Advanced Machine Learning

Differential Calculus

$$f'(x_0) = \left. \frac{\mathrm{d}f(x)}{\mathrm{d}x} \right|_{x=x_0} = \lim_{\epsilon \to 0} \frac{f(x_0+\epsilon) - f(x_0)}{\epsilon}$$

$f(x_0 + \epsilon) - f(x_0)$

ϵ

x_0
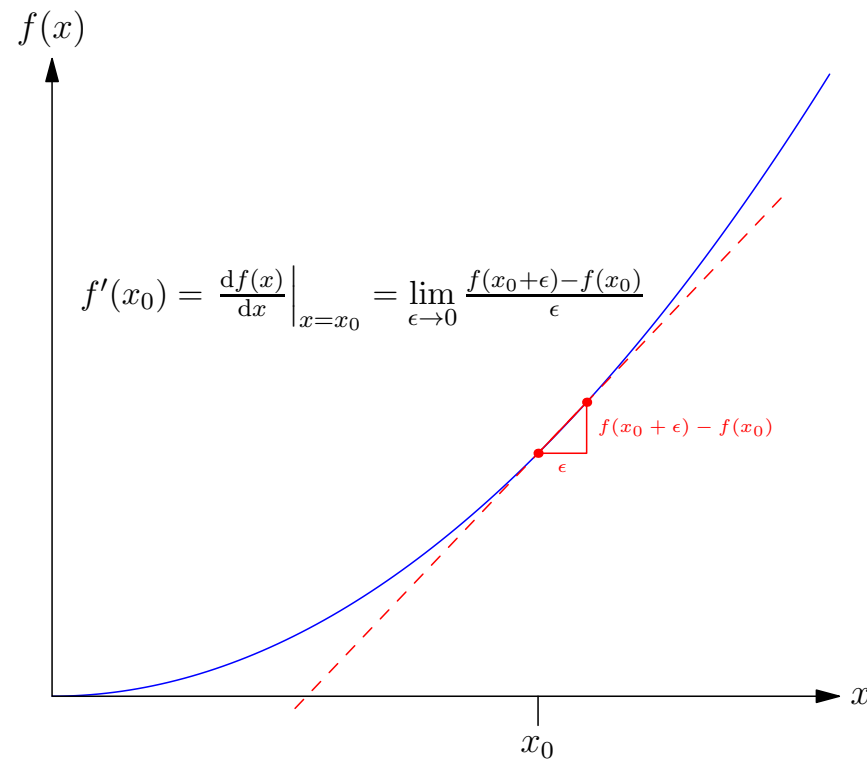
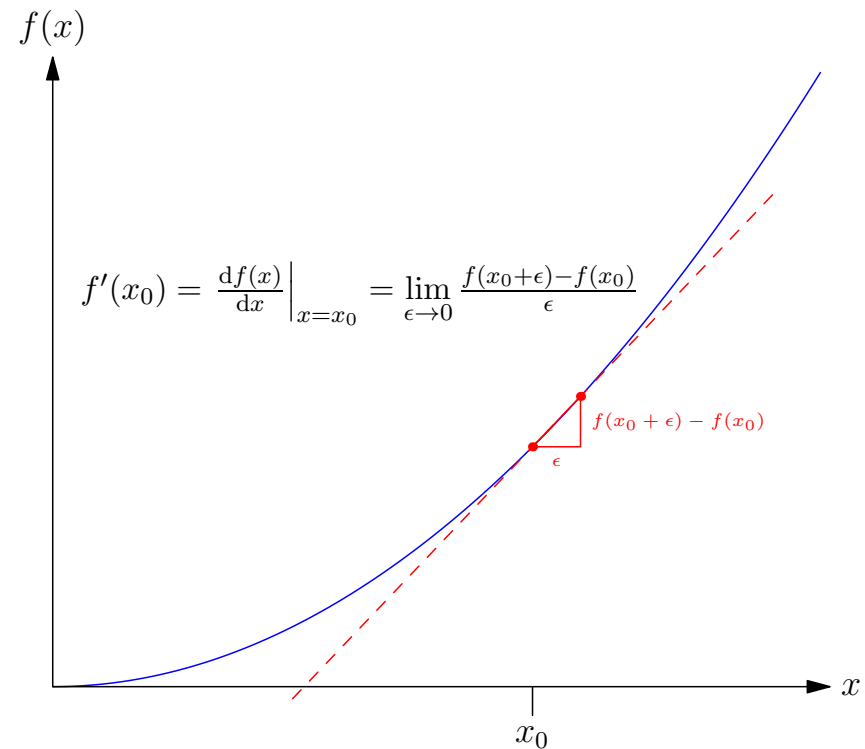*Differentiation, product and chain rules, vectors and matrices*

Outline

1. **Why Calculus?**

2. Differentiation

3. Vector and Matrix Calculus

$f(x)$

$$f'(x_0) = \left.\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|_{x=x_0} = \lim_{\epsilon \to 0} \frac{f(x_0+\epsilon)-f(x_0)}{\epsilon}$$

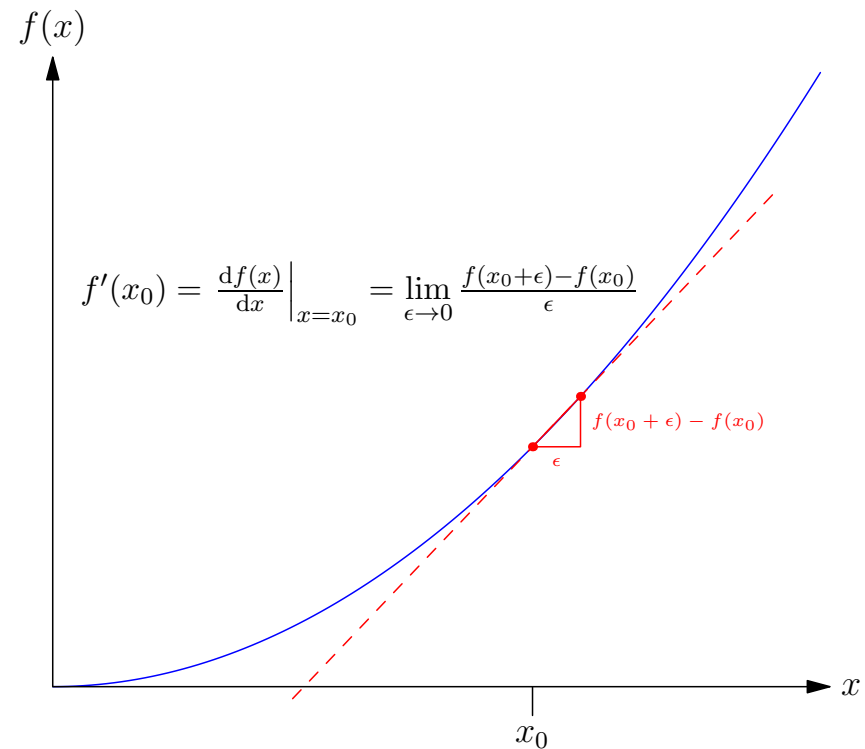$f(x_0 + \epsilon) - f(x_0)$

ϵ

x

x_0

Why Calculus?

- Calculus is a fundamental tool of mathematical analysis

- In machine learning differentiation is fundamental tool in optimisation

- Integration is an essential tool in taking expectations over continuous distributions

- Both differentiation and integration crop up elsewhere

- This material will not be examined explicitly, but I assume elsewhere that you can do calculus
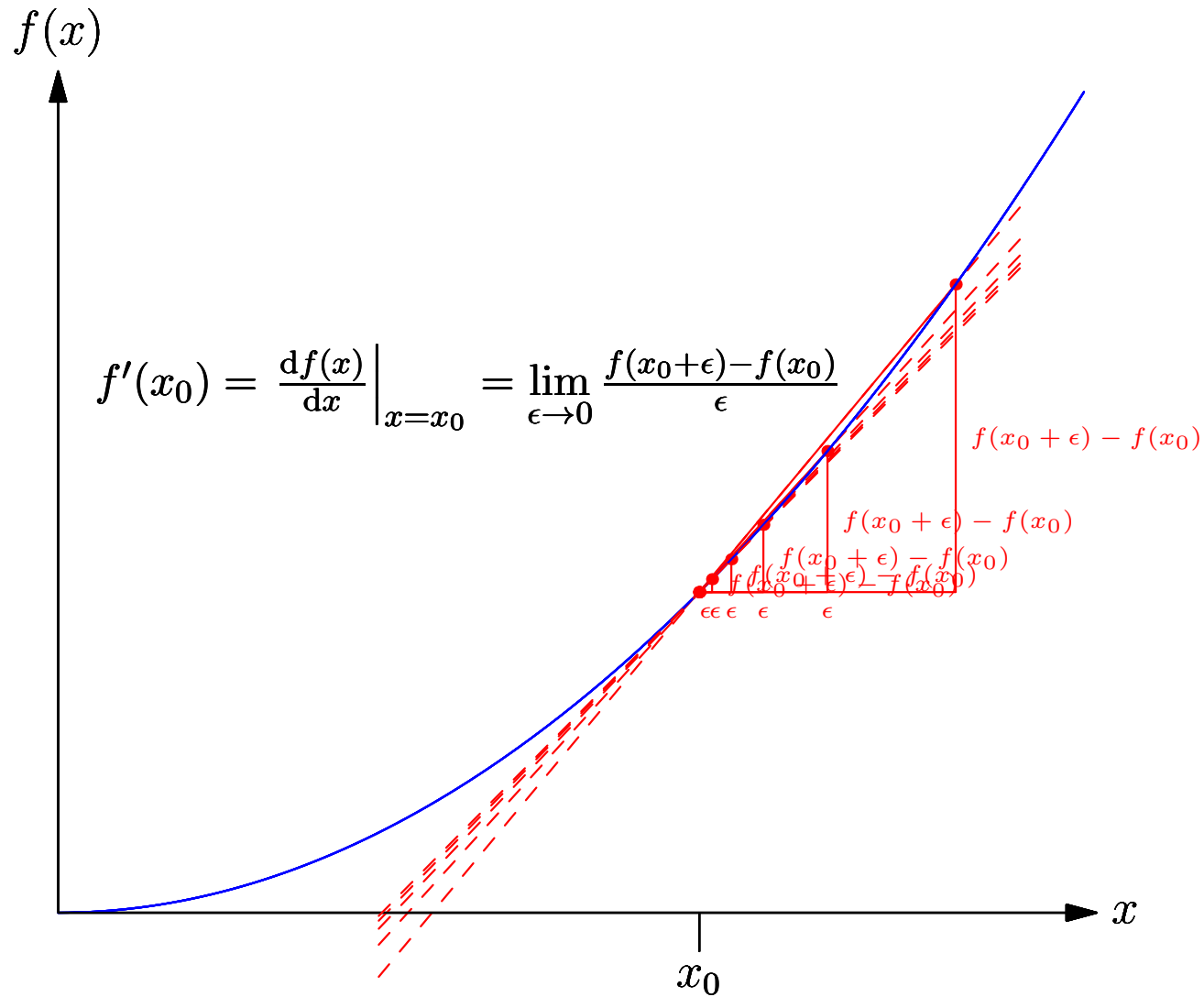
Back to Basics

- You have all done A-level maths so should be familiar with the rules of calculus

- But, it is easy to forget the rules and sometimes we use quite sophisticated tricks

- Although the sophisticated tricks really speed up calculations, it pays to be able to understand where these tricks come from

Outline

1. Why Calculus?

2. **Differentiation**

3. Vector and Matrix Calculus

$$f(x)$$

$$f'(x_0) = \left.\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|_{x=x_0} = \lim_{\epsilon \to 0} \frac{f(x_0+\epsilon)-f(x_0)}{\epsilon}$$

$$f(x_0+\epsilon) - f(x_0)$$

$$\epsilon$$

$$x_0$$

$$x$$

Differentiation

$$f'(x_0) = \frac{\mathrm{d}f(x)}{\mathrm{d}x}\bigg|_{x=x_0} = \lim_{\epsilon \to 0} \frac{f(x_0+\epsilon)-f(x_0)}{\epsilon}$$

$f(x)$

$f(x_0+\epsilon) - f(x_0)$

$f(x_0+\epsilon) - f(x_0)$

$f(x_0+\epsilon) - f(x_0)$

$f(x_0+\epsilon) - f(x_0)$

ϵ ϵ ϵ ϵ ϵ

x

x_0

Polynomials

- $f(x) = x^2$

$$\frac{\mathrm{d}x^2}{\mathrm{d}x} = \lim_{\epsilon \to 0} \frac{(x+\epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \to 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon}$$

$$= \lim_{\epsilon \to 0} 2x + \epsilon = 2x$$

- $(x+\epsilon)^n = (x+\epsilon)(x+\epsilon)\cdots(x+\epsilon) = x^n + n\epsilon x^{n-1} + O(\epsilon^2)$

$$\frac{\mathrm{d}x^n}{\mathrm{d}x} = \lim_{\epsilon \to 0} \frac{(x+\epsilon)^n - x^n}{\epsilon} = \lim_{\epsilon \to 0} nx^{n-1} + O(\epsilon) = nx^{n-1}$$

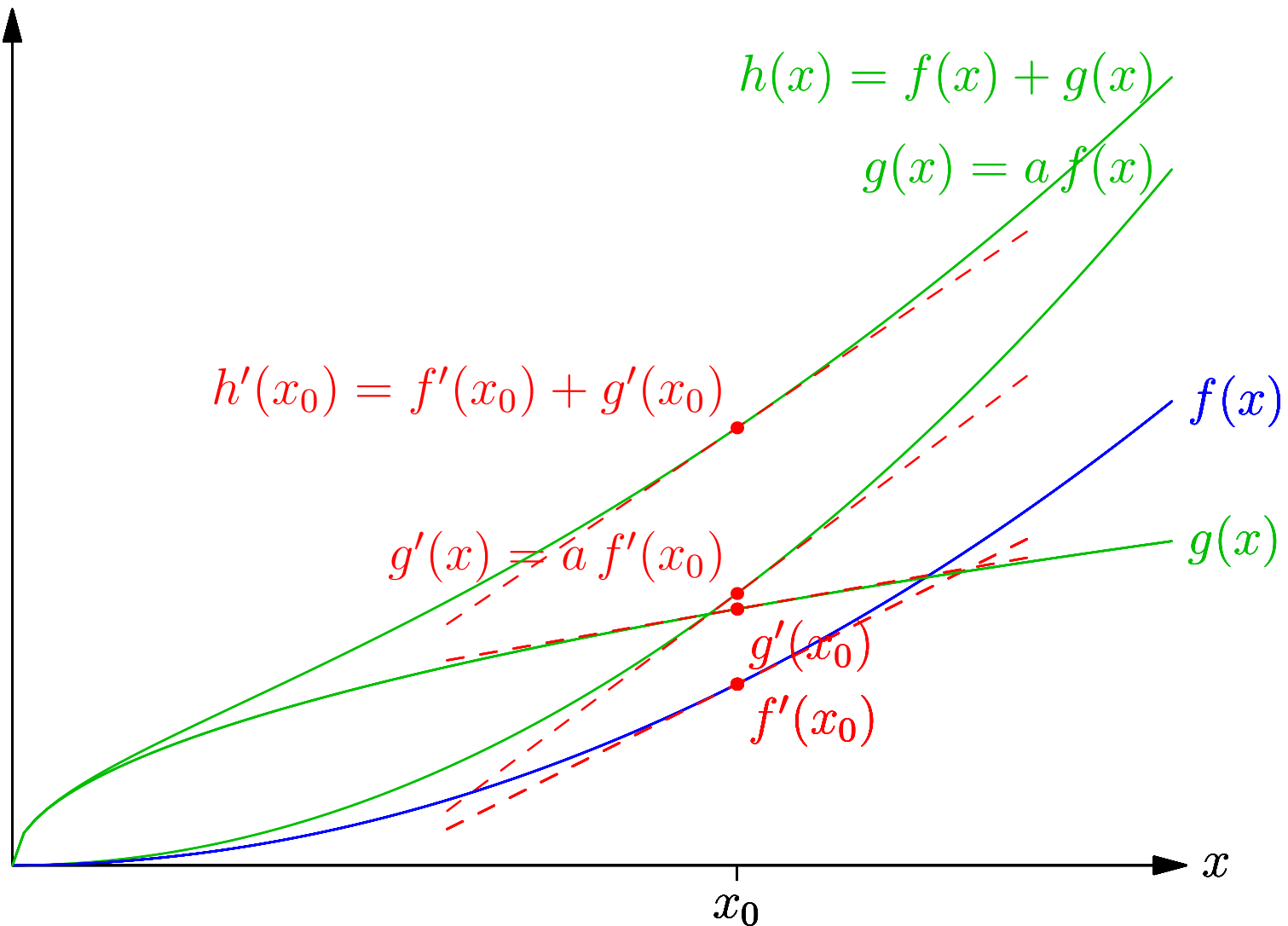# Linearity of derivatives

- Note that $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$ (from the definition of $f'(x)$)

$$\frac{\mathrm{d}(a f(x) + b g(x))}{\mathrm{d}x} = \lim_{\epsilon \to 0} \frac{(a f(x + \epsilon) + b g(x + \epsilon)) - (a f(x) + b g(x))}{\epsilon}$$

$$= \lim_{\epsilon \to 0} \frac{a\epsilon f'(x) + b\epsilon g'(x) + O(\epsilon^2)}{\epsilon}$$

$$= a f'(x) + b g'(x)$$

- **Differentiation is a linear operation!**

Linearity in Pictures

$h(x) = f(x) + g(x)$

$g(x) = a\, f(x)$

$h'(x_0) = f'(x_0) + g'(x_0)$

$f(x)$

$g'(x) = a\, f'(x_0)$

$g(x)$

$g'(x_0)$

$f'(x_0)$

x

x_0

Product Rule

- Recall $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$

- If $h(x) = f(x)g(x)$

$$\begin{aligned}
h'(x) &= \lim_{\epsilon \to 0} \frac{f(x+\epsilon)g(x+\epsilon) - f(x)g(x)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\left(f(x) + \epsilon f'(x) + O(\epsilon^2)\right)\left(g(x) + \epsilon g'(x) + O(\epsilon^2)\right) - f(x)g(x)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\epsilon(f'(x)g(x) + f(x)g'(x)) + O(\epsilon^2)}{\epsilon} = f'(x)g(x) + f(x)g'(x)
\end{aligned}$$

- This is the **product rule**

Chain Rule

- Recall $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$

- Let $h(x) = f(g(x))$

- Then

$$h(x + \epsilon) = f(g(x + \epsilon)) = f\left(g(x) + \epsilon g'(x) + O(\epsilon^2)\right)$$
$$= f(g(x)) + \epsilon g'(x) f'(g(x)) + O(\epsilon^2)$$

- Thus

$$h'(x) = \lim_{\epsilon \to 0} \frac{h(x + \epsilon) - h(x)}{\epsilon} = g'(x) f'(g(x))$$

- This is the famous **chain rule**. Together with the product rule it means you can differentiate almost everything

More on chain rules

- We can also write the chain rule as

$$\frac{\mathrm{d}f(g(x))}{\mathrm{d}x} = \frac{\mathrm{d}f(g)}{\mathrm{d}g}\frac{\mathrm{d}g(x)}{\mathrm{d}x}$$

- Sometimes this is neater or easier to remember

$$\frac{\mathrm{d}\mathrm{e}^{\cos(x^2)}}{\mathrm{d}x} = \frac{\mathrm{d}\mathrm{e}^{\cos(x^2)}}{\mathrm{d}\cos(x^2)}\frac{\mathrm{d}\cos(x^2)}{\mathrm{d}x^2}\frac{\mathrm{d}x^2}{\mathrm{d}x}$$

$$= \mathrm{e}^{\cos(x^2)}\left(-\sin(x^2)\right)2x$$

$$= -2x\sin(x^2)\mathrm{e}^{\cos(x^2)}$$

Inverse functions

- Suppose $g(y) = f^{-1}(y)$ is the inverse of $f(x)$ in the sense that $g(f(x)) = f^{-1}(f(x)) = x$

- Using the chain rule

$$\frac{\mathrm{d}g(f(x))}{\mathrm{d}x} = f'(x)g'(f(x)) = 1$$

since $g(f(x)) = x$

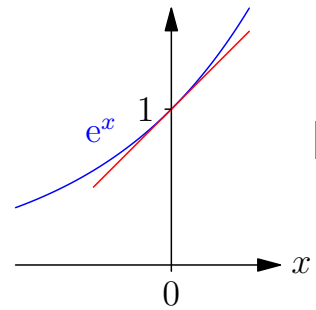- So $g'(f(x)) = 1/f'(x)$

- Writing $y = f(x)$ so that $x = f^{-1}(y) = g(y)$ we find $g'(y) = 1/f'(g(y))$ that is

$$\frac{\mathrm{d}g(y)}{\mathrm{d}y} = \frac{1}{f'(g(y))} \qquad \frac{\mathrm{d}f^{-1}(y)}{\mathrm{d}y} = \frac{1}{f'(f^{-1}(y))}$$

Exponentials

- Note that $a^{b+c} = a^b a^c$ (that is we multiply a together $b+c$ times)

- Now $\mathrm{e}^\epsilon \approx (1 + \epsilon)$

- But $\mathrm{e}^{x+\epsilon} = \mathrm{e}^x \mathrm{e}^\epsilon = \mathrm{e}^x(1 + \epsilon + O(\epsilon^2)) = \mathrm{e}^x + \epsilon \mathrm{e}^x + O(\epsilon^2)$

$$\frac{\mathrm{d}\mathrm{e}^x}{\mathrm{d}x} = \lim_{\epsilon \to 0} \frac{\mathrm{e}^{x+\epsilon} - \mathrm{e}^x}{\epsilon} = \lim_{\epsilon \to 0} \frac{\epsilon \mathrm{e}^x + O(\epsilon^2)}{\epsilon} = \mathrm{e}^x$$

Functions of Exponentials

- What about $f(x) = \mathrm{e}^{cx}$

$$\frac{\mathrm{d}\mathrm{e}^{cx}}{\mathrm{d}x} = \frac{\mathrm{d}\mathrm{e}^{cx}}{\mathrm{d}cx}\frac{\mathrm{d}cx}{\mathrm{d}x} = c\mathrm{e}^{cx}$$

- More generally using the chain rule

$$\frac{\mathrm{d}\mathrm{e}^{g(x)}}{\mathrm{d}x} = g'(x)\mathrm{e}^{g(x)}$$

- Also $a^{bc} = (a^b)^c$ (that is we multiply a together $b \times c$ times)

$$\frac{\mathrm{d}a^x}{\mathrm{d}x} = \frac{\mathrm{d}(\mathrm{e}^{\ln(a)})^x}{\mathrm{d}x} = \frac{\mathrm{d}\mathrm{e}^{\ln(a)x}}{\mathrm{d}x} = \ln(a)\mathrm{e}^{\ln(a)x} = \ln(a)a^x$$

Natural Logarithms

- The natural logarithm is defined as the inverse of e^x

$$\ln(\mathrm{e}^x) = x \qquad\qquad \mathrm{e}^{\ln(y)} = y$$

- Recall that if $g(y) = f^{-1}(y)$ then $g'(y) = 1/f'(g(y))$

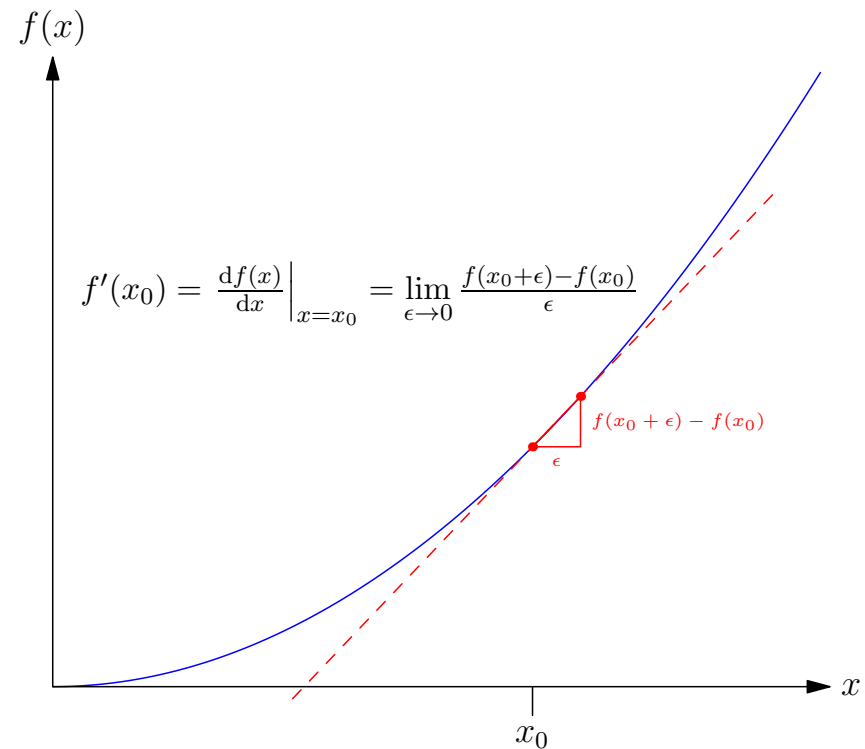- Consider $g(y) = \ln(y)$ and $f(x) = \mathrm{e}^x$ (with $f'(x) = \mathrm{e}^x$)

$$\frac{\mathrm{d}\ln(y)}{\mathrm{d}y} = \frac{1}{\mathrm{e}^{\ln(y)}} = \frac{1}{y}$$

Exponentials and Logarithms

$f(x) = 10^x$

$f(x) = \mathrm{e}^x$

$y = x$

$f(x) = 2^x$

$f^{-1}(x) = \log_{10}(x)$

$f^{-1}(x) = \ln(x)$

$f^{-1}(x) = \log_2(x)$

Outline

1. Why Calculus?

2. Differentiation

3. **Vector and Matrix Calculus**

$$f'(x_0) = \left.\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|_{x=x_0} = \lim_{\epsilon \to 0} \frac{f(x_0+\epsilon)-f(x_0)}{\epsilon}$$

$f(x_0 + \epsilon) - f(x_0)$

ϵ

x_0

$f(x)$

x
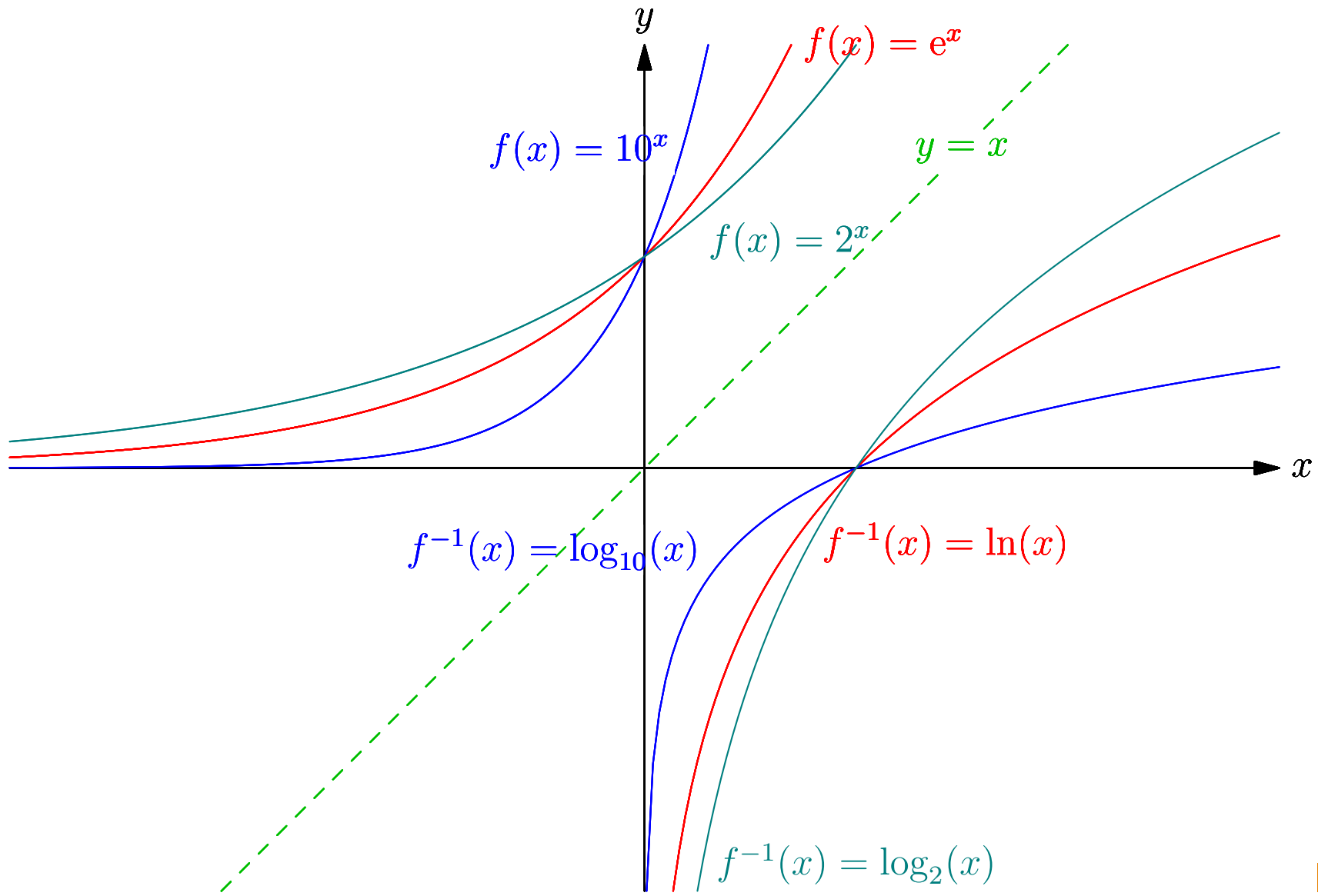
Derivatives in High Dimensions

- When working with functions $f : \mathbb{R}^n \to \mathbb{R}$ in many dimensions then there will typically be different derivative in different directions▮

- To compute the derivative in a direction $\boldsymbol{u} \in \mathbb{R}^n$ (where $\|\boldsymbol{u}\| = 1$) at a point $\boldsymbol{x} \in \mathbb{R}^n$ we use

$$\partial_{\boldsymbol{u}} F(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{f(\boldsymbol{x} + \epsilon \boldsymbol{u}) - f(\boldsymbol{x})}{\epsilon}▮$$

- If $\boldsymbol{u} = \boldsymbol{\delta}_i = (0,\ldots,0,1,0,\ldots,0)$ (i.e. $u_i = 1$) then

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} = \lim_{\epsilon \to 0} \frac{f(\boldsymbol{x} + \epsilon \boldsymbol{\delta}_i) - f(\boldsymbol{x})}{\epsilon}▮$$

Taylor

- If we expand $f(\boldsymbol{x} + \epsilon\boldsymbol{u})$ to first order in ϵ

$$f(\boldsymbol{x} + \epsilon\boldsymbol{u}) = f(\boldsymbol{x}) + \epsilon\boldsymbol{u}^\top \boldsymbol{g}(\boldsymbol{x}) + O(\epsilon^2)$$

then $g_i(\boldsymbol{x}) = \dfrac{\partial f(\boldsymbol{x})}{\partial x_i}$

- Recall we defined the vector of first order derivatives of $f(\boldsymbol{x})$ to be the gradient

$$\boldsymbol{\nabla} f(\boldsymbol{x}) = \begin{pmatrix} \dfrac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \dfrac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \dfrac{\partial f(\boldsymbol{x})}{\partial x_n} \end{pmatrix}$$

- Thus

$$f(\boldsymbol{x} + \epsilon\boldsymbol{u}) = f(\boldsymbol{x}) + \epsilon\boldsymbol{u}^\top \boldsymbol{\nabla} f(\boldsymbol{x}) + O(\epsilon^2)$$

This is the start of the high-dimensional Taylor expansion

Computing Gradients 1

- We can compute the gradient by writing out $f(\boldsymbol{x})$ componentwise and performing the partial derivative with respect to x_i

$$\boldsymbol{\nabla}\boldsymbol{w}^\top\mathbf{M}\boldsymbol{w} = \begin{pmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \frac{\partial}{\partial w_3} \\ \vdots \end{pmatrix} \sum_{i,j} w_i M_{ij} w_j = \begin{pmatrix} \sum_j M_{1j}w_j + \sum_i w_i M_{i1} \\ \sum_j M_{2j}w_j + \sum_i w_i M_{i2} \\ \sum_j M_{3j}w_j + \sum_i w_i M_{i3} \\ \vdots \end{pmatrix}$$

$$= \mathbf{M}\boldsymbol{w} + \mathbf{M}^\top\boldsymbol{w}$$

- It is tedious to compute these things component-wise, but when you need to understand what is going on then go back to the basics

---

Computing Gradients 2

- A slicker way is just to expand $f(x + \epsilon u)$

- Consider $f(x) = x^\top M x + a^\top x$

$$f(x + \epsilon u) = (x + \epsilon u)^\top M (x + \epsilon u) + a^\top (x + \epsilon u)$$
$$= f(x) + \epsilon \left(u^\top M x + x^\top M u + a^\top u \right) + O(\epsilon^2)$$
$$= f(x) + \epsilon u^\top \left(M x + M^\top x + a \right) + O(\epsilon^2)$$

using $x^\top M u = u^\top M^\top x$ and $a^\top u = u^\top a$

- But $f(x + \epsilon u) = f(x) + \epsilon u^\top \nabla f(x) + O(\epsilon^2)$ so

$$\nabla f(x) = M x + M^\top x + a$$

Differentiating Matrices

- Often we have loss functions with respect to a matrix $\boldsymbol{W}$, e.g.

$$L(\boldsymbol{W}) = (\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{b} - c)^2$$

- We might want to find the minimum with respect to $\boldsymbol{W}$

- This occurs at a point $\boldsymbol{W}^*$ where $L(\boldsymbol{W})$ does not increase as we change $\boldsymbol{W}$ in any way

- That is, we seek a $\boldsymbol{W}^*$ such that, for any matrices $\boldsymbol{\mathsf{U}}$

$$L(\boldsymbol{W}^* + \epsilon \boldsymbol{\mathsf{U}}) - L(\boldsymbol{W}^*) = O(\epsilon^2)$$

Generalised Gradient

- We can generalise the idea of gradient to matrices

$$\frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{W}} = \begin{pmatrix} \frac{\partial L(\boldsymbol{W})}{\partial W_{11}} & \frac{\partial L(\boldsymbol{W})}{\partial W_{12}} & \cdots & \frac{\partial L(\boldsymbol{W})}{\partial W_{1m}} \\ \frac{\partial L(\boldsymbol{W})}{\partial W_{21}} & \frac{\partial L(\boldsymbol{W})}{\partial W_{22}} & \cdots & \frac{\partial L(\boldsymbol{W})}{\partial W_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L(\boldsymbol{W})}{\partial W_{n1}} & \frac{\partial L(\boldsymbol{W})}{\partial W_{n2}} & \cdots & \frac{\partial L(\boldsymbol{W})}{\partial W_{nm}} \end{pmatrix} \blacksquare$$

- From an identical argument we used for vectors

$$L(\boldsymbol{W} + \epsilon \mathbf{U}) = L(\boldsymbol{W}) + \epsilon \operatorname{tr} \mathbf{U}^{\mathsf{T}} \frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{W}} + O(\epsilon^2) \blacksquare$$

where

$$\operatorname{tr} \mathbf{U}^{\mathsf{T}} \mathbf{G} = \sum_i \left[\mathbf{U}^{\mathsf{T}} \mathbf{G} \right]_{ii} = \sum_{ij} U_{ji} G_{ji} = \sum_{ij} U_{ij} G_{ij} = \langle \mathbf{U}, \mathbf{G} \rangle \blacksquare$$

Example

- Suppose

$$L(\boldsymbol{W}) = \left(\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{b} - c\right)^2$$

  then

$$L(\boldsymbol{W} + \epsilon \boldsymbol{U}) = \left(\boldsymbol{a}^\top (\boldsymbol{W} + \epsilon \boldsymbol{U})\boldsymbol{b} - c\right)^2 = \left(\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{b} + \epsilon \boldsymbol{a}^\top \boldsymbol{U} \boldsymbol{b} - c\right)^2$$

$$= L(\boldsymbol{W}) + 2\epsilon \left(\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{b} - c\right)\left(\boldsymbol{a}^\top \boldsymbol{U} \boldsymbol{b}\right) + O(\epsilon^2)$$

- Now

$$\boldsymbol{a}^\top \boldsymbol{U} \boldsymbol{b} = \sum_{ij} a_i U_{ij} b_j = \sum_{ij} U_{ji} a_j b_i = \operatorname{tr} \boldsymbol{U}^\top \boldsymbol{a} \boldsymbol{b}^\top$$

Thus $\frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{W}} = 2\left(\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{b} - c\right) \boldsymbol{a} \boldsymbol{b}^\top$

Traces

- The trace of a matrix is the sum of its diagonal elements

$$\mathrm{tr}\mathbf{A} = \mathrm{tr}\mathbf{A}^\top = \sum_i A_{ii}$$

- Clearly $\mathrm{tr}c\mathbf{A} = c\,\mathrm{tr}\mathbf{A}$

- Also $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}\mathbf{A} + \mathrm{tr}\mathbf{B}$

- We note that

$$\mathrm{tr}\mathbf{A}\mathbf{B} = \sum_{i,j} A_{ij}B_{ji} = \sum_{i,j} B_{ij}A_{ji} = \mathrm{tr}\mathbf{B}\mathbf{A}$$

- It follows that

$$\mathrm{tr}\mathbf{A}\mathbf{B}\mathbf{C}\mathbf{D} = \mathrm{tr}\mathbf{D}\mathbf{A}\mathbf{B}\mathbf{C} = \mathrm{tr}\mathbf{C}\mathbf{D}\mathbf{A}\mathbf{B} = \mathrm{tr}\mathbf{B}\mathbf{C}\mathbf{D}\mathbf{A}$$

Quick Matrix Differentiation

- Let

$$\partial_{\mathbf{u}} f(\mathbf{X}) = \lim_{\epsilon \to 0} \frac{f(\mathbf{X} + \epsilon \mathbf{U}) - f(\mathbf{X})}{\epsilon} = \operatorname{tr} \mathbf{U}^\top \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$$

- E.g.

$$\partial_{\mathbf{u}} \operatorname{tr} \mathbf{A}\mathbf{X}\mathbf{B} = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \operatorname{tr} \mathbf{A} \left(\mathbf{X} + \epsilon \mathbf{U} \right) \mathbf{B} - \operatorname{tr} \mathbf{A}\mathbf{X}\mathbf{B}$$

$$= \operatorname{tr} \mathbf{A}\mathbf{U}\mathbf{B} = \operatorname{tr} \mathbf{B}^\top \mathbf{U}^\top \mathbf{A}^\top = \operatorname{tr} \mathbf{U}^\top \mathbf{A}^\top \mathbf{B}^\top$$

thus

$$\frac{\partial \operatorname{tr} \mathbf{A}\mathbf{X}\mathbf{B}}{\partial \mathbf{X}} = \mathbf{A}^\top \mathbf{B}^\top$$

Log Determinants

- We often come across logarithms of determinants of matrices, $\log(|\mathbf{M}|)$ ▮

- For GP we want to choose $\mathbf{K}$ to maximise the marginal likelihood, $\log\big(|\mathbf{K} + \sigma^2\mathbf{I}|\big)$ ▮

- To find the derivative of $\log(|\mathbf{X}|)$ we consider

$$\log(|\mathbf{X} + \epsilon\mathbf{U}|) = \log\big(|\mathbf{X}(\mathbf{I} + \epsilon\mathbf{X}^{-1}\mathbf{U})|\big) \text{▮}$$
$$= \log\big(|\mathbf{X}||\mathbf{I} + \epsilon\mathbf{X}^{-1}\mathbf{U}|\big) \text{▮}$$
$$= \log(|\mathbf{X}|) + \log\big(|\mathbf{I} + \epsilon\mathbf{X}^{-1}\mathbf{U}|\big) \text{▮}$$

⋆ Using $|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|$ ▮
⋆ Using $\log(ab) = \log(a) + \log(b)$ ▮

Determinants

$$|\mathbf{I} + \epsilon\mathbf{M}| = \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} \\ \epsilon M_{21} & 1 + \epsilon M_{22} \end{vmatrix} = (1 + \epsilon M_{11})(1 + \epsilon M_{22}) - \epsilon^2 M_{21} M_{12}$$

$$= 1 + \epsilon(M_{11} + M_{22}) + O(\epsilon^2)$$

Putting it Together

- Recall

$$\log(|\mathbf{X} + \epsilon\mathbf{U}|) - \log(|\mathbf{X}|) = \log\big(|\mathbf{I} + \epsilon\mathbf{X}^{-1}\mathbf{U}|\big)$$
$$= \log\big(1 + \epsilon\,\mathrm{tr}\,\mathbf{X}^{-1}\mathbf{U} + O(\epsilon)^2\big)$$
$$= \epsilon\,\mathrm{tr}\,\mathbf{X}^{-1}\mathbf{U} + O(\epsilon)^2$$
$$= \epsilon\,\mathrm{tr}\,\mathbf{U}^{\mathsf{T}}\big(\mathbf{X}^{-1}\big)^{\mathsf{T}} + O(\epsilon)$$

using $\log(1 + x) = x + \frac{x^2}{2} + \cdots$

- Thus $\partial_{\mathbf{u}}\log(|\mathbf{X}|) = \mathrm{tr}\,\mathbf{U}^{\mathsf{T}}\big(\mathbf{X}^{-1}\big)^{\mathsf{T}}$

- Or

$$\frac{\partial\log(|\mathbf{X}|)}{\partial\mathbf{X}} = \big(\mathbf{X}^{-1}\big)^{\mathsf{T}}$$

Summary

- With care you can differentiate most expressions

- The chain and product rule are incredibly powerful tools

- We can generalise differentiation to vectors and matrices

- There are a number of surprisingly useful results: see **The Matrix Cookbook**

- Next stop: integration