机器学习与深度学习
——导数与链式法则

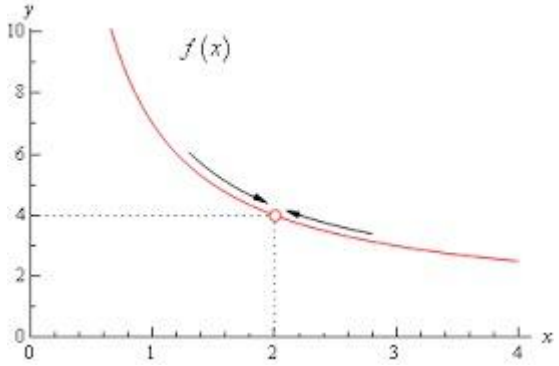
Personal Website: https://www.miaopeng.info/
Email: miaopeng@stu.scu.edu.cn
Github: https://github.com/MMeowhite
Youtube: https://www.youtube.com/@pengmiao-bmm

目录章节

CONTENTS

▶ 极限：引言

- 从“接近”出发：
  - 当我们说‘某物越来越接近某个值’时，这个‘接近’到底意味着什么？
  - 例如：水温逐渐升高，最终接近100℃，这‘接近’是怎么描述的？；一辆车不断减速，速度趋近于0，这个过程如何用数学描述？
- 这些问题的本质：**描述变量如何“无限接近”某个值，是数学上处理“趋近过程”的基础工具。**
- 例如：



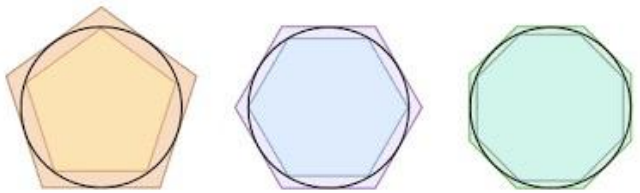
  - 一系列动点沿着数轴缓慢移动，它们从左边或右边不断靠近x=2，虽然从未真正到达x=2，但它们的位置越来越接近 2。
  - 我们关心的是什么？不是动点的出发点，也不是它是否真正到达 2，而是当它们无限逼近 2 时，函数值 f(x)逼近的那个数。
  - 这就是极限：$\lim_{x \to 2} f(x) = L$
  - 它精确刻画了“**当x逼近某点时，f(x)靠近的值**”。

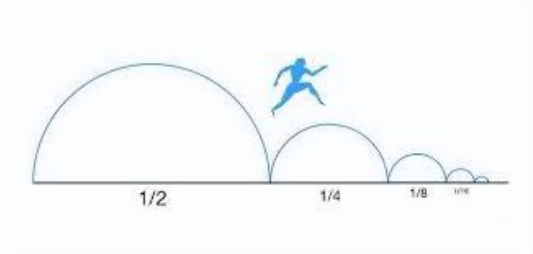
极限是让变化可控、可测的数学语言，如何用数学精确描述‘无限接近’，是理解变化的关键。

▶ 极限：发展历史

➢ 古代萌芽：

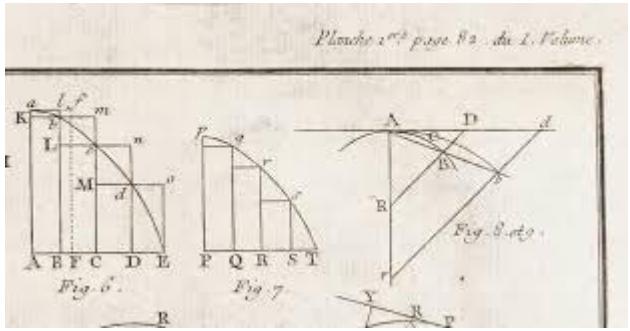
- 阿基米德等人用“穷竭法”计算面积和体积，实际上用到了极限思想，但没有明确极限概念。
- 庄子《庄子·天下篇》：一尺之棰，日取其半，万世不竭；老子《道德经》：道生一，一生二，二生三，三生万物。





➢ 17世纪——微积分的诞生

- 牛顿和莱布尼茨：两者独立创立微积分，使用“无穷小量”概念处理变化率和面积问题。
- 但“无穷小”没有严格定义，导致争议和逻辑问题。（历史上的第一次“数学危机”）。



▶ 极限：发展历史

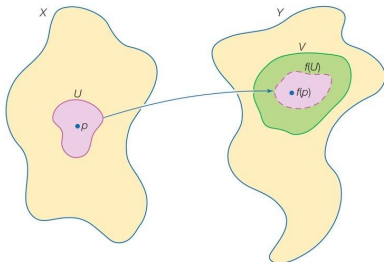
➢ 19世纪——极限的严格定义

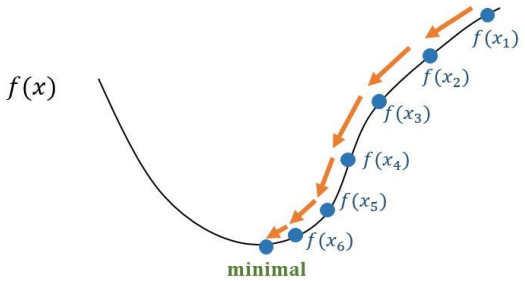
- 柯西和魏尔斯特拉斯，引入了“ϵ-δ”的定义，正式建立了极限理论的严密基础。
- 解决了无穷小量的模糊问题，使微积分走向严格化。

$$\lim_{x \to a} f(x) = L \quad \Longleftrightarrow \quad \forall \varepsilon > 0,\ \exists \delta > 0,\ 0 < |x - a| < \delta \to |f(x) - L| < \varepsilon.$$

➢ 现代发展

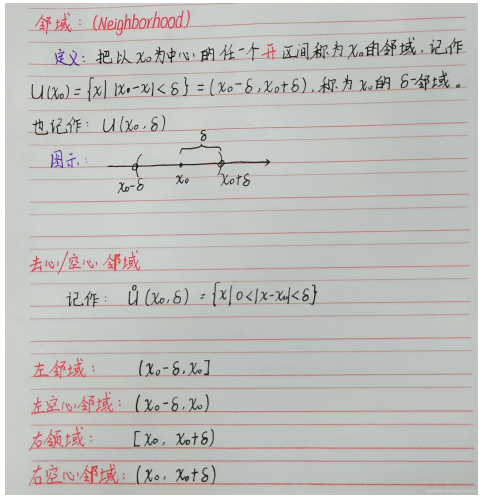
- 极限概念成为数学分析的核心工具，广泛应用于微积分、函数理论、拓扑等领域。
- 计算机科学和机器学习中的许多算法，尤其是优化算法，也基于极限思想。





▶ 什么是极限？

- 极限：当函数的自变量在变化的过程中，逐渐向某一个确定的数值不断地逼近而“永远不能够重合到A”的过程中，此变量描述的变化过程，就是极限。
  - 想象在靠近一个目标：你一步一步走向它，虽然还没有真正到达，但你离它越来越近。
  - 极限描述的就是：**当你无限接近某个点时，函数值会无限接近哪个数**。
- 在数学上，我们用符号lim进行表述极限：$\lim_{x \to a} f(x) = L$
  - 意思是：当x越来越靠近a时，f(x) 越来越靠近L。
- 关键：如何量化“靠近”呢？怎么用数学的语言表述x->a以及f(x) -> L？
  - 用邻域+动态变化的概念进行描述（ϵ-δ语言）：



$$\lim_{x \to a} f(x) = L \iff \forall \varepsilon > 0,\ \exists \delta > 0,\ 0 < |x-a| < \delta \to |f(x) - L| < \varepsilon.$$

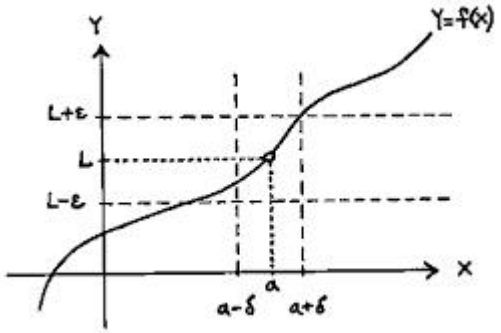
▶ 极限的严谨语言：ϵ-δ语言

- ϵ-δ语言：对于任何正数ε，都能够找到一个正数δ，当x满足0 < |x - a| < δ，对于满足上式的x都有0 < |f(x) - L| < ε。

$$\lim_{x \to a} f(x) = L \iff \forall \varepsilon > 0,\ \exists \delta > 0,\ 0 < |x - a| < \delta \to |f(x) - L| < \varepsilon.$$

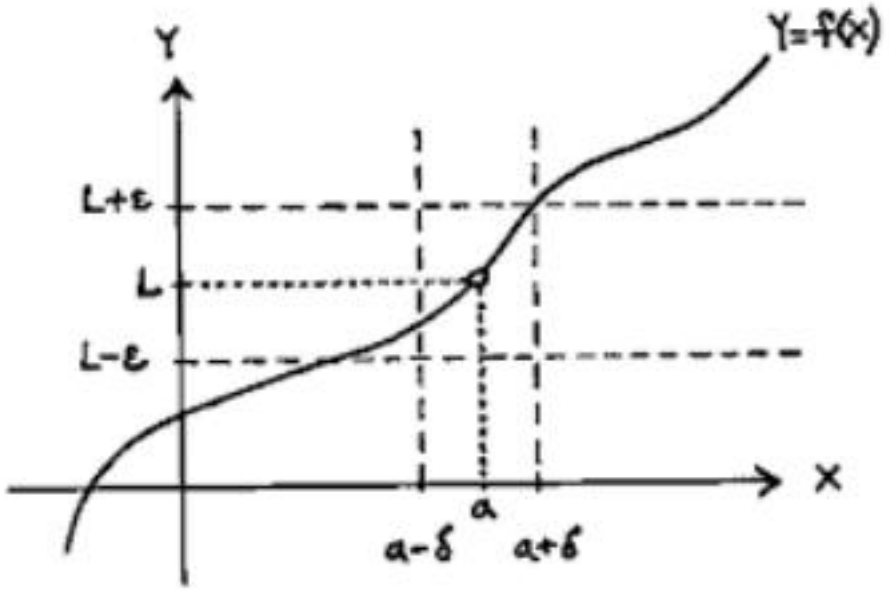
  - f(x)无论距离L有多近，它始终不是L，在f(x)与L之间总是能找到一个数字（而不是无穷小），使这个数字与f(x)与b的差为ε。对于每个ε都存在一个大于零的δ，使得满足上式的x属于a±δ。
  - 其实ε 是"error"(误差)、δ 是"distance"(距离)的首字母。实际上，是柯西在他的著作中用ε来表示"error"的。

- 几何意义：



  - ε－δ语言描述的是：当 x进入点a附近一个足够小的区间时，f(x) 的值会被压进极限L附近一个足够小的区间。
  - ε 控制“我想要多接近”；δ 控制“我得把输入调多小才能达到这个接近”。

▶ 怎么理解ε-δ语言？



- ε（epsilon）：
  - 它是我们想要控制的输出精度。例如：我想让函数值f(x)距离极限L不超过 0.01（这就是ε=0.01）。
- δ（delta）：
  - 它是我们能调节的输入范围。例如：只要把x控制在离a不超过 0.001 的范围内（$|x-a|<\delta$），函数值f(x)就会自动落在我想要的精度范围内。
- 逻辑关系：
  - 无论别人提出多么苛刻的精度要求ε，我们总能找到一个足够小的δ，让x只要落在(a-δ,a+δ)这个范围里，f(x) 就落在(L-ε,L+ε)这个范围里。

▶ ϵ–δ语言的应用

➢ 如何使用ϵ–δ语言证明：针对函数f(x) = x²，对于任何正数ε，都能够找到一个正数δ，当x满足0 < |x − 3| < δ，对于满足上式的x都有0 < |f(x) − 9| < ε。

$$f(x)=x^2, \lim_{x\to 3} f(x)=9 \iff \forall \varepsilon>0,\ \exists \delta>0,\ 0<|x-3|<\delta \to |f(x)-9|<\varepsilon.$$

- 反推法：如果要满足|f(x) − 9| < ε，就需要满足|x² − 9| < ε，即，

$$\sqrt{\epsilon-9}<x<\sqrt{\epsilon+9}$$

- 现在的目标是找到δ，使得当|x − 3| < δ，有上述的不等式成立，两边同时减3，得到：

$$\sqrt{\epsilon-9}-3<x-3<\sqrt{\epsilon+9}-3$$

- 取：

$$\delta=\sqrt{\epsilon+9}-3$$

- 即可成立。因为

$$\forall \varepsilon>0, \text{if } \delta=\sqrt{\epsilon+9}-3 \to |x-3|<\delta=\sqrt{\epsilon+9}-3$$

$$|x^2-9|=|x+3||x-3|<(\delta+6)\delta=(\sqrt{\epsilon+3}+3)(\sqrt{\epsilon+3}-3)=\epsilon$$

极限的四则运算与复合运算

- 两个存在极限的函数，两个函数的和差积商是两个函数各自极限的和差积商：

  - 假设：

$$\lim_{x\to a} f(x) = A, \quad \lim_{x\to a} g(x) = B$$

  - 极限语言表述如下：

$$(1) \quad \lim_{x\to a}\big(f(x)+g(x)\big) = A+B$$

$$(2) \quad \lim_{x\to a}\big(f(x)-g(x)\big) = A-B$$

$$(3) \quad \lim_{x\to a}\big(f(x)\cdot g(x)\big) = A\cdot B$$

$$(4) \quad \lim_{x\to a}\frac{f(x)}{g(x)} = \frac{A}{B} \quad B\neq 0).$$

- 极限的复合运算：

$$\lim_{x\to x_0} f(x) = a, \quad \lim_{x\to a} g(x) = b, \quad \lim_{x\to x_0} g(f(x)) = b$$

▶ 小结

➢ 极限的ε-δ语言：

$$\lim_{x\to a} f(x) = L \quad \Longleftrightarrow \quad \forall \varepsilon > 0,\ \exists \delta > 0,\ 0 < |x-a| < \delta \to |f(x) - L| < \varepsilon.$$

- 无论你要求函数值与极限多接近(ε)，总能找到足够小的邻域(δ)让它成立。

➢ 极限的四则运算：

$$(1) \quad \lim_{x\to a} \big(f(x) + g(x)\big) = A + B$$

$$(2) \quad \lim_{x\to a} \big(f(x) - g(x)\big) = A - B$$

$$(3) \quad \lim_{x\to a} \big(f(x) \cdot g(x)\big) = A \cdot B$$

$$(4) \quad \lim_{x\to a} \frac{f(x)}{g(x)} = \frac{A}{B} \quad B \neq 0).$$

- 极限可拆分做加减乘除，但除法要确保分母极限不为零。

极限是研究函数在“接近某点”时行为的工具，是导数、偏导数的基石。

目录章节

CONTENTS

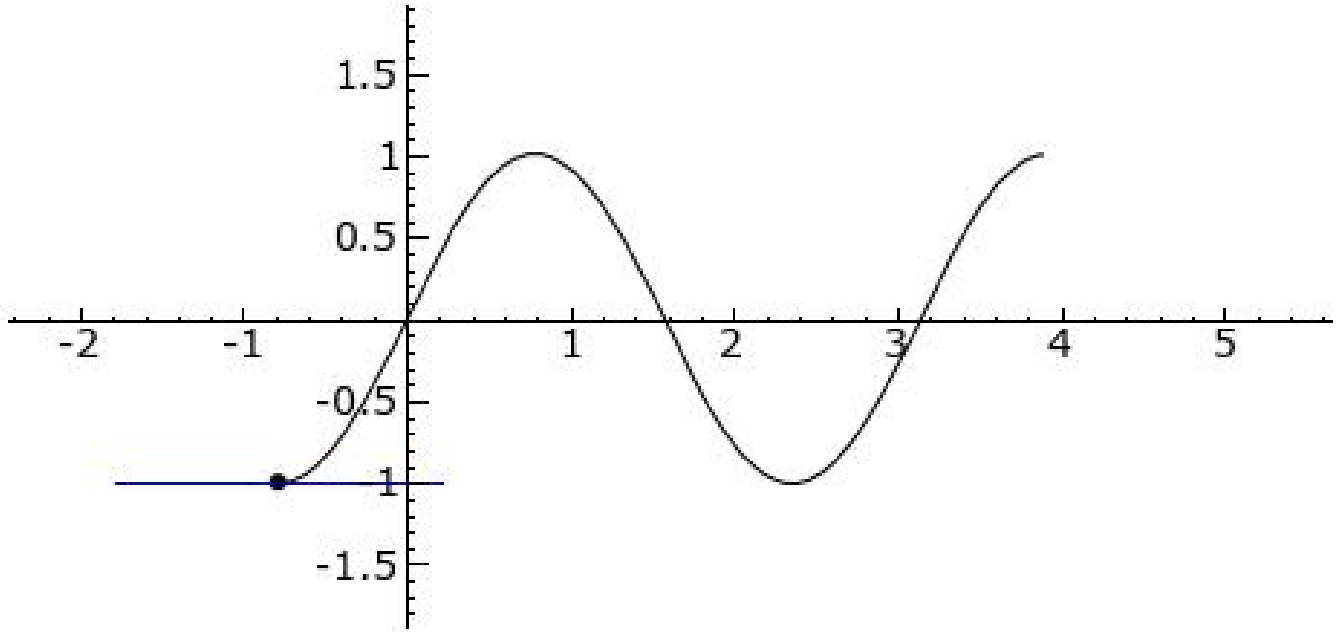
▶ 导数：引言

- 从“变化”出发：
  - 当一个量随着另一个量变化时，我们常常想知道：它到底变化得有多快？
  - 例如：位置随着时间变化 -> 速度；温度随着时间变化 -> 升/降温速度；股价随时间变化 -> 涨跌趋势。
- 这些问题的本质：**在某个瞬间，函数到底以什么速率在变化？**
- 平均变化率 vs 瞬时变化率：
  - 如果你知道一辆车 1 小时跑了 60 公里，平均速度是 60 km/h。
  - 但如果你想知道**某一时刻**它跑得多快？平均变化率（割线斜率）无法精确反映某个时刻的状态，我们需要瞬时变化率。

导数刻画的是函数在某一点的瞬时变化率——也就是它在那一刻的变化方向与快慢。

▶ 什么是导数？

- 导数描述的是函数在某一点的**瞬时**变化率，即它在这一刻变化的方向和快慢。
- 从几何上看，导数是曲线在**该点的切线斜率**。



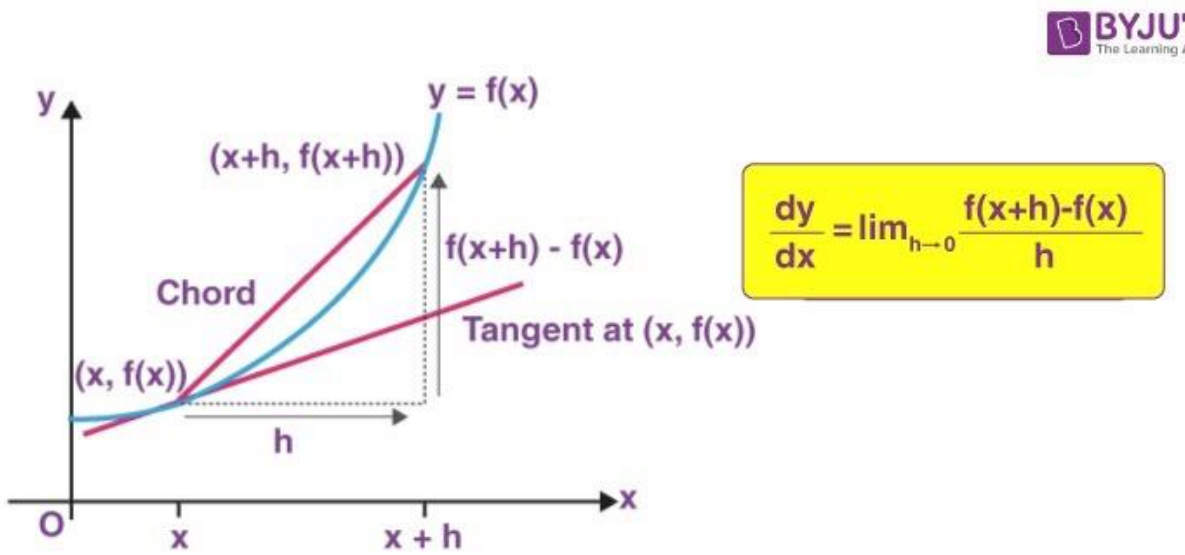
本质上，导数揭示了输入的微小变化如何影响输出，是理解变化最核心的工具。

▶ 导数的核心思想：极限下的变化

- 数学上，瞬时变化率用极限来描述：

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

- 几何意义：这是函数曲线在该点的切线斜率。



割线 → h 变小 → 切线（变化率极限）

- 如果导数是正的，函数在上升；如果是负的，函数在下降；绝对值越大，变化越快。

导数刻画某一点的变化趋势，不仅仅是“变大还是变小”，还告诉我们“变化的快慢”。

▶ 导数如何计算？

- 1）幂函数

$$f(x) = x^n, \quad f'(x) = \lim_{h \to 0} \frac{(x+h)^n - x^n}{h}$$

- 利用二项式定理展开：$(x+h)^n = x^n + nx^{n-1}h + \frac{n(n-1)}{2}x^{n-2}h^2 + \cdots$

- 带回得到：

$$f'(x) = \lim_{h \to 0} \frac{nx^{n-1}h + \frac{n(n-1)}{2}x^{n-2}h^2 + \cdots}{h} = \lim_{h \to 0} (nx^{n-1} + \frac{n(n-1)}{2}x^{n-2}h + \cdots)$$

- 当h -> 0：高阶项消失：

$$f'(x) = nx^{n-1}$$

▶ 导数如何计算？

- 2）指数函数

$$f(x) = e^x \quad f'(x) = \lim_{h \to 0} \frac{e^{x+h} - e^x}{h}$$

- 进一步计算：

$$f'(x) = \lim_{h \to 0} \frac{e^x(e^h - 1)}{h} = e^x \lim_{h \to 0} \frac{e^h - 1}{h}$$

- 对e^h在h=0处进行泰勒展开：

$$e^h = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \cdots$$

- 进一步计算：

$$f'(x) = e^x \lim_{h \to 0} (1 + \frac{h}{2!} + \frac{h^2}{3!} + \cdots)$$

- 所以：

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

导数如何计算？

- 3）对数函数

$$f(x)=\ln x,\quad f'(x)=\lim_{h\to 0}\frac{\ln(x+h)-\ln x}{h}$$

- 根据对数函数的运算法则：$f'(x)=\lim_{h\to 0}\frac{\ln\frac{x+h}{x}}{h}=\lim_{h\to 0}\frac{\ln(1+\frac{h}{x})}{h}$

- 进一步计算：

$$u=\frac{h}{x},\quad f'(x)=\frac{1}{x}\lim_{u\to 0}\frac{\ln(1+u)}{u}$$

- 利用泰勒展开式：$\ln(1+u)=u-\frac{u^2}{2}+\frac{u^3}{3}-\cdots$

- 所以：

$$f'(x)=\frac{1}{x}\lim_{h\to 0}(1-\frac{u}{2}+\frac{u^2}{3})=\frac{1}{x}$$

常见导数

$\frac{d}{dx}(c) = 0$	$\frac{d}{dx}(cx) = c$	$\frac{d}{dx}(x^c) = cx^{c-1}$
$\frac{d}{dx}(c^x) = c^x \ln(c) \quad c > 0$	$\frac{d}{dx}(x^x) = x^x(1 + ln\, x)$	$\frac{d}{dx}(e^x) = e^x$
$\frac{d}{dx}\left(\frac{1}{x}\right) = \frac{-1}{x^2}$	$\frac{d}{dx}\left(\frac{1}{x^2}\right) = \frac{-2}{x^3}$	$\frac{d}{dx}\left(\frac{1}{x^n}\right) = \frac{-n}{x^{n+1}}$
$\frac{d}{dx}(\sqrt{x}) = \frac{1}{2\sqrt{x}} \quad x > 0$	$\frac{d}{dx}(\sqrt[3]{x}) = \frac{1}{3 \cdot \sqrt[3]{x^2}}$	$\frac{d}{dx}(\sqrt[n]{x}) = \frac{1}{n \cdot \sqrt[n]{x^{n-1}}}$
$\frac{d}{dx}\left(\frac{1}{\sqrt{x}}\right) = \frac{-1}{2\sqrt{x^3}}$	$\frac{d}{dx}\left(\frac{1}{\sqrt[3]{x}}\right) = \frac{-1}{3 \cdot \sqrt[3]{x^4}}$	$\frac{d}{dx}\left(\frac{1}{\sqrt[n]{x}}\right) = \frac{-1}{n \cdot \sqrt[n]{x^{n+1}}}$
$\frac{d}{dx}(\ln x) = \frac{1}{x} \quad x > 0$	$\frac{d}{dx}(x \cdot \ln x) = \ln x + 1$	$\frac{d}{dx}(\log_c x) = \frac{1}{x \ln c} \quad c > 0 \quad c \neq 1$
$\frac{d}{dx}\left(\frac{1}{lnx}\right) = \frac{-1}{x(\ln x)^2}$	$\frac{d}{dx}\left(\frac{1}{x \cdot \ln x}\right) = \frac{-(\ln x + 1)}{(x \cdot \ln x)^2}$	$\frac{d}{dx}\left(\frac{1}{\log_c x}\right) = \frac{-1}{x \cdot \ln c \cdot (\log_c x)^2}$
$\frac{d}{dx}\left(\frac{1}{x+1}\right) = \frac{-1}{(x+1)^2}$	$\frac{d}{dx}\left(\frac{1}{(x+1)^2}\right) = \frac{-2}{(x+1)^3}$	$\frac{d}{dx}\left(\frac{1}{(x+1)^n}\right) = \frac{-n}{(x+1)^{n+1}}$
$\frac{d}{dx}\left(\frac{1}{\sqrt{x+1}}\right) = \frac{-1}{2 \cdot \sqrt{(x+1)^3}}$	$\frac{d}{dx}\left(\frac{1}{\sqrt[3]{x+1}}\right) = \frac{-1}{3 \cdot \sqrt[3]{(x+1)^4}}$	$\frac{d}{dx}\left(\frac{1}{\sqrt[n]{x+1}}\right) = \frac{-1}{n \cdot \sqrt[n]{(x+1)^{n+1}}}$

$\frac{d}{dx}\sin x = \cos x$	$\frac{d}{dx}\sinh x = \cosh x$
$\frac{d}{dx}\cos x = -\sin x$	$\frac{d}{dx}\cosh x = \sinh x$
$\frac{d}{dx}\tan x = \sec^2 x = \frac{1}{\cos^2 x}$	$\frac{d}{dx}\tanh x = 1 - \tanh^2 x = \mathrm{sech}^2 x$
$\frac{d}{dx}\cot x = -\csc^2 x = -\frac{1}{\sin^2 x}$	$\frac{d}{dx}coth\, x = -csch^2 x$
$\frac{d}{dx}\csc x = -\csc x \cot x$	$\frac{d}{dx}\mathrm{csch}\, x = -\mathrm{csch}\, x \coth x$
$\frac{d}{dx}\sec x = sec\, x\, tan\, x$	$\frac{d}{dx}\mathrm{sech}\, x = -\mathrm{sech}\, x \tanh x$

https://www.ambrbit.com/Equations/Derivation/Derivation.htm

▶ 导数的基本运算规则

➢ 1）加法/减法

$$(f(x)+g(x))' = f'(x)+g'(x)$$

$$(f(x)-g(x))' = f'(x)-g'(x)$$

➢ 2）乘法法则：

$$(f(x)\cdot g(x))' = f'(x)g(x)+f(x)g'(x)$$

- “一边求导，另一边保持不变，然后交换，再相加”。

➢ 3）除法法则：

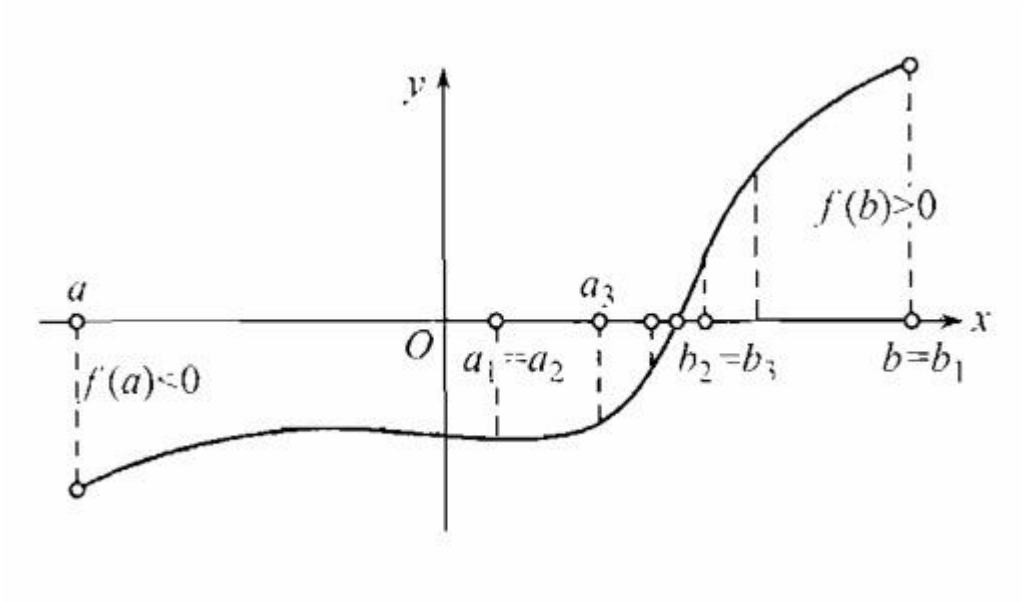
$$\frac{f(x)}{g(x)} = \frac{f'(x)g(x)-f(x)g'(x)}{[g(x)]^2}$$

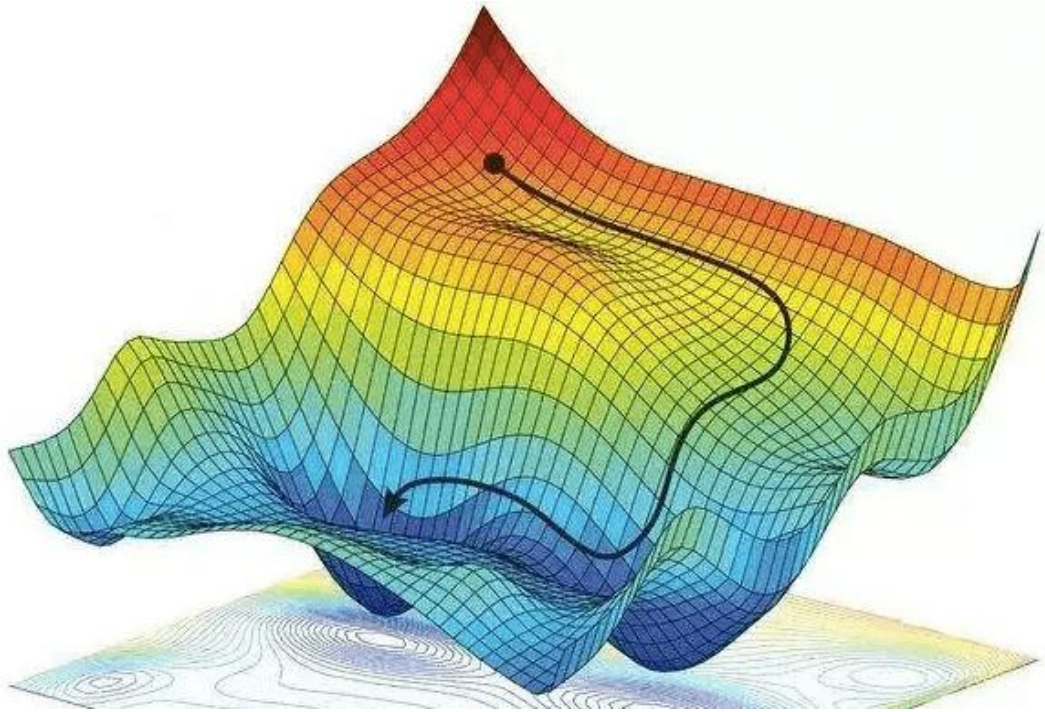
- “分子导×分母 − 分子×分母导，最后除以分母平方”。

加/减 → 各自求导；乘/除 → 用积商法则

▶ 从导数到偏导：为什么需要新的工具？

- 单变量导数：描述一条曲线的斜率。
- **机器学习问题：损失函数L(w1, w2, w3, ...)涉及多个变量。问题：怎么描述函数在多个方向的变化？**





单变量导数告诉我们一条曲线的变化率，偏导数则让我们能够分析高维曲面在每个方向的变化情况，这正是机器学习优化的核心。

▶ 什么是偏导数？

- 偏导数主要针对多元函数f(θ_1, θ_2, ...)。在多元函数里，只让一个变量发生微小变化，其他变量都固定，看函数值变化的快慢。

$$z = f(x, y)$$

- **当y不变的时候（把y视作常量），x改变一点点，z怎么变化？** $\dfrac{\partial f}{\partial x}$

- 严谨定义：设 f(x1, x2,⋯,xn)定义在开集 D⊂R^n，在点a=(a1,⋯,an) 处，对第i个变量的偏导数定义为：

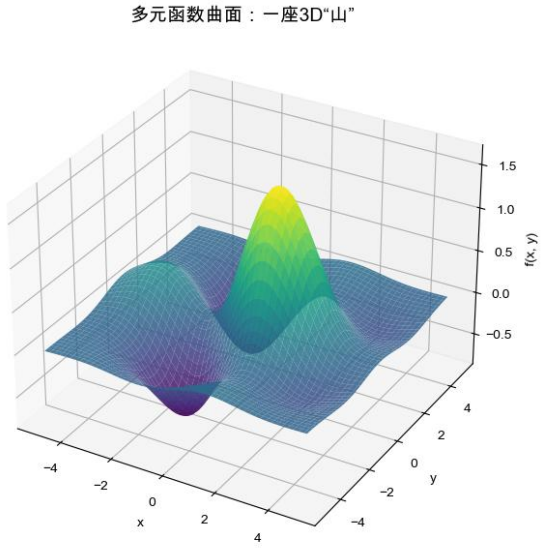
$$\frac{\partial f}{\partial x_i}(\mathbf{a}) = \lim_{h \to 0} \frac{f(a_1, ..., a_i + h, ..., a_n) - f(a_1, ..., a_i, ..., a_n)}{h}$$

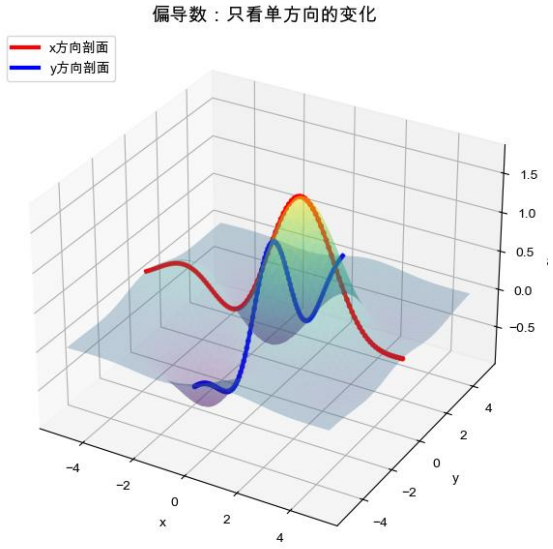
- 如果极限存在，即定义了偏导数。

偏导数就是多元函数在某个方向上的瞬时变化率。

偏导数的核心思想：单向看变率

- **分离变量影响**：多元函数f(x_1, x_2, x_3, ...)依赖多个变量，但我们可以“冻结”其他变量，只研究一个变量的变化对f的影响。
- **局部变化率**：偏导数本质上还是变化率，但是局部的、单方向的变化率。
- **高维曲面的切线斜率**：偏导数描述高维曲面在某个方向上的“切线斜率”。
- 实际应用：在损失函数中，偏导数告诉我们哪个参数该如何调整，是梯度下降的基础。



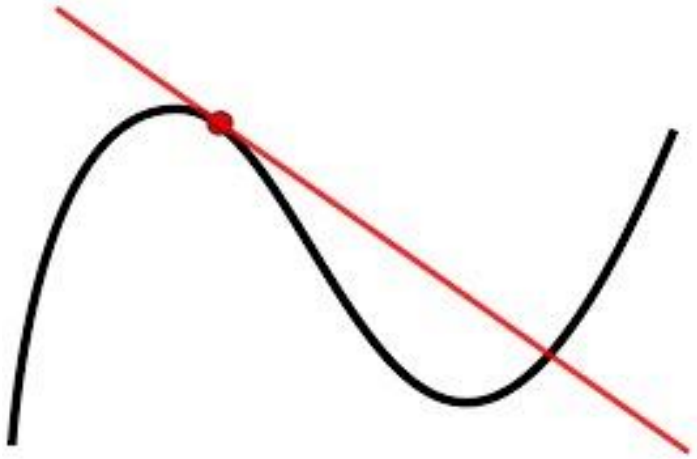


在多元函数中，关注某个变量的变化对函数值的影响，其他变量保持不变。

▶ 小结

➢ 导数的定义：

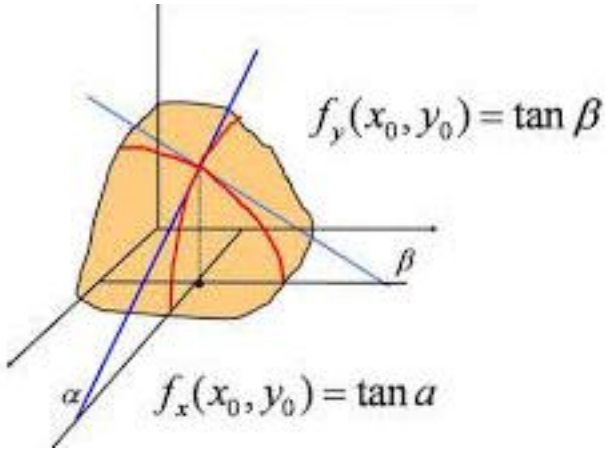
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



- 函数在一点的瞬时变化率（切线斜率）。几何理解：曲线上某点的切线斜率

➢ 偏导数的定义：

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h}$$



- 多元函数中，仅改变一个变量时的变化率。几何理解：高维曲面在某个方向的斜率。

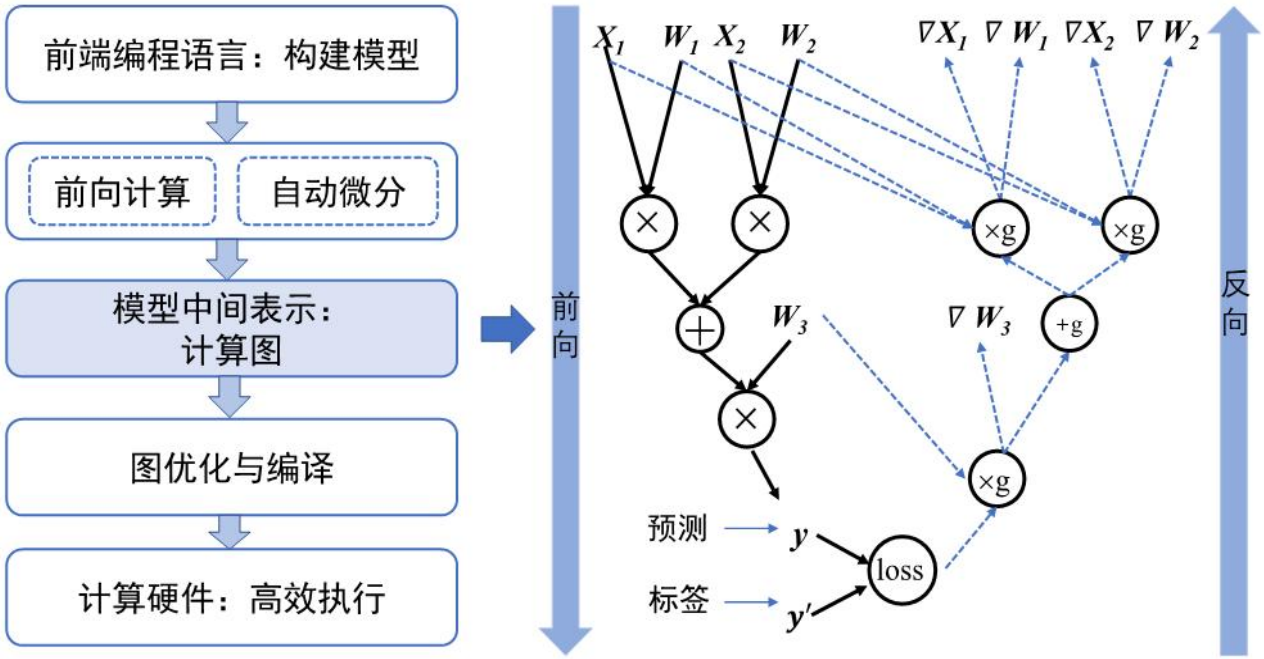
导数：一维变化率；偏导：高维中的单方向变化率。

目录章节

CONTENTS

链式法则：引言与核心思想

- 链式法则：本质就是变化率的传递。
- 假设我们有一个函数，这里以计算图举例：



- 每个节点表示一个基本运算，整个函数是由这些运算层层组合而成的。当输入发生变化时，这个变化会逐层传递，最终影响输出。这时我们需要一种系统的方法来计算这种‘变化率的传递’，这就是链式法则的核心：**通过分解复杂函数，把整体变化率拆成各层局部变化率的乘积或加权和，实现从输入到输出的导数传递**。

链式法则告诉我们：复杂函数的变化率，可以拆解为每一层局部变化率的传递与叠加。

参考资料：https://openmlsys.github.io/chapter_computational_graph/background_and_functionality.html

单变量链式法则

- 一层嵌套：导数的传递

$$y = f(u), u = g(x) \rightarrow \frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

- 例如复合函数：

$$y = (3x^2 + 1)^5$$

  - 将函数进行拆解：

$$y = u^5, \quad u = 3x^2 + 1$$

  - 然后使用链式法则：

$$\frac{dy}{du} = 5u^4, \frac{du}{dx} = 6x \rightarrow \frac{dy}{dx} = 30x(3^2 + 1)^4$$

复合函数的变化率由外函数在内函数处的变化率和内函数自身的变化率共同决定，二者相乘得到总体导数。

▶ 多变量链式法则

➢ 多个中间变量：偏导的传递

$$z = f(u, v), u = g(x, y), v = h(x, y)$$

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x}$$

➢ 例如复合函数：

$$z = (xy)^2 + (x + y^2)^3$$

- 将函数进行拆解：

$$z = u^2 + v^3, \quad u = g(x, y) = xy, \quad v = h(x, y) = x + y^2$$

- 先分别计算外层和内层偏导，然后在利用链式法则即可求出结果：

$$\frac{\partial z}{\partial u} = 2u, \quad \frac{\partial z}{\partial v} = 3v^2$$

$$\frac{\partial u}{\partial x} = y, \quad \frac{\partial u}{\partial y} = x$$

$$\frac{\partial v}{\partial x} = 1, \quad \frac{\partial v}{\partial y} = 2y$$

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} \cdot \frac{\partial u}{\partial x} + \frac{\partial z}{\partial v} \cdot \frac{\partial v}{\partial x}$$

$$\frac{\partial z}{\partial y} = \frac{\partial z}{\partial u} \cdot \frac{\partial u}{\partial y} + \frac{\partial z}{\partial v} \cdot \frac{\partial v}{\partial y}$$

$$\frac{\partial z}{\partial x} = 2u \cdot y + 3v^2 \cdot 1 = 2(xy)y + 3(x + y^2)^2$$

$$\frac{\partial z}{\partial y} = 2u \cdot x + 3v^2 \cdot 2y = 2(xy)x + 6y(x + y^2)^2$$

▶ 扩展：链式法则的推导——单变量

➢ 设

$$f = f(u), u = g(x)$$

➢ 根据导数的意义：

$$\frac{dy}{dx} = \lim_{h \to 0} \frac{f(g(x+h) - f(g(x)))}{h}$$

- 令

$$\Delta u = g(x+h) - g(x)$$

- 则当 h -> 0, Δu -> 0
- 分为两步写：

$$\frac{f(g(x+h) - f(g(x)))}{h} = \frac{f(\Delta u)}{h} = \frac{f(\Delta u)}{\Delta u} \cdot \frac{\Delta u}{h}$$

- 取极限，即可得到：

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

扩展：链式法则的推导——多变量

- 设

$$z = f(u, v), u = g(x, y), v = h(x, y)$$

- 对z做全微分：

$$dz = \frac{\partial f}{\partial u}du + \frac{\partial f}{\partial v}dv$$

  - du和dz可以写作：

$$du = \frac{\partial g}{\partial x}dx + \frac{\partial g}{\partial y}dy$$

$$dv = \frac{\partial h}{\partial x}dx + \frac{\partial h}{\partial y}dy$$

  - 代入即可得到：

$$dz = \frac{\partial f}{\partial u}\left(\frac{\partial g}{\partial x}dx + \frac{\partial g}{\partial y}dy\right) + \frac{\partial f}{\partial v}\left(\frac{\partial h}{\partial x}dx + \frac{\partial h}{\partial y}dy\right)$$

  - 展开即可得到：

$$dz = \left(\frac{\partial f}{\partial u}\frac{\partial g}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial h}{\partial x}\right)dx + \left(\frac{\partial f}{\partial u}\frac{\partial g}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial h}{\partial y}\right)dy$$

  - 对比dy和dx的系数即可得到：

$$\frac{\partial z}{\partial x} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x}, \quad \frac{\partial z}{\partial y} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y}$$

▶ 小结

- 链式法则的基本形式：

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

  - 复合函数求导 = 外层导数 x 内层导数

- 链式法则的多元形式：

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x}$$

  - 多个中间变量时，逐路径相乘，再相加

- 直观理解：像“流水线”：每一步变化都传递到下游，导数=局部变化 × 变化的传递。

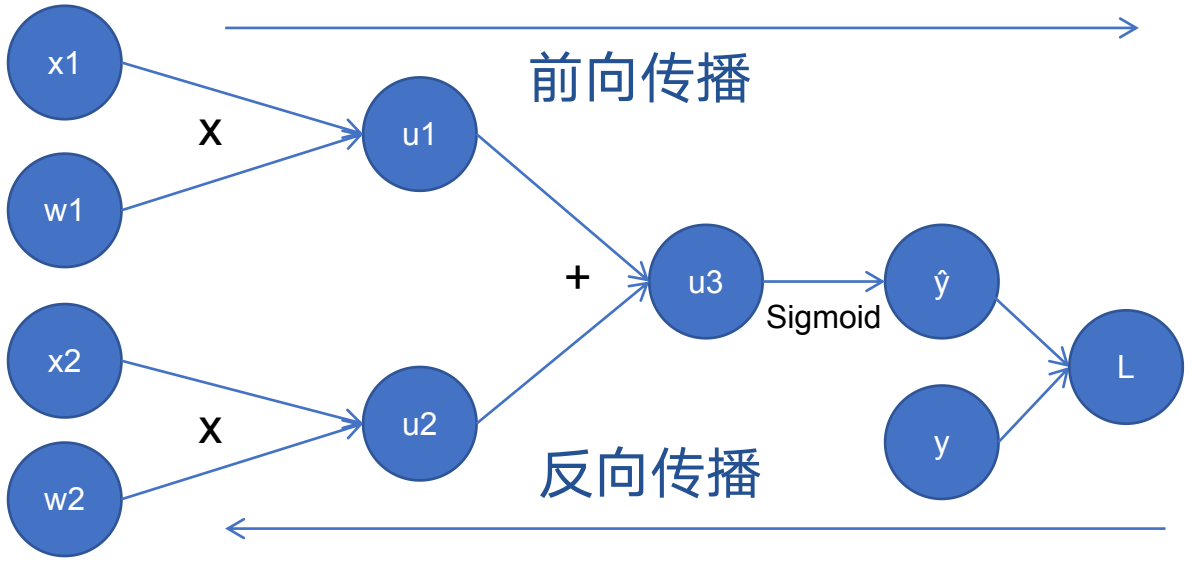
链式法则揭示：复杂系统的整体变化，可分解为每个环节的局部变化乘积——反向传播正是沿着这条链逐层传递梯度。

目录章节

CONTENTS

什么是计算图？

- 计算图：是指将计算过程分解为一系列基本运算，用节点（变量/中间结果）和有向边（运算关系）表示，使得复杂函数的计算与求导可以在图结构上逐步完成。
  - 节点：在神经网络中，输入层、隐藏层和输出层各个操作都可以表示为节点。
  - 边：连接各个节点，表示数据流向或者依赖关系。



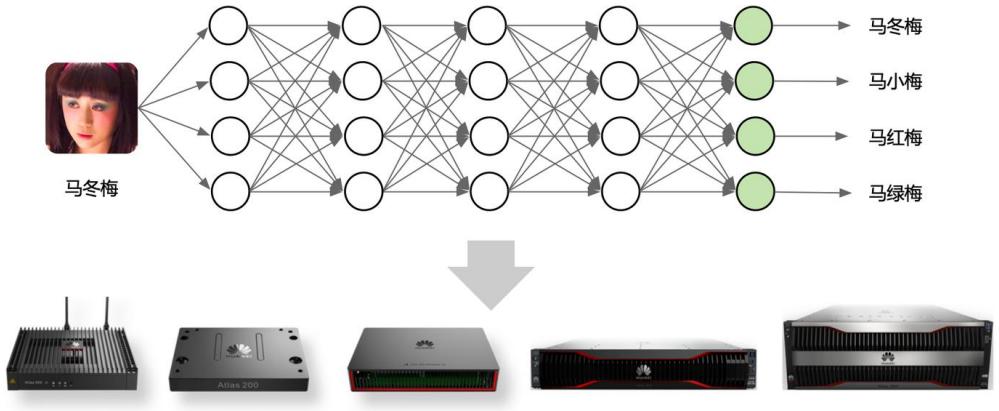
$$L = \frac{1}{2}(\hat{y} - y)^2, \quad \hat{y} = \sigma(u_3), \quad u_3 = u_1 + u_2, \quad u_1 = w_1 \cdot x_1, \quad u_2 = w_2 \cdot x_2$$

- 任务：L是损失函数，我们的任务就是减小损失函数的输出。
- 我们可以将L = f(w_1, w_2)，通过计算图+链式法则求出当w_1或者w_2有一个微小的变化时，y的变化情况如何，这就可以确定到底是增大/减小w_1, w_2。

计算图 = 节点表示变量，边表示依赖，复杂函数拆成简单步骤，便于前向计算与反向求导。

为什么需要计算图？

- 分解复杂问题：复杂函数往往由多个简单函数构成，计算图让它们模块化，便于理解和调试。
- 便于自动求导：链式法则可以在图中自然传播，前向算值，反向算梯度，这是自动微分和反向传播的基础。
- 提升效率：计算图结构可以重用中间结果，避免重复计算，尤其是在深度学习等大模型中非常关键。
- 便于可视化与优化：通过图的结构可以直观理解数据流动，并在图上进行并行化或加速。

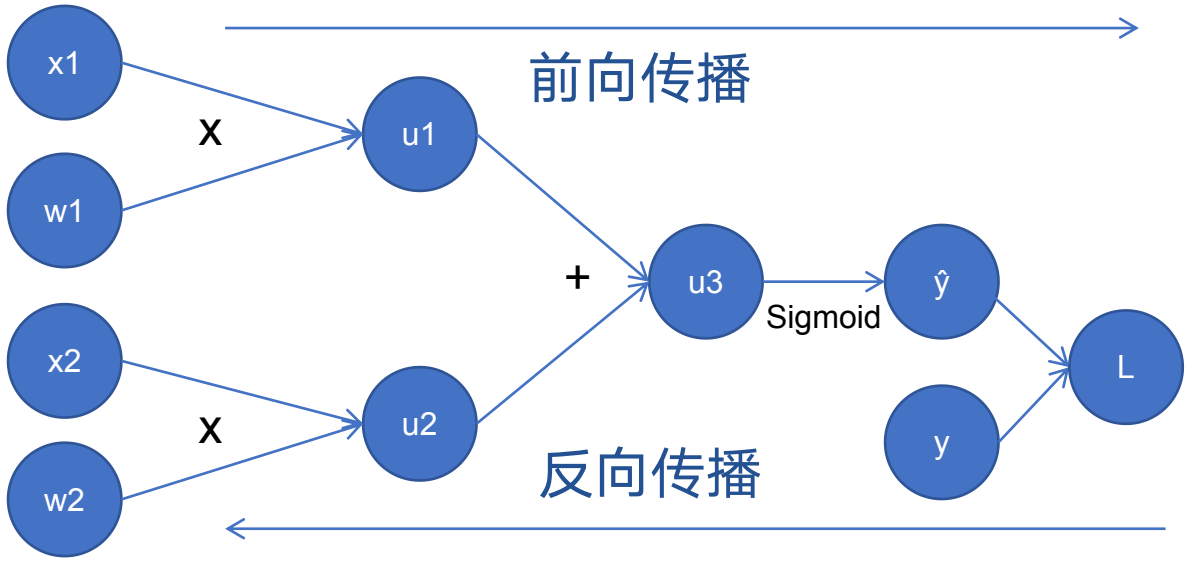


计算图让复杂的函数计算和求导像“搭积木”一样分解与组合，是深度学习中自动微分和梯度下降的基石。

计算图的具体计算步骤

首先根据计算图计算关于w1和w2的偏导数，因为这样就反映出当w1或w2有微小变化时，最后的损失函数有怎么样的改变：

$$L = \frac{1}{2}(\hat{y} - y)^2, \quad \hat{y} = \sigma(u_3), \quad u_3 = u_1 + u_2, \quad u_1 = w_1 \cdot x_1, \quad u_2 = w_2 \cdot x_2$$



- 首先求关于w1的偏导数：

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1}$$

- 然后逐项计算：

$$\frac{\partial L}{\partial \hat{y}} = (\hat{y} - y)$$

$$\frac{\partial \hat{y}}{\partial u_3} = \sigma(u_3)(1 - \sigma(u_3)) = \hat{y}(1 - \hat{y})$$

$$\frac{\partial u_3}{\partial u_1} = 1 \qquad \frac{\partial u_1}{\partial w_1} = x_1$$

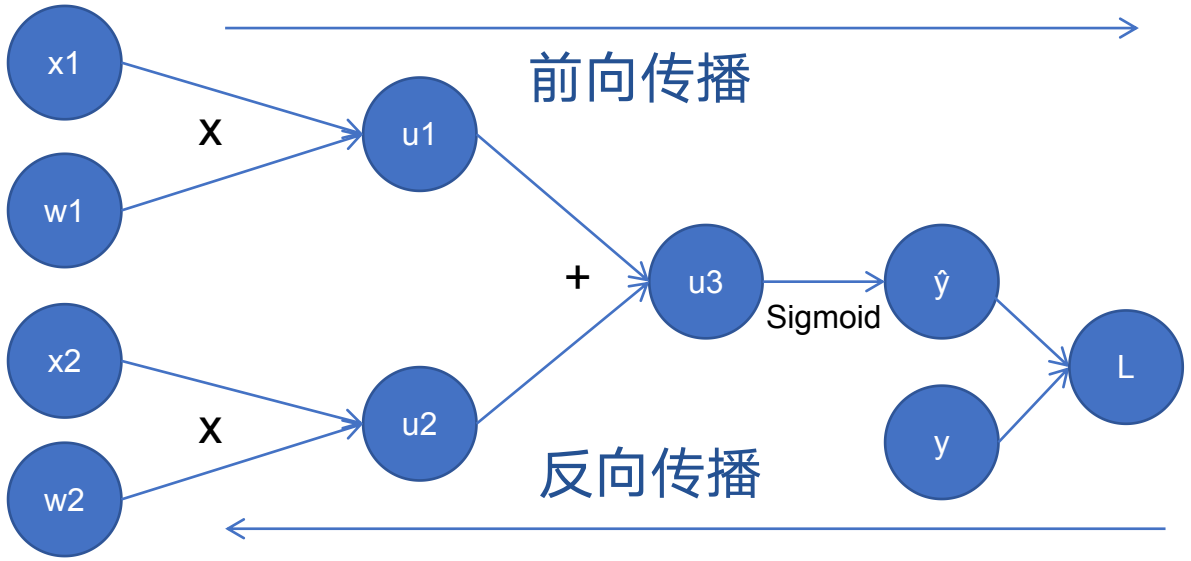
- 综合：

$$\frac{\partial L}{\partial w_1} = (\hat{y} - y) \cdot \hat{y}(1 - \hat{y}) \cdot 1 \cdot x_1$$

计算图的具体计算步骤

- 首先根据计算图计算关于w1和w2的偏导数，因为这样就反映出当w1或w2有微小变化时，最后的损失函数有怎么样的改变：

$$L = \frac{1}{2}(\hat{y} - y)^2, \quad \hat{y} = \sigma(u_3), \quad u_3 = u_1 + u_2, \quad u_1 = w_1 \cdot x_1, \quad u_2 = w_2 \cdot x_2$$



- 然后求关于w2的偏导数：

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_2} \cdot \frac{\partial u_2}{\partial w_2}$$

- 然后逐项计算：

$$\frac{\partial L}{\partial \hat{y}} = (\hat{y} - y)$$

$$\frac{\partial \hat{y}}{\partial u_3} = \sigma(u_3)(1 - \sigma(u_3)) = \hat{y}(1 - \hat{y})$$

$$\frac{\partial u_3}{\partial u_2} = 1 \qquad \frac{\partial u_2}{\partial w_2} = x_2$$

- 综合：

$$\frac{\partial L}{\partial w_2} = (\hat{y} - y) \cdot \hat{y}(1 - \hat{y}) \cdot 1 \cdot x_2$$

计算图计算完成之后的更新

- **从之前的计算步骤可以看出：梯度可以“流动”，即从输出（损失）反向传递到每个参数，得到每一层的梯度。**

- 然后通过梯度下降的方法，由于梯度（偏导数/导数）在某点的方向是函数增长最快的方向，它的反方向就是函数下降最快的方向：

$$w_1 \leftarrow w_1 - \eta \frac{\partial L}{\partial w_1}$$

$$w_2 \leftarrow w_2 - \eta \frac{\partial L}{\partial w_2}$$

- 就可以让参数沿着最陡峭的下降方向移动，这就是训练神经网络的核心。

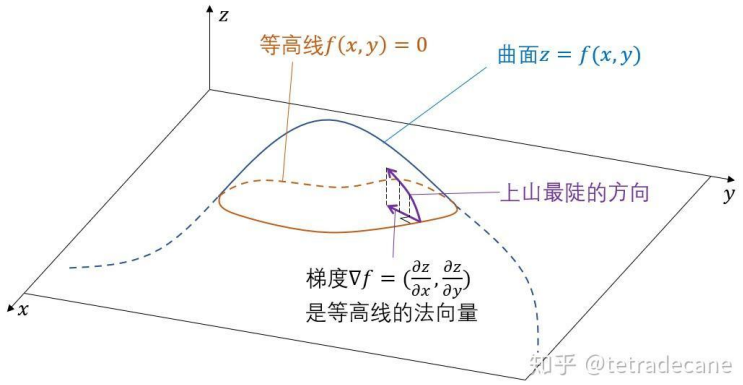
计算图让梯度像水一样从损失流向每个参数——这是反向传播和深度学习优化的核心。

▶ 扩展：梯度

- 定义：梯度是一个向量，包含了多变量函数在各个自变量方向上的偏导数：

$$\nabla f(x_1, x_2, \ldots, x_n) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right)$$

- 几何意义：梯度指向函数在当前点增长最快的方向，梯度的模（大小）表示在这个方向上增长的最快速率。



- 在机器学习中的作用：1）用于优化：通过梯度下降沿着梯度的反方向更新参数，降低损失函数。2）在反向传播中，梯度用于链式法则逐层传递误差信息。

梯度： 多变量函数在一点处增长最快的方向向量，其大小表示最快增长的速率。

▶ 扩展：梯度下降法的公式原理

- 为什么更新是用学习率乘以梯度，而不是其他的值呢？
- 1）梯度的意义：对于一个可微函数L(**w**)，偏导数在某点的方向是函数增长最快的方向，而它的反方向则是函数下降最快的方向。所以如果我们想让损失函数下降，最自然的做法就是，

$$w_i \leftarrow w_i - \eta \frac{\partial L}{\partial w_i}$$

- 2）为什么下降最快的方向就是梯度的方向？

$$L(w_0 + \Delta w) \approx L(w_0) + \nabla L(w_0)^T \Delta w$$

  - 目标：找一个单位方向u，让L下降最快。如果我从当前点w_0朝某个方向u移动一点点【η表示步长，u表示方向（向量）】，损失函数L 会怎么变化？

$$\Delta w = \eta \mathbf{u}, \quad D_{\mathbf{u}} L(w_0) = \nabla L(w_0)^T \mathbf{u}$$

$$|\nabla L(w_0)^T \mathbf{u}| \leq ||\nabla L|| ||\mathbf{u}|| = ||\nabla L||$$

  - 当且仅当u和∇L同方向时取最大值。
    - 最大上升方向：$\mathbf{u} = \frac{\Delta L}{||\Delta L||}$
    - 最大上升方向：$\mathbf{u} = -\frac{\Delta L}{||\Delta L||}$

注：$\nabla L^T u$ 其实就是把梯度投影到方向u上，看沿这个方向 L 的变化速度。

▶ 小结

- 计算图：把复杂函数拆成一系列简单运算节点。
  - 节点：表示变量或运算。输入节点：x,y,z等变量；运算节点：加法、乘法、平方、激活函数等操作；输出节点：最终结果或损失L。
  - 边：表示数据/梯度的传递。前向传播中描述节点之间传递数据（计算结果）；反向传播中表示沿边传递梯度（局部导数的链式乘积）。
- 计算图中的链式法则：梯度沿图反向传播，逐节点相乘，将误差逐层传递至输入层。
  - 梯度沿图反向传播：从输出（损失）往输入方向传递。
  - 逐节点相乘：每个节点的梯度 = 上游梯度 × 当前节点对输入的局部导数。
  - 多路径相加：如果有分支，梯度沿不同路径累加。
  - 误差逐层传递至输入：最终得到每个输入变量对损失的梯度。

计算图中的链式法则：梯度从输出反向传递，逐节点乘局部导数并在分支处累加，最终传到输入。

目录章节

CONTENTS

总结

- **极限（Limit）：**

$$\lim_{x \to a} f(x) = L$$

  - ✓ 极限描述了函数在某点附近趋于某个值。意义：描述变化趋势，是导数、偏导数的基础。

- **导数与偏导数（Derivative and Partial Derivative）：**

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}, \quad \frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h}$$

  - ✓ 导数表示函数在一点的瞬时变化率；偏导数描述多变量函数对单一变量的敏感度。

- **链式法则（Chain Rule）：**

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}, \quad \frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x}$$

  - ✓ 用于处理嵌套函数、多元复合函数。在反向传播中则使用链式法则将误差进行逐层传递。

极限是基础，导数是变化率，偏导是多变量版本，链式法则串联它们。

感谢聆听

Personal Website: https://www.miaopeng.info/

Email: miaopeng@stu.scu.edu.cn

Github: https://github.com/MMeowhite

Youtube: https://www.youtube.com/@pengmiao-bmm