Benign Overfitting of Constant-Stepsize SGD for Linear Regression

Difan Zou* KNOWZOU@CS.UCLA.EDU
Department of Computer Science
University of California, Los Angeles, Los Angeles, CA 90095, USA

Jingfeng Wu* UUUJF@JHU.EDU
Department of Computer Science
Johns Hopkins University, Baltimore, MD 21218, USA

Vladimir Braverman VOVA@CS.JHU.EDU
Department of Computer Science
Johns Hopkins University, Baltimore, MD 21218, USA

Quanquan Gu QGU@CS.UCLA.EDU
Department of Computer Science
University of California, Los Angeles, Los Angeles, CA 90095, USA

Sham M. Kakade SHAM@CS.WASHINGTON.EDU
Department of Computer Science
University of Washington, Seattle & Microsoft Research, Seattle, WA 98195, USA

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

There is an increasing realization that algorithmic inductive biases are central in preventing overfitting; empirically, we often see a *benign overfitting* phenomenon in overparameterized settings for natural learning algorithms, such as stochastic gradient descent (SGD), where little to no *explicit* regularization has been employed. However, such algorithmic aspects of generalization are far less well understood, where we lack a sharp characterization of it and when benign overfitting occurs.

In addition to the existing works that mostly focus on the classical underparameterized regime, the focus of this work is on the overparameterization regime. In particular, we study the generalization ability of SGD in arguably the most basic setting: *constant-stepsize SGD*, with iterate averaging or tail averaging, for overparameterized linear regression. Our main result provides a sharp excess risk bound, stated in terms of the full eigenspectrum of the data covariance matrix, that reveals a bias-variance decomposition characterizing when generalization is possible: (i) the variance bound is characterized in terms of an *effective dimension* (specific for SGD) and (ii) the bias bound provides a sharp geometric characterization in terms of the location of the initial iterate (and how it aligns with the data covariance matrix). In addition, for SGD with iterate averaging (output the average of all iterates), we demonstrate the sharpness of the established excess risk bound by proving a matching lower bound (up to constant factors). For SGD with tail averaging (output the average of tail iterates), we demonstrate its improvements over SGD with iterate averaging by proving a better excess risk bound together with a nearly matching lower bound. Moreover, we reflect on a number of notable differences between the algorithmic regularization afforded by (unregularized) SGD in comparison to ordinary least squares (minimum-norm interpolation) and ridge regression. Experimental results well corroborate our theoretical findings[1].

* Equal Contribution

1. Extended abstract. Full version appears as [ArXiv:2103.12692, v3].

References

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. *Advances in neural information processing systems*, 26:773–781, 2013.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *arXiv preprint arXiv:2006.08212*, 2020.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.

Xi Chen, Qiang Liu, and Xin T Tong. Dimension independent generalization error with regularized online optimization. *arXiv preprint arXiv:2003.11196*, 2020.

Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pages 205–213, 2015.

Aymeric Dieuleveut and Francis R. Bach. Non-parametric stochastic approximation with large step sizes. *The Annals of Statistics*, 2015.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Daniel J. Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017a.

Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1):8258–8299, 2017b.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*. PMLR, 2018.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.