Newton’s Method

CPSC 406 – Computational Optimization

Newton’s Method

- quadratic approximation
- interpretation as scaled descent
- Cholesky factorization

Gradient descent

- suppose f is L-smooth, i.e. $\|\nabla^2 f(x)\| \leq L$ for all x

$$\min_{x\in\mathbb{R}^n} f(x), \quad f_k = f(x^k), \quad g_k = \nabla f(x^k), \quad H_k = \nabla^2 f(x^k)$$

- quadratic approximation of f at x^k

$$\begin{aligned} q_k(x) &:= f_k + g_k^T(x - x^k) + \tfrac{1}{2}(x - x^k)^T H_k (x - x^k) \\ &\leq f_k + g_k^T(x - x^k) + \tfrac{1}{2} L\|x - x^k\|^2 =: \hat{q}_k(x) \end{aligned}$$

- minimizer $\hat{x}$ of upper bound $\hat{q}_k(x)$ satisfies

$$0 = \nabla \hat{q}_k(\hat{x}) = g_k + L(\hat{x} - x^k)$$

- solve for solution $\bar{x}$ to obtain gradient descent with $\alpha = 1/L$

$$\bar{x} = x^k - \frac{1}{L} g_k$$

Newton's method

- 2nd-order approximation of f at x^k

$$q_k(x) = f_k + g_k^T(x - x^k) + \tfrac{1}{2}(x - x^k)^T H_k (x - x^k), \quad g_k = \nabla f(x^k), \quad H_k = \nabla^2 f(x^k) \succ 0$$

- let $\bar{x}$ be the minimizer of $q_k(x)$, ie,

$$0 = \nabla q_k(\bar{x}) = g_k + H_k(\bar{x} - x^k) \quad \Longleftrightarrow \quad \bar{x} = x^k - H_k^{-1} g_k$$

- **pure Newton's method** chooses next iterate $x^{k+1} = \bar{x}$

$$x^{k+1} = x^k + \underbrace{d_N^k}_{=\text{Newton direction}}, \qquad H_k d_N^k = -g_k$$

- **damped Newton's method** chooses next iterate with step $\alpha \le 1$

$$x^{k+1} = x^k + \alpha d_N^k, \qquad H_k d_N^k = -g_k$$

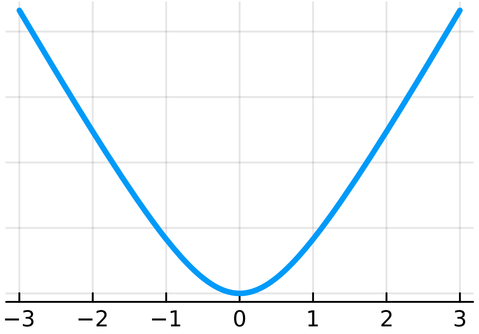
Convergence of Newton's method

- require $\nabla^2 f(x^k) \succ 0$ for all k to ensure **descent**
- may still diverge even if $\nabla^2 f(x^k) \succ 0$ for all k — eg, if $\lambda_{\min}(H_k)$ is small

Example

$$f(x) = \sqrt{1+x^2}, \quad \nabla f(x) = \frac{x}{\sqrt{1+x^2}}, \quad \nabla^2 f(x) = \frac{1}{(1+x^2)^{3/2}}$$

- Newton iteration



$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)} = -(x^k)^3$$

- convergence of iterates depends on initial point

$$x^k \to \begin{cases} 0 & \text{if } |x^0| < 1 \\ \pm 1 & \text{if } |x^0| = 1 \\ \infty & \text{if } |x^0| > 1 \end{cases}$$

Convergence of Newton's method

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and
  - $\nabla^2 f(x^k) \succ \epsilon I$ for some $\epsilon > 0$ and all x
  - $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$ for all x, y for some $L > 0$

- Newton iterations satisfy if x^k is sufficiently close to x^*

$$\|x^{k+1} - x^*\| \leq \frac{L}{2\epsilon}\|x^k - x^*\|^2$$

- in addition, if $\|x^{(0)} - x^*\| \leq \epsilon/L$, then iterates obtain local quadratic convergence

$$\|x^{k+1} - x^*\| \leq \left(\frac{2\epsilon}{L}\right)\left(\frac{1}{4}\right)^{2^k}$$

Example

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

Newton

k	fval
1	1.0100e+02
2	6.7230e+01
3	1.9074e+00
4	1.5506e+00
5	1.1674e+00
6	8.3524e-01
7	6.1188e-01
8	3.8893e-01
9	3.8636e-01
10	1.3032e-01
11	9.0166e-02
12	3.1699e-02
13	2.9670e-02
14	1.3869e-03
15	1.7446e-04

Gradient descent

k	fval
1	1.0100e+02
100	1.4702e+00
200	1.4543e+00
300	1.4345e+00
400	1.4200e+00
500	1.4059e+00
600	1.3918e+00
700	1.3776e+00
800	1.3633e+00
900	1.3490e+00
1000	1.3347e+00

Cholesky factorization

Positive definite matrices

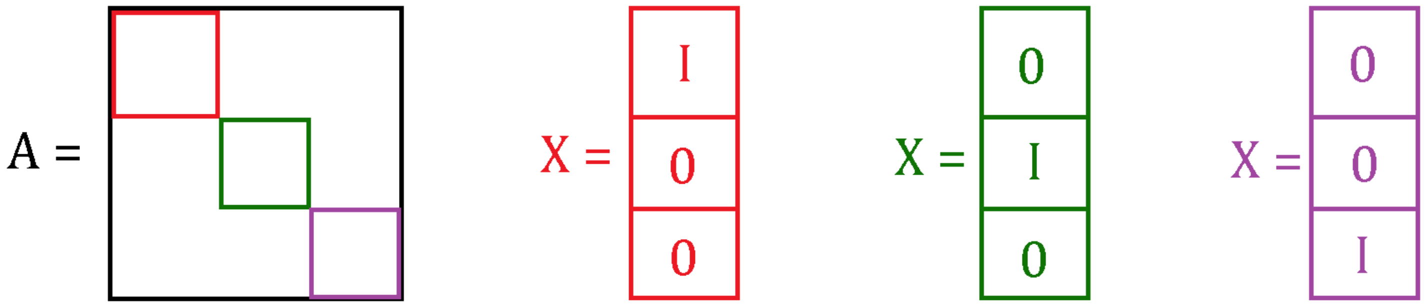
- an $n \times n$ matrix A is positive definite if

$$x^T A x > 0 \quad \text{for all} \quad x \neq 0$$

- all eigenvalues of A are positive

$$0 < x^T A x = x^T(\lambda x) = \lambda x^T x = \lambda \|x\|^2$$

- $A \succ 0 \quad \iff \quad X^T A X \succ 0$ for all X full rank
- every principle submatrix $A_{\mathcal{I},\mathcal{I}}$ is positive definite, eg, diagonals are positive



Cholesky factorization

- if $A \succ 0$, then

$$A = \begin{bmatrix} a_{11} & w^T \\ w & K \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha & \\ w/\alpha & I \end{bmatrix}}_{R_1^T} \underbrace{\begin{bmatrix} 1 & \\ & K - ww^T/\alpha^2 \end{bmatrix}}_{A_1} \underbrace{\begin{bmatrix} \alpha & w^T/\alpha \\ & I \end{bmatrix}}_{R_1} \quad \alpha := \sqrt{a_{11}}$$

- $A \succ 0 \iff K - ww^T/\alpha^2 \succ 0$, thus apply above factorization to $K - ww^T/\alpha^2$:

$$K - ww^T/\alpha^2 = \bar{R}_2^T \bar{A}_2 \bar{R}_2,$$

- recursively apply to obtain $A = R^T R$

$$A = R_1^T \begin{bmatrix} 1 & \\ & \bar{R}_2^T \bar{A}_2 \bar{R}_2 \end{bmatrix} R_1 = R_1^T \underbrace{\begin{bmatrix} 1 & \\ & \bar{R}_2^T \end{bmatrix}}_{R_2^T} \underbrace{\begin{bmatrix} 1 & \\ & \bar{A}_2 \end{bmatrix}}_{A_2} \underbrace{\begin{bmatrix} 1 & \\ & \bar{R}_2 \end{bmatrix}}_{R_2} R_1$$

$$= R_1^T R_2^T A_2 R_2 R_1 = \cdots = \underbrace{(R_1^T R_2^T \cdots R_n^T)}_{R^T} \underbrace{(R_n \cdots R_2 R_1)}_{R}$$

Cholesky factorization (summary)

- an $n \times n$ matrix A is positive definite if and only if

$$A = R^T R \qquad \text{for some nonsingular upper triangular } R$$

- requires $(1/3)n^3$ flops vs $(2/3)n^3$ for LU factorization

```julia
using LinearAlgebra
A = [4 12 -16; 12 37 -43; -16 -43 98]
R = cholesky(A)
R.L
```

```
3×3 LowerTriangular{Float64, Matrix{Float64}}:
  2.0   ⋅    ⋅ 
  6.0  1.0   ⋅ 
 -8.0  5.0  3.0
```

```julia
R.L * R.L' ≈ A
```

```
true
```

```julia
A[3,3] = -1
R = try
  cholesky(A)
catch
  "Matrix is not positive definite"
end
```

```
"Matrix is not positive definite"
```

Solving for Newton direction

- Newton direction d_N^k solves

$$H_k d_N^k = -g_k, \quad \text{where} \quad H_k = \nabla^2 f(x^k), \quad g_k = \nabla f(x^k)$$

- solve for Newton step via

Cholesky

1. $\tau = 0$
2. $(H_k + \tau I) = R^T R$
   - if Cholesky fails, increase τ and repeat
3. solve $R^T R d_N^k = -g_k$

Eigenvalue decomposition

1. choose $\epsilon > 0$ small
2. $H_k = U \Lambda U^T, \quad \Lambda = \mathbf{Diag}(\lambda_1, \ldots, \lambda_n)$
3. $\bar{\lambda}_i = \max(\lambda_i, \epsilon)$
4. $\bar{\Lambda} = \mathbf{Diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_n)$
5. solve $U \bar{\Lambda} U^T d_N^k = -g_k$

Factorizations

- $A = QR$ can be used to solve linear systems or least-squares problems

$$Ax = b \iff Rx = Q^T b$$

- if $A \succ 0$, other factorizations are available:
  - diagonalization: U orthogonal, Λ diagonal

$$A = U\Lambda U^T, \quad Ax = b \iff \Lambda y = U^T b, \quad x = Uy$$

  - Cholesky: $R \succ 0$ lower triangular

$$A = R^T R, \quad Ax = b \iff \underbrace{R^T y = b,}_{\text{backsolve}} \quad \underbrace{Rx = y}_{\text{forward solve}}$$

- why?
  - inverting a matrix can be numerically unstable
  - factorizations can be reused for multiple right-hand sides
  - multiplying orthogonal matrices is numerically stable and solving triangular/diagonal systems is easy