# Optimization Methods (CS1.404), Spring 2024
# Lecture 12

**Naresh Manwani**

Machine Learning Lab, IIIT-H

February 19th, 2024

# Backtracking Line Search With Armijo

## Backtracking

1. **Initialize:** $\alpha^{(0)} \in (0,1)$, $\tau \in (0,1)$, $l = 0$
2. Until $f(\mathbf{x}_k + \alpha^{(l)}\mathbf{d}_k) > f(\mathbf{x}_k) + c_1\alpha^{(l)}\nabla f(\mathbf{x}_k)^T\mathbf{d}_k$
    1. Set $\alpha^{(l+1)} = \tau\alpha^{(l)}$
    2. $l = l + 1$
3. $\alpha_k = \alpha^{(l)}$

In practice the following choices are used

- $\tau \in (0.1, 0.5]$
- $c_1 \in [10^{-5}, 10^{-1}]$

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

- Let $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, where $A$ is a symmetric positive definite matrix.
- For Steepest descent, $\mathbf{d}_k = -\nabla f(\mathbf{x}_k) = -2A\mathbf{x}_k$.
- Exact line search will result in $t_k = \arg\min_{t\geq 0}\ f(\mathbf{x}_k + t\mathbf{d}_k) = \frac{\mathbf{d}_k^T\mathbf{d}_k}{2\mathbf{d}_k^T A\mathbf{d}_k}$. Using this, we get

$$
\begin{aligned}
f(\mathbf{x}_k + t_k\mathbf{d}_k) &= f(\mathbf{x}_k) + t_k^2\mathbf{d}_k^T A\mathbf{d}_k + 2t_k\mathbf{d}_k^T A\mathbf{x}_k \\
&= \mathbf{x}_k^T A\mathbf{x}_k + \frac{(\mathbf{d}_k^T\mathbf{d}_k)^2}{4\mathbf{d}_k^T A\mathbf{d}_k} + \frac{\mathbf{d}_k^T\mathbf{d}_k}{2\mathbf{d}_k^T A\mathbf{d}_k}\mathbf{d}_k^T(-\mathbf{d}_k) \\
&= \mathbf{x}_k^T A\mathbf{x}_k - \frac{1}{4}\frac{(\mathbf{d}_k^T\mathbf{d}_k)^2}{\mathbf{d}_k^T A\mathbf{d}_k} = \mathbf{x}_k^T A\mathbf{x}_k\left(1 - \frac{1}{4}\frac{(\mathbf{d}_k^T\mathbf{d}_k)^2}{(\mathbf{d}_k^T A\mathbf{d}_k)(\mathbf{x}_k^T A\mathbf{x}_k)}\right) \\
&= \mathbf{x}_k^T A\mathbf{x}_k\left(1 - \frac{1}{4}\frac{(\mathbf{d}_k^T\mathbf{d}_k)^2}{(\mathbf{d}_k^T A\mathbf{d}_k)(\mathbf{x}_k^T AA^{-1}A\mathbf{x}_k)}\right) \\
&= \left(1 - \frac{(\mathbf{d}_k^T\mathbf{d}_k)^2}{(\mathbf{d}_k^T A\mathbf{d}_k)(\mathbf{d}_k^T A^{-1}\mathbf{d}_k)}\right) f(\mathbf{x}_k)
\end{aligned}
$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

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

## Kantorovich 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

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

holds.

## 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 $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$. 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 $m = \lambda_{min}(A)$.

# Condition Number and Convergence

## 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 $\mathbf{x}^*$ depend on the condition number of $\nabla^2 f(\mathbf{x}^*)$.

# 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 the optimal value 0.
- The Rosenbrock function is extremely ill conditioned at the optimal solution.

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

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

- $(1, 1)$ is unique stationary point.
- $\nabla^2 f(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$
- Condition number of $\nabla^2 f(1, 1)$ is $2.508 \times 10^3$

# Example: Steepest Descent with Backtracking on Rosenbrock Function

- Starting point $\mathbf{x}_0 = [2, 5]^T$. The run required 6890 iterations. So, ill conditioning of $\nabla^2 f(1, 1)$ has significant impact.

```
iter_number =    1 norm_grad = 118.254478 fun_val = 3.221022
iter_number =    2 norm_grad = 0.723051 fun_val = 1.496586
      :                      :                    :
iter_number = 6889 norm_grad = 0.000019 fun_val = 0.000000
iter_number = 6890 norm_grad = 0.000009 fun_val = 0.000000
```

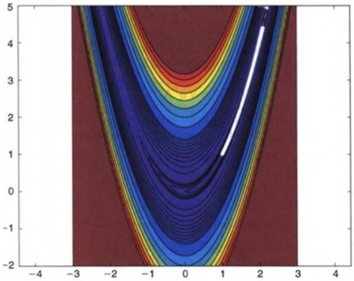



Figure: Banana shaped contour lines of the Rosenbrock function surrounding the unique stationary point $(1, 1)$. Along with it thousands of iterations of steepest descent.

# Convergence Analysis of Gradient Descent

## L-Smooth Functions

An $L$-smooth function is continuously differentiable and that its gradient $\nabla f$ is Lipschitz continuous over $\mathbb{R}^n$, meaning that there exists $L > 0$ for which

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \text{for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

The class of functions with Lipschitz gradient with constant $L$ are denoted by $\mathbb{C}_L^{1,1}$.

**Examples:**

- **Linear Functions:** Given $\mathbf{a} \in \mathbb{R}^n$, the function $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ is in $\mathbb{C}_0^{1,1}$.
- **Quadratic Functions:** Let $A$ be an $n \times n$ symmetric matrix, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then,

$$\|f(\mathbf{x}) - f(\mathbf{y})\| = 2\|A(\mathbf{x} - \mathbf{y})\| \leq 2\|A\|.\|\mathbf{x} - \mathbf{y}\|$$

Thus, the Lipschitz constant of $\nabla f$ is $2\|A\|$.

# Interpretation of $L$-Smoothness

- The gradient of a functions measures how the function changes when we move in a particular direction from a point.
- If the gradient were to change arbitrarily quickly, the old gradient does not give us much information at all even if we take a small step.
- In contrast, smoothness assures us that the gradient cannot change too quickly. Therefore, we have an assurance that the gradient information is informative within a region around where it is taken. The implication is that we can decrease the function's value by moving the direction opposite of the gradient.

**Theorem 4.20 (Chapter 4: Introduction to Nonlinear Optimization by Amir Beck)**

Let $f$ be a twice continuously differentiable function over $\mathbb{R}^n$. Then the following two claims are equivalent.

- $f \in \mathbb{C}_L^{1,1}(\mathbb{R}^n)$
- $\|\nabla^2 f(\mathbf{x})\| \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$.

See the proof in the book.

**Example:** Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = \sqrt{1 + x^2}$. Then,

$$0 \leq f''(x) = \frac{1}{(1 + x^2)^{3/2}} \leq 1$$

for any $x \in \mathbb{R}$. Thus, $f \in \mathbb{C}_1^{1,1}$.

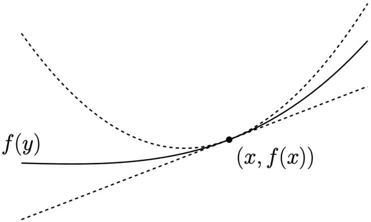
# Descent Property of $L$-Smooth Functions

**Lemma 4.22 (Chapter 4: Introduction to Nonlinear Optimization by Amir Beck)**

Let $D \subseteq \mathbb{R}^n$ and $f \in \mathbb{C}_L^{1,1}(D)$ for some $L > 0$. Then, for any $\mathbf{x}, \mathbf{y} \in D$ satisfying $[\mathbf{x}, \mathbf{y}] \subseteq D$, it holds that

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

See the proof in the book.




Comments:

1. This result shows that an $L$-smooth function can be bounded above by a quadratic function over the entire space.
2. This result is very useful in the convergence proofs of gradient based methods.

# Descent Property of Steepest Descent for $L$-Smooth Functions

**Lemma (Sufficient Decrease of the Gradient Method)**

Suppose that $f \in \mathbb{C}_L^{1,1}(\mathbb{R}^n)$. Let $\{\mathbf{x}_k\}_{k \geq 0}$ be the sequence generated by the gradient method for solving $\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$ with one of the following stepsize strategies:

- constant stepsize $\bar{t} \in (0, \frac{2}{L})$
- exact line search
- backtracking procedure with parameters $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, $\beta \in (0,1)$.

Then for any $\mathbf{x} \in \mathbb{R}^n$ and $t > 0$

$$f(\mathbf{x}) - f(\mathbf{x} - t\nabla f(\mathbf{x})) \geq M\|\nabla f(\mathbf{x})\|^2$$

where

$$M = \begin{cases} \bar{t}\left(1 - \frac{\bar{t}L}{2}\right), & \text{constant stepsize} \\ \frac{1}{2L}, & \text{exact line search} \\ \alpha \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\}, & \text{backtracking} \end{cases}$$

- Above result shows that at each iteration the decrease in the function value is at least a constant times the squared norm of the gradient.

# Convergence of the Steepest Descent for $L$-Smooth Functions

**Lemma (Sufficient Decrease of the Gradient Method)**

Suppose that $f \in \mathbb{C}_L^{1,1}(\mathbb{R}^n)$. Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient method for solving $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$ with one of the following stepsize strategies:

- constant stepsize $\bar{t} \in (0, \frac{2}{L})$
- exact line search
- backtracking procedure with parameters $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, $\beta \in (0,1)$.

Assume that $f$ is bounded below over $\mathbb{R}^n$, that is, there exists $m \in \mathbb{R}$ such that $f(\mathbf{x}) > m$ for all $\mathbf{x} \in \mathbb{R}^n$. Then we have the following:

1. The sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$ is non-increasing. In addition, for any $k \geq 0$, $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ unless $\nabla f(\mathbf{x}_k) = \mathbf{0}$.
2. $\nabla f(\mathbf{x}_k) \to \mathbf{0}$ as $k \to \infty$.