# Subgradient

Acknowledgement: this slides is based on Prof. Lieven Vandenberghes lecture notes

- definition
- subgradient calculus
- duality and optimality conditions
- directional derivative

# Basic inequality

recall basic inequality for convex differentiable $f$:

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

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

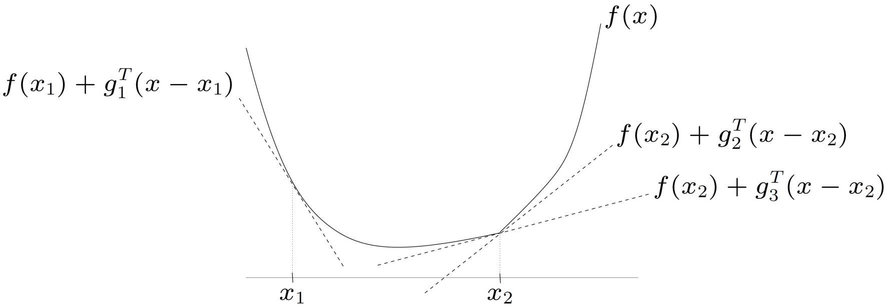
$$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix} \left( \begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall\ (y,t) \in \mathbf{epi}\, f$$

what if $f$ is not differentiable?

# Subgradient

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

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




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

**properties**

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

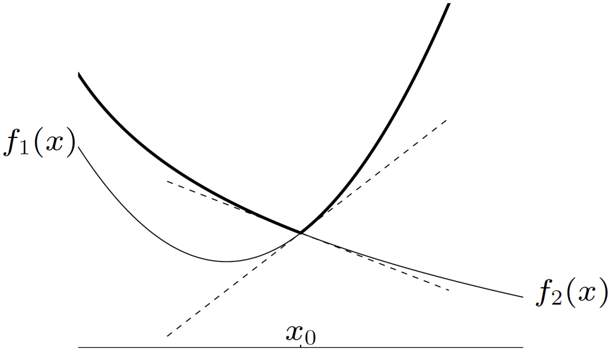
$$\begin{bmatrix} g \\ -1 \end{bmatrix} \left( \begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall\, (y,t) \in \mathbf{epi}\, f$$

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

- algorithms for nondifferentiable convex optimization
- unconstrained optimality: $x$ minimizes $f(x)$ if and only if $0 \in \partial f(x)$
- KKT conditions with nondifferentiable functions

# Example

$$f(x) = \max\{f_1(x),\ f_2(x)\} \quad f_1, f_2 \text{ convex and differentiable}$$




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

# Subdifferential

the **subdifferential** $\partial f(x)$ of $f$ at $x$ is the set of all subgradients:

$$\partial f(x) = \{g | g^\top (y-x) \leq f(y) - f(x)\} \quad \forall\, y \in \mathbf{dom}\, f$$

- $\partial f(x)$ is a closed convex set (possibly empty) (follows from the definition: $\partial f(x)$ is an intersection of halfspaces)
- if $x \in \mathbf{int\ dom}\, f$ then $\partial f(x)$ is nonempty and bounded (proof on next two pages)

*proof*: we show that $\partial f(x)$ is nonempty when $x \in \mathbf{int\,dom}\, f$

- $(x, f(x))$ is in the boundary of the convex set **epi** $f$
- therefore there exists a supporting hyperplane to **epi** $f$ at $(x, f(x))$:

$$\exists\, (a,b) \neq 0, \quad \begin{bmatrix} a \\ b \end{bmatrix}^\top \left( \begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall\, (y,t) \in \mathbf{epi}\, f$$

- $b > 0$ gives a contradiction as $t \to \infty$
- $b = 0$ gives a contradiction for $y = x + \epsilon a$ with small $\epsilon > 0$
- therefore $b < 0$ and $g = a/|b|$ is a subgradient of $f$ at $x$

*proof*: we show that $\partial f(x)$ is empty when $x \in \textbf{int dom} f$

- for small $r > 0$, define a set of $2n$ points

$$B = \{x \pm re_k | k = 1, \cdots, n\} \subset \textbf{dom} f$$

  and define $M = \max_{y \in B} f(y) < \infty$

- for every nonzero $g \in \partial f(x)$, there is a point $y \in B$ with

$$f(y) \geq f(x) + g^\top (y - x) = f(x) + r\|g\|_\infty$$

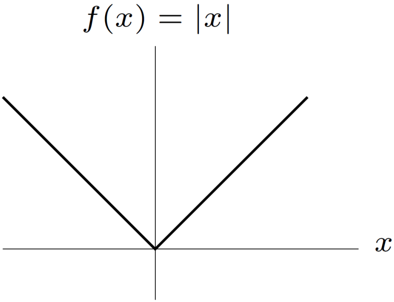
  (choose an index $k$ with $|g_k| = \|g\|_\infty$, and take $y = x + r\textbf{sign}(g_k)e_k$)

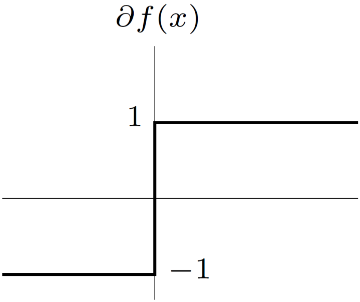
- therefore $\partial f(x)$ is bounded:

$$\sup_{g \in \partial f(x)} \|g\|_\infty \leq \frac{M - f(x)}{r}$$

# Examples

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




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

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

# Monotonicity

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

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

*proof*: by definition

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

combining the two inequalities shows monotonicity

# Examples of non-subdifferentiable functions

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

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

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

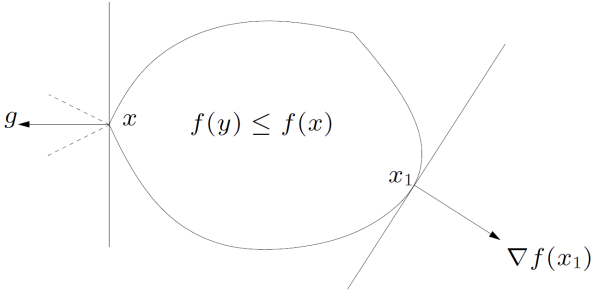
- $f : \mathbf{R} \to \mathbf{R}, \mathbf{dom}\, f = \mathbf{R}_+$

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

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




nonzero subgradients at $x$ define supporting hyperplanes to sublevel set

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

# Outline

- definition
- **subgradient calculus**
- duality and optimality conditions
- directional derivative

# Subgradient calculus

**weak subgradient calculus:** rules for finding *one* subgradient

- sufficient for most nondifferentiable convex optimization algorithms
- 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 entire subdifferential
- can be quite complicated

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

# 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

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

# Pointwise maximum

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

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 $f_k$ at $x$

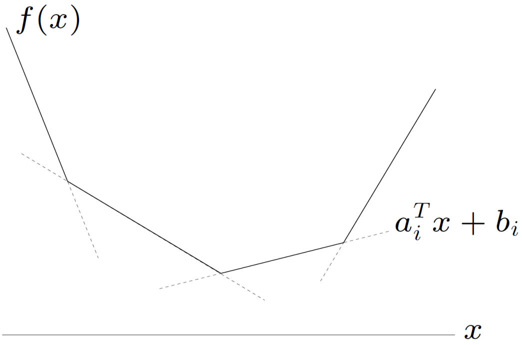
**strong result**

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

- 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: piecewise-linear function

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




the subdifferential at $x$ is a polyhedron

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

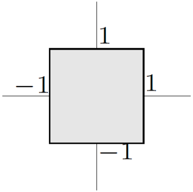
with $I(x) = \{i \mid a_i^\top x + b_i = f(x)\}$

# Example: $\ell_1$-norm

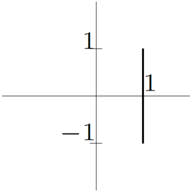
$$f(x) = \|x\|_1 = \max_{s \in \{-1,1\}^n} s^\top x$$

the subdifferential is a product of intervals

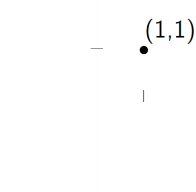
$$\partial f(x) = J_1 \times \cdots \times J_n, \quad J_k = \begin{cases} [-1, 1], & x_k = 0 \\ \{1\}, & x_k > 0 \\ \{-1\}, & x_k < 0 \end{cases}$$




$\partial f(0,0) = [-1,1] \times [-1,1]$




$\partial f(1,0) = \{1\} \times [-1,1]$




$\partial f(1,1) = \{(1,1)\}$

# Pointwise supremum

$$f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x), \quad f_\alpha(x) \text{ convex in } x \text{ for every } \alpha$$

**weak result:** to find a subgradient at $x$,

- find *any* $\beta$ for which $f(\hat{x}) = f_\beta(\hat{x})$ (assuming maximum is attained)
- choose *any* $g \in \partial f_\beta(\hat{x})$

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

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

equality requires extra conditions ($e.g.$, $\mathcal{A}$ compact, $f_\alpha$ continuous in $\alpha$)

# Exercise: maximum eigenvalue

**problem:** explain how to find a subgradient of

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

where $A(x) = A_0 + x_1A_1 + \cdots + x_nA_n$ with symmetric coefficients $A_i$

**solution:** to find a subgradient at $\hat{x}$,

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

$$(y^\top A_1y, \cdots, y^\top A_ny) \in \partial f(\hat{x})$$

# Minimization

$$f(x) = \inf_y h(x, y), \quad h \text{ jointly convex in } (x, y)$$

**weak result:** to find a subgradient at $\hat{x}$

- find $\hat{y}$ that minimizes $h(\hat{x}, y)$ (assuming minimum is attained)
- find subgradient $(g, 0) \in \partial h(\hat{x}, \hat{y})$

*proof*: for all $x, y$

$$\begin{aligned} h(x, y) &\geq h(\hat{x}, \hat{y}) + g^\top (x - \hat{x}) + 0^\top (y - \hat{y}) \\ &= f(\hat{x}) + g^\top (x - \hat{x}) \end{aligned}$$

therefore

$$f(x) = \inf_y h(x, y) \geq f(\hat{x}) + g^\top (x - \hat{x})$$

# Exercise: Euclidean distance to convex set

**problem:** explain how to find a subgradient of

$$f(x) = \inf_{y \in C} \|x - y\|_2$$

where C is a closed convex set

**solution:** to find a subgradient at $\hat{x}$,

- if $f(\hat{x}) = 0$ (that is, $\hat{x} \in C$), take $g = 0$
- if $f(\hat{x}) > 0$, find projection $\hat{y} = P(\hat{x})$ on $C$; take

$$g = \frac{1}{\|\hat{y} - \hat{x}\|_2}(\hat{x} - \hat{y}) = \frac{1}{\|\hat{x} - P(\hat{x})\|_2}(\hat{x} - P(\hat{x}))$$

# Composition

$$f(x) = h(f_1(x), \cdots, f_k(x)), h \text{ convex nondecreasing, } f_i \text{ convex}$$

**weak result:** to find a subgradient at $\hat{x}$,

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

reduces to standard formula for differentiable $h, f_i$

*proof*:

$$\begin{aligned} f(x) &\geq h\left(f_1(\hat{x}) + g_1^\top (x - \hat{x}), \cdots, f_k(\hat{x}) + g_k^\top (x - \hat{x})\right) \\ &\geq h(f_1(\hat{x}), \cdots, f_k(\hat{x})) + z^\top \left(g_1^\top (x - \hat{x}), \cdots, g_k^\top (x - \hat{x})\right) \\ &= f(\hat{x}) + g^\top (x - \hat{x}) \end{aligned}$$

# Optimal value function

define $h(u, v)$ as the optimal value of convex problem

$$\begin{array}{ll} \min & f_0(x) \\ \text{s.t.} & f_i(x) \le u_i,\ i = 1, \cdots, m \\ & Ax = b + v \end{array}$$

(functions $f_i$ are convex; optimization variable is $x$)

**weak result:** suppose $h(\hat{u}, \hat{v})$ is finite, strong duality holds with the dual

$$\begin{array}{ll} \max & \inf_x \left( f_0(x) + \sum_i \lambda_i (f_i(x) - \hat{u}_i) + \nu^\top (Ax - b - \hat{v}) \right) \\ \text{s.t.} & \lambda \succeq 0 \end{array}$$

if $\hat{\lambda}, \hat{\nu}$ are optimal dual variables (for r.h.s. $\hat{u}, \hat{v}$) then $(\hat{\lambda}, \hat{\nu}) \in \partial h(\hat{u}, \hat{v})$

*proof* : by weak duality for problem with r.h.s. $u, v$

$$
\begin{aligned}
h(u,v) &\geq \inf_x \left( f_0(x) + \sum_i \hat{\lambda}_i (f_i(x - u_i) + \hat{\nu}^\top (Ax - b - v) \right) \\
&= \inf_x \left( f_0(x) + \sum_i \hat{\lambda}_i (f_i(x - \hat{u}_i) + \hat{\nu}^\top (Ax - b - \hat{v}) \right) \\
&- \hat{\lambda}^\top (u - \hat{u}) - \hat{\nu}^\top (v - \hat{v}) \\
&= h(\hat{u}, \hat{v}) - \hat{\lambda}^\top (u - \hat{u}) - \hat{\nu}^\top (v - \hat{v})
\end{aligned}
$$

# Expectation

$$f(x) = \mathbf{E}h(x,u) \quad u \text{ random}, \; h \text{ convex in } x \text{ for every } u$$

**weak result:** to find a subgradient at $\hat{x}$

- choose a function $\mapsto g(u)$ with $g(u) \in \partial_x h(\hat{x}, u)$
- then, $g = \mathbf{E}_u \, g(u) \in \partial f(\hat{x})$

*proof* : by convexity of $h$ and definition of $g(u)$,

$$\begin{aligned} f(x) &= \mathbf{E}h(x,u) \\ &\geq \mathbf{E}\left(h(\hat{x},u) + g(u)^\top (x - \hat{x})\right) \\ &= f(\hat{x}) + g^\top (x - \hat{x}) \end{aligned}$$

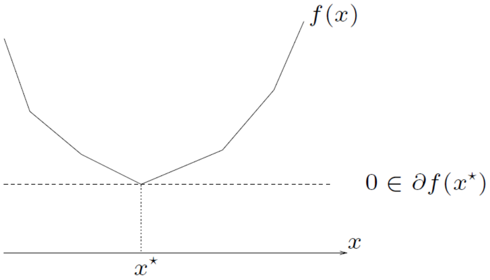
# Outline

- definition
- subgradient calculus
- **duality and optimality conditions**
- directional derivative

# Optimality conditions - unconstrained

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

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




*proof* : by definition

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

# Example: piecewise linear minimization

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

**optimality condition**

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

in other words, $x^*$ is optimal if and only if there is a $\lambda$ with

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

these are the optimality conditions for the equivalent linear program

$$\begin{array}{ll} \min & t \\ \text{s.t.} & Ax + b \leq t\mathbf{1} \end{array} \qquad \begin{array}{ll} \max & b^\top \lambda \\ \text{s.t.} & A^\top \lambda = 0 \\ & \lambda \geq 0, \quad \mathbf{1}^\top \lambda = 1 \end{array}$$

# Optimality conditions - constrained

$$
\begin{aligned}
\min \quad & f_0(x) \\
\text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \cdots, m
\end{aligned}
$$

**from Lagrange duality**

if strong duality holds, then $x^*, \lambda^*$ 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 $i = 1, \cdots, m$
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 $\mathbf{dom}\, f_i = \mathbf{R}^n$)
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 $f_i$

# Outline

- definition
- subgradient calculus
- duality and optimality conditions
- **directional derivative**

# Directional derivative

**Definition** (general $f$): directional derivative of $f$ at $x$ in the direction $y$ is

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

(if the limit exists)

- $f'(x;y)$ is the right derivative of $g(\alpha) = f(x+\alpha y)$ at $\alpha = 0$
- $f'(x;y)$ is homogeneous in $y$:

$$f'(x;\lambda y) = \lambda f'(x;y) \text{ for } \lambda \geq 0$$

# Directional derivative of a convex function

**equivalent definition** (convex $f$): replace lim with inf

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

*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 $t$ (EE236B ex. A2.5); hence

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

# Properties

consequences of the expressions (for convex $f$)

$$f'(x;y) = \lim_{\alpha > 0} \frac{f(x+\alpha y) - f(x)}{\alpha}$$
$$= \lim_{t>0} \left( t(f(x + \frac{1}{t}y) - tf(x) \right)$$

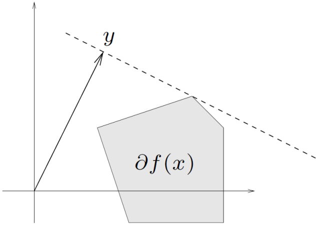
- $f'(x;y)$ is convex in $y$ (partial minimization of a convex function in $y, t$)
- $f'(x;y)$ defines a lower bound on $f$ in the direction $y$:

$$f(x+\alpha y) \geq f(x) + \alpha f'(x;y) \quad \forall \alpha \geq 0$$

# Directional derivative and subgradients

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

$$f'(x; y) = \sup_{g \in \partial f(x)} g^\top y$$




$f'(x; y)$ is *support function* of $\partial f(x)$

- generalizes $f'(x; y) = \nabla f(x)^\top y$ for differentiable functions
- implies that $f'(x; y)$ exists for all $x \in \textbf{int dom} f$, all $y$ (see page 6)

*proof* : if $g \in \partial f(x)$ then from p 35

$$f'(x;y) \geq \inf_{\alpha>0} \frac{f(x) + \alpha g^\top y - f(x)}{\alpha} = g^\top y$$

it remains to show that $f'(x;y) = \hat{g}^\top y$ for at least one $\hat{g} \in \partial f(x)$

- $f'(x;y)$ is convex in $y$ with domain $\mathbf{R}^n$, hence subdifferentiable at all $y$
- let $\hat{g}$ be a subgradient of $f'(x;y)$ at $y$: for all $v, \lambda \geq 0$,

$$\lambda f'(x;v) = f'(x;\lambda v) \geq f'(x;y) + \hat{g}^\top(\lambda v - y)$$

- taking $\lambda \to \infty$ shows $f'(x;v) \geq \hat{g}^\top v$; from the lower bound on p 34

$$f(x+v) \geq f(x) + f'(x;v) \geq f(x) + \hat{g}^\top v \quad \forall v$$

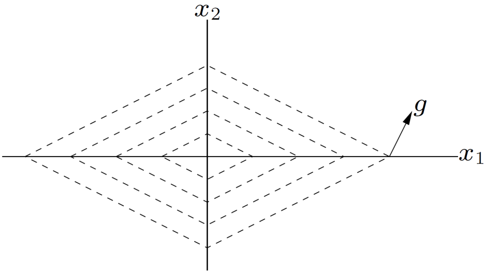
- hence $\hat{g} \in \partial f(x)$; taking $\lambda = 0$ we see that $f'(x;y) \leq \hat{g}^\top y$

# Descent directions and subgradients

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

- negative gradient of differentiable $f$ is descent direction (if $\nabla f(x) \neq 0$)
- negative subgradient 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

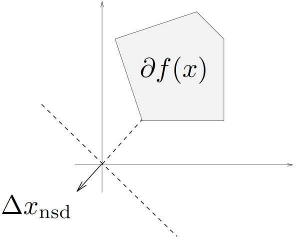
**definition:** (normalized) steepest descent direction at $x \in \textbf{int dom} f$ is

$$\triangle x_{\text{nsd}} = \operatorname*{argmin}_{\|y\|_2 \le 1} f'(x; y)$$

$\triangle x_{\text{nsd}}$ is the primal solution $y$ of the pair of dual problems (BV $\mathcal{S}8.1.3$)

$$\begin{array}{rl} \min\ (\text{over } y) & f'(x;y) \\ \text{s.t.} & \|y\|_2 \le 1 \end{array} \qquad\qquad \begin{array}{rl} \max\ (\text{over } g) & -\|g\|_2 \\ \text{s.t.} & g \in \partial f(x) \end{array}$$

- optimal $g^*$ is subgradient with least norm
- $f'(x; \triangle x_{\text{nsd}}) = -\|g^*\|_2$
- if $0 \notin \partial f(x)$, $\triangle x_{\text{nsd}} = -g^*/\|g^*\|_2$



# Subgradients and distance to sublevel sets

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

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

- $-g$ is descent direction for $\|x - y\|_2$, for **any** $y$ with $f(y) < f(x)$
- in particular, $-g$ is descent direction for distance to any minimizer of $f$

# References


- J.-B. Hiriart-Urruty, C. Lemar echal, *Convex Analysis and Minimization Algoritms* (1993), chapter VI.
- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004), section 3.1.
- B. T. Polyak, *Introduction to Optimization* (1987), section 5.1.