# Risk Comparisons in Linear Regression: Implicit Regularization Dominates Explicit Regularization

**Jingfeng Wu** UUUJF@BERKELEY.EDU
*University of California, Berkeley*

**Peter L. Bartlett*** PETER@BERKELEY.EDU
*University of California, Berkeley & Google DeepMind*

**Sham M. Kakade*** SHAM@SEAS.HARVARD.EDU
*Harvard University & Google DeepMind*

**Jason D. Lee*** JASONDLEE@BERKELEY.EDU
*University of California, Berkeley*

**Bin Yu*** BINYU@BERKELEY.EDU
*University of California, Berkeley*

**Editors:** Steve Hanneke and Tor Lattimore

## Abstract

Existing theory suggests that for linear regression problems categorized by capacity and source conditions, *gradient descent* (GD) is always minimax optimal, while both *ridge regression* and online *stochastic gradient descent* (SGD) are polynomially suboptimal for certain categories of such problems. Moving beyond minimax theory, this work provides *instance-wise* comparisons of the finite-sample risks for these algorithms on any well-specified linear regression problem.

Our analysis yields three key findings. First, GD *dominates* ridge regression: with comparable regularization, the excess risk of GD is *always* within a constant factor of that of ridge, but ridge can be *polynomially* worse even when tuned optimally. Second, GD is *incomparable* with SGD. While it is known that for certain problems GD can be polynomially better than SGD, the reverse is also true: we construct problems, inspired by *benign overfitting* theory, where optimally stopped GD is polynomially worse. Finally, GD dominates SGD for a significant subclass of problems—those with fast and continuously decaying covariance spectra—which includes all problems satisfying the standard capacity condition.[1]

**Keywords:** implicit regularization, gradient descent, early stopping, linear regression, risk bounds, statistical dominance

## References

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pages 1370–1378. PMLR, 2019.

Peter L Bartlett. For valid generalization the size of the weights is more important than the size of the network. *Advances in neural information processing systems*, 9, 1996.

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.

* Alphabetical order.

1. Extended abstract. Full version appears as [arXiv:2509.17251, v2].

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

Peter Bühlmann and Bin Yu. Boosting with the $L_2$ loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52 (6):2879–2912, 2024.

Paramveer S Dhillon, Dean P Foster, Sham M Kakade, and Lyle H Ungar. A risk comparison of ordinary least squares vs ridge regression. *The Journal of Machine Learning Research*, 14(1): 1505–1511, 2013.

Lee H. Dicker, Dean P. Foster, and Daniel Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11(1):1022 – 1047, 2017. doi: 10.1214/17-EJS1258.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363 – 1399, 2016. doi: 10.1214/15-AOS1391.

Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in neural information processing systems*, 32, 2019.

William James and Charles Stein. Estimation with quadratic loss. In *Breakthroughs in statistics: Foundations and basic theory*, pages 443–460. Springer, 1992.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Yicheng Li, Haobo Zhang, and Qian Lin. On the saturation effect of kernel ridge regression. In *The Eleventh International Conference on Learning Representations*, 2023.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

Licong Lin, Jingfeng Wu, and Peter Bartlett. Improved scaling laws in linear regression via data reuse. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator. *arXiv preprint arXiv:2403.08938*, 2024.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

Ryan J Tibshirani. Prediction, generalization, and complexity: Revisiting the view from classical statistics. Lecture Note for Simons MPG Bootcamp, 2024. URL https://www.stat.berkeley.edu/~ryantibs/talks/simons-mpg-2024.pdf.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2 edition, 2026.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International Conference on Machine Learning*, pages 24280–24314. PMLR, 2022a.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *The 36th Conference on Neural Information Processing Systems*, 2022b.

Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Benefits of early stopping in gradient descent for overparameterized logistic regression. In *Forty-second International Conference on Machine Learning*, 2025.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Haihan Zhang, Yuanshi Liu, Qianwen Chen, and Cong Fang. The optimality of (accelerated) sgd for high-dimensional quadratic optimization. *arXiv preprint arXiv:2409.09745*, 2024.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham M Kakade. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34:5456–5468, 2021.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Risk bounds of multi-pass sgd for least squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 35:12909–12920, 2022.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *Journal of Machine Learning Research*, 24(326):1–58, 2023.