IIIT Hyderabad **Scribed By: Bhav Beri (2021111013)**
Team: 12A **Harshit Aggarwal (2021111015)**

Optimization Methods(CS1.404) **Lecture #15**
Instructor: Dr. Naresh Manwani **6^{th} March 2025**

Outline. *This scribe covers the convergence of gradient descent for quadratic and L-smooth functions, including the impact of the condition number and various step-size strategies. It further delves into the properties of L-smooth functions and their descent properties, providing theoretical foundations and examples.*

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

1.1 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 $k = 0, 1, 2, \dots,$

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

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

Note: Proof discussed in last class.

1.2 Condition Number

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

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

- In numerical analysis, the condition number of a function measures how much the output value of the function can change for a small change in the input argument. This is used to measure how sensitive a function is to changes or errors in the input and how much error in the output results from an error in the input.
- Condition number is just the ratio of the maximum and the minimum eigenvalues of A.
- For quadratic functions with large condition numbers, the gradient method might require a large number of iterations to converge.
- Matrices with large condition numbers are called **ill conditioned**.
- Matrices with small condition numbers 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 $\nabla^2 f(\mathbf{x}^*)$.
- In Machine Learning, we do normalization just to make the condition number lesser and the contours of the function as circular as possible, thereby reducing the time/number of iterations to converge.

1.3 Example 1: Condition Number = 1

$$f(x_1, x_2) = (x_1 - 7)^2 + (x_2 - 2)^2$$

$$A = \nabla^2 f(x_1, x_2) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

which implies,

$$\lambda_1 = \lambda_2 = 2$$

$$\chi(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \frac{\lambda_1}{\lambda_2} = \frac{2}{2} = 1$$

We get,

$$\chi(A) = 1$$

By plotting the function, we can directly see that the minimizer of $f(x_1, x_2)$ is $(x_1^*, x_2^*) = (7, 2)$.

Now, we'll perform gradient descent with exact line search on $f(x_1, x_2)$:
Let's take the initial point as $(x_1^{(0)}, x_2^{(0)})$.

$$\nabla f(x_1, x_2) = \begin{bmatrix} 2(x_1^{(0)} - 7) \\ 2(x_2^{(0)} - 2) \end{bmatrix}$$

Let,

$$\overline{X}^{(0)} = \left[x_1^{(0)}, x_2^{(0)} \right]$$

&

$$\overline{X}^{(1)} = \left[x_1^{(1)}, x_2^{(1)} \right]$$

Now,

$$\overline{X}^{(1)} = \overline{X}^{(0)} - \alpha_0 \nabla f(\overline{X}^{(0)}) \tag{1}$$

where, $\alpha_0 = \arg\min_{\alpha > 0} f(\overline{X}^{(0)} + \alpha \bar{d}^{(0)})$ & $\bar{d}^{(0)} = -\nabla f(\overline{X}^{(0)})$

$$\begin{aligned} \overline{X}^{(0)} + \alpha \bar{d}^{(0)} &= \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} - \alpha \begin{bmatrix} 2(x_1^{(0)} - 7) \\ 2(x_2^{(0)} - 2) \end{bmatrix} \\ &= \begin{bmatrix} x_1^{(0)} - 2\alpha(x_1^{(0)} - 7) \\ x_2^{(0)} - 2\alpha(x_2^{(0)} - 2) \end{bmatrix} \\ f(\overline{X}^{(0)} + \alpha \bar{d}^{(0)}) &= \left[x_1^{(0)} - 2\alpha(x_1^{(0)} - 7) - 7 \right]^2 + \left[x_2^{(0)} - 2\alpha(x_2^{(0)} - 2) - 2 \right]^2 \end{aligned}$$

Now, using the exact line search,

$$\alpha_0 = \arg\min_{\alpha>0} f(\overline{X}^{(0)} + \alpha d^{(0)}))$$

$$\begin{aligned}
f'(\overline{X}^{(0)} + \alpha d^{(0)}) &= \frac{df(\overline{X}^{(0)} + \alpha \bar{d}^{(0)})}{d\alpha} \\
&= 2(x_1^{(0)} - 2\alpha(x_1^{(0)} - 7) - 7)(-2)(x_1^{(0)} - 7) + 2(x_2^{(0)} - 2\alpha(x_2^{(0)} - 2) - 2)(-2)(x_2^{(0)} - 2) \\
&= -4(x_1^{(0)} - 2\alpha(x_1^{(0)} - 7) - 7)(x_1^{(0)} - 7) - 4(x_2^{(0)} - 2\alpha(x_2^{(0)} - 2) - 2)(x_2^{(0)} - 2)
\end{aligned}$$

Now, on equating $f'(\overline{X}^{(0)} + \alpha \bar{d}^{(0)}) = 0$, we get

$$\begin{aligned}
0 &= 4(x_1^{(0)} - 7)^2 - 8\alpha(x_1^{(0)} - 7)^2 + 4(x_2^{(0)} - 2)^2 - 8\alpha(x_2^{(0)} - 2)^2 \\
0 &= 4(1 - 2\alpha)[(x_1^{(0)} - 7)^2 + (x_2^{(0)} - 2)^2] \\
\implies 0 &= (1 - 2\alpha) \\
\implies \alpha &= \frac{1}{2}
\end{aligned}$$

Using Equation 1,

$$\begin{aligned}
\overline{X}^{(1)} &= \overline{X}^{(0)} + \alpha_0 \bar{d}^{(0)} \\
&= \begin{bmatrix} x_1^{(0)} - 2 \times \frac{1}{2}(x_1^{(0)} - 7) \\ x_2^{(0)} - 2 \times \frac{1}{2}(x_2^{(0)} - 2) \end{bmatrix} \\
&= \begin{bmatrix} 7 \\ 2 \end{bmatrix}
\end{aligned}$$

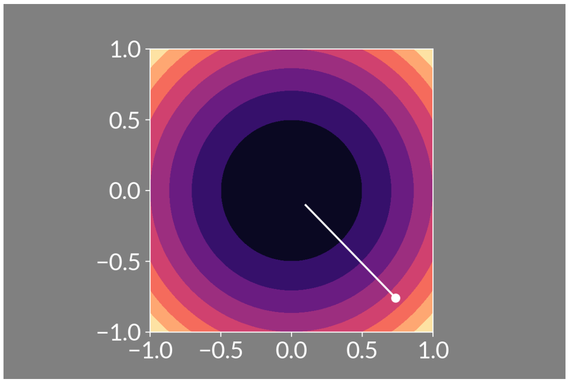

Figure 1: Demonstration of contour optimization plot on gradient descent with Condition Number = 1, i.e. symmetric contours, converging in a single step.

Remark

Gradient descent with exact line search for strictly convex (so that $m > 0$, otherwise condition number will tend to ∞) quadratic functions converge in a single step if the condition number is 1, irrespective of the initial starting point.
Condition number $= 1$ also implies that the contours are circular.

1.4 Example 2: Condition Number $\neq 1$

Let's take $f(x_1, x_2) = 4x_1^2 + x_2^2 - 2x_1x_2$

$$\nabla f(x_1, x_2) = \begin{bmatrix} 8x_1 - 2x_2 \\ 2x_2 - 2x_1 \end{bmatrix}$$

$$A = \nabla^2 f(x_1, x_2) = \begin{bmatrix} 8 & -2 \\ -2 & 2 \end{bmatrix}$$

As we can see that $Tr(A) = 10 > 0, Det(A) = 12 > 9 \implies A$ is positive definite matrix.

$$\lambda_1 = 5 + \sqrt{13}, \lambda_2 = 5 - \sqrt{13}$$

$$\chi(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \frac{\lambda_1}{\lambda_2} = \frac{5 + \sqrt{13}}{5 - \sqrt{13}} \approx 6.171$$

On performing the gradient descent, we get the minimizer as $(x_1^*, x_2^*) = (0, 0)$.

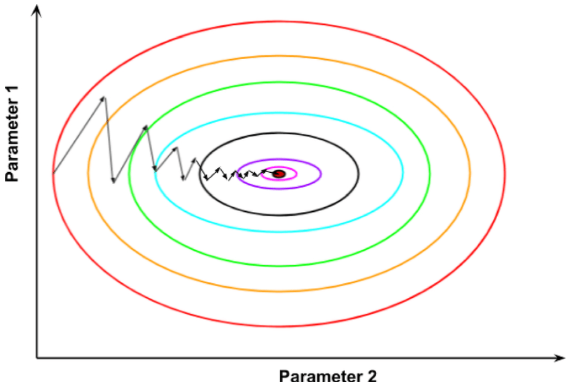

Figure 2: Demonstration of contour optimization plot on gradient descent with Condition Number $\neq 1$

If $\overline{X}^{(0)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, then gradient descent with exact line search will converge in 26 iterations. While if $\overline{X}^{(0)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, then it converges in 4 iterations.
Thus, from this, we can conclude that if condition number $\neq 1$, then the convergence depends on the initial point.
We also observe that the convergence line shows zig-zag patterns of the iterates (Figure 2).

1.5 The Rosenbrock Function

The Rosenbrock function is a well-known test function for optimization algorithms, often used to evaluate the performance of gradient-based methods. It is defined as:

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

This function exhibits a narrow, curved valley containing the global minimum, making it challenging for gradient descent methods, especially when poorly conditioned.

1.5.1 Optimal Solution

The optimal solution of the Rosenbrock function is:

$$(x_1^*, x_2^*) = (1, 1),$$

with the optimal function value:

$$f(1, 1) = 0.$$

1.5.2 Gradient and Hessian of the Rosenbrock Function

The gradient of $f(x)$ is given by:

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

The Hessian matrix, $\nabla^2 f(\mathbf{x})$, is:

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

1.5.3 Conditioning at the Optimal Solution

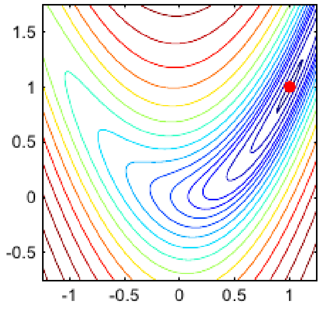

Figure 3: Contour plot of the Rosenbrock banana function with red dot being the optimal solution at $(1, 1)$.

At the optimal point $(1, 1)$, the Hessian matrix evaluates to:

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

The condition number of this Hessian matrix is:

$$\chi(\nabla^2 f(1,1)) = 2.508 \times 10^3.$$

This large condition number indicates that the Rosenbrock function is highly ill-conditioned at the optimal solution.

Importance of Rosenbrock function

Due to its narrow, curved valley leading to the optimal solution, the Rosenbrock function poses significant challenges for first-order optimization methods like gradient descent. The high condition number at the optimal point indicates that the function's landscape stretches more along one direction than another, causing slow convergence and requiring careful step size selection. This makes it a valuable benchmark for testing advanced optimization techniques, such as momentum-based methods, adaptive learning rates, and second-order approaches, which can better handle ill-conditioned problems.

2 L-Smooth Functions

2.1 Continuous functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous at $\overline{X} \in \mathbb{R}^n$ if for every $\epsilon > 0$, there exists $\delta(\overline{X}, \epsilon) > 0$, such that

$$|f(\overline{X}) - f(\overline{Y})| \leq \epsilon \quad \forall \quad \overline{Y} \in B[\overline{X}, \delta(\overline{X}, \epsilon)]$$

2.2 Lipschitz Continuous Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous if there exists $L > 0$ such that

$$|f(\overline{X}) - f(\overline{Y})| \leq L||\overline{X} - \overline{Y}|| \quad \forall \overline{X}, \overline{Y} \in \mathbb{R}^n$$

Note about Lipschitz Continuous Functions

A Lipschitz continuous function provides a stronger form of continuity by ensuring that changes in function values are uniformly bounded by a constant factor of the input variations. The inequality $L\|\overline{X} - \overline{Y}\|$ serves as an upper bound, meaning that the function's rate of change is controlled globally rather than depending on local properties of $\overline{X}$. The slope of function at any point is upper-bounded by L, thus making the variations in the function bounded and function to have smooth curves, ensuring that the function does not oscillate too rapidly.

Note that if ∇f is Lipschitz with constant L, then it is also Lipschitz with constant $\tilde{L}$ for all $\tilde{L} \geq L$. Therefore, there are essentially infinite number of Lipschitz constants for a function with Lipschitz gradient. Frequently, we are interested in the smallest possible Lipschitz constant.

2.3 L-Smooth Functions

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

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

Here, the norm of Hessian is upper-bouneded by L. The class of functions with Lipschitz gradient with constant L is denoted by $\mathcal{C}_L^{1,1}(\mathbb{R}^n)$ or just $\mathcal{C}_L^1$. Occasionally, when the exact value of the Lipschitz constant is unimportant, it maybe denoted directly by $\mathcal{C}^{1,1}$. The following are some simple examples of $\mathcal{C}^{1,1}$ functions:

- **Linear functions:** Given $\mathbf{a} \in \mathbb{R}^n$, the function $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ is in $\mathcal{C}_0^{1.1}$.
- **Sigmoid Function:** Derivative of Sigmoid function ($\sigma(x) = \frac{1}{1+e^{-x}}$) is a $\mathcal{C}^{1,1}$ function.

  Consider the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. Its derivative is $\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)(1 - \sigma(x))$, and the second derivative is $\sigma''(x) = \sigma'(x)(1 - \sigma(x)) - \sigma(x)\sigma'(x) = \sigma'(x)(1 - 2\sigma(x))$.

  Since $0 < \sigma(x) < 1$, we have $|\sigma(x)| < 1$ and $|1 - 2\sigma(x)| < 1$. Also, $\sigma'(x) = \sigma(x)(1 - \sigma(x))$ is maximized when $\sigma(x) = 1/2$ and its maximum value is $1/4$. Therefore $|\sigma''(x)| \leq |\sigma'(x)||1 - 2\sigma(x)| < \frac{1}{4} \cdot 1 = \frac{1}{4}$.

  By the Mean Value Theorem, for any $x, y \in \mathbb{R}$, there exists a c between x and y such that

  $$|\sigma'(x) - \sigma'(y)| = |\sigma''(c)||x - y| < \frac{1}{4}|x - y|$$

  Thus, $\sigma'(x)$ is Lipschitz continuous with Lipschitz constant $L = \frac{1}{4}$ or the $\sigma(x)$ (Sigmoid function) is L-smooth.
- **Quadratic functions:** Let $\mathbf{A}$ be an $n \times n$ symmetric matrix, $\mathbf{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Then the function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$ is a $\mathcal{C}^{1,1}$ function.

$$\begin{aligned} f(\overline{X}) &= \overline{X}^T\mathbf{A}\overline{X} + 2\mathbf{b}^T\overline{X} + c \\ f(\overline{X}) = \nabla f(\overline{X}) &= 2\mathbf{A}\overline{X} + \mathbf{b} \\ \implies ||f(\overline{X}) - f(\overline{Y})|| &= 2||\mathbf{A}(\overline{X} - \overline{Y})|| \\ \text{Using Cauchy-Schwartz Inequality,} & \\ ||f(\overline{X}) - f(\overline{Y})|| &\leq 2||\mathbf{A}||||\overline{X} - \overline{Y}|| \\ \text{here, } ||\mathbf{A}|| \text{ denote the Frobenius norm for matrices.} & \\ \implies L &= ||\mathbf{A}|| \end{aligned}$$

2.4 Interpretation of L-Smoothness

The gradient of a function quantifies the rate of change of the function as we move in a specific direction from a given point. However, if the gradient varies arbitrarily fast, its previous values become unreliable for predicting function behavior, even for infinitesimally small steps.

Smoothness conditions provide a crucial constraint on this variability. If the gradients of a function change very rapidly, then they won't give any information about previous gradients. A function with a Lipschitz-continuous gradient ensures that the gradient does not change too abruptly, guaranteeing a level of predictability in its behavior. This property is fundamental in optimization, as it allows gradient-based methods to make informed updates, ensuring stable convergence. Consequently, smoothness enables us to systematically reduce function values by moving in the opposite direction of the gradient, forming the basis of many descent-based optimization techniques.

2.5 Theorem on Properties of L-Smooth Function

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

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

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, $f \in C_1^{1,1}$.

Proof. (b) ⇒ (a). Suppose that $||\nabla^2 f(x)|| \leq L$ for any $x \in \mathbb{R}^n$. Then by the fundamental theorem of calculus we have for all $x, y \in \mathbb{R}^n$

$$\begin{aligned}\nabla f(y) &= \nabla f(x) + \int_0^1 \nabla^2 f(x + t(y-x))(y-x)dt \\ &= \nabla f(x) + \left(\int_0^1 \nabla^2 f(x + t(y-x))dt\right) \cdot (y-x),\end{aligned}$$

Thus,

$$\begin{aligned}||\nabla f(y) - \nabla f(x)|| &= \left\|\left(\int_0^1 \nabla^2 f(x + t(y-x))dt\right) \cdot (y-x)\right\| \\ &\leq \left\|\int_0^1 \nabla^2 f(x + t(y-x))dt\right\| ||y-x|| \\ &\leq \left(\int_0^1 ||\nabla^2 f(x + t(y-x))||dt\right) ||y-x|| \\ &\leq L||y-x||,\end{aligned}$$

establishing the desired result $f \in C_L^{1,1}$.

(a) ⇒ (b). Suppose now that $f \in C_L^{1,1}$. Then by the fundamental theorem of calculus for any $d \in \mathbb{R}^n$ and $\alpha > 0$ we have

$$\nabla f(x + \alpha d) - \nabla f(x) = \int_0^\alpha \nabla^2 f(x + td)dt.$$

Thus,

$$\left\|\left(\int_0^\alpha \nabla^2 f(x+td)dt\right)d\right\| = ||\nabla f(x+\alpha d) - \nabla f(x)|| \leq \alpha L||d||.$$

Dividing by α and taking the limit $\alpha \to 0^+$, we obtain

$$||\nabla^2 f(x)d|| \leq L||d||,$$

implying that $||\nabla^2 f(x)|| \leq L$.

2.6 Descent Property of L-smooth Functions

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

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

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.

Proof. Using the fundamental theorem of calculus, we have

$$f(y) - f(x) = \int_0^1 \nabla f(x+t(y-x))^T(y-x)dt$$

$$f(y) - f(x) = \nabla f(x)^T(y-x) + \int_0^1 (\nabla f(x+t(y-x)) - \nabla f(x))^T(y-x)dt$$

using cauchy swartz,

$$\leq \nabla f(x)^T(y-x) + \int_0^1 ||\nabla f(x+t(y-x) - \nabla f(x))||||y-x||dt$$

$$= \nabla f(x)^T(y-x) + \int_0^1 Lt||y-x||^2 dt$$

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

2.7 Descent Property of Steepest Descent for L-Smooth Functions

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

- constant stepsize $\bar{t} \in \left(0, \frac{2}{L}\right)$
- exact line search

- backtracking procedure with parameters $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, $\beta \in (0,1)$.

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

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

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.

Proof.

2.7.1 Constant step size

Let $y = x - t\nabla f(x)$.
Using L-smoothness property of ∇f for x and y, we get

$$\begin{aligned} f(\bar{x} - t\nabla f(x)) &= f(x) - t\nabla f(x) + \frac{Lt^2}{2}\|\nabla f(x)\|^2 \\ &= f(x) - t\|\nabla f(x)\|^2 + \frac{Lt^2}{2}\|\nabla f(x)\|^2 \\ &= f(x) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x)\|^2 \quad (1) \end{aligned}$$

For the constant stepsizer, we want that

$$f(x) - f(x - t\nabla f(x)) \geq 0$$

to ensure this,

$$t > 0 \quad \text{and} \quad t \leq \frac{2}{L}$$

For any $t \in (0, 2/L)$, we have

$$f(x) - f(x - t\nabla f(x)) \geq t(1 - \frac{Lt}{2})||\nabla f(x)||^2$$

2.7.2 Exact Line Search

We want the maximum decrement along the direction $-\nabla f(x)$.
We have using equation (1),

$$f(x_k) - f(x_{k+1}) \geq t(1 - \frac{Lt}{2})||\nabla f(x_k)||^2$$

Maximizing $t(1-\frac{Lt}{2})$ w.r.t. t gives $t=\frac{1}{L}$.
For $t=\frac{1}{L}$, we get

$$f(x_k)-f(x_k-\frac{1}{L}\nabla f(x))\geq\frac{1}{2L}||\nabla f(x_k)||^2 \quad (2)$$

Let $t_k=\underset{t>0}{\operatorname{argmin}} f(x_k-t\nabla f(x_k))$ be the step size with exact line search. So, we have
$f(x_{k+1})=f(x_k-t_k\nabla f(x_k))\leq f(x_k-\frac{1}{L}\nabla f(x_k)) \quad \forall t\geq 0.$ (3)
Using equations 2 and 3, $f(x_k)-f(x_{k+1})=f(x_k)-f(x_k-t_k\nabla f(x_k))\geq\frac{1}{2L}||\nabla f(x_k)||^2$
So, for exact line search, we have $M=1/2L$.

2.7.3 Backtracking

Will be covered in next class.

References

[1] Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB by Amir Beck

[2] Wikipedia: Integer Programming

[3] https://blogs.mathworks.com/cleve/2017/07/17/what-is-the-condition-number-of-a-matrix/

[4] https://www.brnt.eu/phd/node10.html

[5] Biswas, S., Nath, S., Dey, S. et al. Tangent-cut optimizer on gradient descent: an approach towards Hybrid Heuristics. Artif Intell Rev 55, 1121–1147 (2022). https://doi.org/10.1007/s10462-021-09984-0