Proceedings of Machine Learning Research vol 134:1–4, 2021 34rd Annual Conference on Learning Theory

On the Minimal Error of Empirical Risk Minimization

Gil Kur GILKUR@MIT.EDU
Massachusetts Institute of Technology

Alexander Rakhlin RAKHLIN@MIT.EDU
Massachusetts Institute of Technology

Editors: Mikhail Belkin and Samory Kpotufe

Keywords: Non parametric Statistics, Regression, Least Squares, Empirical Risk Minimization

We study the minimal error of the Empirical Risk Minimization (ERM) procedure in the task of regression, both in the random and the fixed design settings, with a convex class of regression functions $\mathcal{F}$.[1] We are given n data points $X_1, \dots, X_n$ (distributed i.i.d. according to $\mathbb{P}$ in random design, or chosen deterministically in fixed design) and n observations of $Y_i = f^*(X_i) + \xi_i$, $1 \le i \le n$, where $f^* \in \mathcal{F}$ and $\xi_i \underset{i.i.d.}{\sim} \mathrm{N}(0,1)$. The ERM procedure with respect to the squared loss (equivalently, constrained least squares) is

$$\hat{f}_n := \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

The minimal error of ERM in the random and the fixed designs, is defined, respectively, as

$$\inf_{f^* \in \mathcal{F}} \mathbb{E}_{\xi, X} \int (\hat{f}_n - f^*)^2 d\mathbb{P} \quad \text{and} \quad \inf_{f^* \in \mathcal{F}} \mathbb{E}_{\xi} \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n - f^*)^2 (X_i).$$

These quantities represent the error that ERM will *always* incur for any underlying function $f^* \in \mathcal{F}$, no matter how 'simple' it is. The minimal error should be contrasted with the classical risk formulation for the worst-case regression function $f^* \in \mathcal{F}$.

In this work, we provide sharp lower bounds for the aforementioned quantities. Specifically, in the fixed design setting, we prove the left-hand side of the following inequality:

$$64^{-1}(\mathcal{W}_x(\mathcal{F}) - C_1 n^{-1})^2 \le \inf_{f^* \in \mathcal{F}} \mathbb{E}_{\xi} \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n - f^*)^2 (X_i) \le \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\xi} \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n - f^*)^2 (X_i) \le 4\mathcal{W}_x(\mathcal{F})$$

where $C_1 \ge 0$ is an absolute constant, $\mathcal{W}_x(\mathcal{F}) := \mathbb{E}_{\xi} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(X_i)$ is the Gaussian complexity of the class $\mathcal{F}$ with respect to $X_1, \dots, X_n$, and $\mathcal{F}$ is assumed to be uniformly bounded by 1. Informally speaking, in the fixed design setting, we show that the minimal error is governed by the *global* squared Gaussian complexity of the entire function class. This points to the lack of adaptivity of ERM to a favorable f^* for "rich" function classes.

1. Extended abstract. Full version appears as [arXiv:2102.12066]

In contrast, in the random design setting, ERM may enjoy faster rates of convergence (that is, adapt to simpler f^*), but only if the local neighborhoods around the regression function are nearly as complex as the class itself, a somewhat counter-intuitive conclusion. Specifically, we prove the left-hand side of the following inequality:

$$\forall f^* \in \mathcal{F} \quad c_1 \cdot \min\{\mathcal{W}(\mathcal{F})^2, t_{n,\mathbb{P}}(f^*, \mathcal{F})^2\} \leq \mathbb{E}_{x,\xi} \int (\hat{f}_n - f^*)^2 d\mathbb{P} \leq 64 \cdot \mathcal{W}(\mathcal{F}),$$

where $\mathcal{W}(\mathcal{F}) = \mathbb{E}\mathcal{W}_x(\mathcal{F})$ and the "critical" radius $t_{n,\mathbb{P}}(f^*, \mathcal{F})$ is defined as

$$\min\{t \in (0,2) : \mathcal{W}(B_{\mathbb{P}}(f^*, t)) \leq c_2 \mathcal{W}(\mathcal{F})\} \text{ and } B_{\mathbb{P}}(f^*, t) := \{f \in \mathcal{F} : \int (f^* - f)^2 d\mathbb{P} \leq t^2\}$$

for some absolute constants $c_1, c_2 \in (0,1)$. As an application for our bounds, we provide sharp lower bounds for performance of ERM for both Donsker and non-Donsker classes. We also discuss our results through the lens of recent studies on interpolation in overparameterized models.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *arXiv preprint arXiv:1806.05161*, 2018.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Pierre C Bellec. Optimistic lower bounds for convex regularized least-squares. *arXiv preprint arXiv:1703.01332*, 2017.

Pierre C Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques*, volume 42, pages 273–325, 2006.

Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993.

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

EM Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17(3):393–398, 1976.

Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42 (6):2340–2381, 2014.

Richard M Dudley. *Uniform central limit theorems*. Number 63. Cambridge university press, 1999.

Oliver Y Feng, Adityanand Guntuboyina, Arlene KH Kim, and Richard J Samworth. Adaptation in multivariate log-concave density estimation. *arXiv preprint arXiv:1812.11634*, 2018.

Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. Max-affine regression: Provable, tractable, and near-optimal statistical estimation. *arXiv preprint arXiv:1906.09255*, 2019.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Number 40. Cambridge University Press, 2016.

Qiyang Han. Global empirical risk minimizers with" shape constraints" are rate optimal in general dimensions. *arXiv preprint arXiv:1905.12823*, 2019.

Qiyang Han and Jon A Wellner. Multivariate convex regression: global risk bounds and adaptation. *arXiv preprint arXiv:1601.06844*, 2016.

Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019.

Arlene KH Kim and Richard J Samworth. Global rates of convergence in log-concave density estimation. *The Annals of Statistics*, 44(6):2756–2779, 2016.

Arlene KH Kim, Adityanand Guntuboyina, and Richard J Samworth. Adaptation in log-concave density estimation. *The Annals of Statistics*, 46(5):2279–2306, 2018.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Gil Kur, Yuval Dagan, and Alexander Rakhlin. Optimality of maximum likelihood for log-concave density estimation and bounded convex regression. *arXiv preprint arXiv:1903.05315*, 2019.

Gil Kur, Fuchang Gao, Adityanand Guntuboyina, and Bodhisattva Sen. Convex regression in multidimensions: Suboptimality of least squares estimators. *arXiv preprint arXiv:2006.02044*, 2020a.

Gil Kur, Alexander Rakhlin, and Adityanand Guntuboyina. On suboptimality of least squares with application to estimation of convex bodies. *arXiv preprint arXiv:2006.04046*, 2020b.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

Gilles Pisier. Some applications of the metric entropy condition to harmonic analysis. In *Banach Spaces, Harmonic Analysis, and Probability Theory*, pages 123–154. Springer, 1983.

Richard J Samworth. Recent progress in log-concave density estimation. *Statistical Science*, 33(4):493–509, 2018.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2003.

Sara A van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.