Chain rule

* Chain Rule

$$\frac{d}{dx}[f(g(x))] = \underbrace{f'(g(x))} \cdot g'(x)$$

$$\frac{d}{dx}\left[\underbrace{\ln(\overbrace{\sin(x)}^{g(x)})}_{f(g(x))}\right]$$

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

$$\frac{d}{dx}\left[\overbrace{\ln(x)}^{f(x)}\overbrace{\sin(x)}^{g(x)}\right]$$

Product Rule

Gradient

- Assuming our input is a row vector, that is $\boldsymbol{x} \in \mathbb{R}^{1\times M}$
- The gradient is a vector containing all partial derivatives

$$\frac{dh}{d\boldsymbol{x}} = \nabla_{\boldsymbol{x}} h = [\frac{\partial h}{\partial x_1}, \ldots, \frac{\partial h}{\partial x_M}]$$

- Generalization of the derivative, defined on a univariate function ($M = 1$)

Example

- Often, easier to write things out explicitly
- Let's say $y = \sin(\boldsymbol{x})$, where $\dim(\boldsymbol{x}) = 1 \times M$

$$y = \sin(\boldsymbol{x}), \quad x \in R^{1\times M}$$

$$\frac{dy}{\partial x_j} = \frac{\partial}{\partial x_j}\sin(x_j) \Rightarrow \frac{\partial y}{\partial \boldsymbol{x}} = [\cos(x_1), \ldots, \cos(x_j), \ldots, \cos(x_M)]$$

Jacobian

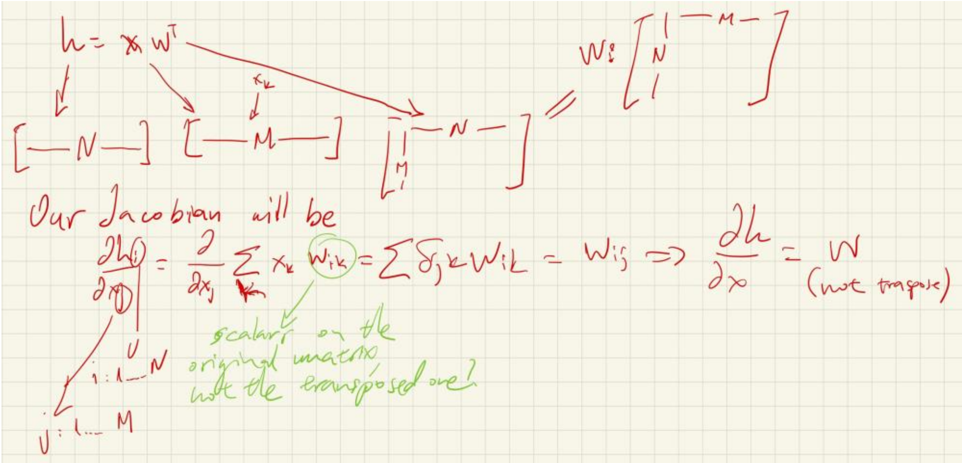
- Generalization of the gradient for vector-valued functions $\boldsymbol{h}(\boldsymbol{x})$
  - all input dimensions contribute to all output dimensions

$$J = \nabla_{\boldsymbol{x}} \boldsymbol{h} = \frac{d\boldsymbol{h}}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_M} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_N}{\partial x_1} & \cdots & \frac{\partial h_N}{\partial x_M} \end{bmatrix}$$

- Single input, single output →
- Multiple input, single output →
- Single input, multiple output →
- Multiple input, multiple output →

Taking gradients with index notation for matrices/vectors…

- Often, output is a vector/matrix/tensor depend on matrix/vector/tensor input
- We still want to see what is the effect of the output w.r.t. the input. How?
- Better use index notation
  - Assign input and output indices and take derivatives with scalar quantities
  - E.g., $\boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{w}^T$, where $\dim(\boldsymbol{y}) = \mathrm{S} \times N, \dim(\boldsymbol{x}) = \mathrm{S} \times M, \dim(\boldsymbol{w}) = \mathrm{N} \times M$

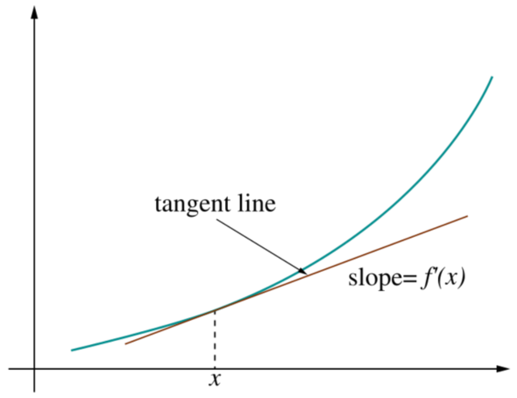


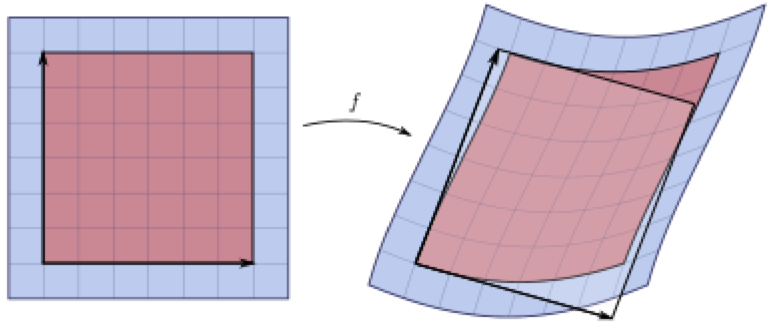
Jacobians, gradients, intuitively

- The Jacobians, gradients and the likes ($\frac{dh}{\mathrm{dx}}$) qualitatively capture the same thing
  - $\underbrace{\text{Change in the output}}_{dh}$ with respect to $\underbrace{\text{change in the input}}_{dx}$
- That is, the final Jacobian/gradient/… is simply a tensor ∇ with the shape
  - $\dim(\nabla) = \mathrm{shape}_{\mathrm{out}} \times \mathrm{shape}_{\mathrm{in}}$
  - If our 'in' is a vector, then we append that shape to the tensor gradient
  - The Einstein notation can be useful (np.einsum) for the computations

Jacobian, geometrically

- The Jacobian represents the best local approximation of how the space changes under a (non-linear) transformation
  - Not unlike derivative being the best linear approximation of a curve (tangent)
- The Jacobian determinant (for square matrices) measures the ratio of areas
  - Similar to what the 'absolute slope' measures in the 1d case (derivative)
  - Used in change of variables (integration by substitution), normalizing flows





Basic rules of partial differentiation

- Product rule
  - $\frac{\partial}{\partial x}\left(f(\boldsymbol{x}) \cdot g(\boldsymbol{x})\right) = f(\boldsymbol{x}) \cdot \frac{\partial}{\partial x} g(\boldsymbol{x}) + g(\boldsymbol{x}) \cdot \frac{\partial}{\partial x} f(\boldsymbol{x})$
- Sum rule
  - $\frac{\partial}{\partial x}\left(f(\boldsymbol{x}) + g(\boldsymbol{x})\right) = \frac{\partial}{\partial x} f(\boldsymbol{x}) + \frac{\partial}{\partial x} g(\boldsymbol{x})$

Computing gradients in complex functions: Chain rule

- Assume a composite function, $h = h_L\left(h_{L-1}\left(\ldots\left(h_1\left(\boldsymbol{x}\right)\right)\right)\right)$, or

$$h = h_L \circ h_{L-1} \circ \cdots \circ h_1\left(\boldsymbol{x}\right)$$

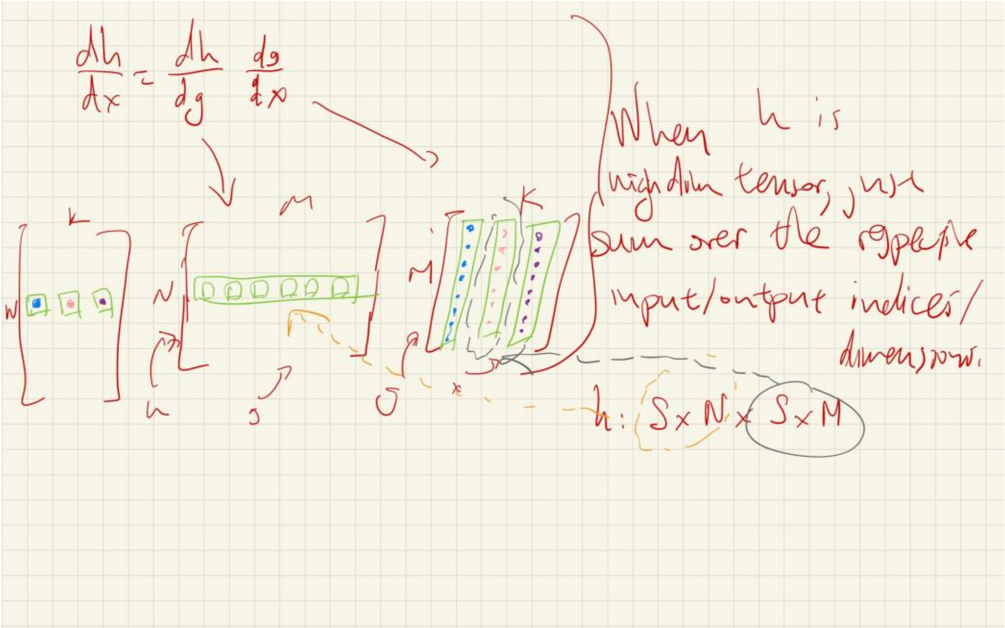
- To compute the derivative/gradient, we can use the chain rule
  - Intuitively, similar to matrix multiplications

$$\frac{dh}{dx} = \frac{dh}{dh_L} \cdot \frac{dh_L}{dh_{L-1}} \cdot \ldots \cdot \frac{dh_1}{dx}$$

- Each $\frac{dh_i}{dh_{i-1}}$ is a Jacobian/gradient/… vector/matrix/tensor

- Make sure each component matches dimensions

Chain rule and tensors, intuitively

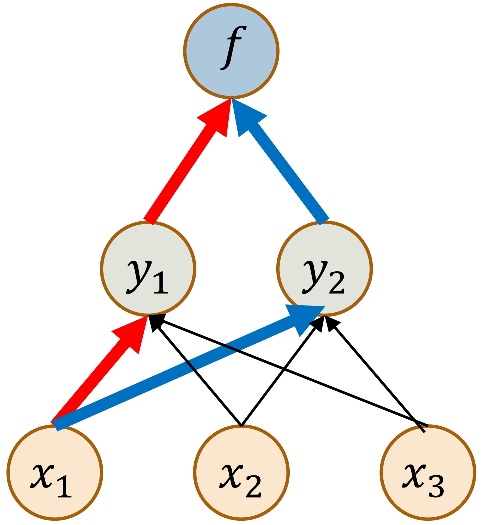


- What does the chain rule stand for with high-dimensional tensors
- Let's keep it simple: $\frac{d\boldsymbol{h}}{d\boldsymbol{x}} = \frac{d\boldsymbol{h}}{d\boldsymbol{g}} \cdot \frac{d\boldsymbol{g}}{d\boldsymbol{x}}$
  - $\boldsymbol{h}(\boldsymbol{g})$ has M inputs, N outputs
  - $\boldsymbol{g}(\boldsymbol{x})$ has K inputs (because of x), M outputs
- We can think of the chain rule as
  - summing over all possible changes
  - caused to $\boldsymbol{h}$ by each element in $\boldsymbol{x}$ via all possible $\boldsymbol{g}$'s
- For high-dim tensors, $\boldsymbol{h}, \boldsymbol{g}, \boldsymbol{x}$, we apply the same logic
  - Replace shape of the vector with shape of tensor
  - Do the summations keeping those shapes fixed
  - Think it in terms of indices, again [Einstein notation]

Example

- For $\mathrm{h} = f \circ y(x)$

$$\frac{dh}{dx} = \frac{df}{dy}\frac{dy}{dx} = \begin{bmatrix} \frac{\partial f}{\partial y_1} & \frac{\partial f}{\partial y_2} \end{bmatrix} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \end{bmatrix}$$



- Focusing on one of the partial derivatives: $\frac{\partial h}{\partial x_1}$

$$\frac{\partial h}{\partial x_1} = \frac{\partial f}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial f}{\partial y_2}\frac{\partial y_2}{\partial x_1}$$

- The partial derivative depends on all paths from f to x_i

Example

- For h $= f \circ y(x)$

$$\frac{dh}{dx} = \frac{df}{dy}\frac{dy}{dx} = \begin{bmatrix} \frac{\partial f}{\partial y_1} & \frac{\partial f}{\partial y_2} \end{bmatrix} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \end{bmatrix}$$

- Focusing on one of the partial derivatives: $\frac{\partial h}{\partial x_2}$

$$\frac{\partial h}{\partial x_2} = \frac{\partial f}{\partial y_1}\frac{dy_1}{dx_2} + \frac{\partial f}{\partial y_2}\frac{\partial y_2}{\partial x_2}$$

- The partial derivative depends on all paths from f to x_i