梯度下降算法优化

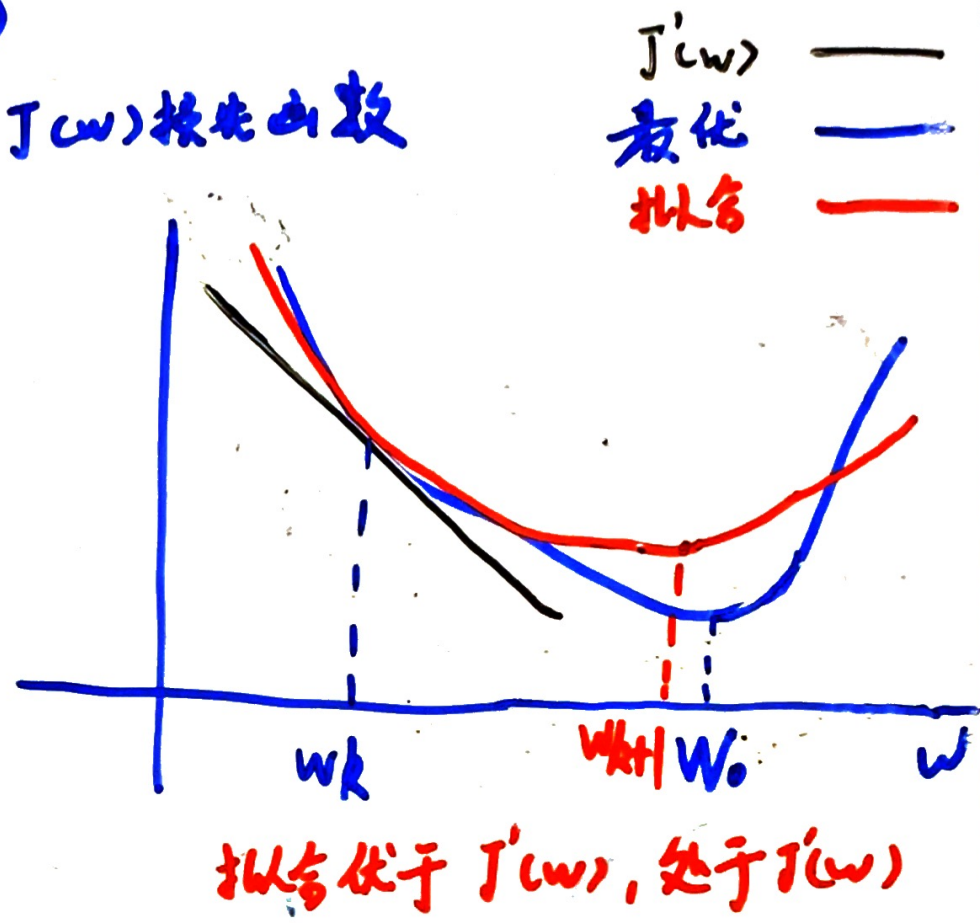
$$W_{k+1} = W_k - \eta J'(W_k)$$

1. 减小计算量

   e.g. 随机梯度下降

2. 优化下降路径

   e.g. 牛顿法 (计算量也大)



拟合优于 $J'(w)$，处于 $J(w)$ 与最优两者之间，手术科

拟合:

$$J(w) \approx J(w_k) + J'(w_k)(w - w_k) + \frac{1}{2} J''(w_k)(w - w_k)^2 + \cdots$$

(泰勒展开)

有, $J'(w_k) \approx \dfrac{J(w) - J(w_k)}{w - w_k}$

$$\Rightarrow J(w) \approx J(w_k) + J'(w_k)(w - w_k)$$

$$\Rightarrow J(w) = J(w_k) + J'(w_k)(w - w_k) + \frac{1}{2} J''(w_k)(w - w_k)^2$$

$$\Downarrow$$

将 $J'(w) = J'(w_k) + J''(w_k)(w - w_k)$ 使 $=0$

推导出 $w_{k+1} = w_k - \dfrac{f'(w_k)}{f''(w_k)}$

补充:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

海森矩阵 (曲率的信息)

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

2. 优化路径

动量法

(历史)

$$W_{(t+1)i} = W_{(t)i} - \eta \frac{\partial J(W_{(t)i})}{\partial W_i}$$

令 $\nabla W_{(t+1)i} = \frac{\partial J(W_{(t)i})}{\partial W_i}$

再令 $V_{(t+1)} = \nabla W_{(t+1)i} + V_{(t)}$

有 $W_{(t+1)i} = W_{(t)i} - \eta V_{(t+1)}$

使历史越久远的数据指数级减弱

影响：指数加权移动平均法

$$V_{(t+1)} = (1-\beta)\nabla W_{(t+1)i} + \beta V_{(t)}$$

0.1 0.9

Nesterov 算法

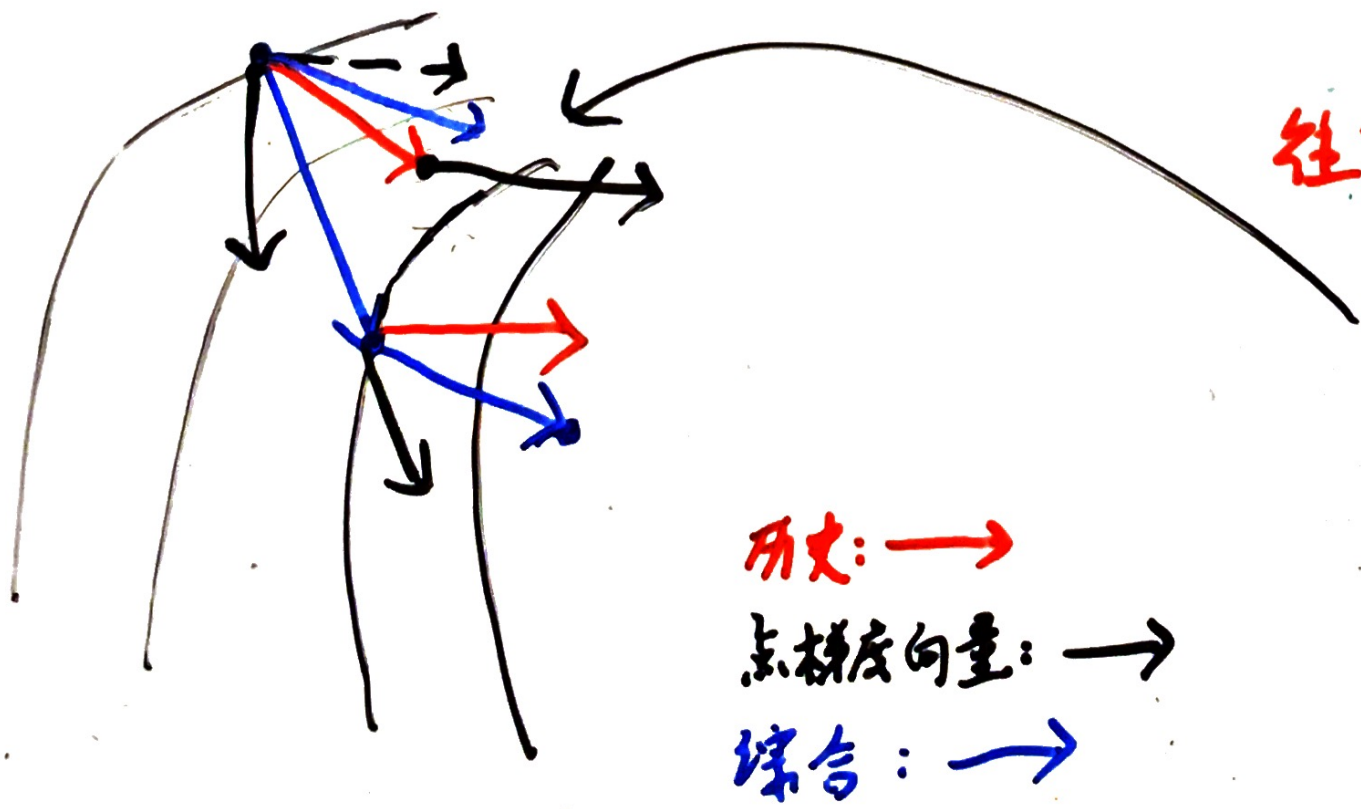
(历史+未来)

梯度等高线

$$\nabla W_{(t+1)i} = \frac{\partial J(W_{(t)i} + \gamma V_{(t)})}{\partial W_i}$$

往前走一步

再求"未来点"的梯度



历史：→

点梯度向量：→

综合：→

2. 优化路径

之前的牛顿法、动量(冲量)法、Nesterov

当优化梯度

现在开始优化学习率

(ci为某一维度)

↳ AdaGrad (自适应)

$$W_{(t+1)i} = W_{(t)i} - \frac{\eta}{\sqrt{S_{(t+1)} + \varepsilon}} \cdot \nabla W_{(t+1)i}$$

↘ 极小量，避免分母为零

$$S_{(t+1)} = S_{(t)} + \nabla W_{(t+1)i} \cdot \nabla W_{(t+1)i}$$

(依赖历史数据)

积累越多，修正越大

缺点：

优势的实现 主要依赖不同特征维度之间的区别

其数据来源为一个样本数据，如有毛没毛

1 0

若更注重某一特征的梯度不一样，即不是，如毛的多少

所以处理 稀疏数据 多零项，可以使用 AdaGrad

锚定问题："平台区"移动很慢，过了"平台区"开始时依然很慢

↳ RMSprop

依然使用 指数加权移动平均法

$$S_{(t+1)} = (1-\beta) \nabla W_{(t+1)i}^2 + \beta S_{(t)}$$

RMSprop + 动量法 ⇒ Adam

RMSprop + Nesterov ⇒ Nadam