梯度

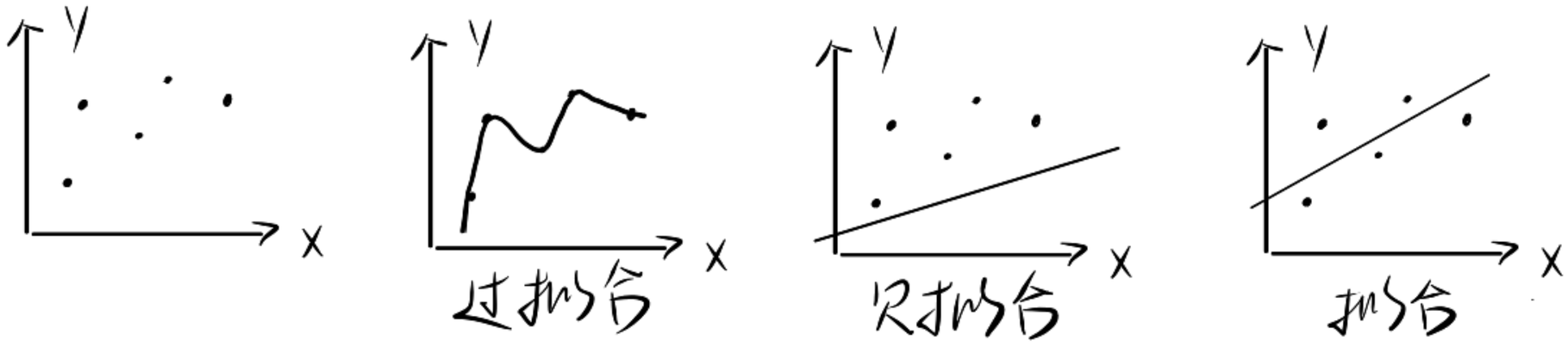
本意是一个向量，表示某一函数在该点处的方向导数沿该方向取得最大值，即函数在该点处沿着该方向（此梯度的方向）变化最快，变化率最大（为该梯度的模）

$$grad\,f(x_0, x_1, \cdots, x_n) = \left(\frac{\partial f}{\partial x_0}, \cdots, \frac{\partial f}{\partial x_i}, \cdots, \frac{\partial f}{\partial x_n}\right)$$

梯度是一个向量（有方向有大小）
梯度的方向是最大方向导数的方向（指向极大值）
梯度的值是最大方向导数的值

回归

我们想探究某两种（或多种）数据的关系，在计算机中，我们实际上是想找出一个这两种数据之间的映射 f



拟合的模型会出现一个问题，不是所有数据对 (x, y) 都完美符合这个模型，实际值 y_i 与预测值 $\hat{y}$ 存在误差，我们用均方误差来判断一个模型的预测是否合理

$$MSE(均方误差) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

在实际操作时，我们会生成无数个可能的模型，这里我们假设都是线性回归（$y = kx + b$）模型，那么一个模型会有自己的 MSE 也就是说 (k_i, b_i) 对应一个 MSE_i，这是一个以 k 和 b 为自变量，MSE 为因变量的二元函数，称为损失函数 $J(\theta)$

θ 是模型参数的统一代称，如对于 $y = \theta_0 x + \theta_1$，$\theta = (\theta_0, \theta_1)$

现在我们的问题是如何求出 $J(\theta)$ 最小值

梯度下降法

在处理损失函数 $J(\theta)$ 时，我们的目标是让它最小化

理论上来说，我们只需要沿当前位置的梯度最大的方向往下走一步，重复这个过程便可逐步逼近最小值

核心公式：对于多元函数 $J(\theta)$

梯度为 $\nabla J(\theta) = \left(\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}, \cdots, \frac{\partial J}{\partial \theta_n}\right)$

梯度下降公式：$\theta_{t+1} = \theta_t - \eta \cdot \nabla J(\theta_t)$

注：θ_t：第 t 次迭代的参数值

η：步长，或称为学习率

$\nabla J(\theta_t)$：$J(\theta)$ 在 θ_t 处的梯度

负号：沿梯度反方向走，即下降

解释：对于参数 θ，不断沿梯度反方向更新参数，逐步逼近 $J(\theta)$ 的最小值

如：对 $y = f(x)$ 来说，梯度下降法可表示为

$$x_{i+1} = x_i - \eta * f'(x_i)$$

$f'(x_i)$ 为 $f(x)$ 在 $x = x_i$ 的梯度

对 $z = f(x, y)$ 来说

$$\begin{cases} x_{i+1} = x_i - \eta * \frac{\partial z}{\partial x}\Big|_{x=x_i, y=y_i} \\ y_{i+1} = y_i - \eta * \frac{\partial z}{\partial y}\Big|_{x=x_i, y=y_i} \end{cases}$$

$\left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right)$ 为 $f(x, y)$ 在 (x_i, y_i) 的梯度

对 $z = f(x_1, \cdots, x_n)$ 来说

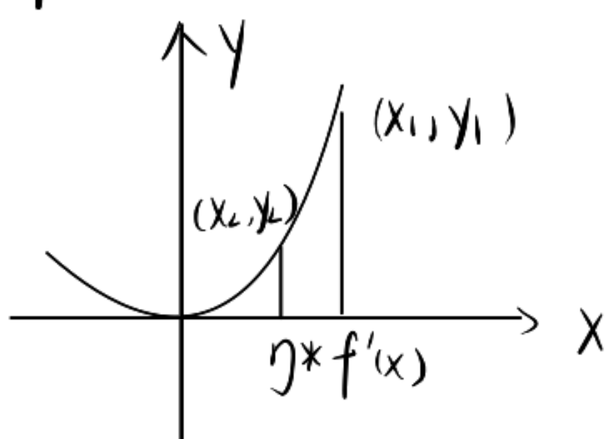
$$\begin{cases} x_{1\,i+1} = x_{1\,i} - \eta * \frac{\partial z}{\partial x_1}\Big|_{x_1=x_{1i}, x_2=x_{2i}, \cdots, x_n=x_{ni}} \\ x_{2\,i+1} = x_{2\,i} - \eta * \frac{\partial z}{\partial x_2}\Big|_{x_1=x_{1i}, x_2=x_{2i}, \cdots, x_n=x_{ni}} \\ \vdots \end{cases}$$

$$x_{n\,i+1} = x_{n\,i} - \eta * \frac{\partial z}{\partial x_n}\Big|_{x_1 = x_{1i},\, x_2 = x_{2i},\, \cdots,\, x_n = x_{ni}}$$

$(\frac{\partial z}{\partial x_1}, \frac{\partial z}{\partial x_2}, \cdots, \frac{\partial z}{\partial x_n})$ 为 f 在 $(x_{1i}, \cdots, x_{ni})$ 的梯度

如：$f(x) = x^2$

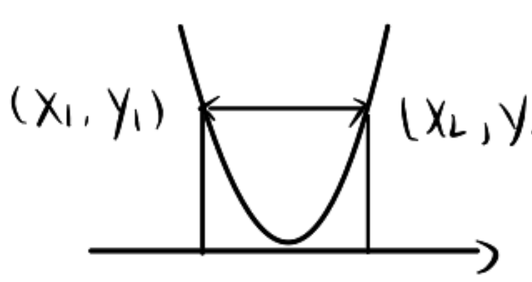
$x_2 = x_1 - 2\eta x_1$



步长(学习率)：影响巨大

① 收敛速度：步长大，收敛快

② 收敛稳定性：步长太大，有可能跨过最低点，导致来回震荡不收敛



如图，步长过大可能在这两点间来回震荡

步长太小，收敛过慢，甚至完全不收敛

③ 最终参数：步长小，容易陷入局部最优

	小步长	大步长
收敛稳定性	稳定	震荡、发散
收敛精度	精细逼近最优解	在最优解附近震荡
收敛速度	慢	快
局部最优	容易陷入	有机会跳出
全局搜索能力	弱，依赖起始点	强，覆盖区域广
计算资源	迭代多，成本高	迭代少，但可能浪费在无效区域