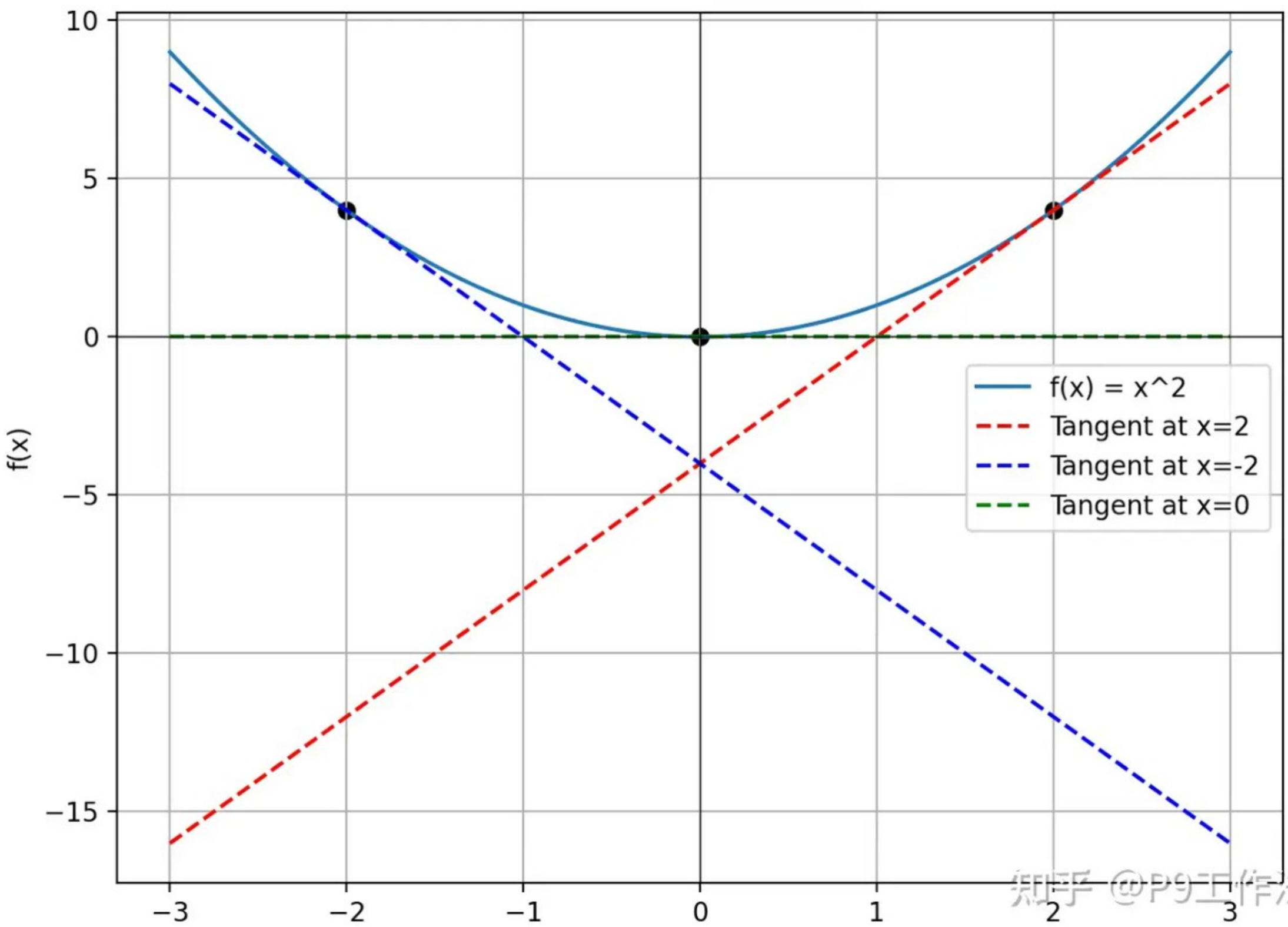
一元函数求极值

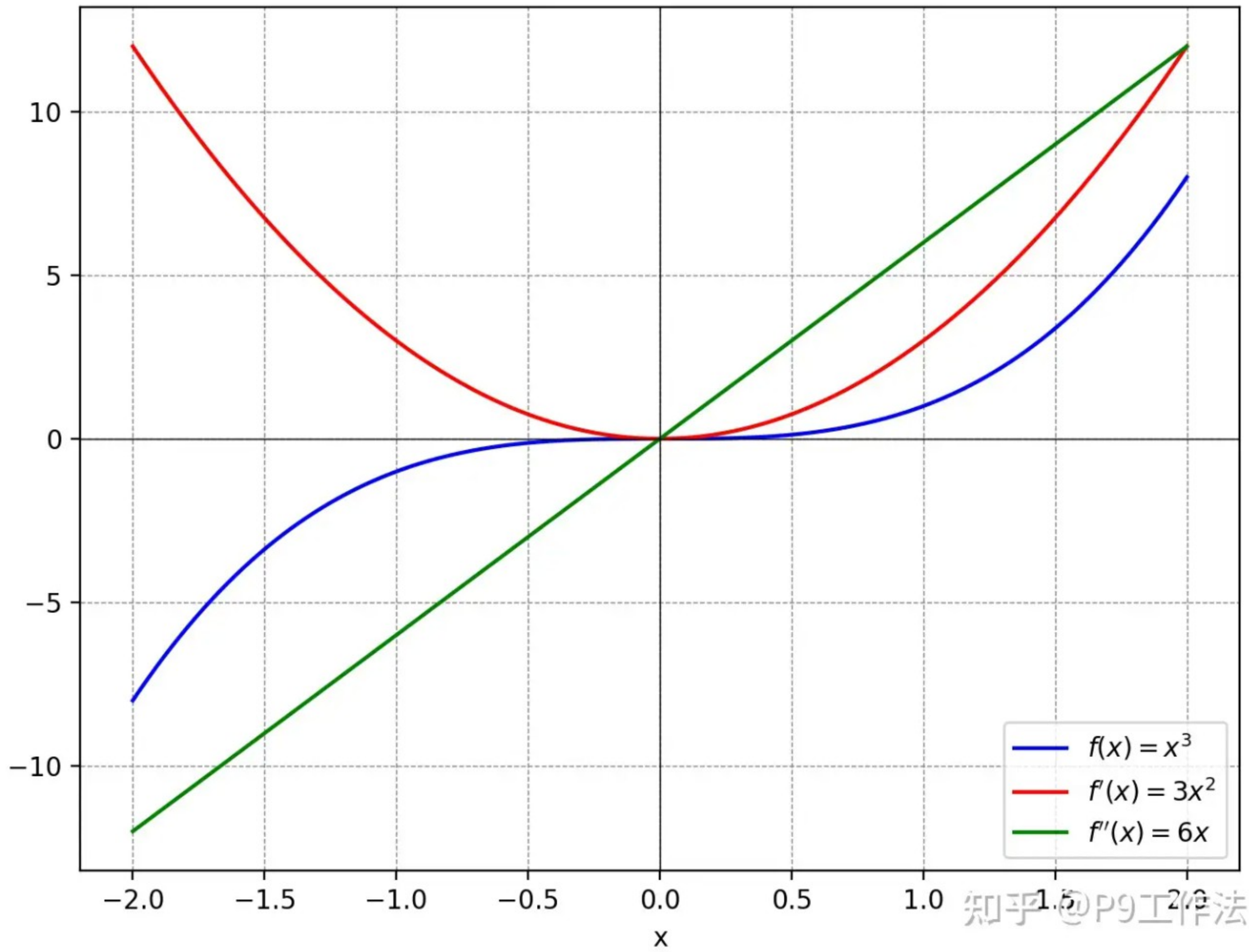
按照一般的数学理解，要求函数 $f(x)$ 的极小值，应该是找导数 $f'(x)$ 为0的地方。因为导数 $f'(x)$ 表示了该点处函数值变化的速率。如 $f(x)=x^2$



如果我们把 $f(x)$ 在x=2,x=-2,x=0的切线画出来。你会发现在x>0时，函数的值是继续变大的，如x=2时；x<0时，函数的值是变小的，如x=-2时；当=0时，函数是取得最小值。这可以被抽象为函数 $f(x)$ 的极值点就是 $f'(x)=0$ 的地方么?

结论是不行的，虽然举例函数 $f(x)=x^2$ 从图形上就能够看出来这个结论，但在数学上并不严谨。

再举例一个函数， $f(x)=x^3$ ，它的导数 $f'(x)=x^2$ ，令 $f'(x)=0$ ，求得x=0。但此时从图上就能够看到出来，x=0并不是函数的极值点。

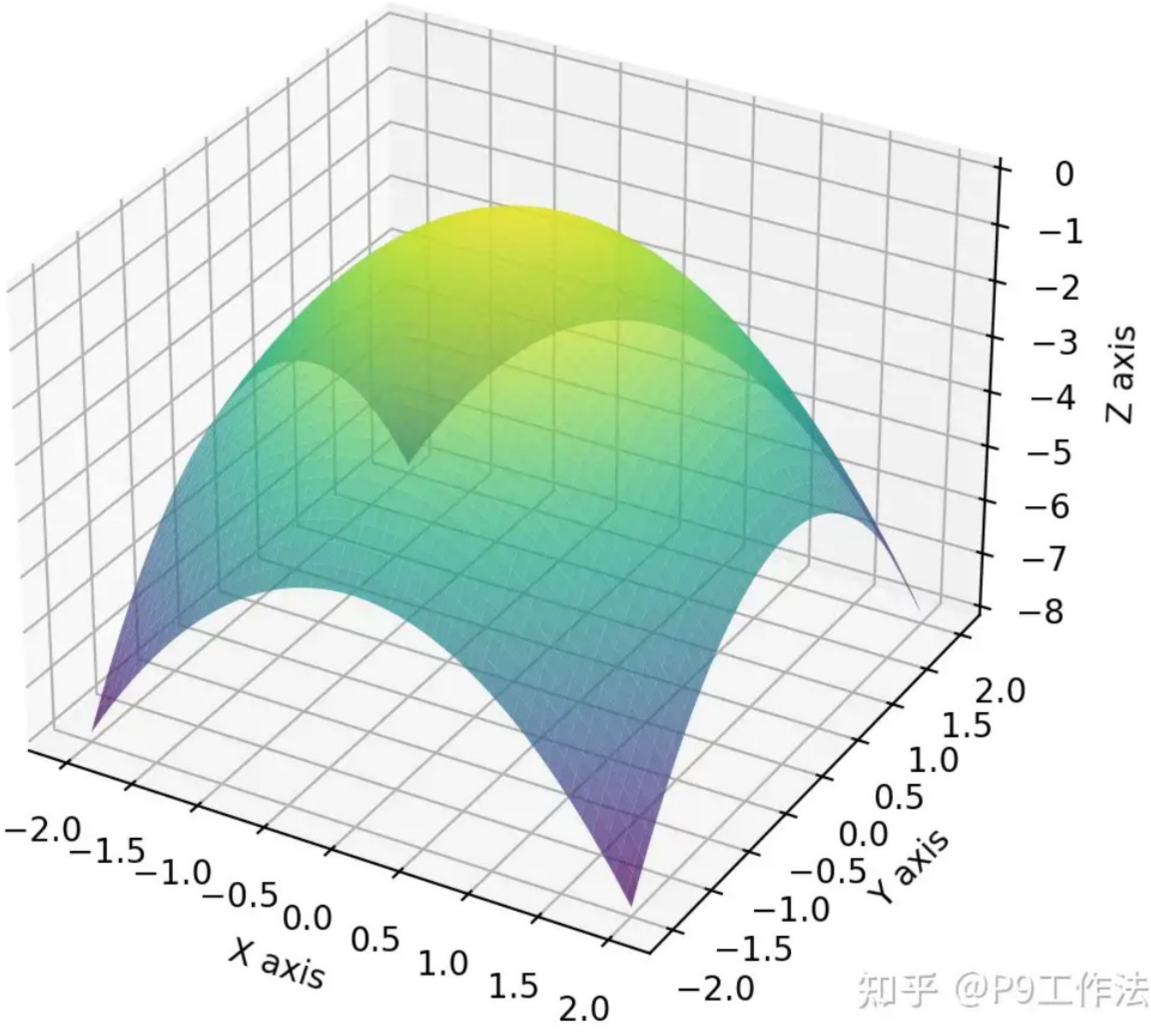


要导数等于0的点是否为极值，还需要判断二阶导数 $f''(x)$ 的取值情况：

- 如果 $f''(x) > 0$，则 $f'(x) = 0$ 的点是一个局部极小值点。
- 如果 $f''(x) < 0$，则 $f'(x) = 0$ 的点是一个局部极大值点。
- 如果 $f''(x) = 0$，则二阶导数测试无法提供明确的信息，该点可能是一个拐点、更高阶的极值点，或者需要进一步分析来确定其性质。

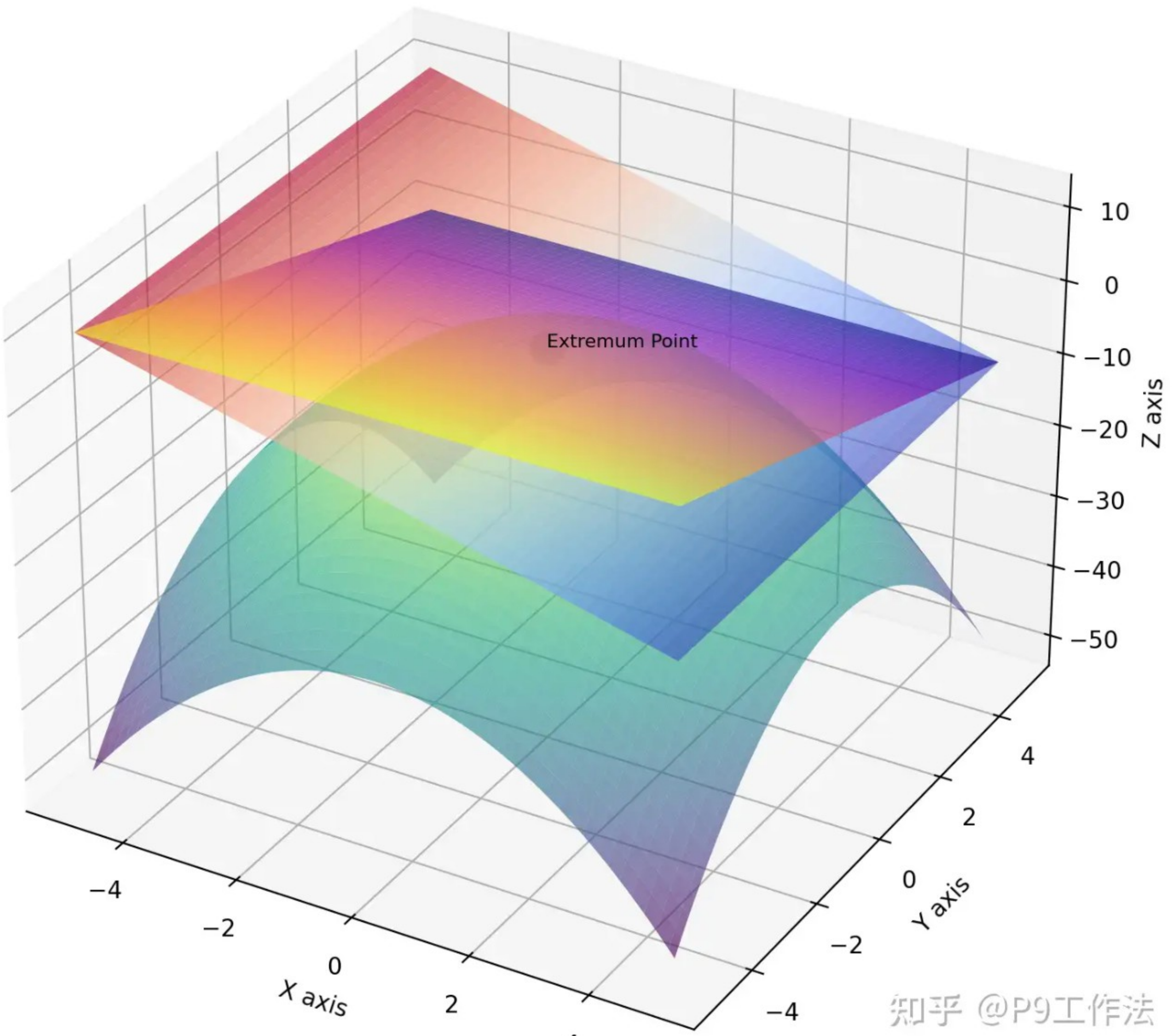
多元函数求极值

假设函数 $f(x, y) = -x^2 - y^2$，它本身的图形如下：



可以看到这个图形应该是有极值点。那么应该如何来求这种多元函数的极值点呢？仍然要用到导数，只不过不再是导数而是偏导数。

- 首先是确定即$f(x,y)$相对于x的偏导数，记为 $\frac{\partial f}{\partial x}$ 。这时要假想y保持不变为常数，那么 $\frac{\partial f}{\partial x}=-2x$ ，那么在三维结构中这就不再是一条直线，而是一个平面。
- 再确定 $f(x,y)$ 相对于y的偏导数，记为 $\frac{\partial f}{\partial y}$ ，同样道理假想x保持不变为常数，那么 $\frac{\partial f}{\partial y}=-2y$,同样这应该是一个平面。
- 极值点应该是在两个偏导数平面与抛物球面相交的地方，那个点应该是 $x=0,y=0$ 处，函数 $f(x,y)=-x^2-y^2=0$



同样道理，二元函数也不能用一阶导数为0来求解是否为极值。仍然面临与一元函数的问题。仍然要用二阶导数来判定。

将 $f(x,y)$ 对 x 的二阶导数表达为 $\frac{\partial^2 f}{\partial x^2}$ ， $\frac{\partial^2 f}{\partial x^2}=\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)=\frac{\partial}{\partial x}(-2x)=-2$ 。

对 y 的二阶导数表达为 $\frac{\partial^2 f}{\partial y^2}$ ， $\frac{\partial^2 f}{\partial y^2}=\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right)=\frac{\partial}{\partial y}(-2y)=-2$

对x求一阶后再求二阶表达为， $\frac{\partial^2 f}{\partial x\partial y}=\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)=\frac{\partial}{\partial x}(-2y)=0$

对y求一阶后再求二阶表达为， $\frac{\partial^2 f}{\partial y\partial x}=\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)=\frac{\partial}{\partial y}(-2x)=0$

再把这些二阶导数写成矩阵，如下

$$\mathcal{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

该矩阵叫海森矩阵✦，这时可以借助矩阵来进行判定

- 如果 $H(f)$ 是正定的（所有特征值都为正），则该临界点是一个局部极小值。
- 如果 $H(f)$ 是负定的（所有特征值都为负），则该临界点是一个局部极大值。
- 如果 $H(f)$ 是不定的（有正有负的特征值），则该临界点是一个鞍点✦。
- 如果 $H(f)$ 的行列式为零，则需要更进一步的分析，因为此时海森矩阵无法提供足够的信息。

在上述案例中

$\mathcal{H}(f) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$，该矩阵所有特征值都是负的（这里两个特征值都是 -2），所以可以断定函数 $f(x, y)$ 有最大值，极值点就为 $x = 0, y = 0$ 处。

为什么直接求解不行

回顾上一篇文章《从结果推导原因-反向传播》，对于损失函数✦ $L = \frac{1}{2}\|\hat{y} - y\|^2$，要求解的就是在什么样的W和b取值下，L的值最小。

按照上面的分析，令偏导数都等于0就能够找到函数的极值点，最多再加上一些分析（海森矩阵判定）就应该可以确定极值点，为什么还会谈到梯度下降算法✦呢?

有以下几点可以直观理解这背后的原因：
1、深度学习的模型中，参数都是上千亿，这么多的参数要去找偏导为0的点，然后联立起来解方程组，这是何其大的工作量，基本上不可行。

2、神经网络的模型训练是一个持续的过程，数据是流式的，需要多次反复用数据来计算，如果每次都是重新求解一次这显然成本太高。

3、在这里 函数是神经网络的本质 提到，神经网络的本质就是函数，一个函数如果弯弯曲曲（想象成空间曲面），那肯定是有多个偏导数为0的地方，也就是按照这样的方法求出来的解可能会非常多个，基本上也是要把所有的可能性探索一遍，这样做也不划算。

为了解决这些个问题，所以提出来了梯度下降法✦来解决这个参数求解问题。