02610
Optimization and Data Fitting

Week 11: Conjugate Gradient Methods & Large-Scale Unconstrained Optimization

Yiqiu Dong

DTU Compute
Technical University of Denmark

16 November 2020

Unconstrained quadratic problems

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}$$

where A is an $n \times n$ symmetric positive definite matrix.

- It is equivalent to solve the linear system of equations: $A\mathbf{x} = \mathbf{b}$.
- The **residual** $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ is the negative gradient: $\mathbf{r} = -\nabla f(\mathbf{x})$.

- **Steepest descent method:** 1 or ∞ iterations.
- **Newton's method:** 1 iteration.
- **Coordinate search method:** n or ∞ iterations.
- **Conjugate gradient method:** n iterations.

Unconstrained quadratic problems

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}$$

where A is an $n \times n$ symmetric positive definite matrix.

- It is equivalent to solve the linear system of equations: $A\mathbf{x} = \mathbf{b}$.
- The **residual** $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ is the negative gradient: $\mathbf{r} = -\nabla f(\mathbf{x})$.

- **Steepest descent method:** 1 or ∞ iterations.
- **Newton's method:** 1 iteration.
- **Coordinate search method:** n or ∞ iterations.
- **Conjugate gradient method:** n iterations.

Conjugate gradient (CG) method

- It was proposed by Hestenes and Stiefel in the 1950s.
- It is the most widely used iterative method for solving $A\mathbf{x} = \mathbf{b}$ with $A \succ 0$
- It was extended to solve nonlinear unconstrained minimization problems in 1960s.

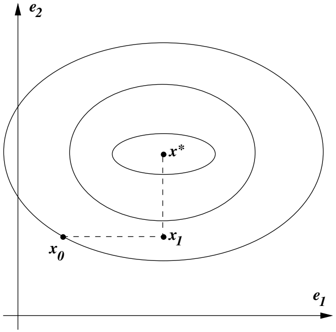
Main advantages of CG:

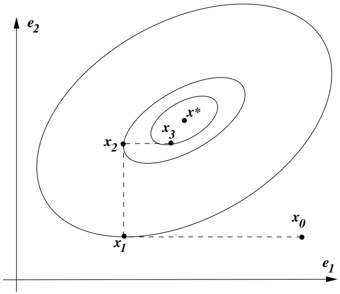
- It takes at most n iterations to the solution (theoretically).
- It does not alter A.
- At each iteration, it only need one computation of the matrix-vector product ($O(n^2)$) and a few vector product and sum ($O(n)$).
- For storage, it only need store a few vectors.
- CG is only used for solving large-scale problems.
- CG is proved with linear convergence rate, but generally much faster than the steepest descent method.

Conjugate directions

Idea:

- If A is diagonal, then the coordinate search method can find the minimizer of $f(\mathbf{x})$ in n iterations.



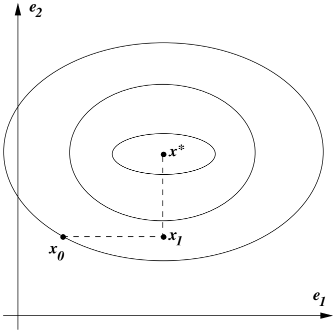


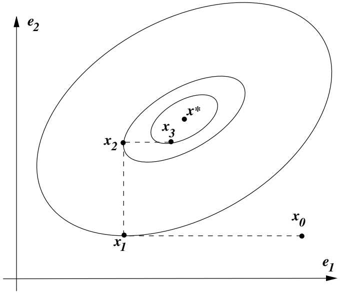
- If A is **NOT** diagonal, we can diagonalize A, that is, accordingly transform the coordinate directions.

Conjugate directions

Idea:

- If A is diagonal, then the coordinate search method can find the minimizer of $f(\mathbf{x})$ in n iterations.





- If A is **NOT** diagonal, we can diagonalize A, that is, accordingly transform the coordinate directions.

Conjugate directions

Suppose that a $n \times n$ matrix $S = [\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{n-1}]$ diagonalizes A, i.e., $S^T A S$ is diagonal. Then, we have

$$\mathbf{p}_i^T A \mathbf{p}_j = 0, \qquad \text{for all } i \neq j,$$

and we call $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ to be **conjugate** with respect to spd. A.

- $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ are conjugate, if and only if they are orthogonal for the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_A = \mathbf{u}^T A \mathbf{v}$.
- If $\mathbf{p}_i \neq \mathbf{0}$ for all i, they are also linearly independent.

If $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ are **conjugate** (*conjugate directions*) and $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$, then the exact line search has a closed-form and gives

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{p}_k}{\mathbf{p}^T A \mathbf{p}_k}.$$

Conjugate directions

Suppose that a $n \times n$ matrix $S = [\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{n-1}]$ diagonalizes A, i.e., $S^T A S$ is diagonal. Then, we have

$$\mathbf{p}_i^T A \mathbf{p}_j = 0, \qquad \text{for all } i \neq j,$$

and we call $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ to be **conjugate** with respect to spd. A.

- $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ are conjugate, if and only if they are orthogonal for the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_A = \mathbf{u}^T A \mathbf{v}$.
- If $\mathbf{p}_i \neq \mathbf{0}$ for all i, they are also linearly independent.

If $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ are **conjugate** (*conjugate directions*) and $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$, then the exact line search has a closed-form and gives

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{p}_k}{\mathbf{p}^T A \mathbf{p}_k}.$$

Conjugate directions

Suppose that a $n \times n$ matrix $S = [\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{n-1}]$ diagonalizes A, i.e., $S^T A S$ is diagonal. Then, we have

$$\mathbf{p}_i^T A \mathbf{p}_j = 0, \qquad \text{for all } i \neq j,$$

and we call $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ to be **conjugate** with respect to spd. A.

- $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ are conjugate, if and only if they are orthogonal for the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_A = \mathbf{u}^T A \mathbf{v}$.
- If $\mathbf{p}_i \neq \mathbf{0}$ for all i, they are also linearly independent.

If $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$ are **conjugate** (*conjugate directions*) and $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$, then the exact line search has a closed-form and gives

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{p}_k}{\mathbf{p}^T A \mathbf{p}_k}.$$

Conjugate direction methods

Algorithm

```
Given $\mathbf{x}_0$ and a set of conjugate directions $\{\mathbf{p}_0, \cdots, \mathbf{p}_{n-1}\}$.
loop
    Compute $\alpha_k = \frac{\mathbf{r}_k^T \mathbf{p}_k}{\mathbf{p}^T A \mathbf{p}_k}$;
    Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$;
end loop
```

Theorem

For any $\mathbf{x}_0 \in \mathbb{R}^n$ the sequence $\{\mathbf{x}_n\}$ generated by the above conjugate direction method converges to the solution $\mathbf{x}^*$ of the linear system $A\mathbf{x} = \mathbf{b}$ in at most n iterations.

Conjugate direction methods

Expanding subspace minimization

Let $\mathbf{x}_0 \in \mathbb{R}^n$ be any starting point and the sequence $\{\mathbf{x}_n\}$ be generated by the conjugate direction method shown in the previous page. Then,

- $\mathbf{r}_k^T \mathbf{p}_i = 0$, for $i = 0, 1, \cdots, k-1$;
- $\mathbf{x}_k$ is the minimizer of $f(\mathbf{x})$ over the set $\{\mathbf{x} | \mathbf{x} = \mathbf{x}_0 + \mathrm{span}\{\mathbf{p}_0, \cdots, \mathbf{p}_{k-1}\}\}$.

- The current residual $\mathbf{r}_k$ is orthogonal to all previous search directions.
- The conjugate direction method minimizes $f(\mathbf{x})$ along one conjugate direction at one iteration.

Conjugate gradient directions

- $\mathbf{p}_k$ is generated by using only the previous vector $\mathbf{p}_{k-1}$.
- $\mathbf{p}_k$ is automatically conjugate to $\{\mathbf{p}_0, \cdots, \mathbf{p}_{k-1}\}$.

Recursion for $\mathbf{p}_k$: We start with $\mathbf{p}_0 = \mathbf{r}_0$ and choose $\mathbf{p}_k$ to be a linear combination of the residual $\mathbf{r}_k$ and the previous direction $\mathbf{p}_{k-1}$:

$$\mathbf{p}_k = \mathbf{r}_k + \beta_k \mathbf{p}_{k-1}.$$

Since $\mathbf{p}_k$ is conjugate to $\mathbf{p}_{k-1}$ w.r.t. A, then we have

$$\beta_k = -\frac{\mathbf{p}_{k-1}^T A \mathbf{r}_k}{\mathbf{p}_{k-1}^T A \mathbf{p}_{k-1}}.$$

Conjugate gradient method (preliminary version)

Algorithm

```
Given $\mathbf{x}_0$;
Set $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$, $\mathbf{p}_0 = \mathbf{r}_0$;
loop
    Compute $\alpha_k = \frac{\mathbf{r}_k^T \mathbf{p}_k}{\mathbf{p}_k^T A \mathbf{p}_k}$;
    Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$;
    Compute $\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1}$;
    Compute $\beta_{k+1} = -\frac{\mathbf{p}_k^T A \mathbf{r}_{k+1}}{\mathbf{p}_k^T A \mathbf{p}_k}$;
    Compute $\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_{k+1} \mathbf{p}_k$;
    Check for convergence;
end loop
Output $\mathbf{x}_{k+1}$.
```

Properties of CG method

Theorem

Suppose that the kth iterate of the CG method is not the solution $\mathbf{x}^*$. Then,

1. $\mathbf{r}_k^T \mathbf{r}_i = 0, \qquad \text{for } i = 0, 1, \cdots, k-1,$
2. $\mathrm{span}\{\mathbf{r}_0, \mathbf{r}_1, \cdots, \mathbf{r}_k\} = \mathrm{span}\{\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_k\} = \mathrm{span}\{\mathbf{r}_0, A\mathbf{r}_0, \cdots, A^k\mathbf{r}_0\},$
3. $\mathbf{p}_k^T A\mathbf{p}_i = 0$, for $i = 0, 1, \cdots, k-1$.

Therefore, the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$ in at most n steps.

- The proof of this theorem relies on the fact that $\mathbf{p}_0 = \mathbf{r}_0$ (the steepest descent direction).
- The result (1) shows that the residuals/gradients at all iterates are orthogonal to each other.
- The result (4) shows that $\{\mathbf{p}_0, \cdots, \mathbf{p}_k\}$ are conjugate directions.
- The result (3) shows that the search directions and the residuals from CG method generate the Krylov subspaces.

Krylov subspaces

Definition: For a linear system $A\mathbf{x} = \mathbf{b}$, a sequence of subspaces

$$\mathcal{K}_0 = \{\mathbf{0}\}, \qquad \mathcal{K}_k = \mathrm{span}\{\mathbf{b}, A\mathbf{b}, \cdots, A^{k-1}\mathbf{b}\} \quad \text{for } k \geq 1.$$

Properties:

- $\mathcal{K}_k = \mathrm{span}\{\mathbf{r}_0, A\mathbf{r}_0, \cdots, A^k\mathbf{r}_0\}$.
- The Krylov subspaces are nested: $\mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \cdots$
- The dimensions of the Krylov subspaces increase by at most one: $\mathrm{dim}\mathcal{K}_{k+1} - \mathrm{dim}\mathcal{K}_k$ is zero or one.
- If $\mathcal{K}_{k+1} = \mathcal{K}_k$, then $\mathcal{K}_i = \mathcal{K}_k$ for all $i \geq k$:

$$\begin{aligned} & A^k\mathbf{b} \in \mathrm{span}\{\mathbf{b}, A\mathbf{b}, \cdots, A^{k-1}\mathbf{b}\} \\ \Longrightarrow \quad & A^i\mathbf{b} \in \mathrm{span}\{\mathbf{b}, A\mathbf{b}, \cdots, A^{k-1}\mathbf{b}\} \quad \text{for } i > k. \end{aligned}$$

Simplified CG method

- Using $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$ and $\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1}$, we obtain

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k. \tag{1}$$

- Using $\mathbf{p}_k = \mathbf{r}_k + \beta_k \mathbf{p}_{k-1}$ and $\mathbf{r}_k^T \mathbf{p}_{k-1} = 0$, we obtain $\mathbf{r}_k^T \mathbf{p}_k = \mathbf{r}_k^T \mathbf{r}_k$, then

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{p}_k}{\mathbf{p}_k^T A\mathbf{p}_k} = \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{p}_k^T A\mathbf{p}_k}. \tag{2}$$

- Using (1), (2) and $\mathbf{r}_{k+1}^T \mathbf{r}_k = 0$, we obtain

$$\beta_{k+1} = -\frac{\mathbf{p}_k^T A\mathbf{r}_{k+1}}{\mathbf{p}_k^T A\mathbf{p}_k} = \frac{\|\mathbf{r}_{k+1}\|_2^2}{\|\mathbf{r}_k\|_2^2}. \tag{3}$$

Conjugate gradient method

Algorithm

```
Given $\mathbf{x}_0$;
Set $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$, $\mathbf{p}_0 = \mathbf{r}_0$;
**loop**
    Compute $\alpha_k = \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{p}_k^T A\mathbf{p}_k}$;
    Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k$;
    Compute $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A\mathbf{p}_k$;
    Compute $\beta_{k+1} = \frac{\|\mathbf{r}_{k+1}\|_2^2}{\|\mathbf{r}_k\|_2^2}$;
    Compute $\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_{k+1}\mathbf{p}_k$;
    Check for convergence;
**end loop**
Output $\mathbf{x}_{k+1}$.
```

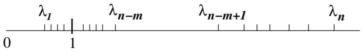
Main computation per iteration is matrix-vector product $A\mathbf{p}_{k+1}$.

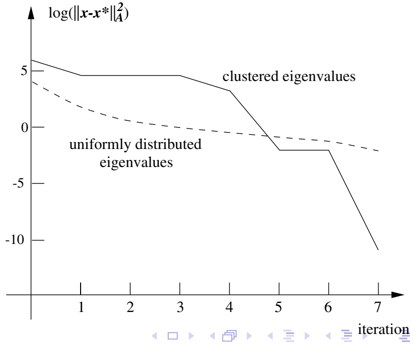
Rate of convergence

- If A has only r distinct eigenvalues, then the CG method will terminate at the solution in at most r iterations.
- If A has eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, we have that

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_A^2 \leq \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 \|\mathbf{x}_0 - \mathbf{x}^*\|_A^2.$$

Example: We apply the CG method to solve $A\mathbf{x} = \mathbf{b}$.





Preconditioning

- **Idea:** Make change of variables $\hat{\mathbf{x}} = C\mathbf{x}$ with C nonsingular, and apply CG to
$$C^{-T}AC^{-1}\hat{\mathbf{x}} = C^{-T}\mathbf{b}.$$
- The spectrum of the new matrix $C^{-T}AC^{-1}$ should be clustered, then PCG converges fast.
- We need consider the trade-off between enhanced convergence and cost of extra computation.
- The matrix $M = C^{T}C$ is called the preconditioner.
- Matlab implementation: `pcg`

Example:

- diagonal $C = \mathrm{diag}(A_{11}, A_{22}, \cdots, A_{nn})$
- incomplete or approximate Cholesky factorization of A
- Good preconditioners are often application-dependent.

Nonlinear conjugate gradient method

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}), \qquad f \text{ is convex and differentiable.}$$

Nonlinear CG methods

- Extend linear CG method to nonquadratic functions.
- Limited global convergence theory.

Modifications needed to extend linear CG method

- Replace $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ with $-\nabla f(\mathbf{x}_k)$.
- Determine the step length α by line search.

Fletcher-Reeves method

Algorithm

```
Given $\mathbf{x}_0$;
Compute $f_0 = f(\mathbf{x}_0)$ and $\nabla f_0 = \nabla f(\mathbf{x}_0)$;
Set $\mathbf{p}_0 = -\nabla f_0$;
loop
    Compute $\alpha_k$ by line search method;
    Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$;
    Evaluate $\nabla f_{k+1}$;
    Compute $\beta_{k+1}^{FR} = \frac{\|\nabla f_{k+1}\|_2^2}{\|\nabla f_k\|_2^2}$;
    Compute $\mathbf{p}_{k+1} = -\nabla f_{k+1} + \beta_{k+1}^{FR} \mathbf{p}_k$;
    Check for convergence;
end loop
Output $\mathbf{x}_{k+1}$.
```

Some observations

Interpretation

- First iteration is a steepest descent step.
- General update is a steepest descent step with momentum term

$$\mathbf{x}_{x+1} = \mathbf{x}_k - \alpha_k \nabla f_k + \frac{\alpha_k \beta_k}{\alpha_{k-1}}(\mathbf{x}_k - \mathbf{x}_{k-1}).$$

- It is common to restart the algorithm every n iterations by taking a steepest descent step to periodically refresh the algorithm.

Line search

- With exact line search, it reduces to linear CG for quadratic f.
- Exact line search in computation of α_k implies that α_k is a local minimizer along $\mathbf{p}_k$, i.e., $\nabla f_{k+1}^T \mathbf{p}_k = 0$. Therefore, $\mathbf{p}_{k+1}$ is a descent direction at $\mathbf{x}_{k+1}$:

$$\nabla f_{k+1}^T \mathbf{p}_{k+1} = -\|\nabla f_{k+1}\|^2 + \beta_{k+1}^{FR} \nabla f_{k+1}^T \mathbf{p}_k = -\|\nabla f_{k+1}\|^2 < 0.$$

- For inexact line search, if α_k satisfies the strong Wolfe conditions, then $\mathbf{p}_{k+1}$ is descent.

Variations

Polak-Ribière method: Compute β_{k+1} from

$$\beta_{k+1} = \frac{\nabla f_{k+1}^T (\nabla f_{k+1} - \nabla f_k)}{\|\nabla f_k\|_2^2}.$$

Hestenes-Stiefel method: Compute β_{k+1} from

$$\beta_{k+1} = \frac{\nabla f_{k+1}^T (\nabla f_{k+1} - \nabla f_k)}{(\nabla f_{k+1} - \nabla f_k)^T \mathbf{p}_k}.$$

- All these formulas are equivalent for quadratic f and exact line search.
- With restarts and the strong Wolfe conditions, all three methods have global convergence.
- Without restarts, FR has global convergence with the strong Wolfe conditions, but PR not.
- In practice, PR is more robust and efficient than FR.

Large-scale unconstrained optimization

- Large-scale problems (today): $10^3 \sim 10^6$ variables.
- When solving large-scale problems, we have to take the storage and computational costs of the optimization algorithm into account.
- In large problems, the following can have a prohibitive cost:
  - factorizing the Hessian (solving for the Newton step)
  - computing the Hessian or multiplying it
  - storing a dense approximate Hessian like in quasi-Newton methods
- Linear/nonlinear conjugate gradient methods can be applied directly to large-scale problems without modification, but not fast.

Inexact Newton methods

Ideas: Use some inexpensive iterative algorithm to *very approximately* solve either

$$\nabla^2 f_k \mathbf{p}_k = -\nabla f_k \qquad \text{(line search)}$$

or

$$\min_{\mathbf{p}\in\mathbb{R}^n} \ m_k(\mathbf{p}) = f_k + \nabla f_k^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \nabla^2 f_k \mathbf{p},$$
$$\text{s. t. } \|\mathbf{p}\|_2 \le \Delta_k, \qquad \text{(trust region)}$$

without ruining global and fast local convergence of exact LS/TR Newton methods.

Stopping criterion for iterative solver:

$$\|\mathbf{r}_k\|_2^2 = \|\nabla^2 f_k \mathbf{p}_k + \nabla f_k\|_2^2 \le \eta_k \|\nabla f_k\|_2^2,$$

where the sequence $\{\eta_k\}$ with $0 \le \eta_k \le 1$ for all k is called the **forcing sequence**.

Local convergence

Convergence theorem

Inexact Newton with unit steps:

- $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$
- $\|\mathbf{r}_k\|_2^2 \leq \eta_k \|\nabla f_k\|_2^2$
- $0 < \eta_k \leq \eta < 1$

Then, if the starting point $\mathbf{x}_0$ is sufficiently near $\mathbf{x}^*$,

- the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$,
- and
$$\|\nabla^2 f(\mathbf{x}*)(\mathbf{x}_{k+1} - \mathbf{x}^*)\|_2 \leq \hat{\eta} \|\nabla^2 f(\mathbf{x}^*)(\mathbf{x}_k - \mathbf{x}^*)\|_2$$
for some constant $\hat{\eta}$ with $\eta < \hat{\eta} < 1$ (linear convergence).

Local convergence

Convergence rate

Inexact Newton with unit steps:

- $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$
- $\|\mathbf{r}_k\|_2^2 \leq \eta_k \|\nabla f_k\|_2^2$
- $0 < \eta_k \leq \eta < 1$

Then,

- if $\eta_k \to 0$, the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$ superlinearly;
- if $\nabla^2 f(\mathbf{x})$ is Lipschitz continuous for $\mathbf{x}$ near $\mathbf{x}^*$ and $\eta_k = O(\|\nabla f_k\|_2)$, then the convergence is quadratic.

Example:

- $\eta_k = \min(0.5, \sqrt{\|\nabla f_k\|_2})$ would yield superlinear convergence;
- $\eta_k = \min(0.5, \|\nabla f_k\|_2)$ would yield quadratic convergence.

Line search Newton-CG method

Algorithm

```
Given $\mathbf{x}_0$;
loop
    Define the forcing sequence $\eta_k = \min(0.5, \sqrt{\|\nabla f_k\|_2})$
    Use CG to solve $\nabla^2 f_k \mathbf{p}_k = -\nabla f_k$ approximately with accuracy $\eta_k$
    Compute $\alpha_k$ by line search method;
    Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$;
end loop
```

Remark: $\nabla^2 f_k$ is not necessarily positive definite, but the CG method is designed to solve positive definite systems. So we need to modify CG loop.

Line search Newton-CG method

```
Given $\mathbf{x}_0$;
for $k = 0, 1, \cdots$ do
    Set $\mathbf{z}_0 = 0$, $\mathbf{r}_0 = -\nabla f_k$, $\mathbf{d}_0 = \mathbf{r}_0$ and $\eta_k = \min(0.5, \sqrt{\|\nabla f_k\|_2})$;
    for $j = 0, 1, \cdots$ do
        if $\mathbf{d}_j^T \nabla^2 f_k \mathbf{d}_j \leq 0$ then
            if $j = 0$ then
                Stop CG and return steepest descent direction: $\mathbf{p}_k = \mathbf{d}_0$;
            else
                Stop CG and return $\mathbf{p}_k = \mathbf{z}_j$;
            end if
        end if
        ⋮ rest of CG loop
    end for
    Compute $\alpha_k$ by line search method;
    Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$;
end for
```

Line search Newton-CG method

- Inner CG loop always produces a direction of descent for f.
- When the Hessian $\nabla^2 f_k$ is nearly singular, the line search Newton-CG direction can be long and of poor quality.
- It does not require explicit knowledge of the Hessian, and it requires only the Hessian-vector products. Finite differencing and automatic differentiation techniques can be used.
- Preconditioning can be introduced to speed up CG.

Limited-memory quasi-Newton methods

Idea: They save only a few vectors of length n that represent the approximation of the Hessian implicitly.

- Useful for solving large problems with costly or nonsparse Hessian.
- Linear convergence but fast rate.

Limited-memory BFGS (L-BFGS):

- It uses curvature information from only the most recent m iterations to construct the Hessian approximation.
- Modest values of m ($\sim 3 - 20$) work fine in practice, but the best m depends on the problem.
- Slow convergence in ill-conditioned problems.

Limited-memory quasi-Newton methods

Idea: They save only a few vectors of length n that represent the approximation of the Hessian implicitly.

- Useful for solving large problems with costly or nonsparse Hessian.
- Linear convergence but fast rate.

Limited-memory BFGS (L-BFGS):

- It uses curvature information from only the most recent m iterations to construct the Hessian approximation.
- Modest values of m ($\sim 3-20$) work fine in practice, but the best m depends on the problem.
- Slow convergence in ill-conditioned problems.

L-BFGS update

Review: BFGS inverse Hessian update:

$$H_{k+1} = V_k^T H_k V_k + \rho_k \mathbf{s}_k \mathbf{s}_k^T$$

where $V_k = I - \rho_k \mathbf{s}_k \mathbf{y}_k^T$, $\rho_k = 1/(\mathbf{y}_k^T \mathbf{s}_k)$, $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \nabla f_{k+1} - \nabla f_k$.

- Since H_k is generally dense, the cost of storing and manipulating it is prohibitive when n is large.
- We store a modified version of H_k implicitly, by storing $m \ll n$ of the vector pairs $\{\mathbf{s}_k, \mathbf{y}_k\}$.
- The product $H_k \nabla f_k$ c an be obtained by performing a sequence of inner products and vector summations.
- After the new iterate is computed, we replace the oldest pair with the new pair.

L-BFGS update

Update algorithm: Compute $H_k \nabla f_k$

```
Given $H_k^0$; Set $\mathbf{q} = \nabla f_k$;
for $i = k-1, k-2, \cdots, k-m$ do
    $\alpha_i = \rho_i \mathbf{s}_i^T \mathbf{q}$;
    $\mathbf{q} = \mathbf{q} - \alpha_i \mathbf{y}_i$;
end for
$\mathbf{r} = H_k^0 \mathbf{q}$;
for $i = k-m, k-m+1, \cdots, k-1$ do
    $\beta = \rho_i \mathbf{y}_i^T \mathbf{r}$;
    $\mathbf{r} = \mathbf{r} + \mathbf{s}_i(\alpha_i - \beta)$;
end for
Output $\mathbf{r}$.
```

- It recursively expands the update with m pairs $\{\mathbf{s}_k, \mathbf{y}_k\}$.
- H_k^0 is allowed to vary from iteration to iteration.
- It requires $4mn$ multiplications and calculation of $H_k^0\mathbf{q}$.

L-BFGS method

```
Given $\mathbf{x}_0$ and $m$;
**loop**
    Choose $H_k^0$;
    Compute $\mathbf{p}_k = -H_k \nabla f_k$ by update algorithm;
    Compute $\alpha_k$ by line search method;
    Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$;
    **if** $k > m$ **then**
        Discard $\{\mathbf{s}_{k-m}, \mathbf{y}_{k-m}\}$ from storage;
    **end if**
    Store $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \nabla f_{k+1} - \nabla f_k$;
**end loop**
```

- A good choice for H_k^0 in practice: $H_k^0 = \gamma_k I$ with $\gamma_k = (\mathbf{s}_{k-1}^T \mathbf{y}_{k-1})/(\mathbf{y}_{k-1}^T \mathbf{y}_{k-1})$.
- The line search based on the (strong) Wolfe conditions makes BFGS stable.
- The first $m-1$ iterates are the same as in BFGS.

Relationship with CG methods

- Limited-memory methods historically evolved as improvements of nonlinear CG methods.
- The Hestenes-Stiefel form of nonlinear CG method:

$$\mathbf{p}_{k+1} = -\nabla f_{k+1} + \frac{\nabla f_{k+1}^T \mathbf{y}_k}{\mathbf{y}_k^T \mathbf{p}_k} \mathbf{p}_k = -\hat{H}_{k+1} \nabla f_{k+1} \quad \text{with } \hat{H}_{k+1} = I - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k},$$

  which resembles quasi-Newton iterates, but $\hat{H}_{k+1}$ is neither symmetric nor positive definite.
- A symmetric positive definite modification, which also satisfies the secant equation, is

$$H_{k+1} = \left(I - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k} \right) \left(I - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k} \right) + \frac{\mathbf{s}_k^T \mathbf{s}_k}{\mathbf{y}_k^T \mathbf{s}_k},$$

  which is exactly the L-BGFS method with $m = 1$ and $H_k^0 = I$ (*memoryless BFGS*).