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

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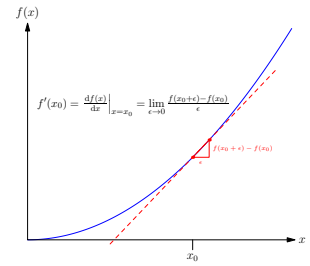

Back to Basics

- 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

Differentiation

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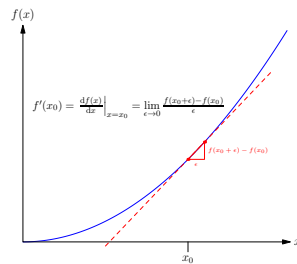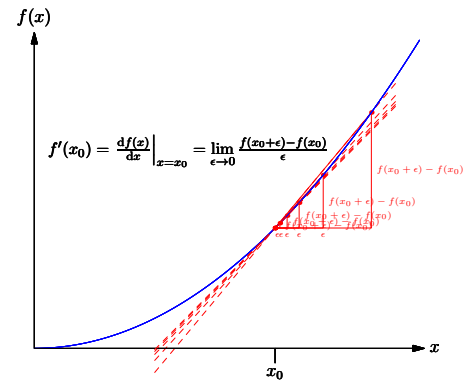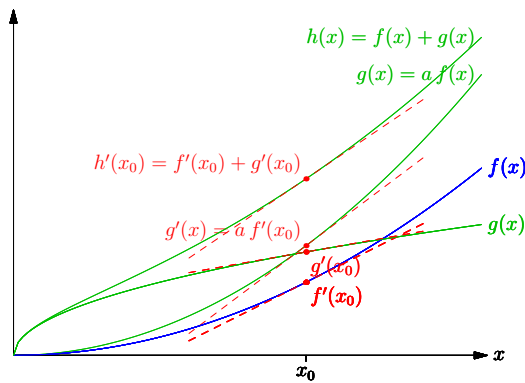

Polynomials

- $f(x) = x^2$

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

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

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

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

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

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

* 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

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

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

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

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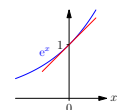
* But $e^{x+\epsilon} = e^x e^\epsilon = e^x(1 + \epsilon + O(\epsilon^2)) = e^x + \epsilon e^x + O(\epsilon^2)$

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

Functions of Exponentials

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

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

* More generally using the chain rule

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

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

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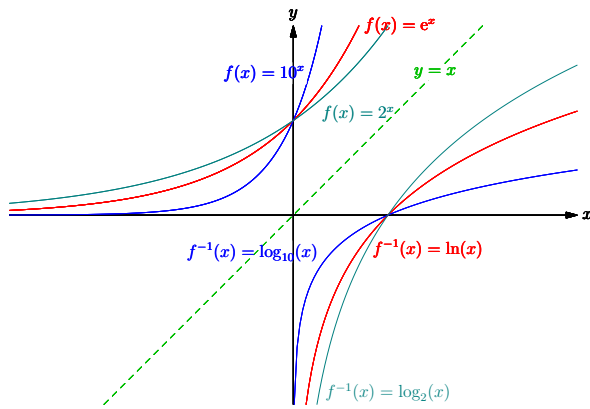$$

Natural Logarithms

* The natural logarithm is defined as the inverse of e^x
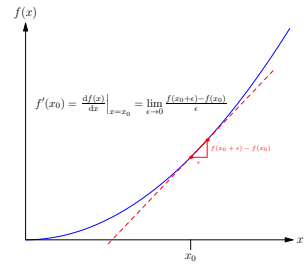
$$
\ln(e^x) = x \qquad\qquad e^{\ln(y)} = y
$$

* 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 $f'(x) = e^x$)

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

Exponentials and Logarithms

Outline

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

Derivatives in High Dimensions

- When working with functions $f : \mathbb{R}^n \to \mathbb{R}$ 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 $g_i(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial x_i}$

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

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

- 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 $f(\boldsymbol{x} + \epsilon \boldsymbol{u})$

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

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

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

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

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

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

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

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 $\mathbf{U}$

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

Generalised Gradient

- We can generalise the idea of gradient to matrices

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

- From an identical argument we used for vectors

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

where

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

Example

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

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

- Now
$$\boldsymbol{a}^\mathsf{T}\mathbf{U}\boldsymbol{b} = \sum_{ij} a_i U_{ij} b_j = \sum_{ij} U_{ji} a_j b_i = \operatorname{tr}\mathbf{U}^\mathsf{T}\boldsymbol{a}\boldsymbol{b}^\mathsf{T}$$

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

Traces

- The trace of a matrix is the sum of its diagonal elements
$$\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{A}^\mathsf{T} = \sum_i A_{ii}$$

- Clearly $\operatorname{tr}c\mathbf{A} = c\operatorname{tr}\mathbf{A}$

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

- We note that
$$\operatorname{tr}\mathbf{A}\mathbf{B} = \sum_{i,j} A_{ij}B_{ji} = \sum_{i,j} B_{ij}A_{ji} = \operatorname{tr}\mathbf{B}\mathbf{A}$$

- It follows that
$$\operatorname{tr}\mathbf{A}\mathbf{B}\mathbf{C}\mathbf{D} = \operatorname{tr}\mathbf{D}\mathbf{A}\mathbf{B}\mathbf{C} = \operatorname{tr}\mathbf{C}\mathbf{D}\mathbf{A}\mathbf{B} = \operatorname{tr}\mathbf{B}\mathbf{C}\mathbf{D}\mathbf{A}$$

Quick Matrix Differentiation

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

- E.g.
$$\partial_{\mathbf{U}}\operatorname{tr}\mathbf{A}\mathbf{X}\mathbf{B} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\operatorname{tr}\mathbf{A}\left(\mathbf{X} + \epsilon\mathbf{U}\right)\mathbf{B} - \operatorname{tr}\mathbf{A}\mathbf{X}\mathbf{B}$$
$$= \operatorname{tr}\mathbf{A}\mathbf{U}\mathbf{B} = \operatorname{tr}\mathbf{B}^\mathsf{T}\mathbf{U}^\mathsf{T}\mathbf{A}^\mathsf{T} = \operatorname{tr}\mathbf{U}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{B}^\mathsf{T}$$

  thus
$$\frac{\partial\operatorname{tr}\mathbf{A}\mathbf{X}\mathbf{B}}{\partial\mathbf{X}} = \mathbf{A}^\mathsf{T}\mathbf{B}^\mathsf{T}$$

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

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

Determinants

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

$$|\mathbf{I} + \epsilon\mathbf{M}| = (1 + \epsilon M_{11})C_{11} + \epsilon M_{21}C_{21} + \epsilon M_{31}C_{31} - \epsilon M_{41}C_{41} + \epsilon M_{51}C_{51}$$

Putting it Together

- Recall
$$\log(|\mathbf{X} + \epsilon\mathbf{U}|) - \log(|\mathbf{X}|) = \log\left(|\mathbf{I} + \epsilon\mathbf{X}^{-1}\mathbf{U}|\right)$$
$$= \log\left(1 + \epsilon\operatorname{tr}\mathbf{X}^{-1}\mathbf{U} + O(\epsilon^2)\right)$$
$$= \epsilon\operatorname{tr}\mathbf{X}^{-1}\mathbf{U} + O(\epsilon)^2$$
$$= \epsilon\operatorname{tr}\mathbf{U}^\mathsf{T}\left(\mathbf{X}^{-1}\right)^\mathsf{T} + O(\epsilon)$$

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

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

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

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