Convex Optimization II

Lecture 8: Subgradient Methods

Hamed Shah-Mansouri

Department of Electrical Engineering

Sharif University of Technology

1400-2

References

[1] S. Boyd and L. Vandenberghe, “Subgradients,” *Notes for EE364b, Stanford University*, May 2014.

[2] S. Boyd, “Subgradient Methods,” *Notes for EE364b, Stanford University*, May 2014.

Thanks to Professor Stephen Boyd for the slides used in this lecture.

BASIC INEQUALITY

- Recall basic inequality for convex differentiable $f : \mathbf{R}^n \to \mathbf{R}$

$$f(z) \geq f(x) + \nabla f(x)^T (z - x)$$

  for all $z \in \textbf{dom } f$

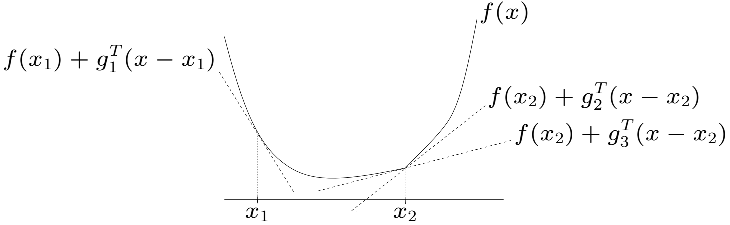
- First-order approximation of f at x is the global underestimator.

- $(\nabla f(x), -1)$ supports epi f at $(x, f(x))$

- What if f is not differentiable?

Subgradient of a function

- A vector $g \in \mathbf{R}^n$ is a subgradient of $f : \mathbf{R}^n \to \mathbf{R}$ (not necessarily convex) at $x \in \mathbf{R}^n$ if
$$f(z) \geq f(x) + g^T(z - x) \qquad \text{for all } z \in \mathbf{dom}\ f$$



- g_2, g_3 are subgradients at x_2
- g_1 is a subgradient at x_1

Subgradient of a Function (cont.)

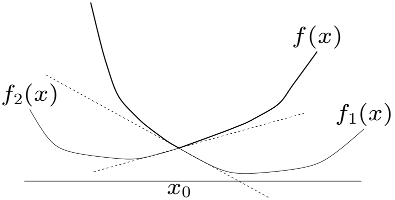
- g is a subgradient of f at x iff $(g, -1)$ supports epi f at $(x, f(x))$
- g is a subgradient iff $f(x) + g^T(z - x)$ is a global (affine) underestimator of f
- if f is convex and differentiable, $\nabla f(x)$ is a subgradient of f at x

Subgradients come up in several contexts:

- algorithms for nondifferentiable convex optimization
- convex analysis, e.g., optimality conditions, duality for nondifferentiable problems

EXAMPLE

$f = \max\{f_1, f_2\}$ with f_1, f_2 convex and differentiable



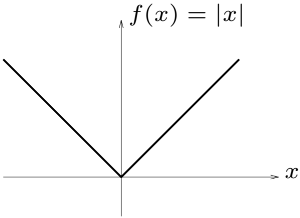
- $f_1(x_0) > f_2(x_0)$: unique subgradient $g = \nabla f_1(x_0)$
- $f_2(x_0) > f_1(x_0)$: unique subgradient $g = \nabla f_2(x_0)$
- $f_1(x_0) = f_2(x_0)$: subgradients form a line segment $[\nabla f_1(x_0), \nabla f_2(x_0)]$

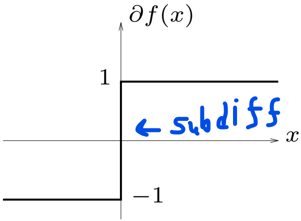
SUBDIFFERENTIAL

- A function f is called subdifferentiable at x if there exists at least one subgradient at x.
- The set of all subgradients of f at x is called the subdifferential of f at x, and is denoted $\partial f(x)$.
- A function f is called subdifferentiable if it is subdifferentiable at all $x \in \mathbf{dom}\ f$.
- The subdifferential $\partial f(x)$ is a closed convex set, even if f is not convex.

Example

$f(x) = |x|$





$$\partial f(x) = \begin{cases} \{-1\}, & x < 0 \\ \{1\}, & x > 0 \\ [-1, 1], & x = 0 \end{cases}$$

PROPERTIES

- Subgradients of differentiable functions
  - If f is convex and differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- The minimum of a nondifferentiable function
  - A point $x^\star$ is a minimizer of a function f (not necessarily convex) if and only if f is subdifferentiable at $x^\star$ and $0 \in \partial f(x^\star)$
  - The condition $0 \in \partial f(x^\star)$ reduces to $\nabla f(x^\star) = 0$ if f is convex and differentiable at $x^\star$

Some Rules for Constructing Subgradients of Convex Functions

- Nonnegative scaling: For $\alpha \geq 0$, $\partial(\alpha f)(x) = \alpha \partial f(x)$.

- Sum: Suppose $f = f_1 + \cdots + f_m$, where $f_1, \ldots, f_m$ are convex functions. We have $\partial f(x) = \partial f_1(x) + \cdots + \partial f_m(x)$.

- Affine transformations of domain: Suppose f is convex, and $h(x) = f(Ax + b)$. Then, $\partial h(x) = A^T \partial f(Ax + b)$.

- Pointwise maximum: Consider $f(x) = \max_{i=1,\ldots,m} f_i(x)$, where f_i are convex and subdifferentiable. Let k be any index for which $f_k(x) = f(x)$ and let $g \in \partial f_k(x)$. Then, $g \in \partial f(x)$. This follows from

$$f(z) \geq f_k(z) \geq f_k(x) + g^T(z - x) = f(x) + g^T(z - x)$$

- Other rules can be found in [1].

Subgradient method

- Simple algorithm to minimize nondifferentiable convex function f.
- Advantage: Allow simple distributed algorithm for a problem when combined with primal or dual decomposition.
- Drawback: Convergence can be slow.
- Consider an unconstrained minimization problem, where function $f : \mathbf{R}^n \to \mathbf{R}$ is convex.
- Basic subgradient method: At each iteration k, take a step in the direction of negative subgradient
$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}, \tag{1}$$
where $x^{(k)}$ is the kth iterate, $g^{(k)} \in \partial f(x^{(k)})$ is *any* subgradient of f at $x^{(k)}$, $\alpha_k > 0$ is the kth step size.
- Not a descent method, so we keep track of the best point so far.
$$f_{\text{best}}^{(k)} = \min_{i=1,\dots,k} f(x^{(i)}) = \min\{f_{\text{best}}^{(k-1)}, f(x^{(k)})\}$$
Since $f_{\text{best}}^{(k)}$ is decreasing, it has a limit.

Basic Step Size Rules

Step sizes are *determined* before the algorithm is run. They do not depend on any data computed during the algorithm.

- Constant step size: $\alpha_k = \alpha$ (positive constant, independent of k)
- Constant step length: $\alpha_k = \frac{\gamma}{||g^{(k)}||_2}$, where $\gamma > 0$ (so $||x^{(k+1)} - x^{(k)}||_2 = \gamma$)
- Square summable but not summable: Step sizes satisfy

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

  Example: $\alpha_k = a/(b+k)$, where $a > 0$ and $b \geq 0$.

- Nonsummable diminishing: Step sizes satisfy

$$\alpha_k \geq 0, \qquad \lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

  Example: $\alpha_k = a/\sqrt{k}$, where $a > 0$.

Assumptions on the Convergence Results

- Assumption A1: There is a minimizer of f, say $x^\star$. We have

$$f^\star = \inf_x f(x) > -\infty, \text{with } f(x^\star) = f^\star.$$

- Assumption A2: Norm of the subgradients is bounded. There exists a G such that $||g^{(k)}||_2 \leq G$, for all k.

- Assumption A3: A number R is known that satisfies $\|x^{(1)} - x^\star\|_2 \leq R$.

The above assumptions are stronger than needed, just to simplify proofs.

Convergence Proof

- Key quantity: *Euclidean distance to the optimal set*, not the function value
- Let $x^\star$ be any minimizer of f.

$$
\begin{aligned}
\|x^{(k+1)} - x^\star\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^\star\|_2^2 \\
&= \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^\star) + \alpha_k^2 \|g^{(k)}\|_2^2 \\
&\leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^\star) + \alpha_k^2 \|g^{(k)}\|_2^2
\end{aligned}
$$

where $f^\star = f(x^\star)$

The last line follows from the definition of subgradient

$$
f(x^\star) \geq f(x^{(k)}) + g^{(k)T}(x^\star - x^{(k)})
$$

Convergence Proof (cont.)

- Apply recursively to get

$$0 \leq ||x^{(k+1)} - x^\star||_2^2 \leq ||x^{(1)} - x^\star||_2^2 - 2\sum_{i=1}^{k} \alpha_i \left(f(x^{(i)}) - f^\star\right) + \sum_{i=1}^{k} \alpha_i^2 ||g^{(i)}||_2^2$$

- Since $||x^{(k+1)} - x^\star||_2^2 \geq 0$ and $||x^{(1)} - x^\star||_2^2 \leq R^2$, we have

$$2\sum_{i=1}^{k} \alpha_i \left(f(x^{(i)}) - f^\star\right) \leq \underbrace{R^2 + \sum_{i=1}^{k} \alpha_i^2 ||g^{(i)}||_2^2}$$

- Now, we use

$$\begin{aligned} \frac{\quad}{2} \geq \sum_{i=1}^{k} \alpha_i (\underbrace{f(x^{(i)})}_{+\infty} - f^\star) \quad &\geq \quad \left(\sum_{i=1}^{k} \alpha_i\right) \min_{i=1,\ldots,k} \left(f(x^{(i)}) - f^\star\right) \\ &= \quad \left(f_{\text{best}}^{(k)} - f^\star\right) \sum_{i=1}^{k} \alpha_i \end{aligned}$$

Convergence Proof (cont.)

- Thus,

$$
\begin{aligned}
f_{\text{best}}^{(k)} - f^\star &= \min_{i=1,\dots,k} f(x^{(i)}) - f^\star \\
&\leq \frac{R^2 + \sum_{i=1}^k \alpha_i^2 ||g^{(i)}||_2^2}{2\sum_{i=1}^k \alpha_i} && (2) \\
&\leq \frac{R^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i} && (3)
\end{aligned}
$$

Convergence Results

Define $\bar{f} = \lim_{k\to\infty} f_{\text{best}}^{(k)}$

- Constant step size: $\bar{f} - f^\star \leq G^2\alpha/2$, i.e., suboptimal
- Constant step length: $\bar{f} - f^\star \leq G\gamma/2$, i.e., suboptimal
- Square summable but not summable: $\bar{f} = f^\star$, i.e., converges
- Diminishing step size rule: $\bar{f} = f^\star$, i.e., converges

Projected Subgradient Method

- solves constrained optimization problem

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

where $f : \mathbf{R}^n \to \mathbf{R}$, $C \subseteq \mathbf{R}^n$ are convex

- projected subgradient method is given by

$$x^{(k+1)} = P(x^{(k)} - \alpha_k g^{(k)}),$$

where P is (Euclidean) projection on C, and $g^{(k)} \in \partial f(x^{(k)})$

- for constant step size, converges to neighborhood of optimal
- for diminishing nonsummable step sizes, converges

Projected Subgradient Method: Convergence Proof

- Consider a standard subgradient update before the projection back onto C

$$z^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- As in the subgradient method, we have

$$\begin{aligned}
||z^{(k+1)} - x^\star||_2^2 &= ||x^{(k)} - \alpha_k g^{(k)} - x^\star||_2^2 \\
&= ||x^{(k)} - x^\star||_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^\star) + \alpha_k^2 ||g^{(k)}||_2^2 \\
&\leq ||x^{(k)} - x^\star||_2^2 - 2\alpha_k (f(x^{(k)}) - f^\star) + \alpha_k^2 ||g^{(k)}||_2^2
\end{aligned}$$

- Now, observe that

$$||x^{(k+1)} - x^\star||_2 = ||P(z^{(k+1)}) - x^\star||_2 \leq ||z^{(k+1)} - x^\star||_2$$

- Thus,

$$||x^{(k+1)} - x^\star||_2^2 \leq ||x^{(k)} - x^\star||_2^2 - 2\alpha_k (f(x^{(k)}) - f^\star) + \alpha_k^2 ||g^{(k)}||_2^2$$

- The rest of proof proceeds exactly as in the ordinary subgradient method.

Projected Subgradient for Dual Problem

- Consider the following convex primal problem

ma - - - -

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1, \dots, m \end{array}$$

- For each $\lambda \succeq 0$, the Lagrangian

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

has a unique minimizer over x, which we denote as $x^\star(\lambda)$.

- The dual function is

sup

$$g(\lambda) = \inf_x L(x, \lambda) = f_0(x^\star(\lambda)) + \sum_{i=1}^{m} \lambda_i f_i(x^\star(\lambda))$$

- The dual problem is

min - - -

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

Projected Subgradient for Dual Problem (cont.)

- Approach: Solve the primal problem by finding an optimal point $\lambda^\star$ of the dual, and then taking $x^\star = x^\star(\lambda^\star)$.

- Via the projected subgradient method, we have

$$\lambda^{(k+1)} = \left(\lambda^{(k)} - \alpha_k h\right)^+, \qquad h \in \partial(-g)(\lambda^{(k)})$$

- Since $-g$ is a supremum of a family of affine functions of λ, index by x, we can find a subgradient by finding one of these functions that achieves the supremum. But there is just one, and its gradient is

$$h = -(f_1(x^\star(\lambda)), \ldots, f_m(x^\star(\lambda))) \in \partial(-g)(\lambda)$$

- The projected subgradient method for the dual has the form

$$x^{(k)} = \arg\min_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^{(k)} f_i(x) \right)$$

$$\lambda_i^{(k+1)} = \left(\lambda_i^{(k)} + \alpha_k f_i(x^{(k)})\right)^+, \;\; i = 1, \ldots, m.$$

Projected Subgradient for Dual Problem (cont.)

- primal iterates $x^{(k)}$ are not feasible, but become feasible only in limit.
- dual function $g(\lambda^{(k)})$ converges to $f^\star = f_0(x^\star)$
- λ_i is the price for a resource with usage measured by $f_i(x)$
- Price update $\lambda_i^{(k+1)} = \left(\lambda_i^{(k)} + \alpha_k f_i(x^{(k)})\right)^+$
- Increase price λ_i if resource i is over-utilized (i.e., $f_i(x) > 0$)
- Decrease price λ_i if resource i is under-utilized (i.e., $f_i(x) < 0$)
- but never let prices be negative

SUMMARY

- Subgradient
- Subdifferential
- Subgradient Method
- Projected Subgradient for Dual Problem