ARE GENERATIVE MODELS THE NEW SPARSITY?

Lenka Zdeborová

(CNRS & CEA Saclay, France)

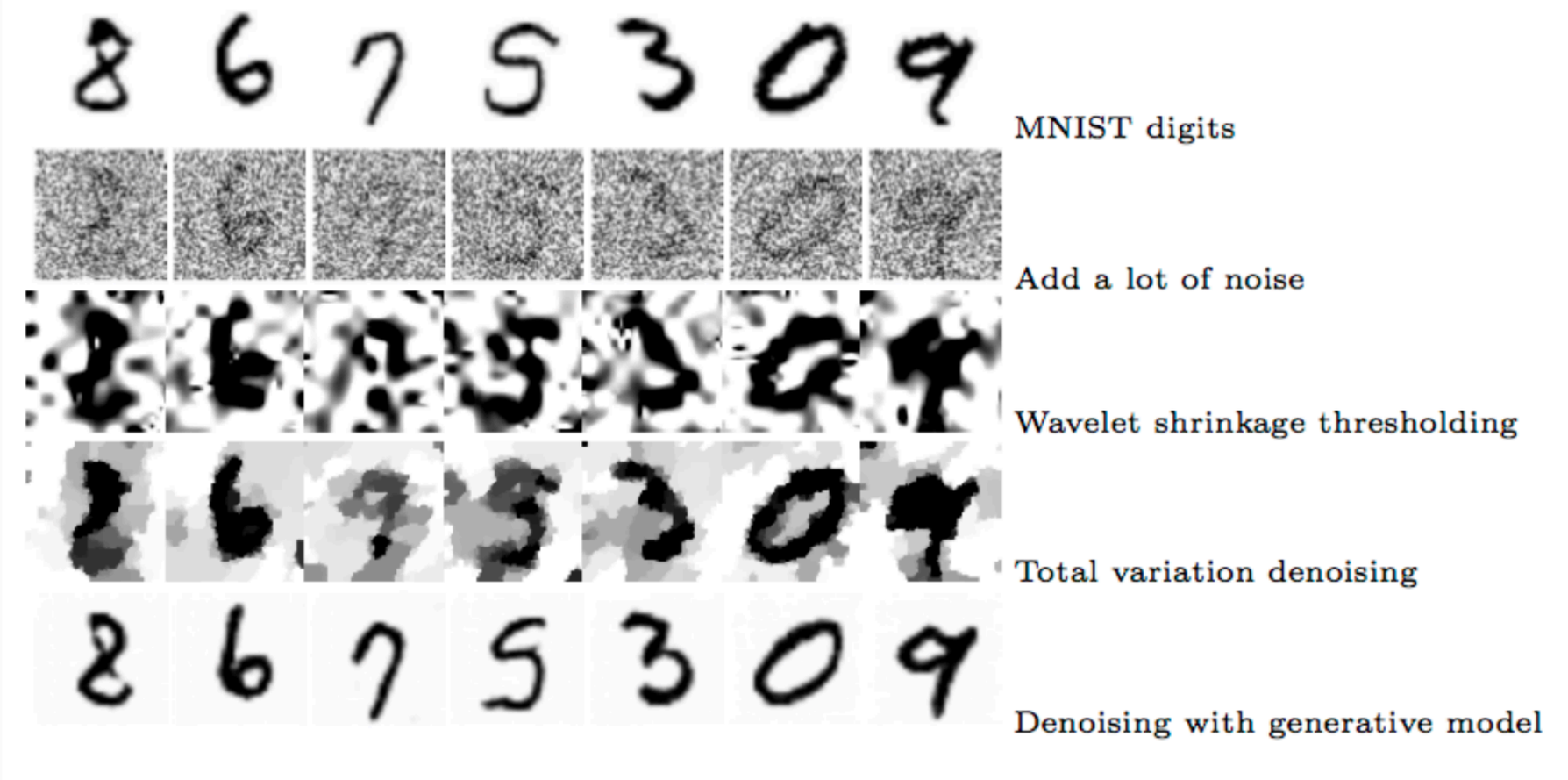
with B. Aubin, B. Loureiro, A. Maillard, F. Krzakala;

SPARS 2019, Toulouse, 1-4 July

Generative models are the new sparsity?

Mar 28, 2018

Posted with : Data science, Data science, Data science



Blog by Soledad Villar, about arXiv:1803.09319
“SUNLayer: Stable denoising with generative networks”

OUTLINE

- I. Introduction and motivation.
- II. Sparse PCA: Reminder of key facts and presentation of the methodology.
- III. Spiked matrix estimation with generative priors: Main results, and two take home messages.

RECOVER SIGNAL V FROM OBSERVATIONS Y

$$y = \Gamma(v) \qquad v \in \mathbb{R}^p$$

observation

noisy transformation

signal

RECOVER SIGNAL v FROM OBSERVATIONS y

$$y = \Gamma(v) \qquad v \in \mathbb{R}^p$$

observation — noisy transformation — signal

Simple examples

noise $\mathcal{N}(0,\Delta)$

- Denoising: $\Gamma(v) = v + \xi$
- Linear inverse problem: $\Gamma(v) = Av + \xi, \qquad A \in \mathbb{R}^{n \times p}$
- Spiked matrix estimation: $\Gamma(v) = vv^T + \xi$

RECOVER SIGNAL v FROM OBSERVATIONS Y

$$y = \Gamma(v) \qquad v \in \mathbb{R}^p$$

observation — noisy transformation — signal

Simple examples

noise $\mathcal{N}(0,\Delta)$

- Denoising: $\Gamma(v) = v + \xi$
- Linear inverse problem: $\Gamma(v) = Av + \xi, \qquad A \in \mathbb{R}^{n \times p}$
- Spiked matrix estimation: $\Gamma(v) = vv^T + \xi$

Basic paradigm of signal processing: Structure in v serves for recovery with better accuracy, larger noise, smaller n, etc.

SPARSITY

$$y = \Gamma(v)$$

There exists a basis W in which the signal v is sparse.

$$v = Wx \qquad \ell_0(x) = k \ll p$$

$$x \in \mathbb{R}^p$$
$$v \in \mathbb{R}^p$$

Simple examples

noise $\mathcal{N}(0,\Delta)$

- Denoising: $\Gamma(v) = v + \xi$
- Compressed sensing: $\Gamma(v) = Av + \xi, \quad A \in \mathbb{R}^{n \times p}$
- Sparse PCA: $\Gamma(v) = vv^T + \xi$

SPARSE CODING
DICTIONARY LEARNING

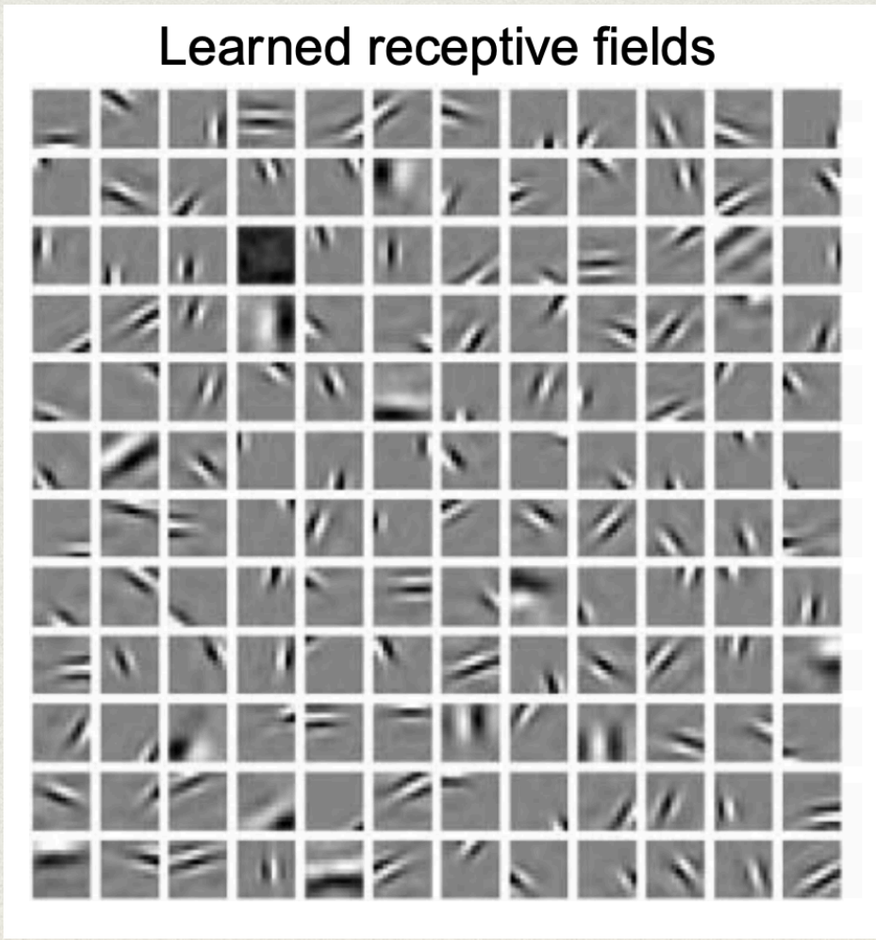
Basis in which the signal is sparse can be learned from examples.

Sparse coding:

Having M examples of signals, learn W so that x is sparse.

$$v_\mu = W x_\mu + \xi$$

$$\mu = 1, \ldots, M$$

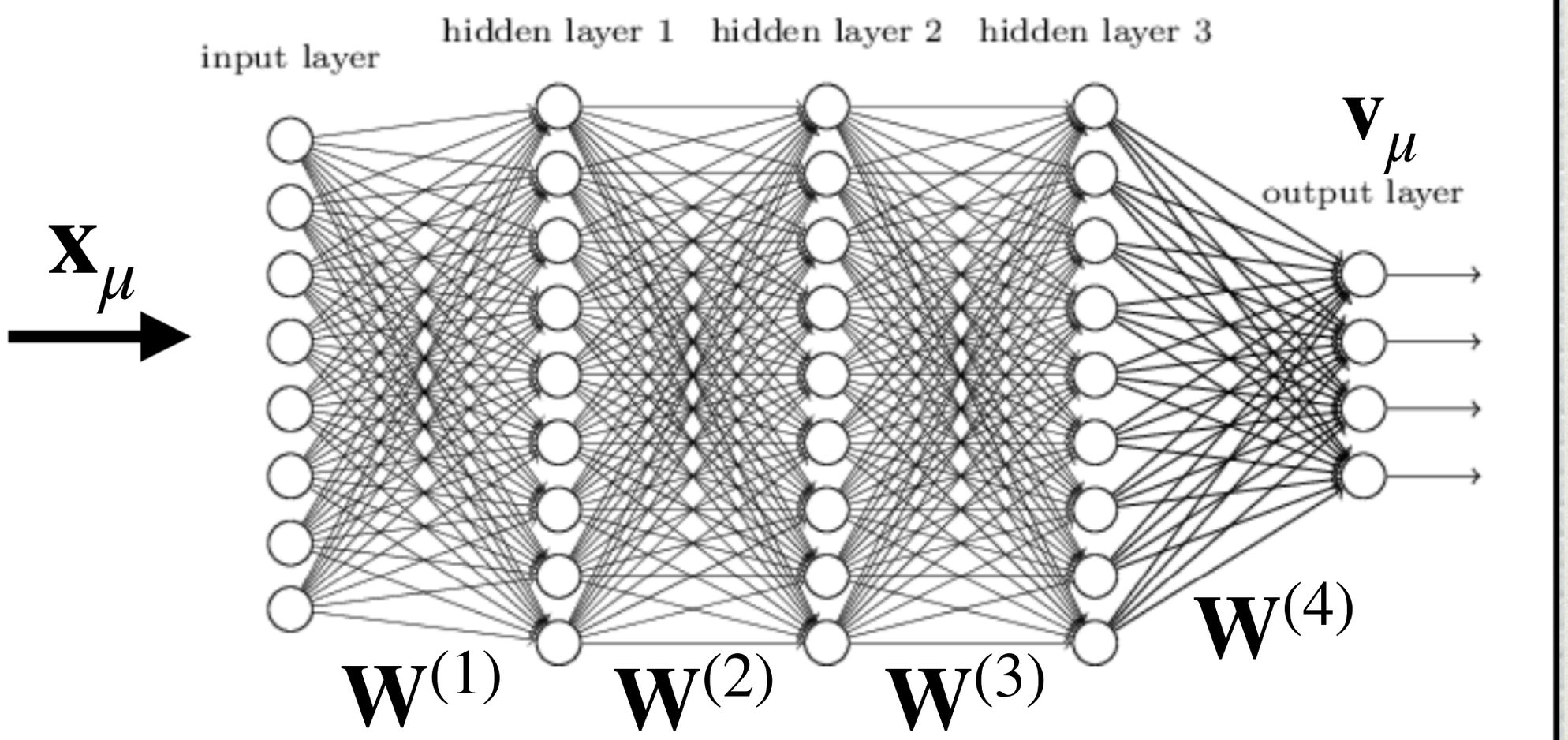


Olshausen, Field'97

LEARNING FROM EXAMPLES

- Generative neural networks (autoencoders, GANs, ...):

$$v_\mu = \varphi^{(4)}(\mathbf{W}^{(4)}\varphi^{(3)}(\mathbf{W}^{(3)}\varphi^{(2)}(\mathbf{W}^{(2)}\varphi^{(1)}(\mathbf{W}^{(1)}\mathbf{x}_\mu))))$$



THIS PERSON DOES NOT EXIST!

GANs generated people.

NVIDIA research

GENERATIVE MODELS AS PRIORS

$$y = \Gamma(v)$$

Recover signal v from observations y, knowing that:

- Sparsity: v is k-sparse.
- A generative model learned from data: There exists $x \in \mathbb{R}^k$ such that

$$v = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}x))))$$

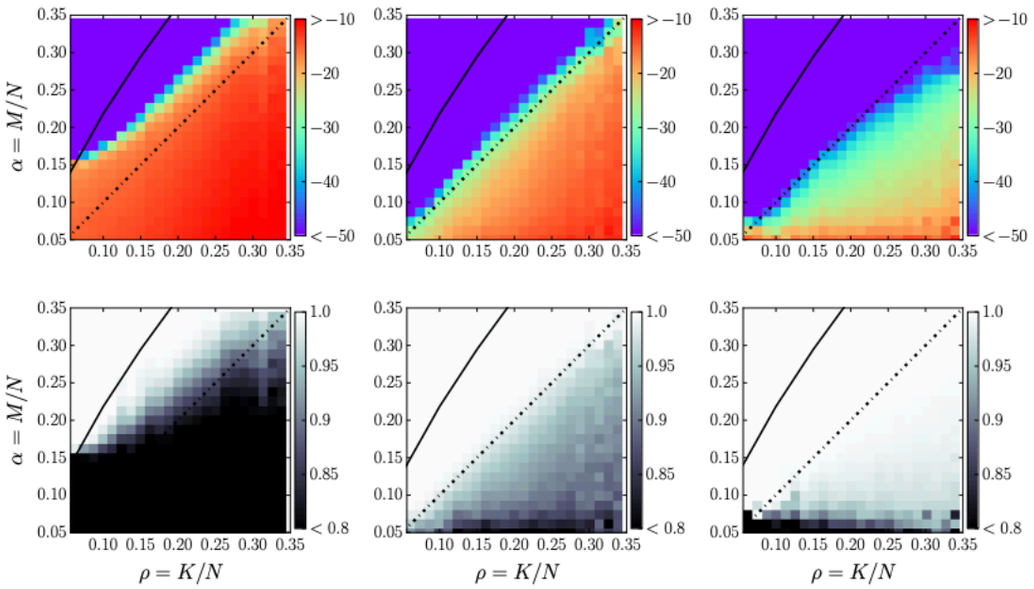
$$\varphi^{(i)}, W^{(i)}, i = 1,\ldots,L \quad \text{known, after training}$$

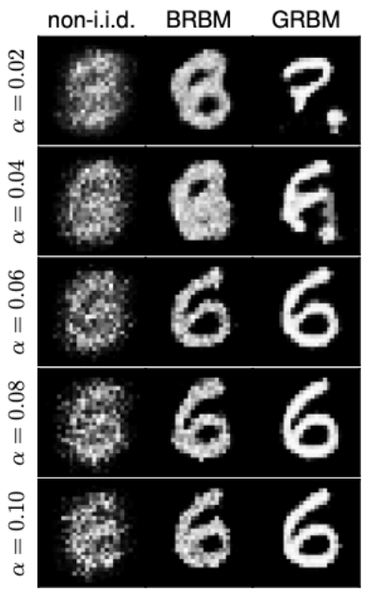
SELECTION OF EXISTING WORKS

Inferring Sparsity: Compressed Sensing using Generalized Restricted Boltzmann Machines

Eric W. Tramel[†], Andre Manoel[†], Francesco Caltagirone[‡], Marylou Gabrié[†] and Florent Krzakala[†§]
[†]Laboratoire de Physique Statistique (CNRS UMR-8550),
École Normale Supérieure, PSL Research University, 24 rue Lhomond, 75005 Paris, France
[§]Université Pierre et Marie Curie, Sorbonne Universités, 75005 Paris, France
[‡]INRIA Paris, 2 rue Simone Iff, 75012 Paris, France

arXiv:1606.03956





SELECTION OF EXISTING WORKS

Semantic Image Inpainting with Deep Generative Models

arXiv:1607.07539

Raymond A. Yeh,* Chen Chen,* Teck Yian Lim,
Alexander G. Schwing, Mark Hasegawa-Johnson, Minh N. Do
University of Illinois at Urbana-Champaign
{yeh17, cchen156, tlim11, aschwing, jhasegaw, minhdo}@illinois.edu

Abstract

Semantic image inpainting is a challenging task where large missing regions have to be filled based on the available visual data. Existing methods which extract information from only a single image generally produce unsatisfactory results due to the lack of high level context. In this paper, we propose a novel method for semantic image inpainting, which generates the missing content by conditioning on the available data. Given a trained generative model, we search for the closest encoding of the corrupted image in the latent image manifold using our context and prior losses. This encoding is then passed through the generative model to infer the missing content. In our method, inference is possible irrespective of how the missing content is structured, while the state-of-the-art learning based method requires specific information about the holes in the training phase. Experiments on three datasets show that our method successfully predicts information in large missing regions and achieves pixel-level photorealism, significantly outperforming the state-of-the-art methods.

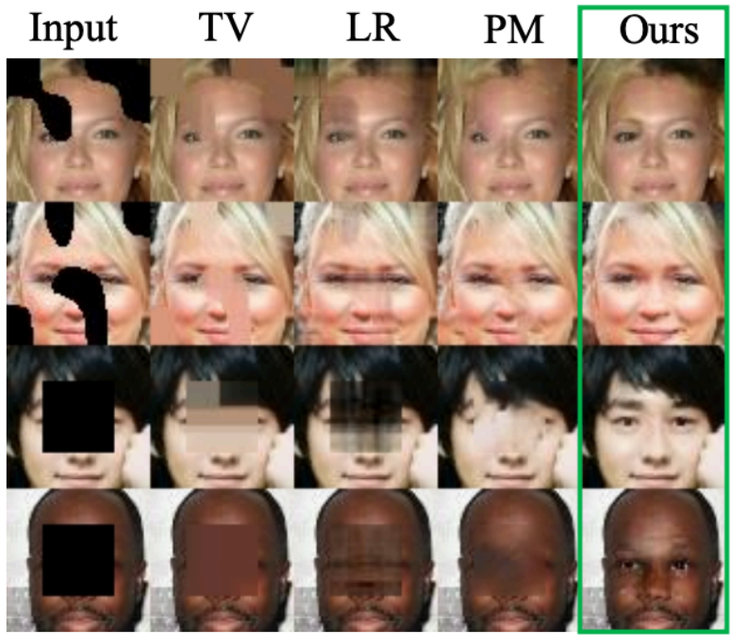


Figure 1. Semantic inpainting results by TV, LR, PM and our method. Holes are marked by black color.

1. Introduction

Hence they are based on the information available in the input image, and exploit image priors to address the ill-posed-ness. For example, total variation (TV) based ap-

SELECTION OF EXISTING WORKS

Compressed Sensing using Generative Models

Ashish Bora* Ajil Jalal† Eric Price‡ Alexandros G. Dimakis§

arXiv:1703.03208

Abstract

The goal of compressed sensing is to estimate a vector from an underdetermined system of noisy linear measurements, by making use of prior knowledge on the structure of vectors in the relevant domain. For almost all results in this literature, the structure is represented by sparsity in a well-chosen basis. We show how to achieve guarantees similar to standard compressed sensing but without employing sparsity at all. Instead, we suppose that vectors lie near the range of a generative model $G : \mathbb{R}^k \to \mathbb{R}^n$. Our main theorem is that, if G is L-Lipschitz, then roughly $O(k \log L)$ random Gaussian measurements suffice for an ℓ_2/ℓ_2 recovery guarantee. We demonstrate our results using generative models from published variational autoencoder and generative adversarial networks. Our method can use 5-10x fewer measurements than Lasso for the same accuracy.

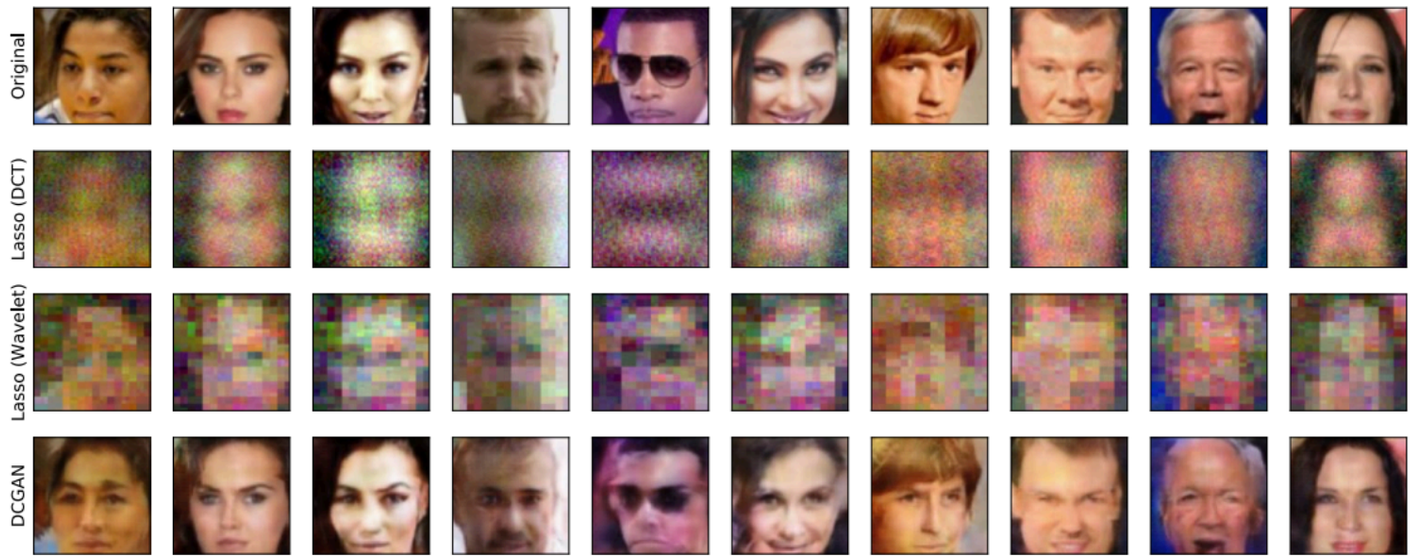


Figure 3: Reconstruction results on celebA with $m = 500$ measurements (of $n = 12288$ dimensional vector). We show original images (top row), and reconstructions by Lasso with DCT basis (second row), Lasso with wavelet basis (third row), and our algorithm (last row).

INKLINGS OF THEORY

- ⭐ We all love the math behind compressed sensing. Analog for learned neural networks is so far a challenge.

- ☑ Interesting results for NNs with random weights: Manoel, Krzakala, Mezard, LZ, arXiv:1701.06981; Hand, Voroninsky, arXiv:1705.07576, Huang, Hand, Heckel, Voroninsky, arXiv:1812.04176 and others.

- ☑ Random rotationally invariant weight matrices: Fletcher, Rangan, arXiv:1706.06549v1, Reeves, arXiv:1710.04580, Gabrié, Manoel, Luneau, Barbier, Macris, Krzakala, LZ, arXiv:1805.09785

- 📌 A lot remains do be done: how many samples/measures needed; sharper analysis; more insights; weaker assumptions on the weights; etc.

FOCUS ON SPIKED MATRIX ESTIMATION

Simple examples

- Denoising: $\Gamma(v) = v + \xi$ (noise $\mathcal{N}(0,\Delta)$)
- Compressed sensing: $\Gamma(v) = Av + \xi, \quad A \in \mathbb{R}^{n \times p}$
- Spiked matrix estimation: $\Gamma(v) = vv^T + \xi$

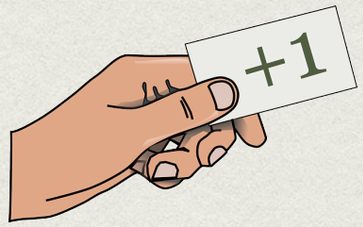
Aubin, Loureiro, Maillard, Krzakala, LZ, arXiv:1905.12385

The spiked matrix model with generative priors

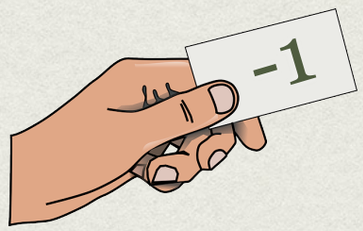
OUTLINE

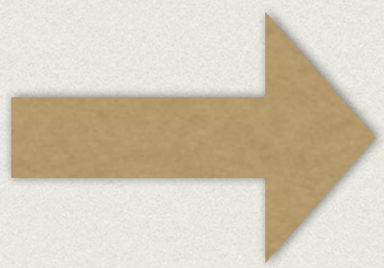
- I. Introduction and motivation.
- II. Sparse PCA: Reminder of key facts and presentation of the methodology.
- III. Spiked matrix estimation with generative priors: Main results, and two take home messages.

LET'S PLAY A GAME

$\frac{\rho}{2}$ 

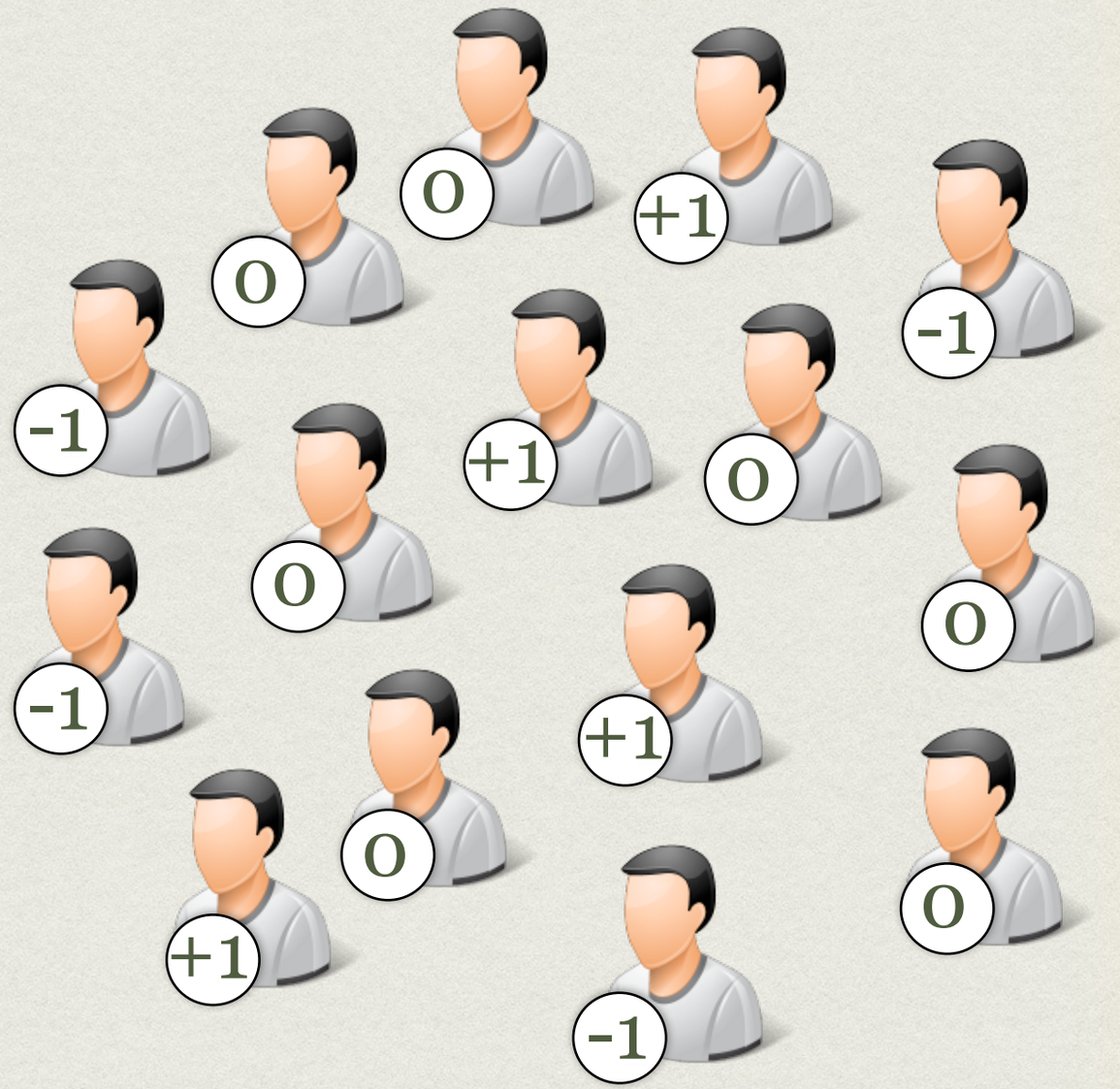
$1-\rho$ 

$\frac{\rho}{2}$ 

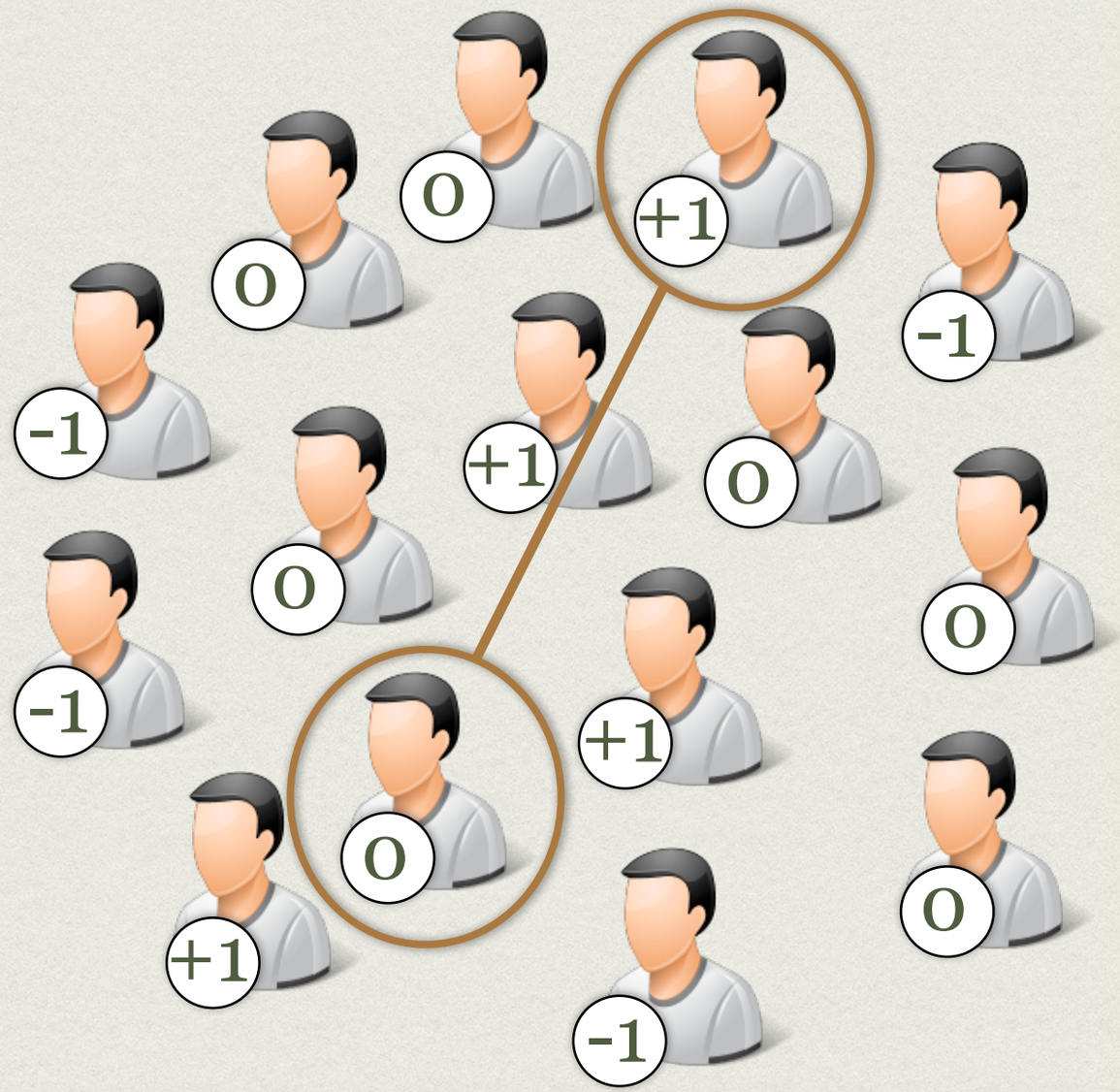


p=15 people

LET'S PLAY A GAME



LET'S PLAY A GAME



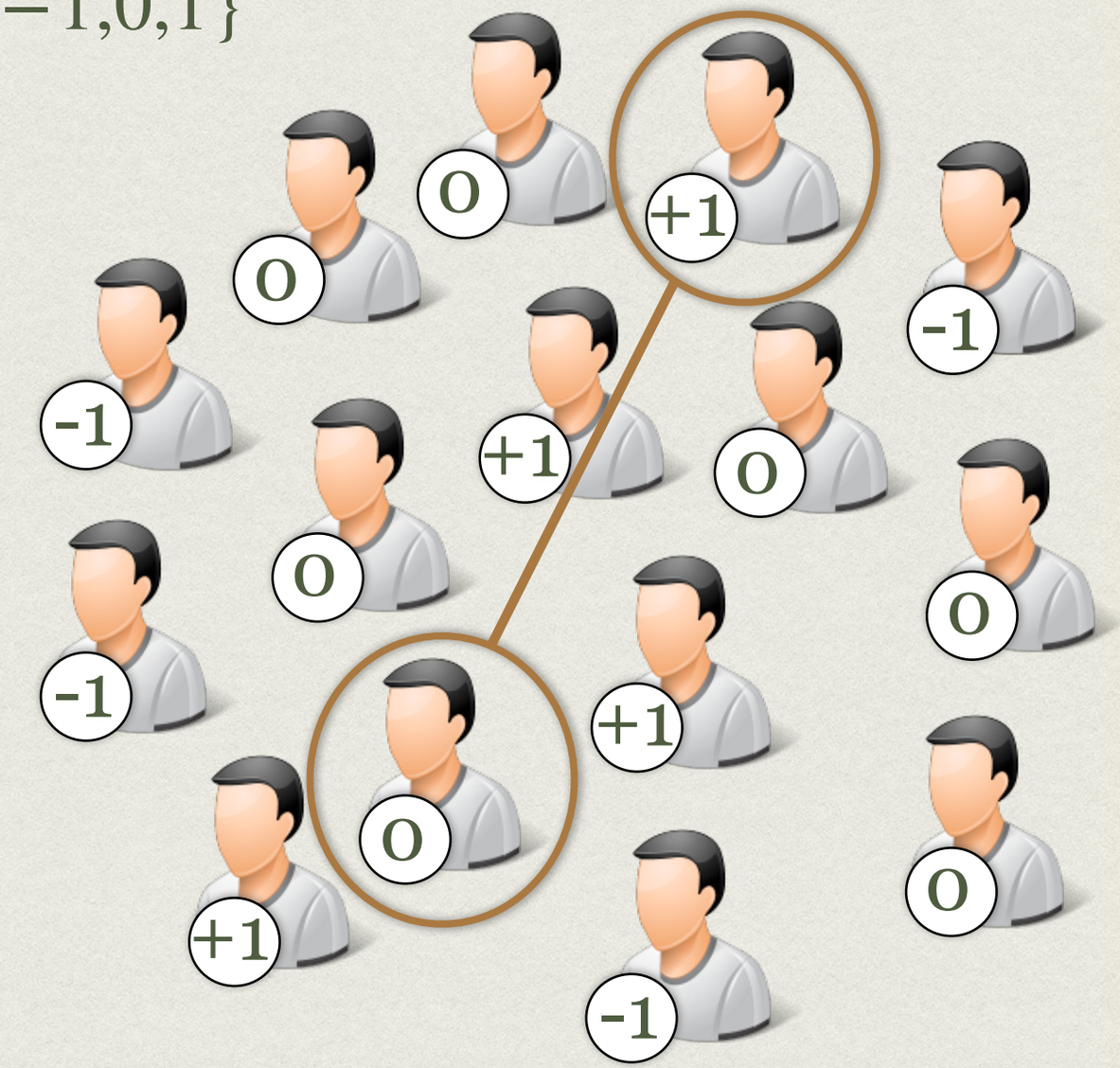
LET'S PLAY A GAME

- Each pair reports: $x_i^* \in \{-1,0,1\}$
  - $Y_{ij} = Z_{ij} + x_i^* x_j^* / \sqrt{p}$

$$Z_{ij} \sim \mathcal{N}(0, \Delta)$$

Collect Y_{ij} for every pair (ij).

Goal: Recover cards (up to symmetry) purely from the knowledge of $\mathbf{Y} = \{Y_{ij}\}_{i<j}$



HOW TO SOLVE THIS?

$$Y_{ij} = \frac{1}{\sqrt{p}} x_i^* x_j^* + Z_{ij}$$

true values of cards: $x_i^* \in \{-1, 0, 1\}$

$$Z_{ij} \sim \mathcal{N}(0, \Delta)$$

x_{PCA} = leading eigenvector of Y estimates x* (up to a sign).

BBP phase transition: $\Delta > \rho^2$ $\quad x_{\mathrm{PCA}} \cdot x^* \approx 0$

$\Delta < \rho^2$ $\quad |x_{\mathrm{PCA}} \cdot x^*| > 0$

Watkin, Nadal'94
Baik, BenArous, Pechet'04

HOW TO SOLVE THIS?

$$Y_{ij} = \frac{1}{\sqrt{p}} x_i^* x_j^* + Z_{ij}$$

true values of cards: $x_i^* \in \{-1, 0, 1\}$

$$Z_{ij} \sim \mathcal{N}(0, \Delta)$$

x_{PCA} = leading eigenvector of Y estimates x* (up to a sign).

BBP phase transition: $\Delta > \rho^2$ $\quad x_{\mathrm{PCA}} \cdot x^* \approx 0$

Watkin, Nadal'94
Baik, BenArous, Pechet'04 $\quad \Delta < \rho^2$ $\quad |x_{\mathrm{PCA}} \cdot x^*| > 0$

PCA: not optimal error value (does not maximise the number of correctly assigned cards)

BAYESIAN INFERENCE

$$P(x|Y) = \frac{P(x)P(Y|x)}{P(Y)}$$

Posterior distribution:

$$P(x\,|\,Y) = \frac{1}{Z(Y,\Delta)} \prod_{i=1}^{p} P_X(x_i) \prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - x_i x_j/\sqrt{p}\right)^2}$$

$$P_X(x_i) = (1-\rho)\delta(x_i) + \frac{\rho}{2}\left[\delta(x_i - 1) + \delta(x_i + 1)\right]$$

Bayes-optimal inference = computation of marginals
(argmax maximizes the number of correctly assigned values,
mean of marginals minimises the mean-squared error).

BAYESIAN INFERENCE

$$P(x|Y) = \frac{P(x)P(Y|x)}{P(Y)}$$

Posterior distribution:

$$P(x\,|\,Y) = \frac{1}{Z(Y,\Delta)} \prod_{i=1}^{p} P_X(x_i) \prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - x_i x_j/\sqrt{p}\right)^2}$$

$$P_X(x_i) = (1-\rho)\delta(x_i) + \frac{\rho}{2}\left[\delta(x_i - 1) + \delta(x_i + 1)\right]$$

Bayes-optimal inference = computation of marginals
(argmax maximizes the number of correctly assigned values,
mean of marginals minimises the mean-squared error).

Computationally costly.

HOW SIMPLE TO ANALYZE?

- High-dimensional (non-convex) problem.
- No statistical consistency as $p \to \infty$.
- We want errors including constants.

Outside the box of “traditional” statistics.

RECENT PROGRESS

(by my group and colleagues)

- Solution of spiked matrix and tensor estimation for any noise distribution, any (separable) prior and rank. (Lesieur, Krzakala, LZ'15-17)

- Rigorous proof that the replica solution for Bayes-optimal inference is correct. (Krzakala, Xu, LZ'16 and Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16)

- Approximate message passing algorithm matching the predicted performance. (Rangan, Fletcher'12, Matsushita, Tanaka'13, Deshpande, Montanari'14, Lesieur, Krzakala, LZ'15-17)

RECENT PROGRESS

(by my group and colleagues)

- Solution of spiked matrix and tensor estimation for any noise distribution, any (separable) prior and rank. (Lesieur, Krzakala, LZ'15-17)

- Rigorous proof that the replica solution for Bayes-optimal inference is correct. (Krzakala, Xu, LZ'16 and Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16)

- Approximate message passing algorithm matching the predicted performance. (Rangan, Fletcher'12, Matsushita, Tanaka'13, Deshpande, Montanari'14, Lesieur, Krzakala, LZ'15-17)

SPIKED MATRIX (TENSOR) MODEL "GENERALISED GAME"

Bayes-optimal inference for generic prior, output, and rank

$$P(x|Y) = \frac{1}{Z(Y)} \prod_{i=1}^{N} P_X(x_i) \prod_{i<j} P_{\text{out}}(Y_{ij}|x_i^T x_j/\sqrt{N}) \qquad x_i \in \mathbb{R}^r$$

or

$$P(u,v|Y) = \frac{1}{Z(Y)} \prod_{i=1}^{N} P_U(u_i) \prod_{j=1}^{M} P_V(v_j) \prod_{i,j} P_{\text{out}}(Y_{ij}|u_i^T v_j/\sqrt{N})$$

or

$$P(x|Y) = \frac{1}{Z(Y)} \prod_{i=1}^{N} P_X(x_i) \prod_{i_1<\ldots<i_p} P_{\text{out}}(Y_{i_1\ldots i_p}|\frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} x_{i_1}\ldots x_{i_p})$$

Generate ground-truth x_i^* from P_X. Generate Y_{ij} from P_{out}.

Goal: Infer x^* from Y.

LOW-RANK MATRIX ESTIMATION

- Symmetric
  - Stochastic Block Model
  - Matrix completion.
  - Submatrix localization.
  - Z2 synchronization.
  - Spiked Wigner models.
- Tensor
  - Spiked tensor model
  - Hyper-graph clustering
  - Tensor completion.
  - Sub-tensor localisation
- Non-symmetric
  - Gaussian mixture clustering.
  - Biclustering.
  - Dawid-Skene model for crowdsourcing.
  - Johnstone's spiked covariance model.
  - Restricted Boltzmann machine with random weights.

RECENT PROGRESS

(by my group and colleagues)

- Solution of spiked matrix and tensor estimation for any noise distribution, any (separable) prior and rank. (Lesieur, Krzakala, LZ'15-17)

- Rigorous proof that the replica solution for Bayes-optimal inference is correct. (Krzakala, Xu, LZ'16 and Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16)

- Approximate message passing algorithm matching the predicted performance. (Rangan, Fletcher'12, Matsushita, Tanaka'13, Deshpande, Montanari'14, Lesieur, Krzakala, LZ'15-17)

PERFORMANCE OF THE BAYES-OPTIMAL ESTIMATOR

Theorem 1: As $p \to \infty$

$\frac{1}{p} \log Z(Y, \Delta)$ concentrates around the maximum of $\Phi(m)$

$$\Phi(m) = \mathbb{E}_{x,w}\left[\log \mathcal{Z}\left(\frac{m}{\Delta}, \frac{m}{\Delta}x + \sqrt{\frac{m}{\Delta}}w\right)\right] - \frac{m^2}{4\Delta} \qquad \begin{array}{l} m \in \mathbb{R} \\ x \sim P_X \\ w \sim \mathcal{N}(0,1) \end{array}$$

= replica symmetric free entropy

$\mathcal{Z}(A, B)$ auxiliary function defined by:

$$\mathcal{P}(x; A, B) = \frac{1}{\mathcal{Z}(A, B)} P_X(x) e^{Bx - Ax^2/2}$$

Proofs: Krzakala, Xu, LZ, ITW'16, Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16 & 18; Lelarge, Miolane'16; El-Alaoui, Krzakala'17

PERFORMANCE OF THE BAYES-OPTIMAL ESTIMATOR

Theorem 1: As $p \to \infty$

$\frac{1}{p} \log Z(Y, \Delta)$ concentrates around the maximum of $\Phi(m)$

$$\Phi(m) = \mathbb{E}_{x,w}\left[\log \mathcal{Z}\left(\frac{m}{\Delta}, \frac{m}{\Delta}x + \sqrt{\frac{m}{\Delta}}w\right)\right] - \frac{m^2}{4\Delta}$$

$m \in \mathbb{R}$

$x \sim P_X$

$w \sim \mathcal{N}(0,1)$

Theorem 2: mean-squared-error of the Bayes-optimal estimator

$$\text{MMSE} = \mathbb{E}_{P_X}(x^2) - \operatorname{argmax} \Phi(m)$$

Proofs: Krzakala, Xu, LZ, ITW'16, Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16 & 18; Lelarge, Miolane'16; El-Alaoui, Krzakala'17

RECENT PROGRESS

(by my group and colleagues)

- Solution of spiked matrix and tensor estimation for any noise distribution, any (separable) prior and rank. (Lesieur, Krzakala, LZ'15-17)

- Rigorous proof that the replica solution for Bayes-optimal inference is correct. (Krzakala, Xu, LZ'16 and Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16)

- Approximate message passing algorithm matching the predicted performance. (Rangan, Fletcher'12, Matsushita, Tanaka'13, Deshpande, Montanari'14, Lesieur, Krzakala, LZ'15-17)

APPROXIMATE MESSAGE PASSING

AMP algorithm estimates means and variances of the marginals:

$$a_i^{t+1} = f(A^t, B_i^t) \qquad v_i^{t+1} = \partial_B f(A^t, B_i^t)$$

$$B_i^t = \frac{1}{\Delta\sqrt{N}} \sum_{l=1}^{N} Y_{il} a_l^t - \frac{1}{\Delta}\left(\frac{1}{N}\sum_{l=1}^{N} v_l^t\right) a_i^{t-1} \qquad A^t = \frac{1}{N\Delta}\sum_{l=1}^{N} \left(a_l^t\right)^2$$

$f(A, B)$ auxiliary function defined by:

$$\mathscr{P}(x; A, B) = \frac{1}{\mathscr{Z}(A, B)} P_X(x) e^{Bx - Ax^2/2} \qquad f(A, B) = \mathbb{E}_{\mathscr{P}}(x)$$

Derived in: Rangan, Fletcher'12; Matsushita, Tanaka'13; Javanmard, Montanari'13; Deshpande, Montanari'14; Lesieur, Krzakala, LZ'15

Traces back to: Thouless, Anderson, Palmer'76

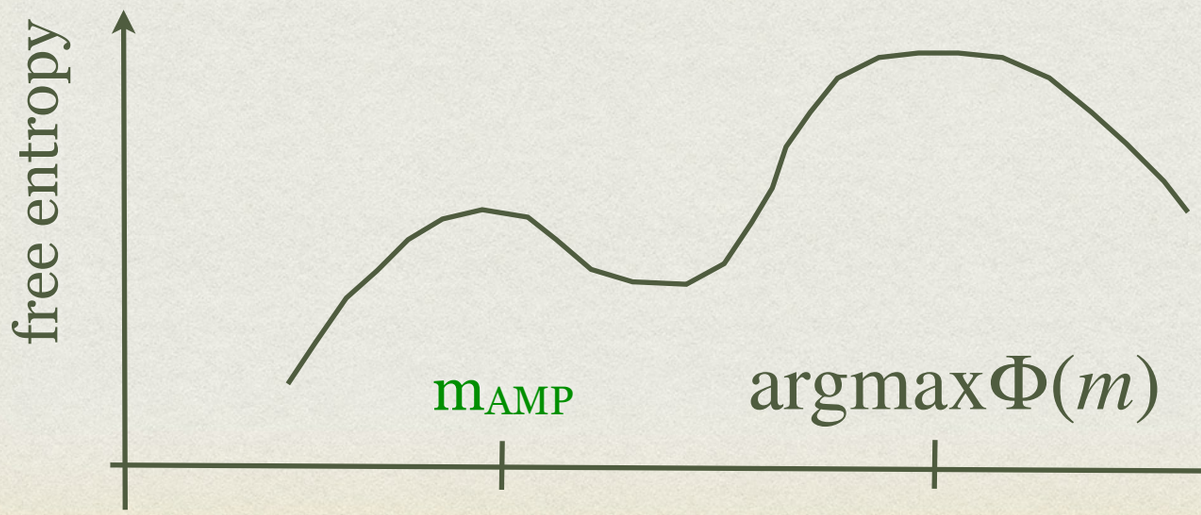
STATE EVOLUTION

$$\Phi(m) = \mathbb{E}_{x,w}\left[\log \mathcal{Z}\left(\frac{m}{\Delta}, \frac{m}{\Delta}x + \sqrt{\frac{m}{\Delta}}w\right)\right] - \frac{m^2}{4\Delta}$$

As $p \to \infty$:

- AMP-MSE given by the local maximum of the free entropy reached ascent starting from small m/large MSE. (Proofs: Rangan, Fletcher'12, Javanmard, Montanari'12, Deshpande, Montanari'14)

- MMSE is given by the global maximum of the free entropy.



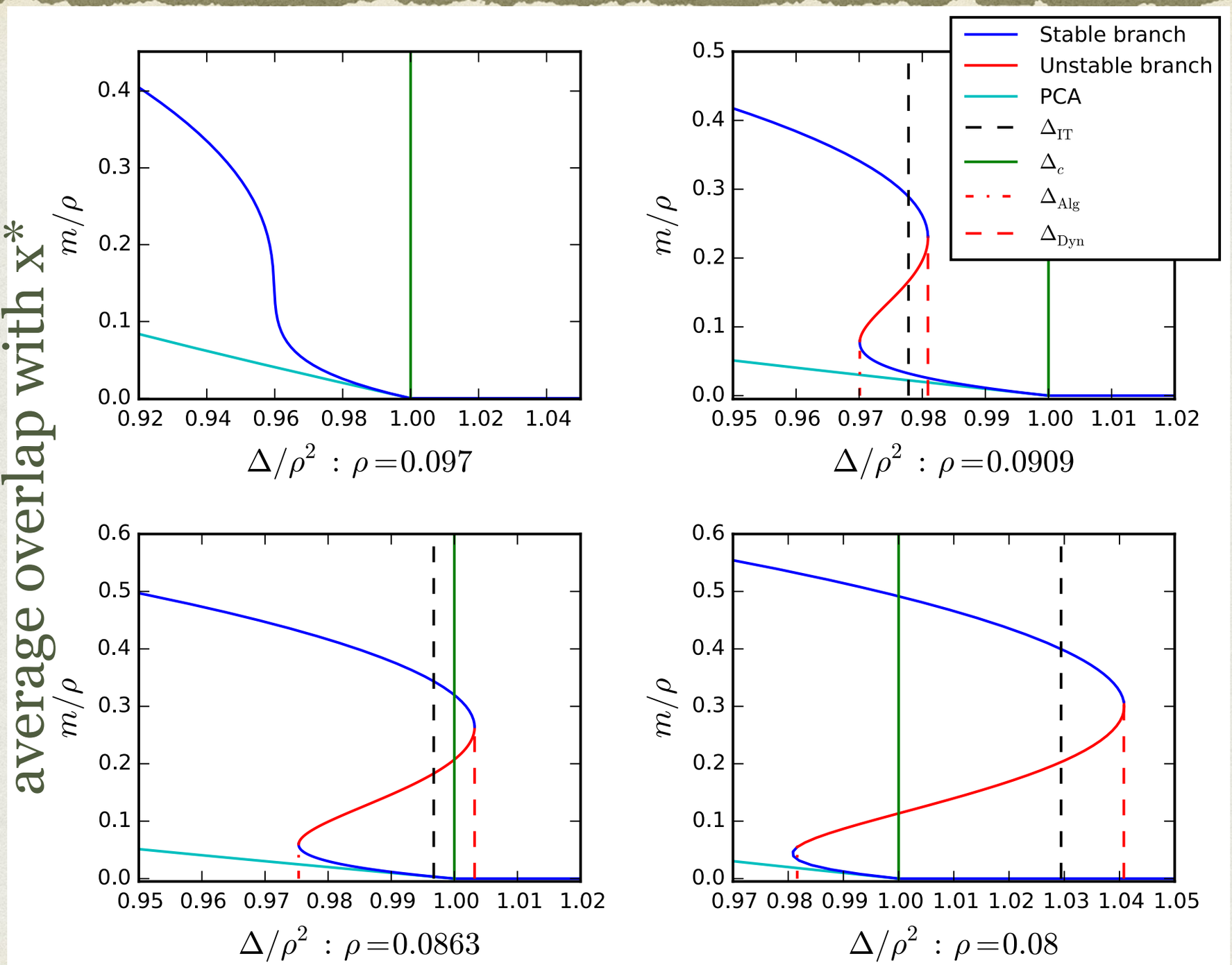
$$\text{MMSE} = \mathbb{E}_{P_X}(x^2) - \operatorname{argmax} \Phi(m)$$

$$\text{MSE}_{\text{AMP}} = \mathbb{E}_{P_X}(x^2) - m_{\text{AMP}}$$

What does this theory imply for Sparse PCA?

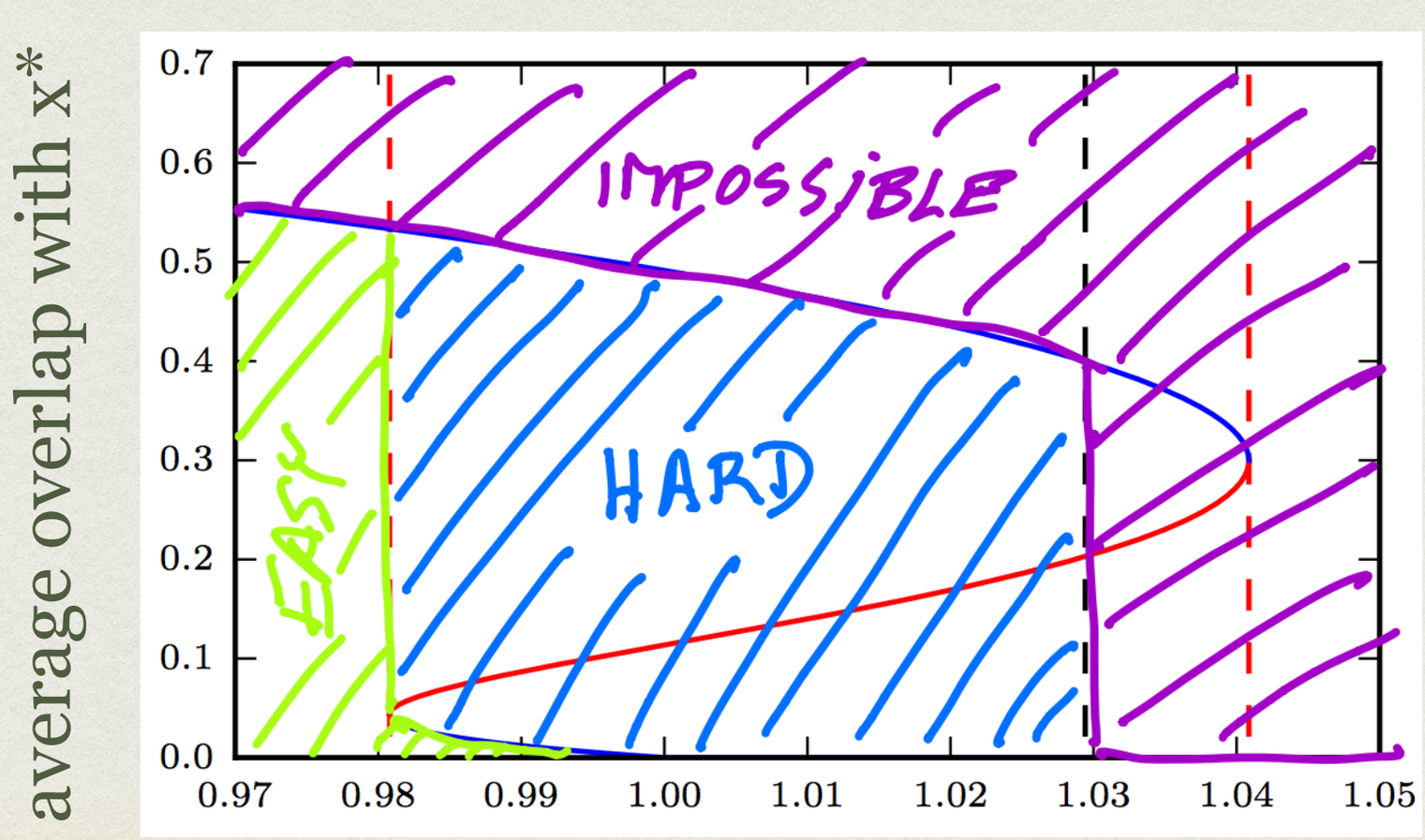
FROM FIXED POINTS TO PHASE TRANSITIONS

$$P_X(x_i) = (1-\rho)\delta(x_i) + \rho\left[\delta(x_i - 1) + \delta(x_i + 1)\right]$$



ALGORITHMIC INTERPRETATION

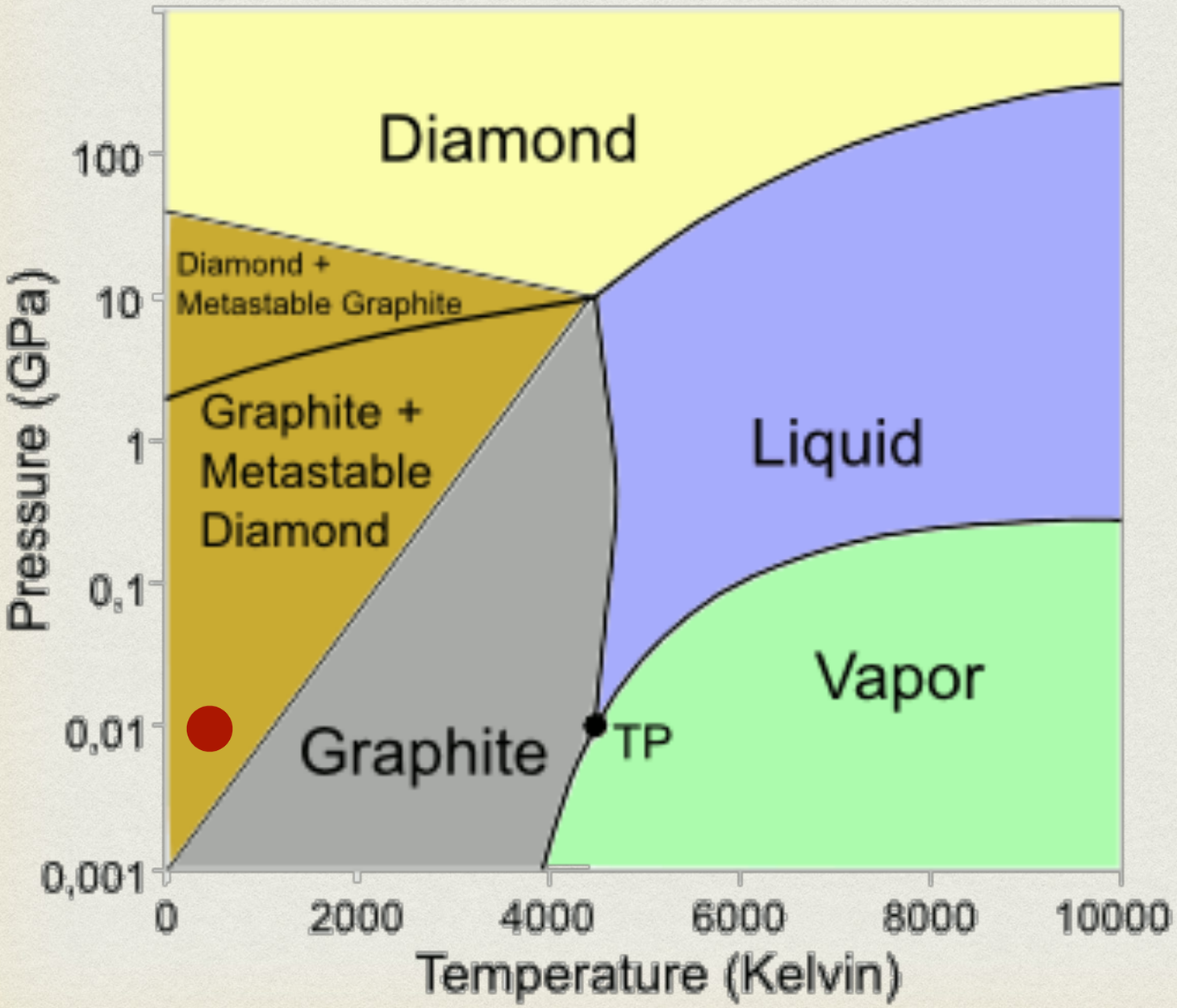
- Easy by approximate message passing.
- Impossible information theoretically.
- Hard phase: coming with *first order phase transition.*



$\rho = 0.08$

HARD PHASE

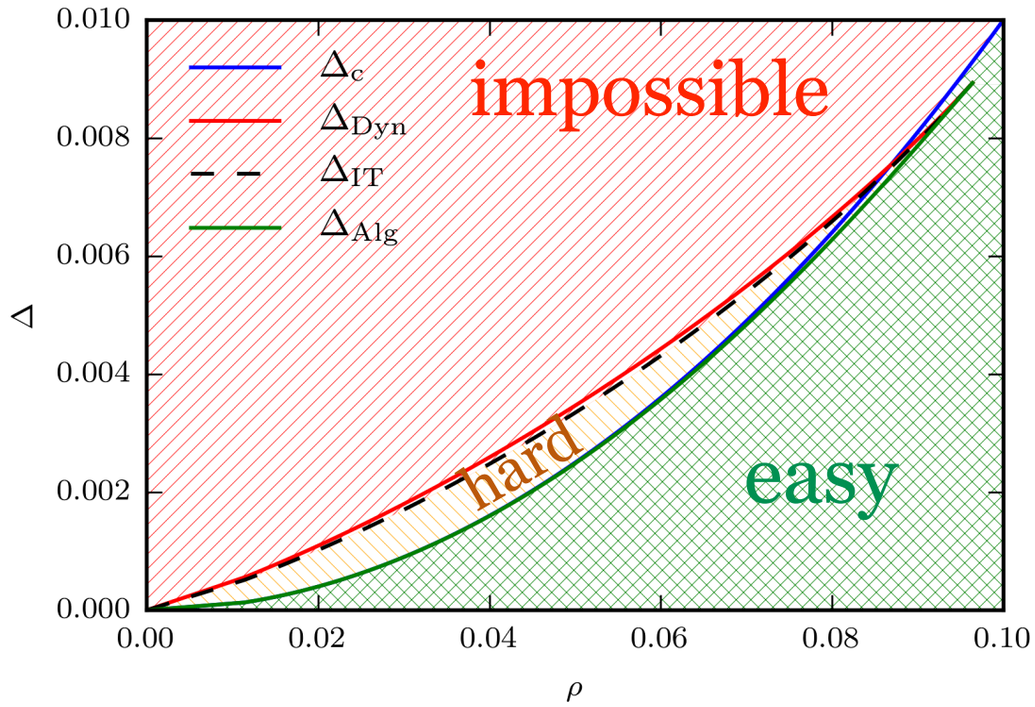
Hard phase: Algorithms “stuck” at low accuracy for exponential time.

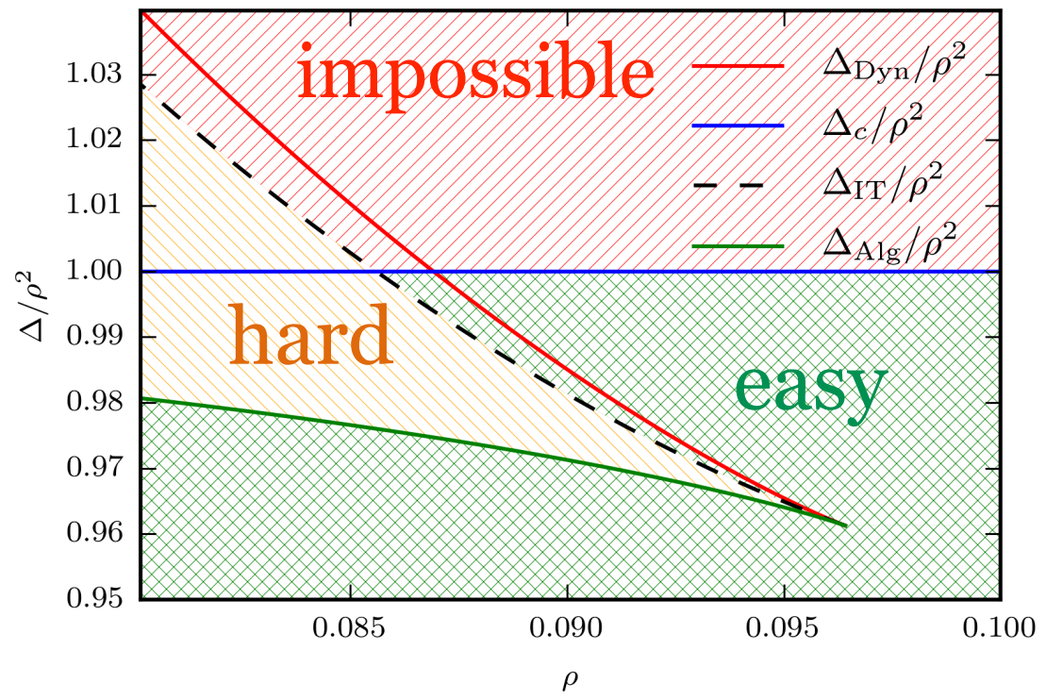


Metastable diamond
= low accuracy.

Equilibrium graphite
= high accuracy.

PHASE DIAGRAM





- Algorithmic gap at small ρ: $\Delta_{\mathrm{Alg}} = \Delta_{\mathrm{PCA}} = \rho^2$, $\Delta_{\mathrm{IT}} \sim_{\rho\to 0} \dfrac{-\rho}{4\log\rho}$
- PCA threshold optimal known at $\rho = \Theta(1)$. For $\rho = o(1)$ better algorithms than PCA (Amini, Wainwright'08; Deshpande, Montanari'14)
- Proof of computational hardness assuming hardness of planted clique problem (Berthet, Rigollet'13).

HARD PHASE

Hard phase = spinodal region of first order phase transitions.

Algorithmic threshold shared by spectral methods and SDPs.

Conjecture:

AMP achieves (in the large N limit) the lowest error among all polynomial algorithms.

Hard phase identified in:

- dense planted sub-matrix;
- sparse principal component analysis;
- Gaussian mixture clustering;
- low-rank tensor completion;

- stochastic block model
- planted constraint satisfaction;
- low-density parity check error correcting codes;

- generalised linear regression;
- compressed sensing;
- learning in binary perceptron;
- phase retrieval;
- committee machine; ...

Computational Threshold Phenomena for Average-Case Problems in Statistics, Machine Learning, and Combinatorial Optimization

STOC 2018 Workshop. June 29, 2018. Los Angeles, CA.

STOC = Symposium of the theory of computing

(Leading conference in computational complexity.)

SPARSE PCA

Facts to recall:

- For $\rho = \Theta(1)$ we have no known algorithms with threshold better than PCA.

- At small ρ, large gap between information-theoretic and best-known-algorithmic performance.

OUTLINE

- I. Introduction and motivation.
- II. Sparse PCA: Reminder of key facts and presentation of the methodology.
- III. Spiked matrix estimation with generative priors: Main results, and two take home messages.

SPIKED MATRIX MODEL WITH GENERATIVE PRIORS

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi \qquad v^* \in \mathbb{R}^p \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

$$v^* = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}x^*)))) \qquad x^* \in \mathbb{R}^k$$

- I. Theory for $W^{(i)}, i = 1,\dots,L$ with **random iid components**.
- II. Approximate message passing reaching optimality (hard phase vanishes).
- III. Spectral algorithms improving over PCA.

BAYESIAN INFERENCE

$$P(v \mid Y) = \frac{1}{Z(Y,\Delta)} P_v(v) \prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - v_i v_j/\sqrt{p}\right)^2}$$

Mutual information: $I(Y; v^*) = -\mathbb{E}_Y[\log Z(Y,\Delta)] + \frac{\rho_v p}{4\Delta}$ $\qquad \rho_v \equiv \frac{1}{p}\mathbb{E}(v^T v)$

Main Theorem: $\lim_{p\to\infty} \frac{I(Y; v^\star)}{p} = \inf_{\rho_v \geq q_v \geq 0} i_{\mathrm{RS}}(\Delta, q_v)$

$$\mathrm{MMSE}_v = \rho_v - \operatorname{arginf} i_{\mathrm{RS}}(q_v)$$

where $i_{\mathrm{RS}}(\Delta, q_v) \equiv \frac{(\rho_v - q_v)^2}{4\Delta} + \frac{1}{p} \lim_{p\to\infty} I\left(v; v + \sqrt{\frac{\Delta}{q_v}}\xi\right)$

Proof: By Guerra interpolation from original to the denoising problem (Aubin, Loureiro, Maillard, Krzakala, LZ, arXiv:1905.12385).

BAYESIAN INFERENCE

$$P(v \mid Y) = \frac{1}{Z(Y,\Delta)} P_v(v) \prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - v_i v_j/\sqrt{p}\right)^2}$$

Mutual information: $I(Y; v^*) = -\mathbb{E}_Y[\log Z(Y,\Delta)] + \frac{\rho_v p}{4\Delta}$ $\qquad \rho_v \equiv \frac{1}{p}\mathbb{E}(v^T v)$

Main Theorem: $\lim_{p\to\infty} \frac{I(Y; v^\star)}{p} = \inf_{\rho_v \geq q_v \geq 0} i_{\mathrm{RS}}(\Delta, q_v)$

$$\mathrm{MMSE}_v = \rho_v - \operatorname{arginf} i_{\mathrm{RS}}(q_v)$$

where $i_{\mathrm{RS}}(\Delta, q_v) \equiv \frac{(\rho_v - q_v)^2}{4\Delta} + \frac{1}{p} \lim_{p\to\infty} I\left(v; v + \sqrt{\frac{\Delta}{q_v}}\xi\right)$

Proof: By Guerra interpolation from original to the denoising problem (Aubin, Loureiro, Maillard, Krzakala, LZ, arXiv:1905.12385).

BAYESIAN INFERENCE

$$P(v \mid Y) = \frac{1}{Z(Y,\Delta)} P_v(v) \prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - v_i v_j/\sqrt{p}\right)^2}$$

Mutual information: $I(Y; v^*) = -\mathbb{E}_Y[\log Z(Y,\Delta)] + \frac{\rho_v p}{4\Delta} \qquad \rho_v \equiv \frac{1}{p}\mathbb{E}(v^T v)$

Main Theorem: $\lim_{p\to\infty} \frac{I(Y; v^\star)}{p} = \inf_{\rho_v \geq q_v \geq 0} i_{\mathrm{RS}}(\Delta, q_v)$

$$\mathrm{MMSE}_v = \rho_v - \operatorname{arginf} i_{\mathrm{RS}}(q_v)$$

where $i_{\mathrm{RS}}(\Delta, q_v) \equiv \frac{(\rho_v - q_v)^2}{4\Delta} + \frac{1}{p}\lim_{p\to\infty} I\left(v; v + \sqrt{\frac{\Delta}{q_v}}\xi\right)$

Proof: By Guerra interpolation from original to the denoising problem (Aubin, Loureiro, Maillard, Krzakala, LZ, arXiv:1905.12385).

PRIOR-MODEL DENOISING

$$v = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}x^*)))) + \frac{\Delta}{q_v}\xi$$

Multi-Layer Generalized Linear Estimation

Andre Manoel
Neurospin, CEA
Université Paris-Saclay

Florent Krzakala
LPS ENS, CNRS
PSL, UPMC & Sorbonne Univ.

Marc Mézard
Ecole Normale Supérieure
PSL Research University

Lenka Zdeborová
IPhT, CNRS, CEA
Université Paris-Saclay

ISIT'17

***Abstract*—We consider the problem of reconstructing a signal from multi-layered (possibly) non-linear measurements. Using non-rigorous but standard methods from statistical physics we present the Multi-Layer Approximate Message Passing (ML-AMP) algorithm for computing marginal probabilities of the corresponding estimation problem and derive the associated state evolution equations to analyze its performance. We also give the expression of the asymptotic free energy and the minimal information-theoretically achievable reconstruction error. Finally, we present some applications of this measurement model for compressed sensing and perceptron learning with structured matrices/patterns, and for a simple model of estimation of latent variables in an auto-encoder.**

components of each of these matrices are drawn independently at random, from a probability distribution $P_{W^{(\ell)}}$ having zero mean and variance $1/n_\ell$. We consider a signal $\boldsymbol{x} \in \mathbb{R}^{n_L}$ with elements x_i, $i = 1, \ldots, n_L$ sampled independently from a distribution $P_X(x_i)$. We then collect n_0 observations $\boldsymbol{y} \in \mathbb{R}^{n_0}$ of the signal $\boldsymbol{x}$ as

$$\boldsymbol{y} = f^{(1)}_{\xi^1}(W^{(1)} f^{(2)}_{\xi^2}(W^{(2)} \cdots f^{(L)}_{\xi^L}(W^{(L)}\boldsymbol{x}))), \qquad (1)$$

where the so-called *activation functions* $f^{(\ell)}_{\xi^\ell}$, $\ell = 1, \ldots, L$, are applied element-wise. These functions can be deterministic or stochastic and are, in general, non-linear. Assuming $f^{(\ell)}_{\xi}(z)$

PRIOR-MODEL DENOISING

Multi-Layer Generalized Linear Estimation

Andre Manoel
Neurospin, CEA
Université Paris-Saclay

Florent Krzakala
LPS ENS, CNRS
PSL, UPMC & Sorbonne Univ.

Marc Mézard
Ecole Normale Supérieure
PSL Research University

Lenka Zdeborová
IPhT, CNRS, CEA
Université Paris-Saclay

ISIT'17

***Abstract*—We consider the problem of reconstructing a signal from multi-layered (possibly) non-linear measurements. Using non-rigorous but standard methods from statistical physics we present the Multi-Layer Approximate Message Passing (ML-AMP) algorithm for computing marginal probabilities of the corresponding estimation problem and derive the associated state evolution equations to analyze its performance. We also give the expression of the asymptotic free energy and the minimal information-theoretically achievable reconstruction error. Finally, we present some applications of this measurement model for compressed sensing and perceptron learning with structured matrices/patterns, and for a simple model of estimation of latent variables in an auto-encoder.**

components of each of these matrices are drawn independently at random, from a probability distribution $P_{W^{(\ell)}}$ having zero mean and variance $1/n_\ell$. We consider a signal $\boldsymbol{x} \in \mathbb{R}^{n_L}$ with elements x_i, $i = 1, \ldots, n_L$ sampled independently from a distribution $P_X(x_i)$. We then collect n_0 observations $\boldsymbol{y} \in \mathbb{R}^{n_0}$ of the signal $\boldsymbol{x}$ as

$$\boldsymbol{y} = f^{(1)}_{\xi^1}(W^{(1)} f^{(2)}_{\xi^2}(W^{(2)} \cdots f^{(L)}_{\xi^L}(W^{(L)}\boldsymbol{x}))), \qquad (1)$$

where the so-called *activation functions* $f^{(\ell)}_{\xi^\ell}$, $\ell = 1, \ldots, L$, are applied element-wise. These functions can be deterministic or stochastic and are, in general, non-linear. Assuming $f^{(\ell)}_{\xi}(z)$

- Proof for single layer prior: Barbier, Krzakala, Macris, Miolane, Krzakala, LZ, COLT'18, PNAS'19
- Proof for two-layer prior: Gabrié, Manoel, Luneau, Macris, Krzakala, LZ, NeurIPS'18.

EXAMPLE OF A RESULT

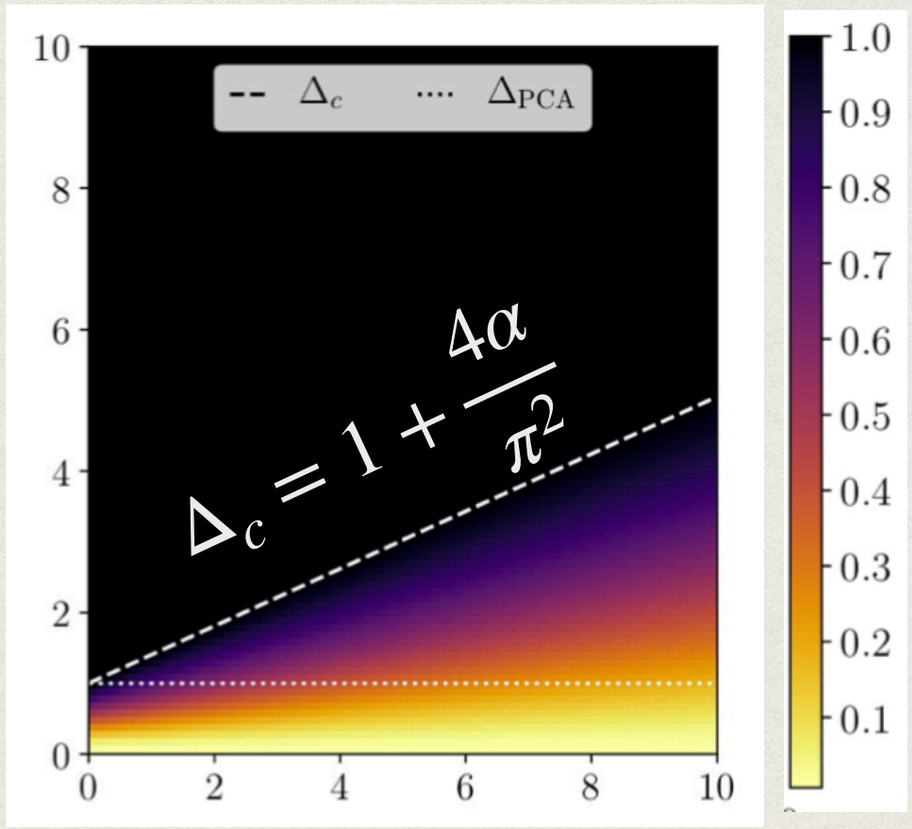
$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi$$

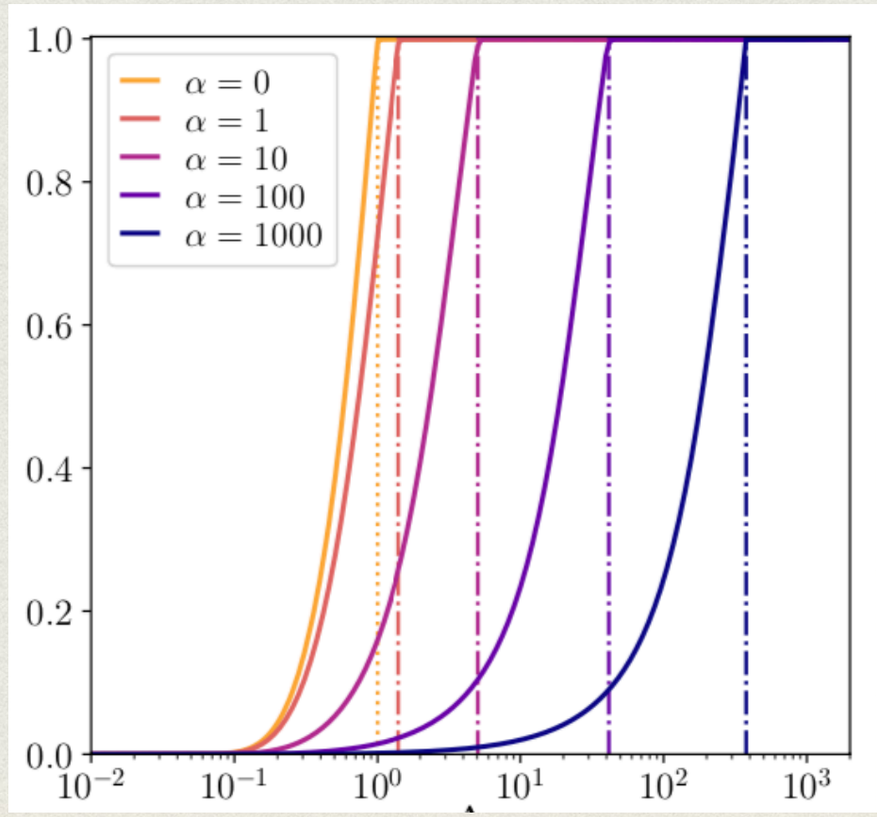
$$v^* = \mathrm{sign}(Wx^*)$$

$v^* \in \mathbb{R}^p$ $\quad \xi_{ij} \sim \mathcal{N}(0,\Delta)$

$x^* \in \mathbb{R}^k$ $\quad x_i^* \sim \mathcal{N}(0,1)$

$W \in \mathbb{R}^{p\times k}$ $\quad W_{ij} \sim \mathcal{N}(0,1/p)$





APPROXIMATE MESSAGE PASSING

Input: $Y \in \mathbb{R}^{p\times p}$ and $W \in \mathbb{R}^{p\times k}$:
Initialize to zero: $(\mathbf{g}, \hat{\mathbf{v}}, \mathbf{B}_v, A_v)^{t=0}$.
Initialize with: $\hat{\mathbf{v}}^{t=1} = \mathcal{N}(0, \sigma^2)$, $\hat{\mathbf{z}}^{t=1} = \mathcal{N}(0, \sigma^2)$, and $\hat{\mathbf{c}}_v^{t=1} = \mathbb{1}_p$, $\hat{\mathbf{c}}_z^{t=1} = \mathbb{1}_k$, $t = 1$.
repeat
Spiked layer:

$$\mathbf{B}_v^t = \frac{1}{\Delta}\frac{Y}{\sqrt{p}}\hat{\mathbf{v}}^t - \frac{1}{\Delta}\frac{\left(\mathbb{1}_p^\intercal \hat{\mathbf{c}}_v^t\right)}{p}\hat{\mathbf{v}}^{t-1} \quad \text{and} \quad A_v^t = \frac{1}{\Delta p}\|\hat{\mathbf{v}}^t\|_2^2 \mathrm{I}_p.$$

Generative layer:

$$V^t = \frac{1}{k}\left(\mathbb{1}_k^\intercal \hat{\mathbf{c}}_z^t\right)\mathrm{I}_p, \quad \boldsymbol{\omega}^t = \frac{1}{\sqrt{k}}W\hat{\mathbf{z}}^t - V^t\mathbf{g}^{t-1} \quad \text{and} \quad \mathbf{g}^t = f_{\text{out}}\left(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t\right),$$

$$\Lambda^t = \frac{1}{k}\|\mathbf{g}^t\|_2^2 \mathrm{I}_k \quad \text{and} \quad \boldsymbol{\gamma}^t = \frac{1}{\sqrt{k}}W^\intercal \mathbf{g}^t + \Lambda^t \hat{\mathbf{z}}^t.$$

Update of the estimated marginals:

$$\hat{\mathbf{v}}^{t+1} = f_v(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t) \quad \text{and} \quad \hat{\mathbf{c}}_v^{t+1} = \partial_B f_v(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t),$$

$$\hat{\mathbf{z}}^{t+1} = f_z(\boldsymbol{\gamma}^t, \Lambda^t) \quad \text{and} \quad \hat{\mathbf{c}}_z^{t+1} = \partial_\gamma f_z(\boldsymbol{\gamma}^t, \Lambda^t),$$

$t = t + 1$.
until Convergence.
Output: $\hat{\mathbf{v}}, \hat{\mathbf{z}}$.

Similar to D-AMP of Metzler, Maleki, Baraniuk'14

STATE EVOLUTION

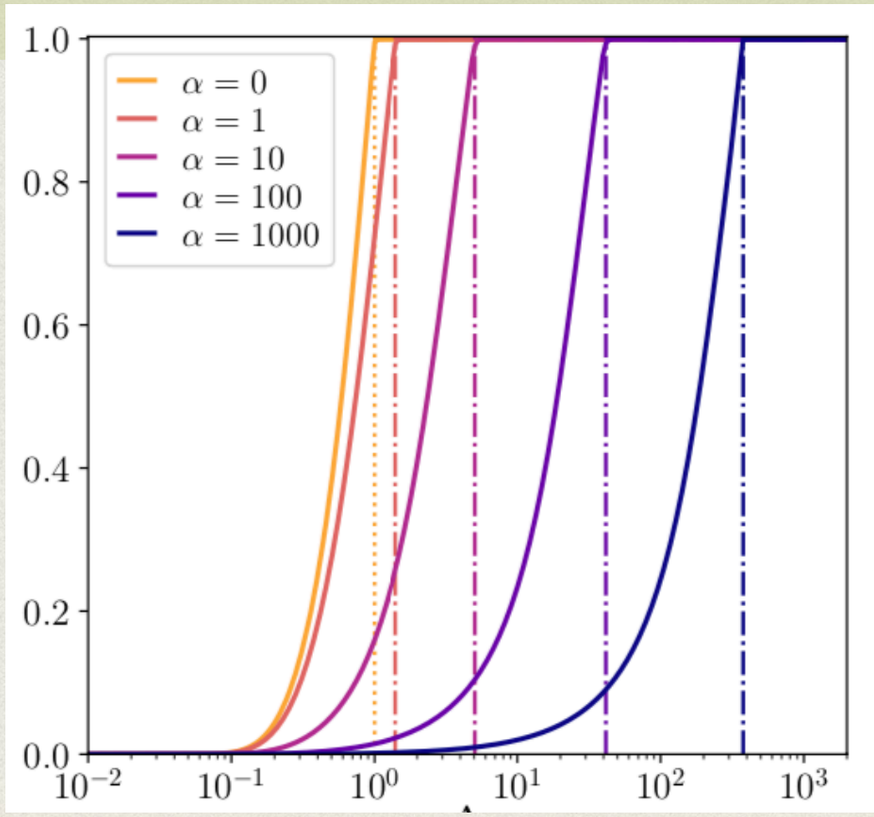
$$i_{\mathrm{RS}}(\Delta, q_v) \equiv \frac{(\rho_v - q_v)^2}{4\Delta} + \frac{1}{p} \lim_{p\to\infty} I\left(v; v + \sqrt{\frac{\Delta}{q_v}}\xi\right)$$

- As long as $i_{\mathrm{RS}}(q_v)$ has a unique minimiser, AMP matches the optimal performance as $p \to \infty$

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi$$

$$v^* = \mathrm{sign}(Wx^*) \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

$$W \in \mathbb{R}^{p\times k} \qquad \alpha = p/k$$



TAKE-HOME MESSAGE I

- Sparse prior: At small ρ, large gap between information-theoretic and best-known-algorithmic performance.

- Generative prior: No gap between information-theoretic and best-known-algorithmic performance.

Generative priors are better than sparsity.

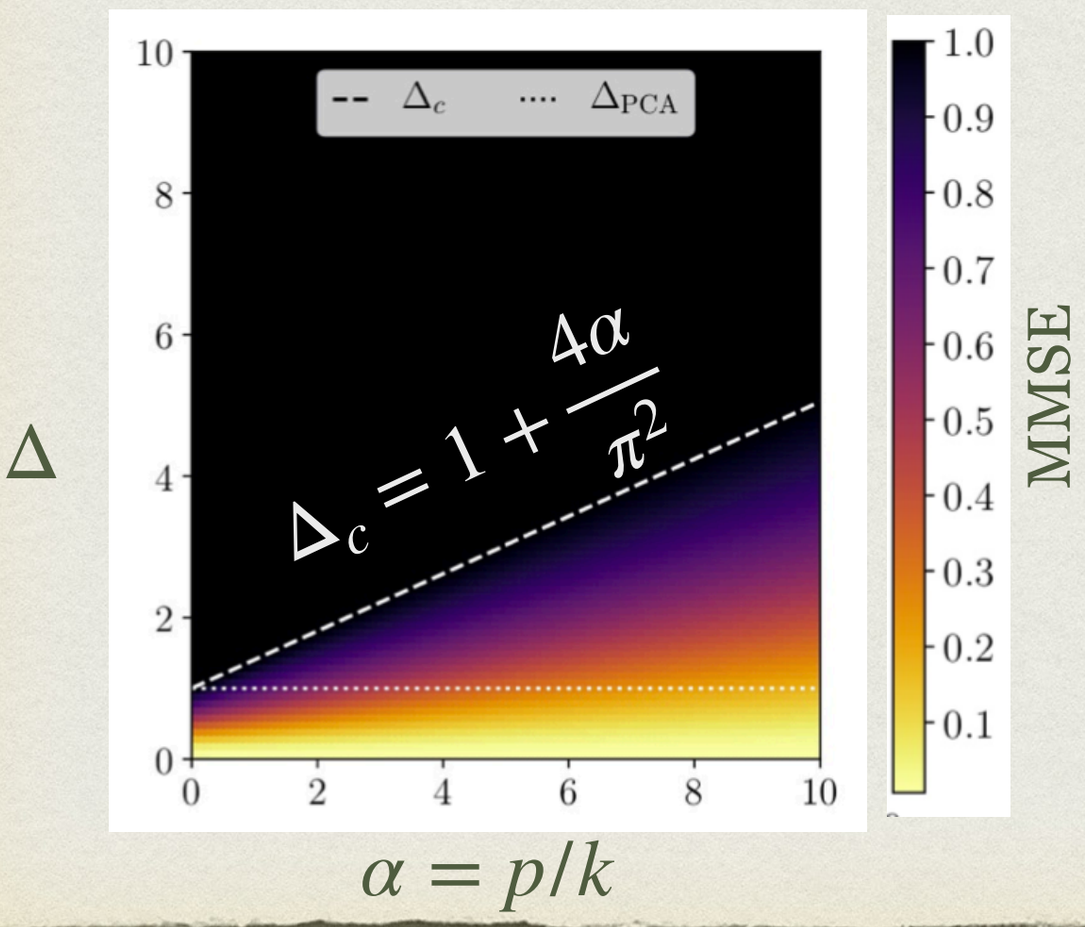
SPECTRAL ALGORITHMS

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi$$

$$v^* = \text{sign}(Wx^*)$$

$$v^* \in \mathbb{R}^p \qquad x^* \in \mathbb{R}^k \qquad W \in \mathbb{R}^{p \times k}$$

$$\xi_{ij} \sim \mathcal{N}(0,\Delta) \qquad x_i^* \sim \mathcal{N}(0,1) \qquad W_{ij} \sim \mathcal{N}(0,1/p)$$

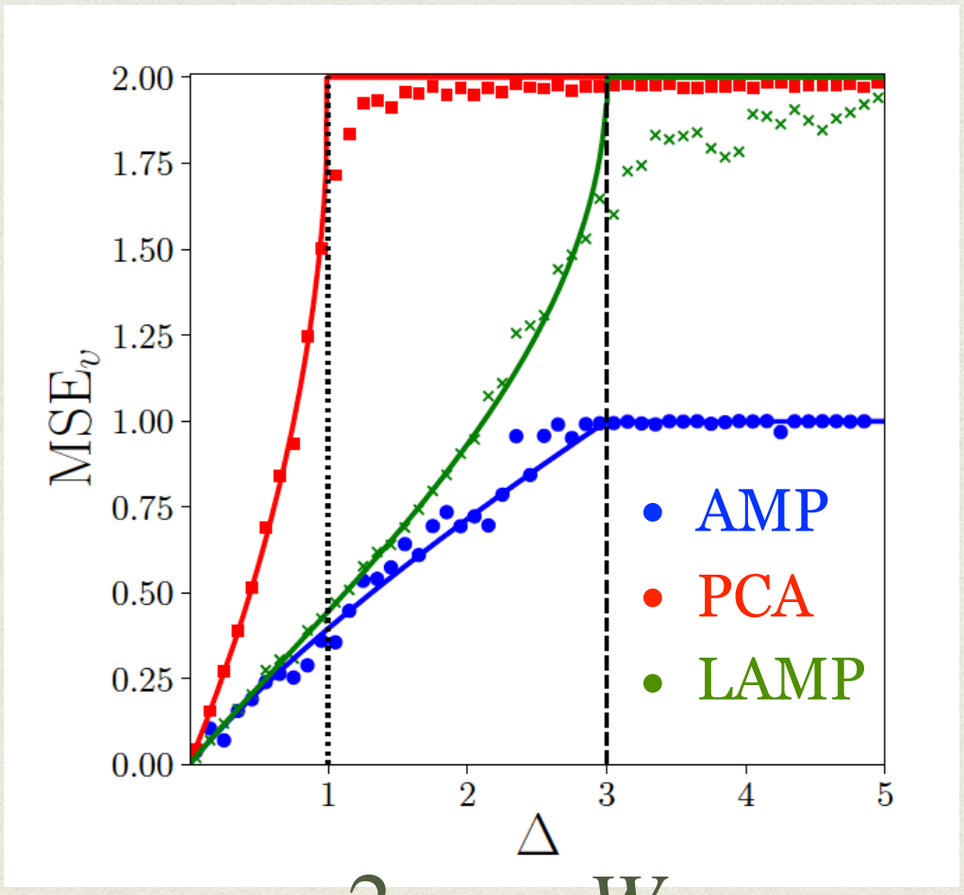


- AMP works for $\Delta < 1 + \frac{4\alpha}{\pi^2}$
- PCA works for $\Delta < 1$

Better spectral algorithms?

OPTIMAL AMONG SPECTRAL ALGORITHMS

- Strategy: Linearize approximate message passing or belief propagation (from Krzakala, Mossel, Moore, Neeman, Sly, LZ, Zhang, PNAS'13)
- Resulting conjecture: Optimal spectral algorithm LAMP uses



$\alpha = 2, v = Wx$

$$\Gamma = K_p \left[Y - \sqrt{p}\, \mathrm{I}_p \right]$$

$$K_p = \mathbb{E}(vv^T)$$

FOR RANDOM MATRIX THEORY LOVERS

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi$$

$$v^* = Wx^*$$

$$v^* \in \mathbb{R}^p \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

$$x^* \in \mathbb{R}^k \qquad x_i^* \sim \mathcal{N}(0,1)$$

$$W \in \mathbb{R}^{p \times k} \qquad W_{ij} \sim \mathcal{N}(0,1/p)$$

$$\alpha = p/k$$

- Theorem: The leading eigenvector of Γ correlates with signal iff $\Delta < 1 + \alpha$

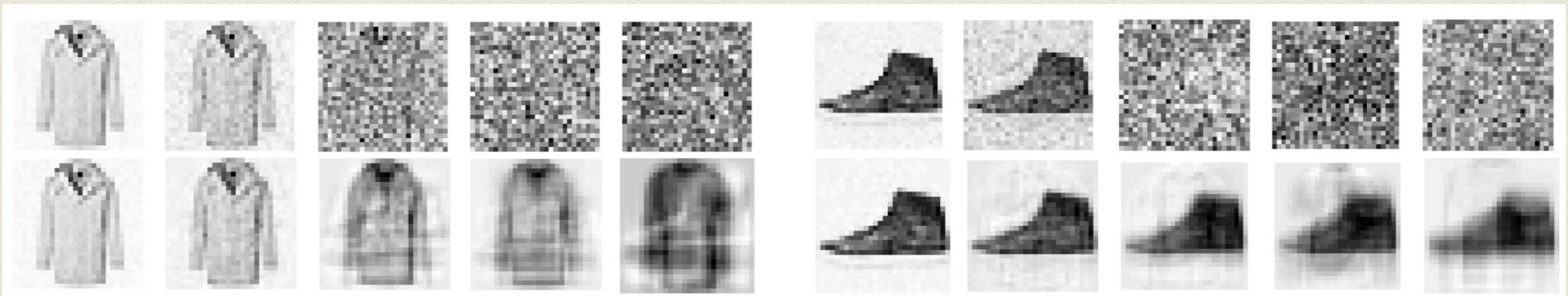
$$\Gamma = WW^T \left[Y - \sqrt{p}\, \mathrm{I}_p \right]$$

- Open problem for any other φ, with $v^* = \varphi(Wx^*)$

LAMP IMPROVES PCA WITHOUT TRAINING ON DATA

PCA (up) versus LAMP (bottom) on spiked matrix estimation

$\Delta = 0.01, 0.1, 1, 2, 10$



$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

LAMP: $\Gamma = K_p \left[Y - \sqrt{p}\, \mathrm{I}_p \right]$ K_p: empirical covariance

TAKE-HOME MESSAGE II

- Sparse prior: For $\rho = \Theta(1)$ no known algorithms with threshold better than PCA.

- Generative prior: spectral LAMP algorithm is better than PCA. Has the same threshold as AMP, conjectured optimal.

$$\Gamma = K_p \left[Y - \sqrt{p}\, \mathrm{I}_p \right]$$

$$K_p = \mathbb{E}(vv^T)$$

CONCLUSION

- Generative models provide generic way to exploit structure of data for better signal processing.

- Spiked matrix estimation with generative priors:
  - Tights analysis for generative neural network with random weights.
  - Absence of algorithmic gaps, contrasting with sparse-PCA.
  - Improved generic-purpose spectral algorithm: LAMP.

TALK BASED ON

- Aubin, Loureiro, Maillard, Krzakala, LZ, *The spiked matrix model with generative priors*, arXiv:1905.12385

- Krzakala, Xu, LZ, *Mutual information in rank-one matrix estimation*, ITW'16
- Barbier, Dia, Macris, Krzakala, Lesieur, LZ *Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula*, NIPS'16
- Lesieur, Krzakala, LZ, *Constrained Low-rank Matrix Estimation: Phase Transitions, Approximate Message Passing and Applications*, J. Stat. Mech.'17
- Manoel, Krzakala, Mezard, LZ, *Multi-layer generalized linear estimation*, ISIT'17.
- Barbier, Krzakala, Macris, Miolane, LZ, *Optimal errors and phase transitions in high-dimensional generalized linear models*, COLT'18, PNAS'19.