ORF522 – Linear and Nonlinear Optimization

12. Subgradients

Bartolomeo Stellato — Fall 2024

Today's lecture

[Chapter 3 and 8, FMO][ee364b][Chapter 1 LSCO][Chapter 3, ILCO]

Gradient descent

- Line search

Subgradients

- Geometric definitions
- Subgradients
- Subgradient calculus
- Optimality conditions based on subgradients

Line search

Exact line search

Choose the best step along the descent direction

$$t_k = \underset{t \geq 0}{\operatorname{argmin}}\, f(x^k - t\nabla f(x^k))$$

Used when

- computational cost very low or
- there exist closed-form solutions

In general, impractical to perform exactly

Backtracking line search

Condition

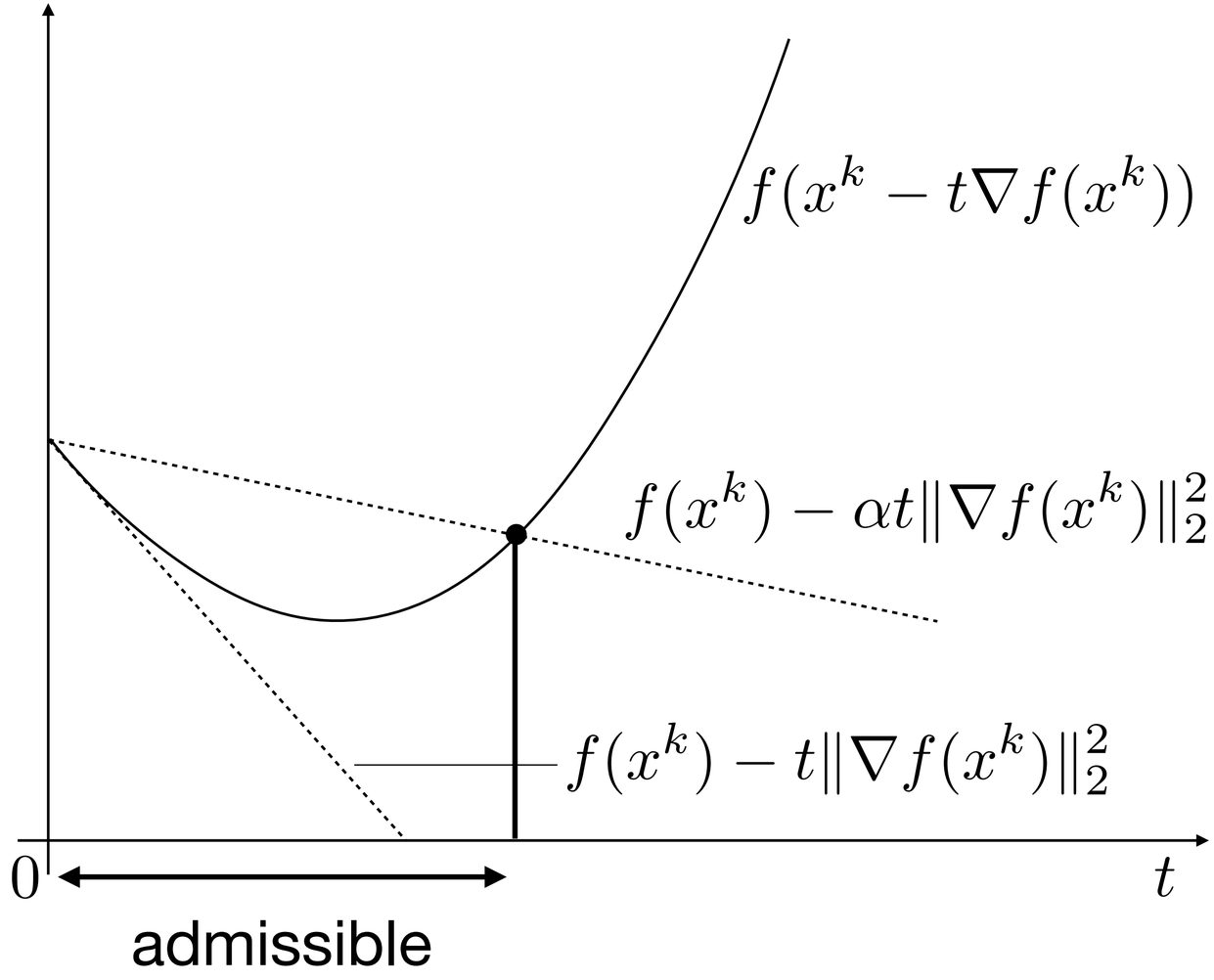
Armijo condition: for some $0 < \alpha \leq 1/2$

$$f(x^k + td^k) < f(x^k) + \alpha t \nabla f(x^k)^T d^k$$

where $d^k = -\nabla f(x^k)$

$$f(x^k - t\nabla f(x^k)) < f(x^k) - \alpha t \|\nabla f(x^k)\|_2^2$$

Guarantees
sufficient decrease
in objective value



Backtracking line search

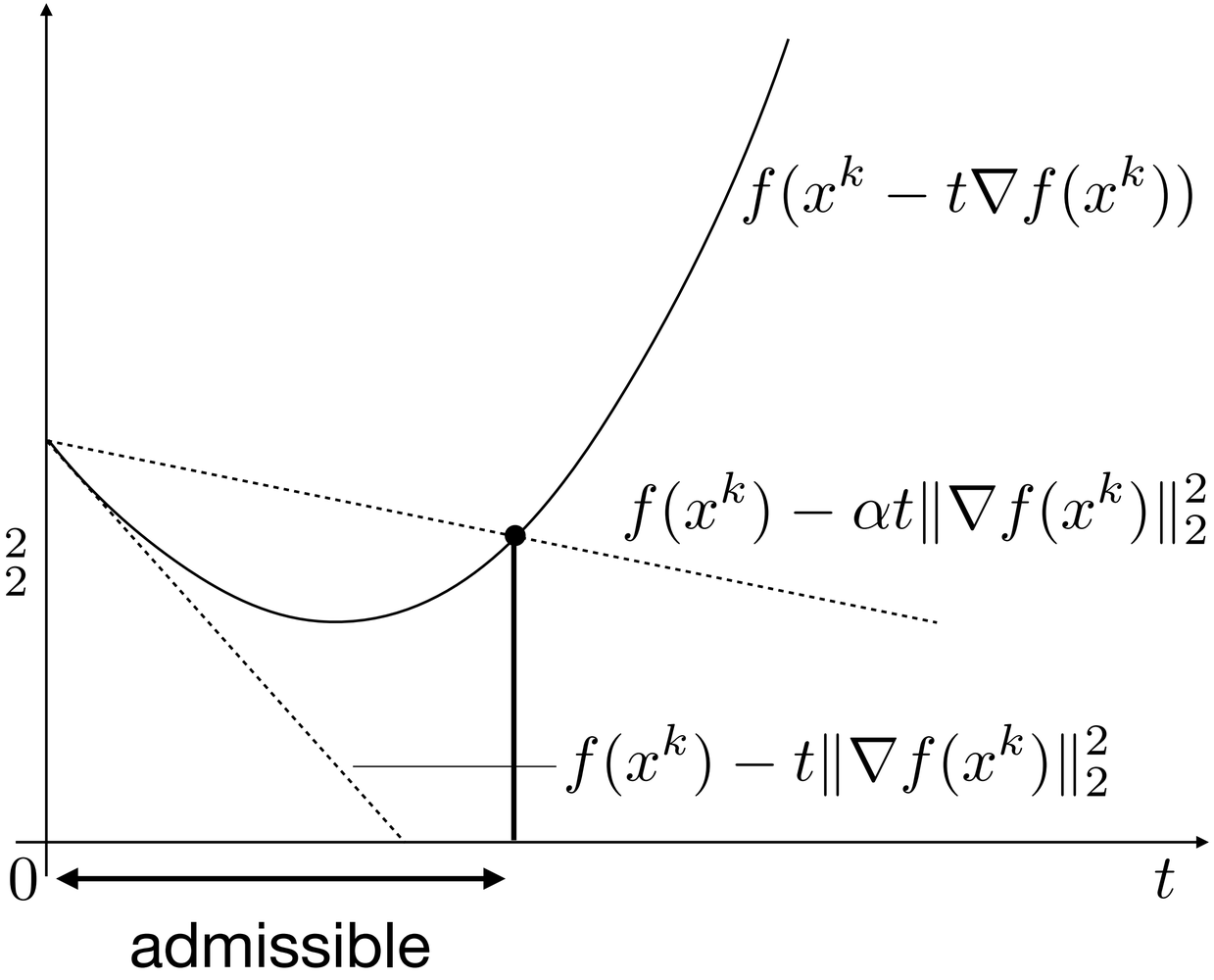
Iterations

initialization

$t = 1, \quad 0 < \alpha \leq 1/2, \quad 0 < \beta < 1$

while $f(x^k - t\nabla f(x^k)) > f(x^k) - \alpha t \|\nabla f(x^k)\|_2^2$

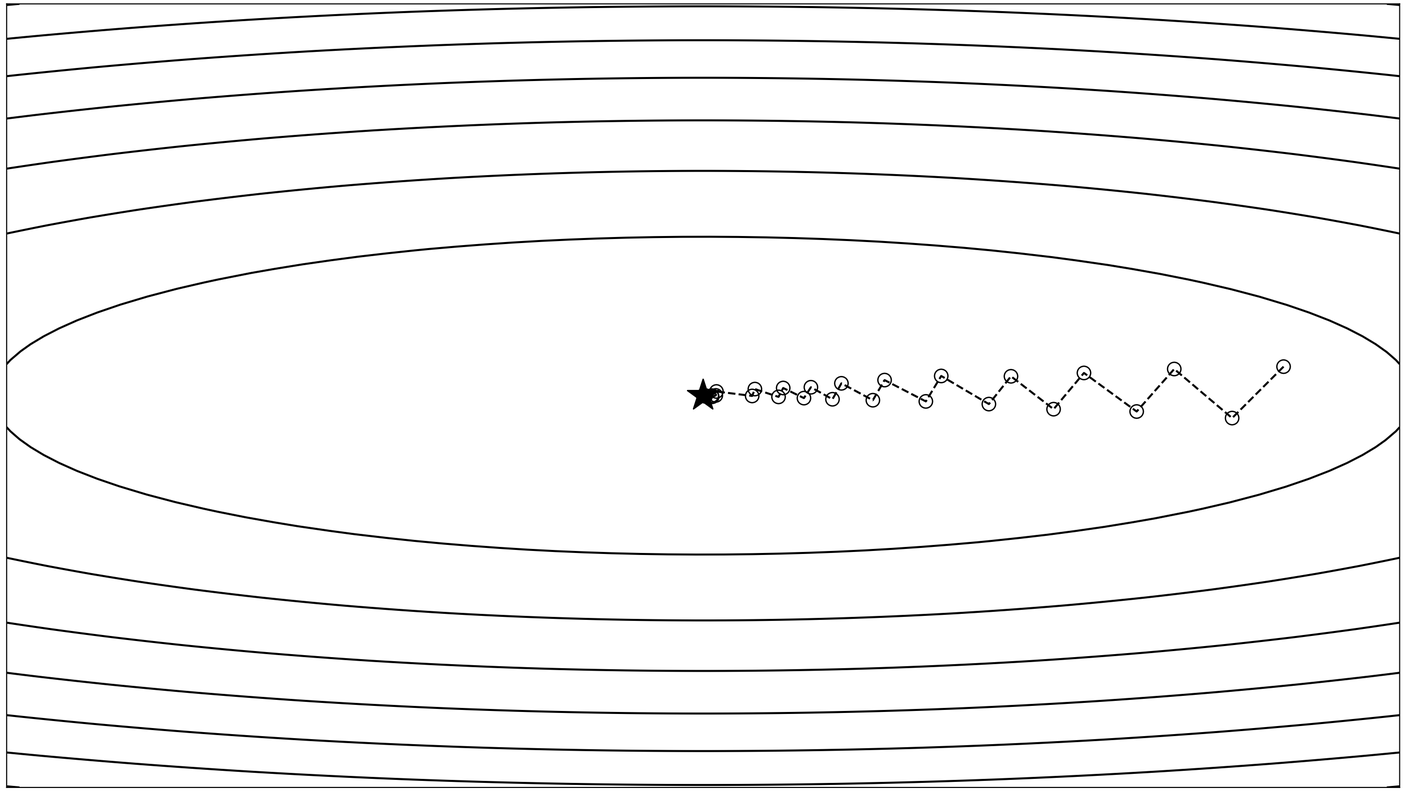
$t \leftarrow \beta t$



Backtracking line search

$$f(x) = (x_1^2 + 20x_2^2)/2$$

$$x^0 = (20, 1)$$



Backtracking line search

Converges in 31 iterations

Backtracking line search convergence

Theorem

Let f be L-smooth. Gradient descent with backtracking line search satisfies

$$f(x^k) - f(x^\star) \leq \frac{\|x^0 - x^\star\|_2^2}{2t_{\min}k}$$

where $t_{\min} = \min\{1, \beta/L\}$

Proof almost identical to fixed step case

Remarks

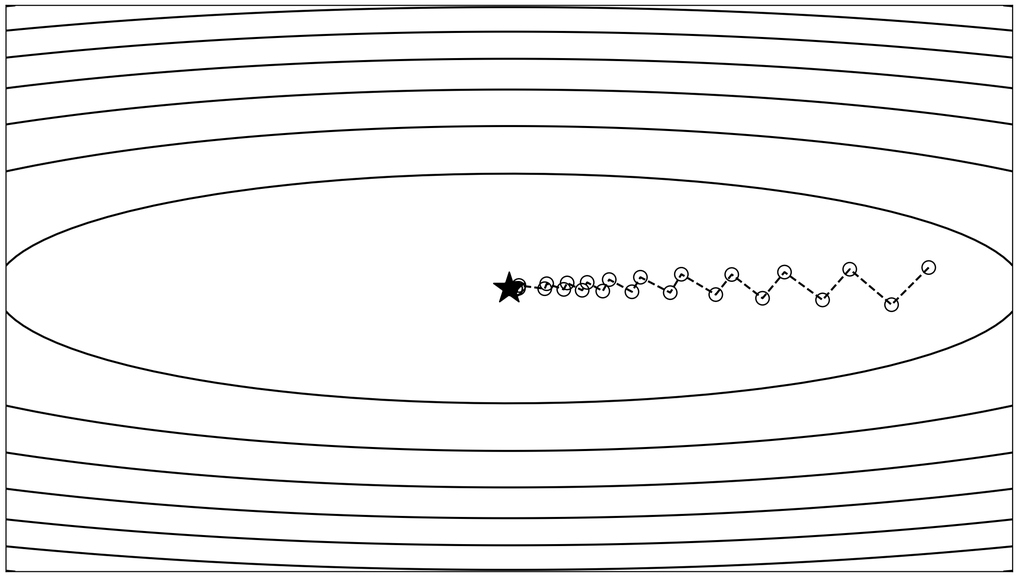
- If $\beta \approx 1$, similar to optimal step-size (β/L vs $1/L$)
- Still convergence rate $O(1/\epsilon)$ iterations (can be very slow!)

Gradient descent issues

Slow convergence

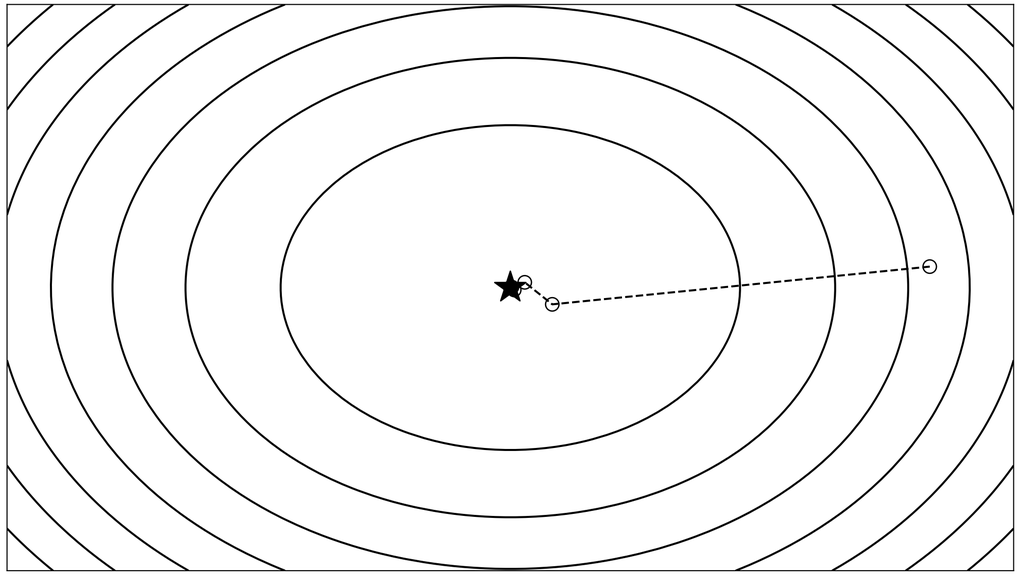
Very dependent on scaling

$$f(x) = (x_1^2 + 20x_2^2)/2$$



Slow convergence

$$f(x) = (x_1^2 + 2x_2^2)/2$$

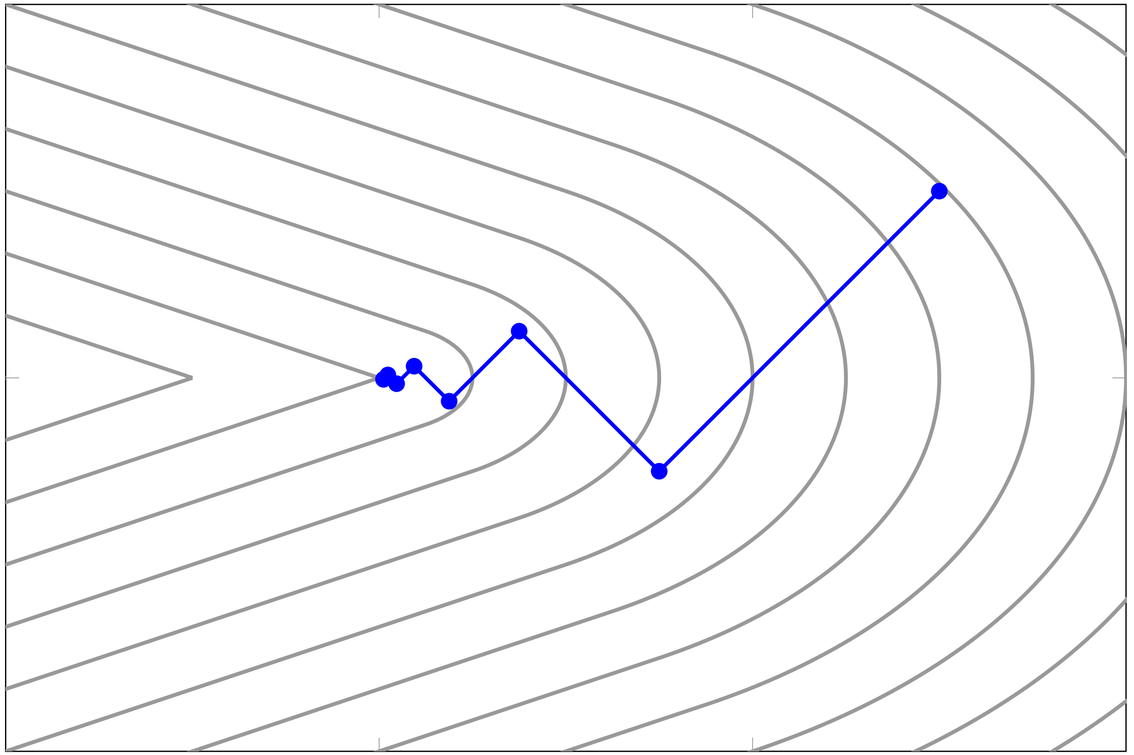


Faster

Non-differentiability

Wolfe's example

$$f(x) = \begin{cases} \sqrt{x_1^2 + \gamma x_2^2} & |x_2| \leq x_1 \\ \dfrac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}} & |x_2| > x_1 \end{cases}$$



Gradient descent with *exact line search* gets stuck at $x = (0, 0)$

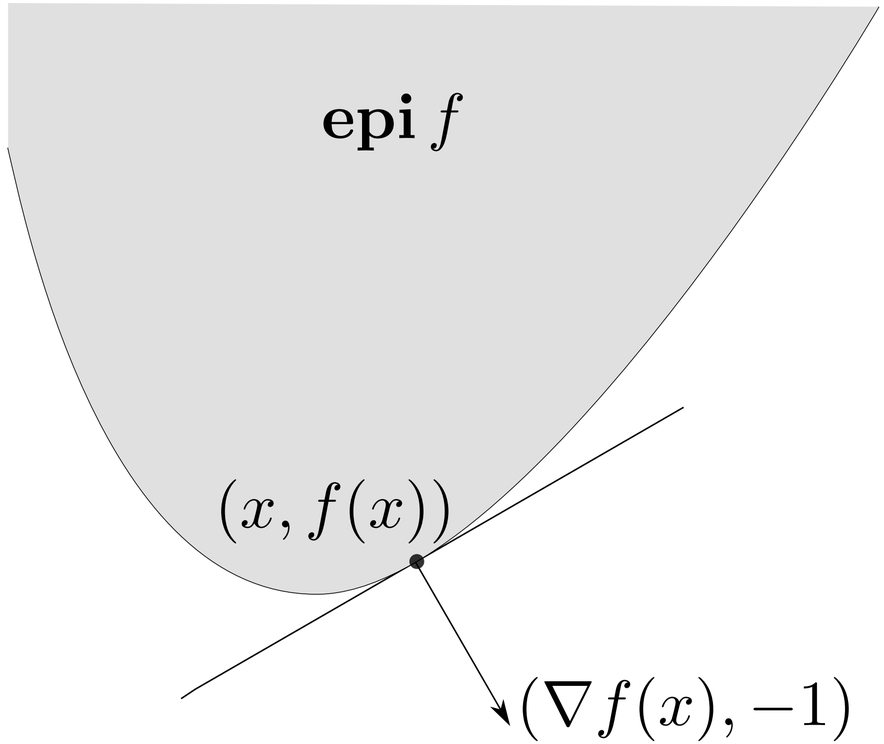
In general: gradient descent cannot handle non-differentiable functions and constraints

Subgradients

Gradients and epigraphs

For a convex differentiable function f, i.e.

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



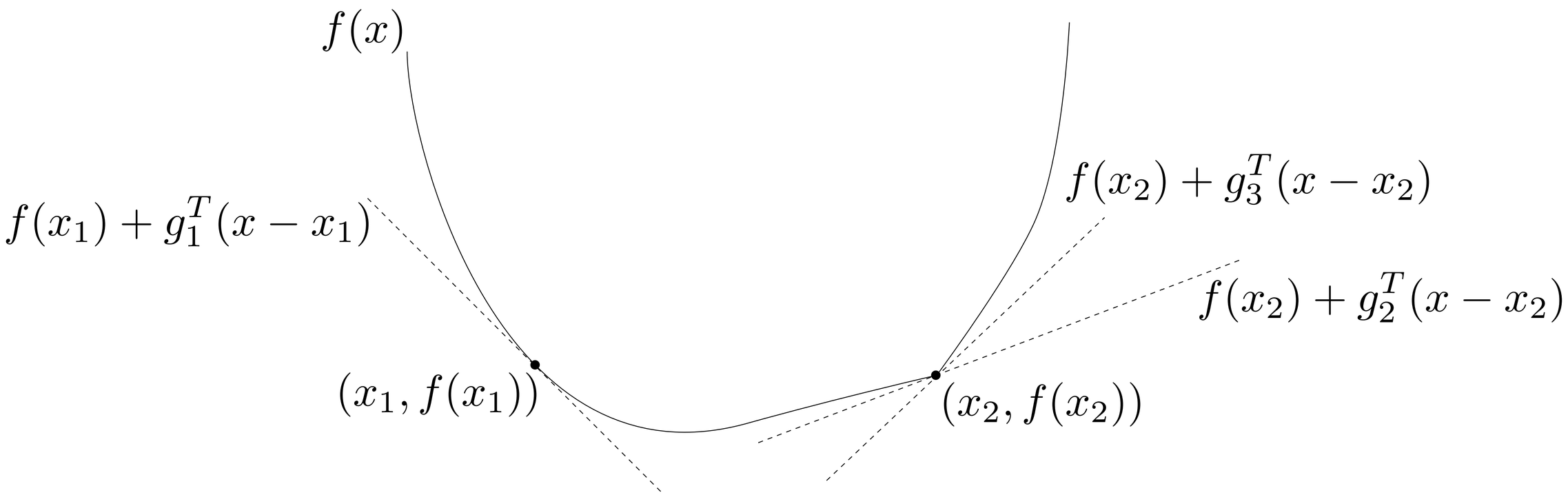
$(\nabla f(x), -1)$ defines a **supporting hyperplane** to epigraph of f at $(x, f(x))$

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

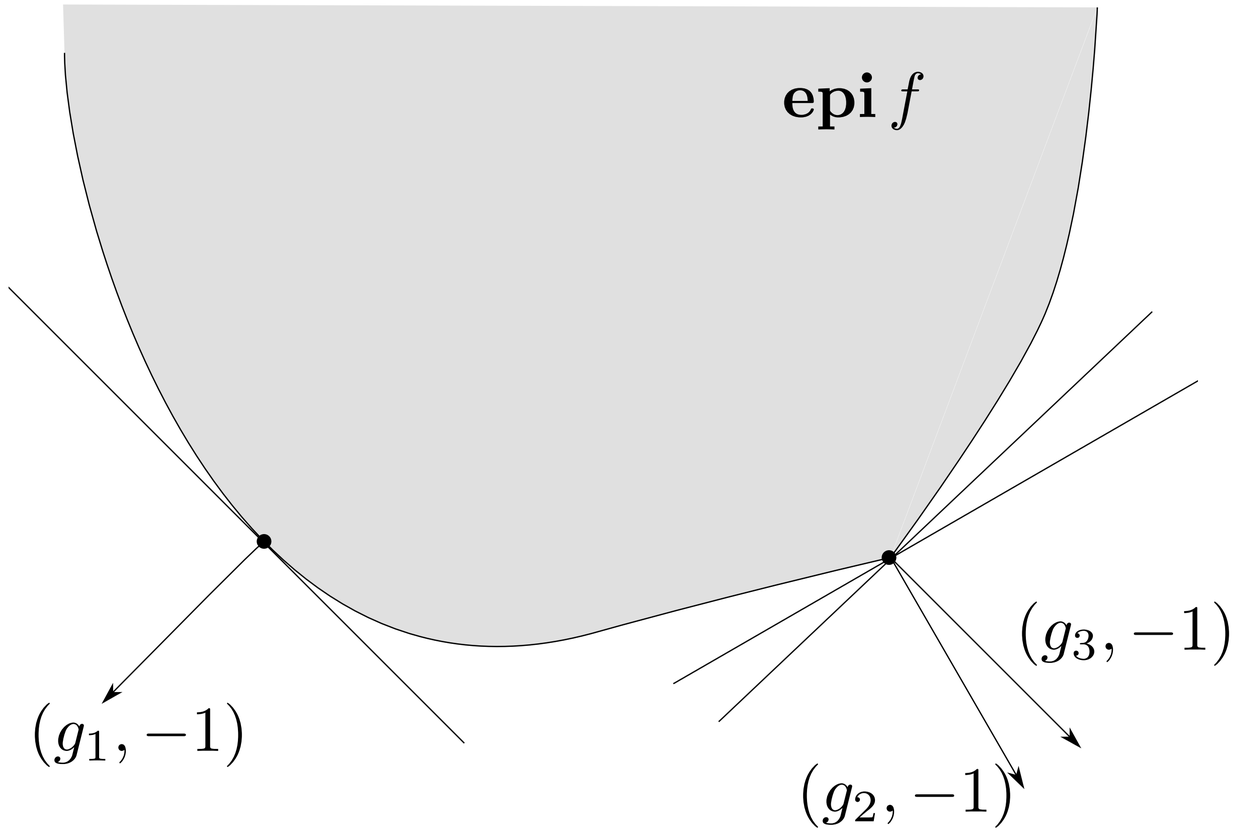
Subgradient

We say that g is a **subgradient** of function f at point x if

$$f(y) \geq f(x) + g^T(y - x), \quad \forall y$$



Subgradient properties



g is a subgradient of f at x iff $(g, -1)$ supports $\mathbf{epi}\, f$ at $(x, f(x))$

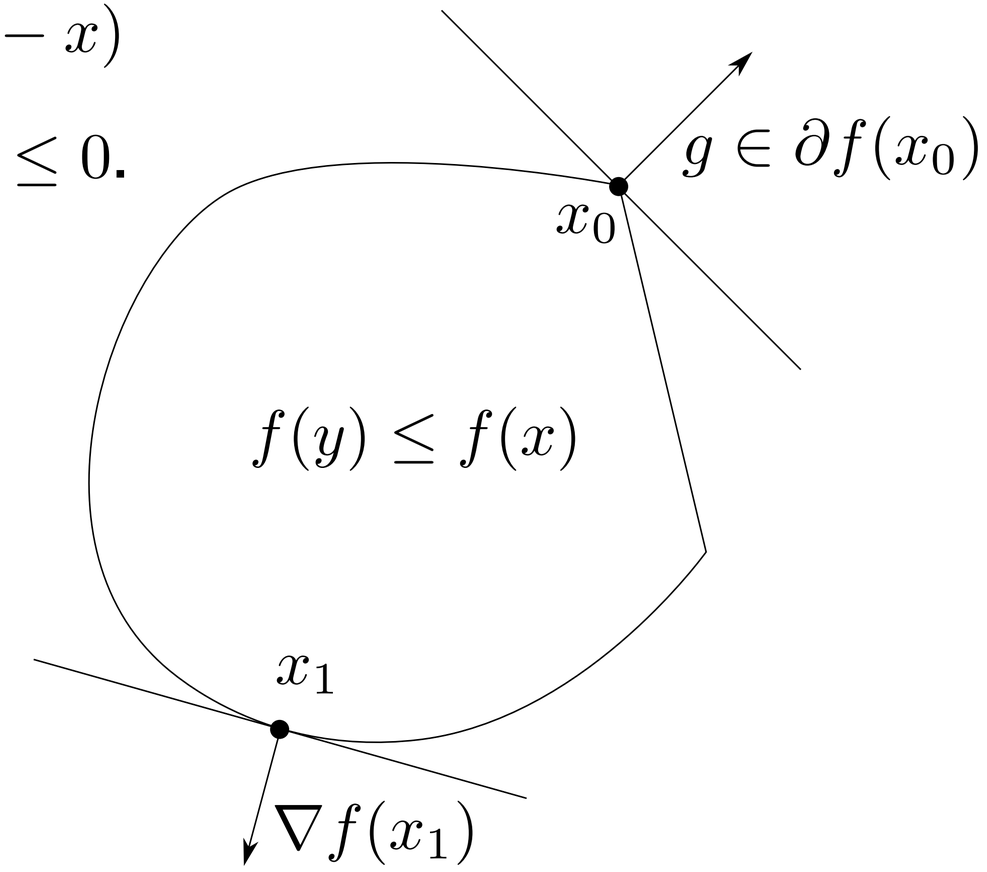
g is a subgradient of f iff $f(x) + g^T(y - x)$ is a global underestimator of f

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

(Sub)gradients and sublevel sets

g being a subgradient of f means $f(y) \geq f(x) + g^T(y - x)$

Therefore, if $f(y) \leq f(x)$ (sublevel set), then $g^T(y - x) \leq 0$.



f differentiable at x
$\nabla f(x)$ is normal to the sublevel set $\{y \mid f(y) \leq f(x)\}$

f nondifferentiable at x
subgradients define supporting hyperplane to sublevel set throgh x

Subdifferential

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

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

Properties

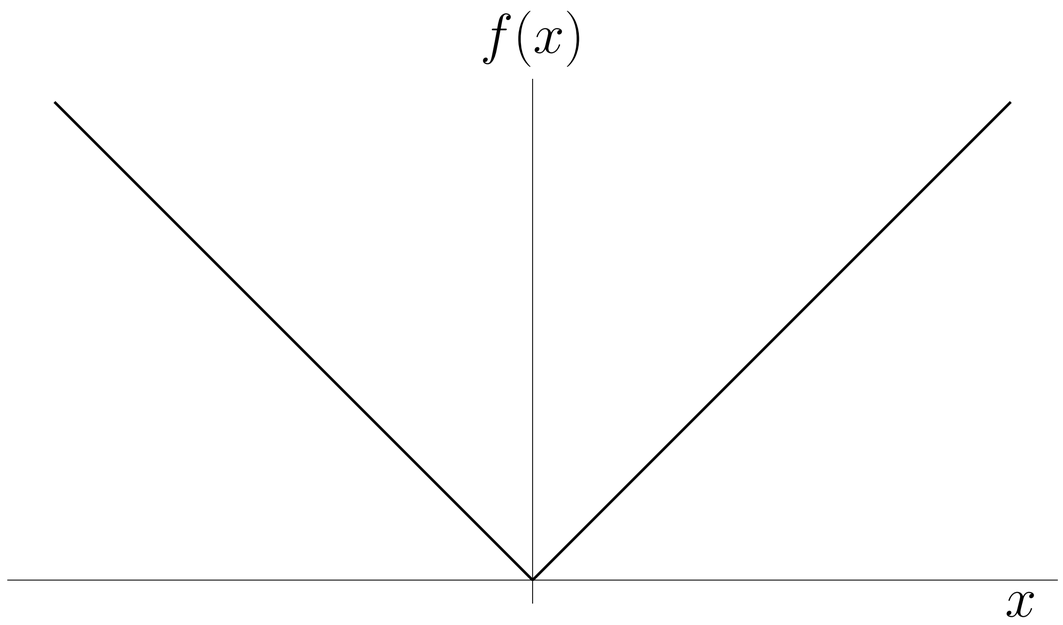
- $\partial f(x)$ is always closed and convex, also for nonconvex f.
  (intersection of halfspaces)
- If f is convex and differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- If f is convex and $\partial f(x) = \{g\}$, then f is differentiable at x and $g = \nabla f(x)$

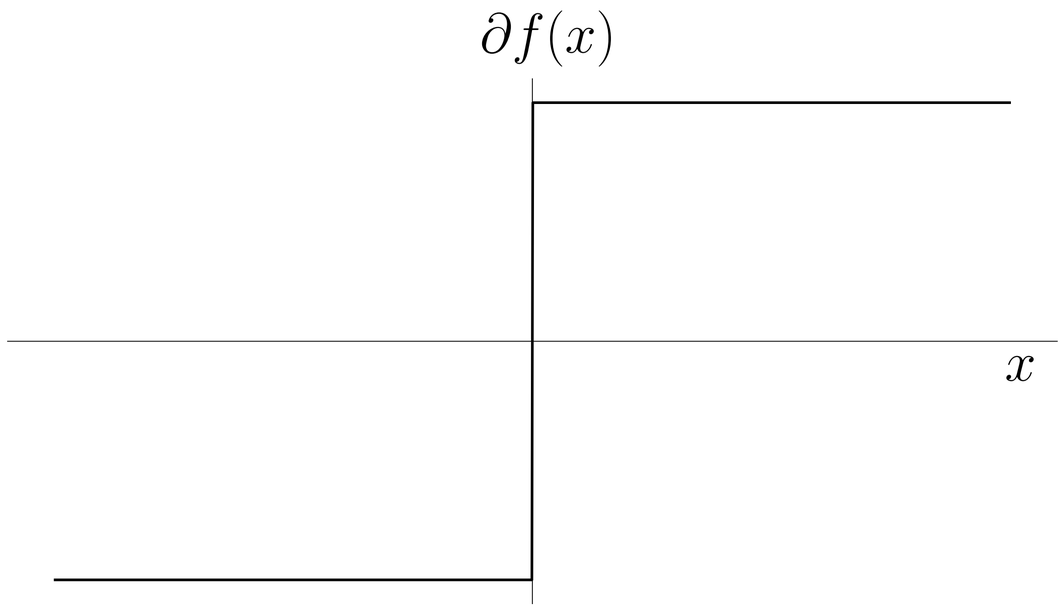
Example

Absolute value

$$f(x) = |x|$$

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





Subgradient calculus

Subgradient calculus

Strong subgradient calculus
Formulas for finding the whole subdifferential $\partial f(x)$ ⟶ **Hard**

Weak subgradient calculus
Formulas for finding *one* subgradient $g \in \partial f(x)$ ⟶ **Easy**

In practice, most algorithms require only *one* subgradient g at point x

Basic rules

Nonnegative scaling: $\partial(\alpha f) = \alpha \partial f$ with $\alpha > 0$

Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$

Affine transformation: $f(x) = h(Ax + b)$, then

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

Basic rules

Pointwise maxima

Finite pointwise maximum $f(x) = \max_{i=1,\dots,m} f_i(x)$, then

$$\partial f(x) = \mathbf{conv}\left(\bigcup\{\partial f_i(x) \mid f_i(x) = f(x)\}\right)$$ (convex hull of active functions)

General pointwise maximum (supremum) $f(x) = \max_{s \in S} f_s(x)$, then

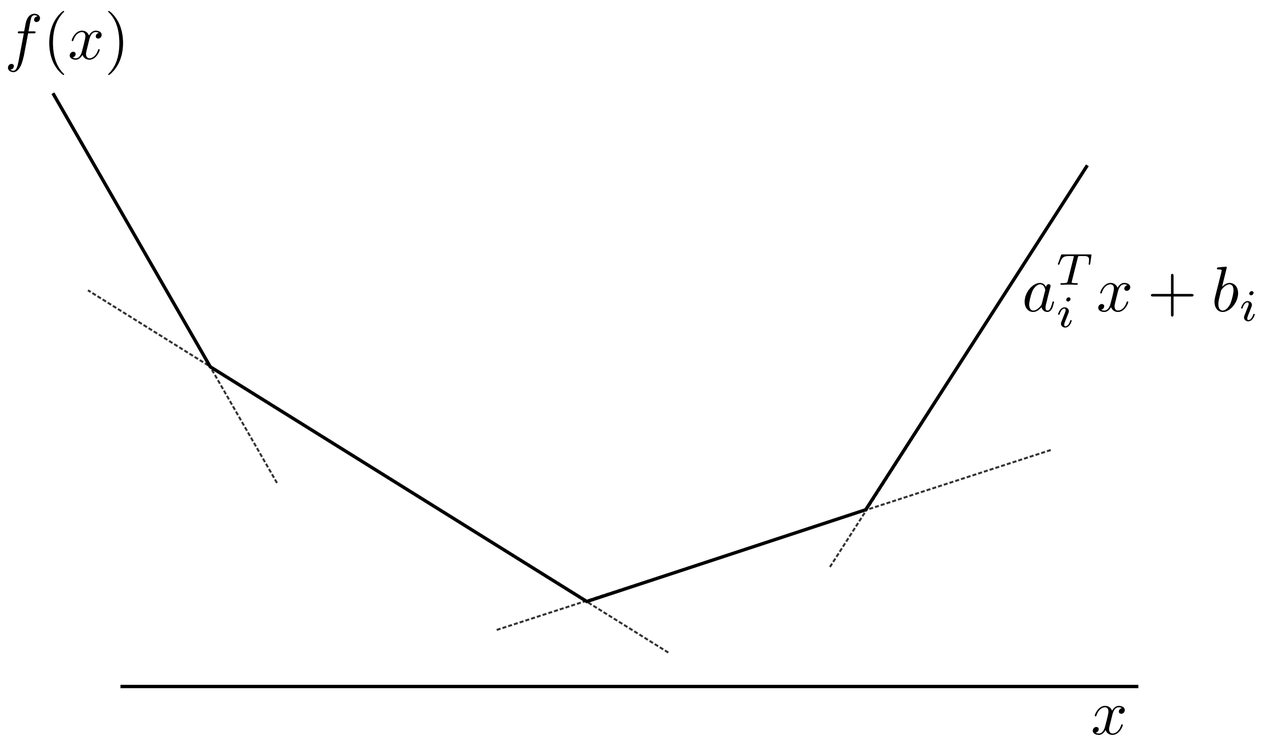
$$\partial f(x) \supseteq \mathbf{conv}\left(\bigcup\{\partial f_s(x) \mid f_s(x) = f(x)\}\right)$$

Note: Equality requires some regularity assumptions
(e.g. S compact and f_s is continuous in s)

Example

Piecewise linear function

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



Subdifferential is a polyhedron

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

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

Example

Norms

Given $f(x) = \|x\|_p$ we can express it as

$$\|x\|_p = \max_{\|z\|_q \leq 1} z^T x,$$

where q such that $1/p + 1/q = 1$ defines the **dual norm**. Therefore,

$$\partial f(x) = \operatorname*{argmax}_{\|z\|_q \leq 1} z^T x$$

Example: $f(x) = \|x\|_1 = \max_{\|s\|_\infty \leq 1} s^T x$

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

weak result

$$\mathbf{sign}(x) \in \partial f(x)$$

Basic rules

Composition

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

$$g = q_1 g_1 + \cdots + q_k g_k \in \partial f(x)$$

where $q \in \partial h(f_1(x), \ldots, f_k(x))$ and $g_i \in \partial f_i(x)$

Proof

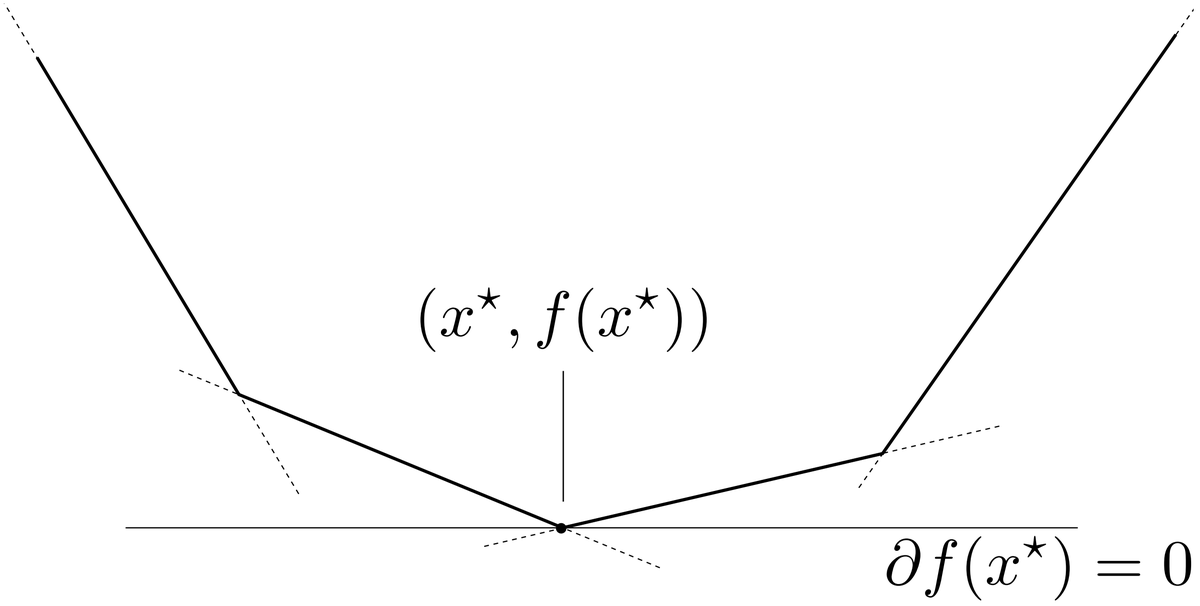
$$\begin{aligned} f(y) &= h(f_1(y), \ldots, f_k(y)) \\ &\geq h(f_1(x) + g_1^T(y-x), \ldots, f_k(x) + g_k^T(y-x)) \\ &\geq h(f_1(x), \ldots, f_k(x)) + q^T(g_1^T(y-x), \ldots, g_k^T(y-x)) \\ &= f(x) + g^T(y-x) \end{aligned}$$

Optimality conditions

Fermat's optimality condition

For a convex function f, then
$x^\star$ is a global minimizer if and only if

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



Proof
A subgradient $g = 0$ means that, for all y

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star) \quad \blacksquare$$

Note differentiable case with $\partial f(x) = \{\nabla f(x)\}$

Example: piecewise linear function

Optimality condition

$$f(x) = \max_{i=1,\ldots,m} (a_i^T x + b_i) \longrightarrow 0 \in \partial f(x) = \mathbf{conv}\{a_i \mid a_i^T x + b_i = f(x)\}$$

In other words, $x^\star$ is optimal if and only if $\exists \lambda$ such that

$$\lambda \geq 0, \quad \mathbf{1}^T \lambda = 1, \quad \sum_{i=1}^m \lambda_i a_i = 0 \quad \leftarrow (0 \in \partial f(x))$$

where $\lambda_i = 0$ if $a_i^T x^\star + b_i < f(x^\star)$

Same KKT optimality conditions as the primal-dual problems

minimize $\quad t$

subject to $\quad Ax + b \leq t\mathbf{1}$

maximize $\quad b^T \lambda$

subject to $\quad A^T \lambda = 0$

$\lambda \geq 0, \quad \mathbf{1}^T \lambda = 1$

Constrained optimization

Indicator function
of a convex set

$$\mathcal{I}_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

Constrained form

minimize $f(x)$

subject to $x \in C$

$\longleftrightarrow$

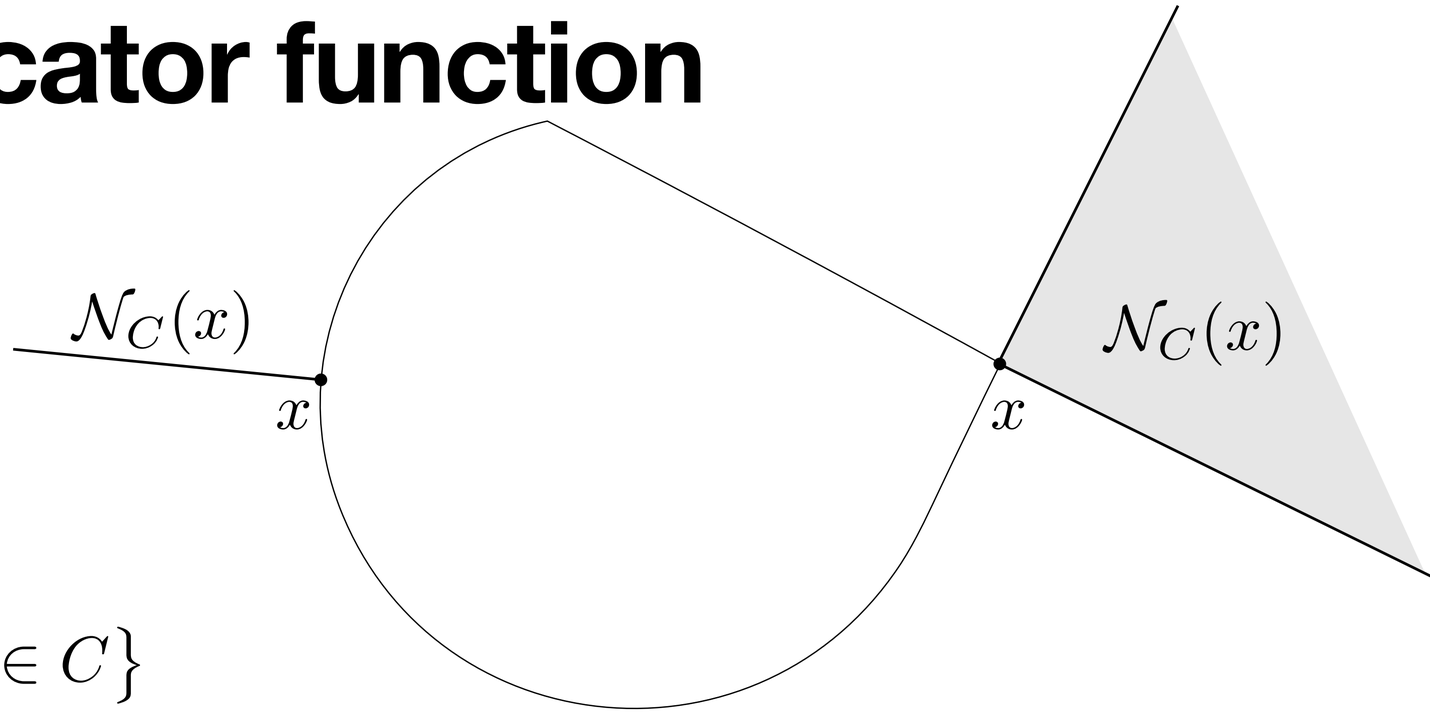
Unconstrained form

minimize $f(x) + \mathcal{I}_C(x)$

Subgradient of indicator function

The subdifferential of the **indicator functon** is the **normal cone**

$$\partial \mathcal{I}_C(x) = \mathcal{N}_C(x)$$



where,

$$\mathcal{N}_C(x) = \{g \mid g^T(y-x) \leq 0, \quad \text{for all } y \in C\}$$

Proof

By definition of subgradient g, $\mathcal{I}_C(y) \geq \mathcal{I}_C(x) + g^T(y-x), \quad \forall y$

$$y \notin C \implies \mathcal{I}_C(y) = \infty$$

$$y \in C \implies 0 \geq g^T(y-x)$$

First-order optimality conditions from subdifferentials

minimize $\quad f(x) + \mathcal{I}_C(x)$

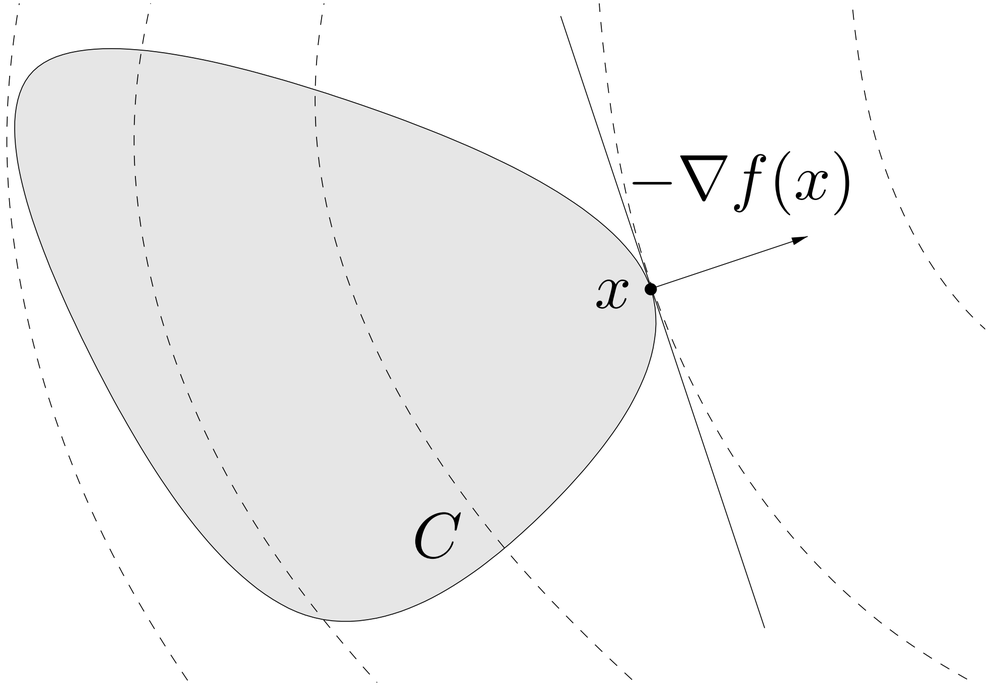
f convex smooth,
C convex

Fermat's optimality condition

$$0 \in \partial(f(x) + \mathcal{I}_C(x))$$

$$\iff \; 0 \in \{\nabla f(x)\} + \mathcal{N}_C(x)$$

$$\iff \; -\nabla f(x) \in \mathcal{N}_C(x)$$



Equivalent to

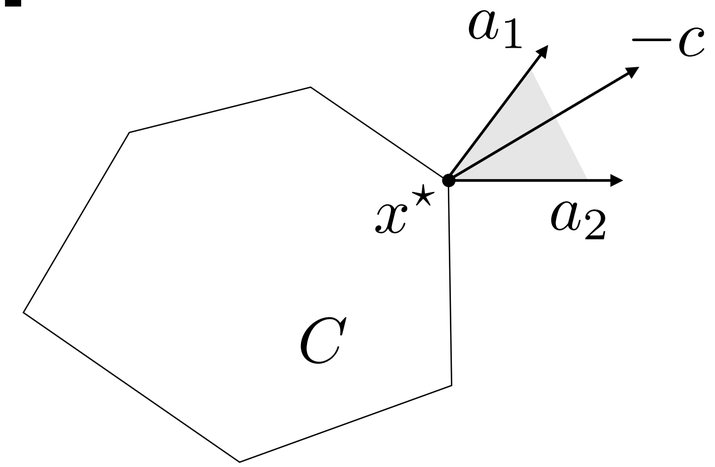
$$\nabla f(x)^T (y - x) \geq 0, \quad \forall y \in C$$

Normal cone condition

Linear program example

minimize $\quad c^T x$

subject to $\quad Ax \leq b$



Recap from Lecture 8

Two active constraints at optimum: $a_1^T x^\star = b_1, \quad a_2^T x^\star = b_2$

Optimal dual solution y satisfies:

$$A^T y + c = 0, \quad y \geq 0, \quad y_i = 0 \text{ for } i \neq \{1, 2\}$$

In other words, $-c = a_1 y_1 + a_2 y_2$ with $y_1, y_2 \geq 0$

Normal cone to polyhedron

$$-c \in \mathcal{N}_{\{Ax \leq b\}}(x^\star) = \{A^T y \mid y \geq 0 \quad \text{and} \quad y_i(a_i^T x^\star - b_i) = 0\}$$

Example: KKT of a quadratic program

minimize $(1/2)x^T Px + q^T x$
subject to $Ax \leq b$

$\longrightarrow$ minimize $(1/2)x^T Px + q^T x + \mathcal{I}_{\{Ax \leq b\}}(x)$

Gradient

$$\nabla f(x) = Px + q$$

Normal cone to polyhedron Proof: [Theorem 6.46, Variational Analysis, Rockafellar & Wets]

$$\mathcal{N}_{\{Ax \leq b\}}(x) = \{A^T y \mid y \geq 0 \quad \text{and} \quad y_i(a_i^T x - b_i) = 0\}$$

First-order optimality condition

$$-\nabla f(x) \in \partial \mathcal{I}_{\{Ax \leq b\}}(x) = \mathcal{N}_{\{Ax \leq b\}}(x)$$

$\longleftrightarrow$

KKT Optimality conditions

$$\begin{aligned} Px + q + A^T y &= 0 \\ y &\geq 0 \\ Ax - b &\leq 0 \\ y_i(a_i^T x - b_i) &= 0, \quad i = 1, \ldots, m \end{aligned}$$

Subgradient methods

Today, we learned to:

- **Analyze** gradient descent with line search
- **Understand** issues with gradient descent
- **Define** subgradients
- **Apply** subgradient calculus
- **Derive** optimality conditions from subgradients

Next lecture

- Subrgradient method
- Proximal algorithms