Subgradients

S. Boyd and L. Vandenberghe
Notes for EE364b, Stanford University, Winter 2006-07

April 6, 2011

1 Definition

We say a vector $g \in \mathbf{R}^n$ is a *subgradient* of $f : \mathbf{R}^n \to \mathbf{R}$ at $x \in \mathbf{dom}\, f$ if for all $z \in \mathbf{dom}\, f$,

$$f(z) \geq f(x) + g^T(z - x). \tag{1}$$

If f is convex and differentiable, then its gradient at x is a subgradient. But a subgradient can exist even when f is not differentiable at x, as illustrated in figure 1. The same example shows that there can be more than one subgradient of a function f at a point x.

There are several ways to interpret a subgradient. A vector g is a subgradient of f at x if the affine function (of z) $f(x) + g^T(z - x)$ is a global underestimator of f. Geometrically, g is a subgradient of f at x if $(g, -1)$ supports $\mathbf{epi}\, f$ at $(x, f(x))$, as illustrated in figure 2.

A function f is called *subdifferentiable* at x if there exists at least one subgradient at x. The set of subgradients of f at the point 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 $x \in \mathbf{dom}\, f$.

> **Example.** *Absolute value.* Consider $f(z) = |z|$. For $x < 0$ the subgradient is unique: $\partial f(x) = \{-1\}$. Similarly, for $x > 0$ we have $\partial f(x) = \{1\}$. At $x = 0$ the subdifferential is defined by the inequality $|z| \geq gz$ for all z, which is satisfied if and only if $g \in [-1, 1]$. Therefore we have $\partial f(0) = [-1, 1]$. This is illustrated in figure 3.

2 Basic properties

The subdifferential $\partial f(x)$ is always a closed convex set, even if f is not convex. This follows from the fact that it is the intersection of an infinite set of halfspaces:

$$\partial f(x) = \bigcap_{z \in \mathbf{dom}\, f} \{g \mid f(z) \geq f(x) + g^T(z - x)\}.$$

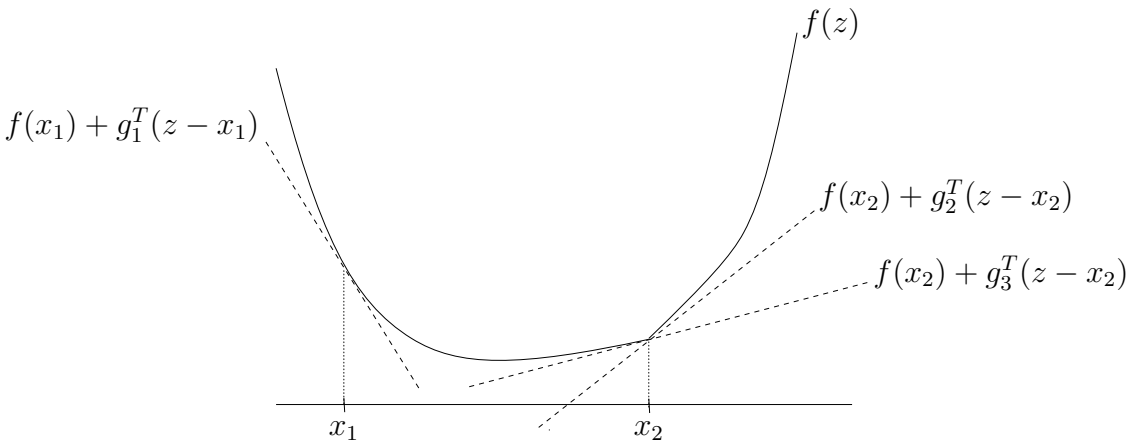

Figure 1: At x_1, the convex function f is differentiable, and g_1 (which is the derivative of f at x_1) is the unique subgradient at x_1. At the point x_2, f is not differentiable. At this point, f has many subgradients: two subgradients, g_2 and g_3, are shown.

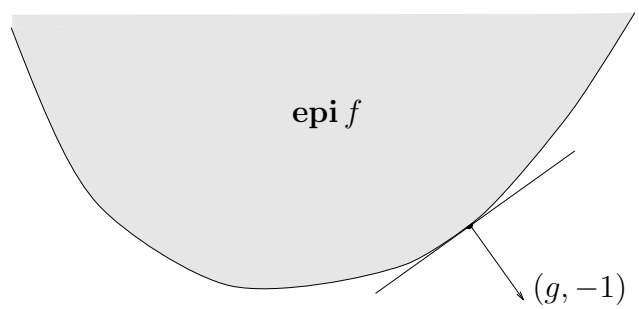

Figure 2: A vector $g \in \mathbf{R}^n$ is a subgradient of f at x if and only if $(g, -1)$ defines a supporting hyperplane to **epi** f at $(x, f(x))$.

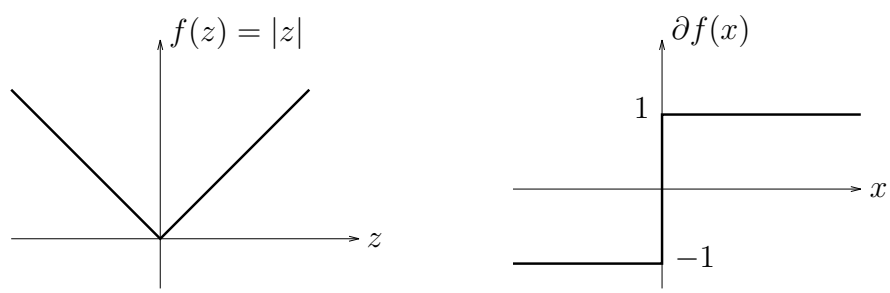

Figure 3: The absolute value function (left), and its subdifferential $\partial f(x)$ as a function of x (right).

2.1 Existence of subgradients

If f is convex and $x \in \mathbf{int}\,\mathbf{dom}\, f$, then $\partial f(x)$ is nonempty and bounded. To establish that $\partial f(x) \neq \emptyset$, we apply the supporting hyperplane theorem to the convex set $\mathbf{epi}\, f$ at the boundary point $(x, f(x))$, to conclude the existence of $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$, not both zero, such that

$$\begin{bmatrix} a \\ b \end{bmatrix}^T \left(\begin{bmatrix} z \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) = a^T(z - x) + b(t - f(x)) \leq 0$$

for all $(z, t) \in \mathbf{epi}\, f$. This implies $b \leq 0$, and that

$$a^T(z - x) + b(f(z) - f(x)) \leq 0$$

for all z. If $b \neq 0$, we can divide by b to obtain

$$f(z) \geq f(x) - (a/b)^T(z - x),$$

which shows that $-a/b \in \partial f(x)$. Now we show that $b \neq 0$, *i.e.*, that the supporting hyperplane cannot be vertical. If $b = 0$ we conclude that $a^T(z - x) \leq 0$ for all $z \in \mathbf{dom}\, f$. This is impossible since $x \in \mathbf{int}\,\mathbf{dom}\, f$.

This discussion shows that a convex function has a subgradient at x if there is at least one nonvertical supporting hyperplane to $\mathbf{epi}\, f$ at $(x, f(x))$. This is the case, for example, if f is continuous. There are pathological convex functions which do not have subgradients at some points, but we will assume in the sequel that all convex functions are subdifferentiable (at every point in $\mathbf{dom}\, f$).

2.2 Subgradients of differentiable functions

If f is convex and differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$, *i.e.*, its gradient is its only subgradient. Conversely, if f is convex and $\partial f(x) = \{g\}$, then f is differentiable at x and $g = \nabla f(x)$.

2.3 The minimum of a nondifferentiable function

A point $x^\star$ is a minimizer of a convex function f if and only if f is subdifferentiable at $x^\star$ and

$$0 \in \partial f(x^\star),$$

i.e., $g = 0$ is a subgradient of f at $x^\star$. This follows directly from the fact that $f(x) \geq f(x^\star)$ for all $x \in \mathbf{dom}\, f$.

This condition $0 \in \partial f(x^\star)$ reduces to $\nabla f(x^\star) = 0$ if f is differentiable at $x^\star$.

3 Calculus of subgradients

In this section we describe rules for constructing subgradients of convex functions. We will distinguish two levels of detail. In the 'weak' calculus of subgradients the goal is to produce one subgradient, even if more subgradients exist. This is sufficient in practice, since subgradient, localization, and cutting-plane methods require only *a* subgradient at any point.

A second and much more difficult task is to describe the complete set of subgradients $\partial f(x)$ as a function of x. We will call this the 'strong' calculus of subgradients. It is useful in theoretical investigations, for example, when describing the precise optimality conditions.

3.1 Nonnegative scaling

For $\alpha \geq 0$, $\partial(\alpha f)(x) = \alpha \partial f(x)$.

3.2 Sum and integral

Suppose $f = f_1 + \cdots + f_m$, where $f_1, \ldots, f_m$ are convex functions. Then we have

$$\partial f(x) = \partial f_1(x) + \cdots + \partial f_m(x).$$

This property extends to infinite sums, integrals, and expectations (provided they exist).

3.3 Affine transformations of domain

Suppose f is convex, and let $h(x) = f(Ax + b)$. Then $\partial h(x) = A^T \partial f(Ax + b)$.

3.4 Pointwise maximum

Suppose f is the pointwise maximum of convex functions $f_1, \ldots, f_m$, *i.e.*,

$$f(x) = \max_{i=1,\ldots,m} f_i(x),$$

where the functions f_i are subdifferentiable. We first show how to construct a subgradient of f at x.

Let k be any index for which $f_k(x) = f(x)$, and let $g \in \partial f_k(x)$. Then $g \in \partial f(x)$. In other words, to find a subgradient of the maximum of functions, we can choose one of the functions that achieves the maximum at the point, and choose any subgradient of that function at the point. This follows from

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

More generally, we have

$$\partial f(x) = \mathbf{Co} \cup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

i.e., the subdifferential of the maximum of functions is the convex hull of the union of subdifferentials of the 'active' functions at x.

Example. *Maximum of differentiable functions.* Suppose $f(x) = \max_{i=1,\ldots,m} f_i(x)$, where f_i are convex and differentiable. Then we have

$$\partial f(x) = \mathbf{Co}\{\nabla f_i(x) \mid f_i(x) = f(x)\}.$$

At a point x where only one of the functions, say f_k, is active, f is differentiable and has gradient $\nabla f_k(x)$. At a point x where several of the functions are active, $\partial f(x)$ is a polyhedron.

Example. *ℓ_1-norm.* The ℓ_1-norm

$$f(x) = \|x\|_1 = |x_1| + \cdots + |x_n|$$

is a nondifferentiable convex function of x. To find its subgradients, we note that f can expressed as the maximum of 2^n linear functions:

$$\|x\|_1 = \max\{s^T x \mid s_i \in \{-1, 1\}\},$$

so we can apply the rules for the subgradient of the maximum. The first step is to identify an active function $s^T x$, *i.e.*, find an $s \in \{-1, +1\}^n$ such that $s^T x = \|x\|_1$. We can choose $s_i = +1$ if $x_i > 0$, and $s_i = -1$ if $x_i < 0$. If $x_i = 0$, more than one function is active, and both $s_i = +1$, and $s_i = -1$ work. The function $s^T x$ is differentiable and has a unique subgradient s. We can therefore take

$$g_i = \begin{cases} +1 & x_i > 0 \\ -1 & x_i < 0 \\ -1 \text{ or } +1 & x_i = 0. \end{cases}$$

The subdifferential is the convex hull of all subgradients that can be generated this way:

$$\partial f(x) = \{g \mid \|g\|_\infty \leq 1, g^T x = \|x\|_1\}.$$

3.5 Supremum

Next we consider the extension to the supremum over an infinite number of functions, *i.e.*, we consider

$$f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x),$$

where the functions f_α are subdifferentiable. We only discuss the weak property.

Suppose the supremum in the definition of $f(x)$ is attained. Let $\beta \in \mathcal{A}$ be an index for which $f_\beta(x) = f(x)$, and let $g \in \partial f_\beta(x)$. Then $g \in \partial f(x)$. If the supremum in the definition is not attained, the function may or may not be subdifferentiable at x, depending on the index set $\mathcal{A}$.

Assume however that $\mathcal{A}$ is compact (in some metric), and that the function $\alpha \mapsto f_\alpha(x)$ is upper semi-continuous for each x. Then

$$\partial f(x) = \mathbf{Co} \cup \{\partial f_\alpha(x) \mid f_\alpha(x) = f(x)\}.$$

Example. *Maximum eigenvalue of a symmetric matrix.* Let $f(x) = \lambda_{\max}(A(x))$, where $A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n$, and $A_i \in \mathbf{S}^m$. We can express f as the pointwise supremum of convex functions,

$$f(x) = \lambda_{\max}(A(x)) = \sup_{\|y\|_2=1} y^T A(x) y.$$

Here the index set $\mathcal{A}$ is $\mathcal{A} = \{y \in \mathbf{R}^n \mid \|y_2\|_1 \leq 1\}$.

Each of the functions $f_y(x) = y^T A(x) y$ is affine in x for fixed y, as can be easily seen from

$$y^T A(x) y = y^T A_0 y + x_1 y^T A_1 y + \cdots + x_n y^T A_n y,$$

so it is differentiable with gradient $\nabla f_y(x) = (y^T A_1 y, \ldots, y^T A_n y)$.

The active functions $y^T A(x) y$ are those associated with the eigenvectors corresponding to the maximum eigenvalue. Hence to find a subgradient, we compute an eigenvector y with eigenvalue $\lambda_{\max}$, normalized to have unit norm, and take

$$g = (y^T A_1 y, y^T A_2 y, \ldots, y^T A_n y).$$

The 'index set' in this example is $\{y \mid \|y\| = 1\}$ is a compact set. Therefore

$$\partial f(x) = \mathbf{Co}\left\{\nabla f_y \mid A(x) y = \lambda_{\max}(A(x)) y,\ \|y\| = 1\right\}.$$

3.6 Minimization over some variables

The next subgradient calculus rule concerns functions of the form

$$f(x) = \inf_y F(x, y)$$

where $F(x, y)$ is subdifferentiable and jointly convex in x and y. Again we only discuss the weak property.

Suppose the infimum over y in the definition of $f(\hat{x})$ is attained at $y = \hat{y}$, *i.e.*, $f(\hat{x}) = F(\hat{x}, \hat{y})$. Then there exists a g such that $(g, 0) \in \partial F(\hat{x}, \hat{y})$, and any such g is a subgradient of f at $\hat{x}$.

Strong property. Let x_2 be such that $f(x_1) = \inf_{x_2} F(x_1, x_2)$. Then $\partial f(x_1) = \{g_1 \mid (g_1, 0) \in \partial F(x_1, x_2)\}$, (and the resulting subdifferential is independent of the choice of x_2).

3.7 Optimal value function of a convex optimization problem

Suppose $f : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$ is defined as the optimal value of a convex optimization problem in standard form, with $z \in \mathbf{R}^n$ as optimization variable,

$$\begin{array}{ll} \text{minimize} & f_0(z) \\ \text{subject to} & f_i(z) \leq x_i, \quad i = 1, \ldots, m \\ & Az = y. \end{array} \tag{2}$$

In other words, $f(x,y) = \inf_z F(x,y,z)$ where

$$F(x,y,z) = \begin{cases} f_0(z) & f_i(z) \le x_i,\ i = 1,\ldots,m,\ Az = y \\ +\infty & \text{otherwise,} \end{cases}$$

which is jointly convex in x, y, z. Subgradients of f can be related to the dual problem of (2) as follows.

Suppose we are interested in subdifferentiating f at $(\hat{x},\hat{y})$. We can express the dual problem of (2) as

$$\begin{array}{ll} \text{maximize} & g(\lambda) - x^T\lambda - y^T\nu \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{3}$$

where

$$g(\lambda) = \inf_z \left(f_0(z) + \sum_{i=1}^m \lambda_i f_i(z) + \nu^T Az \right).$$

Suppose strong duality holds for problems (2) and (3) at $x = \hat{x}$ and $y = \hat{y}$, and that the dual optimum is attained at $\lambda^\star$, $\nu^\star$ (for example, because Slater's condition holds). From the global perturbation inequalities we know that

$$f(x,y) \ge f(\hat{x},\hat{y}) - \lambda^{\star T}(x - \hat{x}) - \nu^{\star T}(y - \hat{y})$$

In other words, the dual optimal solution provides a subgradient:

$$-(\lambda^\star, \nu^\star) \in \partial f(\hat{x},\hat{y}).$$

4 Quasigradients

If $f(x)$ is quasiconvex, then g is a *quasigradient* at x_0 if

$$g^T(x - x_0) \ge 0 \Rightarrow f(x) \ge f(x_0),$$

Geometrically, g defines a supporting hyperplane to the sublevel set $\{x \mid f(x) \le f(x_0)\}$. Note that the set of quasigradients at x_0 form a cone.

Example. *Linear fractional function.* $f(x) = \frac{a^T x + b}{c^T x + d}$. Let $c^T x_0 + d > 0$. Then $g = a - f(x_0)c$ is a quasigradient at x_0. If $c^T x + d > 0$, we have

$$a^T(x - x_0) \ge f(x_0)c^T(x - x_0) \Longrightarrow f(x) \ge f(x_0).$$

Example. *Degree of a polynomial.* Define $f : \mathbf{R}^n \to \mathbf{R}$ by

$$f(a) = \min\{i \mid a_{i+2} = \cdots = a_n = 0\},$$

i.e., the degree of the polynomial $a_1 + a_2 t + \cdots + a_n t^{n-1}$. Let $a \ne 0$, and $k = f(a)$, then $g = \mathrm{sign}(a_{k+1})e_{k+1}$ is a quasigradient at a

To see this, we note that

$$g^T(b - a) = \mathrm{sign}(a_{k+1})b_{k+1} - |a_{k+1}| \ge 0$$

implies $b_{k+1} \ne 0$.