

5. Duality

- Lagrange dual problem
- weak and strong duality
- geometric interpretation
- optimality conditions
- perturbation and sensitivity analysis
- examples
- generalized inequalities

Perturbation and sensitivity analysis

(unperturbed) optimization problem and its dual

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

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

perturbed problem and its dual

$$\begin{array}{ll}\text{min.} & f_0(x) \\ \text{s.t.} & f_i(x) \leq u_i, \quad i = 1, \dots, m \\ & h_i(x) = v_i, \quad i = 1, \dots, p\end{array}$$

$$\begin{array}{ll}\text{max.} & g(\lambda, \nu) - u^T \lambda - v^T \nu \\ \text{s.t.} & \lambda \succeq 0\end{array}$$

- x is primal variable; u, v are parameters
- $p^*(u, v)$ is optimal value as a function of u, v
- we are interested in information about $p^*(u, v)$ that we can obtain from the solution of the unperturbed problem and its dual

global sensitivity result

assume strong duality holds for unperturbed problem, and that λ^* , ν^* are dual optimal for unperturbed problem

apply weak duality to perturbed problem:

$$\begin{aligned} p^*(u, v) &\geq g(\lambda^*, \nu^*) - u^T \lambda^* - v^T \nu^* \\ &= p^*(0, 0) - u^T \lambda^* - v^T \nu^* \end{aligned}$$

sensitivity interpretation

- if λ_i^* large: p^* increases greatly if we tighten constraint i ($u_i < 0$)
- if λ_i^* small: p^* does not decrease much if we loosen constraint i ($u_i > 0$)
- if ν_i^* large and positive: p^* increases greatly if we take $v_i < 0$;
if ν_i^* large and negative: p^* increases greatly if we take $v_i > 0$
- if ν_i^* small and positive: p^* does not decrease much if we take $v_i > 0$;
if ν_i^* small and negative: p^* does not decrease much if we take $v_i < 0$

local sensitivity: if (in addition) $p^*(u, v)$ is differentiable at $(0, 0)$, then

$$\lambda_i^* = -\frac{\partial p^*(0, 0)}{\partial u_i}, \quad \nu_i^* = -\frac{\partial p^*(0, 0)}{\partial v_i}$$

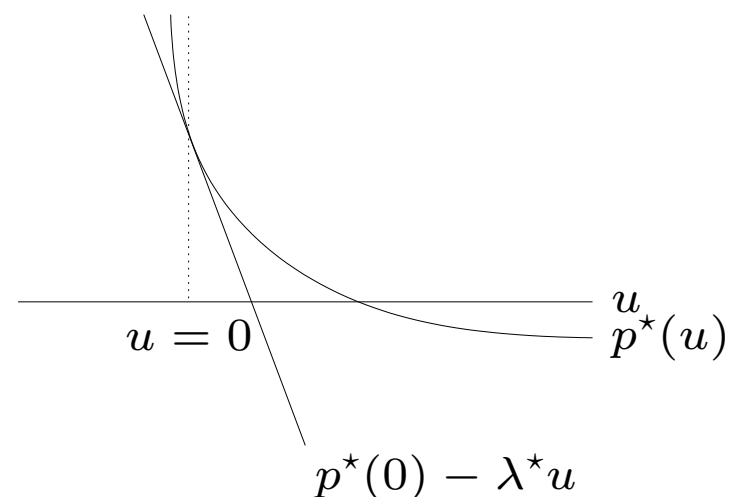
proof (for λ_i^*): from global sensitivity result,

$$\frac{\partial p^*(0, 0)}{\partial u_i} = \lim_{t \searrow 0} \frac{p^*(te_i, 0) - p^*(0, 0)}{t} \geq -\lambda_i^*$$

$$\frac{\partial p^*(0, 0)}{\partial u_i} = \lim_{t \nearrow 0} \frac{p^*(te_i, 0) - p^*(0, 0)}{t} \leq -\lambda_i^*$$

hence, equality

$p^*(u)$ for a problem with one (inequality) constraint:



Duality and problem reformulations

- equivalent formulations of a problem can lead to very different duals
- reformulating the primal problem can be useful when the dual is difficult to derive, or uninteresting

common reformulations

- introduce new variables and equality constraints
- make explicit constraints implicit or vice-versa
- transform objective or constraint functions
e.g., replace $f_0(x)$ by $\phi(f_0(x))$ with ϕ convex, increasing

Introducing new variables and equality constraints

$$\text{minimize } f_0(Ax + b)$$

- dual function is constant: $g = \inf_x L(x) = \inf_x f_0(Ax + b) = p^*$
- we have strong duality, but dual is quite useless

reformulated problem and its dual

$$\begin{array}{ll} \text{minimize} & f_0(y) \\ \text{subject to} & Ax + b - y = 0 \end{array}$$

$$\begin{array}{ll} \text{maximize} & b^T \nu - f_0^*(\nu) \\ \text{subject to} & A^T \nu = 0 \end{array}$$

dual function follows from

$$\begin{aligned} g(\nu) &= \inf_{x,y} (f_0(y) - \nu^T y + \nu^T Ax + b^T \nu) \\ &= \begin{cases} -f_0^*(\nu) + b^T \nu & A^T \nu = 0 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

norm approximation problem: minimize $\|Ax - b\|$

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

can look up conjugate of $\|\cdot\|$, or derive dual directly

$$\begin{aligned}g(\nu) &= \inf_{x,y} (\|y\| + \nu^T y - \nu^T Ax + b^T \nu) \\ &= \begin{cases} b^T \nu + \inf_y (\|y\| + \nu^T y) & A^T \nu = 0 \\ -\infty & \text{otherwise} \end{cases} \\ &= \begin{cases} b^T \nu & A^T \nu = 0, \quad \|\nu\|_* \leq 1 \\ -\infty & \text{otherwise} \end{cases}\end{aligned}$$

(see page 5–4)

dual of norm approximation problem

$$\begin{array}{ll}\text{maximize} & b^T \nu \\ \text{subject to} & A^T \nu = 0, \quad \|\nu\|_* \leq 1\end{array}$$

Implicit constraints

LP with box constraints: primal and dual problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & -\mathbf{1} \preceq x \preceq \mathbf{1} \end{array} \qquad \begin{array}{ll} \text{maximize} & -b^T \nu - \mathbf{1}^T \lambda_1 - \mathbf{1}^T \lambda_2 \\ \text{subject to} & c + A^T \nu + \lambda_1 - \lambda_2 = 0 \\ & \lambda_1 \succeq 0, \quad \lambda_2 \succeq 0 \end{array}$$

reformulation with box constraints made implicit

$$\begin{array}{ll} \text{minimize} & f_0(x) = \begin{cases} c^T x & -\mathbf{1} \preceq x \preceq \mathbf{1} \\ \infty & \text{otherwise} \end{cases} \\ \text{subject to} & Ax = b \end{array}$$

dual function

$$\begin{aligned} g(\nu) &= \inf_{-\mathbf{1} \preceq x \preceq \mathbf{1}} (c^T x + \nu^T (Ax - b)) \\ &= -b^T \nu - \|A^T \nu + c\|_1 \end{aligned}$$

dual problem: maximize $-b^T \nu - \|A^T \nu + c\|_1$

Problems with generalized inequalities

$$\begin{array}{ll}\text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \preceq_{K_i} 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p\end{array}$$

\preceq_{K_i} is generalized inequality on \mathbf{R}^{k_i}

definitions are parallel to scalar case:

- Lagrange multiplier for $f_i(x) \preceq_{K_i} 0$ is vector $\lambda_i \in \mathbf{R}^{k_i}$
- Lagrangian $L : \mathbf{R}^n \times \mathbf{R}^{k_1} \times \dots \times \mathbf{R}^{k_m} \times \mathbf{R}^p \rightarrow \mathbf{R}$, is defined as

$$L(x, \lambda_1, \dots, \lambda_m, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i^T f_i(x) + \sum_{i=1}^p \nu_i h_i(x)$$

- dual function $g : \mathbf{R}^{k_1} \times \dots \times \mathbf{R}^{k_m} \times \mathbf{R}^p \rightarrow \mathbf{R}$, is defined as

$$g(\lambda_1, \dots, \lambda_m, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda_1, \dots, \lambda_m, \nu)$$

lower bound property: if $\lambda_i \succeq_{K_i^*} 0$, then $g(\lambda_1, \dots, \lambda_m, \nu) \leq p^*$

proof: if \tilde{x} is feasible and $\lambda \succeq_{K_i^*} 0$, then

$$\begin{aligned} f_0(\tilde{x}) &\geq f_0(\tilde{x}) + \sum_{i=1}^m \lambda_i^T f_i(\tilde{x}) + \sum_{i=1}^p \nu_i h_i(\tilde{x}) \\ &\geq \inf_{x \in \mathcal{D}} L(x, \lambda_1, \dots, \lambda_m, \nu) \\ &= g(\lambda_1, \dots, \lambda_m, \nu) \end{aligned}$$

minimizing over all feasible \tilde{x} gives $p^* \geq g(\lambda_1, \dots, \lambda_m, \nu)$

dual problem

$$\begin{aligned} &\text{maximize} && g(\lambda_1, \dots, \lambda_m, \nu) \\ &\text{subject to} && \lambda_i \succeq_{K_i^*} 0, \quad i = 1, \dots, m \end{aligned}$$

- weak duality: $p^* \geq d^*$ always
- strong duality: $p^* = d^*$ for convex problem with constraint qualification (for example, Slater's: primal problem is strictly feasible)

Semidefinite program

primal SDP ($F_i, G \in \mathbf{S}^k$)

$$\begin{array}{ll}\text{minimize} & c^T x \\ \text{subject to} & x_1 F_1 + \cdots + x_n F_n \preceq G\end{array}$$

- Lagrange multiplier is matrix $Z \in \mathbf{S}^k$
- Lagrangian $L(x, Z) = c^T x + \mathbf{tr}(Z(x_1 F_1 + \cdots + x_n F_n - G))$
- dual function

$$g(Z) = \inf_x L(x, Z) = \begin{cases} -\mathbf{tr}(GZ) & \mathbf{tr}(F_i Z) + c_i = 0, \quad i = 1, \dots, n \\ -\infty & \text{otherwise} \end{cases}$$

dual SDP

$$\begin{array}{ll}\text{maximize} & -\mathbf{tr}(GZ) \\ \text{subject to} & Z \succeq 0, \quad \mathbf{tr}(F_i Z) + c_i = 0, \quad i = 1, \dots, n\end{array}$$

$p^* = d^*$ if primal SDP is strictly feasible ($\exists x$ with $x_1 F_1 + \cdots + x_n F_n \prec G$)

10. Unconstrained minimization

- terminology and assumptions
- gradient descent method
- steepest descent method
- Newton's method
- self-concordant functions
- implementation

Unconstrained minimization

$$\text{minimize } f(x)$$

- f convex, twice continuously differentiable (hence $\text{dom } f$ open)
- we assume optimal value $p^* = \inf_x f(x)$ is attained (and finite)

unconstrained minimization methods

- produce sequence of points $x^{(k)} \in \text{dom } f$, $k = 0, 1, \dots$ with

$$f(x^{(k)}) \rightarrow p^*$$

- can be interpreted as iterative methods for solving optimality condition

$$\nabla f(x^*) = 0$$

Initial point and sublevel set

algorithms in this chapter require a starting point $x^{(0)}$ such that

- $x^{(0)} \in \mathbf{dom} f$
- sublevel set $S = \{x \mid f(x) \leq f(x^{(0)})\}$ is closed

2nd condition is hard to verify, except when *all* sublevel sets are closed:

- equivalent to condition that $\mathbf{epi} f$ is closed
- true if $\mathbf{dom} f = \mathbf{R}^n$
- true if $f(x) \rightarrow \infty$ as $x \rightarrow \mathbf{bd} \mathbf{dom} f$

examples of differentiable functions with closed sublevel sets:

$$f(x) = \log\left(\sum_{i=1}^m \exp(a_i^T x + b_i)\right), \quad f(x) = -\sum_{i=1}^m \log(b_i - a_i^T x)$$

Strong convexity and implications

f is strongly convex on S if there exists an $m > 0$ such that

$$\nabla^2 f(x) \succeq mI \quad \text{for all } x \in S$$

implications

- for $x, y \in S$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|x - y\|_2^2$$

hence, S is bounded

- $p^\star > -\infty$, and for $x \in S$,

$$f(x) - p^\star \leq \frac{1}{2m} \|\nabla f(x)\|_2^2$$

useful as stopping criterion (if you know m)

Descent methods

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \quad \text{with } f(x^{(k+1)}) < f(x^{(k)})$$

- other notations: $x^+ = x + t\Delta x$, $x := x + t\Delta x$
- Δx is the *step*, or *search direction*; t is the *step size*, or *step length*
- from convexity, $f(x^+) < f(x)$ implies $\nabla f(x)^T \Delta x < 0$
(*i.e.*, Δx is a *descent direction*)

General descent method.

given a starting point $x \in \text{dom } f$.

repeat

1. Determine a descent direction Δx .
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

Line search types

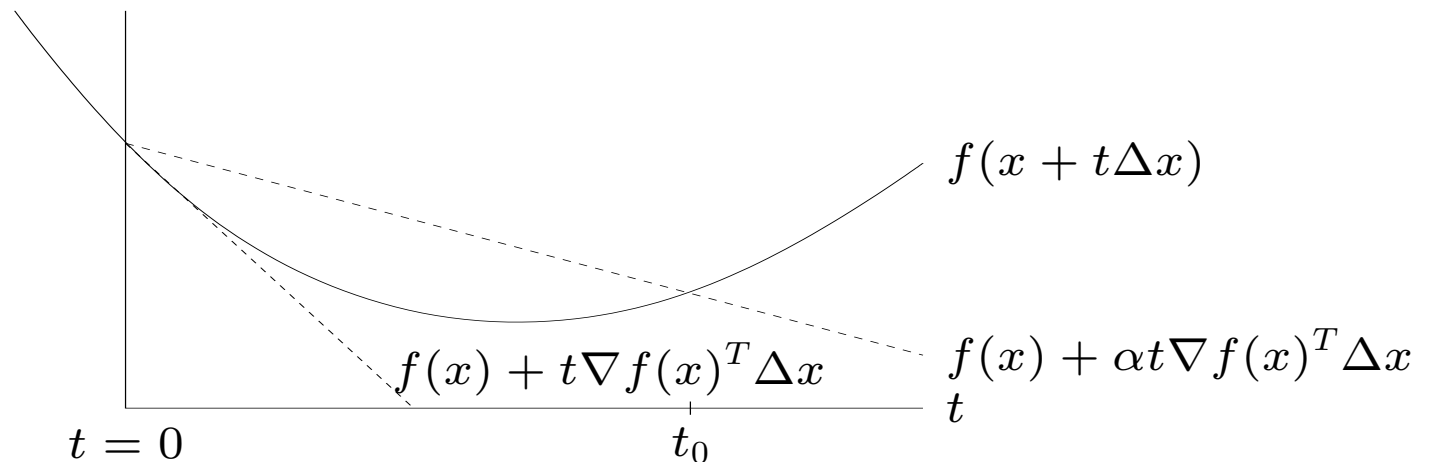
exact line search: $t = \operatorname{argmin}_{t>0} f(x + t\Delta x)$

backtracking line search (with parameters $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$)

- starting at $t = 1$, repeat $t := \beta t$ until

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x$$

- graphical interpretation: backtrack until $t \leq t_0$



Gradient descent method

general descent method with $\Delta x = -\nabla f(x)$

given a starting point $x \in \text{dom } f$.

repeat

1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

- stopping criterion usually of the form $\|\nabla f(x)\|_2 \leq \epsilon$
- convergence result: for strongly convex f ,

$$f(x^{(k)}) - p^* \leq c^k (f(x^{(0)}) - p^*)$$

$c \in (0, 1)$ depends on m , $x^{(0)}$, line search type

- very simple, but often very slow; rarely used in practice

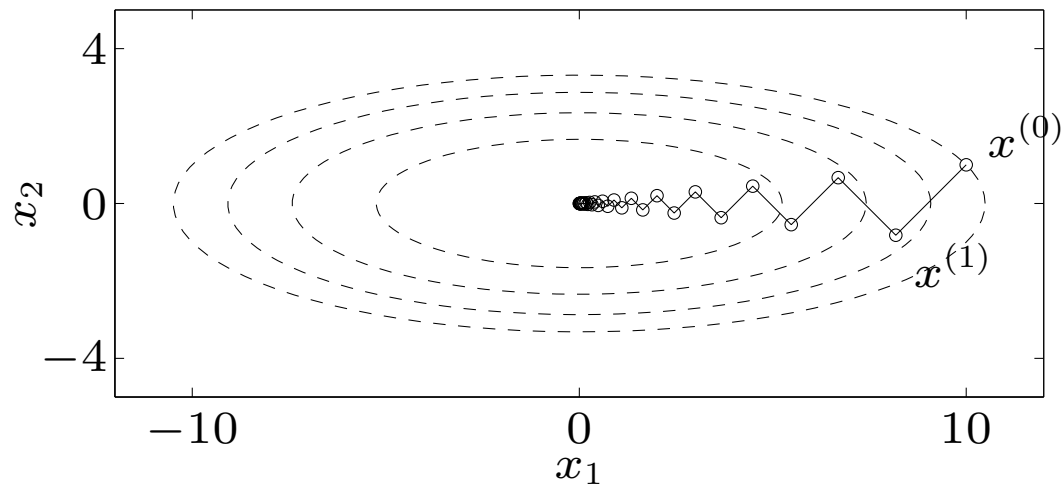
quadratic problem in \mathbf{R}^2

$$f(x) = (1/2)(x_1^2 + \gamma x_2^2) \quad (\gamma > 0)$$

with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

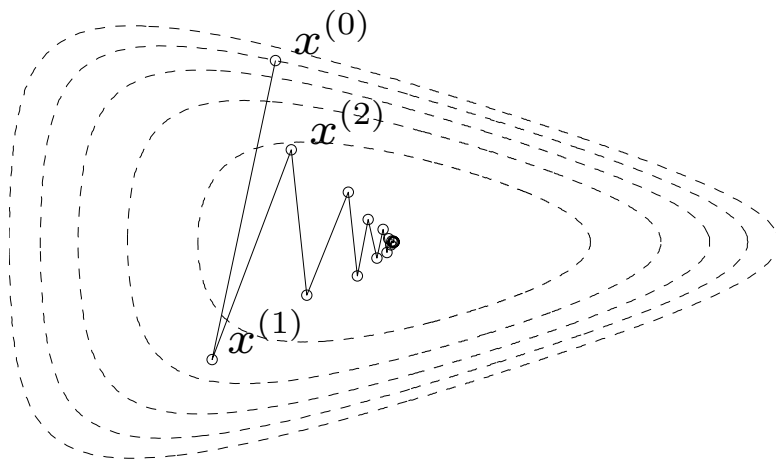
$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1} \right)^k, \quad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1} \right)^k$$

- very slow if $\gamma \gg 1$ or $\gamma \ll 1$
- example for $\gamma = 10$:

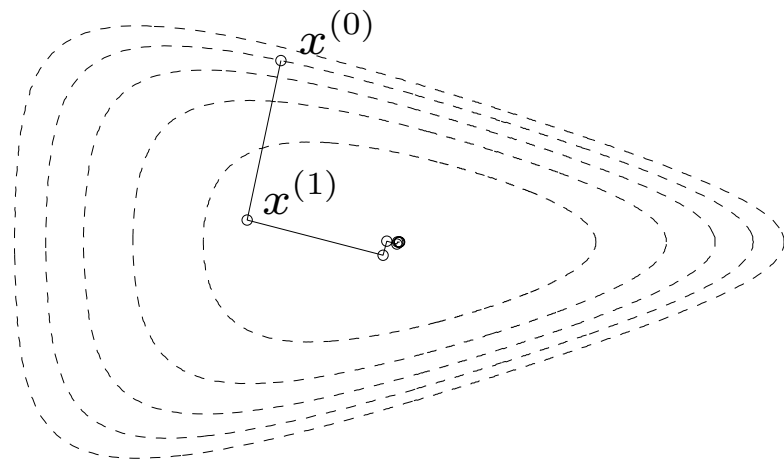


nonquadratic example

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$



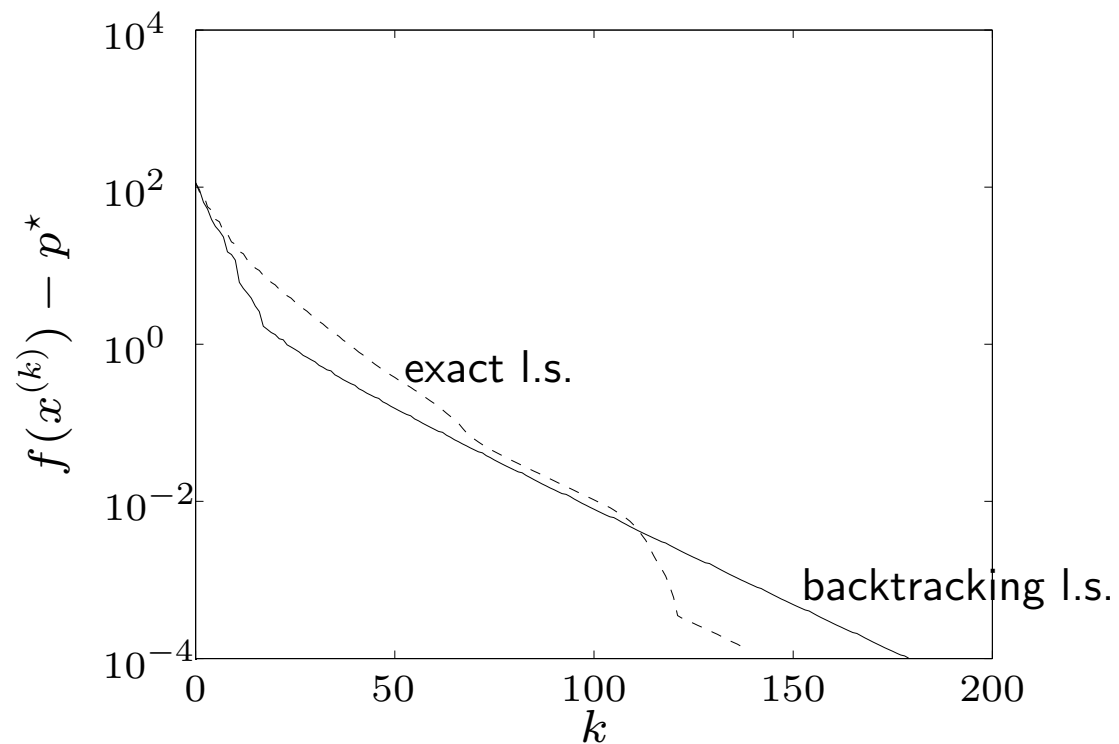
backtracking line search



exact line search

a problem in \mathbf{R}^{100}

$$f(x) = c^T x - \sum_{i=1}^{500} \log(b_i - a_i^T x)$$



‘linear’ convergence, *i.e.*, a straight line on a semilog plot

Steepest descent method

normalized steepest descent direction (at x , for norm $\|\cdot\|$):

$$\Delta x_{\text{nsd}} = \operatorname{argmin}\{\nabla f(x)^T v \mid \|v\| = 1\}$$

interpretation: for small v , $f(x + v) \approx f(x) + \nabla f(x)^T v$;
direction Δx_{nsd} is unit-norm step with most negative directional derivative

(unnormalized) steepest descent direction

$$\Delta x_{\text{sd}} = \|\nabla f(x)\|_* \Delta x_{\text{nsd}}$$

satisfies $\nabla f(x)^T \Delta x_{\text{sd}} = -\|\nabla f(x)\|_*^2$

steepest descent method

- general descent method with $\Delta x = \Delta x_{\text{sd}}$
- convergence properties similar to gradient descent

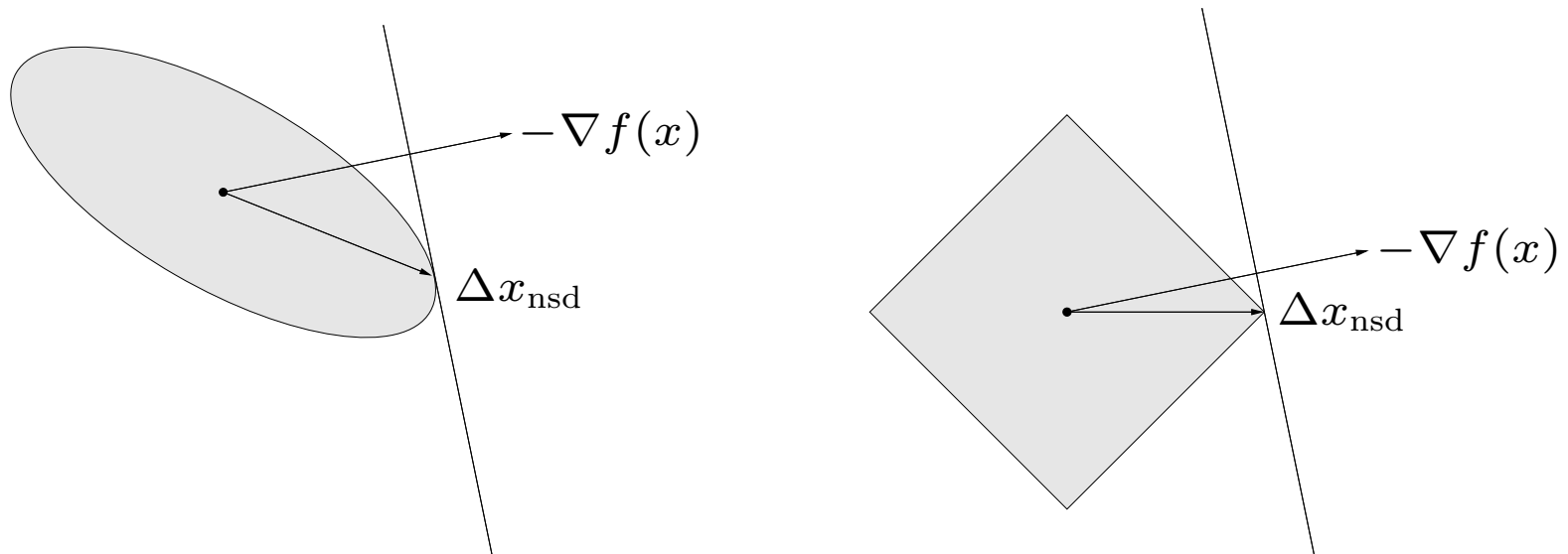
examples

- Euclidean norm: $\Delta x_{\text{sd}} = -\nabla f(x)$
- quadratic norm $\|x\|_P = (x^T P x)^{1/2}$ ($P \in \mathbf{S}_{++}^n$): $\Delta x_{\text{sd}} = -P^{-1} \nabla f(x)$

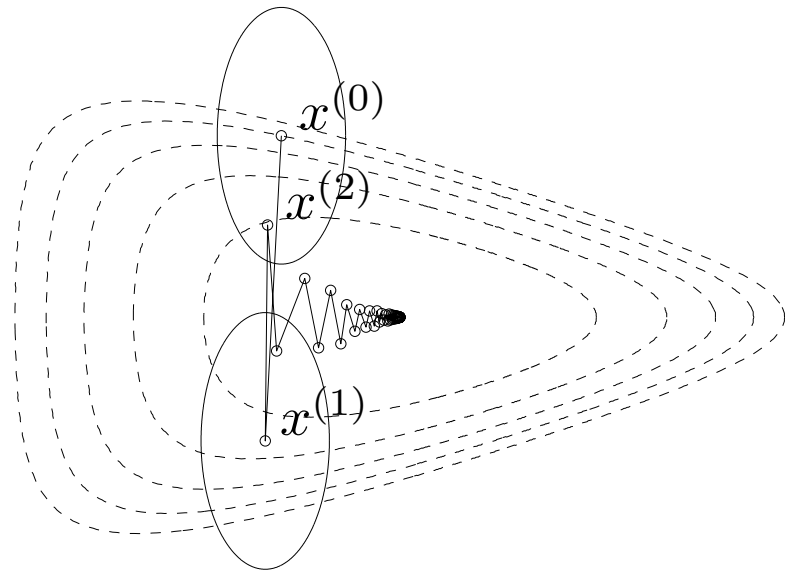
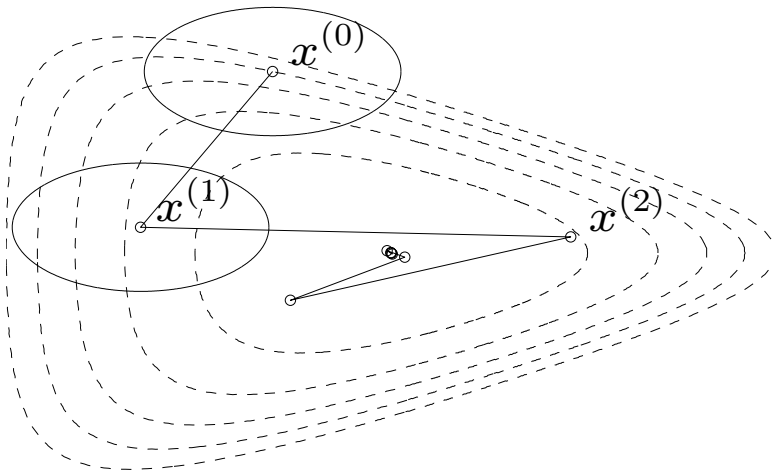
examples

- Euclidean norm: $\Delta x_{\text{sd}} = -\nabla f(x)$
- quadratic norm $\|x\|_P = (x^T P x)^{1/2}$ ($P \in \mathbf{S}_{++}^n$): $\Delta x_{\text{sd}} = -P^{-1} \nabla f(x)$
- ℓ_1 -norm: $\Delta x_{\text{sd}} = -(\partial f(x)/\partial x_i) e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

unit balls and normalized steepest descent directions for a quadratic norm and the ℓ_1 -norm:



choice of norm for steepest descent



- steepest descent with backtracking line search for two quadratic norms
- ellipses show $\{x \mid \|x - x^{(k)}\|_P = 1\}$
- equivalent interpretation of steepest descent with quadratic norm $\|\cdot\|_P$: gradient descent after change of variables $\bar{x} = P^{1/2}x$

shows choice of P has strong effect on speed of convergence

Newton step

$$\Delta x_{\text{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

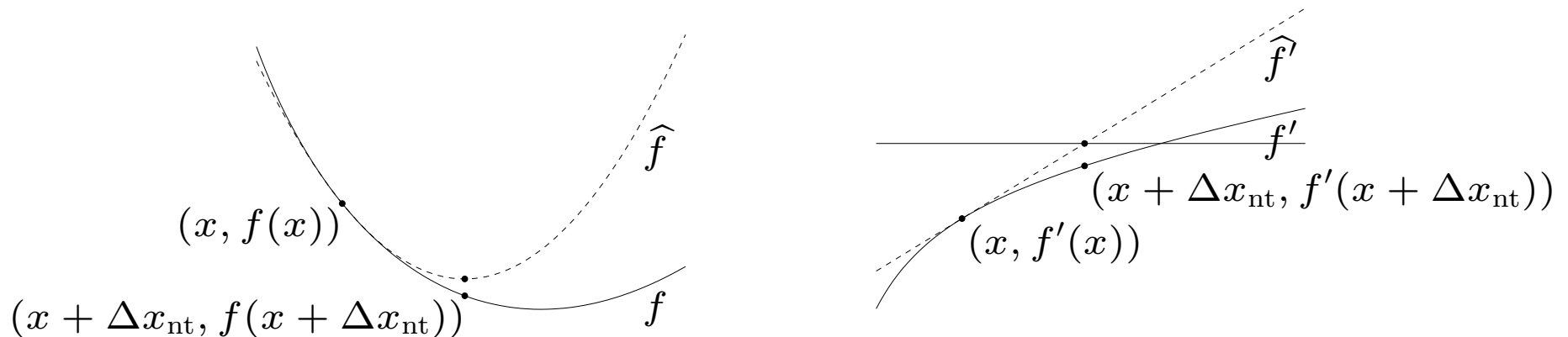
interpretations

- $x + \Delta x_{\text{nt}}$ minimizes second order approximation

$$\hat{f}(x + v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

- $x + \Delta x_{\text{nt}}$ solves linearized optimality condition

$$\nabla f(x + v) \approx \nabla \hat{f}(x + v) = \nabla f(x) + \nabla^2 f(x) v = 0$$



- Δx_{nt} is steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = (u^T \nabla^2 f(x) u)^{1/2}$$



dashed lines are contour lines of f ; ellipse is $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$

arrow shows $-\nabla f(x)$

Newton decrement

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \right)^{1/2}$$

a measure of the proximity of x to x^\star

properties

- gives an estimate of $f(x) - p^\star$, using quadratic approximation \hat{f} :

$$f(x) - \inf_y \hat{f}(y) = \frac{1}{2} \lambda(x)^2$$

Newton decrement

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \right)^{1/2}$$

a measure of the proximity of x to x^*

properties

- gives an estimate of $f(x) - p^*$, using quadratic approximation \hat{f} :

$$f(x) - \inf_y \hat{f}(y) = \frac{1}{2} \lambda(x)^2$$

- equal to the norm of the Newton step in the quadratic Hessian norm

$$\lambda(x) = \left(\Delta x_{\text{nt}}^T \nabla^2 f(x) \Delta x_{\text{nt}} \right)^{1/2}$$

- directional derivative in the Newton direction: $\nabla f(x)^T \Delta x_{\text{nt}} = -\lambda(x)^2$
- affine invariant (unlike $\|\nabla f(x)\|_2$)

Newton's method

given a starting point $x \in \text{dom } f$, tolerance $\epsilon > 0$.

repeat

1. *Compute the Newton step and decrement.*

$$\Delta x_{\text{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x); \quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x).$$

2. *Stopping criterion. quit* if $\lambda^2/2 \leq \epsilon$.

3. *Line search.* Choose step size t by backtracking line search.

4. *Update.* $x := x + t\Delta x_{\text{nt}}$.

affine invariant, *i.e.*, independent of linear changes of coordinates:

Newton iterates for $\tilde{f}(y) = f(Ty)$ with starting point $y^{(0)} = T^{-1}x^{(0)}$ are

$$y^{(k)} = T^{-1}x^{(k)}$$

Classical convergence analysis

assumptions

- f strongly convex on S with constant m
- $\nabla^2 f$ is Lipschitz continuous on S , with constant $L > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2$$

(L measures how well f can be approximated by a quadratic function)

outline: there exist constants $\eta \in (0, m^2/L)$, $\gamma > 0$ such that

- if $\|\nabla f(x)\|_2 \geq \eta$, then $f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$
- if $\|\nabla f(x)\|_2 < \eta$, then

$$\frac{L}{2m^2} \|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2} \|\nabla f(x^{(k)})\|_2 \right)^2$$

damped Newton phase ($\|\nabla f(x)\|_2 \geq \eta$)

- most iterations require backtracking steps
- function value decreases by at least γ
- if $p^* > -\infty$, this phase ends after at most $(f(x^{(0)}) - p^*)/\gamma$ iterations

quadratically convergent phase ($\|\nabla f(x)\|_2 < \eta$)

- all iterations use step size $t = 1$
- $\|\nabla f(x)\|_2$ converges to zero quadratically: if $\|\nabla f(x^{(k)})\|_2 < \eta$, then

$$\frac{L}{2m^2} \|\nabla f(x^l)\|_2 \leq \left(\frac{L}{2m^2} \|\nabla f(x^k)\|_2 \right)^{2^{l-k}} \leq \left(\frac{1}{2} \right)^{2^{l-k}}, \quad l \geq k$$

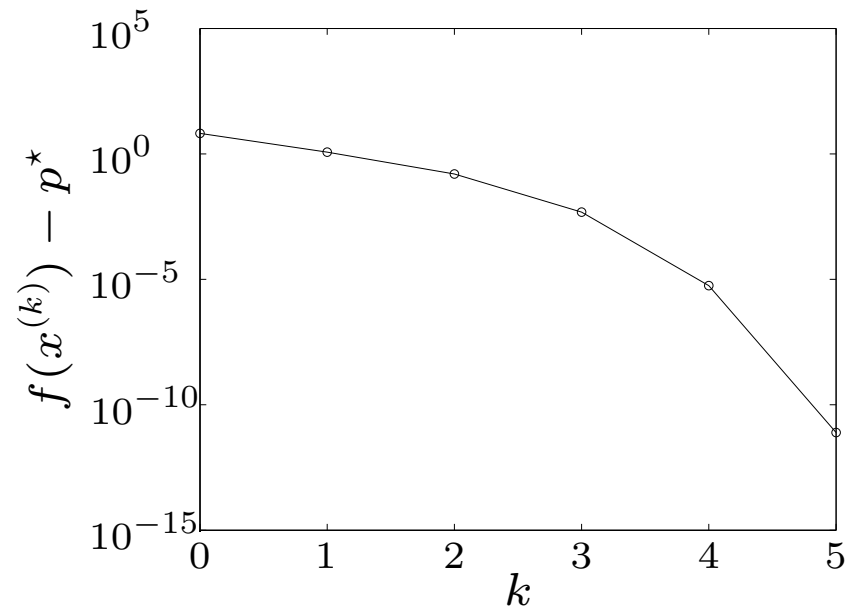
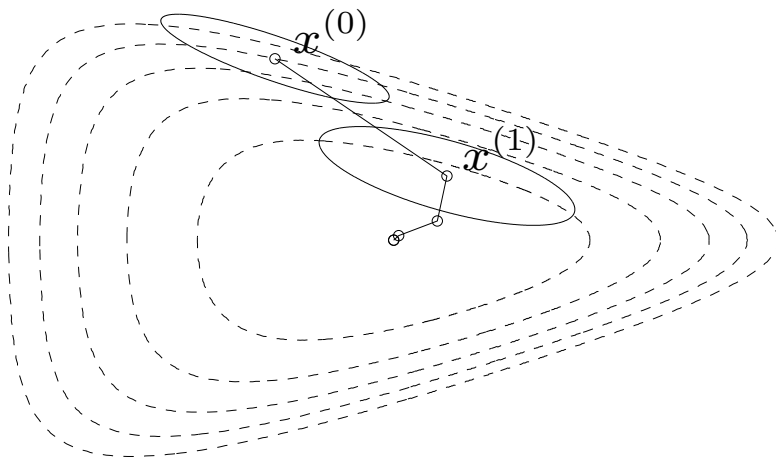
conclusion: number of iterations until $f(x) - p^* \leq \epsilon$ is bounded above by

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

- γ, ϵ_0 are constants that depend on $m, L, x^{(0)}$
- second term is small (of the order of 6) and almost constant for practical purposes
- in practice, constants m, L (hence γ, ϵ_0) are usually unknown
- provides qualitative insight in convergence properties (*i.e.*, explains two algorithm phases)

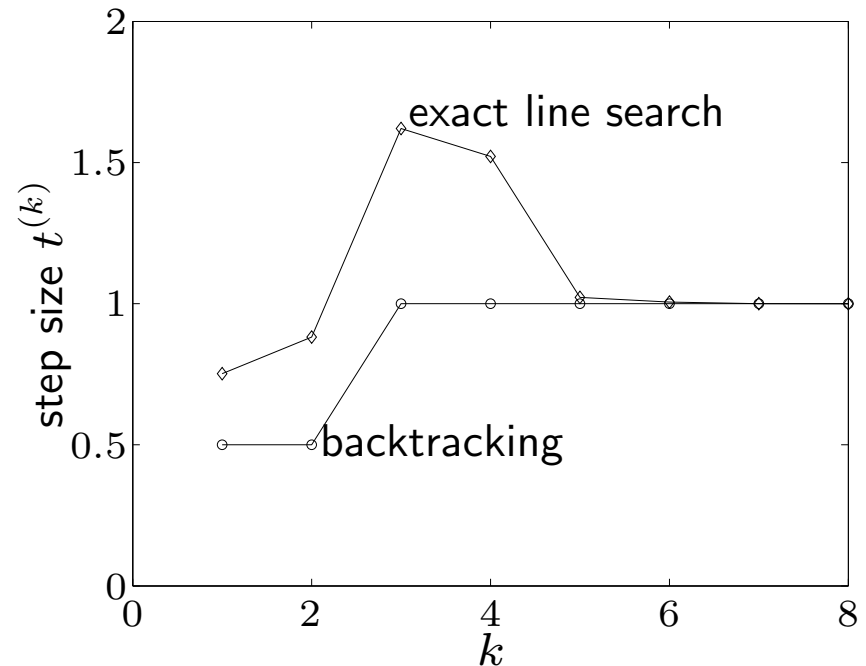
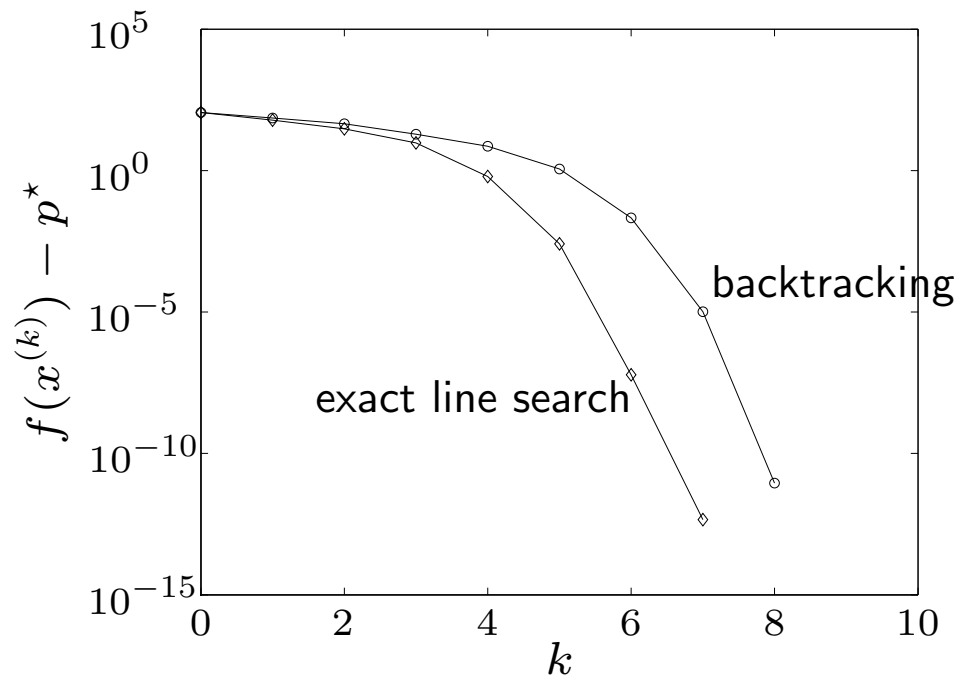
Examples

example in \mathbf{R}^2 (page 10–9)



- backtracking parameters $\alpha = 0.1$, $\beta = 0.7$
- converges in only 5 steps
- quadratic local convergence

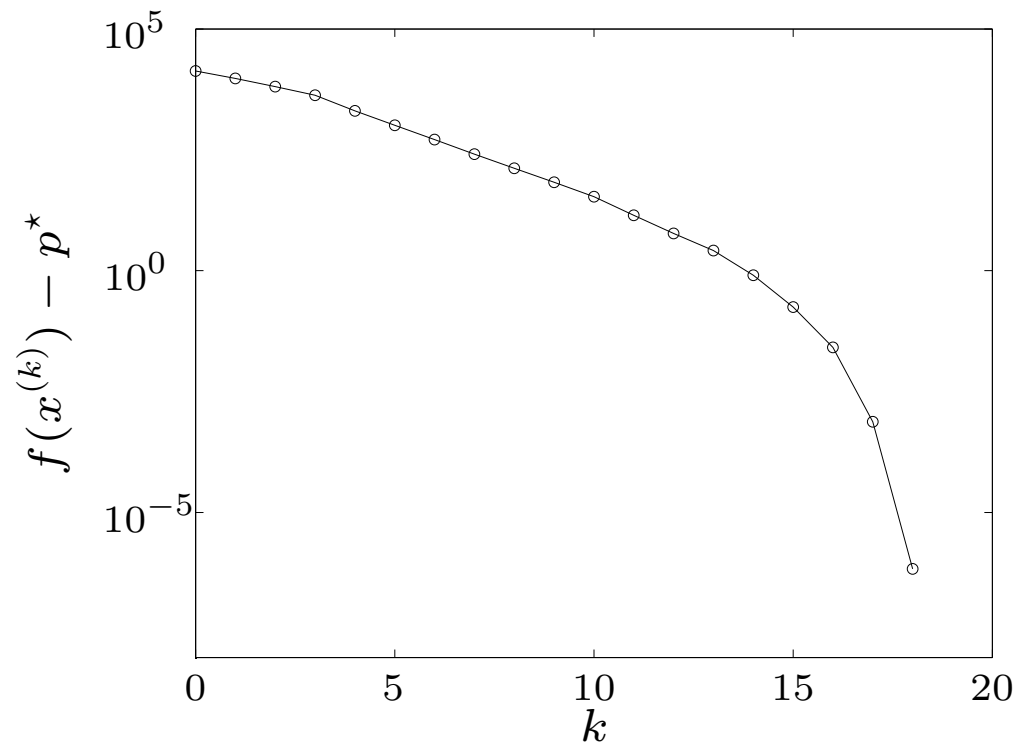
example in \mathbf{R}^{100} (page 10–10)



- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$
- backtracking line search almost as fast as exact l.s. (and much simpler)
- clearly shows two phases in algorithm

example in \mathbf{R}^{10000} (with sparse a_i)

$$f(x) = - \sum_{i=1}^{10000} \log(1 - x_i^2) - \sum_{i=1}^{100000} \log(b_i - a_i^T x)$$



- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$.
- performance similar as for small examples

Summary