Neural Network Optimization – Stochastic gradient descent

Algorithm 1 Stochastic gradient descent - $O(n)$

1: **procedure** WEIGHT UPDATE (θ initial weights, ϵ_i learning rate in iteration k, m batch size)
2: $k \leftarrow 1$
3: **while** stopping criterion not met **do**
4: Estimate average batch gradient: $\hat{\mathbf{g}} = \frac{1}{m} \nabla_\theta \sum_i^m L(f(\mathbf{x}_i; \theta); \mathbf{y}_i)$
5: Update the weights: $\theta' = \theta - \epsilon_k \hat{\mathbf{g}}(\theta)$
6: $k \leftarrow k + 1$

Neural Network Optimization – Stochastic gradient descent

Stochastic (gradient of a batch) as opposed to deterministic (gradient of the whole dataset)

Standard error of the mean.
Unbiased estimate of the gradient.
Computational effort

Neural Network Optimization – Stochastic gradient descent

Stochastic

Standard error of the mean $\frac{\sigma}{\sqrt{m}}$. Decreases only by $\sqrt{m}$.
With *m* samples in a batch.

Unbiased estimate of the gradient.
Computational effort

Neural Network Optimization – Stochastic gradient descent

Stochastic

Standard error of the mean.

Randomly selected set of *m* training samples for a batch achieves an **unbiased estimate of the gradient**.

Computational effort

Neural Network Optimization – Stochastic gradient descent

Stochastic

Standard error of the mean.

Unbiased estimate of the gradient.

Limiting number of *m* samples per batch, sets and upper bound to the **computational effort** during the update (growing datasets, growing sample size)