迭代优化

确定性优化和随机优化

确定性优化算法在给定相同的输入和初始条件时总是**产生相同的输出**

随机优化算法由于其**随机**性质每次运行时**产生不同的输出。**

确定性优化算法通常用于解决具有明确目标函数和约束条件的问题，而随机优化算法则用于解决复杂且难以数学建模的问题，或者目标函数是嘈杂或非凸的情况下。

确定性优化方法

一阶方法：梯度下降

目的:最小化一阶泰勒展开式近似式f

$$\min_x f(x) \approx \min_x f(x_t) + \nabla f(x_t)^T (x - x_t)$$

- Update rule:

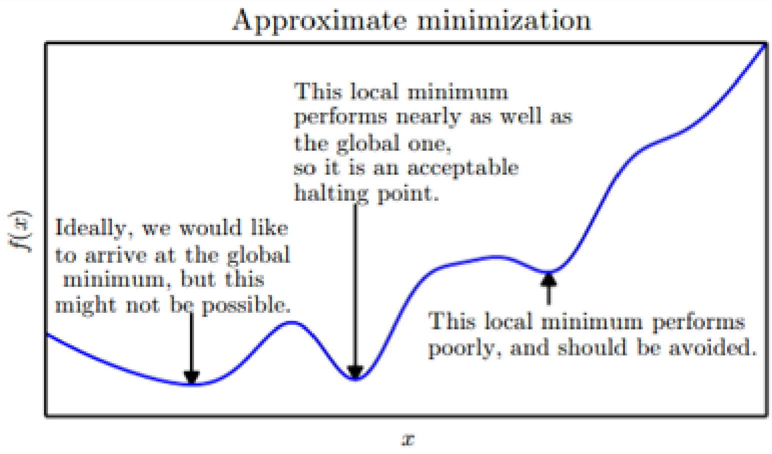
  $$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

  Where $\eta_t > 0$ is the step-size (learning rate).

全局最优解，局部最优解

全局最小值：获得的绝对最低值$f(x)$.

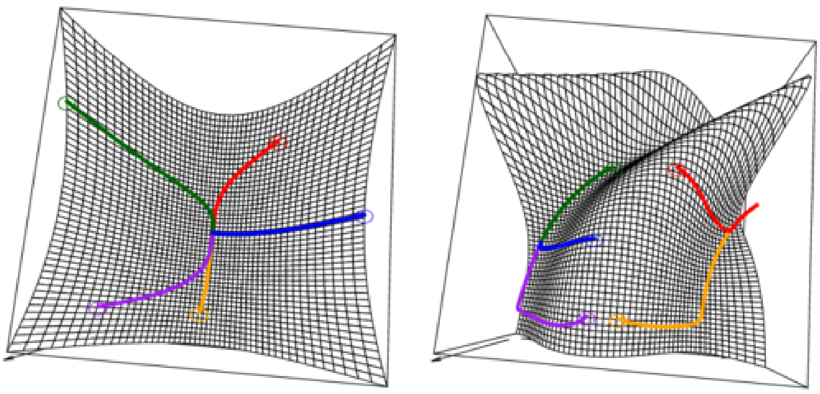
局部最小值：$f(x)$ 高于所有相邻点



梯度下降从不同的位置开始会得到不同的结果

Different Starting Points

- Gradient Descent with different starting points are illustrated in different colors.

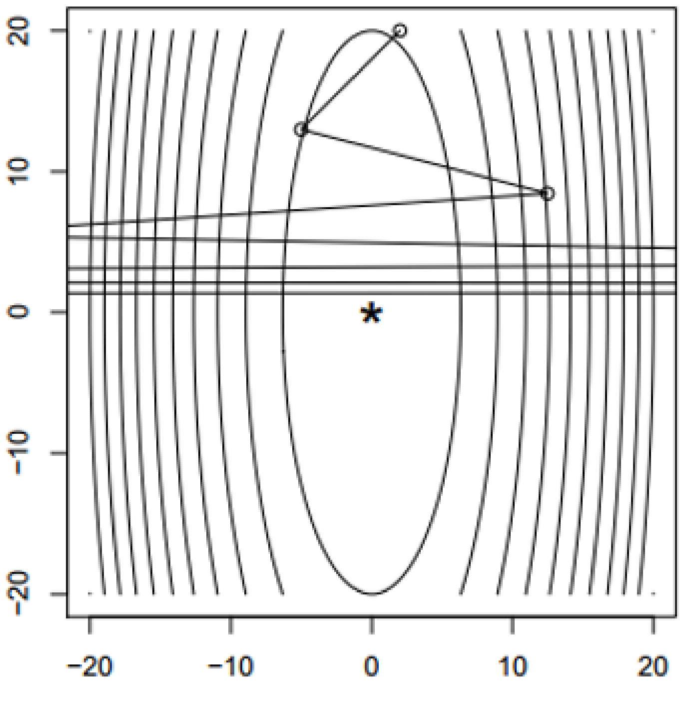


(a) Convex function (b) Non-convex function

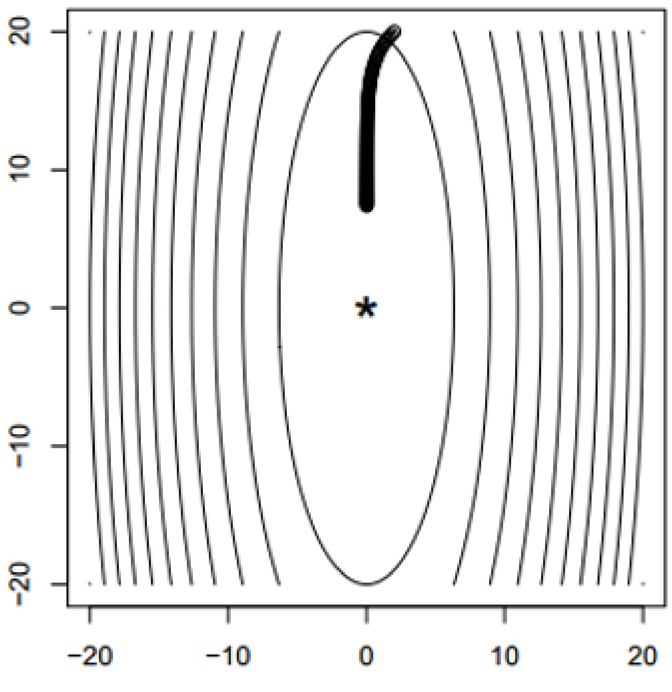
- (a): Strictly convex function: Converge to the global optimum.
- (b): Non-convex function: Different paths may end up at different local optima.

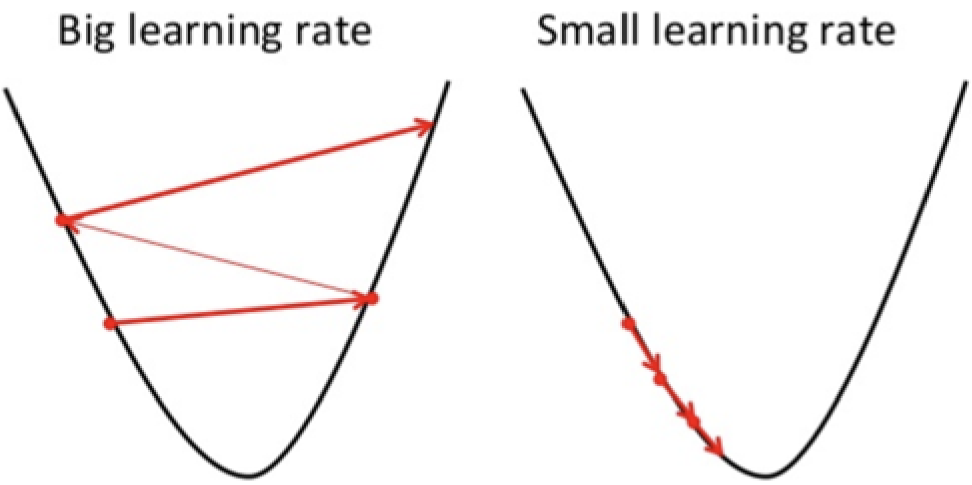
学习率的影响

学习率太大，会导致在最优解旁边震荡，不能收敛



学习率太小，函数收敛需要的时间更长





二阶方法：使用梯度下降和黑森矩阵

黑森矩阵

$$H = \nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

, or

$$H_{ij} = \frac{\partial^2 f}{\partial x_i x_j}$$

Newton's Methods

- Motivation: to minimize the local second-order Taylor approximation of f.

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \approx \min_{\boldsymbol{x}} f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^T(\boldsymbol{x} - \boldsymbol{x}_t) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_t)^T \nabla^2 f(\boldsymbol{x}_t)(\boldsymbol{x} - \boldsymbol{x}_t)$$

- Take the derivative of $\boldsymbol{x}$ on both side, we have,

$$\frac{df(\boldsymbol{x})}{d\boldsymbol{x}} = \nabla f(x_t) + \nabla^2 f(\boldsymbol{x}_t)(\boldsymbol{x} - \boldsymbol{x}_t) = \mathbf{0}$$

- Update rule: suppose $\nabla^2 f(x_t)$ is positive definite,

$$\boldsymbol{x} = \boldsymbol{x}_t - [\nabla^2 f(x_t)]^{-1} \nabla f(x_t)$$

Newton’s Methods

- **Advantage:**
  - More accurate local approximation of the objective, 更准确的物理局部近似
  - The convergence is much faster.收敛速度快得多
- **Disadvantage:**
  - Need to compute the second derivatives需要计算二阶导数
  - Need to compute the inverse of Hessian (time/storage consuming)需要计算Hessian的逆（时间/存储消耗）