机器学习第九周作业

樊泽羲 2200010816

Q1:给出(7.33) 和(7.34)中的函数ψ,ϕ在不同优化算法下的具体形式

A1:

1.学习率调整

1.固定衰减、周期性：相较SGD仅改变了α_t,因此ψ与ϕ与SGD相同

$$\begin{aligned}\psi &= 1-\epsilon \\ \phi &= g_t\end{aligned} \tag{1}$$

2.自适应：

2.1. AdaGrad:

$$\begin{aligned}\psi &= \sum_{\tau=1}^{t} g_\tau \odot g_\tau \\ \phi &= g_t\end{aligned} \tag{2}$$

2.2. RMSProp:

$$\begin{aligned}\psi &= (1-\beta)\sum_{\tau=1}^{t} \beta^{t-\tau} g_\tau \odot g_\tau \\ \phi &= g_t\end{aligned} \tag{3}$$

2.3. AdaDelta: 相较RMSProp仅改变了α_t,因此ψ与ϕ与RMSProp相同

$$\begin{aligned}\psi &= (1-\beta)\sum_{\tau=1}^{t} \beta^{t-\tau} g_\tau \odot g_\tau \\ \phi &= g_t\end{aligned} \tag{4}$$

2.梯度估计修正：

1. 动量法：

$$\begin{aligned}\psi &= 1-\epsilon \\ \phi &= \sum_{\tau=1}^{t} \rho^{t-\tau} g_\tau(\theta_{\tau-1})\end{aligned} \tag{5}$$

2. Nesterov加速动量法：

$$\begin{aligned}\psi &= 1-\epsilon \\ \phi &= \sum_{\tau=1}^{t} \rho^{t-\tau} g_\tau(\theta_{\tau-1}+\rho\Delta\theta_{\tau-1})\end{aligned} \tag{6}$$

3. 梯度截断：

3.1. 按值截断：

$$\begin{aligned}\psi &= 1-\epsilon \\ \phi &= max(min(g_t, b), a)\end{aligned} \tag{7}$$

3.2. 按模截断：

$$\begin{aligned}\psi &= 1-\epsilon \\ \phi &= \frac{b}{||g_t||} g_t\end{aligned} \tag{8}$$

3.综合方法

Adam:

$$\begin{aligned} \psi &= \frac{1-\beta_2}{1-\beta_2^t}\sum_{\tau=0}^{t-1}\beta_2^\tau(g_{t-\tau}\odot g_{t-\tau}) \\ \phi &= \frac{1-\beta_1}{1-\beta_1^t}\sum_{\tau=0}^{t-1}\beta_1^\tau g_{t-\tau} \end{aligned} \tag{9}$$

Q2:给出标签平滑正则化方法下的交叉熵损失函数

A2:

Cross-entropy without label smoothing:

$$H(p,q) = -\sum_i p(i)\log q(i)$$

where

$$\begin{aligned} p(i) &= \text{true probability of class } i, \\ q(i) &= \text{predicted probability of class } i. \end{aligned}$$

With label smoothing, the true distribution $p'(i)$ is modified:

$$p'(i) = \begin{cases} 1-\epsilon & \text{if } i \text{ is the correct class,} \\ \frac{\epsilon}{K-1} & \text{otherwise.} \end{cases}$$

The cross-entropy with label smoothing then is:

$$H(p',q) = -\left((1-\epsilon)\log q(\text{correct class}) + \sum_{j\neq\text{correct class}}\frac{\epsilon}{K-1}\log q(j)\right)$$

where

$$\begin{aligned} \epsilon &= \text{smoothing parameter}, \\ K &= \text{number of classes}. \end{aligned}$$