

# boosting

## Authors

Alexander Jung\*, Department of Computer Science, Aalto University, Espoo, Finland

\*Corresponding author: alex.jung@aalto.fi

## Abstract

Boosting is an iterative optimization method that learns an accurate hypothesis by sequentially combining less accurate base learners, referred to as weak learners. Starting from an initialization, each iteration adds the output of a base learner, scaled by a learning rate, to the current hypothesis. This update generalizes the gradient step of gradient-based methods for empirical risk minimization (ERM): the output of the base learner acts as a negative generalized gradient. The fixed points of the update are hypotheses at which no base learner can reduce the training error further. Gradient boosting instantiates this scheme with decision trees as base learners.

## Definition

An online advertiser must predict, for each impression, whether the user will click on the shown advertisement. Widely deployed predictors for this task are built by boosting: an iterative optimization method that learns an accurate hypothesis map (or strong learner) by sequentially combining less accurate base learners (referred to as weak learners) (Drucker, 1997; Hastie et al., 2009, Chap. 10; Ridgeway et al., 1999; Schapire, 1999). Boosting can be understood as a generalization of gradient-based methods for empirical risk minimization (ERM) using parametric models and smooth loss functions (Friedman, 2001).

In particular, starting from an initialization  $\tilde{h}$ , boosting methods construct a sequence of hypotheses  $\tilde{h}^{(t)}$ ,  $t = 1, \dots$ , via a generalized gradient step

$$\tilde{h}^{(t)} = \tilde{h}^{(t-1)} + \eta^{(t)} \hat{h}^{(t)}.$$

Here,  $\eta^{(t)}$  denotes a learning rate and  $\hat{h}^{(t)}$  is provided by the  $t$ th base learner (see Fig. 1). Comparing the above update with the plain gradient step suggests viewing  $\hat{h}^{(t)}$  as a (negative) generalized gradient.

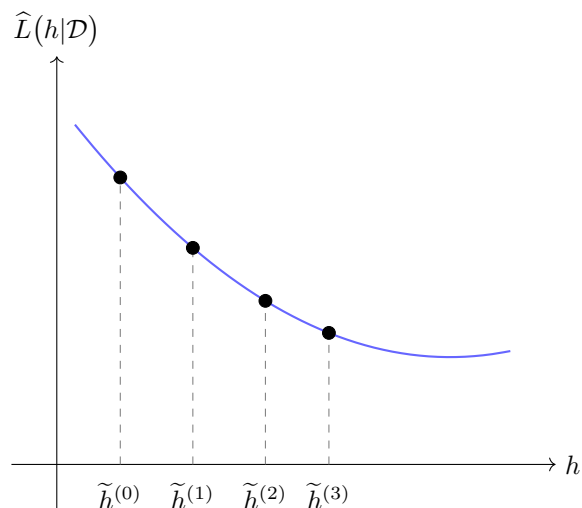


Figure 1: Boosting methods construct a sequence of hypothesis maps via a generalized gradient step. This generalized gradient step uses the output of base learners

The update is a fixed-point iteration on the hypothesis space: its fixed points are hypotheses at which the base learners provide no further descent direction, so the sequence stalls exactly where no base learner can reduce the training error further. Boosting methods differ in their choice of base learners for computing the generalized gradients  $\hat{h}^{(t)}$ . For instance, gradient boosting — behind many deployed click-through predictors — most commonly uses decision trees as base learners.

## Cross References

ensemble, AdaBoost, gradient boosting, base learner, gradient-based method

## References

1. Drucker (1997). Improving Regressors using Boosting Techniques. In: Proc. 14th Int. Conf. Mach. Learn., pp. 107–115.
2. Friedman (2001). Greedy Function Approximation: A Gradient Boosting Machine. Ann. Statist., 29(5), pp. 1189–1232.
3. Schapire (1999). A brief introduction to boosting. In: Proc. 16th Int. Joint Conf. Artif. Intell., pp. 1401–1406.
4. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.

5. Ridgeway et al. (1999). Boosting Methodology for Regression Problems.  
In: Proc. 7th Int. Workshop Artif. Intell. Statist.