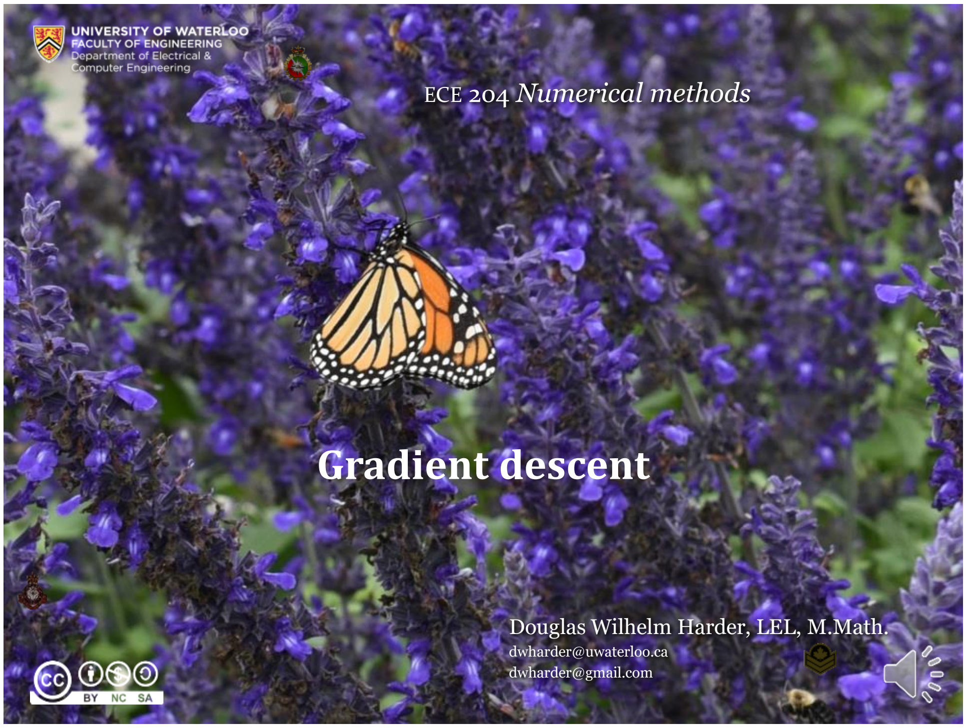


1

Introduction

- In this topic, we will
  - Describe why we might not want to use one of the two previous algorithms
  - Discuss the idea of the gradient
  - See how the gradient can indicate which direction to search
  - Explain how we can convert an *n*-dimensional minimization problem into a one-dimensional minimization problem
  - Discuss issues with estimating the gradient

2

The problem

- We have seen algorithms to approximate the minimum of a real-valued function of a real variable
- We have also seen how Newton's method can be used if we can calculate the gradient and the Hessian
- If we cannot calculate the Hessian, we could revert to the Hooke-Jeeves method
- What if, however, the function is sufficiently differentiable
  - Can we develop a better approach?

3

3

The gradient

- Given a sufficiently differentiable real-valued function of a vector variable,
  the gradient is defined as

$$\vec{\nabla} f(\mathbf{u}) = \begin{pmatrix} \frac{\partial}{\partial u_1} f(\mathbf{u}) \\ \frac{\partial}{\partial u_2} f(\mathbf{u}) \\ \vdots \\ \frac{\partial}{\partial u_n} f(\mathbf{u}) \end{pmatrix}$$

- We can normalize this gradient vector by dividing by its 2-norm,
  and denote a normalized vector by a hat:

$$\hat{\vec{\nabla}} f(\mathbf{u}) \stackrel{\text{def}}{=} \frac{\vec{\nabla} f(\mathbf{u})}{\left\|\vec{\nabla} f(\mathbf{u})\right\|_2}$$

4

4

Gradient descent

The gradient

- Evaluating the gradient at a point $\mathbf{u}_k$ gives us the direction of maximum increase
  - Thus for sufficiently small $\varepsilon > 0$,

$$f\left(\mathbf{u}_k + \varepsilon \vec{\nabla} f\left(\mathbf{u}_k\right)\right) \geq f\left(\mathbf{u}_k + \varepsilon \hat{\mathbf{u}}\right)$$

- Similarly, for a sufficiently differentiable function, the opposite direction gives the direction of maximum decrease:

$$f\left(\mathbf{u}_k - \varepsilon \vec{\nabla} f\left(\mathbf{u}_k\right)\right) \leq f\left(\mathbf{u}_k - \varepsilon \hat{\mathbf{u}}\right)$$

5

5

Gradient descent

The gradient

- Mathematically, we say this as follows:
  - The slope in the direction of the gradient is greater than or equal to the slope in any other direction

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f\left(\mathbf{u}_k + \alpha \vec{\nabla} f\left(\mathbf{u}_k\right)\right) \geq \frac{\mathrm{d}}{\mathrm{d}\alpha} f\left(\mathbf{u}_k + \alpha \hat{\mathbf{u}}\right)$$

- Also, $\frac{\mathrm{d}}{\mathrm{d}\alpha} f\left(\mathbf{u}_k + \alpha \vec{\nabla} f\left(\mathbf{u}_k\right)\right) = \left\|\vec{\nabla} f\left(\mathbf{u}_k\right)\right\|_2$

6

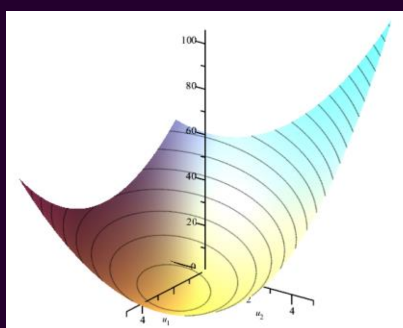
6

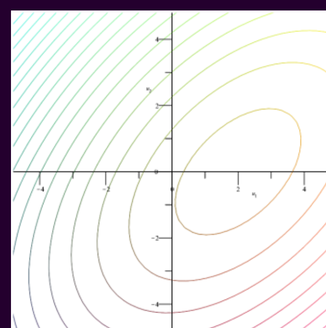
The gradient

- For example, consider the bivariate function:

$$f\begin{pmatrix} x \\ y \end{pmatrix} \overset{\text{def}}{=} x^2 - xy + y^2 - 3x + y + 1$$

  – This function has a unique minimum at the point $\mathbf{u} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$





7

7

The gradient

- For this function, $f\begin{pmatrix} x \\ y \end{pmatrix} \overset{\text{def}}{=} x^2 - xy + y^2 - 3x + y + 1$

$$\vec{\nabla} f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x - y - 3 \\ 2y - x + 1 \end{pmatrix}$$

  – At the point $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, we have

$$\vec{\nabla} f\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 - 2 - 3 \\ 4 - 1 + 1 \end{pmatrix} = \begin{pmatrix} -3 \\ 4 \end{pmatrix}$$

$$\vec{\nabla} f\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -0.6 \\ 0.8 \end{pmatrix}$$

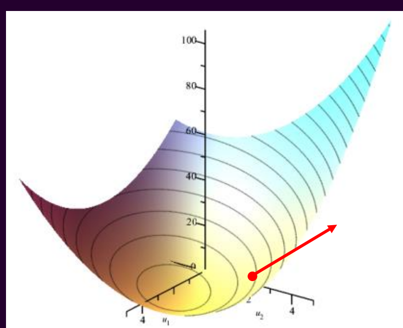
8

8

The gradient

- Viewing this visually, we have

$$\vec{\nabla} f\begin{pmatrix}1\\2\end{pmatrix}=\begin{pmatrix}2-2-3\\4-1+1\end{pmatrix}=\begin{pmatrix}-3\\4\end{pmatrix}$$



9

9

The gradient

- Now, if $\mathbf{u}_k+\alpha\vec{\nabla} f\left(\mathbf{u}_k\right)$ for $\alpha>0$ moves in the direction of steepest ascent,

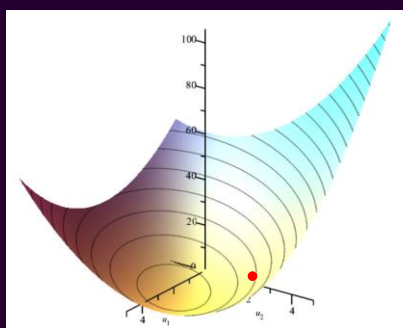
  then $\mathbf{u}_k-\alpha\vec{\nabla} f\left(\mathbf{u}_k\right)$ for $\alpha>0$ moves in the direction of steepest descent

10

10

The gradient

- The opposite direction is the direction of steepest descent, but it does not point directly at the minimum
  - It does, however, move us in the direction of that minimum



A real-valued function of a real variable

- Notice now that $\mathbf{u}_k - \alpha\vec{\nabla} f(\mathbf{u}_k)$ has one real variable α
- Thus, $f\left(\mathbf{u}_k - \alpha\vec{\nabla} f(\mathbf{u}_k)\right)$ is a real-value function of a real variable
- In our example, we had $\begin{pmatrix}1\\2\end{pmatrix} - \alpha\begin{pmatrix}-0.6\\0.8\end{pmatrix} = \begin{pmatrix}1+0.6\alpha\\2-0.8\alpha\end{pmatrix}$
- Substituting this into the function, we have:

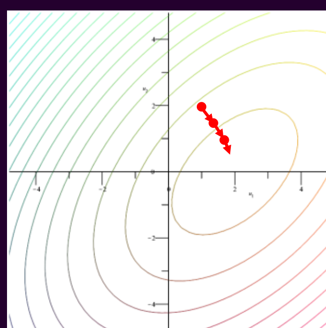
$$f\left(\mathbf{u} + \alpha\vec{\nabla} f(\mathbf{u})\right) = f\begin{pmatrix}1+0.6\alpha\\2-0.8\alpha\end{pmatrix}$$

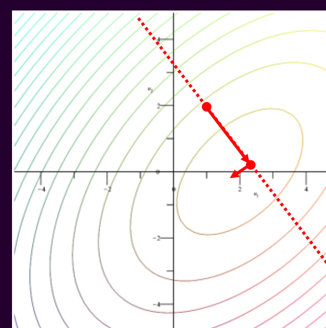
$$= (1+0.6\alpha)^2 - (1+0.6\alpha)(2-0.8\alpha)$$

$$+ (2-0.8\alpha)^2 - 3(1+0.6\alpha) + (2-0.8\alpha) + 1$$

Possible strategies

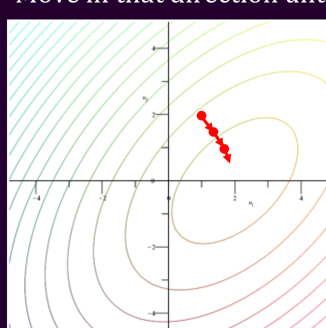
- Once we find the gradient at $\mathbf{u}_k$, we have one of two strategies:
  - Move one step in that direction and try again
  - Find a local minimum in that direction and only then try again

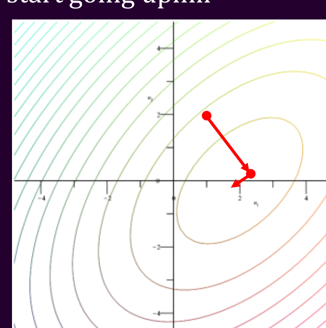




Possible strategies

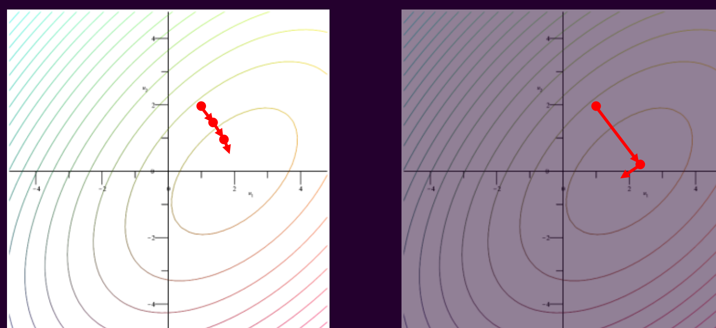
- Analogies are as follows: determine the direction of steepest descent and:
  - Take one step in that direction
  - Move in that direction until you start going uphill





Possible strategies

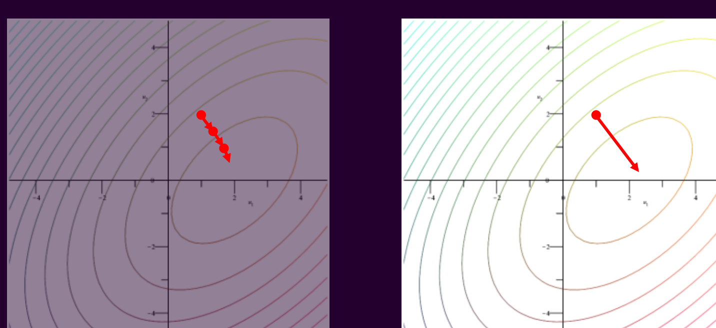
- Taking one step and then recalculating the gradient may require significant computational effort at each step
  - Also, how do we pick the *optimal* step size?



Gradient descent

Possible strategies

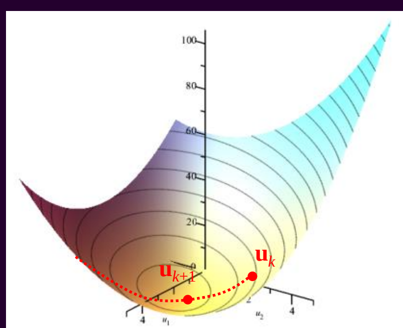
- The second strategy, searching in the direction of the gradient, we calculate the gradient once, but then use a solver for minimizing a real-valued function of a real variable

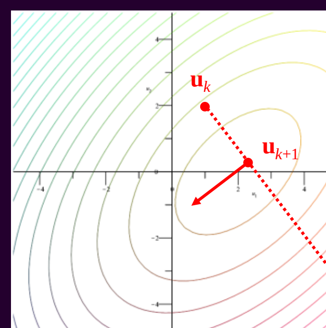


Minimizing in the direction of the gradient

- Suppose we adopt this second strategy
  - Given $\mathbf{u}_k$, we calculate the gradient and find $\mathbf{u}_{k+1}$
  - If we calculate the gradient at $\mathbf{u}_{k+1}$, you will find that

$$\vec{\nabla} f(\mathbf{u}_k) \perp \vec{\nabla} f(\mathbf{u}_{k+1})$$



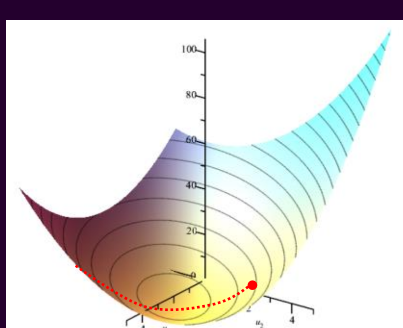


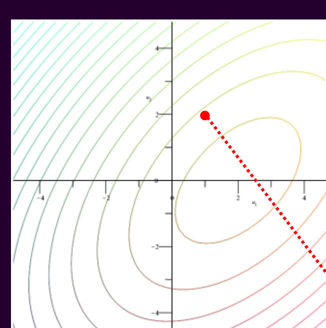
17

17

Minimizing in the direction of the gradient

- How do we find a minimum in this direction?
  - Previously, we assumed we had an idea as to where the minimum was
  - Strategies vary, but we will focus on one additional assumption



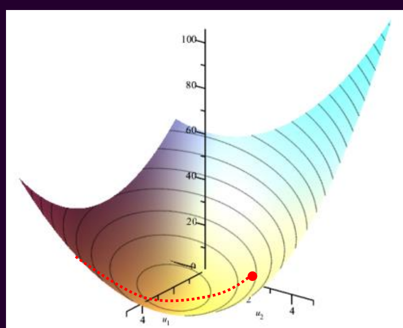


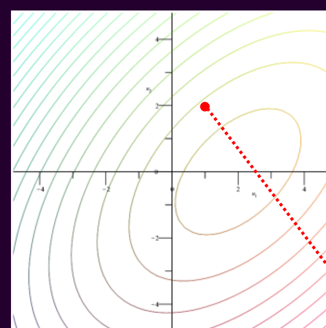
18

18

Minimizing in the direction of the gradient

- We will assume that:
  - The function has a unique minimum and is concave up
  - The function, in any direction, ultimately goes to infinity
- If both these conditions are satisfied,
  both these conditions are also satisfied on any line





19

19

Minimizing in the direction of the gradient

- Now, begin calculating

$$\phi = \frac{1+\sqrt{5}}{2} \approx 1.618$$

$$f\left(\mathbf{u}_k - \phi^m \vec{\nabla} f\left(\mathbf{u}_k\right)\right)$$

  starting with $m = 0$
- Begin incrementing or decrementing m until you find three points such that

$$f\left(\mathbf{u}_k - \phi^M \vec{\nabla} f\left(\mathbf{u}_k\right)\right) < f\left(\mathbf{u}_k - \phi^{M-1} \vec{\nabla} f\left(\mathbf{u}_k\right)\right), f\left(\mathbf{u}_k - \phi^{M+1} \vec{\nabla} f\left(\mathbf{u}_k\right)\right)$$

- In this case, you then continue with the golden-ratio search with

$$f\left(\mathbf{u}_k - \alpha \vec{\nabla} f\left(\mathbf{u}_k\right)\right)$$

  starting with $\phi^{M-1} \le \alpha \le \phi^{M+1}$ and continuing with the Brent-Dekker method

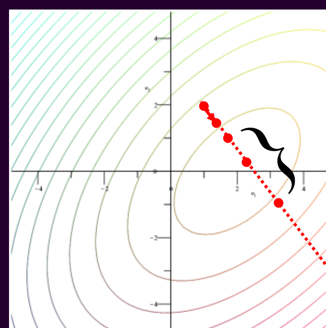
20

20

Gradient descent

Minimizing in the direction of the gradient

- In our example, we'd proceed as follows
  - Thus, start searching between $\phi \leq \alpha \leq \phi^3$



21

21

Gradient descent

Minimizing in the direction of the gradient

- Some caveats:
  - Should you always start with $m = 0$?
    - No, after the first iteration,
      you should probably start with the previous M you found
  - As you approach the minimum, it may happen that a range of powers may be equal (and very close to the minimum value):

$$f\left(\mathbf{u}_k - \phi^{M_1-1}\vec{\nabla} f\left(\mathbf{u}_k\right)\right) > f\left(\mathbf{u}_k - \phi^{M_1}\vec{\nabla} f\left(\mathbf{u}_k\right)\right) = \cdots = f\left(\mathbf{u}_k - \phi^{M_2}\vec{\nabla} f\left(\mathbf{u}_k\right)\right) < f\left(\mathbf{u}_k - \phi^{M_2+1}\vec{\nabla} f\left(\mathbf{u}_k\right)\right)$$

22

22

Calculating or estimating the gradient

- In general:
  - If automatic differentiation is being used to calculate the gradient, automatic differentiation can be used to calculate the Hessian, so you should consider using Newton's method
  - If the function *f* is not sufficiently differentiable to calculate the Hessian, but you can still calculate the gradient, use this technique
  - If automatic differentiation is not available, we can still estimate the gradient
    - This is possible because of the properties of extrema if the solution is differentiable near the extremum
    - An approximation of the gradient will still move us in the direction of the minimum

Calculating or estimating the gradient

- To approximate the gradient:
  - Recall that $\mathbf{e}_k$ is the k^{th} canonical unit vector:
    - All entries are zero except for the k^{th} entry, which is one
  - In this case,

$$\left(\vec{\nabla} f(\mathbf{u})\right)_k \approx \frac{f(\mathbf{u}+h\mathbf{e}_k) - f(\mathbf{u}-h\mathbf{e}_k)}{2h}$$

Calculating or estimating the gradient

- Recall that for $f\begin{pmatrix}x\\y\end{pmatrix} \stackrel{\text{def}}{=} x^2 - xy + y^2 - 3x + y + 1$,

  we had $\vec{\nabla} f\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}2x - y - 3\\2y - x + 1\end{pmatrix}$ and $\vec{\nabla} f\begin{pmatrix}1\\2\end{pmatrix} = \begin{pmatrix}2 - 2 - 3\\4 - 1 + 1\end{pmatrix} = \begin{pmatrix}-3\\4\end{pmatrix}$

- Letting $h = 0.1$, we can approximate the two entries:

$$\left(\vec{\nabla} f\begin{pmatrix}x\\y\end{pmatrix}\right)_1 \approx \frac{f\begin{pmatrix}1.01\\2\end{pmatrix} - f\begin{pmatrix}0.99\\2\end{pmatrix}}{0.02} = -3 \qquad \left(\vec{\nabla} f\begin{pmatrix}x\\y\end{pmatrix}\right)_2 \approx \frac{f\begin{pmatrix}1\\2.01\end{pmatrix} - f\begin{pmatrix}1\\1.99\end{pmatrix}}{0.02} = 4$$

25

25

Calculating or estimating the gradient

- Is it safe to use an approximation of the gradient?
  - Recall that at a minimum, there will be an entire region where the truncated floating-point values will be equal
  - There will also be a much larger region where the value of the function is close to the minimum value

- Thus, we are still likely to find a good approximation of the minimum value, even if we don't have that ideal an approximation of exactly where that minimum is

26

26

Summary

- Following this topic, you now
  - Understand the idea of gradient descent
    - If the gradient points in the direction of maximum increase, the opposite direction points in the direction of maximum decrease
  - Understand that you should move in that direction until you find a local minimum
  - Are aware that this reduces the number of times we must actually calculate the gradient—an expensive operation
  - Know that you can approximate the gradient by using finite difference formulas and that these are likely sufficiently accurate to help find that minimum

27

27

References

[1] https://en.wikipedia.org/wiki/Gradient_descent

28

28

Acknowledgments

None so far.

Colophon

These slides were prepared using the Cambria typeface. Mathematical equations use Times New Roman, and source code is presented using `Consolas`. Mathematical equations are prepared in MathType by Design Science, Inc.

Examples may be formulated and checked using Maple by Maplesoft, Inc.

The photographs of flowers and a monarch butter appearing on the title slide and accenting the top of each other slide were taken at the Royal Botanical Gardens in October of 2017 by Douglas Wilhelm Harder. Please see

https://www.rbg.ca/

for more information.

Disclaimer

These slides are provided for the ECE 204 *Numerical methods* course taught at the University of Waterloo. The material in it reflects the author's best judgment in light of the information available to them at the time of preparation. Any reliance on these course slides by any party for any other purpose are the responsibility of such parties. The authors accept no responsibility for damages, if any, suffered by any party as a result of decisions made or actions based on these course slides for any other purpose than that for which it was intended.