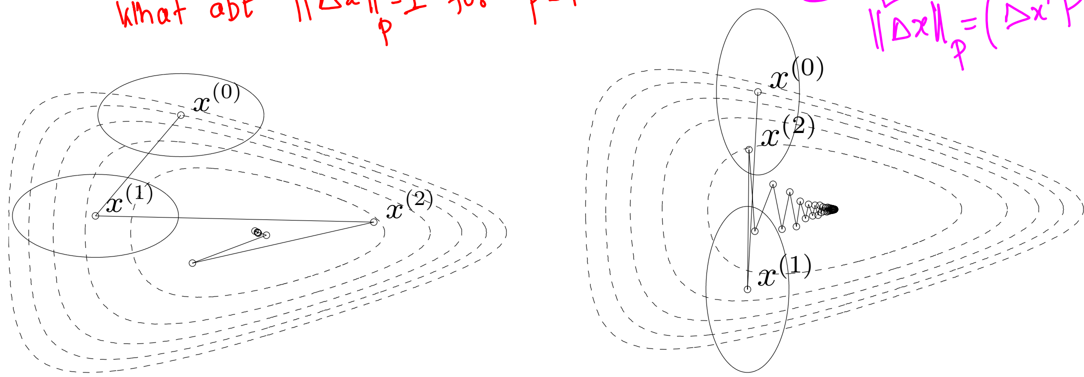
choice of norm for steepest descent

What abt $\|\Delta x\|_p = 1$ for $p=1$ or ∞ or matrix induced norm?

$\|\Delta x\|_P = (\Delta x^T P \Delta x)^{1/2}$



- steepest descent with backtracking line search for two quadratic norms
- ellipses show $\{x \mid \|x - x^{(k)}\|_P = 1\}$: Ellipses show search space for Δx
- 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

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

Adaptive steepest descent
(adaptive in curvature)
adaptive through $\nabla^2 f(x)$

interpretations

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

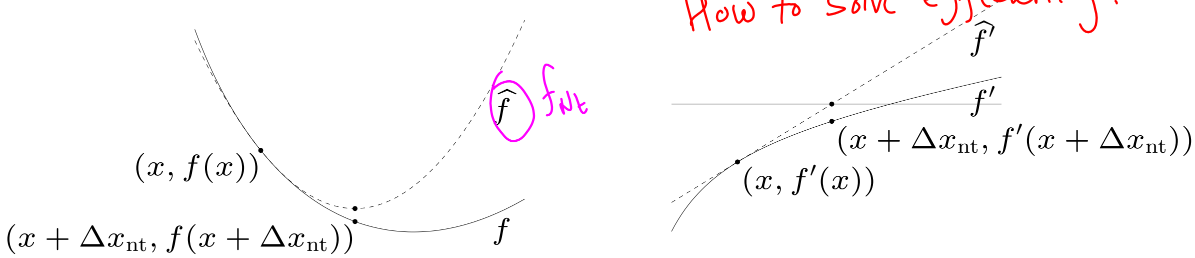
$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

exact Taylor series has $\nabla^2 f(y)$ for some $y = x + \alpha v$, $\alpha \in (0,1)$

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

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

How to solve efficiently?



Q: How to efficiently solve for Δx in Newton update?

$$\nabla f(x^k) + \nabla^2 f(x^k)\,\Delta x = 0$$

Ans ① Dont invert $\nabla^2 f(x)$ though $\Delta x = -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x^k)$

② Gauss Elimination $\doteq n^3$: You have not accounted for positive semi-definiteness (& symmetry) of $\nabla^2 f(x)$ since f is convex

③ Cholesky decomposition: $\nabla^2 f(x^k) = L L^T$ $L =$ 

$n^3/6 \longrightarrow$ constant factors matter since Δx is computed in every iteration

$$L \underbrace{L^T \Delta x}_{y} = -\nabla f(x^k)$$

$L y = -\nabla f(x^k)$ . . . Solve by forward substitution in $O(n^2)$

$L^T \Delta x = y$ . . . $O(n^2)$

Quasi Newton methods find even $n^3/6$ unacceptable & try & avoid computing $\nabla^2 f(x)$ by moving in a subspace that potentially spans the $\nabla^2 f(x)$

Is $\Delta x^k = -\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$ a (valid) descent direction?

Q is $\nabla^T f(x^k)\, \Delta x^k < 0$

ie is $-\nabla^T f(x^k) \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k) < 0$ → Newton decrement

if f is strictly convex at x^k

$x^{k+1} = x^k + t^k \Delta x^k$: Q what if $\nabla^2 f(x^k)$ is not invertible / positive definite

→ You could still solve $\nabla^2 f(x^k) \Delta x^k = -\nabla f(x^k)$

→ But since $\nabla^2 f(x^k)$ is p.s.d, Δx^k does not guarantee decrement

→ Line search over t might yield $t^k = 0$ if no descent is possible along Δx^k

Approx soln You can solve $\Delta x^k = -\left(\nabla^2 f(x^k) + \sigma I\right)^{-1} \nabla f(x^k)$

Problem: Find $H^k \approx \nabla^2 f(x^k)$ but positive definite

Quasi Newton: $H^k \approx \left(\nabla^2 f(x^k)\right)^{-1}$ if $\nabla^2 f(x^k)$ is positive definite

& $H^k \approx \left(\nabla^2 f(x^k) + \sigma I\right)^{-1}$ o/w

→ Finding successive approx to inverse of Hessian

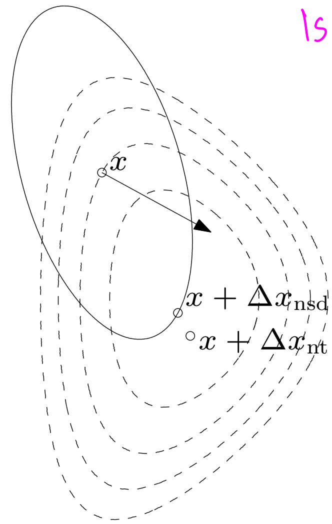
- $\Delta x_{\rm nt}$ is steepest descent direction at x in local Hessian norm

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

Is Δx_{Nt} a descent direction?

$\Delta x_{Nt}^T \cdot \nabla f(x) < 0$

Newton decrement



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

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

- directional derivative in the Newton direction: $\nabla f(x)^T \Delta x_{\rm nt} = -\lambda(x)^2$
- 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.* $x := x + t\Delta x_{\mathrm{nt}}$.

Newton step could be exact or approx (if psd)

→ Decrement for convergence works for Newton since Quad approx is more robust with $\nabla^2 f(\cdot)$

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

$x^{(k)} = Ty^{(k)} \leftarrow y^{(k)} = T^{-1}x^{(k)}$ $\qquad x^{(0)} = Ty^{(0)}$

$f(x) \Longleftrightarrow f(Ty)$

$Ax = b \Leftrightarrow x = x_{part} + x_{Nullspace}$ if $b \in \mathrm{Col}(A)$

ie $x = \underbrace{\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} x_{part} \\ x_{Null} \end{bmatrix}}_{Y}$

eg: $\log(\det(X)) = \log(\det(TY))$

Classical convergence analysis

assumptions

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

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

(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 $\|\nabla f(x)\|_2 \ge \eta$, then $f(x^{(k+1)}) - f(x^{(k)}) \le -\gamma$ $\qquad$ ∴ Gradient is sufficiently non-zero ⇒ Sublinear
- if $\|\nabla f(x)\|_2 < \eta$, then

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

∴ Quadratic

$$f_{Q_{N_k}}(y) = f(x^k) + \cancel{\nabla^T f(x^k)}\,\cancel{(y - x^k)} + \frac{1}{2}(y - x^k)^T \nabla^2 f(x^k)(y - x^k)$$

0

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

when $\|\nabla f(x^k)\|_2 < \eta$

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

Need to eliminate

$$\Delta x^k = \underset{\Delta x}{\operatorname{argmin}}\ f(x^k) + \nabla^T f(x^k)\Delta x + \frac{1}{2}\Delta x^T \nabla^2 f(x^k)\Delta x$$

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

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

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

H/W

- 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)

(# of iterations for $f(x^k) - f(x^*) \leq \epsilon$)

Gradient descent:

f is Lipschitz: $k = O(1/\epsilon^2)$

$\nabla f(x)$ is Lipschitz: $k = O(1/\epsilon)$

f is strongly convex & Lipschitz: $k = O(\log(1/\epsilon))$

Newton:

$\nabla^2 f$ is Lipschitz & f is strongly convex → Damped sublinear; → Quad conv. $O(\log\log(1/\epsilon))$

Stochastic Gradient: $f = E(f_i(x))$

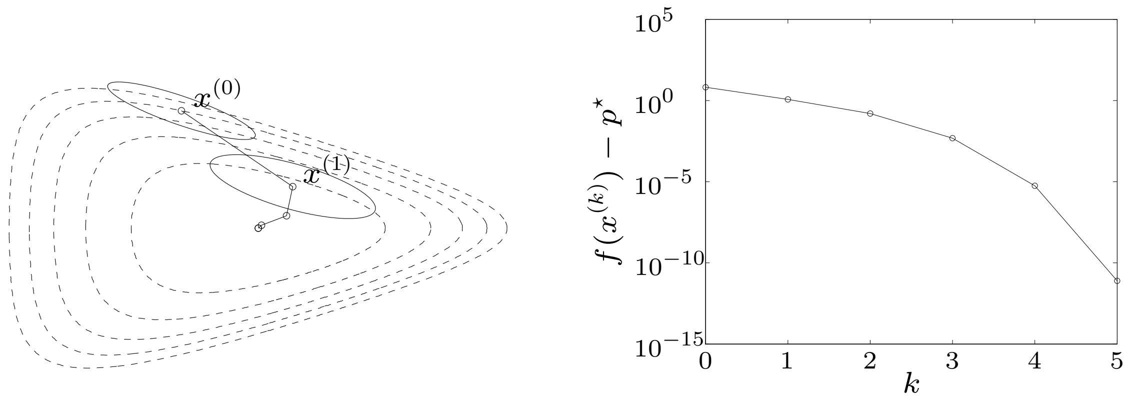
At i^{th} iteration: $x^{k+1} = x^k - \nabla f_i(x^k)$

f is Lipschitz: $O(1/\epsilon^2)$

f is strongly convex & ∇f is Lipschitz: $O\left(\frac{\log(1/\epsilon)}{\epsilon}\right)$

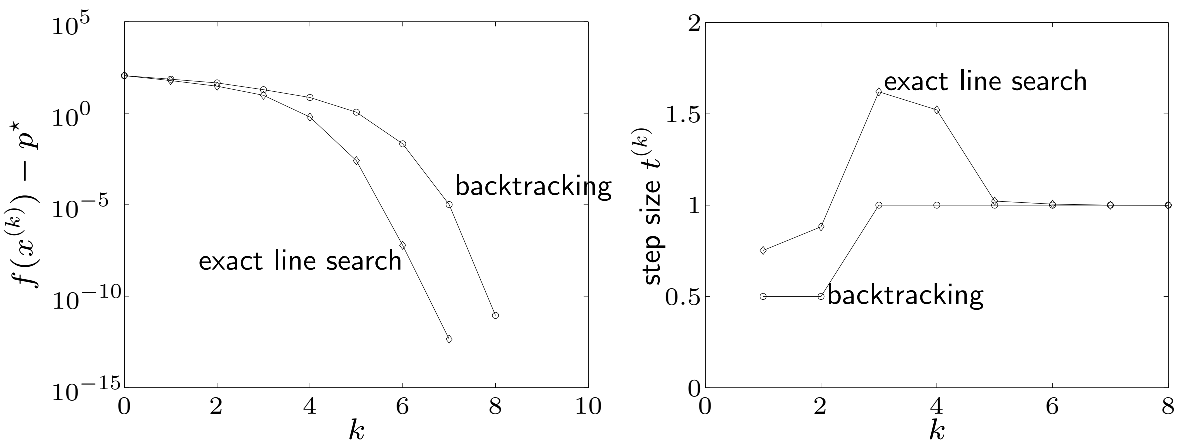
Examples

example in $\mathbf{R}^2$ (page 10–9)



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

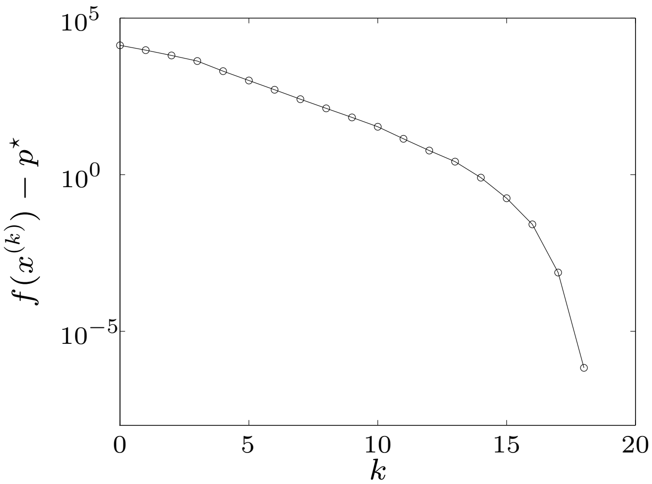
example in $\mathbf{R}^{100}$ (page 10–10)



- 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

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

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 $A \in \mathbf{R}^{p\times n}$, 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 (page 9–15): 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 Δx from first equation; compute w and Δx 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 $L_0^T A D^{-1} A^T L_0$)