提出了动量算法的一个变种。这种情况的更新规则如下:

$$oldsymbol{v}$$
一个变种。这种情况的更新规则如下:  
 $oldsymbol{v} \leftarrow lpha oldsymbol{v} - \epsilon 
abla_{oldsymbol{\theta}} \left[ \frac{1}{m} \sum_{i=1}^{m} L(oldsymbol{f}(oldsymbol{x}^{(i)}; oldsymbol{\theta} + lpha oldsymbol{v}), oldsymbol{y}^{(i)}) \right],$  (8.21)  
 $oldsymbol{\theta} \leftarrow oldsymbol{\theta} + oldsymbol{v},$  (8.22)

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{v},$$
 (8.22)

其中参数  $\alpha$  和  $\epsilon$  发挥了和标准动量方法中类似的作用。Nesterov 动量和标准动量之 间的区别体现在梯度计算上。Nesterov 动量中,梯度计算在施加当前速度之后。因此, Nesterov 动量可以解释为往标准动量方法中添加了一个校正因子。完整的 Nesterov 动量算法如算法8.3所示。

## 算法 8.3 使用 Nesterov动量的随机梯度下降(SGD)

Require: 学习速率  $\epsilon$ , 动量参数  $\alpha$ 

Require: 初始参数  $\theta$ , 初始速度 v

while 没有达到停止准则 do

从训练集中采包含 m 个样本  $\{ {m x}^{(1)}, \ldots, {m x}^{(m)} \}$  的minibatch,对应目标为  ${m y}^{(i)}$ 。

应用临时更新:  $\tilde{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta} + \alpha \boldsymbol{v}$  发友一岛

计算梯度 ( 在临时点 ):  $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\tilde{\boldsymbol{\theta}}} \sum_{i} L(f(\mathbf{x}^{(i)}; \tilde{\boldsymbol{\theta}}), \mathbf{y}^{(i)})$ 

计算速度更新:  $\mathbf{v} \leftarrow \alpha \mathbf{v} - \epsilon \mathbf{g}$ 

应用更新: $\theta \leftarrow$	heta+v
end while	
	·   ·
	.
	.
	.



