ be convex functions with nonempty domains

- **Max-Function** The function $f(x) = \max_{1 \le i \le m} f_i(x)$ is convex and

$$\partial f(x) = \operatorname{conv}\left(\{\partial f_i(x) \mid i \in I(x)\}\right) \qquad \text{for all } x \in \operatorname{int}(\operatorname{dom} f)$$

where $I(x) = \{i \mid f_i(x) = f(x)\}$ and $\operatorname{dom} f = \cap_{i=1}^{m} \operatorname{dom} f_i$

Example $f(x) = \max\{1 - x, 1 + x\}$, we have

$$f(x) = \begin{cases} f_1(x) & \text{for } x < 0 \\ f_2(x) & \text{for } x > 0 \\ f_1(x) \text{ or } f_2(x) & \text{for } x = 0 \end{cases} \qquad I(x) = \begin{cases} \{1\} & \text{for } x < 0 \\ \{2\} & \text{for } x > 0 \\ \{1, 2\} & \text{for } x = 0 \end{cases}$$

$\partial f(x) = -1$ for $x < 0$, $\partial f(x) = 1$ for $x > 0$, and $\partial f(0) = [-1, 1]$

Examples

- $f(x) = \sum_{i=1}^{m} |a_i^T x - b_i|$.
  Let

$$I_-(x) = \{i \mid a_i^T x - b_i < 0\}$$
$$I_+(x) = \{i \mid a_i^T x - b_i > 0\}$$
$$I_0(x) = \{i \mid a_i^T x - b_i = 0\}$$

  Then

$$\partial f(x) = \sum_{i \in I_+(x)} a_i - \sum_{i \in I_-(x)} a_i + \sum_{i \in I_0(x)} \operatorname{conv}(\{-a_i, a_i\})$$

- **Euclidian Norm** $f(x) = \|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$. Then

$$\partial f(x) = \frac{x}{\|x\|} \qquad \text{for } x \neq 0$$
$$\partial f(0) = \{x \mid \|x\| \leq 1\} = B_2(0, 1)$$

Sup-Type Functions

Let $\phi(x, z)$ be function convex in x for every $z \in Z$, with $Z \neq \emptyset$

- **Sup-Function** The function $f(x) = \sup_{z \in Z} \phi(x, z)$ is convex and

$$\text{conv}\left(\{\partial_x \phi(x, z) \mid z \in Z^*(x)\}\right) \subset \partial f(x) \qquad \text{for all } x \in \text{dom } f$$

$$Z^*(x) = \{z \mid \phi(x, z) = f(x)\} \qquad \text{dom } f = \left\{x \mid \sup_{z \in Z} \phi(x, z) < \infty\right\}$$

Proof Let $z \in Z^*(\hat{x})$ and $s \in \partial_x \phi(\hat{x}, z)$. Then, for any $x \in \text{dom } f$:

$$f(x) \geq \phi(x, z) \geq \phi(\hat{x}, z) + s^T(x - \hat{x}) = f(\hat{x}) + s^T(x - \hat{x})$$

- When $\phi(x, z)$ is differentiable with respect to x for every $z \in Z$:

$$\text{conv}\left(\{\nabla_x \phi(x, z) \mid z \in Z^*(x)\}\right) \subset \partial f(x) \qquad \text{for all } x \in \text{dom } f$$

Optimality Conditions: Unconstrained Case

Unconstrained optimization

$$\text{minimize } f(x)$$

Assumption

- The function f is convex (non-differentiable) and *proper*
  [f proper means $f(x) > -\infty$ for all x and $\operatorname{dom} f \neq \emptyset$]

Theorem Under this assumption, a vector x^* **minimizes f over $\mathbb{R}^n$ if and only if**

$$0 \in \partial f(x^*)$$

- The result is a generalization of $\nabla f(x^*) = 0$
- Proof x^* is optimal if and only if $f(x) \geq f(x^*)$ for all x, or equivalently
  $$f(x) \geq f(x^*) + 0^T(x - x^*) \qquad \text{for all } x \in \mathbb{R}^n$$
  Thus, x^* is optimal if and only if $0 \in \partial f(x^*)$

Examples

- The function $f(x) = |x|$

$$\partial f(0) = \begin{cases} \operatorname{sign}(x) & \text{for } x \neq 0 \\ [-1, 1] & \text{for } x = 0 \end{cases}$$

The minimum is at $x^* = 0$, and evidently $0 \in \partial f(0)$

- The function $f(x) = \|x\|$

$$\partial f(x) = \begin{cases} \frac{x}{\|x\|} & \text{for } x \neq 0 \\ \{s \mid \|s\| \leq 1\} & \text{for } x = 0 \end{cases}$$

Again, the minimum is at $x^* = 0$ and $0 \in \partial f(0)$

- The function $f(x) = \max\{x^2 + 2x - 3, x^2 - 2x - 3, 4\}$

$$f(x) = \begin{cases} x^2 - 2x - 3 & \text{for } x < -1 \\ 4 & \text{for } x \in [-1, 1] \\ x^2 + 2x - 3 & \text{for } x > 1 \end{cases}$$

$$\partial f(x) = \begin{cases} 2x - 2 & \text{for } x > 1 \\ [-4, 0] & \text{for } x = -1 \\ 0 & \text{for } x \in (-1, 1) \\ [0, 4] & \text{for } x = 1 \\ 2x + 2 & \text{for } x > 1 \end{cases}$$

- The optimal set is $X^* = [-1, 1]$
- For every $x^* \in X^*$, we have $0 \in \partial f(x^*)$

Optimality Conditions: Constrained Case

Constrained optimization

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in C \end{aligned}$$

Assumption

- The function f is convex (non-differentiable) and $proper$
- The set C is nonempty and convex

Theorem Under this assumption, a vector $x^* \in C$ minimizes f over the set C if and only if there exists a subgradient $d \in \partial f(x^*)$ such that

$$d^T(x - x^*) \geq 0 \qquad \text{for all } x \in C$$

- The result is a generalization of $\nabla f(x^*)^T(x - x^*) \geq 0$ for $x \in C$