ET4350
Applied Convex Optimization
Lecture 10

Subgradients

- subgradients
- strong and weak subgradient calculus
- optimality conditions via subgradients
- directional derivatives

Basic inequality

recall basic inequality for convex differentiable f:

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

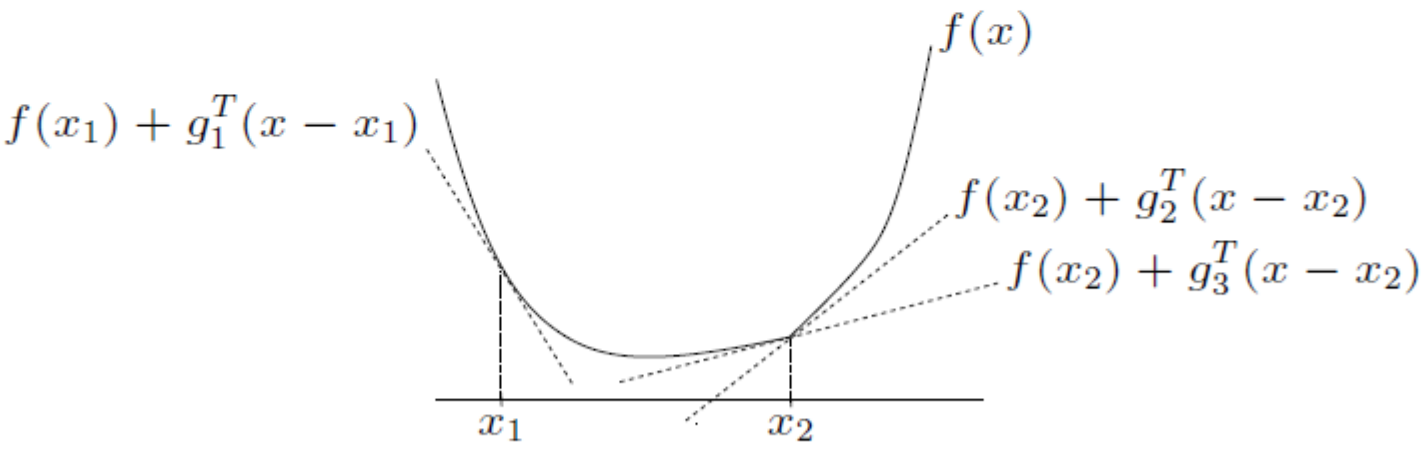
- first-order approximation of f at x is global underestimator
- $(\nabla f(x), -1)$ supports $\mathbf{epi}\, f$ at $(x, f(x))$

what if f is not differentiable?

Subgradient of a function

g is a **subgradient** of f (not necessarily convex) at x if

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



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

- g is a subgradient of f at x iff $(g, -1)$ supports $\mathbf{epi}\, f$ at $(x, f(x))$
- g is a subgradient iff $f(x) + g^T(y - x)$ is a global (affine) underestimator of f
- if f is convex and differentiable, $\nabla f(x)$ is a subgradient of f at x

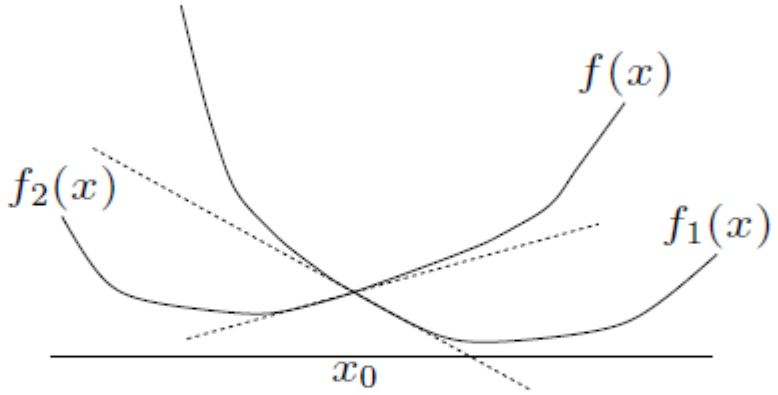
subgradients come up in several contexts:

- algorithms for nondifferentiable convex optimization
- convex analysis, $e.g.$, optimality conditions, duality for nondifferentiable problems

(if $f(y) \leq f(x) + g^T(y - x)$ for all y, then g is a **supergradient**)

Example

$f = \max\{f_1, f_2\}$, with f_1, f_2 convex and differentiable



- $f_1(x_0) > f_2(x_0)$: unique subgradient $g = \nabla f_1(x_0)$
- $f_2(x_0) > f_1(x_0)$: unique subgradient $g = \nabla f_2(x_0)$
- $f_1(x_0) = f_2(x_0)$: subgradients form a line segment $[\nabla f_1(x_0), \nabla f_2(x_0)]$

Subdifferential

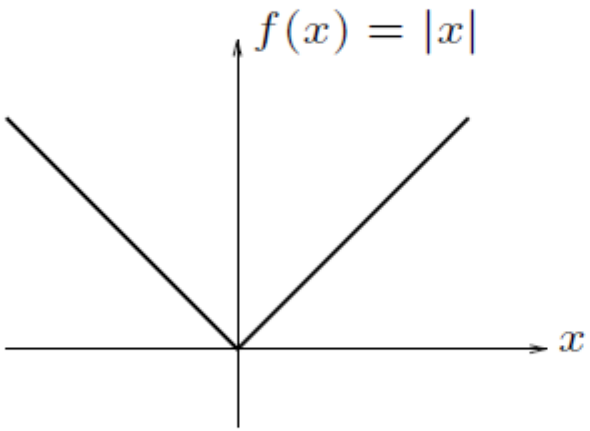
- set of all subgradients of f at x is called the **subdifferential** of f at x, denoted $\partial f(x)$
- $\partial f(x)$ is a closed convex set (can be empty)

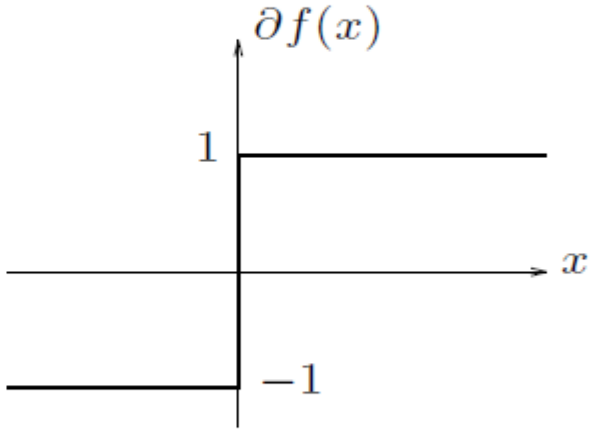
if f is convex,

- $\partial f(x)$ is nonempty, for $x \in \mathbf{relint}\,\mathbf{dom}\, f$
- $\partial f(x) = \{\nabla f(x)\}$, if f is differentiable at x
- if $\partial f(x) = \{g\}$, then f is differentiable at x and $g = \nabla f(x)$

Example

$f(x) = |x|$





righthand plot shows $\bigcup \{(x, g) \mid x \in \mathbf{R},\ g \in \partial f(x)\}$

Subgradient calculus

- **weak subgradient calculus**: formulas for finding *one* subgradient $g \in \partial f(x)$
- **strong subgradient calculus**: formulas for finding the whole subdifferential $\partial f(x)$, *i.e.*, *all* subgradients of f at x
- many algorithms for nondifferentiable convex optimization require only *one* subgradient at each step, so weak calculus suffices
- some algorithms, optimality conditions, etc., need whole subdifferential
- roughly speaking: if you can compute $f(x)$, you can usually compute a $g \in \partial f(x)$
- we'll assume that f is convex, and $x \in \mathbf{relint}\,\mathbf{dom}\, f$

Some basic rules

- $\partial f(x) = \{\nabla f(x)\}$ if f is differentiable at x
- **scaling:** $\partial(\alpha f) = \alpha \partial f$ (if $\alpha > 0$)
- **addition:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$ (RHS is addition of point-to-set mappings)
- **affine transformation of variables:** if $g(x) = f(Ax + b)$, then $\partial g(x) = A^T \partial f(Ax + b)$
- **finite pointwise maximum:** if $f = \max_{i=1,\ldots,m} f_i$, then

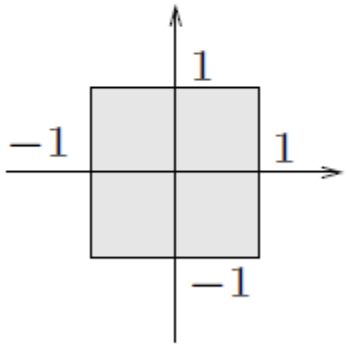
$$\partial f(x) = \mathbf{Co} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

i.e., convex hull of union of subdifferentials of 'active' functions at x

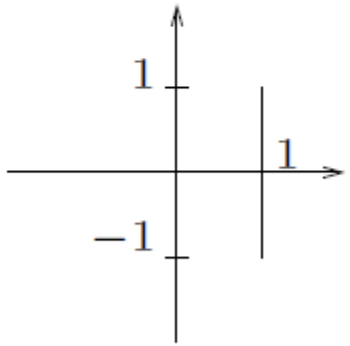
$f(x) = \max\{f_1(x), \ldots, f_m(x)\}$, with $f_1, \ldots, f_m$ differentiable

$$\partial f(x) = \mathbf{Co}\{\nabla f_i(x) \mid f_i(x) = f(x)\}$$

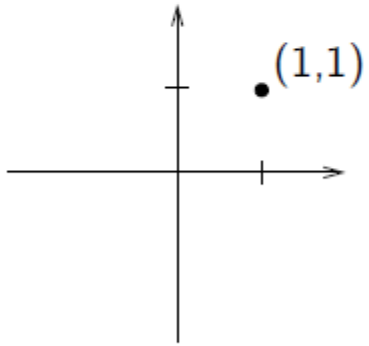
example: $f(x) = \|x\|_1 = \max\{s^T x \mid s_i \in \{-1, 1\}\}$



$\partial f(x)$ at $x = (0, 0)$



at $x = (1, 0)$



at $x = (1, 1)$

Pointwise supremum

if $f = \sup_{\alpha \in \mathcal{A}} f_\alpha$,

$$\mathbf{cl}\,\mathbf{Co} \bigcup \{\partial f_\beta(x) \mid f_\beta(x) = f(x)\} \subseteq \partial f(x)$$

(usually get equality, but requires some technical conditions to hold, *e.g.*, $\mathcal{A}$ compact, f_α cts in x and α)

roughly speaking, $\partial f(x)$ is closure of convex hull of union of subdifferentials of active functions

Weak rule for pointwise supremum

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

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

example

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

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

- f is pointwise supremum of $g_y(x) = y^T A(x) y$ over $\|y\|_2 = 1$
- g_y is affine in x, with $\nabla g_y(x) = (y^T A_1 y, \ldots, y^T A_n y)$
- hence, $\partial f(x) \supseteq \mathbf{Co}\{\nabla g_y \mid A(x)y = \lambda_{\max}(A(x))y,\ \|y\|_2 = 1\}$
  (in fact equality holds here)

to find **one** subgradient at x, can choose **any** unit eigenvector y associated with $\lambda_{\max}(A(x))$; then

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

Composition

- $f(x) = h(f_1(x), \ldots, f_k(x))$, with h convex nondecreasing, f_i convex
- find $q \in \partial h(f_1(x), \ldots, f_k(x))$, $g_i \in \partial f_i(x)$
- then, $g = q_1 g_1 + \cdots + q_k g_k \in \partial f(x)$
- reduces to standard formula for differentiable h, f_i

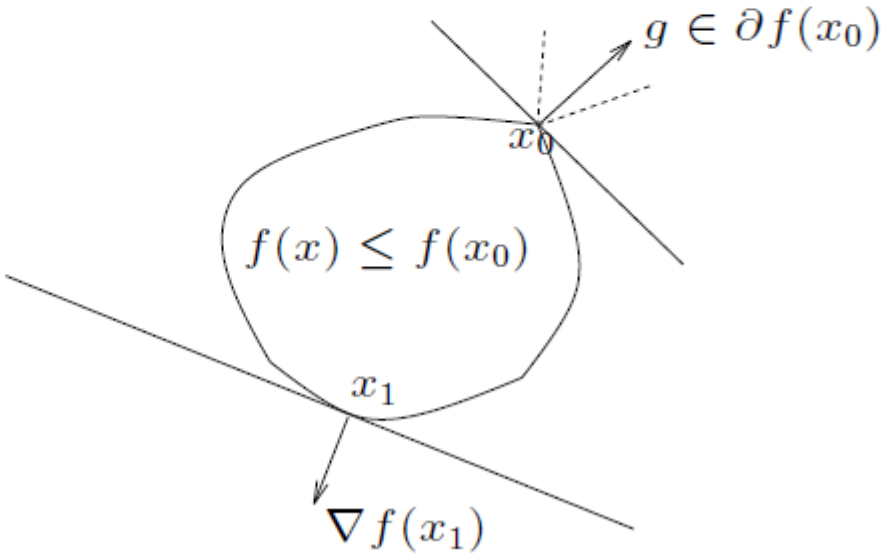
proof:

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

Subgradients and sublevel sets

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

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



- f differentiable at x_0: $\nabla f(x_0)$ is normal to the sublevel set $\{x \mid f(x) \leq f(x_0)\}$

- f nondifferentiable at x_0: subgradient defines a supporting hyperplane to sublevel set through x_0

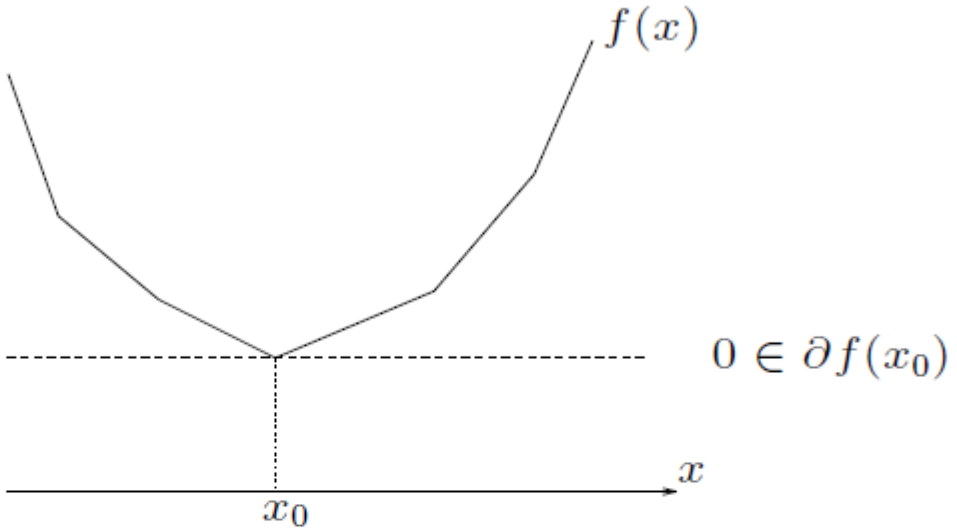
Optimality conditions — unconstrained

recall for f convex, differentiable,

$$f(x^\star) = \inf_x f(x) \iff 0 = \nabla f(x^\star)$$

generalization to nondifferentiable convex f:

$$f(x^\star) = \inf_x f(x) \iff 0 \in \partial f(x^\star)$$



proof. by definition (!)

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

. . . seems trivial but isn't

Example: piecewise linear minimization

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

$x^\star$ minimizes $f \iff 0 \in \partial f(x^\star) = \mathbf{Co}\{a_i \mid a_i^T x^\star + b_i = f(x^\star)\}$

$\iff$ there is a λ with

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

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

. . . but these are the KKT conditions for the epigraph form

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & a_i^T x + b_i \leq t, \quad i = 1, \ldots, m \end{array}$$

with dual

$$\begin{array}{lll} \text{maximize} & b^T \lambda & \\ \text{subject to} & \lambda \succeq 0, \qquad A^T \lambda = 0, & \qquad \mathbf{1}^T \lambda = 1 \end{array}$$

Optimality conditions — constrained

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

we assume

- f_i convex, defined on $\mathbf{R}^n$ (hence subdifferentiable)
- strict feasibility (Slater's condition)

$x^\star$ is primal optimal ($\lambda^\star$ is dual optimal) iff

$$\begin{aligned} & f_i(x^\star) \leq 0, \quad \lambda_i^\star \geq 0 \\ & 0 \in \partial f_0(x^\star) + \textstyle\sum_{i=1}^m \lambda_i^\star \partial f_i(x^\star) \\ & \lambda_i^\star f_i(x^\star) = 0 \end{aligned}$$

. . . generalizes KKT for nondifferentiable f_i

Directional derivative

directional derivative of f at x in the direction δx is

$$f'(x;\delta x) \stackrel{\Delta}{=} \lim_{h \searrow 0} \frac{f(x+h\delta x) - f(x)}{h}$$

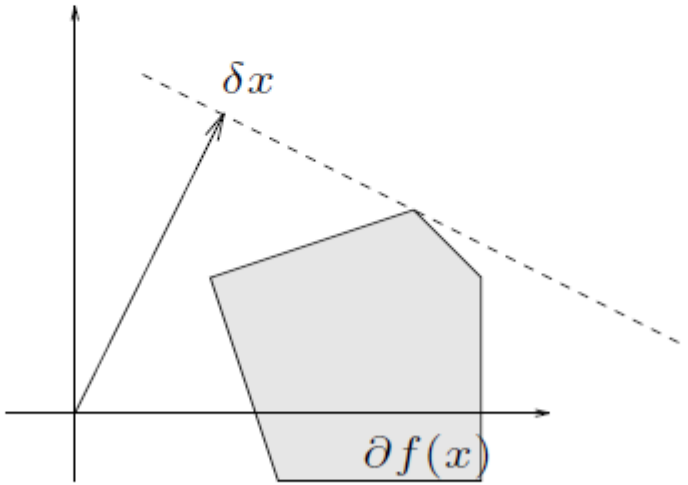
can be $+\infty$ or $-\infty$

- f convex, finite near $x \Longrightarrow f'(x;\delta x)$ exists

- f differentiable at x if and only if, for some g $(= \nabla f(x))$ and all δx, $f'(x;\delta x) = g^T \delta x$ $(i.e.,\ f'(x;\delta x)$ is a linear function of $\delta x)$

Directional derivative and subdifferential

general formula for convex f: $f'(x;\delta x) = \sup_{g \in \partial f(x)} g^T \delta x$



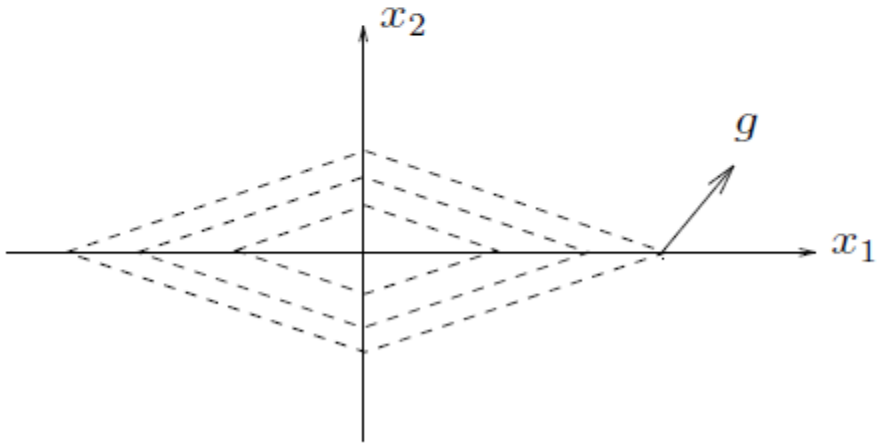
Descent directions

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

for differentiable f, $\delta x = -\nabla f(x)$ is always a descent direction (except when it is zero)

warning: for nondifferentiable (convex) functions, $\delta x = -g$, with $g \in \partial f(x)$, need not be descent direction

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



Subgradients and distance to sublevel sets

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

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

thus $-g$ is descent direction for $\|x - z\|_2$, for **any** z with $f(z) < f(x)$ $(e.g.,\ x^\star)$

negative subgradient is descent direction for distance to optimal point

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

Descent directions and optimality

fact: for f convex, finite near x, either

- $0 \in \partial f(x)$ (in which case x minimizes f), or
- there is a descent direction for f at x

i.e., x is optimal (minimizes f) iff there is no descent direction for f at x

proof: define $\delta x_{\mathrm{sd}} = -\operatorname*{argmin}_{z \in \partial f(x)} \|z\|_2$

if $\delta x_{\mathrm{sd}} = 0$, then $0 \in \partial f(x)$, so x is optimal; otherwise
$f'(x; \delta x_{\mathrm{sd}}) = -\left(\inf_{z \in \partial f(x)} \|z\|_2\right)^2 < 0$, so δx_{sd} is a descent direction

Subgradient Methods

- subgradient method and stepsize rules
- convergence results and proof
- optimal step size and alternating projections
- speeding up subgradient methods

Subgradient method

subgradient method is simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- $x^{(k)}$ is the kth iterate
- $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- $\alpha_k > 0$ is the kth step size

not a descent method, so we keep track of best point so far

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

Step size rules

step sizes are fixed ahead of time

- *constant step size:* $\alpha_k = \alpha$ (constant)
- *constant step length:* $\alpha_k = \gamma/\|g^{(k)}\|_2$ (so $\|x^{(k+1)} - x^{(k)}\|_2 = \gamma$)
- *square summable but not summable:* step sizes satisfy

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

- *nonsummable diminishing:* step sizes satisfy

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

Convergence results

define $\bar{f} = \lim_{k\to\infty} f_{\text{best}}^{(k)}$

- *constant step size:* $\bar{f} - f^\star \leq G^2\alpha/2$, *i.e.*,
  converges to $G^2\alpha/2$-suboptimal
  (converges to $f^\star$ if f differentiable, α small enough)

- *constant step length:* $\bar{f} - f^\star \leq G\gamma/2$, *i.e.*,
  converges to $G\gamma/2$-suboptimal

- *diminishing step size rule:* $\bar{f} = f^\star$, *i.e.*, **converges**

Example: Piecewise linear minimization

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

to find a subgradient of f: find index j for which

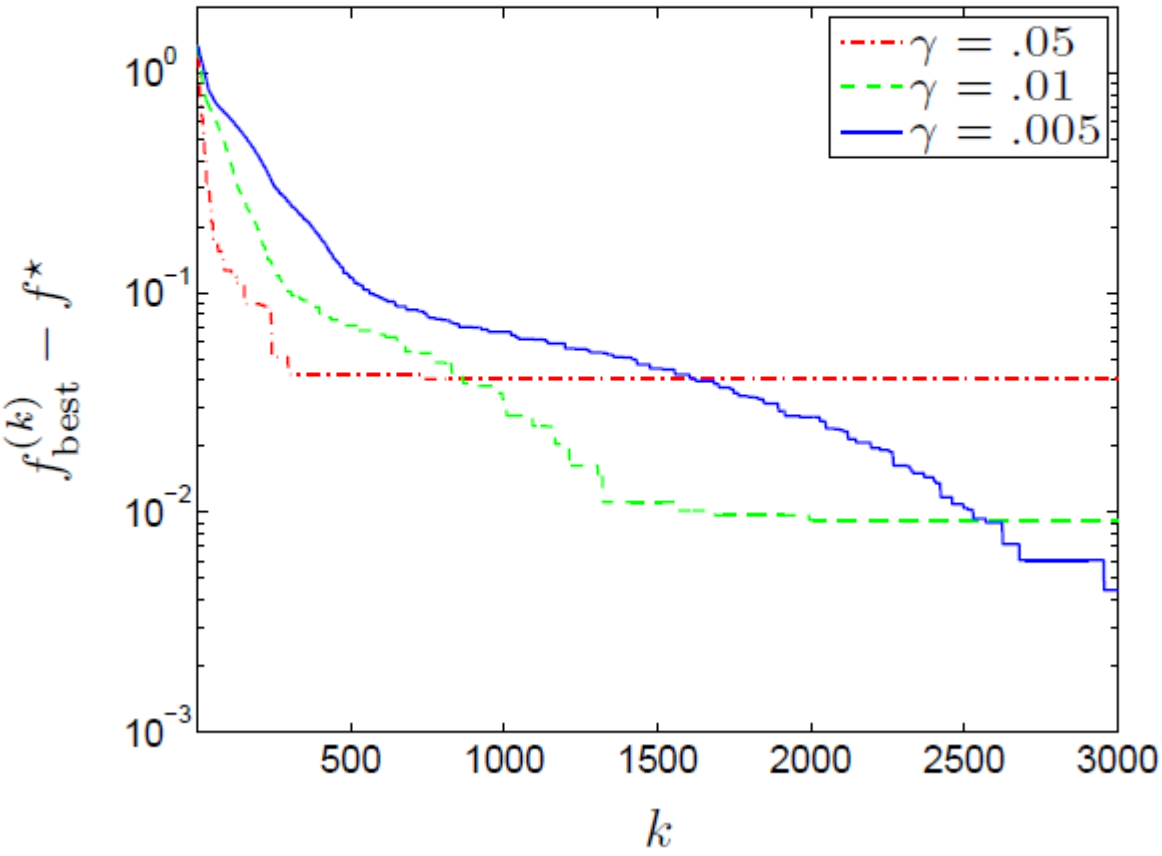
$$a_j^T x + b_j = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

and take $g = a_j$

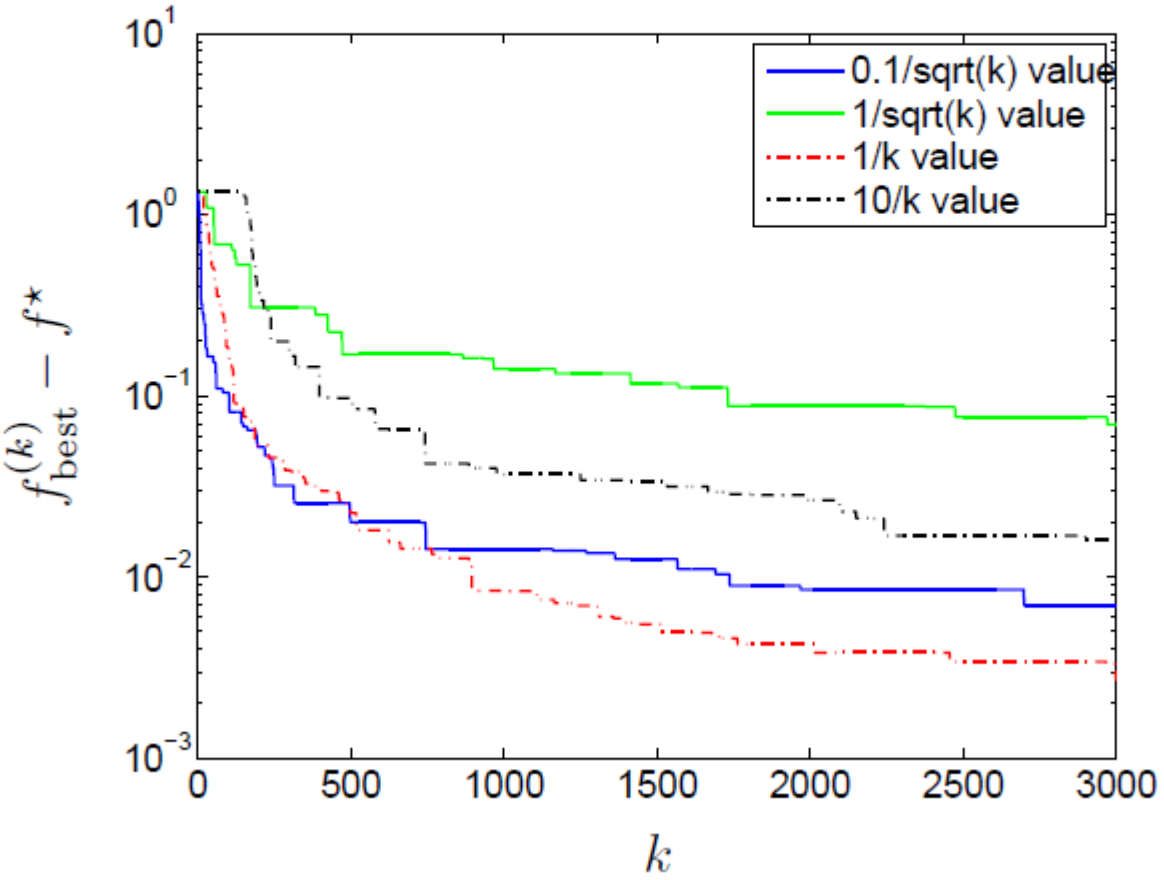
subgradient method: $x^{(k+1)} = x^{(k)} - \alpha_k a_j$

problem instance with $n = 20$ variables, $m = 100$ terms, $f^\star \approx 1.1$

$f_{\text{best}}^{(k)} - f^\star$, constant step length $\gamma = 0.05, 0.01, 0.005$



diminishing step rules $\alpha_k = 0.1/\sqrt{k}$ and $\alpha_k = 1/\sqrt{k}$, square summable step size rules $\alpha_k = 1/k$ and $\alpha_k = 10/k$



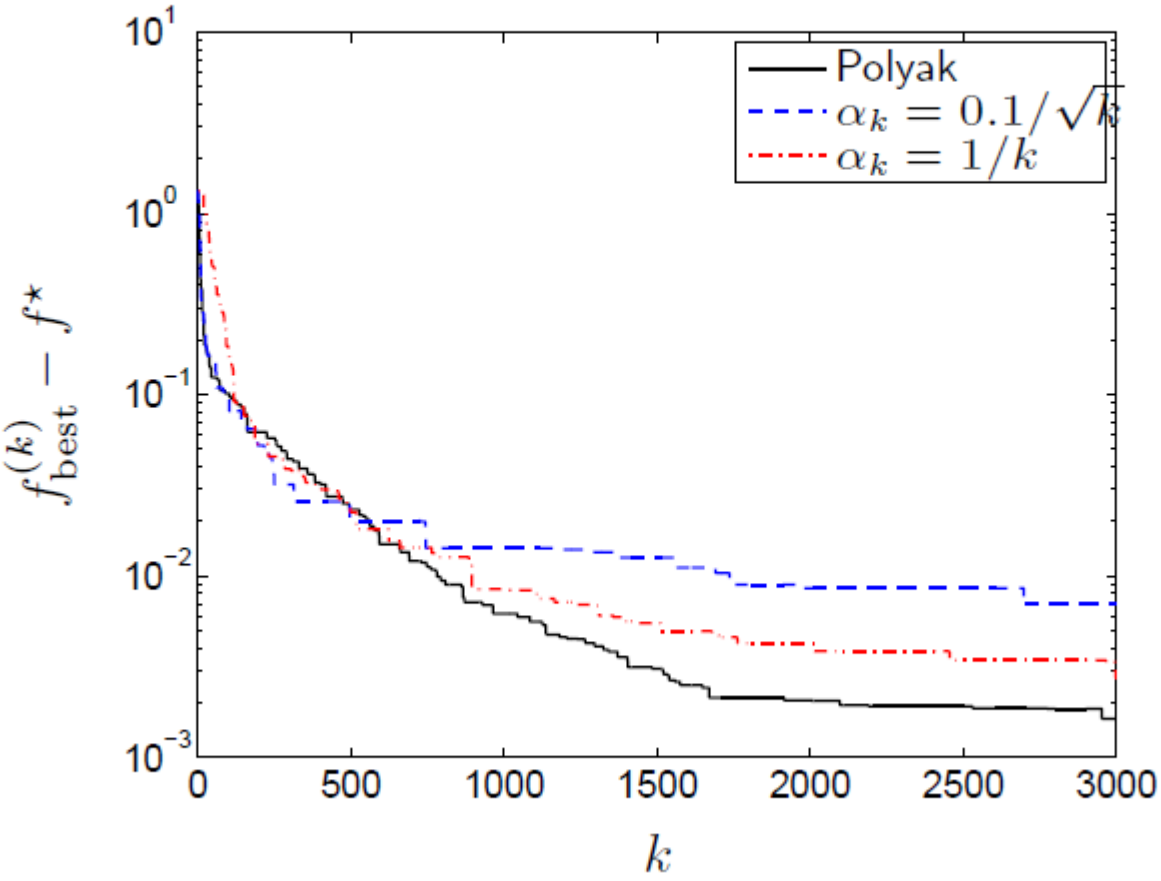
Optimal step size when $f^\star$ is known

- choice due to Polyak:

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

  (can also use when optimal value is estimated)

PWL example with Polyak's step size, $\alpha_k = 0.1/\sqrt{k}$, $\alpha_k = 1/k$



Finding a point in the intersection of convex sets

$C = C_1 \cap \cdots C_m$ is nonempty, $C_1, \ldots, C_m \subseteq \mathbf{R}^n$ closed and convex

find a point in C by minimizing

$$f(x) = \max\{\mathbf{dist}(x, C_1), \ldots, \mathbf{dist}(x, C_m)\}$$

with $\mathbf{dist}(x, C_j) = f(x)$, a subgradient of f is

$$g = \nabla \mathbf{dist}(x, C_j) = \frac{x - P_{C_j}(x)}{\|x - P_{C_j}(x)\|_2}$$

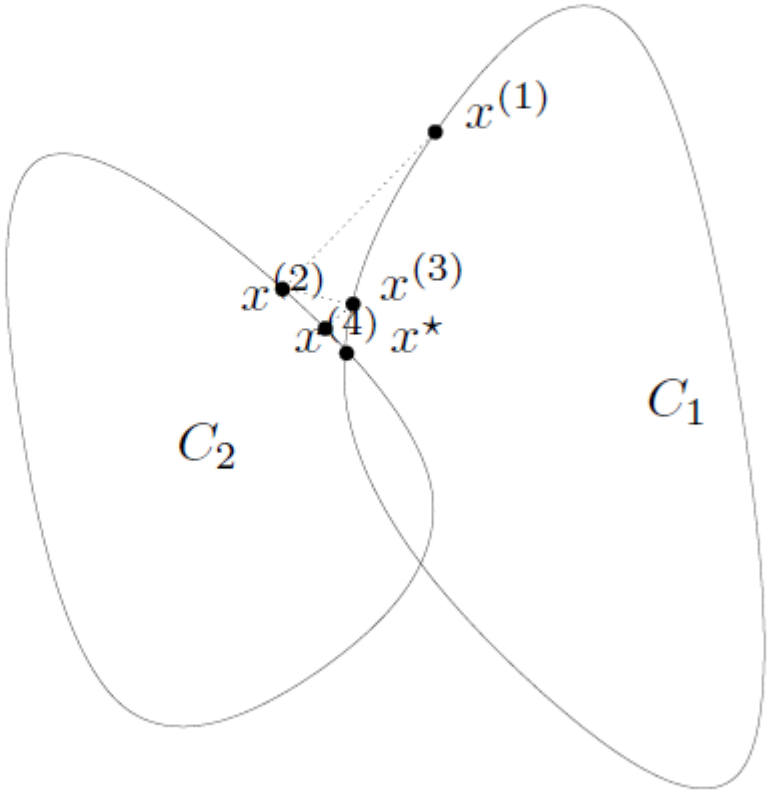
subgradient update with optimal step size:

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha_k g^{(k)} \\ &= x^{(k)} - f(x^{(k)}) \frac{x - P_{C_j}(x)}{\|x - P_{C_j}(x)\|_2} \\ &= P_{C_j}(x^{(k)}) \end{aligned}$$

- a version of the famous *alternating projections* algorithm
- at each step, project the current point onto the farthest set
- for $m = 2$ sets, projections alternate onto one set, then the other
- convergence: $\mathbf{dist}(x^{(k)}, C) \to 0$ as $k \to \infty$

Alternating projections

first few iterations:



. . . $x^{(k)}$ eventually converges to a point $x^\star \in C_1 \cap C_2$

Subgradient Methods for Constrained Problems

- projected subgradient method
- projected subgradient for dual
- subgradient method for constrained optimization

Projected subgradient method

solves constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C}, \end{array}$$

where $f : \mathbf{R}^n \to \mathbf{R}$, $\mathcal{C} \subseteq \mathbf{R}^n$ are convex

projected subgradient method is given by

$$x^{(k+1)} = \Pi(x^{(k)} - \alpha_k g^{(k)}),$$

Π is (Euclidean) projection on $\mathcal{C}$, and $g^{(k)} \in \partial f(x^{(k)})$

same convergence results:

- for constant step size, converges to neighborhood of optimal (for f differentiable and h small enough, converges)
- for diminishing nonsummable step sizes, converges

key idea: projection does not increase distance to $x^\star$

Linear equality constraints

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

projection of z onto $\{x \mid Ax = b\}$ is

$$\begin{aligned} \Pi(z) &= z - A^T(AA^T)^{-1}(Az - b) \\ &= (I - A^T(AA^T)^{-1}A)z + A^T(AA^T)^{-1}b \end{aligned}$$

projected subgradient update is (using $Ax^{(k)} = b$)

$$\begin{aligned} x^{(k+1)} &= \Pi(x^{(k)} - \alpha_k g^{(k)}) \\ &= x^{(k)} - \alpha_k(I - A^T(AA^T)^{-1}A)g^{(k)} \\ &= x^{(k)} - \alpha_k \Pi_{\mathcal{N}(A)}(g^{(k)}) \end{aligned}$$

Example: Least l_1-norm

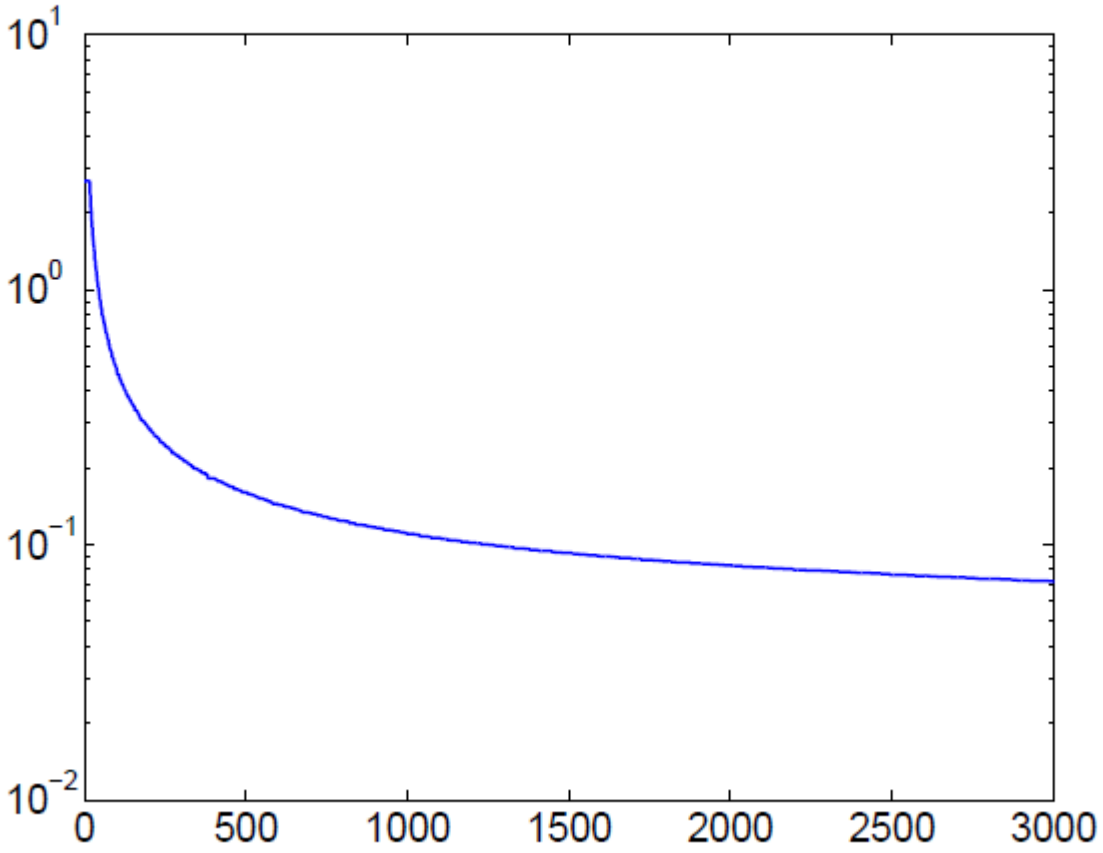
$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = b \end{array}$$

subgradient of objective is $g = \mathbf{sign}(x)$

projected subgradient update is

$$x^{(k+1)} = x^{(k)} - \alpha_k (I - A^T (AA^T)^{-1} A)\, \mathbf{sign}(x^{(k)})$$

problem instance with $n = 1000$, $m = 50$, step size $\alpha_k = 0.1/k$, $f^\star \approx 3.2$



Projected subgradient for dual problem

(convex) primal:

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

solve dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

via projected subgradient method:

$$\lambda^{(k+1)} = \left(\lambda^{(k)} - \alpha_k h\right)_+, \qquad h \in \partial(-g)(\lambda^{(k)})$$

Subgradient of negative dual function

assume f_0 is strictly convex, and denote, for $\lambda \succeq 0$,

$$x^*(\lambda) = \operatorname*{argmin}_z \left(f_0(z) + \lambda_1 f_1(z) + \cdots + \lambda_m f_m(z)\right)$$

so $g(\lambda) = f_0(x^*(\lambda)) + \lambda_1 f_1(x^*(\lambda)) + \cdots + \lambda_m f_m(x^*(\lambda))$

a subgradient of $-g$ at λ is given by $h_i = -f_i(x^*(\lambda))$

projected subgradient method for dual:

$$x^{(k)} = x^*(\lambda^{(k)}), \qquad \lambda_i^{(k+1)} = \left(\lambda_i^{(k)} + \alpha_k f_i(x^{(k)})\right)_+$$

Example

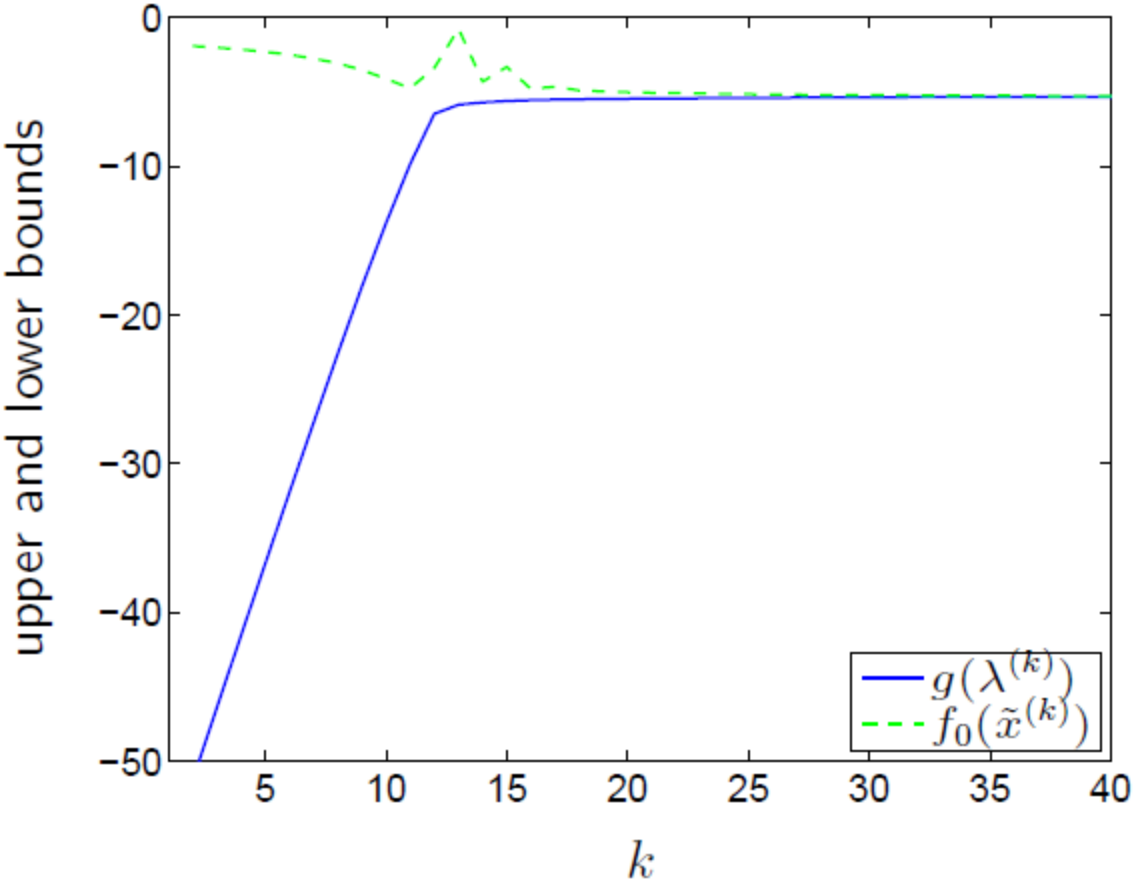
minimize strictly convex quadratic ($P \succ 0$) over unit box:

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P x - q^T x \\ \text{subject to} & x_i^2 \leq 1, \quad i = 1, \ldots, n \end{array}$$

- $L(x, \lambda) = (1/2)x^T(P + \mathbf{diag}(2\lambda))x - q^T x - \mathbf{1}^T\lambda$
- $x^*(\lambda) = (P + \mathbf{diag}(2\lambda))^{-1}q$
- projected subgradient for dual:

$$x^{(k)} = (P + \mathbf{diag}(2\lambda^{(k)}))^{-1}q, \quad \lambda_i^{(k+1)} = \left(\lambda_i^{(k)} + \alpha_k((x_i^{(k)})^2 - 1)\right)_+$$

problem instance with $n = 50$, fixed step size $\alpha = 0.1$, $f^\star \approx -5.3$; $\tilde{x}^{(k)}$ is a nearby feasible point for $x^{(k)}$



Subgradient method for constrained optimization

solves constrained optimization problem

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

where $f_i : \mathbf{R}^n \to \mathbf{R}$ are convex

same update $x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$, but we have

$$g^{(k)} \in \begin{cases} \partial f_0(x) & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ \partial f_j(x) & f_j(x) > 0 \end{cases}$$

define $f_{\text{best}}^{(k)} = \min\{f_0(x^{(i)}) \mid x^{(i)} \text{ feasible}, \ i = 1, \ldots, k\}$

Convergence

assumptions:

- there exists an optimal $x^\star$; Slater's condition holds
- $\|g^{(k)}\|_2 \leq G$; $\|x^{(1)} - x^\star\|_2 \leq R$

typical result: for $\alpha_k > 0$, $\alpha_k \to 0$, $\sum_{i=1}^{\infty} \alpha_i = \infty$, we have $f_{\text{best}}^{(k)} \to f^\star$