13. Cholesky & Quasi Newton

- Cholesky decomposition
- PD check.
- Quasi- Newton.

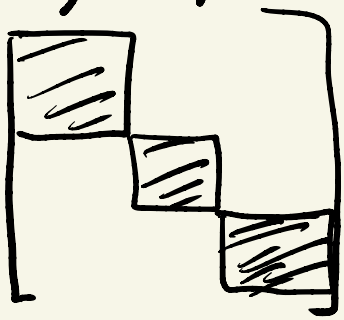
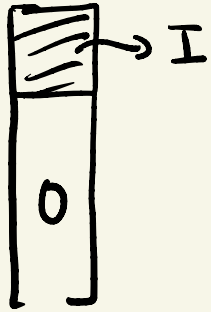
Cholesky Decomposition.

Cholesky Decomposition can verify if a matrix is Positive definite and help solving the corresponding linear system.

A $n \times n$ symmetric matrix is positive definite if

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

- $A \succeq 0 \iff X^T A X \succeq 0$ if X has full column rank.
- Every principal submatrix $A_{I,I}$ of A is PD.

$A =$  , $X =$ 

$\Rightarrow X^T A X \succeq 0.$

Cholesky factorization

$$A = LL^T$$

- How to calculate

$$A = \begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} l_{11} & L_{21}^T \\ 0 & L_{22}^T \end{bmatrix}$$

$$= \begin{bmatrix} l_{11}^2 & l_{11} L_{21}^T \\ l_{11} L_{21} & L_{21} L_{21}^T + L_{22} L_{22}^T \end{bmatrix}$$

so,

$$l_{11} = \frac{1}{\sqrt{a_{11}}}, \quad L_{21} = \frac{1}{\sqrt{a_{11}}} A_{21}, \quad \underbrace{L_{22} L_{22}^T = A_{22} - L_{21} L_{21}^T}_{\text{next Cholesky system}}$$

Q: What happens if $a_{11} \leq 0$?

Example

$$A = \begin{pmatrix} 9 & 3 & 3 \\ 3 & 17 & 21 \\ 3 & 21 & 107 \end{pmatrix}$$

$$L = \begin{pmatrix} l_{11} & & \\ l_{21} & l_{22} & \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \Rightarrow l_{11} = 3, \quad \begin{pmatrix} l_{21} \\ l_{22} \end{pmatrix} = \frac{1}{\sqrt{9}} \begin{pmatrix} 3 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\Rightarrow L_{22} L_{22}^T = \begin{bmatrix} 17 & 21 \\ 21 & 107 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 16 & 20 \\ 20 & 106 \end{bmatrix}$$

$$\Rightarrow l_{22} = 4, \quad l_{32} = \frac{1}{\sqrt{16}} \cdot 20 = 5,$$

Lastly, we find Cholesky factorization of $106 - \frac{1}{16}(20 \cdot 20) = 81$

so, $l_{33} = 9$. $\Rightarrow L = \begin{bmatrix} 3 & & \\ 1 & 4 & \\ 1 & 5 & 9 \end{bmatrix}$ //

Positive definite?

Is the matrix $A = \begin{bmatrix} 2 & 4 & 7 \\ 4 & 6 & 7 \\ 7 & 7 & 4 \end{bmatrix}$ positive definite?

$l_{11} = \sqrt{2}$, $\begin{pmatrix} l_{21} \\ l_{31} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 4 \\ 7 \end{pmatrix}$

$$L_{22} L_{22}^T = A_{22} - L_{21} L_{21}^T = \begin{bmatrix} 6 & 7 \\ 7 & 4 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 16 & 28 \\ 28 & 49 \end{bmatrix}$$

$$= \begin{bmatrix} -2 & -7 \\ -7 & -\frac{41}{2} \end{bmatrix}$$

So, A is not positive definite.

Drawbacks of Newton's method.

Newton update:

$$x_{k+1} = x_k - \alpha_k \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k).$$

- Newton's method is sensitive to initialization. Newton direction $\left(-\nabla^2 f(x_k)^{-1} \nabla f(x_k)\right)$ is a descent direction if $\nabla^2 f(x_k) \succ 0$.
  Could take us to saddle point, maximum. Only appropriate for convex optimization.
- Newton's method is computationally expensive.
  $\left(O(n^2) + O(n^3)\right)$

Low rank update.

Let $P \in \mathbb{R}^{n \times n}$ be invertible. What is the inverse of $P + vv^T$, $v \in \mathbb{R}^n$?

Rank 1 update of the inverse:

$$(P+vv^T)^{-1} = P^{-1} + \frac{1}{1+\tilde{v}^T\tilde{v}}\tilde{v}\tilde{v}^T, \quad \tilde{v} = P^{-1}v.$$

cost: $O(n^2)$.

In general: If P is $n \times n$ invertible, and L is low rank symmetric matrix, $(P+L)^{-1}$ can be computed in $O(rn^2)$ (r - rank of L). Sherman-Morrison-Woodbury identity.

Quasi-Newton method (Approximate Hessian)

Quadratic model around current iterate x_k

$$\tilde{f}_k(x_k + d) = f(x_k) + d^T g_k + d^T H_k d$$

minimizing approximation $\Rightarrow$ $d = -H_k^{-1} g_k$

For quasi-Newton, we keep $g_k = \nabla f(x_k)$ but choose H_k so that the gradients match at current and previous iterate:

$$\nabla \tilde{f}_k(x_k) = \nabla f(x_k) \quad \text{and} \quad \nabla \tilde{f}_k(x_{k-1}) = \nabla f(x_{k-1})$$

so, we need $\nabla \tilde{f}_k(x_k - d_k) = \nabla f(x_{k-1})$, $d_k = x_k - x_{k-1}$

$$\Rightarrow \quad \nabla f(x_k) - d_k H_k d_k = \nabla f(x_{k-1})$$

$$\Rightarrow \quad H_k d_k = \nabla f(x_k) - \nabla f(x_{k-1})$$

Let $s = x_k - x_{k-1}$, $y = \nabla f(x_k) - \nabla f(x_{k-1})$

$\Rightarrow$ H_k needs to satisfy $H_k s = y$ $\rightarrow$ secant condition

- $\|H_{k+1} - H_k\|$ is small. Quadratic models are similar
- Symmetric i.e. $H_{k+1} = H_{k+1}^T$
- Secant condition: H_{k+1}

BFGS: $H_k = H_{k-1} + \frac{1}{y^T s} yy^T - \frac{1}{s^T H_{k-1} s} H_{k-1} ss^T H_{k-1}$

Inverse: $H_k^{-1} = \left(I - \frac{1}{y^T s} sy^T\right) H_{k-1}^{-1} \left(I - \frac{1}{y^T s} ys^T\right) + \frac{1}{y^T s} ss^T$

Limited memory BFGS.

- Recompute H_k^{-1} each time using past $k-L+1, \ldots, k$ iterates for s and y.
  - estimate the Hessian estimate.
  - storage $O(n^2)$ vs $O(nL)$.
- When implemented correctly, no matrices are formed.

Non-convex optimization

- Gradient descent can get stuck on local min but all steps are descent steps.
- Newton's method can get "sucked into" local max / saddle points.
  - If $\nabla^2 f(x) \not\succ 0$, then Newton step is not necessarily a descent step.
- In practice,
  - we can use BFGS (and L-BFGS).
  - switch between gradient and newton step.
  - Trust region method.