Proceedings of Machine Learning Research vol 195:1–9, 2023 36th Annual Conference on Learning Theory

Implicit Balancing and Regularization: Generalization and Convergence Guarantees for Overparameterized Asymmetric Matrix Sensing (Extended Abstract)

Mahdi Soltanolkotabi SOLTANOL@USC.EDU
University of Southern California

Dominik Stöger DOMINIK.STOEGER@KU.DE
KU Eichstätt-Ingolstadt

Changzhi Xie CHANGZHI@USC.EDU
University of Southern California

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Recently, there has been significant progress in understanding the convergence and generalization properties of gradient-based methods for training overparameterized learning models. However, many aspects including the role of small random initialization and how the various parameters of the model are coupled during gradient-based updates to facilitate good generalization remain largely mysterious. A series of recent papers have begun to study this role for non-convex formulations of symmetric Positive Semi-Definite (PSD) matrix sensing problems which involve reconstructing a low-rank PSD matrix from a few linear measurements. The underlying symmetry/PSDness is crucial to existing convergence and generalization guarantees for this problem. In this paper, we study a general overparameterized low-rank matrix sensing problem where one wishes to reconstruct an asymmetric rectangular low-rank matrix from a few linear measurements. We prove that an overparameterized model trained via factorized gradient descent converges to the low-rank matrix generating the measurements. We show that in this setting, factorized gradient descent enjoys two implicit properties: (1) coupling of the trajectory of gradient descent where the factors are coupled in various ways throughout the gradient update trajectory and (2) an algorithmic regularization property where the iterates show a propensity towards low-rank models despite the overparameterized nature of the factorized model. These two implicit properties in turn allow us to show that the gradient descent trajectory from small random initialization moves towards solutions that are both globally optimal and generalize well. [1]

Keywords: asymmetric matrix sensing, factorized gradient descent, overparameterization, generalization with small random initialization, non-convex optimization

Acknowledgments

MS is supported by the Packard Fellowship in Science and Engineering, a Sloan Fellowship in Mathe- matics, an NSF-CAREER under award 1846369, DARPA Learning with Less Labels (LwLL) and FastNICS programs, and NSF-CIF awards 1813877 and 2008443. Furthermore, the authors want to thank Rene Vidal for fruitful discussions about matrix factorization.

1. Extended abstract. Full version appears as arXiv:2303.14244, v2.

References

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=wt7cd9m2cz2.

Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory*, 57(4):2342–2359, 2011. ISSN 0018-9448.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

Lijun Ding, Zhen Qin, Liwei Jiang, Jinxin Zhou, and Zhihui Zhu. A validation approach to over-parameterized matrix and image recovery. *arXiv preprint arXiv:2209.10675*, 2022.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f79921bbae40a577928b76d2fc3edc2a-Paper.pdf.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*, 31, 2018.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6152–6160, 2017.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. 31, 2018. URL https://proceedings.neurips.cc/paper/2018/file/0e98aeeb54acf612b9eb4e48a269814c-Paper.pdf.

Liwei Jiang, Yudong Chen, and Lijun Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *arXiv preprint arXiv:2203.02839*, 2022.

Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon S Du, and Jason D Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. *arXiv preprint arXiv:2301.11500*, 2023.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. volume 75 of *Proceedings of Machine Learning Research*, pages 2–47. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/li18a.html.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=AHOs7Sm5H7R.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.

Michael O'Neill and Stephen J Wright. A line-search descent algorithm for strict saddle functions with complexity guarantees. *Journal of Machine Learning Research*, 24(10):1–34, 2023.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21174–21187. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f21e255f89e0f258accbe4e984eef486-Paper.pdf.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.

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

Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. *arXiv preprint arXiv:2302.01186*, 2023.

Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34:1429–1439, 2021.

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