Proceedings of Machine Learning Research vol 75:1–4, 2018 31st Annual Conference on Learning Theory

Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models

Jean Barbier JEAN.BARBIER@EPFL.CH
Laboratoire de Théorie des Communications, Faculté Informatique et Communications, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Suisse
Probability and Applications Group, School of Mathematical Sciences, Queen Mary University of London, E14NS London, United-Kingdom

Florent Krzakala FLORENT.KRZAKALA@ENS.FR
Laboratoire de Physique Statistique, CNRS & Sorbonne Universits & Ecole Normale Supérieure & PSL University, 75005 Paris, France

Nicolas Macris NICOLAS.MACRIS@EPFL.CH
Laboratoire de Théorie des Communications, Faculté Informatique et Communications, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Suisse

Léo Miolane LEO.MIOLANE@INRIA.FR
Département d'Informatique de l'ENS, Ecole Normale Supérieure & CNRS & PSL University & Inria, 75005 Paris, France

Lenka Zdeborová LENKA.ZDEBOROVA@IPHT.FR
Institut de Physique Thorique, CNRS & CEA & Universit Paris-Saclay, 91191 Gif-sur-Yvette, France

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Generalized linear models (GLMs) arise in high-dimensional machine learning, statistics, communications and signal processing. In this paper we analyze GLMs when the data matrix is random, as relevant in problems such as compressed sensing, error-correcting codes or benchmarks models in neural networks. We evaluate the mutual information (or "free entropy") from which we deduce the Bayes-optimal inference and generalization errors. Our analysis applies to the high-dimensional limit where both the number of samples and dimensions are large and their ratio is fixed. Non-rigorous predictions for the optimal inference and generalization errors existed for special cases of GLMs, e.g. for the perceptron in the field of statistical physics based on the so-called replica method. Our present paper rigorously establishes those decades old conjectures and brings forward their algorithmic interpretation in terms of performance of the generalized approximate message-passing algorithm. Furthermore, we tightly characterize, for many learning problems, regions of parameters for which this algorithm achieves the optimal performance, and locate the associated sharp phase transitions separating learnable and non-learnable regions[1].

Keywords: high-dimensional inference | generalized linear model | Bayesian inference | perceptron | phase transitions | approximate message-passing algorithm

1. Extended abstract. Full version appears as arXiv:1708.03395

We discuss generalized linear estimation models (GLMs) Nelder and Baker (1972); McCullagh (1984) where data are generated as follows: Given a n-dimensional vector $\mathbf{X}^*$, hidden to the statistician, he/she observes instead an m-dimensional vector $\mathbf{Y}$ where each component reads

$$Y_\mu = \varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_\mu\Big)\,, \qquad 1 \le \mu \le m\,, \tag{1}$$

where $\mathbf{\Phi}$ is a $m \times n$ "measurement" or "data" matrix. The model is "linear" because the output Y_μ depends on a *linear* combination of the data $z_\mu = [\mathbf{\Phi X}^*]_\mu = \sum_{i=1}^n \Phi_{\mu i} X_i^*$. The GLM generalizes the ordinary linear regression by allowing the output function $\varphi(z)$ to be non-linear and/or stochastic. There are two main learning problems in GLMs: $i)$ The *estimation* task requires, knowing the measured vector $\mathbf{Y}$ and the matrix $\mathbf{\Phi}$, to infer the unknown vector $\mathbf{X}^*$; $ii)$ the *prediction* or *generalization* task instead requires, again knowing $\mathbf{Y}$ and $\mathbf{\Phi}$, to predict accurately new values $Y_{\rm new}$ when new rows (i.e. data-points) are added to the matrix $\mathbf{\Phi}$.

In the present work, we build a rigorous theory for these tasks for *random instances* of the GLM. In this setting each element $\Phi_{\mu i}$ of the matrix is sampled i.i.d. from a probability distribution of zero mean and unit variance, and the unknown vector $\mathbf{X}^*$ has been also created randomly from a probability distribution P_0, with each of its iid components $X_1^*, \dots, X_n^* \sim P_0$. We assume that P_0 and φ are known to the statistician: if they are not, the task can only be harder. Our results are derived in the challenging and interesting high-dimensional limit where $m, n \to \infty$ while $m/n \to \alpha$ a constant. Random instances of GLMs are both practically and theoretically relevant in many different contexts: In *compressed sensing* Donoho and Tanner (2005); Candes and Tao (2006); Donoho et al. (2009); Rangan (2011); Zdeborová and Krzakala (2016); In *statistical learning*: Bayati and Montanari (2012); El Karoui et al. (2013); Donoho and Montanari (2016); In *artificial neural networks*: Gardner and Derrida (1989); Seung et al. (1992); Watkin et al. (1993); In *communications*: Shannon (1948); Tanaka (2002); Guo and Verdú (2005); Barbier and Krzakala (2017).

Many previous studies rely on the algorithmic performance of the so-called generalized approximate message-passing algorithm (GAMP) Mézard (1989); Donoho et al. (2009); Rangan (2011). GAMP is remarkable in that its asymptotic ($n, m \to \infty$, $m/n \to \alpha$) performance can be analyzed rigorously using the so-called state evolution Bolthausen (2014); Bayati and Montanari (2011); Bayati et al. (2015). However, GAMP is not expected to be always information-theoretically optimal. Most results giving information-theoretic predictions (except for the linear case Barbier et al. (a,b); Reeves and Pfister) are based on powerful and sophisticated but *non-rigorous* techniques originating in statistical physics of disordered systems, such as the cavity and replica methods Mézard et al. (1987). Historically, the first of these non-rigorous, yet correct, results on information-theoretic limitations of learning was for the perceptron with binary weights and was established using the replica method in Gardner and Derrida (1989); Györgyi (1990); Seung et al. (1992).

We closed the above gap between mathematically rigorous work and conjectures (some of them several decades old) from statistical mechanics. In particular, we prove that the results for GLMs stemming from the replica method are indeed correct and imply the optimal value of both the estimation and generalization error. The proof is based on a powerful evolution of the interpolation method Guerra and Toninelli (2002) called the *adaptive interpolation method*, and recently developed in Barbier and Macris (2017). We compute in particular the asymptotic mutual information (or free energy in the statistical mechanics language) between the unknown variable $\mathbf{X}^*$ and the measurement $\mathbf{Y}$. We also compute the minimal mean-square error on the reconstruction of $\mathbf{X}^*$ and the generalization error in the so-called teacher-student scenario.

A second object of focus is the algorithmic complexity: When is it possible to *efficiently* perform these optimal estimations? To answer this question, we compare our information-theoretic results to the performance of the GAMP algorithm and its state evolution Rangan (2011). We determine regions of parameters where this algorithm is or is not information-theoretically optimal. Up to technical assumptions, our results apply to all activation functions φ and priors P_0, thus unifying a large volume of previous work where many particular functions have been analyzed on a case by case basis. This generality allows us to provide a unifying understanding of the types of phase transitions and phase diagrams that we can encounter in GLMs. Among other, we discuss the perceptron problem, one-bit compressed sensing, real valued-phase retrievial (or sign-less compressed sensing) and Relu-type measurements.

Acknowledgments

This work has been supported by funding from the SNSF (grant 200021-156672), from the ERC under the European Unions FP7 Grant Agreement 307087-SPARCS and the European Union's Horizon 2020 Research and Innovation Program 714608-SMiLe, as well as by the French Agence Nationale de la Recherche under grant ANR-17-CE23-0023-01 PAIL. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU and the Chaire de recherche sur les modèles et sciences des données, Fondation CFM pour la Recherche-ENS. Part of this work was done while Léo Miolane was visiting EPFL

References

J. Barbier and F. Krzakala. Approximate message-passing decoder and capacity achieving sparse superposition codes. *IEEE Transactions on Information Theory*, 63(8):4894–4927, 2017.

J. Barbier, M. Dia, N. Macris, and F. Krzakala. The mutual information in random linear estimation. In *54th Annual Allerton Conf. on Communication, Control, and Computing*, page 625, a.

Jean Barbier and Nicolas Macris. The adaptive interpolation method: A simple scheme to prove replica formulas in bayesian inference. *arXiv:1705.02780[v3]*, 2017.

Jean Barbier, Nicolas Macris, Mohamad Dia, and Florent Krzakala. Mutual information and optimality of approximate message-passing in random linear estimation. *arXiv:1701.05823*, b.

M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

M. Bayati and A. Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333366, 2014.

Emmanuel J. Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406, 2006.

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

David L Donoho and Jared Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Nat. Acad. Sci.*, 102(27):9446–9451, 2005.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proc. Nat. Acad. Sci.*, 106(45):18914–18919, Nov 2009.

N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proc. Nat. Acad. Sci.*, 110(36):14557, 2013.

Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Francesco Guerra and Fabio Lucio Toninelli. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, 230(1):71–79, 2002.

Dongning Guo and Sergio Verdú. Randomly spread cdma: Asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010, June 2005. ISSN 0018-9448.

Géza Györgyi. First-order transition to perfect generalization in a neural network with binary synapses. *Physical Review A*, 41(12):7097, 1990.

P. McCullagh. Generalized linear models. *Euro. Journal of Operational Research*, 16(3):285, 1984.

M. Mézard, G. Parisi, and MA Virasoro. *Spin glass theory and beyond.* World Sci. Publish., 1987.

Marc Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22(12):2181–2190, 1989.

John Ashworth Nelder and R Jacob Baker. *Generalized linear models*. Wiley Online Library, 1972.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *IEEE ISIT*, pages 2168–2172, July 2011.

G. Reeves and H. Pfister. The replica-symmetric prediction for compressed sensing with gaussian matrices is exact. In *Inf. Theory (ISIT), 2016 IEEE International Symposium on*, page 665.

Sebastian H. Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45:6056–6091, Apr 1992.

Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623, 1948.

Toshiyuki Tanaka. A statistical-mechanics approach to large-system analysis of cdma multiuser detectors. *IEEE Transactions on Information Theory*, 48(11):2888–2910, Nov 2002.

Timothy L. H. Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65:499–556, Apr 1993.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.