Spectral Estimators for Structured Generalized Linear Models via Approximate Message Passing (Extended Abstract)

Yihan Zhang ZEPHYR.Z798@GMAIL.COM
Institute of Science and Technology Austria

Hong Chang Ji HONGCHANG.JI@IST.AC.AT
Institute of Science and Technology Austria

Ramji Venkataramanan RV285@CAM.AC.UK
University of Cambridge

Marco Mondelli MARCO.MONDELLI@IST.AC.AT
Institute of Science and Technology Austria

Editors: Shipra Agrawal and Aaron Roth

Abstract

We consider the problem of parameter estimation in a high-dimensional generalized linear model. Spectral methods obtained via the principal eigenvector of a suitable data-dependent matrix provide a simple yet surprisingly effective solution. However, despite their wide use, a rigorous performance characterization, as well as a principled way to preprocess the data, are available only for unstructured (i.i.d. Gaussian and Haar orthogonal) designs. In contrast, real-world data matrices are highly structured and exhibit non-trivial correlations. To address the problem, we consider correlated Gaussian designs capturing the anisotropic nature of the features via a covariance matrix Σ. Our main result is a precise asymptotic characterization of the performance of spectral estimators. This allows us to identify the optimal preprocessing that minimizes the number of samples needed for parameter estimation. Surprisingly, such preprocessing is universal across a broad set of statistical models, which partly addresses a conjecture on optimal spectral estimators for rotationally invariant designs. Our principled approach vastly improves upon previous heuristic methods, including for designs common in computational imaging and genetics. The proposed methodology, based on approximate message passing, is broadly applicable and opens the way to the precise characterization of spiked matrices and of the corresponding spectral methods in a variety of settings.[1]

Keywords: High-dimensional estimation, structured data, generalized linear models, spectral estimator, approximate message passing

Acknowledgments

Y.Z. and M.M. are partially supported by the 2019 Lopez-Loreta Prize and by the Interdisciplinary Projects Committee (IPC) at ISTA. H.C.J. is supported by the ERC Advanced Grant "RMTBeyond" No. 101020331.

1. Extended abstract. Full version appears as [arXiv:2308.14507, v2].

References

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Information Theory*, 57:764–785, 2011.

Serban T. Belinschi, Hari Bercovici, Mireille Capitaine, and Maxime Février. Outliers in the spectrum of large deformed unitarily invariant models. *The Annals of Probability*, 45(6A):3571–3625, 2017.

Florent Benaych-Georges. Rectangular random matrices, related convolution. *Probability Theory and Related Fields*, 144(3-4):471–515, 2009.

Florent Benaych-Georges. On a surprising relation between the Marchenko-Pastur law, rectangular and square free convolutions. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 46(3):644–652, 2010.

Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

Petros Boufounos and Richard G. Baraniuk. 1-bit compressive sensing. In *42nd Annual Conference on Information Sciences and Systems (CISS)*, pages 16–21, 2008.

Zhiqi Bu, Jason M. Klusowski, Cynthia Rush, and Weijie J. Su. Characterizing the SLOPE trade-off: a variational perspective and the Donoho-Tanner limit. *The Annals of Statistics*, 51(1):33–61, 2023.

Collin Cademartori and Cynthia Rush. A non-asymptotic analysis of generalized approximate message passing algorithms with right rotationally invariant designs. *arXiv preprint arXiv:2302.00088*, 2023.

Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015a.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Information Theory*, 61(4):1985–2007, 2015b.

Michael Celentano and Andrea Montanari. CAD: Debiasing the lasso with inaccurate covariate model. *arXiv preprint arXiv:2107.14172*, 2021.

Michael Celentano, Andrea Montanari, and Yuting Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 2023.

Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *Conference on Learning Theory (COLT)*, pages 1161–1227, 2020.

Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5): 822–883, 2017.

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *The Annals of Statistics*, 47(4):2204–2235, 2019.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: A statistical perspective. *Foundations and Trends in Machine Learning*, 14(5):566–806, 2021.

Romain Couillet and Walid Hachem. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices. Theory and Applications*, 3(4): 1450016, 23, 2014.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7:1–46, 1970.

Abhishek Dhawan, Cheng Mao, and Ashwin Pananjady. Sharp analysis of EM for learning mixtures of pairwise differences. In *Conference on Learning Theory (COLT)*, pages 4384–4428, 2023.

Xiucai Ding and Hong Chang Ji. Spiked multiplicative random matrices and principal components. *Stochastic Processes and their Applications*, 163:25–60, 2023.

Xiucai Ding and Fan Yang. Spiked separable covariance matrices and principal components. *The Annals of Statistics*, 49(2):1113–1138, 2021.

David Donoho and Andrea Montanari. High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

David L. Donoho, Arian Maleki, and Andrea Montanari. Message Passing Algorithms for Compressed Sensing. *Proceedings of the National Academy of Sciences*, 106:18914–18919, 2009.

David L. Donoho, Adel Javanmard, and Andrea Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. Information Theory*, 59(11):7434–7464, 2013.

Rishabh Dudeja, Milad Bakhshizadeh, Junjie Ma, and Arian Maleki. Analysis of spectral methods for phase retrieval with random orthogonal matrices. *IEEE Trans. Information Theory*, 66(8): 5182–5203, 2020a.

Rishabh Dudeja, Junjie Ma, and Arian Maleki. Information theoretic limits for phase retrieval with subsampled haar sensing matrices. *IEEE Trans. Information Theory*, 66(12):8002–8045, 2020b.

Rishabh Dudeja, Yue M Lu, and Subhabrata Sen. Universality of approximate message passing with semi-random matrices. *The Annals of Probability*, 2023.

Zhou Fan. Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197–224, 2022.

Zhou Fan, Yi Sun, and Zhichao Wang. Principal components in linear mixed models with general bulk. *The Annals of Statistics*, 49(3):1489–1513, 2021.

Albert Fannjiang and Thomas Strohmer. The numerics of phase retrieval. *Acta Numerica*, 29: 125–228, 2020.

Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A unifying tutorial on approximate message passing. *Foundations and Trends in Machine Learning*, 15(4):335–536, 2022.

J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, Aug 1982.

Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.

Ulf Grenander and Gábor Szegő. *Toeplitz forms and their applications*. Chelsea Publishing Co., New York, second edition, 1984.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014a.

Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014b.

Adel Javanmard and Andrea Montanari. Debiasing the Lasso: optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

Yoshiyuki Kabashima. A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111–11121, Oct 2003.

M. Kac, W. L. Murdock, and G. Szegö. On the eigenvalues of certain Hermitian forms. *Journal of Rational Mechanics and Analysis*, 2:767–800, 1953.

Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

Gen Li, Yuantao Gu, and Yue M. Lu. Phase retrieval using iterative projections: Dynamics in the large systems limit. In *53rd Annual Allerton Conference on Communication, Control, and Computing*, pages 1114–1118, 2015.

Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

Yue M Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541, 2020.

Wangyu Luo, Wael Alghamdi, and Yue M. Lu. Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Trans. Signal Processing*, 67(9):2347–2356, 2019.

Junjie Ma, Rishabh Dudeja, Ji Xu, Arian Maleki, and Xiaodong Wang. Spectral method for phase retrieval: an expectation propagation perspective. *IEEE Trans. Information Theory*, 67(2):1332–1355, 2021a.

Junjie Ma, Ji Xu, and Arian Maleki. Analysis of sensing spectral for signal recovery under a generalized linear model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 22601–22613, 2021b.

Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11071–11082, 2020.

Antoine Maillard, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. Construction of optimal spectral methods in phase retrieval. In *Mathematical and Scientific Machine Learning (MSML)*, pages 693–720, 2022.

P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1989.

Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, 19(3):703–773, 2019.

Marco Mondelli and Ramji Venkataramanan. Approximate message passing with spectral initialization for generalized linear models. *Journal of Statistical Mechanics: Theory and Experiment*, (11), 2022.

Marco Mondelli, Christos Thrampoulidis, and Ramji Venkataramanan. Optimal combination of linear and spectral estimators for generalized linear models. *Foundations of Computational Mathematics*, pages 1–54, 2021.

Andrea Montanari and Yuchen Wu. Adversarial examples in random neural networks with general activations. *Mathematical Statistics and Learning*, 6(1):143–200, 2023.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *IEEE Trans. Signal Processing*, 63(18):4814–4826, 2015.

Jonathan Novak. Three lectures on free probability. In *Random matrix theory, interacting particle systems, and integrable systems*, volume 65, pages 309–383. Cambridge Univ. Press, 2014.

Aleksandr Pak, Justin Ko, and Florent Krzakala. Optimal algorithms for the inhomogeneous spiked wigner model. In *Advances in Neural Information Processing Systems*, 2023.

Debashis Paul and Jack W. Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1): 37–57, 2009.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning. *Book draft*, 2022.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symposium on Inf. Theory (ISIT)*, pages 2168–2172, 2011.

Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Pragya Sur. *A modern maximum likelihood theory for high-dimensional logistic regression*. PhD thesis, Stanford University, 2019.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Christos Thrampoulidis and Ankit Singh Rawat. Lifting high-dimensional non-linear models with gaussian regressors. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3206–3215, 2019.

William F. Trench. Numerical solution of the eigenvalue problem for symmetric rationally generated Toeplitz matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(2):291–303, 1988.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Ramji Venkataramanan, Kevin Kögler, and Marco Mondelli. Estimation in rotationally invariant generalized linear models via approximate message passing. In *International Conference on Machine Learning (ICML)*, pages 22120–22144, 2022.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55(5):2183–2202, 2009.

Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, MaxCut and complex semidefinite programming. *Mathematical Programming*, 149:47–81, 2015.

Gang Wang, Georgios B. Giannakis, and Yonina C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Trans. Information Theory*, 64(2):773–794, 2018.

Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*, 2022.

Ke Wei. Solving systems of phaseless equations via Kaczmarz methods: a proof of concept study. *Inverse Problems*, 31(12):125008, 23, 2015.

Yihong Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW (UIUC)*, 16, 2017.

Yuchen Wu and Kangjie Zhou. Lower bounds for the convergence of tensor power iteration on random overcomplete models. In *Conference on Learning Theory (COLT)*, pages 3783–3820, 2023.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning (ICML)*, pages 613–621, 2014.

Y. Q. Yin, Z. D. Bai, and P. R. Krishnaiah. On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78(4):509–521, 1988.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 76(1):217–242, 2014.

Lixin Zhang. *Spectral analysis of large dimentional random matrices*. PhD thesis, National University of Singapore, 2007.

Yihan Zhang, Marco Mondelli, and Ramji Venkataramanan. Precise asymptotics for spectral methods in mixed generalized linear models. *arXiv preprint arXiv:2211.11368*, 2022.

Qian Zhao, Pragya Sur, and Emmanuel J. Candès. The asymptotic distribution of the MLE in high-dimensional logistic models: arbitrary covariance. *Bernoulli*, 28(3):1835–1861, 2022.