# Optimisation Combinatoire et Convexe

## Interior Point Methods

# Today

**Interior point methods.**

- Unconstrained minimization
- Barrier method
- Primal dual methods

# Unconstrained minimization

# 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 $\textbf{dom}\, f$ open)
- we assume optimal value $p^\star = \inf_x f(x)$ is attained (and finite)

**unconstrained minimization methods**

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

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

- can be interpreted as iterative methods for solving optimality condition

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

# 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 $\mathbf{dom}\, f = \mathbb{R}^n$
- true if $f(x) \to \infty$ as $x \to \mathbf{bd}\, \mathbf{dom}\, f$

examples of differentiable functions with closed sublevel sets:

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

# Strong convexity and implications

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

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

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

*General descent method.*

**given** a starting point $x \in \mathbf{dom}\, f$.
**repeat**
1. Determine a descent direction $\Delta x$.
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

**until** stopping criterion is satisfied.

## Line search types

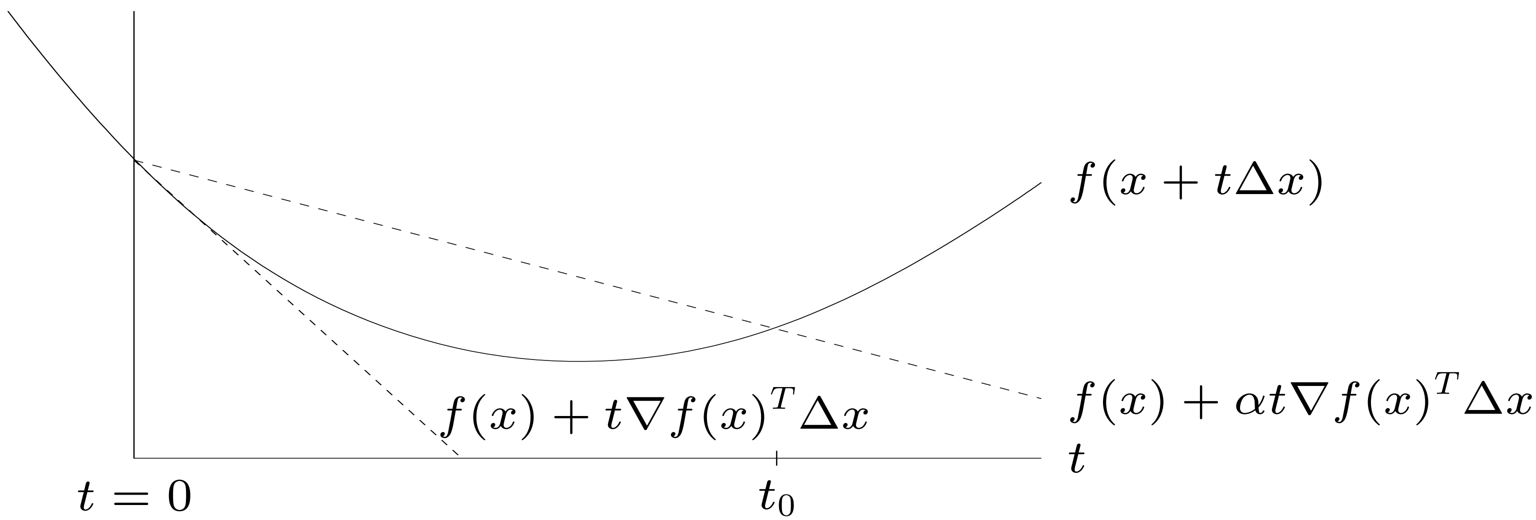
**exact line search:** $t = \mathrm{argmin}_{t>0} f(x + t\Delta x)$

**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 $t \le t_0$



# 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 $f$,

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

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

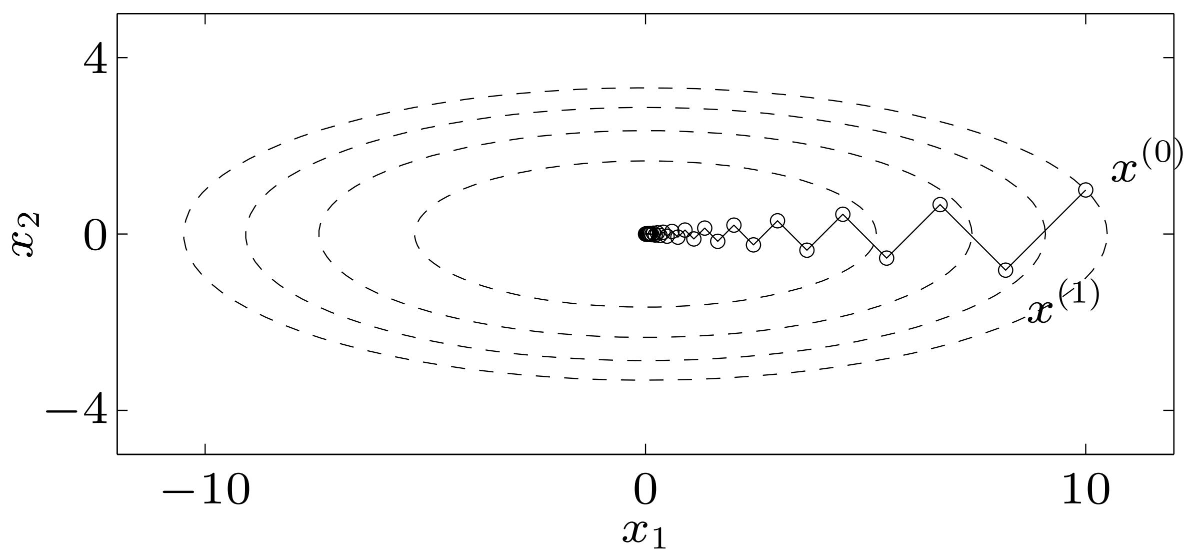
## quadratic problem in $\mathbb{R}^2$

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

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

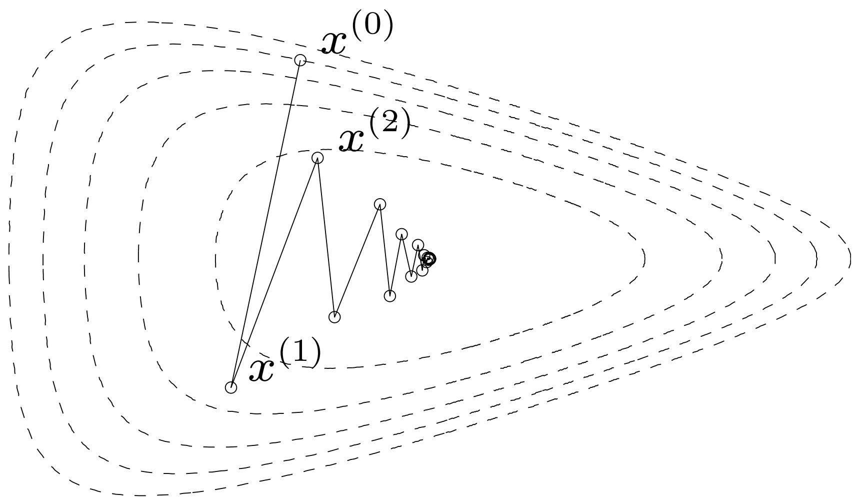
$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1}\right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1}\right)^k$$

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

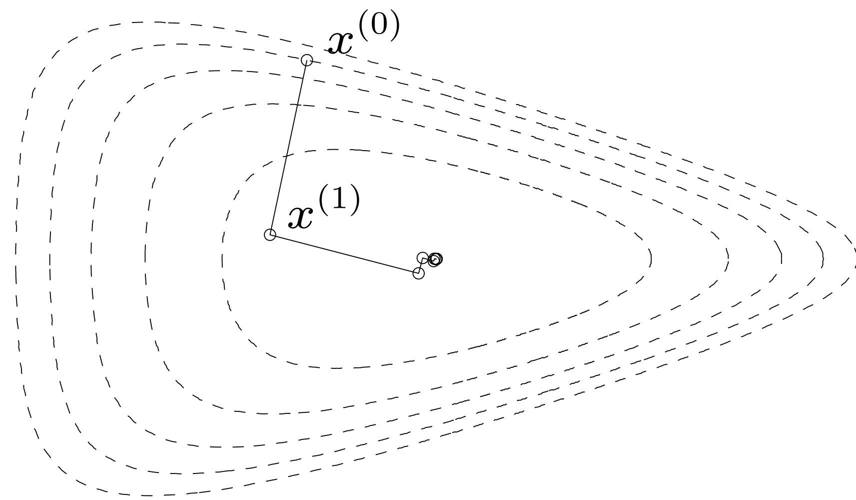


## nonquadratic example

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



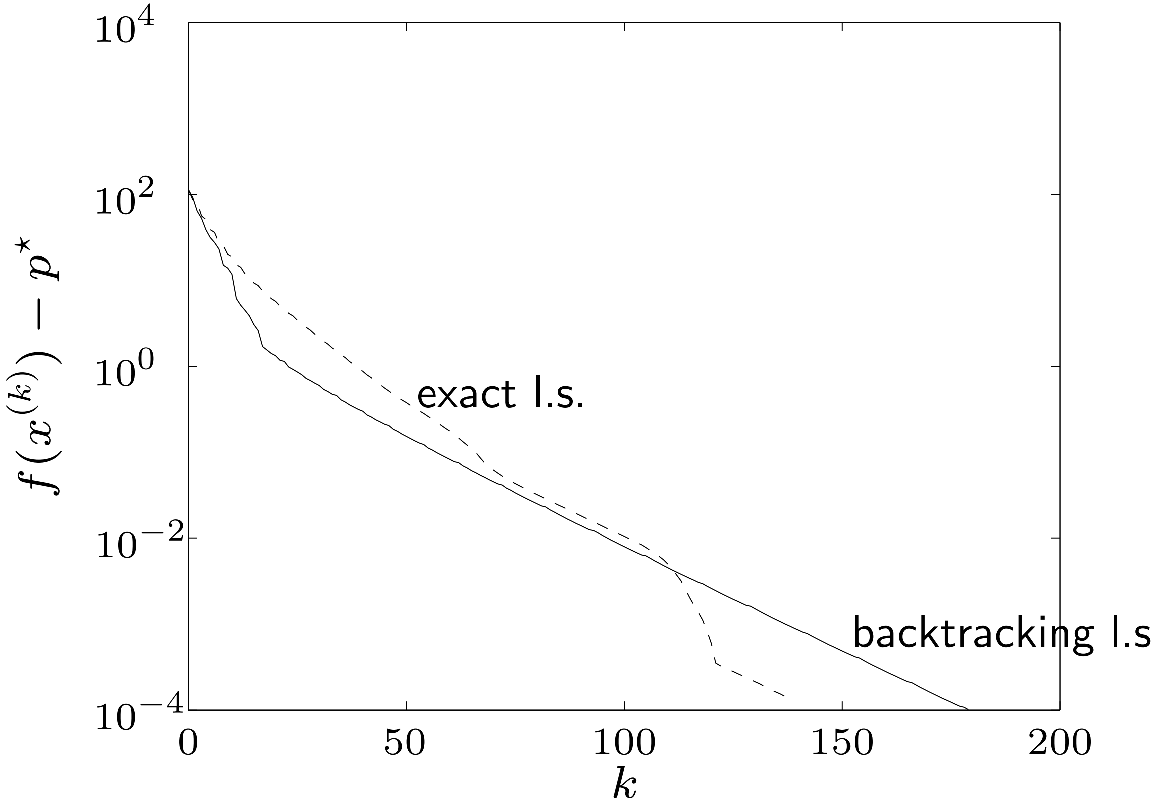

backtracking line search




exact line search

## a problem in $\mathbb{R}^{100}$

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




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

# Steepest descent method

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

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

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

**(unnormalized) steepest descent direction**

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

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

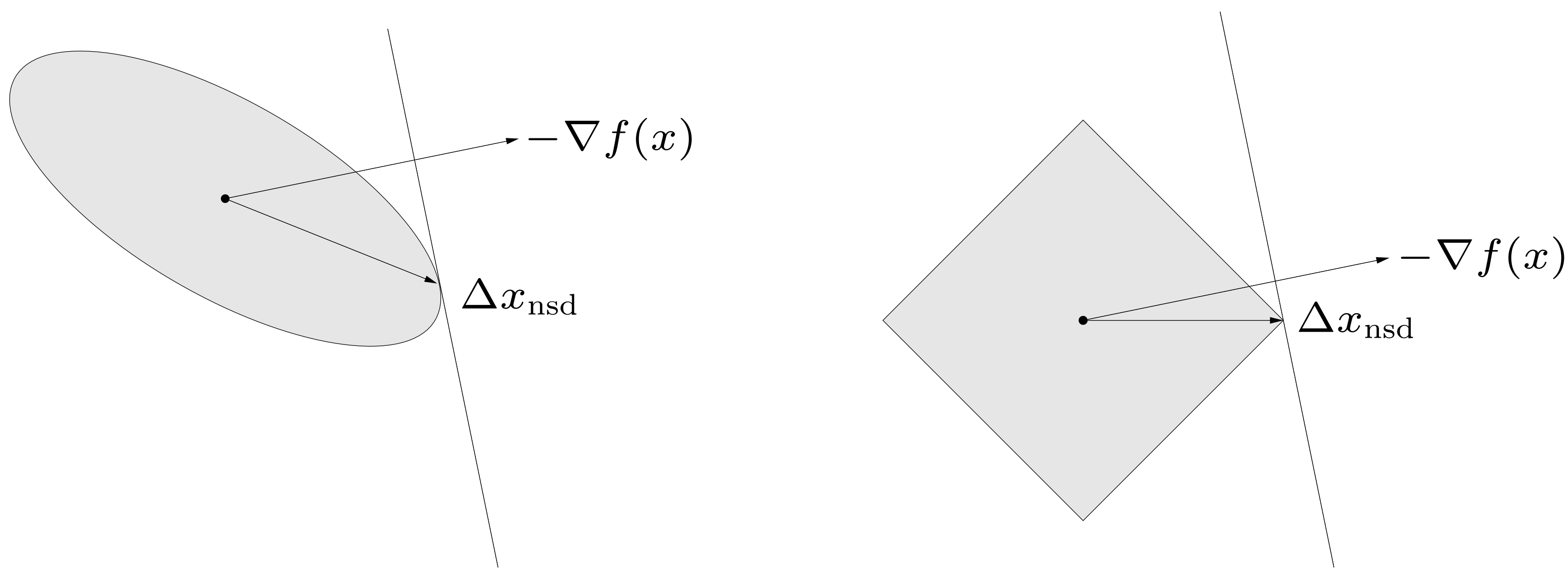
**steepest descent method**

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

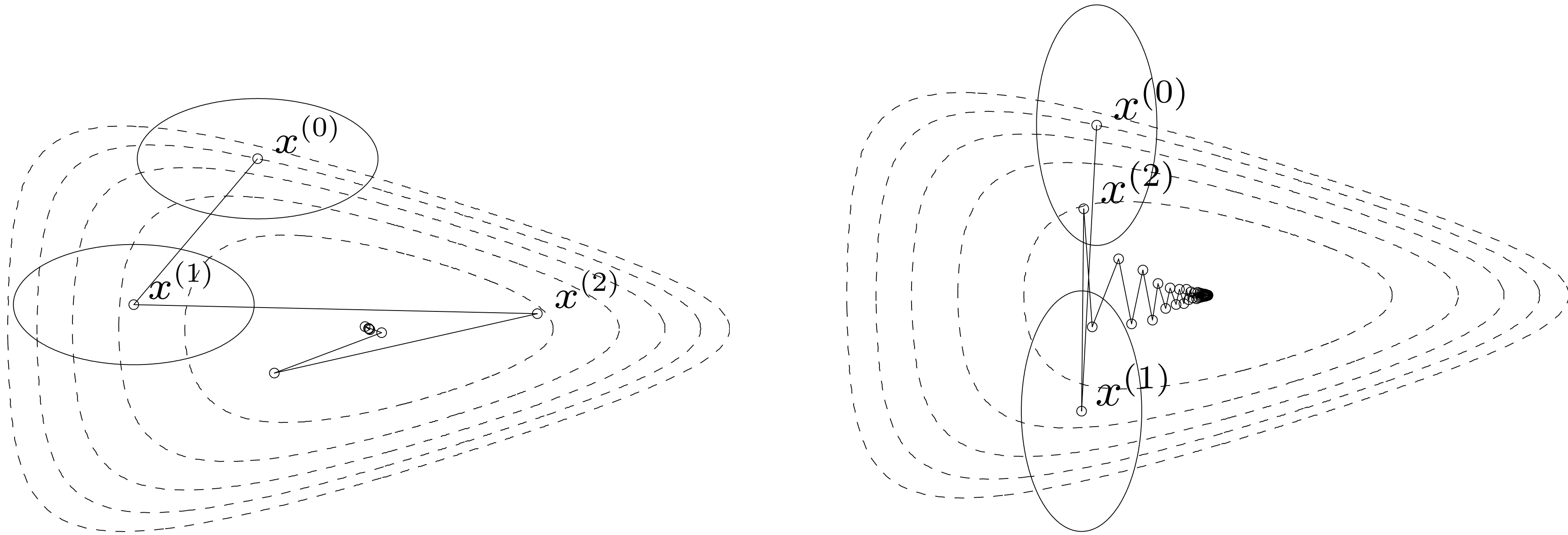
## examples

- Euclidean norm: $\Delta x_{\rm sd} = -\nabla f(x)$
- quadratic norm $\|x\|_P = (x^T P x)^{1/2}$ ($P \in \mathbf{S}^n_{++}$): $\Delta x_{\rm sd} = -P^{-1}\nabla f(x)$
- $\ell_1$-norm: $\Delta x_{\rm sd} = -(\partial f(x)/\partial x_i) e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

unit balls and normalized steepest descent directions for a quadratic norm and the $\ell_1$-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

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

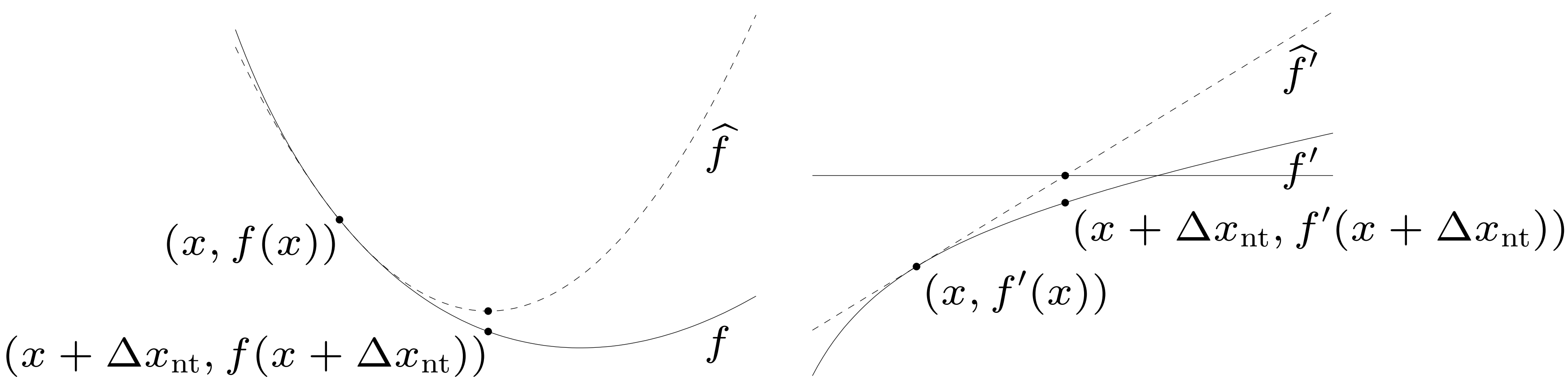
**interpretations**

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

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

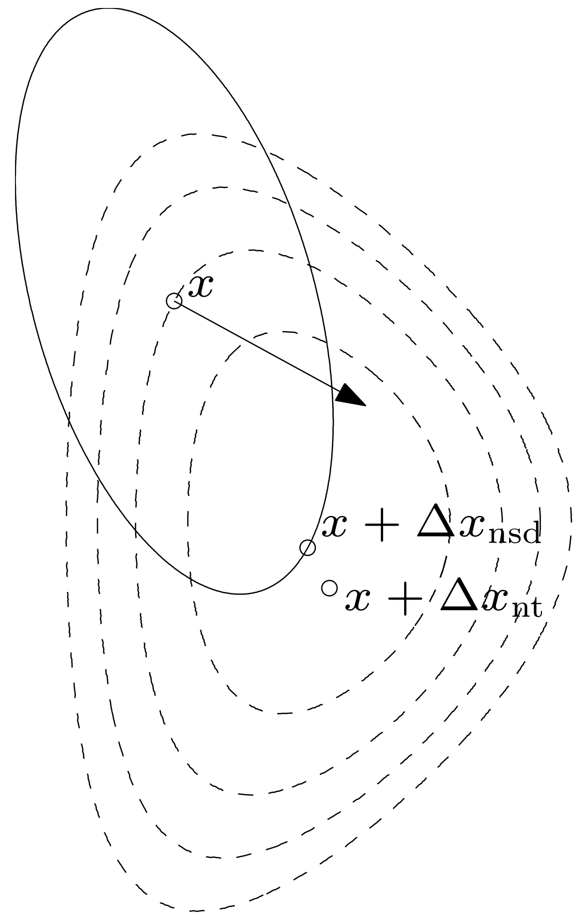
- $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition

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



- $\Delta x_{\mathrm{nt}}$ 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 $\widehat{f}$:

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

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

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

- 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 $\lambda^2/2 \leq \epsilon$.
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

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

# Classical convergence analysis

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

- most iterations require backtracking steps
- function value decreases by at least $\gamma$
- 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

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

# Classical convergence analysis

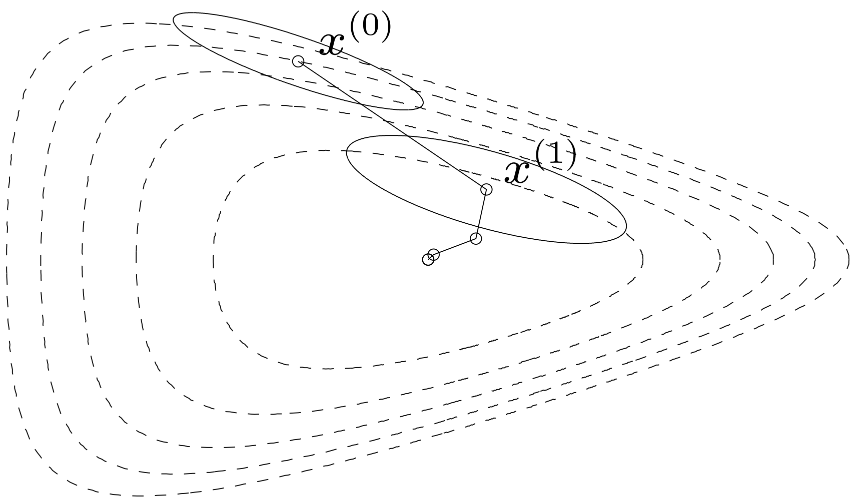
**conclusion:** number of iterations until $f(x) - p^\star \leq \epsilon$ is bounded above by

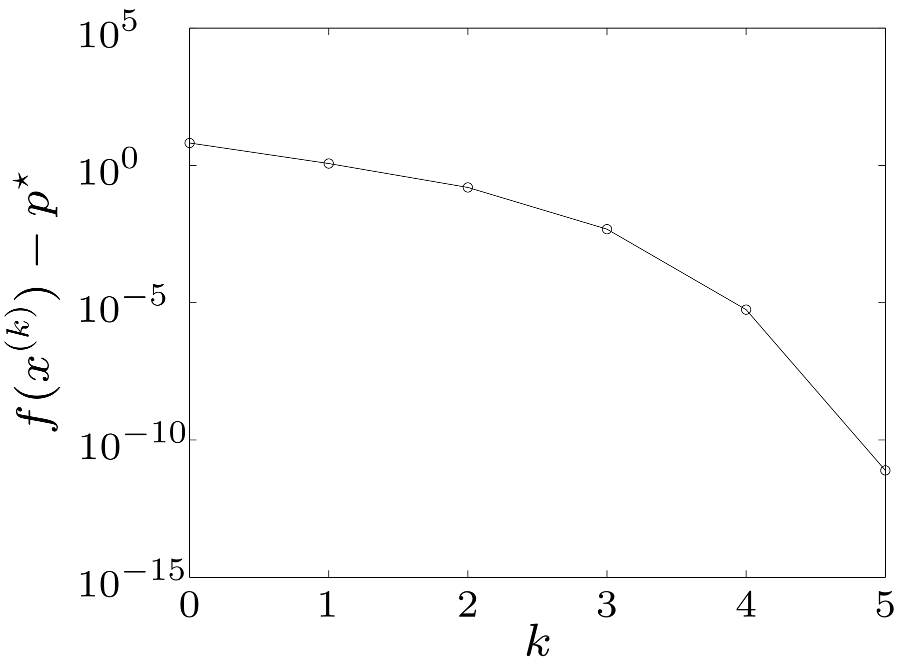
$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

- $\gamma$, $\epsilon_0$ 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 $\gamma$, $\epsilon_0$) are usually unknown
- provides qualitative insight in convergence properties ($i.e.$, explains two algorithm phases)

# Examples

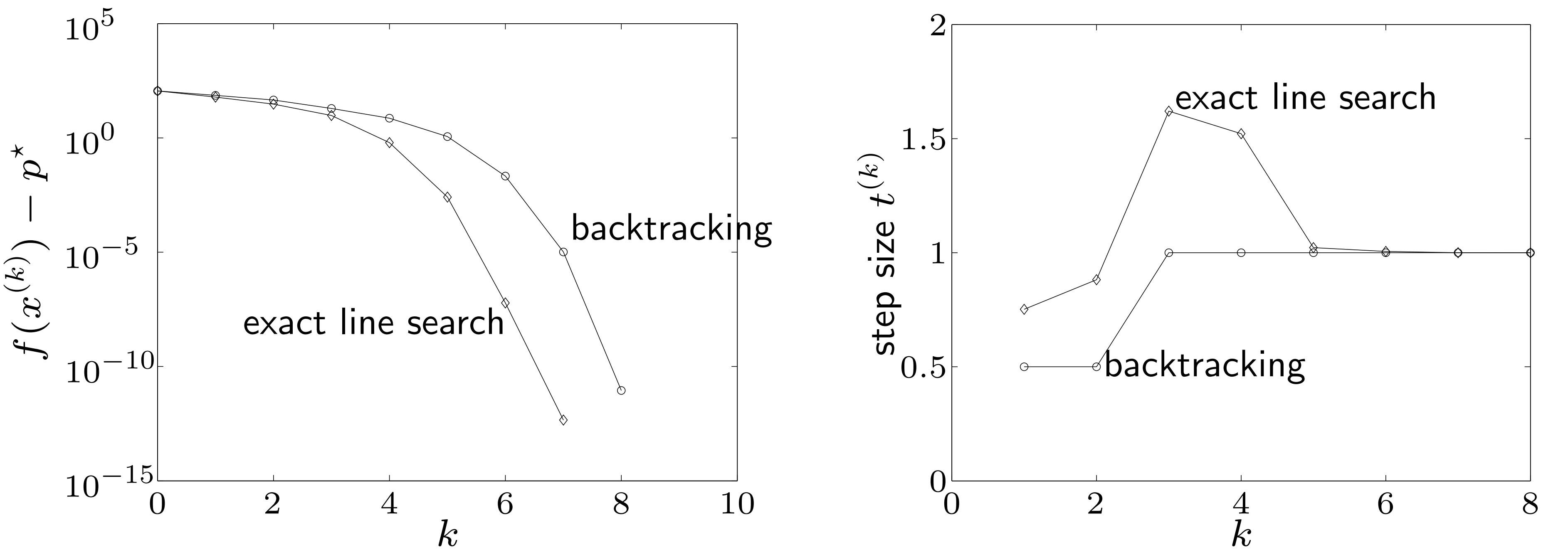
**example in** $\mathbb{R}^2$ (page 12)




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

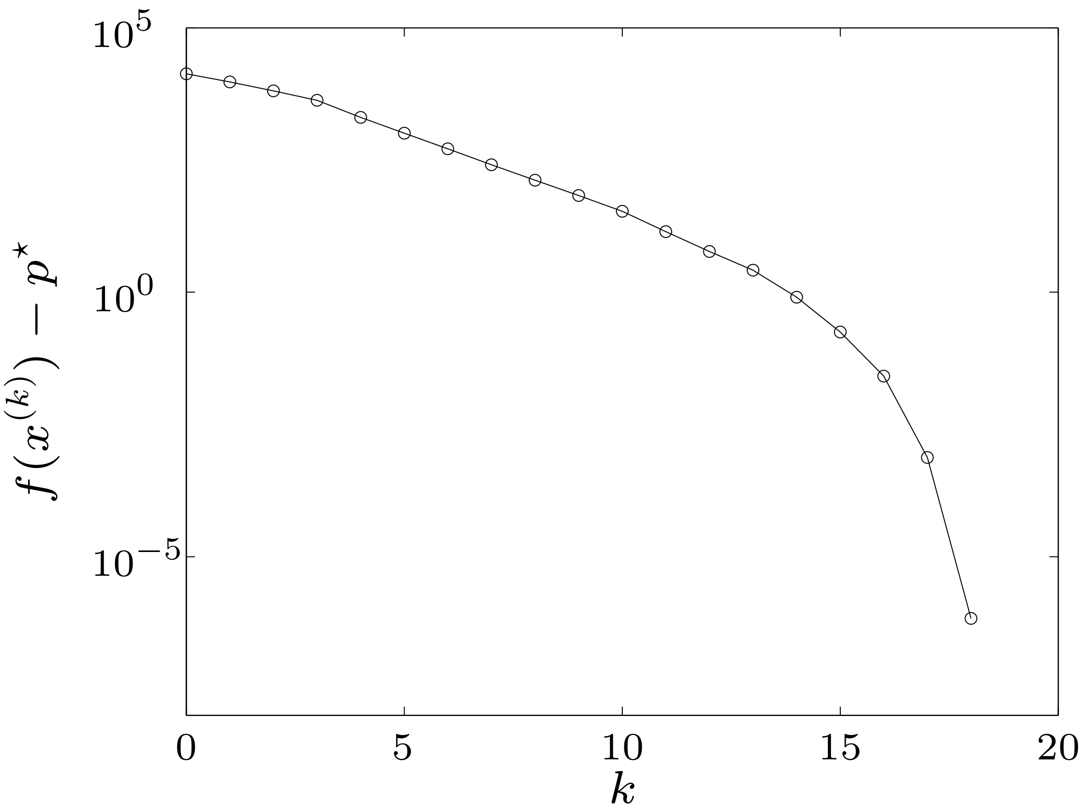
## **example in** $\mathbb{R}^{100}$ (page 13)




- 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 $\mathbb{R}^{10000}$** (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

# Implementation

main effort in each iteration: evaluate derivatives and solve Newton system

$$H\Delta x = g$$

where $H = \nabla^2 f(x)$, $g = -\nabla f(x)$

**via Cholesky factorization**

$$H = LL^T, \qquad \Delta x_{\mathrm{nt}} = L^{-T}L^{-1}g, \qquad \lambda(x) = \|L^{-1}g\|_2$$

- cost $(1/3)n^3$ flops for unstructured system
- cost $\ll (1/3)n^3$ if $H$ sparse, banded

## example of dense Newton system with structure

$$f(x) = \sum_{i=1}^{n} \psi_i(x_i) + \psi_0(Ax+b), \qquad H = D + A^T H_0 A$$

- assume $A \in \mathbb{R}^{p \times n}$, dense, with $p \ll n$
- $D$ diagonal with diagonal elements $\psi_i''(x_i)$; $H_0 = \nabla^2 \psi_0(Ax+b)$

**method 1**: form $H$, solve via dense Cholesky factorization: (cost $(1/3)n^3$)

**method 2**: factor $H_0 = L_0 L_0^T$; write Newton system as

$$D\Delta x + A^T L_0 w = -g, \qquad L_0^T A \Delta x - w = 0$$

eliminate $\Delta x$ from first equation; compute $w$ and $\Delta x$ from

$$(I + L_0^T A D^{-1} A^T L_0) w = -L_0^T A D^{-1} g, \qquad D\Delta x = -g - A^T L_0 w$$

cost: $2p^2 n$ (dominated by computation of $L_0^T A D^{-1} A L_0$)

# Self-concordance

**shortcomings of classical convergence analysis**

- depends on unknown constants ($m$, $L$, . . . )
- bound is not affinely invariant, although Newton's method is

**convergence analysis via self-concordance** (Nesterov and Nemirovski)

- does not depend on any unknown constants
- gives affine-invariant bound
- applies to special class of convex functions ('self-concordant' functions)
- developed to analyze polynomial-time interior-point methods for convex optimization

# Self-concordant functions

## definition

- $f : \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|f'''(x)| \leq 2f''(x)^{3/2}$$

  for all $x \in \mathbf{dom}\, f$

- $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant if $g(t) = f(x + tv)$ is self-concordant for all $x \in \mathbf{dom}\, f$, $v \in \mathbb{R}^n$

**examples on $\mathbb{R}$**

- linear and quadratic functions
- negative logarithm $f(x) = -\log x$
- negative entropy plus negative logarithm: $f(x) = x \log x - \log x$

**affine invariance:** if $f : \mathbb{R} \to \mathbb{R}$ is s.c., then $\tilde{f}(y) = f(ay + b)$ is s.c.:

$$\tilde{f}'''(y) = a^3 f'''(ay + b), \qquad \tilde{f}''(y) = a^2 f''(ay + b)$$

# Self-concordant calculus

**properties**

- preserved under positive scaling $\alpha \geq 1$, and sum
- preserved under composition with affine function
- if $g$ is convex with $\mathbf{dom}\, g = \mathbb{R}_{++}$ and $|g'''(x)| \leq 3g''(x)/x$ then

$$f(x) = \log(-g(x)) - \log x$$

  is self-concordant

**examples**: properties can be used to show that the following are s.c.

- $f(x) = -\sum_{i=1}^m \log(b_i - a_i^T x)$ on $\{x \mid a_i^T x < b_i,\ i = 1, \ldots, m\}$
- $f(X) = -\log\det X$ on $\mathbf{S}^n_{++}$
- $f(x) = -\log(y^2 - x^T x)$ on $\{(x, y) \mid \|x\|_2 < y\}$

# Self-concordance: complexity analysis

**Newton's method for self-concordant functions.**

**Convergence proof:**

- Affine invariant bounds on Hessian
- Newton decrement and bounds on suboptimality
- Damped Newton phase
- Quadratic Newton phase

We often only consider univariate functions to simplify analysis. . .

# Self-concordance: complexity analysis

**Affine invariant bounds on the Hessian.** Replace Lipschitz bounds and strong convexity in classical analysis.

**Lemma**

**Hessian bounds.** *Suppose $f : \mathbb{R} \to \mathbb{R}$ is a convex self-concordant function, either $f''(x) = 0$ for all $x \in \mathbf{dom}\, f$, or $f''(x) > 0$ for all $x \in \mathbf{dom}\, f$.*

**Proof.** Suppose $f''(0) > 0$, $f''(\bar{x}) = 0$ for $\bar{x} > 0$, and $f''(x) > 0$ on the interval between $0$ and $\bar{x}$. We have

$$\frac{d}{dx} f''(x)^{-1/2} = (-1/2) \frac{f'''(x)}{f''(x)^{3/2}},$$

this means we can write the self-concordance inequality $|f'''(x)| \leq 2f''(x)^{3/2}$ for all $x \in \mathbf{dom}\, f$ as

$$\left| \frac{d}{dt} \left( f''(t)^{-1/2} \right) \right| \leq 1 \tag{1}$$

for all $t \in \mathbf{dom}\, f$. This holds for $x$ between $0$ and $\bar{x}$. Integrating gives

$$f''(\bar{x})^{-1/2} - f''(0)^{-1/2} \leq \bar{x}$$

which contradicts $f''(\bar{x}) = 0$. ■

# Self-concordance: complexity analysis

**Proposition**

**Hessian bounds.** *Suppose $f:\mathbb{R}\to\mathbb{R}$ is a strictly convex self-concordant function. We have*

$$\frac{f''(0)}{\left(1+tf''(0)^{1/2}\right)^2} \leq f''(t) \leq \frac{f''(0)}{\left(1-tf''(0)^{1/2}\right)^2}. \tag{2}$$

*The lower bound is valid for all nonnegative $t\in\mathbf{dom}\, f$, the upper bound is valid if $t\in\mathbf{dom}\, f$ and $0\leq t< f''(0)^{-1/2}$.*

**Proof.** Assuming $t\geq 0$ and the interval between $0$ and $t$ is in $\mathbf{dom}\, f$, we can integrate (1) between $0$ and $t$ to obtain

$$-t \leq \int_0^t \frac{d}{d\tau}\left(f''(\tau)^{-1/2}\right) d\tau \leq t,$$

$i.e.$, $-t\leq f''(t)^{-1/2}-f''(0)^{-1/2}\leq t$. From this we obtain lower and upper bounds on $f''(t)$. ■

# Self-concordance: complexity analysis

**Lemma**

**Newton Decrement.** *Let* $\lambda(x)$ *be the Newton decrement*

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

*We have, for any nonzero* $v$

$$\frac{-v^T \nabla f(x)}{(v^T \nabla^2 f(x) v)^{1/2}} \leq \lambda(x) \tag{3}$$

*with equality for* $v = \Delta x_{\mathrm{nt}}$.

**Proof.** The Newton decrement can also be expressed as

$$\lambda(x) = \sup_{v \neq 0} \frac{-v^T \nabla f(x)}{(v^T \nabla^2 f(x) v)^{1/2}}$$

using $\|w\|_2 = \sup_{\|x\|_2=1} w^T x$, after setting $y = (\nabla^2 f(x))^{1/2} v$. ∎

# Self-concordance: complexity analysis

**Proposition**

**Bounds on suboptimality.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strictly convex self-concordant function. We have*

$$p^\star \geq f(x) - \lambda(x)^2 \tag{4}$$

*which is valid for $\lambda(x) \leq 0.68$.*

**Proof.** Let $v$ be a descent direction ($i.e.$, any direction satisfying $v^T \nabla f(x) < 0$, not necessarily the Newton direction). Define $\tilde{f} : \mathbb{R} \to \mathbb{R}$ as $\tilde{f}(t) = f(x + tv)$. By definition, the function $\tilde{f}$ is self-concordant.

Integrating the lower bound in (2) yields a lower bound on $\tilde{f}'(t)$:

$$\tilde{f}'(t) \geq \tilde{f}'(0) + \tilde{f}''(0)^{1/2} - \frac{\tilde{f}''(0)^{1/2}}{1 + t\tilde{f}''(0)^{1/2}}. \tag{5}$$

Integrating again yields a lower bound on $\tilde{f}(t)$:

$$\tilde{f}(t) \geq \tilde{f}(0) + t\tilde{f}'(0) + t\tilde{f}''(0)^{1/2} - \log(1 + t\tilde{f}''(0)^{1/2}). \qquad (6)$$

The righthand side reaches its minimum at

$$\bar{t} = \frac{-\tilde{f}'(0)}{\tilde{f}''(0) + \tilde{f}''(0)^{1/2}\tilde{f}'(0)},$$

and evaluating at $\bar{t}$ provides a lower bound on $\tilde{f}$:

$$\begin{aligned} \inf_{t \geq 0} \tilde{f}(t) \quad &\geq \quad \tilde{f}(0) + \bar{t}\tilde{f}'(0) + \bar{t}\tilde{f}''(0)^{1/2} - \log(1 + \bar{t}\tilde{f}''(0)^{1/2}) \\ &= \quad \tilde{f}(0) - \tilde{f}'(0)\tilde{f}''(0)^{-1/2} + \log(1 + \tilde{f}'(0)\tilde{f}''(0)^{-1/2}). \end{aligned}$$

The inequality (3) can be expressed as

$$\lambda(x) \geq -\tilde{f}'(0)\tilde{f}''(0)^{-1/2}$$

(with equality when $v = \Delta x_{\rm nt}$), since we have

$$\tilde{f}'(0) = v^T \nabla f(x), \qquad \tilde{f}''(0) = v^T \nabla^2 f(x) v.$$

Now using the fact that $u + \log(1-u)$ is a monotonically decreasing function of $u$, and the inequality above, we get

$$\inf_{t \geq 0} \tilde{f}(t) \geq \tilde{f}(0) + \lambda(x) + \log(1 - \lambda(x)).$$

This inequality holds for any descent direction $v$. Therefore

$$p^\star \geq f(x) + \lambda(x) + \log(1 - \lambda(x)) \qquad (7)$$

provided $\lambda(x) < 1$. The function $-(\lambda + \log(1-\lambda))$ satisfies

$$-(\lambda + \log(1-\lambda)) \approx \lambda^2/2,$$

for small $\lambda$, and the bound

$$-(\lambda + \log(1-\lambda)) \leq \lambda^2$$

for $\lambda \leq 0.68$. Thus, we have the bound on suboptimality

$$p^\star \geq f(x) - \lambda(x)^2,$$

valid for $\lambda(x) \leq 0.68$. ■

## Self-concordance: complexity analysis

**Newton's method with backtracking line search.** Assume,

- $f$ strictly convex self-concordant function
- A starting point $x^{(0)}$
- Sublevel set $S = \{x \mid f(x) \leq f(x^{(0)})\}$ is closed
- $f$ is bounded below (has a minimizer).

We show that there are numbers $\eta$ and $\gamma > 0$, with $0 < \eta \leq 1/4$, that depend only on the line search parameters $\alpha$ and $\beta$, such that

- If $\lambda(x^{(k)}) > \eta$, then
$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma. \tag{8}$$
- If $\lambda(x^{(k)}) \leq \eta$, then the backtracking line search selects $t = 1$ and
$$2\lambda(x^{(k+1)}) \leq \left(2\lambda(x^{(k)})\right)^2. \tag{9}$$

# Self-concordance: complexity analysis

## Proposition

**Damped phase** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strictly convex self-concordant function. After one step of Newton's method with backtracking line search*

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\alpha\beta\frac{\eta^2}{1+\eta}. \tag{10}$$

**Proof.** Let $\tilde{f}(t) = f(x + t\Delta x_{\mathrm{nt}})$, so we have

$$\tilde{f}'(0) = -\lambda(x)^2, \qquad \tilde{f}''(0) = \lambda(x)^2.$$

If we integrate the upper bound in (2) twice, we obtain an upper bound for $\tilde{f}(t)$:

$$\begin{aligned}\tilde{f}(t) &\leq \tilde{f}(0) + t\tilde{f}'(0) - t\tilde{f}''(0)^{1/2} - \log\left(1 - t\tilde{f}''(0)^{1/2}\right) \\ &= \tilde{f}(0) - t\lambda(x)^2 - t\lambda(x) - \log(1 - t\lambda(x)), \end{aligned} \tag{11}$$

valid for $0 \leq t < 1/\lambda(x)$.

We can use this bound to show the backtracking line search always results in a step size $t \geq \beta/(1+\lambda(x))$. To prove this we note that the point $\hat{t} = 1/(1+\lambda(x))$ satisfies the exit condition of the line search:

$$\begin{aligned} \tilde{f}(\hat{t}) &\leq \tilde{f}(0) - \hat{t}\lambda(x)^2 - \hat{t}\lambda(x) - \log(1-\hat{t}\lambda(x)) \\ &= \tilde{f}(0) - \lambda(x) + \log(1+\lambda(x)) \\ &\leq \tilde{f}(0) - \alpha\frac{\lambda(x)^2}{1+\lambda(x)} \\ &= \tilde{f}(0) - \alpha\lambda(x)^2\hat{t}. \end{aligned}$$

The second inequality follows from the fact that

$$-x + \log(1+x) + \frac{x^2}{2(1+x)} \leq 0$$

for $x \geq 0$. Since $t \geq \beta/(1+\lambda(x))$, we have

$$\tilde{f}(t) - \tilde{f}(0) \leq -\alpha\beta\frac{\lambda(x)^2}{1+\lambda(x)}. \quad \blacksquare$$

# Self-concordance: complexity analysis

**Lemma**

**Newton decrement: quadratic phase** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strictly convex self-concordant function. Suppose $\lambda(x) < 1$, and define $x^+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$, then*

$$\lambda(x^+) \leq \frac{\lambda(x)^2}{(1-\lambda(x))^2}.$$

**Proof.** Let $v = -\nabla^2 f(x)^{-1} \nabla f(x)$. From exercise 9.17, part (c), which generalizes the affine lower and upper bounds on the Hessian, we have

$$(1 - t\lambda(x))^2 \nabla^2 f(x) \preceq \nabla^2 f(x + tv) \preceq \frac{1}{(1 - t\lambda(x))^2} \nabla^2 f(x).$$

We can assume without loss of generality that $\nabla^2 f(x) = I$ (hence, $v = -\nabla f(x)$), so

$$(1 - \lambda(x))^2 I \preceq \nabla^2 f(x^+) \preceq \frac{1}{(1 - \lambda(x))^2} I,$$

and write $\lambda(x^+)$ as

$$\begin{aligned}
\lambda(x^+) &= \|\nabla^2 f(x^+)^{-1}\nabla f(x^+)\|_2 \\
&\leq (1-\lambda(x))^{-1}\|\nabla f(x^+)\|_2 \\
&= (1-\lambda(x))^{-1}\left\|\left(\int_0^1 \nabla^2 f(x+tv)v\,dt + \nabla f(x)\right)\right\|_2 \\
&= (1-\lambda(x))^{-1}\left\|\left(\int_0^1 (\nabla^2 f(x+tv)-I)\,dt\right)v\right\|_2 \\
&\leq (1-\lambda(x))^{-1}\left\|\left(\int_0^1 (\frac{1}{(1-t\lambda(x))^2}-1)\,dt\right)v\right\|_2 \\
&\leq \|v\|_2(1-\lambda(x))^{-1}\int_0^1 (\frac{1}{(1-t\lambda(x))^2}-1)\,dt \\
&= \frac{\lambda(x)^2}{(1-\lambda(x))^2}.
\end{aligned}$$

which is the desired result ■

# Self-concordance: complexity analysis

**Proposition**

**Quadratic phase** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strictly convex self-concordant function. If $\lambda(x^{(k)}) \leq \eta$, where $\eta = (1-2\alpha)/4$, after each step of Newton's method with backtracking line search*

$$2\lambda(x^{(k+1)}) \leq \left(2\lambda(x^{(k)})\right)^2.$$

**Proof.** Picking $\eta = (1-2\alpha)/4$ (which satisfies $0 < \eta < 1/4$, since $0 < \alpha < 1/2$), $i.e.$, if $\lambda(x^{(k)}) \leq (1-2\alpha)/4$, we show that the backtracking line search accepts the unit step and (9) holds.

Note that the upper bound (11) implies that a unit step $t = 1$ yields a point in $\mathbf{dom}\, f$ if $\lambda(x) < 1$.

Moreover, if $\lambda(x) \leq (1-2\alpha)/2$, we have, using (11),

$$\begin{array}{rcl} \tilde{f}(1) & \leq & \tilde{f}(0) - \lambda(x)^2 - \lambda(x) - \log(1-\lambda(x)) \\ & \leq & \tilde{f}(0) - \frac{1}{2}\lambda(x)^2 + \lambda(x)^3 \\ & \leq & \tilde{f}(0) - \alpha\lambda(x)^2, \end{array}$$

so the unit step satisfies the condition of sufficient decrease. (The second line follows from the fact that $-x - \log(1-x) \leq \frac{1}{2}x^2 + x^3$ for $0 \leq x \leq 0.81$.)

The result follows from the previous lemma: If $\lambda(x) < 1$, and $x^+ = x - \nabla^2 f(x)^{-1}\nabla f(x)$, then

$$\lambda(x^+) \leq \frac{\lambda(x)^2}{(1-\lambda(x))^2}. \tag{12}$$

In particular, if $\lambda(x) \leq 1/4$,

$$\lambda(x^+) \leq 2\lambda(x)^2,$$

which proves that the result we seek holds when $\lambda(x^{(k)}) \leq \eta$. $\blacksquare$

# Convergence analysis for self-concordant functions

**Summary.** There exist constants $\eta \in (0, 1/4]$, $\gamma > 0$ such that

- if $\lambda(x) > \eta$, then
$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$$
- if $\lambda(x) \leq \eta$, then
$$2\lambda(x^{(k+1)}) \leq \left(2\lambda(x^{(k)})\right)^2$$

($\eta$ and $\gamma$ only depend on backtracking parameters $\alpha$, $\beta$)

**Complexity bound.** Number of Newton iterations bounded by

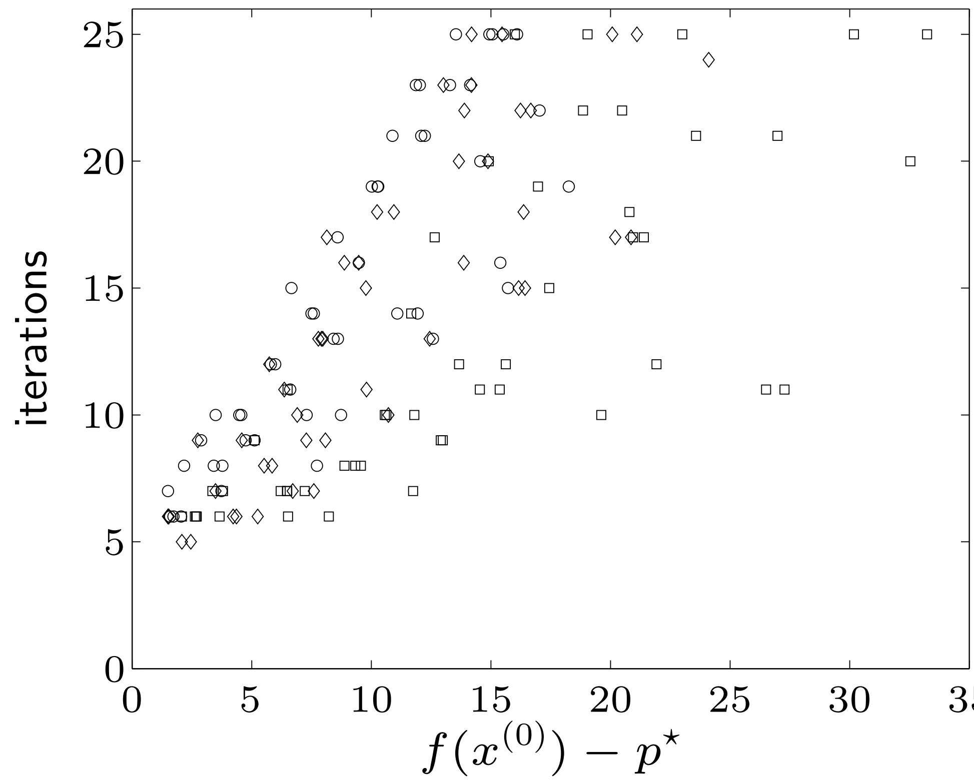
$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2(1/\epsilon)$$

for $\alpha = 0.1$, $\beta = 0.8$, $\epsilon = 10^{-10}$, bound evaluates to $375(f(x^{(0)}) - p^\star) + 6$.
**Independent of the problem dimension!**

**numerical example:** 150 randomly generated instances of

$$\text{minimize} \quad f(x) = -\sum_{i=1}^m \log(b_i - a_i^T x)$$

$\bigcirc$: $m = 100$, $n = 50$
$\square$: $m = 1000$, $n = 500$
$\diamond$: $m = 1000$, $n = 50$




- number of iterations much smaller than $375(f(x^{(0)}) - p^\star) + 6$
- bound of the form $c(f(x^{(0)}) - p^\star) + 6$ with smaller $c$ (empirically) valid
- Dimension independence verified empirically.

# Equality Constraints

# Equality Constraints

- equality constrained minimization
- eliminating equality constraints
- Newton's method with equality constraints
- infeasible start Newton method
- implementation

# Equality constrained minimization

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

- $f$ convex, twice continuously differentiable
- $A \in \mathbb{R}^{p \times n}$ with $\mathbf{Rank}\, A = p$
- we assume $p^\star$ is finite and attained

**optimality conditions:** $x^\star$ is optimal iff there exists a $\nu^\star$ such that

$$\nabla f(x^\star) + A^T \nu^\star = 0, \qquad Ax^\star = b$$

**equality constrained quadratic minimization** (with $P \in \mathbf{S}^n_+$)

$$\begin{array}{ll} \text{minimize} & (1/2)x^TPx + q^Tx + r \\ \text{subject to} & Ax = b \end{array}$$

optimality condition:

$$\left[\begin{array}{cc} P & A^T \\ A & 0 \end{array}\right] \left[\begin{array}{c} x^\star \\ \nu^\star \end{array}\right] = \left[\begin{array}{c} -q \\ b \end{array}\right]$$

- coefficient matrix is called KKT matrix
- KKT matrix is nonsingular if and only if

$$Ax = 0, \quad x \neq 0 \qquad \Longrightarrow \qquad x^TPx > 0$$

- equivalent condition for nonsingularity: $P + A^TA \succ 0$

# Eliminating equality constraints

represent solution of $\{x \mid Ax = b\}$ as

$$\{x \mid Ax = b\} = \{Fz + \hat{x} \mid z \in \mathbb{R}^{n-p}\}$$

- $\hat{x}$ is (any) particular solution
- range of $F \in \mathbb{R}^{n\times(n-p)}$ is nullspace of $A$ ($\mathbf{Rank}\, F = n - p$ and $AF = 0$)

**reduced or eliminated problem**

$$\text{minimize} \quad f(Fz + \hat{x})$$

- an unconstrained problem with variable $z \in \mathbb{R}^{n-p}$
- from solution $z^\star$, obtain $x^\star$ and $\nu^\star$ as

$$x^\star = Fz^\star + \hat{x}, \qquad \nu^\star = -(AA^T)^{-1}A\nabla f(x^\star)$$

**example:** optimal allocation with resource constraint

$$\begin{array}{ll} \text{minimize} & f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) \\ \text{subject to} & x_1 + x_2 + \cdots + x_n = b \end{array}$$

eliminate $x_n = b - x_1 - \cdots - x_{n-1}$, $i.e.$, choose

$$\hat{x} = be_n, \qquad F = \begin{bmatrix} I \\ -\mathbf{1}^T \end{bmatrix} \in \mathbf{R}^{n\times(n-1)}$$

reduced problem:

$$\text{minimize} \quad f_1(x_1) + \cdots + f_{n-1}(x_{n-1}) + f_n(b - x_1 - \cdots - x_{n-1})$$

(variables $x_1$, . . . , $x_{n-1}$)

# Newton step

Newton step of $f$ at feasible $x$ is given by (1st block) of solution of

$$\left[\begin{array}{cc} \nabla^2 f(x) & A^T \\ A & 0 \end{array}\right]\left[\begin{array}{c} \Delta x_{\rm nt} \\ w \end{array}\right] = \left[\begin{array}{c} -\nabla f(x) \\ 0 \end{array}\right]$$

**interpretations**

- $\Delta x_{\rm nt}$ solves second order approximation (with variable $v$)

$$\begin{array}{ll} \text{minimize} & \widehat{f}(x+v) = f(x) + \nabla f(x)^T v + (1/2) v^T \nabla^2 f(x) v \\ \text{subject to} & A(x+v) = b \end{array}$$

- equations follow from linearizing optimality conditions

$$\nabla f(x + \Delta x_{\rm nt}) + A^T w = 0, \qquad A(x + \Delta x_{\rm nt}) = b$$

# Newton decrement

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

**properties**

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

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

- directional derivative in Newton direction:

$$\left.\frac{d}{dt} f(x + t\Delta x_{\rm nt})\right|_{t=0} = -\lambda(x)^2$$

- in general, $\lambda(x) \neq \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}$

# Newton's method with equality constraints

**given** starting point $x \in \mathbf{dom}\, f$ with $Ax = b$, tolerance $\epsilon > 0$.
**repeat**

1. Compute the Newton step and decrement $\Delta x_{\rm nt}$, $\lambda(x)$.
2. *Stopping criterion.* **quit** if $\lambda^2/2 \leq \epsilon$.
3. *Line search.* Choose step size $t$ by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\rm nt}$.

- a feasible descent method: $x^{(k)}$ feasible and $f(x^{(k+1)}) < f(x^{(k)})$
- affine invariant

# Newton's method and elimination

**Newton's method for reduced problem**

$$\text{minimize} \quad \tilde{f}(z) = f(Fz + \hat{x})$$

- variables $z \in \mathbb{R}^{n-p}$
- $\hat{x}$ satisfies $A\hat{x} = b$; $\textbf{Rank}\, F = n - p$ and $AF = 0$
- Newton's method for $\tilde{f}$, started at $z^{(0)}$, generates iterates $z^{(k)}$

**Newton's method with equality constraints**

when started at $x^{(0)} = Fz^{(0)} + \hat{x}$, iterates are

$$x^{(k+1)} = Fz^{(k)} + \hat{x}$$

hence, don't need separate convergence analysis

# Newton step at infeasible points

2nd interpretation of page 55 extends to infeasible $x$ (*i.e.*, $Ax \neq b$)

linearizing optimality conditions at infeasible $x$ (with $x \in \mathbf{dom}\, f$) gives

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\mathrm{nt}} \\ w \end{bmatrix} = - \begin{bmatrix} \nabla f(x) \\ Ax - b \end{bmatrix} \tag{13}$$

**primal-dual interpretation**

- write optimality condition as $r(y) = 0$, where

$$y = (x, \nu), \qquad r(y) = (\nabla f(x) + A^T \nu, Ax - b)$$

- linearizing $r(y) = 0$ gives $r(y + \Delta y) \approx r(y) + Dr(y)\Delta y = 0$:

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\mathrm{nt}} \\ \Delta \nu_{\mathrm{nt}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x) + A^T \nu \\ Ax - b \end{bmatrix}$$

  same as (13) with $w = \nu + \Delta \nu_{\mathrm{nt}}$

# Infeasible start Newton method

**given** starting point $x \in \mathbf{dom}\, f$, $\nu$, tolerance $\epsilon > 0$, $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$.
**repeat**
1. Compute primal and dual Newton steps $\Delta x_{\rm nt}$, $\Delta \nu_{\rm nt}$.
2. *Backtracking line search on* $\|r\|_2$.
   $t := 1$.
   **while** $\|r(x + t\Delta x_{\rm nt}, \nu + t\Delta \nu_{\rm nt})\|_2 > (1 - \alpha t)\|r(x, \nu)\|_2$, $\quad t := \beta t$.
3. *Update.* $x := x + t\Delta x_{\rm nt}$, $\nu := \nu + t\Delta \nu_{\rm nt}$.

**until** $Ax = b$ and $\|r(x, \nu)\|_2 \leq \epsilon$.

- not a descent method: $f(x^{(k+1)}) > f(x^{(k)})$ is possible
- directional derivative of $\|r(y)\|_2^2$ in direction $\Delta y = (\Delta x_{\rm nt}, \Delta \nu_{\rm nt})$ is

$$\frac{d}{dt}\|r(y + \Delta y)\|_2 \bigg|_{t=0} = -\|r(y)\|_2$$

# Solving KKT systems

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = - \begin{bmatrix} g \\ h \end{bmatrix}$$

**solution methods**

- LDL$^{\mathsf{T}}$ factorization
- elimination (if $H$ nonsingular)

$$AH^{-1}A^T w = h - AH^{-1}g, \qquad Hv = -(g + A^T w)$$

- elimination with singular $H$: write as

$$\begin{bmatrix} H + A^T Q A & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = - \begin{bmatrix} g + A^T Q h \\ h \end{bmatrix}$$

with $Q \succeq 0$ for which $H + A^T Q A \succ 0$, and apply elimination

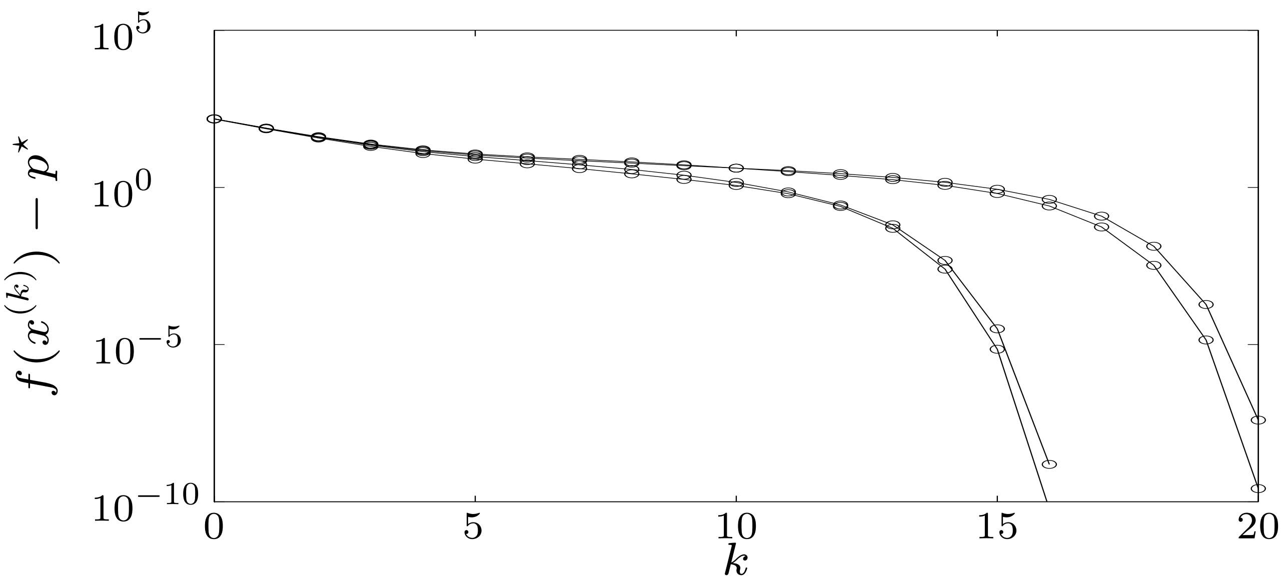
# Equality constrained analytic centering

**primal problem:** minimize $-\sum_{i=1}^{n} \log x_i$ subject to $Ax = b$

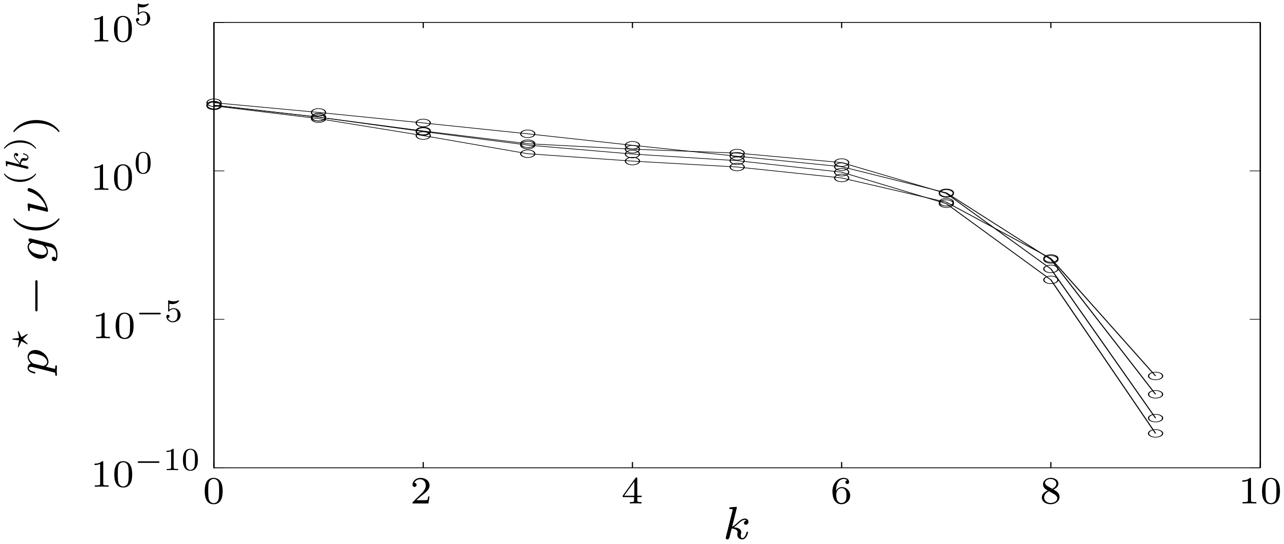
**dual problem:** maximize $-b^T\nu + \sum_{i=1}^{n} \log(A^T\nu)_i + n$

three methods for an example with $A \in \mathbb{R}^{100\times 500}$, different starting points

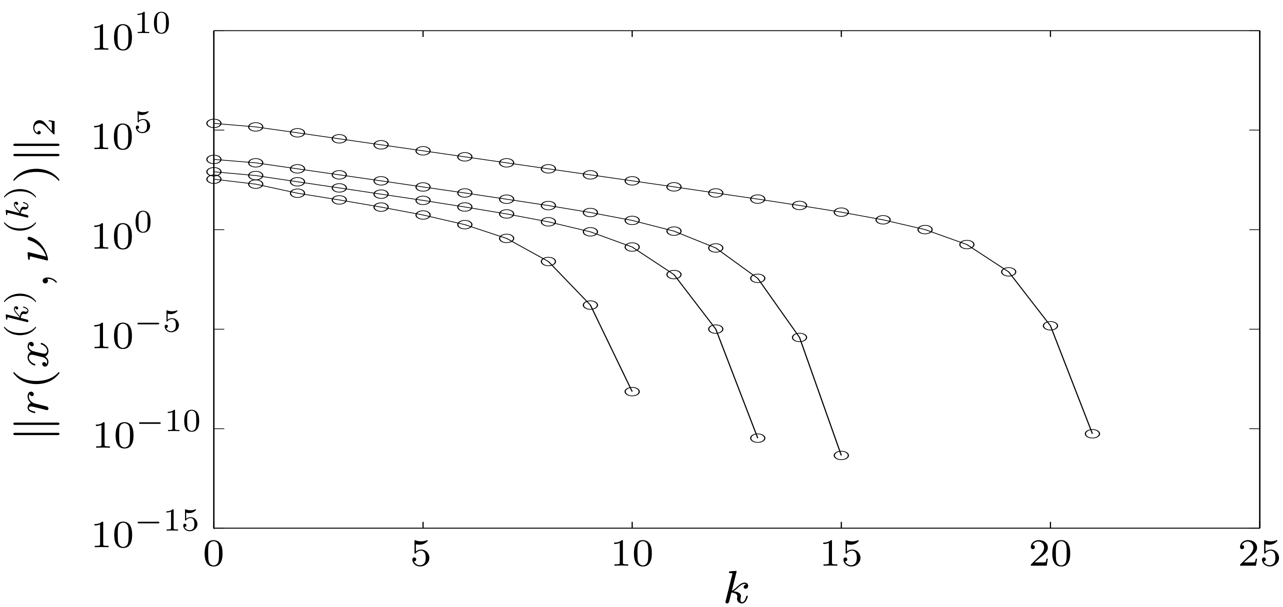
1. Newton method with equality constraints (requires $x^{(0)} \succ 0$, $Ax^{(0)} = b$)



2. Newton method applied to dual problem (requires $A^T\nu^{(0)} \succ 0$)




3. infeasible start Newton method (requires $x^{(0)} \succ 0$)



## complexity per iteration of three methods is identical

1. use block elimination to solve KKT system

$$\left[\begin{array}{cc} \mathbf{diag}(x)^{-2} & A^T \\ A & 0 \end{array}\right]\left[\begin{array}{c} \Delta x \\ w \end{array}\right] = \left[\begin{array}{c} \mathbf{diag}(x)^{-1}\mathbf{1} \\ 0 \end{array}\right]$$

   reduces to solving $A\,\mathbf{diag}(x)^2 A^T w = b$

2. solve Newton system $A\,\mathbf{diag}(A^T\nu)^{-2}A^T\Delta\nu = -b + A\,\mathbf{diag}(A^T\nu)^{-1}\mathbf{1}$

3. use block elimination to solve KKT system

$$\left[\begin{array}{cc} \mathbf{diag}(x)^{-2} & A^T \\ A & 0 \end{array}\right]\left[\begin{array}{c} \Delta x \\ \Delta\nu \end{array}\right] = \left[\begin{array}{c} \mathbf{diag}(x)^{-1}\mathbf{1} \\ Ax - b \end{array}\right]$$

   reduces to solving $A\,\mathbf{diag}(x)^2 A^T w = 2Ax - b$

**conclusion:** in each case, solve $ADA^T w = h$ with $D$ positive diagonal. It helps if this linear system is **structured**.

# Network flow optimization

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \phi_i(x_i) \\ \text{subject to} & Ax = b \end{array}$$

- directed graph with $n$ arcs, $p+1$ nodes
- $x_i$: flow through arc $i$; $\phi_i$: cost flow function for arc $i$ (with $\phi_i''(x) > 0$)
- node-incidence matrix $\tilde{A} \in \mathbb{R}^{(p+1)\times n}$ defined as

$$\tilde{A}_{ij} = \begin{cases} 1 & \text{arc } j \text{ leaves node } i \\ -1 & \text{arc } j \text{ enters node } i \\ 0 & \text{otherwise} \end{cases}$$

- reduced node-incidence matrix $A \in \mathbb{R}^{p\times n}$ is $\tilde{A}$ with last row removed
- $b \in \mathbb{R}^p$ is (reduced) source vector
- $\mathbf{Rank}\, A = p$ if graph is connected

# KKT system

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = - \begin{bmatrix} g \\ h \end{bmatrix}$$

- $H = \mathbf{diag}(\phi_1''(x_1), \ldots, \phi_n''(x_n))$, positive diagonal
- solve via elimination:

$$AH^{-1}A^Tw = h - AH^{-1}g, \qquad Hv = -(g + A^Tw)$$

sparsity pattern of coefficient matrix is given by graph connectivity

$$\begin{aligned} (AH^{-1}A^T)_{ij} \neq 0 \quad &\Longleftrightarrow \quad (AA^T)_{ij} \neq 0 \\ &\Longleftrightarrow \quad \text{nodes } i \text{ and } j \text{ are connected by an arc} \end{aligned}$$

# Analytic center of linear matrix inequality

$$\begin{array}{ll} \text{minimize} & -\log\det X \\ \text{subject to} & \mathbf{Tr}(A_i X) = b_i, \quad i = 1, \ldots, p \end{array}$$

variable $X \in \mathbf{S}^n$

**optimality conditions**

$$X^\star \succ 0, \qquad -(X^\star)^{-1} + \sum_{j=1}^{p} \nu_j^\star A_i = 0, \qquad \mathbf{Tr}(A_i X^\star) = b_i, \quad i = 1, \ldots, p$$

**Newton equation at feasible $X$:**

$$X^{-1}\Delta X X^{-1} + \sum_{j=1}^{p} w_j A_i = X^{-1}, \qquad \mathbf{Tr}(A_i \Delta X) = 0, \quad i = 1, \ldots, p$$

- follows from linear approximation $(X + \Delta X)^{-1} \approx X^{-1} - X^{-1}\Delta X X^{-1}$
- $n(n+1)/2 + p$ variables $\Delta X$, $w$

## solution by block elimination

- eliminate $\Delta X$ from first equation: $\Delta X = X - \sum_{j=1}^{p} w_j X A_j X$
- substitute $\Delta X$ in second equation

$$\sum_{j=1}^{p} \mathbf{Tr}(A_i X A_j X) w_j = b_i, \quad i = 1, \ldots, p \tag{14}$$

a dense positive definite set of linear equations with variable $w \in \mathbb{R}^p$

flop count (dominant terms) using Cholesky factorization $X = LL^T$:

- form $p$ products $L^T A_j L$: $(3/2)pn^3$
- form $p(p+1)/2$ inner products $\mathbf{Tr}((L^T A_i L)(L^T A_j L))$: $(1/2)p^2n^2$
- solve (14) via Cholesky factorization: $(1/3)p^3$

# Barrier Method

# Barrier Method

- inequality constrained minimization
- logarithmic barrier function and central path
- barrier method
- feasibility and phase I methods
- complexity analysis via self-concordance
- generalized inequalities

# Inequality constrained minimization

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array} \tag{15}$$

- $f_i$ convex, twice continuously differentiable
- $A \in \mathbb{R}^{p \times n}$ with $\mathbf{Rank}\, A = p$
- we assume $p^\star$ is finite and attained
- we assume problem is strictly feasible: there exists $\tilde{x}$ with

$$\tilde{x} \in \mathbf{dom}\, f_0, \qquad f_i(\tilde{x}) < 0, \quad i = 1, \ldots, m, \qquad A\tilde{x} = b$$

hence, strong duality holds and dual optimum is attained

# Examples

- LP, QP, QCQP, GP
- entropy maximization with linear inequality constraints

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Fx \preceq g \\ & Ax = b \end{array}$$

  with $\mathbf{dom}\, f_0 = \mathbb{R}^n_{++}$
- differentiability may require reformulating the problem, $e.g.$, piecewise-linear minimization or $\ell_\infty$-norm approximation via LP
- SDPs and SOCPs are better handled as problems with generalized inequalities (see later)

# Logarithmic barrier

**reformulation of (15) via indicator function:**

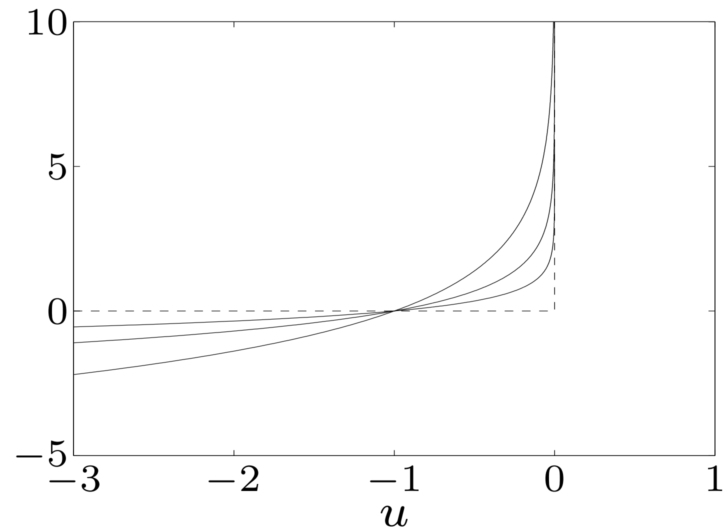
$$\begin{array}{ll} \text{minimize} & f_0(x) + \sum_{i=1}^m I_-(f_i(x)) \\ \text{subject to} & Ax = b \end{array}$$

where $I_-(u) = 0$ if $u \leq 0$, $I_-(u) = \infty$ otherwise (indicator function of $\mathbb{R}_-$)

**approximation via logarithmic barrier**

$$\begin{array}{ll} \text{minimize} & f_0(x) - (1/t)\sum_{i=1}^m \log(-f_i(x)) \\ \text{subject to} & Ax = b \end{array}$$

- an equality constrained problem
- for $t > 0$, $-(1/t)\log(-u)$ is a smooth approximation of $I_-$
- approximation improves as $t \to \infty$



## logarithmic barrier function

$$\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x)), \quad \mathbf{dom}\, \phi = \{x \mid f_1(x) < 0, \ldots, f_m(x) < 0\}$$

- convex (follows from composition rules)
- twice continuously differentiable, with derivatives

$$\begin{aligned} \nabla \phi(x) &= \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla f_i(x) \\ \nabla^2 \phi(x) &= \sum_{i=1}^{m} \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x) \end{aligned}$$

# Central path

- for $t > 0$, define $x^\star(t)$ as the solution of

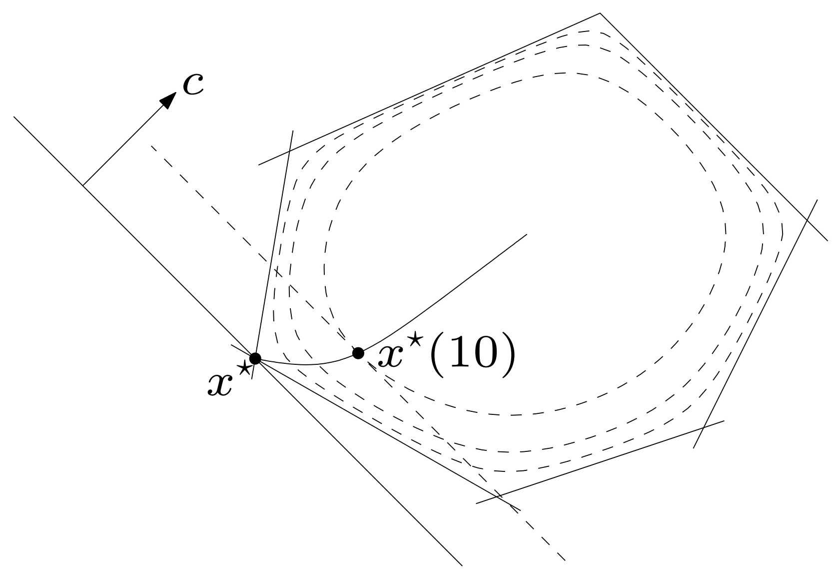
$$\begin{array}{ll} \text{minimize} & t f_0(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

  (for now, assume $x^\star(t)$ exists and is unique for each $t > 0$)

- central path is $\{x^\star(t) \mid t > 0\}$

**example:** central path for an LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a_i^T x \leq b_i, \quad i = 1, \ldots, 6 \end{array}$$




hyperplane $c^T x = c^T x^\star(t)$ is tangent to level curve of $\phi$ through $x^\star(t)$

# Dual points on central path

$x = x^\star(t)$ if there exists a $w$ such that

$$t\nabla f_0(x) + \sum_{i=1}^m \frac{1}{-f_i(x)}\nabla f_i(x) + A^T w = 0, \qquad Ax = b$$

- therefore, $x^\star(t)$ minimizes the Lagrangian

$$L(x, \lambda^\star(t), \nu^\star(t)) = f_0(x) + \sum_{i=1}^m \lambda_i^\star(t) f_i(x) + \nu^\star(t)^T(Ax - b)$$

  where we define $\lambda_i^\star(t) = 1/(-tf_i(x^\star(t))$ and $\nu^\star(t) = w/t$. We get **dual points for free.**

- this confirms the intuitive idea that $f_0(x^\star(t)) \to p^\star$ if $t \to \infty$:

$$\begin{aligned} p^\star &\geq g(\lambda^\star(t), \nu^\star(t)) \\ &= L(x^\star(t), \lambda^\star(t), \nu^\star(t)) \\ &= f_0(x^\star(t)) - m/t \end{aligned}$$

# Interpretation via KKT conditions

$x = x^\star(t)$, $\lambda = \lambda^\star(t)$, $\nu = \nu^\star(t)$ satisfy

1. primal constraints: $f_i(x) \leq 0$, $i = 1, \ldots, m$, $Ax = b$
2. dual constraints: $\lambda \succeq 0$
3. approximate complementary slackness: $-\lambda_i f_i(x) = 1/t$, $i = 1, \ldots, m$
4. gradient of Lagrangian with respect to $x$ vanishes:

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + A^T \nu = 0$$

difference with KKT is that condition 3 replaces $\lambda_i f_i(x) = 0$

# Force field interpretation

**centering problem** (for problem with no equality constraints)

$$\text{minimize} \quad t f_0(x) - \textstyle\sum_{i=1}^m \log(-f_i(x))$$

**force field interpretation**

- $t f_0(x)$ is potential of force field $F_0(x) = -t\nabla f_0(x)$
- $-\log(-f_i(x))$ is potential of force field $F_i(x) = (1/f_i(x))\nabla f_i(x)$

the forces balance at $x^\star(t)$:

$$F_0(x^\star(t)) + \sum_{i=1}^m F_i(x^\star(t)) = 0$$

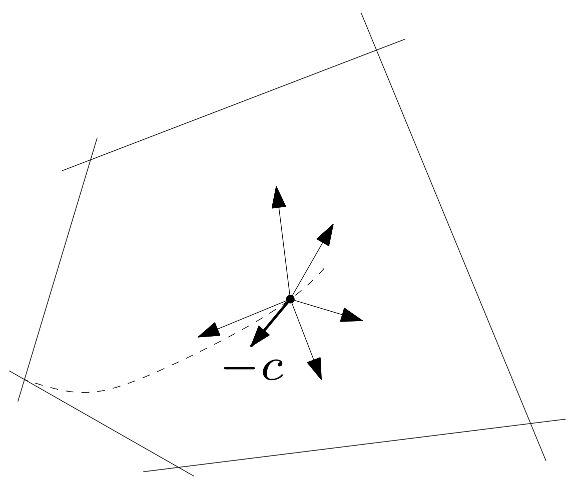
**example**

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

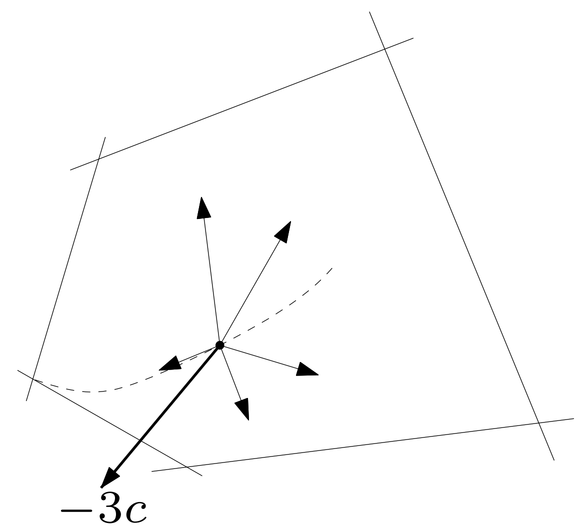
- objective force field is constant: $F_0(x) = -tc$
- constraint force field decays as inverse distance to constraint hyperplane:

$$F_i(x) = \frac{-a_i}{b_i - a_i^T x}, \qquad \|F_i(x)\|_2 = \frac{1}{\mathbf{dist}(x, \mathcal{H}_i)}$$

where $\mathcal{H}_i = \{x \mid a_i^T x = b_i\}$




$t = 1$




$t = 3$

# Barrier method

**given** strictly feasible $x$, $t := t^{(0)} > 0$, $\mu > 1$, tolerance $\epsilon > 0$.

**repeat**

1. *Centering step.* Compute $x^\star(t)$ by minimizing $tf_0 + \phi$, subject to $Ax = b$.
2. *Update.* $x := x^\star(t)$.
3. *Stopping criterion.* **quit** if $m/t < \epsilon$.
4. *Increase* $t$. $t := \mu t$.

- terminates with $f_0(x) - p^\star \leq \epsilon$ (stopping criterion follows from $f_0(x^\star(t)) - p^\star \leq m/t$)
- centering usually done using Newton's method, starting at current $x$
- choice of $\mu$ involves a trade-off: large $\mu$ means fewer outer iterations, more inner problem minimization iterations (i.e. Newton steps); typical values: $\mu = 10$–$20$
- several heuristics for choice of $t^{(0)}$

# Convergence analysis

**number of outer (centering) iterations:** exactly

$$\left\lceil \frac{\log(m/(\epsilon t^{(0)}))}{\log \mu} \right\rceil$$

plus the initial centering step (to compute $x^\star(t^{(0)})$)

**centering problem**

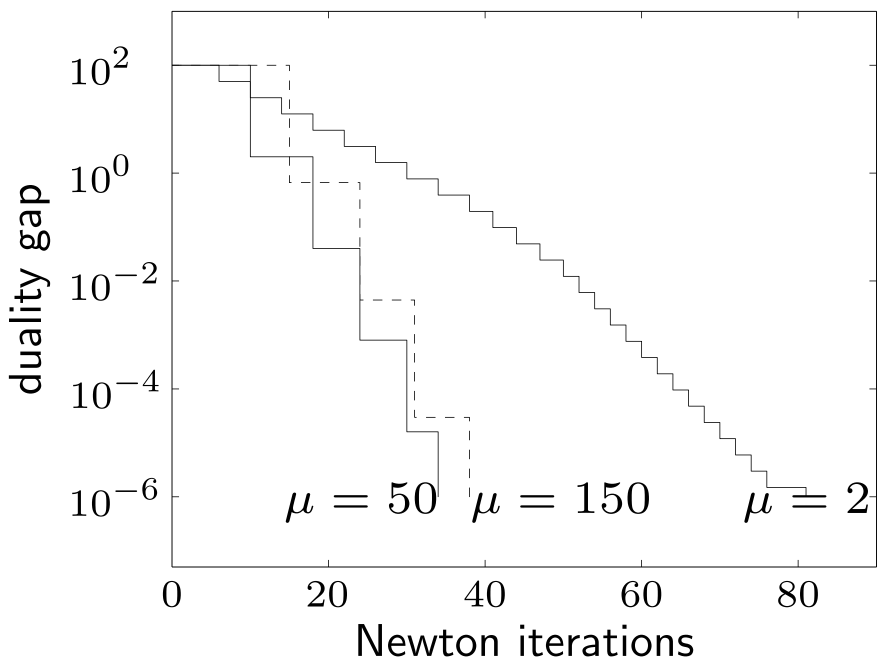
$$\text{minimize} \quad t f_0(x) + \phi(x)$$

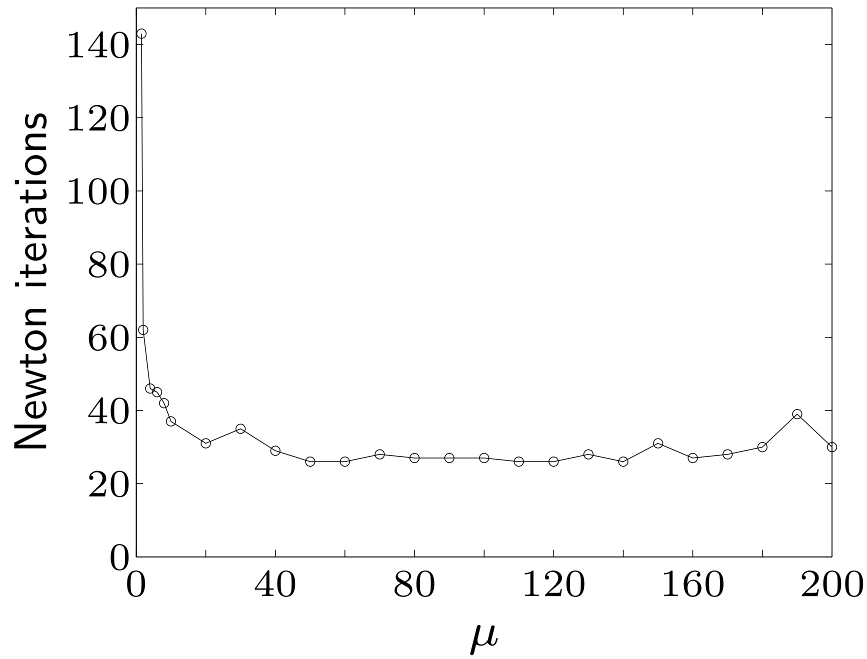
see convergence analysis of Newton's method

- $t f_0 + \phi$ must have closed sublevel sets for $t \geq t^{(0)}$
- classical analysis requires strong convexity, Lipschitz condition
- analysis via self-concordance requires self-concordance of $t f_0 + \phi$

# Examples

**inequality form LP** ($m = 100$ inequalities, $n = 50$ variables)

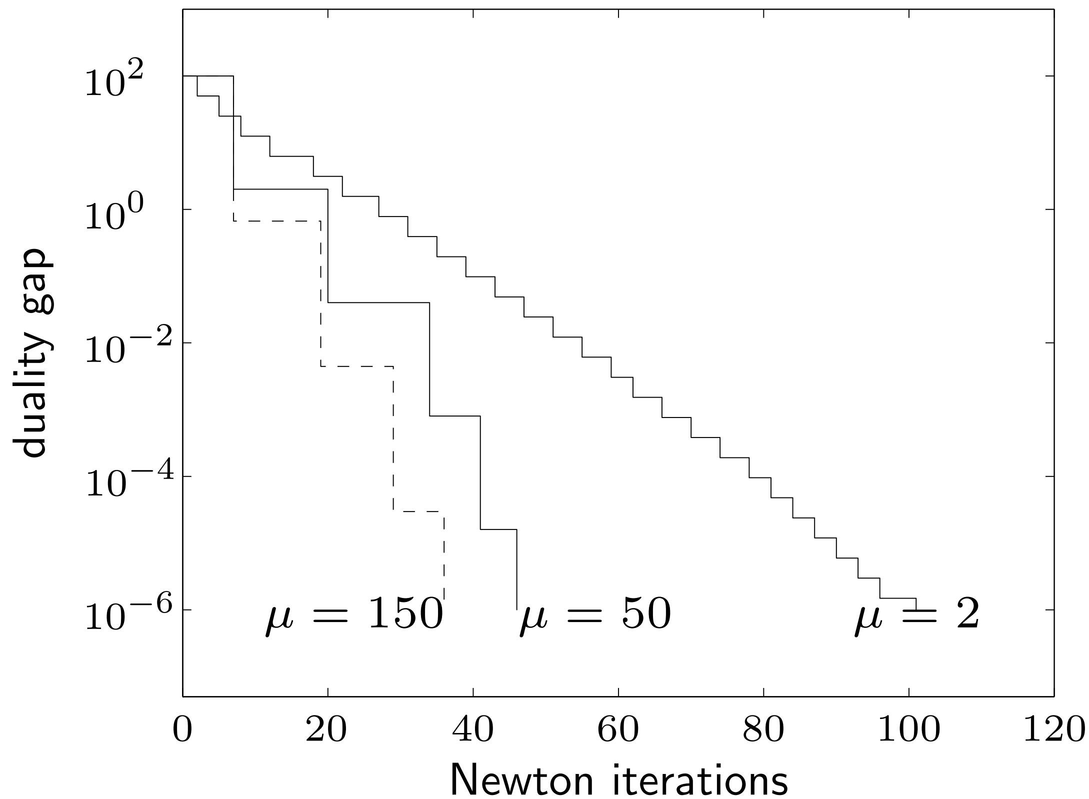



- starts with $x$ on central path ($t^{(0)} = 1$, duality gap 100)
- terminates when $t = 10^8$ (gap $10^{-6}$)
- centering uses Newton's method with backtracking
- total number of Newton iterations not very sensitive for $\mu \geq 10$

**geometric program** ($m = 100$ inequalities and $n = 50$ variables)

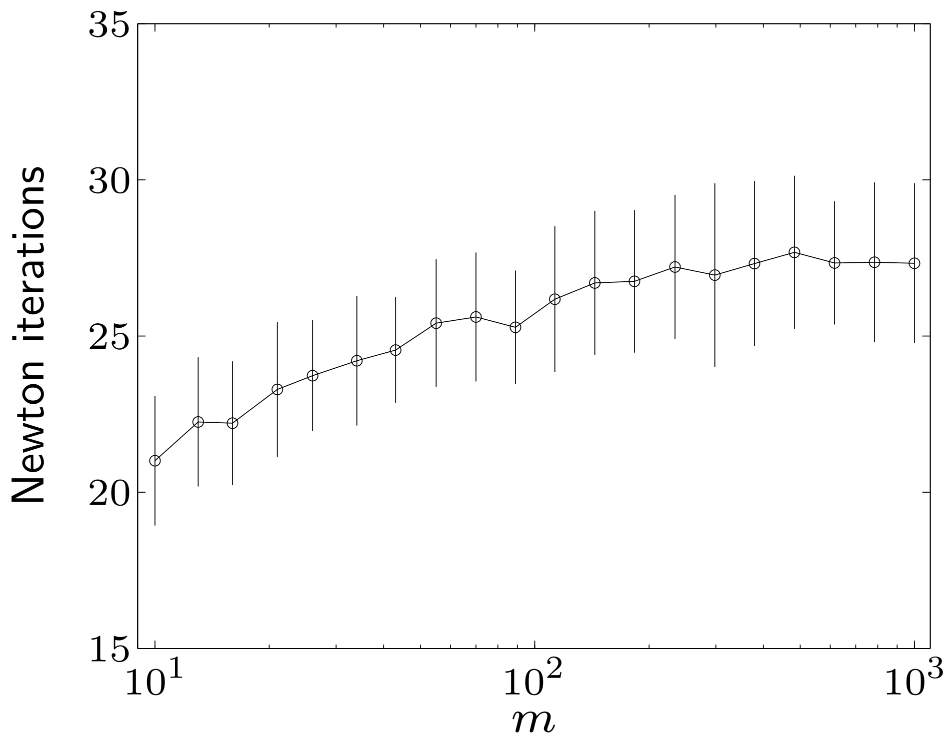
$$\begin{array}{ll} \text{minimize} & \log\left(\sum_{k=1}^{5} \exp(a_{0k}^T x + b_{0k})\right) \\ \text{subject to} & \log\left(\sum_{k=1}^{5} \exp(a_{ik}^T x + b_{ik})\right) \leq 0, \quad i = 1, \ldots, m \end{array}$$



**family of standard LPs** ($A \in \mathbb{R}^{m \times 2m}$)

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \quad x \succeq 0 \end{array}$$

$m = 10, \ldots, 1000$; for each $m$, solve 100 randomly generated instances




number of iterations grows very slowly as $m$ ranges over a $100 : 1$ ratio

# Feasibility and phase I methods

**feasibility problem:** find $x$ such that

$$f_i(x) \leq 0, \quad i = 1, \ldots, m, \qquad Ax = b \tag{16}$$

**phase I**: computes strictly feasible starting point for barrier method

**basic phase I method**

$$\begin{array}{ll} \text{minimize (over } x, s) & s \\ \text{subject to} & f_i(x) \leq s, \quad i = 1, \ldots, m \\ & Ax = b \end{array} \tag{17}$$

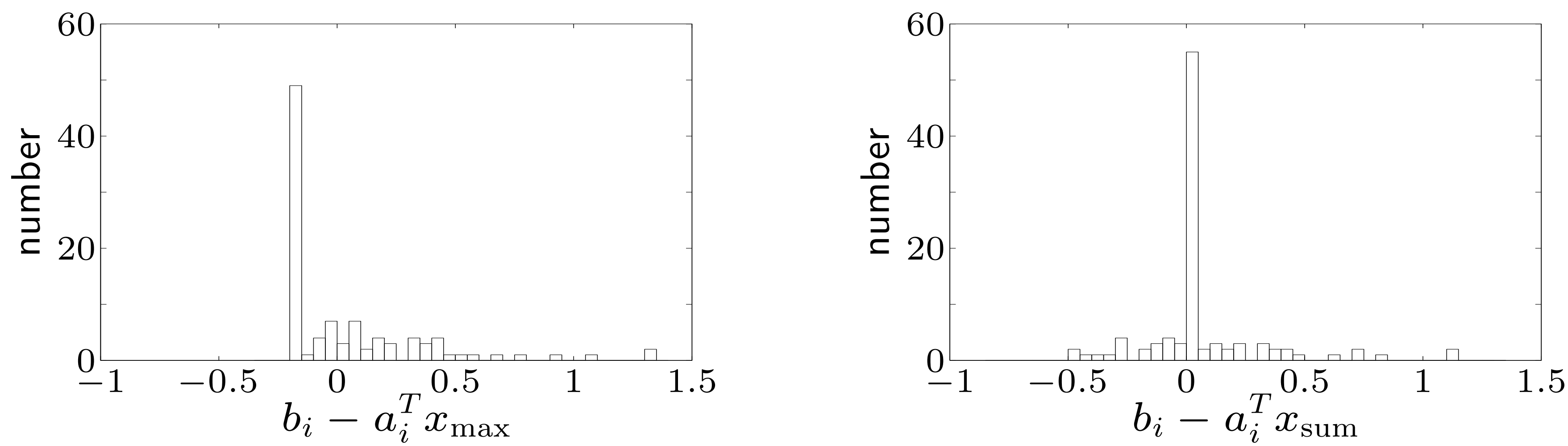
- if $x$, $s$ feasible, with $s < 0$, then $x$ is strictly feasible for (16)
- if optimal value $\bar{p}^\star$ of (17) is positive, then problem (16) is infeasible
- if $\bar{p}^\star = 0$ and attained, then problem (16) is feasible (but not strictly);
  if $\bar{p}^\star = 0$ and not attained, then problem (16) is infeasible

## sum of infeasibilities phase I method

$$\begin{array}{ll}\text{minimize} & \mathbf{1}^T s \\ \text{subject to} & s \succeq 0, \quad f_i(x) \leq s_i, \quad i = 1, \ldots, m \\ & Ax = b\end{array}$$

for infeasible problems, produces a solution that satisfies many more inequalities than basic phase I method

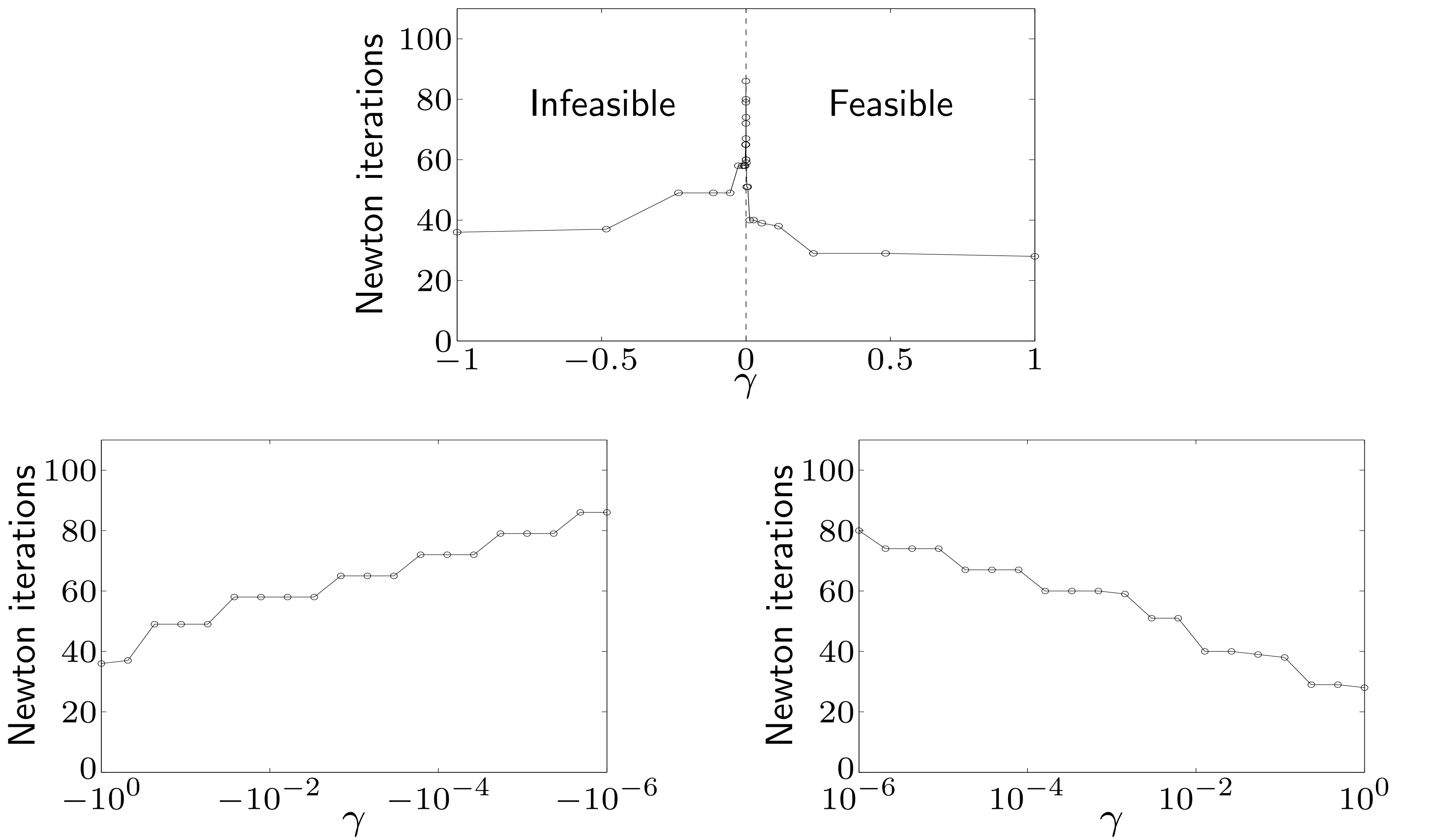
**example** (infeasible set of 100 linear inequalities in 50 variables)




left: basic phase I solution; satisfies 39 inequalities
right: sum of infeasibilities phase I solution; satisfies 79 inequalities

**example:** family of linear inequalities $Ax \preceq b + \gamma \Delta b$

- data chosen to be strictly feasible for $\gamma > 0$, infeasible for $\gamma \leq 0$
- use basic phase I, terminate when $s < 0$ or dual objective is positive




number of iterations roughly proportional to $\log(1/|\gamma|)$

# Complexity analysis via self-concordance

same assumptions as on page 71, plus:

- sublevel sets (of $f_0$, on the feasible set) are bounded
- $tf_0 + \phi$ is self-concordant with closed sublevel sets

second condition

- holds for LP, QP, QCQP
- may require reformulating the problem, *e.g.*,

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Fx \preceq g \end{array} \quad \longrightarrow \quad \begin{array}{ll} \text{minimize} & \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Fx \preceq g, \quad x \succeq 0 \end{array}$$

- needed for complexity analysis; barrier method works even when self-concordance assumption does not apply

**Newton iterations per centering step:** from self-concordance theory

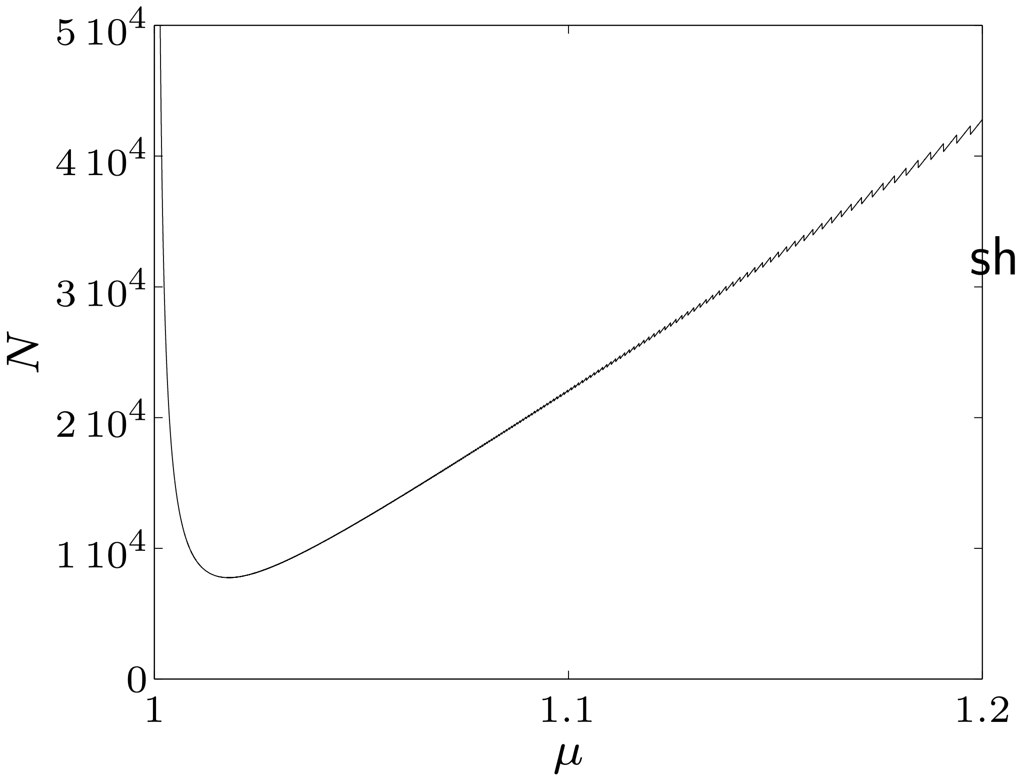
$$\#\text{Newton iterations} \leq \frac{\mu t f_0(x) + \phi(x) - \mu t f_0(x^+) - \phi(x^+)}{\gamma} + c$$

- bound on effort of computing $x^+ = x^\star(\mu t)$ starting at $x = x^\star(t)$
- **Note:** The complexity of Newton's method is independent of $m$, but the precision target is not in this case. $\gamma$, $c$ are constants (line search params).
- from duality (with $\lambda = \lambda^\star(t)$, $\nu = \nu^\star(t)$):

$$\begin{aligned}
&\mu t f_0(x) + \phi(x) - \mu t f_0(x^+) - \phi(x^+) \\
&\quad= \mu t f_0(x) - \mu t f_0(x^+) + \sum_{i=1}^m \log(-\mu t \lambda_i f_i(x^+)) - m \log \mu \\
&\quad\leq \mu t f_0(x) - \mu t f_0(x^+) - \mu t \sum_{i=1}^m \lambda_i f_i(x^+) - m - m \log \mu \\
&\quad\leq \mu t f_0(x) - \mu t g(\lambda, \nu) - m - m \log \mu \\
&\quad= m(\mu - 1 - \log \mu)
\end{aligned}$$

**total number of Newton iterations** (excluding first centering step)

$$\#\text{Newton iterations} \leq N = \left\lceil \frac{\log(m/(t^{(0)}\epsilon))}{\log \mu} \right\rceil \left( \frac{m(\mu - 1 - \log \mu)}{\gamma} + c \right)$$




shows $N$ for typical values of $\gamma$, $c$,

$$m = 100, \qquad \frac{m}{t^{(0)}\epsilon} = 10^5$$

- confirms trade-off in choice of $\mu$
- in practice, #iterations is in the tens; not very sensitive for $\mu \geq 10$

## polynomial-time complexity of barrier method

- for $\mu = 1 + 1/\sqrt{m}$:

$$N = O\left(\sqrt{m}\log\left(\frac{m/t^{(0)}}{\epsilon}\right)\right)$$

- number of Newton iterations for fixed gap reduction is $O(\sqrt{m})$
- multiply with cost of one Newton iteration (solving a linear system: cost is a polynomial function of problem dimensions), to get bound on number of flops

this choice of $\mu$ optimizes worst-case complexity; in practice we choose $\mu$ fixed ($\mu = 10, \ldots, 20$)

# Generalized inequalities

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

- $f_0$ convex, $f_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$, $i = 1, \ldots, m$, convex with respect to proper cones $K_i \in \mathbb{R}^{k_i}$
- $f_i$ twice continuously differentiable
- $A \in \mathbb{R}^{p \times n}$ with $\textbf{Rank}\, A = p$
- we assume $p^\star$ is finite and attained
- we assume problem is strictly feasible; hence strong duality holds and dual optimum is attained

Very useful **generalization of linear programming**. Examples of greatest interest: SOCP, SDP

# Generalized logarithm for proper cone

$\psi : \mathbb{R}^q \to \mathbb{R}$ is generalized logarithm for proper cone $K \subseteq \mathbb{R}^q$ if:

- $\mathbf{dom}\,\psi = \mathbf{int}\,K$ and $\nabla^2\psi(y) \prec 0$ for $y \succ_K 0$
- $\psi(sy) = \psi(y) + \theta \log s$ for $y \succ_K 0$, $s > 0$ ($\theta$ is the degree of $\psi$)

**examples**

- nonnegative orthant $K = \mathbb{R}^n_+$: $\psi(y) = \sum_{i=1}^n \log y_i$, with degree $\theta = n$
- positive semidefinite cone $K = \mathbf{S}^n_+$:

$$\psi(Y) = \log\det Y \qquad (\theta = n)$$

- second-order cone $K = \{y \in \mathbb{R}^{n+1} \mid (y_1^2 + \cdots + y_n^2)^{1/2} \leq y_{n+1}\}$:

$$\psi(y) = \log(y_{n+1}^2 - y_1^2 - \cdots - y_n^2) \qquad (\theta = 2)$$

**properties** (without proof): for $y \succ_K 0$,

$$\nabla \psi(y) \succeq_{K^*} 0, \qquad y^T \nabla \psi(y) = \theta$$

- nonnegative orthant $\mathbb{R}^n_+$: $\psi(y) = \sum_{i=1}^n \log y_i$

$$\nabla \psi(y) = (1/y_1, \ldots, 1/y_n), \qquad y^T \nabla \psi(y) = n$$

- positive semidefinite cone $\mathbf{S}^n_+$: $\psi(Y) = \log \det Y$

$$\nabla \psi(Y) = Y^{-1}, \qquad \mathbf{Tr}(Y \nabla \psi(Y)) = n$$

- second-order cone $K = \{y \in \mathbb{R}^{n+1} \mid (y_1^2 + \cdots + y_n^2)^{1/2} \le y_{n+1}\}$:

$$\psi(y) = \frac{2}{y_{n+1}^2 - y_1^2 - \cdots - y_n^2} \begin{bmatrix} -y_1 \\ \vdots \\ -y_n \\ y_{n+1} \end{bmatrix}, \qquad y^T \nabla \psi(y) = 2$$

# Logarithmic barrier and central path

**logarithmic barrier** for $f_1(x) \preceq_{K_1} 0, \ldots, f_m(x) \preceq_{K_m} 0$:

$$\phi(x) = -\sum_{i=1}^{m} \psi_i(-f_i(x)), \qquad \mathbf{dom}\,\phi = \{x \mid f_i(x) \prec_{K_i} 0,\ i = 1, \ldots, m\}$$

- $\psi_i$ is generalized logarithm for $K_i$, with degree $\theta_i$
- $\phi$ is convex, twice continuously differentiable

**central path:** $\{x^\star(t) \mid t > 0\}$ where $x^\star(t)$ solves

$$\begin{array}{ll} \text{minimize} & tf_0(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

# Dual points on central path

$x = x^\star(t)$ if there exists $w \in \mathbb{R}^p$,

$$t\nabla f_0(x) + \sum_{i=1}^m Df_i(x)^T \nabla\psi_i(-f_i(x)) + A^T w = 0$$

($Df_i(x) \in \mathbb{R}^{k_i \times n}$ is derivative matrix of $f_i$)

- therefore, $x^\star(t)$ minimizes Lagrangian $L(x, \lambda^\star(t), \nu^\star(t))$, where

$$\lambda_i^\star(t) = \frac{1}{t}\nabla\psi_i(-f_i(x^\star(t))), \qquad \nu^\star(t) = \frac{w}{t}$$

- from properties of $\psi_i$: $\lambda_i^\star(t) \succ_{K_i^*} 0$, with duality gap

$$f_0(x^\star(t)) - g(\lambda^\star(t), \nu^\star(t)) = (1/t)\sum_{i=1}^m \theta_i$$

**example: semidefinite programming** (with $F_i \in \mathbf{S}^p$)

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & F(x) = \sum_{i=1}^n x_i F_i + G \preceq 0 \end{array}$$

- logarithmic barrier: $\phi(x) = \log\det(-F(x)^{-1})$
- central path: $x^\star(t)$ minimizes $tc^T x - \log\det(-F(x))$; hence

$$tc_i - \mathbf{Tr}(F_i F(x^\star(t))^{-1}) = 0, \quad i = 1, \ldots, n$$

- dual point on central path: $Z^\star(t) = -(1/t)F(x^\star(t))^{-1}$ is feasible for

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

- duality gap on central path: $c^T x^\star(t) - \mathbf{Tr}(GZ^\star(t)) = p/t$

# Barrier method

**given** strictly feasible $x$, $t := t^{(0)} > 0$, $\mu > 1$, tolerance $\epsilon > 0$.

**repeat**

1. *Centering step.* Compute $x^\star(t)$ by minimizing $tf_0 + \phi$, subject to $Ax = b$.
2. *Update.* $x := x^\star(t)$.
3. *Stopping criterion.* **quit** if $(\sum_i \theta_i)/t < \epsilon$.
4. *Increase* $t$*.* $t := \mu t$.

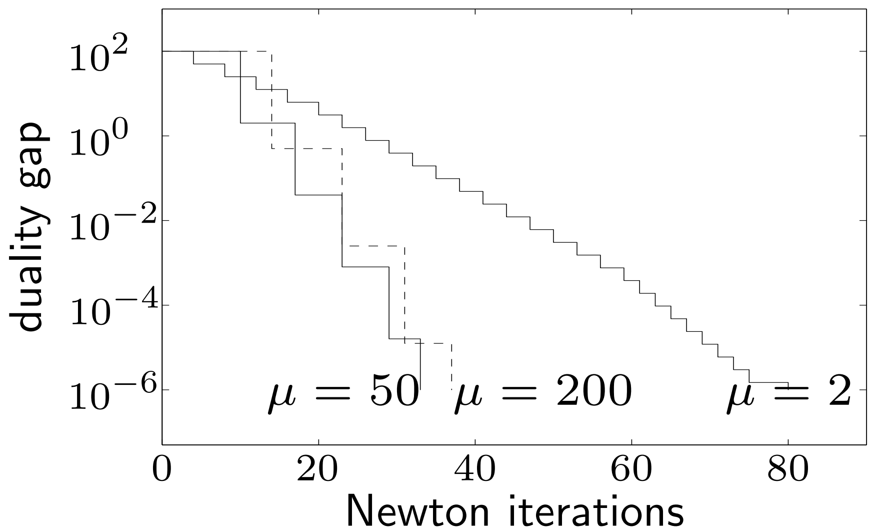
- only difference is duality gap $m/t$ on central path is replaced by $\sum_i \theta_i / t$
- number of outer iterations:

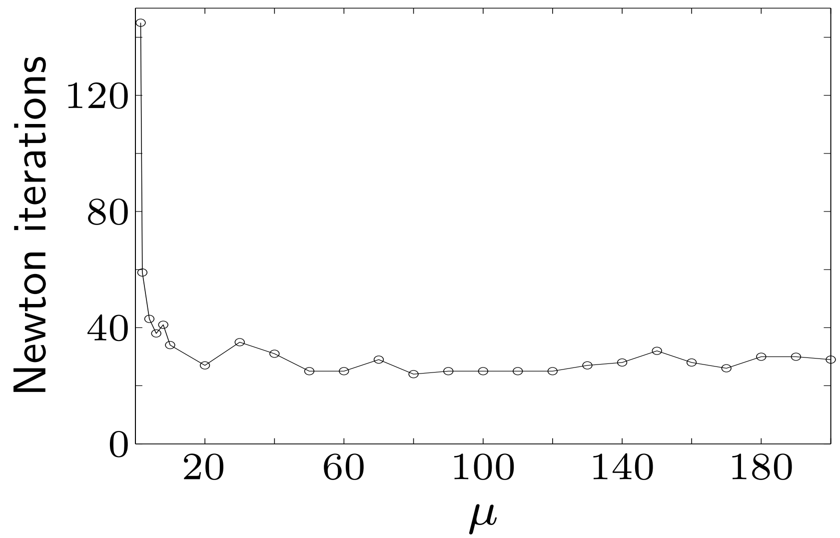
$$\left\lceil \frac{\log((\sum_i \theta_i)/(\epsilon t^{(0)}))}{\log \mu} \right\rceil$$

- complexity analysis via self-concordance applies to SDP, SOCP

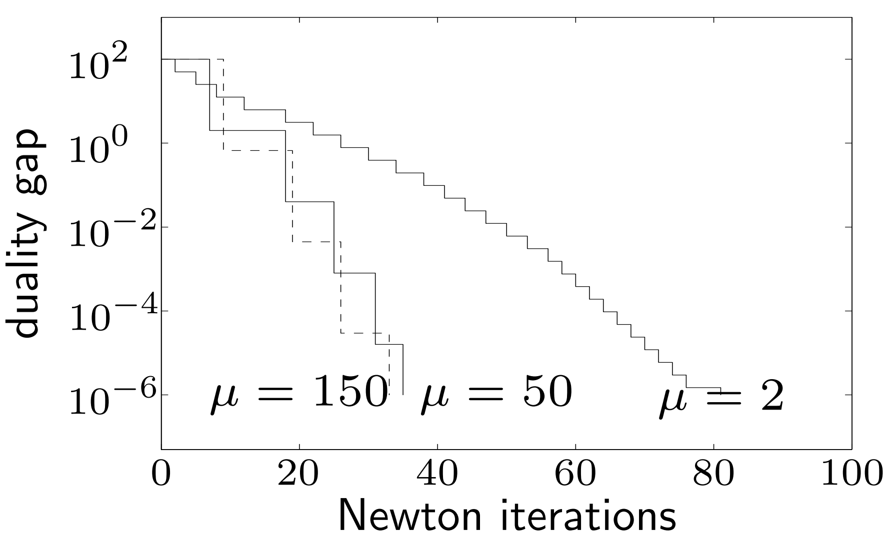
# Examples

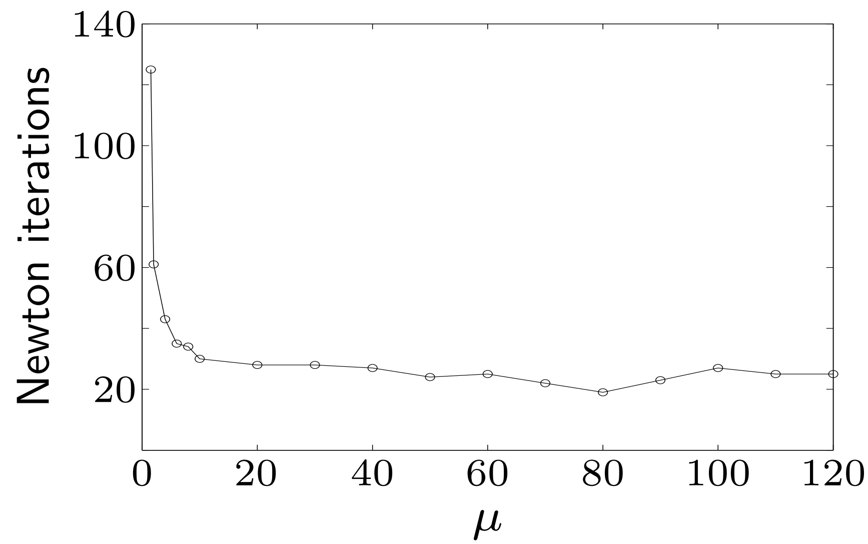
**second-order cone program** ($50$ variables, $50$ SOC constraints in $\mathbb{R}^6$)




**semidefinite program** ($100$ variables, LMI constraint in $\mathbf{S}^{100}$)

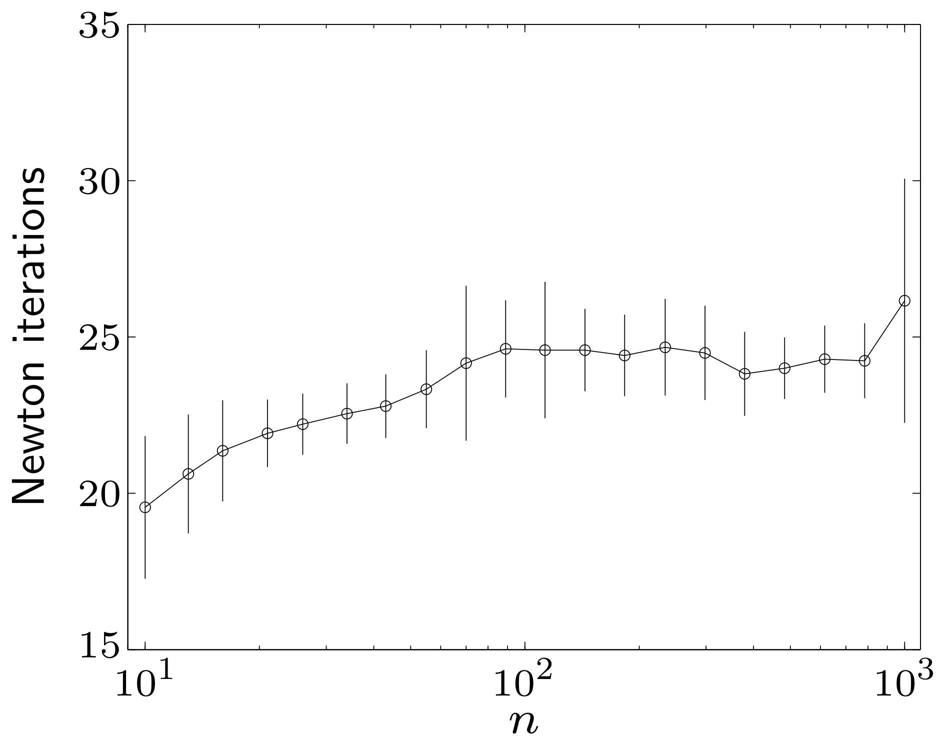


**family of SDPs** ($A \in \mathbf{S}^n$, $x \in \mathbb{R}^n$)

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T x \\ \text{subject to} & A + \mathbf{diag}(x) \succeq 0 \end{array}$$

$n = 10, \ldots, 1000$, for each $n$ solve $100$ randomly generated instances



# Primal-dual interior-point methods

more efficient than barrier method when high accuracy is needed

- update primal and dual variables at each iteration; no distinction between inner and outer iterations
- often exhibit superlinear asymptotic convergence
- search directions can be interpreted as Newton directions for modified KKT conditions
- can start at infeasible points
- cost per iteration same as barrier method

# Interior-point methods: summary

- Interior point methods (IPM) are very reliable on small scale problems.
  - Example: SDP of dimension 100, SOCP with less than a thousand variables.
  - Most conic problems with a couple of hundred variables can formulated and solved very quickly using preprocessors such as CVX.
- IPM often efficient on larger problems if KKT system has some structure (sparsity, blocks, etc).
  - Large scale linear programs with thousands of variables are routinely solved by free or commercial solvers using IPM (e.g. SDPT3, MOSEK, GLPK, CPLEX, etc.).
  - Much larger sparse LPs can also be solved efficiently using the same techniques.
- Not workable for very large problems.
  - For some problems, e.g. semidefinite programs, exploiting structure in IPM is hard.
  - First order methods (using the gradient only) seem to be the only option for extremely large problems

# Semidefinite programming: CVX

Solving the maxcut relaxation

$$\begin{array}{ll} \text{max.} & \mathbf{Tr}(XC) \\ \text{s.t.} & \mathbf{diag}(X) = \mathbf{1} \\ & X \succeq 0, \end{array}$$

is written as follows in CVX/MATLAB

```
cvx_begin
.  variable X(n,n) symmetric
.  maximize trace(C*X)
.  subject to
.    diag(X)==1
.    X==semidefinite(n)
cvx_end
```

*

References