MPI - Lecture 1

Optimization

Introduction

Examples

Duration of a text processing program (1 of 6)

Problem 1. *Imagine the following situation: You have created a program that processes a text input by a user. You know, from theoretical analysis of the source code and algorithms used within the program, that it is impossible to determine the exact time needed to process a text of length* k*. However, you know that it is* ***approximately*** *proportional to the length of the text.*

Mathematically*: Denote* $t(k)$ *the "average" number of seconds needed to process a text of length* k*. We know that*

$$t(k) \approx \alpha k \qquad \text{for some } \alpha \in \mathbb{R}.$$

Problem*: The proportionality constant* α *is unknown. How would you reasonably estimate its value?*

Duration of a text processing program (2 of 6)

Sketch of a solution:

1. Run the program for several, say n, texts of various lengths and measure the actual running times. This gives us n couples of measurements $(k_1, t_1), (k_2, t_2), \ldots, (k_n, t_n)$.

2. For a given α, we can measure the approximation error $t(k) \approx \alpha k$ by this function:

$$e(\alpha) = (t_1 - \alpha k_1)^2 + (t_2 - \alpha k_2)^2 + \cdots + (t_n - \alpha k_n)^2 = \sum_{i=1}^{n} (t_i - \alpha k_i)^2.$$

3. In order to find the best approximating proportionality constant α, we find the value of α for which the error $e(\alpha)$ is minimal:

an optimal value of α is a minimum point of the function $e(\alpha)$.

Duration of a text processing program (3 of 6)

How to find a minimum point of $e(\alpha)$:

1. Find the first derivative $e'(\alpha)$:

$$e'(\alpha) = \left(\sum_{i=1}^{n} (t_i - \alpha k_i)^2 \right)' = \sum_{i=1}^{n} -2k_i(t_i - \alpha k_i)\,.$$

2. Find the critical points, i.e., the points α_0 where $e'(\alpha_0)$ is zero or does not exist:

$$e'(\alpha_0) = 0 \Leftrightarrow \sum_{i=1}^{n} -2k_i(t_i - \alpha_0 k_i) = 0 \Leftrightarrow \sum_{i=1}^{n} k_i t_i = \alpha_0 \sum_{i=1}^{n} k_i^2 \Leftrightarrow \alpha_0 = \frac{\sum_{i=1}^{n} k_i t_i}{\sum_{i=1}^{n} k_i^2}$$

3. The critical points are our candidates for the points of (local) minimal or maximal values of the function e. To be sure that the value of α we found is a minimum we need the second derivative:

$$e''(\alpha) = \left(\sum_{i=1}^{n} -2k_i(t_i - \alpha k_i) \right)' = \sum_{i=1}^{n} 2k_i^2\,.$$

...`continues`...

Duration of a text processing program (4 of 6)

We know that if $e''(\alpha_0) > 0$ (resp. $e''(\alpha_0) < 0$), then the critical point α_0 is a local strict minimum (resp. strict maximum) point.

If $e''(\alpha_0) = 0$, then α is neither of these two cases (maybe an inflexion point?).

Solution: based on our measurements $(k_1, t_1), (k_2, t_2), \ldots, (k_n, t_n)$, we get the best approximation $t(k) \approx \alpha k$ for

$$\alpha = \alpha_0 = \frac{\sum_{i=1}^{n} k_i t_i}{\sum_{i=1}^{n} k_i^2}.$$

Indeed, this α_0 is the unique (why unique?) global (why global?) minimum point of the approximation error function $e(\alpha)$ since the second derivative

$$e''(\alpha_0) = \sum_{i=1}^{n} 2k_i^2 \quad \text{is positive.}$$

Duration of a text processing program (5 of 6)

Problem 2 (slight modification)**.** *Imagine the following situation: You have created a program that processes a text input by a user. You know, from theoretical analysis of the source code and algorithms used within the program, that it is impossible to determine precisely the time needed to process a text of length k. However, you know that it is **approximately** proportional to the length of the text **and to the frequency of the processor**.*

***Mathematically**: Denote by $t(k, f)$ the "average" number of seconds needed to process a text of length k, and the frequency of the processor by f. We know that*

$$t(k, f) \approx \alpha k + \beta f \qquad \text{for some } \alpha, \beta \in \mathbb{R}.$$

***Problem**: The constants α and β are unknown. How would you reasonably estimate their values?*

Duration of a text processing program (6 of 6)

Sketch of solution:

1. Run the program for several, say n, texts of various lengths on computers with different frequencies and measure the actual running times. This gives us n triplets of measurements $(k_1, t_1, f_1), (k_2, t_2, f_2), \ldots, (k_n, t_n, f_n)$.

2. For a given α and β, we can measure the approximation error $t(k, f) \approx \alpha k + \beta f$ by this **two-variable** function:

$$e(\alpha, \beta) = (t_1 - \alpha k_1 - \beta f_1)^2 + (t_2 - \alpha k_2 - \beta f_2)^2 + \cdots + (t_n - \alpha k_n - \beta f_n)^2 =$$
$$= \sum_{i=1}^{n} (t_i - \alpha k_i - \beta f_i)^2.$$

3. In order to find the best approximating constants α and β, we find values of α and β for which the error $e(\alpha, \beta)$ is minimal: an optimal value of α and β is the "two-dimensional" minimum point of $e(\alpha, \beta)$.

Comments

Comments

Why "optimization"?

A typical situation in physics, engineering, economy, chemistry, etc. is that you have a function that measures your profit, your loss, the energy of something, etc.

The value of such function is given by one or more inputs and the relation between inputs and the resulting value is usually stated as a mathematical formula since all these sciences uses mathematical models to understand and quantify their subject of interest.

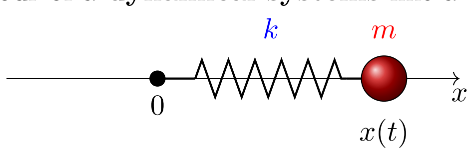
An example of such function is our function $e(\alpha, \beta)$ that measures the approximation error.

Typically, we want to maximize or minimize such functions (maximize the profit, the energy, minimize the loss, the error) which leads to the problem of finding **optimal** values of the inputs. Therefore the name "optimization".

Comments

There is **another very important usage** of the derivative.

Derivatives measure the rate of change of a function. This helps us to describe the behaviour of a **dynamical systems** like a ball on a spring:



The position of the ball at time t is a function $x(t)$ satisfying the differential equation

$$x''(t) + \omega^2 x(t) = 0\,.$$

The solution of this equation is

$$x(t) = x_0 \cos(\omega t) + \frac{v_0}{\omega} \sin(\omega t), \qquad t \in \mathbb{R},$$

where $x_0 = x(0)$ and v_0 are the position and the speed of the ball at time $t = 0$. This model is known as harmonic oscillator.

Univariate optimization

Derivative

How do we differentiate?

Example 3. *Find the first derivative of $f(x)$, where*

(a) $f(x) = x^3 + 4x^2 + 6$,

(b) $f(x) = \sin(x^3)$,

(c) $f(x) = e^x \sin x$.

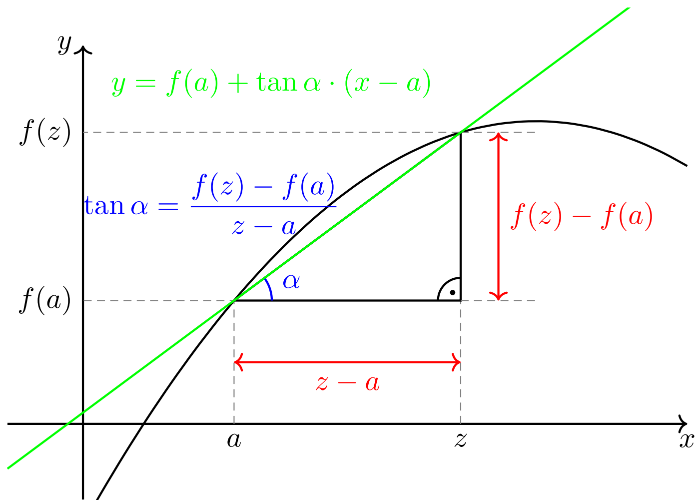
Solutions:

(a) $f'(x) = 3x^2 + 8x$,

(b) $f'(x) = 3x^2 \cos(x^3)$,

(c) $f'(x) = e^x \sin x + e^x \cos x$.

Geometrical meaning of derivative: tangent line (1 of 2)



Geometrical meaning of derivative: tangent line (2 of 2)

- The tangent line to the graph of a function $f(x)$ at a point x_0 is a straight line that "just touches" the curve at that point.
- Any straight line has equation $y = ax + b$, where a is the slope of the line.
- The slope of the tangent line to $f(x)$ at the point x_0 equals the first derivative evaluated at x_0: $f'(x_0)$.
- The tangent line at the point x_0 satisfies the equation

$$y = f'(x_0)(x - x_0) + f(x_0).$$

Derivative and optimization

With this geometrical explanation it is easy to see that the following statements are true:

- If $f'(x_0)$ is positive, then $f(x)$ is increasing at x_0.
- If $f'(x_0)$ is negative, then $f(x)$ is decreasing at x_0.
- If x_0 is a local minimum/maximum point of $f(x)$, then $f'(x_0) = 0$ or $f'(x_0)$ does not exists. Such points are called **critical points**.

Example 4. *Find all critical points of*

$$f(x) = \frac{x^3}{3} + 2x^2 + 3x + 6\,.$$

Second derivative

What does it mean that the second derivative $f''(x)$ is positive?

- The second derivative is a derivative of the first derivative; therefore the fact that $f''(x)$ is positive implies that $f'(x)$ is increasing (at the point x).
- If $f'(x)$ is increasing, then the function $f(x)$ is more and more increasing (if $f'(x) > 0$) or less and less decreasing (if $f'(x) < 0$).
- An illustrative example of function with positive second derivative is $f(x) = x^2$.

Second derivative as a criterion for extremal values

Again, if we understand the geometrical meaning of the second derivative, we can easily see that the following statements are true:

Theorem 5. *Let x_0 be a critical point of a function $f(x)$ such that $f'(x_0) = 0$ and $f''(x_0)$ exists.*

- *If $f''(x_0) > 0$, then the function is convex at x_0 and x_0 is a point of a (strict) minimum.*
- *If $f''(x_0) < 0$, then the function is concave at x_0 and x_0 is a point of a (strict) maximum.*

Question: what can happen if $f''(x_0) = 0$?

Universal cookbook of univariate optimization

Given a function $f(x)$, we want to find its extremal values.

1. Find the first derivative $f'(x)$.
2. Find the critical points: solve the equation $f'(x) = 0$ and find the points where the derivative **does not exist**.
3. Find the second derivative $f''(x)$.

4. If possible, for all critical points x_0 evaluate $f''(x_0)$ and decide whether this point is a point of minimum or maximum or whether it is an inflexion point. (Other critical points have to be treated by hand.)

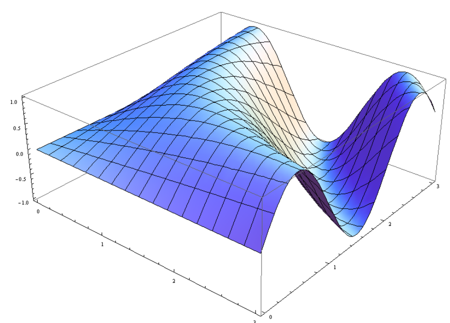
The goal of this and the next lecture is to understand what happens when we have more than 1 variable. We shall build a similar cookbook for such functions.

Multivariate optimization

Graph of multivariate functions (1 of 2)

For a univariate function $f(x)$, its graph is the set of points $(x, f(x))$ which can be depicted in Cartesian coordinate system (typically with x- and y-axis). [2mm]

What if the function depends on more variables? For instance: $f(x, y)$.

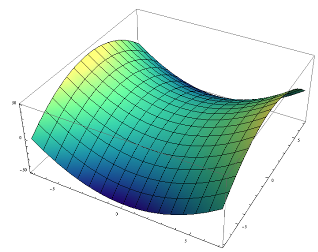


Graph of a two-variable function $\sin(x \cdot y)$: the set of points $(x, y, \sin(x \cdot y))$.

Graph of multivariate functions (2 of 2)

- To depict a graph of a two-variable function we need a third axis (typically z-axis) and a 3-dimensional figure. Such graph is in general some surface.
- It is impossible to (directly) depict graphs of functions of more than 2 variables since we cannot make 4 or more dimensional figures.

Example 6. *How does the graph of $f(x, y) = x^2 - y^2$ look like?*



Partial derivative – introduction

Given the function $f(x, y) = x^2 + xy + y^2$.

- If we fix the value of the variable y to 3, we obtain a univariate function $f(x) = x^2 + 3x + 9$ having its derivative equal to $2x + 3$.
- We can fix the value of y not only to a specific number: **we just treat y as a constant**. Then we get univariate function $f(x) = x^2 + xy + y^2$ and its derivative is $2x + y$.
- This derivative is called **partial derivative of (x, y) with respect to** x and denoted by
$$\frac{\partial f}{\partial x}(x, y) = 2x + y\,.$$
- In the same way we define the partial derivative of $f(x, y)$ with respect to y:
$$\frac{\partial f}{\partial y}(x, y) = x + 2y\,.$$
- In general $\frac{\partial f}{\partial x}(x, y)$ and $\frac{\partial f}{\partial y}(x, y)$ are two-variate functions.

Partial derivative – definition

The derivative of a (single variate) function $f(x)$ is the following limit (if it exists):
$$f'(x) = \lim_{\delta \to 0} \frac{f(x + \delta) - f(x)}{\delta}.$$

Partial derivatives are defined similarly:

Definition 7. *The partial derivative of* $f(x_1, x_2, \ldots, x_n)$ *with respect to* x_i *at the point* $(x_1, x_2, \ldots, x_n)$ *is defined by (if the limit below exists)*

$$\frac{\partial f}{\partial x_i}(x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) =$$
$$= \lim_{\delta \to 0} \frac{f(x_1, x_2, \ldots, x_{i-1}, x_i + \delta, x_{i+1}, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{\delta}.$$

Since the definition is similar, even the geometrical meaning is analogous.

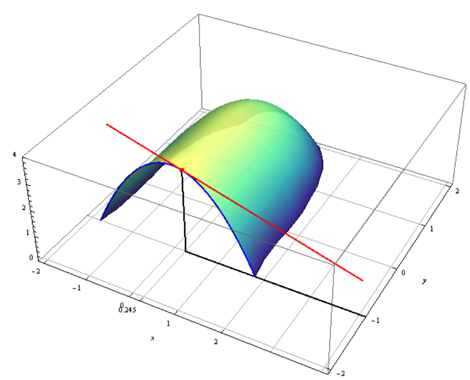
Partial derivative – definition

The partial derivatives of $f(x, y)$ can be in short denoted by

$$f_x(x, y) = \frac{\partial f}{\partial x}(x, y) \quad \text{and} \quad f_y(x, y) = \frac{\partial f}{\partial y}(x, y).$$

The number $f_x(x, y)$ for given values of x and y is again the slope of a tangent line, but a surface has infinitely many tangent lines in all possible directions at any point, so which one is this one?

It is the only tangent line which is parallel to the x-axis.



Second partial derivatives

Definition 8. *For a function* $f(x_1, x_2, \ldots, x_n)$ *we define second partial derivatives*

$$f_{x_j x_i}(x_1, x_2, \ldots, x_n) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x_1, x_2, \ldots, x_n) = \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}(x_1, x_2, \ldots, x_n)\right),$$

in particular, for $i = j$ we have

$$f_{x_i x_i}(x_1, x_2, \ldots, x_n) = \frac{\partial^2 f}{\partial x_i^2}(x_1, x_2, \ldots, x_n) = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}(x_1, x_2, \ldots, x_n)\right).$$

Partial derivatives – exercises

Example 9. *Find partial derivatives with respect to all variables*

(a) $f(x, y) = xy + e^x \cos y$,

(b) $f(x, y) = x^2y^3 + x^3y^4 - e^{xy^2}$,

(c) $f(x, y, z) = \sin(xy/z)$.

Example 10. *Find all second partial derivatives of the functions*

(a) $f(x, y) = x^2 + xy^2 + 3x^3y$,

(b) $f(x, y, z) = e^{xz} + y \cos x$,

(c) $f(x, y, z) = z \cos(xy) + x \sin(yz)$.

Equality of mixed partial derivatives

The fact that the mixed partial derivatives are equal is not a coincidence:

Theorem 11. *If a function $f(x, y)$ has continuous second partial derivatives, then the mixed second derivatives are equal, i.e.,*

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}.$$

This theorem is not true in general, a counterexample is the function

$$f(x, y) = \begin{cases} 0 & \text{at point } (0, 0) \\ \dfrac{xy(x^2 - y^2)}{x^2 + y^2} & \text{otherwise.} \end{cases}$$