LECTURE 18

LECTURE OUTLINE

- Approximate subgradient methods
- ϵ-subdifferential
- ϵ-subgradient methods
- Incremental subgradient methods

APPROXIMATE SUBGRADIENT METHODS

- Consider minimization of

$$f(x) = \sup_{z \in Z} \phi(x, z)$$

where $Z \subset \Re^m$ and $\phi(\cdot, z)$ is convex for all $z \in Z$ (dual minimization is a special case).

- To compute subgradients of f at $x \in \mathrm{dom}(f)$, we find $z_x \in Z$ attaining the supremum above. Then

$$g_x \in \partial\phi(x, z_x) \qquad \Rightarrow \qquad g_x \in \partial f(x)$$

- Two potential areas of difficulty:
  - For subgradient method, we need to solve exactly the above maximization over $z \in Z$.
  - For steepest descent, we need all the subgradients, and then there are convergence difficulties to contend with.

- In this lecture we address the first difficulty, in the next lecture the second.

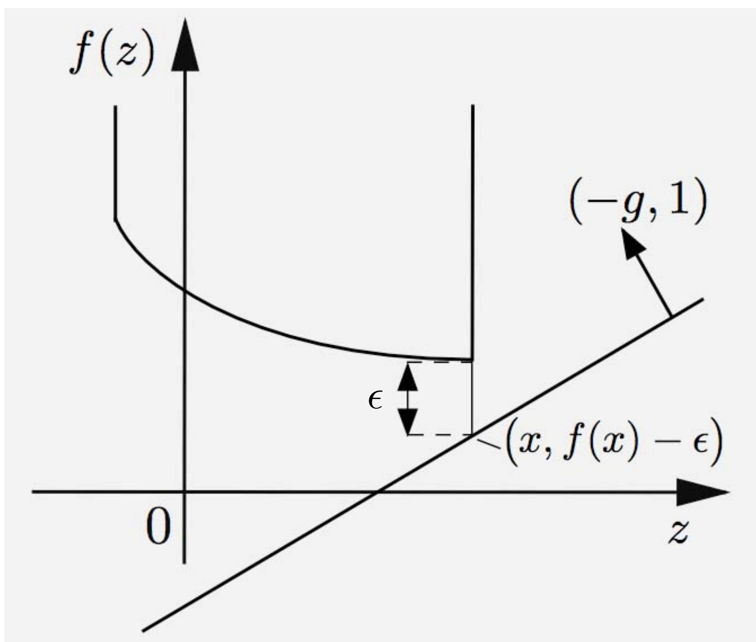
- We consider methods that use "approximate" subgradients.

ϵ-SUBDIFFERENTIAL

- We enlarge $\partial f(x)$ so that we take into account "nearby" subgradients.

- Fot a proper convex $f : \Re^n \mapsto (-\infty, \infty]$ and $\epsilon > 0$, we say that a vector g is an *ϵ-subgradient* of f at a point $x \in \text{dom}(f)$ if

$$f(z) \geq f(x) + (z - x)'g - \epsilon, \qquad \forall\, z \in \Re^n$$



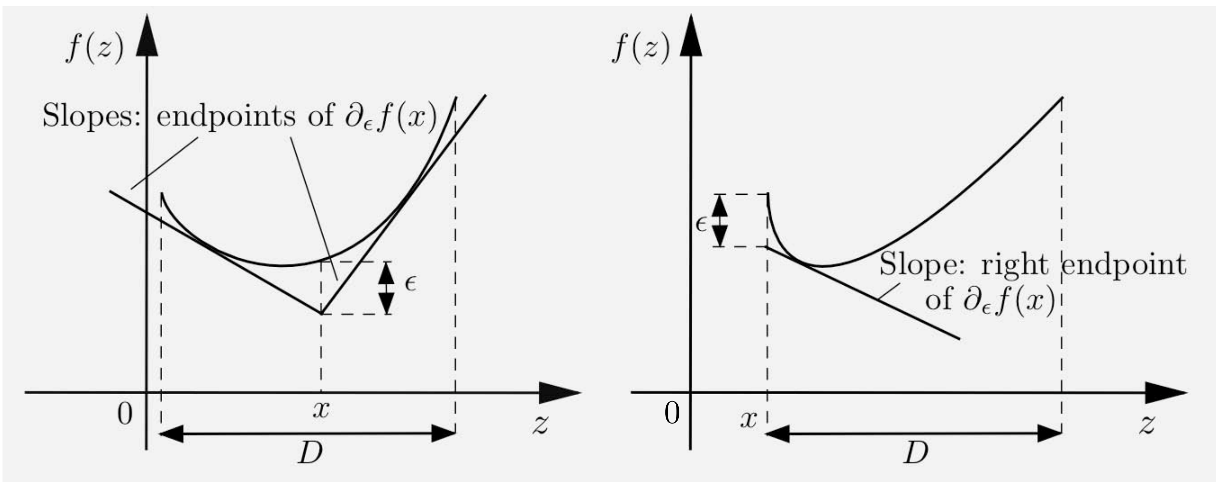
- The *ϵ-subdifferential* $\partial_\epsilon f(x)$ is the set of all ϵ-subgradients of f at x. By convention, $\partial_\epsilon f(x) = \emptyset$ for $x \notin \text{dom}(f)$.

- We have $\cap_{\epsilon \downarrow 0} \partial_\epsilon f(x) = \partial f(x)$ and

$$\partial_{\epsilon_1} f(x) \subset \partial_{\epsilon_2} f(x) \qquad \text{if } 0 < \epsilon_1 < \epsilon_2$$

PROPERTIES OF ϵ-SUBDIFFERENTIALS

- Assume that f is closed proper convex, $\epsilon > 0$.
- $\partial_\epsilon f(x)$ is **nonempty** and closed for all $x \in \mathrm{dom}(f)$. (Use nonvertical separating hyperplane theorem.)



- $\partial_\epsilon f(x)$ is compact iff $x \in \mathrm{int}\big(\mathrm{dom}(f)\big)$. True in particular, if f is real-valued.
- **Neighborhood/continuity property:** Subgradients at nearby points are ϵ-subgradients at given point (for sufficiently large ϵ).
- The support function of $\partial_\epsilon f(x)$ is

$$\sigma_{\partial_\epsilon f(x)}(y) = \sup_{g \in \partial_\epsilon f(x)} y'g = \inf_{\alpha > 0} \frac{f(x + \alpha y) - f(x) + \epsilon}{\alpha}$$

CALCULATION OF AN ϵ-SUBGRADIENT

- Consider minimization of

$$f(x) = \sup_{z \in Z} \phi(x, z), \tag{1}$$

where $x \in \Re^n$, $z \in \Re^m$, Z is a subset of $\Re^m$, and $\phi : \Re^n \times \Re^m \mapsto (-\infty, \infty]$ is a function such that $\phi(\cdot, z)$ is convex and closed for each $z \in Z$.

- How to calculate ϵ-subgradient at $x \in \mathrm{dom}(f)$?
- Let $z_x \in Z$ attain the supremum within $\epsilon \geq 0$ in Eq. (1), and let g_x be some subgradient of the convex function $\phi(\cdot, z_x)$.
- For all $y \in \Re^n$, using the subgradient inequality,

$$\begin{aligned} f(y) = \sup_{z \in Z} \phi(y, z) &\geq \phi(y, z_x) \\ &\geq \phi(x, z_x) + g_x'(y - x) \geq f(x) - \epsilon + g_x'(y - x) \end{aligned}$$

i.e., g_x is an ϵ-subgradient of f at x, so

$$\phi(x, z_x) \geq \sup_{z \in Z} \phi(x, z) - \epsilon \text{ and } g_x \in \partial \phi(x, z_x)$$

$$\Rightarrow \qquad g_x \in \partial_\epsilon f(x)$$

ϵ-SUBGRADIENT METHOD

- Can be viewed as an approximate subgradient method, using an ϵ-subgradient in place of a subgradient.

- **Problem:** Minimize convex $f : \Re^n \mapsto \Re$ over a closed convex set X.

- **Method:**

$$x_{k+1} = P_X(x_k - \alpha_k g_k)$$

where g_k is an ϵ_k-subgradient of f at x_k, α_k is a positive stepsize, and $P_X(\cdot)$ denotes projection on X.

- Can be viewed as subgradient method with “errors”.

CONVERGENCE ANALYSIS

- **Basic inequality:** If $\{x_k\}$ is the ϵ-subgradient method sequence, for all $y \in X$ and $k \geq 0$

$$\|x_{k+1}-y\|^2 \leq \|x_k-y\|^2 - 2\alpha_k\big(f(x_k)-f(y)-\epsilon_k\big)+\alpha_k^2\|g_k\|$$

- Replicate the entire convergence analysis for subgradient methods, but carry along the ϵ_k terms.

- **Example:** Constant $\alpha_k \equiv \alpha$, constant $\epsilon_k \equiv \epsilon$. Assume $\|g_k\| \leq c$ for all k. For any optimal x^*,

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - 2\alpha\big(f(x_k)-f^*-\epsilon\big)+\alpha^2 c^2,$$

so the distance to x^* decreases if

$$0 < \alpha < \frac{2\big(f(x_k)-f^*-\epsilon\big)}{c^2}$$

or equivalently, if x_k is outside the level set

$$\left\{x \;\middle|\; f(x) \leq f^* + \epsilon + \frac{\alpha c^2}{2}\right\}$$

- **Example:** If $\alpha_k \to 0$, $\sum_k \alpha_k \to \infty$, and $\epsilon_k \to \epsilon$, we get convergence to the ϵ-optimal set.

INCREMENTAL SUBGRADIENT METHODS

- Consider minimization of sum

$$f(x) = \sum_{i=1}^{m} f_i(x)$$

- Often arises in duality contexts with m: **very large** (e.g., separable problems).

- Incremental method **moves x along a subgradient g_i of a component function f_i** NOT the (expensive) subgradient of f, which is $\sum_i g_i$.

- View an iteration as a cycle of m subiterations, one for each component f_i.

- Let x_k be obtained after k cycles. To obtain x_{k+1}, do one more cycle: Start with $\psi_0 = x_k$, and set $x_{k+1} = \psi_m$, after the m steps

$$\psi_i = P_X(\psi_{i-1} - \alpha_k g_i), \qquad i = 1, \ldots, m$$

with g_i being a subgradient of f_i at ψ_{i-1}.

- **Motivation is faster convergence.** A cycle can make much more progress than a subgradient iteration with essentially the same computation.

CONNECTION WITH ϵ-SUBGRADIENTS

- **Neighborhood property:** If x and x are "near" each other, then subgradients at x can be viewed as ϵ-subgradients at x, with ϵ "small."

- If $g \in \partial f(x)$, we have for all $z \in \Re^n$,

$$\begin{aligned} f(z) &\geq f(x) + g'(z - x) \\ &\geq f(x) + g'(z - x) + f(x) - f(x) + g'(x - x) \\ &\geq f(x) + g'(z - x) - \epsilon, \end{aligned}$$

where $\epsilon = |f(x) - f(x)| + \|g\| \cdot \|x - x\|$. Thus, $g \in \partial_\epsilon f(x)$, with ϵ: small when x is near x.

- The incremental subgradient iter. is an ϵ-subgradient iter. with $\epsilon = \epsilon_1 + \cdots + \epsilon_m$, where ϵ_i is the "error" in ith step in the cycle (ϵ_i: Proportional to α_k).

- Use

$$\partial_{\epsilon_1} f_1(x) + \cdots + \partial_{\epsilon_m} f_m(x) \subset \partial_\epsilon f(x),$$

where $\epsilon = \epsilon_1 + \cdots + \epsilon_m$, to approximate the ϵ-subdifferential of the sum $f = \sum_{i=1}^m f_i$.

- Convergence to optimal if $\alpha_k \to 0$, $\sum_k \alpha_k \to \infty$.

CONVERGENCE OF INCREMENTAL SUBGR.

- Problem

$$\min_{x\in X} \sum_{i=1}^{m} f_i(x)$$

- Incremental subgradient method

$$x_{k+1} = \psi_{m,k}, \quad \psi_{i,k} = [\psi_{i-1,k} - \alpha_k g_{i,k}]^+, \; i = 1, \ldots, m$$

starting with $\psi_{0,k} = x_k$, where $g_{i,k}$ is a subgradient of f_i at $\psi_{i-1,k}$.

- Analysis parallels/extends the one for nonincremental subgradient methods

- **Key Lemma:** For all $y \in X$ and k,

$$||x_{k+1}-y||^2 \le ||x_k-y||^2 - 2\alpha_k\big(f(x_k)-f(y)\big) + \alpha_k^2 C^2,$$

where $C = \sum_{i=1}^m C_i$ and

$$C_i = \sup_k \big\{ ||g|| \mid g \in \partial f_i(x_k) \cup \partial f_i(\psi_{i-1,k}) \big\}$$

ERROR BOUND: CONSTANT STEPSIZE

- For $\alpha_k \equiv \alpha$, we have

$$\inf_{k\geq 0} f(x_k) \leq f^* + \frac{\alpha C^2}{2} \leq f^* + \frac{\alpha m^2 C_0^2}{2}$$

where

$$C_0 = \max\{C_1, \ldots, C_m\}$$

is the max component subgradient bound. (Comparable error to the nonincremental method.)

- **Sharpness of the estimate:** There are problems for which the upper bound is (almost) sharp with cyclic order of processing the component functions (see the end-of-chapter problems).

- **Lower bound on the error:** There is a problem, where even with best processing order,

$$f^* + \frac{\alpha m C_0^2}{2} \leq \inf_{k\geq 0} f(x_k)$$

where

$$C_0 = \max\{C_1, \ldots, C_m\}$$

- **Question:** Is it possible to improve the upper bound by optimizing the order of processing the component functions?

RANDOMIZED ORDER METHODS

$$x_{k+1} = \left[x_k - \alpha_k g(\omega_k, x_k)\right]^+$$

where ω_k is a random variable taking equiprobable values from the set $\{1, \ldots, m\}$, and $g(\omega_k, x_k)$ is a subgradient of the component f_{ω_k} at x_k.

- Assumptions:

  (a) $\{\omega_k\}$ is a sequence of independent random variables. Furthermore, the sequence $\{\omega_k\}$ is independent of the sequence $\{x_k\}$.

  (b) The set of subgradients $\big\{g(\omega_k, x_k) \mid k = 0, 1, \ldots\big\}$ is bounded, i.e., there exists a positive constant C_0 such that with prob. 1

$$||g(\omega_k, x_k)|| \leq C_0, \qquad \forall\, k \geq 0$$

- Stepsize Rules:
  - Constant: $\alpha_k \equiv \alpha$
  - Diminishing: $\sum_k \alpha_k = \infty$, $\sum_k (\alpha_k)^2 < \infty$
  - Dynamic

RANDOMIZED METHOD W/ CONSTANT STEP

- With probability 1

$$\inf_{k\geq 0} f(x_k) \leq f^* + \frac{\alpha m C_0^2}{2}$$

A better/sharp error bound!

Proof: By adapting key lemma, for all $y \in X$, k

$$||x_{k+1}-y||^2 \leq ||x_k-y||^2 - 2\alpha\big(f_{\omega_k}(x_k) - f_{\omega_k}(y)\big) + \alpha^2 C_0^2$$

Take conditional expectation with $\mathcal{F}_k = \{x_0, \ldots, x_k\}$

$$\begin{aligned} E\big\{||x_{k+1}-y||^2 \mid \mathcal{F}_k\big\} &\leq ||x_k - y||^2 \\ &\qquad - 2\alpha E\big\{f_{\omega_k}(x_k) - f_{\omega_k}(y) \mid \mathcal{F}_k\big\} + \alpha^2 C_0^2 \\ &= ||x_k - y||^2 - 2\alpha \sum_{i=1}^{m} \frac{1}{m}\big(f_i(x_k) - f_i(y)\big) + \alpha^2 C_0^2 \\ &= ||x_k - y||^2 - \frac{2\alpha}{m}\big(f(x_k) - f(y)\big) + \alpha^2 C_0^2, \end{aligned}$$

where the first equality follows since ω_k takes the values $1, \ldots, m$ with equal probability $1/m$.

PROOF CONTINUED I

- Fix $\gamma > 0$, consider the level set L_γ defined by

$$L_\gamma = \left\{ x \in X \mid f(x) < f^* + \frac{2}{\gamma} + \frac{\alpha m C_0^2}{2} \right\}$$

and let $y_\gamma \in L_\gamma$ be such that $f(y_\gamma) = f^* + \frac{1}{\gamma}$. Define a new process $\{\hat{x}_k\}$ as follows

$$\hat{x}_{k+1} = \begin{cases} \left[\hat{x}_k - \alpha g(\omega_k, \hat{x}_k)\right]^+ & \text{if } \hat{x}_k \notin L_\gamma, \\ y_\gamma & \text{otherwise,} \end{cases}$$

where $\hat{x}_0 = x_0$. We argue that $\{\hat{x}_k\}$ (and hence also $\{x_k\}$) will eventually enter each of the sets L_γ.

Using key lemma with $y = y_\gamma$, we have

$$E\left\{ ||\hat{x}_{k+1} - y_\gamma||^2 \mid \mathcal{F}_k \right\} \le ||\hat{x}_k - y_\gamma||^2 - z_k,$$

where

$$z_k = \begin{cases} \frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_\gamma)\big) - \alpha^2 C_0^2 & \text{if } \hat{x}_k \notin L_\gamma, \\ 0 & \text{if } \hat{x}_k = y_\gamma. \end{cases}$$

PROOF CONTINUED II

- If $\hat{x}_k \notin L_\gamma$, we have

$$
\begin{aligned}
z_k &= \frac{2\alpha}{m}\big(f(\hat{x}_k) - f(y_\gamma)\big) - \alpha^2 C_0^2 \\
&\geq \frac{2\alpha}{m}\left(f^* + \frac{2}{\gamma} + \frac{\alpha m C_0^2}{2} - f^* - \frac{1}{\gamma}\right) - \alpha^2 C_0^2 \\
&= \frac{2\alpha}{m\gamma}.
\end{aligned}
$$

Hence, as long as $\hat{x}_k \notin L_\gamma$, we have

$$
E\big\{||\hat{x}_{k+1} - y_\gamma||^2 \mid \mathcal{F}_k\big\} \leq ||\hat{x}_k - y_\gamma||^2 - \frac{2\alpha}{m\gamma}
$$

This, cannot happen for an infinite number of iterations, so that $\hat{x}_k \in L_\gamma$ for sufficiently large k (the Supermartingale Convergence Theorem is used here; see the notes.) Hence, in the original process we have

$$
\inf_{k\geq 0} f(x_k) \leq f^* + \frac{2}{\gamma} + \frac{\alpha m C_0^2}{2}
$$

with probability 1. Letting $\gamma \to \infty$, we obtain $\inf_{k\geq 0} f(x_k) \leq f^* + \alpha m C_0^2/2$. **Q.E.D.**

A CONVERGENCE RATE RESULT

- Let $\alpha_k \equiv \alpha$ in the randomized method. Then, for any positive scalar ϵ, we have with prob. 1

$$\min_{0 \leq k \leq N} f(x_k) \leq f^* + \frac{\alpha m C_0^2 + \epsilon}{2},$$

where N is a random variable with

$$E\{N\} \leq \frac{m\big(d(x_0, X^*)\big)^2}{\alpha\epsilon}$$

where $d(x_0, X^*)$ is the min distance of x_0 to the optimal set X^*.

- Compare w/ the deterministic method. It is guaranteed to reach after processing no more than

$$K = \frac{m\big(d(x_0, X^*)\big)^2}{\alpha\epsilon}$$

components the level set

$$\left\{x \;\middle|\; f(x) \leq f^* + \frac{\alpha m^2 C_0^2 + \epsilon}{2}\right\}$$