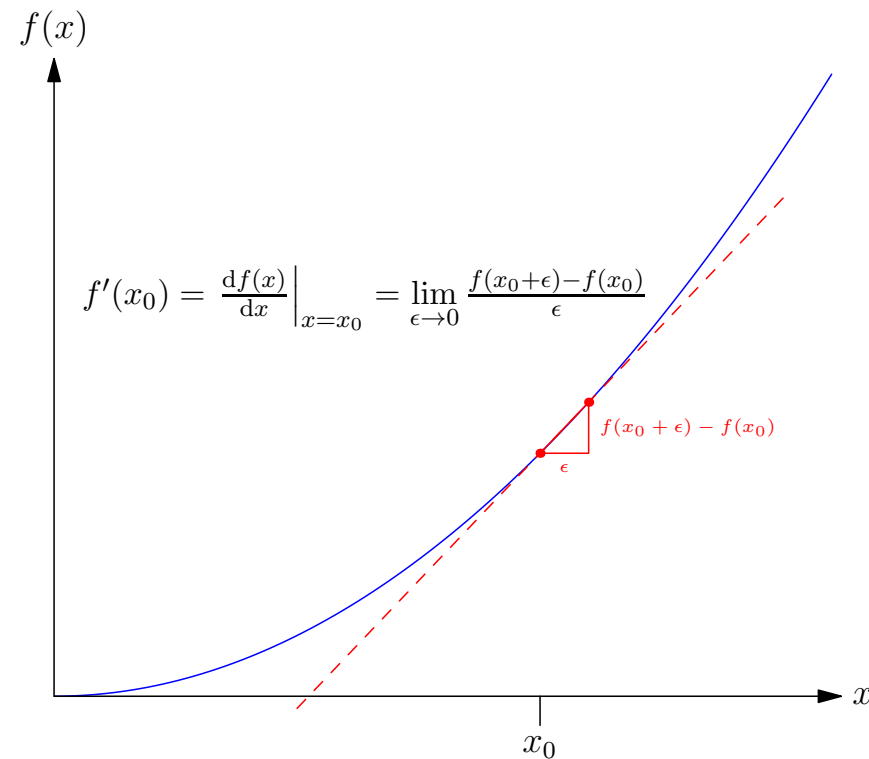


Advanced Machine Learning

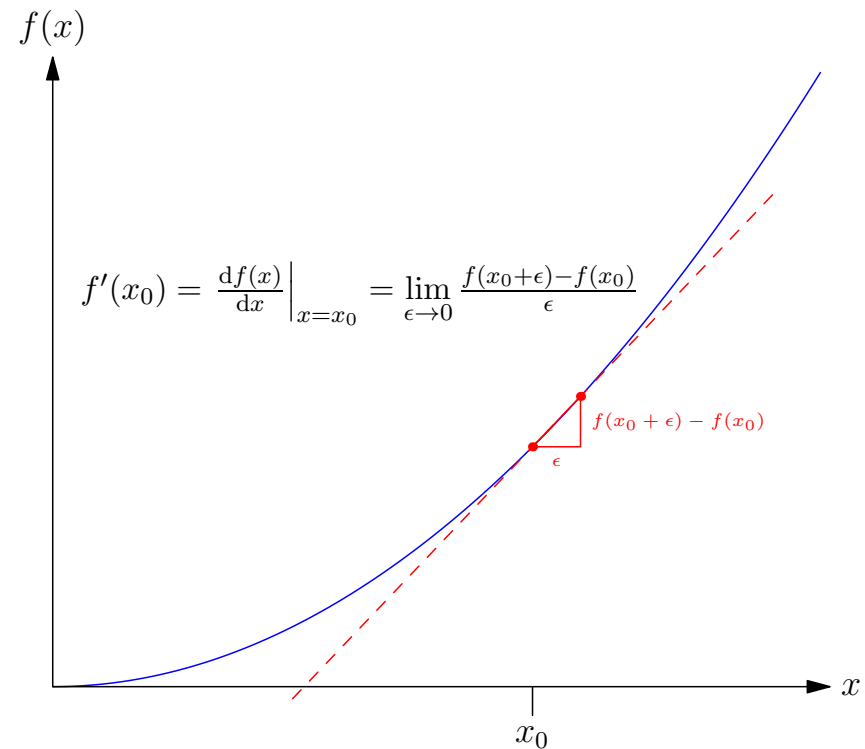
Differential Calculus



Differentiation, product and chain rules, vectors and matrices

Outline

1. **Why Calculus?**
2. Differentiation
3. Vector and Matrix Calculus



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

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

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

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

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

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

Back to Basics

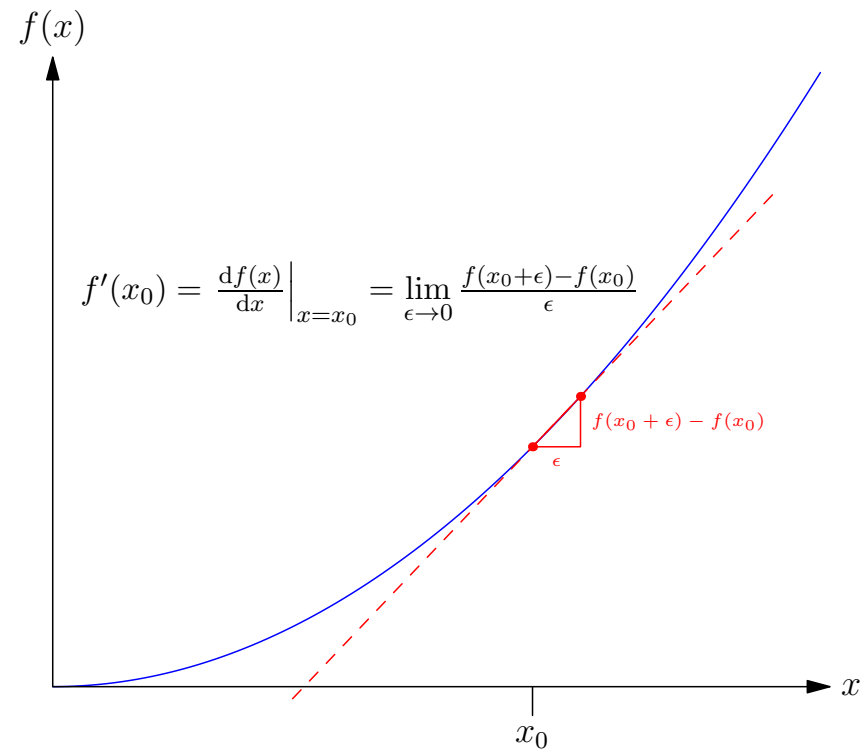
- You have all done A-level maths so should be familiar with the rules of calculus
- But, it is easy to forget the rules and sometimes we use quite sophisticated tricks
- Although the sophisticated tricks really speed up calculations, it pays to be able to understand where these tricks come from

Back to Basics

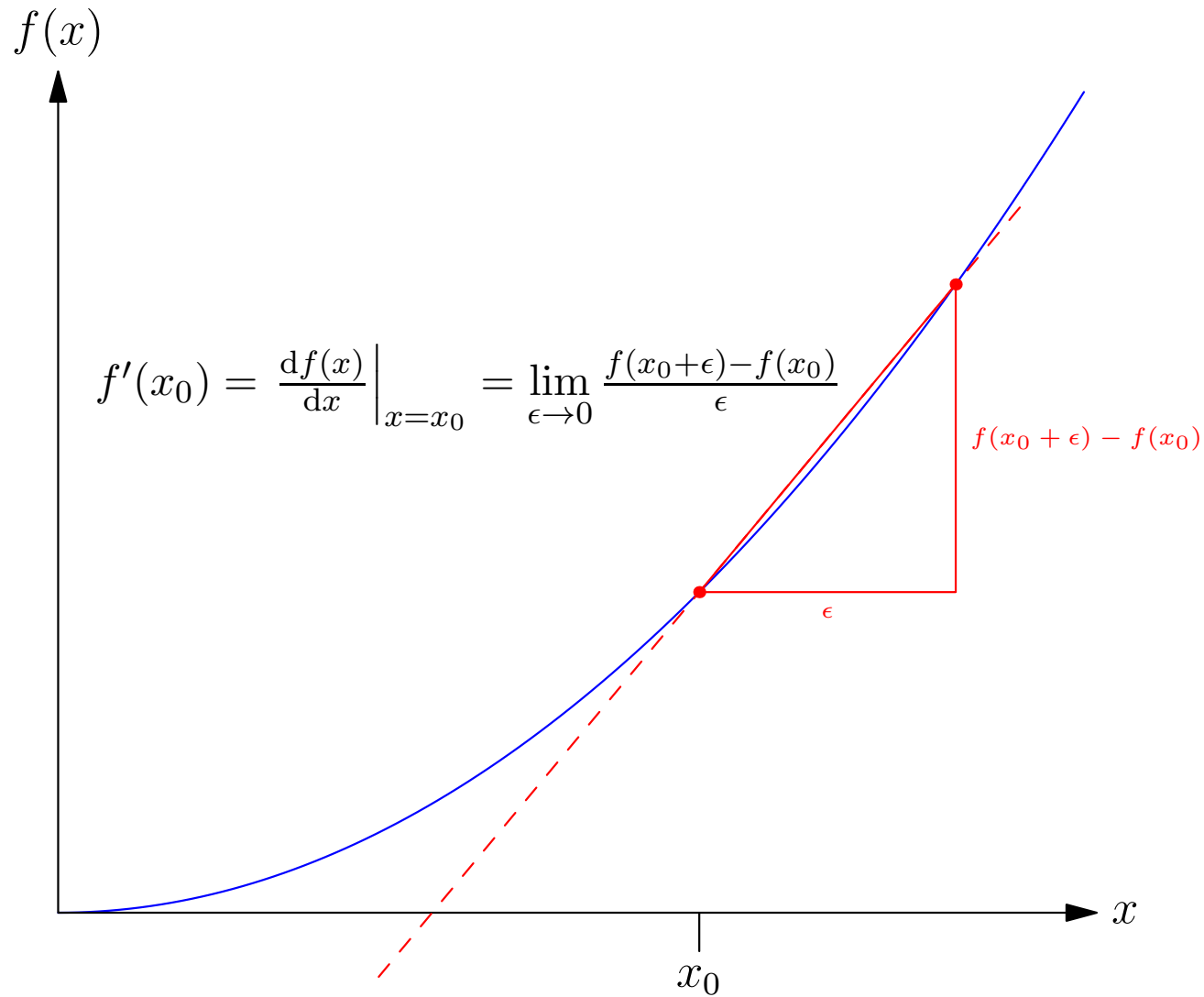
- You have all done A-level maths so should be familiar with the rules of calculus
- But, it is easy to forget the rules and sometimes we use quite sophisticated tricks
- Although the sophisticated tricks really speed up calculations, it pays to be able to understand where these tricks come from

Outline

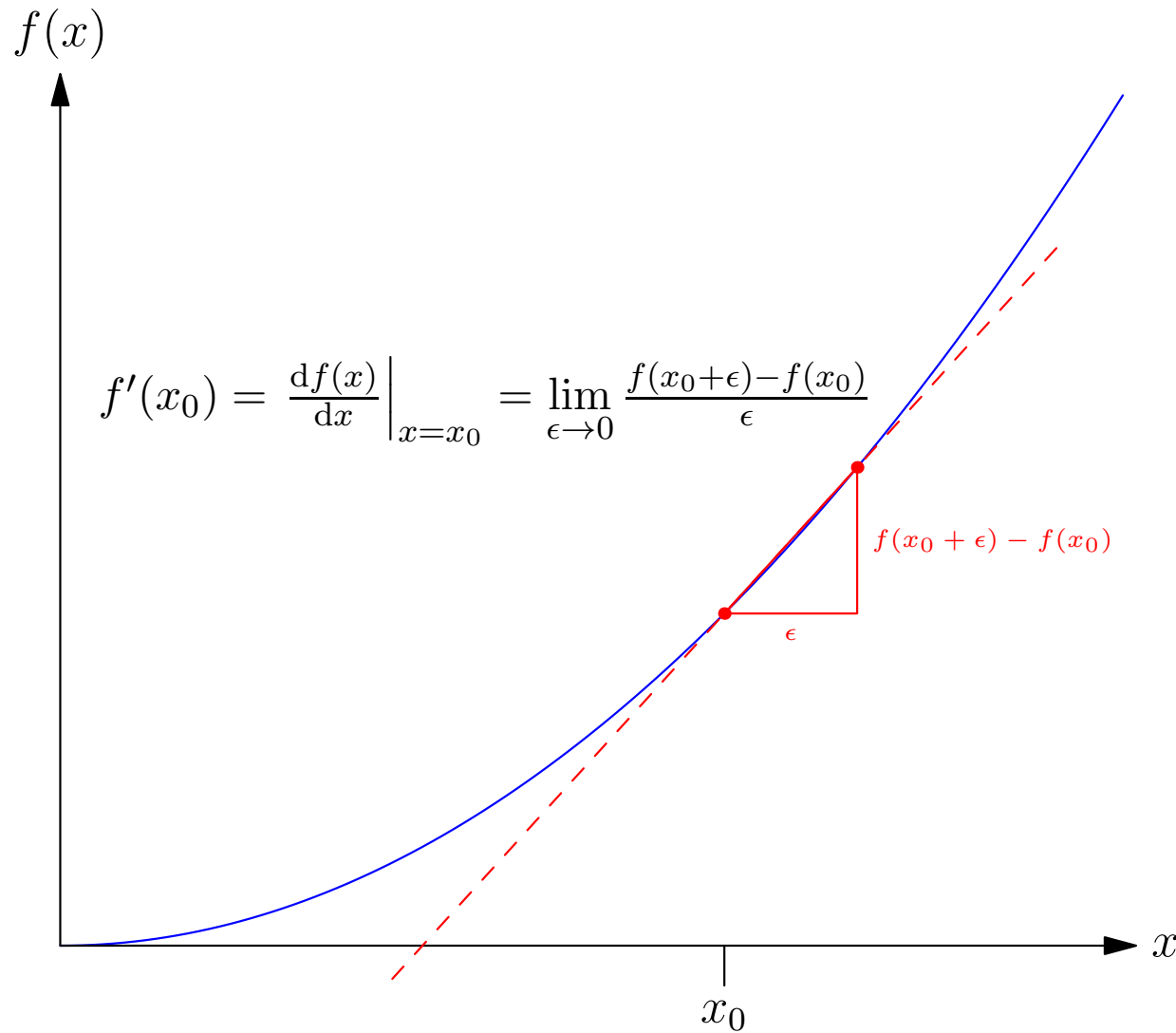
1. Why Calculus?
2. **Differentiation**
3. Vector and Matrix Calculus



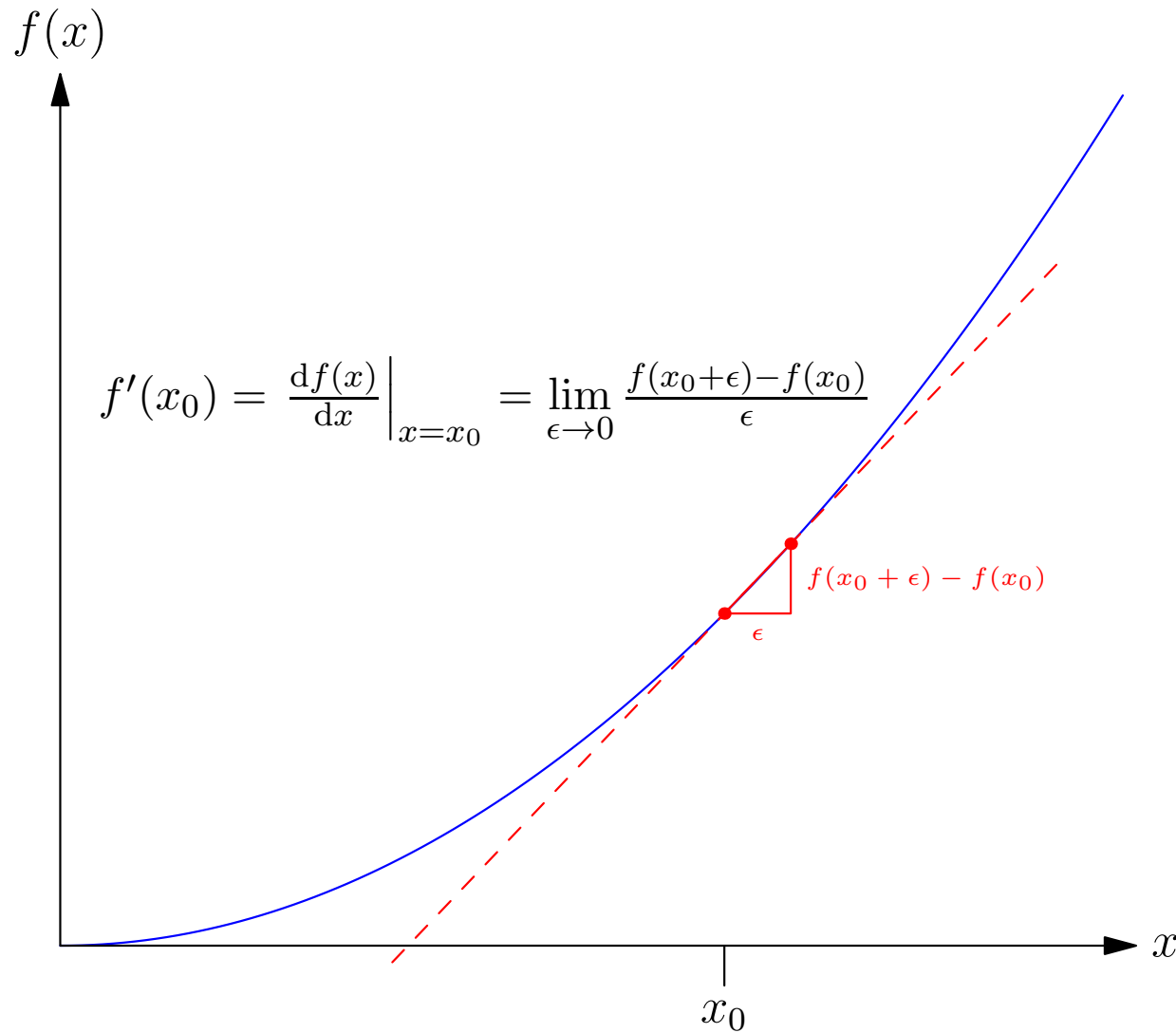
Differentiation



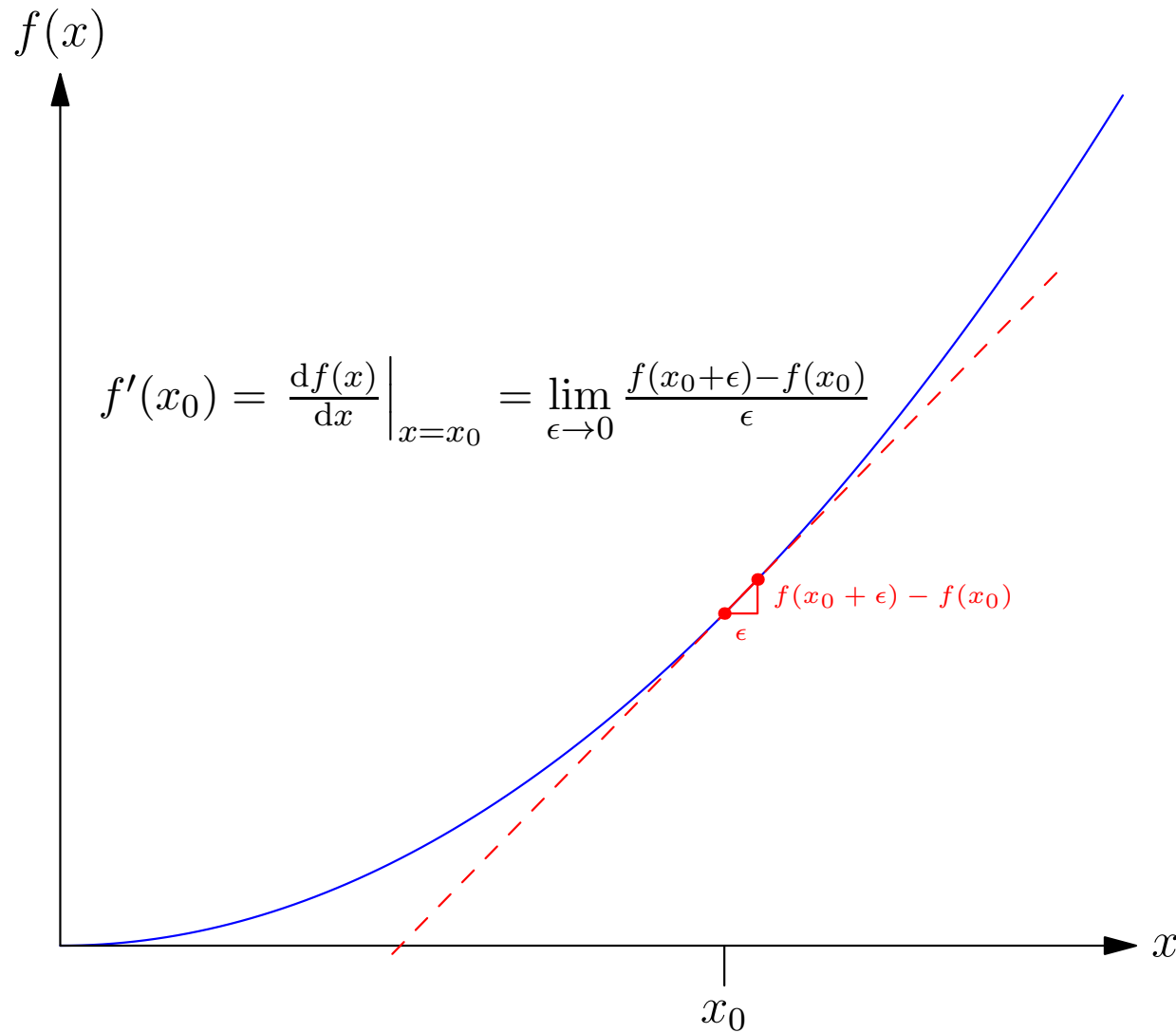
Differentiation



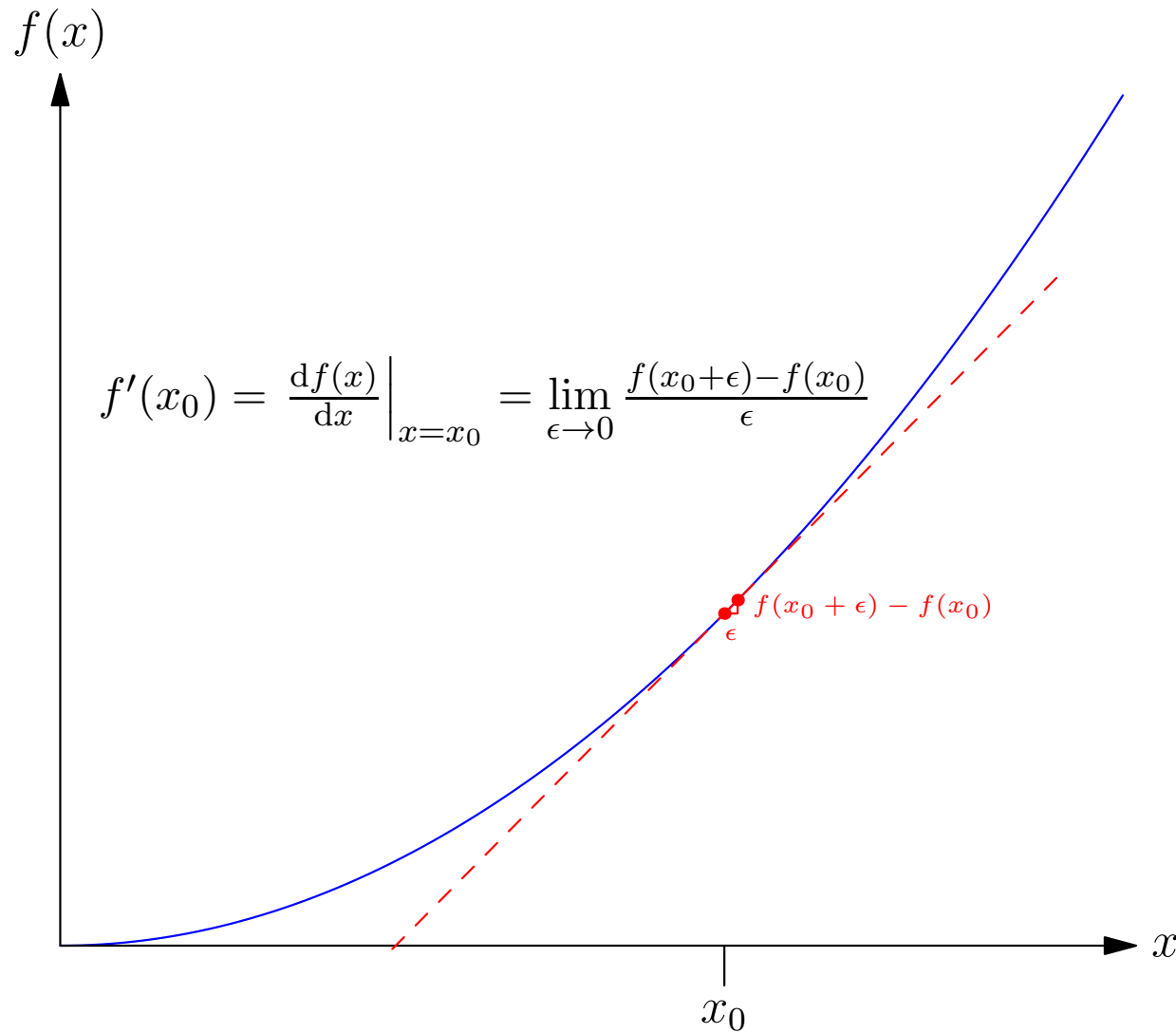
Differentiation



Differentiation



Differentiation



Polynomials

- $f(x) = x^2$

$$\frac{dx^2}{dx} = \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon}$$

Polynomials

- $f(x) = x^2$

$$\frac{dx^2}{dx} = \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon}$$

Polynomials

- $f(x) = x^2$

$$\begin{aligned}\frac{dx^2}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} 2x + \epsilon\end{aligned}$$

Polynomials

- $f(x) = x^2$

$$\begin{aligned}\frac{dx^2}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} 2x + \epsilon = 2x\end{aligned}$$

Polynomials

- $f(x) = x^2$

$$\begin{aligned}\frac{dx^2}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} 2x + \epsilon = 2x\end{aligned}$$

- $(x + \epsilon)^n = (x + \epsilon)(x + \epsilon) \cdots (x + \epsilon)$

Polynomials

- $f(x) = x^2$

$$\begin{aligned}\frac{dx^2}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} 2x + \epsilon = 2x\end{aligned}$$

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

Polynomials

- $f(x) = x^2$

$$\begin{aligned}\frac{dx^2}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} 2x + \epsilon = 2x\end{aligned}$$

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

$$\frac{dx^n}{dx} = \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^n - x^n}{\epsilon}$$

Polynomials

- $f(x) = x^2$

$$\begin{aligned}\frac{dx^2}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} 2x + \epsilon = 2x\end{aligned}$$

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

$$\frac{dx^n}{dx} = \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^n - x^n}{\epsilon} = \lim_{\epsilon \rightarrow 0} nx^{n-1} + O(\epsilon)$$

Polynomials

- $f(x) = x^2$

$$\begin{aligned}\frac{dx^2}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(x^2 + 2\epsilon x + \epsilon^2) - x^2}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} 2x + \epsilon = 2x\end{aligned}$$

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

$$\frac{dx^n}{dx} = \lim_{\epsilon \rightarrow 0} \frac{(x + \epsilon)^n - x^n}{\epsilon} = \lim_{\epsilon \rightarrow 0} nx^{n-1} + O(\epsilon) = nx^{n-1}$$

Linearity of derivatives

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

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

Linearity of derivatives

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

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

Linearity of derivatives

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

$$\begin{aligned} \frac{d(a f(x) + b g(x))}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(a f(x + \epsilon) + b g(x + \epsilon)) - (a f(x) + b g(x))}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{a \epsilon f'(x) + b \epsilon g'(x) + O(\epsilon^2)}{\epsilon} \end{aligned}$$

Linearity of derivatives

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

$$\begin{aligned}\frac{d(a f(x) + b g(x))}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(a f(x + \epsilon) + b g(x + \epsilon)) - (a f(x) + b g(x))}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{a \epsilon f'(x) + b \epsilon g'(x) + O(\epsilon^2)}{\epsilon} \\ &= a f'(x) + b g'(x)\end{aligned}$$

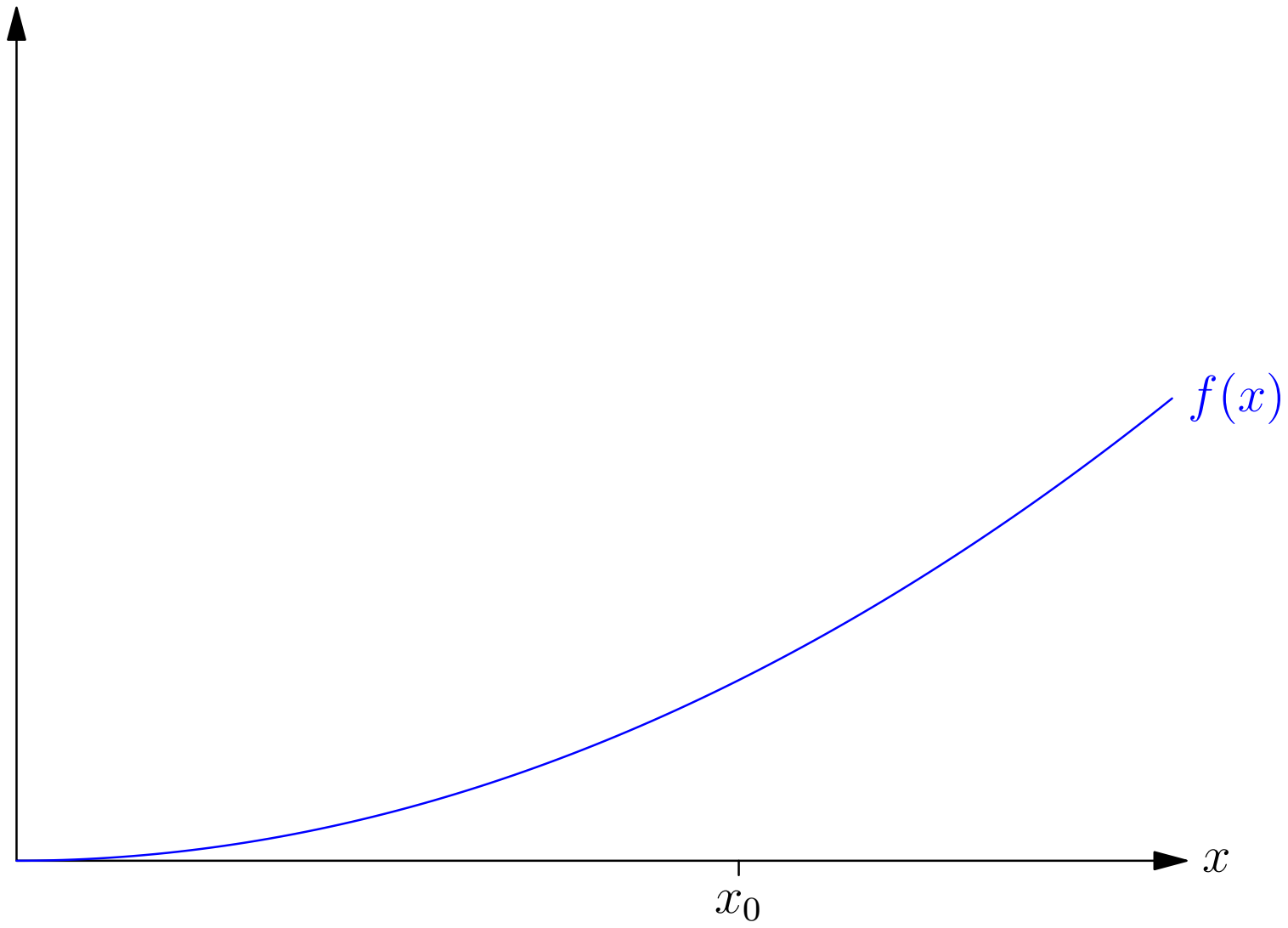
Linearity of derivatives

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

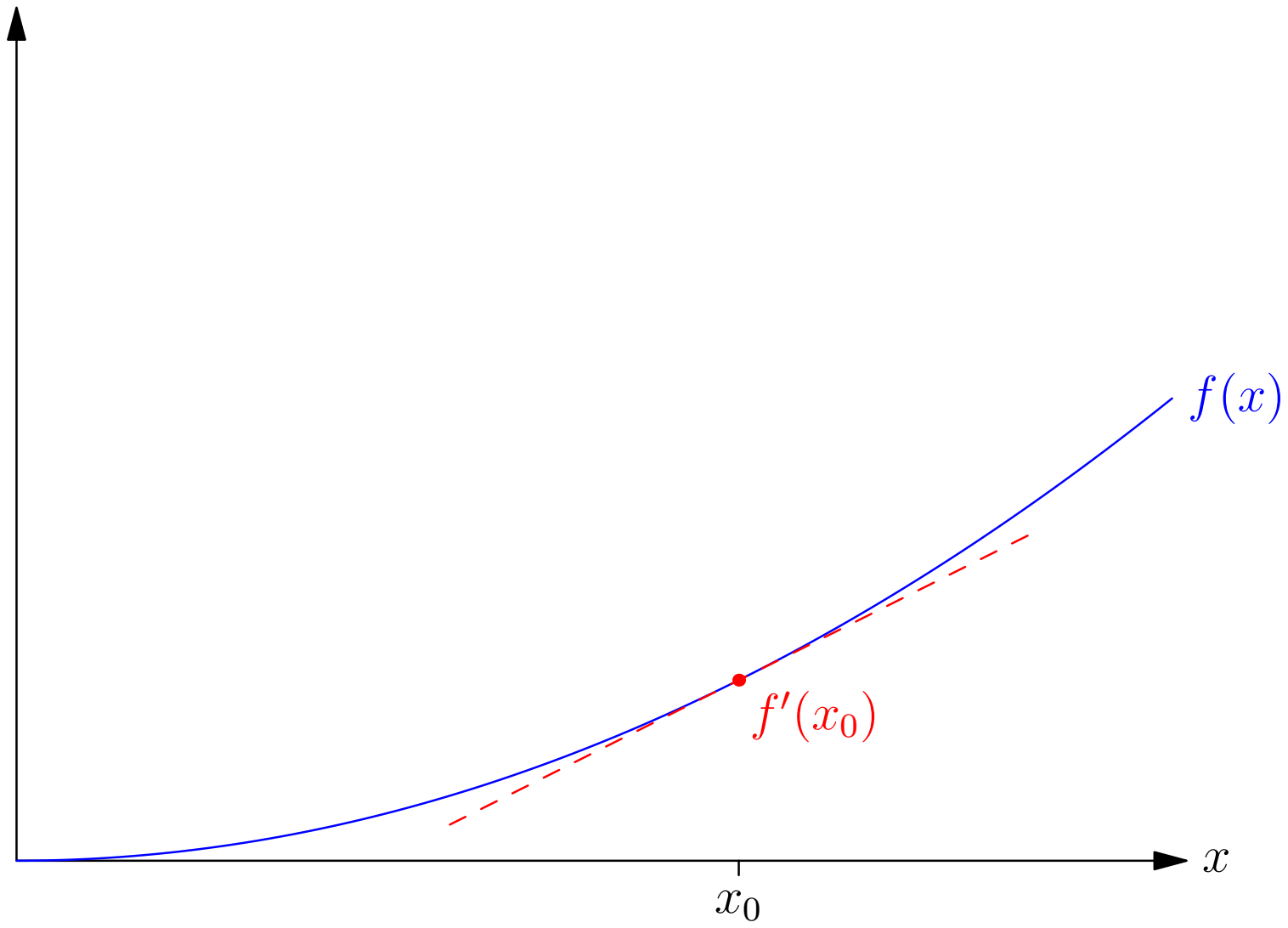
$$\begin{aligned}\frac{d(a f(x) + b g(x))}{dx} &= \lim_{\epsilon \rightarrow 0} \frac{(a f(x + \epsilon) + b g(x + \epsilon)) - (a f(x) + b g(x))}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{a \epsilon f'(x) + b \epsilon g'(x) + O(\epsilon^2)}{\epsilon} \\ &= a f'(x) + b g'(x)\end{aligned}$$

- **Differentiation is a linear operation!**

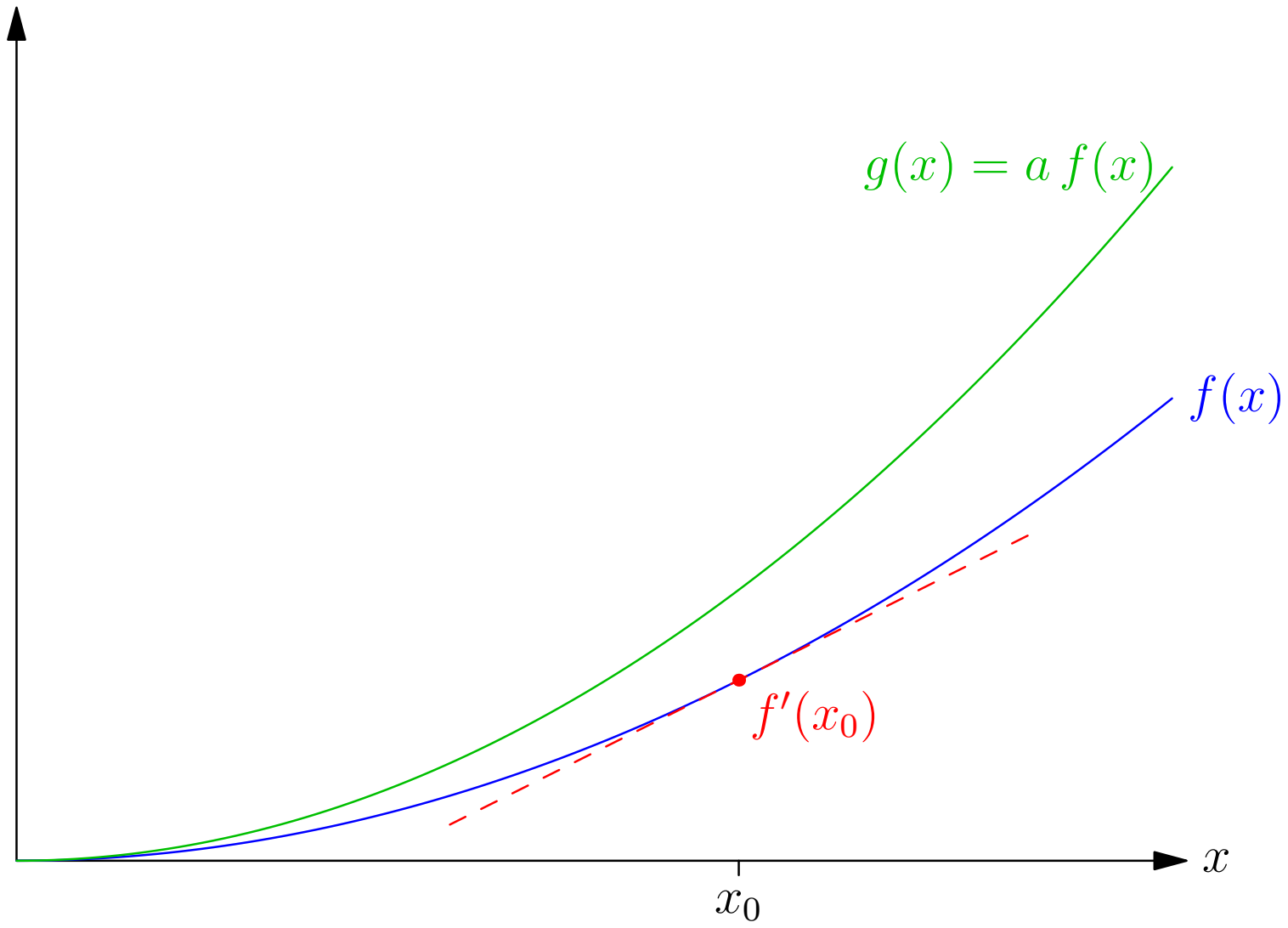
Linearity in Pictures



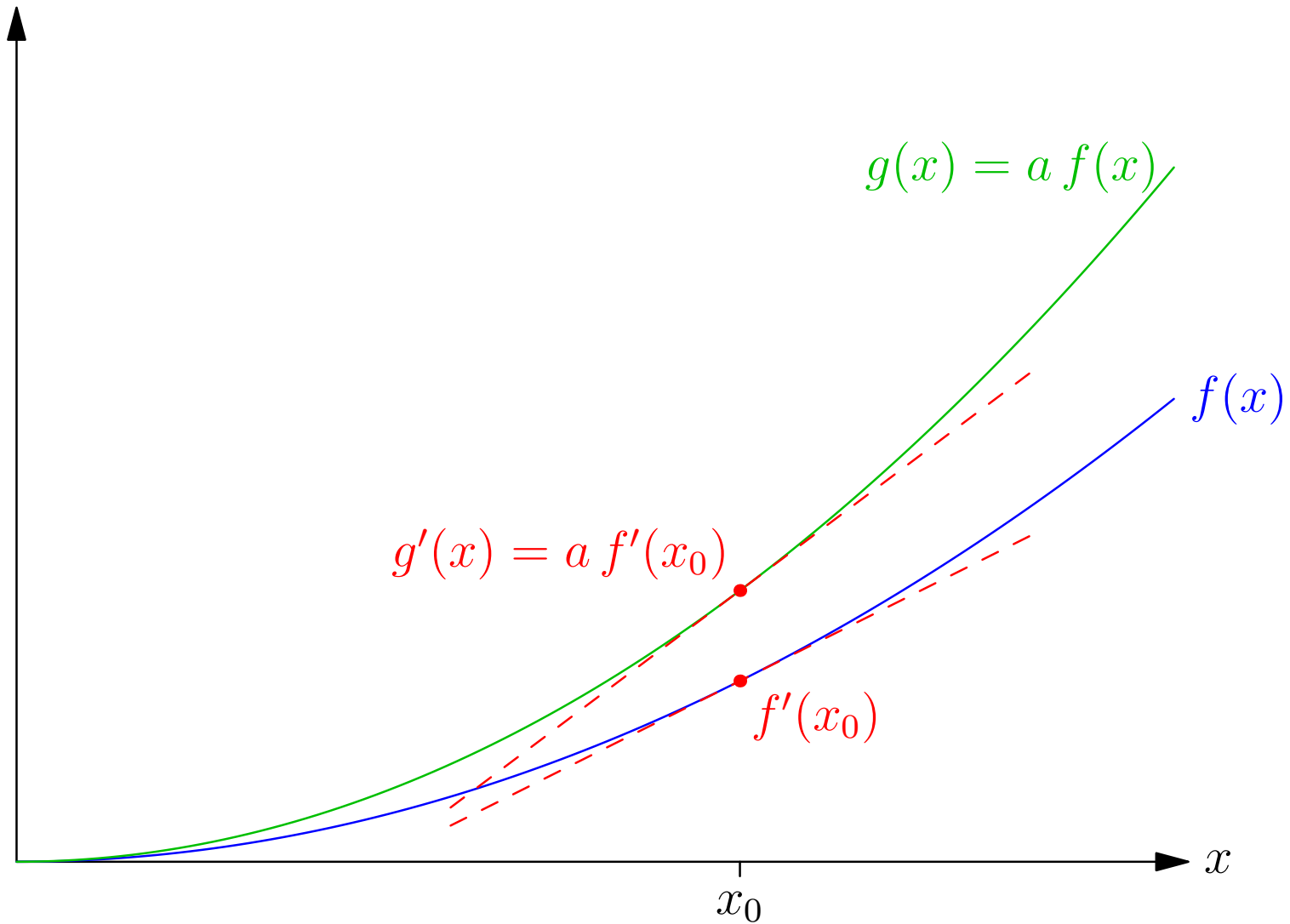
Linearity in Pictures



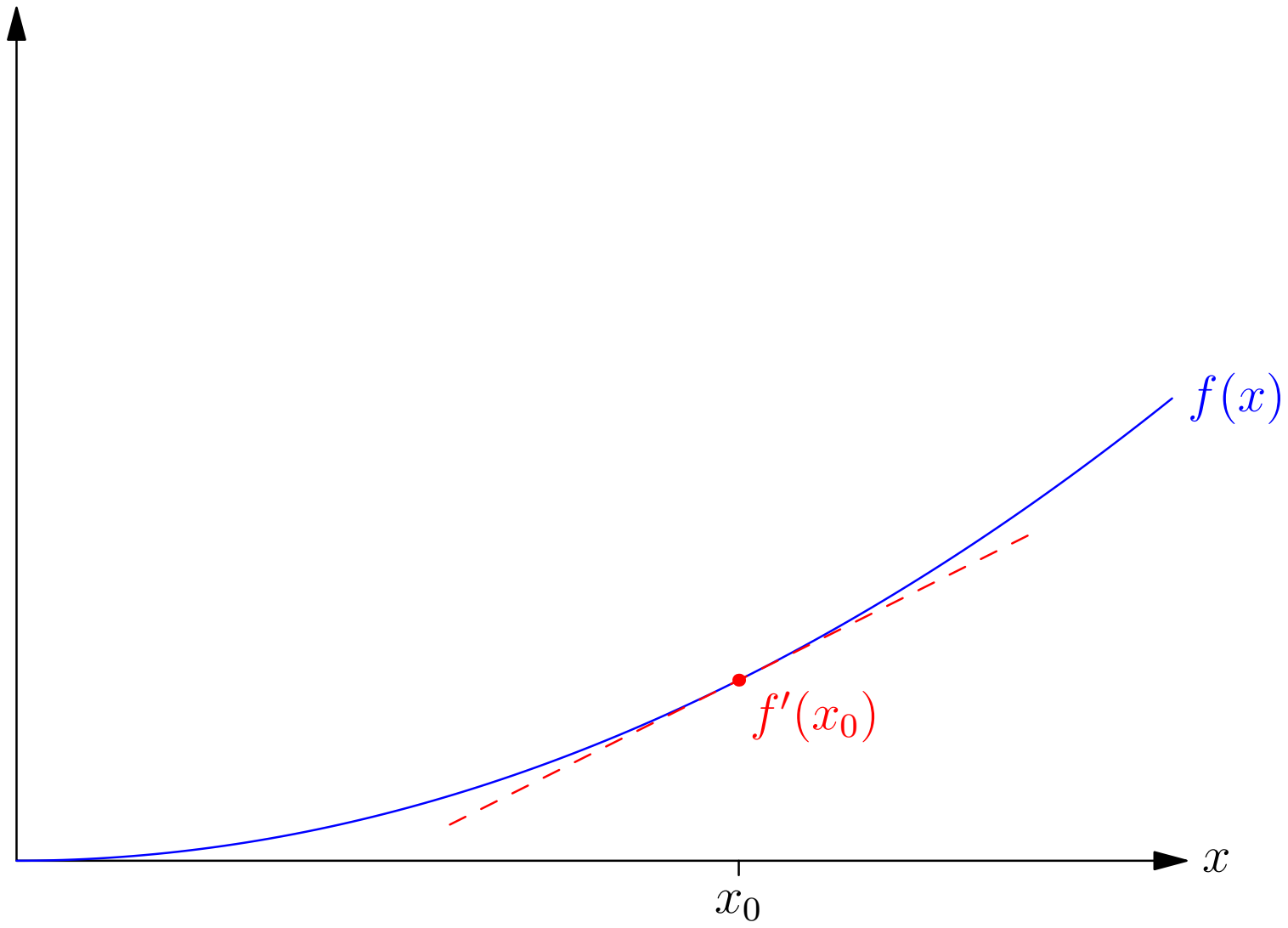
Linearity in Pictures



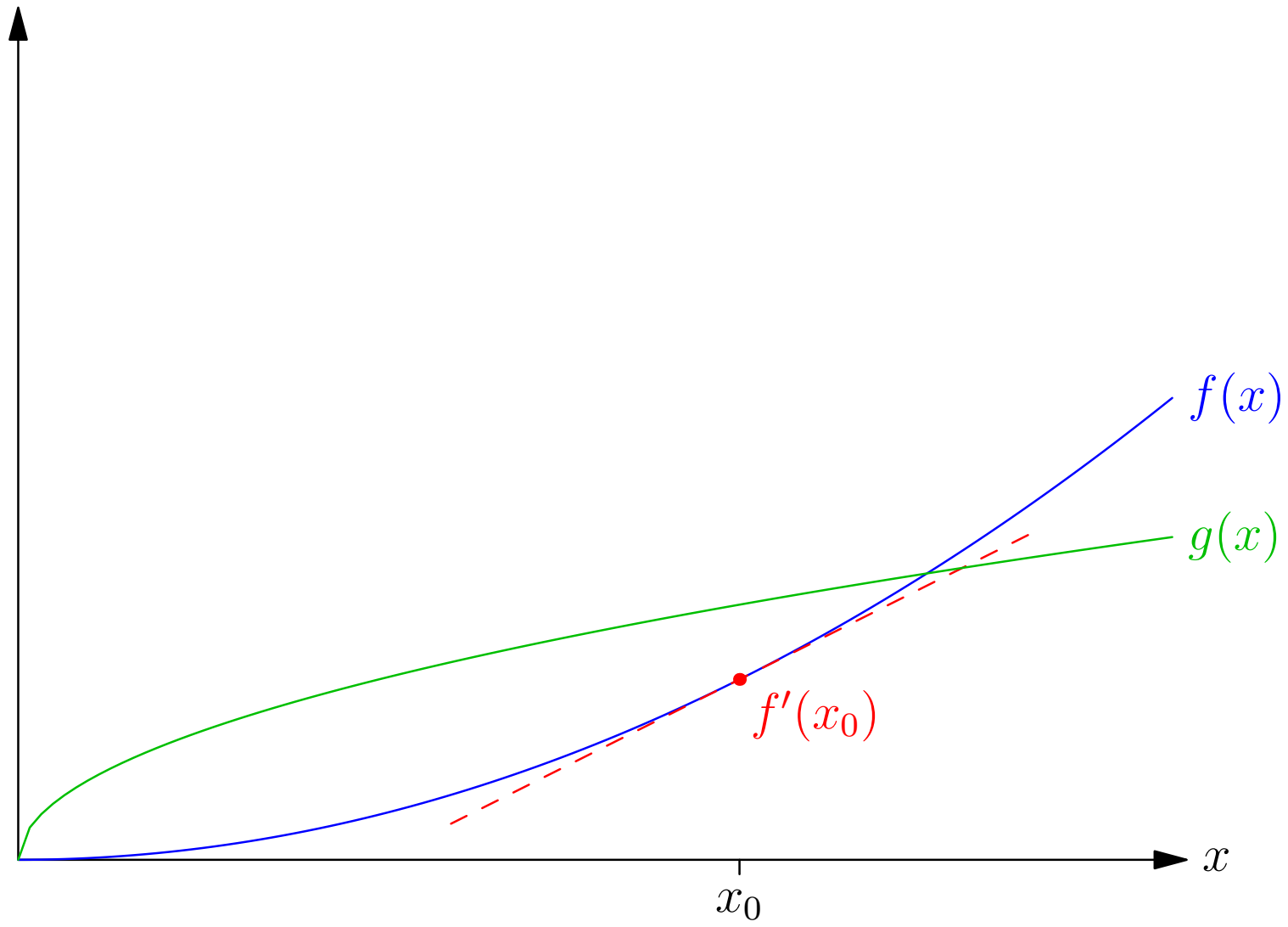
Linearity in Pictures



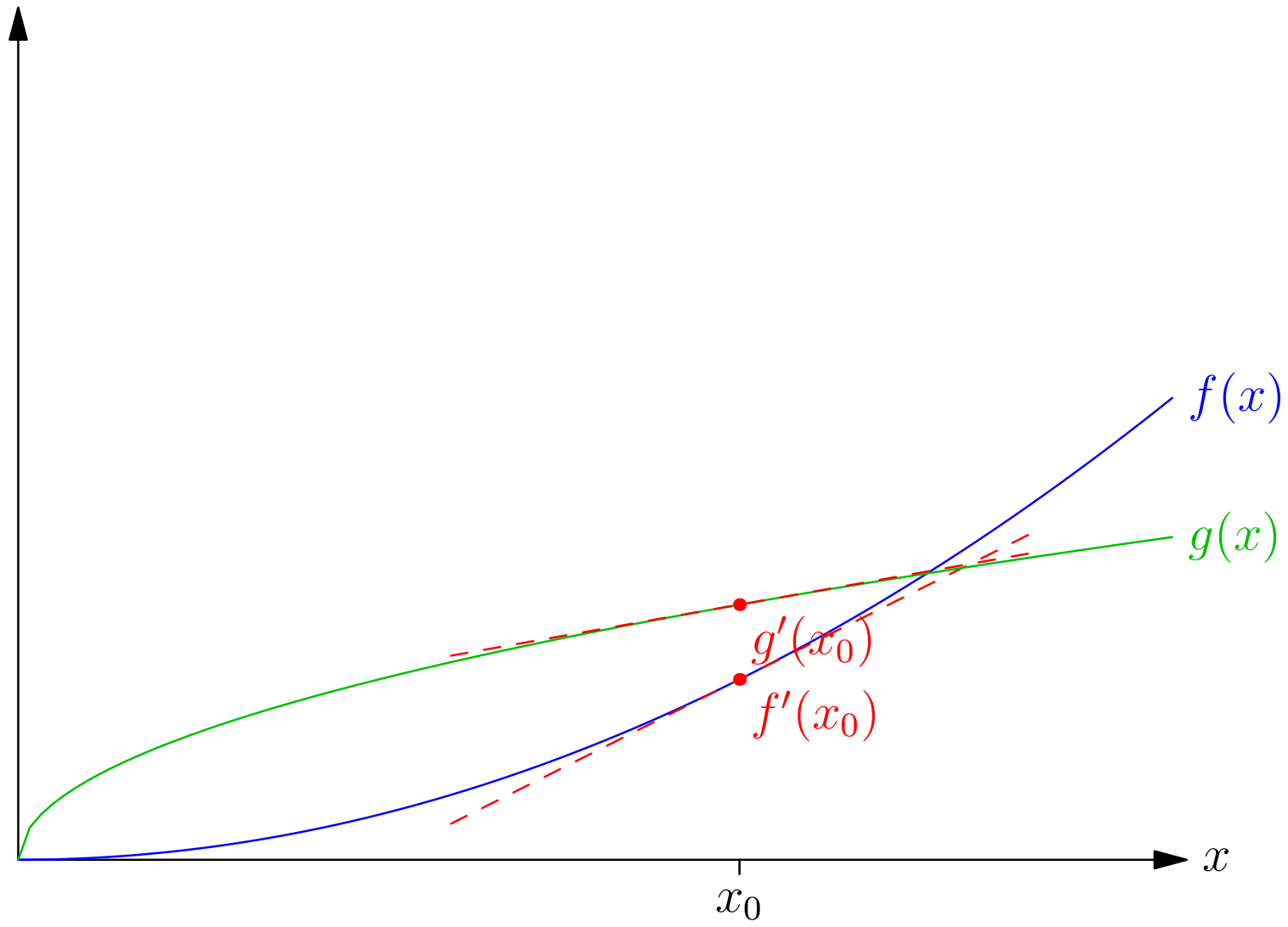
Linearity in Pictures



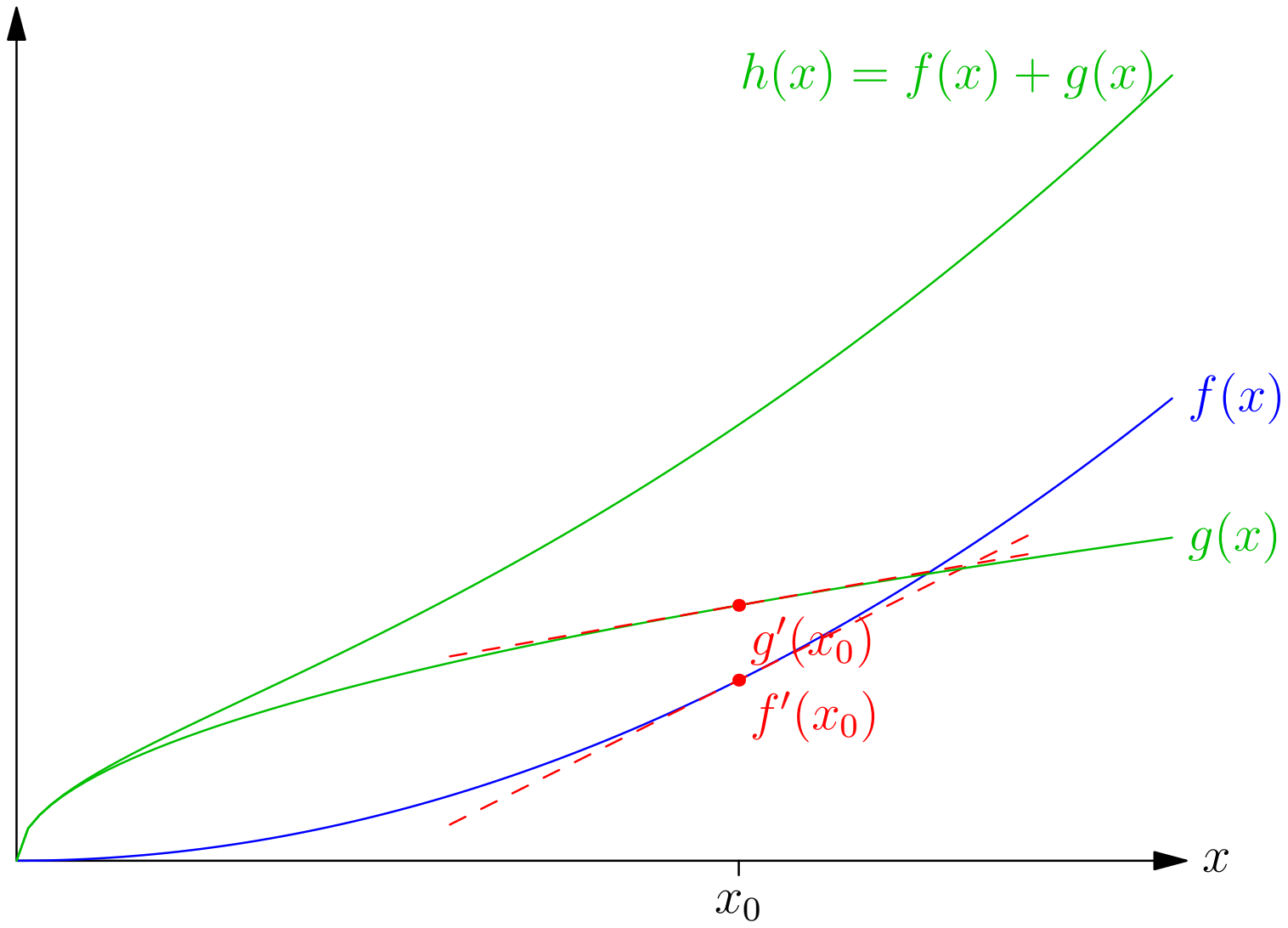
Linearity in Pictures



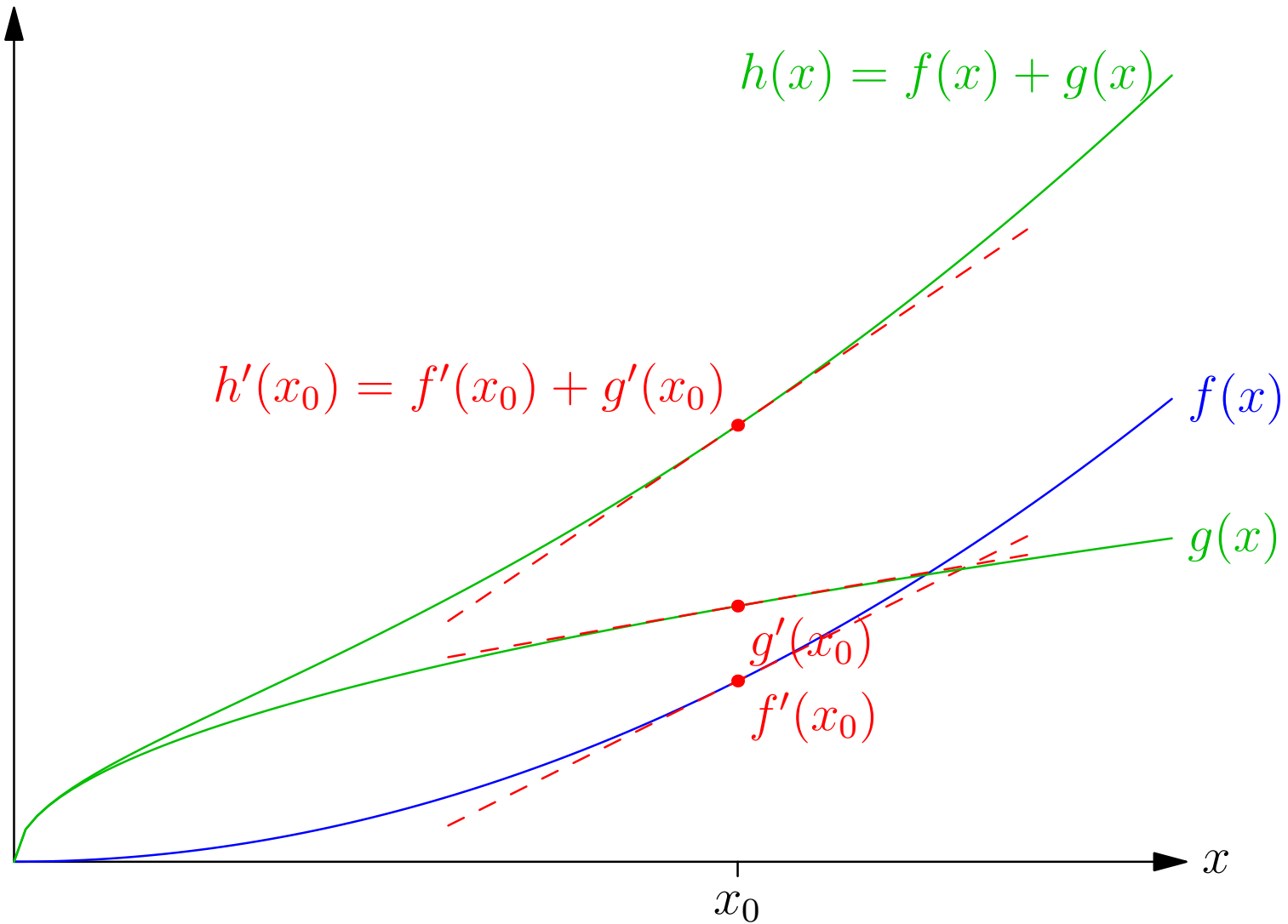
Linearity in Pictures



Linearity in Pictures



Linearity in Pictures



Product Rule

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

Product Rule

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

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

Product Rule

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

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

Product Rule

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

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

Product Rule

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

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

Product Rule

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

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

- 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

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

Chain Rule

- Recall $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$
- Let $h(x) = f(g(x))$
- Then

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

Chain Rule

- Recall $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$
- Let $h(x) = f(g(x))$
- Then

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

Chain Rule

- Recall $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$
- Let $h(x) = f(g(x))$
- Then

$$\begin{aligned} h(x + \epsilon) &= f(g(x + \epsilon)) = f(g(x) + \epsilon g'(x) + O(\epsilon^2)) \\ &= f(g(x)) + \epsilon g'(x) f'(g(x)) + O(\epsilon^2) \end{aligned}$$

Chain Rule

- Recall $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$
- Let $h(x) = f(g(x))$
- Then

$$\begin{aligned} h(x + \epsilon) &= f(g(x + \epsilon)) = f(g(x) + \epsilon g'(x) + O(\epsilon^2)) \\ &= f(g(x)) + \epsilon g'(x) f'(g(x)) + O(\epsilon^2) \end{aligned}$$

- Thus

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

Chain Rule

- Recall $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$
- Let $h(x) = f(g(x))$
- Then

$$\begin{aligned} h(x + \epsilon) &= f(g(x + \epsilon)) = f(g(x) + \epsilon g'(x) + O(\epsilon^2)) \\ &= f(g(x)) + \epsilon g'(x) f'(g(x)) + O(\epsilon^2) \end{aligned}$$

- Thus

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

- This is the famous **chain rule**

Chain Rule

- Recall $f(x + \epsilon) = f(x) + \epsilon f'(x) + O(\epsilon^2)$
- Let $h(x) = f(g(x))$
- Then

$$\begin{aligned} h(x + \epsilon) &= f(g(x + \epsilon)) = f(g(x) + \epsilon g'(x) + O(\epsilon^2)) \\ &= f(g(x)) + \epsilon g'(x) f'(g(x)) + O(\epsilon^2) \end{aligned}$$

- Thus

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

- 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

$$\frac{df(g(x))}{dx} = \frac{df(g)}{dg} \frac{dg(x)}{dx}$$

- Sometimes this is neater or easier to remember

$$\frac{de^{\cos(x^2)}}{dx} = \frac{de^{\cos(x^2)}}{d\cos(x^2)} \frac{d\cos(x^2)}{dx^2} \frac{dx^2}{dx}$$

More on chain rules

- We can also write the chain rule as

$$\frac{df(g(x))}{dx} = \frac{df(g)}{dg} \frac{dg(x)}{dx}$$

- Sometimes this is neater or easier to remember

$$\frac{de^{\cos(x^2)}}{dx} = \frac{de^{\cos(x^2)}}{d\cos(x^2)} \frac{d\cos(x^2)}{dx^2} \frac{dx^2}{dx}$$

More on chain rules

- We can also write the chain rule as

$$\frac{df(g(x))}{dx} = \frac{df(g)}{dg} \frac{dg(x)}{dx}$$

- Sometimes this is neater or easier to remember

$$\begin{aligned} \frac{de^{\cos(x^2)}}{dx} &= \frac{de^{\cos(x^2)}}{d\cos(x^2)} \frac{d\cos(x^2)}{dx^2} \frac{dx^2}{dx} \\ &= e^{\cos(x^2)} (-\sin(x^2)) 2x \end{aligned}$$

More on chain rules

- We can also write the chain rule as

$$\frac{df(g(x))}{dx} = \frac{df(g)}{dg} \frac{dg(x)}{dx}$$

- Sometimes this is neater or easier to remember

$$\begin{aligned} \frac{de^{\cos(x^2)}}{dx} &= \frac{de^{\cos(x^2)}}{d\cos(x^2)} \frac{d\cos(x^2)}{dx^2} \frac{dx^2}{dx} \\ &= e^{\cos(x^2)} (-\sin(x^2)) 2x \\ &= -2x \sin(x^2) e^{\cos(x^2)} \end{aligned}$$

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

$$\frac{dg(f(x))}{dx} = f'(x)g'(f(x))$$

- 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

$$\frac{dg(y)}{dy} = \frac{1}{f'(g(y))}$$

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

$$\frac{dg(f(x))}{dx} = f'(x)g'(f(x))$$

- 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

$$\frac{dg(y)}{dy} = \frac{1}{f'(g(y))}$$

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

$$\frac{dg(f(x))}{dx} = f'(x)g'(f(x)) = 1$$

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

$$\frac{dg(y)}{dy} = \frac{1}{f'(g(y))}$$

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

$$\frac{dg(f(x))}{dx} = f'(x)g'(f(x)) = 1$$

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

$$\frac{dg(y)}{dy} = \frac{1}{f'(g(y))}$$

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

$$\frac{dg(f(x))}{dx} = f'(x)g'(f(x)) = 1$$

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

$$\frac{dg(y)}{dy} = \frac{1}{f'(g(y))}$$

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

$$\frac{dg(f(x))}{dx} = f'(x)g'(f(x)) = 1$$

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

$$\frac{dg(y)}{dy} = \frac{1}{f'(g(y))}$$

$$\frac{df^{-1}(y)}{dy} = \frac{1}{f'(f^{-1}(y))}$$

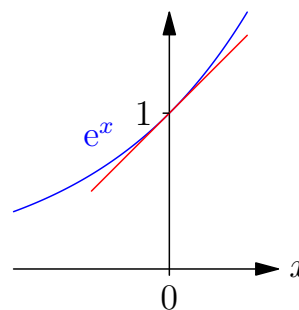
Exponentials

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

Exponentials

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

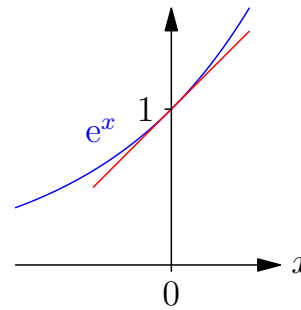
- Now $e^\epsilon \approx (1 + \epsilon)$



Exponentials

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

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

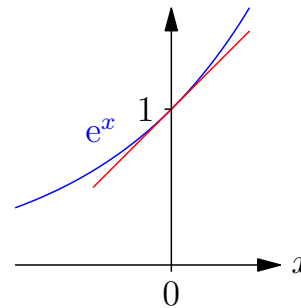


- But $e^{x+\epsilon} = e^x e^\epsilon = e^x (1 + \epsilon + O(\epsilon^2))$

Exponentials

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

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

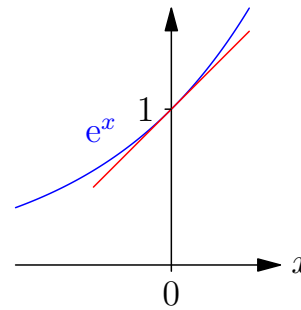


- But $e^{x+\epsilon} = e^x e^\epsilon = e^x (1 + \epsilon + O(\epsilon^2)) = e^x + \epsilon e^x + O(\epsilon^2)$

Exponentials

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

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



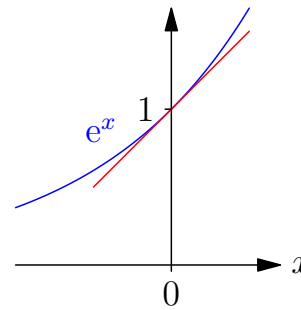
- But $e^{x+\epsilon} = e^x e^\epsilon = e^x (1 + \epsilon + O(\epsilon^2)) = e^x + \epsilon e^x + O(\epsilon^2)$

$$\frac{de^x}{dx} = \lim_{\epsilon \rightarrow 0} \frac{e^{x+\epsilon} - e^x}{\epsilon}$$

Exponentials

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

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



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

$$\frac{de^x}{dx} = \lim_{\epsilon \rightarrow 0} \frac{e^{x+\epsilon} - e^x}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{\epsilon e^x + O(\epsilon^2)}{\epsilon} = e^x$$

Functions of Exponentials

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

$$\frac{de^{cx}}{dx} = \frac{de^{cx}}{dcx} \frac{dcx}{dx}$$

- More generally using the chain rule

$$\frac{de^{g(x)}}{dx} = g'(x)e^{g(x)}$$

Functions of Exponentials

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

$$\frac{de^{cx}}{dx} = \frac{de^{cx}}{dcx} \frac{dcx}{dx} = ce^{cx}$$

- More generally using the chain rule

$$\frac{de^{g(x)}}{dx} = g'(x)e^{g(x)}$$

Functions of Exponentials

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

$$\frac{de^{cx}}{dx} = \frac{de^{cx}}{dcx} \frac{dcx}{dx} = ce^{cx}$$

- More generally using the chain rule

$$\frac{de^{g(x)}}{dx} = g'(x)e^{g(x)}$$

Functions of Exponentials

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

$$\frac{de^{cx}}{dx} = \frac{de^{cx}}{dcx} \frac{dcx}{dx} = ce^{cx}$$

- More generally using the chain rule

$$\frac{de^{g(x)}}{dx} = g'(x)e^{g(x)}$$

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

Functions of Exponentials

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

$$\frac{de^{cx}}{dx} = \frac{de^{cx}}{dcx} \frac{dcx}{dx} = ce^{cx}$$

- More generally using the chain rule

$$\frac{de^{g(x)}}{dx} = g'(x)e^{g(x)}$$

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

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

Functions of Exponentials

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

$$\frac{de^{cx}}{dx} = \frac{de^{cx}}{dcx} \frac{dcx}{dx} = ce^{cx}$$

- More generally using the chain rule

$$\frac{de^{g(x)}}{dx} = g'(x)e^{g(x)}$$

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

$$\frac{da^x}{dx} = \frac{d(e^{\ln(a)})^x}{dx} = \frac{de^{\ln(a)x}}{dx}$$

Functions of Exponentials

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

$$\frac{de^{cx}}{dx} = \frac{de^{cx}}{dcx} \frac{dcx}{dx} = ce^{cx}$$

- More generally using the chain rule

$$\frac{de^{g(x)}}{dx} = g'(x)e^{g(x)}$$

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

$$\frac{da^x}{dx} = \frac{d(e^{\ln(a)})^x}{dx} = \frac{de^{\ln(a)x}}{dx} = \ln(a)e^{\ln(a)x}$$

Functions of Exponentials

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

$$\frac{de^{cx}}{dx} = \frac{de^{cx}}{dcx} \frac{dcx}{dx} = ce^{cx}$$

- More generally using the chain rule

$$\frac{de^{g(x)}}{dx} = g'(x)e^{g(x)}$$

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

$$\frac{da^x}{dx} = \frac{d(e^{\ln(a)})^x}{dx} = \frac{de^{\ln(a)x}}{dx} = \ln(a)e^{\ln(a)x} = \ln(a)a^x$$

Natural Logarithms

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

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

- Recall that if $g(y) = f^{-1}(y)$ then $g'(y) = 1/f'(g(y))$
- Consider $g(y) = \ln(y)$ and $f(x) = e^x$ (with $f'(x) = e^x$)

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

Natural Logarithms

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

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

- Recall that if $g(y) = f^{-1}(y)$ then $g'(y) = 1/f'(g(y))$
- Consider $g(y) = \ln(y)$ and $f(x) = e^x$ (with $f'(x) = e^x$)

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

Natural Logarithms

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

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

- Recall that if $g(y) = f^{-1}(y)$ then $g'(y) = 1/f'(g(y))$
- Consider $g(y) = \ln(y)$ and $f(x) = e^x$ (with $f'(x) = e^x$)

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

Natural Logarithms

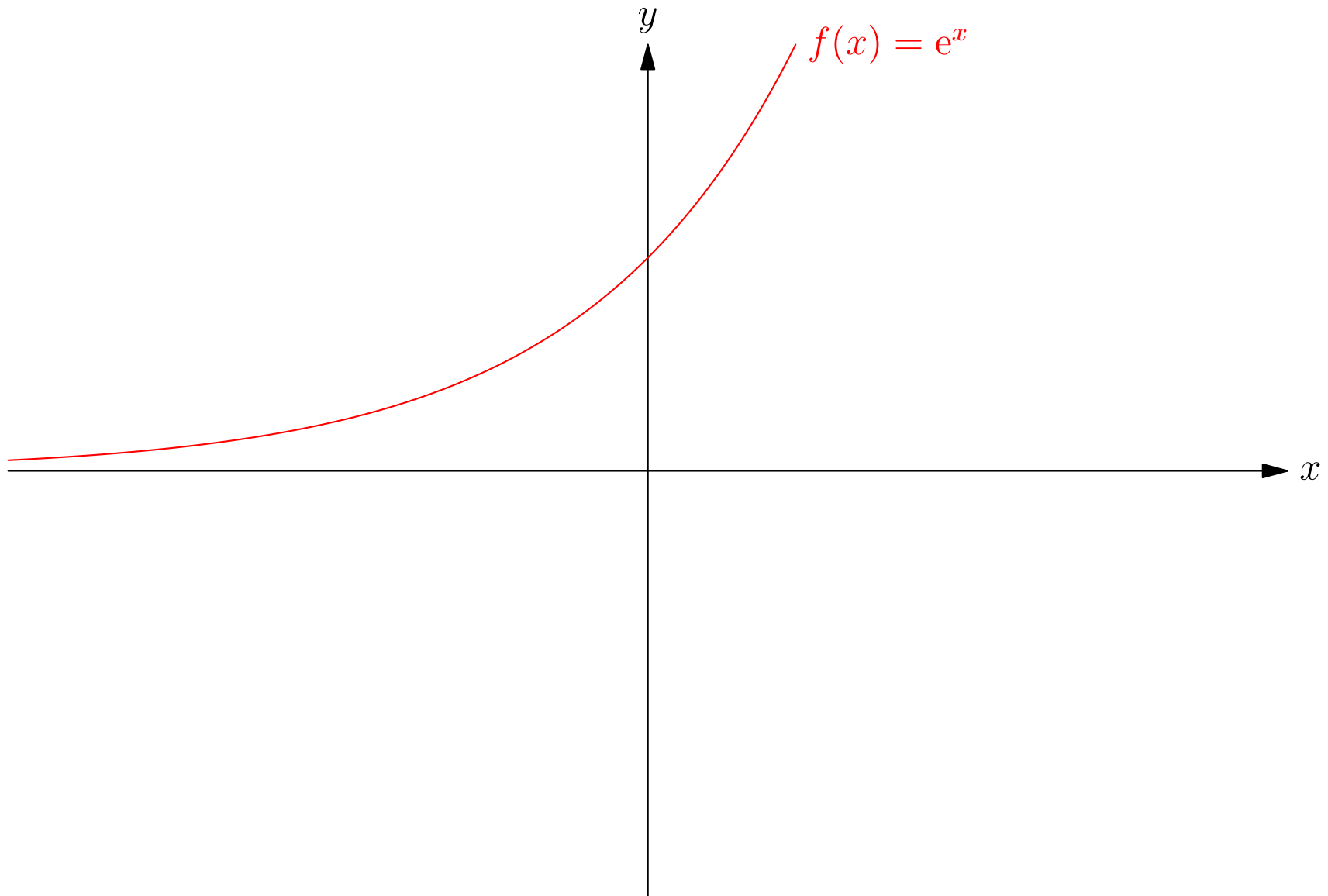
- The natural logarithm is defined as the inverse of e^x

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

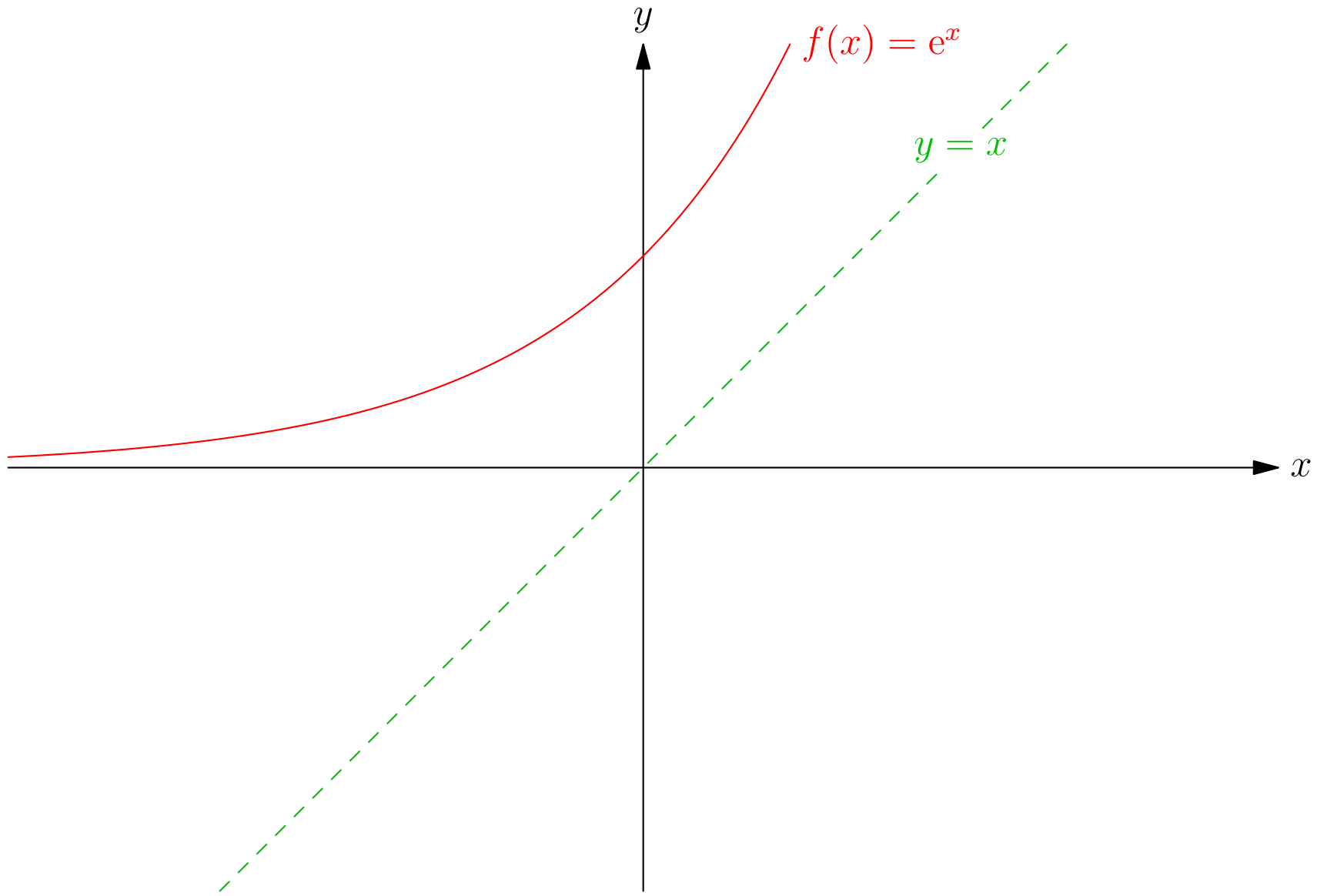
- Recall that if $g(y) = f^{-1}(y)$ then $g'(y) = 1/f'(g(y))$
- Consider $g(y) = \ln(y)$ and $f(x) = e^x$ (with $f'(x) = e^x$)

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

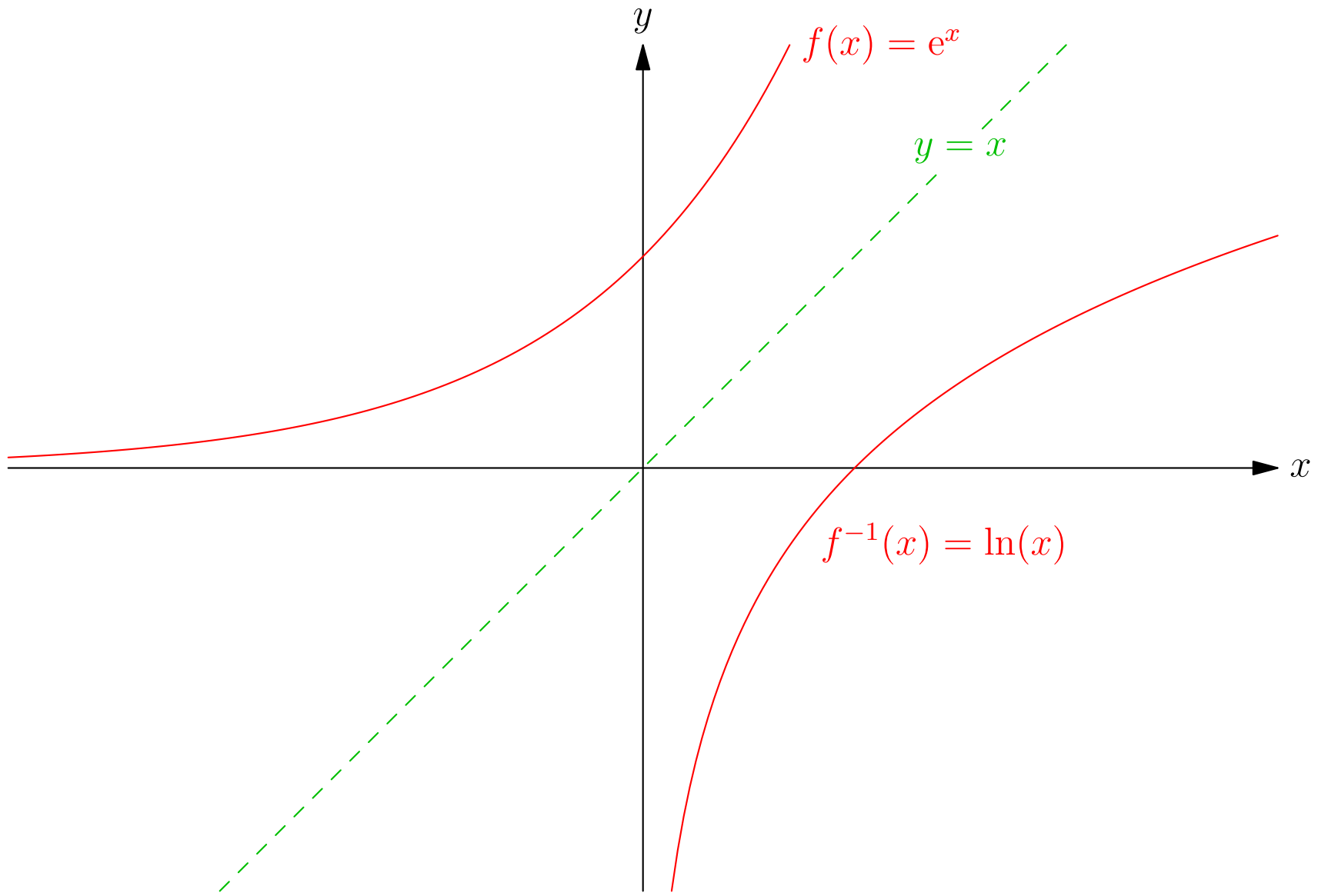
Exponentials and Logarithms



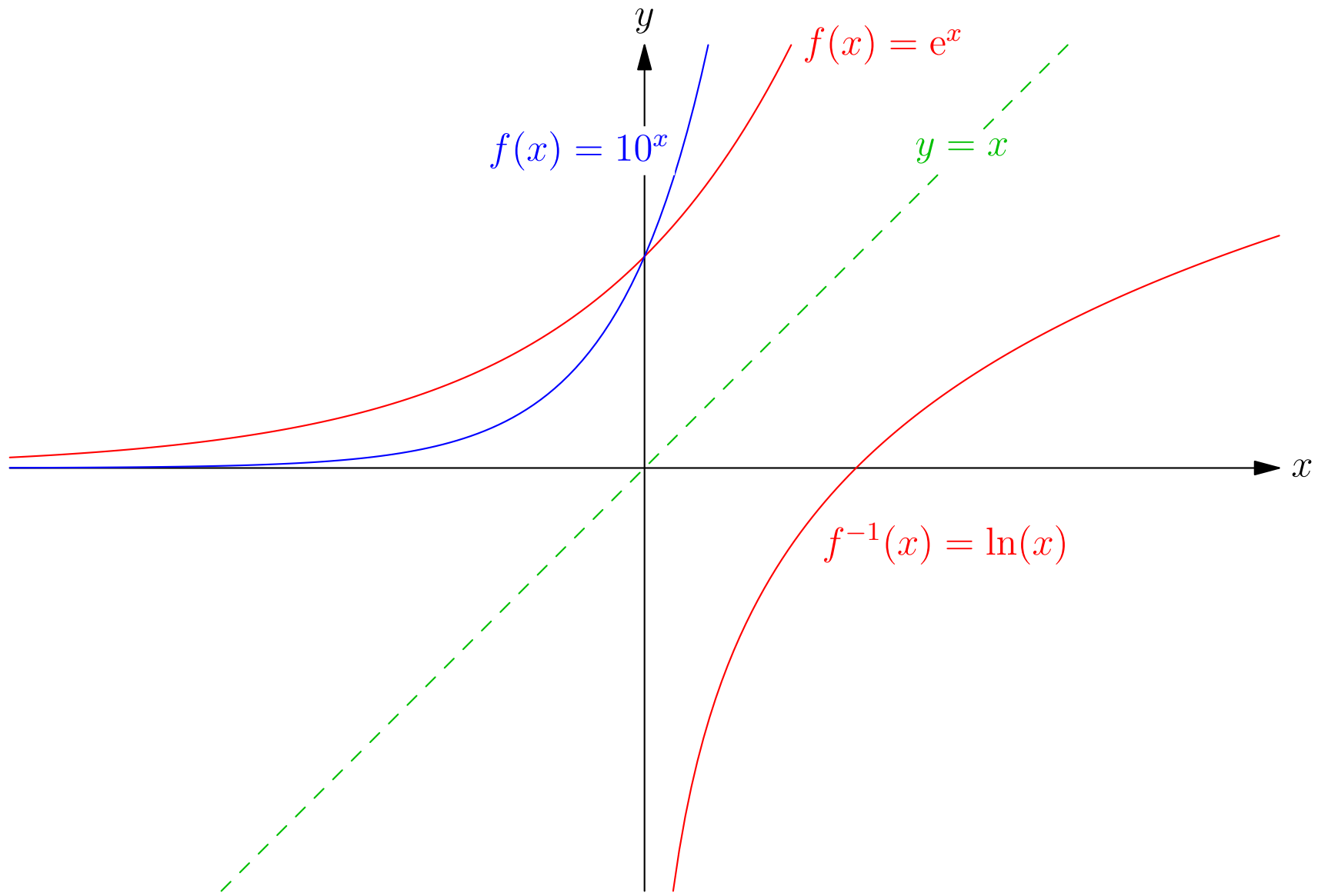
Exponentials and Logarithms



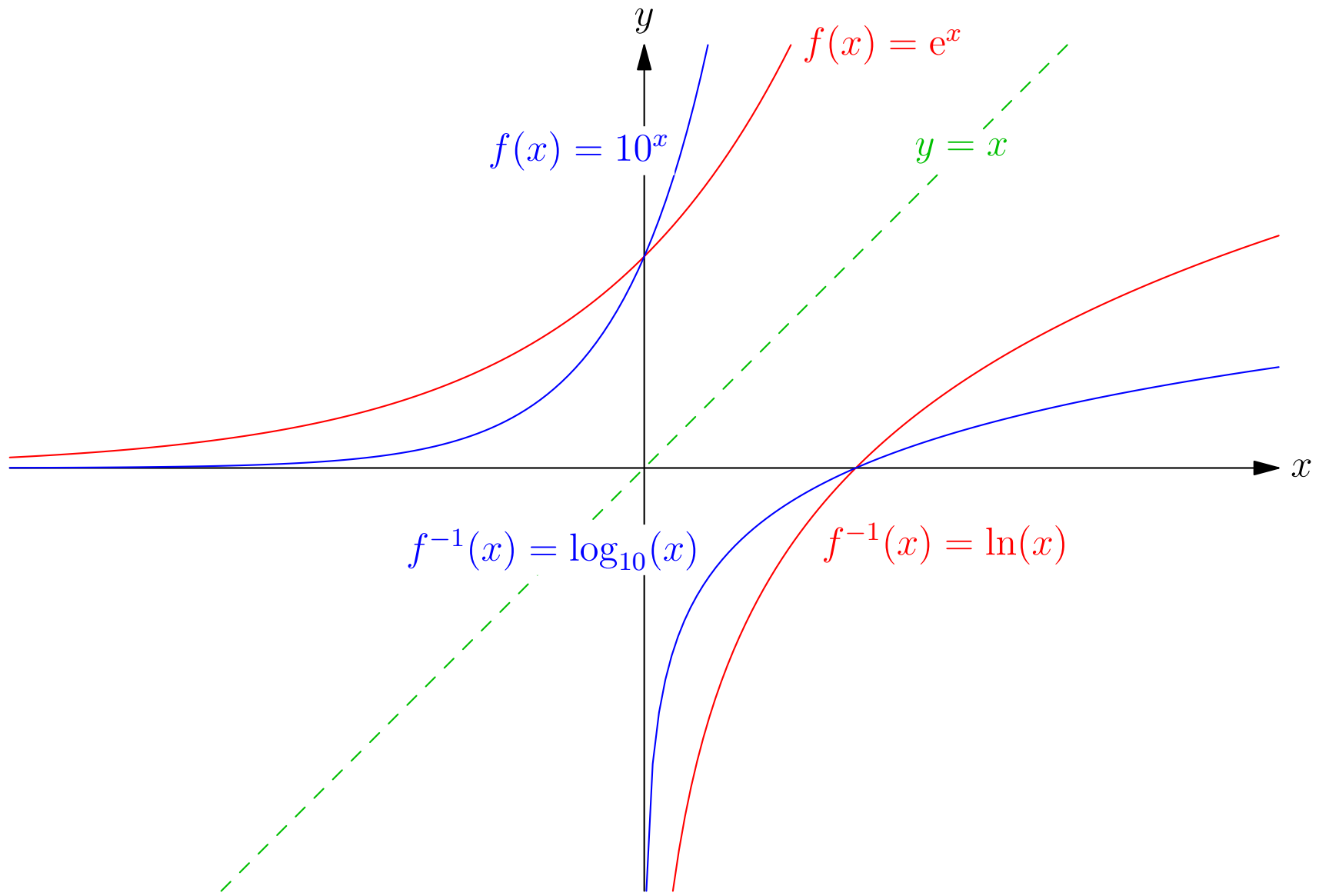
Exponentials and Logarithms



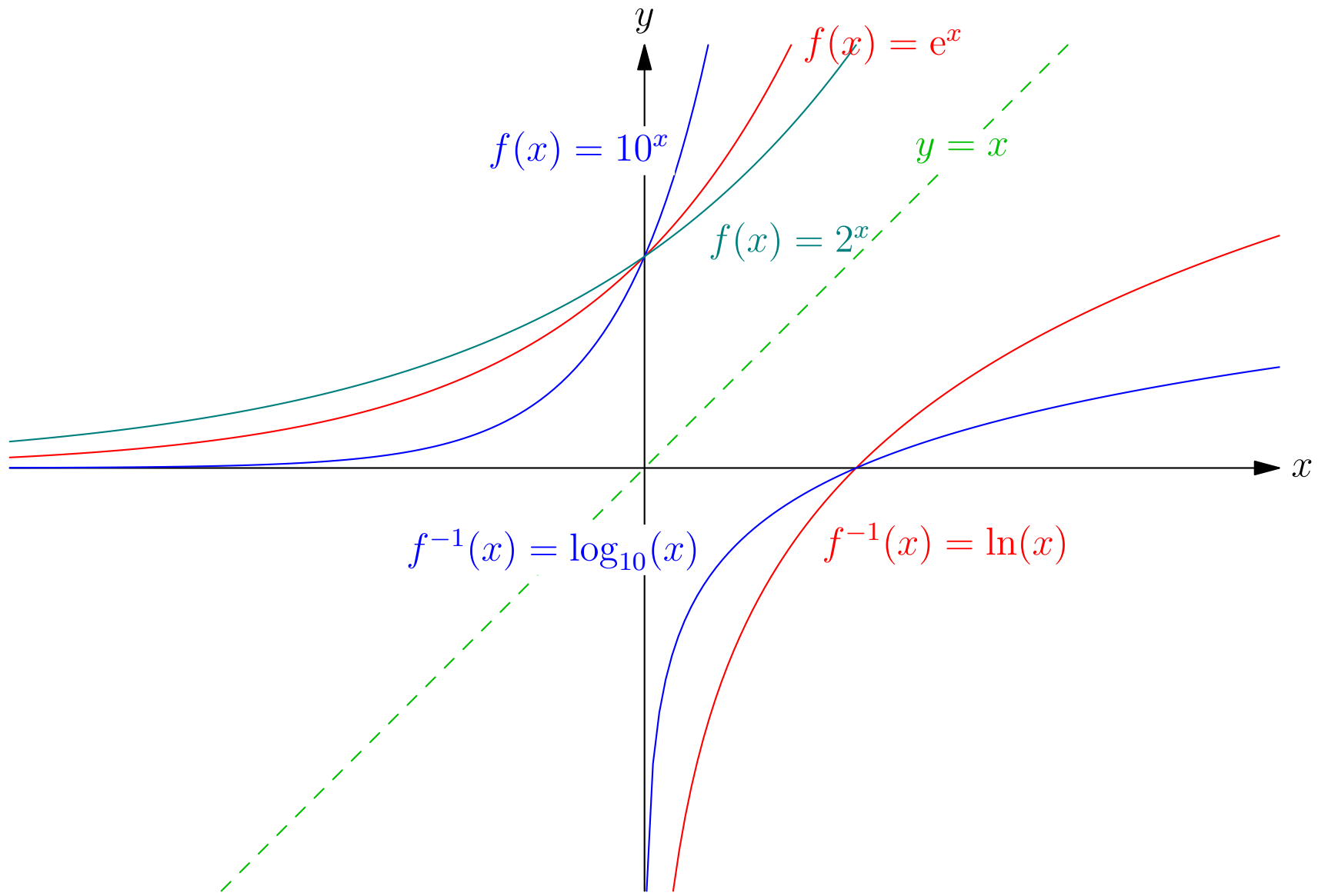
Exponentials and Logarithms



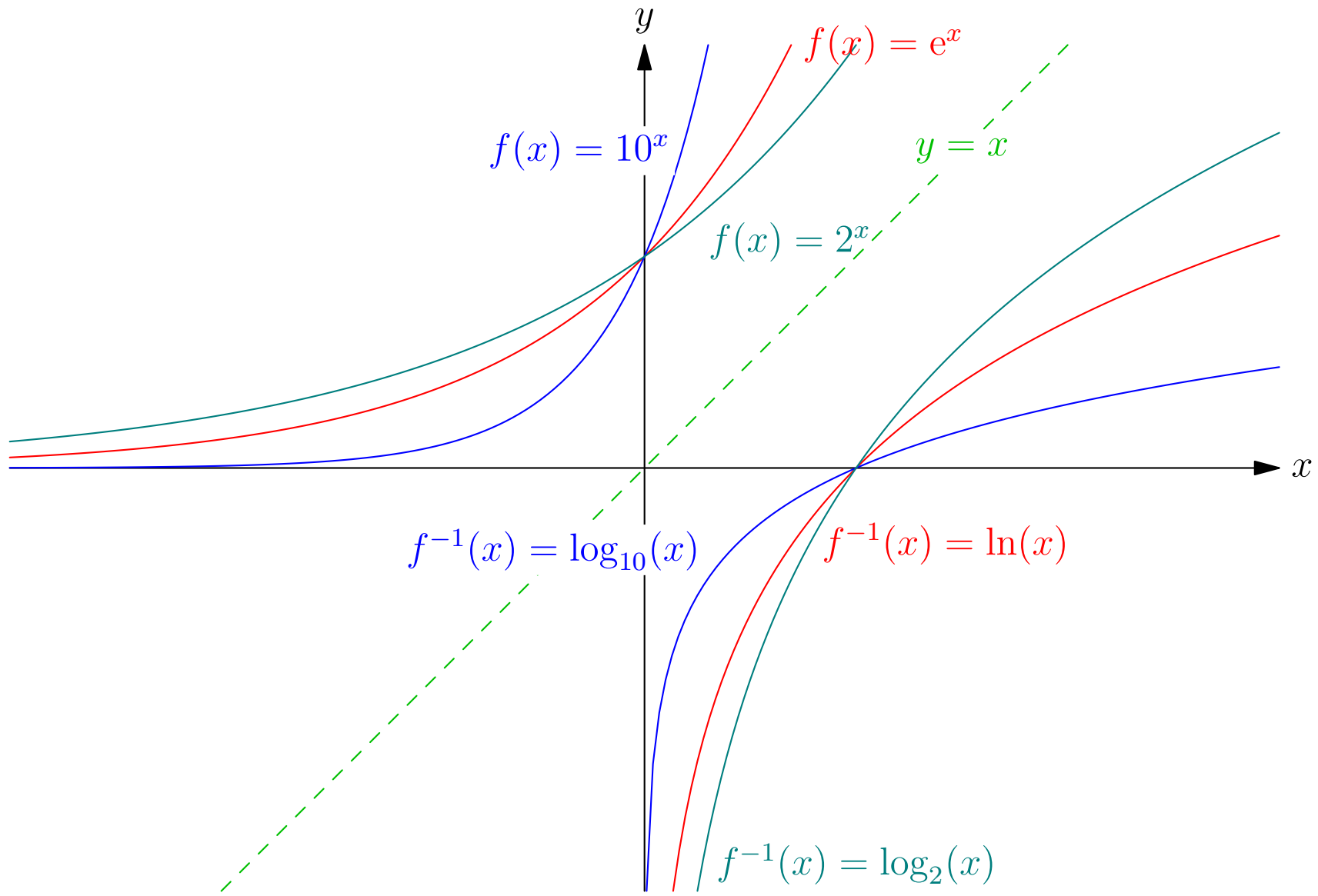
Exponentials and Logarithms



Exponentials and Logarithms

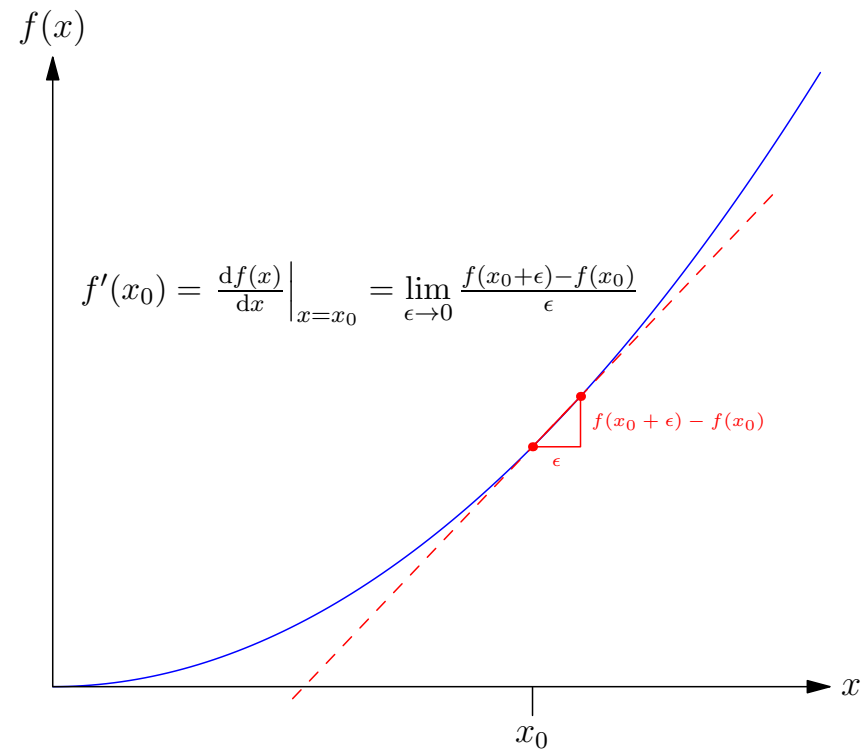


Exponentials and Logarithms



Outline

1. Why Calculus?
2. Differentiation
3. **Vector and Matrix Calculus**



Derivatives in High Dimensions

- When working with functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ in many dimensions then there will typically be different derivative in different directions
- To compute the derivative in a direction $\mathbf{u} \in \mathbb{R}^n$ (where $\|\mathbf{u}\| = 1$) at a point $\mathbf{x} \in \mathbb{R}^n$ we use

$$\partial_{\mathbf{u}} F(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{x} + \epsilon \mathbf{u}) - f(\mathbf{x})}{\epsilon}$$

- If $\mathbf{u} = \delta_i = (0, \dots, 0, 1, 0, \dots, 0)$ (i.e. $u_i = 1$) then

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{x} + \epsilon \delta_i) - f(\mathbf{x})}{\epsilon}$$

Derivatives in High Dimensions

- When working with functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ in many dimensions then there will typically be different derivative in different directions
- To compute the derivative in a direction $\mathbf{u} \in \mathbb{R}^n$ (where $\|\mathbf{u}\| = 1$) at a point $\mathbf{x} \in \mathbb{R}^n$ we use

$$\partial_{\mathbf{u}} F(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{x} + \epsilon \mathbf{u}) - f(\mathbf{x})}{\epsilon}$$

- If $\mathbf{u} = \delta_i = (0, \dots, 0, 1, 0, \dots, 0)$ (i.e. $u_i = 1$) then

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{x} + \epsilon \delta_i) - f(\mathbf{x})}{\epsilon}$$

Derivatives in High Dimensions

- When working with functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ in many dimensions then there will typically be different derivative in different directions
- To compute the derivative in a direction $\mathbf{u} \in \mathbb{R}^n$ (where $\|\mathbf{u}\| = 1$) at a point $\mathbf{x} \in \mathbb{R}^n$ we use

$$\partial_{\mathbf{u}} F(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{x} + \epsilon \mathbf{u}) - f(\mathbf{x})}{\epsilon}$$

- If $\mathbf{u} = \delta_i = (0, \dots, 0, 1, 0, \dots, 0)$ (i.e. $u_i = 1$) then

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{x} + \epsilon \delta_i) - f(\mathbf{x})}{\epsilon}$$

Taylor

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

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

then $g_i(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_i}$

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

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

- Thus

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

Taylor

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

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

then $g_i(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_i}$

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

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

- Thus

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

Taylor

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

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

then $g_i(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_i}$

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

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

- Thus

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

Taylor

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

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

then $g_i(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_i}$

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

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

- Thus

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

This is the start of the high-dimensional Taylor expansion

Computing Gradients 1

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

$$\nabla w^\top M w$$

Computing Gradients 1

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

$$\nabla w^T M w$$

Computing Gradients 1

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

$$\nabla \mathbf{w}^\top \mathbf{M} \mathbf{w} = \begin{pmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \frac{\partial}{\partial w_3} \\ \vdots \end{pmatrix} \sum_{i,j} w_i M_{ij} w_j$$

Computing Gradients 1

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

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

Computing Gradients 1

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

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

Computing Gradients 1

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

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

- 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 $f(\mathbf{x} + \epsilon \mathbf{u})$
- Consider $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{M} \mathbf{x} + \mathbf{a}^\top \mathbf{x}$

$$f(\mathbf{x} + \epsilon \mathbf{u}) = (\mathbf{x} + \epsilon \mathbf{u})^\top \mathbf{M} (\mathbf{x} + \epsilon \mathbf{u}) + \mathbf{a}^\top (\mathbf{x} + \epsilon \mathbf{u})$$

Computing Gradients 2

- A slicker way is just to expand $f(\mathbf{x} + \epsilon \mathbf{u})$
- Consider $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{M} \mathbf{x} + \mathbf{a}^\top \mathbf{x}$

$$f(\mathbf{x} + \epsilon \mathbf{u}) = (\mathbf{x} + \epsilon \mathbf{u})^\top \mathbf{M} (\mathbf{x} + \epsilon \mathbf{u}) + \mathbf{a}^\top (\mathbf{x} + \epsilon \mathbf{u})$$

Computing Gradients 2

- A slicker way is just to expand $f(\mathbf{x} + \epsilon \mathbf{u})$
- Consider $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{M} \mathbf{x} + \mathbf{a}^\top \mathbf{x}$

$$\begin{aligned} f(\mathbf{x} + \epsilon \mathbf{u}) &= (\mathbf{x} + \epsilon \mathbf{u})^\top \mathbf{M} (\mathbf{x} + \epsilon \mathbf{u}) + \mathbf{a}^\top (\mathbf{x} + \epsilon \mathbf{u}) \\ &= f(\mathbf{x}) + \epsilon (\mathbf{u}^\top \mathbf{M} \mathbf{x} + \mathbf{x}^\top \mathbf{M} \mathbf{u} + \mathbf{a}^\top \mathbf{u}) + O(\epsilon^2) \end{aligned}$$

Computing Gradients 2

- A slicker way is just to expand $f(x + \epsilon u)$
- Consider $f(x) = x^\top M x + a^\top x$

$$\begin{aligned} f(x + \epsilon u) &= (x + \epsilon u)^\top M (x + \epsilon u) + a^\top (x + \epsilon u) \\ &= f(x) + \epsilon (u^\top M x + x^\top M u + a^\top u) + O(\epsilon^2) \\ &= f(x) + \epsilon u^\top (M x + M^\top x + a) + O(\epsilon^2) \end{aligned}$$

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

Computing Gradients 2

- A slicker way is just to expand $f(\mathbf{x} + \epsilon \mathbf{u})$
- Consider $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{M} \mathbf{x} + \mathbf{a}^\top \mathbf{x}$

$$\begin{aligned} f(\mathbf{x} + \epsilon \mathbf{u}) &= (\mathbf{x} + \epsilon \mathbf{u})^\top \mathbf{M} (\mathbf{x} + \epsilon \mathbf{u}) + \mathbf{a}^\top (\mathbf{x} + \epsilon \mathbf{u}) \\ &= f(\mathbf{x}) + \epsilon (\mathbf{u}^\top \mathbf{M} \mathbf{x} + \mathbf{x}^\top \mathbf{M} \mathbf{u} + \mathbf{a}^\top \mathbf{u}) + O(\epsilon^2) \\ &= f(\mathbf{x}) + \epsilon \mathbf{u}^\top (\mathbf{M} \mathbf{x} + \mathbf{M}^\top \mathbf{x} + \mathbf{a}) + O(\epsilon^2) \end{aligned}$$

using $\mathbf{x}^\top \mathbf{M} \mathbf{u} = \mathbf{u}^\top \mathbf{M}^\top \mathbf{x}$ and $\mathbf{a}^\top \mathbf{u} = \mathbf{u}^\top \mathbf{a}$

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

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

Differentiating Matrices

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

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

- We might want to find the minimum with respect to \mathbf{W}
- This occurs at a point \mathbf{W}^* where $L(\mathbf{W})$ does not increase as we change \mathbf{W} in any way
- That is, we seek a \mathbf{W}^* such that, for any matrices \mathbf{U}

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

Differentiating Matrices

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

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

- We might want to find the minimum with respect to \mathbf{W}
- This occurs at a point \mathbf{W}^* where $L(\mathbf{W})$ does not increase as we change \mathbf{W} in any way
- That is, we seek a \mathbf{W}^* such that, for any matrices \mathbf{U}

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

Differentiating Matrices

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

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

- We might want to find the minimum with respect to \mathbf{W}
- This occurs at a point \mathbf{W}^* where $L(\mathbf{W})$ does not increase as we change \mathbf{W} in any way
- That is, we seek a \mathbf{W}^* such that, for any matrices \mathbf{U}

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

Differentiating Matrices

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

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

- We might want to find the minimum with respect to \mathbf{W}
- This occurs at a point \mathbf{W}^* where $L(\mathbf{W})$ does not increase as we change \mathbf{W} in any way
- That is, we seek a \mathbf{W}^* such that, for any matrices \mathbf{U}

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

Generalised Gradient

- We can generalise the idea of gradient to matrices

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

- From an identical argument we used for vectors

$$L(\mathbf{W} + \epsilon \mathbf{U}) = L(\mathbf{W}) + \epsilon \text{tr} \mathbf{U}^\top \frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} + O(\epsilon^2)$$

where

$$\text{tr} \mathbf{U}^\top \mathbf{G} = \sum_i [\mathbf{U}^\top \mathbf{G}]_{ii} = \sum_{ij} U_{ji} G_{ji} = \sum_{ij} U_{ij} G_{ij} = \langle \mathbf{U}, \mathbf{G} \rangle$$

Generalised Gradient

- We can generalise the idea of gradient to matrices

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

- From an identical argument we used for vectors

$$L(\mathbf{W} + \epsilon \mathbf{U}) = L(\mathbf{W}) + \epsilon \text{tr} \mathbf{U}^\top \frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} + O(\epsilon^2)$$

where

$$\text{tr} \mathbf{U}^\top \mathbf{G} = \sum_i [\mathbf{U}^\top \mathbf{G}]_{ii} = \sum_{ij} U_{ji} G_{ji} = \sum_{ij} U_{ij} G_{ij} = \langle \mathbf{U}, \mathbf{G} \rangle$$

Generalised Gradient

- We can generalise the idea of gradient to matrices

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

- From an identical argument we used for vectors

$$L(\mathbf{W} + \epsilon \mathbf{U}) = L(\mathbf{W}) + \epsilon \text{tr} \mathbf{U}^\top \frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} + O(\epsilon^2)$$

where

$$\text{tr} \mathbf{U}^\top \mathbf{G} = \sum_i [\mathbf{U}^\top \mathbf{G}]_{ii} = \sum_{ij} U_{ji} G_{ji} = \sum_{ij} U_{ij} G_{ij} = \langle \mathbf{U}, \mathbf{G} \rangle$$

Example

- Suppose

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

then

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

Example

- Suppose

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

then

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

Example

- Suppose

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

then

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

Example

- Suppose

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

then

$$\begin{aligned} L(\mathbf{W} + \epsilon \mathbf{U}) &= (\mathbf{a}^\top (\mathbf{W} + \epsilon \mathbf{U}) \mathbf{b} - c)^2 = (\mathbf{a}^\top \mathbf{W} \mathbf{b} + \epsilon \mathbf{a}^\top \mathbf{U} \mathbf{b} - c)^2 \\ &= L(\mathbf{W}) + 2\epsilon (\mathbf{a}^\top \mathbf{W} \mathbf{b} - c) (\mathbf{a}^\top \mathbf{U} \mathbf{b}) + O(\epsilon^2) \end{aligned}$$

Example

- Suppose

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

then

$$\begin{aligned} L(\mathbf{W} + \epsilon \mathbf{U}) &= (\mathbf{a}^\top (\mathbf{W} + \epsilon \mathbf{U}) \mathbf{b} - c)^2 = (\mathbf{a}^\top \mathbf{W} \mathbf{b} + \epsilon \mathbf{a}^\top \mathbf{U} \mathbf{b} - c)^2 \\ &= L(\mathbf{W}) + 2\epsilon (\mathbf{a}^\top \mathbf{W} \mathbf{b} - c) (\mathbf{a}^\top \mathbf{U} \mathbf{b}) + O(\epsilon^2) \end{aligned}$$

- Now

$$\mathbf{a}^\top \mathbf{U} \mathbf{b} = \sum_{ij} a_i U_{ij} b_j$$

Example

- Suppose

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

then

$$\begin{aligned} L(\mathbf{W} + \epsilon \mathbf{U}) &= (\mathbf{a}^\top (\mathbf{W} + \epsilon \mathbf{U}) \mathbf{b} - c)^2 = (\mathbf{a}^\top \mathbf{W} \mathbf{b} + \epsilon \mathbf{a}^\top \mathbf{U} \mathbf{b} - c)^2 \\ &= L(\mathbf{W}) + 2\epsilon (\mathbf{a}^\top \mathbf{W} \mathbf{b} - c) (\mathbf{a}^\top \mathbf{U} \mathbf{b}) + O(\epsilon^2) \end{aligned}$$

- Now

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

Example

- Suppose

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

then

$$\begin{aligned} L(\mathbf{W} + \epsilon \mathbf{U}) &= (\mathbf{a}^\top (\mathbf{W} + \epsilon \mathbf{U}) \mathbf{b} - c)^2 = (\mathbf{a}^\top \mathbf{W} \mathbf{b} + \epsilon \mathbf{a}^\top \mathbf{U} \mathbf{b} - c)^2 \\ &= L(\mathbf{W}) + 2\epsilon (\mathbf{a}^\top \mathbf{W} \mathbf{b} - c) (\mathbf{a}^\top \mathbf{U} \mathbf{b}) + O(\epsilon^2) \end{aligned}$$

- Now

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

Example

- Suppose

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

then

$$\begin{aligned} L(\mathbf{W} + \epsilon \mathbf{U}) &= (\mathbf{a}^\top (\mathbf{W} + \epsilon \mathbf{U}) \mathbf{b} - c)^2 = (\mathbf{a}^\top \mathbf{W} \mathbf{b} + \epsilon \mathbf{a}^\top \mathbf{U} \mathbf{b} - c)^2 \\ &= L(\mathbf{W}) + 2\epsilon (\mathbf{a}^\top \mathbf{W} \mathbf{b} - c) (\mathbf{a}^\top \mathbf{U} \mathbf{b}) + O(\epsilon^2) \end{aligned}$$

- Now

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

$$\text{Thus } \frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} = 2 (\mathbf{a}^\top \mathbf{W} \mathbf{b} - c) \mathbf{a} \mathbf{b}^\top$$

Traces

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

$$\text{tr} \mathbf{A} = \text{tr} \mathbf{A}^T = \sum_i A_{ii}$$

- Clearly $\text{tr} c\mathbf{A} = c\text{tr} \mathbf{A}$
- Also $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr} \mathbf{A} + \text{tr} \mathbf{B}$
- We note that

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

- It follows that

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

Traces

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

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

- Clearly $\text{trc}\mathbf{A} = \text{ctr}\mathbf{A}$
- Also $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}\mathbf{A} + \text{tr}\mathbf{B}$
- We note that

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

- It follows that

$$\text{tr}\mathbf{ABCD} = \text{tr}\mathbf{DABC}$$

Traces

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

$$\text{tr}\mathbf{A} = \text{tr}\mathbf{A}^T = \sum_i A_{ii}$$

- Clearly $\text{tr}c\mathbf{A} = c\text{tr}\mathbf{A}$
- Also $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}\mathbf{A} + \text{tr}\mathbf{B}$
- We note that

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

- It follows that

$$\text{tr}\mathbf{ABCD} = \text{tr}\mathbf{DABC}$$

Traces

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

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

- Clearly $\text{tr}c\mathbf{A} = c\text{tr}\mathbf{A}$
- Also $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}\mathbf{A} + \text{tr}\mathbf{B}$
- We note that

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

- It follows that

$$\text{tr}\mathbf{ABCD} = \text{tr}\mathbf{DABC}$$

Traces

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

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

- Clearly $\text{tr}c\mathbf{A} = c\text{tr}\mathbf{A}$
- Also $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}\mathbf{A} + \text{tr}\mathbf{B}$
- We note that

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

- It follows that

$$\text{tr}\mathbf{ABCD} = \text{tr}\mathbf{DABC}$$

Traces

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

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

- Clearly $\text{tr}c\mathbf{A} = c\text{tr}\mathbf{A}$
- Also $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}\mathbf{A} + \text{tr}\mathbf{B}$
- We note that

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

- It follows that

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

Traces

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

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

- Clearly $\text{tr}c\mathbf{A} = c\text{tr}\mathbf{A}$
- Also $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}\mathbf{A} + \text{tr}\mathbf{B}$
- We note that

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

- It follows that

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

Quick Matrix Differentiation

- Let

$$\partial_{\mathbf{U}} f(\mathbf{X}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{X} + \epsilon \mathbf{U}) - f(\mathbf{X})}{\epsilon}$$

- E.g.

$$\begin{aligned} \partial_{\mathbf{U}} \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \text{tr} \mathbf{A} (\mathbf{X} + \epsilon \mathbf{U}) \mathbf{B} - \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} \\ &= \text{tr} \mathbf{A} \mathbf{U} \mathbf{B} \end{aligned}$$

Quick Matrix Differentiation

- Let

$$\partial_{\mathbf{U}} f(\mathbf{X}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{X} + \epsilon \mathbf{U}) - f(\mathbf{X})}{\epsilon} = \text{tr} \mathbf{U}^T \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$$

- E.g.

$$\begin{aligned} \partial_{\mathbf{U}} \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \text{tr} \mathbf{A} (\mathbf{X} + \epsilon \mathbf{U}) \mathbf{B} - \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} \\ &= \text{tr} \mathbf{A} \mathbf{U} \mathbf{B} \end{aligned}$$

Quick Matrix Differentiation

- Let

$$\partial_{\mathbf{U}} f(\mathbf{X}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{X} + \epsilon \mathbf{U}) - f(\mathbf{X})}{\epsilon} = \text{tr} \mathbf{U}^T \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$$

- E.g.

$$\begin{aligned} \partial_{\mathbf{U}} \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \text{tr} \mathbf{A} (\mathbf{X} + \epsilon \mathbf{U}) \mathbf{B} - \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} \\ &= \text{tr} \mathbf{A} \mathbf{U} \mathbf{B} \end{aligned}$$

Quick Matrix Differentiation

- Let

$$\partial_{\mathbf{U}} f(\mathbf{X}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{X} + \epsilon \mathbf{U}) - f(\mathbf{X})}{\epsilon} = \text{tr} \mathbf{U}^T \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$$

- E.g.

$$\begin{aligned} \partial_{\mathbf{U}} \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \text{tr} \mathbf{A} (\mathbf{X} + \epsilon \mathbf{U}) \mathbf{B} - \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} \\ &= \text{tr} \mathbf{A} \mathbf{U} \mathbf{B} = \text{tr} \mathbf{B}^T \mathbf{U}^T \mathbf{A}^T \end{aligned}$$

Quick Matrix Differentiation

- Let

$$\partial_{\mathbf{U}} f(\mathbf{X}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{X} + \epsilon \mathbf{U}) - f(\mathbf{X})}{\epsilon} = \text{tr} \mathbf{U}^T \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$$

- E.g.

$$\begin{aligned} \partial_{\mathbf{U}} \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \text{tr} \mathbf{A} (\mathbf{X} + \epsilon \mathbf{U}) \mathbf{B} - \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} \\ &= \text{tr} \mathbf{A} \mathbf{U} \mathbf{B} = \text{tr} \mathbf{B}^T \mathbf{U}^T \mathbf{A}^T = \text{tr} \mathbf{U}^T \mathbf{A}^T \mathbf{B}^T \end{aligned}$$

Quick Matrix Differentiation

- Let

$$\partial_{\mathbf{U}} f(\mathbf{X}) = \lim_{\epsilon \rightarrow 0} \frac{f(\mathbf{X} + \epsilon \mathbf{U}) - f(\mathbf{X})}{\epsilon} = \text{tr} \mathbf{U}^T \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$$

- E.g.

$$\begin{aligned} \partial_{\mathbf{U}} \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \text{tr} \mathbf{A} (\mathbf{X} + \epsilon \mathbf{U}) \mathbf{B} - \text{tr} \mathbf{A} \mathbf{X} \mathbf{B} \\ &= \text{tr} \mathbf{A} \mathbf{U} \mathbf{B} = \text{tr} \mathbf{B}^T \mathbf{U}^T \mathbf{A}^T = \text{tr} \mathbf{U}^T \mathbf{A}^T \mathbf{B}^T \end{aligned}$$

thus

$$\frac{\partial \text{tr} \mathbf{A} \mathbf{X} \mathbf{B}}{\partial \mathbf{X}} = \mathbf{A}^T \mathbf{B}^T$$

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, $\log(|\mathbf{K} + \sigma^2 \mathbf{I}|)$
- To find the derivative of $\log(|\mathbf{X}|)$ we consider

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

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, $\log(|\mathbf{K} + \sigma^2 \mathbf{I}|)$
- To find the derivative of $\log(|\mathbf{X}|)$ we consider

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

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, $\log(|\mathbf{K} + \sigma^2 \mathbf{I}|)$
- To find the derivative of $\log(|\mathbf{X}|)$ we consider

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

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, $\log(|\mathbf{K} + \sigma^2 \mathbf{I}|)$
- To find the derivative of $\log(|\mathbf{X}|)$ we consider

$$\begin{aligned}\log(|\mathbf{X} + \epsilon \mathbf{U}|) &= \log(|\mathbf{X}(\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U})|) \\ &= \log(|\mathbf{X}| |\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U}|)\end{aligned}$$

★ Using $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$

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, $\log(|\mathbf{K} + \sigma^2 \mathbf{I}|)$
- To find the derivative of $\log(|\mathbf{X}|)$ we consider

$$\begin{aligned}\log(|\mathbf{X} + \epsilon \mathbf{U}|) &= \log(|\mathbf{X}(\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U})|) \\ &= \log(|\mathbf{X}| |\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U}|) \\ &= \log(|\mathbf{X}|) + \log(|\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U}|)\end{aligned}$$

- ★ Using $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$
- ★ Using $\log(ab) = \log(a) + \log(b)$

Determinants

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

Determinants

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

Determinants

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

Determinants

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

Determinants

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

$$|\mathbf{I} + \epsilon \mathbf{M}| = \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{21} & \epsilon M_{31} & \epsilon M_{41} & \epsilon M_{51} \\ \epsilon M_{12} & 1 + \epsilon M_{22} & \epsilon M_{32} & \epsilon M_{42} & \epsilon M_{52} \\ \epsilon M_{13} & \epsilon M_{23} & 1 + \epsilon M_{33} & \epsilon M_{43} & \epsilon M_{53} \\ \epsilon M_{14} & \epsilon M_{24} & \epsilon M_{34} & 1 + \epsilon M_{44} & \epsilon M_{54} \\ \epsilon M_{15} & \epsilon M_{25} & \epsilon M_{35} & \epsilon M_{45} & 1 + \epsilon M_{55} \end{vmatrix}$$

Determinants

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

$$|\mathbf{I} + \epsilon \mathbf{M}| = \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\ = (1 + \epsilon M_{11}) C_{11}$$

Determinants

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

$$|\mathbf{I} + \epsilon \mathbf{M}| = \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\ = (1 + \epsilon M_{11}) C_{11} - \epsilon M_{21} C_{21}$$

Determinants

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

$$|\mathbf{I} + \epsilon \mathbf{M}| = \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\ = (1 + \epsilon M_{11}) C_{11} - \epsilon M_{21} C_{21} + \epsilon M_{31} C_{31}$$

Determinants

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

$$|\mathbf{I} + \epsilon \mathbf{M}| = \begin{vmatrix} 1 & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\ = (1 + \epsilon M_{11}) C_{11} - \epsilon M_{21} C_{21} + \epsilon M_{31} C_{31} - \epsilon M_{41} C_{41}$$

Determinants

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

$$|\mathbf{I} + \epsilon \mathbf{M}| = \begin{vmatrix} 1 & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\ = (1 + \epsilon M_{11}) C_{11} - \epsilon M_{21} C_{21} + \epsilon M_{31} C_{31} - \epsilon M_{41} C_{41} + \epsilon M_{51} C_{51}$$

Determinants

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

$$\begin{aligned}
 |\mathbf{I} + \epsilon \mathbf{M}| &= \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\
 &= (1 + \epsilon M_{11}) C_{11} + O(\epsilon^2)
 \end{aligned}$$

Determinants

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

$$|\mathbf{I} + \epsilon \mathbf{M}| = \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\ = (1 + \epsilon M_{11}) C_{11} + O(\epsilon^2)$$

Determinants

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

$$|\mathbf{I} + \epsilon \mathbf{M}| = \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\ = (1 + \epsilon M_{11})(1 + \epsilon M_{22}) C'_{22} + O(\epsilon^2)$$

Determinants

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

$$\begin{aligned}
 |\mathbf{I} + \epsilon \mathbf{M}| &= \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\
 &= \prod_i (1 + \epsilon M_{ii}) + O(\epsilon^2)
 \end{aligned}$$

Determinants

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

$$\begin{aligned}
 |\mathbf{I} + \epsilon \mathbf{M}| &= \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{12} & \epsilon M_{13} & \epsilon M_{14} & \epsilon M_{15} \\ \epsilon M_{21} & 1 + \epsilon M_{22} & \epsilon M_{23} & \epsilon M_{24} & \epsilon M_{25} \\ \epsilon M_{31} & \epsilon M_{32} & 1 + \epsilon M_{33} & \epsilon M_{34} & \epsilon M_{35} \\ \epsilon M_{41} & \epsilon M_{42} & \epsilon M_{43} & 1 + \epsilon M_{44} & \epsilon M_{45} \\ \epsilon M_{51} & \epsilon M_{52} & \epsilon M_{53} & \epsilon M_{54} & 1 + \epsilon M_{55} \end{vmatrix} \\
 &= (1 + \epsilon \sum_i M_{ii}) + O(\epsilon^2)
 \end{aligned}$$

Determinants

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

$$\begin{aligned}
 |\mathbf{I} + \epsilon \mathbf{M}| &= \begin{vmatrix} 1 + \epsilon M_{11} & \epsilon M_{21} & \epsilon M_{31} & \epsilon M_{41} & \epsilon M_{51} \\ \epsilon M_{12} & 1 + \epsilon M_{22} & \epsilon M_{32} & \epsilon M_{42} & \epsilon M_{52} \\ \epsilon M_{13} & \epsilon M_{23} & 1 + \epsilon M_{33} & \epsilon M_{43} & \epsilon M_{53} \\ \epsilon M_{14} & \epsilon M_{24} & \epsilon M_{34} & 1 + \epsilon M_{44} & \epsilon M_{54} \\ \epsilon M_{15} & \epsilon M_{25} & \epsilon M_{35} & \epsilon M_{45} & 1 + \epsilon M_{55} \end{vmatrix} \\
 &= (1 + \epsilon \operatorname{tr} \mathbf{M}) + O(\epsilon^2)
 \end{aligned}$$

Putting it Together

- Recall

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

Putting it Together

- Recall

$$\begin{aligned}\log(|\mathbf{X} + \epsilon \mathbf{U}|) - \log(|\mathbf{X}|) &= \log(|\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U}|) \\ &= \log(1 + \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2)\end{aligned}$$

Putting it Together

- Recall

$$\begin{aligned}\log(|\mathbf{X} + \epsilon \mathbf{U}|) - \log(|\mathbf{X}|) &= \log(|\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U}|) \\ &= \log(1 + \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2) \\ &= \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2\end{aligned}$$

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

Putting it Together

- Recall

$$\begin{aligned}\log(|\mathbf{X} + \epsilon \mathbf{U}|) - \log(|\mathbf{X}|) &= \log(|\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U}|) \\ &= \log(1 + \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2) \\ &= \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2 \\ &= \epsilon \operatorname{tr} \mathbf{U}^{\top} (\mathbf{X}^{-1})^{\top} + O(\epsilon)\end{aligned}$$

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

Putting it Together

- Recall

$$\begin{aligned}\log(|\mathbf{X} + \epsilon \mathbf{U}|) - \log(|\mathbf{X}|) &= \log(|\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U}|) \\ &= \log(1 + \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2) \\ &= \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2 \\ &= \epsilon \operatorname{tr} \mathbf{U}^T (\mathbf{X}^{-1})^T + O(\epsilon)\end{aligned}$$

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

- Thus $\partial_{\mathbf{U}} \log(|\mathbf{X}|) = \operatorname{tr} \mathbf{U}^T (\mathbf{X}^{-1})^T$

Putting it Together

- Recall

$$\begin{aligned}\log(|\mathbf{X} + \epsilon \mathbf{U}|) - \log(|\mathbf{X}|) &= \log(|\mathbf{I} + \epsilon \mathbf{X}^{-1} \mathbf{U}|) \\ &= \log(1 + \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2) \\ &= \epsilon \operatorname{tr} \mathbf{X}^{-1} \mathbf{U} + O(\epsilon)^2 \\ &= \epsilon \operatorname{tr} \mathbf{U}^T (\mathbf{X}^{-1})^T + O(\epsilon)\end{aligned}$$

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

- Thus $\partial_{\mathbf{U}} \log(|\mathbf{X}|) = \operatorname{tr} \mathbf{U}^T (\mathbf{X}^{-1})^T$

- Or

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

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

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

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

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

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**

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