

2. Subgradients

- definition
- subgradient calculus
- optimality conditions via subgradients
- directional derivative

2-1

Basic inequality

recall basic inequality for convex differentiable f :

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

- first-order approximation of f at x is global lower bound
- $\nabla f(x)$ defines non-vertical supporting hyperplane to $\mathbf{epi} f$ at $(x, f(x))$

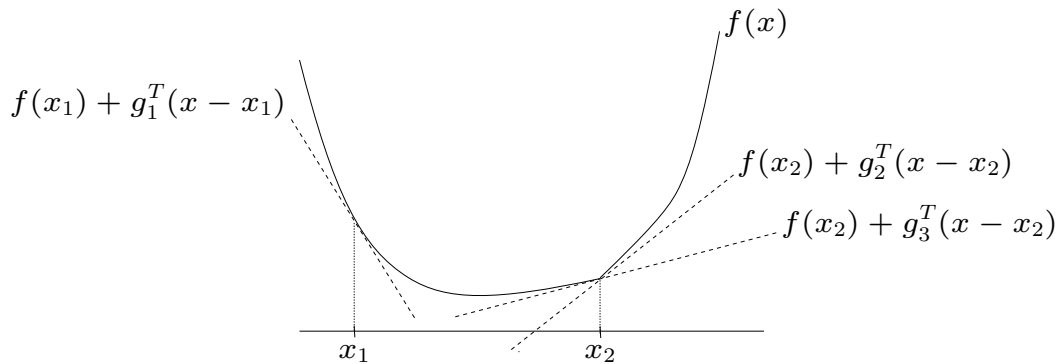
$$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall (y, t) \in \mathbf{epi} f$$

what if f is not differentiable?

Subgradient of a function

definition: g is a subgradient of a convex function f at $x \in \mathbf{dom} f$ if

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y \in \mathbf{dom} f$$



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

Subgradients

2-3

properties

- $f(x) + g^T(y - x)$ is a global lower bound on f
- g defines non-vertical supporting hyperplane to $\mathbf{epi} f$ at $(x, f(x))$

$$\begin{bmatrix} g \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall (y, t) \in \mathbf{epi} f$$

- if f is convex and differentiable, then $\nabla f(x)$ is a subgradient of f at x

applications

- algorithms for nondifferentiable convex optimization
- optimality conditions, duality for nondifferentiable problems

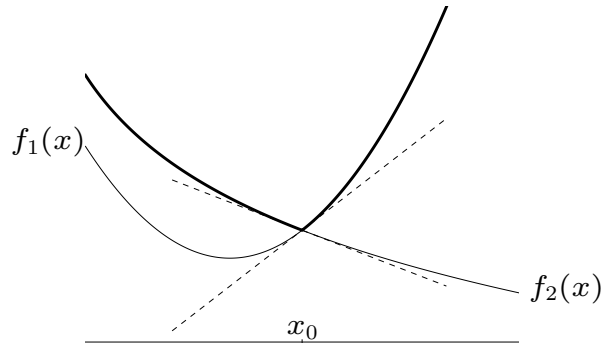
Subgradients

2-4

Example

$$f(x) = \max\{f_1(x), f_2(x)\}$$

f_1, f_2 convex and differentiable



- subgradients at x_0 form line segment $[\nabla f_1(x_0), \nabla f_2(x_0)]$
- if $f_1(\hat{x}) > f_2(\hat{x})$, subgradient of f at \hat{x} is $\nabla f_1(\hat{x})$
- if $f_1(\hat{x}) < f_2(\hat{x})$, subgradient of f at \hat{x} is $\nabla f_2(\hat{x})$

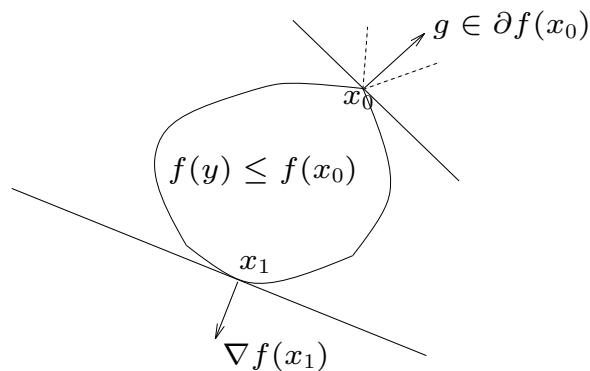
Subgradients

2-5

Subgradients and sublevel sets

if g is a subgradient of f at x , then

$$f(y) \leq f(x) \implies g^T(y - x) \leq 0$$



nonzero subgradients at x define supporting hyperplanes to sublevel set

$$\{y \mid f(y) \leq f(x)\}$$

Subgradients

2-6

Subdifferential

subdifferential of f at $x \in \mathbf{dom} f$ is the set of all subgradients of f at x

notation: $\partial f(x)$

properties

- $\partial f(x)$ is a closed convex set (possibly empty)

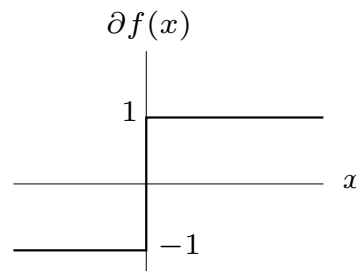
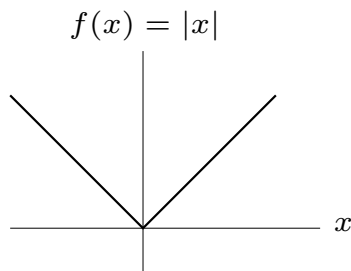
proof: $\partial f(x)$ is an intersection of halfspaces

$$\partial f(x) = \{g \mid f(x) + g^T(y - x) \leq f(y) \forall y \in \mathbf{dom} f\}$$

- if $x \in \mathbf{int} \mathbf{dom} f$ then $\partial f(x)$ is nonempty and bounded

Examples

absolute value $f(x) = |x|$



Euclidean norm $f(x) = \|x\|_2$

$$\partial f(x) = \frac{1}{\|x\|_2} x \quad \text{if } x \neq 0, \quad \partial f(x) = \{g \mid \|g\|_2 \leq 1\} \quad \text{if } x = 0$$

Examples of non-subdifferentiable functions

the following functions are not subdifferentiable at $x = 0$

- $f : \mathbf{R} \rightarrow \mathbf{R}$, $\text{dom } f = \mathbf{R}_+$

$$f(x) = 1 \quad \text{if } x = 0, \quad f(x) = 0 \quad \text{if } x > 0$$

- $f : \mathbf{R} \rightarrow \mathbf{R}$, $\text{dom } f = \mathbf{R}_+$

$$f(x) = -\sqrt{x}$$

the only supporting hyperplane to $\text{epi } f$ at $(0, f(0))$ is vertical

Subgradient calculus

weak subgradient calculus: rules for finding *one* subgradient

- sufficient for many algorithms for nondifferentiable convex optimization
- if you can evaluate $f(x)$, you can usually compute a subgradient

strong subgradient calculus: rules for finding $\partial f(x)$ (*all* subgradients)

- some algorithms, optimality conditions, etc., need whole subdifferential
- can be quite complicated

we will assume that $x \in \text{int dom } f$

Some basic rules

differentiable functions: $\partial f(x) = \{\nabla f(x)\}$ if f is differentiable at x

nonnegative combination

if $h(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ with $\alpha_1, \alpha_2 \geq 0$, then

$$\partial h(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x)$$

(r.h.s. is addition of sets)

affine transformation of variables: if $h(x) = f(Ax + b)$, then

$$\partial h(x) = A^T \partial f(Ax + b)$$

Pointwise maximum

$$f(x) = \max\{f_1(x), \dots, f_m(x)\}$$

define $I(x) = \{i \mid f_i(x) = f(x)\}$, the 'active' functions at x

weak result: to compute a subgradient at x ,

choose any $k \in I(x)$, and any subgradient of $f_k(x)$

strong result

$$\partial f(x) = \mathbf{conv} \bigcup_{i \in I(x)} \partial f_i(x)$$

- convex hull of the union of subdifferentials of 'active' functions at x
- if f_i 's are differentiable, $\partial f(x) = \mathbf{conv}\{\nabla f_i(x) \mid i \in I(x)\}$

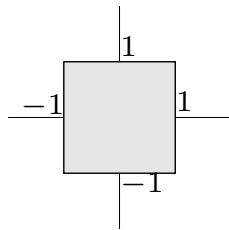
example

$$f(x) = \max_{i=1,\dots,m} a_i^T x + b_i$$

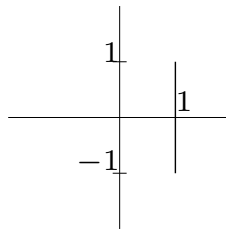
the subdifferential is a polyhedron $\partial f(x) = \mathbf{conv}\{a_i \mid i \in I(x)\}$

example

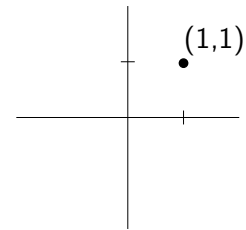
$$f(x) = \|x\|_1 = \max_{s \in \{-1,1\}^n} s^T x$$



$\partial f(0,0)$



$\partial f(1,0)$



$\partial f(1,1)$

Pointwise supremum

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

with $f_\alpha(x)$ convex for every α

weak result: to find a subgradient at x ,

- find *any* β for which $f(x) = f_\beta(x)$ (assuming supremum is achieved)
- choose *any* $g \in \partial f_\beta(x)$

(partial) strong result: define $\mathcal{I}(x) = \{\alpha \in \mathcal{A} \mid f_\alpha(x) = f(x)\}$

$$\mathbf{conv} \bigcup_{\alpha \in \mathcal{I}(x)} \partial f_\alpha(x) \subseteq \partial f(x)$$

equality requires some technical conditions

example

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

where $A(x) = A_0 + x_1 A_1 + \dots + x_n A_n$, $A_i \in \mathbf{S}^k$

to find *one* subgradient at x :

- choose *any* unit eigenvector y associated with $\lambda_{\max}(A(x))$
- the gradient of $y^T A(x) y$ is a subgradient of f :

$$(y^T A_1 y, \dots, y^T A_n y) \in \partial f(x)$$

Expectation

$$f(x) = \mathbf{E} h(x, u)$$

with h convex in x for each u , u a random variable

weak result: to find a subgradient at x

- for each u , choose *any* $g_u \in \partial_x h(x, u)$ (so $u \mapsto g_u$ is a function)
- then, $g = \mathbf{E} g_u \in \partial f(x)$

proof: by convexity of h ,

$$h(y, u) \geq h(x, u) + g_u^T (y - x) \quad \forall y$$

therefore

$$f(y) = \mathbf{E} h(y, u) \geq \mathbf{E} h(x, u) + \mathbf{E} g_u^T (y - x) = f(x) + g^T (y - x)$$

Minimization

define $h(y)$ as the optimal value of

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

(f_i convex; variable x)

if strong duality holds and λ^* is an optimal dual variable, then

$$h(z) \geq h(y) - \sum_{i=1}^m \lambda_i^*(z_i - y_i)$$

i.e., $-\lambda^*$ is a subgradient of h at y

Composition

$$f(x) = h(f_1(x), \dots, f_k(x))$$

with h convex nondecreasing, f_i convex

weak result: to find a subgradient at x ,

- find $z \in \partial h(f_1(x), \dots, f_k(x))$, and $g_i \in \partial f_i(x)$
- then, $g = z_1 g_1 + \dots + z_k g_k \in \partial f(x)$

reduces to standard formula for differentiable h , f_i

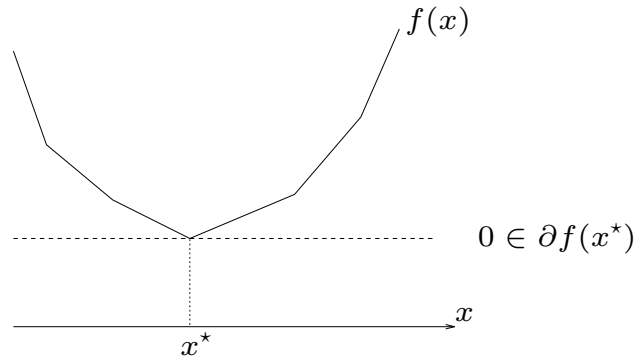
proof:

$$\begin{aligned} f(y) &= h(f_1(y), \dots, f_k(y)) \\ &\geq h(f_1(x) + g_1^T(y - x), \dots, f_k(x) + g_k^T(y - x)) \\ &\geq h(f_1(x), \dots, f_k(x)) + z^T(g_1^T(y - x), \dots, g_k^T(y - x)) \\ &= f(x) + g^T(y - x) \end{aligned}$$

Optimality conditions — unconstrained

x^* minimizes $f(x)$ if and only

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



proof: by definition (!)

$$f(y) \geq f(x^*) + 0^T(y - x^*) \text{ for all } y \iff 0 \in \partial f(x^*)$$

Example: piecewise linear minimization

$$f(x) = \max_{i=1, \dots, m} (a_i^T x + b_i)$$

optimality condition

$$0 \in \text{conv}\{a_i \mid i \in I(x^*)\} \quad (I(x) = \{i \mid a_i^T x + b_i = f(x)\})$$

in other words, there is a λ with

$$\lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1, \quad \sum_{i=1}^m \lambda_i a_i = 0, \quad \lambda_i = 0 \text{ for } i \notin I(x^*)$$

these are the KKT conditions for the equivalent LP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & Ax + b \preceq t\mathbf{1} \end{array} \qquad \begin{array}{ll} \text{maximize} & b^T \lambda \\ \text{subject to} & A^T \lambda = 0 \\ & \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1 \end{array}$$

Optimality conditions — constrained

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

from Lagrange duality

if strong duality holds, then x^* , λ^* are primal, dual optimal if and only if

1. x^* is primal feasible
2. $\lambda^* \succeq 0$
3. $\lambda_i^* f_i(x^*) = 0$ for $i = 1, \dots, m$
4. x^* is a minimizer of

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

Karush-Kuhn-Tucker conditions (if $\text{dom } f_i = \mathbf{R}^n$)

conditions 1,2,3 and

$$0 \in \partial L_x(x^*, \lambda^*) = \partial f_0(x^*) + \sum_{i=1}^m \lambda_i^* \partial f_i(x^*)$$

this generalizes the condition

$$0 = \nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*)$$

for differentiable f_i

Directional derivative

the directional derivative of f at x in the direction y is defined as

$$f'(x; y) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha}$$

properties (for convex f)

- if $x \in \text{int dom } f$, then $f'(x; y)$ exists for all y
- homogeneous in y : $f'(x; \lambda y) = \lambda f'(x; y)$ for $\lambda \geq 0$

infimum characterization

$$\begin{aligned} f'(x; y) &= \inf_{t > 0} \left(t f\left(x + \frac{1}{t}y\right) - t f(x) \right) \\ &= \inf_{\alpha > 0} \frac{f(x + \alpha y) - f(x)}{\alpha} \end{aligned}$$

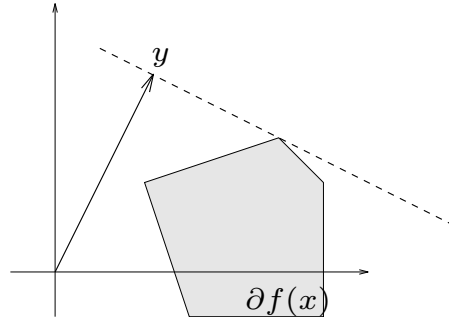
- the infimum over t of the perspective of $h(y) = f(x + y) - f(x)$
- follows from fact that $th(y/t)$ is nonincreasing in t if $h(0) \leq 0$
- important consequence: $f'(x; y)$ is convex in y
- also implies that

$$f(x + \alpha y) \geq f(x) + \alpha f'(x; y) \quad \text{for } \alpha \geq 0$$

Directional derivative and subdifferential

for convex f and $x \in \text{int dom } f$

$$f'(x; y) = \sup_{g \in \partial f(x)} g^T y$$



- generalizes $f'(x; y) = \nabla f(x)^T y$ for differentiable functions
- the directional derivative is the support function of the subdifferential

$$f'(x; y) = S_{\partial f(x)}(y)$$

proof

1. if $g \in \partial f(x)$, then $f(x + \alpha y) \geq f(x) + \alpha g^T y$ for all α, y ; therefore

$$f'(x; y) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha} \geq g^T y$$

this shows that

$$f'(x; y) \geq \sup_{g \in \partial f(x)} g^T y$$

2. if $\hat{g} \in \partial_y f'(x; y)$, then for all $v, \lambda \geq 0$

$$\lambda f'(x; v) = f'(x; \lambda v) \geq f'(x; y) + \hat{g}^T (\lambda v - y)$$

taking $\lambda \rightarrow \infty$ we get $f'(x; v) \geq \hat{g}^T v$, and therefore

$$f(x + v) \geq f(x) + f'(x; v) \geq f(x) + \hat{g}^T v$$

this means that $\hat{g} \in \partial f(x)$, and from 1, $f'(x; y) \geq \hat{g}^T y$

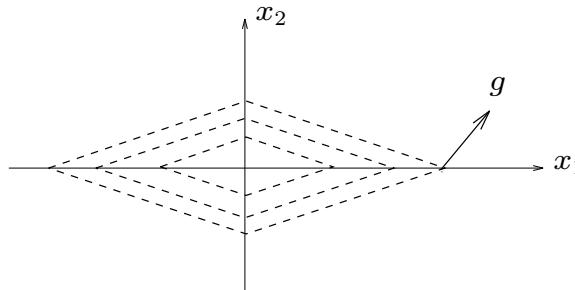
taking $\lambda = 0$ we see that $f'(x; y) = \hat{g}^T y$

Descent directions

y is a **descent direction** for f at x if $f'(x; y) < 0$

- if f is differentiable, then $-\nabla f(x)$ is a descent direction (if $\nabla f(x) \neq 0$)
- if f is nondifferentiable, then $-g$, with $g \in \partial f(x)$, is **not** always a descent direction

example: $f(x_1, x_2) = |x_1| + 2|x_2|$



$g = (1, 2) \in \partial f(1, 0)$, but $y = (-1, -2)$ is not a descent direction at $(1, 0)$

Steepest descent direction

normalized steepest descent direction at $x \in \text{int dom } f$

$$\Delta x_{\text{nsd}} = \underset{\|y\|_2 \leq 1}{\operatorname{argmin}} f'(x; y)$$

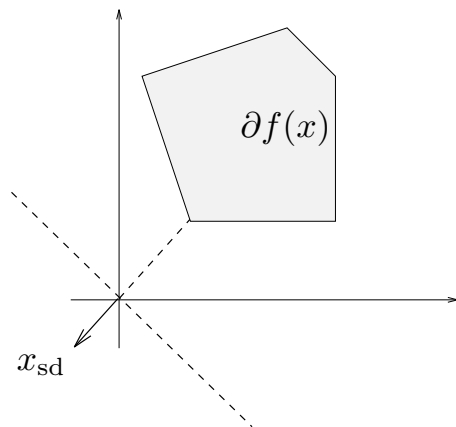
- can use same definition for steepest descent direction in other norms
- from duality (BV §8.1.3)

$$f'(x; \Delta x_{\text{nsd}}) = - \inf_{g \in \partial f(x)} \|g\|_2$$

- $\Delta x_{\text{sd}} = 0$ if and only if x is optimal; otherwise $f'(x; \Delta x_{\text{nsd}}) < 0$

unnormalized steepest descent direction

$$\Delta x_{\text{sd}} = - \operatorname{argmin}_{g \in \partial f(x)} \|g\|_2^2, \quad f'(x, \Delta x_{\text{sd}}) = - \inf_{g \in \partial f(x)} \|g\|_2^2$$



Subgradients and distance to sublevel sets

if f is convex, $f(z) < f(x)$, $g \in \partial f(x)$, then for small $t > 0$,

$$\|x - tg - z\|_2 < \|x - z\|_2$$

- $-g$ is descent direction for $\|x - z\|_2$, for **any** z with $f(z) < f(x)$
- negative subgradient is descent direction for distance to optimal point

proof:

$$\begin{aligned} \|x - tg - z\|_2^2 &= \|x - z\|_2^2 - 2tg^T(x - z) + t^2\|g\|_2^2 \\ &\leq \|x - z\|_2^2 - 2t(f(x) - f(z)) + t^2\|g\|_2^2 \\ &< \|x - z\|_2^2 \end{aligned}$$