Proceedings of Machine Learning Research vol 195:1–2, 2023 36th Annual Conference on Learning Theory

Local Risk Bounds for Statistical Aggregation

Jaouad Mourtada JAOUAD.MOURTADA@ENSAE.FR
CREST, ENSAE, Institut Polytechnique de Paris, France

Tomas Vaškevičius TOMAS.VASKEVICIUS@EPFL.CH
Institute of Mathematics, EPFL, Switzerland

Nikita Zhivotovskiy ZHIVOTOVSKIY@BERKELEY.EDU
Department of Statistics, UC Berkeley, USA

Editors: Gergely Neu and Lorenzo Rosasco

Consider a family of predictors $\{f_\theta : \theta \in \Theta\}$ indexed by an abstract set Θ. In the aggregation problem, the aim is to combine the given predictors to achieve predictions nearly as accurate as the best one. Given a prior probability measure π on Θ, classical *global* bounds in aggregation take the form

$$\mathbf{E}R(\widehat{f}) \leqslant \mathsf{C}_{\mathsf{glob}}(n/c, \pi) := -\frac{c}{n} \log \left(\int_\Theta \exp \left[-\frac{n}{c} R(f_\theta) \right] \pi(\mathrm{d}\theta) \right),$$

where $\widehat{f}$ is the output of the aggregation procedure, $R(f_\theta)$ measures the risk of the predictor f_θ, c is a universal constant, and n is the size of the sample. An important feature of aggregation theory is that bounds of the above form can be obtained without imposing any restrictions on the set Θ. In particular, this allows to handle non-convex classes of predictors.

Aggregation has been studied in both sequential and statistical contexts. The classical results in both cases feature the same *global* complexity measure $\mathsf{C}_{\mathsf{glob}}(n/c, \pi)$, despite the sequential aggregation problem being intrinsically harder than its statistical counterpart. In this paper, we revisit and tighten classical results in the theory of aggregation in the statistical setting by replacing the global complexity with a smaller, *local* one. This local complexity measure can be seen as the global complexity computed with respect to a prior distribution $\pi_{-\frac{n}{c}R}$ that is more tightly concentrated around functions with low population risk; it can be defined as follows:

$$\mathsf{C}_{\mathsf{loc}}(n/c, \pi) \asymp \mathsf{C}_{\mathsf{glob}}(n/c, \pi_{-\frac{n}{c}R}), \quad \text{where} \quad \pi_{-\frac{n}{c}R}(\mathrm{d}\theta) = \frac{\exp\left(-\frac{n}{c}R(f_\theta)\right)\pi(\mathrm{d}\theta)}{\int_\Theta \exp\left(-\frac{n}{c}R(f_{\theta'})\right)\pi(\mathrm{d}\theta')}.$$

We discuss the nature of this improvement in some special cases, including finite classes and Gaussian process priors on functions. Our approach is inspired by the technique of PAC-Bayesian localization introduced by Catoni (2007).

Among other results, we prove localized versions of the classical bound for the exponential weights estimator due to Leung and Barron (2006) and deviation-optimal bounds for the Q-aggregation estimator. Our bounds refine the results of Dai, Rigollet, and Zhang (2012) in the fixed design setting and the results of Lecué and Rigollet (2014) for random design.

Acknowledgments

J. M. and N. Z. acknowledge the support of the France-Berkeley Fund.

Extended abstract. Full version appears as [arXiv:2306.17151, v1], (Mourtada, Vaškevičius, and Zhivotovskiy, 2023)

References

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007.

Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.

Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.

Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Local risk bounds for statistical aggregation. *arXiv preprint arXiv:2306.17151*, 2023.