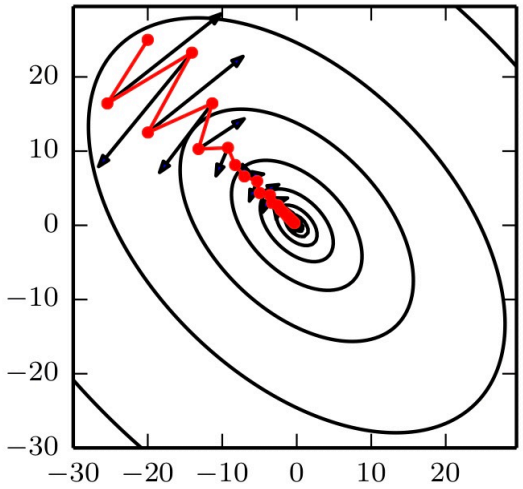

动量克服Hessian矩阵病态的难题.

图 8.5: 动量的主要目的是解决两个问题：Hessian矩阵的不良条件数和随机梯度的方差。我们通过此图说明动量如何克服这两个问题的第一个。轮廓线描绘了一个二次损失函数（具有不良条件数的Hessian矩阵）。横跨轮廓的红色路径表示动量学习规则所遵循的路径，它使该函数最小化。我们在每个步骤画一个箭头，指示梯度下降将在该点采取的步骤。我们可以看到，一个条件数较差的二次目标函数看起来像一个长而窄的山谷或陡峭的峡谷。动量正确地纵向穿过峡谷，而梯度步骤则会浪费时间在峡谷的窄轴上来回移动。比较图4.6，它也显示了没有动量的梯度下降的行为。

更新规则如下：

$$\boldsymbol{v} \leftarrow \alpha \boldsymbol{v} - \epsilon \nabla_{\boldsymbol{\theta}} \left(\frac{1}{m} \sum_{i=1}^{m} L(\boldsymbol{f}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)}) \right), \tag{8.15}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{v}. \tag{8.16}$$

速度 $\boldsymbol{v}$ 累积了梯度元素 $\nabla_{\boldsymbol{\theta}}(\frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{f}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}),\boldsymbol{y}^{(i)}))$。相对于 ϵ 的 α 越大，之前梯度对现在方向的影响也越大。带动量的SGD算法如算法8.2所示。

算法 8.2 使用动量的随机梯度下降(SGD)

Require: 学习速率 ϵ，动量参数 α

Require: 初始参数 $\boldsymbol{\theta}$，初始速度 $\boldsymbol{v}$

while 没有达到停止准则 **do**

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

计算梯度估计：$\boldsymbol{g} \leftarrow \frac{1}{m}\nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}),\boldsymbol{y}^{(i)})$

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

应用更新：$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{v}$

end while