A Unifying Tutorial on Approximate Message Passing

Other titles in Foundations and Trends® in Machine Learning

Dynamical Variational Autoencoders: A Comprehensive Review
Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber and Xavier Alameda-Pineda
ISBN: 978-1-68083-912-8

Machine Learning for Automated Theorem Proving: Learning to Solve SAT and QSATe
Sean B. Holden
ISBN: 978-1-68083-898-5

Spectral Methods for Data Science: A Statistical Perspective
Yuxin Chen, Yuejie Chi, Jianqing Fan and Cong Ma
ISBN: 978-1-68083-896-1

Tensor Regression
Jiani Liu, Ce Zhu, Zhen Long and Yipeng Liu
ISBN: 978-1-68083-886-2

Minimum-Distortion Embedding
Akshay Agrawal, Alnur Ali and Stephen Boyd
ISBN: 978-1-68083-888-6

Graph Kernels: State-of-the-Art and Future Challenges
Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray and Bastian Rieck
ISBN: 978-1-68083-770-4

A Unifying Tutorial on Approximate Message Passing

Oliver Y. Feng
University of Cambridge
UK
o.feng@statslab.cam.ac.uk

Ramji Venkataramanan
University of Cambridge
UK
ramji.v@eng.cam.ac.uk

Cynthia Rush
Columbia University
USA
cgr2130@columbia.edu

Richard J. Samworth
University of Cambridge
UK
r.samworth@statslab.cam.ac.uk

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

O. Y. Feng *et al.*. *A Unifying Tutorial on Approximate Message Passing*. Foundations and Trends® in Machine Learning, vol. 15, no. 4, pp. 335–536, 2022.

ISBN: 978-1-63828-005-7

Foundations and Trends® in Machine Learning
Volume 15, Issue 4, 2022

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2022, Volume 15, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

1 Introduction **2**
1.1 Notation and Preliminaries 9

2 Master Theorems for Abstract AMP Recursions **13**
2.1 Symmetric AMP 13
2.2 Asymmetric AMP 21

3 Low-Rank Matrix Estimation **25**
3.1 An AMP Algorithm for Estimating a Symmetric Rank-One Matrix 25
3.2 Spectral Initialisation 33
3.3 Choosing the Functions g_k 37
3.4 Confidence Intervals and p-Values 47
3.5 AMP for More General Low-Rank Matrix Estimation Problems 48

4 GAMP for Generalised Linear Models **55**
4.1 Master Theorem for GAMP 56
4.2 Choosing the Functions f_k, g_k, and Inference for β 63
4.3 AMP for the Linear Model 66
4.4 GAMP Algorithms for Convex Optimisation 68
4.5 AMP for the Lasso 75

4.6 AMP for M-Estimation in the Linear Model 80
4.7 GAMP for Logistic Regression 85

5 Conclusions and Extensions **91**

Acknowledgements **95**

Appendices **96**

A Proofs and Technical Remarks **97**
A.1 Technical Remarks on the Master Theorems in Section 2.1 97
A.2 Conditional Distributions for Symmetric AMP 102
A.3 Proofs of Results in Section A.2 106
A.4 Proof Outline for the AMP Master Theorems in Section 2.1 112
A.5 Proofs for Sections 2.1 and A.1 118
A.6 Auxiliary Results and Proofs for Section 2 137
A.7 AMP with Matrix-Valued Iterates 139
A.8 Proofs for Section 3 140
A.9 Proofs for Section 4 152

B Supplementary Mathematical Background **156**
B.1 Basic Properties of Complete Convergence 156
B.2 Regular Conditional Distributions and Conditional Independence 159
B.3 Auxiliary Probabilistic Results 167
B.4 Wasserstein Convergence and Pseudo-Lipschitz Functions . 175

References **188**

A Unifying Tutorial on Approximate Message Passing

Oliver Y. Feng[1], Ramji Venkataramanan[2], Cynthia Rush[3] and Richard J. Samworth[1]

[1]*Statistical Laboratory, University of Cambridge, UK; o.feng@statslab.cam.ac.uk; r.samworth@statslab.cam.ac.uk*
[2]*Department of Engineering, University of Cambridge, UK; ramji.v@eng.cam.ac.uk*
[3]*Department of Statistics, Columbia University, USA; cgr2130@columbia.edu*

ABSTRACT

Over the last decade or so, Approximate Message Passing (AMP) algorithms have become extremely popular in various structured high-dimensional statistical problems. Although the origins of these techniques can be traced back to notions of belief propagation in the statistical physics literature, our goals in this work are to present the main ideas of AMP from a statistical perspective and to illustrate the power and flexibility of the AMP framework. Along the way, we strengthen and unify many of the results in the existing literature.

Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush and Richard J. Samworth (2022), "A Unifying Tutorial on Approximate Message Passing", Foundations and Trends® in Machine Learning: Vol. 15, No. 4, pp 335–536. DOI: 10.1561/2200000092.

1

Introduction

Approximate Message Passing (AMP) refers to a class of iterative algorithms that have been successfully applied to a number of statistical estimation tasks such as linear regression (Bayati and Montanari, 2011; Donoho *et al.*, 2009; Krzakala *et al.*, 2012), generalised linear models (Mondelli and Venkataramanan, 2020; Rangan, 2011; Schniter and Rangan, 2014) and low-rank matrix estimation (Deshpande and Montanari, 2014; Deshpande *et al.*, 2016; Kabashima *et al.*, 2016; Lesieur *et al.*, 2017; Matsushita and Tanaka, 2013; Montanari and Richard, 2016; Montanari and Venkataramanan, 2021; Rangan and Fletcher, 2018). Moreover, these techniques are also popular and practical in a variety of engineering and computer science applications such as imaging (Fletcher and Rangan, 2014; Metzler *et al.*, 2017; Vila *et al.*, 2015), communications (Barbier and Krzakala, 2017; Jeon *et al.*, 2015; Rush *et al.*, 2017; Schniter, 2011) and machine learning (El Alaoui *et al.*, 2018; Emami *et al.*, 2020; Manoel *et al.*, 2017; Pandit *et al.*, 2020; Yang, 2019). AMP algorithms have two features that make them particularly attractive. First, they can easily be tailored to take advantage of prior information on the structure of the signal, such as sparsity or other constraints. Second, under suitable assumptions on a design or data matrix, AMP

theory provides precise asymptotic guarantees for statistical procedures in the high-dimensional regime where the ratio of the number of observations n to dimensions p converges to a constant (Bayati and Montanari, 2012; Donoho and Montanari, 2016; Donoho *et al.*, 2013; Sur *et al.*, 2017). More generally, AMP has been used to obtain lower bounds on the estimation error of first-order methods (Celentano *et al.*, 2020). In generalised linear models, low-rank matrix estimation and neural network models, it also plays a fundamental role in understanding the performance gap between information-theoretically optimal and computationally feasible estimators (Aubin *et al.*, 2019, 2020; Barbier *et al.*, 2019; Lelarge and Miolane, 2019; Reeves and Pfister, 2019). In these settings, it is conjectured that AMP achieves the optimal asymptotic estimation error among all polynomial-time algorithms (cf. Celentano and Montanari, 2022).

The purpose of this tutorial is to give a comprehensive and rigorous introduction to what AMP can offer, as well as to unify and formalise the core concepts within the large body of recent work in the area. We mention here that many of the original ideas of AMP were developed in the physics and engineering literature, and involved notions such as "loopy belief propagation" (e.g., Koller and Friedman, 2009, Section 11.3) and the "replica method" (e.g., Guo and Verdú, 2005; Krzakala *et al.*, 2012; Mézard and Montanari, 2009; Tanaka, 2002; Rangan *et al.*, 2009). Our starting point, however, will be an abstract AMP recursion, whose form depends on whether or not the data matrix is symmetric; we will study the symmetric case in detail, and then present the asymmetric version, which can be handled via a reduction argument. The striking and crucial feature of this recursion is that when the dimension is large, the empirical distribution of the coordinates of each iterate is approximately Gaussian, with limiting variance given by a scalar iteration called "state evolution".

Rigorous formulations of the key AMP property are given in Theorems 2.1 and 2.3 (for the symmetric case) and Theorem 2.5 (for the asymmetric case), which can be found in Sections 2.1 and 2.2 respectively. Here, we both strengthen earlier related results, and seek to make the underlying arguments more transparent. These "master theorems", which can be viewed as asymptotic results on Gaussian random matrices,

can be adapted to analyse variants of the original AMP recursion that are geared towards more statistical problems. In this aspect, we focus on two canonical statistical settings, namely estimation of low-rank matrices in Section 3, and estimation in generalised linear models (GLMs) in Section 4. The former encompasses Sparse Principal Component Analysis (Deshpande and Montanari, 2014; Gataric *et al.*, 2020; Jolliffe *et al.*, 2003; Wang *et al.*, 2016; Zou *et al.*, 2006), submatrix detection (Ma and Wu, 2015), hidden clique detection (Alon *et al.*, 1998; Deshpande and Montanari, 2015), spectral clustering (von Luxburg, 2007), matrix completion (Candès and Recht, 2009; Zhu *et al.*, 2019), topic modelling (Blei *et al.*, 2003) and collaborative filtering (Su and Khoshgoftaar, 2009). The latter provides a holistic approach to studying a suite of popular modern statistical methods, including penalised M-estimators such as the Lasso (Tibshirani, 1996) and SLOPE (Bogdan *et al.*, 2015), as well as more traditional techniques such as logistic regression. A novel aspect of our presentation in Section 4 is that we formalise the connection between AMP and a broad class of convex optimisation problems, and then show how to systematically derive exact expressions for the asymptotic risk of estimators in GLMs. We expect that our general recipe can be applied to a wider class of GLMs than have been studied in the AMP literature to date.

To preview the statistical content in this tutorial and highlight some recurring themes, we now discuss two prototypical applications of AMP that form the basis of Sections 3 and 4 respectively. First, suppose that we wish to estimate an unknown signal $v \in \mathbb{R}^n$ based on an observation

$$A = \frac{\lambda}{n} vv^\top + W,$$

where $\lambda > 0$ is fixed and $W \in \mathbb{R}^{n \times n}$ is a symmetric Gaussian noise matrix. In this so-called spiked Wigner model (see Section 3.1 and the references therein), a popular and well-studied estimator of v is the leading eigenvector $\hat{\varphi}$ of A, which can be approximated via the power method, with iterates

$$v^{k+1} = \frac{Av^k}{\|Av^k\|}.$$

An AMP algorithm in this context can be interpreted as a generalised power method that produces a sequence of estimates $\hat{v}^k$ of v via iterative updates of the form

$$\hat{v}^k = g_k(v^k), \qquad v^{k+1} = A\hat{v}^k - b_k\hat{v}^{k-1}$$

for $k \in \mathbb{N}_0$, where we emphasise the following two characteristic features:

(i) Each "denoising" function $g_k : \mathbb{R} \to \mathbb{R}$ is applied componentwise to vectors, and can be chosen appropriately to exploit different types of prior information about the structure of v (e.g., to encourage $\hat{v}^k$ to be sparse).

(ii) In the "memory" term $-b_k\hat{v}^{k-1}$, which is called an "Onsager" correction in the AMP literature (e.g., Bayati and Montanari, 2011; Donoho *et al.*, 2009), the scalar b_k is defined as a specific function of v^k to ensure that the iterates v^{k+1} have desirable statistical properties; see (3.3).

One way to incorporate additional structural information on v into the spiked model is to assume that its entries are drawn independently from some prior distribution π on $\mathbb{R}$; for example, we can enforce sparsity through priors that place strictly positive mass at 0. Then under appropriate conditions, AMP theory guarantees that, for each k, the components of the estimate $\hat{v}^k$ have approximately the same empirical distribution as those of $g_k(\mu_k v + \sigma_k \xi)$; here, $\xi \sim N_n(0, I_n)$ is a "noise" vector that is independent of the signal $v \in \mathbb{R}^n$, and the "signal" and "noise" parameters $\mu_k \in \mathbb{R}$, $\sigma_k > 0$ are determined by a scalar state evolution recursion that depends on (g_k) and the prior distribution π; see (3.6). This distributional characterisation effectively reduces the analysis of the high-dimensional $\hat{v}^k$ to a much simpler univariate denoising problem, where the aim is to reconstruct $V \sim \pi$ based on a single corrupted observation of the form $\mu_k V + \sigma_k G$ with $G \sim N(0, 1)$ representing independent Gaussian noise. The functions g_k can then be chosen in such a way that the "effective signal-to-noise ratios" $(\mu_k/\sigma_k)^2$ are large and $g_k(\mu_k V + \sigma_k G)$ accurately estimates V. This ensures that the resulting AMP estimates $\hat{v}^k = g_k(v^k)$ have low asymptotic estimation error as $n \to \infty$.

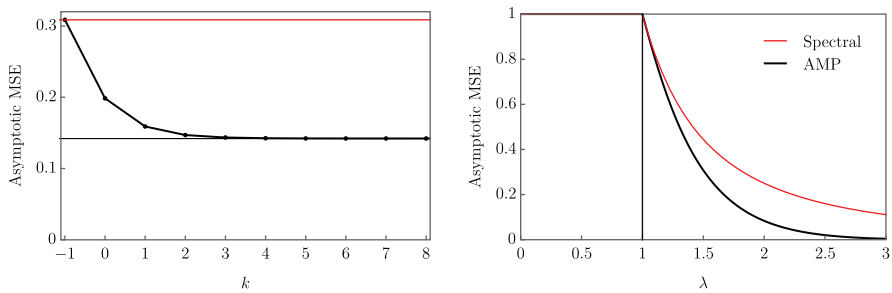

Figure 1.1: Asymptotic mean-squared error plots for estimation of a signal $v \in \mathbb{R}^n$ with i.i.d. $U\{-1,1\}$ entries in the rank-one spiked model, based on an AMP algorithm with denoising functions $g_k \colon x \mapsto \tanh(\mu_k x/\sigma_k^2)$ and spectral initialisation ($v^0 = \hat{\varphi}$ and $\hat{v}^{-1} = \lambda^{-1}\hat{\varphi}$ with $\|\hat{\varphi}\| = \sqrt{n\lambda^2(\lambda^2-1)_+}$). See Sections 3.2–3.3, where we also discuss how to consistently estimate λ when it is unknown (Remark 3.12).
Left: Plot of $\mathrm{AMSE}_k(\lambda) := \lim_{n\to\infty} \|\hat{v}^k - v\|^2/n$ against the iteration number k for the AMP estimates $\hat{v}^k \equiv \hat{v}^k_\lambda(n)$, when $\lambda = 1.7$. $\mathrm{AMSE}_k(\lambda)$ decreases monotonically to some $\mathrm{AMSE}_\infty(\lambda)$ as $k \to \infty$; see Theorem 3.10(c). *Right*: Plots of $\mathrm{AMSE}_{-1}(\lambda) = 1 \wedge \lambda^{-2}$ for the pilot spectral estimator $\hat{v}^{-1}$ and $\mathrm{AMSE}_\infty(\lambda)$ for AMP, with $\lambda \in [0,3]$. The spectral estimator undergoes the so-called BBP phase transition at $\lambda = 1$; see Section 3.1.

For instance, suppose that the entries of v are drawn uniformly at random from $\{-1,1\}$. Then provided we initialise the AMP algorithm with $v^0 = \hat{\varphi}$ and $\hat{v}^{-1} = \lambda^{-1}\hat{\varphi}$, where $\|\hat{\varphi}\| = \sqrt{n\lambda^2(\lambda^2-1)_+}$, it turns out that the asymptotic mean squared error (MSE) of $\hat{v}^k$ is minimised by choosing g_k to be the function $x \mapsto \tanh(\mu_k x/\sigma_k^2)$; see Section 3.3. Figure 1.1 illustrates that the limiting MSE of the AMP estimates $\hat{v}^k$ decreases with the iteration number k, and in particular that they improve on the pilot spectral estimator $\hat{v}^{-1}$ (which is agnostic to the structure of v).

As a second example, consider the linear model $y = X\beta + \varepsilon$, where $\beta \in \mathbb{R}^p$ is the target of inference, $\varepsilon \in \mathbb{R}^n$ is a noise vector, and $X \in \mathbb{R}^{n\times p}$ is a random design matrix with independent $N(0,1/n)$ entries. In high-dimensional regimes where p is comparable in magnitude to, or even much larger than n, a popular (sparse) estimator is the Lasso (Tibshirani, 1996), which for $\lambda > 0$ is defined by

$$\hat{\beta}^{\mathrm{L},\lambda} \in \operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|y - X\tilde{\beta}\|^2 + \lambda\|\tilde{\beta}\|_1 \right\}.$$

In the literature on high-dimensional estimation, upper bounds on the prediction and estimation error of the Lasso have been obtained under suitable conditions on the design matrix X, such as the restricted isometry property or compatibility conditions (e.g., Bühlmann and van de Geer, 2011). AMP offers complementary guarantees by providing exact formulae for the asymptotic risk in the "large system limit" where $n, p \to \infty$ with $n/p \to \delta \in (0, \infty)$, and with the components of β drawn independently from a prior distribution on $\mathbb{R}$. To motivate the form of the AMP algorithm in this setting, first consider the iterative soft thresholding algorithm (ISTA) for solving the Lasso optimisation problem, whose update steps can be written as

$$\hat{r}^k = y - X\hat{\beta}^k, \qquad \hat{\beta}^{k+1} = \mathrm{ST}_{\lambda\eta_k}(\hat{\beta}^k + \eta_k X^\top \hat{r}^k) \qquad \text{for } k \in \mathbb{N}_0; \tag{1.1}$$

here, $\hat{r}^k$ is the current residual, $\eta_k > 0$ is a deterministic step size, and for $t > 0$, the soft-thresholding function $\mathrm{ST}_t \colon w \mapsto \mathrm{sgn}(w)(|w| - t)_+$ is applied componentwise to vectors. This is an instance of the general-purpose proximal gradient method (Parikh and Boyd, 2013, Sections 4.2 and 4.3). An "accelerated" version of (1.1) called FISTA (Beck and Teboulle, 2009) bears a closer resemblance to an AMP algorithm, where the iterates of the latter are given by

$$\hat{r}^k = y - X\hat{\beta}^k + \frac{\|\hat{\beta}^k\|_0}{n}\,\hat{r}^{k-1}, \quad \hat{\beta}^{k+1} = \mathrm{ST}_{t_{k+1}}(\hat{\beta}^k + X^\top \hat{r}^k) \quad \text{for } k \in \mathbb{N}_0. \tag{1.2}$$

Here, each $t_k > 0$ is a deterministic threshold and $\|\hat{\beta}^k\|_0$ denotes the number of non-zero entries of $\hat{\beta}^k \in \mathbb{R}^p$. By comparison with (1.1), we observe that $\hat{r}^k$ in (1.2) is a corrected residual, whose definition includes an additional memory term that is crucial for ensuring that the empirical distribution of the iterates can be characterised exactly. Indeed, for each fixed $k \in \mathbb{N}$, the entries of the AMP estimate $\hat{\beta}^k$ of β have approximately the same empirical distribution as those of $\mathrm{ST}_{t_k}(\beta + \sigma_k \xi)$ when p is large; here $\xi \sim N_p(0, I_p)$ is a noise vector that is independent of β, the noise level $\sigma_k > 0$ is determined by the state evolution recursion defined in (4.41) below, and the scalar denoising function ST_{t_k} induces sparsity.

Bayati and Montanari (2012) proved that in the asymptotic regime above, the AMP iterates $(\hat{r}^k, \hat{\beta}^k)$ converge in a suitable sense to a fixed point $(\hat{r}^*, \hat{\beta}^*)$, and a key property of (1.2) is that for any such fixed point,

$\hat{\beta}^*$ is a Lasso solution; see (4.42) below. It follows that the performance of the Lasso is precisely characterised by a fixed point of the state evolution recursion (4.41); see Theorem 4.5. Since the above properties are proved under a Gaussian design, the main utility of AMP in this setting is not so much as an efficient Lasso computational algorithm, but rather as a device for gaining insight into the statistical properties of the estimator. In Section 4, the above theory is developed as part of an overarching AMP framework for linear models and generalised linear models (GLMs).

Note that in both of the examples above, the limiting empirical distributions of the entries of the AMP iterates can be decomposed into independent "signal" and "noise" components, and the effective signal strength and noise level are determined by a state evolution recursion. In Sections 3 and 4, we show how to derive these asymptotic guarantees by applying the master theorems in Section 2 to suitable abstract recursions, which track the evolution of the asymptotically Gaussian "noise" components of the AMP iterates. We discuss various extensions in Section 5, and provide proofs in the Appendix (Section A), with supplementary mathematical background deferred to Section B. As a guide to the reader, we remark that rigorous formulations of the results in this monograph require a number of technical conditions. While we take care to state these precisely, and discuss them at appropriate places, we emphasise that these should generally be regarded as mild. We therefore recommend that the reader initially focuses on the main conclusions of the results.

The statistical roots of AMP lie in compressed sensing (Donoho *et al.*, 2009, 2013). A reader approaching the subject from this perspective can consult Montanari (2012), Tramel *et al.* (2014) and Schniter (2020) for accessible expositions of the motivating ideas and the connections with message passing algorithms on dense graphs. Alternatively, for comprehensive reviews of AMP from a statistical physics perspective, see Zdeborová and Krzakala (2016), Krzakala *et al.* (2012) and Lesieur *et al.* (2017).

In spin glass theory, an AMP algorithm was proposed as an iterative scheme for solving the Thouless–Anderson–Palmer (TAP) equations

corresponding to a Sherrington–Kirkpatrick model with specific parameters (Bolthausen, 2014; Mézard and Montanari, 2009; Mézard *et al.*, 1987; Talagrand, 2011). The estimation problem here is equivalent to one of reconstructing a symmetric rank-one matrix in a Gaussian spiked model. Bolthausen (2014) proved a rigorous state evolution result for AMP in this specific setting, by introducing a conditioning argument that became an essential ingredient in subsequent analyses of AMP (Bayati and Montanari, 2011; Berthier *et al.*, 2020; Javanmard and Montanari, 2013; Fan, 2022). See Section A.2 for a detailed discussion of this proof technique.

In this tutorial, we restrict our focus to AMP recursions in which the random matrices are Gaussian. However, as we discuss in Section 5, several recent works have extended AMP and its state evolution recursion to more general non-Gaussian settings. For matrices with independent sub-Gaussian entries, results on the "universality" of AMP were first established by Bayati *et al.* (2015) and later in greater generality by Chen and Lam (2021). In addition, to accommodate the class of rotationally invariant random matrices, a number of extensions of the original AMP framework have recently been proposed, including Orthogonal AMP (Ma and Ping, 2017; Takeuchi, 2020) and Vector AMP (Schniter *et al.*, 2016; Rangan *et al.*, 2019b), as well as the general iterative schemes of Opper *et al.* (2016), Çakmak and Opper (2019) and Fan (2022). Some of these are closely related to expectation propagation (Opper and Winther, 2005; Kabashima and Vehkaperä, 2014). In all of the above variants of AMP, the recursion is tailored to the spectrum of the random matrix.

1.1 Notation and Preliminaries

Here, we introduce some notation used throughout this tutorial, and present basic properties of Wasserstein distances, pseudo-Lipschitz functions, as well as the complete convergence of random sequences.

General notation: For $n \in \mathbb{N}$, let $e_1, \dots, e_n$ be the standard basis vectors in $\mathbb{R}^n$. For $r \in [1, \infty]$, we write $\|x\|_r$ for the ℓ_r norm of $x \equiv (x_1, \dots, x_n) \in \mathbb{R}^n$, so that $\|x\|_r = (\sum_{i=1}^n |x_i|^r)^{1/r}$ when $r \in [1, \infty)$ and $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$. We also define $\|x\|_{n,r} := n^{-1/r}\|x\|_r = (n^{-1}\sum_{i=1}^n |x_i|^r)^{1/r}$ for $r \in (1, \infty)$. Let $\langle \cdot, \cdot \rangle$ and $\|\cdot\| := \|\cdot\|_2$ be the

standard Euclidean inner product and norm on $\mathbb{R}^n$ respectively, and define $\langle \cdot, \cdot \rangle_n$ to be the scaled Euclidean inner product on $\mathbb{R}^n$ given by $\langle x, y \rangle_n := n^{-1} \langle x, y \rangle$ for $x, y \in \mathbb{R}^n$, which induces the norm $\|\cdot\|_n := \|\cdot\|_{n,2}$. We denote by $\mathbf{1}_n := (1, \ldots, 1) \in \mathbb{R}^n$ the all-ones vector and write $\langle x \rangle_n := \langle x, \mathbf{1}_n \rangle_n = n^{-1} \sum_{i=1}^n x_i$ for each $x \in \mathbb{R}^n$.

For $D \in \mathbb{N}$ and $x^1, \ldots, x^D \in \mathbb{R}^n$, we denote by $\nu_n(x^1, \ldots, x^D) := n^{-1} \sum_{i=1}^n \delta_{(x_i^1, \ldots, x_i^D)}$ the joint empirical distribution of their components, and for a function $f \colon \mathbb{R}^D \to \mathbb{R}$, write $f(x^1, \ldots, x^D) := (f(x_i^1, \ldots, x_i^D) \colon 1 \leq i \leq n) \in \mathbb{R}^n$ for the row-wise application of f to $(x^1 \ \cdots \ x^D)$.

By a *Euclidean space* $(E, \|\cdot\|_E)$ we mean a finite-dimensional inner product space over $\mathbb{R}$, equipped with the norm induced by its inner product; examples include $(\mathbb{R}^n, \|\cdot\|)$ for $n \in \mathbb{N}$ and $(\mathbb{R}^{k \times \ell}, \|\cdot\|_{\mathrm{F}})$ for $k, \ell \in \mathbb{N}$, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm induced by the trace inner product $(A, B) \mapsto \mathrm{tr}(A^\top B)$.

Gaussian orthogonal ensemble: We write $W \sim \mathrm{GOE}(n)$ if $W = (W_{ij})_{1 \leq i,j \leq n}$ takes values in the space of all symmetric $n \times n$ matrices, and has the property that $(W_{ij})_{1 \leq i \leq j \leq n}$ are independent, with $W_{ij} \sim N(0, 1/n)$ for $1 \leq i < j \leq n$ and $W_{ii} \sim N(0, 2/n)$ for $i = 1, \ldots, n$. Writing $\mathbb{O}_n$ for the set of all $n \times n$ orthogonal matrices, we note the orthogonal invariance property of the $\mathrm{GOE}(n)$ distribution: if $Q \in \mathbb{O}_n$ and $W \sim \mathrm{GOE}(n)$, then $Q^\top W Q \sim \mathrm{GOE}(n)$.

Complete convergence of random sequences: The asymptotic results below are formulated in terms of the notion of *complete convergence* (e.g., Hsu and Robbins, 1947; Serfling, 1980, Chapter 1.3). This is a stronger mode of stochastic convergence than almost sure convergence, and is denoted throughout using the symbol $\xrightarrow{c}$. In Definition 1.1 and Proposition 1.2 below, we give two equivalent characterisations of complete convergence and introduce some associated stochastic O symbols.

Definition 1.1. Let (X_n) be a sequence of random elements taking values in a Euclidean space $(E, \|\cdot\|_E)$. We say that X_n *converges completely* to a deterministic limit $x \in E$, and write $X_n \xrightarrow{c} x$ or $\text{c-lim}_{n \to \infty} X_n = x$, if $Y_n \to x$ almost surely for any sequence of E-valued random elements (Y_n) with $Y_n \overset{d}{=} X_n$ for all n.

We write $X_n = o_c(1)$ if $X_n \xrightarrow{c} 0$, and write $X_n = O_c(1)$ if $Y_n = O_{a.s.}(1)$ (i.e., $\limsup_{n\to\infty} \|Y_n\|_E < \infty$ almost surely) for any sequence of E-valued random elements (Y_n) with $Y_n \overset{d}{=} X_n$ for all n.

Proposition 1.2. For a sequence (X_n) of random elements taking values in a Euclidean space $(E, \|\cdot\|_E)$, we have

(a) $X_n = o_c(1)$ if and only if $\sum_n \mathbb{P}(\|X_n\|_E > \varepsilon) < \infty$ for all $\varepsilon > 0$;

(b) $X_n = O_c(1)$ if and only if there exists $C > 0$ such that $\sum_n \mathbb{P}(\|X_n\|_E > C) < \infty$.

For a deterministic $x \in E$, we see that $X_n \xrightarrow{c} x$ if and only if $\sum_n \mathbb{P}(\|X_n - x\|_E > \varepsilon) < \infty$ for all $\varepsilon > 0$. Moreover, if $X_n \xrightarrow{c} x$, then $X_n = O_c(1)$. The proof of Proposition 1.2, along with various other properties of complete convergence and a calculus for $o_c(1)$ and $O_c(1)$ notation, is given in Section B.1; see also Remark A.1.

Wasserstein distances and pseudo-Lipschitz functions: For $D \in \mathbb{N}$ and $r \in [1, \infty)$, we write $\mathcal{P}(r) \equiv \mathcal{P}_D(r)$ for the set of all Borel probability measures P on $\mathbb{R}^D$ with $\int_{\mathbb{R}^D} \|x\|^r \, dP(x) < \infty$. For $P, Q \in \mathcal{P}_D(r)$, the *$r$-Wasserstein distance* between P and Q is defined by

$$d_r(P, Q) := \inf_{(X,Y)} \mathbb{E}(\|X - Y\|^r)^{1/r},$$

where the infimum is taken over all pairs of random vectors (X, Y) defined on a common probability space with $X \sim P$ and $Y \sim Q$. For $P, P_1, P_2, \ldots \in \mathcal{P}_D(r)$, we have $d_r(P_n, P) \to 0$ if and only if both $\int_{\mathbb{R}^D} \|x\|^r \, dP_n(x) \to \int_{\mathbb{R}^D} \|x\|^r \, dP(x)$ and $P_n \to P$ weakly (e.g., Villani, 2003, Theorem 7.12). Furthermore, for $L > 0$, we write $\mathrm{PL}_D(r, L)$ for the set of functions $\psi \colon \mathbb{R}^D \to \mathbb{R}$ such that

$$|\psi(x) - \psi(y)| \le L\|x - y\| \, (1 + \|x\|^{r-1} + \|y\|^{r-1}) \tag{1.3}$$

for all $x, y \in \mathbb{R}^D$, and denote by $\mathrm{PL}_D(r) := \bigcup_{L>0} \mathrm{PL}_D(r, L)$ the class of *pseudo-Lipschitz functions* $f \colon \mathbb{R}^D \to \mathbb{R}$ *of order* r. Note that $\mathrm{PL}_D(1, L)$ is precisely the class of all $(3L)$-Lipschitz functions on $\mathbb{R}^D$, and that $\mathrm{PL}_D(s) \subseteq \mathrm{PL}_D(r)$ for any $1 \le s \le r$. Moreover, for any probability measure $P \in \mathcal{P}_D(r)$, we have $|\int_{\mathbb{R}^d} \psi \, dP| \le L \int_{\mathbb{R}^D} (\|x\| + \|x\|^r) \, dP(x) +$

$|\psi(0)| < \infty$ for all $\psi \in \mathrm{PL}_D(r, L)$. Now for $P, Q \in \mathcal{P}_D(r)$, we define

$$\widetilde{d}_r(P, Q) := \sup_{\psi \in \mathrm{PL}_D(r,1)} \left| \int_{\mathbb{R}^D} \psi \, dP - \int_{\mathbb{R}^D} \psi \, dQ \right|. \tag{1.4}$$

In Section B.4, we show (among other things) that $\widetilde{d}_r, d_r$ are metrics on $\mathcal{P}_D(r)$ that induce the same topology (Remark B.18).

References

Advani, M. S., A. M. Saxe, and H. Sompolinsky (2020). "High-dimensional dynamics of generalization error in neural networks". *Neural Netw.* 132: 428–446.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. New Jersey: Wiley.

Albert, A. and J. A. Anderson (1984). "On the existence of maximum likelihood estimates in logistic regression models". *Biometrika*. 71: 1–10.

Aliprantis, C. D. and O. Burkinshaw (1998). *Principles of Real Analysis*. 3rd edn. San Diego: Academic Press.

Alon, N., M. Krivelevich, and B. Sudakov (1998). "Finding a large hidden clique in a random graph". *Random Struct. Algorithms*. 13: 457–466.

Anderson, G., A. Guionnet, and O. Zeitouni (2010). *An Introduction to Random Matrices*. Cambridge: Cambridge University Press.

Aubin, B., B. Loureiro, A. Maillard, F. Krzakala, and L. Zdeborová (2020). "The spiked matrix model with generative priors". *IEEE Trans. Inf. Theory*. 67: 1156–1181.

Aubin, B., A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová (2019). "The committee machine: Computational to statistical gaps in learning a two-layers neural network". *J. Stat. Mech. Theory Exp.* 124023.

Bai, Z. and J. Silverstein (2010). *Spectral Analysis of Large Dimensional Random Matrices*. 2nd edn. New York: Springer.

Baik, J., G. Ben Arous, and S. Péché (2005). "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". *Ann. Probab.* 33: 1643–1697.

Baik, J. and J. W. Silverstein (2006). "Eigenvalues of large sample covariance matrices of spiked population models". *J. Multivariate Anal.* 97: 1382–1408.

Bakhshizadeh, M., A. Maleki, and V. H. de la Pena (2020). "Sharp concentration results for heavy-tailed distributions". URL: https://arxiv.org/pdf/2003.13819.pdf.

Barata, J. C. A. and M. S. Hussein (2012). "The Moore–Penrose pseudoinverse: A tutorial review of the theory". *Braz. J. Phys.* 42: 146–165.

Barbier, J., M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová (2016). "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula". *Advances in Neural Information Processing Systems*. 29: 424–432.

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

Barbier, J., F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová (2019). "Optimal errors and phase transitions in high-dimensional generalized linear models". *Proc. Natl. Acad. Sci. U.S.A.* 116: 5451–5460.

Barbier, J., N. Macris, and C. Rush (2020). "All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation". URL: https://arxiv.org/pdf/2006.07971.pdf.

Bartlett, P. L., P. M. Long, G. Lugosi, and A. Tsigler (2020). "Benign overfitting in linear regression". *Proc. Natl. Acad. Sci. U.S.A.* 117: 30063–30070.

Bayati, M., M. Lelarge, and A. Montanari (2015). "Universality in polytope phase transitions and message passing algorithms". *Ann. Appl. Probab.* 25: 753–822.

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

Bayati, M. and A. Montanari (2012). "The LASSO risk for Gaussian matrices". *IEEE Trans. Inf. Theory.* 58: 1997–2017.

Beck, A. and M. Teboulle (2009). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". *SIAM J. Imaging Sci.* 2: 183–202.

Belkin, M., D. Hsu, S. Ma, and S. Mandal (2019). "Reconciling modern machine-learning practice and the classical bias-variance trade-off". *Proc. Natl. Acad. Sci. U.S.A.* 116: 15849–15854.

Belkin, M., D. Hsu, and J. Xu (2020). "Two models of double descent for weak features". *SIAM J. Math. Data Sci.* 2: 1167–1180.

Bellec, P. C., G. Lecué, and A. B. Tsybakov (2018). "SLOPE meets LASSO: Improved oracle bounds and optimality". *Ann. Statist.* 46: 3603–3642.

Benaych-Georges, F. and R. R. Nadakuditi (2011). "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". *Adv. Math.* 227: 494–521.

Berthier, R., A. Montanari, and P.-M. Nguyen (2020). "State evolution for approximate message passing with non-separable functions". *Inf. Inference.* 9: 33–79.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). "Latent Dirichlet allocation". *J. Mach. Learn. Res.* 3: 993–1022.

Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. Candès (2015). "SLOPE—Adaptive variable selection via convex optimization". *Ann. Appl. Stat.* 9: 1103–1140.

Bolthausen, E. (2014). "An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model". *Comm. Math. Phys.* 325: 333–366.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford: Oxford University Press.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). "Distributed optimization and statistical learning via the alternating direction method of multipliers". *Found. Trends Mach. Learn.* 3: 1–122.

Brown, L. D. and R. Purves (1973). "Measurable selections of extrema". *Ann. Statist.* 1: 902–912.

Bu, Z., J. Klusowski, C. Rush, and W. Su (2021). "Algorithmic analysis and statistical estimation of SLOPE via approximate message passing". *IEEE Trans. Inf. Theory.* 67: 506–537.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Berlin: Springer.

Çakmak, B. and M. Opper (2019). "Memory-free dynamics for the Thouless–Anderson–Palmer equations of Ising models with arbitrary rotation-invariant ensembles of random coupling matrices". *Phys. Rev. E.* 99: 062140.

Candès, E. J. and B. Recht (2009). "Exact matrix completion via convex optimization". *Found. Comput. Math.* 9: 717–772.

Candès, E. J. and P. Sur (2020). "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression". *Ann. Statist.* 48: 27–42.

Capitaine, M., C. Donati-Martin, and D. Féral (2009). "The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations". *Ann. Probab.* 37: 1–47.

Celentano, M. and A. Montanari (2022). "Fundamental barriers to high-dimensional regression with convex penalties". *Ann. Statist.* 50: 170–196.

Celentano, M., A. Montanari, and Y. Wu (2020). "The estimation error of general first order methods". *Proc. Mach. Learn. Res.* 125: 1–64.

Chen, W.-K. and W.-K. Lam (2021). "Universality of approximate message passing algorithms". *Electron. J. Probab.* 26: 1–44.

Dar, Y., V. Muthukumar, and R. Baraniuk (2021). "A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning". URL: https://arxiv.org/pdf/2109.02355.pdf.

d'Ascoli, S., M. Refinetti, G. Biroli, and F. Krzakala (2020). "Double trouble in double descent: Bias and variance(s) in the lazy regime". *Proc. Mach. Learn. Res.* 119: 2280–2290.

Deng, Z., A. Kammoun, and C. Thrampoulidis (2019). "A model of double descent for high-dimensional binary linear classification". URL: https://arxiv.org/pdf/1911.05822.pdf.

Deshpande, Y., E. Abbe, and A. Montanari (2016). "Asymptotic mutual information for the balanced binary stochastic block model". *Inf. Inference.* 6: 125–170.

Deshpande, Y. and A. Montanari (2014). "Information-theoretically optimal sparse PCA". In: *2014 IEEE International Symposium on Information Theory.* 2197–2201.

Deshpande, Y. and A. Montanari (2015). "Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time". *Found. Comput. Math.* 15: 1069–1128.

Donoho, D. L., A. Javanmard, and A. Montanari (2013). "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing". *IEEE Trans. Inf. Theory.* 59: 7434–7464.

Donoho, D. L. and I. M. Johnstone (1994). "Minimax risk over l_p balls for l_q error". *Prob. Theory Related Fields.* 99: 277–303.

Donoho, D. L. and I. M. Johnstone (1998). "Minimax estimation via wavelet shrinkage". *Ann. Statist.* 26: 879–921.

Donoho, D. L., A. Maleki, and A. Montanari (2009). "Message-passing algorithms for compressed sensing". *Proc. Natl. Acad. Sci. U.S.A.* 106: 18914–18919.

Donoho, D. and A. Montanari (2015). "Variance breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$". URL: https://arxiv.org/pdf/1503.02106.pdf.

Donoho, D. and A. Montanari (2016). "High dimensional robust M-estimation: Asymptotic variance via approximate message passing". *Probab. Theory Related Fields.* 166: 935–969.

Dudley, R. M. (2002). *Real Analysis and Probability.* 2nd edn. Cambridge: Cambridge University Press.

Dümbgen, L., R. J. Samworth, and J. A. Wellner (2021). "Bounding distributional errors via density ratios". *Bernoulli.* 27: 818–852.

Dümbgen, L., R. Samworth, and D. Schuhmacher (2011). "Approximation by log-concave distributions, with applications to regression". *Ann. Statist.* 39: 702–730.

Efron, B. (2011). "Tweedie's formula and selection bias". *J. Amer. Statist. Assoc.* 106: 1602–1614.

El Alaoui, A., A. Ramdas, F. Krzakala, Zdeborová, and M. I. Jordan (2018). "Decoding from pooled data: Phase transitions of message passing". *IEEE Trans. Inf. Theory.* 65: 572–585.

Emami, M., M. Sahraee-Ardakan, P. Pandit, S. Rangan, and A. K. Fletcher (2020). "Generalization error of generalized linear models in high dimensions". *Proc. Mach. Learn. Res.* 119: 2892–2901.

Fan, Z. (2022). "Approximate message passing algorithms for rotationally invariant matrices". *Ann. Statist.* 50: 197–224.

Federer, H. (1996). *Geometric Measure Theory.* New York: Springer-Verlag.

Féral, D. and S. Péché (2007). "The largest eigenvalue of rank one deformation of large Wigner matrices". *Comm. Math. Phys.* 272: 185–228.

Fletcher, A. K. and S. Rangan (2014). "Scalable inference for neuronal connectivity from calcium imaging". *Advances in Neural Information Processing Systems.* 27: 2843–2851.

Fletcher, A. K., S. Rangan, and P. Schniter (2018). "Inference in deep networks in high dimensions". In: *2018 IEEE International Symposium on Information Theory.* 1884–1888.

Fourdrinier, D., W. E. Strawderman, and M. T. Wells (2018). *Shrinkage Estimation.* New York: Springer.

Gataric, M., T. Wang, and R. J. Samworth (2020). "Sparse principal component analysis via axis-aligned random projections". *J. Roy. Statist. Soc., Ser. B.* 82: 329–359.

Geiger, M., A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart (2019). "Scaling description of generalization with number of parameters in deep learning". URL: https://arxiv.org/pdf/1901.01608.pdf.

Gerbelot, C., A. Abbara, and F. Krzakala (2020a). "Asymptotic errors for convex penalized linear regression beyond Gaussian matrices". *Proc. Mach. Learn. Res.* 125: 1682–1713.

Gerbelot, C., A. Abbara, and F. Krzakala (2020b). "Asymptotic errors for teacher-student convex generalized linear models (or: How to prove Kabashima's replica formula)". URL: https://arxiv.org/pdf/2006.06581.pdf.

Gordon, L. (1994). "A stochastic approach to the gamma function". *Am. Math. Mon.* 101: 858–865.

Guo, D. and S. Verdú (2005). "Randomly spread CDMA: Asymptotics via statistical physics". *IEEE Trans. Inf. Theory.* 51: 1983–2010.

Han, Q. (2022). "Noisy linear inverse problems under convex constraints: Exact risk asymptotics in high dimensions". URL: https://arxiv.org/pdf/2201.08435.pdf.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). "Surprises in high-dimensional ridgeless least squares interpolation". *Ann. Statist.* 50: 949–986.

Hopkins, S. B., J. Shi, and D. Steurer (2015). "Tensor principal component analysis via sum-of-square proofs". *Proc. Mach. Learn. Res.* 40: 956–1006.

Hsu, P. L. and H. Robbins (1947). "Complete convergence and the law of large numbers". *Proc. Natl. Acad. Sci. U.S.A.* 33: 25–31.

Huber, P. J. (1964). "Robust estimation of a location parameter". *Ann. Math. Statist.* 35: 73–101.

Huber, P. J. (1973). "Robust regression: Asymptotics, conjectures and Monte Carlo". *Ann. Statist.* 1: 799–821.

Huber, P. J. and E. Ronchetti (2009). *Robust Statistics*. 2nd edn. New York: Wiley.

Javanmard, A. and A. Montanari (2013). "State evolution for general approximate message passing algorithms, with applications to spatial coupling". *Inf. Inference.* 2: 115–144.

Jeon, C., R. Ghods, A. Maleki, and C. Studer (2015). "Optimality of large MIMO detection via approximate message passing". In: *2015 IEEE International Symposium on Information Theory.* 1227–1231.

Johnstone, I. M. (2006). "High dimensional statistical inference and random matrices". In: *Proceedings of the International Congress of Mathematicians, Madrid 2006.* 307–333.

Johnstone, I. M. and A. Y. Lu (2009). "On consistency and sparsity for principal components analysis in high dimensions". *J. Amer. Statist. Assoc.* 104: 682–693.

Johnstone, I. M. and D. Paul (2018). "PCA in high dimensions: An orientation". *Proc. IEEE*. 106: 1277–1292.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). "A modified principal component technique based on the LASSO". *J. Comput. Graph. Statist.* 12: 531–547.

Kabashima, Y., F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová (2016). "Phase transitions and sample complexity in Bayes optimal matrix factorization". *IEEE Trans. Inf. Theory*. 62: 4228–4265.

Kabashima, Y. and M. Vehkaperä (2014). "Signal recovery using expectation consistent approximation for linear observations". In: *2014 IEEE International Symposium on Information Theory*. 226–230.

Kallenberg, O. (1997). *Foundations of Modern Probability*. New York: Springer–Verlag.

Kini, G. R. and C. Thrampoulidis (2020). "Analytic study of double descent in binary classification: The impact of loss". In: *2020 IEEE International Symposium on Information Theory*. 2527–2532.

Knowles, A. and J. Yin (2013). "The isotropic semicircle law and deformation of Wigner matrices". *Comm. Pure Appl. Math.* 66: 1663–1749.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, Massachusetts: MIT Press.

Krzakala, F., M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová (2012). "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices". *J. Stat. Mech. Theory Exp.* P08009.

Kuchibhotla, A. and A. Chakrabortty (2018). "Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression". URL: https://arxiv.org/pdf/1804.02605.pdf.

Lelarge, M. and L. Miolane (2019). "Fundamental limits of symmetric low-rank matrix estimation". *Probab. Theory Related Fields*. 173: 859–929.

Lesieur, T., F. Krzakala, and L. Zdeborová (2017). "Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications". *J. Stat. Mech. Theory Exp.* 073403.

Li, Y. and Y. Wei (2021). "Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent". URL: https://arxiv.org/pdf/2110.09502.pdf.

Liang, T. and A. Rakhlin (2020). "Just interpolate: Kernel 'ridgeless' regression can generalize". *Ann. Statist.* 48: 1329–1347.

Liang, T. and P. Sur (2022). "A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers". *Ann. Statist.* to appear.

Liu, L., S. Huang, and B. M. Kurkoski (2021). "Memory approximate message passing". In: *2021 IEEE International Symposium on Information Theory*. 1379–1384.

Ma, J. and L. Ping (2017). "Orthogonal AMP". *IEEE Access*. 5: 2020–2033.

Ma, J., J. Xu, and A. Maleki (2019a). "Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2 regularization". *IEEE Trans. Inf. Theory*. 65: 3600–3629.

Ma, J., J. Xu, and A. Maleki (2021). "Impact of the sensing spectrum on signal recovery in generalized linear models". URL: https://arxiv.org/pdf/2111.03237.pdf.

Ma, Y., C. Rush, and D. Baron (2019b). "Analysis of approximate message passing with non-separable denoisers and Markov random field priors". *IEEE Trans. Inf. Theory*. 65: 7367–7389.

Ma, Z. and Y. Wu (2015). "Computational barriers in minimax submatrix detection". *Ann. Statist.* 43: 1089–1116.

Manoel, A., F. Krzakala, M. Mézard, and L. Zdeborová (2017). "Multi-layer generalized linear estimation". In: *2017 IEEE International Symposium on Information Theory*. 2098–2102.

Matsushita, R. and T. Tanaka (2013). "Low-rank matrix reconstruction and clustering via approximate message passing". *Advances in Neural Information Processing Systems*. 26: 917–925.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. 2nd edn. Boca Raton: Chapman & Hall/CRC.

Mehta, M. L. (2004). *Random Matrices*. 3rd edn. San Diego: Elsevier.

Mei, S. and A. Montanari (2020). "The generalization error of random features regression: Precise asymptotics and double descent curve". *Comm. Pure Appl. Math.* 75: 667–766.

Metzler, C., A. Mousavi, and R. Baraniuk (2017). "Learned D-AMP: Principled neural network based compressive image recovery". *Advances in Neural Information Processing Systems*. 30: 1772–1783.

Mézard, M. and M. Montanari (2009). *Information, Physics, and Computation*. Oxford: Oxford University Press.

Mézard, M., G. Parisi, and M. A. Virasoro (1987). *Spin Glass Theory and Beyond*. Vol. 9. World Scientific Lecture Notes in Physics.

Miolane, L. and A. Montanari (2021). "The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning". *Ann. Statist.* 49: 2313–2335.

Mondelli, M., C. Thrampoulidis, and R. Venkataramanan (2021). "Optimal combination of linear and spectral estimators for generalized linear models". *Found. Comput. Math.* 2021.

Mondelli, M. and R. Venkataramanan (2020). "Approximate message passing with spectral initialization for generalized linear models". *Proc. Mach. Learn. Res.* 130: 397–405.

Mondelli, M. and R. Venkataramanan (2021). "PCA initialization for approximate message passing in rotationally invariant models". URL: https://arxiv.org/pdf/2106.02356.pdf.

Montanari, A. (2012). "Graphical models concepts in compressed sensing". In: *Compressed Sensing: Theory and Applications*. Ed. by Y. Eldar and G. Kutyniok. Cambridge: Cambridge University Press.

Montanari, A. and E. Richard (2014). "A statistical model for tensor PCA". *Advances in Neural Information Processing Systems*. 27: 2897–2905.

Montanari, A. and E. Richard (2016). "Non-negative principal component analysis: Message passing algorithms and sharp asymptotics". *IEEE Trans. Inf. Theory*. 62: 1458–1484.

Montanari, A. and R. Venkataramanan (2021). "Estimation of low-rank matrices via approximate message passing". *Ann. Statist.* 49: 321–345.

Mousavi, A., A. Maleki, and R. G. Baraniuk (2018). "Consistent parameter estimation for LASSO and approximate message passing". *Ann. Statist.* 46: 119–148.

Nakkiran, P., G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever (2021). "Deep double descent: Where bigger models and more data hurt". *J. Stat. Mech. Theory Exp.* 124003.

Opper, M., B. Çakmak, and O. Winther (2016). "A theory of solving TAP equations for Ising models with general invariant random matrices". *J. Phys. A.* 49: 114002.

Opper, M. and O. Winther (2005). "Expectation consistent approximate inference". *J. Mach. Learn. Res.* 6: 2177–2204.

Pace, L. and A. Salvan (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific.

Panaretos, V. M. and Y. Zemel (2020). *An Invitation to Statistics in Wasserstein Space*. New York: Springer–Verlag.

Pandit, P., M. Sahraee-Ardakan, S. Rangan, P. Schniter, and A. K. Fletcher (2020). "Inference with deep generative priors in high dimensions". *IEEE J. Sel. Areas Inf. Theory.* 1: 336–347.

Parikh, N. and S. Boyd (2013). "Proximal algorithms". *Found. Trends Optim.* 1: 123–231.

Parker, J. T., P. Schniter, and V. Cevher (2014a). "Bilinear generalized approximate message passing—Part I: Derivation". *IEEE Trans. Signal Process.* 62: 5839–5853.

Parker, J. T., P. Schniter, and V. Cevher (2014b). "Bilinear generalized approximate message passing—Part II: Applications". *IEEE Trans. Signal Process.* 62: 5854–5867.

Paul, D. (2007). "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model". *Statist. Sinica.* 17: 1617–1642.

Peng, M. (2012). "Eigenvalues of deformed random matrices". URL: https://arxiv.org/pdf/1205.0572.pdf.

Perry, A., A. S. Wein, A. S. Bandeira, and A. Moitra (2018). "Optimality and sub-optimality of PCA I: Spiked random matrix models". *Ann. Statist.* 46: 2416–2451.

Portnoy, S. (1984). "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency". *Ann. Statist.* 12: 1298–1309.

Portnoy, S. (1985). "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal approximation". *Ann. Statist.* 13: 1403–1417.

Portnoy, S. (1988). "Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity". *Ann. Statist.* 16: 356–366.

Prékopa, A. (1980). "Logarithmic concave measures and related topics". In: *Stochastic Programming (Proc. Internat. Conf., Univ. Oxford, Oxford, 1974, M. A. H. Dempster ed.)* London: Academic Press. 63–82.

Rangan, S. (2011). "Generalized approximate message passing for estimation with random linear mixing". In: *2011 IEEE International Symposium on Information Theory*. 2168–2172.

Rangan, S. and A. K. Fletcher (2012). "Iterative estimation of constrained rank-one matrices in noise". In: *2012 IEEE International Symposium on Information Theory*. 1246–1250.

Rangan, S. and A. K. Fletcher (2018). "Iterative reconstruction of rank-one matrices in noise". *Inf. Inference*. 7: 1246–1250.

Rangan, S., A. K. Fletcher, and V. K. Goyal (2009). "Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing". *Advances in Neural Information Processing Systems*. 22: 1545–1553.

Rangan, S., P. Schniter, and A. K. Fletcher (2019b). "Vector approximate message passing". *IEEE Trans. Inf. Theory*. 65: 6664–6684.

Rangan, S., P. Schniter, A. K. Fletcher, and S. Sarkar (2019a). "On the convergence of approximate message passing with arbitrary matrices". *IEEE Trans. Inf. Theory*. 65: 5339–5351.

Rangan, S., P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher (2016). "Fixed points of generalized approximate message passing with arbitrary matrices". *IEEE Trans. Inf. Theory*. 62: 7464–7474.

Reeves, G. and H. D. Pfister (2019). "The replica-symmetric prediction for random linear estimation with Gaussian matrices is exact". *IEEE Trans. Inf. Theory*. 65: 2252–2283.

Robbins, H. (1956). "An empirical Bayes approach to statistics". *Proc. Third Berkeley Symp. Math. Statist. Prob.* 1: 157–163.

Rockafellar, R. T. (1997). *Convex Analysis*. Princeton: Princeton University Press.

Rush, C., A. Greig, and R. Venkataramanan (2017). "Capacity-achieving sparse superposition codes via approximate message passing decoding". *IEEE Trans. Inf. Theory.* 63: 1476–1500.

Rush, C. and R. Venkataramanan (2018). "Finite sample analysis of approximate message passing algorithms". *IEEE Trans. Inf. Theory.* 64: 7264–7286.

Schniter, P. (2011). "A message-passing receiver for BICM-OFDM over unknown clustered-sparse channels". *IEEE J. Sel. Top. Signal Process.* 5: 1462–1474.

Schniter, P. (2020). "A simple derivation of AMP and its state evolution via first-order cancellation". *IEEE Trans. Signal Process.* 68: 4283–4292.

Schniter, P. and S. Rangan (2014). "Compressive phase retrieval via generalized approximate message passing". *IEEE Trans. Signal Process.* 63: 1043–1055.

Schniter, P., S. Rangan, and A. K. Fletcher (2016). "Vector approximate message passing for the generalized linear model". In: *50th Asilomar Conference on Signals, Systems and Computers*. 1525–1529.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Su, W., M. Bogdan, and E. Candès (2017). "False discoveries occur early on the LASSO path". *Ann. Statist.* 45: 2133–2150.

Su, W. and E. Candès (2016). "SLOPE is adaptive to unknown sparsity and asymptotically minimax". *Ann. Statist.* 44: 1038–1068.

Su, X. and T. M. Khoshgoftaar (2009). "A survey of collaborative filtering techniques". *Adv. Artif. Intelligence.* 2009: 1–19.

Sur, P. and E. J. Candès (2019a). "A modern maximum-likelihood theory for high-dimensional logistic regression". *Proc. Natl. Acad. Sci. U.S.A.* 116: 14516–14525.

Sur, P. and E. J. Candès (2019b). "Additional supplementary materials for 'A modern maximum-likelihood theory for high-dimensional logistic regression'". URL: https://sites.fas.harvard.edu/~prs499/papers/proofs_LogisticAMP.pdf.

Sur, P., Y. Chen, and E. J. Candès (2017). "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square". *Probab. Theory Related Fields*. 175: 487–558.

Takeuchi, K. (2021a). "Bayes-optimal convolutional AMP". *IEEE Trans. Inf. Theory*. 67: 4405–4428.

Takeuchi, K. (2021b). "On the convergence of orthogonal/vector AMP: Long-memory message-passing strategy". URL: https://arxiv.org/pdf/2111.05522.pdf.

Takeuchi, K. (2020). "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements". *IEEE Trans. Inf. Theory*. 66: 368–386.

Talagrand, M. (2011). *Mean Field Models for Spin Glasses, Vol I: Basic Examples*. New York: Springer.

Tanaka, T. (2002). "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors". *IEEE Trans. Inf. Theory*. 48: 2888–2910.

Thrampoulidis, C., E. Abbasi, and B. Hassibi (2018). "Precise error analysis of regularized M-estimators in high dimensions". *IEEE Trans. Inf. Theory*. 64: 5592–5628.

Thrampoulidis, C., S. Oymak, and B. Hassibi (2015). "Regularized linear regression: A precise analysis of the estimation error". *Proc. Mach. Learn. Res.* 40: 1683–1709.

Tian, F., L. Liu, and X. Chen (2021). "Generalized memory approximate message passing". URL: https://arxiv.org/pdf/2110.06069.pdf.

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso". *J. Roy. Statist. Soc., Ser. B*. 58: 267–288.

Tramel, E. W., S. Kumar, A. Giurgiu, and A. Montanari (2014). "Statistical estimation: From denoising to sparse regression and hidden cliques". URL: https://arxiv.org/pdf/1409.5557.pdf.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer–Verlag.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Venkataramanan, R., K. Kögler, and M. Mondelli (2021). "Estimation in rotationally invariant generalized linear models via approximate message passing". URL: https://arxiv.org/pdf/2112.04330.pdf.

Vila, J., P. Schniter, and J. Meola (2015). "Hyperspectral unmixing via turbo bilinear approximate message passing". *IEEE Trans. Comput. Imaging*. 1: 143–158.

Villani, C. (2003). *Topics in Optimal Transportation. Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society.

Villani, C. (2009). *Optimal Transport, Old and New*. New York: Springer–Verlag.

von Luxburg, U. (2007). "A tutorial on spectral clustering". *Statist. Comput.* 17: 395–416.

Vu, V. Q. and J. Lei (2013). "Minimax sparse principal subspace estimation in high dimensions". *Ann. Statist.* 41: 2905–2947.

Wang, T., Q. Berthet, and R. J. Samworth (2016). "Statistical and computational trade-offs in estimation of sparse principal components". *Ann. Statist.* 44: 1896–1930.

Wein, A. S., A. El Alaoui, and C. Moore (2019). "The Kikuchi hierarchy and tensor PCA". In: *2019 IEEE Annual Symposium on Foundations of Computer Science*. 1446–1468.

Yang, G. (2019). "Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation". URL: https://arxiv.org/pdf/1902.04760.pdf.

Zdeborová, L. and F. Krzakala (2016). "Statistical physics of inference: Thresholds and algorithms". *Adv. Phys.* 65: 453–552.

Zhong, X., T. Wang, and Z. Fan (2021). "Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization". URL: https://arxiv.org/pdf/2110.02318.pdf.

Zhu, Z., T. Wang, and R. J. Samworth (2019). "High-dimensional principal component analysis with heterogeneous missingness". URL: https://arxiv.org/pdf/1906.12125.pdf.

Zou, H., T. Hastie, and R. Tibshirani (2006). "Sparse principal component analysis". *J. Comput. Graph. Statist.* 15: 265–286.