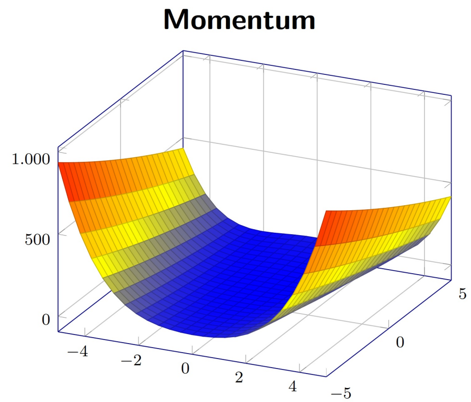
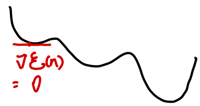
陈翊乔 YiqiaoChen 2024/10/20



- Gradient Descent would move quickly down the walls, but very slowly through the valley floor

Advantages:

1. Help network out of local minima



At the local minimum point, the current gradient is 0, making it difficult for the NN to escape. The Momentum introduces the previous gradient values to help the NN learning.

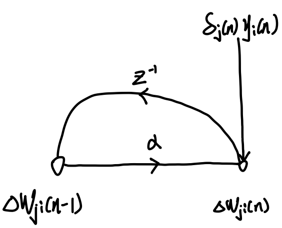
2. Accelerate learning using SGD.

3. When the gradient keeps changing direction, momentum will smooth out the variations

① Derivation

control the decay rate by $|\alpha| \in [0, 1)$, for the time series to be convergent

$$\Delta W_{ij}(n) = \underbrace{\alpha \Delta W_{ij}(n-1)}_{\text{Momentum term}} + \eta \delta_j(n) y_i(n)$$



expand the difference equation as:

$$\begin{aligned}
\Delta W_{ij}(n) &= \alpha \Delta W_{ij}(n-1) + \eta \delta_j(n) y_i(n) \\
&= \alpha(\alpha \Delta W_{ij}(n-2) + \eta \delta_j(n-1) y_i(n-1)) + \eta \delta_j(n) y_i(n) \\
&\vdots \\
&= \alpha^n \Delta W_{ij}(0) + \alpha^{n-1} \eta \delta_j(1) y_i(1) + \cdots + \eta \delta_j(n) y_i(n) \\
&= \sum_{t=0}^{n} \alpha^{n-t} \eta \delta_j(t) y_i(t) \\
&= -\eta \sum_{t=0}^{n} \alpha^{n-t} \frac{\partial \mathcal{E}(t)}{\partial W_{ji}(t)}
\end{aligned}$$

② Conclusion

I. When the partial derivative has same algebraic sign on consecutive iterations, ΔW_{ji} grows in magnitude, $W_{ji}(n)$ adjusted by large amount, SGD is tends to accelerate descent.

II. When the partial derivative has opposite signs on consecutive iterations, ΔW_{ji} shinks in magnitude, $W_{ji}(n)$ adjusted by small amount, SGD is tends to stabilizing effect.