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

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array} \qquad \begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

perturbed problem and its dual

$$\begin{array}{ll} \text{min.} & f_0(x) \\ \text{s.t.} & f_i(x) \leq u_i, \quad i = 1, \ldots, m \\ & h_i(x) = v_i, \quad i = 1, \ldots, p \end{array} \qquad \begin{array}{ll} \text{max.} & g(\lambda, \nu) - u^T\lambda - v^T\nu \\ \text{s.t.} & \lambda \succeq 0 \end{array}$$

- x is primal variable; u, v are parameters
- $p^\star(u, v)$ is optimal value as a function of u, v
- we are interested in information about $p^\star(u, v)$ 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 $\lambda^\star$, $\nu^\star$ are dual optimal for unperturbed problem

apply weak duality to perturbed problem:

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

sensitivity interpretation

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

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

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

proof (for $\lambda_i^\star$): from global sensitivity result, $u = t \cdot e_i$, $v = 0$

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

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

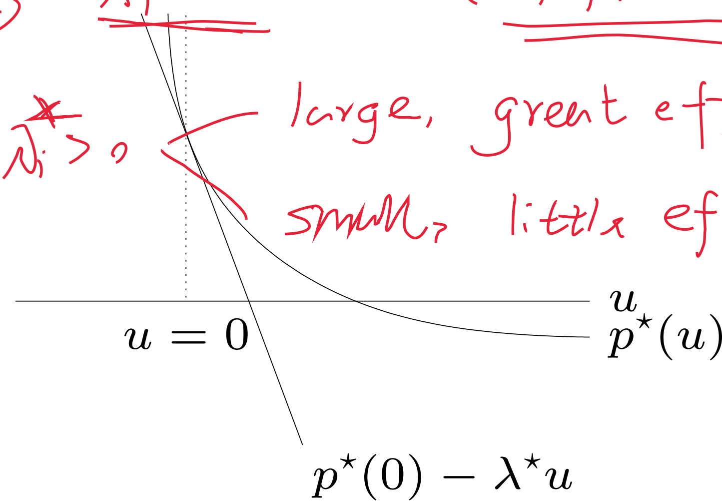
hence, equality

- $f_i(x^\star) < 0$, inactive $\Rightarrow$ $\lambda_i^\star = 0$ $\quad (\lambda_i^\star \cdot f_i(x^\star) = 0)$
- $f_i(x^\star) = 0$, active $\Rightarrow$ $\lambda_i^\star > 0$

large, great effect

small, little effect

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



$p^\star(u) \geq p^\star(0) - \lambda^\star u$

$$\begin{tikzcd} P \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & D \arrow[d, leftrightarrow] \\ \tilde{P} \arrow[r, leftrightarrow] & \tilde{D} \end{tikzcd}$$

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

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

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

reformulated problem and its dual

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

dual function follows from

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

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

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

$l_\infty \leftrightarrow l_1$

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

Problems with generalized inequalities

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

$\preceq_{K_i}$ 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 $L : \mathbf{R}^n \times \mathbf{R}^{k_1} \times \cdots \times \mathbf{R}^{k_m} \times \mathbf{R}^p \to \mathbf{R}$, is defined as

$$L(x, \lambda_1, \cdots, \lambda_m, \nu) = f_0(x) + \sum_{i=1}^{m} \underbrace{\lambda_i^T}_{\in R^{k_i}} f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

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

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

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

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

$K^* = \{y \mid y^T x \geq 0,\ x \in K\}$

$$
\begin{aligned}
f_0(\tilde{x}) &\geq f_0(\tilde{x}) + \sum_{i=1}^{m} \underbrace{\overset{\succeq_{K^*} 0}{\lambda_i^T} \overset{\preceq_{K_i} 0}{f_i(\tilde{x})}}_{\leq 0} + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \\
&\geq \inf_{x \in \mathcal{D}} L(x, \lambda_1, \ldots, \lambda_m, \nu) \\
&= g(\lambda_1, \ldots, \lambda_m, \nu)
\end{aligned}
$$

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

dual problem

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

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

$\langle a, b\rangle = a^T b$

$\langle A, B\rangle = \mathrm{Tr}(A^T B)$

Semidefinite program

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

$K^* = K = (\mathbf{S}^n_+)$

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

- 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

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

dual SDP

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

$p^\star = d^\star$ 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

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

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

unconstrained minimization methods

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

$$f(x^{(k)}) \to p^\star$$

- can be interpreted as iterative methods for solving optimality condition

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

optimality condition.

Initial point and sublevel set

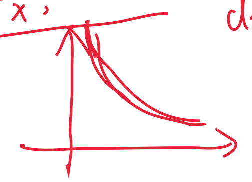
algorithms in this chapter require a starting point $x^{(0)}$ such that

- $x^{(0)} \in \mathbf{dom}\, f$
- 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 $f(x) \to \infty$ as $x \to \mathbf{bd}\, \mathbf{dom}\, f$

$f(x) = \frac{1}{x}, (x > 0)$ $\quad$ $\mathrm{dom} f = (0, +\infty)$



examples of differentiable functions with closed sublevel sets: $\quad -\log(x)$

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

$\mathrm{dom} f = R^n$ $\qquad$ $\mathrm{dom} f = \{x \mid b_i - a_i^T x > 0, \forall i\}$

$f: R^n \to R$ $\nabla f \in R^n$ $\nabla^2 f \in R^{n \times n}$. Hessen matrix (S^n)

$x \in R^n$

Strong convexity and implications

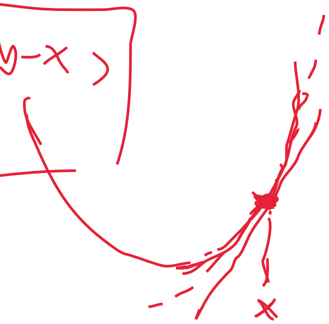
f is strongly convex on S if there exists an $m > 0$ such that $\left(f(x) = \frac{1}{x},\ x > 0\right)$ ×

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

$(K = S^n_+)$ convex: $f(y) \geq f(x) + \nabla f^T(x)(y-x)$

implications

$f(y) = f(x) + \nabla f^T(x)(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$ $(z \in [x, y])$



- 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 $x \in S$,

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

$\hat{f}(x) \approx f(x) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(a)(x-a)^n$

  useful as stopping criterion (if you know m)

$x \in R^n$

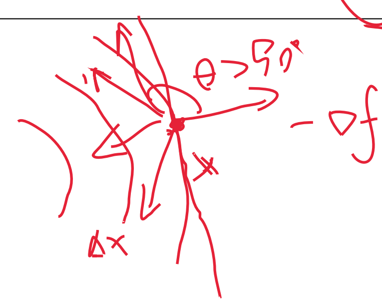
Descent methods

f: convex

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

- 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* $(t>0)$
- from convexity, $f(x^+) < f(x)$ implies $\nabla f(x)^T \Delta x < 0$
  (*i.e.*, Δx is a *descent direction*)

$$f(x^+) \geq f(x) + \boxed{\nabla f^T(x) \underbrace{(x^+ - x)}_{t \cdot \Delta x}} < 0$$



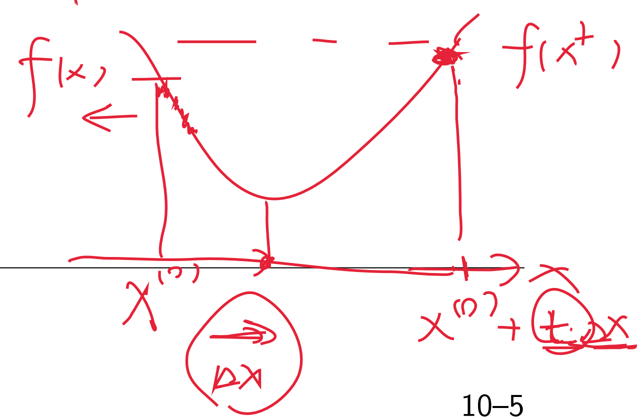
$f(x) = x^2$

General descent method.

given a starting point $x \in \mathbf{dom}\, f$.
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. ← $\|\nabla f\|_2$ small



$\min_x f(x)$

Line search types

$f(x): \mathbb{R}^n \to \mathbb{R}$

$\tilde{f}(t): \mathbb{R} \to \mathbb{R}$

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

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

$(\alpha = 0.1, 0.01, \ \beta = \frac{1}{2})$

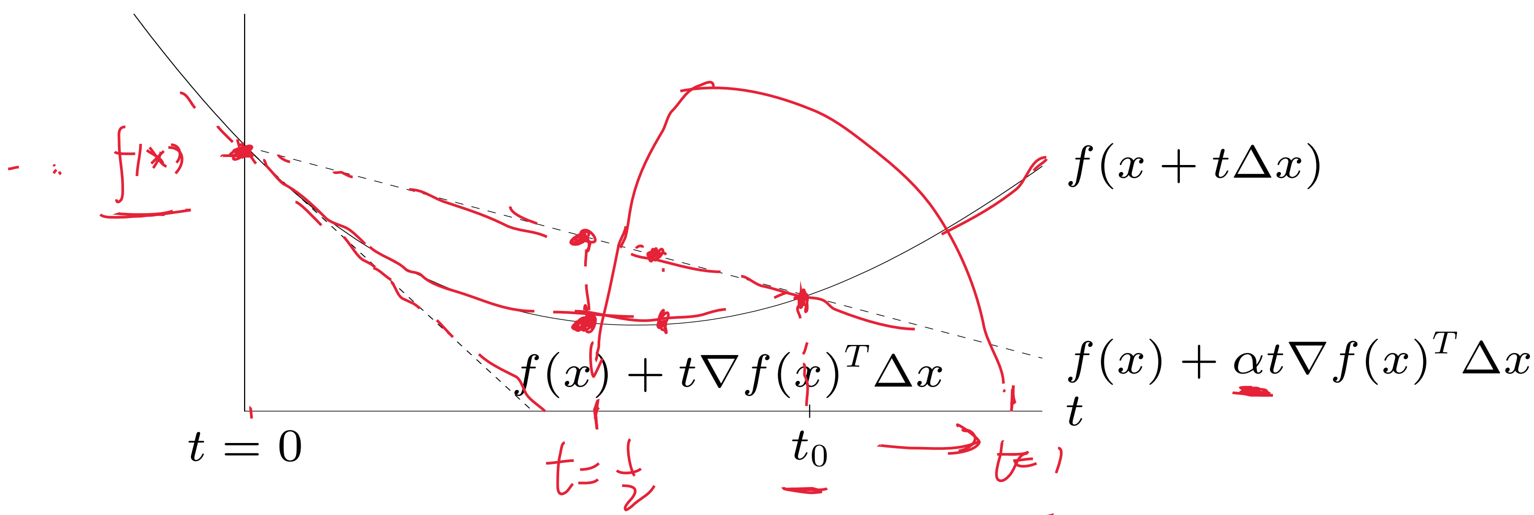
- starting at $t = 1$, repeat $t := \beta t$ until

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

$x^+ = x + t \cdot \Delta x$

$f(x + t\Delta x) \geq \hat{f}(x + t\Delta x) = f(x) + t \nabla f(x)^T \Delta x$

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



$f(x + t_0 \Delta x) = f(x) + \alpha t_0 \nabla f(x)^T \Delta x$

(GD)

Gradient descent method

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

---

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

$$\log \quad f(x^{(k)}) - p^\star \leq c^k (f(x^{(0)}) - p^\star)$$

k log c

linear - convergence

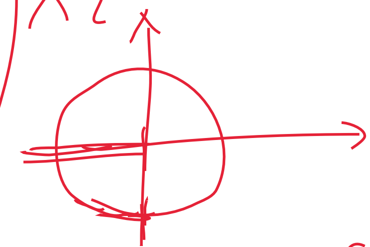
  $c \in (0,1)$ depends on m, $x^{(0)}$, line search type
- very simple, but often very slow; rarely used in practice

$$\underbrace{mI \preceq \nabla^2 f \preceq MI}_{K = S_{+}^{n}} \quad \Big| \quad \{ K(\nabla^2 f) \le \frac{M}{m} \}$$

$f: R^2 \to R$

$f(x) = \|x\|_2^2 \le 1$

$K = 1$



$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$$

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

$f(x) = x^T P x \le 1$

$$K = \frac{\lambda_{max}}{\lambda_{min}}$$

GD $\quad \Delta x = -\nabla f(x), \quad x^+ = x + t\Delta x$

$$= x - t\nabla f(x)$$

$$\|\nabla f(x)\|_2^2$$

$$f(x^+) \le f(x) - t\nabla f(x)^T \nabla f(x) + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$$

$$f(x^+) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$$

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

$$\left(1 - \frac{m}{M}\right)$$

Exact l.s.: $\quad t_{exact} = \arg\min_{t \ge 0} f(x - t\nabla f(x))$

$$f(x^+) = \tilde{f}(t_{exact}) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$$

$$k \le \frac{\log((f(x^{(0)}) - p^*)/\epsilon)}{\log(1/c)} \simeq \frac{M}{m} \downarrow$$

$$f(x^+) - p^* \le (f(x) - p^*) - \frac{1}{2M}\|\nabla f\|_2^2$$

$$f(x^+) - p^* \le \underbrace{\left(1 - \frac{m}{M}\right)}_{c}(f(x) - p^*)$$

$\nabla^2 f(x)$ { $K \uparrow$, #iters $\uparrow$; $K \downarrow$, #iters $\downarrow$ }

quadratic problem in $\mathbf{R}^2$

$$f(x) = \frac{1}{2}\underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{P}$$

$$\kappa(P) = \begin{cases} \gamma, & \gamma \geq 1 \\ \frac{1}{\gamma}, & \gamma < 1 \end{cases} = \max\{\gamma, \tfrac{1}{\gamma}\}$$

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

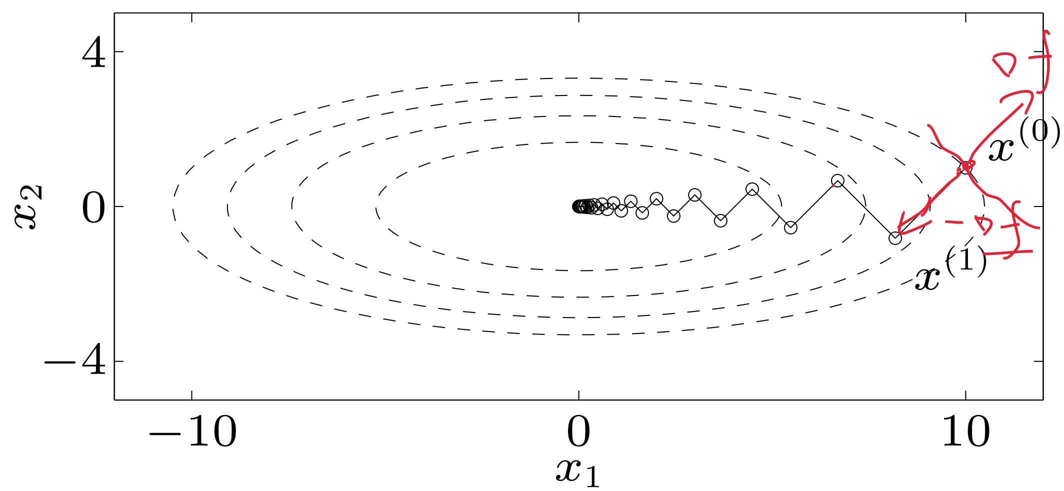
with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

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

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

$\gamma = 1 = \kappa$

$\kappa = 10$, $\gamma = 10{,}000 = \kappa$



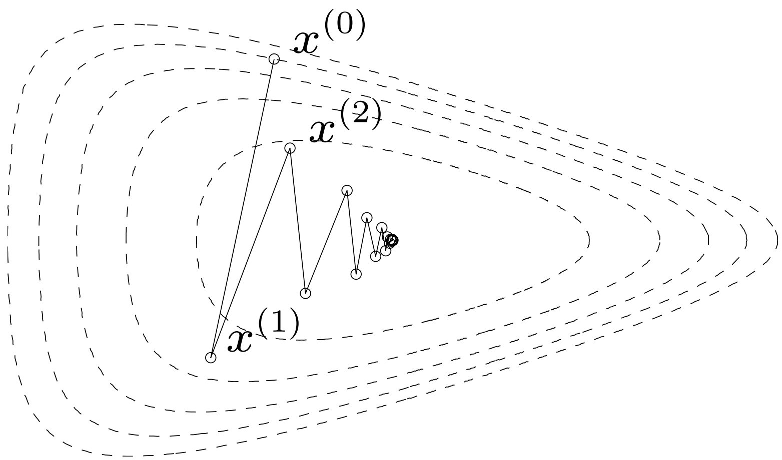
GD + exact l.s.

$x^+ = x - t\nabla f$

nonquadratic example R^2

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

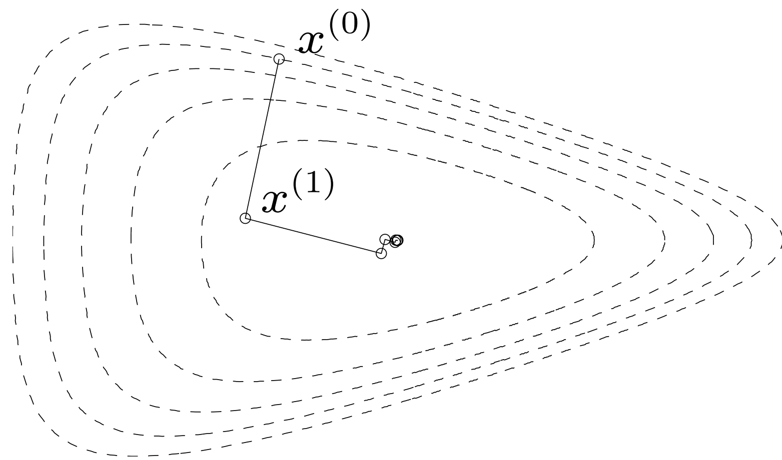
$x^+ = x - t\nabla f(x)$



backtracking line search

#iter = 20 : $\varepsilon \approx 10^{-7}$

(0.4)



exact line search

#iter = 15, $\varepsilon \approx 10^{-11}$

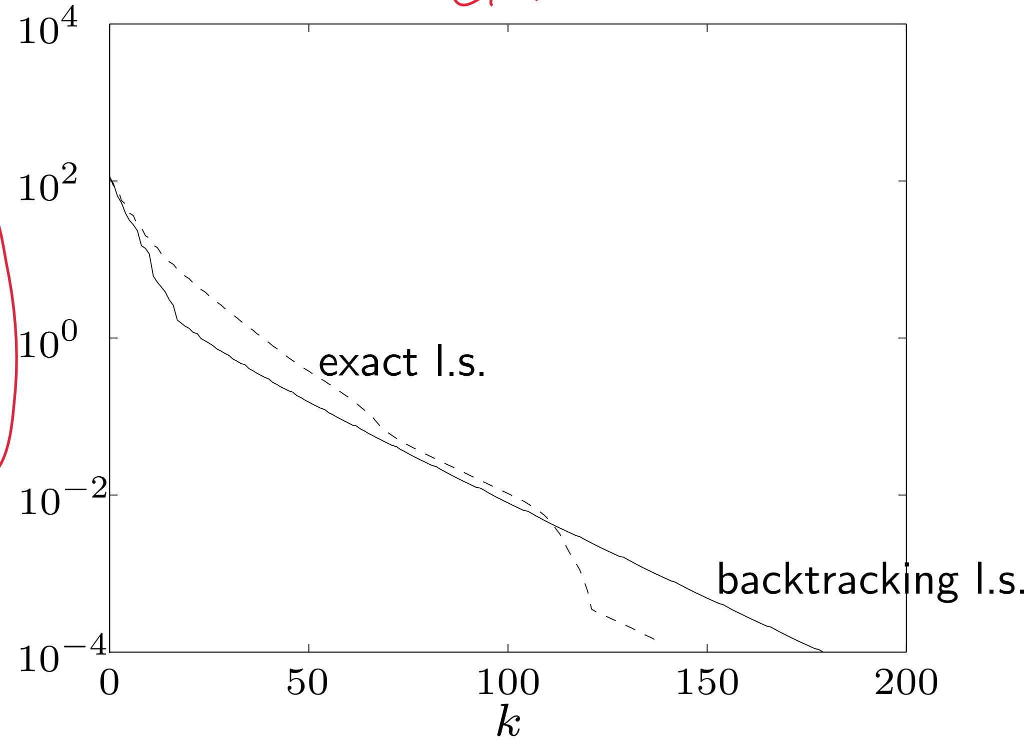
(0.2)

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

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

GD

log-scale



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

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

$(x^+ = x + v)$

$f(\overset{x^+}{x+v}) < f(x)$

$\boxed{\min_v \nabla f(x)^T v \quad \text{s.t. } \|v\| = 1} \Rightarrow \Delta x_{nsd} = \operatorname{argmin} \nabla f(x)^T v$, $\|v\| = 1$ $(\|v\| \le 1)$

Steepest descent method

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

$\|z\|_* = \{\sup_z z^T x \mid \|x\| \le 1\}$

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

$v \in \mathbb{R}^n$

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

$\nabla f(x)^T \cdot \Delta x_{nsd} = -\|\nabla f(x)\|_*$

(unnormalized) steepest descent direction

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

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

$\nabla f(x)^T v$

$= \min_{\|v\| \le 1} \langle \nabla f(x), v \rangle$

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

$= -\max_{\|v\| \le 1} \langle -\nabla f(x), v \rangle = -\|-\nabla f(x)\|_*$

$= -\|\nabla f(x)\|_*$

steepest descent method

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

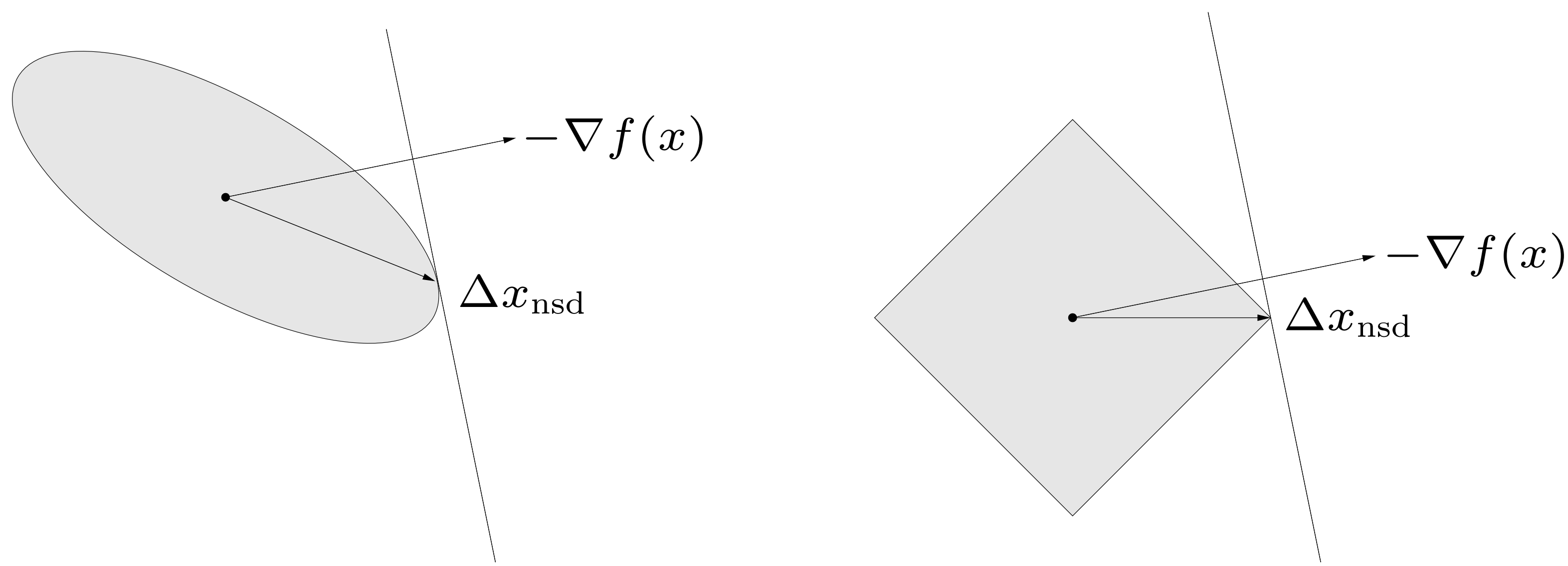
examples

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

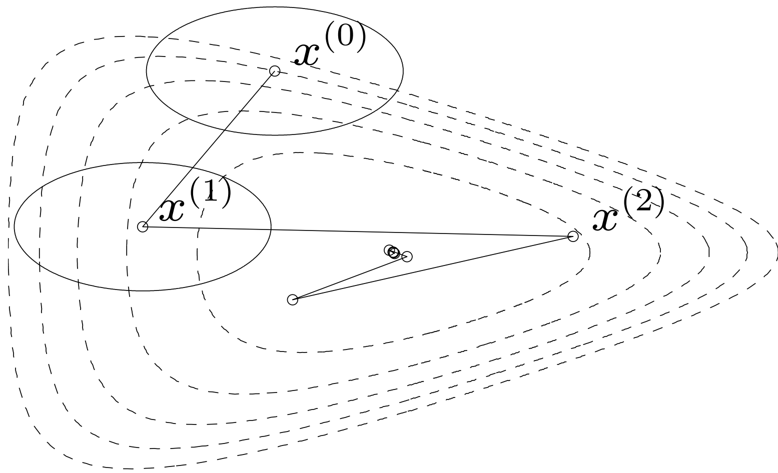
examples

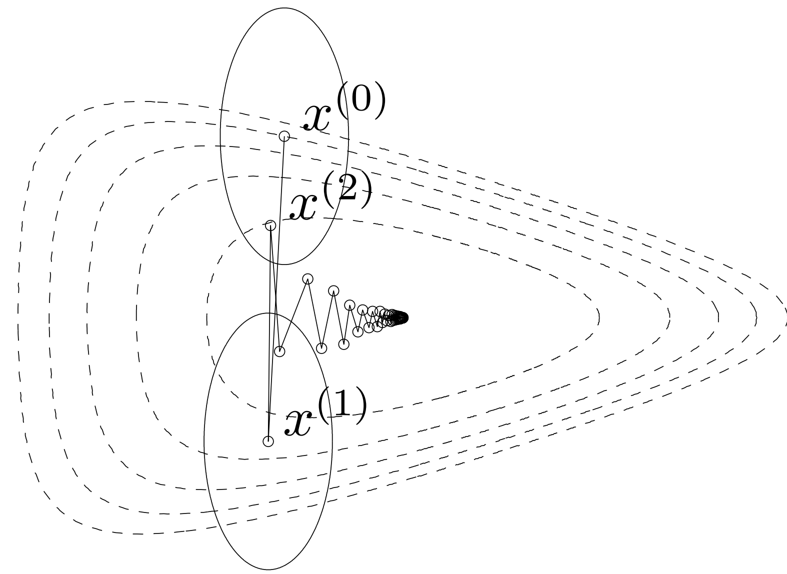
- Euclidean norm: $\Delta x_{\mathrm{sd}} = -\nabla f(x)$
- quadratic norm $\|x\|_P = (x^T P x)^{1/2}$ ($P \in \mathbf{S}^n_{++}$): $\Delta x_{\mathrm{sd}} = -P^{-1}\nabla f(x)$
- ℓ_1-norm: $\Delta x_{\mathrm{sd}} = -(\partial f(x)/\partial x_i)e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

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

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

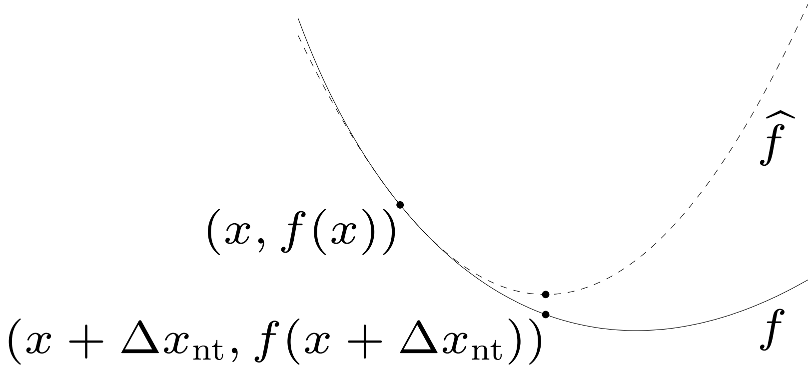
interpretations

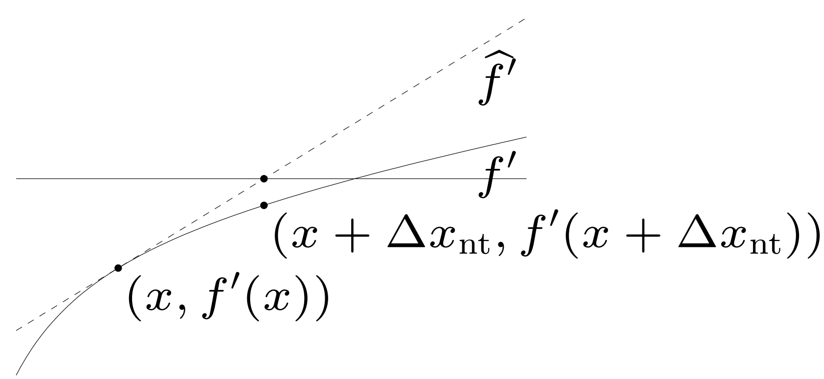
- $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

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

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

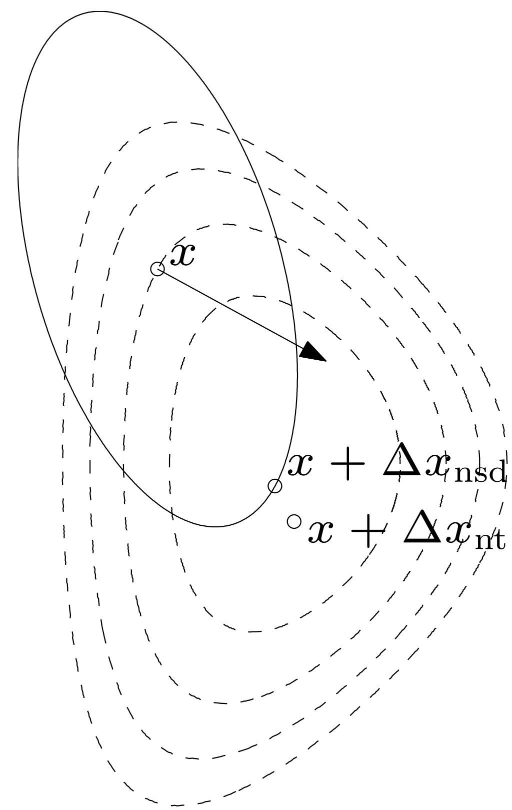
$$\nabla f(x+v) \approx \nabla \widehat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$





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

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



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 $\widehat{f}$:

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

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 $x^\star$

properties

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

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

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

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

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

Newton's method

---

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

1. *Compute the Newton step and decrement.*
   $\Delta x_{\mathrm{nt}} := -\nabla^2 f(x)^{-1}\nabla f(x); \quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1}\nabla f(x).$
2. *Stopping criterion.* **quit** if $\lambda^2/2 \le \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\mathrm{nt}}$.

---

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 $p^\star > -\infty$, this phase ends after at most $(f(x^{(0)}) - p^\star)/\gamma$ 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

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

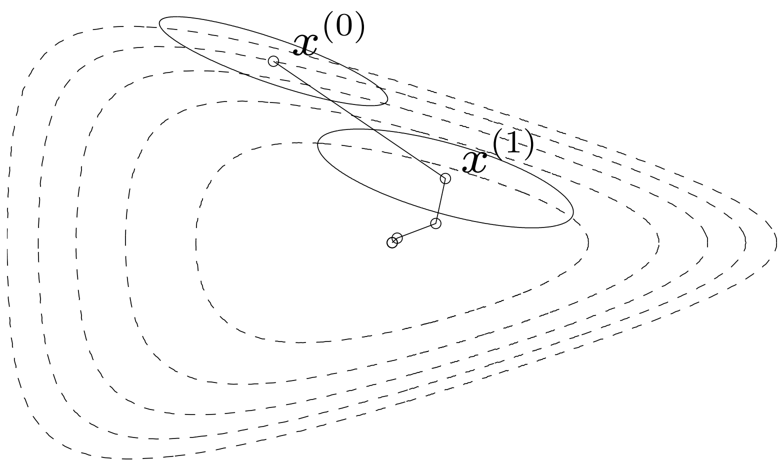
conclusion: number of iterations until $f(x) - p^\star \leq \epsilon$ is bounded above by

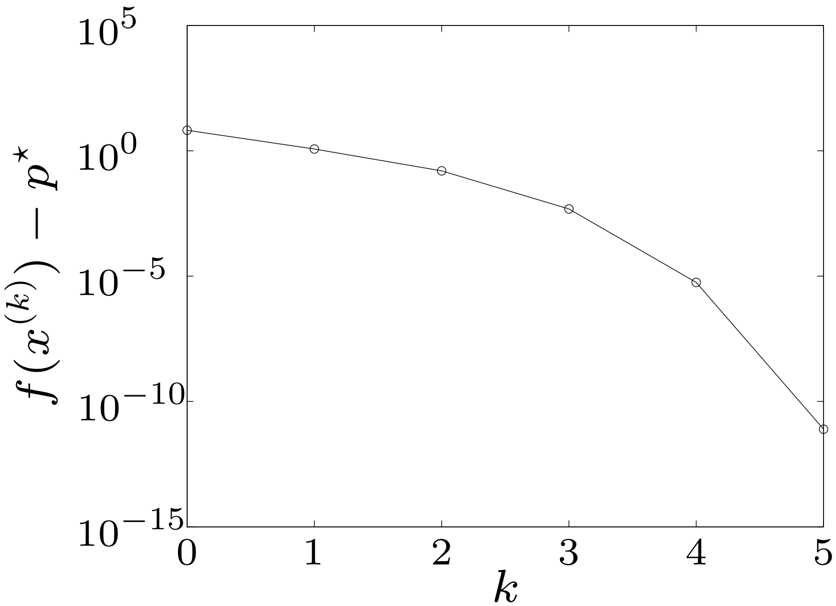
$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

- 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

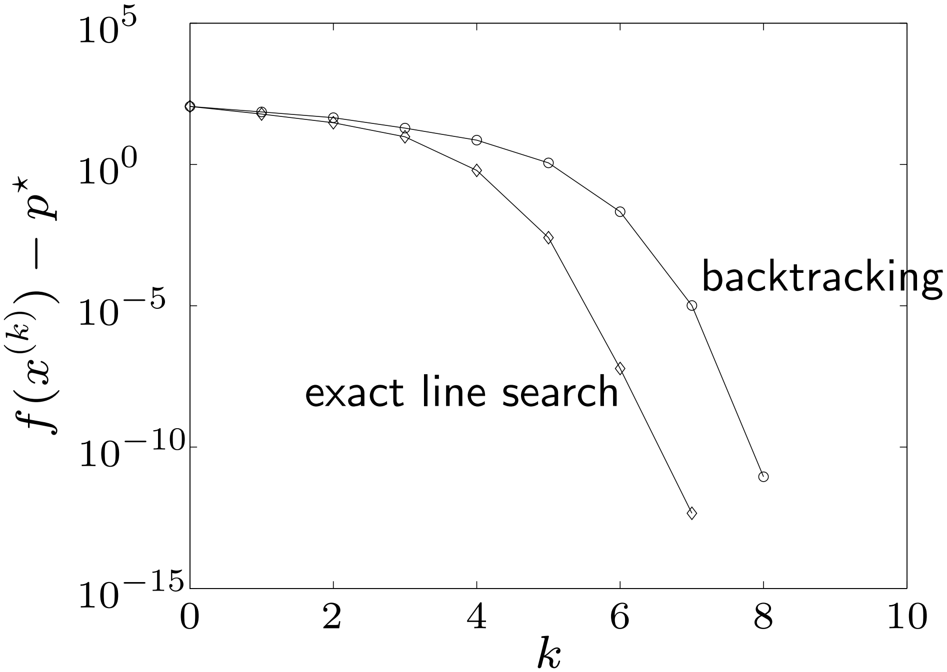
example in $\mathbf{R}^2$ (page 10–9)

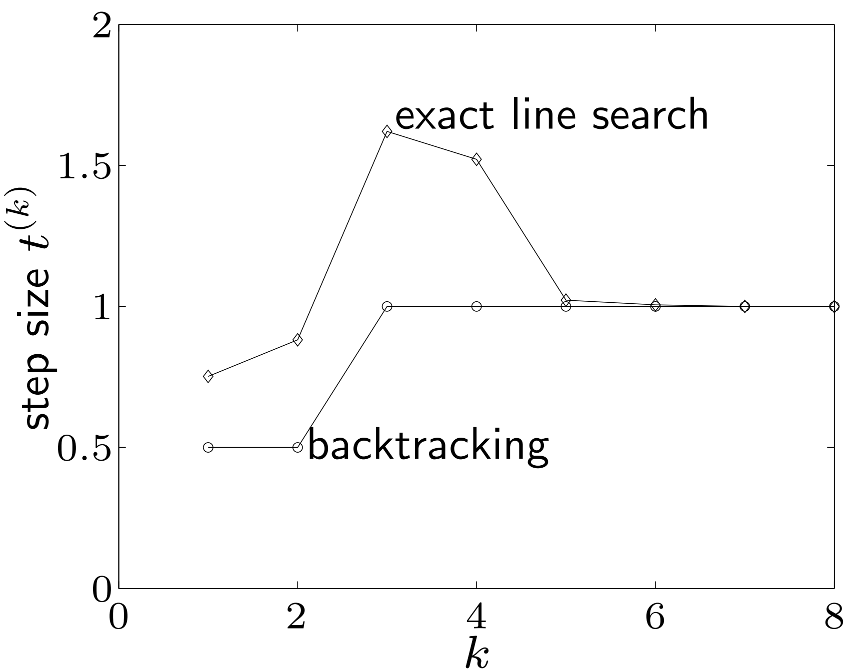




- backtracking parameters $\alpha = 0.1$, $\beta = 0.7$
- converges in only 5 steps
- quadratic local convergence

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

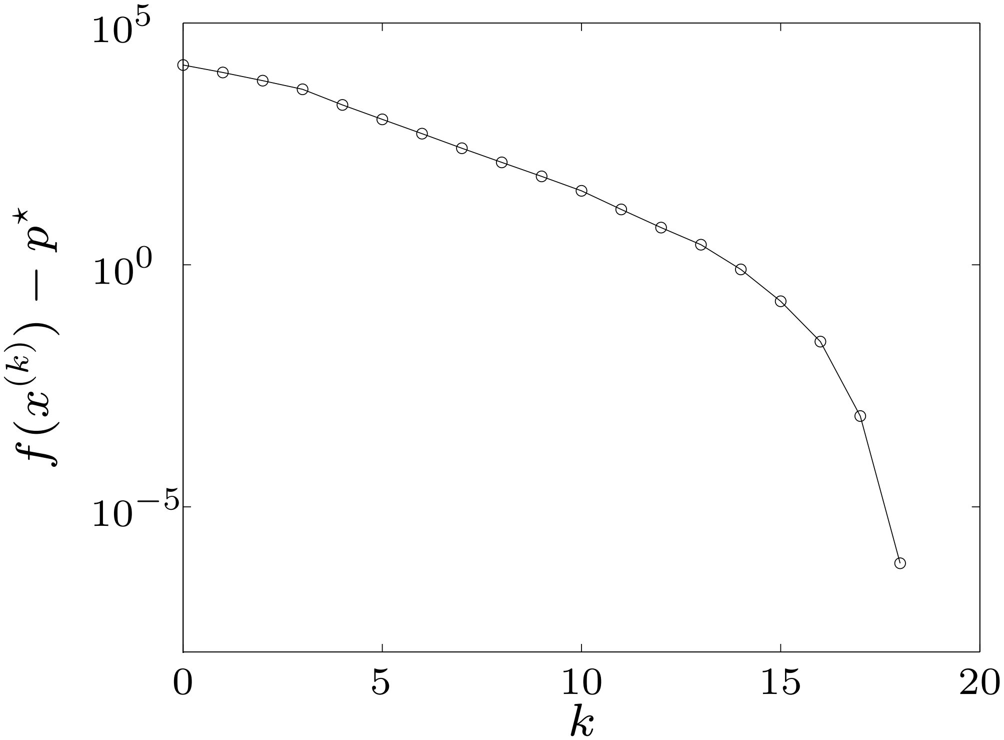




- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$
- backtracking line search almost as fast as exact l.s. (and much simpler)
- clearly shows two phases in algorithm

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