4. Subgradients

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

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 $(x, f(x))$

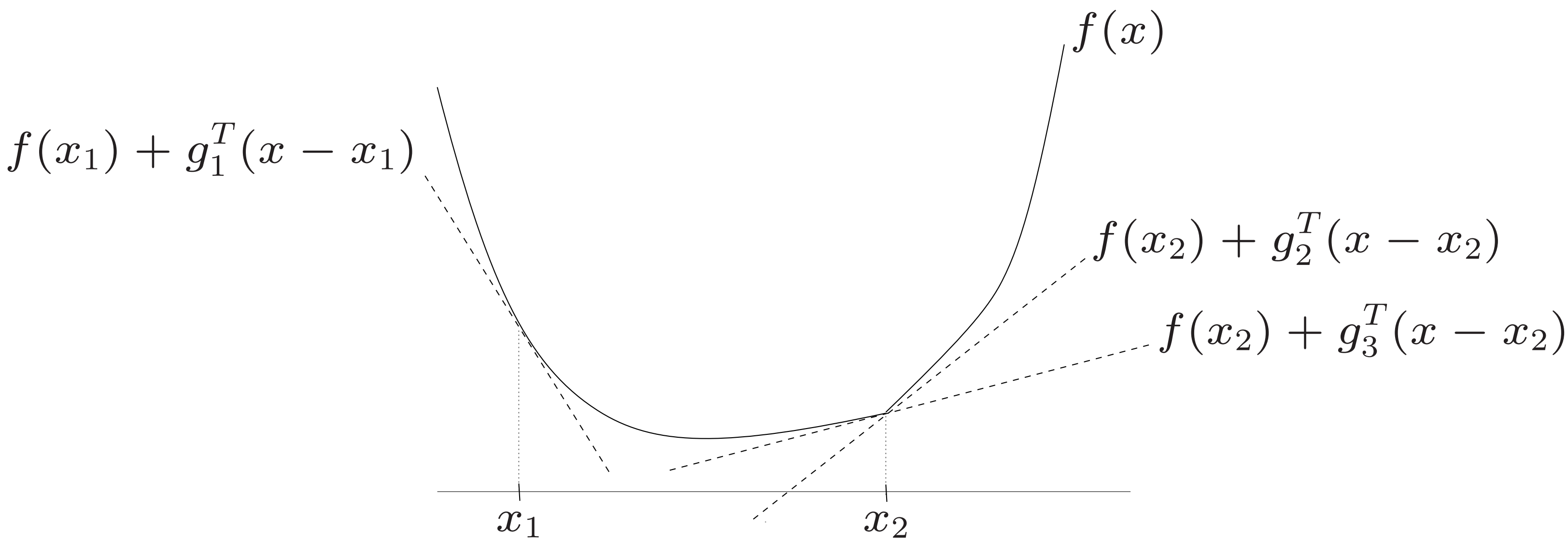
$$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall (y,t) \in \mathbf{epi}\, f$$

($\mathbf{epi}\, f$ denotes the epigraph of 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

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \mathbf{dom}\, f$$



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

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

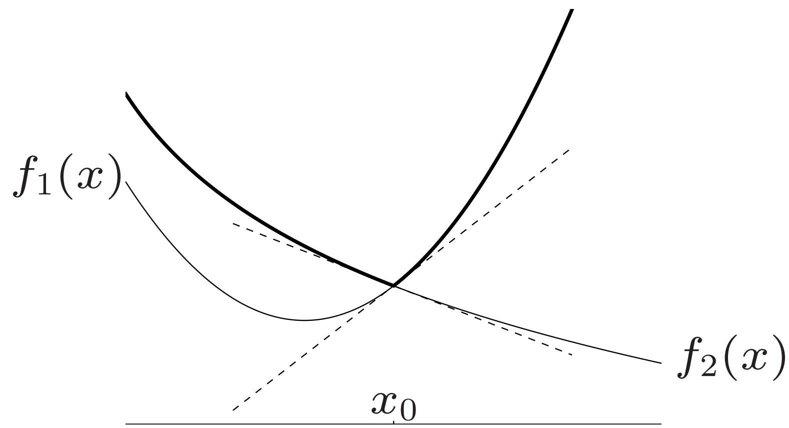
applications

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

Example

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

f_1, f_2 convex and differentiable; $x \in \mathbf{R}$



- 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 $\nabla f_2(\hat{x})$

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)

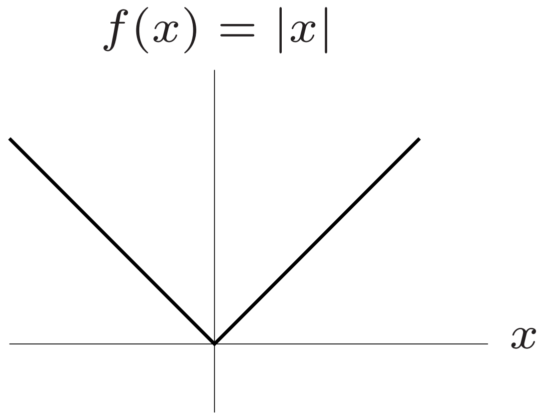
  proof: $\partial f(x)$ is an intersection of halfspaces

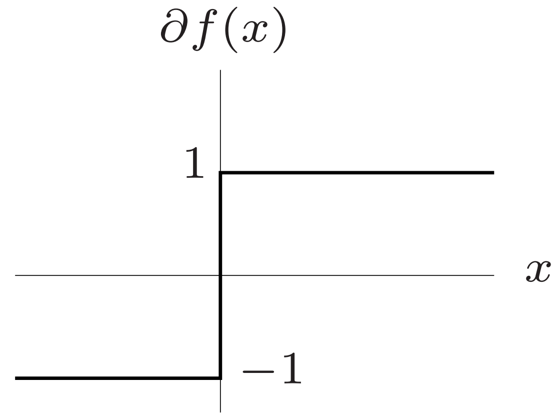
$$\partial f(x) = \left\{ g \mid f(x) + g^T(y - x) \leq f(y) \ \forall y \in \mathbf{dom}\, f \right\}$$

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

Examples

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





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

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

Monotonicity

subdifferential of a convex function is a **monotone operator:**

$$(u-v)^T(x-y) \geq 0 \quad \forall u \in \partial f(x),\ v \in \partial f(y)$$

proof: by definition

$$f(y) \geq f(x) + u^T(y-x), \qquad f(x) \geq f(y) + v^T(x-y)$$

combining the two inequalities shows monotonicity

Examples of non-subdifferentiable functions

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

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

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

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

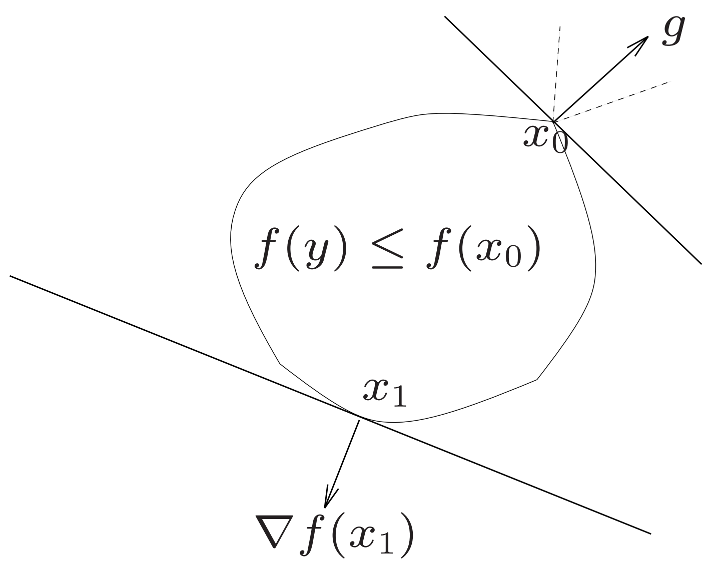
$$f(x) = -\sqrt{x}$$

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

Subgradients and sublevel sets

if g is a subgradient of f at x, then

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



nonzero subgradients at x define supporting hyperplanes to sublevel set

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

Outline

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

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 $x \in \mathbf{int}\,\mathbf{dom}\, f$

Some basic rules

(suppose all f_i are convex unless otherwise stated)

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

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

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

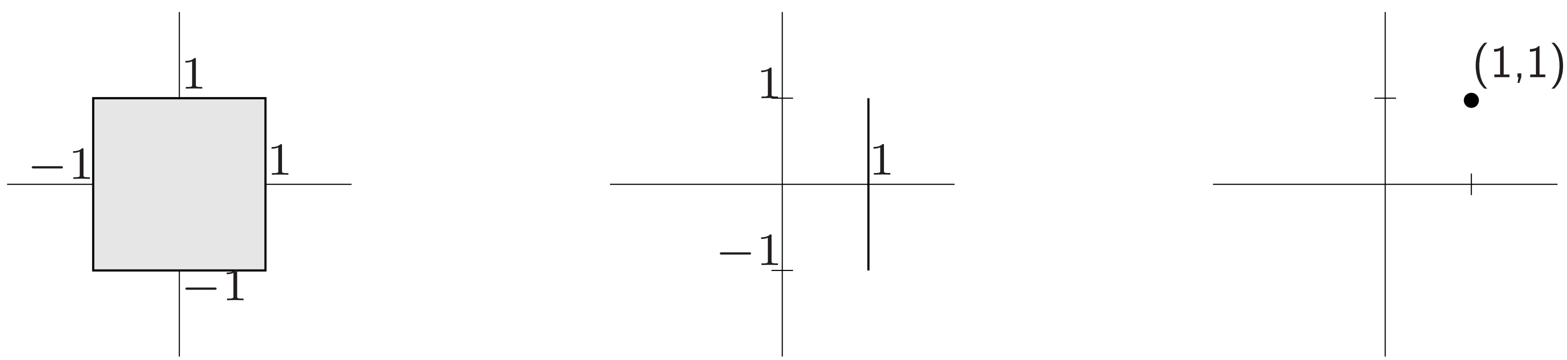
example

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

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: maximum eignenvalue

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

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

how to find *one* subgradient at $\hat{x}$?

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

$$(y^T A_1 y, \ldots, y^T A_n y) \in \partial f(\hat{x})$$

similarly can find a subgradient of $\|A(x)\| = \sigma_{\max}(A(x))$

Expectation

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

with h convex in x for each u, expectation is over random variable u

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 and definition of g_u,

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

therefore

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

(will use this in stochastic gradient methods)

Optimal value function

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

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

(f_i convex; variable x)

if strong duality holds and $\hat{\lambda}$ is an optimal dual variable, then

$$h(u) \geq h(\hat{u}) - \sum_{i=1}^{m} \hat{\lambda}_i (u_i - \hat{u}_i)$$

$i.e.$, $-\hat{\lambda}$ is a subgradient of h at y

Composition

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

with h convex nondecreasing, f_i convex

weak result: to find a subgradient at x,

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

reduces to standard formula for differentiable h, f_i

proof:

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

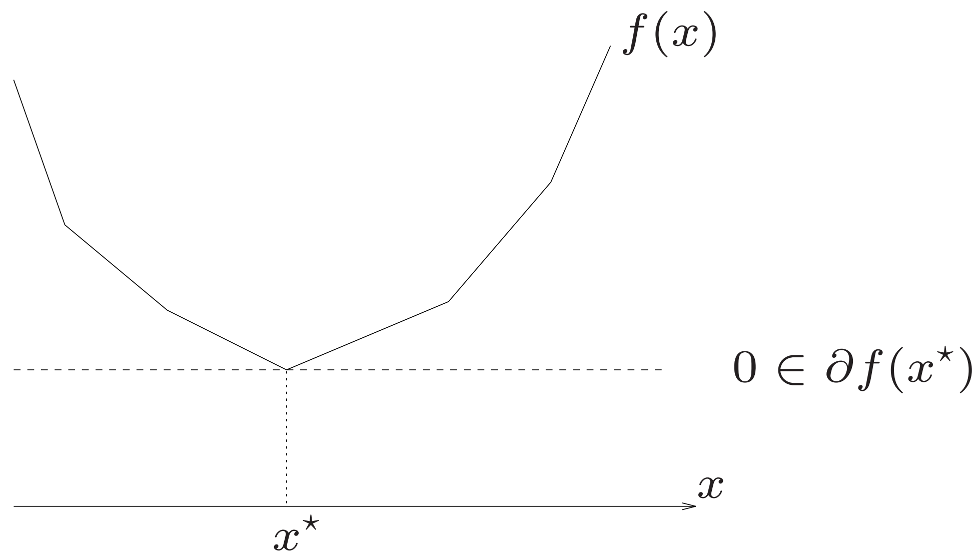
Outline

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

Optimality conditions — unconstrained

$x^\star$ minimizes $f(x)$ if and only

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



proof: by definition

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

Example: piecewise linear minimization

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

optimality condition

$$0 \in \mathbf{conv}\{a_i \mid i \in I(x^\star)\} \qquad (I(x) = \{i \mid a_i^T x + b_i = f(x)\})$$

in other words, there is a λ with

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

these are the KKT conditions for the equivalent LP

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

Optimality conditions — constrained

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

from Lagrange duality

if strong duality holds, then $x^\star$, $\lambda^\star$ are primal, dual optimal if and only if

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

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

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

conditions 1, 2, 3 and

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

this generalizes the condition

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

for differentiable f_i

Outline

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

Directional derivative

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

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

(if the limit exists)

properties (for convex f)

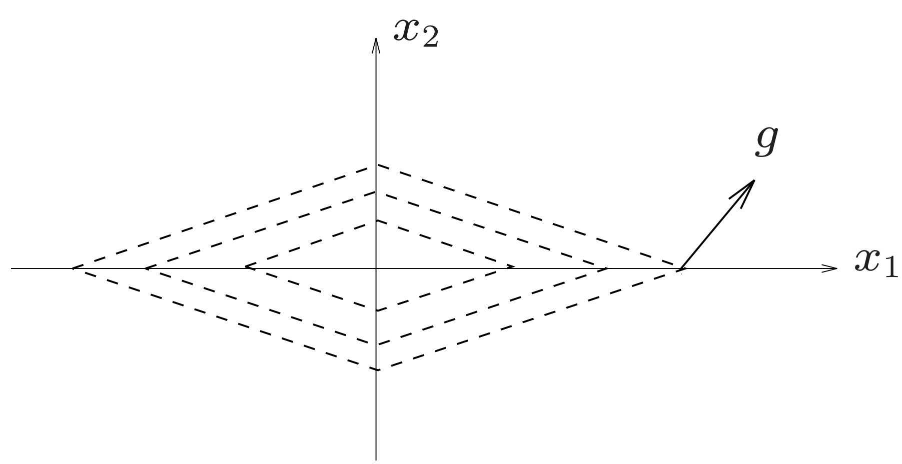
- if $x \in \mathbf{int\,dom}\, f$, then $f'(x; y)$ exists for all y
- homogeneous in y: $f'(x; \lambda y) = \lambda f'(x; y)$ for $\lambda \geq 0$

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

Descent directions and subgradients

- 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 $(1, 0)$

Directional derivative for convex f

equivalent definition: can replace lim with inf

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

proof:

- the function $h(y) = f(x+y) - f(x)$ is convex in y with $h(0) = 0$
- its perspective $th(y/t)$ is nonincreasing in t (an exercise from EE578), hence

$$f'(x;y) = \lim_{t \searrow 0} th(y/t) = \inf_{t>0} th(y/t)$$

consequences of expressions

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

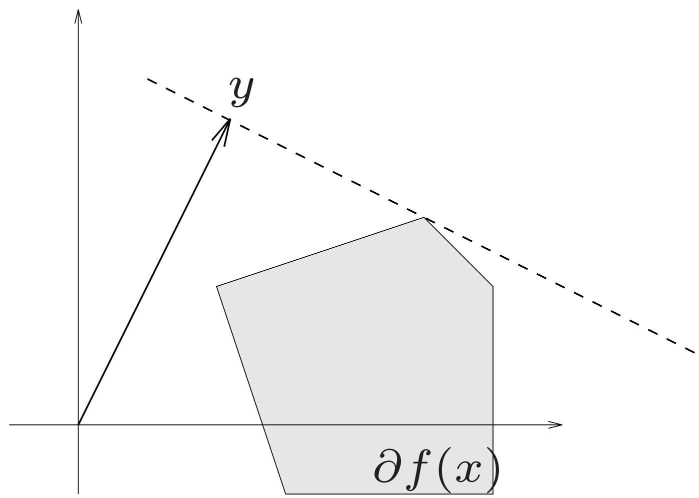
- $f'(x;y)$ is convex in y (partial minimization of convex fct in y, t)
- $f'(x;y)$ defines a lower bound on f in the direction y:

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

Directional derivative and subdifferential

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

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



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

proof

1. suppose $g \in \partial f(x)$, *i.e.*, $f(x + \alpha y) \geq f(x) + \alpha g^T y$ for all α, y; then

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

   this shows that

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

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

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

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

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

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

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

Subgradients and distance to sublevel sets

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

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

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

proof:

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

References

- L. Vandenberghe, *Lecture notes for EE236C - Optimization Methods for Large-Scale Systems* (Spring 2011 and Spring 2014), UCLA.
- J.-B. Hiriart-Urruty, C. Lemaréchal, *Convex Analysis and Minimization Algorithms* (1993), chapter VI.
- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004), section 3.1.