IIIT Hyderabad **Scribed By: Sanchit Jalan (2022101070)**
- **Hardik Kalia (2022101092)**

Optimization Methods(CS1.404) **Lecture #11**
Instructor: Dr. Naresh Manwani 3^{rd} **Mar 2025**

Outline. *The scribe covers key optimization methods, including the Armijo-Wolfe conditions for step size selection, the exact line search method, and the steepest gradient descent algorithm. It also discusses the Kantorovich inequality, condition numbers, and their impact on convergence rates, with examples such as the Rosenbrock function.*

1 Armijo-Wolfe Conditions

The Armijo Wolfe condition is used to rule out unacceptably short steps, called the curvature condition, and ensure sufficent decrease. The conditions are

Armijo-Wolfe Condition

Let $\mathbf{x}_k$ be the estimate at iteration k and $\mathbf{d}_k$ be a descent direction at $\mathbf{x}_k$. Then the step size α_k must satisfy

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \leq f(\mathbf{x}_k) + c_1 \alpha_k \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k$$

$$\nabla f(\mathbf{x}_k + \alpha_k \mathbf{d}_k)^\top \mathbf{d}_k \geq c_2 \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k$$

for some constants $0 < c_2 < c_1 < 1$

For a given $\mathbf{x}_k$ and $\mathbf{d}_k$, taking $\phi(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{d}_k)$, we get $\phi'(\alpha) = \nabla f(\mathbf{x}_k + \alpha \mathbf{d}_k)^\top \mathbf{d}_k$. The curvature condition states that

$$\phi'(\alpha_k) \geq c_2 \phi'(0)$$

This restriction ensures that if $\phi'(\alpha)$ is strongly negative, then α is not chosen as a step size as f can be further reduced along the chosen direction. On the other hand, if $\phi'(\alpha)$ is slightly negative or positive, then we cannot expect more decrease in f in this direction, so the line search can be terminated. Thus, the Wolfe condition ensures sufficient rate of decrease

Issue: A step length may satisfy the Armijo-Wolfe condition without being particularly close to a minimizer of ϕ

2 Exact Line Search Method

The Exact Line Search Method chooses step size α_k as the exact minimizer of f along the ray $\mathbf{x}_k + \alpha \mathbf{d}_k$

$$\alpha = \arg\min_{\alpha \geq 0} f(\mathbf{x}_k + \alpha \mathbf{d}_k)$$

2.1 Exact Line Search for Quadratic Function

Result

- Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$, where A is an $n \times n$ symmetric positive definite matrix, $b \in \mathbb{R}^n, c \in \mathbb{R}$
- Let $\mathbf{x} \in \mathbb{R}^n$, and d be a descent direction of f at $\mathbf{x}$

Then

$$\arg\min_{t \geq 0} f(\mathbf{x} + t\mathbf{d}) = -\frac{\nabla f(\mathbf{x})^\top \mathbf{d}}{\mathbf{d}^T A \mathbf{d}}$$

Proof.

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c \\ \implies \nabla f &= Ax + b \\ \implies \nabla^2 f &= A \end{aligned}$$

Let $\phi(\alpha) = f(\mathbf{x} + \alpha\mathbf{d})$

$$\begin{aligned} \phi(\alpha) &= f(\mathbf{x} + \alpha\mathbf{d}) \\ \implies \phi'(\alpha) &= \nabla f(\mathbf{x} + \alpha\mathbf{d})^\top \mathbf{d} \\ \implies \phi''(\alpha) &= \mathbf{d}^\top \nabla^2 f(\mathbf{x} + \alpha\mathbf{d})\mathbf{d} \\ &= \mathbf{d}^\top A\mathbf{d} \end{aligned}$$

Since A is given to be a positive semidefinite matrix, this implies that $\mathbf{d}^\top A\mathbf{d} \geq 0$ for all $\mathbf{d}$ i.e. $\phi''(\alpha) \geq 0 \, \forall \alpha$, and using the second order characterization of convex functions, we conclude that $\phi(\alpha)$ is a convex function
Since ϕ is a convex function, its minimizer can be found by solving $\phi' = 0$

$$\begin{aligned} \phi'(\alpha) &= \mathbf{d}_k^\top \nabla f(\mathbf{x}_k + \alpha\mathbf{d}_k) &&= 0 \\ \implies \mathbf{d}_k^\top (A(\mathbf{x}_k + \alpha\mathbf{d}_k) + b) &= 0 \\ \implies \mathbf{d}_k^\top (A\mathbf{x}_k + b) + \alpha\mathbf{d}_k^\top A\mathbf{d_k} &= 0 \\ \implies \mathbf{d}_k^\top \nabla f(\mathbf{x}_k) + \alpha\mathbf{d}_k^\top A\mathbf{d_k} &= 0 \\ \implies \alpha &= -\frac{\nabla f(\mathbf{x})^\top \mathbf{d}}{\mathbf{d}^T A\mathbf{d}} \end{aligned}$$

□

3 Steepest Gradient Descent

3.1 Optimality of Steepest Gradient Descent Direction

Lemma

Let f be a continuously differentiable function, and let $\mathbf{x} \in \mathbb{R}^n$ be a non-stationary point, i.e.,

$$\nabla f(\mathbf{x}) \neq 0.$$

Then, the optimal solution to the following problem:

$$\min_{\mathbf{d} \in \mathbb{R}^n} \nabla f(\mathbf{x})^\top \mathbf{d}$$

subject to the constraint

$$\|\mathbf{d}\| = 1$$

is given by

$$\mathbf{d} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}.$$

Proof. Using the **Cauchy-Schwarz inequality**, we have:

$$|\nabla f(\mathbf{x})^\top \mathbf{d}| \leq \|\nabla f(\mathbf{x})\| \|\mathbf{d}\|.$$

Since $\|\mathbf{d}\| = 1$, it follows that:

$$|\nabla f(\mathbf{x})^\top \mathbf{d}| \leq \|\nabla f(\mathbf{x})\|.$$

Removing the modulus, we obtain:

$$-\|\nabla f(\mathbf{x})\| \leq \nabla f(\mathbf{x})^\top \mathbf{d} \leq \|\nabla f(\mathbf{x})\|.$$

To minimize $\nabla f(\mathbf{x})^\top \mathbf{d}$, we need to attain the lower bound $-\|\nabla f(\mathbf{x})\|$.
Consider the choice of $\mathbf{d}$ that achieves this lower bound. Call it $\mathbf{d}_{\min}$. Then,

$$\nabla f(\mathbf{x})^\top \mathbf{d}_{\min} = -\|\nabla f(\mathbf{x})\|.$$

This happens when $\mathbf{d}_{\min}$ is aligned in the opposite direction of $\nabla f(\mathbf{x})$, i.e.,

$$\mathbf{d}_{\min} = -\lambda \nabla f(\mathbf{x}).$$

Since $\|\mathbf{d}\| = 1$, we get:

$$\|\mathbf{d}_{\min}\| = \| - \lambda \nabla f(\mathbf{x})\| = \lambda \|\nabla f(\mathbf{x})\| = 1.$$

$$\lambda = \frac{1}{\|\nabla f(\mathbf{x})\|}.$$

Thus,

$$\mathbf{d}_{\min} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}.$$

□

3.2 Steepest Gradient Descent Algorithm

Algorithm 1 Steepest Descent Algorithm

1: **Input:** Tolerance parameter $\epsilon > 0$
2: **Initialization:** Pick $x_0 \in \mathbb{R}^n$ arbitrarily, set $k = 0$
3: **repeat**
4: Compute descent direction $d_k = -\nabla f(x_k)$
5: Find stepsize t_k by a line search on

$$g(t) = f(x_k + td_k)$$

6: Update $x_{k+1} = x_k + t_k d_k$
7: $k \leftarrow k + 1$
8: **until** $\|\nabla f(x_{k+1})\| \leq \epsilon$
9: **Return:** x_{k+1} as the optimal solution

3.3 Some examples of Gradient Descent with different Line Search Algorithms and Step sizes

3.3.1 Gradient Descent with Exact Line Search on Quadratic Function

Consider the function $f(x, y) = x^2 + 2y^2$, whose optimal solution is $(0, 0)$ with an optimal value of 0.
Let $(x_0, y_0) = (2, 1)$, and $\epsilon = 10^{-5}$.
The gradient descent approach stops in 13 iterations and finds a solution that is close to the optimal value.

$$(x^*, y^*) = (0.1254 \times 10^{-5}, -0.0627 \times 10^{-5})$$

```
iter_number =  1  norm_grad = 1.885618  fun_val = 0.666667
iter_number =  2  norm_grad = 0.628539  fun_val = 0.074074
iter_number =  3  norm_grad = 0.209513  fun_val = 0.008230
iter_number =  4  norm_grad = 0.069838  fun_val = 0.000914
iter_number =  5  norm_grad = 0.023279  fun_val = 0.000102
iter_number =  6  norm_grad = 0.007760  fun_val = 0.000011
iter_number =  7  norm_grad = 0.002587  fun_val = 0.000001
iter_number =  8  norm_grad = 0.000862  fun_val = 0.000000
iter_number =  9  norm_grad = 0.000290  fun_val = 0.000000
iter_number = 10  norm_grad = 0.000097  fun_val = 0.000000
iter_number = 11  norm_grad = 0.000032  fun_val = 0.000000
iter_number = 12  norm_grad = 0.000011  fun_val = 0.000000
iter_number = 13  norm_grad = 0.000004  fun_val = 0.000000
```

Optimal solution: $[1.25445095 \times 10^{-6}, -6.27225474 \times 10^{-7}]$
Optimal function value: $2.3604707743147666 \times 10^{-12}$

3.3.2 Gradient Descent with Constant Step Size on Quadratic Function

Consider the function $f(x, y) = x^2 + 2y^2$, whose optimal solution is $(0, 0)$ with an optimal value of 0.
Let $(x_0, y_0) = (2, 1)$, $\epsilon = 10^{-5}$, $t_k = 0.1$..
The gradient descent approach stops in 58 iterations. The step size was too small, which causes slow convergence.However, taking a stepsize which is large might lead to divergence of the iterates. For example, taking the constant stepsize to be 100 results in a divergent sequence.

```
iter_number =  1  norm_grad = 4.000000  fun_val = 3.280000
iter_number =  2  norm_grad = 2.937210  fun_val = 1.897600
iter_number =  3  norm_grad = 2.222791  fun_val = 1.141888
iter_number =  4  norm_grad = 1.718457  fun_val = 0.704681
iter_number =  5  norm_grad = 1.347120  fun_val = 0.441590
     .
     .
     .
iter_number = 54  norm_grad = 0.000023  fun_val = 0.000000
iter_number = 55  norm_grad = 0.000019  fun_val = 0.000000
iter_number = 56  norm_grad = 0.000015  fun_val = 0.000000
iter_number = 57  norm_grad = 0.000012  fun_val = 0.000000
iter_number = 58  norm_grad = 0.000010  fun_val = 0.000000
```

Optimal solution: $[4.78904857 \times 10^{-6}, 1.35760217 \times 10^{-13}]$
Optimal function value: $2.2934986159900767 \times 10^{-11}$

3.3.3 Example 1: Gradient Descent with Backtracking Line Search on Quadratic Function

Consider the function $f(x, y) = x^2 + 2y^2$, whose optimal solution is $(0, 0)$ with an optimal value of 0.
Let $(x_0, y_0) = (2, 1)$, $\epsilon = 10^{-5}$, $\tau = 0.5$, $s = 2$, and $c_1 = 0.25$.
The gradient descent approach stops in 2 iterations and outputs the exact optimal solution.
For this example, inexact line search performs better than exact line search.

```
iter_number =  1 norm_grad = 2.000000 fun_val = 1.000000
iter_number =  2 norm_grad = 0.000000 fun_val = 0.000000
```

Optimal solution: $[0.0, 0.0]$
Optimal function value: 0.0

3.3.4 Example 2: Gradient Descent with Backtracking Line Search on Quadratic Function

Consider the function $f(x, y) = x^2 + \frac{1}{100}y^2$, whose optimal solution is $(0, 0)$ with an optimal value of 0.
Let $(x_0, y_0) = \left(\frac{1}{100}, 1\right)$, $\epsilon = 10^{-5}$, $\tau = 0.5$, $s = 2$, and $c_1 = 0.25$.
The gradient descent approach stops in 201 iterations.
Iteration details:

```
iter_number =   1, norm_grad = 0.028003, fun_val = 0.009704
iter_number =   2, norm_grad = 0.027730, fun_val = 0.009324
iter_number =   3, norm_grad = 0.027465, fun_val = 0.008958
iter_number =   4, norm_grad = 0.027209, fun_val = 0.008608
iter_number =   5, norm_grad = 0.026960, fun_val = 0.008271
     .
     .
     .
iter_number = 197, norm_grad = 0.000012, fun_val = 0.000000
iter_number = 198, norm_grad = 0.000011, fun_val = 0.000000
iter_number = 199, norm_grad = 0.000011, fun_val = 0.000000
iter_number = 200, norm_grad = 0.000010, fun_val = 0.000000
iter_number = 201, norm_grad = 0.000010, fun_val = 0.000000
```

Optimal solution: $[0.0, 0.00049166]$
Optimal function value: $2.417269699979995 \times 10^{-9}$

3.3.5 Conclusion

- For different quadratic functions, we observe that the convergence time varies for gradient descent.
- One such measure which can partially answer the above question is condition number.

3.4 Convergence of Steepest Gradient Descent with Exact Line Search for Quadratic Function

- Let $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$, where A is a symmetric positive definite matrix.
- For Steepest descent, the search direction is given by:

$$\mathbf{d}_k = -\nabla f(\mathbf{x}_k) = -2A\mathbf{x}_k.$$

- Exact line search results in:

$$t_k = \arg\min_{t \geq 0} f(\mathbf{x}_k + t\mathbf{d}_k) = \frac{\mathbf{d}_k^\top \mathbf{d}_k}{2\mathbf{d}_k^\top A\mathbf{d}_k}.$$

  Now, expanding $f(\mathbf{x}_k + t_k\mathbf{d}_k)$:

$$f(\mathbf{x}_k + t_k\mathbf{d}_k) = (\mathbf{x}_k + t_k\mathbf{d}_k)^\top A(\mathbf{x}_k + t_k\mathbf{d}_k).$$
$$= \mathbf{x}_k^\top A\mathbf{x}_k + t_k^2\mathbf{d}_k^\top A\mathbf{d}_k + 2t_k\mathbf{d}_k^\top A\mathbf{x}_k.$$

Substituting $t_k = \frac{\mathbf{d}_k^\top \mathbf{d}_k}{2\mathbf{d}_k^\top A\mathbf{d}_k}$:

$$= \mathbf{x}_k^\top A\mathbf{x}_k + \left(\frac{\mathbf{d}_k^\top \mathbf{d}_k}{2\mathbf{d}_k^\top A\mathbf{d}_k}\right)^2 \mathbf{d}_k^\top A\mathbf{d}_k + 2\left(\frac{\mathbf{d}_k^\top \mathbf{d}_k}{2\mathbf{d}_k^\top A\mathbf{d}_k}\right)\mathbf{d}_k^\top A\mathbf{x}_k.$$

Since $\mathbf{d}_k = -2A\mathbf{x}_k$, we have:

$$\mathbf{d}_k^\top A\mathbf{x}_k = (-2A\mathbf{x}_k)^\top A\mathbf{x}_k = -2\mathbf{x}_k^\top AA\mathbf{x}_k.$$

$$= \mathbf{x}_k^\top A\mathbf{x}_k + \frac{(\mathbf{d}_k^\top \mathbf{d}_k)^2}{4\mathbf{d}_k^\top A\mathbf{d}_k} - \frac{\mathbf{d}_k^\top \mathbf{d}_k}{\mathbf{d}_k^\top A\mathbf{d}_k}\mathbf{x}_k^\top A\mathbf{x}_k.$$

Factoring $\mathbf{x}_k^\top A\mathbf{x}_k$:

$$= \mathbf{x}_k^\top A\mathbf{x}_k \left(1 - \frac{1}{4}\frac{(\mathbf{d}_k^\top \mathbf{d}_k)^2}{\mathbf{d}_k^\top A\mathbf{d}_k \cdot \mathbf{x}_k^\top A\mathbf{x}_k}\right).$$

Using the property $A^{-1}A = I$, we can rewrite:

$$= \mathbf{x}_k^\top A\mathbf{x}_k \left(1 - \frac{1}{4}\frac{(\mathbf{d}_k^\top A\mathbf{d}_k)^2}{\mathbf{d}_k^\top AA^{-1}A\mathbf{x}_k}\right).$$

$$= \left(1 - \frac{(\mathbf{d}_k^\top A\mathbf{d}_k)^2}{(\mathbf{d}_k^\top A\mathbf{d}_k)(\mathbf{d}_k^\top A^{-1}\mathbf{d}_k)}\right) f(\mathbf{x}_k).$$

This establishes the reduction ratio in the objective function value after one iteration of steepest descent with exact line search.

4 Kantorovich Inequality

Kantarovich Inequality

Let A be a positive definite $n \times n$ matrix. Then for any $\mathbf{x} \in \mathbb{R}^n$ ($\mathbf{x} \neq \mathbf{0}$), the inequality

$$\frac{(\mathbf{x}^\top \mathbf{x})^2}{(\mathbf{x}^\top A\mathbf{x})(\mathbf{x}^\top A^{-1}\mathbf{x})} \geq \frac{4\lambda_{\max}(A)\lambda_{\min}(A)}{(\lambda_{\max}(A) + \lambda_{\min}(A))^2}$$

holds.

4.1 Eigenvalues of $X = A + mMA^{-1}$

Theorem

Let A be a positive definite $n \times n$ matrix. The eigenvalues of

$$X = A + mMA^{-1}$$

are given by

$$\lambda_i + \frac{mM}{\lambda_i}$$

where:
- λ_i are the eigenvalues of A,
- $m = \lambda_{\min}(A)$ (smallest eigenvalue of A),
- $M = \lambda_{\max}(A)$ (largest eigenvalue of A).

Proof. Since A is positive definite, it has a full set of eigenvectors forming a basis.
If v is an eigenvector of A corresponding to eigenvalue λ_i, we have:

$$Av = \lambda_i v.$$

Since A is invertible, we can apply its inverse:

$$A^{-1}v = \frac{1}{\lambda_i}v.$$

$$X = A + mMA^{-1}.$$

Multiplying this with v:

$$Xv = \left(A + mMA^{-1}\right)v.$$

Using $Av = \lambda_i v$ and $A^{-1}v = \frac{1}{\lambda_i}v$, we substitute:

$$Xv = \lambda_i v + mM\frac{1}{\lambda_i}v.$$

$$Xv = \left(\lambda_i + \frac{mM}{\lambda_i}\right)v.$$

This shows that the eigenvalues of X are:

$$\mu_i = \lambda_i + \frac{mM}{\lambda_i}.$$

□

4.2 Maximum Eigenvalue of $X = A + mMA^{-1}$

Maximum Eigenvalue of $X = A + mMA^{-1}$

We are given a positive definite $n \times n$ matrix A. We define a new matrix X as:

$$X = A + mMA^{-1}$$

where:

- $m = \lambda_{\min}(A)$ (smallest eigenvalue of A).
- $M = \lambda_{\max}(A)$ (largest eigenvalue of A).

Prove that the maximum eigenvalue of X is $m + M$.

Proof:
From Section 4.1, the eigenvalues of X are

$$\mu_i = \lambda_i + \frac{mM}{\lambda_i}, \quad i = 1, 2, \ldots, n.$$

To determine the maximum eigenvalue, consider the function

$$f(\lambda) = \lambda + \frac{mM}{\lambda}, \quad \lambda \in (m, M).$$

Differentiating, we obtain

$$f'(\lambda) = 1 - \frac{mM}{\lambda^2}.$$

Setting $f'(\lambda) = 0$ gives

$$1 = \frac{mM}{\lambda^2} \quad \Rightarrow \quad \lambda^2 = mM \quad \Rightarrow \quad \lambda = \sqrt{mM}.$$

Evaluating $f(\lambda)$ at the endpoints,

$$f(m) = m + \frac{mM}{m} = m + M, \quad f(M) = M + \frac{mM}{M} = M + m.$$

Since $f(m) = f(M) = m + M$ and by convexity of $f(\lambda)$, we conclude that the maximum eigenvalue of X is

$$\lambda_{\max}(X) = m + M.$$

Hence,Proved.

4.3 Matrix Inequality Proof

Theorem

Let A be a positive definite matrix with eigenvalues satisfying

$$m \leq \lambda_i(A) \leq M.$$

Define the matrix

$$B = A + mMA^{-1}.$$

Then, for any vector x, we have

$$x^T Bx \leq (m + M)\|x\|^2.$$

Proof. Since A is symmetric and positive definite, it has an orthonormal eigenbasis. Let λ_i be its eigenvalues, and consider A^{-1}, which has eigenvalues $\frac{1}{\lambda_i}$.
Expanding the quadratic form:

$$x^T Ax = \sum_i \lambda_i c_i^2$$

for some coefficients c_i, where x is expressed in terms of the eigenvectors of A. Similarly,

$$x^T A^{-1} x = \sum_i \frac{c_i^2}{\lambda_i}.$$

$$x^T (A + mMA^{-1})x = x^T Ax + mMx^T A^{-1} x.$$

Substituting the eigenvalue expressions:

$$= \sum_i \lambda_i c_i^2 + mM \sum_i \frac{c_i^2}{\lambda_i}.$$

Using the eigenvalue bounds $m \leq \lambda_i \leq M$, we apply the inequality from Section 4.2:

$$\lambda_i + \frac{mM}{\lambda_i} \leq m + M.$$

Summing over all i, we get:

$$\sum_i \left(\lambda_i + \frac{mM}{\lambda_i} \right) c_i^2 \leq (m + M) \sum_i c_i^2.$$

Since $\sum_i c_i^2 = \|x\|^2$

$$x^T (A + mMA^{-1})x \leq (m + M)\|x\|^2.$$

□

4.4 Proof of Kantorovich Inequality

Proof. Define the quadratic forms:

$$a = \mathbf{x}^T A\mathbf{x}, \quad b = \mathbf{x}^T A^{-1}\mathbf{x}.$$

Applying the AM-GM inequality to a and mMb, we have

$$a + mMb \geq 2\sqrt{a \cdot mMb} \tag{1}$$

From Section 4.3, we have:

$$\begin{aligned} \mathbf{x}^T(A + mMA^{-1})\mathbf{x} &\leq (m+M)\mathbf{x}^T\mathbf{x}. \\ \implies a + mMb &\leq (m+M)\mathbf{x}^T\mathbf{x}. \end{aligned} \tag{2}$$

From (1) and (2), we have

$$(m+M)\mathbf{x}^\top\mathbf{x} \geq 2\sqrt{abmM}$$

Dividing by $2\sqrt{ab}$ on both sides:

$$\frac{(m+M)\mathbf{x}^T\mathbf{x}}{2\sqrt{ab}} \geq \sqrt{mM}.$$

Squaring both sides:

$$\frac{(m+M)^2(\mathbf{x}^T\mathbf{x})^2}{4ab} \geq mM.$$

Thus, we obtain:

$$\frac{(\mathbf{x}^T\mathbf{x})^2}{ab} \geq \frac{4mM}{(m+M)^2}.$$

Since $m = \lambda_{\min}(A)$ and $M = \lambda_{\max}(A)$, substituting these values gives the desired result:

$$\frac{(\mathbf{x}^T\mathbf{x})^2}{(\mathbf{x}^TA\mathbf{x})(\mathbf{x}^TA^{-1}\mathbf{x})} \geq \frac{4\lambda_{\max}(A)\lambda_{\min}(A)}{(\lambda_{\max}(A)+\lambda_{\min}(A))^2}.$$

□

4.5 Lemma

Lemma

Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient descent method with exact line search for finding the minimizer of $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$. Then, for any $k = 0, 1, 2, \ldots$,

$$f(\mathbf{x}_{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(\mathbf{x}_k)$$

where $M = \lambda_{\max}(A)$ and $m = \lambda_{\min}(A)$.

Proof.

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k + t_k \mathbf{d}_k)$$

Using the result from Section 3.4, we have

$$f(\mathbf{x}_{k+1}) = \left(1 - \frac{(\mathbf{d}_k^\top A \mathbf{d}_k)^2}{(\mathbf{d}_k^\top A \mathbf{d}_k)(\mathbf{d}_k^\top A^{-1} \mathbf{d}_k)}\right) f(\mathbf{x}_k).$$

Using Kantarovich Inequality, we have

$$\begin{aligned}
\frac{(\mathbf{d}_k^\top A \mathbf{d}_k)^2}{(\mathbf{d}_k^\top A \mathbf{d}_k)(\mathbf{d}_k^\top A^{-1} \mathbf{d}_k)} &\geq \frac{4mM}{m+M} \\
\implies f(\mathbf{x}_{k+1}) &\leq \left(1 - \left(\frac{4mM}{m+M}\right)^2\right) f(\mathbf{x}_k) \\
\implies f(\mathbf{x}_{k+1}) &\leq \left(\frac{M-m}{M+m}\right)^2 f(\mathbf{x}_k)
\end{aligned}$$

□

5 Condition Number

Condition Number

Let A be an $n \times n$ positive definite matrix. Then the **condition number** of A is defined as

$$\chi(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

- For quadratic functions with large condition number, gradient method might require large number of iterations to converge.
- Matrices with large condition number are called **ill conditioned**.
- Matrices with small condition number are called **well conditioned**.
- In case of non-quadratic functions, the rate of convergence of $\mathbf{x}_k$ to a given stationary point $\mathbf{x}^*$ depends on the condition number of $\nabla^2 f(\mathbf{x}^*)$.

From the lemma in section 4.5, we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq \left(\frac{M-m}{M+m}\right)^2 f(\mathbf{x}_k) \\
&= \left(\frac{\frac{M}{m}-1}{\frac{M}{m}+1}\right)^2 f(\mathbf{x}_k) \\
&= \left(\frac{\chi-1}{\chi+1}\right)^2 f(\mathbf{x}_k)
\end{aligned}$$

Applying this recursively, we get

$$f(\mathbf{x}_{k+1}) \leq \left(\frac{\chi - 1}{\chi + 1}\right)^{2k} f(\mathbf{x}_0)$$

Let the function $g(\chi) = \frac{\chi-1}{\chi+1}$. Since $g(\chi)$ is increasing in χ, a smaller value of χ results in a smaller value of $g(\chi)$, which leads to faster convergence.
Thus, a smaller condition number leads to faster convergence

5.1 Example: Rosenbrock Function

The Rosenbrock function is the following function:

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

The optimal solution is $(1, 1)$ with optimal value 0
The gradient and hessian are given by

$$\nabla f = \begin{pmatrix} -400x_1(x_2 - x_1)^2 - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix}$$
$$\nabla^2 f = \begin{pmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}$$

At the stationary point $(1, 1)$, the hessian is

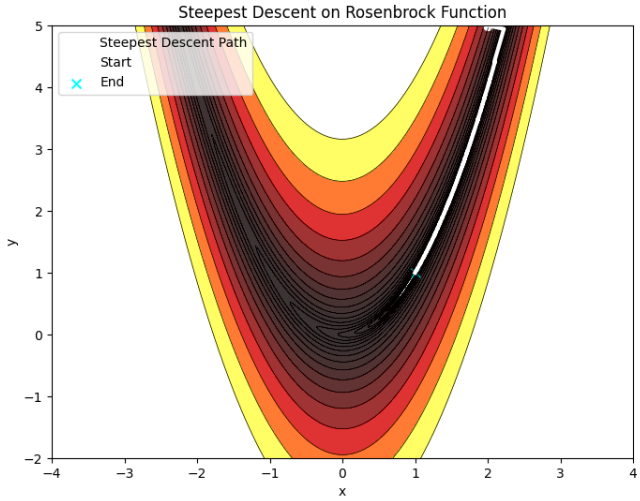
$$\nabla^2 f(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$$

The condition number of $\nabla^2 f(1, 1)$ is 2508

5.2 Steepest Descent with Backtracking on Rosenbrock Function

- Starting point $x_0 = [2, 5]^T$. The run required 6891 iterations. So, ill-conditioning of $\nabla^2 f(1, 1)$ has a significant impact.

```
iter_number =    1 norm_grad = 822.68098 fun_val = 101.00000
iter_number =    2 norm_grad = 118.25448 fun_val = 3.22102
iter_number =    3 norm_grad =   0.72305 fun_val = 1.49659
iter_number =    4 norm_grad =   0.73237 fun_val = 1.49645
iter_number =    5 norm_grad =   0.74206 fun_val = 1.49631
     .
     .
     .
iter_number = 6887 norm_grad =   0.00001 fun_val = 0.00000
iter_number = 6888 norm_grad =   0.00027 fun_val = 0.00000
iter_number = 6889 norm_grad =   0.00001 fun_val = 0.00000
iter_number = 6890 norm_grad =   0.00002 fun_val = 0.00000
iter_number = 6891 norm_grad =   0.00001 fun_val = 0.00000
```



[H]

5.3 Theorem

Theorem

Gradient descent with exact line search for strictly convex quadratic functions converges in a single step if the condition number is 1, irrespective of the initial point

Proof. Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^T\mathbf{x}$ The gradient and hessian are given by

$$\nabla f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$$
$$\nabla^2 f(\mathbf{x}) = A$$

Since f is strictly convex, using the second order characterization of convex functions we can say that A is symmetric and positive definite.
Since the condition number of A is 1, this implies that all eigenvalues are equal to (say) λ. Since A is symmetric positive definite, it will have an orthonormal eigenbasis
Consider the eigendecomposition of the matrix A.

$$A = Q\Lambda Q^\top$$

Since all the eigenvalues are equal, we have

$$\begin{aligned} & \Lambda = \lambda I \\ \implies & A = Q(\lambda I)Q^\top \\ & = \lambda(QQ^\top) \\ & = \lambda I \end{aligned}$$

Now consider gradient descent with exact line search on f, with the intial point $\mathbf{x}_0$. The descent direction $\mathbf{d}_0$ is given by

$$\mathbf{d}_0 = -\nabla f(\mathbf{x}_0)$$

To find the step size α_0, we need to minimize

$$\phi(\alpha) = f(\mathbf{x}_0 + \alpha \mathbf{d}_0)$$

Since f is strictly convex, ϕ will also be strictly convex, and the minimizer will satisfy

$$\begin{aligned} \phi'(\alpha_0) &= 0 \\ \implies \nabla f(\mathbf{x}_0 + \alpha \mathbf{d}_0)^\top \mathbf{d}_0 &= 0 \\ (\lambda I(\mathbf{x}_0 + \alpha \mathbf{d}_0) + \mathbf{b})^\top \mathbf{d}_0 &= 0 \end{aligned}$$

Substituting $\mathbf{d}_0 = -\nabla f(\mathbf{x}_0) = -(A\mathbf{x}_0 + \mathbf{b})$ and solving, we get

$$\alpha_0 = \frac{1}{\lambda}$$

Substituting and solving for $\mathbf{x}_1$, we get

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{x}_0 + \alpha_0 \mathbf{d}_0 \\ &= \mathbf{x}_0 - \frac{1}{\lambda} \nabla f(\mathbf{x}_0) \\ &= x_0 - \frac{1}{\lambda}(\lambda I \mathbf{x}_0 + \mathbf{b}) \\ &= -\frac{1}{\lambda}\mathbf{b} \end{aligned}$$

$$\begin{aligned} \nabla f(\mathbf{x}_1) &= A\mathbf{x}_1 + \mathbf{b} \\ &= \lambda I \left(-\frac{1}{\lambda}\mathbf{b} \right) + \mathbf{b} \\ &= \mathbf{0} \end{aligned}$$

Since the gradient at $\mathbf{x}_1$ is $\mathbf{0}$, the search terminates after 1 step. Also the strong convexity of f implies that $\mathbf{x}_1$ is the global minima □

References

[1] Introduction to Linear Optimization by Dimitris Bertsimas, John N. Tsitsiklis

[2] Wikipedia Integer Programming

[3] Introduction to NonLinear Optimization by Amir Beck