Lecture 7: Unconstrained Minimization

Xiugang Wu

University of Delaware

Fall 2019

Outline

- Introduction
- Gradient Descent Method
- Steepest Descent Method
- Newton's Method
- Self-Concordant Functions

Outline

- Introduction
- Gradient Descent Method
- Steepest Descent Method
- Newton's Method
- Self-Concordant Functions

Unconstrained Minimization

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

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

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

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

- 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 $S = \{x | f(x) \leq f(x^{(0)})\}$ is closed

2nd condition is hard to verify, except when all sublevel sets are closed:
- true if $\mathbf{dom}\ f = \mathbf{R}^n$
- true if $f(x) \to \infty$ as $x \to \mathbf{bd}\ \mathbf{dom}\ f$

examples of differentiable functions with closed sublevel sets:

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

Strong Convexity and Implications

f is strongly convex on S if there exists some $m \geq 0$ such that $\nabla^2 f(x) \succeq mI$. For any $x, y \in S$ we have

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

for some z on the line segment $[x, y]$. Combining this with strong convexity:

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

- When $m = 0$, we recover the basic inequality characterizing convexity; when $m > 0$ we obtain a better bound on $f(y)$.

- S is bounded

- $f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|_2^2$; $\|\nabla f(x)\|_2 \leq \sqrt{2m\epsilon} \Rightarrow f(x) - p^* \leq \epsilon$

- $\|x^* - x\|_2 \leq \frac{2}{m}\|\nabla f(x)\|_2$; the optimal point is unique

Descent Method

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

- other notations: $x^+ = x + t\Delta x$ or $x := x + t\Delta x$

- $\Delta x^{(k)}$ is the step, or search direction; $t^{(k)} > 0$ is the step size, or step length

- from convexity, $f(x^{(k+1)}) < f(x^{(k)})$ implies $\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$

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.

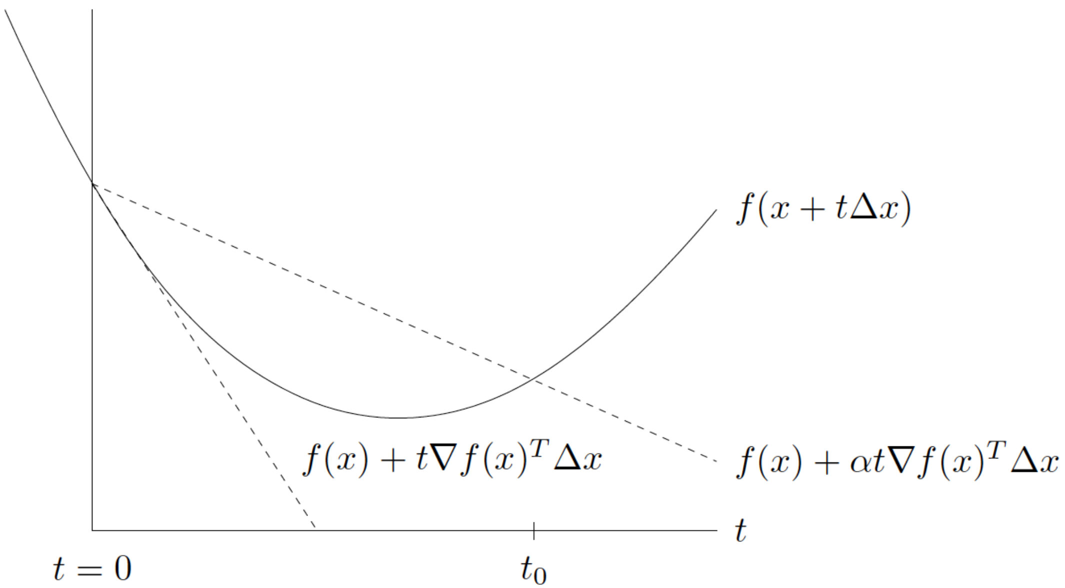
Line Search Types

Exact line search: $t = \arg\min_{s \geq 0} f(x + s\Delta x)$

Backtracking line search:
- Given a descent direction Δx, and $\alpha \in (0, 0.5)$ and $\beta \in (0, 1)$
- Start at $t = 1$, repeat $t := \beta t$ until

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

- such a t always exists as long as it is small enough.



Outline

- Introduction
- Gradient Descent Method
- Steepest Descent Method
- Newton's Method
- Self-Concordant Functions

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,

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

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

Example

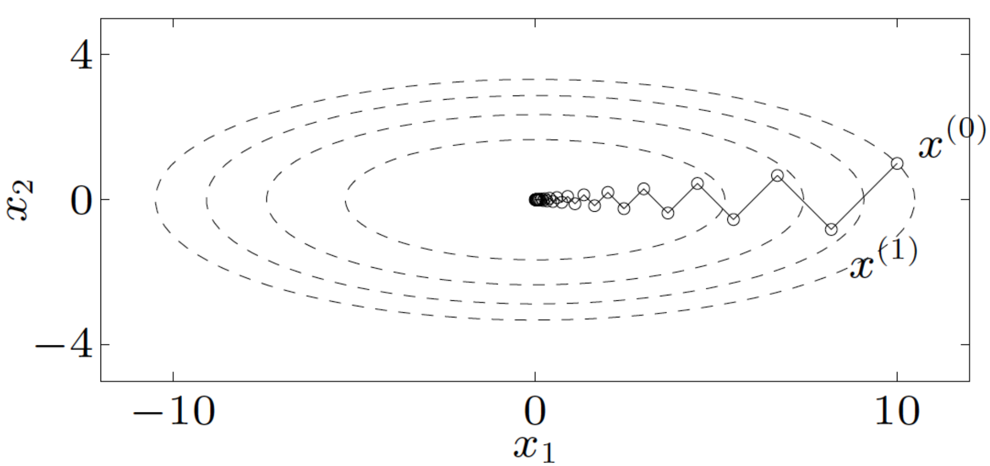
Quadratic problem in $\mathbf{R}^2$

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

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

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

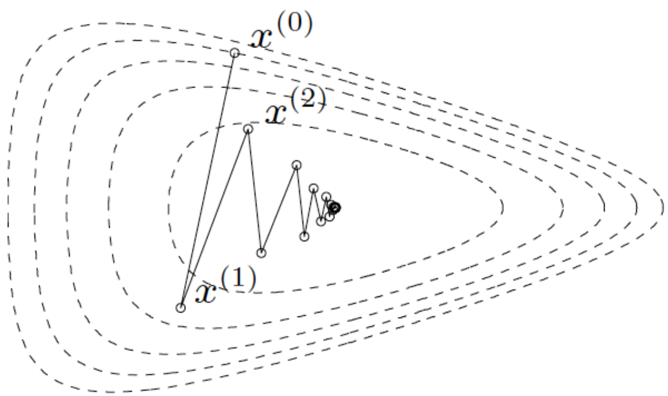
- error reduced by a factor of $\frac{\gamma-1}{\gamma+1}$ at each iteration
- very slow if $\gamma \gg 1$ or $\gamma \ll 1$
- example for $\gamma = 10$



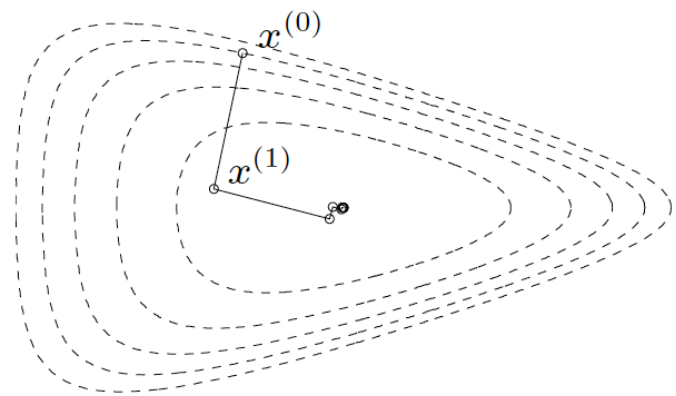
Example

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

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



backtracking line search

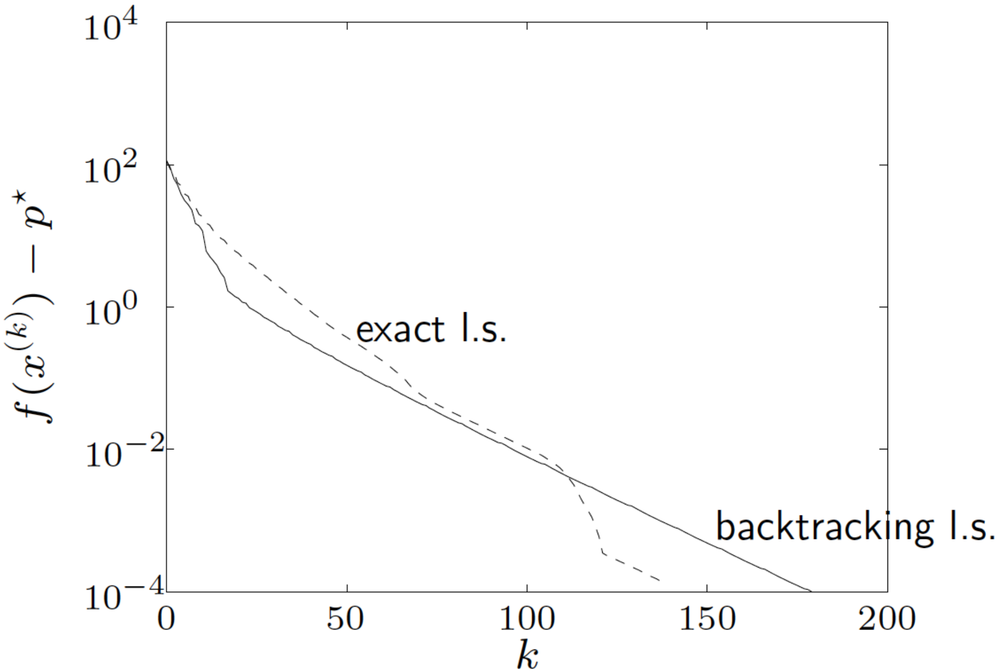


exact line search

Example

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

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



"linear" convergence—a straight line on a semilog plot

Outline

- Introduction
- Gradient Descent Method
- Steepest Descent Method
- Newton's Method
- Self-Concordant Functions

Steepest Descent Method

The first order Taylor approximation:

$$f(x+v) \approx f(x) + \nabla f(x)^T v$$

- How to choose v to make the $\nabla f(x)^T v$ as negative as possible?
- To make the question sensible, we limit the size of v

normalized steepest descent direction (at x w.r.t. the norm $\|\cdot\|$):

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

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

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

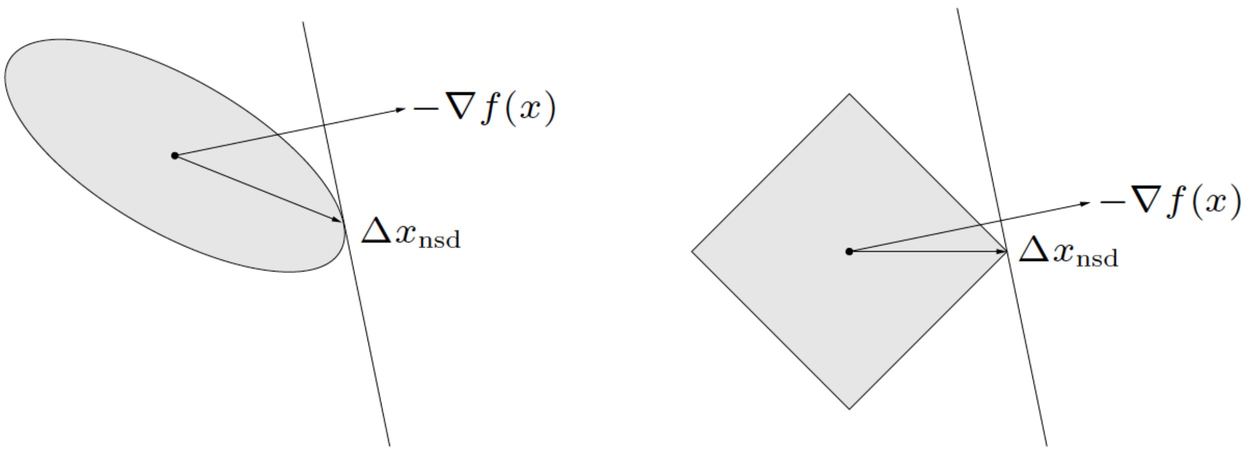
Examples

- Euclidean norm: $\Delta x_{\mathrm{sd}} = -\nabla f(x)$

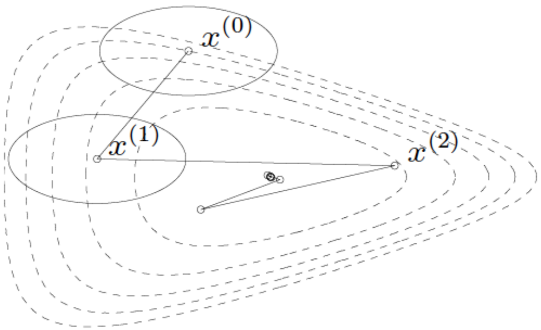
- Quadratic norm $\|x\|_P = (x^T P x)^{1/2} = \|P^{1/2} x\|_2$ where $P \succ 0$:

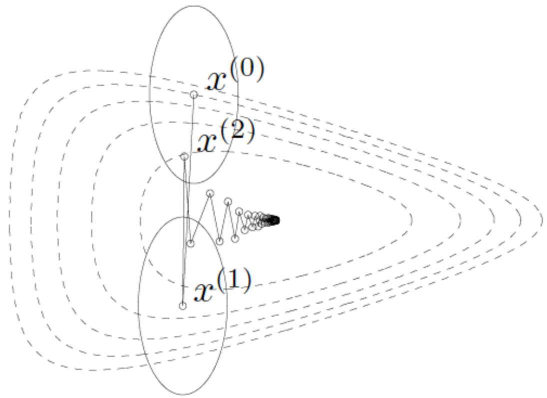
$$\Delta x_{\mathrm{sd}} = -P^{-1} \nabla f(x)$$

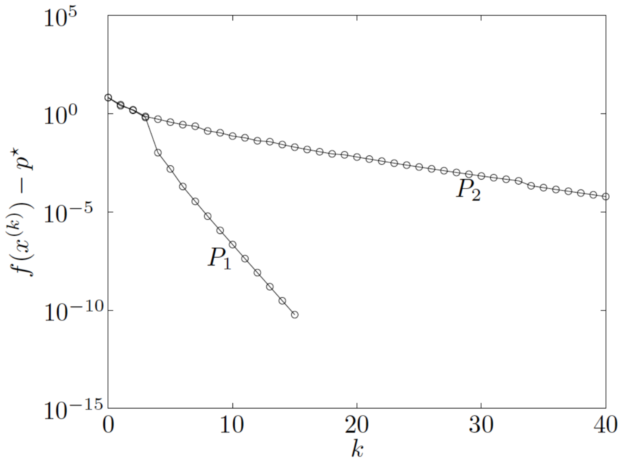
- ℓ_1-norm: $\Delta x_{\mathrm{sd}} = -\frac{\partial f(x)}{\partial x_i} e_i$ where $\frac{\partial f(x)}{\partial x_i} = \|\nabla f(x)\|_\infty$



Choice of Norm for Steepest Descent







- steepest descent with backtracking line search for two norms P_1 and P_2

- ellipses show $\{x \mid \|x - x^{(k)}\|_P = 1\}$

- choice of P has strong effect on speed of convergence; optimist vs. pessimist

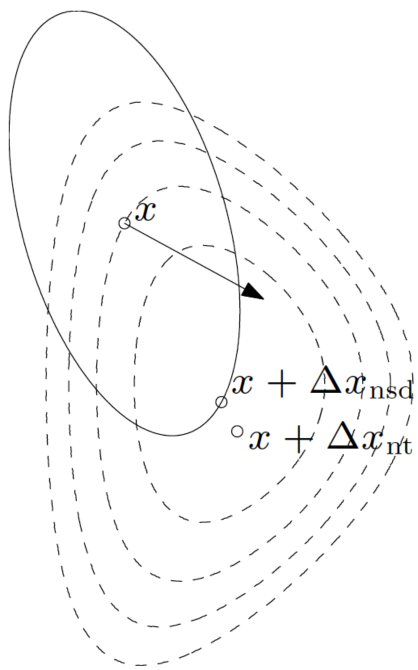
Outline

- Introduction
- Gradient Descent Method
- Steepest Descent Method
- Newton's Method
- Self-Concordant Functions

Newton Step

$\Delta x_{\rm nt}$ is steepest descent direction at x w.r.t. local Hessian norm:

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



- dashed lines: contour lines of f
- ellipse: $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$
- arrow: $-\nabla f(x)$

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

affine invariant: Consider $f(x)$ and $\bar{f}(y) = f(Ty)$ with nonsingular T.

$$x + \Delta x_{\rm nt} = T(y + \Delta y_{\rm nt})$$

Interpretations

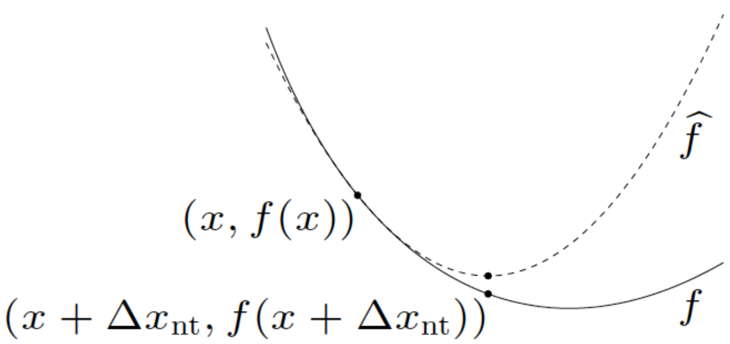
$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

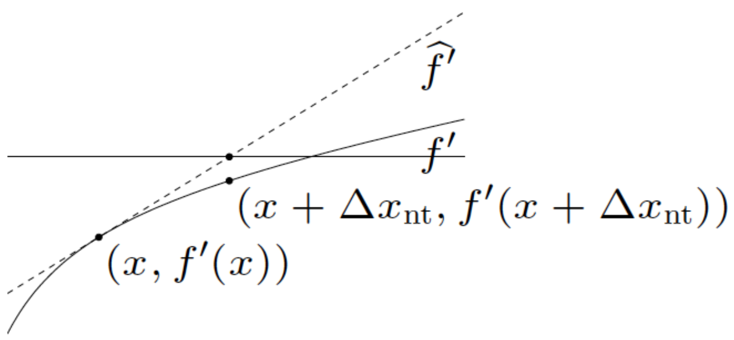
- $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

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

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

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





Newton Decrement

a measure of the proximity of x to x^*:

$$\begin{aligned} \lambda(x) &= (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{1/2} \\ &= \|\Delta x_{\mathrm{nt}}\|_{\nabla^2 f(x)} \end{aligned}$$

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

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

- as in general steepest descent,

$$\nabla f(x)^T \Delta x_{\mathrm{nt}} = -\|\Delta x_{\mathrm{nt}}\|_{\nabla^2 f(x)}^2 = -\lambda(x)^2$$

therefore it comes up in backtracking line search

- affine invariant (unlike $\|\nabla f(x)\|_2$):

$$f(x) = \bar{f}(y) \text{ for } x = Ty \Rightarrow \lambda_f(x) = \lambda_{\bar{f}}(y) \text{ for } x = Ty$$

Newton's Method

---

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

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

- backtracking line search: repeat $t := \beta t$ until

$$\begin{aligned} f(x + t\Delta x_{\rm nt}) &\leq f(x) + \alpha t \nabla f(x)^T \Delta x_{\rm nt} \\ &= f(x) - \alpha t \lambda(x)^2 \end{aligned}$$

- progress independent of affine change of coordinates. Newton iterates for $\bar{f}(y) = f(Ty)$ with starting point $y^{(0)} = T^{-1}x^{(0)}$ are:

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

Convergence Analysis

Assumptions:

- f strongly convex on S with constant m

- $\nabla^2 f$ is Lipschitz continuous on S, with constant L:

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

I.e., L measures how well f can be approximated by a quadratic function

Result: there exists 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

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

Two-Phase Convergence

Damped Newton Phase ($\|\nabla f(x)\|_2 \geq \eta$):
- most iterations requires backtracking steps
- at each iteration, function value decreases by at least γ
- 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

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

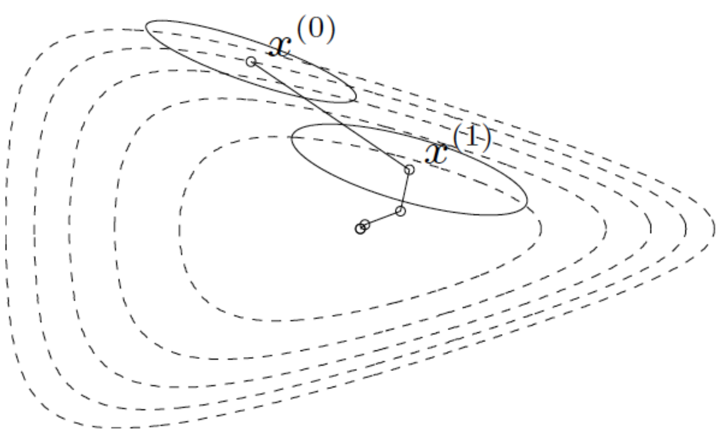
Conclusion: total number of iterations until $f(x) - p^* \leq \epsilon$ is upper bounded by

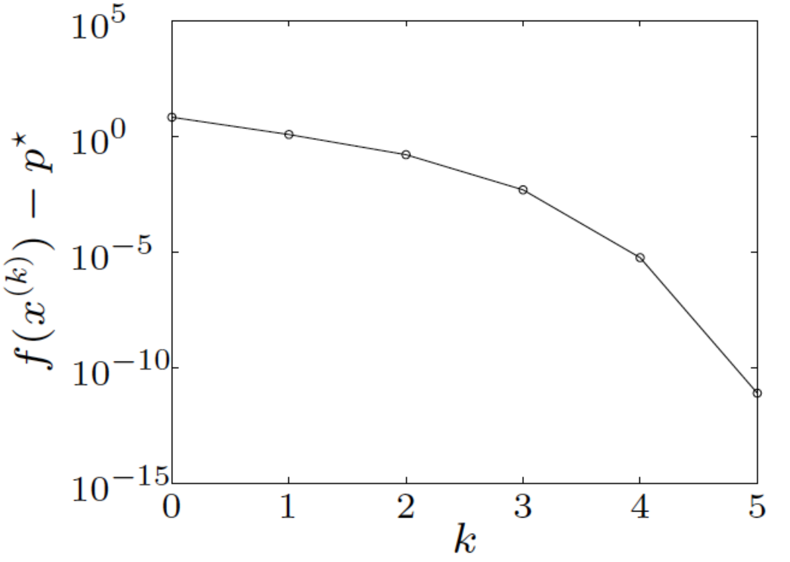
$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2 \left(\frac{\epsilon_0}{\epsilon}\right)$$

- γ, ϵ_0 are constants that depend on $m, L, x^{(0)}$
- second terms small (of the order of 6); almost constant for practical purposes
- in practice, constants m, L (hence γ, ϵ_0) are usually unknown
- provides qualitative insight in two-phase convergence

Examples

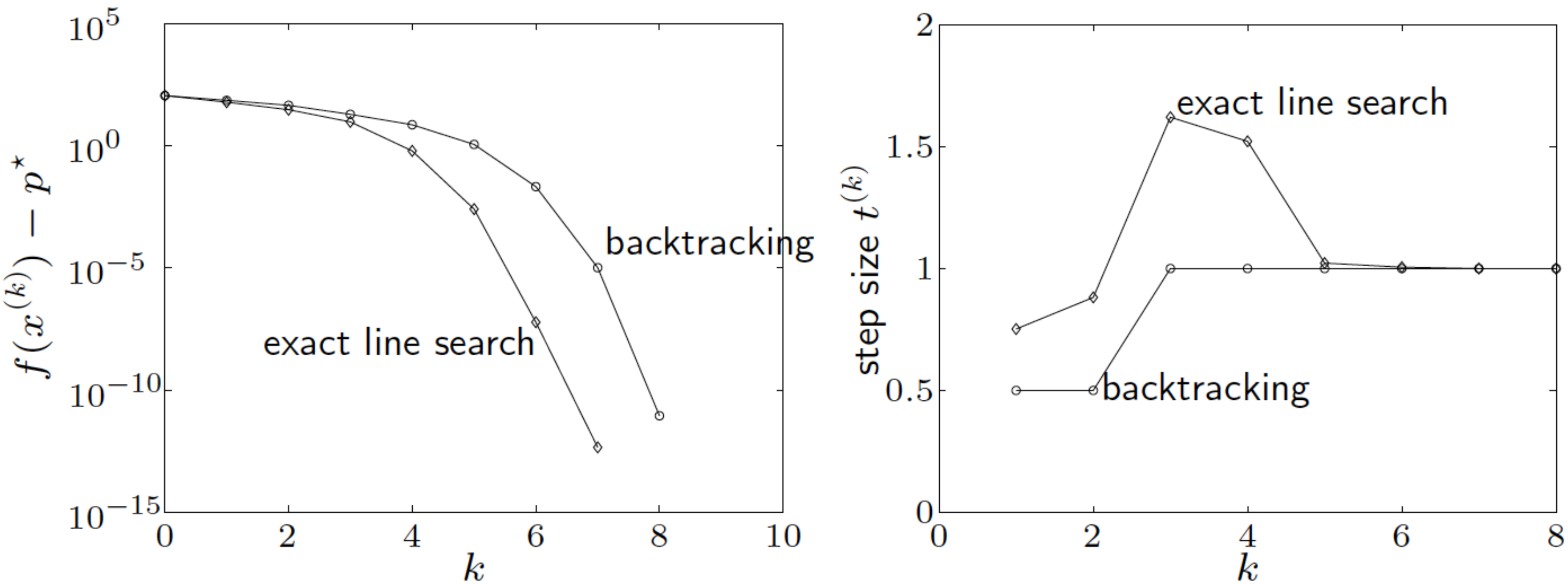
Example in $\mathbf{R}^2$
- backtracking parameters $\alpha = 0.1$, $\beta = 0.7$
- converges in only 5 steps
- quadratic local convergence





Examples

Example in $\mathbf{R}^{100}$
- backtracking parameters $\alpha = 0.01$, $\beta = 0.5$
- backtracking line search almost as fast as exact line search
- clearly shows two phases in algorithm



Outline

- Introduction
- Gradient Descent Method
- Steepest Descent Method
- Newton's Method
- Self-Concordant Functions

Self-Concordance

shortcomings of classical convergence analysis
- depends on unknown constants $(m, L, \dots)$
- 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:
- convex $f : \mathbf{R} \to \mathbf{R}$ is self-concordant if $|f'''(x)| \leq 2f''(x)^{3/2}$ for all $x \in \text{dom } f$
- convex $f : \mathbf{R}^n \to \mathbf{R}$ is $g(t) = f(x + tv)$ is self-concordant for all $x \in \text{dom } f$ and $v \in \mathbf{R}^n$

examples:
- linear and quadratic functions
- $f(x) = -\log x$
- $f(x) = x \log x - \log x$

affine invariance: If $f : \mathbf{R} \to \mathbf{R}$ is self-concordant, then $\bar{f}(y) = f(ay + b)$ is self-concordant:

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

Convergence Analysis For Self-Concordant Functions

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

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

Here η, γ only depend on backtracking parameters α, β.

Complexity bound: number of Newton iterations bounded by

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2 \left(\frac{1}{\epsilon}\right)$$