The power of differentiation

Jonathon Hare

Vision, Learning and Control
University of Southampton

Topics

- The big idea: optimisation by following gradients
- Recap: what are gradients and how do we find them?
- Recap: Singular Value Decomposition and its applications
- Example: Computing SVD using gradients - The Netflix Challenge

The big idea: optimisation by following gradients

- Fundamentally, we're interested in machines that we train by optimising parameters

The big idea: optimisation by following gradients

- Fundamentally, we're interested in machines that we train by optimising parameters
  - How do we select those parameters?

The big idea: optimisation by following gradients

- Fundamentally, we're interested in machines that we train by optimising parameters
  - How do we select those parameters?
- In deep learning/differentiable programming we typically define an objective function that we *minimise* (or *maximise*) with respect to those parameters

The big idea: optimisation by following gradients

- Fundamentally, we're interested in machines that we train by optimising parameters
  - How do we select those parameters?
- In deep learning/differentiable programming we typically define an objective function that we *minimise* (or *maximise*) with respect to those parameters
- This implies that we're looking for points at which the gradient of the objective function is zero w.r.t the parameters

The big idea: optimisation by following gradients

A simple 1D example

The big idea: optimisation by following gradients
A simple 2D example

The big idea: optimisation by following gradients

A more indicative example

Recap: what are gradients and how do we find them?

The derivative of $y = x^2$ from first principles

$$y = x^2$$

Recap: what are gradients and how do we find them?

The derivative of $y = x^2$ from first principles

$$
\begin{aligned}
y &= x^2 \\
\frac{dy}{dx} &= \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h}
\end{aligned}
$$

Recap: what are gradients and how do we find them?

The derivative of $y = x^2$ from first principles

$$y = x^2$$

$$\frac{dy}{dx} = \lim_{h\to 0} \frac{(x+h)^2 - x^2}{h}$$

$$\frac{dy}{dx} = \lim_{h\to 0} \frac{x^2 + h^2 + 2hx - x^2}{h}$$

Recap: what are gradients and how do we find them?

The derivative of $y = x^2$ from first principles

$$
\begin{aligned}
y &= x^2 \\
\frac{dy}{dx} &= \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} \\
\frac{dy}{dx} &= \lim_{h \to 0} \frac{x^2 + h^2 + 2hx - x^2}{h} \\
\frac{dy}{dx} &= \lim_{h \to 0} \frac{h^2 + 2hx}{h}
\end{aligned}
$$

Recap: what are gradients and how do we find them?

The derivative of $y = x^2$ from first principles

$$
\begin{aligned}
y &= x^2 \\
\frac{dy}{dx} &= \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} \\
\frac{dy}{dx} &= \lim_{h \to 0} \frac{x^2 + h^2 + 2hx - x^2}{h} \\
\frac{dy}{dx} &= \lim_{h \to 0} \frac{h^2 + 2hx}{h} \\
\frac{dy}{dx} &= \lim_{h \to 0} h + 2x)
\end{aligned}
$$

Recap: what are gradients and how do we find them?

The derivative of $y = x^2$ from first principles

$$y = x^2$$

$$\frac{dy}{dx} = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h}$$

$$\frac{dy}{dx} = \lim_{h \to 0} \frac{x^2 + h^2 + 2hx - x^2}{h}$$

$$\frac{dy}{dx} = \lim_{h \to 0} \frac{h^2 + 2hx}{h}$$

$$\frac{dy}{dx} = \lim_{h \to 0} h + 2x)$$

$$\frac{dy}{dx} = 2x$$

- For numerical computation of gradients it is better to use a "centralised" definition of the gradient:
  - $f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a-h)}{2h}$

Recap: what are gradients and how do we find them?

Aside: numerical approximation of the derivative

- For numerical computation of gradients it is better to use a "centralised" definition of the gradient:
  - $f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a-h)}{2h}$
  - This is known as the *symmetric difference quotient*

- For numerical computation of gradients it is better to use a "centralised" definition of the gradient:
  - $f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a-h)}{2h}$
  - This is known as the *symmetric difference quotient*
  - For small values of h this has less error than the standard one-sided difference quotient.

- For numerical computation of gradients it is better to use a "centralised" definition of the gradient:
  - $f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a-h)}{2h}$
  - This is known as the *symmetric difference quotient*
  - For small values of h this has less error than the standard one-sided difference quotient.
- If you are going to use this to estimate gradients you need to be aware of potential rounding errors due to floating point representations.
  - Calculating gradients this way using less than 64-bit precision is rarely going to be useful. (Numbers are not represented exactly, so even if h is represented exactly, $x + h$ will probably not be)
  - You need to pick an appropriate h - too small and the subtraction will have a large rounding error!

- Deep learning is all about optimising deeper functions; functions that are compositions of other functions
    - e.g. $z = f \circ g(x) = f(g(x))$

Recap: what are gradients and how do we find them?

Derivatives of deeper functions

- Deep learning is all about optimising deeper functions; functions that are compositions of other functions
  - e.g. $z = f \circ g(x) = f(g(x))$
- The chain rule of calculus tells us how to differentiate compositions of functions:
  - $\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{x}$

Recap: what are gradients and how do we find them?

Example: differentiating $z = x^4$

Note that this is a silly example that just serves to demonstrate the principle!

$$z = x^4$$

Recap: what are gradients and how do we find them?

Example: differentiating $z = x^4$

Note that this is a silly example that just serves to demonstrate the principle!

$$z = x^4$$

$$z = (x^2)^2 = y^2 \quad \text{where} \quad \mathrm{y} = \mathrm{x}^2$$

Note that this is a silly example that just serves to demonstrate the principle!

$$z = x^4$$

$$z = (x^2)^2 = y^2 \quad \text{where} \quad \mathrm{y = x^2}$$

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} = (2y)(2x) = (2x^2)(2x) = 4x^3$$

Note that this is a silly example that just serves to demonstrate the principle!

$$
\begin{aligned}
z &= x^4 \\
z &= (x^2)^2 = y^2 \quad \text{where} \quad \mathrm{y} = \mathrm{x}^2 \\
\frac{dz}{dx} &= \frac{dz}{dy}\frac{dy}{dx} = (2y)(2x) = (2x^2)(2x) = 4x^3
\end{aligned}
$$

Equivalently, from first principles:

$$
\begin{aligned}
z &= x^4 \\
\frac{dz}{dx} &= \lim_{h \to 0} \frac{(x+h)^4 - x^4}{h} \\
\frac{dz}{dx} &= \lim_{h \to 0} \frac{h^4 + 4h^3x + 6h^2x^2 + 4hx^3 + x^4 - x^4}{h} \\
\frac{dz}{dx} &= \lim_{h \to 0} h^3 + 4h^2x + 6hx^2 + 4x^3 = 4x^3
\end{aligned}
$$

Recap: what are gradients and how do we find them?

Functions of vectors and matrices: partial differentiation

Recap: Singular Value Decomposition and its applications

Example: Computing SVD using gradients - The Netflix Challenge