learning rate

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

A learning rate is the parameter of an iterative machine learning (ML) method that scales the change applied to the current model parameters in one iteration. In the gradient step of gradient descent (GD), the learning rate η multiplies the negative gradient of the objective function, so the update is $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)})$. A learning rate that is too large lets the iterates overshoot, so the objective function value increases, while one that is too small makes too little progress within the available number of iterations. The learning rate is therefore a hyperparameter, chosen by validation or by a schedule that decreases it over the iterations.

Definition

Consider an iterative machine learning (ML) method for finding or learning a useful hypothesis $h \in \mathcal{H}$. Such an iterative method repeats similar computational (update) steps that adjust or modify the current hypothesis to obtain an improved hypothesis. A key parameter of an iterative method is the learning rate. The learning rate controls the extent to which the current hypothesis can be modified during a single iteration. Consider, for example, the gradient step (Jung, 2022, Ch. 5)

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)}) \tag{1}$$

of a gradient-based method for empirical risk minimization (ERM), where the objective function $f(\mathbf{w})$ is the empirical risk incurred by $h^{(\mathbf{w})}$ on a training set. Given the current model parameters $\mathbf{w}^{(t)}$ at iteration t, the gradient step produces updated model parameters $\mathbf{w}^{(t+1)}$ by moving in the opposite direction of the gradient $\nabla f(\mathbf{w}^{(t)})$.

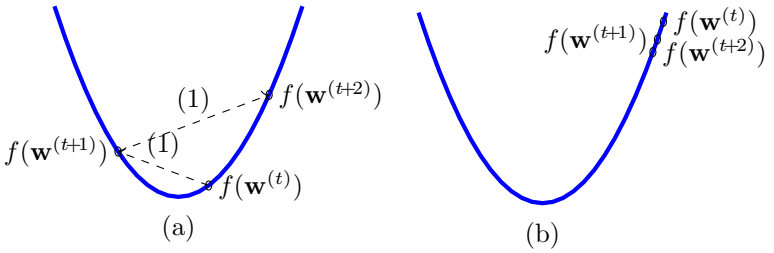

Figure 1: Effect of using an inadequate learning rate η in the gradient step (1). (a) If η is too large, the gradient steps can "overshoot" such that the iterates $\mathbf{w}^{(t)}$ diverge away from the optimum, i.e., $f(\mathbf{w}^{(t+1)}) > f(\mathbf{w}^{(t)})$. (b) If η is too small, the gradient steps make too little progress towards the optimum within the available number of iterations (due to limited computational budget).

Cross References

ML, hypothesis, parameter, GD, SGD, projected GD, step size

References

1. Jung (2022). Machine Learning: The Basics. Springer Nature, Singapore, Singapore.