SOLVABLE MODEL OF UNSUPERVISED FEATURE LEARNING

Lenka Zdeborová

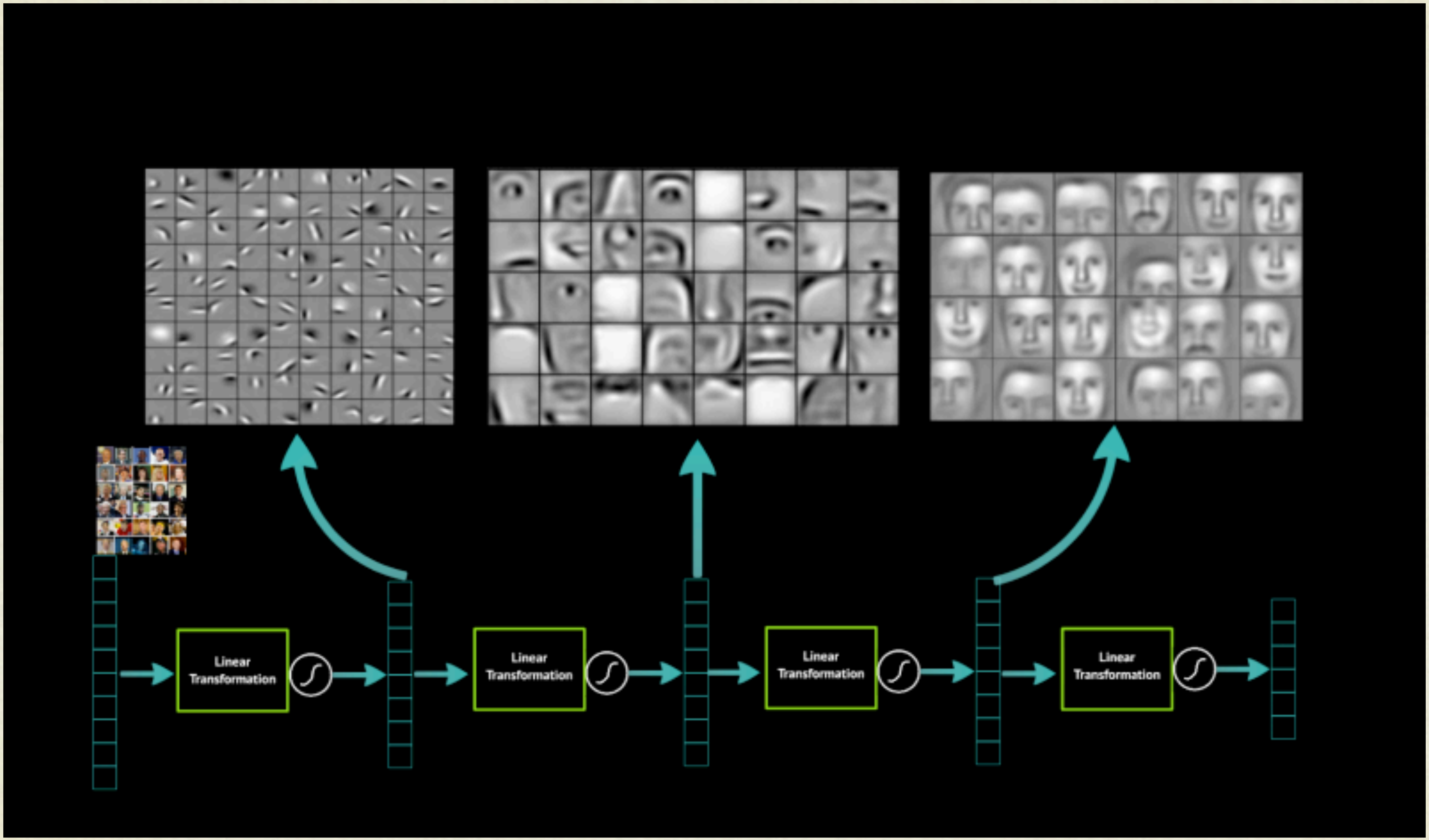
(CNRS and CEA Saclay, France)

with Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata

Simon's Institute, Random Instances workshop

May 2, 2016

FEATURE LEARNING

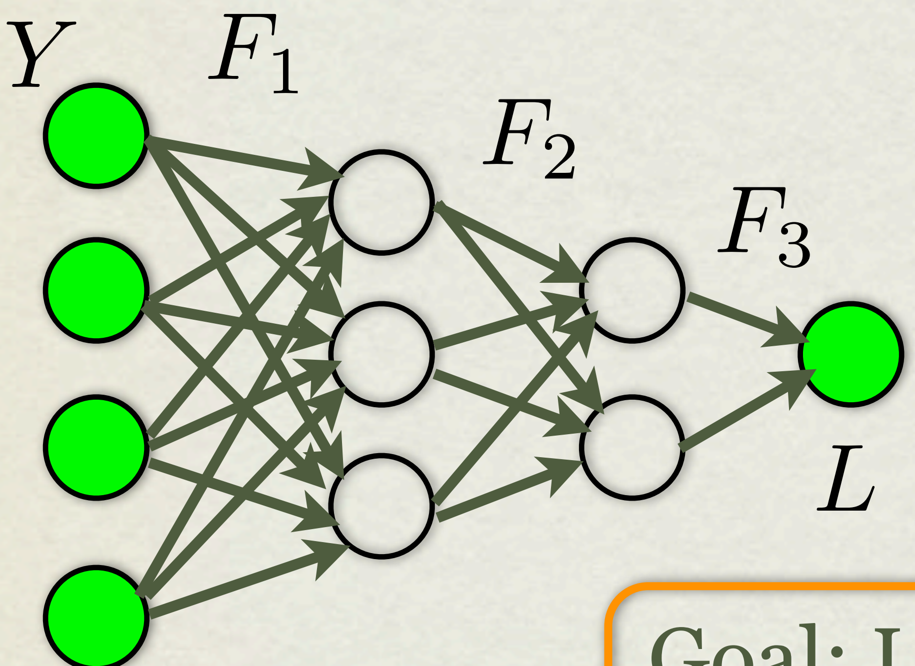


MULTILAYER PERCEPTRON (SUPERVISED)

(Rosenblatt'61)

$Y \in \mathbb{R}^{N \times P}$ P samples of N-dimensional data (known)

$L \in \mathbb{R}^{P}$ Samples are labeled (labels L known).



$N = 4, R_2 = 3, R_2 = 2$

Hierarchy of features /synaptic weights (unknown):

$F_1 \in \mathbb{R}^{R_1 \times N}$

$F_2 \in \mathbb{R}^{R_2 \times R_1}$

$F_3 \in \mathbb{R}^{R_2}$

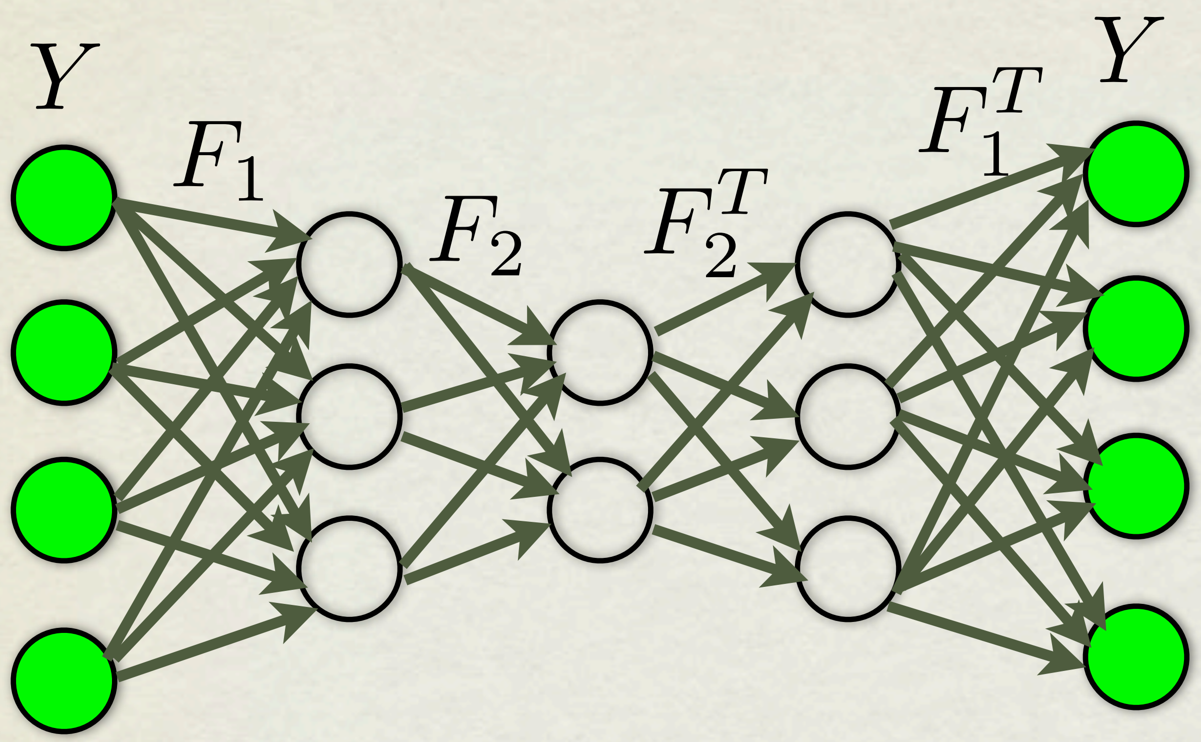
Goal: Learn F_1, F_2, F_3, such that

$$L = g_3\left(F_3\, g_2\left(F_2\, g_1(F_1 Y)\right)\right)$$

g_3, g_2, g_1 activation functions (element-wise) , e.g. sign

AUTO-ENCODER (UNSUPERVISED) (Rumelhart, Hinton, Williams'86)

$Y \in \mathbb{R}^{N \times P}$ P samples of N-dimensional data (known)



$N = 4, R_2 = 3, R_2 = 2$

Hierarchy of features (unknown):

$$F_1 \in \mathbb{R}^{R_1 \times N}$$
$$F_2 \in \mathbb{R}^{R_2 \times R_1}$$

Goal: Learn F_1, F_2, such that

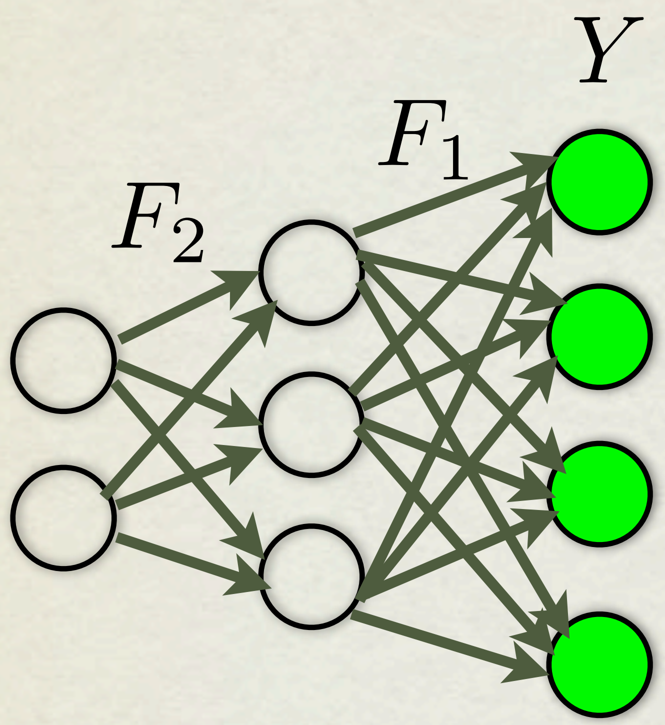
$$Y = g_1\left(F_1^T g_2\left(F_2^T \tilde{g}_2\left(F_2 \tilde{g}_1(F_1 Y)\right)\right)\right)$$

$\tilde{g}_2, \tilde{g}_1, g_2, g_1$ activation functions (element-wise).

INVERTRON (UNSUPERVISED)

(Baldassi, Krzakala, Mezard, LZ, Zecchina, in preparation)

$Y \in \mathbb{R}^{N \times P}$ P samples of N-dimensional data (known)



Hierarchy of features (unknown): $F_1 \in \mathbb{R}^{N \times R_1}$ $F_2 \in \mathbb{R}^{R_1 \times R_2}$

Representation/ compression (sparse, or low-dimensional): $X \in \mathbb{R}^{R_2 \times P}$

Goal: Learn F_1, F_2, X, such that

$$Y = g_1 \left(F_1 g_2 \left(F_2 X \right) \right)$$

g_2, g_1 activation functions (element-wise).

HOW TO BUILD A THEORY?

- Y = some real data, say a database of images. What can be done theoretically!? Not much (with our techniques) ....

HOW TO BUILD A THEORY?

- Y = some real data, say a database of images. What can be done theoretically!? Not much (with our techniques) ....
- Y = random iid elements. For this we have replicas/cavity. Studied for perceptron (Gardner, Derrida, Sompolinsky, ... 80s). But random data do not have features!

HOW TO BUILD A THEORY?

- Y = some real data, say a database of images. What can be done theoretically!? Not much (with our techniques) ....
- Y = random iid elements. For this we have replicas/cavity. Studied for perceptron (Gardner, Derrida, Sompolinsky, ... 80s). But random data do not have features!
- Y = data created by planting iid random features. Now we can talk!

$$Y = g_1\left(F_1^* g_2\left(F_2^* X^*\right)\right)$$

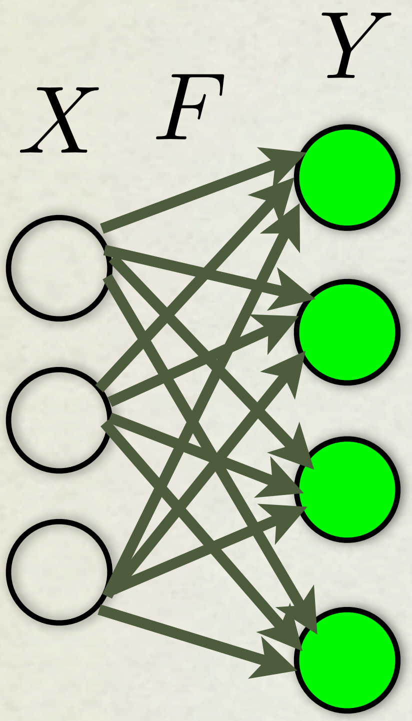
- Planted Invertron: Learn F_1, F_2, X, such that

$$Y = g_1\left(F_1 g_2\left(F_2 X\right)\right)$$

SIMPLEST CASE TO STUDY

(Kabashima, Krzakala, Mezard, Sakata, LZ, Trans. Inf. Theory'16)

$Y \in \mathbb{R}^{N \times P}$ P samples of N-dimensional data (known)



Features (unknown): $F \in \mathbb{R}^{N \times R}$

Sparse representation (unknown): $X \in \mathbb{R}^{R \times P}$

$g(\cdot)$ activation functions (element-wise).

Goal: Learn F, and X, such that $Y = g(FX)$

Known also as (Olshausen, Field'97):
Dictionary learning, sparse coding, matrix factorization ...

MATRIX FACTORIZATION

- = the smallest non-trivial piece of feature learning.
- Represent P-samples of N-dimensional data (Y, known) by features (F, unknown), and weights (X, unknown) trough a (non-linear) activation function f(.)

$$Y_{\mu i} = f\Big(\sum_{\alpha=1}^{R} X_{\mu\alpha} F_{\alpha i}\Big) \qquad \begin{array}{l} \mu = 1, \ldots, P \\ i = 1, \ldots, N \end{array}$$

- Dictionary learning: The dictionary (F) has R "atoms", we typically look for F such that the data Y can be explained with sparse weights X (think of sound expressed with Fourier, images in wavelets ...).
- Related to talks by F. Krzakala (with R=O(N)), D. Steurer (k=2)

SOME KNOWN RESULTS

- Algorithms: MOD, K-SVD, alternate minimization with L_1 regularization. But all require many samples P. What is the minimal number of samples needed?

- Theory: Interesting statistical results assuming incoherence of F, o(N) sparsity of X. For O(N) sparsity existing results not satisfactory. So far O(N log(N)) samples needed, MMSE unknown.

[30] Lewicki M. S. & Sejnowski T. J. Learning overcomplete representations. *Neural computation* **12**, 337–365 (2000).

[31] Engan K., Aase S. O. & Husoy J. H. Method of optimal directions for frame design. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2443–2446 (IEEE, 1999).

[32] Aharon M., Elad M. & Bruckstein A. M. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing* **54**, 4311 (2006).

[33] Michal Aharon, Michael Elad A. M. B. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra and its Applications* **416**, 48–67 (2006).

[34] Vainsencher D., Mannor S. & Bruckstein A. M. The Sample Complexity of Dictionary Learning. *Journal of Machine Learning Research* **12**, 3259–3281 (2011).

[35] Jenatton R., Gribonval R. & Bach F. Local stability and robustness of sparse dictionary learning in the presence of noise. *arXiv:1210.0685* (2012).

[36] Spielman D. A., Wang H. & Wright J. Exact recovery of sparsely-used dictionaries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 3087–3090 (AAAI Press, 2013).

[37] Arora S., Ge R. & Moitra A. New algorithms for learning incoherent and overcomplete dictionaries. *arXiv preprint arXiv:1308.6273* (2013).

[38] Agarwal A., Anandkumar A. & Netrapalli P. Exact Recovery of Sparsely Used Overcomplete Dictionaries. *arXiv preprint arXiv:1309.1952* (2013).

[39] Gribonval R., Jenatton R., Bach F., Kleinsteuber M. & Seibert M. Sample complexity of dictionary learning and other matrix factorizations. *arXiv preprint arXiv:1312.3790* (2013).

"PLANTED" MATRIX FACTORIZATION

Teacher creates data Y as:

$$\mu = 1, \dots, P$$
$$i = 1, \dots, N$$
$$\alpha = 1, \dots, R$$

$$Y_{\mu i} = f\left(\sum_{\alpha=1}^{R} X^*_{\mu\alpha} F^*_{\alpha i}\right) \qquad \begin{aligned} X^*_{\mu\alpha} &\sim P_X(X^*_{\mu\alpha}) \\ F^*_{\alpha i} &\sim P_F(F^*_{\alpha i}) \end{aligned}$$

Student estimates F, X from Y, f(.), P_X and P_F.

- Y known data (P samples of N-dimensional data)
- F unknown dictionary, features
- X unknown coefficients (typically sparse).
- f(.) known "output channel", e.g. $f(Z) = Z + W, \quad W \sim \mathcal{N}(0, \Delta)$
  nonlinear f(.) relevant in neural nets.

BAYES-OPTIMAL STUDENT

- Posterior probability distribution

$$P(X_{\mu\alpha}, F_{\alpha i}|Y_{\mu i}) = \frac{1}{Z} \prod_{\alpha i} P_F(F_{\alpha i}) \prod_{\mu\alpha} P_X(X_{\mu\alpha}) \prod_{\mu i} P_{\text{out}}(Y_{\mu i}| \sum_{\alpha} F_{\alpha i} X_{\mu\alpha})$$

- Marginal probabilities

$$\mu_X(X_{\mu\alpha}), \mu_F(F_{\alpha i})$$

- Bayes-optimal estimator minimizes the mean-squared error, i.e. squared distance to the ground-truth

$$\hat{X}_{\mu\alpha} = \mathbb{E}_{\mu_X}(X_{\mu\alpha}) \qquad \hat{F}_{\alpha i} = \mathbb{E}_{\mu_F}(F_{\alpha i})$$

SOLVABLE WITH REPLICAS

Exact (but non-rigorous) computation of the performance (MMSE) of the Bayes-optimal student.

Posterior probability distribution:

$$P(X_{\mu\alpha}, F_{\alpha i}|Y_{\mu i}) = \frac{1}{Z}\prod_{\alpha i} P_F(F_{\alpha i})\prod_{\mu\alpha} P_X(X_{\mu\alpha})\prod_{\mu i} P_{\rm out}(Y_{\mu i}|\sum_{\alpha} F_{\alpha i}X_{\mu\alpha})$$

$$\mu = 1,\dots,P$$
$$i = 1,\dots,N$$
$$\alpha = 1,\dots,R$$

The thermodynamic limit and scaling of quantities:

$$N, P, R \to \infty \qquad \alpha = N/R = O(1) \qquad \pi = P/R = O(1)$$

$$Y_{\mu i} = O(1) \qquad X_{\mu\alpha} = O(1) \qquad F_{\alpha i} = O(1/\sqrt{R})$$

$$\mathbb{E}_{P_F}(F_{\alpha i}) = 0$$

THE REPLICA METHOD

$$P(X_{\mu\alpha}, F_{\alpha i}|Y_{\mu i}) = \frac{1}{Z}\prod_{\alpha i} P_F(F_{\alpha i}) \prod_{\mu\alpha} P_X(X_{\mu\alpha}) \prod_{\mu i} P_{\text{out}}(Y_{\mu i}|\sum_{\alpha} F_{\alpha i} X_{\mu\alpha})$$

$$N, P, R \to \infty \qquad \alpha = N/R = O(1) \qquad \pi = P/R = O(1)$$

1) Compute average of Z^n over realizations of $X^*_{\mu\alpha}, F^*_{\alpha i}$ and noise for $n \in \mathbb{N}$

2) Use the following identity: $\overline{\log Z} = \lim_{n\to 0} \frac{\overline{Z^n} - 1}{n}$

3) After (a bit of) work: $\overline{\log Z} \propto \int \mathrm{d}m_F \, \mathrm{d}m_X \, \mathrm{d}\hat{m} \, e^{N^2 \Phi(m_X, m_F, \hat{m})}$

FREE ENERGY

Replica free energy of the planted dictionary learning:

$$\phi(m_F, m_X, \hat{m}_F = \pi m_X \hat{m}, \hat{m}_X = \alpha m_F \hat{m}) =$$

$$\alpha\pi \int \mathrm{d}y\, \mathcal{D}\xi\, \mathcal{D}u^0 P_{\mathrm{out}}\left(y|\sqrt{\Gamma - m_F m_X}u^0 + \sqrt{m_F m_X}\xi\right) \log\left(\int \mathcal{D}u\, P_{\mathrm{out}}\left(y|\sqrt{\Gamma - m_F m_X}u + \sqrt{m_F m_X}\xi\right)\right)$$
$$+\alpha\left(-\frac{\hat{m}_F m_F}{2} + \int \mathcal{D}\xi\, \mathrm{d}F^0 e^{-\frac{R\hat{m}_F}{2}(F^0)^2 + \sqrt{R\hat{m}_F}\xi F^0} P_F(F^0) \log\left(\int \mathrm{d}F e^{-\frac{R\hat{m}_F}{2}F^2 + \sqrt{R\hat{m}_F}\xi F} P_F(F)\right)\right)$$
$$+\pi\left(-\frac{\hat{m}_X m_X}{2} + \int \mathcal{D}\xi\, \mathrm{d}X^0 e^{-\frac{\hat{m}_X}{2}(X^0)^2 + \sqrt{\hat{m}_X}\xi X^0} P_X(X^0) \log\left(\int \mathrm{d}X e^{-\frac{\hat{m}_X}{2}X^2 + \sqrt{\hat{m}_X}\xi X} P_X(X)\right)\right),$$

$$\Gamma = R\mathbb{E}_{P_X}(X^2)\mathbb{E}_{P_F}(F^2)$$

Global maximum of $\phi(m_F, m_X, \hat{m})$ gives the MMSE:

$$E_X = \mathrm{MMSE}(X) = \mathbb{E}_{P_X}(X^2) - m_X$$

$$E_F = \mathrm{MMSE}(F) = R\mathbb{E}_{P_F}(F^2) - m_F$$

STATIONARITY CONDITIONS

$$m_X = \frac{1}{\sqrt{\alpha m_F \hat{m}}} \int \mathrm{d}t \frac{\left[f_1^X\left(\frac{t}{\sqrt{\alpha m_F \hat{m}}}, \frac{1}{\alpha m_F \hat{m}}\right)\right]^2}{f_0^X\left(\frac{t}{\sqrt{\alpha m_F \hat{m}}}, \frac{1}{\alpha m_F \hat{m}}\right)}$$

$$m_F = \frac{1}{\sqrt{\pi m_X \hat{m}}} \int \mathrm{d}t \frac{\left[f_1^F\left(\frac{t}{\sqrt{\pi m_X \hat{m}}}, \frac{1}{\pi m_X \hat{m}}\right)\right]^2}{f_0^F\left(\frac{t}{\sqrt{\pi m_X \hat{m}}}, \frac{1}{\pi m_X \hat{m}}\right)}$$

$$\hat{m} = \frac{1}{m_X m_F} \int \mathrm{d}y \int \mathcal{D}t \frac{\left[\partial_t f_0^Y(y|\sqrt{m_X m_F} t, \Gamma - m_X m_F)\right]^2}{f_0^Y(y|\sqrt{m_X m_F} t, \Gamma - m_X m_F)}$$

Generic

Problem dependent functions

$$f_n^X(T, \Sigma) \equiv \frac{1}{\sqrt{2\pi\Sigma}} \int \mathrm{d}X \, X^n P_X(X) e^{-\frac{(X-T)^2}{2\Sigma}}$$

$$f_n^F(W, Z) \equiv \frac{1}{\sqrt{2\pi Z}} \int \mathrm{d}F \, (\sqrt{R}F)^n P_F(F) e^{-\frac{(\sqrt{R}F-W)^2}{2Z}}$$

$$f_n^Y(y|\omega, V) \equiv \frac{1}{\sqrt{2\pi V}} \int \mathrm{d}t \, (t-\omega)^n P_{\mathrm{out}}(y|t) e^{-\frac{(t-\omega)^2}{2V}}$$

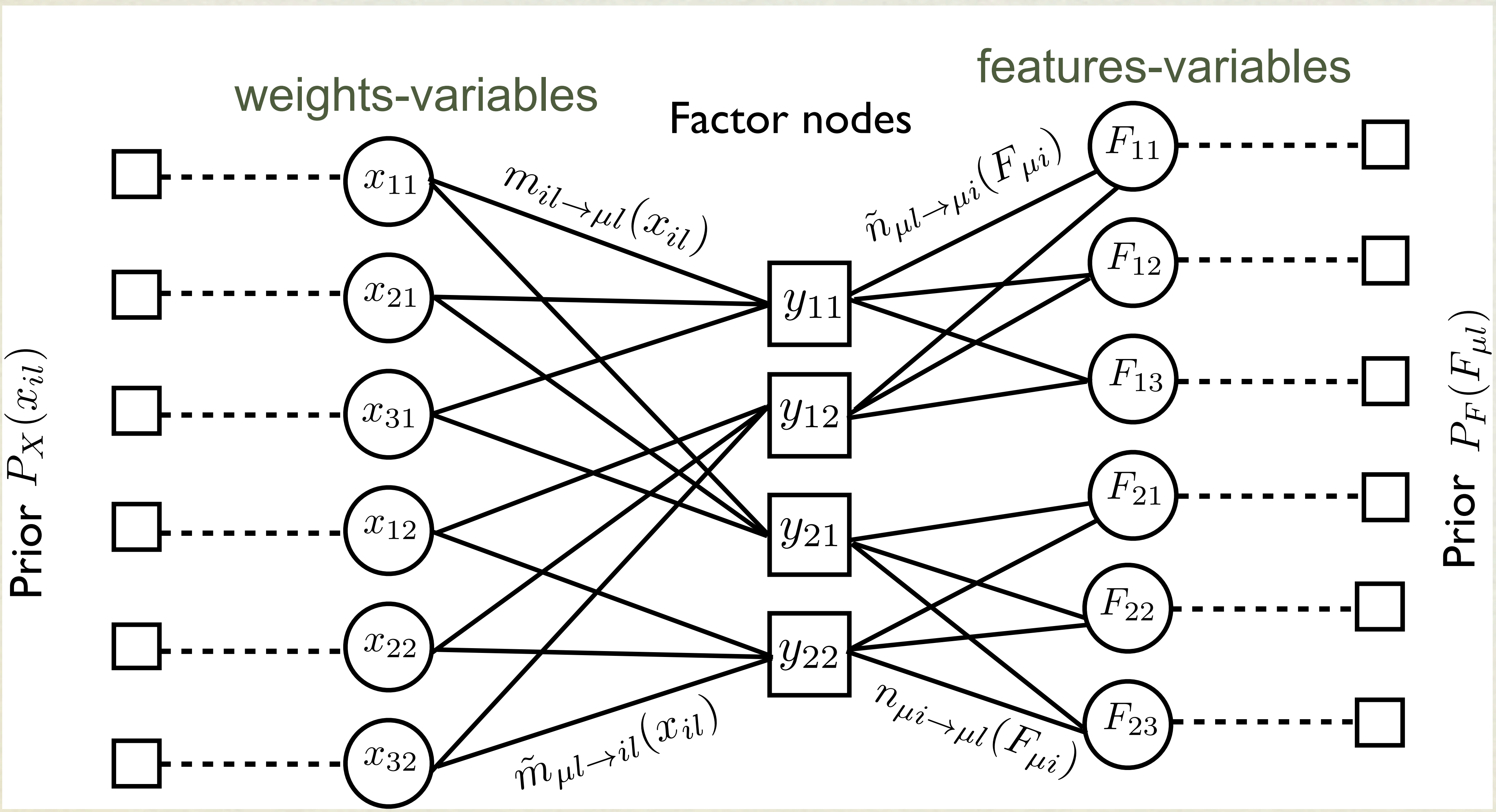
AND ALGORITHMS?

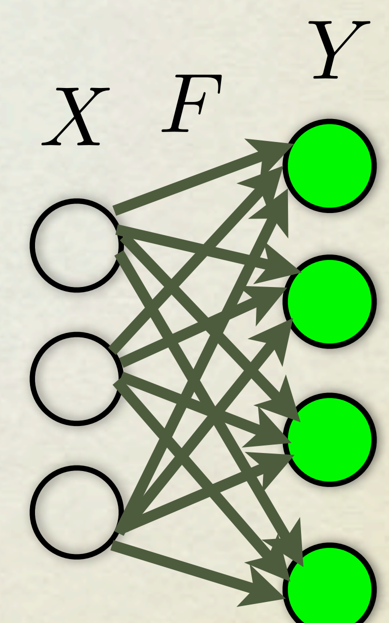
- Andrea Montanari on Monday:

  For dense models do approximate message passing.

GRAPHICAL MODEL

$$P(X_{\mu\alpha}, F_{\alpha i}|Y_{\mu i}) = \frac{1}{Z}\prod_{\alpha i} P_F(F_{\alpha i}) \prod_{\mu\alpha} P_X(X_{\mu\alpha}) \prod_{\mu i} P_{\text{out}}(Y_{\mu i}|\sum_{\alpha} F_{\alpha i} X_{\mu\alpha})$$





BELIEF PROPAGATION

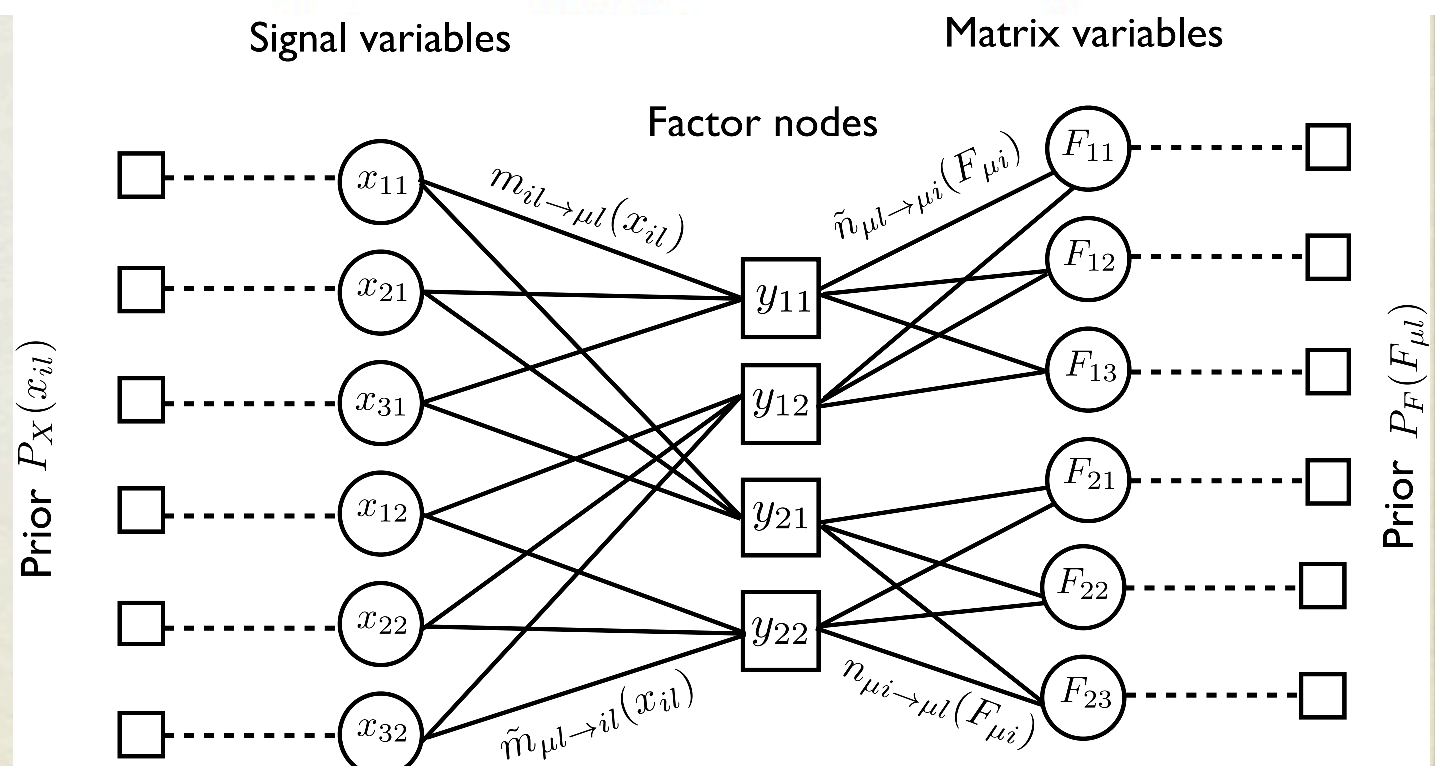
$$m_{il\to\mu l}(t+1,X_{il}) = \frac{1}{\mathcal{Z}_{il\to\mu l}} P_X(X_{il}) \prod_{\nu(\neq\mu)}^{N} \tilde{m}_{\nu l\to il}(t,X_{il})\,,$$

$$n_{\mu i\to\mu l}(t+1,F_{\mu i}) = \frac{1}{\mathcal{Z}_{\mu i\to\mu l}} P_F(F_{\mu i}) \prod_{n(\neq l)}^{P} \tilde{n}_{\mu n\to\mu i}(t,F_{\mu i})\,,$$

$$\tilde{m}_{\mu l\to il}(t,X_{il}) = \frac{1}{\mathcal{Z}_{\mu l\to il}} \int \prod_{j(\neq i)}^{R} \mathrm{d}X_{jl} \prod_{k}^{R} dF_{\mu k}\, P_{\mathrm{out}}(y_{\mu l}|\sum_{k}^{R} F_{\mu k}X_{kl}) \prod_{k}^{R} n_{\mu k\to\mu l}(t,F_{\mu k}) \prod_{j(\neq i)}^{R} m_{jl\to\mu l}(t,X_{jl})\,,$$

$$\tilde{n}_{\mu l\to\mu i}(t,F_{\mu i}) = \frac{1}{\mathcal{Z}_{\mu l\to\mu i}} \int \prod_{j}^{R} \mathrm{d}X_{jl} \prod_{k(\neq i)}^{R} dF_{\mu k}\, P_{\mathrm{out}}(y_{\mu l}|\sum_{k}^{R} F_{\mu k}X_{kl}) \prod_{k(\neq i)}^{R} n_{\mu k\to\mu l}(t,F_{\mu k}) \prod_{j}^{R} m_{jl\to\mu l}(t,X_{jl})$$

Not tractable .... each node many neighbors, incoming messages independent (by assumption), smells central limit theorem ....



Approximate message passing

- Physics-wise: AMP = TAP (Thouless, Anderson, Palmer'77) equations generalized to the present graphical model. Kabashima'04 for CDMA & perceptron (linear estimation).

- Approximate Message Passing (AMP) for linear estimation (firm rigorous foundations, non-Bayesian, continuous variables) by Donoho, Maleki, Montanari'09, Bayati, Montanari'11, Rangan'10, and many followers since.

- AMP in the present problem different from the one of linear estimation of low-rank factorization. Notably, not much known rigorously.

- For very nice applications-oriented work on AMP for matrix factorization see: BiG-AMP by Schniter, Parker, Cevher'13.

AMP FOR MATRIX FACTORIZATION

$$V^t_{\mu l} = \frac{1}{N}\sum_j [c_{jl}(t)s_{\mu j}(t) + c_{jl}(t)\hat{f}^2_{\mu j}(t) + \hat{x}^2_{jl}(t)s_{\mu j}(t)] \,,$$

$$\omega^t_{\mu l} = \frac{1}{\sqrt{N}}\sum_j \hat{x}_{jl}(t)\hat{f}_{\mu j}(t) - g_{\text{out}}(\omega^{t-1}_{\mu l}, y_{\mu l}, V^{t-1}_{\mu l})\frac{1}{N}\sum_j \left[\hat{f}_{\mu j}(t)\hat{f}_{\mu j}(t-1)c_{jl}(t) + \hat{x}_{jl}(t)\hat{x}_{jl}(t-1)s_{\mu j}(t)\right]$$

$$(\Sigma^t_{il})^{-1} = \frac{1}{N}\sum_\mu \left\{-\partial_\omega g_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l})\,\left[\hat{f}^2_{\mu i}(t) + s_{\mu i}(t)\right] - g^2_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l})\, s_{\mu i}(t)\right\} \,,$$

$$\begin{aligned} T^t_{il} = \Sigma^t_{il}\Big\{ &\frac{1}{\sqrt{N}}\sum_\mu g_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l})\,\hat{f}_{\mu i}(t) - \hat{x}_{il}(t)\frac{1}{N}\sum_\mu \hat{f}^2_{\mu i}(t)\partial_\omega g_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) \\ &-\hat{x}_{il}(t-1)\frac{1}{N}\sum_\mu s_{\mu i}(t) g_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) g_{\text{out}}(\omega^{t-1}_{\mu l}, y_{\mu l}, V^{t-1}_{\mu l})\Big\} \,, \end{aligned}$$

$$(Z^t_{\mu i})^{-1} = \frac{1}{N}\sum_l \left\{-\partial_\omega g_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l})\,\left[\hat{x}^2_{il}(t) + c_{il}(t)\right] - g^2_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l})\, c_{il}(t)\right\} \,,$$

$$\begin{aligned} W^t_{\mu i} = Z^t_{il}\Big\{ &\frac{1}{\sqrt{N}}\sum_l g_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l})\,\hat{x}_{il}(t) - \hat{f}_{\mu i}(t)\frac{1}{N}\sum_l \hat{x}^2_{il}(t)\partial_\omega g_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) \\ &-\hat{f}_{\mu i}(t-1)\frac{1}{N}\sum_l c_{il}(t) g_{\text{out}}(\omega^t_{\mu l}, y_{\mu l}, V^t_{\mu l}) g_{\text{out}}(\omega^{t-1}_{\mu l}, y_{\mu l}, V^{t-1}_{\mu l})\Big\} \,, \end{aligned}$$

$$\hat{x}_{il}(t+1) = f_X(\Sigma^t_{il}, T^t_{il}) \,, \qquad c_{il}(t+1) = f_c(\Sigma^t_{il}, T^t_{il}) \,,$$

$$\hat{f}_{\mu i}(t+1) = f_F(Z^t_{\mu i}, W^t_{\mu i}) \,, \qquad s_{\mu i}(t+1) = f_s(Z^t_{\mu i}, W^t_{i\mu}) \,.$$

GENERALITY OF AMP

- Only 3 quantities (function) in AMP are problem dependent:
- Input functions

$$f_X(\Sigma, T) \equiv \frac{\int \mathrm{d}X\, X\, P_X(X) e^{-\frac{(X-T)^2}{2\Sigma}}}{\int \mathrm{d}X\, P_X(X) e^{-\frac{(X-T)^2}{2\Sigma}}}$$

$$f_F(Z, W) \equiv \frac{\int \mathrm{d}F\, \sqrt{R}F\, P_F(F) e^{-\frac{(\sqrt{R}F-W)^2}{2Z}}}{\int \mathrm{d}F\, P_F(F) e^{-\frac{(F-W)^2}{2Z}}}$$

- Output function

$$g_{\mathrm{out}}(\omega, y, V) \equiv \frac{\int \mathrm{d}z P_{\mathrm{out}}(y|z)\,(z-\omega)\, e^{-\frac{(z-\omega)^2}{2V}}}{V \int \mathrm{d}z P_{\mathrm{out}}(y|z) e^{-\frac{(z-\omega)^2}{2V}}}$$

STATE EVOLUTION

- Physics-wise: Cavity method to derive RS solution from TAP.
- Rigorous for linear estimation, low rank factorization in Bayati, Montanari'11, Bayati, Lelarge, Montanari'15, Javanmard, Montanari'13. No proof yet for the present model.
- Define order parameters:

$$m_X^t \equiv \frac{1}{RP} \sum_{jl} \hat{x}_{jl}(t) X^*_{jl}$$

$$m_F^t \equiv \frac{1}{N\sqrt{R}} \sum_{\mu i} \hat{f}_{\mu i}(t) F^*_{\mu i} ,$$

- Track their evolution as AMP is iterated.

STATE EVOLUTION

$$m_X = \frac{1}{\sqrt{\alpha m_F \hat{m}}} \int \mathrm{d}t \frac{\left[f_1^X\left(\frac{t}{\sqrt{\alpha m_F \hat{m}}}, \frac{1}{\alpha m_F \hat{m}}\right)\right]^2}{f_0^X\left(\frac{t}{\sqrt{\alpha m_F \hat{m}}}, \frac{1}{\alpha m_F \hat{m}}\right)}$$

$$m_F = \frac{1}{\sqrt{\pi m_X \hat{m}}} \int \mathrm{d}t \frac{\left[f_1^F\left(\frac{t}{\sqrt{\pi m_X \hat{m}}}, \frac{1}{\pi m_X \hat{m}}\right)\right]^2}{f_0^F\left(\frac{t}{\sqrt{\pi m_X \hat{m}}}, \frac{1}{\pi m_X \hat{m}}\right)}$$

$$\hat{m} = \frac{1}{m_X m_F} \int \mathrm{d}y \int \mathcal{D}t \frac{\left[\partial_t f_0^Y(y|\sqrt{m_X m_F} t, \Gamma - m_X m_F)\right]^2}{f_0^Y(y|\sqrt{m_X m_F} t, \Gamma - m_X m_F)}$$

Generic

Problem dependent functions

$$f_n^X(T, \Sigma) \equiv \frac{1}{\sqrt{2\pi\Sigma}} \int \mathrm{d}X \, X^n P_X(X) e^{-\frac{(X-T)^2}{2\Sigma}}$$

$$f_n^F(W, Z) \equiv \frac{1}{\sqrt{2\pi Z}} \int \mathrm{d}F \, (\sqrt{R}F)^n P_F(F) e^{-\frac{(\sqrt{R}F-W)^2}{2Z}}$$

$$f_n^Y(y|\omega, V) \equiv \frac{1}{\sqrt{2\pi V}} \int \mathrm{d}t \, (t-\omega)^n P_{\mathrm{out}}(y|t) e^{-\frac{(t-\omega)^2}{2V}}$$

BOTTOM LINE

- State evolution of AMP gives the same expressions as the replica method.
- AMP-MSE is the local maximum of the free energy reached by state evolution initialized uninformatively.
- MMSE is the global maximum of the free energy.

EXAMPLE: DICTIONARY LEARNING

Also known as sparse coding:

$$Y \in \mathbb{R}^{N \times P}$$
$$\alpha = N/R$$
$$\pi = P/R$$

$$Y_{\mu i} = \sum_{\alpha=1}^{R} X_{\mu\alpha} F_{\alpha i} + W_{\mu i}$$

$$P_{\text{out}}(Y|Z) = \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{(Y-Z)^2}{2\Delta}}$$ Gaussian additive noise

$$P_X(X) = (1-\rho)\delta(X) + \rho\mathcal{N}(0,1)$$ Gauss-Bernoulli weights

$$P_F(F) = \mathcal{N}(0, 1/R)$$ Gaussian features

Free energy of dictionary learning

AMP-MSE is the local maximum of $\Phi(E_X, E_F)$ with largest EX, EF.

MMSE is the global maximum of $\Phi(E_X, E_F)$.

$$\Phi(E_X, E_F) = -\frac{\alpha}{2}\log\left(\Delta + E_X + E_F(\rho - E_X)\right) - \frac{\alpha(\Delta+\rho)}{\Delta + E_X + E_F(\rho - E_X)} + \frac{\alpha}{2}$$
$$+ \left[\int \mathcal{D}z \log\left[e^{-\frac{\hat{m}_x}{2}x^2 + \hat{m}_x x x^0 + z\sqrt{\hat{m}_x}x}\right]_{P_X(x)}\right]_{P_X(x^0)}$$
$$+ \frac{\alpha}{\pi}\left[\int \mathcal{D}z \log\left[e^{-\frac{R\hat{m}_F F^2}{2} + R\hat{m}_F F F^0 + z\sqrt{N\hat{m}_F}F}\right]_{P_F(F)}\right]_{P_F(F^0)}$$

$$\hat{m}_x = \frac{\alpha(1 - E_F)}{\Delta + E_X + \rho E_F - E_X E_F} \qquad \hat{m}_F = \frac{\pi(\rho - E_X)}{\Delta + E_X + \rho E_F - E_X E_F}$$

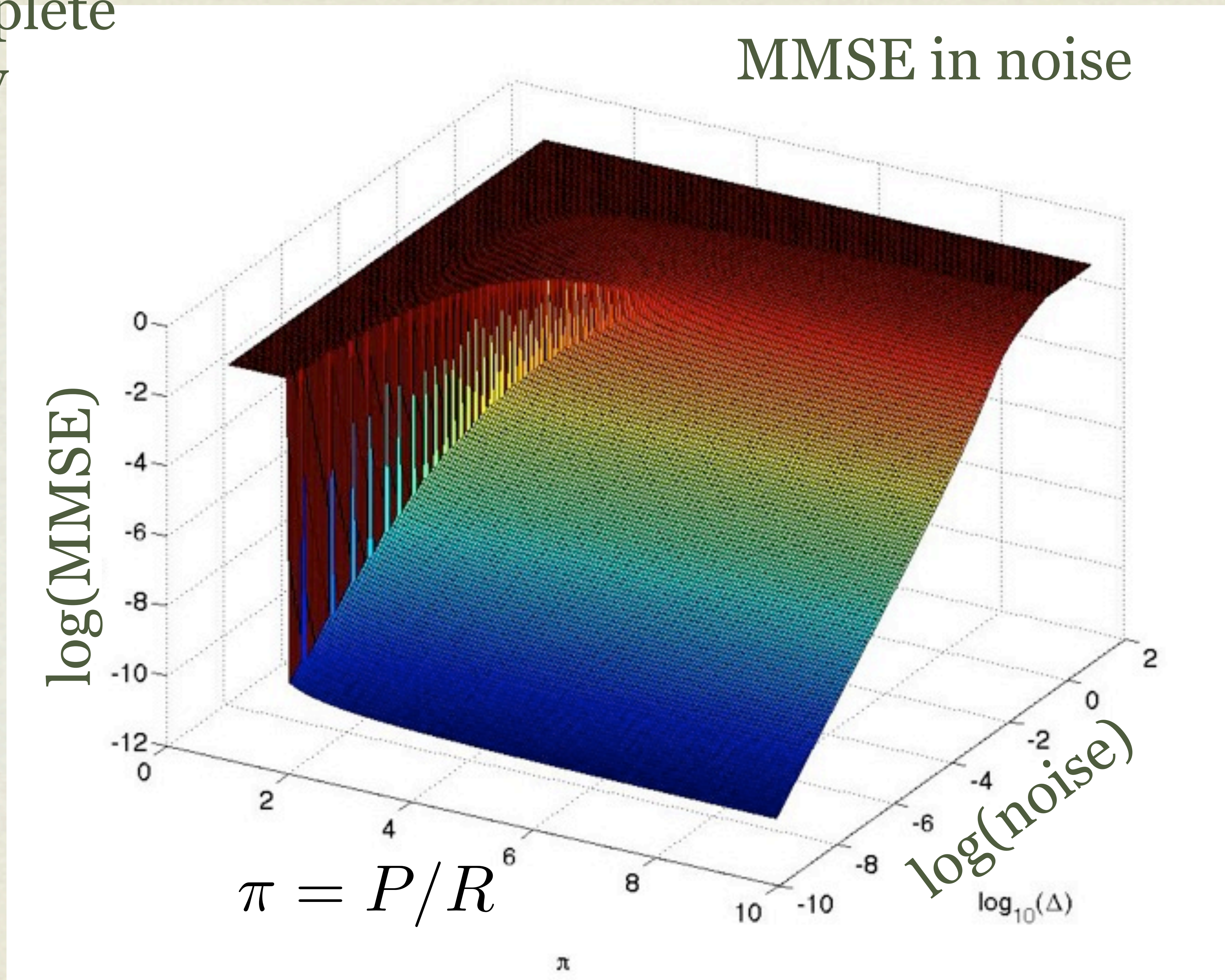
The phase diagram of dictionary learning

over-complete dictionary

$R = 2N$

sparsity

$\rho = 0.2$



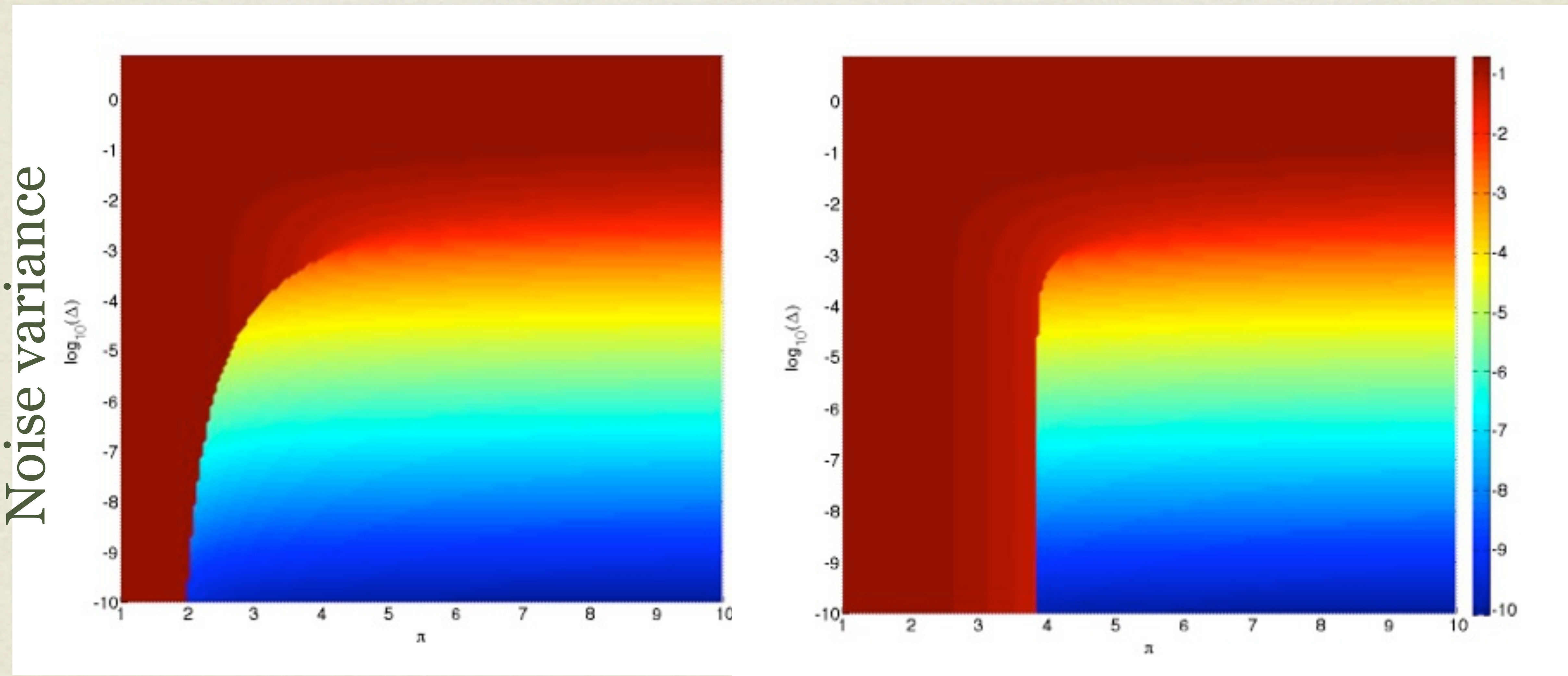
The phase diagram of dictionary learning

$R = 2N \qquad \rho = 0.2$

MMSE

AMP-MSE

Noise variance



Sample complexity $P/(2N)$

Lower bound for the noiseless case and continuous variables

$$Y_{\mu i} = \sum_{\alpha=1}^{R} X_{\mu\alpha} F_{\alpha i}$$

$$P_X(X) = (1-\rho)\delta(X) + \rho\mathcal{N}(0,1)$$

$$P_F(F) = \mathcal{N}(0, 1/R)$$

Number of knowns >= number of unknowns

$$PN \geq \rho PR + RN$$

$$\alpha\pi \geq \rho\pi + \alpha$$

$$\pi \geq \frac{\alpha}{\alpha - \rho}$$

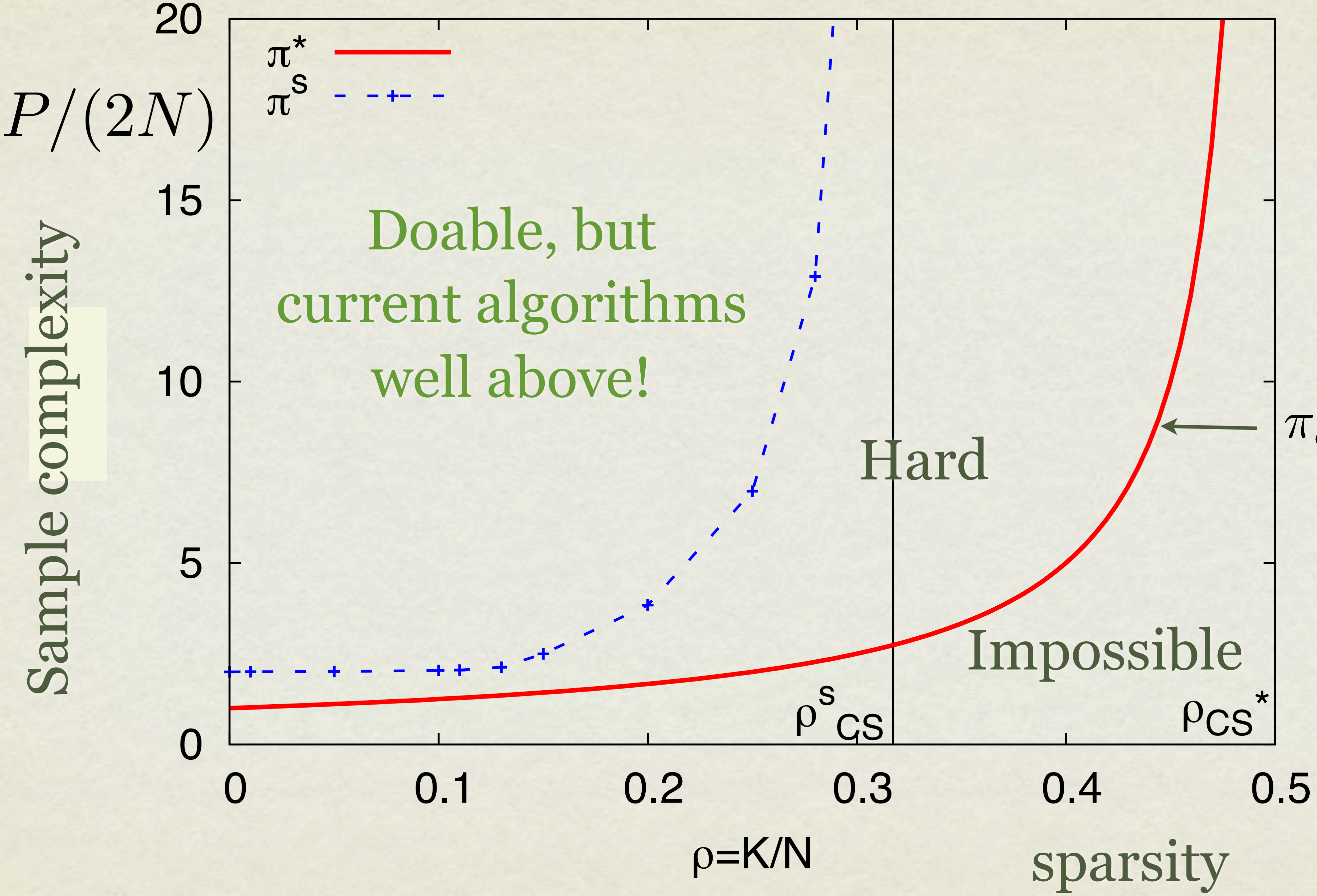
Sample complexity of dictionary learning

$R = 2N$

$\rho = 0.2$

$\Delta = 0$

$\pi = P/(2N)$



$\pi_c = \frac{\alpha}{\alpha - \rho}$

CONCLUSIONS

- Teacher-student matrix factorization with general output as a simple model for feature learning. Also model for dictionary learning, blind source separation, sparse PCA, robust PCA, ...

- Invertron: Model for structured data. Useful for benchmarking of algorithms, and as insight into theoretical understanding of feature learning.

- Exact formula for the MMSE. Its evaluation suggests that current state-of-the-art algorithms have large gap to optimality.

- Reading: Kabashima, Krzakala, Mezard, Sakata, LZ, arXiv:1402.1298. Schniter, Parker, Cevher'13 for the algorithmic applications.

TO DO LIST

- Math: Prove that the state evolution is correct.
- Math: Proving the detectability lower bound is tight in the noiseless planted matrix factorization.
- Math, CS: Which other algorithms (provably and empirically) work down to the AMP phase transition?
- Ph, CS: Robust and simple implementation of AMP (so far convergence issues, instabilities ...)
- Ph: Replica symmetry breaking when prior does not match the model, or when we want a ground state.
- Ph: Generalize to non-separable priors, more layers, tensors, ...

AMP for matrix factorization