北京師範大學
BEIJING NORMAL UNIVERSITY

计算语言学的数学基础

(2) 函数与优化

王予沛

2025 年 3 月 10 日

北京师范大学数字人文系

1. 函数
2. 优化

函数

一元函数的定义

给定两个实数集 X 和 Y, 若有对应法则 f, 使对 X 内每一个数 x, 都有唯一的一个数 $y \in Y$ 与它相对应, 则称 f 是定义在数集 X 上的**函数**，记作

$$\begin{aligned} f &: X \to Y, \\ x &\mapsto y. \end{aligned} \tag{1}$$

其中 $\to$ 表示两个集合间的整体映射方向，$\mapsto$ 表示两个集合中单个元素间具体的映射规则。数集 X 称为函数 f 的**定义域**(domain)，x 所对应的数 y 称为 f 在点 x 的**函数值**(value)，常记为 $f(x)$. 全体函数值的集合

$$f(X) = \{y \mid y = f(x), x \in X\} (\subset Y)$$

称为函数 f 的**值域**(codomain/range).

线性函数

简单来说，线性函数是其图像为直线的函数，即零次或一次的多项式函数。

严格来说，线性函数就是线性映射，即从一个向量空间 V 到另一个向量空间 W 的映射且保持加法运算和数量乘法的运算。

函数的性质

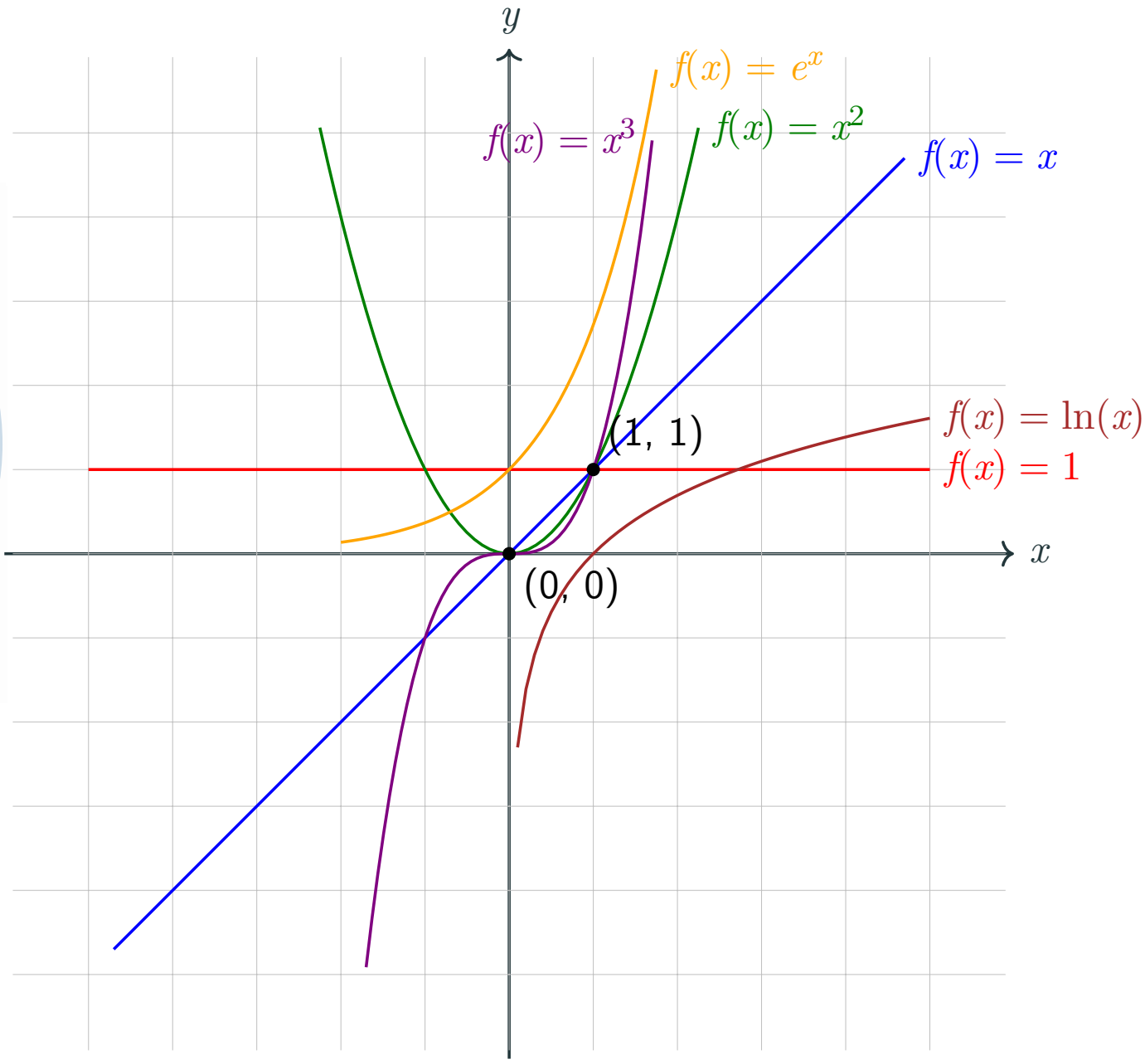
- 有界性
  - 有上界函数：能找到一个数 M，对所有 $x \in X$，有 $f(x) \leq M$;
  - 有下界函数：能找到一个数 L，对所有 $x \in X$，有 $f(x) \geq L$;
  - 有界函数：能找到一个正数 B，对所有 $x \in X$，有 $|f(x)| \leq B$.
- 单调性：若对任何 $x_1, x_2 \in X$, 当 $x_1 < x_2$ 时, 总有
  - $f(x_1) \leqslant f(x_2)$, 则称 f 为 X 上的**增函数**, 不能取等号时, 称 f 为 X 上的**严格增函数**;
  - $f(x_1) \geqslant f(x_2)$, 则称 f 为 X 上的**减函数**, 不能取等号时, 称 f 为 X 上的**严格减函数**.
- 奇偶性
  - 奇函数：对所有 $x \in X$，有 $f(-x) = -f(x) \Rightarrow$ 关于原点对称;
  - 偶函数：对所有 $x \in X$，有 $f(-x) = f(x) \Rightarrow$ 关于 y 轴对称.
- 周期性
  - 周期函数：对所有 $x \in X$，存在正数 T，满足 $x \pm T \in X$ 且 $f(x \pm T) = f(x)$，则称 f 为周期函数，T 称为 f 的周期.

基本初等函数

表 1: 基本初等函数分类表

函数类型	函数表达式
常量函数	$y = c \quad (c \text{ 是常数 })$
幂函数	$y = x^u \quad (u \text{ 为实数 })$
指数函数	$y = a^x \quad (a > 0, a \neq 1)$
对数函数	$y = \log_a x \quad (a > 0, a \neq 1)$
三角函数	$y = \sin x$ (正弦函数) $y = \cos x$ (余弦函数) $y = \tan x$ (正切函数) $y = \cot x$ (余切函数)
反三角函数	$y = \arcsin x$ (反正弦函数) $y = \arccos x$ (反余弦函数) $y = \arctan x$ (反正切函数) $y = \operatorname{arccot} x$ (反余切函数)

基本初等函数图像



基本初等函数的图像

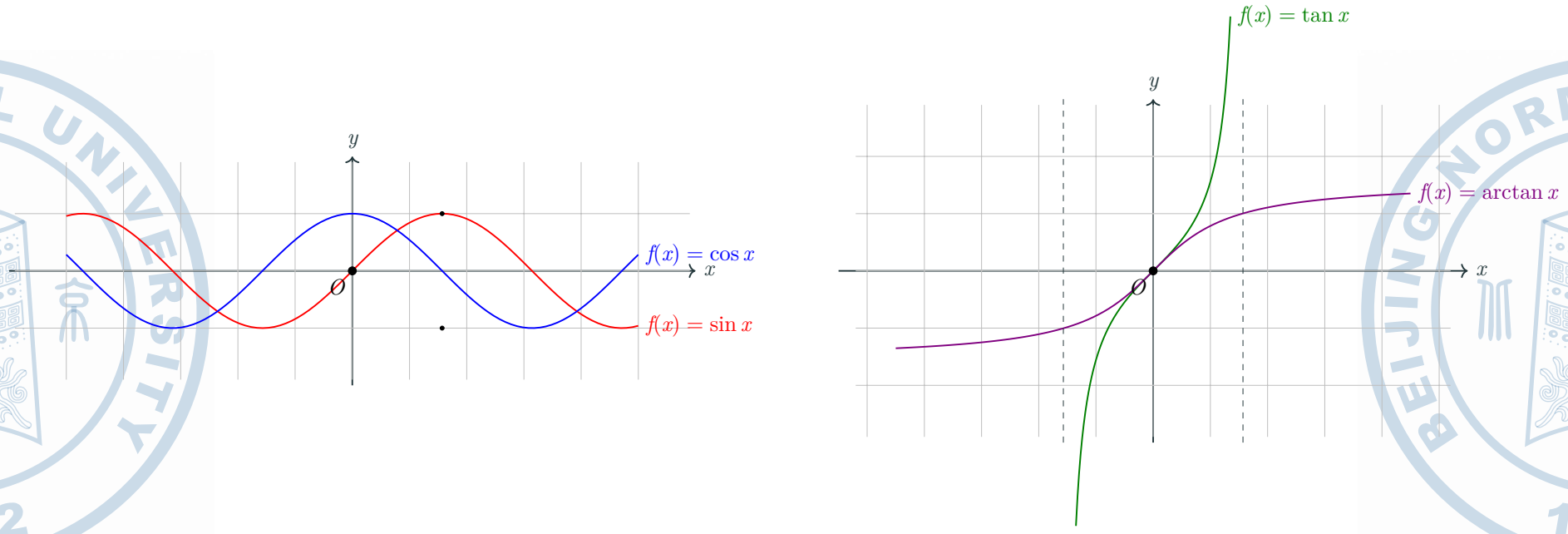


图 2: 三角函数和反三角函数的图像

反函数

设函数 $f: X \to Y$ 是双射，则存在函数 $g: Y \to X$，使得对任意 $y \in Y$，有 $g(f(x)) = x$，则称 g 为 f 的**反函数**，记作 f^{-1}。

理解：

- 反函数是函数的逆变换，即 $f^{-1}(f(x)) = x$；
- 反函数与原函数关于 $y = x$ 对称；
- 并不是所有函数都有反函数，只有双射函数才有反函数。

复合函数

复合函数 (composite function)，又称作合成函数，在数学中是指逐点地把一个函数作用于另一个函数的结果，所得到的第三个函数。

设函数 $f\colon X \to Y$ 和 $g\colon Y \to Z$，则复合函数 $g \circ f\colon X \to Z$ 定义为：

$$(g \circ f)(x) = g(f(x)) \tag{2}$$

直观地说，复合两个函数是把两个函数链接在一起的过程，内函数的输出就是外函数的输入。

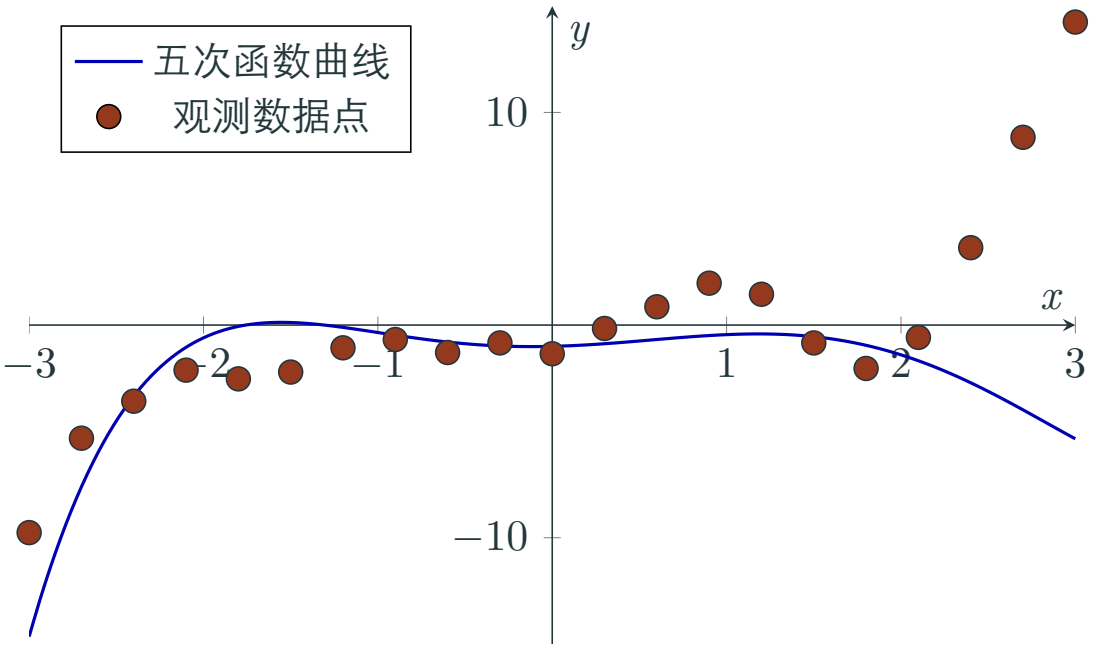
多元函数

类似于一元函数的定义，我们把能从 $\mathbb{R}^n$ 中的一个点 $(x_1, \cdots, x_n)$ 对应到 $\mathbb{R}$ 中唯一的一个实数的映射称为 n 元函数。该函数同样有对应的定义域和值域，只不过一般来说定义域不再是某一段区间，而是 $\mathbb{R}^n$ 上的某块区域。

优化

五次函数拟合

我们想要找到一个五次函数 $f(x) = k_5x^5 + k_4x^4 + k_3x^3 + k_2x^2 + k_1x + k_0$ 使其最好地拟合观测数据点。这是一个**回归**问题。



如果我们试图找到一个指标 $\mathcal{L}$ 来衡量 “拟合优度”(goodness of fit)，我们需要注意到这个指标仅仅是五次函数的系数（也称**参数**）$k_5, k_4, k_3, k_2, k_1, k_0$ 的函数，即 $\mathcal{L}$ 是一个六元函数：

$$\mathcal{L} = \mathcal{L}(k_5, k_4, k_3, k_2, k_1, k_0)$$

损失函数

回归问题的损失函数往往定义为拟合函数对观测数据点的预测值和观测值之间的误差的平方和，即：

$$\text{Loss} = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{3}$$

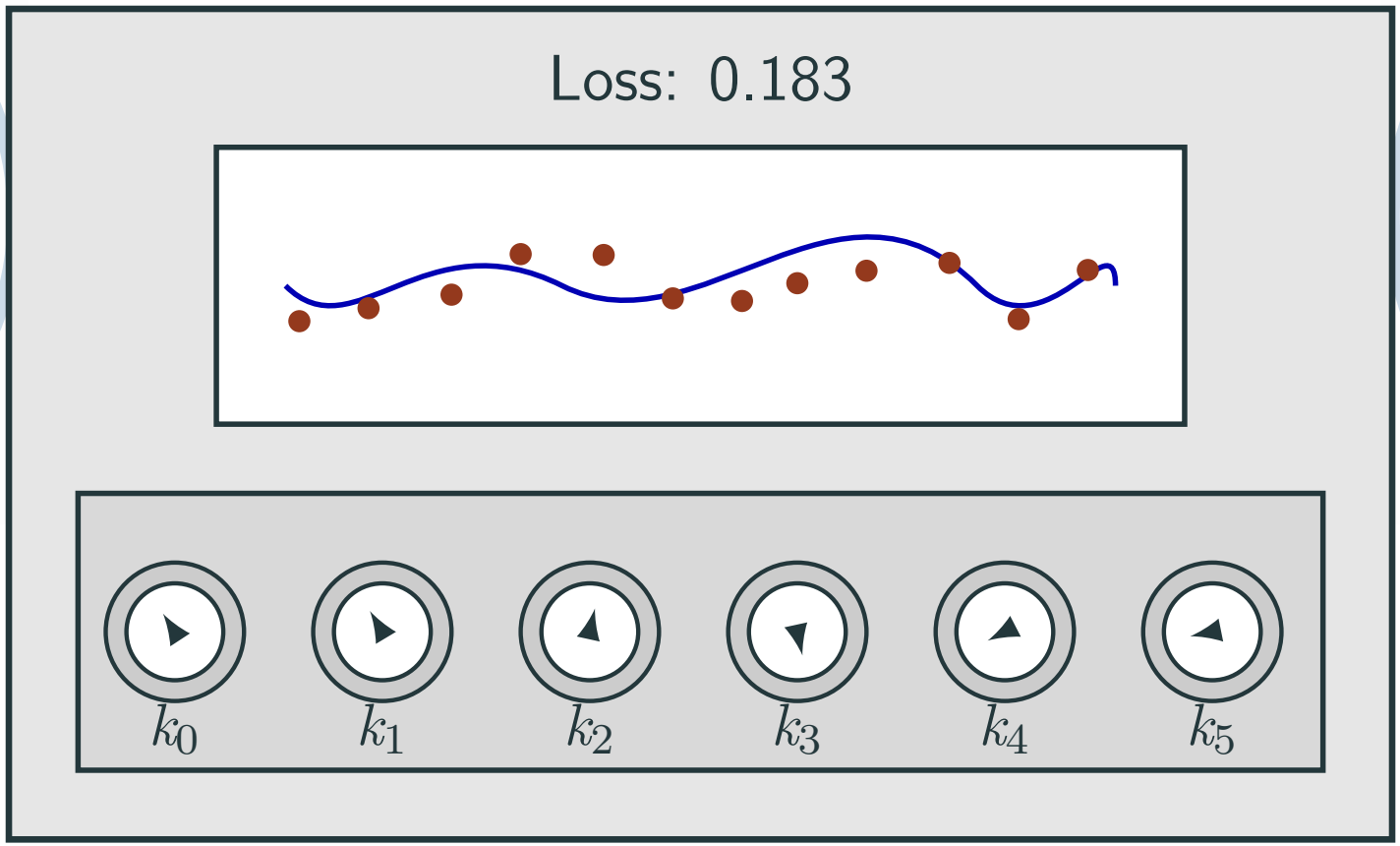
其中 n 是观测数据点的数量，x_i 和 y_i 是第 i 个观测数据点的自变量和因变量。我们可以将式 (3) 作为 $\mathcal{L}$，因其仅是关于 $k_5, k_4, k_3, k_2, k_1, k_0$ 的函数（x_i 和 y_i 已知）。

想要找到能够最好地拟合观测数据点的五次函数，我们需要找到使损失函数 $\mathcal{L}$ 取值最小的参数 $k_5, k_4, k_3, k_2, k_1, k_0$。这是一个**优化**问题。

> *Mathematical optimization or mathematical programming is the selection of a best element, with regard to some criteria, from some set of available alternatives.*
>
> —*Wikipedia*

黑箱优化器

想象一个黑箱，上方有一个屏幕，屏幕上显示了函数的图像和损失函数的值。黑箱上还有六个" 手柄"，我们可以通过调节这六个" 手柄" 来控制多项式各项系数 $f(x) = k_5x^5 + k_4x^4 + k_3x^3 + k_2x^2 + k_1x + k_0$.



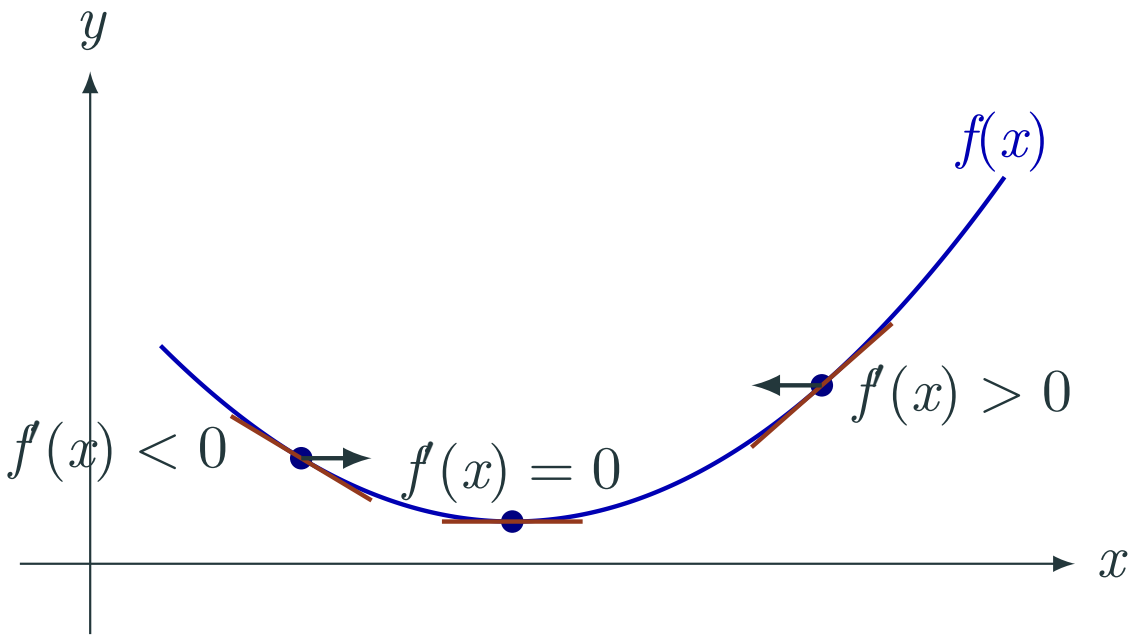
拨动一个手柄

假设我们先给参数 $k_5, k_4, k_3, k_2, k_1, k_0$ 随机赋上值，然后固定其他参数，只拨动 k_2 手柄，观察损失函数的变化。显然我们在拨动 k_2 手柄时，损失函数会发生变化，可能变大也可能变小。

此时，损失函数 $\mathcal{L}$ 是关于 k_2 的函数，即 $\mathcal{L}(k_2)$，这是一个一元函数。因此，找到使损失函数 $\mathcal{L}$ 取值最小的 k_2 的问题，就是一个一元函数优化问题。

一元函数优化问题，需要借助我们在高中学习过的导数来解决。

导数的几何意义



导数表示函数在某点的斜率

- 导数为负 ($f'(x) < 0$): 向右移动可以减小函数值
- 导数为零 ($f'(x) = 0$): 局部极值点 (可能是最小值)
- 导数为正 ($f'(x) > 0$): 向左移动可以减小函数值

与导数相关的四个概念

连续性

- 定义：若 $\lim_{x \to x_0} f(x) = f(x_0)$，则 $f(x)$ 在 x_0 处连续
- 直观：函数图像无间断点，可一笔画出

可导性

- 定义：若导数 $f'(x_0) = \lim_{h \to 0} \frac{f(x_0+h)-f(x_0)}{h}$ 存在
- 直观：函数在该点有唯一切线，无"尖点"

可微性

- 定义[a]：$f(x_0 + \Delta x) = f(x_0) + f'(x_0)\Delta x + o(\Delta x)$
- 直观：函数可用线性函数良好近似

光滑性

- 定义：函数具有连续的各阶导数
- C^k：具有 k 阶连续导数
- C^∞：具有任意阶连续导数
- 直观：函数曲线平滑，无任何"棱角"

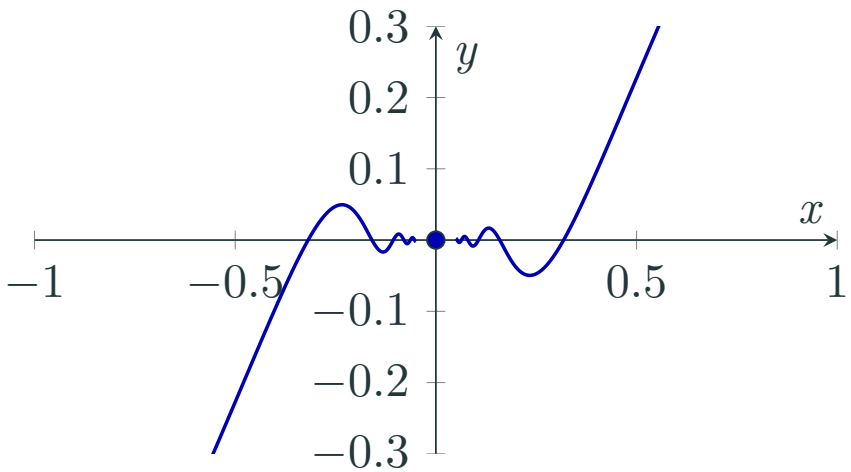
[a] $o(\Delta x)$ 是 Δx 的高阶无穷小，意思是 $o(\Delta x)$ 越小趋于 0 的速度比 Δx 更快，因此相较 Δx 可以忽略。

四个概念间的关系

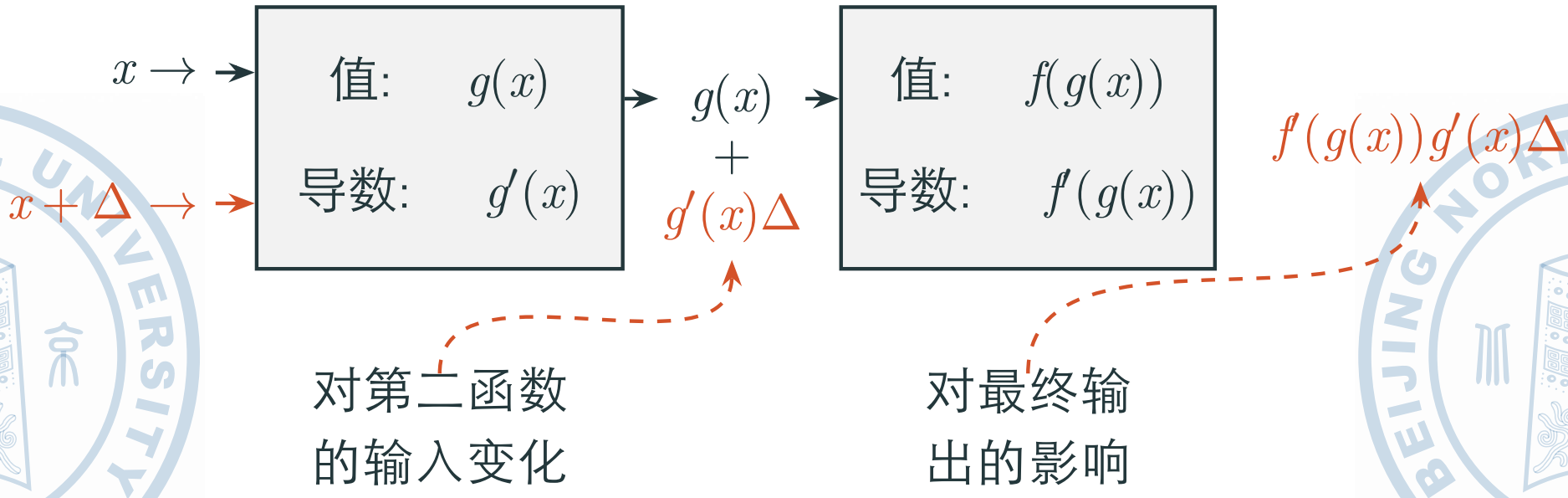
- **连续与可导的关系**
  - 可导必连续，连续不一定可导
  - 经典反例：$f(x) = |x|$ 在 $x = 0$ 处连续但不可导
- **可导与可微的关系**
  - 对于一元函数，可导与可微等价
  - 对于多元函数，可导不等价于可微
- **可导与光滑的关系**
  - 光滑函数要求不仅可导，还要求各阶导数连续
  - 例如：$f(x) = x^2 \sin(1/x)$ $(x \neq 0)$，$f(0) = 0$
  - 在 $x = 0$ 处可导但二阶导数不存在，不是光滑函数
- **从条件强弱排序**
  - 光滑（C^∞）$\subset \cdots \subset$ 二阶导连续（C^2）$\subset$ 一阶导连续（C^1）$\subset$ 可导（可微）$\subset$ 连续（C^0）.

可导函数不光滑的例子

光滑函数必可导，但可导函数未必光滑。例如，$f(x) = x^2 \sin(1/x)$（补充定义 $f(0) = 0$）在原点可导，但其导函数在该点不连续，故不属于 C^1，更非光滑。



链式法则



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

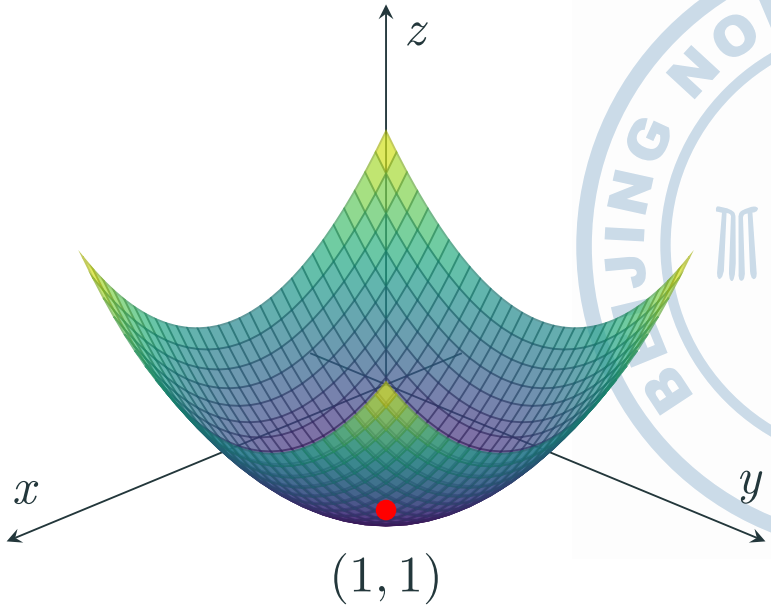
拨动多个手柄

若要一次性拨动多个手柄，则 $\mathcal{L}$ 是关于多个自变量的多元函数。对于多元函数 $f(x_1, x_2, \cdots, x_n)$，可对其各个自变量分别求导，得到**偏导数**。对 x_i 求偏导写作 $\frac{\partial f}{\partial x_i}$。此时视其他自变量 $x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$ 为常数。

$$f(x, y) = (x-1)^2 + (y-1)^2$$

$$\frac{\partial f}{\partial x} = 2x - 2$$

$$\frac{\partial f}{\partial y} = 2y - 2$$



$\frac{\partial f}{\partial x}$ 也写作 $f_x(x, y)$ 或 $D_x f(x, y)$，$\frac{\partial f}{\partial y}$ 也写作 $f_y(x, y)$ 或 $D_y f(x, y)$。

方向导数

现在，我们可以求出标准基向量方向上的导数了，那么如何求出任意方向 $\boldsymbol{u} = \cos\theta \boldsymbol{i} + \sin\theta \boldsymbol{j}$ 上的导数呢[1]?

按照导数定义考虑

$$\lim_{t\to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{u}) - f(\boldsymbol{x})}{t}$$

令 $g(t) = f(x_0 + t\cos\theta, y_0 + t\sin\theta)$。根据链式法则，$g(t)$ 的导数为：

$$g'(t) = f_x(x_0 + t\cos\theta, y_0 + t\sin\theta)\cos\theta + f_y(x_0 + t\cos\theta, y_0 + t\sin\theta)\sin\theta$$

在 $t = 0$ 时，得到：

$$g'(0) = f_x(x_0, y_0)\cos\theta + f_y(x_0, y_0)\sin\theta \tag{4}$$

这正是**方向导数**的值。

[1] $\boldsymbol{i}$ 和 $\boldsymbol{j}$ 是标准基向量，θ 是向量 $\boldsymbol{u}$ 与 x 轴的夹角，$\boldsymbol{u}$ 的模为 1。

方向导数与梯度

式（4）可以进一步改写为向量点积的形式：

$$D_u f(x_0, y_0) = (f_x(x_0, y_0), f_y(x_0, y_0)) \cdot \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

将 $(f_x(x_0, y_0), f_y(x_0, y_0))$ 记作 $\nabla f(x_0, y_0)$，则式（4）可以改写为：

$$D_u f(x_0, y_0) = \nabla f(x_0, y_0) \cdot \boldsymbol{u} = \|\nabla f(x_0, y_0)\|\|\boldsymbol{u}\| \cos\varphi = \|\nabla f(x_0, y_0)\| \cos\varphi$$

其中 φ 是 $\nabla f(x_0, y_0)$ 与 $\boldsymbol{u}$ 之间的夹角。

可见，想要使方向导数 $D_u f(x_0, y_0)$ 最大，只需使 $\cos\varphi$ 最大，即 $\varphi = 0$，即当 $\boldsymbol{u}$ 与 $\nabla f(x_0, y_0)$ 方向一致时，函数值在 $\boldsymbol{u}$ 方向上的变化率最大。我们称 $\nabla f(x_0, y_0)$ 为 $f(x, y)$ 在 (x_0, y_0) 处的**梯度**。

由于 $g(t)$ 是一元函数，我们直接借助在一元函数上的观察给出，梯度方向是函数值增长最快的方向，负梯度方向是函数值下降最快的方向。

梯度下降

现在，我们已经知道了负梯度方向是函数值下降最快的方向。那么，我们就可以使用**梯度下降法**来求解最优化问题。

对于 $\mathcal{L}$，我们使用下面的更新公式来使函数值沿着负梯度方向逐渐变小：

$$\boldsymbol{k}^{\text{new}} \leftarrow \boldsymbol{k}^{\text{old}} - \alpha \nabla \mathcal{L}(\boldsymbol{k}^{\text{old}})$$

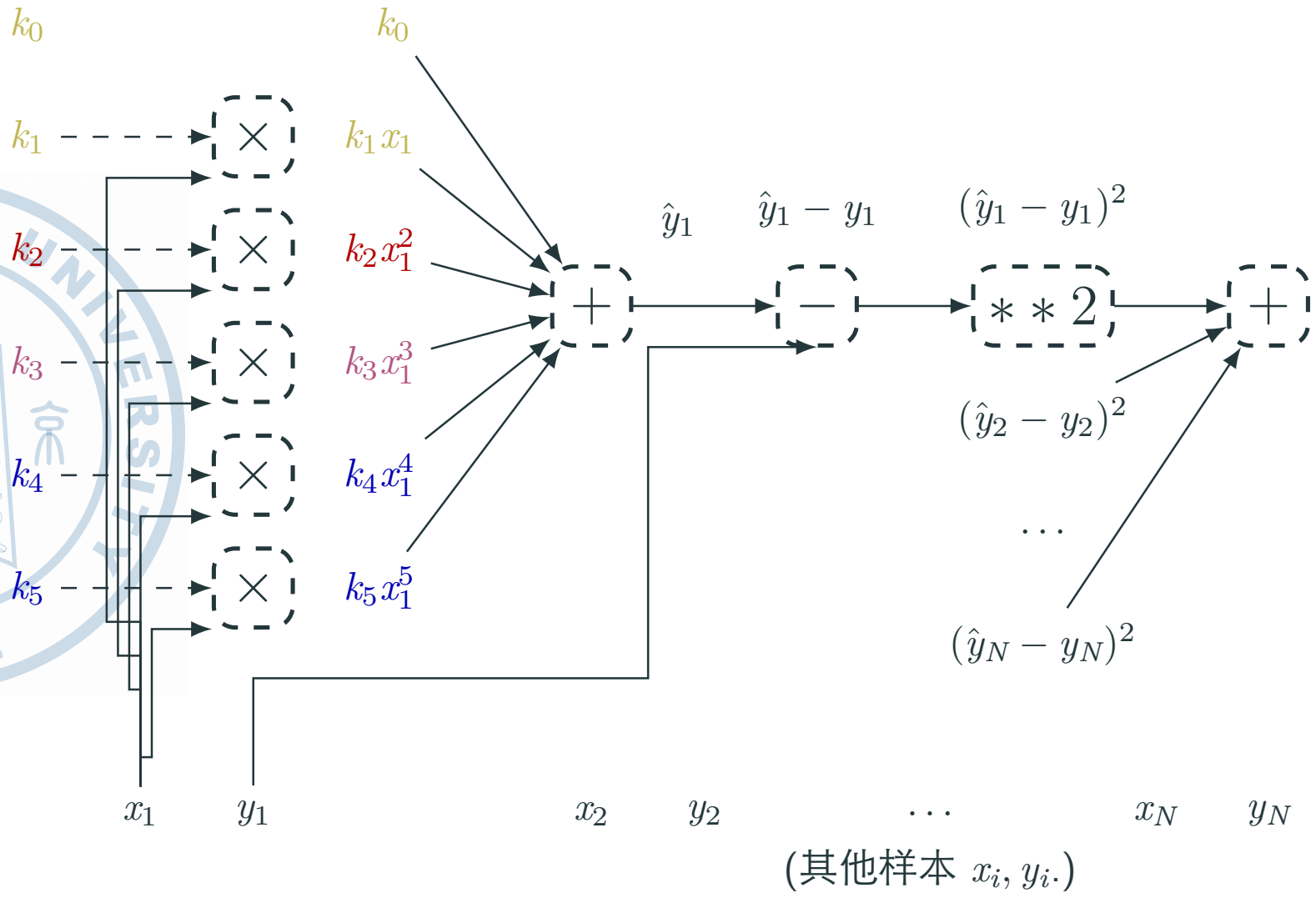
其中，α 是学习率（控制步幅），$\nabla \mathcal{L}(\boldsymbol{k})$ 是损失函数对参数 $\boldsymbol{k}$ 的梯度。

对于每一个手柄 k_i，这意味着借助其偏导数来调整其位置，使得函数值下降。具体来说，对于每一个手柄 k_i，我们可以使用梯度下降法的更新公式：

$$k_i \leftarrow k_i - \alpha \frac{\partial \mathcal{L}}{\partial k_i}$$

其中，$\frac{\partial \mathcal{L}}{\partial k_i}$ 是损失函数对参数 k_i 的偏导数。

回归问题的计算图


k_0
k_1
k_2
k_3
k_4
k_5
k_0
$k_1 x_1$
$k_2 x_1^2$
$k_3 x_1^3$
$k_4 x_1^4$
$k_5 x_1^5$
$\hat{y}_1$
$\hat{y}_1 - y_1$
$(\hat{y}_1 - y_1)^2$
$* * 2$
$(\hat{y}_2 - y_2)^2$
$\cdots$
$(\hat{y}_N - y_N)^2$
Total Loss
x_1
y_1
x_2
y_2
$\cdots$
x_N
y_N
(其他样本 x_i, y_i.)

回顾

- 前向过程（forward pass）：计算损失函数
- 反向过程（backward pass）：计算梯度
- 更新参数：$\boldsymbol{k}^{\text{new}} \leftarrow \boldsymbol{k}^{\text{old}} - \alpha \nabla \mathcal{L}(\boldsymbol{k}^{\text{old}})$
- 重复前向和反向过程，直到损失函数收敛

我们将这个算法称为**反向传播算法**（back-propagation algorithm）。