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

- Gradient based optimisation is a *big* field!
- With deep learning we're primarily interested in first-order methods.

The big idea: optimisation by following gradients

- Gradient based optimisation is a *big* field!
  - First order methods, second order methods, subgradient methods...
- With deep learning we're primarily interested in first-order methods[1].

[1] Second order gradient optimisers are potentially better, but for systems with many variables are currently impractical as they require computing the Hessian.

The big idea: optimisation by following gradients

- Gradient based optimisation is a *big* field!
  - First order methods, second order methods, subgradient methods...
- With deep learning we're primarily interested in first-order methods[1].
  - Primarily using variants of gradient descent: a function $F(\boldsymbol{x})$ has *a* minima[2] at a point $\boldsymbol{x} = \boldsymbol{a}$ where $\boldsymbol{a}$ is given by applying $\boldsymbol{a}_{n+1} = \boldsymbol{a} - \alpha \nabla F(\boldsymbol{a}_n)$ until convergence.

[1] Second order gradient optimisers are potentially better, but for systems with many variables are currently impractical as they require computing the Hessian.

[2] not necessarily global or unique

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

The derivative in 1D

- Recall that the gradient of a straight line is $\frac{\Delta x}{\Delta y}$.

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

The derivative in 1D

- Recall that the gradient of a straight line is $\frac{\Delta x}{\Delta y}$.
- For an arbitrary real-valued function, $f(a)$, we can approximate the derivative, $f'(a)$ using the gradient of the *secant line* defined by $(a, f(a))$ and a point a small distance, h, away $(a+h, f(a+h))$: $f'(a) \approx \frac{f(a+h)-f(a)}{h}$.

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

The derivative in 1D

- Recall that the gradient of a straight line is $\frac{\Delta x}{\Delta y}$.
- For an arbitrary real-valued function, $f(a)$, we can approximate the derivative, $f'(a)$ using the gradient of the *secant line* defined by $(a, f(a))$ and a point a small distance, h, away $(a+h, f(a+h))$: $f'(a) \approx \frac{f(a+h)-f(a)}{h}$.
  - This expression is 'Newton's Difference Quotient'.

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

The derivative in 1D

- Recall that the gradient of a straight line is $\frac{\Delta x}{\Delta y}$.
- For an arbitrary real-valued function, $f(a)$, we can approximate the derivative, $f'(a)$ using the gradient of the *secant line* defined by $(a, f(a))$ and a point a small distance, h, away $(a+h, f(a+h))$: $f'(a) \approx \frac{f(a+h)-f(a)}{h}$.
    - This expression is 'Newton's Difference Quotient'.
    - As h becomes smaller, the approximated derivative becomes more accurate.

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

The derivative in 1D

- Recall that the gradient of a straight line is $\frac{\Delta x}{\Delta y}$.
- For an arbitrary real-valued function, $f(a)$, we can approximate the derivative, $f'(a)$ using the gradient of the *secant line* defined by $(a, f(a))$ and a point a small distance, h, away $(a+h, f(a+h))$: $f'(a) \approx \frac{f(a+h)-f(a)}{h}$.
    - This expression is 'Newton's Difference Quotient'.
    - As h becomes smaller, the approximated derivative becomes more accurate.
    - If we take the limit as $h \to 0$, then we have an exact expression for the derivative: $\frac{df}{da} = f'(a) = \lim_{h \to 0} \frac{f(a+h)-f(a)}{h}$.

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

$$y = x^2$$

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

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

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

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

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

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

$$y = x^2$$

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

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

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

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

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

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

Aside: numerical approximation of the derivative

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

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

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

Aside: numerical approximation of the derivative

- For numerical computation of derivatives it is better to use a "centralised" definition of the derivative:
  - $f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a-h)}{2h}$
  - The bit inside the limit 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 derivatives it is better to use a "centralised" definition of the derivative:
  - $f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a-h)}{2h}$
  - The bit inside the limit 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 derivatives you need to be aware of potential rounding errors due to floating point representations.
  - Calculating derivatives 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!

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. $z = f \circ g(x) = f(g(x))$

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:
  - $\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$

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!

$$z = x^4$$

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

$$z = x^4$$
$$z = (x^2)^2 = y^2 \quad \text{where} \quad y = x^2$$

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!

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

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!

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

Equivalently, from first principles:

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

- What if we're dealing with a *vector* function, $\boldsymbol{y}(t)$?

- What if we're dealing with a *vector* function, $\boldsymbol{y}(t)$?
  - This can be split into its constituent coordinate functions: $\boldsymbol{y}(t) = (y_1(t), \ldots, y_n(t))$.

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

Vector functions

- What if we're dealing with a *vector* function, $\boldsymbol{y}(t)$?
  - This can be split into its constituent coordinate functions: $\boldsymbol{y}(t) = (y_1(t), \ldots, y_n(t))$.
  - Thus the derivative is a vector (the 'tangent vector'), $\boldsymbol{y}'(t) = (y_1'(t), \ldots, y_n'(t))$, which consists of the derivatives of the coordinate functions.

- What if we're dealing with a *vector* function, $\boldsymbol{y}(t)$?
  - This can be split into its constituent coordinate functions: $\boldsymbol{y}(t) = (y_1(t), \dots, y_n(t))$.
  - Thus the derivative is a vector (the 'tangent vector'), $\boldsymbol{y}'(t) = (y_1'(t), \dots, y_n'(t))$, which consists of the derivatives of the coordinate functions.
  - Equivalently, $\boldsymbol{y}'(t) = \lim_{h\to 0} \frac{\boldsymbol{y}(t+h)-\boldsymbol{y}(t)}{h}$ if the limit exists.

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

Functions of multiple variables: partial differentiation

- What if the function we're trying to deal with has multiple variables[3] (e.g. $f(x, y) = x^2 + xy + y^2$)?
  - This expression has a pair of *partial derivatives*, $\frac{\partial f}{\partial x} = 2x + y$ and $\frac{\partial f}{\partial y} = x + 2y$, computed by differentiating with respect to each variable x and y whilst holding the other(s) constant.

[3] A multivariate function

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

Functions of multiple variables: partial differentiation

- What if the function we're trying to deal with has multiple variables[3] (e.g. $f(x, y) = x^2 + xy + y^2$)?
  - This expression has a pair of *partial derivatives*, $\frac{\partial f}{\partial x} = 2x + y$ and $\frac{\partial f}{\partial y} = x + 2y$, computed by differentiating with respect to each variable x and y whilst holding the other(s) constant.
- In general, the partial derivative of a function $f(x_1, \ldots, x_n)$ at a point $(a_1, \ldots, a_n)$ is given by:
  $\frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) = \lim_{h \to 0} \frac{f(a_1 \ldots, a_i + h, \ldots, a_n) - f(a_1 \ldots, a_i, \ldots, a_n)}{h}$.

[3] A multivariate function

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

Functions of multiple variables: partial differentiation

- What if the function we're trying to deal with has multiple variables[3] (e.g. $f(x, y) = x^2 + xy + y^2$)?
  - This expression has a pair of *partial derivatives*, $\frac{\partial f}{\partial x} = 2x + y$ and $\frac{\partial f}{\partial y} = x + 2y$, computed by differentiating with respect to each variable x and y whilst holding the other(s) constant.
- In general, the partial derivative of a function $f(x_1, \ldots, x_n)$ at a point $(a_1, \ldots, a_n)$ is given by:
  $\frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) = \lim_{h \to 0} \frac{f(a_1 \ldots, a_i + h, \ldots, a_n) - f(a_1 \ldots, a_i, \ldots, a_n)}{h}$.
- The vector of partial derivatives of a scalar-value multivariate function, $f((x_1, \ldots, x_n)$ at a point $(a_1, \ldots, a_n)$, can be arranged into a vector: $\nabla f(a_1, \ldots, a_n) = (\frac{\partial f}{\partial x_1}(a_1, \ldots, a_n), \ldots, \frac{\partial f}{\partial x_n}(a_1, \ldots, a_n))$.

[3] A multivariate function

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

Functions of multiple variables: partial differentiation

- What if the function we're trying to deal with has multiple variables[3] (e.g. $f(x, y) = x^2 + xy + y^2$)?
  - This expression has a pair of *partial derivatives*, $\frac{\partial f}{\partial x} = 2x + y$ and $\frac{\partial f}{\partial y} = x + 2y$, computed by differentiating with respect to each variable x and y whilst holding the other(s) constant.
- In general, the partial derivative of a function $f(x_1, \ldots, x_n)$ at a point $(a_1, \ldots, a_n)$ is given by: $\frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) = \lim_{h \to 0} \frac{f(a_1 \ldots, a_i + h, \ldots, a_n) - f(a_1 \ldots, a_i, \ldots, a_n)}{h}$.
- The vector of partial derivatives of a scalar-value multivariate function, $f((x_1, \ldots, x_n)$ at a point $(a_1, \ldots, a_n)$, can be arranged into a vector: $\nabla f(a_1, \ldots, a_n) = (\frac{\partial f}{\partial x_1}(a_1, \ldots, a_n), \ldots, \frac{\partial f}{\partial x_n}(a_1, \ldots, a_n))$.
  - This is the **gradient** of f at a.

[3] A multivariate function

- What if the function we're trying to deal with has multiple variables[3] (e.g. $f(x, y) = x^2 + xy + y^2$)?
  - This expression has a pair of *partial derivatives*, $\frac{\partial f}{\partial x} = 2x + y$ and $\frac{\partial f}{\partial y} = x + 2y$, computed by differentiating with respect to each variable x and y whilst holding the other(s) constant.
- In general, the partial derivative of a function $f(x_1, \ldots, x_n)$ at a point $(a_1, \ldots, a_n)$ is given by:
  $\frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) = \lim_{h \to 0} \frac{f(a_1 \ldots, a_i + h, \ldots, a_n) - f(a_1 \ldots, a_i, \ldots, a_n)}{h}$.
- The vector of partial derivatives of a scalar-value multivariate function, $f((x_1, \ldots, x_n)$ at a point $(a_1, \ldots, a_n)$, can be arranged into a vector: $\nabla f(a_1, \ldots, a_n) = (\frac{\partial f}{\partial x_1}(a_1, \ldots, a_n), \ldots, \frac{\partial f}{\partial x_n}(a_1, \ldots, a_n))$.
  - This is the **gradient** of f at a.
- In the case of a vector-valued multivariate function, the partial derivatives form a matrix called the **Jacobian**.

[3] A multivariate function

- For the kinds of functions (and programs) that we'll look at *optimising* in this course have a number of typical properties:

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

Functions of vectors and matrices: partial differentiation

- For the kinds of functions (and programs) that we'll look at *optimising* in this course have a number of typical properties:
  - They are scalar-valued
    - We'll look at programs with *multiple losses*, but ultimately we can just consider optimising with respect to the *sum* of the losses.

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

Functions of vectors and matrices: partial differentiation

- For the kinds of functions (and programs) that we'll look at *optimising* in this course have a number of typical properties:
  - They are scalar-valued
    - We'll look at programs with *multiple losses*, but ultimately we can just consider optimising with respect to the *sum* of the losses.
  - They involve multiple variables, which are often wrapped up in the form of vectors or matrices, and more generally tensors.

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

Functions of vectors and matrices: partial differentiation

- For the kinds of functions (and programs) that we'll look at *optimising* in this course have a number of typical properties:
  - They are scalar-valued
    - We'll look at programs with *multiple losses*, but ultimately we can just consider optimising with respect to the *sum* of the losses.
  - They involve multiple variables, which are often wrapped up in the form of vectors or matrices, and more generally tensors.
  - **How will we find the gradients of these?**

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

Functions of tensors

Recap: Singular Value Decomposition and its applications

Example: Computing SVD using gradients - The Netflix Challenge