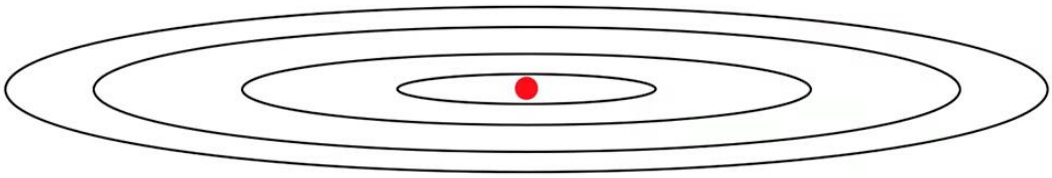
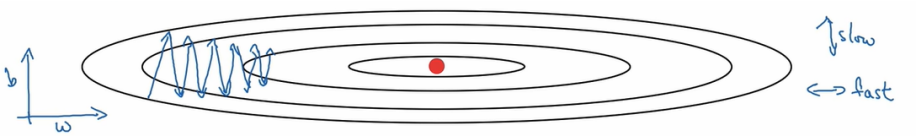
Gradient Descent with Momentum:



这里我们会试图去优化上图所示的损失函数，红点代表着全局最小值，



它具有两个维度，假设为b 和 ω，显然，我们可以看到由于函数沿b方向的 gradient 更大 (larger db) 这导致在梯度下降的过程中会出现上图蓝线样式的波动，使得我们需要更多的时间来达到全局最小值。我们可以通过 **Gradient Descent with Momentum** 来解决这个问题，算法细节如下:

On interaction t, compute $d\omega, db$ on current mini batch, the learning rate is α:

Initial with $S_{d\omega} = S_{db} = 0$

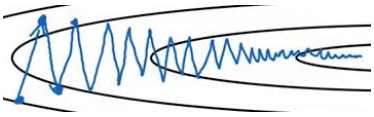
$$S_{d\omega} = \beta S_{d\omega} + (1-\beta)d\omega$$
$$S_{db} = \beta S_{db} + (1-\beta)db$$

Herer S means the exponentially weighted average (please refer to the Exponentially Weighted Averages.pdf), and the position will be updated using the above two equations:

$$\omega := \omega - \alpha S_{d\omega} \qquad b := b - \alpha S_{db}$$

Usually, $\beta = 0.9$, which is like averaging the gradients of the last 10 iterations. Therefore, it smooths out the steps of gradient descent

这样一来，如下图所示，在 b 方向的 gradient 会相互抵消，而在 ω 方向的 gradient 则不会出现这种情况，从而减小了波动，使得模型更快收敛



Note: 这个算法收到了基础物理中的启发，我们可以把算法中的 $d\omega$，db比作加速度(当前状态)，把 $S_{d\omega}$，S_{db}看作速度