Generalization for Streaming Data

Michael Spece[1,⋆]

1. Departments of Machine Learning & Statistics, Carnegie Mellon University
⋆Contact author: spece@cmu.edu

Keywords: generalization, online learning, streaming data, adversarial data, non-convex loss

An estimator's generalization error is an L^1 norm of the difference in the estimator's empirical and expected losses. The empirical loss is the arithmetic mean of instantaneous losses of the estimation applied to each observation. For online learning, the estimation is normally taken to be the arithmetic mean (under an appropriate functional calculus) of intra-period estimates, and each period in the stream is treated as a separate observation [2] [1]. However, doing so loses the information encoded in the ordering of the intra-period estimates and restricts the notion of empirical loss to that of the arithmetic mean, when more general notions may be preferred, such as approachability distance (game playing), calibration score (forecasting), or non-additive global cost (job scheduling). In contrast, viewing the stream as a single vector-valued observation for which there is a single instantaneous loss[1] preserves the ordering and the ability to apply any of the mentioned notions of empirical loss. Despite the diversity among these notions, there is a martingale bound on the generalization error that applies to all of them.

References

[1] Agarwal, A. and J. C. Duchi (2013). The generalization ability of online algorithms for dependent data. *Information Theory, IEEE Transactions on 59*(1), 573–587.

[2] Cesa-Bianchi, N., A. Conconi, and C. Gentile (2004). On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on 50*(9), 2050–2057.

[1] Hence, the arithmetic mean of the instantaneous losses, with respect to which generalization error is defined, is that single loss.