stochastic gradient descent (SGD)

Authors

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

Abstract

Stochastic gradient descent (SGD) is a variant of gradient descent (GD) in which the gradient of the objective function is replaced by a computationally cheaper stochastic approximation. Its main application in machine learning (ML) is empirical risk minimization (ERM) on a training set that is large or stored in a distributed database: the gradient of the empirical risk is a sum of one gradient per data point, so its exact evaluation requires a pass over the entire training set. SGD approximates this sum by the sum over a randomly chosen subset of data points, referred to as a batch. The batch size is an important parameter of SGD: it trades the computational cost of a single update against the accuracy of the gradient approximation.

Definition

SGD is obtained from gradient descent (GD) by replacing the gradient of the objective function with a stochastic approximation. A main application of SGD is to train a parameterized model via empirical risk minimization (ERM) on a training set $\mathcal{D}$ that is either large or not readily available (e.g., when data points are stored in a database distributed globally). To evaluate the gradient of the empirical risk (as a function of the model parameters $\mathbf{w}$), it is necessary to compute a sum $\sum_{r=1}^{m} \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(r)}, \mathbf{w}\right)$ over all data points in the training set. A stochastic approximation to the gradient is obtained by replacing the sum $\sum_{r=1}^{m} \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(r)}, \mathbf{w}\right)$ with a sum $\sum_{r \in \mathcal{B}} \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(r)}, \mathbf{w}\right)$ over a randomly chosen subset $\mathcal{B} \subseteq \{1, \ldots, m\}$ (see Fig. 1). These randomly chosen data points are often referred to as a batch. The batch size $|\mathcal{B}|$ is an important parameter of SGD. SGD with $|\mathcal{B}| > 1$ is referred to as mini-batch SGD (Bottou, 1999).

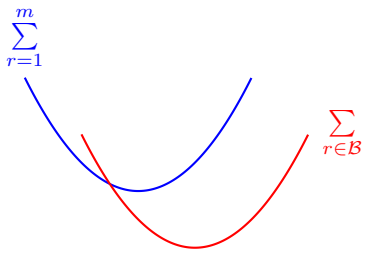

Figure 1: SGD for ERM approximates the gradient by replacing the sum $\sum_{r=1}^{m} \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(r)}, \mathbf{w}\right)$ over all data points in the training set (indexed by $r = 1, \ldots, m$) with a sum over a randomly chosen subset $\mathcal{B} \subseteq \{1, \ldots, m\}$

Cross References

GD, gradient, objective function, stochastic, model, ERM, training set, data point, empirical risk, function, model parameter, batch, parameter

References

1. Bottou (1999). On-line learning and stochastic approximations. In: On-Line Learning in Neural Networks, pp. 9–42. Cambridge Univ. Press.