ORF522 – Linear and Nonlinear Optimization

15. Subgradient methods

Bartolomeo Stellato — Fall 2020

Ed forum

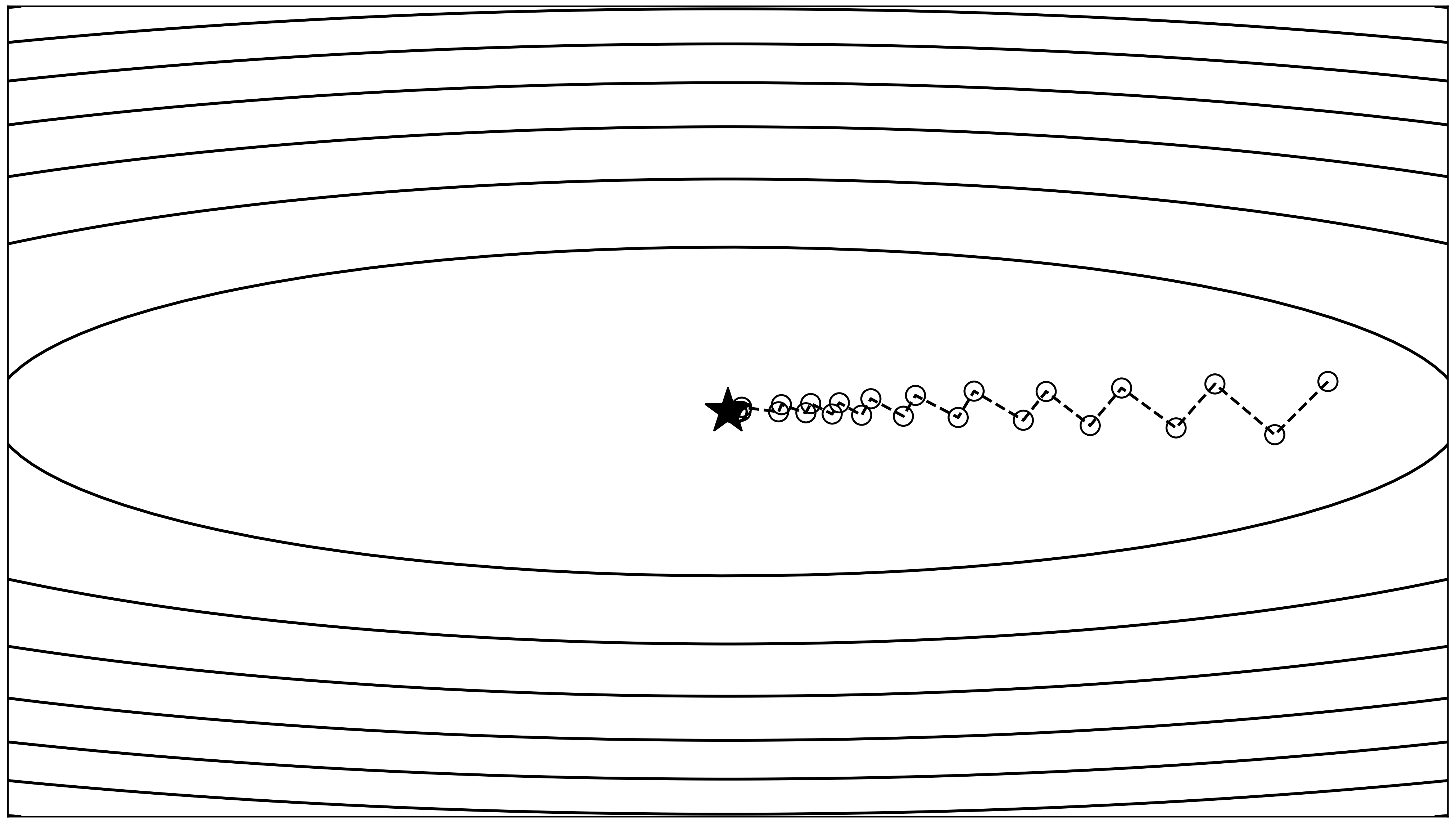
- Convex functions rule: nonincreasing (instead of decreasing) and nondecreasing (instead of increasing)
- Without strong convexity, objective error acts better than variable error. Does this mean that generally, we choose objective error in our algorithms for functions without strong convexity and objective error for functions with strong convexity? Or can we always look at the objective error for different functions?
(We can look at both and The truth: we look at what's easiest to prove :))
- If we don't have L-smooth and strong convex in the whole domain, but within some subset of the domain, and the initial point is close to the optimal, can we still get linear convergence? It seems to me that as long as the quadratic approximation is relatively accurate, we can achieve similar result.
(Yes, local strong convexity. There are also other conditions: regularity condition, Polyak-Lojasiewicz (PL) condition)

Recap

Slow convergence

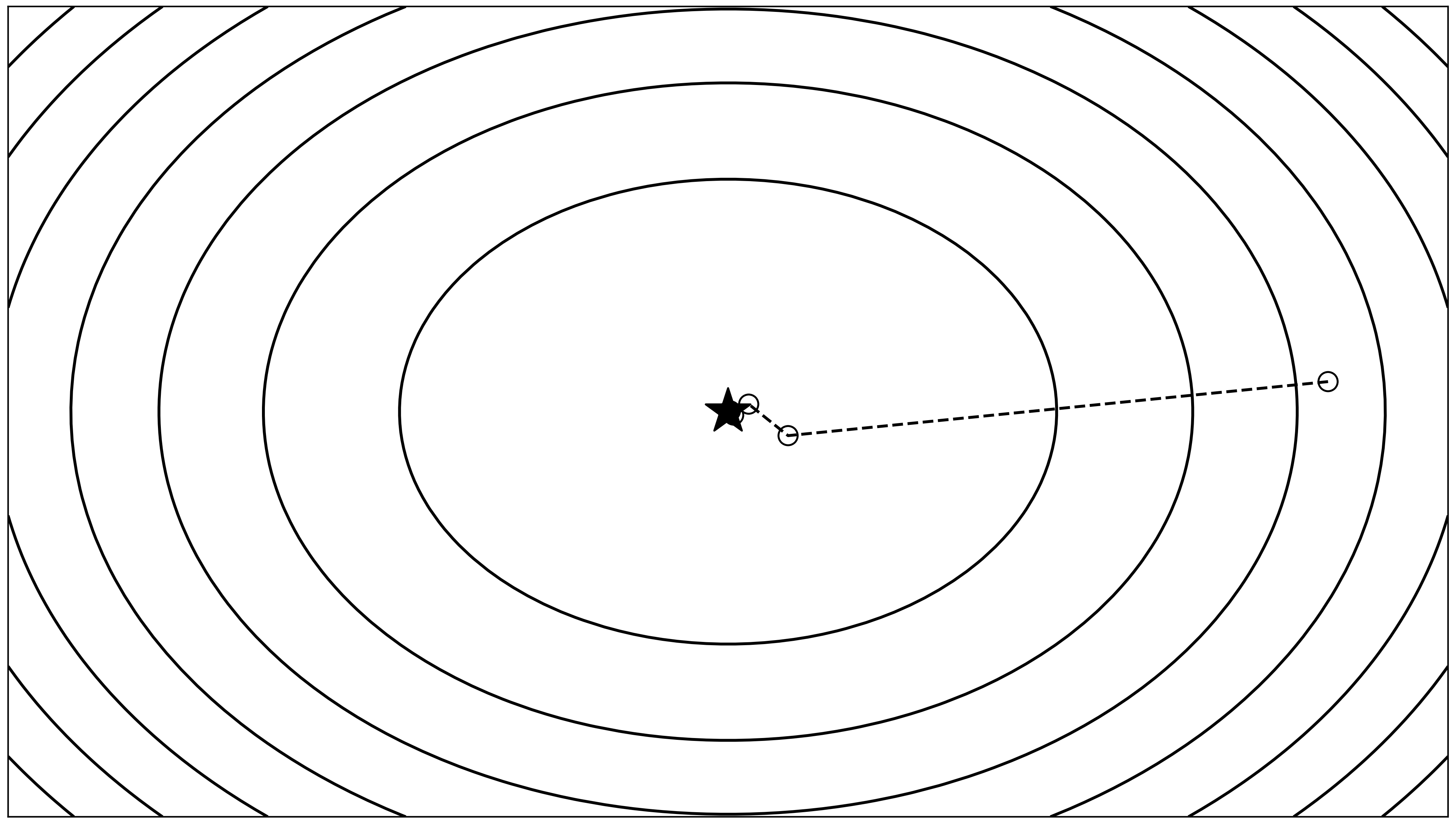
Very dependent on scaling

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



Slow convergence

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

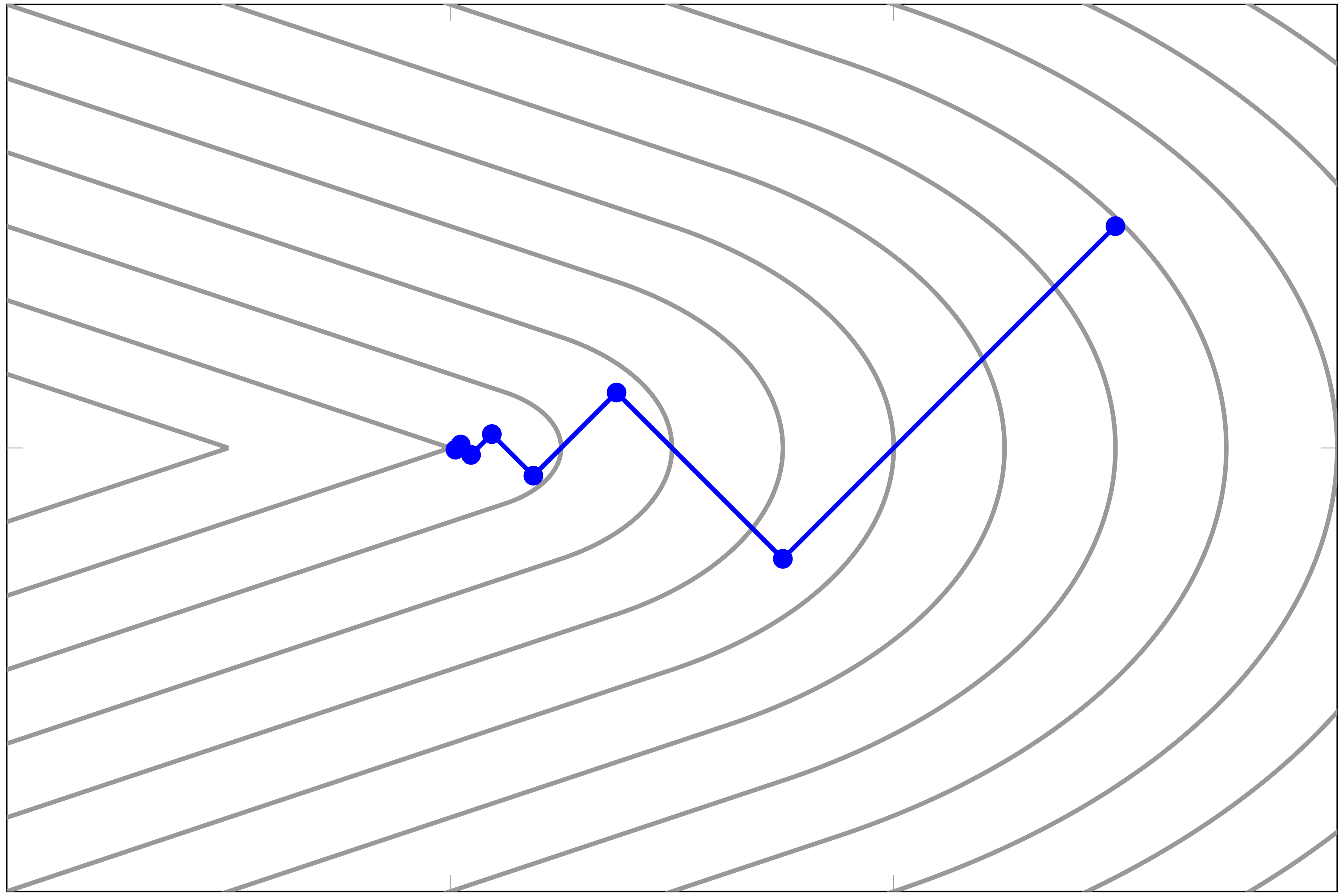


Faster

Non-differentiability

Wolfe's example

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



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

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

Today's lecture

[Chapter 3 and 8, Beck]
[ee364b Lecture notes, Boyd]
[Chapter 3, Lectures on Convex Optimization, Nesterov]

Subgradient methods

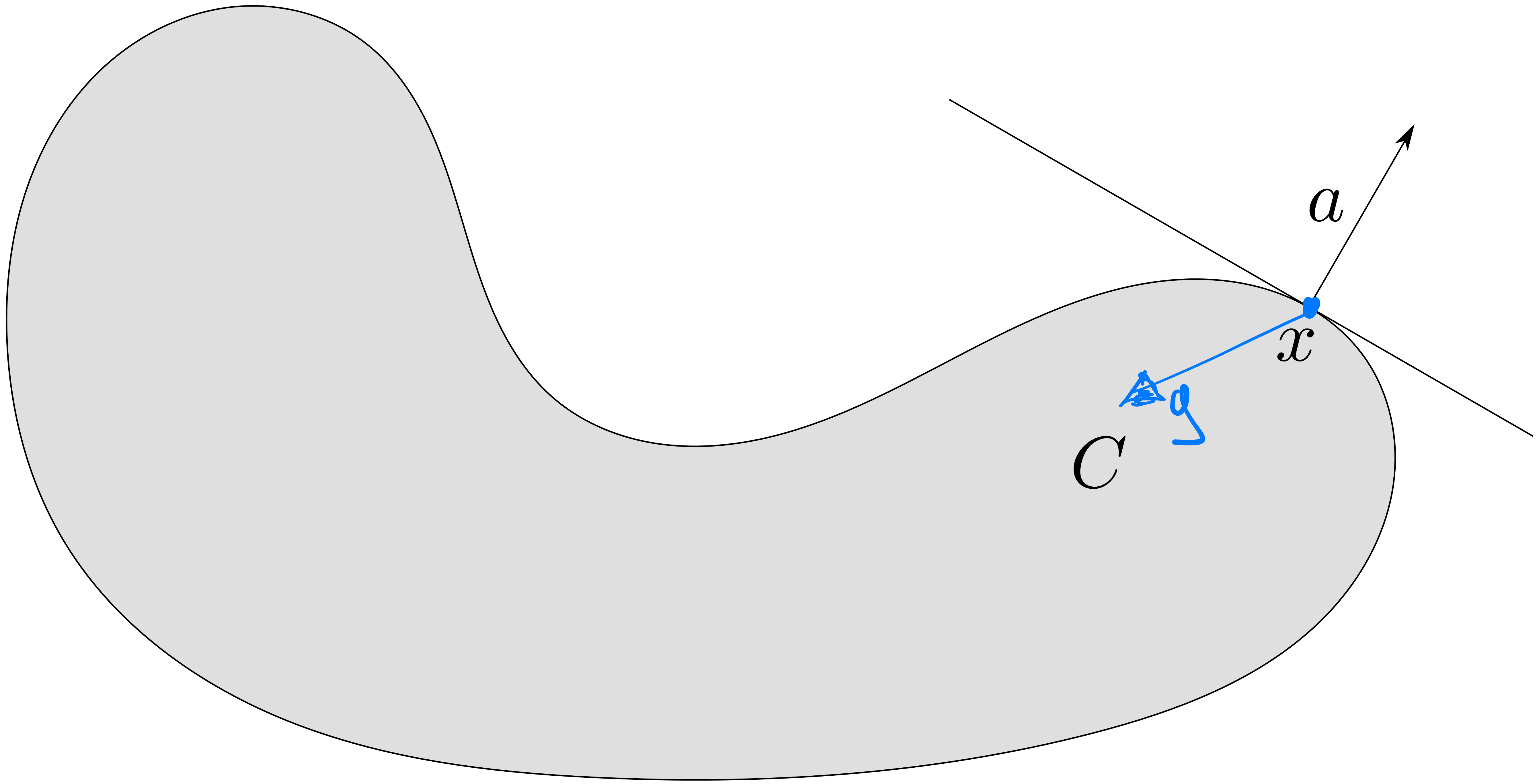
- Geometric definitions
- Subgradients
- Subgradient calculus
- Optimality conditions based on subgradients
- Subgradient methods

Geometric definitions

Supporting hyperplanes

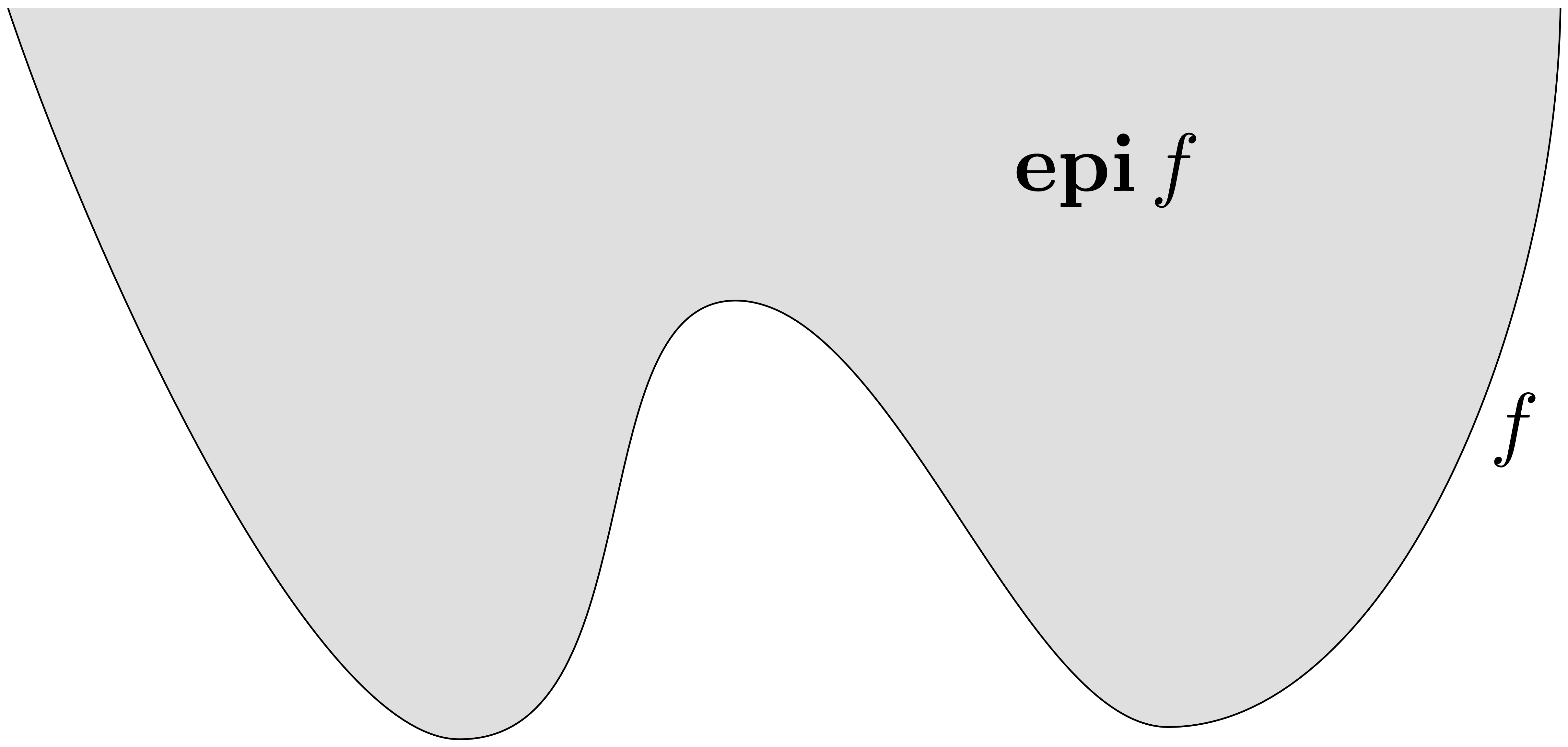
Given a set C point x at the boundary of C
a hyperplane $\{z \mid a^T z = a^T x\}$ is a **supporting hyperplane** if

$$a^T(y - x) \leq 0, \quad \forall y \in C$$



Function epigraph

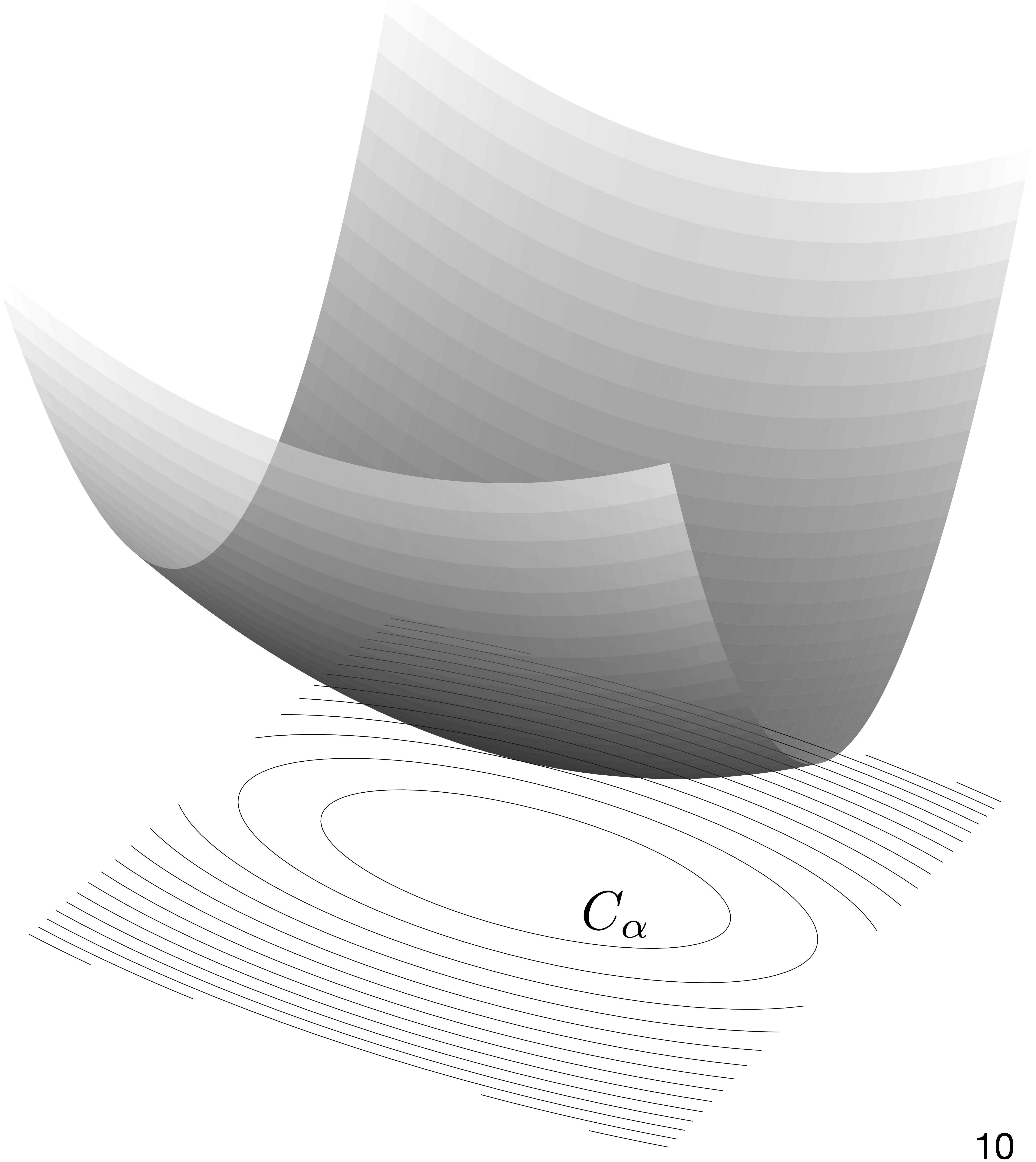
$$\mathbf{epi}\, f = \{(x,t) \mid x \in \mathbf{dom}\, f,\ f(x) \le t\}$$



f is convex if and only if [redacted] $\mathbf{epi}\, f$ is a convex set

Sublevel sets

$$C_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \le \alpha\}$$



If f is convex, then C_α is convex $\forall \alpha$

Note converse not true, e.g., $f(x) = -e^x$

Subgradients

Gradients and epigraphs

For a convex differentiable function f, i.e.

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

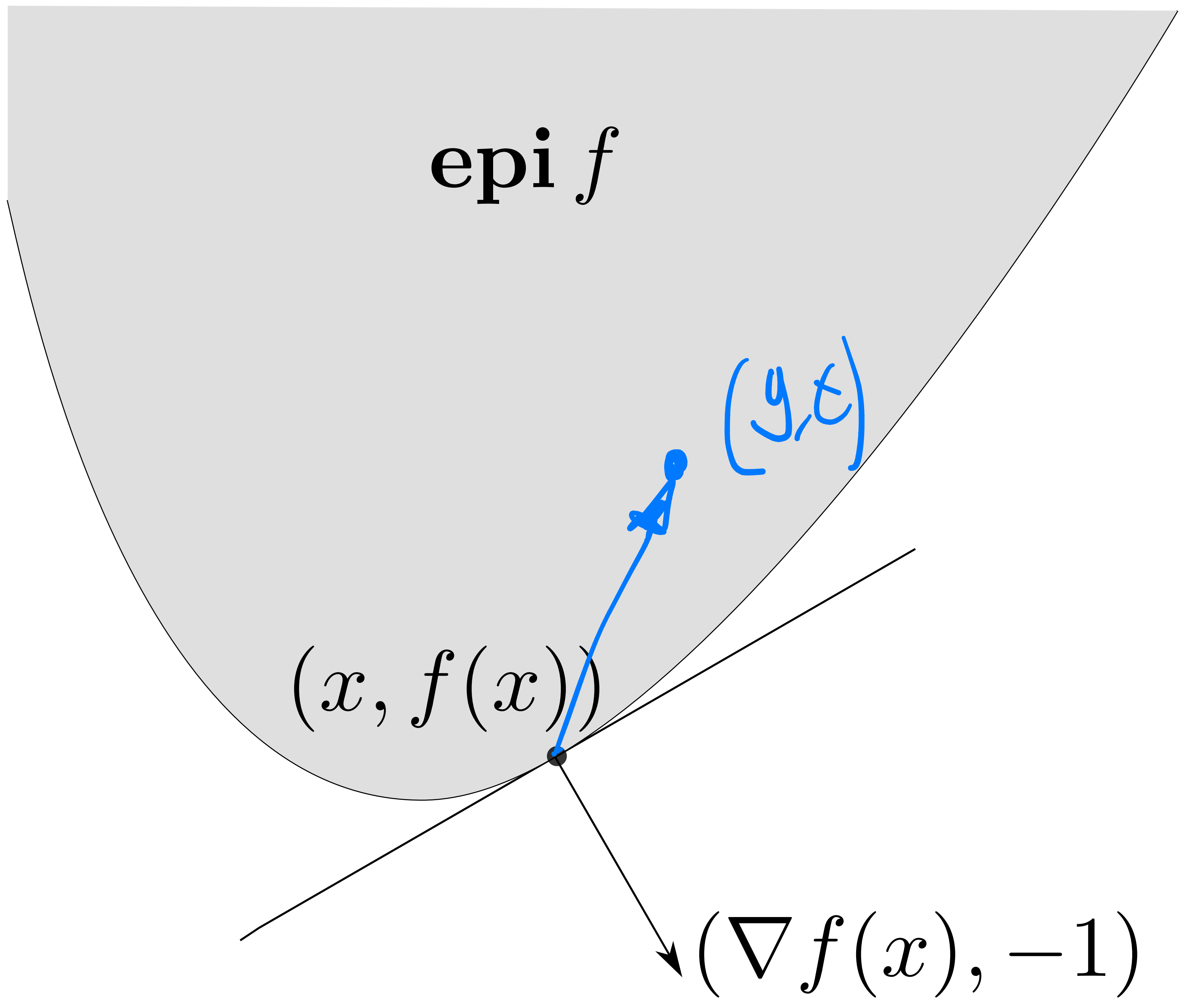
Gradients and epigraphs

For a convex differentiable function f, i.e.

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

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

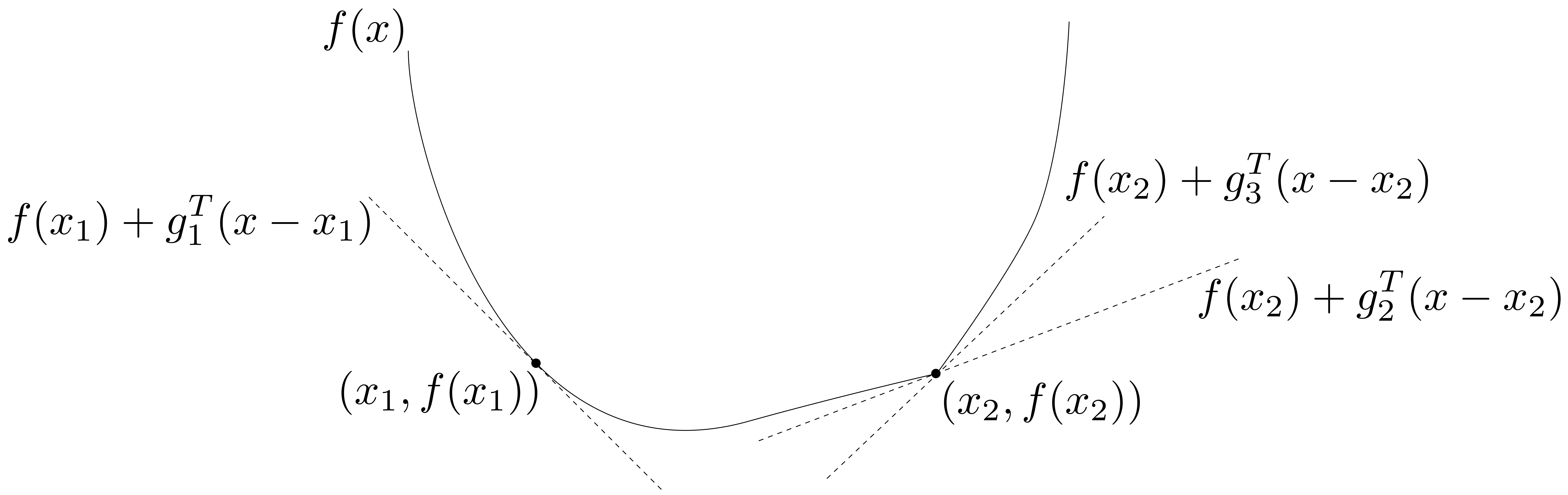
$$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0, \quad \forall (y, t) \in \textbf{epi}\, f$$



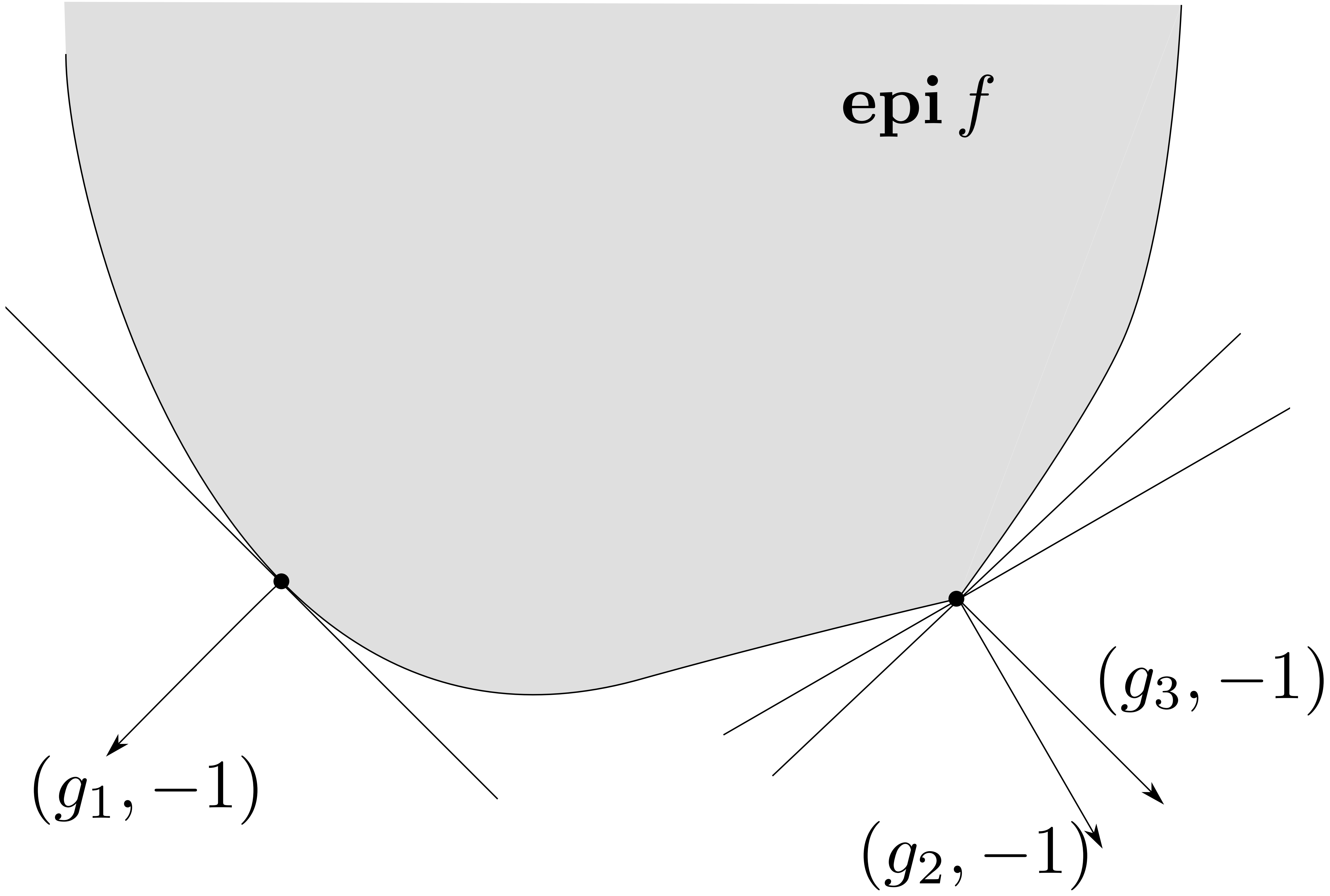
Subgradient

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

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

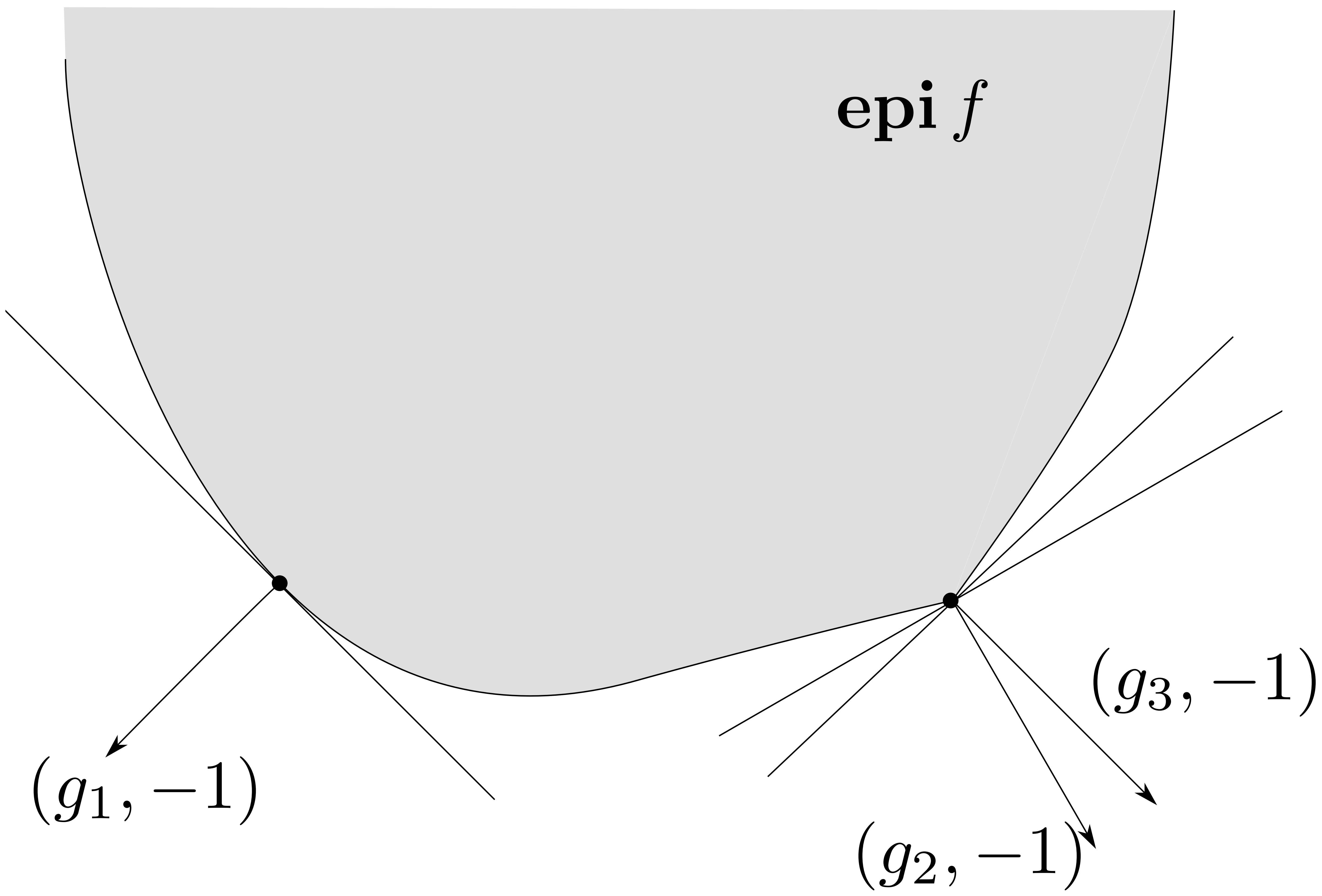


Subgradient properties



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

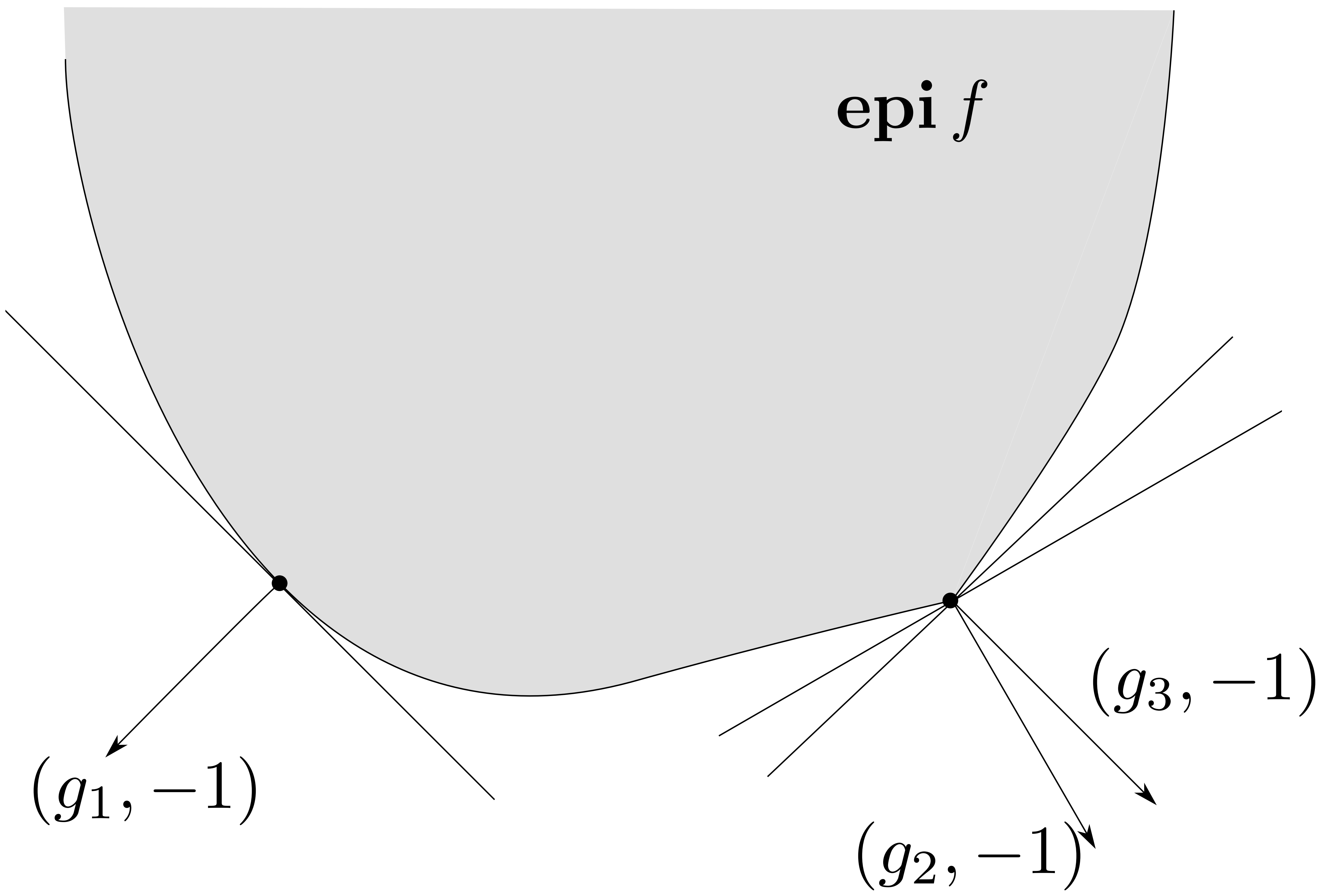
Subgradient properties



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

Subgradient properties



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

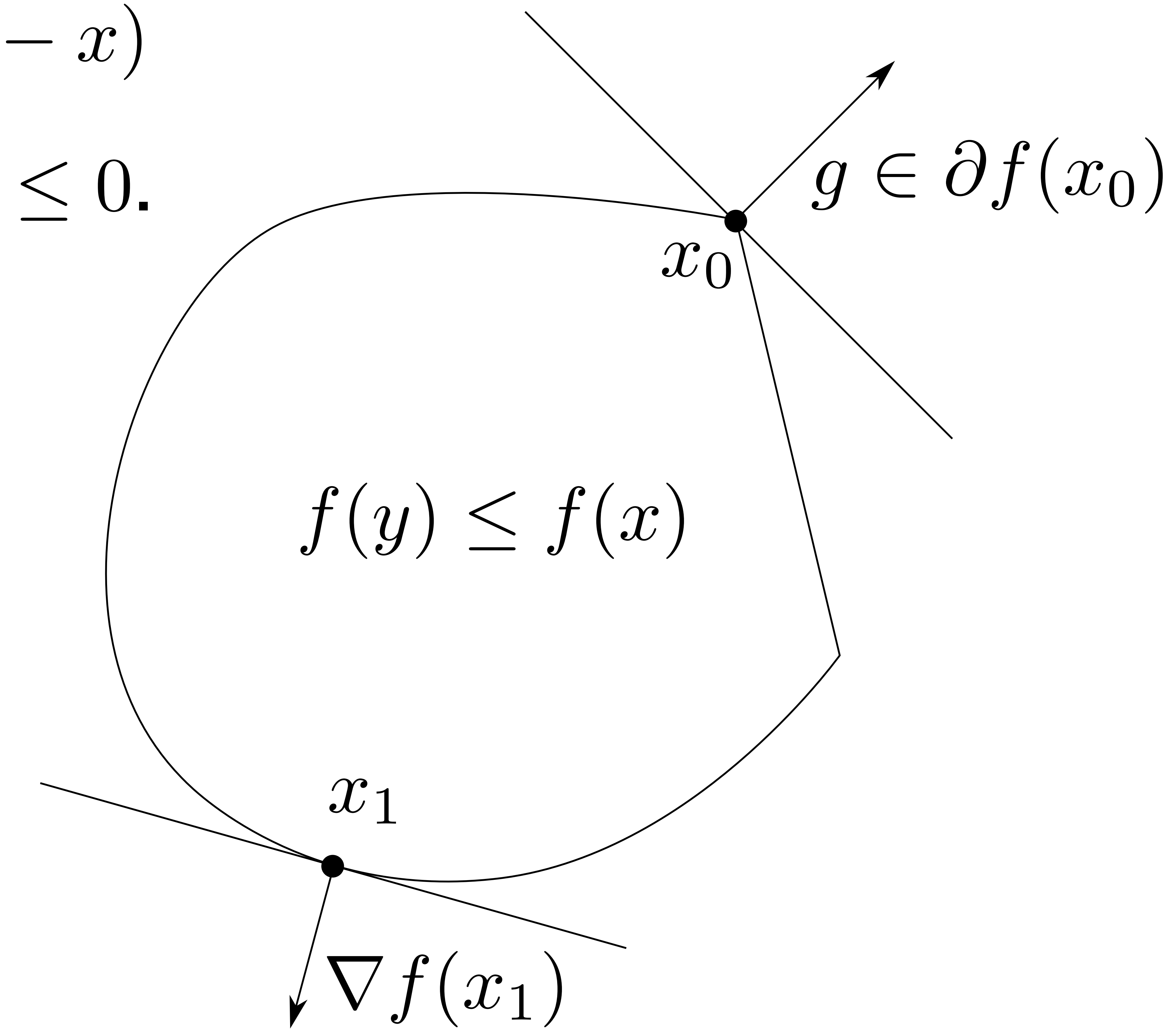
g is a subgradient of f iff $f(x) + g^T(y - x)$ is a global underestimator of f

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

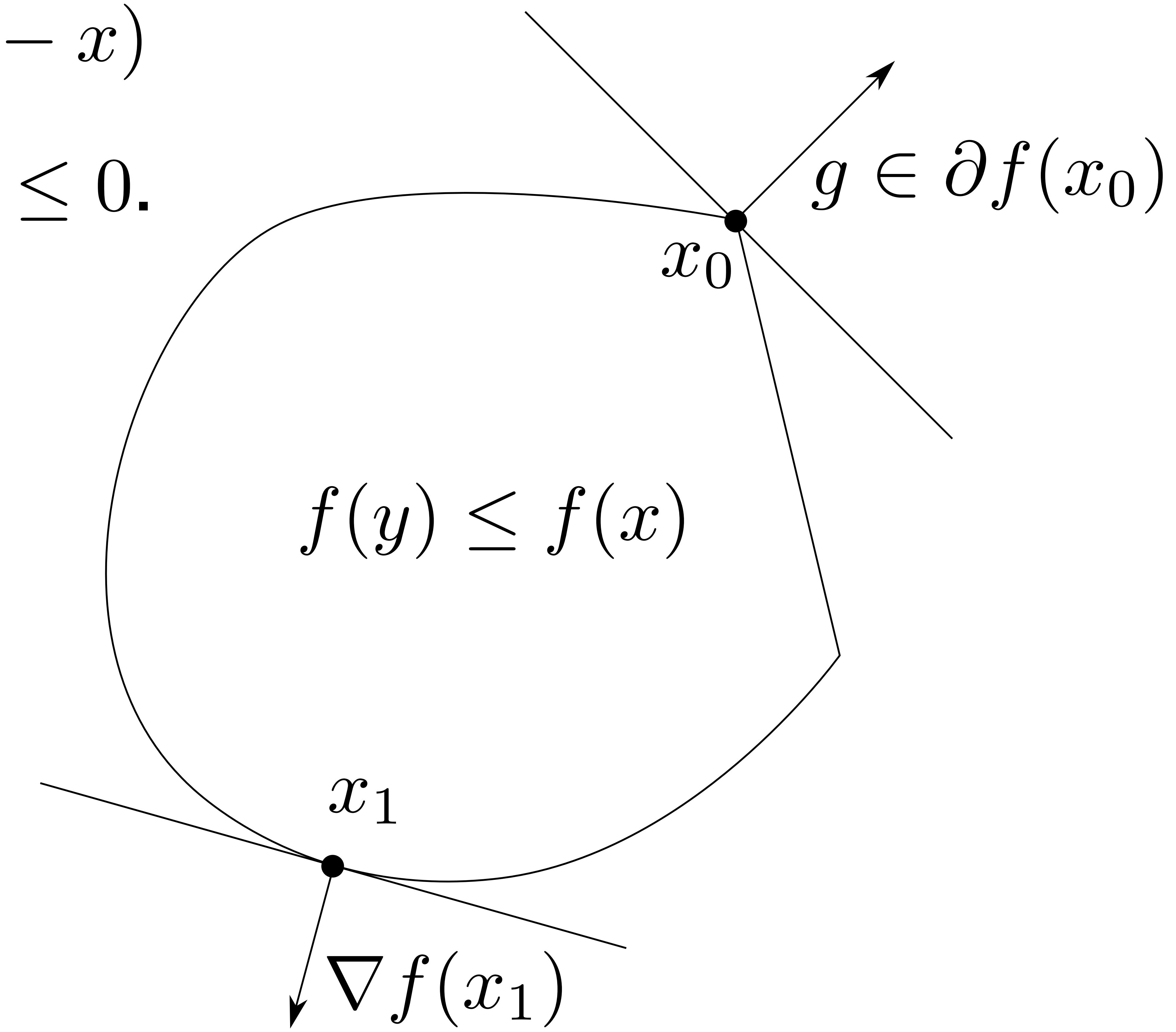
Therefore, if $f(y) \leq f(x)$ (sublevel set), then $g^T(y - x) \leq 0$.



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

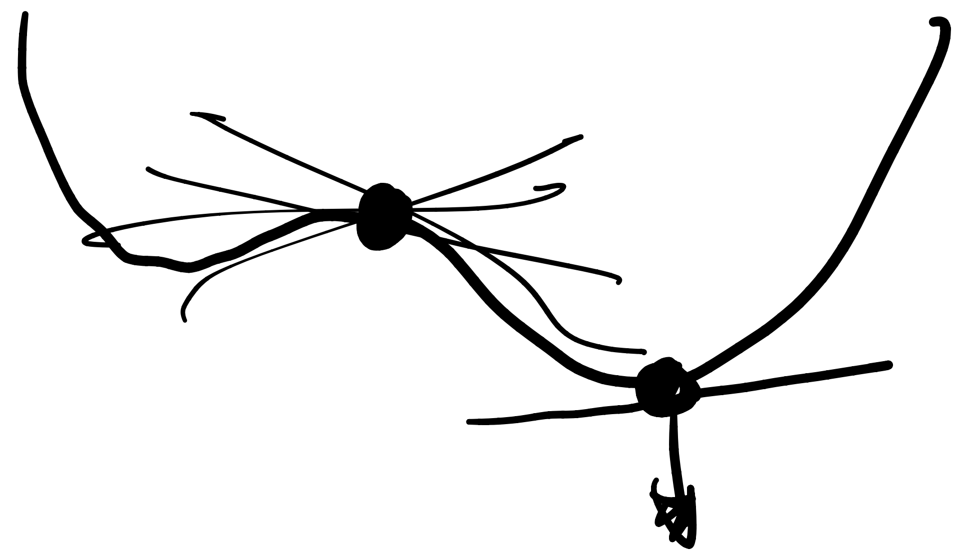
f nondifferentiable at x
subgradients define supporting hyperplane to sublevel set throgh x

Subdifferential

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

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

Subdifferential



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

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

Properties

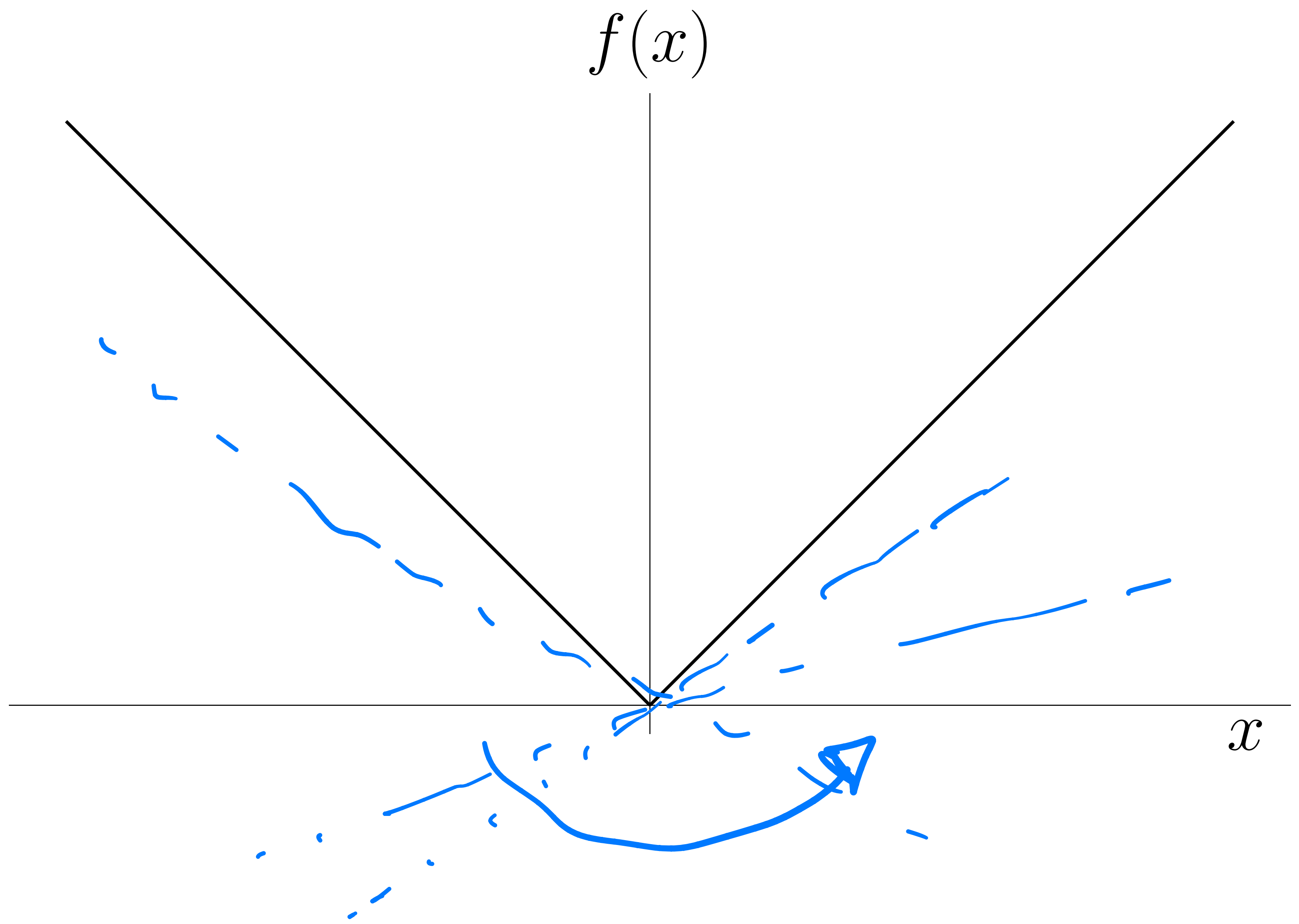
- $\partial f(x)$ is always closed and convex, also for nonconvex f. (intersection of halfspaces)
- If $\partial f(x) \neq \emptyset$ then f is convex (converse not true)
- If f is convex and differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- If f is convex and $\partial f(x) = \{g\}$, then f is differentiable at x and $g = \nabla f(x)$

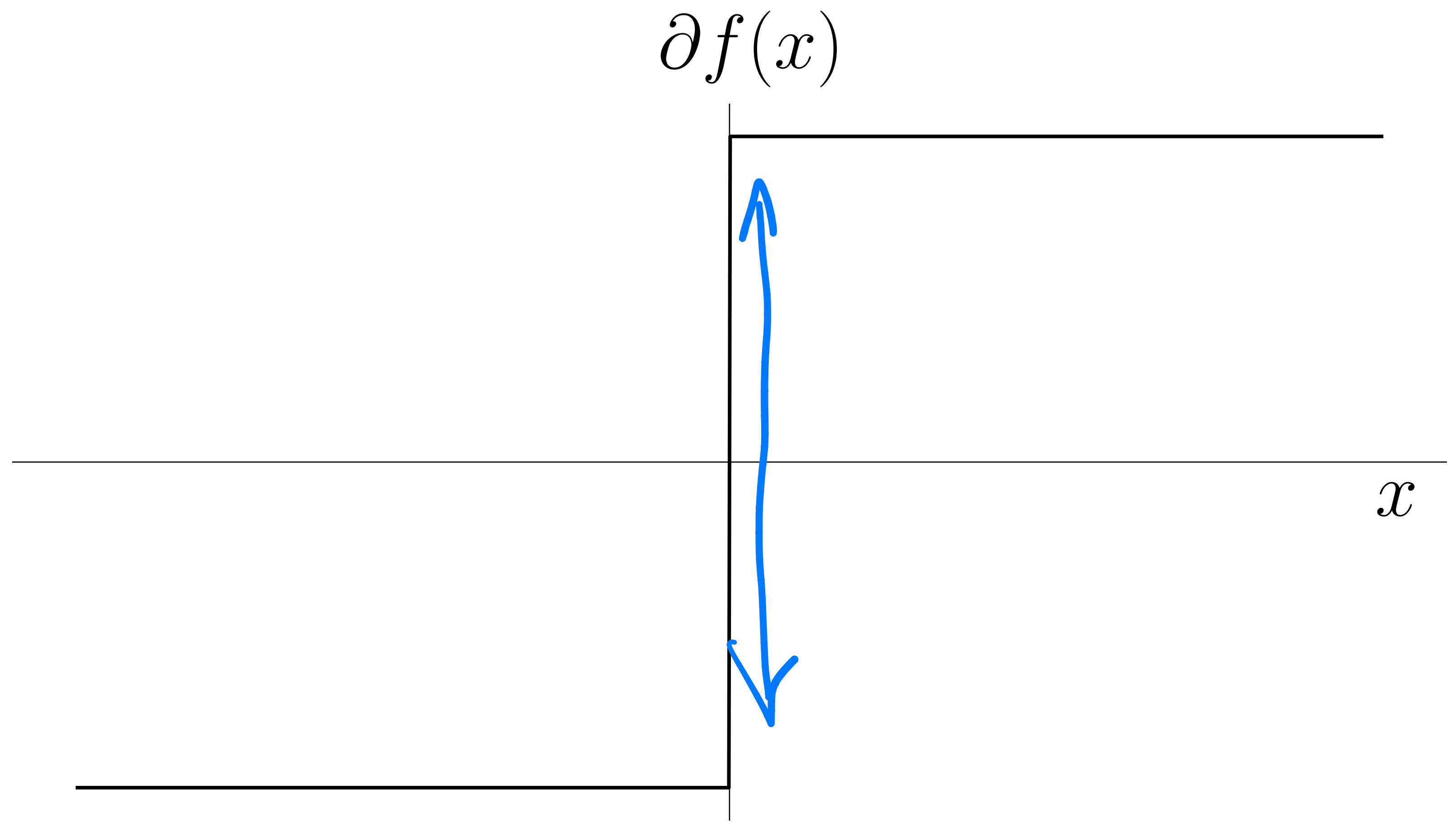
Example

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 $\partial f(x)$

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

Subgradient calculus

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

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

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

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

Basic rules

Pointwise maxima

Finite pointwise maximum $f(x) = \max_{i=1,\dots,m} f(x)$, 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 $f(x) = \max_{s \in S} f_s(x)$, then (SUPREMUM)

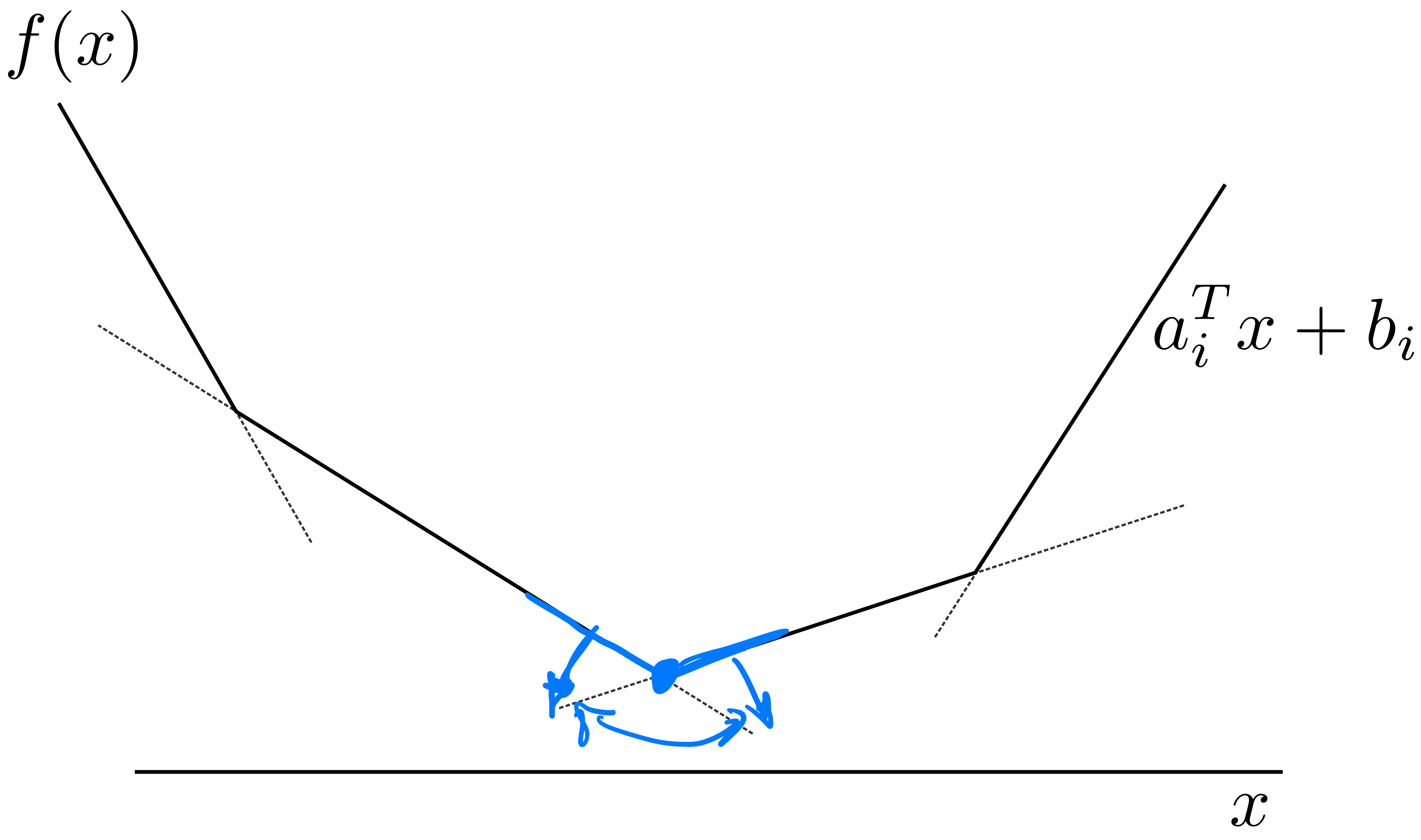
$$\partial f(x) = \mathbf{cl}\left(\mathbf{conv}\left(\bigcup \{\partial f_s(x) \mid f_s(x) = f(x)\}\right)\right)$$ (closure of the hull)

Note: Equality requires some regularity assumptions (otherwise $\supseteq$)
(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

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

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

DUAL NORMS

$1 \to \infty$

$2 \to 2$

$p \quad q$

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,

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

Example

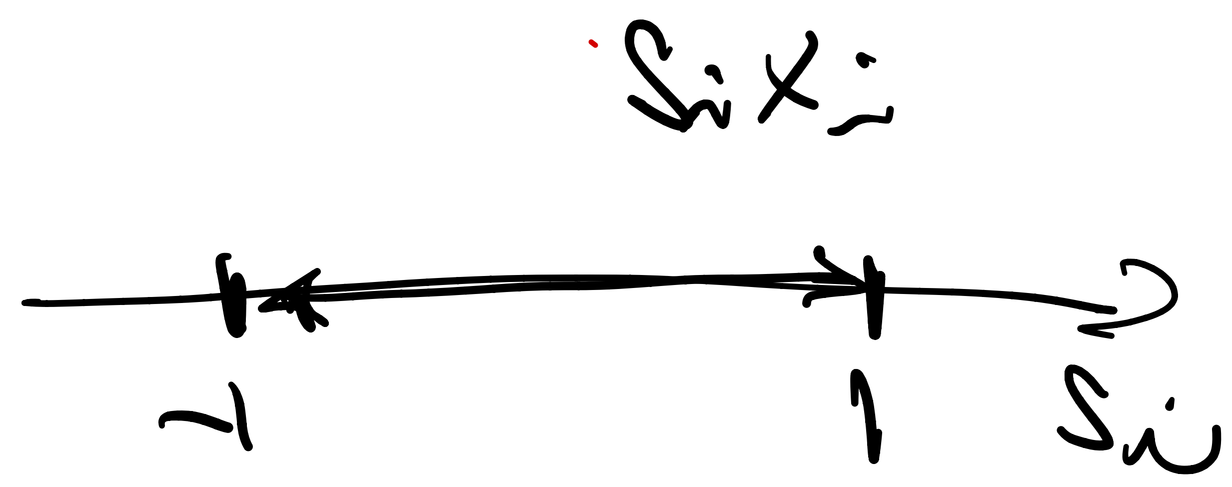
Norms

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

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

Weak result alternative

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

↳ COMPONENTWISE

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 $g_i \in \partial f_i(x)$

Basic rules

Composition

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

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

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

Proof

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

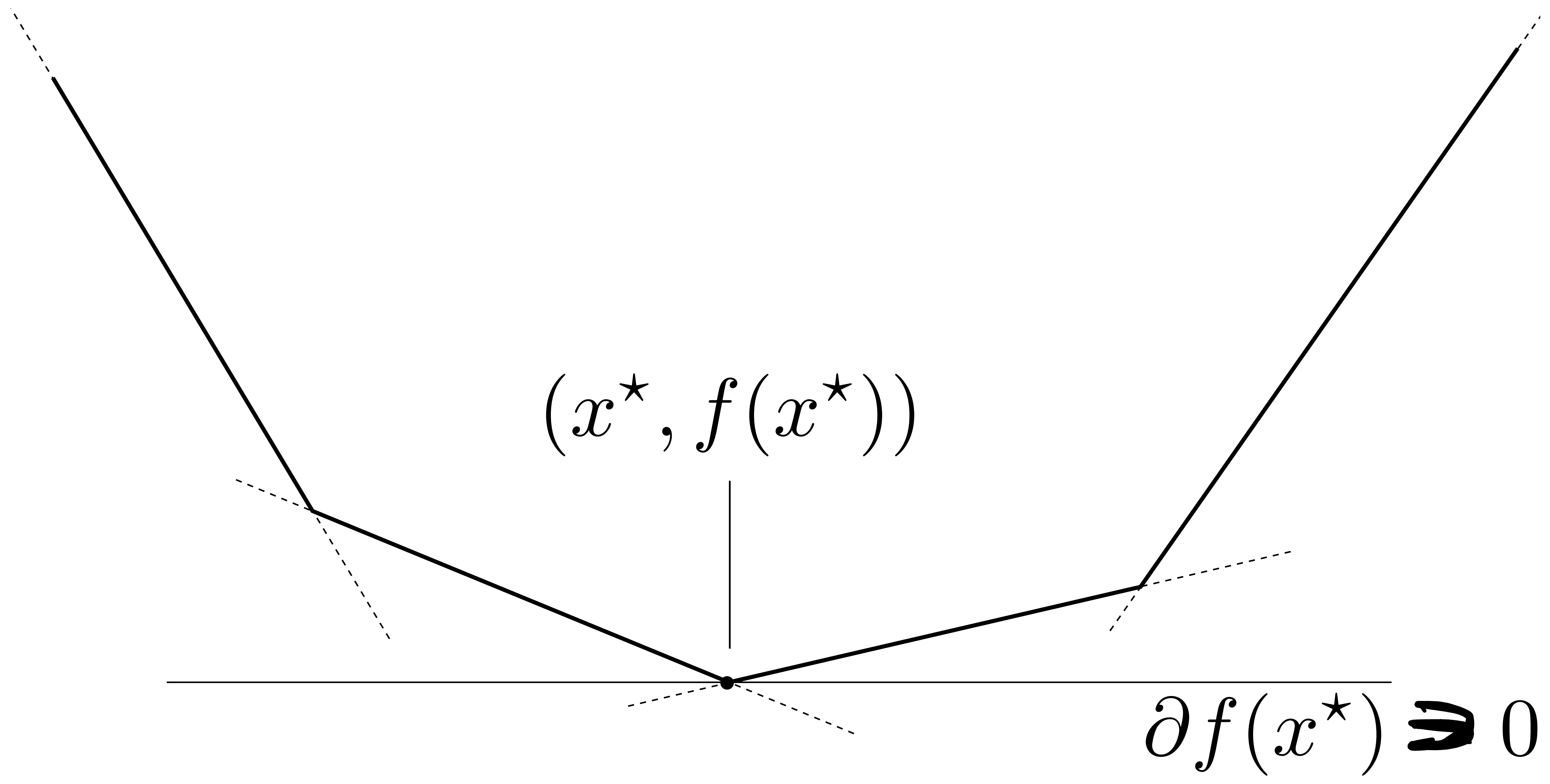
■

Optimality conditions

Fermat's optimality condition

For any convex f, $x^\star$ is a local minimizer if and only if

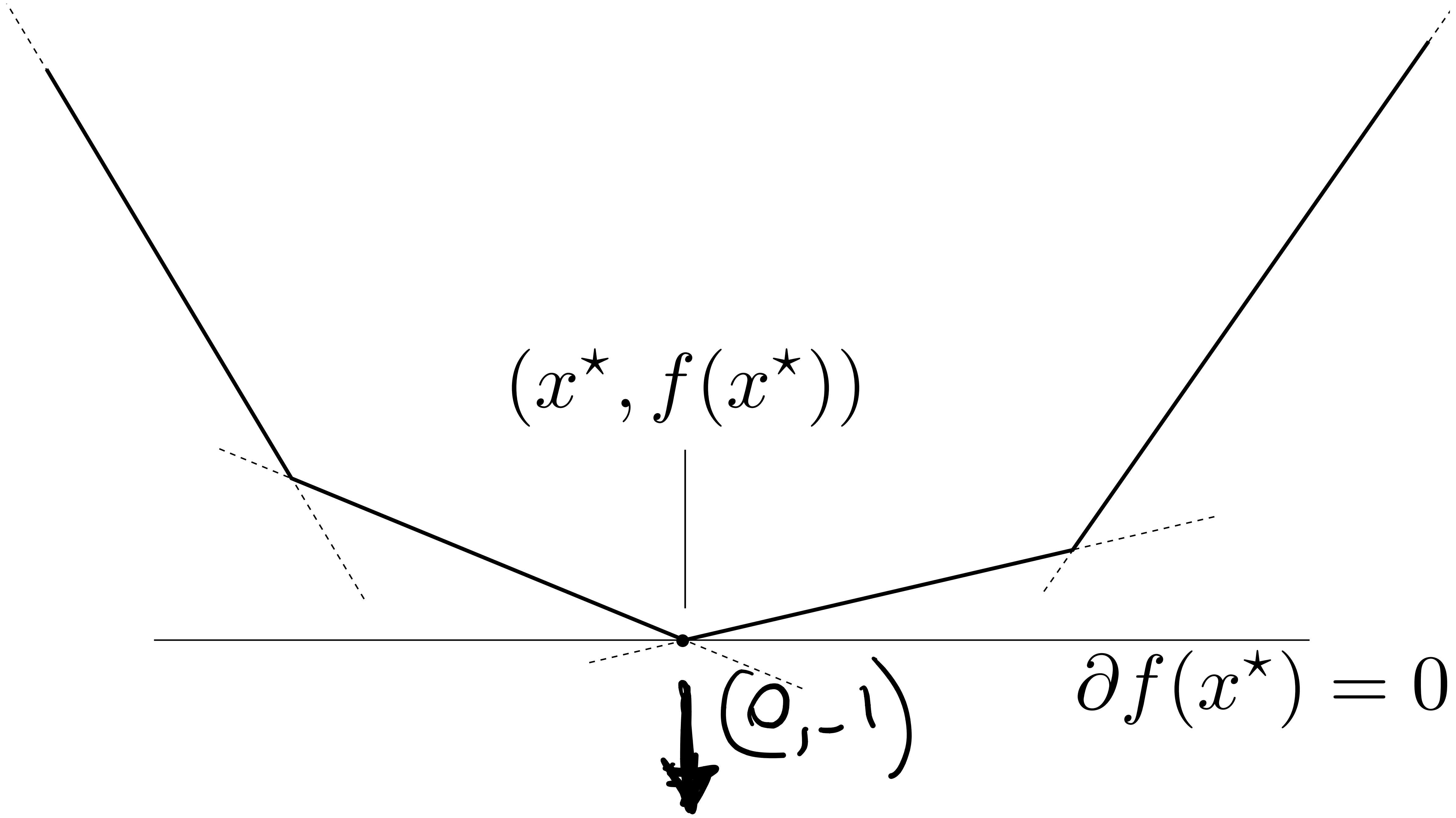
$$0 \in \partial f(x^\star)$$



Fermat's optimality condition

For any convex f, $x^\star$ is a local 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

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

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

$$\lambda \geq 0, \quad \mathbf{1}^T \lambda = 1, \quad \sum_{i=1}^{m} \lambda_i a_i = 0$$

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

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

$$\lambda \geq 0, \quad \mathbf{1}^T \lambda = 1, \quad \sum_{i=1}^{m} \lambda_i a_i = 0$$

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

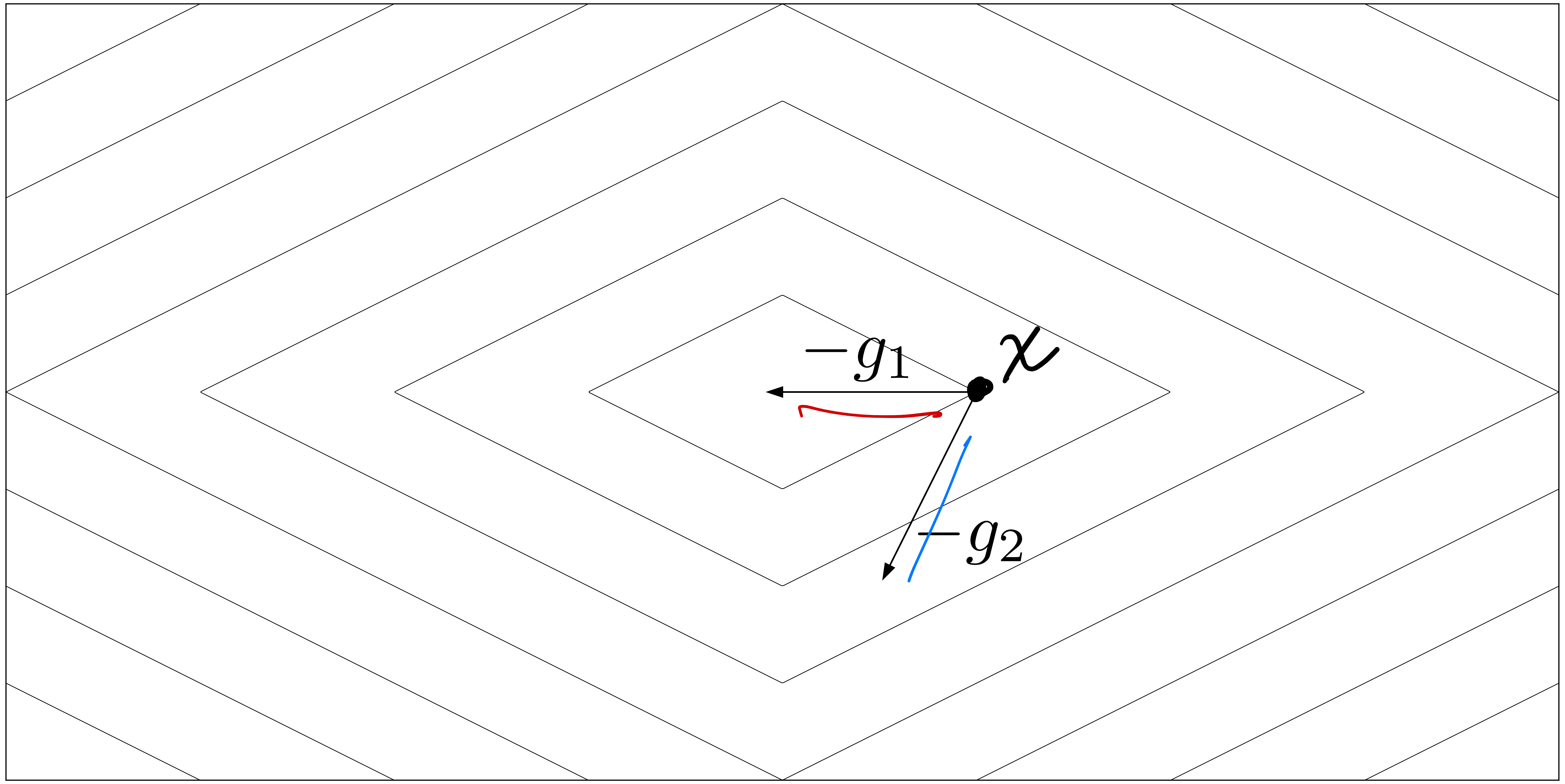
maximize $b^T \lambda$
subject to $A^T \lambda = 0$
$\lambda \geq 0, \quad \mathbf{1}^T \lambda = 1$

Subgradient method

Negative subgradients are not necessarily descent directions

$$f(x) = |x_1| + 2|x_2|$$

$$x = (1, 0)$$



$g_1 = (1, 0) \in \partial f(x)$ and
$-g_1$ is a descent direction

$g_2 = (1, 2) \in \partial f(x)$ and
$-g_2$ is not a descent direction

Subgradient method

Convex optimization problem

$$\text{minimize} \quad f(x) \qquad \text{(optimal cost } f^\star\text{)}$$

Subgradient method

Convex optimization problem

minimize $f(x)$ (optimal cost $f^\star$)

Iterations

$$x^{k+1} = x^k - t_k g^k, \qquad g^k \in \partial f(x^k)$$

g^k is **any subgradient** of f at x^k

Subgradient method

Convex optimization problem

minimize $\quad f(x) \qquad$ (optimal cost $f^\star$)

Iterations

$$x^{k+1} = x^k - t_k g^k, \qquad g^k \in \partial f(x^k)$$

g^k is **any subgradient** of f at x^k

Not a descent method, keep track of the best point

$$f_{\rm best}^k = \min_{i=1,\dots,k} f(x^i)$$

Step sizes

Line search can lead to **suboptimal points**

Step sizes ***pre-specified***, not adaptively computed
(different than gradient descent)

Step sizes

Line search can lead to **suboptimal points**

Step sizes ***pre-specified***, not adaptively computed
(different than gradient descent)

Fixed: $t_k = t$ for $k = 0, \ldots$

Diminishing: $$\sum_{k=0}^{\infty} t_k^2 < \infty, \quad \sum_{k=0}^{\infty} t_k = \infty$$

Square summable but not summable
(goes to 0 but not too fast)

e.g., $t_k = O(1/k)$

Convergence

Assumptions

- f is convex with $\mathbf{dom}\, f = \mathbf{R}^n$
- $f(x^\star) > -\infty$ (finite optimal value)
- f is Lipschitz continuous with constant $G > 0$, i.e.

$$|f(x) - f(y)| \leq G\|x - y\|_2, \quad \forall x, y$$

which is equivalent to $\|g\|_2 \leq G, \quad \forall g \in \partial f(x)$, $\forall x$

Convergence

Lipschitz continuity equivalence

f is Lipschitz continuous with constant $G > 0$, i.e.

$$|f(x) - f(y)| \leq G\|x - y\|_2, \quad \forall x, y$$

which is equivalent to $\|g\|_2 \leq G, \quad \forall g \in \partial f(x),\ \forall x$

Convergence

Lipschitz continuity equivalence

f is Lipschitz continuous with constant $G > 0$, i.e.

$$|f(x) - f(y)| \leq G\|x - y\|_2, \quad \forall x, y$$

which is equivalent to $\|g\|_2 \leq G, \quad \forall g \in \partial f(x),\ \forall x$

Proof

If $\|g\| \leq G$ for all subgradients, pick $x, g_x \in \partial f(x)$ and $y, g_y \in \partial f(y)$. Then,

$$\begin{aligned} & g_x^T(x - y) \geq f(x) - f(y) \geq g_y^T(x - y) \\ \implies \quad & G\|x - y\|_2 \geq f(x) - f(y) \geq -G\|x - y\|_2 \end{aligned}$$

Convergence

Lipschitz continuity equivalence

f is Lipschitz continuous with constant $G > 0$, i.e.

$$|f(x) - f(y)| \leq G\|x - y\|_2, \quad \forall x, y$$

which is equivalent to $\|g\|_2 \leq G, \quad \forall g \in \partial f(x),\ \forall x$

Proof

If $\|g\| \leq G$ for all subgradients, pick $x, g_x \in \partial f(x)$ and $y, g_y \in \partial f(y)$. Then,

$$g_x^T(x - y) \geq f(x) - f(y) \geq g_y^T(x - y)$$
$$\implies \quad G\|x - y\|_2 \geq f(x) - f(y) \geq -G\|x - y\|_2$$

If $\|g\|_2 > G$ for some $g \in \partial f(x)$. Take $y = x + g/\|g\|_2$ such that $\|x - y\|_2 = 1$:

$$f(y) \geq f(x) + g^T(y - x) = f(x) + \|g\|_2 > f(x) + G$$

■

Convergence

Theorem

Given a convex, G-Lipschitz continuous f with finite optimal value, the subgradient method obeys

$$f_{\text{best}}^k - f^\star \leq \frac{R^2 + G^2 \sum_{i=0}^k t_i^2}{2 \sum_{i=0}^k t_i}$$

where $\|x^0 - x^\star\|_2 \leq R$

Convergence

Proof

Key quantity: euclidean distance to optimal set (not function value)

$$
\begin{aligned}
\|x^{k+1} - x^\star\|_2^2 &= \|x^k - t_k g^k - x^\star\|_2^2 \\
&= \|x^k - x^\star\|_2^2 - 2t_k (g^k)^T (x^k - x^\star) + t_k^2 \|g^k\|_2^2 \\
&\leq \|x^k - x^\star\|_2^2 - 2t_k (f(x^k) - f^\star) + t_k^2 \|g^k\|_2^2
\end{aligned}
$$

using $f^\star = f(x^\star) \geq f(x^k) + (g^k)^T (x^\star - x^k)$

Convergence

Proof (continued)

Apply inequality recursively, obtaining

$$\begin{aligned}\|x^{k+1} - x^\star\|_2^2 &\leq \|x^0 - x^\star\|_2^2 - 2\sum_{i=0}^{k} t_i(f(x^i) - f^\star) + \sum_{i=0}^{k} t_i^2 \|g^i\|_2^2 \\ &\leq R^2 - 2\sum_{i=0}^{k} t_i(f(x^i) - f^\star) + G^2 \sum_{i=0}^{k} t_i^2\end{aligned}$$

Convergence

Proof (continued)

Apply inequality recursively, obtaining

$$\|x^{k+1} - x^\star\|_2^2 \leq \|x^0 - x^\star\|_2^2 - 2\sum_{i=0}^{k} t_i(f(x^i) - f^\star) + \sum_{i=0}^{k} t_i^2 \|g^i\|_2^2$$

$$\leq R^2 - 2\sum_{i=0}^{k} t_i(f(x^i) - f^\star) + G^2 \sum_{i=0}^{k} t_i^2$$

Using $\|x^{k+1} - x^\star\|_2^2 \geq 0$ we get

$$2\sum_{i=0}^{k} t_i(f(x^i) - f^\star) \leq R^2 + G^2 \sum_{i=0}^{k} t_i^2$$

Convergence

Proof (continued)

$$2\sum_{i=0}^{k} t_i(f(x^i) - f^\star) \leq R^2 + G^2 \sum_{i=0}^{k} t_i^2$$

Convergence

Proof (continued)

$$2\sum_{i=0}^{k} t_i(f(x^i) - f^\star) \leq R^2 + G^2 \sum_{i=0}^{k} t_i^2$$

Combine it with

$$\sum_{i=0}^{k} t_i(f(x^i) - f(x^\star)) \geq \left(\sum_{i=0}^{k} t_i\right) \min_{i=0,\ldots,k} (f(x^i) - f^\star) = \left(\sum_{i=0}^{k} t_i\right) (f_{\text{best}}^k - f^\star)$$

Convergence

Proof (continued)

$$2\sum_{i=0}^{k} t_i(f(x^i) - f^\star) \le R^2 + G^2 \sum_{i=0}^{k} t_i^2$$

Combine it with

$$\sum_{i=0}^{k} t_i(f(x^i) - f(x^\star)) \ge \left(\sum_{i=0}^{k} t_i\right) \min_{i=0,\ldots,k} (f(x^i) - f^\star) = \left(\sum_{i=0}^{k} t_i\right)(f_{\text{best}}^k - f^\star)$$

to get

$$f_{\text{best}}^k - f^\star \le \frac{R^2 + G^2 \sum_{i=0}^{k} t_i^2}{2\sum_{i=0}^{k} t_i}$$

■

Implications for step size rules

$$f^k_{\text{best}} - f^\star \leq \frac{R^2 + G^2 \sum_{i=0}^k t_i^2}{2 \sum_{i=0}^k t_i}$$

Implications for step size rules

$$f^k_{\text{best}} - f^\star \le \frac{R^2 + G^2 \sum_{i=0}^k t_i^2}{2 \sum_{i=0}^k t_i}$$

Fixed: $t_k = t$ for $k = 0, \ldots$

$$f^k_{\text{best}} - f^\star \le \frac{R^2 + G^2(k+1)t^2}{2(k+1)t}$$

May be suboptimal

$$\lim_{k \to \infty} f^k_{\text{best}} \le f^\star + \frac{G^2 t}{2}$$

Implications for step size rules

$$f_{\text{best}}^k - f^\star \le \frac{R^2 + G^2 \sum_{i=0}^k t_i^2}{2\sum_{i=0}^k t_i}$$

Fixed: $t_k = t$ for $k = 0, \ldots$

$$f_{\text{best}}^k - f^\star \le \frac{R^2 + G^2(k+1)t^2}{2(k+1)t}$$

May be suboptimal

$$\lim_{k\to\infty} f_{\text{best}}^k \le f^\star + \frac{G^2 t}{2}$$

Diminishing: $$\sum_{k=0}^{\infty} t_k^2 < \infty, \quad \sum_{k=0}^{\infty} t_k = \infty$$

e.g., $t_k = \tau/(k+1)$ or $t_k = \tau/\sqrt{k+1}$

Optimal

$$\lim_{k\to\infty} f_{\text{best}}^k = f^\star$$

Optimal step size and convergence rate

For a tolerance $\epsilon > 0$, let's find the optimal t_k for a fixed k:

$$\frac{R^2 + G^2 \sum_{i=0}^{k} t_i^2}{2 \sum_{i=0}^{k} t_i} \leq \epsilon$$

Optimal step size and convergence rate

For a tolerance $\epsilon > 0$, let's find the optimal t_k for a fixed k:

$$\frac{R^2 + G^2 \sum_{i=0}^{k} t_i^2}{2 \sum_{i=0}^{k} t_i} \leq \epsilon$$

Convex and symmetric in $(t_0, \ldots, t_k)$
Hence, minimum when $t_i = t$

$$\longrightarrow \quad \frac{R^2 + G^2(k+1)t^2}{2(k+1)t}$$

Optimal choice $\quad t = \dfrac{R}{G\sqrt{k+1}}$

Optimal step size and convergence rate

For a tolerance $\epsilon > 0$, let's find the optimal t_k for a fixed k:

$$\frac{R^2 + G^2 \sum_{i=0}^k t_i^2}{2 \sum_{i=0}^k t_i} \leq \epsilon$$

Convex and symmetric in $(t_0, \dots, t_k)$
Hence, minimum when $t_i = t$

$$\longrightarrow \quad \frac{R^2 + G^2 (k+1) t^2}{2(k+1)t}$$

Optimal choice $\quad t = \frac{R}{G\sqrt{k+1}}$

Convergence rate

$$f_{\text{best}}^k - f^\star \leq \frac{RG}{\sqrt{k+1}}$$

Iterations required

$$k = O(1/\epsilon^2)$$

(gradient descent $k = O(1/\epsilon)$)

Stopping criterion

Terminating when

$$\frac{R^2 + G^2 \sum_{i=0}^{k} t_i^2}{2 \sum_{i=0}^{k} t_i} \leq \epsilon$$

is really, really slow.

Bad news

There is not really a good stopping criterion for the subgradient method

Optimal step size when $f^\star$ is known

Polyak step size

$$t_k = \frac{f(x^k) - f^\star}{\|g^k\|_2^2}$$

Optimal step size when $f^\star$ is known

Polyak step size

$$t_k = \frac{f(x^k) - f^\star}{\|g^k\|_2^2}$$

Motivation: minimize righthand side of

$$\|x^{k+1} - x^\star\|_2^2 \leq \|x^k - x^\star\|_2^2 - 2t_k(f(x^k) - f^\star) + t_k^2\|g^k\|_2^2$$

Obtaining $\quad (f(x^k) - f^\star)^2 \leq \left(\|x^{k+1} - x^\star\|_2^2 - \|x^k - x^\star\|_2^2\right) G^2$

Optimal step size when $f^\star$ is known

Polyak step size

$$t_k = \frac{f(x^k) - f^\star}{\|g^k\|_2^2}$$

Motivation: minimize righthand side of

$$\|x^{k+1} - x^\star\|_2^2 \leq \|x^k - x^\star\|_2^2 - 2t_k(f(x^k) - f^\star) + t_k^2\|g^k\|_2^2$$

Obtaining $\quad (f(x^k) - f^\star)^2 \leq \left(\|x^{k+1} - x^\star\|_2^2 - \|x^k - x^\star\|_2^2\right) G^2$

Applying recursively, $\quad f_{\text{best}}^k - f^\star \leq \dfrac{GR}{\sqrt{k+1}}$

Optimal step size when $f^\star$ is known

Polyak step size

$$t_k = \frac{f(x^k) - f^\star}{\|g^k\|_2^2}$$

Motivation: minimize righthand side of

$$\|x^{k+1} - x^\star\|_2^2 \leq \|x^k - x^\star\|_2^2 - 2t_k(f(x^k) - f^\star) + t_k^2\|g^k\|_2^2$$

Obtaining $\quad (f(x^k) - f^\star)^2 \leq \left(\|x^{k+1} - x^\star\|_2^2 - \|x^k - x^\star\|_2^2\right) G^2$

Applying recursively, $\quad f_{\text{best}}^k - f^\star \leq \dfrac{GR}{\sqrt{k+1}}$

Iterations required

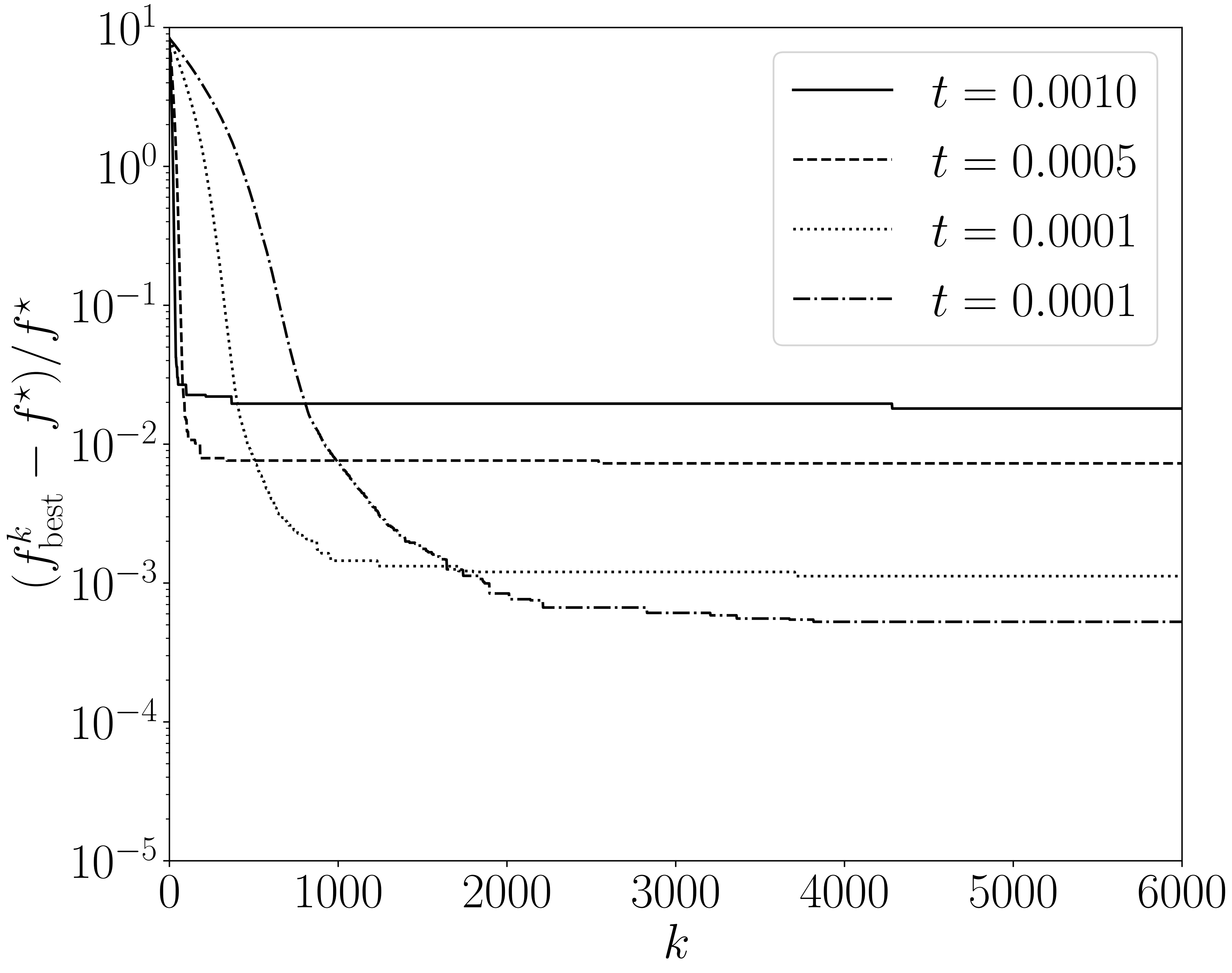
$$k = O(1/\epsilon^2)$$

still not great

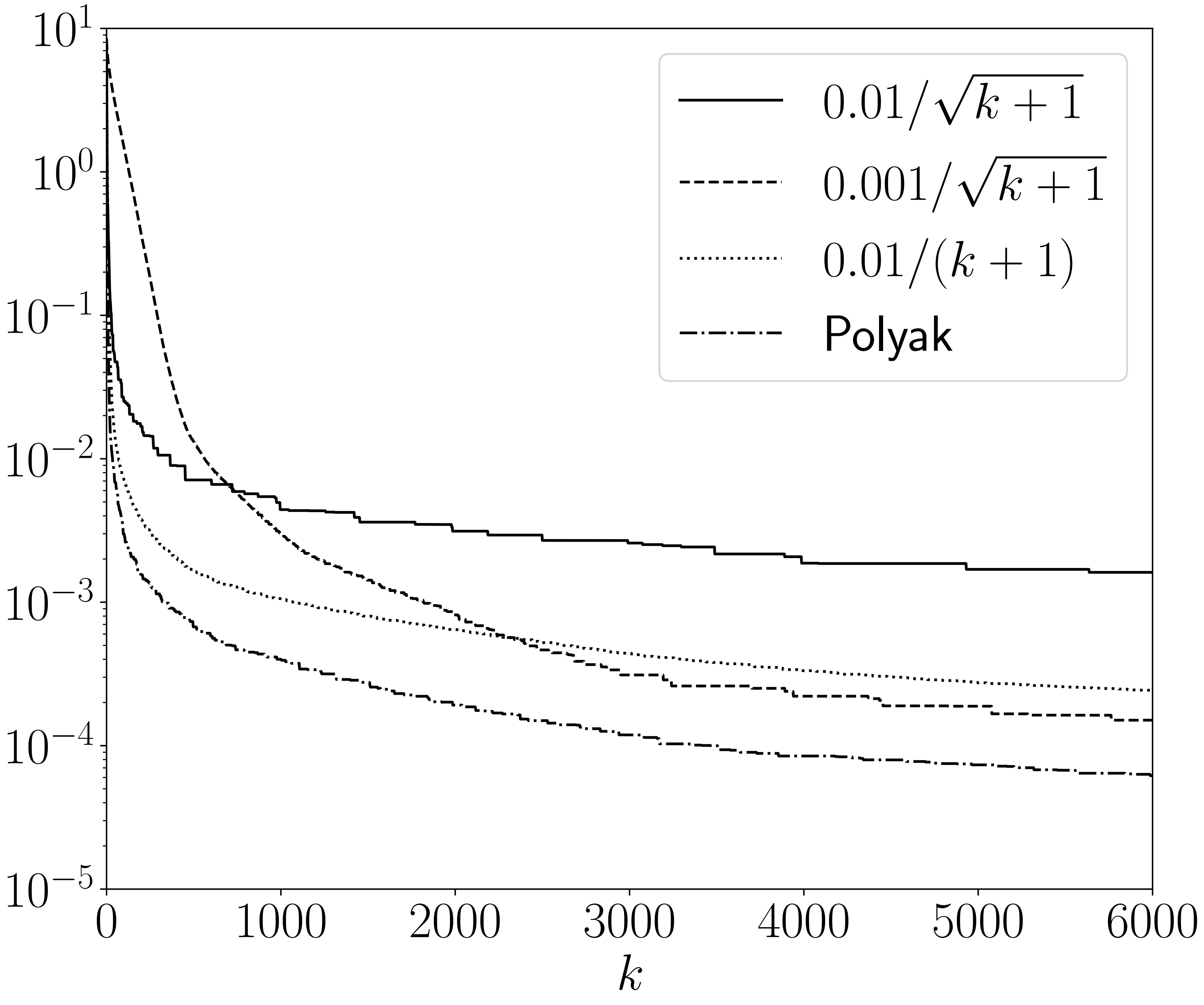
Example: 1-norm minimization

minimize $\quad f(x) = \|Ax - b\|_1 \qquad\qquad g = A^T \, \mathbf{sign}(Ax - b) \in \partial f(x)$

Fixed step size

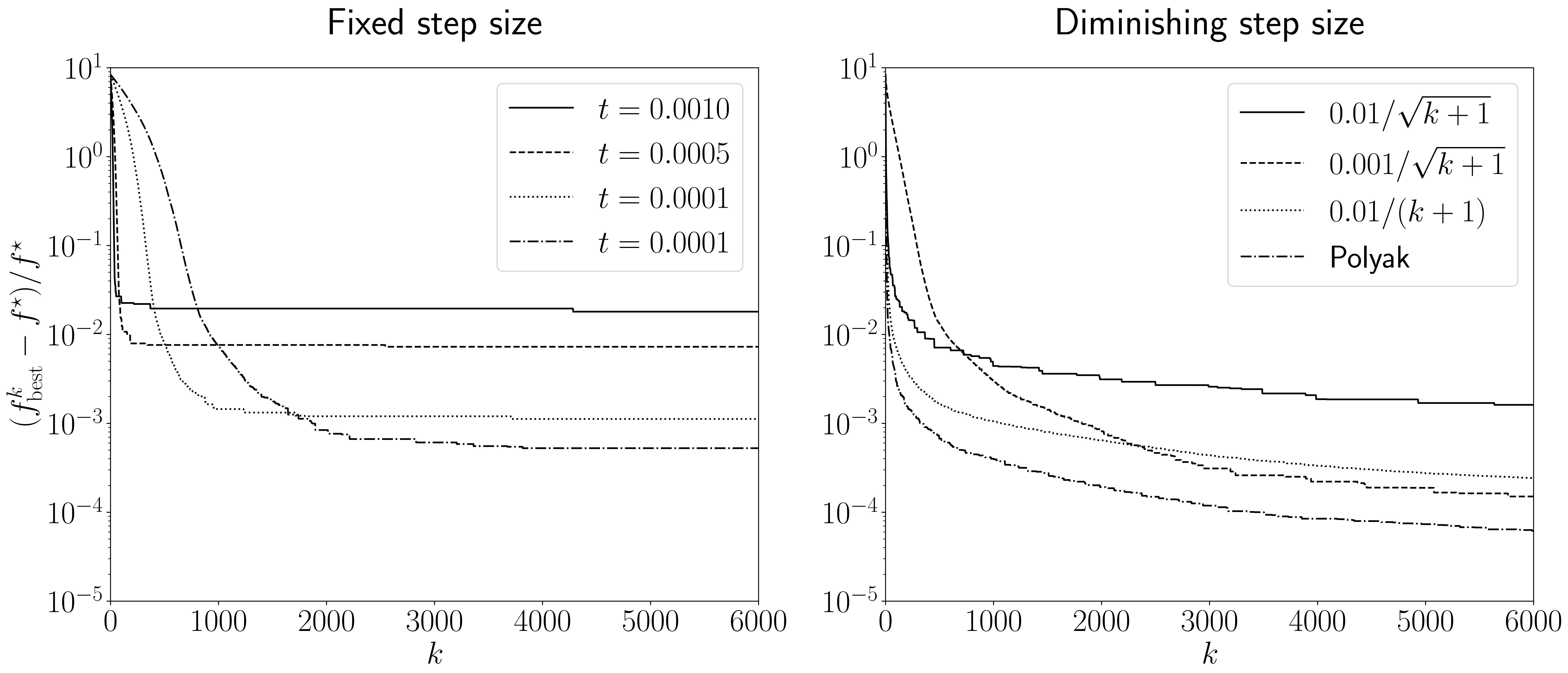


Diminishing step size



Example: 1-norm minimization

minimize $\quad f(x) = \|Ax - b\|_1 \qquad\qquad g = A^T \,\mathbf{sign}(Ax - b) \in \partial f(x)$



Efficient packages to automatically compute (sub)gradients:
Python: JAX, PyTorch
Julia: Zygote.jl, ForwardDiff.jl, ReverseDiff.jl

Summary subgradient method

- Simple
- Handles general nondifferentiable convex functions
- Very slow convergence $O(1/\epsilon^2)$
- No good stopping criterion

Summary subgradient method

- Simple
- Handles general nondifferentiable convex functions
- Very slow convergence $O(1/\epsilon^2)$
- No good stopping criterion

Can we do better?

Can we incorporate constraints?

Subgradient methods

Today, we learned to:

- **Define** subgradients
- **Apply** subgradient calculus
- **Derive** optimality conditions from subgradients
- **Define** subgradient method and **analyze** its convergence

Next lecture

- Proximal algorithms