



线性回归的数学、算法 与现代人工智能

回归、概率分类、泛化与可运行实践

三课时递进教学 90 页 逐页讲稿与离线实验

看清变量 走完推导 运行代码 解释结果

一次学习过程的完整地图

先分清模型、损失、求解与评价分别回答什么

01 表示规定模型能看到什么

输入 $x \rightarrow$ 特征 $\varphi(x) \rightarrow$ 预测

02 损失规定训练中的好坏标准

观测 $y \rightarrow$ 损失 $J \rightarrow$ 梯度 $\rightarrow$ 参数更新

03 评价检验对新样本的有效性

训练拟合 $\rightarrow$ 验证选择 $\rightarrow$ 冻结 $\rightarrow$ 测试

推导 2/3 增加训练轮数与增加 x^2 特征, 改变的是同一件事吗?

三课时路线：推导帧、重点页与实践

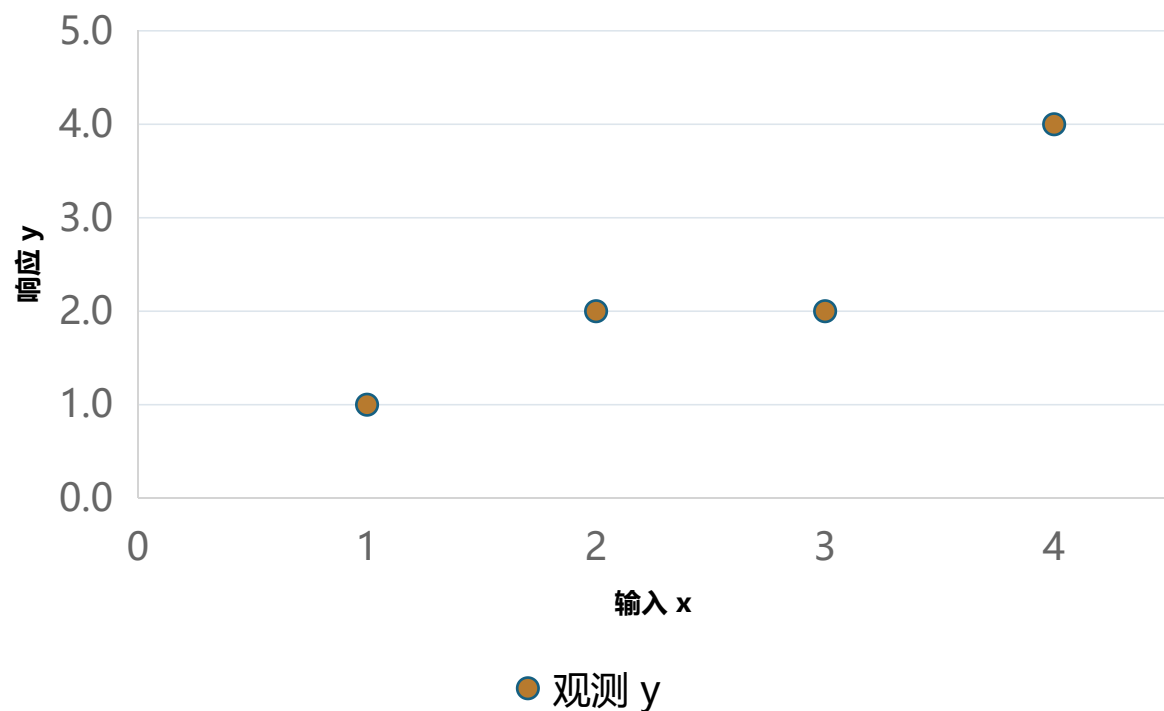
连续推导帧可快速播放，手算与实验页需要停留

课时与页码	主线及现场活动	退出时的学生产出
第1课时 45分钟 1–24页	四点模型、逐项求导 单步演示与学生手算	给出 e 、 g 、 w^+ 、 J
第2课时 45分钟 25–58页	矩阵、估计与泛化 矩阵操作、验证选模	解释维度与正规方程
第3课时 45分钟 59–90页（含选读）	分类25分钟，现代案例20分钟 逻辑/Softmax过程、在线案例	解释更新与评价条件

基础推导逐帧展开，减少跳步 每课时安排可核验的学生产出 分钟为讲授路径建议，扩展页可选讲

贯穿算例：四个有误差的观测点

先固定数据；后续每个数都能从这张表重算



$$\mathbf{x}=(1,2,3,4)^T$$

$$\mathbf{y}=(1,2,2,4)^T$$

$N=4$ ；先用 $\hat{y}=wx$ ，不含截距

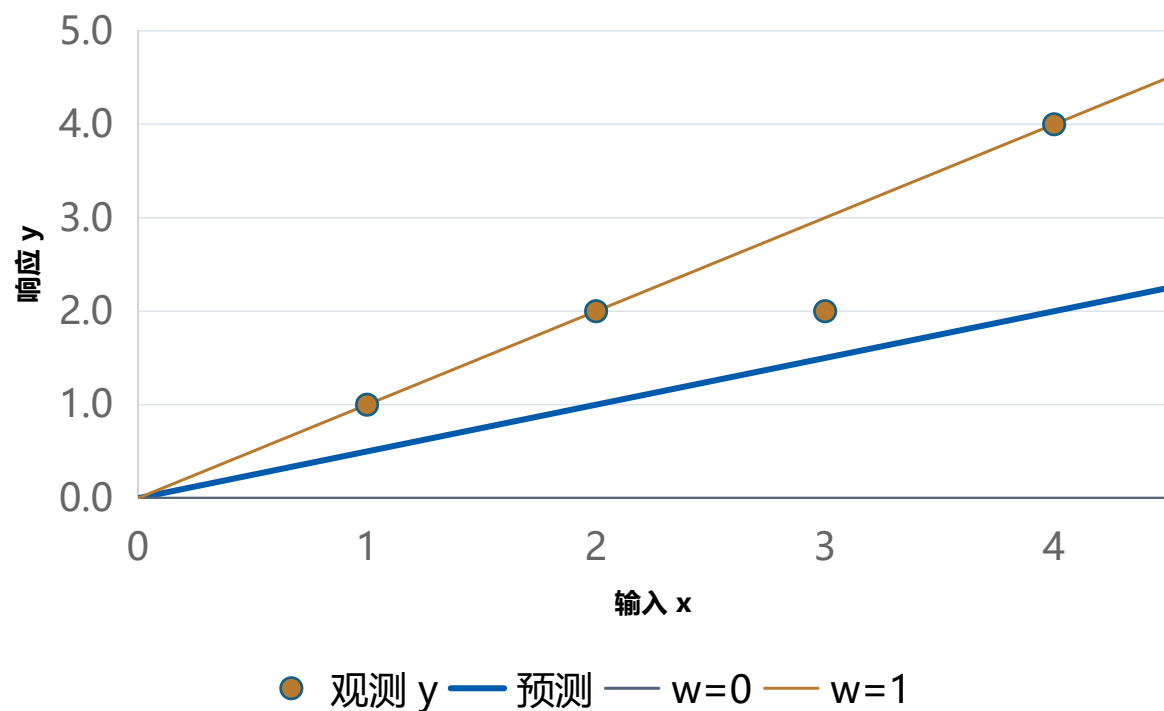
同一输入对应观测与预测两个量

标签来自测量，参数来自学习

教学数据用于计算，不代表总体规律

斜率 w 控制直线怎样倾斜

把参数变化翻译成图上可见的预测变化



$$\hat{y} = wx$$

x 增加1时，预测增加 w

$w=0$ 、 0.5 、 1 ：三条候选线

同一条线同时预测全部样本

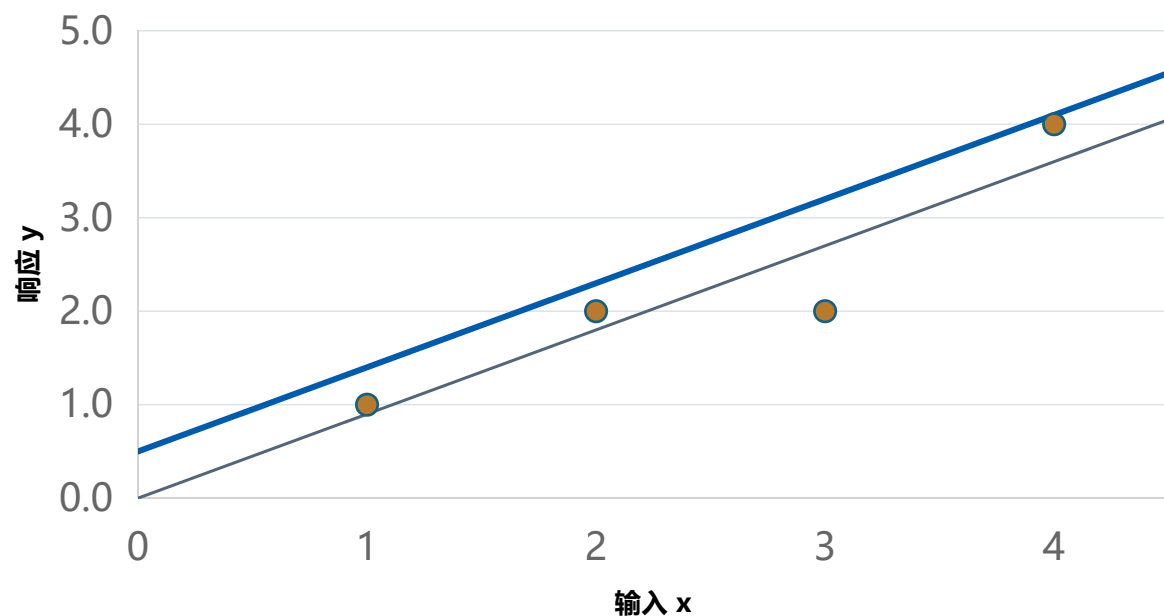
斜率带有“输出单位/输入单位”

修改 w 会联动所有预测

推导 2/11 w 从0.5变成0.6时， $x=4$ 的预测增加多少？

截距让直线整体平移

先认识完整模型；四点手算仍保持 $b=0$



● 观测 y — $b=0.5, w=0.9$ — $b=0, w=0.9$

$$\hat{y} = b + wx$$

$b = \hat{y}(0)$: 输入为0时的模型预测

本轮手算固定 $b=0$ ，只学习 w

改变 b : 所有预测增加同样的量

改变 w : 不同输入的变化不同

是否强制过原点需要领域依据

推导 3/11 把截距增加0.2，四个训练点的预测分别增加多少？

预测不是观测：同一输入上的两个数

先算预测，再与已有观测比较

x	观测y	预测 $\hat{y}$ (w=.5)
1	1	0.5
2	2	1
3	2	1.5
4	4	2

$$w=0.5; \hat{y}_i=0.5x_i$$

$$\hat{y}=(0.5,1,1.5,2)^T$$

$$y=(1,2,2,4)^T$$

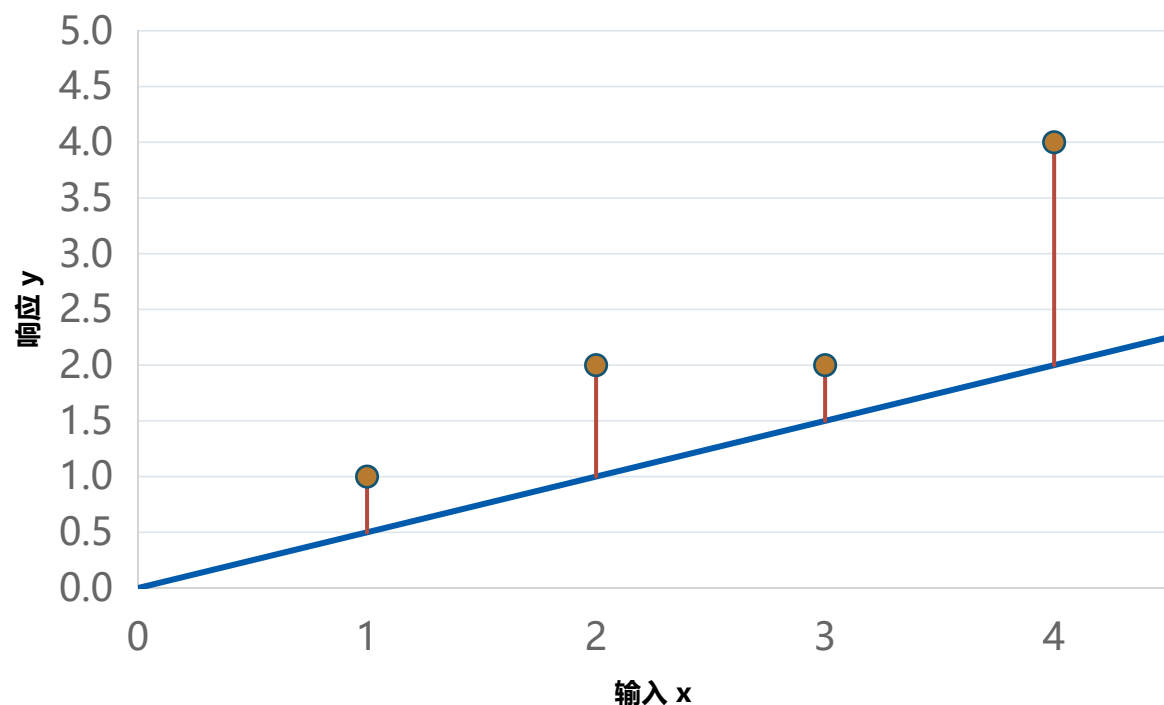
观测y固定，预测随参数改变

预测阶段通常不知道真实y

训练阶段用已有y提供监督

残差是观测减预测

用带符号的竖直距离描述低估与高估



$$r_i = y_i - \hat{y}_i$$

$$w = .5: r = (.5, 1, .5, 2)^T$$

$$e_i = \hat{y}_i - y_i = -r_i$$

$r > 0$ 表示低估, $r < 0$ 表示高估

本课残差方向始终不变

求导中使用 $e = \text{预测} - \text{观测}$

推导 5/11 第三点若改用 $w=1$, 残差是多少?

把四个残差变成可比较的平方误差

平方避免正负抵消，同时放大较大偏差

样本	r ($w=.5$)	r^2
1	.5	.25
2	1	1
3	.5	.25
4	2	4

$$SSE = \sum_i r_i^2 = \sum_i e_i^2$$

$$w=.5: SSE = .25 + 1 + .25 + 4 = 5.5$$

$$w=1: SSE = 0 + 0 + 1 + 0 = 1$$

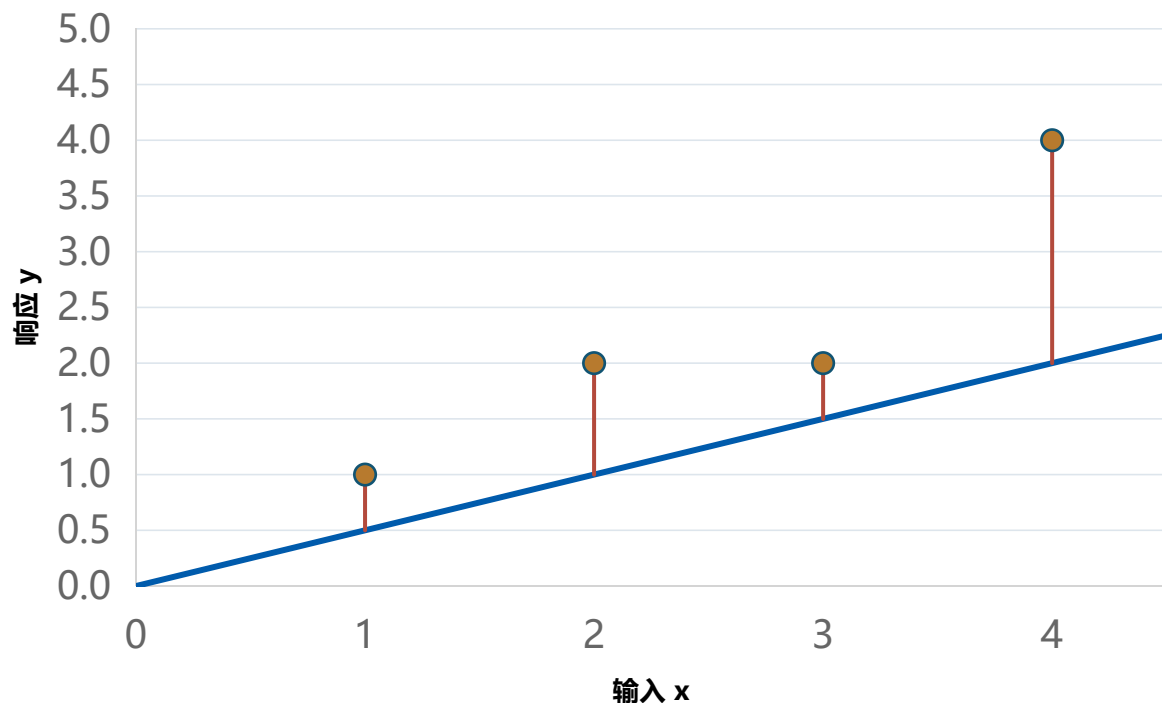
先逐样本平方，再对样本求和

大残差承担更大权重

平方误差不是唯一可选标准

平均与二分之一：明确目标的数值尺度

本课优化J，报告MSE时不带1/2



$$\text{MSE} = (1/N) \sum_i r_i^2$$

$$J(w) = 1/(2N) \sum_i (wx_i - y_i)^2$$

$$N=4: J = \text{SSE}/8 = \text{MSE}/2$$

$$w=.5: \text{MSE}=1.375, J=.6875$$

除以N便于比较不同样本量

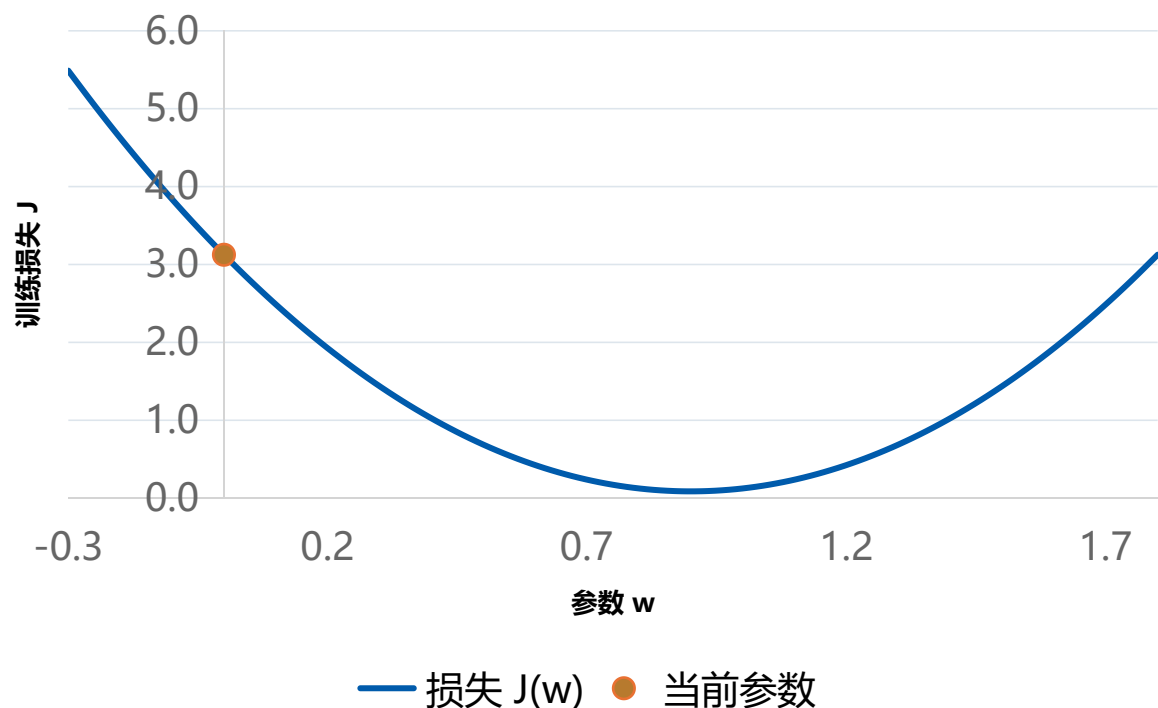
1/2使求导后的系数简洁

缩放目标会改变梯度与有效步长

推导 7/11 把J乘以2，最优斜率会改变吗？梯度呢？

损失曲线：每一个w对应一个J

横轴现在是参数，不再是输入x



$$J(w) = (30w^2 - 54w + 25)/8$$

$$J(w) = 3.75(w - 0.9)^2 + 0.0875$$

$$w^* = 0.9; J^* = 0.0875$$

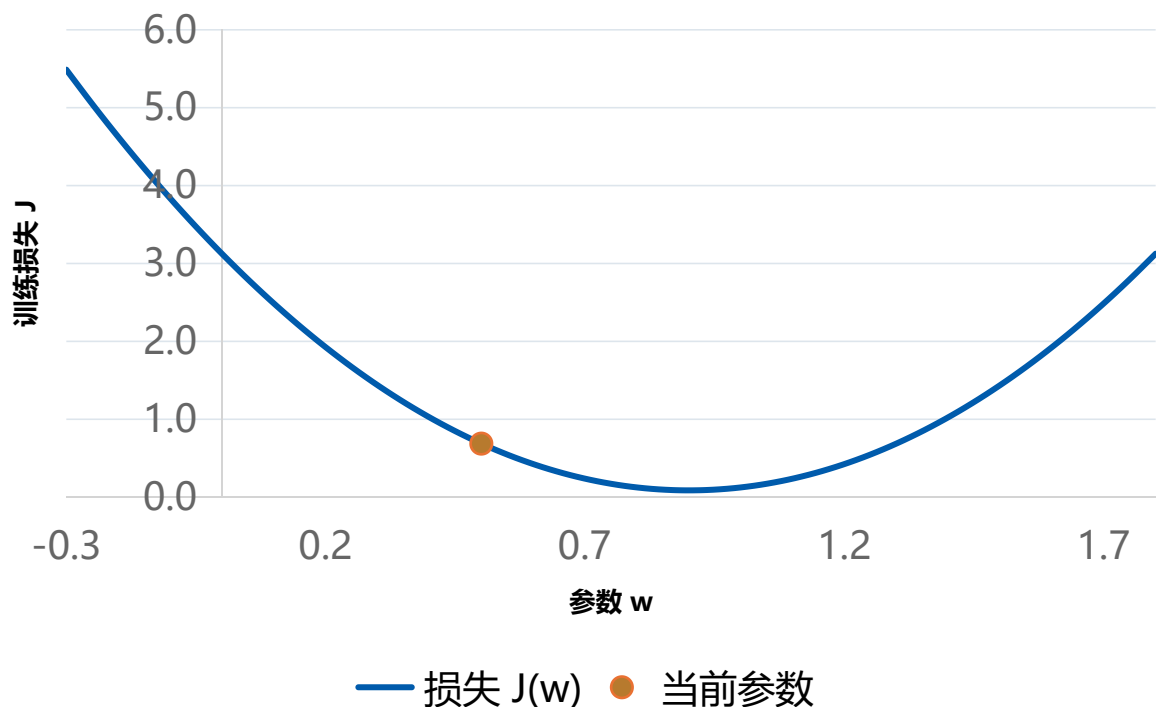
曲线上的点代表一整条候选拟合线

最低点对应本目标最小值

损失不是第五个数据观测

数据空间与参数空间：一条线对应一个点

来回翻译，避免把两幅图的轴混淆



参数 $w=0.5 \leftrightarrow$ 数据空间直线 $\hat{y}=0.5x$

预测向量 $\hat{y}=(0.5, 1, 1.5, 2)^T$

目标点 $(w, J)=(0.5, 0.6875)$

数据图看每个样本怎么被预测

损失图看参数整体表现如何

一维参数也能同时产生四维预测

推导 9/11 参数只有一个，为什么预测向量有四个分量？

人工试参数：可以比较，但怎样知道往哪走

用候选值表引出梯度，而非靠反复猜测

w	J(w)	相对观察
0	3.125	明显低估
.5	.6875	仍偏小
1	.125	接近低点
.9	.0875	本目标最优

$$J(w) = 3.75(w - .9)^2 + .0875$$

$$w = .5 \rightarrow .6875; \quad w = 1 \rightarrow .125$$

$$w = .9 \rightarrow .0875$$

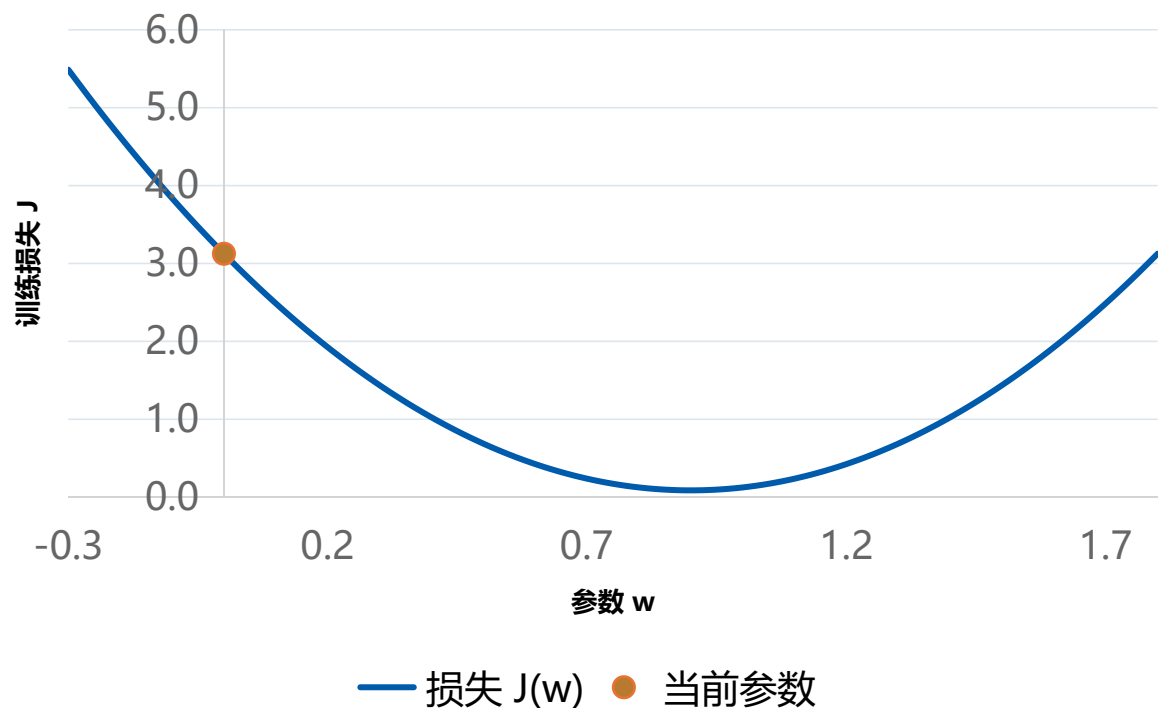
试值能比较已选出的候选

参数很多时穷举成本迅速增加

梯度提供局部变化方向

四点算例的公式总览

保留这页作速查，下一组逐行求导



$$\hat{y}_i = wx_i; \quad r_i = y_i - \hat{y}_i; \quad e_i = -r_i$$

$$J = (1/8) \sum_i e_i^2$$

$$J = (30w^2 - 54w + 25)/8$$

$$J = 3.75(w - .9)^2 + .0875$$

从逐样本式到展开式再到配方式

三种写法是同一个目标

求导时数据固定， w 是变量



求导第一步：平方项怎样变化

先对单个样本求导，再汇总所有样本

计算图：每条边表示一次局部变化

参数 w

误差 $e=wx-y$

损失 $\ell=e^2/2$

$$e_i(w)=wx_i-y_i$$

$$\ell_i(w)=1/2e_i(w)^2$$

$$d\ell_i/de_i=e_i$$

先暂时把 e 当作一个中间变量

平方的导数消去系数 $1/2$

损失变快慢仍需乘内层导数

外层变化： $d\ell/de = e$

推导 1/6 当 $e=-2$ 时， $d\ell/de$ 是多少？这是否就是 $d\ell/dw$ ？

求导第二步：链式法则带来特征x

梯度是预测误差与输入敏感度的乘积

计算图：每条边表示一次局部变化

参数 w

误差 $e=wx-y$

损失 $\ell=e^2/2$

$$de/dw = x$$

$$e_i=wx_i-y_i; \quad \ell_i=1/2e_i^2$$

$$de_i/dw=x_i$$

$$d\ell_i/dw=(d\ell_i/de_i)(de_i/dw)$$

$$d\ell_i/dw=(wx_i-y_i)x_i$$

x 来自预测对参数的导数

符号给方向，大小给局部敏感度

所有样本随后对同一 w 求和

链式法则： $d\ell/dw = e \cdot x$

推导 2/6 同样 $e=-1$ ， $x=1$ 与 $x=4$ 的参数导数分别是多少？

四个样本分别贡献多少梯度

在 $w_0 = 0$ 处，把每一项都算出来

x	y	e (w=0)	e•x
1	1	-1	-1
2	2	-2	-4
3	2	-2	-6
4	4	-4	-16

$$d\ell_i/dw = (wx_i - y_i)x_i$$

$$w_0 = 0 \Rightarrow e_i = -y_i$$

单样本贡献：(-1, -4, -6, -16)

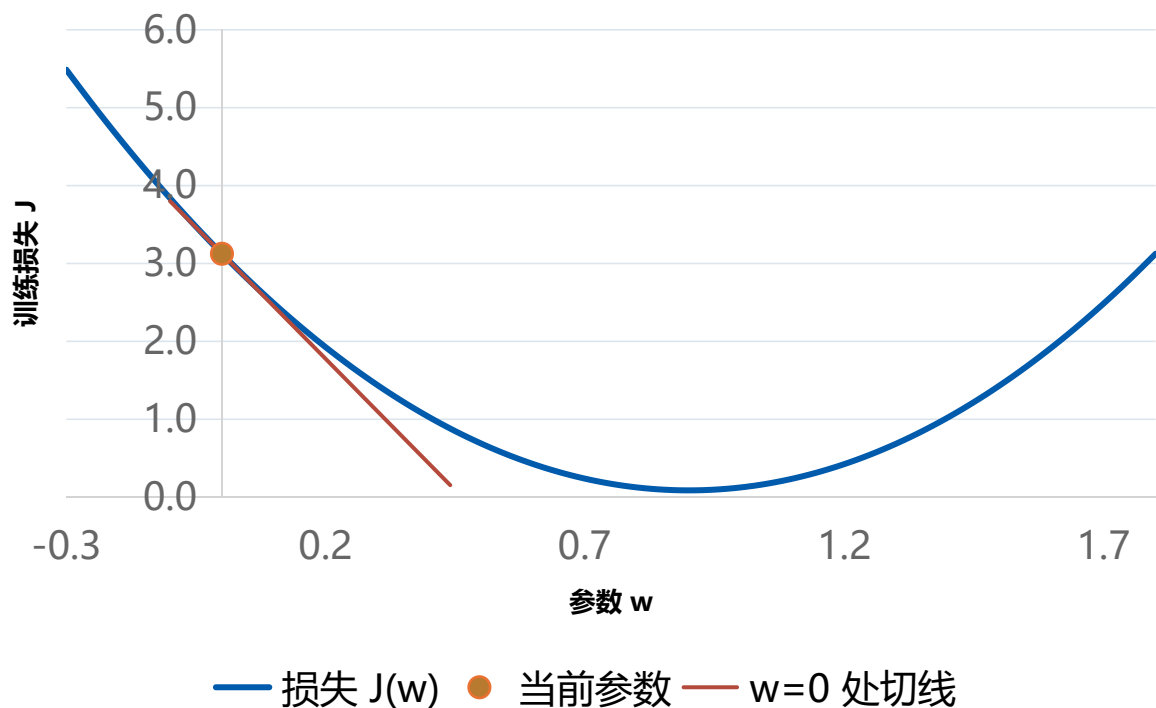
每一行都对同一个参数求导

第四样本的梯度贡献绝对值更大

先列逐项贡献，避免漏乘x

把逐项梯度平均：得到 $g_0 = -6.75$

归一化来自目标定义，不是更新时临时决定



$$J = (1/N) \sum_i \ell_i$$

$$g(w) = (1/N) \sum_i (wx_i - y_i) x_i$$

$$g(0) = (-1 - 4 - 6 - 16) / 4$$

$$g_0 = -27/4 = -6.75$$

导数可以逐项求和再平均

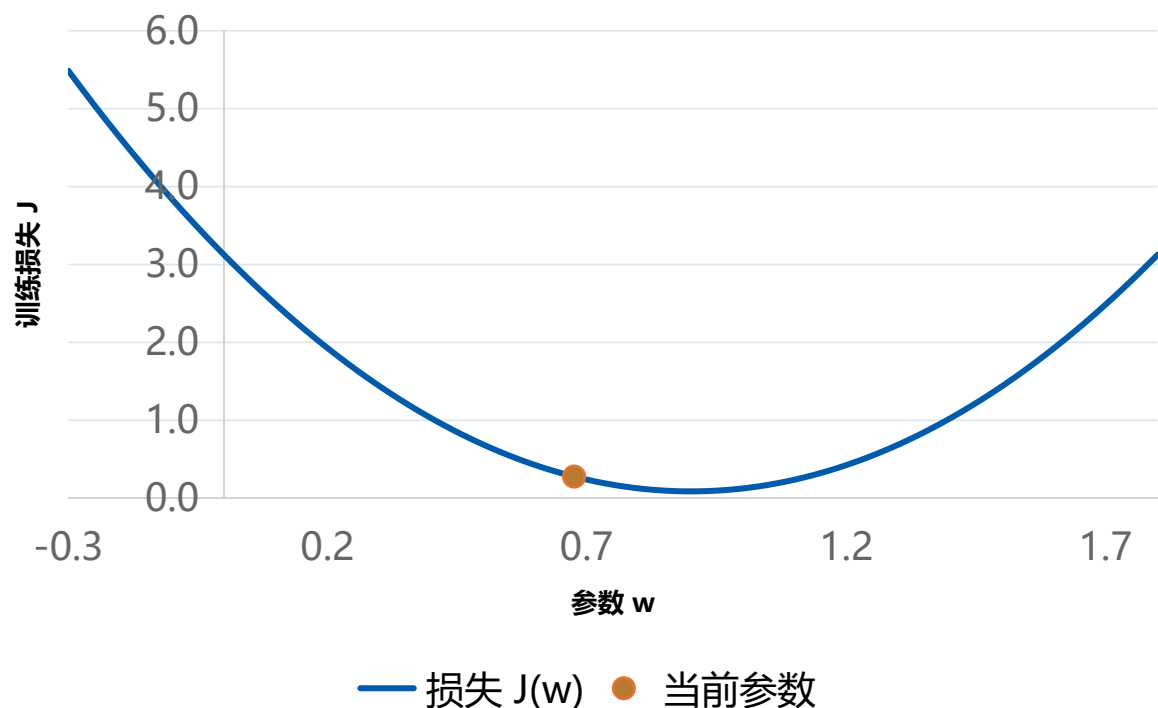
平均与 J 的 $1/(2N)$ 约定一致

负梯度指向增大 w 的一侧

推导 4/6 若误把梯度写成 -27 ，同样学习率会产生什么后果？

沿负梯度更新: $w_1 = 0.675$

更新方向与学习率共同决定下一步位置



$$g_0 = -6.75; \quad \eta = 0.1$$

$$w_1 = w_0 - \eta g_0$$

$$w_1 = 0 - 0.1 \times (-6.75) = 0.675$$

学习率控制走多远

负梯度方向与梯度方向相反

一个正确步骤通常不是精确最优点

推导 5/6 梯度是负数, 为什么 w 反而增加了?

更新后重新计算：损失是否下降

预测、残差与损失都必须由新参数重算

x	预测 $\hat{y}$	残差 r	r^2
1	.675	.325	.105625
2	1.350	.650	.422500
3	2.025	-.025	.000625
4	2.700	1.300	1.690000

$$w_1 = .675; \hat{y} = (.675, 1.35, 2.025, 2.7)$$

$$r = (.325, .65, -.025, 1.3)$$

$$MSE = 2.21875 / 4 = .5546875$$

$$J_1 = .27734375 < J_0 = 3.125$$

第三样本此时已经轻微高估

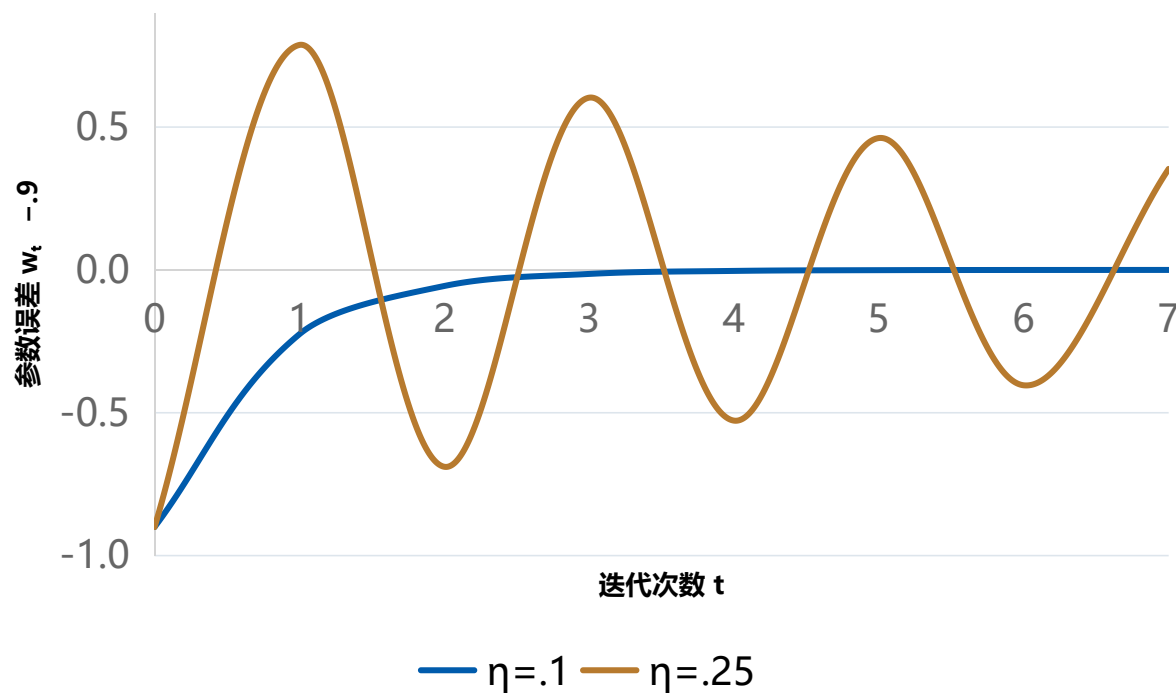
其他样本仍推动斜率继续增大

损失下降不等于所有残差都同向

推导 6/6 J下降时，是否每个样本的绝对残差都必须下降？

学习率为什么有上限

本例可以精确写出误差如何递推



$$g(w) = 7.5w - 6.75 = 7.5(w - 0.9)$$

$$w_{t+1} - 0.9 = (1 - 7.5\eta)(w_t - 0.9)$$

收敛要求 $|1 - 7.5\eta| < 1$

所以 $0 < \eta < 2/7.5 \approx 0.2667$

$\eta = 0.1$: 误差每步乘.25

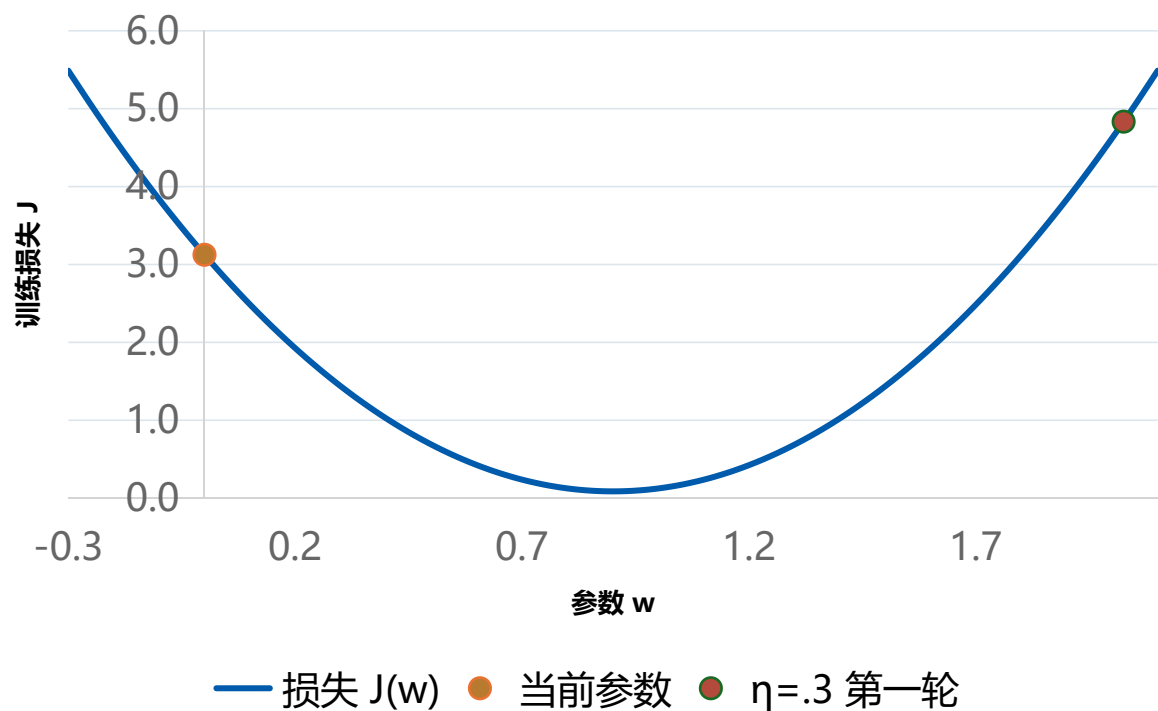
$\eta = 0.25$: 误差变号且绝对值乘.875

凸性不保证任意步长都收敛

推导 1/3 $\eta = 0.25$ 时参数跨过最优点, 是否说明算法一定失败?

大步长反例：凸函数上仍会越走越远

$\eta=0.3$ 超过本例稳定上界



$$\eta=0.3 \Rightarrow 1-7.5\eta=-1.25$$

$$w_0=0 \rightarrow w_1=2.025 \rightarrow w_2=-.50625$$

$$J_0=3.125 \rightarrow J_1=4.83359375$$

$|w_t - 0.9|$ 每步放大1.25倍

方向正确也可能因为步子太大而失败

检查损失与参数，不只看是否有输出

上界属于当前目标与特征尺度

推导 2/3 为什么不能把0.2667当作所有梯度下降问题的步长上界？



过程演示：四点拟合与损失曲线同步移动

先预测更新方向，再用界面与代码核验

```
x = np.array([1., 2., 3., 4.])
y = np.array([1., 2., 2., 4.])
w, eta = 0., .1
for step in range(5):
    e = w*x - y
    J = np.mean(e*e)/2
    g = np.mean(e*x)
    print(step, w, J, g)
    w -= eta*g
```

操作与观察

①四点页签固定 $w_0 = 0$ 、 $\eta = .1$

③损失曲线页查看J并改变步长

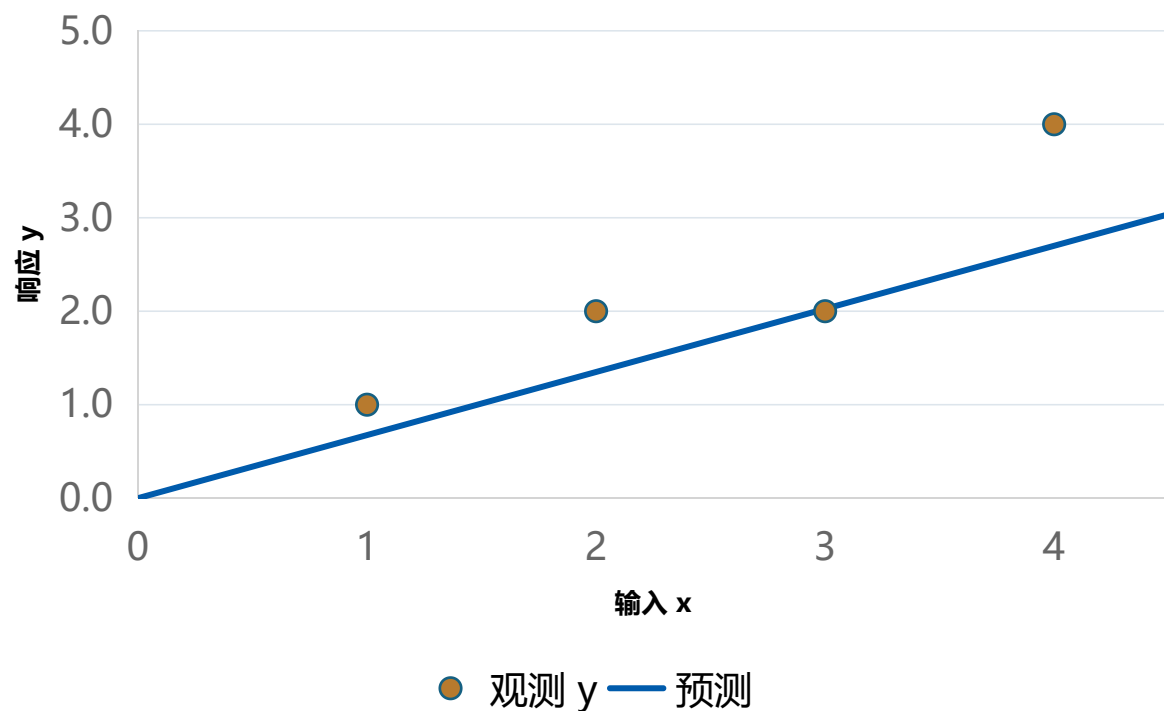
使用隐藏/揭示答案先让学生计算

$\eta = .1$: $0 \rightarrow .675 \rightarrow .84375 \rightarrow \dots$

推导 3/3 第二次 $\eta = .1$ 更新后的w是多少?

学生验收：把一次更新完整交出来

先独立计算，再交换检查符号与归一化



已知 $w=.675$, $\eta=.1$

任务：算 $\hat{y}$ 、 e 、 g 、 w^+ 、 $J(w^+)$

核对： $g=-1.6875$, $w^+=.84375$

$J(.84375)=.099365234375$

先判断方向，再写数值

每一步保留至少一个可检查中间量

区分残差 r 与预测误差 e

推导 1/1 如果有人得到 $w^+ = 1.35$ ，最可能先检查什么？

多特征模型：一个参数对应一种可用信息

从单斜率扩展到多个系数，预测仍是加权相加

输入数据

共享参数

$$\hat{y} = b + w_1 x_1 + w_2 x_2$$

距离 x_1 $\times w_1$

坡度 x_2 $\times w_2$

例：用距离与坡度预测骑行时间

参数共享；每个样本有自己的特征值

特征由任务决定，不由求解器凭空发现

交互项可以表达组合影响

预测时不可获得的信息不得进入输入

相加并加截距 b

输出：预测骑行时间 $\hat{y}$

推导 1/2 距离乘坡度可以作为一个特征吗？加入它是否仍对参数线性？

把截距与特征装进同一个向量

常数特征1让截距进入统一的内积

样本 i	常数 1	输入 x	观测 y
1	1	1	1
2	1	2	2
3	1	3	2
4	1	4	4

Φ : 常数与输入两列
 y : 末列监督信号

$$\varphi(x) = [1, x_1, \dots, x_p]^T \in \mathbb{R}^d$$

$$\mathbf{w} = [b, w_1, \dots, w_p]^T, \quad d = p + 1$$

$$\hat{y} = \varphi(x)^T \mathbf{w} = \mathbf{w}^T \varphi(x)$$

第一维是截距特征，恒等于1

向量默认用列向量

内积结果是一个标量预测

设计矩阵第一行：先把一个样本写对

回到四点数据，此组加入截距作矩阵示范

样本 i	常数 1	输入 x	观测 y
1	1	1	1
2	1	2	2
3	1	3	2
4	1	4	4

Φ : 常数与输入两列
 y : 末列监督信号

$$\varphi(\mathbf{x}_1)^T = [1, 1]$$

$$\mathbf{w} = [b, a]^T = [.5, .8]^T$$

$$\hat{y}_1 = [1, 1][.5, .8]^T = 1.3$$

这里a是斜率，b是截距

矩阵示范参数不同于前面的无截距手算

一行对应一个样本

推导 1/4 第一行中的两个1意义相同吗？

把四个特征行堆起来：得到 Φ

列代表特征，行代表样本

样本 i	常数 1	输入 x	观测 y
1	1	1	1
2	1	2	2
3	1	3	2
4	1	4	4

Φ : 常数与输入两列
 y : 末列监督信号

$$\Phi = [[1, 1], [1, 2], [1, 3], [1, 4]]$$

$$\mathbf{w} = [0.5, 8]^T; \quad \mathbf{y} = [1, 2, 2, 4]^T$$

$$\Phi \in \mathbb{R}^{4 \times 2}, \quad \mathbf{w} \in \mathbb{R}^2, \quad \mathbf{y} \in \mathbb{R}^4$$

第一列全1，对应共享截距

第二列保存四个输入

增加样本加行，增加特征加列

推导 2/4 交换 Φ 第二行与第三行，却不交换 y ，会不会一定出现维度报错？

矩阵乘法逐元素：四个内积同时计算

Φw 是一套共享参数对所有样本的批量预测

样本 i	常数 1	输入 x	观测 y
1	1	1	1
2	1	2	2
3	1	3	2
4	1	4	4

Φ : 常数与输入两列
 y : 末列监督信号

$$\hat{y}_1 = .5 + .8 \times 1 = 1.3$$

$$\hat{y}_2 = .5 + .8 \times 2 = 2.1$$

$$\hat{y}_3 = .5 + .8 \times 3 = 2.9$$

$$\hat{y}_4 = .5 + .8 \times 4 = 3.7$$

矩阵乘法没有引入新预测规则

每一行使用同一个参数列

$$e = (.3, .1, .9, -.3)^T, \quad r = -e$$

推导 3/4 第二行预测2.1对应的残差与预测误差分别是什么？

维度核对：写代码以前先判断乘法是否有意义

用维度排除一大类静默广播错误

对象	维度	含义
Φ	$N \times d$	样本 $\times$ 特征
w	$d \times 1$	每种特征一个参数
$\hat{y}$ 、 y 、 e	$N \times 1$	每个样本一个值
$\Phi^T e$	$d \times 1$	回到参数空间

$$\Phi: N \times d; w: d \times 1$$

$$\hat{y} = \Phi w: N \times 1; e = \hat{y} - y: N \times 1$$

$$\Phi^T e: (d \times N)(N \times 1) = d \times 1$$

预测与观测必须按样本对齐

参数与梯度必须同形

NumPy一维数组规则需额外留意

推导 4/4 4维一维数组减 4×1 数组，可能得到什么形状？

矩阵损失：把四项平方写成一个范数

先定义误差，再写内积，不跳过中间对象

x_i	e_i	$e_i \times 1$	$e_i \times x_i$
1	.3	.3	.3
2	.1	.1	.2
3	.9	.9	2.7
4	-.3	-.3	-1.2

示例 $b=.5$, $a=.8$

两列贡献取平均： $\nabla J = (.25, .5)^T$

$$e = \Phi w - y; \quad r = -e$$

$$\sum_i e_i^2 = e^T e = \|e\|^2$$

$$J(w) = e^T e / (2N)$$

$$J = (\Phi w - y)^T (\Phi w - y) / (2N)$$

范数平方与逐项平方和完全一致

目标必须是一个标量

矩阵记号压缩重复运算而非省略含义

推导 1/6 若误写 ee^T ，结果和 $e^T e$ 相同吗？

微分展开：保留两项，再利用标量相等

使用微分把矩阵求导拆成可检查的小步骤

x_i	e_i	$e_i \times 1$	$e_i \times x_i$
1	.3	.3	.3
2	.1	.1	.2
3	.9	.9	2.7
4	-.3	-.3	-1.2

示例 $b=.5$, $a=.8$
两列贡献取平均: $\nabla J = (.25, .5)^T$

$$J = e^T e / (2N), \quad de = \Phi dw$$

$$dJ = [(de)^T e + e^T de] / (2N)$$

$(de)^T e = e^T de$: 均为同一标量

$$dJ = e^T \Phi dw / N$$

Φ 与 y 是固定数据, 只有 w 变化

乘积微分先产生两项

标量转置不改变数值

识别梯度：从dJ读出 ∇J

梯度必须与参数具有相同维度

x_i	e_i	$e_i \times 1$	$e_i \times x_i$
1	.3	.3	.3
2	.1	.1	.2
3	.9	.9	2.7
4	-.3	-.3	-1.2

示例 $b=.5$, $a=.8$
两列贡献取平均: $\nabla J = (.25, .5)^T$

$$dJ = e^T \Phi dw / N$$

$$dJ = (\nabla J)^T dw$$

$$\nabla J = \Phi^T e / N = \Phi^T (\Phi w - y) / N$$

$$\nabla J_j = (1/N) \sum_i e_i \Phi_{ij}$$

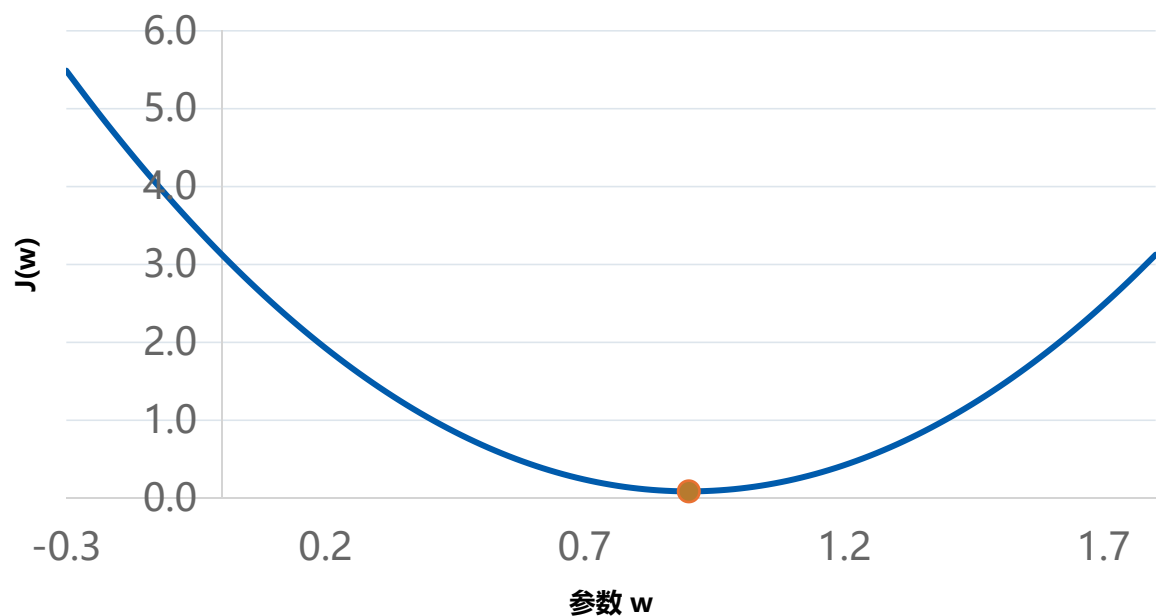
截距梯度就是误差平均值

矩阵式可逐分量还原

推导 3/6 示范参数 $b=.5$ 、 $a=.8$ 时，两维梯度分别是多少？

Hessian与凸性：最小点为什么没有局部陷阱

半正定保证凸性，满列秩进一步保证唯一参数



$$\nabla J = \Phi^T(\Phi w - y)/N$$

$$H = \nabla^2 J = \Phi^T \Phi / N$$

$$v^T H v = \|\Phi v\|^2 / N \geq 0$$

Φ 满列秩 $\Rightarrow H > 0 \Rightarrow$ 参数唯一

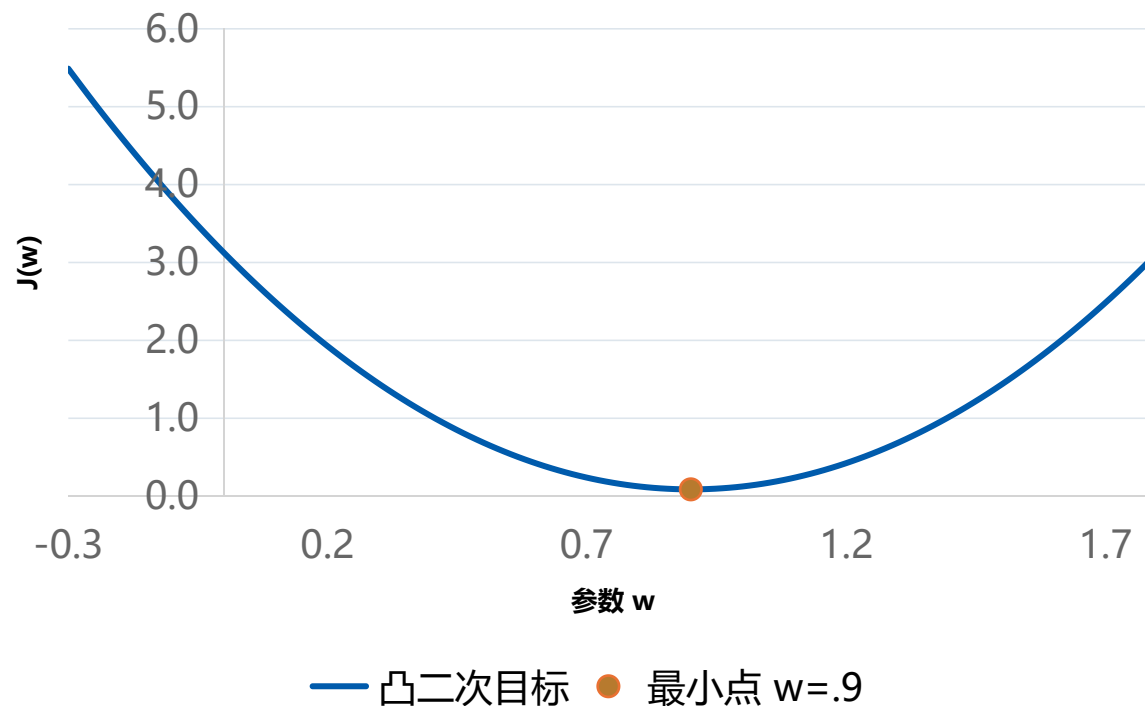
平方损失对固定特征参数是凸函数

秩亏时可能沿某些方向完全平坦

凸性与学习率稳定性是两个问题

正规方程：把最优点的梯度设为零

由优化条件进入线性方程求解



$$\nabla J(\hat{w}) = \Phi^T(\Phi\hat{w} - y)/N = 0$$

$$\Phi^T\Phi\hat{w} = \Phi^Ty$$

满列秩时： $\hat{w} = (\Phi^T\Phi)^{-1}\Phi^Ty$

实际计算优先QR/SVD，不显式求逆

凸目标的一阶条件足以刻画最优解

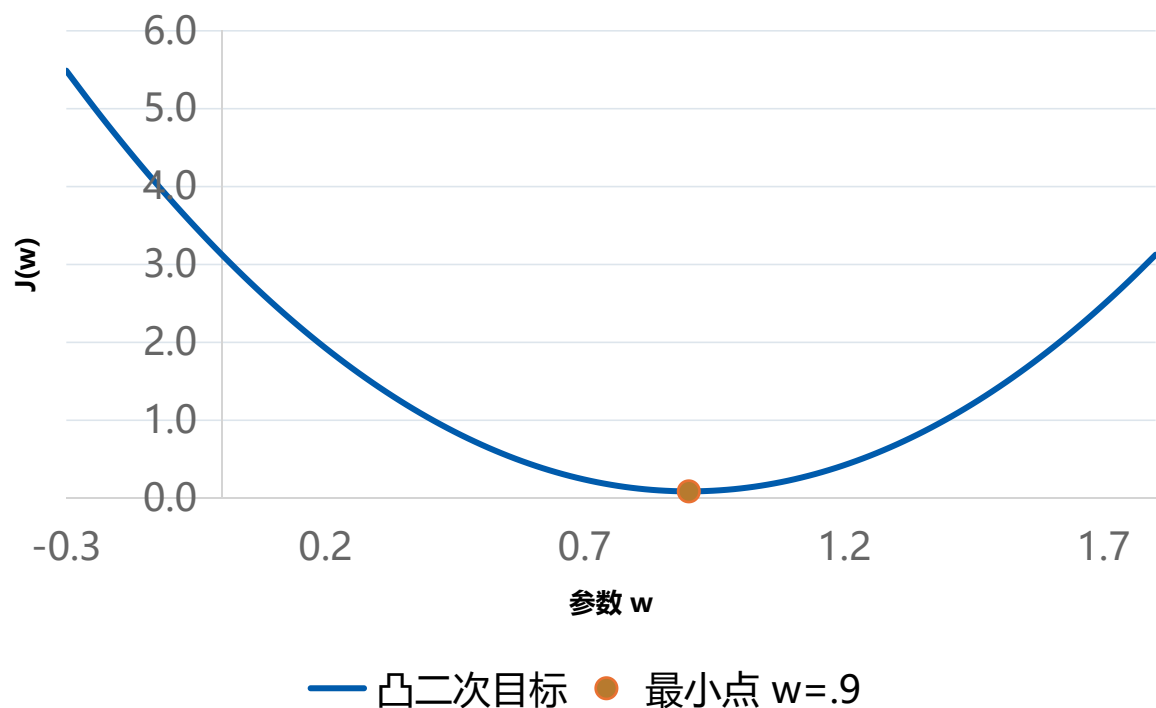
逆矩阵公式需要满列秩

正规方程表达结构，不规定实现方式

推导 5/6 梯度为零只是必要条件，为什么这里也足够？

四点解析解：与梯度下降终点核对

无截距时，正规方程只剩一个标量等式



$$\sum x_i^2 \cdot \hat{w} = \sum x_i y_i$$

$$(1+4+9+16)\hat{w} = 1+4+6+16$$

$$30\hat{w} = 27 \Rightarrow \hat{w} = .9$$

$$\mathbf{r} = (.1, .2, -.7, .4)^T, J^* = .0875$$

回到最初无截距模型作核对

最优残差仍可非零

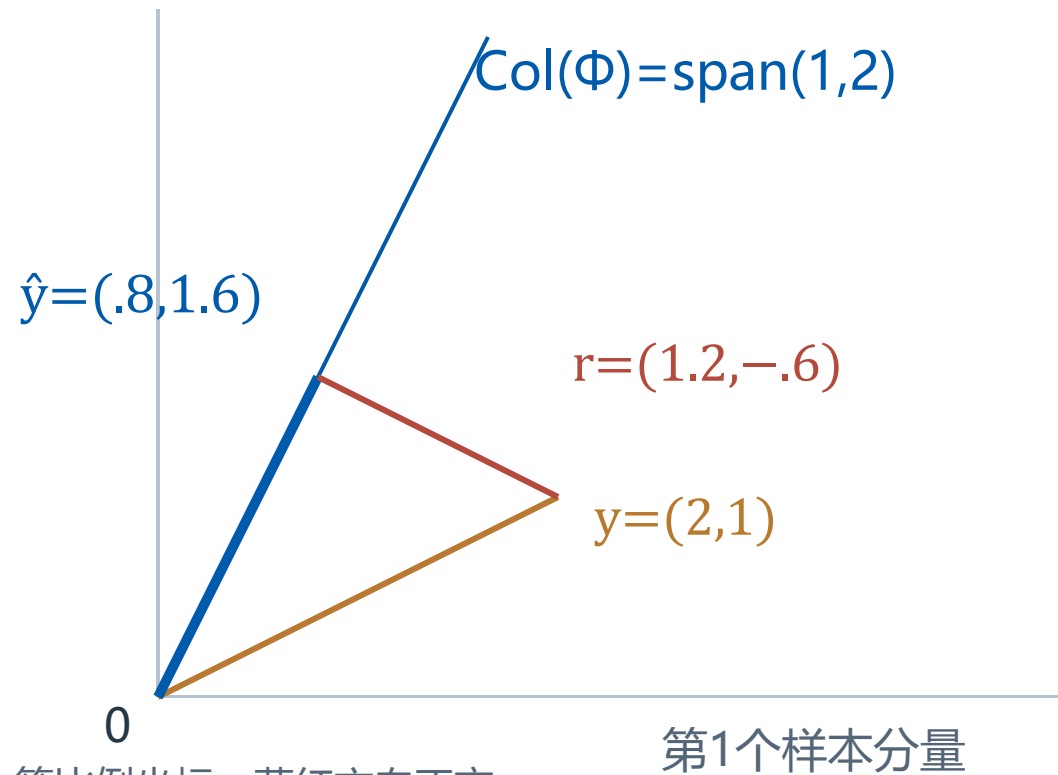
迭代法与解析法应求解同一个目标

推导 6/6 最优损失不是零，是否说明求解器不够精确？

向量投影：两样本预测构成一条子空间直线

独立二维几何例：坐标表示两个样本的预测分量

第2个样本分量



等比例坐标，蓝红方向正交

$$\Phi = (1, 2)^T, \quad y = (2, 1)^T; \quad \text{无截距}$$

$$\hat{w} = \Phi^T y / (\Phi^T \Phi) = 4/5 = .8$$

$$\hat{y} = \Phi \hat{w} = (.8, 1.6)^T$$

$$\text{Col}(\Phi) = \text{span}\{(1, 2)^T\}$$

横纵坐标是样本1与样本2的分量

所有预测都在 $\text{span}\{\Phi\}$ 这条直线上

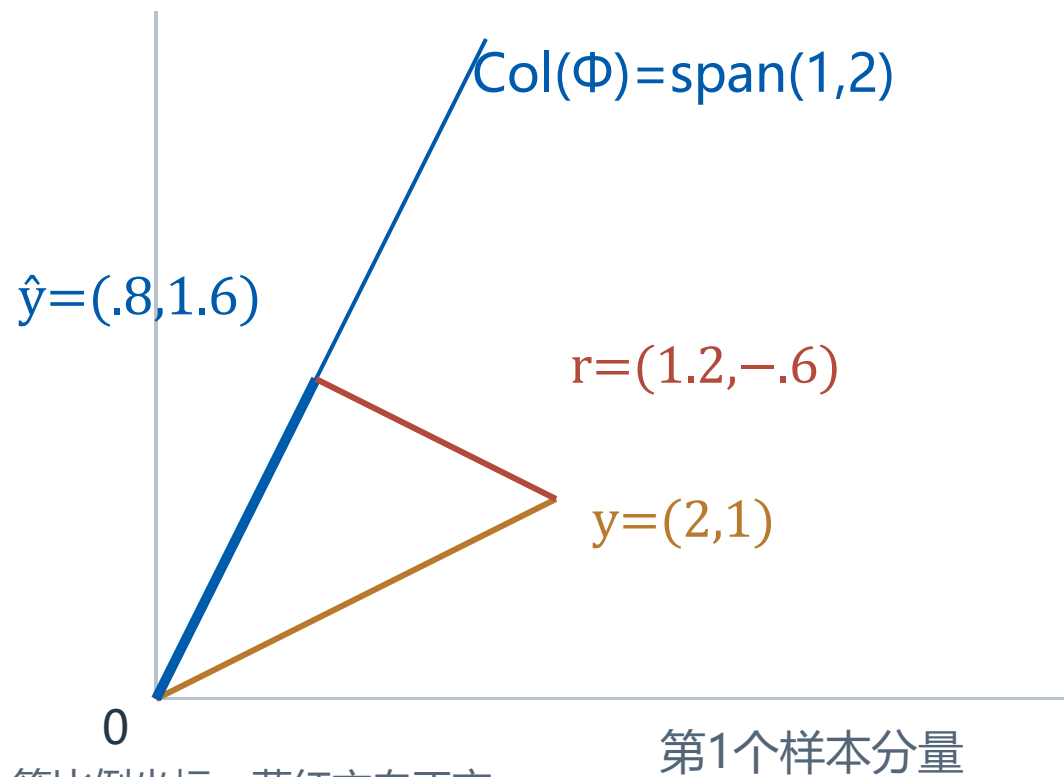
本页独立两样本例，不是原四点散点图

推导 1/3 为什么所有可能预测只能沿同一条直线移动？

正交条件：残差不能再被现有特征解释

最优性方程与投影垂直性是同一关系的两种表达

第2个样本分量



$$r = y - \hat{y} = (1.2, -0.6)^T$$

$$\Phi^T r = 1 \times 1.2 + 2 \times (-0.6) = 0$$

$$r \perp \text{Col}(\Phi)$$

$$\text{一般式: } \Phi^T (y - \Phi \hat{w}) = 0$$

本例残差非零，但与特征方向正交

无截距时不保证残差和为零

含截距的无正则最小二乘才有 $\sum r = 0$

推导 2/3 残差与当前特征正交，是否说明它与所有输入函数都无关？

秩亏：不同参数可以给出相同预测

冗余特征使某些参数方向不可识别

01 只能识别组合 $a+2b$

$$\Phi=[x,2x], \quad w=(a,b)^T$$

02 最优训练预测唯一，参数可能不唯一

$$\Phi w=x(a+2b)$$

$$(a,b)=(.9,0) \text{ 与 } (.1,.4)$$

03 伪逆选择最小范数解，不创造新信息

$$\text{两者都给出 } \hat{y}=.9x$$

推导 3/3 如果把a减少0.2、b增加0.1，预测改变吗？



QR与SVD：按原问题求解更稳定

正规方程用于理解，最小二乘接口用于计算

```
w, rss, rank, singular = np.linalg.lstsq(
    Phi, y, rcond=None)
yhat = Phi @ w
print("rank:", rank)
print("singular values:", singular)
print("MSE:", np.mean((yhat-y)**2))
```

操作与观察

形成 $\Phi^T\Phi$ 可能放大病态

`np.linalg.lstsq`直接接收 Φ 与 y

预测稳定不代表每个系数都稳定

满列秩： $\kappa_2(\Phi^T\Phi)=\kappa_2(\Phi)^2$

推导 1/3 最小二乘函数成功返回参数，是否保证每个系数都有稳定科学解释？

标准化：改变坐标尺度，不增加新信息

统计量只在训练数据拟合，并随模型保存

01 尺度差异影响优化与正则化

$$\mu_j = \text{mean}(X_{\text{train},j}), \quad s_j = \text{std}(X_{\text{train},j})$$

02 常数列和零方差特征需单独处理

$$\mathbf{x}'_j = (\mathbf{x}_j - \mu_j) / s_j$$

验证与测试使用同一 μ_j 、 s_j

03 标准化不能自动消除特征共线性

推导 2/3 在测试集上重新计算自己的均值和标准差，是否与训练规则一致？



矩阵代码演示：每个@都能解释成内积

用设计矩阵页签核对，再运行批量实现

```
x = np.array([1., 2., 3., 4.])
y = np.array([1., 2., 2., 4.])
Phi = np.c_[np.ones(4), x]
w = np.array([.5, .8])
pred = Phi @ w
e = pred - y
grad = Phi.T @ e / len(y)
print(Phi.shape, pred, grad)
```

操作与观察

先核对shape，再核对数值

网页逐行展示，代码一次计算

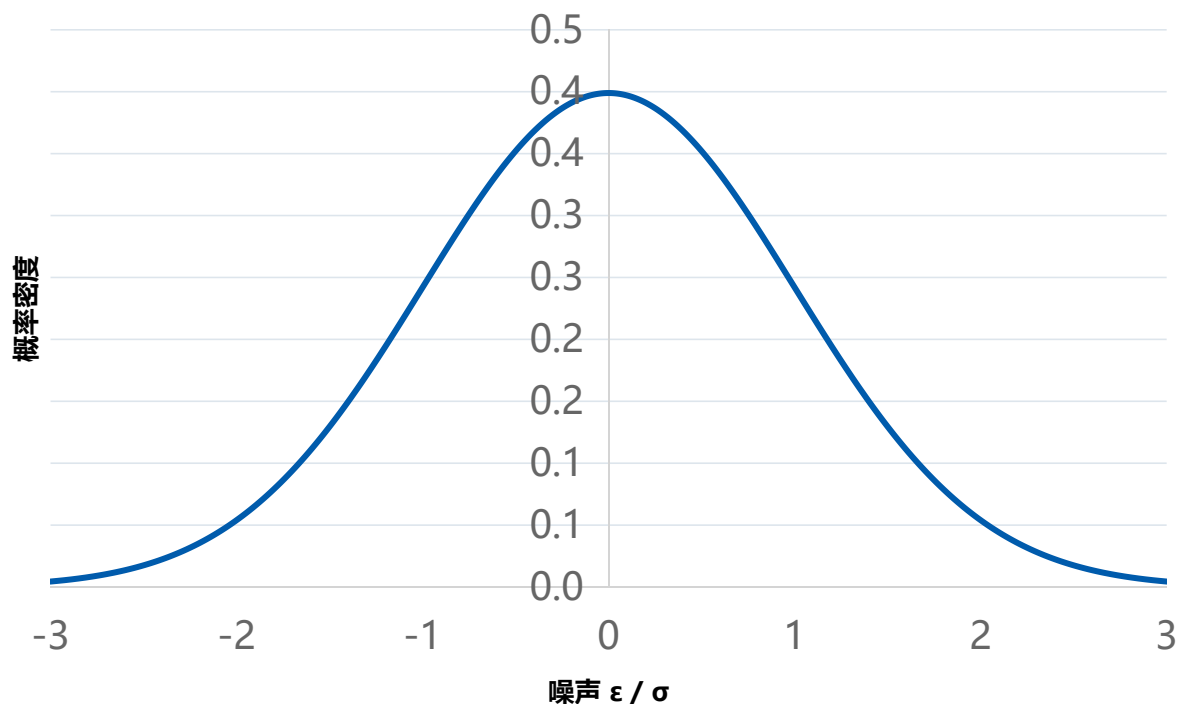
任何差异先检查截距与数据顺序

默认 $b=.5$, $a=.8$

推导 3/3 将 b 增加0.1，所有预测与截距梯度怎样改变？

噪声与残差：生成假设不同于拟合后的量

残差包含随机扰动，也可能包含模型与估计误差



$$y = f^*(x) + \varepsilon, \quad E[\varepsilon|x] = 0 \quad (\text{假设})$$

$$r = y - \hat{f}(x)$$

$$r = \varepsilon + [f^*(x) - \hat{f}(x)]$$

真实函数通常未知，噪声通常不可直接观测

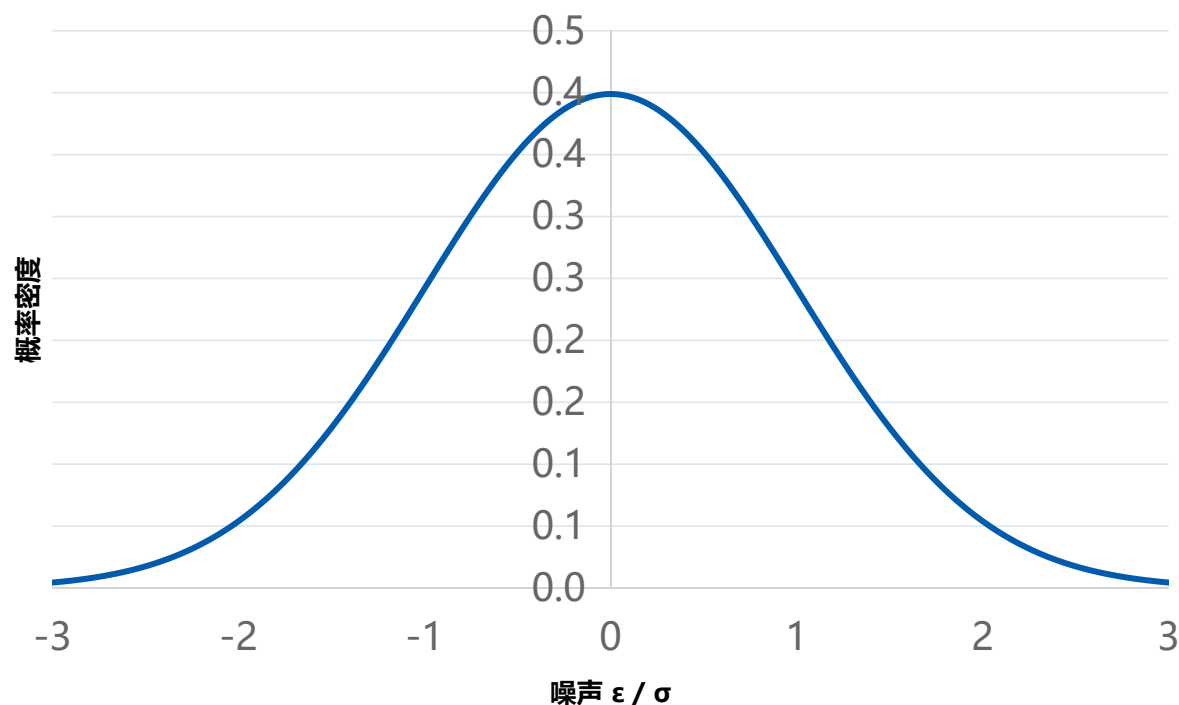
残差由数据与已拟合模型计算

残差有结构时应检查模型和假设

推导 1/1 总体平均噪声为零，是否等价于每个输入条件下平均噪声都为零？

高斯密度：固定输入后，观测围绕均值波动

平方残差先出现在概率密度的指数里



$$y_i | x_i, w \sim N(\varphi_i^T w, \sigma^2)$$

$$p(y_i | x_i, w) = (2\pi\sigma^2)^{-1/2} \times \exp[-r_i^2 / (2\sigma^2)]$$

$$r_i = y_i - \varphi_i^T w$$

均值由模型预测，方差先视为固定正数

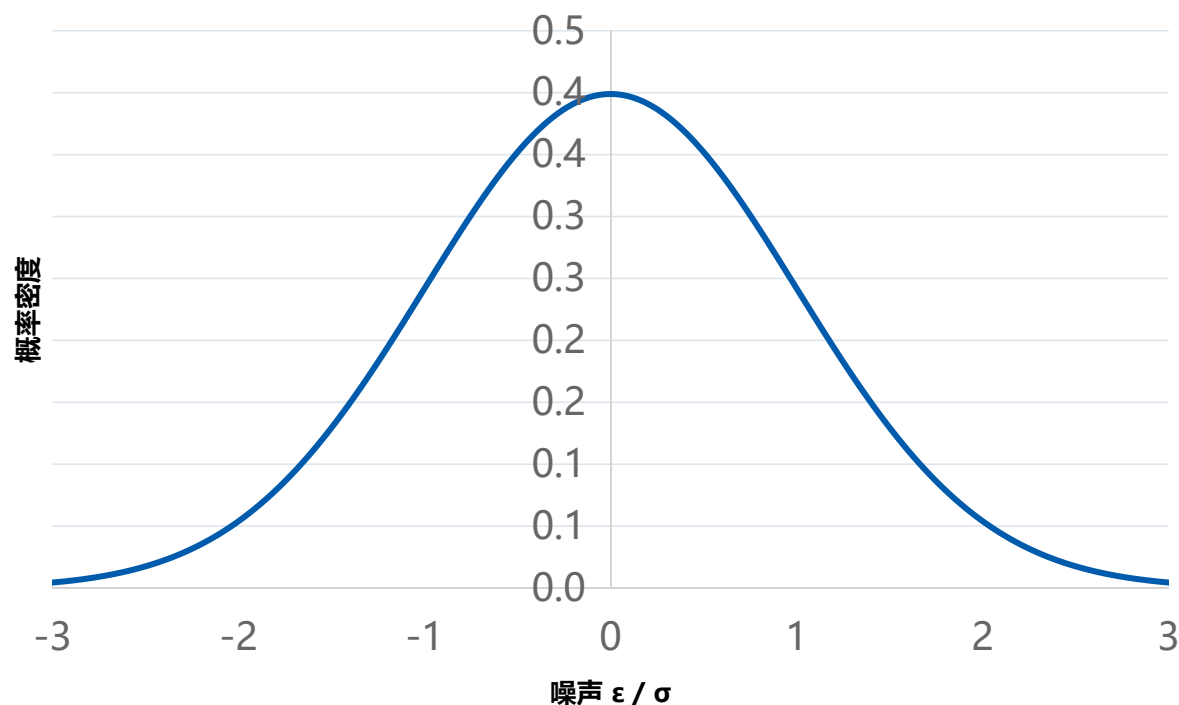
连续变量使用密度，不是单点概率

离均值越远，密度通常越低

推导 1/4 高斯密度值能超过1吗？是否违反概率不超过1？

联合似然：同一参数怎样解释全部观测

条件独立使联合密度成为逐样本密度的乘积



$$L(w) = \prod_i p(y_i | x_i, w)$$

$$L(w) = (2\pi\sigma^2)^{-N/2}$$

$$\times \exp[-\sum_i r_i^2 / (2\sigma^2)]$$

所有样本使用同一组参数 w

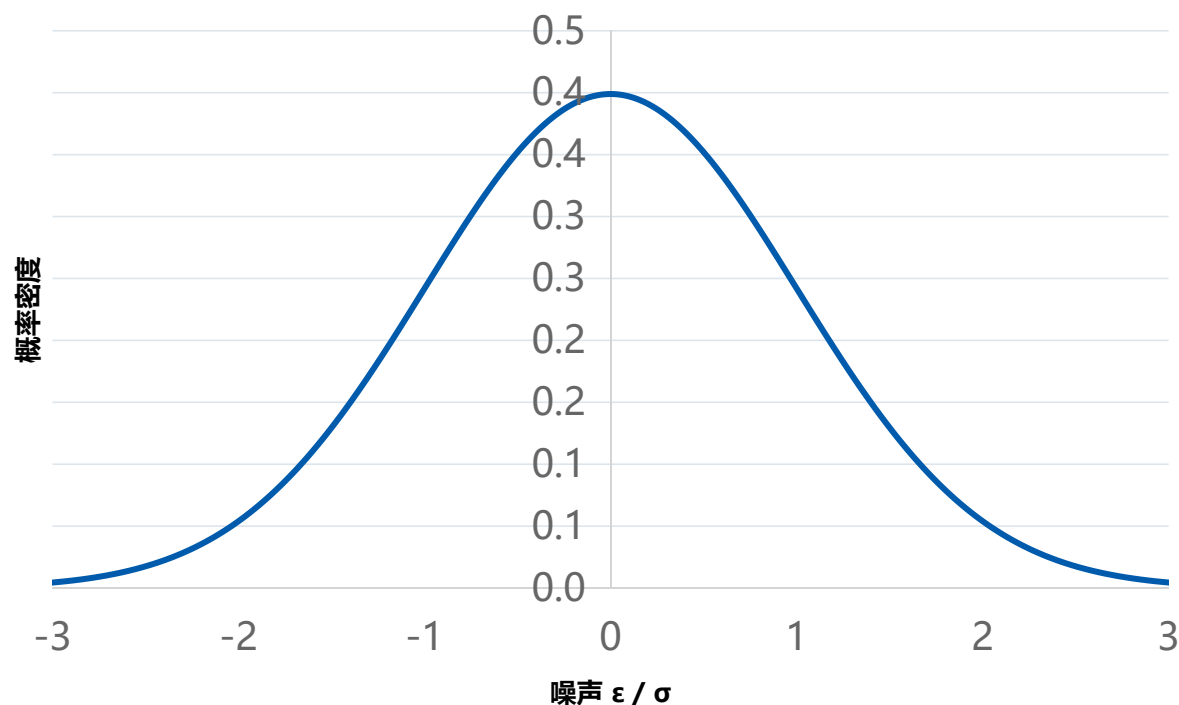
相乘需要条件独立等建模假设

似然把数据固定，将参数看成变量

推导 2/4 若同一设备的相邻观测明显相关，直接乘独立密度是否自动合理？

取负对数：乘积变求和，最大化变最小化

单调性保留最优参数，同时让表达更易计算



$$-\log L(w) = N/2 \cdot \log(2\pi\sigma^2) + \sum_i r_i^2 / (2\sigma^2)$$

固定 σ^2 ：第一项与 w 无关

$$\operatorname{argmax} L = \operatorname{argmin} \sum_i r_i^2$$

log把多个密度相乘转为相加

负号把最大似然改写成最小化

仅删除与所优化参数无关的项

推导 3/4 把 σ^2 也作为待估参数时， $N/2 \cdot \log(2\pi\sigma^2)$ 还能删除吗？

最小二乘的概率解释有明确前提

计算算法可以使用，似然解释必须说明假设

假设	在推导中的作用
均值= $\varphi^T w$	把参数放进残差
条件独立	联合密度写成乘积
共同固定方差	各平方残差同权
取负对数	最大化转为最小化

条件高斯+独立+共同固定 σ^2

最大似然 $\hat{w}$ = 最小二乘 $\hat{w}$

$$J(w)=\Sigma_i r_i^2/(2N)$$

J不一定等于未缩放的负对数似然

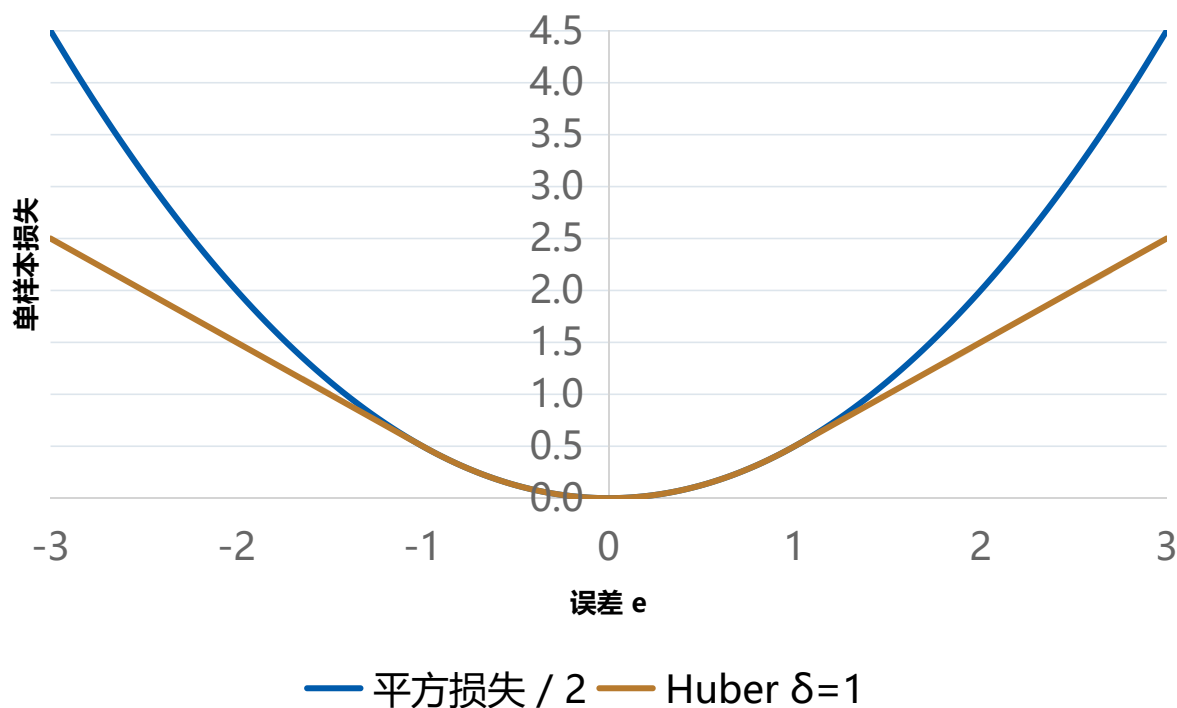
平方损失并不要求先证明严格高斯

高斯假设支持特定概率解释

异方差可能导出加权最小二乘

异常值与Huber: 平方不是唯一误差标准

根据误差含义选择目标, 不自动删除难点



$$\ell_2(r)=r^2/2; \ell_1(r)=|r|$$

$$\text{Huber}\delta(r)=r^2/2, |r|\leq\delta$$

$$\text{Huber}\delta(r)=\delta(|r|-\delta/2), |r|>\delta$$

大残差在平方损失下影响更强

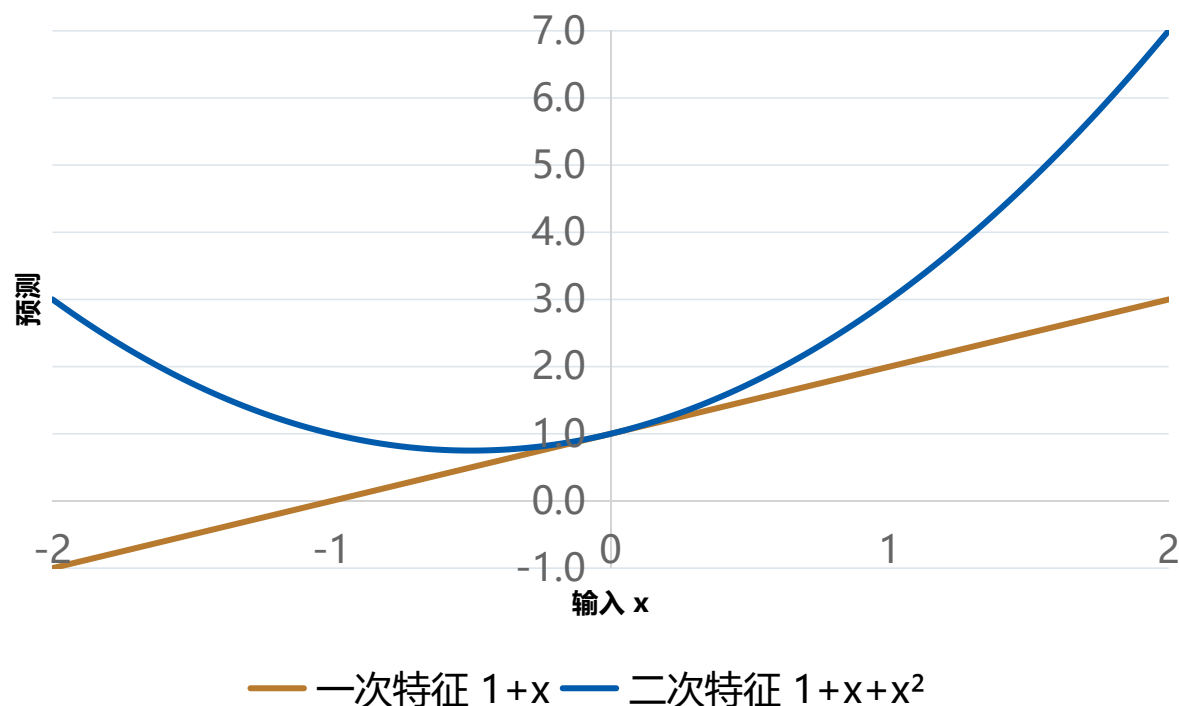
Huber小残差二次、大残差线性

异常点先核查来源, 再决定处理

推导 1/1 $\delta=1$ 、残差 $r=3$ 时, Huber损失是多少?

加入 x^2 ：弯曲关系仍可对参数线性

特征映射改变可表达的函数，矩阵框架保持不变



$$\varphi(x)=[1,x,x^2]^T$$

$$f_w(x)=w_0+w_1x+w_2x^2$$

$$\nabla J=\Phi^T(\Phi w-y)/N \text{ 仍然适用}$$

曲线对 x 非线性，对参数仍线性

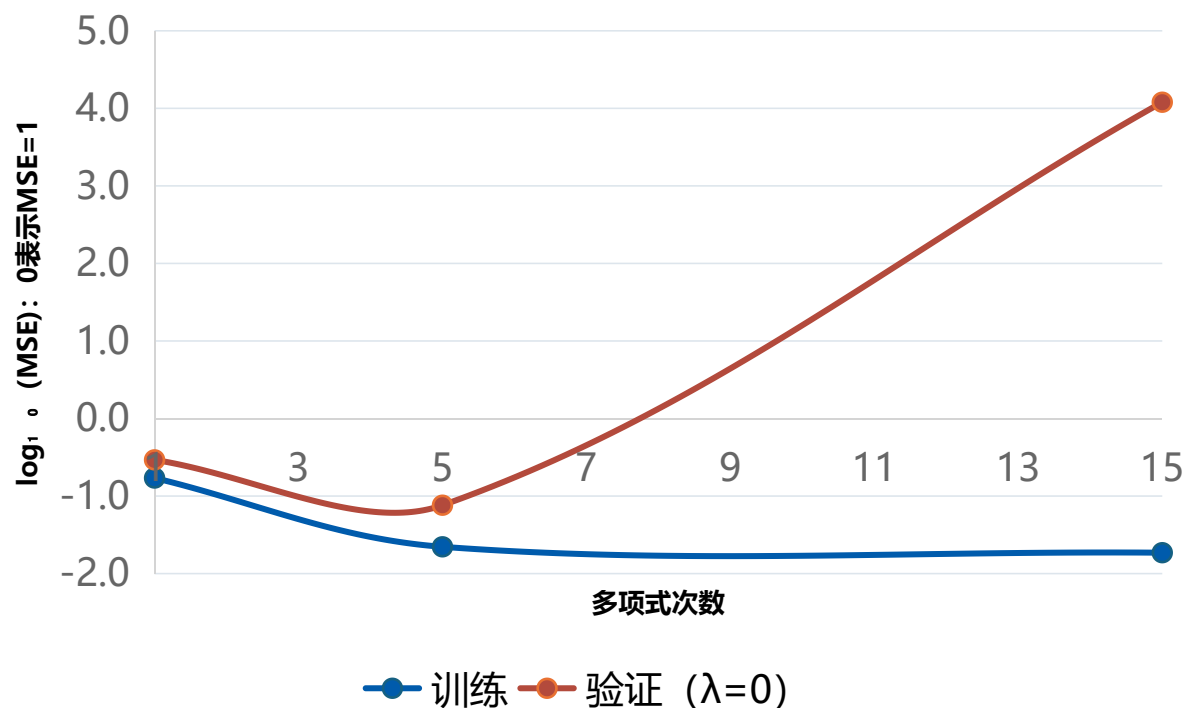
每加一列特征，增加一个系数

特征选择要有任务依据与验证证据

推导 1/3 模型 $w_0 + w_1 x + w_2 x^2$ 属于线性回归吗？

复杂度与泛化：训练更好，不必然预测更好

同一数据协议下比较表达不足、合理拟合与追随噪声



训练目标：已见样本上的平均误差

泛化目标：目标分布中新样本的误差

偏差、估计方差与噪声共同影响结果

训练误差不能独立选择复杂度

高阶特征也可能带来数值病态

一次验证波动不等于确定规律

推导 2/3 训练与验证误差都很高，应直接判断为过拟合吗？

训练、验证、测试：把选择与评价分开

所有从数据学出的步骤都属于训练流程

01 标准化只在训练数据拟合

训练：拟合预处理与模型参数

02 测试结果不能反过来用于调参

验证：选择特征、 λ 、阈值、轮次

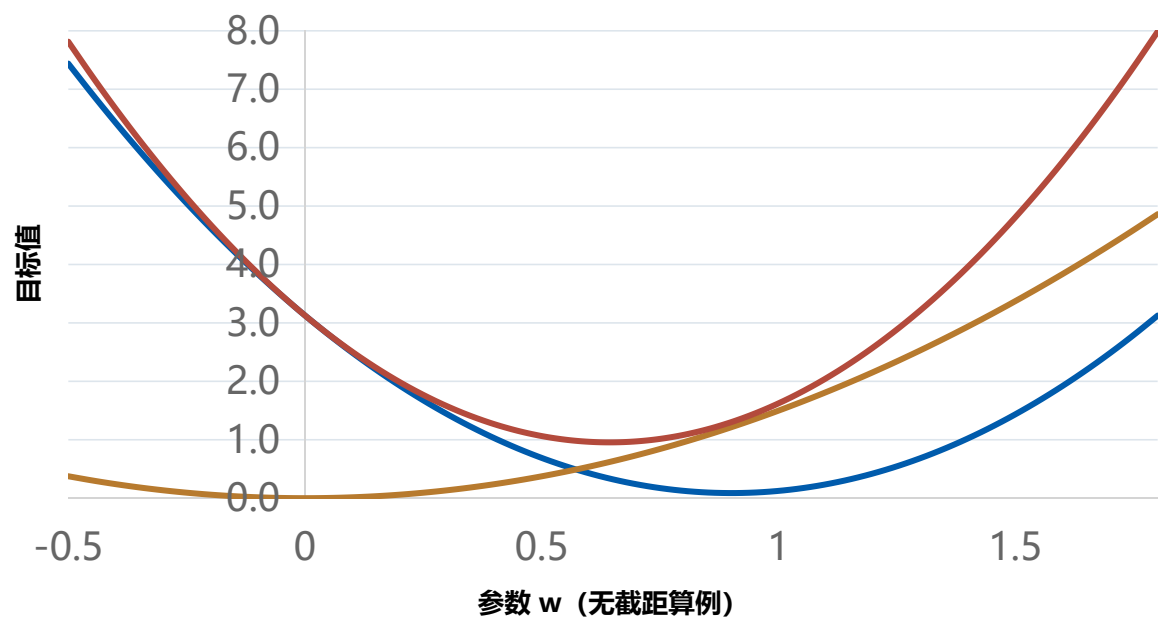
03 相关主体或时间数据按组或时间划分

冻结：固定全部规则后再测试

推导 3/3 不使用测试标签，只用全部输入计算标准化统计量，就一定没有泄漏吗？

L2目标：在拟合与参数规模之间权衡

保持本课平均损失约定，截距不受惩罚



— 数据损失 J — 惩罚项 $\lambda w^2/2$, $\lambda=3$ — 总目标

$$J_\lambda(w) = \|\Phi w - y\|^2 / (2N) + (\lambda/2) w^T D w$$

$$D = \text{diag}(0, 1, \dots, 1), \quad \lambda \geq 0$$

$$w^T D w = \sum_{j \geq 1} w_j^2$$

第一项要求解释观测

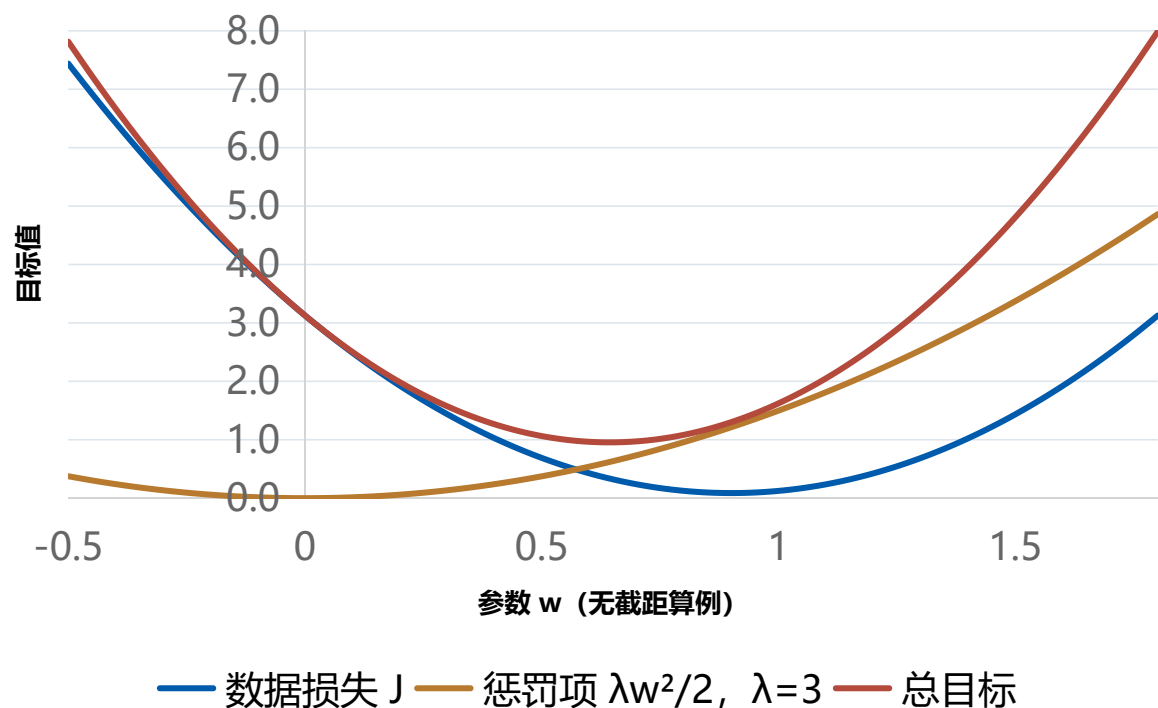
第二项惩罚非截距系数过大

λ 必须由验证选择，越大不一定越好

推导 1/4 为什么同样 λ 不能不加检查地跨特征单位或软件接口比较？

岭梯度：正则化怎样参与每一步更新

在原梯度上增加 λDw ，截距项保持原样



$$J_\lambda = J + (\lambda/2)w^T D w$$

$$\nabla J = \Phi^T(\Phi w - y)/N$$

$$\nabla[(\lambda/2)w^T D w] = \lambda D w$$

$$\nabla J_\lambda = \Phi^T(\Phi w - y)/N + \lambda D w$$

D 对称，二次项求导得到 $2Dw$

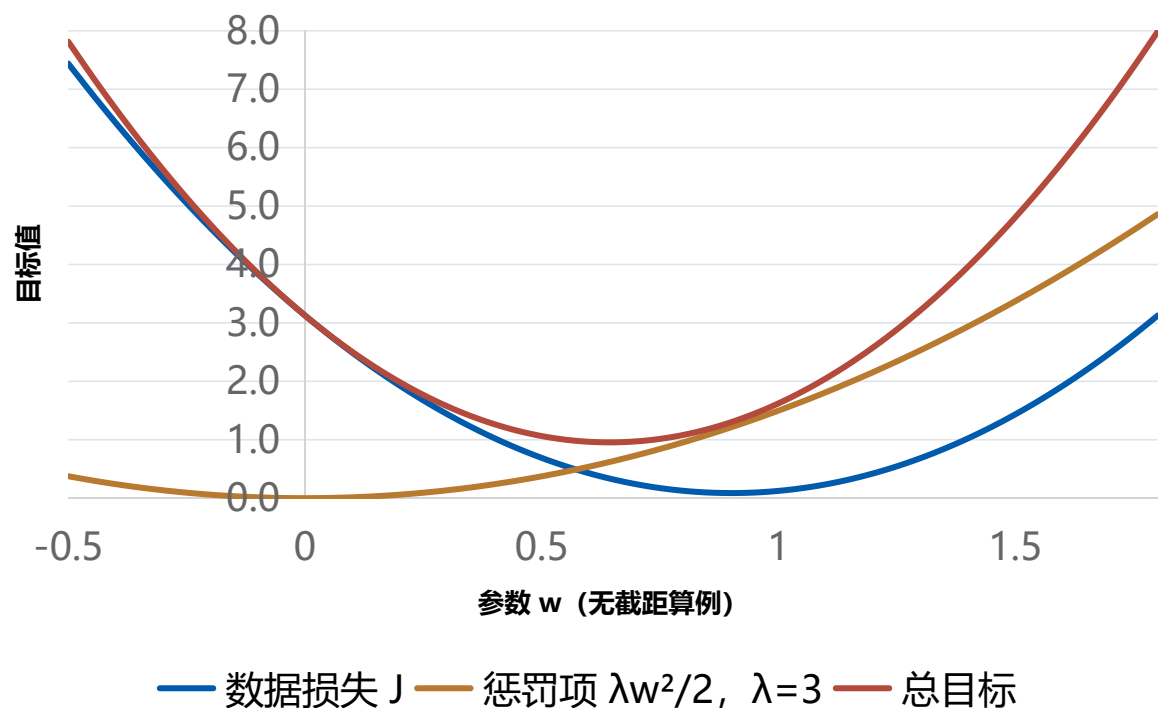
$1/2$ 消去导数中的 2

先求总梯度，再执行一次更新

推导 2/4 λDw 是不是还要除以 N ?

岭正规方程： $N\lambda$ 来自哪里

平均损失的归一化决定方程中的系数



$$0 = \Phi^T(\Phi w - y)/N + \lambda D w$$

$$0 = \Phi^T \Phi w - \Phi^T y + N \lambda D w$$

$$(\Phi^T \Phi + N \lambda D) \hat{w} = \Phi^T y$$

$$\text{无截距四点: } \hat{w} \lambda = 27 / (30 + 4 \lambda)$$

两侧乘 N 时，正则项同时乘 N

$\lambda=0$ 恢复普通最小二乘

本例 $\lambda=1$ 时斜率约.7941

推导 3/4 若误写成 $27/(30+\lambda)$ ，相对于本课定义意味着什么？

增广最小二乘：无需显式形成 $\Phi^T\Phi$

把L2惩罚写成额外平方项，交给稳定求解器

```
D = np.eye(Phi.shape[1])
D[0, 0] = 0
A = np.vstack([Phi, np.sqrt(N*lam)*D])
b = np.r_[y, np.zeros(Phi.shape[1])]
w = np.linalg.lstsq(A, b, rcond=None)[0]
```

操作与观察

堆叠额外行表达参数惩罚

增广后直接使用lstsq

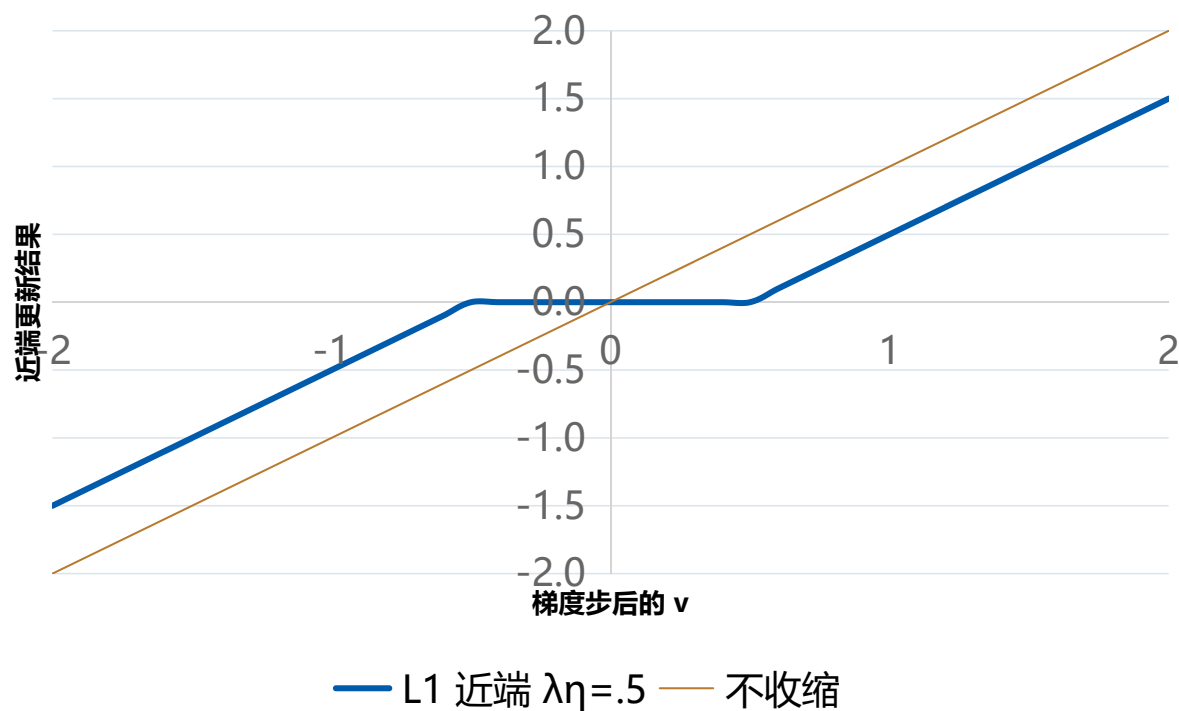
一般惩罚矩阵需满足 $R^TR=D$

$$A=[\Phi; \sqrt{(N\lambda)D}], \quad b=[y; 0]$$

推导 4/4 为什么代码堆叠的是 $\sqrt{(N\lambda)D}$ 而不是 $N\lambda D$?

L1近端：把小系数压到精确的零

稀疏是另一种偏好，不能直接当成因果解释



$$J_\lambda = J + \lambda \sum_j |w_j|$$

先做数据步： $v = w - \eta \nabla J(w)$

$$w_j^+ = \text{sign}(v_j) \max(|v_j| - \eta \lambda, 0)$$

截距不阈值： $w_0^+ = v_0$

L1在零点不可微

近端更新把部分小系数置零

相关特征之间的选择不稳定

推导 1/1 小系数.04经过阈值.1后为何不是-.06?

交叉验证：每一折都重新拟合整条管道

验证模型选择过程，同时保护最终测试集

- 01 先全数据标准化再分折会泄漏
- 02 时间或主体相关数据需相应划分
- 03 候选越多，越应重视选择不确定性

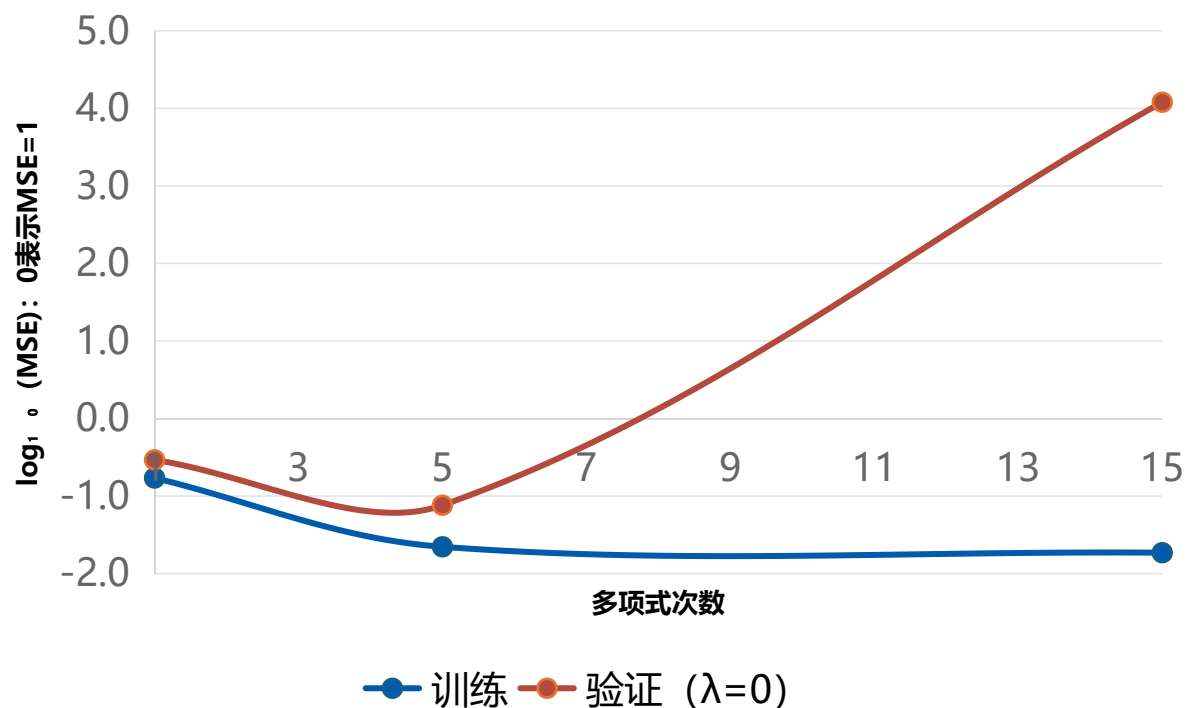
每折：fit预处理(train_fold)

再fit模型(train_fold) $\rightarrow$ score(val_fold)

对候选配置汇总各折验证指标

泛化演示：先选择，再冻结，最后测试

让学生说明为什么选择某个候选，而非只报最优分数



候选：次数1、5、15；六档 λ ，共18组

验证选择：5次， $\lambda=0$

冻结后测试 $\text{MSE} \approx .07667$

逐步展示训练、验证、冻结、测试四阶段

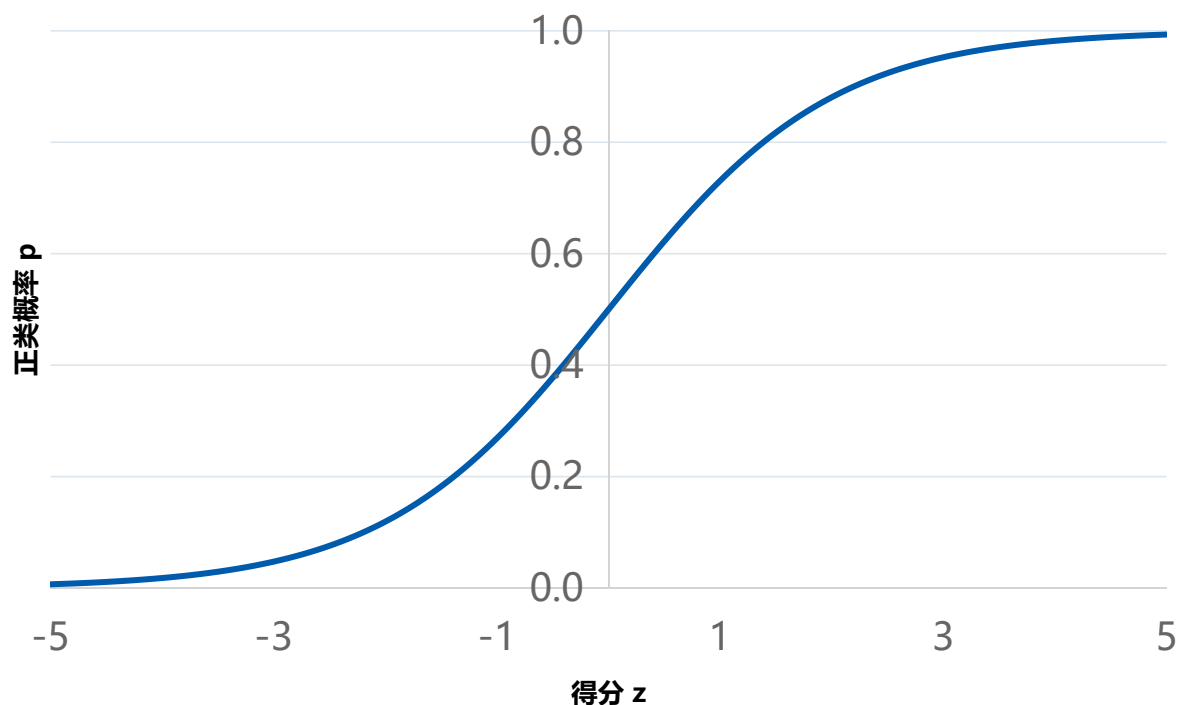
比较训练误差与验证误差的不同角色

正则化不保证每次都要选非零

推导 2/2 看完测试误差后发现另一个模型更好，能否立即回头更换并仍称为独立测试？

得分、概率与类别：分类系统有三层输出

保留线性得分，更换输出语义和决策规则



$$z = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) \in \mathbb{R}$$

$$p = P(y=1 | \mathbf{x}; \mathbf{w}) \in (0, 1)$$

$$\hat{y} = 1[p \geq \tau]$$

得分不是概率，类别不是概率

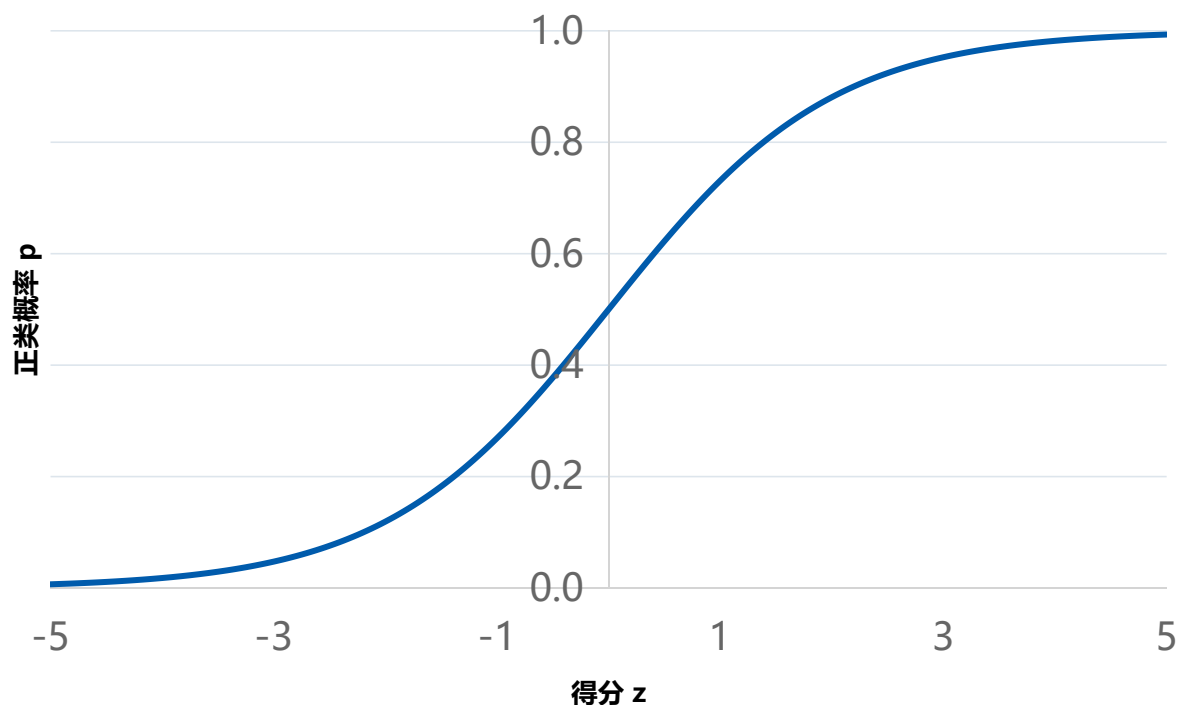
阈值决定动作，不改变已训练权重

先说明 $y=1$ 究竟代表什么事件

推导 1/9 把阈值从.5改成.8，模型输出概率会跟着改变吗？

Sigmoid与logit: 概率和对数几率相互转换

默认边界来自 $z=0$, 概率尺度还取决于参数大小



$$p = \sigma(z) = 1 / (1 + \exp(-z))$$

$$\text{logit}(p) = \log[p / (1 - p)] = z$$

$$\sigma'(z) = p(1 - p)$$

$$p \geq \tau \Leftrightarrow z \geq \text{logit}(\tau)$$

$\sigma(0) = .5$, 变换保持得分排序

整体正缩放保持.5边界, 但改变概率

非线性特征可以产生弯曲输入边界

推导 2/9 $z=.7$ 时概率是.7吗? 扩大五倍后默认类别改变吗?

Bernoulli概率：用一个式子覆盖两种标签

把真实标签出现的概率写成训练目标的起点

01 y选择保留哪一个概率因子

$$y \in \{0, 1\}, \quad p = P(y=1|x)$$

02 单个标签是观测，p是模型预测

$$P(y|x, w) = p^y (1-p)^{1-y}$$

$$y=1 \Rightarrow p; \quad y=0 \Rightarrow 1-p$$

03 最大似然偏好真实标签获得较大概率

推导 3/9 若 $p=.8$ 而真实标签为0，该样本的似然贡献是多少？

BCE推导：对真实标签概率取负对数

从单样本到全数据，保留平均与正负号

01 采用条件独立样本的似然模型

$$L(w) = \prod_i p_i^{y_i} (1 - p_i)^{1 - y_i}$$

02 自信且错误会受到很大惩罚

$$\ell_i = -y_i \log p_i - (1 - y_i) \log(1 - p_i)$$

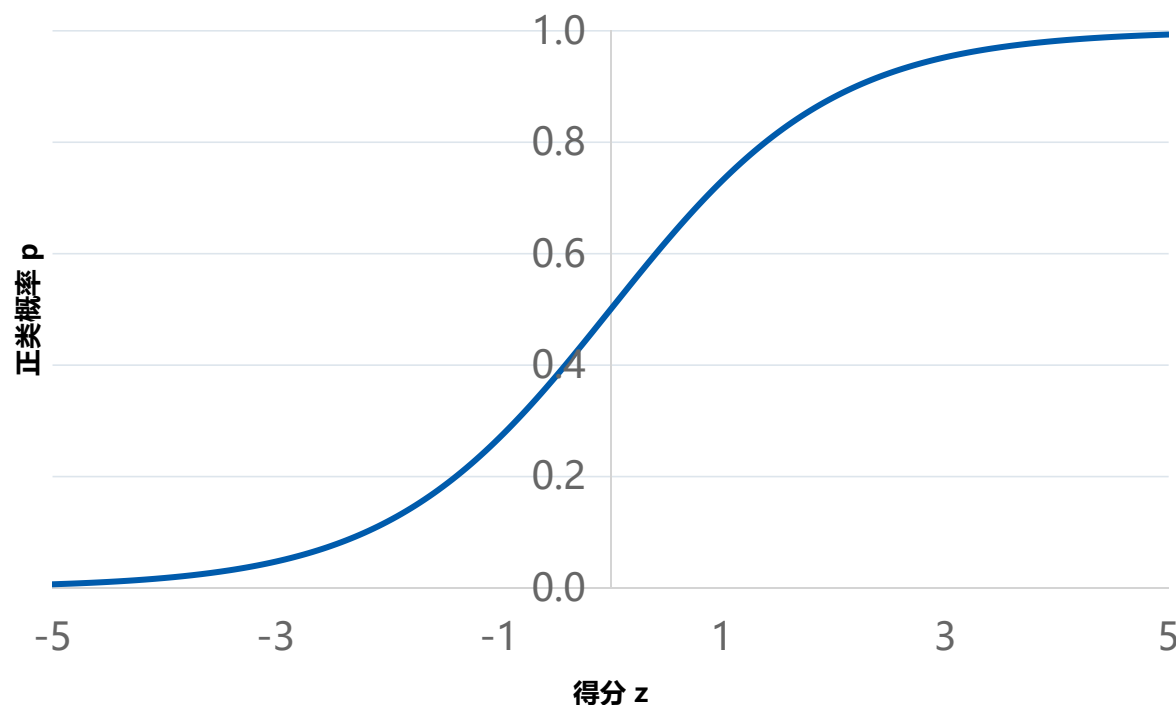
$$J(w) = \sum_i \ell_i / N$$

03 交叉熵与准确率回答不同问题

$$y=1: \quad p=.5 \rightarrow \ell=.693; \quad p=.9 \rightarrow .105$$

链式求导：为什么最后只剩 $p-y$

两个分母与Sigmoid导数相消，特征仍在外层



$$\partial \ell / \partial p = (p - y) / [p(1 - p)]$$

$$\partial p / \partial z = p(1 - p)$$

$$\partial \ell / \partial z = p - y$$

$$\nabla \mathbf{w} \ell = (p - y) \boldsymbol{\phi}; \quad \nabla J = \Phi^T (p - y) / N$$

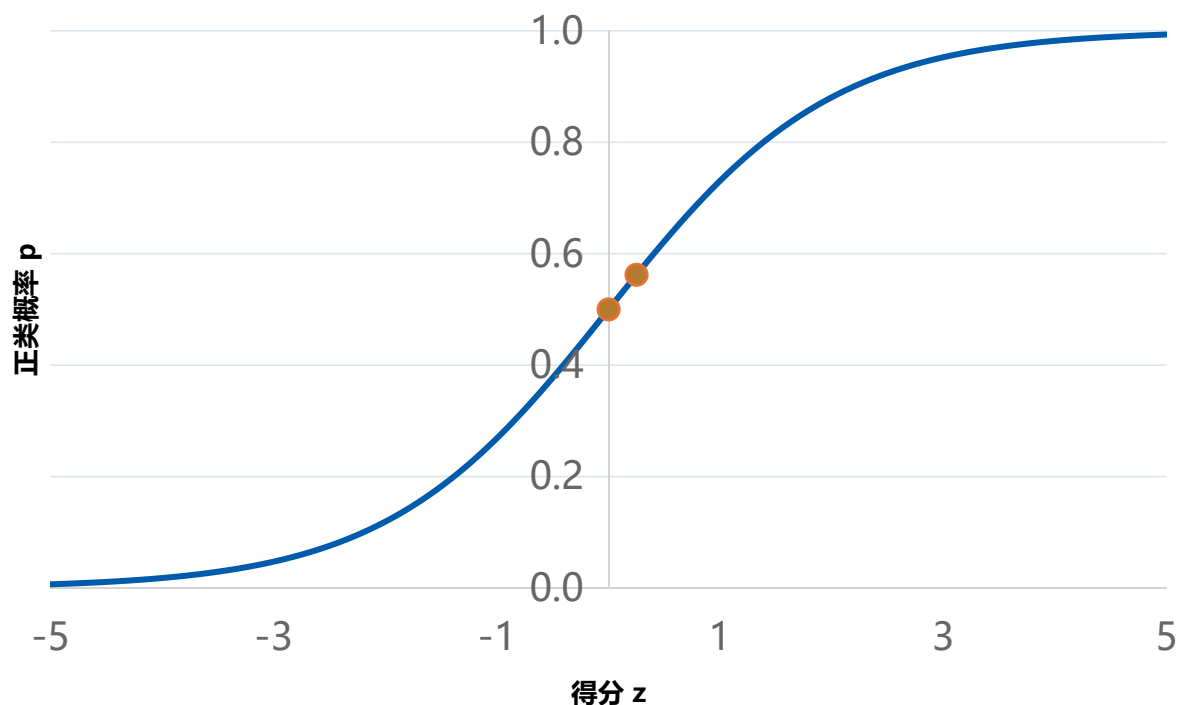
先对概率求导，再通过得分回传

$p - y$ 是对logit的导数，不是对概率的
梯度维度仍与参数一致

推导 5/9 当 $y=1$ 且 $p=.2$ 时， $\partial \ell / \partial z$ 是多少，更新希望怎样改变 z ？

单样本逻辑手算：从z到参数逐步核验

使用同一输入、初值与学习率完成七步



$$\varphi=(1,2)^T, \quad y=1, \quad w_0=(0,0)^T$$

$$z_0=0, \quad p_0=.5, \quad \ell_0 \approx .693147$$

$$g_0=(p-y)\varphi=(-.5,-1)^T$$

$$\eta=.1 \Rightarrow w_1=(.05,.10)^T, \quad z_1=.25$$

先算概率，再算误差信号

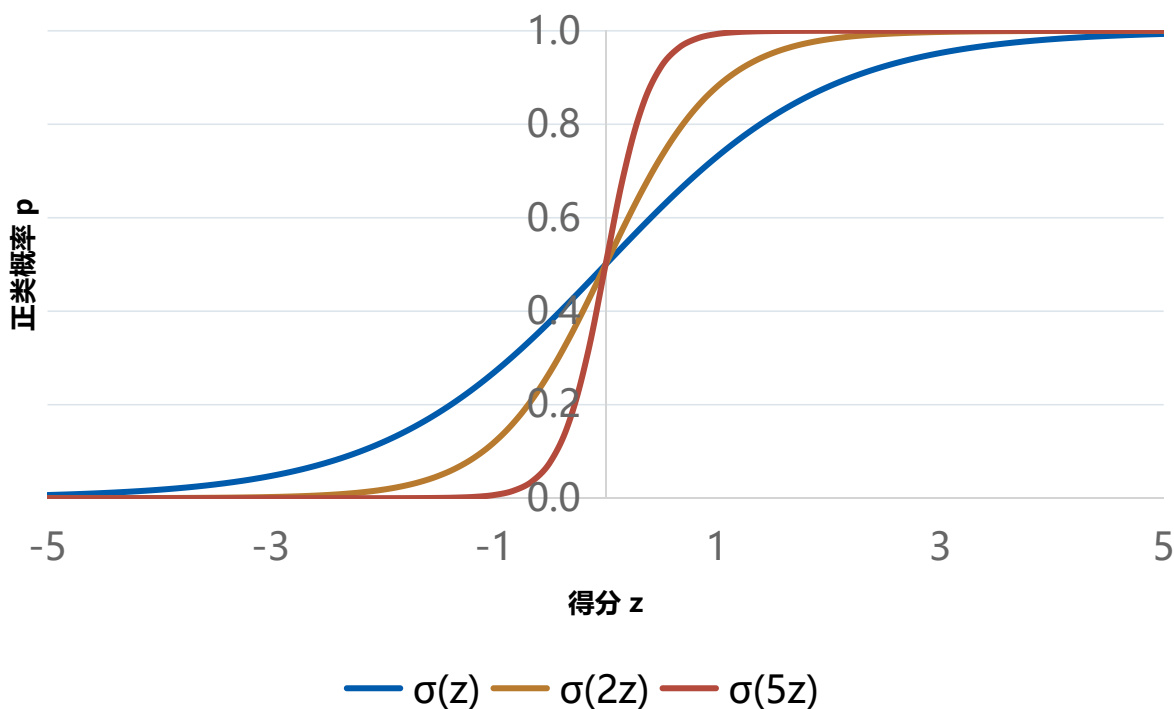
截距与特征系数共同更新

更新后 $p \approx .562$, $BCE \approx .576$

推导 6/9 若y改为0而其余初值相同，梯度与更新怎样改变？

可分数据：凸损失也可能没有有限最优参数

准确率已不变，参数范数仍可能继续增长



$\tilde{y}_i = 2y_i - 1$; 存在 w 使 $\tilde{y}_i \phi_i^T w > 0$

放大 cw , $c \rightarrow \infty$ 时真实类概率 $\rightarrow 1$

未正则 $BCE(cw) \rightarrow 0$, 但有限 w 达不到 0

$$H = \Phi^T \text{diag}[p_i(1-p_i)] \Phi / N \geq 0$$

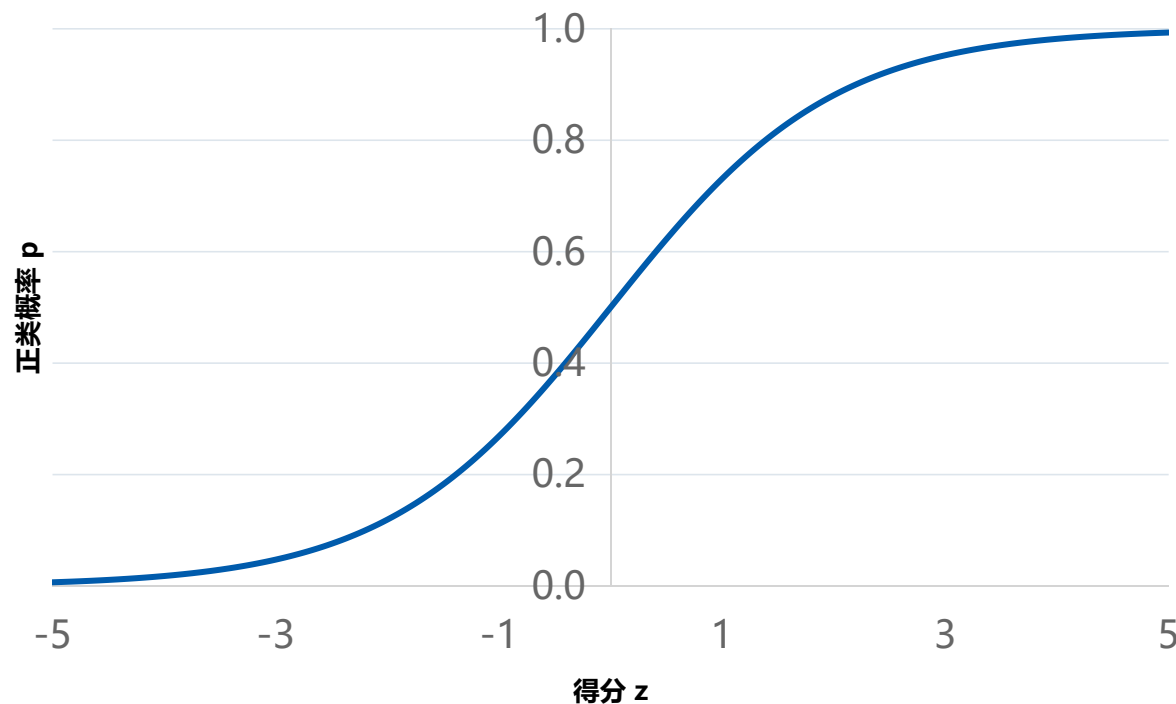
凸性与有限最优解存在性不同

边界不变也可能越来越自信

正则化或验证早停限制参数规模

成本阈值：概率怎样决定行动

先明确误报、漏报与校准条件，再求交点



$$R(\text{判}1|x) = C_{FP}(1-p)$$

$$R(\text{判}0|x) = C_{FN} \cdot p$$

$$\text{判}1 \Leftrightarrow p \geq C_{FP} / (C_{FP} + C_{FN})$$

$$C_{FP}=1, C_{FN}=10 \Rightarrow \tau^*=1/11$$

正确决策成本设为零

公式要求可用的条件后验概率

概率失准、基率改变或容量约束需再验证

混淆矩阵：指标必须说明正类与阈值

不同分母回答不同问题，平均正确率会掩盖稀有事件

真实/预测	预测1	预测0
真实1	TP=18	FN=2
真实0	FP=6	TN=74

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / N$$

例：TP=18, FP=6, FN=2, TN=74

精确率关注预测正类的纯度

召回率关注真实正类找到多少

分母为零时需说明处理规则

互斥多类与多标签：先确定标签结构

同样K个输出，可以表达完全不同的任务

方式	输出含义	典型结构
Softmax	一个K类分布	互斥单标签
多个Sigmoid	每标签出现概率	可共存多标签
One-vs-Rest	第k类对其余类	独立二元比较

互斥：一个样本恰属一个类别

多标签：多个标签可以同时成立

OvR：每类各拟合一个二元模型

Softmax让类别竞争同一单位概率

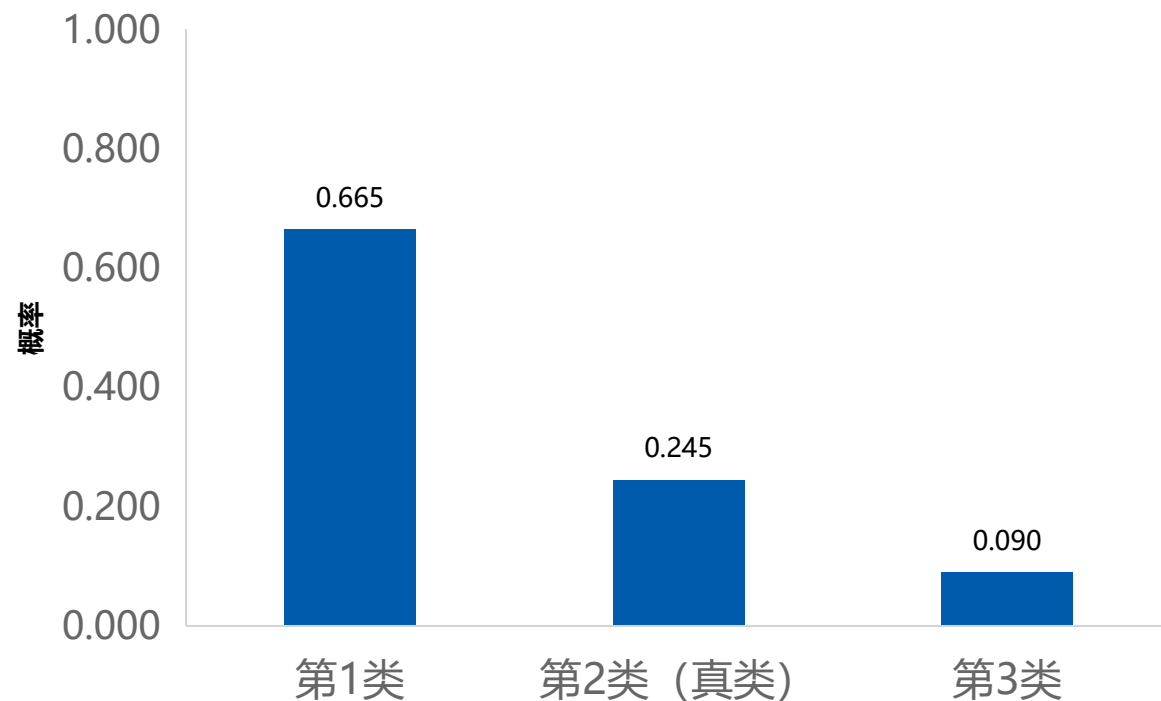
多个Sigmoid允许多个标签同时为正

OvR输出不天然组成和为1的分布

推导 1/5 同一文章可同时属于医学和机器学习，应直接用互斥Softmax吗？

Softmax手算：移位、指数、求和、归一化

保持四步中间量，检查第二类概率而非只看赢家



$$z=(2,1,0), \quad s=z-\max(z)=(0,-1,-2)$$

$$\exp(s) \approx (1, .367879, .135335)$$

$$S=\Sigma \exp(s) \approx 1.503214$$

$$p \approx (.665241, .244728, .090031)$$

共同平移得分不改变概率

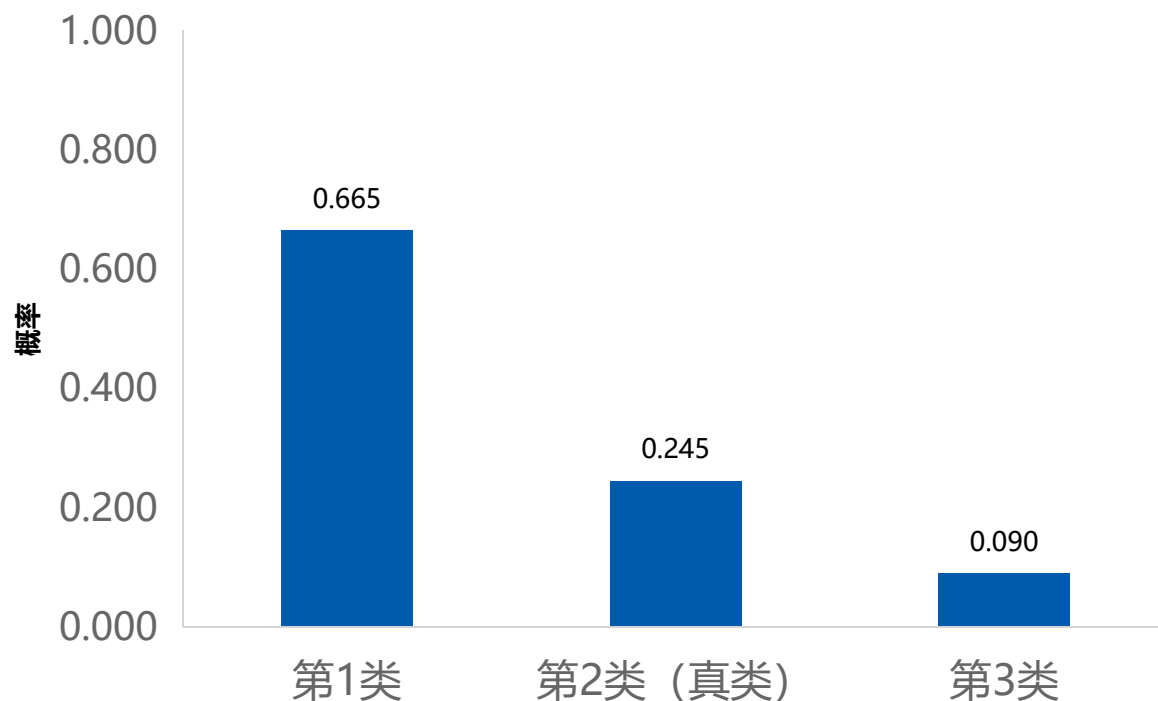
减最大值使指数输入不为正

真实第二类对应代码标签 $y=1$

推导 2/5 共同加1000后为什么可以得到相同概率，而直接计算 $\exp(1002)$ 可能失败？

多类交叉熵与梯度：真实类取对数，所有类共同更新

用one-hot矩阵统一样本轴与类别轴



$$\ell = -\sum_k y_k \log p_k = -\log p_{\text{真实类}}$$

本例真实第2类： $\ell \approx 1.407606$

$$\mathbf{Z} = \Phi \mathbf{W}, \mathbf{P} = \text{softmax}(\mathbf{Z}), \partial \ell / \partial \mathbf{z} = \mathbf{p} - \mathbf{y}$$

$$\nabla J(\mathbf{W}) = \Phi^T (\mathbf{P} - \mathbf{Y}) / N$$

$\mathbf{W}$ 为 $d \times K$, $\mathbf{P}$ 与 $\mathbf{Y}$ 为 $N \times K$

每行概率与one-hot标签都和为1

共同logit平移导致参数冗余

推导 3/5 真实第二类时，为什么第一类梯度为正而第二类为负？

稳定分类代码：直接从logit计算损失

数值有限不等于接口语义正确，先核对轴与标签

```
S = Z - Z.max(axis=1, keepdims=True)
logp = S - np.log(np.exp(S).sum(1, keepdims=True))
loss = -logp[np.arange(N), y].mean()
P = np.exp(logp)
G = P.copy()
G[np.arange(N), y] -= 1
grad = Phi.T @ G / N
# Binary logits: np.mean(np.logaddexp(0,z)-y*z)
```

操作与观察

BCE若要求logit，勿先转概率

减最大值避免指数溢出

概率行和与有限差分可核验实现

$BCE = \text{mean}(\text{logaddexp}(0, z) - y \cdot z)$

推导 4/5 把概率传给一个要求logit的BCE函数，即使没有报错，目标还正确吗？

共同梯度框架：特征转置把误差信号送回参数

统一计算结构，同时保留输出语义与假设差别

模型	输出与目标	误差信号
线性回归	连续 $\hat{y}$ ；平方损失	$\hat{y}-y$
二元逻辑回归	Sigmoid；BCE	$p-y$
Softmax回归	K类概率；CE	$P-Y$

线性： $\nabla J = \Phi^T(\hat{y}-y)/N$

二元： $\nabla J = \Phi^T(p-y)/N$

多类： $\nabla J = \Phi^T(P-Y)/N$

加入L2： 再加 λD_w 或 λDW

误差信号来自各自模型与损失的推导
固定特征时可分析凸性与可识别性
更复杂表示仍需数据协议与独立评价

线性估计在现代系统中的三种位置

基础模型保留，输入表示、数据到达方式和科学约束继续扩展

01 新表示让线性读出处理复杂输入

固定表示: $\hat{y} = h(x)^T w$

02 在线估计把参数变成可更新状态

在线记忆: w_t 随新数据更新

03 候选函数库把方程发现写成回归

科学发现: $dX/dt \approx \Theta(X)\epsilon$

这三条路线中，哪一个对象发生了变化？



案例一：系统漂移与在线回归

场景：传感器标定关系在 $t=400$ 发生变化，标签在预测后到达

01 先用旧参数预测，再揭示当前 y

$$\varphi_t = (1, u_t)^T, \quad \hat{y}_t = \varphi_t^T w_{t-1}$$

前段 $y \approx .4 + 1.5u$; 后段 $y \approx -.6 - u$

02 $\rho=1$ 保留历史同等权重

$$J_t = \frac{1}{2} \sum_{s \leq t} \rho^{t-s} (\varphi_s^T w - y_s)^2$$

03 $\rho=.98$ 逐渐降低旧样本的影响

程序另含初始先验项 $\delta \rho^{t+1} \|w\|^2 / 2$, $\delta=.001$ (含截距); t 从0计数。

递归最小二乘的一次更新

保留逆加权Gram矩阵P，避免每个样本都重新解全部历史数据

01 r 是当前预测残差

$$q_t = P_{t-1} \varphi_t, \quad k_t = q_t / (\rho + \varphi_t^T q_t)$$

02 k 按已有特征信息调节更新幅度

$$r_t = y_t - \varphi_t^T w_{t-1}$$

$$w_t = w_{t-1} + k_t r_t$$

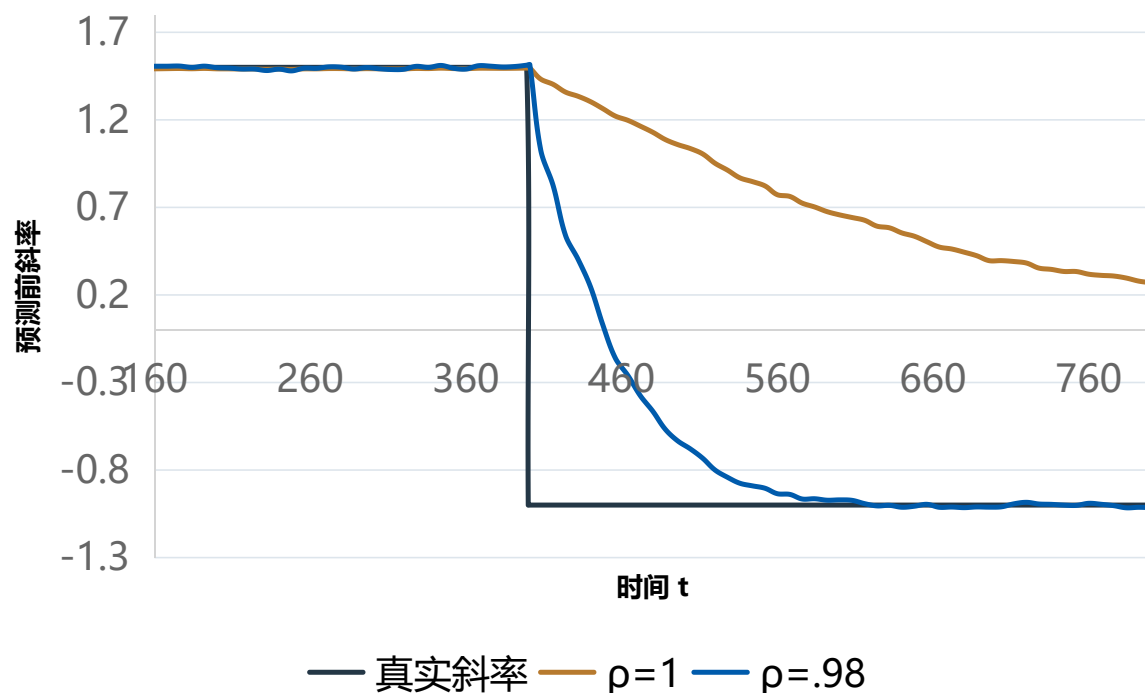
03 $P_{-1} = I / .001, \quad w_{-1} = 0$ (t 从0计)

$$P_t = (P_{t-1} - k_t \varphi_t^T P_{t-1}) / \rho$$

φ 和 w 是 d 维列向量， k 和 P 的维度分别是多少？

在线适应：观察斜率怎样追上系统变化

图取自800步真实运行日志；比较预测前的参数状态



漂移后评价区间： $500 \leq t < 800$

冻结批量： $MSE=7.23134$

$\rho=1$: 3.10332; $\rho=.98$: .03587

先观察 $t=400$ 附近的过渡过程

比较最终误差与适应速度

ρ 较小也会提高估计波动

适应后的低MSE能否说明变化发生的瞬间也预测得好？

在线回归代码：预测、揭示、更新

对应现代案例/modern_cases.py中的rls_trace

```
before = w.copy()
pred = float(xt @ before)
px = P @ xt
gain = px / (rho + xt @ px)
error = yt - pred
w = w + gain * error
P = (P - np.outer(gain, xt) @ P) / rho
P = (P + P.T) / 2
```

操作与观察

运行“课堂单步演示.py”

输入400，先判断再按h揭示

n/p前后单步，q退出

命令：python 课堂单步演示.py --mode rls --start 399

如果把pred移到w更新之后，哪条验收原则被破坏？

前沿联系：TTT把学习器参数作为记忆状态

Learning to (Learn at Test Time), ICML 2025, 首发2024

01 共同点：学习误差推动状态更新

隐藏状态可以是一组模型参数 W_t

02 TTT可用线性或MLP学习器

$$W_t = W_{t-1} - \eta \nabla \ell(W_{t-1}; x_t)$$

当前输入提供自监督学习信号

03 本课RLS有监督，不是完整TTT实现

本课RLS与论文TTT之间最关键的差异之一是什么？

案例二：冻结表示与线性读出

DINOv3 (2025) 用冻结表示上的线性评估检验任务信息

01 表示负责变换复杂输入

$h = \text{frozen_encoder}(x)$

02 读出负责适配当前监督任务

连续目标: $\hat{y} = h^T w + b$

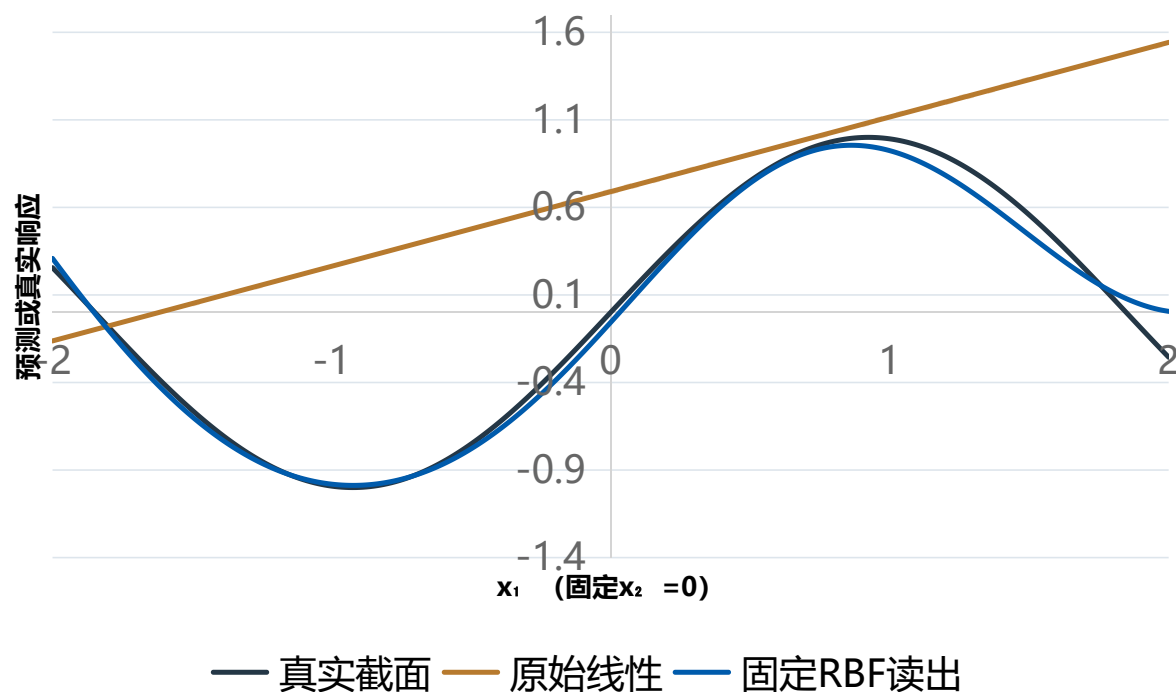
类别目标: $p = \text{softmax}(W^T h + b)$

03 整体可非线性，末层仍可线性

只训练线性读出，是否意味着最终预测必须是输入的直线函数？

固定特征改变了可表达的函数

同一训练/验证/测试划分，只更换表示形式



$$h_j(x) = \exp(-\|x - c_j\|^2 / (2\ell^2))$$

64个固定中心, $\ell = .8$

测试RMSE: 原始线性 .9951

固定RBF读出: .1481

图为 $x_2 = 0$ 截面, 非测试散点

中心与宽度事先固定

标准化仅在训练集拟合

图中的曲线拟合得好，为什么仍要看独立二维测试样本？

线性读出仍由同一正规方程控制

冻结特征后，把h替换进设计矩阵即可

01 $D = \text{diag}(0, 1, \dots, 1)$ ，不惩罚截距

02 参数与梯度维度保持一致

03 λ 控制读出层的复杂度

$H = [1, h_1, \dots, h_{64}]$ ，形状 $N \times 65$

$J = \|Hw - y\|^2 / (2N) + \lambda w^T D w / 2$

$(H^T H + N\lambda D)w = H^T y$

实现：增广最小二乘，避免显式求逆

如果编码器也参与更新，关于凸优化的哪条前提发生变化？



固定表示实验的复现与互动

完整入口run_all.py; 输出保存候选配置与最终测试结果

```
dist2 = ((X[:, None] - centers[None])**2).sum(2)
H = np.exp(-dist2 / (2 * lengthscale**2))
mu, sd = H[tr].mean(0), H[tr].std(0)
sd = np.where(sd < 1e-12, 1., sd)
Phi = np.c_[np.ones(len(H)), (H-mu)/sd]
w = ridge(Phi[tr], y[tr], lam)
val_mse = np.mean((Phi[va] @ w - y[va])**2)
```

操作与观察

先运行默认64中心实验

比较训练误差与验证误差

改中心数后重新执行完整协议

冻结中心检查: True; 训练/验证/测试互斥: True

能否用全部720个样本计算特征标准差?

案例三：用回归发现动力学方程

场景：通过多条阻尼振子轨迹估计恢复力与阻尼系数

01 每列候选函数代表一种可能机制

02 每个状态导数对应一个回归任务

03 稀疏结构帮助保留少量解释项

状态 $X=(x,v)$, $\dot{x}=v$, $\dot{v}=-kx-cv$

候选库 $\Theta=[1,x,v,x^2,xv,v^2,...]$

回归目标: $X \approx \Theta(X)E$

为什么包含 x^2 和 xv 后, 仍可以对 E 做线性回归?

稀疏识别：拟合、阈值筛选、重新拟合

STLSQ用重复最小二乘寻找简洁的候选方程

01 先按训练库RMS缩放特征

02 阈值按验证轨迹重建误差选择

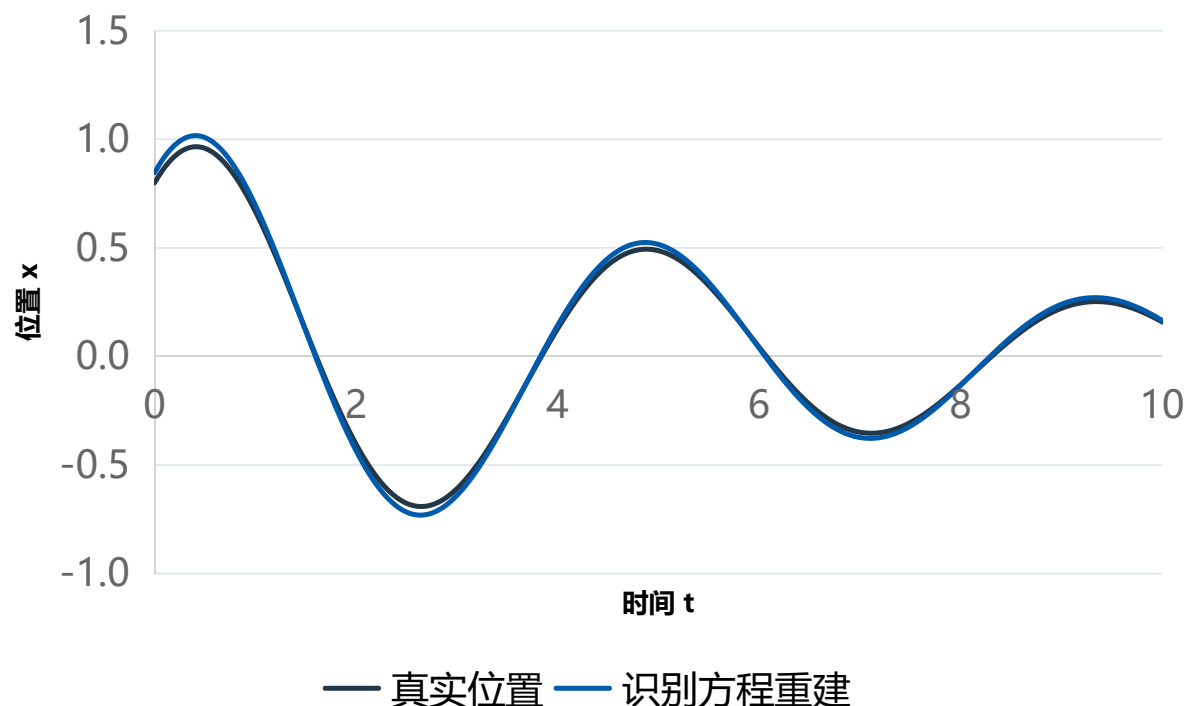
03 STLSQ与L1近端是不同算法

- ① 用全部候选列拟合 Ξ
- ② 把小系数标为0
- ③ 只在保留列重新拟合
- ④ 重复至结构稳定或达到迭代上限

为什么不能随机打散同一轨迹的所有时间点再切训练测试？

恢复方程以后，还要检验新轨迹

$\sigma=.03$ 的观测噪声；图示独立测试轨迹6的位置分量



真实 $k=2$, $c=.3$

估计 $k=2.00280$, $c=.29639$

独立两轨迹状态RMSE=.02266

训练4条、验证2条、测试2条

起点使用测试轨迹首个有噪观测

干净真值只用于模拟验收

为什么局部导数拟合误差小，不足以保证长期轨迹准确？

科学数据的难点：数值求导放大噪声

2024弱形式稀疏发现研究提供了更进一步的处理方向

01 边界项为0需测试函数端点为0

中心差分： $(x_{t+1}-x_{t-1})/(2\Delta t)$

02 弱形式把导数移到已知测试函数

独立噪声项方差 $\approx \sigma^2/(2\Delta t^2)$

弱形式： $\int \psi X dt = - \int \dot{\psi} X dt$

03 当前程序实现差分STLSQ

Δt 减半时，上式中的噪声方差大约变为多少倍？

科学发现实验：可执行步骤与失败检查

modern_cases.py中library、stlsq、case3_sindy组成完整流程

```
A = Theta / scale
Xi = np.linalg.lstsq(A, dX, rcond=None)[0]
for _ in range(12):
    small = np.abs(Xi) < threshold
    Xi[small] = 0
    for j in range(dX.shape[1]):
        keep = ~small[:, j]
        if keep.any():
            Xi[keep, j] = np.linalg.lstsq(
                A[:, keep], dX[:, j], rcond=None)[0]
Xi = Xi / scale[:, None]
```

操作与观察

生成并按完整轨迹划分

训练库缩放、阈值选模

读系数、导数误差与测试重建

默认实验同时比较 $\sigma=0$ 、.01、.03、.10

输入Theta来自训练轨迹，scale为各列RMS（做零值保护）；选模与积分检查见完整源码。

共同框架：现代系统仍需可解释的中间量

表示、目标、梯度或求解器、评价协议需要逐一明确

问题	基础模型	现代系统中的对应对象
输入如何表示	$\varphi(x)$ 与设计矩阵	冻结编码器、函数库、流式特征
误差怎样学习	平方损失与梯度	自监督更新、读出、稀疏估计
结论如何可信	验证与独立测试	因果流式评价、独立轨迹、任务边界

看到一个复杂模型，首先可以追问哪些可检查对象？

课堂验收：一段可核对的解释与一次复现

提交中间过程，说明结论依赖什么条件

任务	提交物	通过条件
手算一次更新	e 、 g 、 w^+ 、 J	符号、维度、归一化一致
复现一个现代案例	命令、配置、输出数值	固定种子与划分，能重复得到
改变一个条件	预判、结果、解释	不用测试集反复选择配置
说清一个边界	具体失效情形	区分课堂类比与论文系统

“程序运行成功”与“建模结论可信”分别需要什么证据？

已核验来源与延伸阅读

论文依据用于解释联系，课堂实验采用自包含合成数据

主题	论文与标识	本课阅读重点
测试时学习	TTT, ICML 2025 arXiv:2407.04620	§2.1 学习器状态 §2.6 等价性条件
冻结视觉表示	DINOv3, 2025 arXiv:2508.10104	§6.1.2 与 §6.2.1 线性读出评估协议
动力学发现	Weak-form modified SINDy arXiv:2410.17838	§2 弱形式 §4.2.4 Duffing 实例

读论文时怎样判断某种“等价”结论能否直接迁移？