Fundamentals of Machine Learning

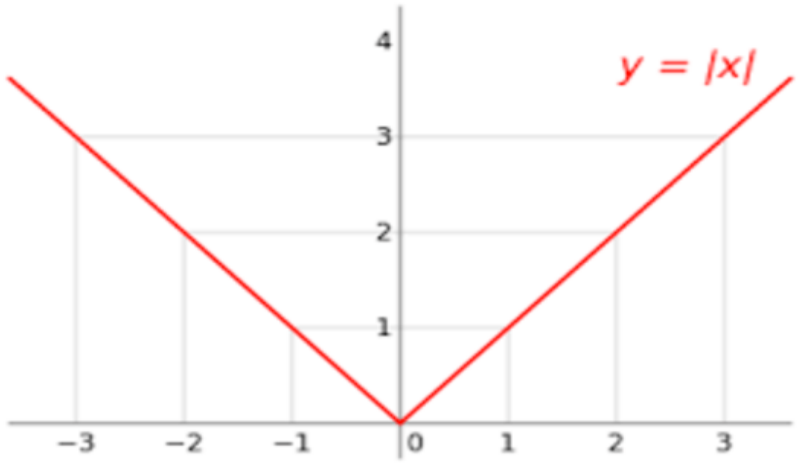
Vector calculus: linearization and multivariate taylor series in machine learning and Automatic Differentiation

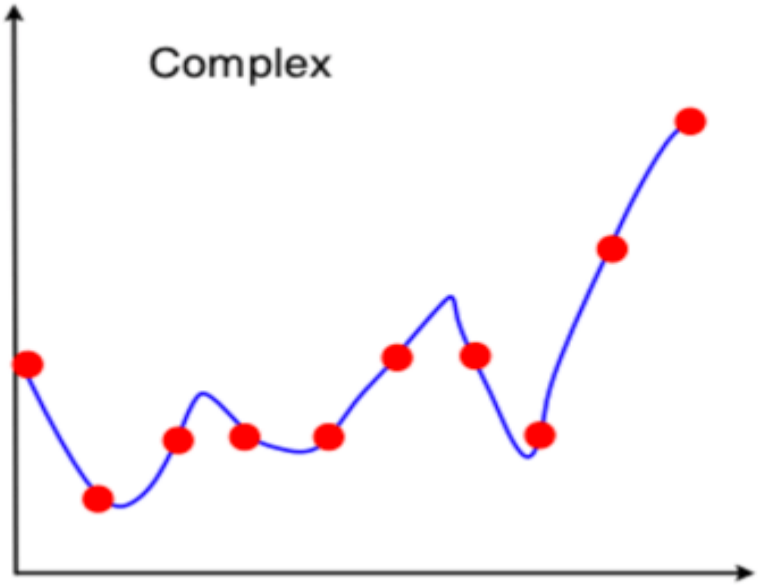
Approximation

- Real applications have high dimensions which are difficult to visualize or are complex to differentiate everywhere .
- How to minimize functions that we *can't* differentiate everywhere and *can't* visualize?

Approximation

- We used derivatives (called *differentiating* the function) to find flat points on cost functions that might minimize the cost.
- But not all functions are differentiable everywhere.
- In given figures, it's global minimum at x = 0, but we *can't differentiate the function* at x = 0 because it's pointy there rather than flat.





Solution: Approximation

- Find a point where function can be differentiated and use it to find an approximation of …..a minimum.
  - Automatic Differentiation
  - Linearization
  - Taylor series

Linear approximation/ Linearization

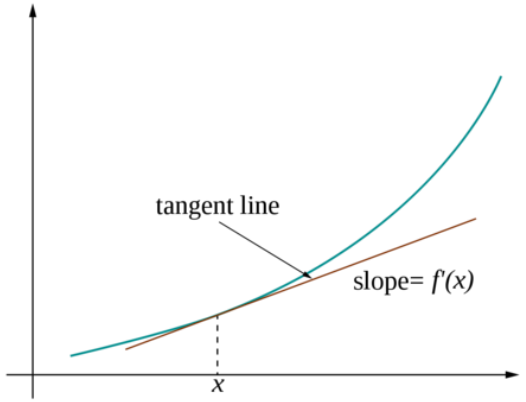
Linear approximation is a method of estimating the value of a function, f(x), near a point, x = a, using the following formula:

$$y = f(a) + f'(a)(x - a)$$

The formula is known as the linearization of f at $x = a$, but this formula is identical to the equation of the tangent line to f at $x = a$.

The value of the derivative at a specific point, $x = a$, measures the slope of the curve, $y = f(x)$, at that point. In other words, $f'(a)$ = slope of the tangent line at a.

The tangent line is the one line that matches the direction of the curve most closely, at the specific x-value we are interested in..



Formula for Linearization

How to find the linearization of a function f at a point $x = a$?

The equation of a line can be determined if we know
1.The slope of the line, m
2.Any single point that the line goes through, (a, b).

Plug these pieces of info into the point-slope form, and this gives us the equation of the line.
$y - b = m(x-a)$

Firstly $m = f'(a)$, because the derivative measures the slope, and secondly, $b = f(a)$, because the original function measures y-values.
Putting it all together and solving for y:

$$\begin{aligned} y - b &= m(x - a) \\ y &= b + m(x - a) \\ y &= f(a) + f'(a)(x - a) \end{aligned}$$

The last line is precisely the linearization of the function f at the point $x = a$.

Practice Example

Find the approximate value of (natural log function)
$f(x) = \ln x$. Find the linearization at x=1. Use the linearization to approximate f(1.1) and f(0.9) using linear approximation.

Solution:

Given function $f(x) = \ln x$, **to find the linearization at 1, we need to find f(1) and f ' (1).**

$f'(x) = 1/x$.

$f'(a) = f'(1) = 1$.

Furthermore, $f(a) = \ln 1 = 0$.

Therefore, the linear approximation of f at $x = 1$ is

$y = f(1) + f'(1)(x - 1)$

$y = 0 + 1(x - 1)$

$y = x - 1$

The linear approximation for f(1.1):

$f(1.1) = \ln 1.1$ can be approximated by

$y = 1.1 - 1 = 0.1$.

The linear approximation for f(0.9):

f(0.9) can be approximated by

$y = 0.9 - 1 = -0.1$.

.

Exercise :Find the linearization of f(x) =cos(x) at a= $\frac{\pi}{2}$

$$L(x) = f(a) + f'(a)(x-a)$$

$$f(a) = f(\pi/2) = \cos(\pi/2) = 0$$

$$f'(a) = f'(\pi/2) = -\sin(\pi/2)$$

$$= -1$$

$$L(x) = 0 + -1(x - \pi/2)$$

$$= -1(x - \pi/2)$$

$$\boxed{L(x) = -x + \pi/2}$$

Approximating the change in a function

Differentials can be used to estimate the change in the value of a function resulting from a small change in input values.

Consider a function **f** that is differentiable at point **a** .
Suppose the **input x** changes by a small amount, then how much the **output y** changes.
If **x** changes from **a** to **a+ *dx*** then the change in **y** is given by

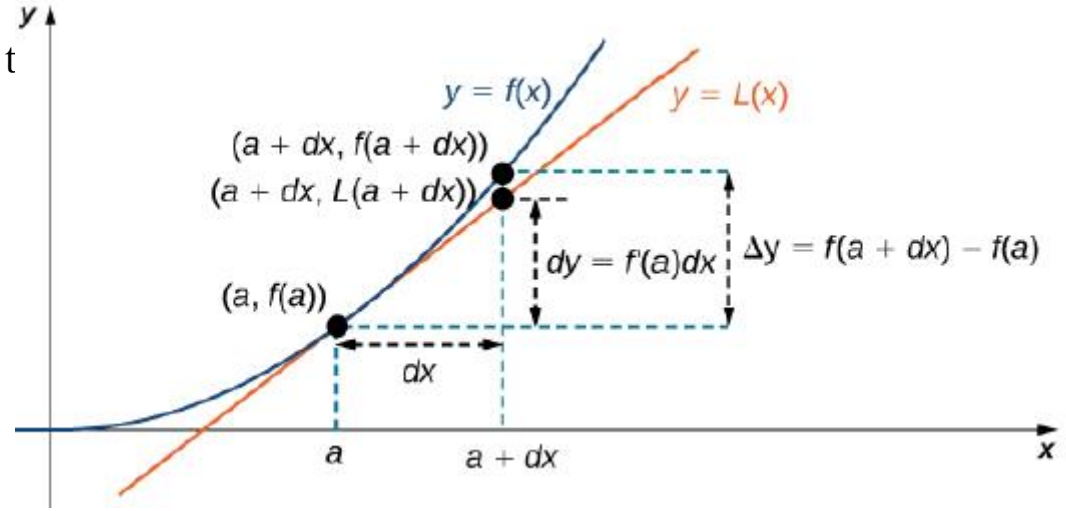
$$y = f(a + dx) - f(a).$$

Instead of calculating the exact change in **y**, it is easier to approximate t change in **y** by using a linear approximation. For **x** near **a** can be approximated by the linear approximation

$$L(x) = f(a) + f\prime(a)(x - a).$$

Therefore, if dx is small,

$$f(a+dx) \approx L(a+dx) = f(a) + f'(a)(a+dx-a).$$



$$\Delta y = f(a+dx) - f(a) \approx L(a+dx) - f(a) = f'(a)dx = dy.$$

Therefore, we can use the differential $dy = f'(a)dx$ to approximate the change in y if x increases from $x = a$ to $x = a+dx$.

Calculating the Amount of Error

Any type of measurement is prone to a certain amount of error.

Consider a function **f** with an input is a measured quantity. Suppose the exact value of the **f** is **a** , but the measured value is **a+dx** . Then the measurement error is **dx** (orΔx). As a result, an error occurs in the calculated quantity **f(x) .**

The error is:

$$\Delta y = f(a+dx) - f(a).$$

if f is a differentiable function at a,the propagated error is

$$\Delta y \approx dy = f'(a)dx\,.$$

Since the value of **a** is not known, the measured value **a+dx** can be used to estimate **a**

$$\Delta y \approx dy \approx f'(a+dx)dx\,.$$

- If a number a is an approximation to the number A, then the error in approximating A is the absolute value of the difference,

$$|a - A|.$$

- The percentage error in the approximation of A by a is the quantity

$$100 \times \frac{|a - A|}{|A|}.$$

Use Linear approximation to find the value of $\sqrt{17}$. Find the error in the approximation.
Also find the percentage of error

$L(x) = f(a) + f'(a)\cdot(x-a)$

$a = ?$

a value that is close enough whose precise value we already know.

$a = 16 \qquad \sqrt{16} = 4$

$f(x) = \sqrt{x} = (x)^{1/2}$

$f'(x) = \frac{1}{2} x^{-1/2}$

$= \frac{1}{2\sqrt{x}}$

$f(17) \approx \sqrt{16} + \frac{1}{2\sqrt{16}}(17-16)$

$\approx 4 + \frac{1}{8}(1)$

$\approx 4 + \frac{1}{8} \approx 4\frac{1}{8} = 4.125$

Calculator $\sqrt{17} = 4.123$

Error $= |\text{approx} - \text{exact}|$

$= 4.123 - 4.125$

Calculator $\sqrt{17} = 4.123$

Error $= |\text{approx} - \text{exact}|$

$= 4.123 - 4.125$

$= 0.00189$

percent error $= \left|\frac{\text{approx} - \text{exact}}{\text{exact}}\right| \times 100$

$= \frac{0.00189}{4.123}$

$= 0.045\%$

Example. Let $f(x) = \sqrt{1+2x}$ and use the linearization to approximate $f(4.3)$.

Find the error in the approximation of $f(4.3)$, the percentage error in the approximation of $f(4.3)$ and the percentage error in the approximation of Δf.

Solution

It it is easy to compute f(4) = √ 9 = 3. So compute the linearization of f at 4.

$$L(x) = f(a) + f'(a)(x-a).$$

To do this we need to find f' (4).

$$\begin{aligned}\frac{d}{dx}\sqrt{1+2x} &= \frac{d}{dx}(1+2x)^{1/2} \\ &= \frac{1}{2}(1+2x)^{-1/2}2 \\ &= \frac{1}{\sqrt{1+2x}}.\end{aligned}$$

Evaluating at x = 4, give $f'(4) = 1/3$. Thus the linearization of f at 4 is

$$L(x) = f(4) + f'(4)(x-4) = 3 + \frac{1}{3}(x-4).$$

To make the approximation we write

$$f(4.3) \approx L(4.3) = f(4) + \frac{1}{3}0.3 = 3.1.$$

A calculator gives the more accurate approximation,

$$f(4.3) \approx 3.0984.$$

The errors are

$$|f(4.3) - 3.1| \approx 0.00161, \qquad 100\% \times \frac{|f(4.3) - 3.1|}{|f(4.3)|} \approx 0.0521\%.$$

Finally the error in the change of f is given by

$$100\% \times \frac{|f(4.3) - f(4) - f'(4)(4.3 - 4)|}{|f(4.3) - f(4)|} \approx 1.6\%.$$

Linear approximation of a multivariate function

Linear approximation of a multivariate function

Given a differentiable scalar multivariate function f(x), where $x = [x_1, x_2]^T$, the linear approximation of the function at point $a = [a_1, a_2]^T$ is

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{\partial}{\partial x_1} f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}(x_1 - a_1)$$

$$+ \frac{\partial}{\partial x_2} f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}(x_2 - a_2)$$

$$= f(\mathbf{a}) + (\mathbf{x} - \mathbf{a})^T \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}$$

where

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

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{\partial}{\partial x_1} f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}(x_1 - a_1)$$

$$+ \frac{\partial}{\partial x_2} f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}(x_2 - a_2)$$

$$= f(\mathbf{a}) + (\mathbf{x} - \mathbf{a})^T \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}$$

Example: Suppose

$$f(\mathbf{x}) = x_1^2 + 2x_2^2 - 5x_1 - 8x_2 + 7$$

is to be linearised at x =[5 4] T

Solution:

$$f([5\ 4]^T) = 25 + 2 \times 16 - 5 \times 5 - 8 \times 4 + 7 = 7$$

$$\frac{\partial}{\partial x_1} f(\mathbf{x}) = 2x_1 - 5$$

$$\frac{\partial}{\partial x_1} f(\mathbf{x})|_{\mathbf{x}=[5\ 4]^T} = 10 - 5 = 5$$

$$\frac{\partial}{\partial x_2} f(\mathbf{x}) = 4x_2 - 8$$

$$\frac{\partial}{\partial x_2} f(\mathbf{x})|_{\mathbf{x}=[5\ 4]^T} = 16 - 8 = 8$$

$$f(x) \approx 7 + (x_1 - 5) * 5 + (x_2 - 4) * 8$$

$$= 5x_1 + 8x_2 - 50$$

Hence can also say that $g(\mathbf{x}) = 5x_1 + 8x_2 - 50$ is equation of tangent to surface $f(\mathbf{x}) = x_1^2 + 2x_2^2 - 5x_1 - 8x_2 + 7$ at $x = [5\ \ 4]^T$.

Approximation using Taylor Series

Taylor Series

The Taylor Series can be used to approximate a function, f(x). The Taylor approximation has a symmetric definition that uses the first, second, third, and so on, derivatives, and also the factorial function.

In various ML applications, working directly with some function f(x) is very difficult, but working with the Taylor approximation to f(x), i.e P(x), is easier.

Given a differentiable scalar function f(x) of the real variable x, its Taylor series reads

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2!}f''(a)(x-a)^2 + \dots = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(a)(x-a)^k$$

where $f''(a) = \frac{d^2 f(x)}{dx^2}|_{x=a}$, $f^{(n)}(a) = \frac{d^n f(x)}{dx^n}|_{x=a}$.

If $n = 1$, this reduces to the linear approximation.

Example: let f(x) $= x^2 e^x$ Show that the Taylor expansion of f(x) around 0, and up to the 4th order, is

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2!}f''(a)(x-a)^2 + \ldots$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(a)(x-a)^k$$

Solution :

$$x^2 e^x = x^2 \sum_{k=0}^{\infty} \frac{x^k}{k!} = \sum_{k=0}^{\infty} \frac{x^{k+2}}{k!}$$

$$f(x) = x^2 + x^3 + \frac{1}{2}x^4 + \ldots$$

$$f(0) = 0^2 e^0 = 0$$

$$f'(x) = 2xe^x + x^2 e^x = (2x + x^2)e^x$$

$$f'(0) = 0$$

$$f''(x) = (2 + 2x)e^x + (2x + x^2)e^x$$

$$= (2 + 4x + x^2)e^x$$

$$f''(0) = 2$$

$$f^{(3)}(x) = (4 + 2x)e^x + (2 + 4x + x^2)e^x$$

$$= (6 + 6x + x^2)e^x$$

$$f^{(3)}(0) = 6$$

$$f^{(4)}(x) = (6 + 2x)e^x + (6 + 6x + x^2)e^x$$

$$= (12 + 8x + x^2)e^x$$

$$f^{(4)}(0) = 12$$

$$f(x) = \frac{2}{2!}x^2 + \frac{6}{3!}x^3 + \frac{12}{4!}x^4 + \ldots$$

$$= x^2 + x^3 + \frac{1}{2}x^4 + \ldots$$

Single variable Taylor series:

Let f be an infinitely differentiable function in some open interval around x = a.

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

Linear approximation in one variable:

Take the constant and linear terms from the Taylor series. In an open interval around x = a,

$$f(x) \approx f(a) + f'(a)(x-a) \quad \text{linear approximation}$$

Quadratic approximation in one variable:

Take the constant, linear, and quadratic terms from the Taylor series. In an open interval around x = a,

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 \quad \text{quadradic approximation}$$

Taylor expansion of multivariate functions

Taylor expansion of multivariate functions:

Given a differentiable scalar multivariate function f(x), where $x = [x_1, ..., x_m]^T$, the Taylor series of the function at point $a = [a_1, ..., a_m]^T$

$$f(\mathbf{x}) = f(\mathbf{a}) + \frac{\partial}{\partial x_1} f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}(x_1 - a_1) + \ldots$$

$$+ \frac{\partial}{\partial x_m} f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}(x_m - a_m)$$

$$+ \frac{1}{2}\frac{\partial^2}{\partial x_1^2} f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}(x_1 - a_1)^2 + \ldots$$

$$+ \frac{1}{2}\frac{\partial^2}{\partial x_m^2} f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}(x_m - a_m)^2 + \ldots$$

$$= f(\mathbf{a}) + (\mathbf{x} - \mathbf{a})^T \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}}$$

$$+ \frac{1}{2}(\mathbf{x} - \mathbf{a})^T \mathbf{H} (\mathbf{x} - \mathbf{a})^T + \ldots$$

where

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

$$H = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}|_{\mathbf{x}=\mathbf{a}}$$

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

Note : The second term is an inner product of a gradient of f(x) with the vector x−a and the third term is a quadratic form with the Hessian matrix H(f)

Example:let $f(\mathbf{x}) = (x_1^2+2x_2^2-5x_1-8x_2+7)e^{x_1}$.
Show that the Taylor expansion of $f(\mathbf{x})$ around $\mathbf{0}$, and up to the 2nd order,is

$$f(\mathbf{x}) = f(\mathbf{a}) + (\mathbf{x}-\mathbf{a})^T \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}} + \frac{1}{2}(\mathbf{x}-\mathbf{a})^T \mathbf{H} (\mathbf{x}-\mathbf{a})^T + \ldots$$

Solution:

$$f(\mathbf{x}) = 7 + 2x_1 - 8x_2 - \frac{1}{2}x_1^2 - 8x_1x_2 + 2x_2^2 + \ldots$$

$$f([0\ 0]^T) = 7$$

$$\frac{\partial}{\partial x_1} f(\mathbf{x}) = (2x_1-5)e^{x_1} + (x_1^2+2x_2^2-5x_1-8x_2+7)e^{x_1}$$

$$= (x_1^2 + 2x_2^2 - 3x_1 - 8x_2 + 2)e^{x_1}$$

$$\frac{\partial}{\partial x_1} f(\mathbf{x})|_{\mathbf{x}=[0\ 0]^T} = 2$$

$$\frac{\partial}{\partial x_2} f(\mathbf{x}) = (4x_2 - 8)e^{x_1}$$

$$\frac{\partial}{\partial x_2} f(\mathbf{x})|_{\mathbf{x}=[0\ 0]^T} = -8$$

$$\frac{\partial^2}{\partial x_1^2} f(\mathbf{x}) = (2x_1 - 3)e^{x_1} + (x_1^2 + 2x_2^2 - 3x_1 - 8x_2 + 2)e^{x_1}$$

$$= (x_1^2 + 2x_2^2 - x_1 - 8x_2 - 1)e^{x_1}$$

$$\frac{\partial^2}{\partial x_1^2} f(\mathbf{x})|_{\mathbf{x}=[0\ 0]^T} = -1$$

$$f(\mathbf{x}) = f(\mathbf{a}) + (\mathbf{x} - \mathbf{a})^T \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}} + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T \mathbf{H} (\mathbf{x} - \mathbf{a})^T + \ldots$$

$$f(\mathbf{x}) = f(\mathbf{a}) + (\mathbf{x}-\mathbf{a})^T \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{a}} + \frac{1}{2}(\mathbf{x}-\mathbf{a})^T \mathbf{H} (\mathbf{x}-\mathbf{a})^T + \ldots$$

$$\frac{\partial^2}{\partial x_1 \partial x_2} f(\mathbf{x}) = (4x_2 - 8)e^{x_1}$$

$$\frac{\partial^2}{\partial x_1 \partial x_2} f(\mathbf{x})|_{\mathbf{x}=[0\ 0]^T} = -8$$

$$\frac{\partial^2}{\partial x_2^2} f(\mathbf{x}) = 4e^{x_1}$$

$$\frac{\partial^2}{\partial x_2 \partial x_1} f(\mathbf{x}) = (4x_2 - 8)e^{x_1}$$

$$\frac{\partial^2}{\partial x_2^2} f(\mathbf{x})_{\mathbf{x}=[0\ 0]^T} = 4$$

$$\frac{\partial^2}{\partial x_2 \partial x_1} f(\mathbf{x})|_{\mathbf{x}=[0\ 0]^T} = -8$$

Thus

$$f(x) \approx 7 + [x_1\ x_2] \times [2\ -8]^T + \frac{1}{2}[x_1\ x_2] \times \begin{pmatrix} -1 & -8 \\ -8 & 4 \end{pmatrix} \times [x_1\ x_2]^T$$

Automatic Differentiation

- *Automatic Differentiation lets you compute exact derivatives in constant time*

Basic problem in machine learning is **function approximation**.

Given some inputs ˆx and some outputs ˆy, and fit some function f such that it predicts ˆy well.

The basic idea of neural networks is very simple.

- **Make up a big nonlinear function f (x,$\boldsymbol{\theta}$), parameterized by some vector $\boldsymbol{\theta}$.**
- **pick some loss function, measuring how well f predicts y.**
- **Finally, use some local optimization algorithm to minimize the empirical risk**

$$\sum_{\{\hat{\mathbf{x}},\hat{\mathbf{y}}\}} L\big(\mathrm{f}(\hat{\mathbf{x}};\boldsymbol{\theta}),\hat{\mathbf{y}}\big).$$

A simple example. Imagine that we have some functions f, g, and h, and we are interested in computing the derivative of the function $f(g(h(x)))$ with respect to x. By the chain rule,

$$\frac{d}{dx} f(g(h(x))) = f'(g(h(x))) \cdot g'(h(x)) \cdot h'(x).$$

Consider the function

$$f_k(f_{k-1}(\cdots f_2(f_1(x))\cdots)).$$

The chain rule will give us

$$\frac{d}{dx} f_k(f_{k-1}(\cdots f_2(f_1(x))\cdots)) = f_k'(f_{k-1}(\cdots f_2(f_1(x))\cdots)) \cdot f_{k-1}'(f_{k-2}(\cdots f_2(f_1(x))\cdots)) \cdots f_2'(f_1(x)) \cdot f_1'(x).$$

Deep Learning Scenario

Gradient descent is used to find good model parameters by computing the gradient(using calculus and applying the chain rule) of a learning objective with respect to the parameters of the model.

In certain problems computing Gradients is more expensive than computing the function

Chain rule is used extremely in deep learning algorithms

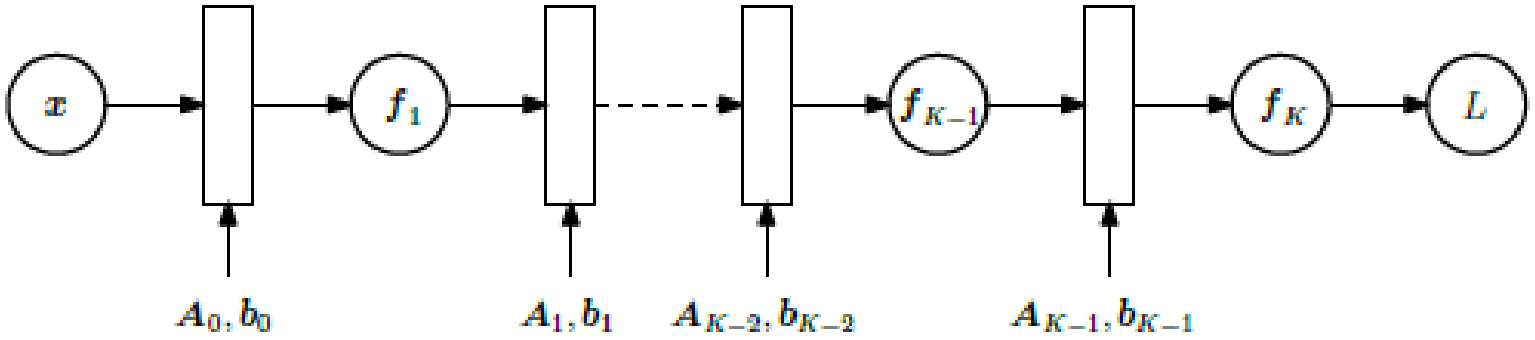


Figure shows the forward pass in a multi-layer neural network to compute the loss L as a function of the inputs x and the parameters A_i; b_i.

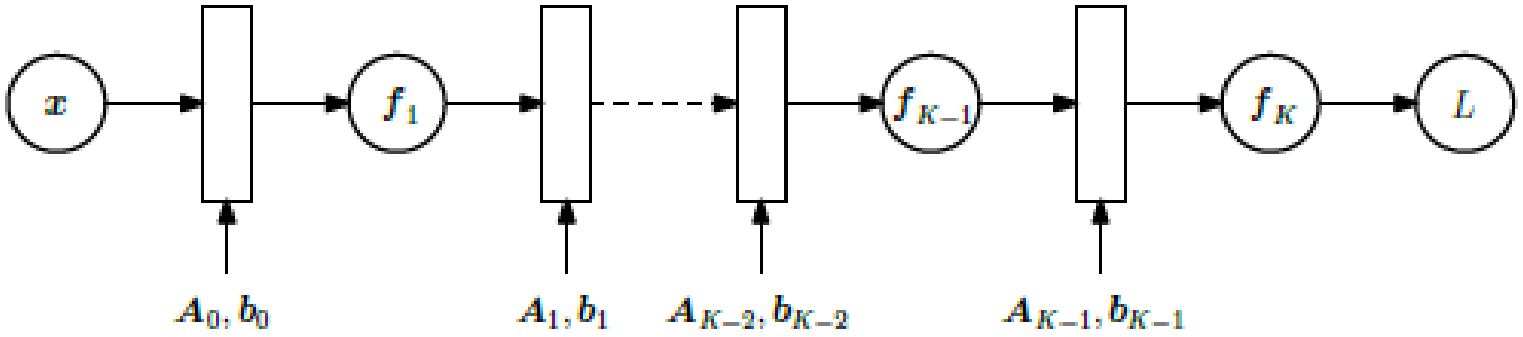
$$f_i\,(x_{i-1}) = \sigma(A_{i-1}x_{i-1}+b_{i-1})$$

where f_i is the function in the layer i , x_{i-1} is the output of i-1 layer, σ is the activation function.

To train these models , gradient of a loss function L is computed with respect to all model parameters Aj, bj, for all j=1,…K

A network structure with inputs x and observations y is defined as

$$f_0 := x$$
$$f_i := \sigma_i(A_{i-1} f_{i-1} + b_{i-1}), \quad i = 1, \ldots, K,$$



To find the values of parameters, Loss(L) is minimized as

$$L(\theta) = \|y - f_K(\theta, x)\|^2$$

Where $\theta = \{A_0, b_0, \ldots, A_{K-1}, b_{K-1}\}$

- For find the gradients with respect to θ, partial derivatives are computed with respect to A_j, b_j, of each layer

$$\frac{\partial L}{\partial \theta_{K-1}} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial \theta_{K-1}}$$

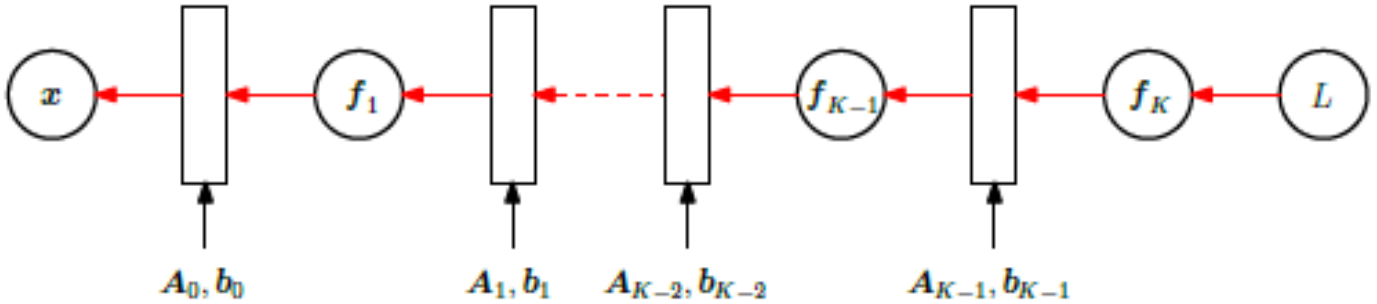
$$\frac{\partial L}{\partial \theta_{K-2}} = \frac{\partial L}{\partial f_K}\boxed{\frac{\partial f_K}{\partial f_{K-1}}\frac{\partial f_{K-1}}{\partial \theta_{K-2}}}$$

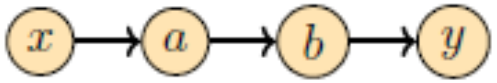
$$\frac{\partial L}{\partial \theta_{K-3}} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial f_{K-1}}\boxed{\frac{\partial f_{K-1}}{\partial f_{K-2}}\frac{\partial f_{K-2}}{\partial \theta_{K-3}}}$$

$$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial f_{K-1}}\cdots\boxed{\frac{\partial f_{i+2}}{\partial f_{i+1}}\frac{\partial f_{i+1}}{\partial \theta_i}}$$

The **orange** terms are partial derivatives of the output of a layer with respect to its inputs, whereas the **blue** terms are partial derivatives of the output of a layer with respect to its parameters.

Backpropagation



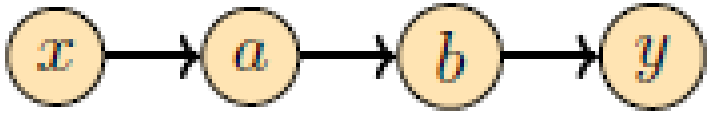


Automatic Differentiation

- Backpropagation is a special case of numerical analysis technique called as Automatic Differentiation
- It is defined as a set of techniques to numerically evaluate the exact (up to machine precision) gradient of a function by working with intermediate variables and applying the chain rule.
- It applies a series of elementary arithmetic operations like addition, and multiplication and elementary functions e.f sin, cos, exp, log .
- By applying the chain rule to these operations, the gradient of quite complicated functions can be computed automatically

Automatic Differentiation

- Automatic differentiation applies to general computer programs and has forward and reverse modes.



$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}b}\frac{\mathrm{d}b}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}x}$$

- Forward mode and reverse mode differ in order of multiplication, due to associative property of multiplication

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \left(\frac{\mathrm{d}y}{\mathrm{d}b}\frac{\mathrm{d}b}{\mathrm{d}a}\right)\frac{\mathrm{d}a}{\mathrm{d}x}, \quad \textit{reverse mode}$$

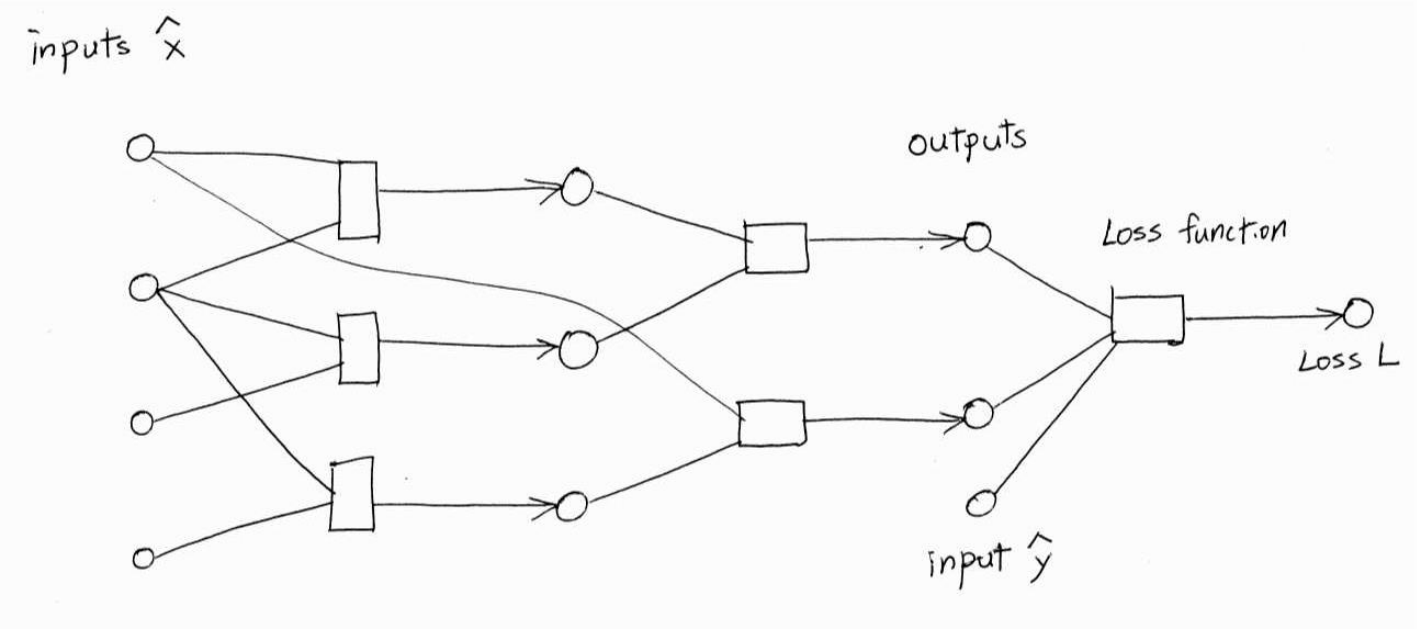
Backpropagation or

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}b}\left(\frac{\mathrm{d}b}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}x}\right) \quad \textit{forward mode},$$

- Where gradients flow with the data from left to right through the graph

Automatic Differentiation

- In the context of neural networks, where the input dimensionality is often much higher than the dimensionality of the labels, the reverse mode is computationally significantly cheaper than the forward mode.



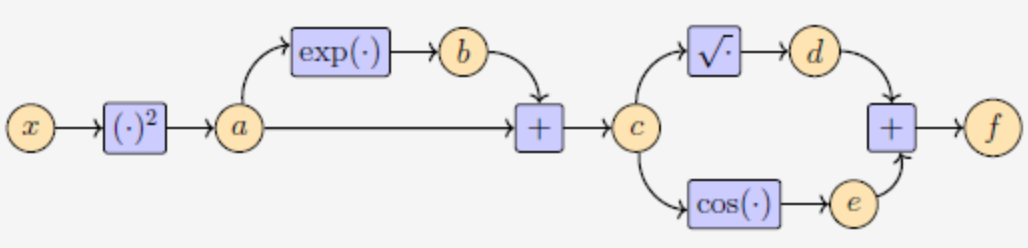
Example

- Consider the function $f(x) = \sqrt{x^2 + \exp(x^2)} + \cos\left(x^2 + \exp(x^2)\right)$
- To implement a function f on a computer, save some computation by using *intermediate variables:*

$$
\begin{aligned}
a &= x^2\,, \\
b &= \exp(a)\,, \\
c &= a + b\,, \\
d &= \sqrt{c}\,, \\
e &= \cos(c)\,, \\
f &= d + e\,.
\end{aligned}
$$

Example

- Figure shows the flow of data and computations required to obtain the function value f.

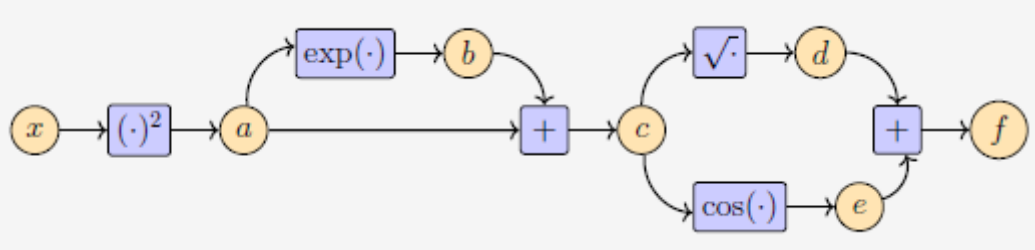


$$\frac{\partial a}{\partial x} = 2x$$
$$\frac{\partial b}{\partial a} = \exp(a)$$

$$\frac{\partial c}{\partial a} = 1 = \frac{\partial c}{\partial b}$$
$$\frac{\partial d}{\partial c} = \frac{1}{2\sqrt{c}}$$
$$\frac{\partial e}{\partial c} = -\sin(c)$$
$$\frac{\partial f}{\partial d} = 1 = \frac{\partial f}{\partial e}\,.$$

Example

- By looking at the computation graph in Figure, we can compute derivative of f by working backward from the output and obtain



$$\frac{\partial f}{\partial c} = \frac{\partial f}{\partial d}\frac{\partial d}{\partial c} + \frac{\partial f}{\partial e}\frac{\partial e}{\partial c}$$
$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c}\frac{\partial c}{\partial b}$$
$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b}\frac{\partial b}{\partial a} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial a}$$
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial x}.$$

Example

- we implicitly applied the chain rule to obtain derivative of f . By substituting the results of the derivatives of the elementary functions, we get

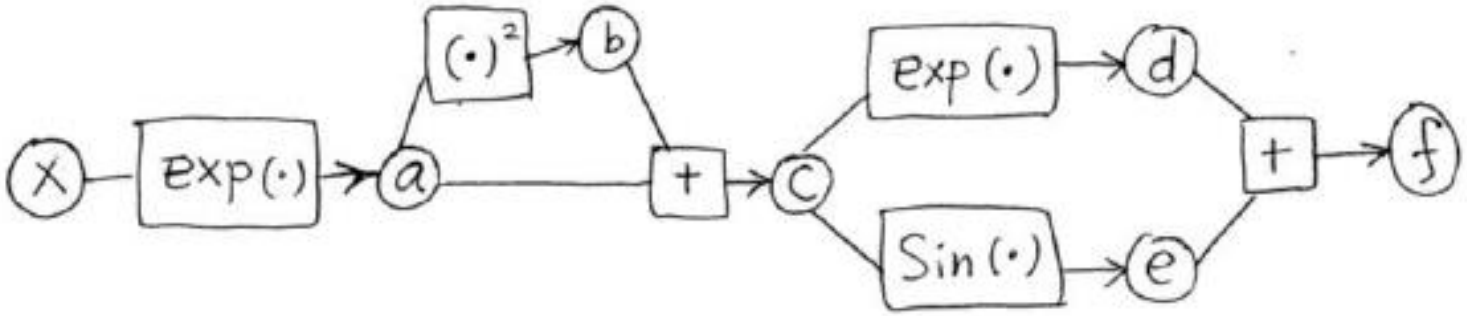
$$\frac{\partial f}{\partial c} = 1 \cdot \frac{1}{2\sqrt{c}} + 1 \cdot (-\sin(c))$$
$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c} \cdot 1$$
$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b} \exp(a) + \frac{\partial f}{\partial c} \cdot 1$$
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a} \cdot 2x\,.$$

Example : Automatic Differentiation

Consider the scalar function

$$f = \exp\left(\exp(x) + \exp(x)^2\right) + \sin\left(\exp(x) + \exp(x)^2\right)$$

$$\begin{aligned} a &= \exp(x) \\ b &= a^2 \\ c &= a + b \\ d &= \exp(c) \\ e &= \sin(c) \\ f &= d + e. \end{aligned} \tag{1}$$



$$
\begin{aligned}
a &= \exp(x) \\
b &= a^2 \\
c &= a + b \\
d &= \exp(c) \\
e &= \sin(c) \\
f &= d + e.
\end{aligned}
$$

$$
\begin{aligned}
\frac{df}{dd} &= 1 \\
\frac{df}{de} &= 1 \\
\frac{df}{dc} &= \frac{df}{dd}\frac{dd}{dc} + \frac{df}{de}\frac{de}{dc} \\
\frac{df}{db} &= \frac{df}{dc}\frac{dc}{db} \\
\frac{df}{da} &= \frac{df}{dc}\frac{dc}{da} + \frac{df}{db}\frac{db}{da} \\
\frac{df}{dx} &= \frac{df}{da}\frac{da}{dx}
\end{aligned}
$$

$$
\begin{aligned}
\frac{df}{dd} &= 1 \\
\frac{df}{de} &= 1 \\
\frac{df}{dc} &= \frac{df}{dd}\exp(c) + \frac{df}{de}\cos(c) \\
\frac{df}{db} &= \frac{df}{dc} \\
\frac{df}{da} &= \frac{df}{dc} + \frac{df}{db}2a \\
\frac{df}{dx} &= \frac{df}{da}\exp(x).
\end{aligned}
$$

Back propogation works backwards from the end of the graph, computing the derivative of each variable, making the use of the derivatives of the children of that variable.

Note : The function is represented as a sequence of basic operations, as in Eq. 1 computing derivates of the same

Automatic Differentiation

- By considering each of the derivatives above as a variable, the computation required for calculating the derivative is of similar complexity as the computation of the function itself.
- The automatic differentiation approach above works whenever we have a function that can be expressed as a computation graph, where the elementary functions are differentiable.
- The function may be a mathematical function or a computer program.
- However, not all computer programs can be automatically differentiated, e.g., if we cannot find differential elementary functions. Programming structures, such as for loops and if statements, require more care as well.

References

- https://chelseatroy.com/2017/02/23/machine-learning-intuition-understanding-taylor-series-approximation/
- http://www.ms.uky.edu/~rbrown/courses/ma113.f.12/l24-linear.pdf
- https://medium.com/@isaacng/mathematics-for-machine-learning-multivariate-calculus-7102c7a586c6
- https://web.ma.utexas.edu/users/m408m/Display14-4-3.shtml#:~:text=The%20Linearization%20of%20a%20function,term%20for%20the%20second%20variable.
- https://study.com/academy/lesson/linear-approximation-in-calculus-formula-examples.html