Computational Graphs, and Backpropagation

Michael Collins, Columbia University

A Key Problem: Calculating Derivatives

$$p(y|x;\theta,v) = \frac{\exp\left(v(y)\cdot\phi(x;\theta)+\gamma_y\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(v(y')\cdot\phi(x;\theta)+\gamma_{y'}\right)} \quad (1)$$

where

$$\phi(x;\theta) = g(Wx+b)$$

and

- m is an integer specifying the number of hidden units
- $W \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$ are the parameters in θ.
  $g : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is the transfer function
- Key question, given a training example (x^i, y^i), define

$$L(\theta,v) = -\log p(y_i|x_i;\theta,v)$$

How do we calculate derivatives such as $\frac{dL(\theta,v)}{dW_{k,j}}$?

A Simple Version of Stochastic Gradient Descent (Continued)

Algorithm:

- For $t = 1 \dots T$
  - Select an integer i uniformly at random from $\{1 \dots n\}$
  - Define $L(\theta, v) = -\log p(y_i | x_i; \theta, v)$
  - For each parameter θ_j, $\theta_j = \theta_j - \eta^t \times \frac{dL(\theta,v)}{d\theta_j}$
  - For each label y, for each parameter $v_k(y)$, $v_k(y) = v_k(y) - \eta^t \times \frac{dL(\theta,v)}{dv_k(y)}$
  - For each label y, $\gamma_y = \gamma_y - \eta^t \times \frac{dL(\theta,v)}{d\gamma_y}$

Output: parameters θ and v

Overview

- Introduction
- The chain rule
- Derivatives in a single-layer neural network
- Computational graphs
- Backpropagation in computational graphs
- Justification for backpropagation

Partial Derivatives

Assume we have scalar variables $z_1, z_2 \ldots z_n$, and y, and a function f, and we define

$$y = f(z_1, z_2, \ldots z_n)$$

Then the *partial derivative* of f with respect to z_i is written as

$$\frac{\partial f(z_1, z_2, \ldots z_n)}{\partial z_i}$$

We will also write the partial derivative as

$$\left.\frac{\partial y}{\partial z_i}\right|_{z_1 \ldots z_m}^{f}$$

which can be read as "the partial derivative of y with respect to z_i, under function f, at values $z_1 \ldots z_m$"

Partial Derivatives (continued)

We will also write the partial derivative as

$$\left.\frac{\partial y}{\partial z_i}\right|^{f}_{z_1 \ldots z_m}$$

which can be read as "the partial derivative of y with respect to z_i, under function f, at values $z_1 \ldots z_m$"

The notation including f is non-standard, but helps to alleviate a lot of potential confusion...

We will sometimes drop f and/or $z_1 \ldots z_m$ when this is clear from context

The Chain Rule

Assume we have equations

$$y = f(z), \qquad z = g(x)$$

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

Then

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

Or equivalently,

$$\left.\frac{\partial y}{\partial x}\right|_{x}^{h} = \left.\frac{\partial y}{\partial z}\right|_{g(x)}^{f} \times \left.\frac{\partial z}{\partial x}\right|_{x}^{g}$$

The Chain Rule

Assume we have equations

$$y = f(z), \qquad z = g(x)$$

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

then

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

For example, assume $f(z) = z^2$ and $g(x) = x^3$. Assume in addition that $x = 2$. Then:

$$z = x^3 = 8, \quad \frac{dg(x)}{dx} = 3x^2 = 12, \quad f(z) = z^2 = 64, \quad \frac{df(z)}{dz} = 2z = 16$$

from which it follows that $\frac{dh(x)}{dx} = 12 \times 16 = 192$

The Chain Rule (continued)

Assume we have equations

$$y = f(z)$$

$$z_1 = g_1(x), z_2 = g_2(x), \ldots, z_n = g_n(x)$$

For some functions f, $g_1 \ldots g_n$, where z is a vector $z \in \mathbb{R}^n$, and x is a vector $x \in \mathbb{R}^m$.
Define the function

$$h(x) = f(g_1(x), g_2(x), \ldots g_n(x))$$

Then we have

$$\frac{\partial h(x)}{\partial x_j} = \sum_i \frac{\partial f(z)}{\partial z_i} \frac{\partial g_i(x)}{\partial x_j}$$

where z is the vector $g_1(x), g_2(x), \ldots g_n(x)$.

The Jacobian Matrix

Assume we have a function $f : \mathbb{R}^n \to \mathbb{R}^m$ that takes some vector $x \in \mathbb{R}^n$ and then returns a vector $y \in \mathbb{R}^m$:

$$y = f(x)$$

The *Jacobian* $J \in \mathbb{R}^{m \times n}$ is defined as the matrix with entries

$$J_{i,j} = \frac{\partial f_i(x)}{\partial x_j}$$

Hence the Jacobian contains all partial derivatives of the function.

The Jacobian Matrix

Assume we have a function $f : \mathbb{R}^n \to \mathbb{R}^m$ that takes some vector $x \in \mathbb{R}^n$ and then returns a vector $y \in \mathbb{R}^m$:

$$y = f(x)$$

The *Jacobian* $J \in \mathbb{R}^{m \times n}$ is defined as the matrix with entries

$$J_{i,j} = \frac{\partial f_i(x)}{\partial x_j}$$

Hence the Jacobian contains all partial derivatives of the function. We will also use

$$\left.\frac{\partial y}{\partial x}\right|_x^f$$

for vectors y and x to refer to the Jacobian matrix with respect to a function f mapping x to y, evaluated at x

An Example of the Jacobian: The LOG-SOFTMAX Function

We define $\mathsf{LS} : \mathbb{R}^K \rightarrow \mathbb{R}^K$ to be the function such that for $k = 1 \ldots K$,

$$\mathsf{LS}_k(l) = \log\left(\frac{\exp\{l_k\}}{\sum_{k'} \exp\{l_{k'}\}}\right) = l_k - \log\sum_{k'} \exp\{l_{k'}\}$$

The Jacobian then has entries

$$\left[\frac{\partial \mathsf{LS}(l)}{\partial l}\right]_{k,k'} = \frac{\partial \mathsf{LS}_k(l)}{\partial l_{k'}} = [[k = k']] - \frac{\exp\{l_{k'}\}}{\sum_{k''} \exp\{l_{k''}\}}$$

where $[[k = k']] = 1$ if $k = k'$, 0 otherwise.

The Chain Rule (continued)

Assume we have equations

$$y = f(z^1, z^2, \ldots z^n)$$

$$z^i = g^i(x^1, x^2, \ldots x^m)$$

for $i = 1 \ldots n$ where y is a vector, z^i for all i are vectors, and x^j for all j are vectors. Define $h(x^1 \ldots x^m)$ to be the composition of f and g, so $y = h(x^1 \ldots x^m)$. Then

$$\underbrace{\left.\frac{\partial y}{\partial x^j}\right|^{h}}_{d(y)\times d(x^j)} = \sum_{i=1}^{n} \underbrace{\left.\frac{\partial y}{\partial z^i}\right|^{f}}_{d(y)\times d(z^i)} \times \underbrace{\left.\frac{\partial z^i}{\partial x_j}\right|^{g^i}}_{d(z^i)\times d(x^j)}$$

where $d(v)$ is the dimensionality of vector v.

Overview

- Introduction
- The chain rule
- Derivatives in a single-layer neural network
- Computational graphs
- Backpropagation in computational graphs
- Justification for backpropagation

Derivatives in a Feedforward Network

Definitions: The set of possible labels is $\mathcal{Y}$. We define $K = |\mathcal{Y}|$.
$g : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a transfer function. We define
LS = LOG-SOFTMAX.

Inputs: $x^i \in \mathbb{R}^d, y^i \in \mathcal{Y}$, $W \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$, $V \in \mathbb{R}^{K \times m}$, $\gamma \in \mathbb{R}^K$.

Equations:

$$
\begin{aligned}
z \in \mathbb{R}^m &= W x^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= V h + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}
$$

Jacobian Involving Matrices

Equations:

$$
\begin{aligned}
z \in \mathbb{R}^m &= Wx^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}
$$

If $W \in \mathbb{R}^{m \times d}$, $z \in \mathbb{R}^m$, the Jacobian

$$\frac{\partial z}{\partial W}$$

is a matrix of dimension $m \times m'$ where $m' = (m \times d)$ is the number of entries in W. So we treat W as a vector with $(m \times d)$ elements.

Local Functions

Equations:

$$
\begin{aligned}
z \in \mathbb{R}^m &= Wx^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}
$$

Leaf variables: W, x^i, b, V, γ, y_i

Intermediate variables: z, h, l, q

Output variable: o

Each intermediate variable has a "Local" function:

$$f^z(W, x^i, b) = Wx^i + b, \quad f^h(z) = g(z), \quad f^l(h) = Vh + \gamma, \quad \ldots$$

Global Functions

Equations:

$$
\begin{aligned}
z \in \mathbb{R}^m &= Wx^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}
$$

Leaf variables: W, x^i, b, V, γ, y_i

Intermediate variables: z, h, l, q

Output variable: o

Global functions: for the output variable o, we define $\bar{f}^o$ to be the function that maps the leaf values W, x^i, b, V, γ, y_i to the output value $o = \bar{f}^o(W, x^i, b, V, \gamma, y_i)$. We use similar definitions for $\bar{f}^z(W, x^i, b, V, \gamma, y_i)$, $\bar{f}^h(W, x^i, b, V, \gamma, y_i)$, etc.

Applying the Chain Rule

Equations:

$$\begin{aligned} z \in \mathbb{R}^m &= Wx^i + b \\ h \in \mathbb{R}^m &= g(z) \\ l \in \mathbb{R}^K &= Vh + \gamma \\ q \in \mathbb{R}^K &= \mathsf{LS}(l) \\ o \in \mathbb{R} &= -q_{y_i} \end{aligned}$$

Derivative:

$$\left.\frac{\partial o}{\partial W}\right|^{\bar{f}^o} =$$

Applying the Chain Rule

Derivative:

$$\left.\frac{\partial o}{\partial W}\right|^{\bar{f}^o} = \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial W}\right|^{\bar{f}^q}$$

Equations:

$$\begin{aligned}
z \in \mathbb{R}^m &= Wx^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}$$

Applying the Chain Rule

Equations:

$$\begin{aligned} z \in \mathbb{R}^m &= W x^i + b \\ h \in \mathbb{R}^m &= g(z) \\ l \in \mathbb{R}^K &= V h + \gamma \\ q \in \mathbb{R}^K &= \mathsf{LS}(l) \\ o \in \mathbb{R} &= -q_{y_i} \end{aligned}$$

Derivative:

$$\begin{aligned} \frac{\partial o}{\partial W}\bigg|^{\bar{f}^o} &= \frac{\partial o}{\partial q}\bigg|^{f^o} \times \frac{\partial q}{\partial W}\bigg|^{\bar{f}^q} \\ &= \frac{\partial o}{\partial q}\bigg|^{f^o} \times \frac{\partial q}{\partial l}\bigg|^{f^q} \times \frac{\partial l}{\partial W}\bigg|^{\bar{f}^l} \end{aligned}$$

Applying the Chain Rule

Equations:

$$
\begin{aligned}
z \in \mathbb{R}^m &= Wx^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}
$$

Derivative:

$$
\begin{aligned}
\left.\frac{\partial o}{\partial W}\right|^{\bar{f}^o} &= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial W}\right|^{\bar{f}^q} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial W}\right|^{\bar{f}^l} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial h}\right|^{f^l} \times \left.\frac{\partial h}{\partial W}\right|^{\bar{f}^h}
\end{aligned}
$$

Applying the Chain Rule

Equations:

$$
\begin{aligned}
z \in \mathbb{R}^m &= Wx^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}
$$

Derivative:

$$
\begin{aligned}
\left.\frac{\partial o}{\partial W}\right|^{\bar{f}^o} &= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial W}\right|^{\bar{f}^q} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial W}\right|^{\bar{f}^l} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial h}\right|^{f^l} \times \left.\frac{\partial h}{\partial W}\right|^{\bar{f}^h} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial h}\right|^{f^l} \times \left.\frac{\partial h}{\partial z}\right|^{f^h} \times \left.\frac{\partial z}{\partial W}\right|^{\bar{f}^z}
\end{aligned}
$$

Applying the Chain Rule

Equations:

$$
\begin{aligned}
z \in \mathbb{R}^m &= Wx^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}
$$

Derivative:

$$
\begin{aligned}
\left.\frac{\partial o}{\partial W}\right|^{\bar{f}^o} &= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial W}\right|^{\bar{f}^q} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial W}\right|^{\bar{f}^l} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial h}\right|^{f^l} \times \left.\frac{\partial h}{\partial W}\right|^{\bar{f}^h} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial h}\right|^{f^l} \times \left.\frac{\partial h}{\partial z}\right|^{f^h} \times \left.\frac{\partial z}{\partial W}\right|^{\bar{f}^z} \\
&= \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial h}\right|^{f^l} \times \left.\frac{\partial h}{\partial z}\right|^{f^h} \times \left.\frac{\partial z}{\partial W}\right|^{f^z}
\end{aligned}
$$

Another Derivative

Equations:

$$
\begin{aligned}
z \in \mathbb{R}^m &= Wx^i + b \\
h \in \mathbb{R}^m &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \mathsf{LS}(l) \\
o \in \mathbb{R} &= -q_{y_i}
\end{aligned}
$$

$$\left.\frac{\partial o}{\partial V}\right|^{\bar{f}^o} = \left.\frac{\partial o}{\partial q}\right|^{f^o} \times \left.\frac{\partial q}{\partial l}\right|^{f^q} \times \left.\frac{\partial l}{\partial v}\right|^{f^l}$$

A Computational Graph

Equations:

$$\begin{aligned} z \in \mathbb{R}^m &= Wx^i + b \\ h \in \mathbb{R}^m &= g(z) \\ l \in \mathbb{R}^K &= Vh + \gamma \\ q \in \mathbb{R}^K &= \mathsf{LS}(l) \\ o \in \mathbb{R} &= -q_{y_i} \end{aligned}$$

Derivatives:

$$\begin{aligned} \frac{\partial o}{\partial V}\bigg|^{\bar{f}^o} &= \frac{\partial o}{\partial q}\bigg|^{f^o} \times \frac{\partial q}{\partial l}\bigg|^{f^q} \times \frac{\partial l}{\partial v}\bigg|^{f^l} \\ \frac{\partial o}{\partial W}\bigg|^{\bar{f}^o} &= \frac{\partial o}{\partial q}\bigg|^{f^o} \times \frac{\partial q}{\partial l}\bigg|^{f^q} \times \frac{\partial l}{\partial h}\bigg|^{f^l} \times \frac{\partial h}{\partial z}\bigg|^{f^h} \times \frac{\partial z}{\partial W}\bigg|^{f^z} \end{aligned}$$

Overview

- Introduction
- The chain rule
- Derivatives in a single-layer neural network
- Computational graphs
- Backpropagation in computational graphs
- Justification for backpropagation

Computational Graphs: a Formal Definition

A computational graph consists of:

- An integer n specifying the number of vertices in the graph. An integer $l < n$ specifying the number of leaves in the graph. Vertices $1 \ldots l$ are leaves in the graph. Vertex n is a special "output" vertex.
- A set of directed edges E. Each member of E is an ordered pair (j, i) where $j \in \{1 \ldots n\}$, $i \in \{(l+1) \ldots n\}$, and $i > j$. For any i we define $\pi(i)$ to be the set of parents of i in the graph:

$$\pi(i) = \{j : (j, i) \in E\}$$

Computational Graphs (continued)

- A variable $u^i \in \mathbb{R}^{d_i}$ is associated with each vertex in the graph. Here d_i for $i = 1 \ldots n$ specifies the dimensionality of u^i. We assume $d_n = 1$, hence the output variable is a scalar.
- A function f^i is associated with each non-leaf vertex in the graph ($i \in \{(l+1) \ldots n\}$). The function maps a vector A^i defined as

  $$A^i = \langle u^j | j \in \pi(i) \rangle$$

  to a vector $f^i(A^i) \in \mathbb{R}^{d_i}$

An Example

- Define $n = 4$, $l = 2$
- Define $d_i = 1$ for all i (all variables are scalars)
- Define $E = \{(1,3),(2,3),(2,4),(3,4)\}$
- Define

$$f^3(u^1, u^2) = u^1 + u^2$$

$$f^4(u^2, u^3) = u^2 \times u^3$$

Two Questions

- Note that the computational graph defines a function, which we call $\bar{f}^n$, from the values of the leaf variables to the output variable:

$$u^n = \bar{f}^n(u^1 \ldots u^l)$$

- Given a computational graph, and values for the leaf variables $u^1 \ldots u^l$:
  1. How do we compute the output u^n?
  2. How do we compute the partial derivatives

$$\left.\frac{\partial u^n}{\partial u^i}\right|^{\bar{f}^n}$$

  for all $i \in \{1 \ldots l\}$?

Forward Computation

Input: Values for leaf variables $u^1 \ldots u^l$
Algorithm:

- For $i = (l + 1) \ldots n$

$$u^i = f^i(A^i)$$

  where

$$A^i = \langle u^j | j \in \pi(i) \rangle$$

An Example

- Define $n = 4$, $l = 2$
- Define $d_i = 1$ for all i (all variables are scalars)
- Define $E = \{(1,3),(2,3),(2,4),(3,4)\}$
- Define

$$f^3(u^1, u^2) = u^1 + u^2$$

$$f^4(u^2, u^3) = u^2 \times u^3$$

Defining and Calculating Derivatives

- For any $k \in \{(l+1) \ldots n\}$, there is a function $\bar{f}^k$ such that

$$u^k = \bar{f}^k(u^1, u^2, \ldots u^l)$$

- We want to calculate

$$\left.\frac{\partial u^n}{\partial u^j}\right|^{\bar{f}^n}_{u^1 \ldots u^l}$$

for $j = 1 \ldots l$

Computational Graphs (continued)

- A function $J^{j \to i}$ is associated with each edge $(j, i) \in E$. The function maps a vector A^i defined as

$$A^i = \langle u^j | j \in \pi(i) \rangle$$

to a matrix $J^{j \to i}(A^i) \in \mathbb{R}^{d_i \times d_j}$.

$$J^{j \to i}(A^i) = \frac{\partial f^i(A^i)}{\partial u^j} = \frac{\partial u^i}{\partial u^j}\bigg|_{A^i}^{f^i}$$

Forward Pass

Input: Values for leaf variables $u^1 \dots u^l$
Algorithm:

- For $i = (l+1) \dots n$

$$A^i = \langle u^j | j \in \pi(i) \rangle$$

$$u^i = f^i(A^i)$$

Backward Pass

- $p^n = 1$
- For $j = (n-1)\ldots 1$:

$$p^j = \sum_{i:(j,i)\in E} p^i J^{j\to i}(A^i)$$

- **Output:** p^i for $i = 1\ldots l$ satisfying

$$p^i = \frac{\partial o}{\partial u^i}\bigg|_{u^1\ldots u^l}^{\bar{f}^n}$$

An Example

$p^n = 1$

For $j = (n-1) \ldots 1$:

$$p^j = \sum_{i:(j,i) \in E} p^i J^{j \to i}(A^i)$$

Overview

- Introduction
- The chain rule
- Derivatives in a single-layer neural network
- Computational graphs
- Backpropagation in computational graphs
- Justification for backpropagation

Products of Jacobians over Paths in the Graph

- A *directed path* between vertices j and k is a sequence of edges $(i_1, i_2), (i_2, i_3), \ldots (i_{n-1}, i_n)$ with $n \geq 2$ such that each edge is in E, and $i_1 = j$, and $i_n = k$.
- For any j, k, we write $\mathcal{P}(j, k)$ to denote the set of all directed paths between j and k
- For convenience we define $D^{a \to b} = J^{a \to b}(A^b)$ for all edges (a, b).
- *Theorem:* for any $j \in \{1 \ldots l\}$, $k \in \{(l+1) \ldots n\}$,

$$\left.\frac{\partial u^k}{\partial u^j}\right|^{\bar{f}^k} = \sum_{p \in \mathcal{P}(j,k)} \prod_{(a,b) \in p} D^{a \to b}$$

An Example

$$\left.\frac{\partial u^k}{\partial u^j}\right|^{\bar{f}^k} = \sum_{p \in \mathcal{P}(j,k)} \prod_{(a,b) \in p} D^{a \to b}$$

Proof Sketch

- For any j, j', k, we write $\mathcal{P}(j, j', k)$ to denote the set of all directed paths between j and k such that the last edge in the sequence is (j', k).
- *Proof sketch:* By induction over the graph. By the chain rule we have

$$
\begin{aligned}
\frac{\partial u^k}{\partial u^j}\bigg|^{\bar{f}^k} &= \sum_{j':(j',k)\in E} D^{j'\to k} \times \frac{\partial u^{j'}}{\partial u^j}\bigg|^{\bar{f}^{j'}} \\
&= \sum_{j':(j',k)\in E} D^{j'\to k} \times \sum_{p\in\mathcal{P}(j,j')} \prod_{(a,b)\in p} D^{a\to b} \\
&= \sum_{j':(j',k)\in E} \sum_{p\in\mathcal{P}(j,j',k)} \prod_{(a,b)\in p} D^{a\to b} \\
&= \sum_{p\in\mathcal{P}(j,k)} \prod_{(a,b)\in p} D^{a\to b}
\end{aligned}
$$

Backward Pass

- $p^n = 1$
- For $j = (n-1)\ldots 1$:

$$p^j = \sum_{i:(j,i)\in E} p^i D^{j\to i}$$

- **Output:** p^i for $i = 1\ldots l$ satisfying

$$p^i = \left.\frac{\partial o}{\partial u^i}\right|_{u^1,u^2,\ldots u^l}^{\bar{f}^o}$$

Correctness of the Backward Pass

- *Theorem:* For all p^i we have

$$p^i = \sum_{p \in \mathcal{P}(i,n)} \prod_{(a,b) \in p} D^{a \to b}$$

It follows that for any $i \in \{1 \ldots l\}$,

$$p^i = \left. \frac{\partial u^n}{\partial u^i} \right|^{\bar{f}^n}$$

Proof

- *Theorem:* For all p^i we have

$$p^i = \sum_{p \in \mathcal{P}(i,n)} \prod_{(a,b) \in p} D^{a \rightarrow b}$$

- *Proof sketch:* by induction on $i = n, i = (n-1), i = (n-2), \ldots i = 1$. For $i = n$ we have $p^n = 1$ so the proposition is true. For $j = (n-1) \ldots 1$ we have

$$\begin{aligned} p^j &= \sum_{i:(j,i) \in E} p^i D^{j \rightarrow i} \\ &= \sum_{i:(j,i) \in E} \left(\sum_{p \in \mathcal{P}(i,n)} \prod_{(a,b) \in p} D^{a \rightarrow b} \right) D^{j \rightarrow i} = \sum_{p \in \mathcal{P}(j,n)} \prod_{(a,b) \in p} D^{a \rightarrow b} \end{aligned}$$