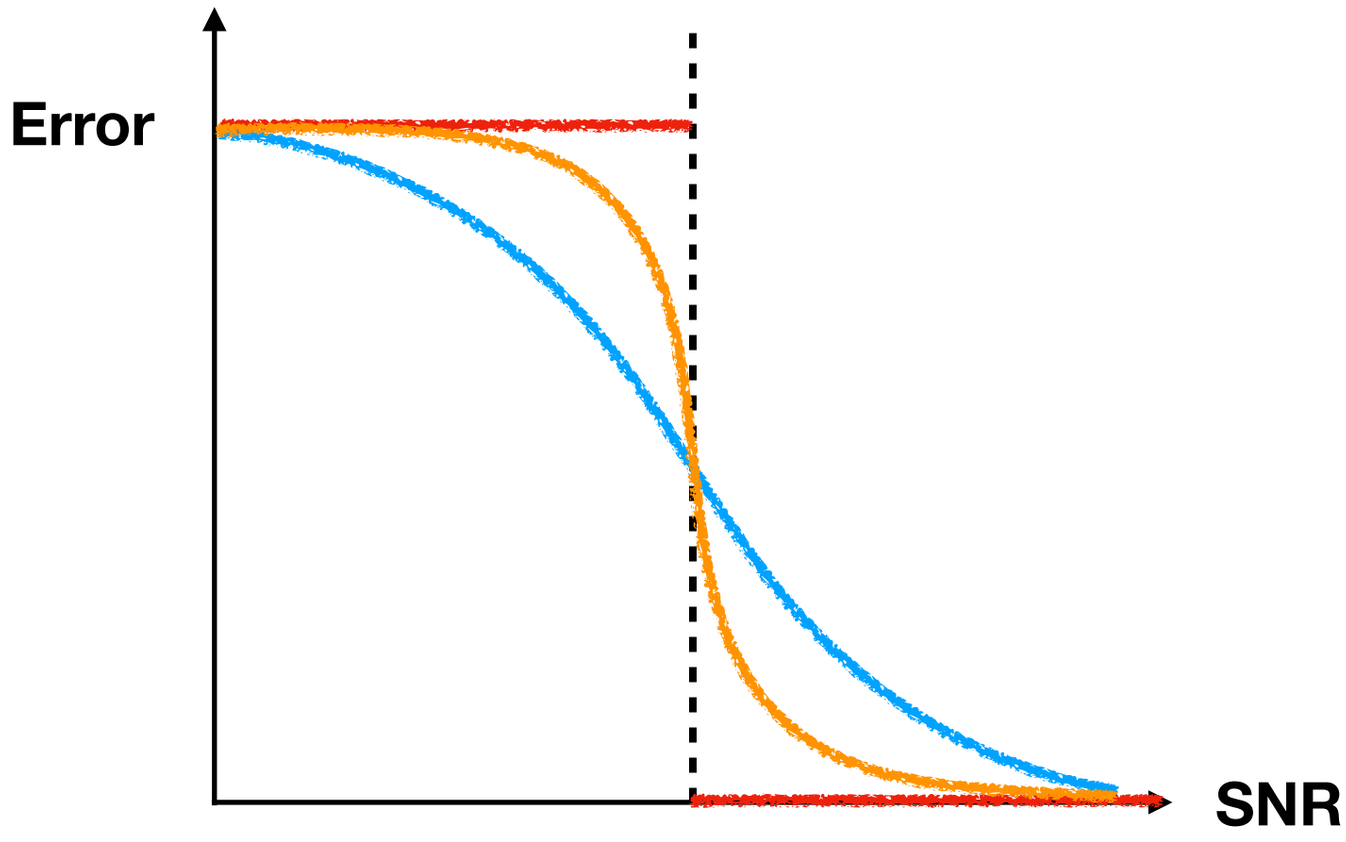
The « all-or-nothing phenomenon »

A peculiar phase transition in sparse high-dimensional inference & learning

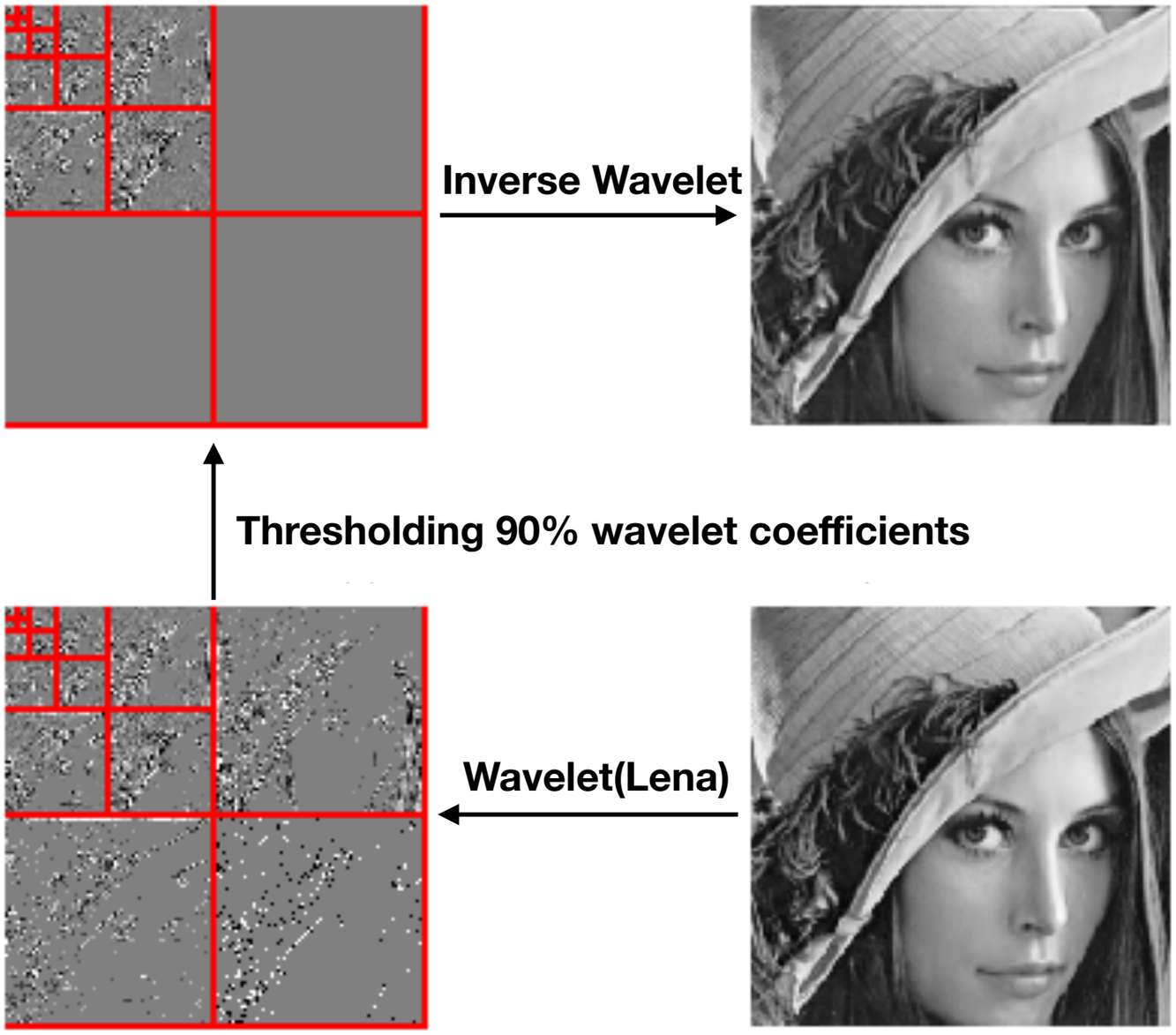
Jean Barbier

International Center for Theoretical Physics, Trieste



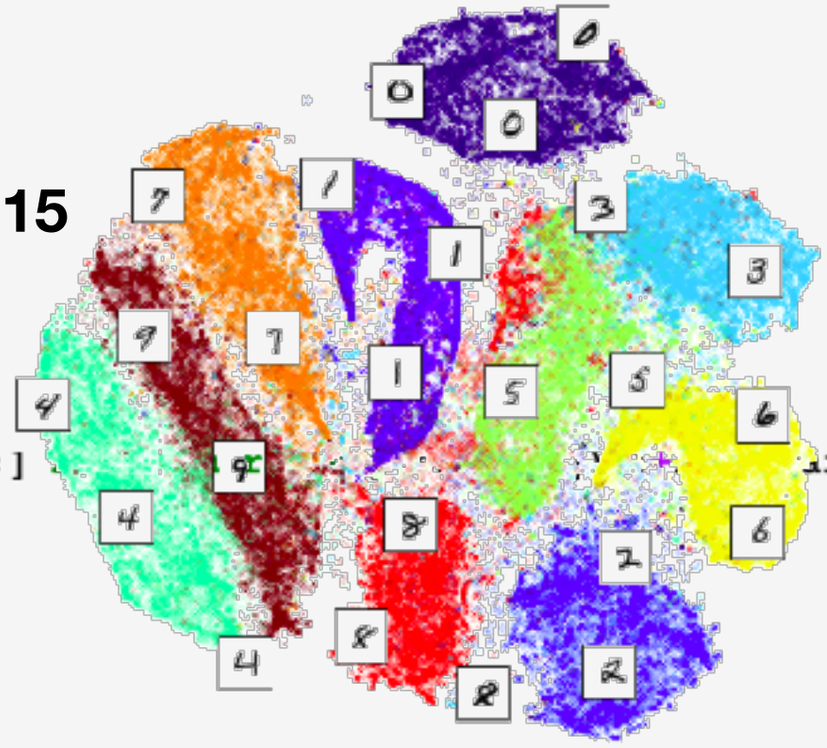
Joint with **Nicolas Macris** (EPFL), **Clément Luneau** (EPFL) & **Cynthia Rush** (Columbia University)

Low-dimensional manifolds underlying high-dimensional data



Natural Images
d~1e5
Intrinsic d~100

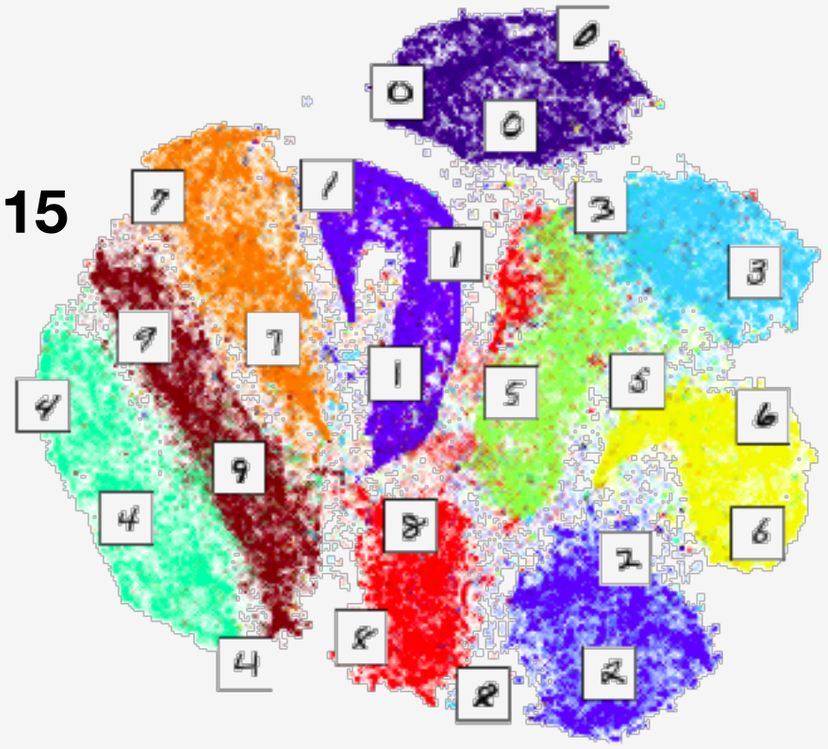
MNIST
d=784
Intrinsic d~10—15



```python
# logistic function
def logistic(x):
    return 1/(1+np.exp(-x))

# classification error
def class_error(X,y,params):
    return 100*np.mean(np.abs((logistic(np.dot(X,params))>0.5)-y))

# gradient of logistic regression with L2 penalization
def grad_log(X,y,params,lamb):
    return np.sum(np.array([(logistic(np.dot(X[i,:],params))-y[i])*X[i,:]

# divide the data in mini batches
def mini_batch(X,y,size_batch,number_batch):
    perm=np.random.permutation(X.shape[0])
    randX=X[perm,:]
    randy=y[perm]

    X_mini=np.array([randX[i*size_batch:min(i*size_batch+size_batch,X.shape[0]),:] for i in range(number_batch)])
    y_mini=np.array([randy[i*size_batch:min(i*size_batch+size_batch,X.shape[0])] for i in range(number_batch)])
    return X_mini,y_mini

# SGD with momentum
def SGD(X,y,grad,init,epochs=20,learn_rate=10**-4,momentum=0.9,L2_pen=0.01,size_batch=10,ground_truth=0):
    params=np.array(init)
    v=0
    number_batch=int(np.ceil(X.shape[0]/size_batch))

    for j in range(epochs):
        X_mini,y_mini=mini_batch(X,y,size_batch,number_batch)

        for k in range(number_batch):
            v=momentum*v+learn_rate*np.array(grad(X_mini[k],y_mini[k],params,L2_pen))
            params=params-v

        train_er=class_error(X,y,params)
        print(['epoch',j,'train error',train_er])

    return params
```

```
n_train=5000
n_test=8000
learn_rate=1e-4
size_batch=20
L2_pen=0.01
epochs=20

# select subset of two numbers
number_1='1'
number_2='9'

##################################

N=784 # number of pixels/features

# select subset of two numbers
select=np.where((y_mnist==number_1)|(y_mnist==number_2))
X,y=np.array(X_mnist[select[0],:]),np.array(y_mnist[select[0]])
y[y==number_1]=0
y[y==number_2]=1
y=y.astype('float')

# pick training and test data sets
perm=np.random.permutation(X.shape[0])
X,y=X[perm,:],y[perm]
X_train,X_test,y_train,y_test=train_test_split(X,y,train_size=n_train,test_size=n_test)

# standardize data
scaler=StandardScaler()
X_train=scaler.fit_transform(X_train)
X_test=scaler.transform(X_test)

# SGD
params=SGD(X_train,y_train,grad_log,np.zeros(N),epochs=epochs,learn_rate=learn_rate,momentum=0.9,L2_pen=L2_pen,size

# check accuracy
print(['train: % misclassified', class_error(X_train,y_train,params)])
print(['test: % misclassified', class_error(X_test,y_test,params)])
```

MNIST
d=784
Intrinsic d~10—15



```
['epoch', 17, 'train error', 0.24]
['epoch', 18, 'train error', 0.24]
['epoch', 19, 'train error', 0.24]
['train: % misclassified', 0.24]
['test: % misclassified', 0.525]
```

“Bet on sparsity principle”

Intrinsic / effective low-dimensionality is often a crucial ingredient for the interpretability of high-dimensional data

Statistical physics approach:

Idealised models of high-dimensional data with low intrinsic / effective dimension

- Average case / ensemble analysis -> typical behaviour
- Information-theoretic and algorithmic phase transitions
- Exact asymptotic thresholds and formulas in « high-dimensional » regime

  free energy / mutual information, minimum mean-square and/or generalisation errors

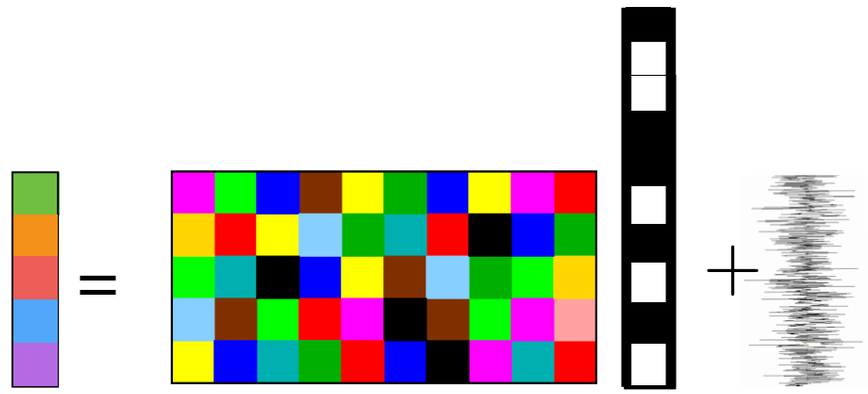
Paradigmatic models for high-dimensional inference & learning

Random linear estimation / compressive sensing

[Barbier Macris Dia Krzakala 16], [Reeves Pfister 16]

$$\mathbf{W} = \sqrt{\frac{\lambda}{n}}\mathbf{\Phi}\mathbf{X} + \mathbf{Z}$$

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



$\mathbf{W}$ $\mathbf{\Phi}$ $\mathbf{X}$?

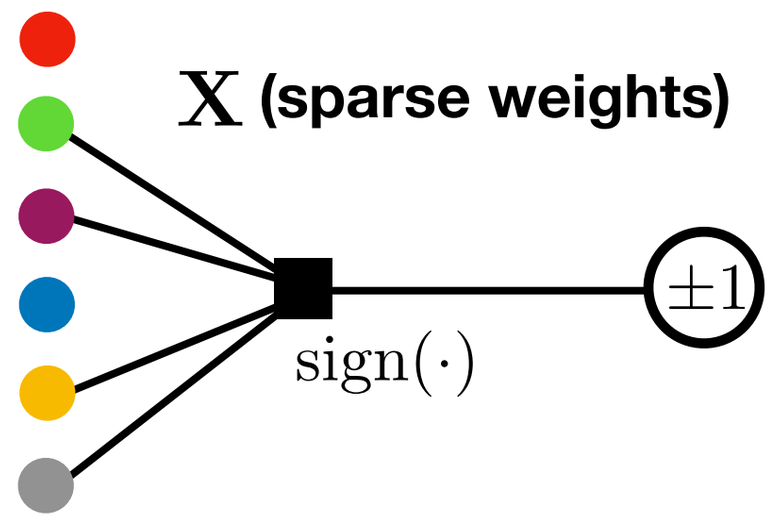
Generalized linear models

[Barbier Miolane Macris Krzakala Zdeborovà 18]

$$\mathbf{W} = \varphi\left(\sqrt{\frac{1}{n}}\mathbf{\Phi}\mathbf{X}\right)$$

Perceptron neural net $\mathbf{W} = \mathrm{sign}(\mathbf{\Phi}\mathbf{X})$

« Teacher-student » scenario
= realisable rule setting in ML



Signal processing

$$\mathbf{Y} = \mathbf{\Phi X} + \mathbf{Z}\sqrt{\Delta}$$

Compressed sensing (signal processing)
CDMA (multi-user communication theory)
Superposition codes (coding theory)

$$\mathbf{Y} = |\mathbf{\Phi X}|$$

Phase retrieval (signal processing)

$$\mathbf{Y} = \mathrm{sign}(\mathbf{\Phi X} + \mathbf{Z}\sqrt{\Delta})$$

1-bit compressed sensing (signal processing)
Perceptron (statistics/ML)

Learning

$$\mathbf{Y} = \max(0, \mathbf{\Phi X})$$

Rectified Linear Unit (ReLU) (ML/neural nets)

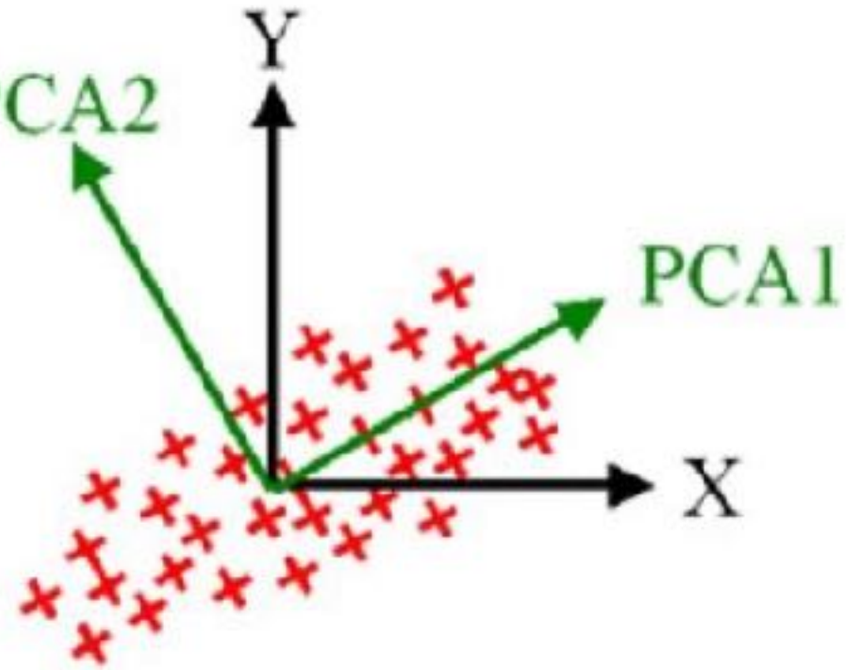
$$P(Y_\mu = 1) = \frac{1}{1 + \exp(-\lambda \mathbf{\Phi}_\mu \cdot \mathbf{X})}$$

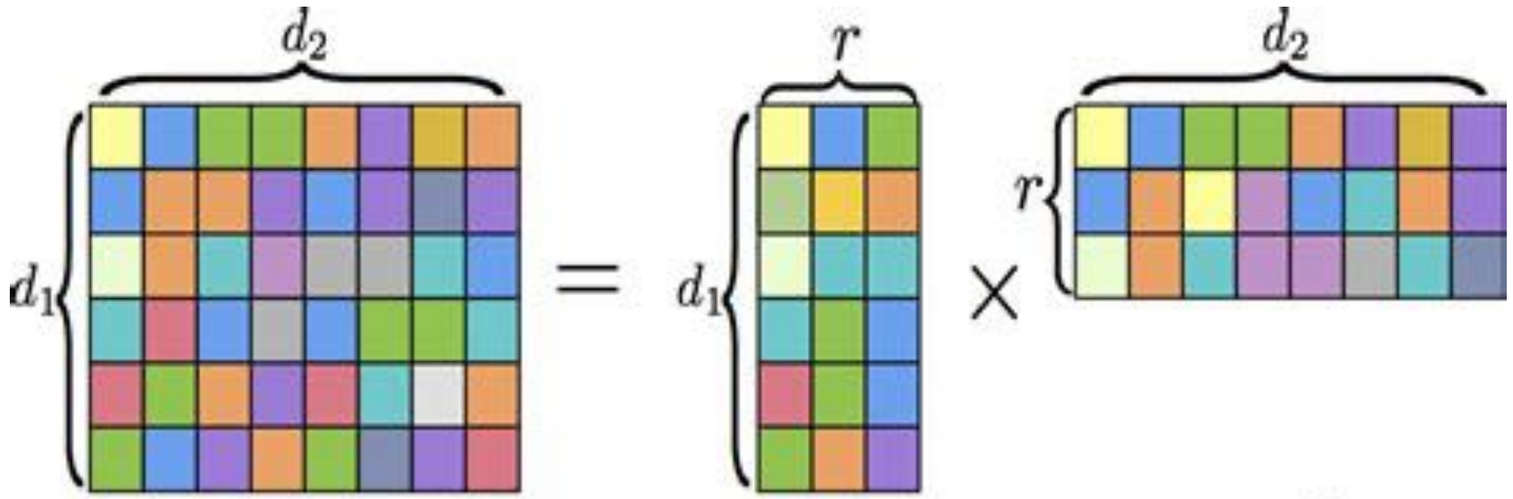
Sigmoid/Logistic regression (ML/neural nets)

Low-rank matrix factorisation
Sparse principal components analysis

[Korada Macris 09]
[Deshpande Montanari 14]
[Barbier Macris Dia Lesieur Krzakala Zdeborová 16]
[Miolane Lelarge 17]

$$\mathbf{W} = \sqrt{\frac{\lambda}{n}}\mathbf{X}\mathbf{X}^\intercal + \mathbf{Z}$$



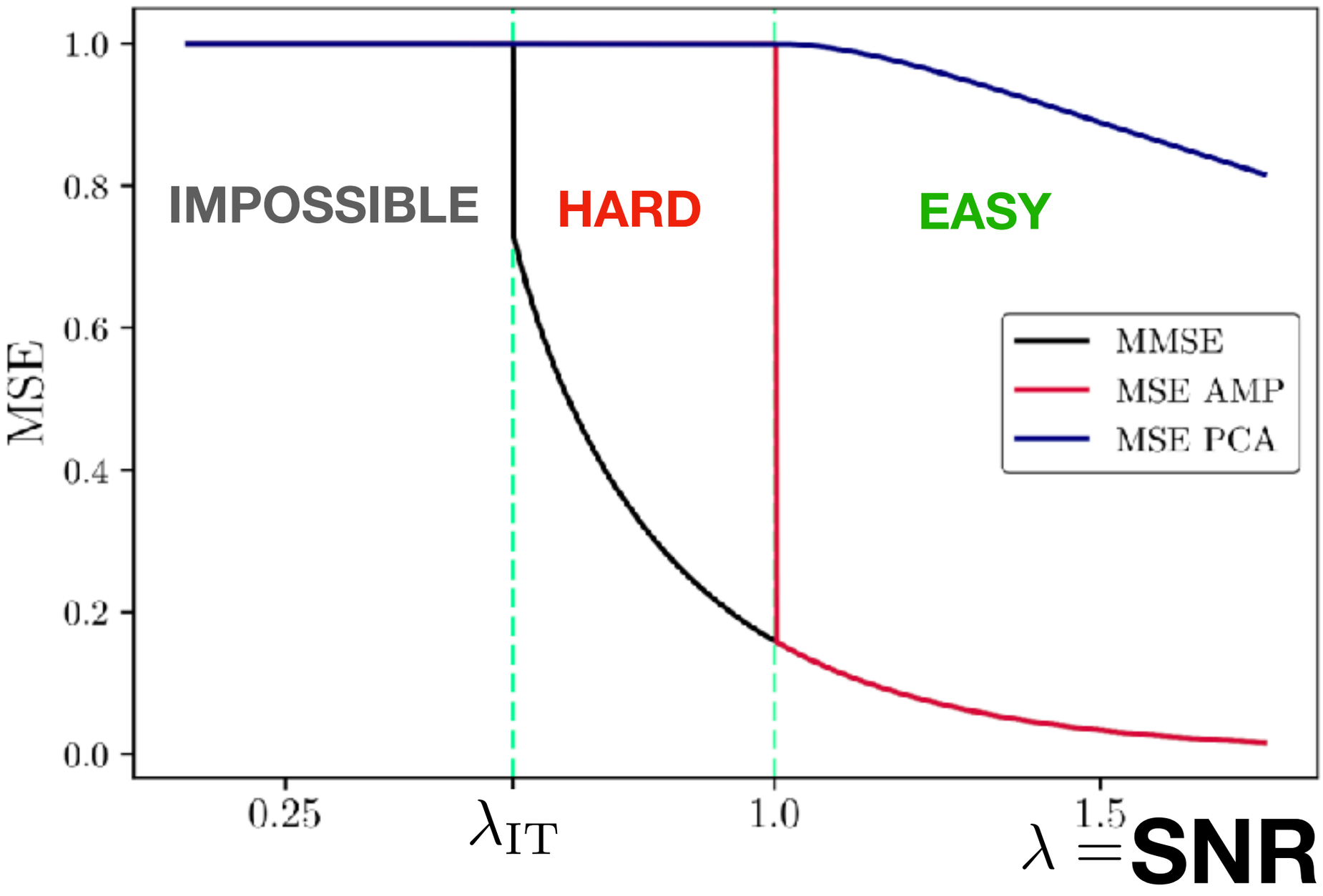


Generic scenario: discontinuous phase transition & computational gap

$$\mathbf{W} = \sqrt{\frac{\lambda}{n}}\mathbf{X}\mathbf{X}^\intercal + \mathbf{Z}$$

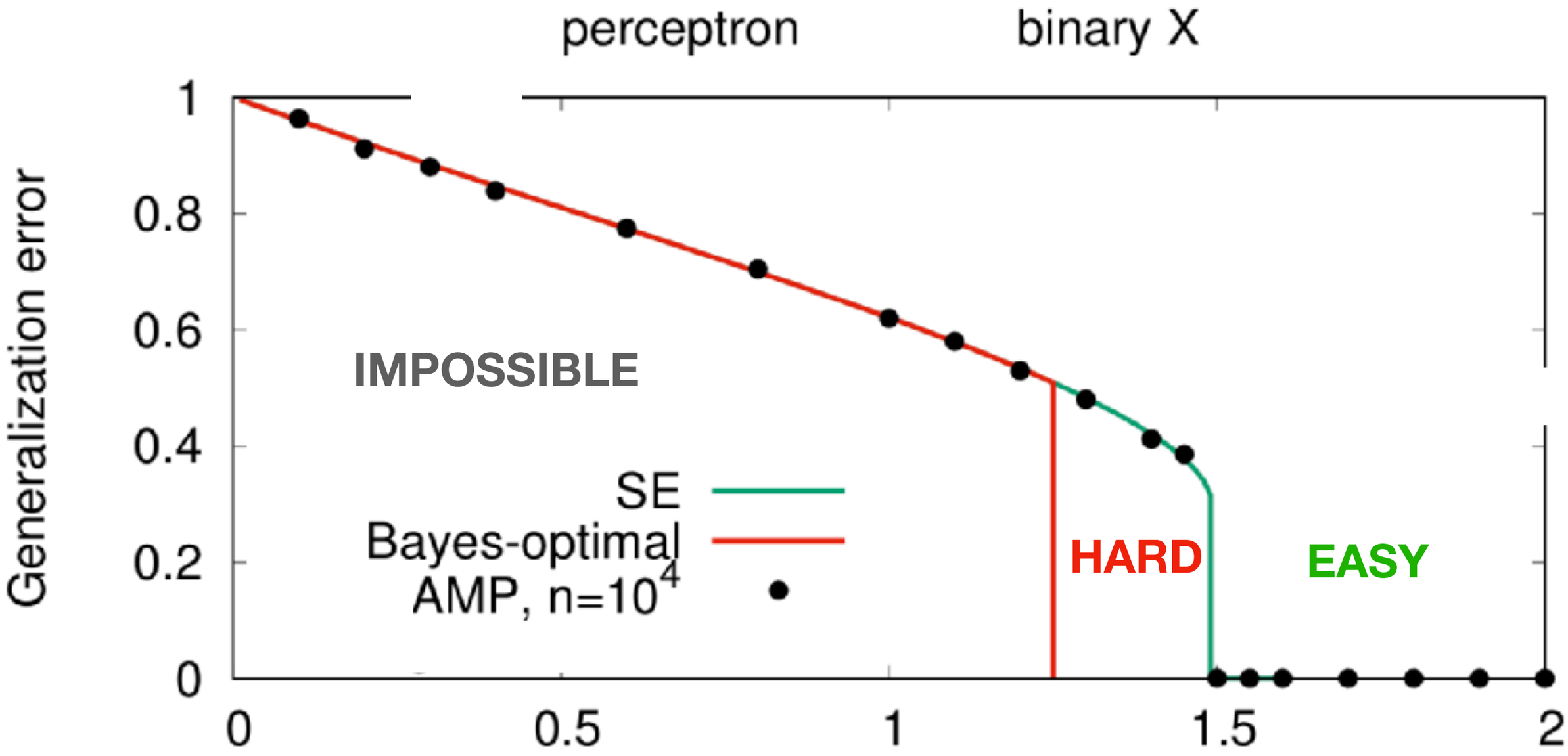
Binary X

$$\lim_{n\to\infty} \frac{1}{n^2}\mathbb{E}\|\mathbf{X}\mathbf{X}^\intercal - \mathbb{E}[\mathbf{X}\mathbf{X}^\intercal|\mathbf{W}]\|_{\mathrm{F}}^2$$

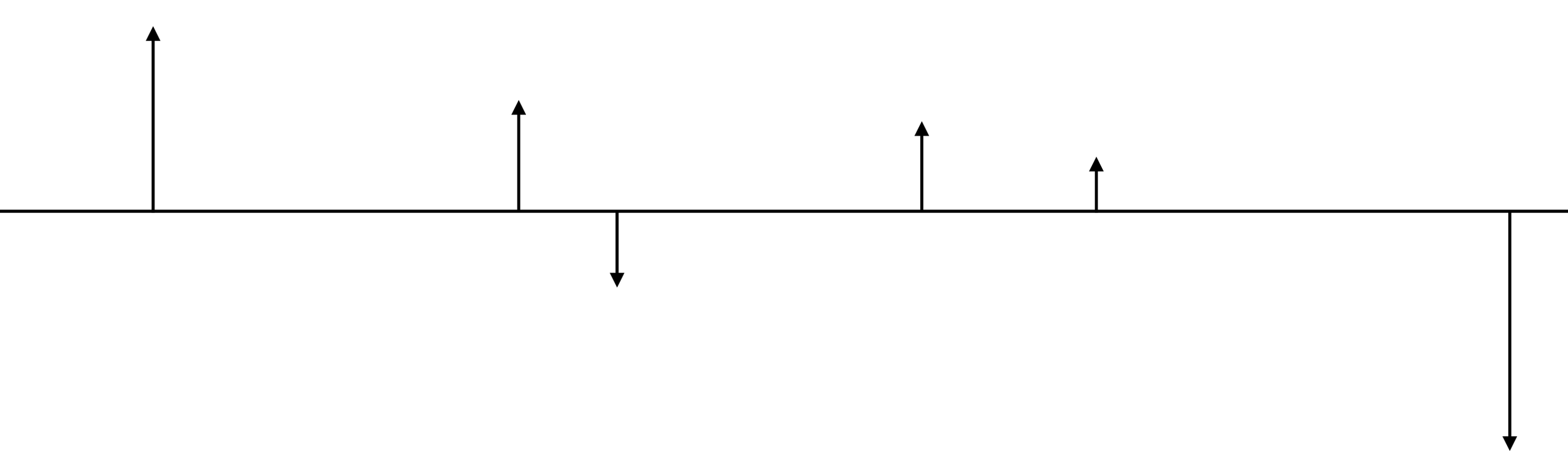


Generic scenario: discontinuous phase transition & computational gap

$$\mathbf{W} = \mathrm{sign}(\mathbf{\Phi X}) \qquad \lim_{n\to\infty} \mathbb{E}\left[\left(W_{\mathrm{new}} - \mathbb{E}[W_{\mathrm{new}}|\boldsymbol{W}, \mathbf{\Phi}, \mathbf{\Phi}_{\mathrm{new}}]\right)^2\right]$$



Anything special in the very high sparsity / low effective dimension regime?



Compressed sensing: low measurement-rate regime

$$W_\mu = \sqrt{\frac{\lambda}{n}} \mathbf{\Phi}_\mu \cdot \mathbf{X} + Z_\mu \quad \mu = 1, \dots, m$$

$$X_i \sim \mathrm{Ber}(\rho) \quad i = 1, \dots, n$$

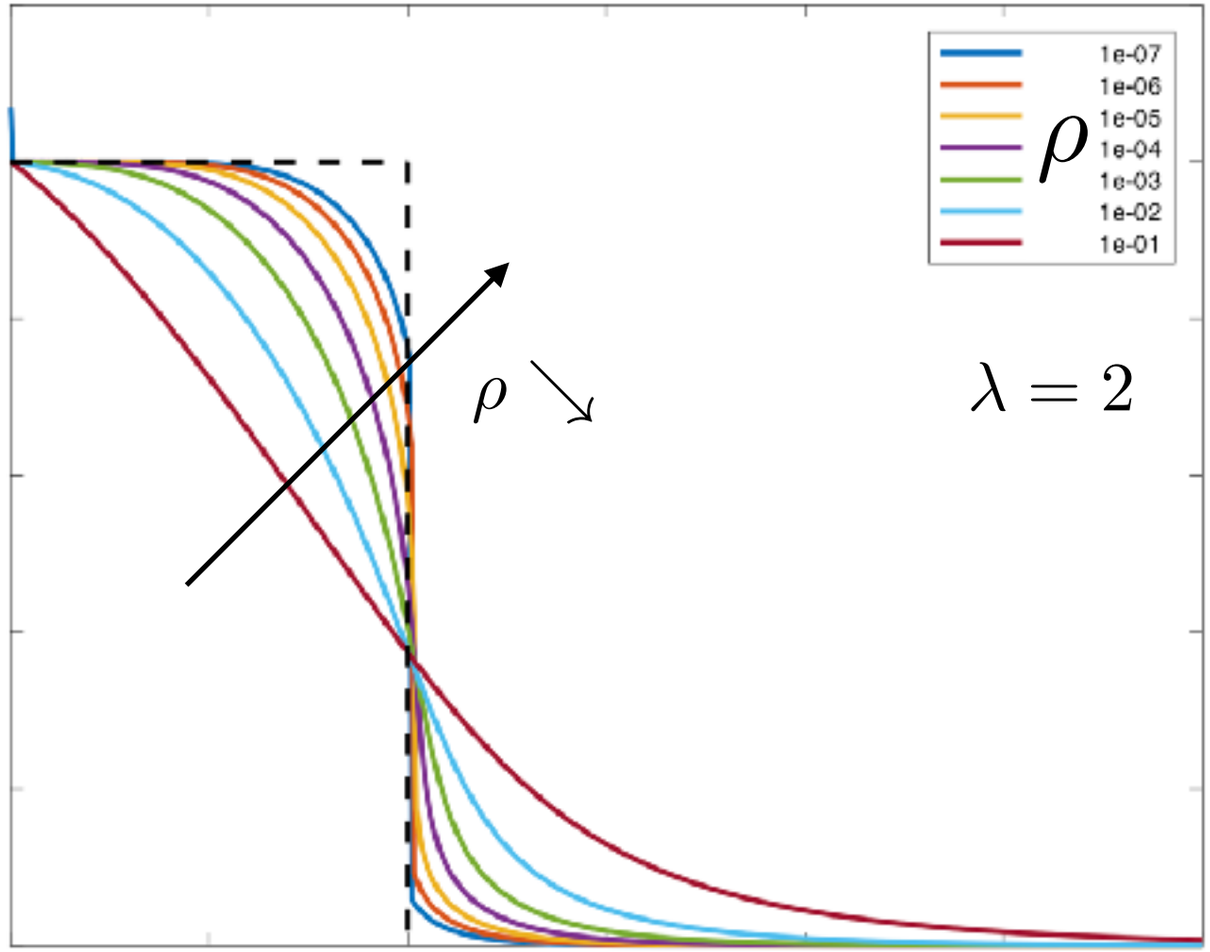
[Gamarnik Zadik 17]

[Reeves Xu Zadik 19]

(informal) Theorem [Barbier Macris Dia Krzakala 16] [Reeves Pfister 16]:

$$\lim_{n\to\infty} \frac{1}{\rho n} \mathbb{E} \|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{W}]\|^2 = \frac{1}{\rho} \mathrm{mmse}(q^*, \lambda)$$

$$\lim_{n\to\infty} \frac{1}{\rho n} \mathbb{E} \|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{W}]\|^2$$



$$W_\mu = \sqrt{\frac{\lambda_n}{\rho_n n}}\mathbf{\Phi}_\mu \cdot \mathbf{X} + Z_\mu \quad \mu = 1, \dots, \alpha_n n$$

$$X_i \sim \mathrm{Ber}(\rho_n) \quad i = 1, \dots, n \qquad \lambda_n = \Omega(1)$$

(informal) Theorem [Reeves Xu Zadik 19]: « All-or-nothing » phenomenon

$$\rho_n = o(n^{-1/2}) \text{ and } \alpha < \alpha_{\mathrm{IT},n} \text{ with } \alpha_{\mathrm{IT},n} = \frac{2\rho_n |\ln \rho_n|}{\ln(1+\lambda_n)}$$

⟶ **Weak recovery impossible:** $\liminf_{n\to\infty} \frac{1}{\rho_n n}\mathbb{E}\|\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{opt}}\|^2 = 1$

$$\rho_n = o(1) \text{ and } \alpha > \alpha_{\mathrm{IT},n}$$

⟶ **Strong recovery possible:** $\limsup_{n\to\infty} \frac{1}{\rho_n n}\mathbb{E}\|\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{opt}}\|^2 = 0$

All-or-nothing in sparse principal components analysis

$$\mathbf{W} = \sqrt{\frac{\lambda_n}{n}}\mathbf{X}\otimes\mathbf{X} + \mathbf{Z}$$

$$X_i \sim \mathrm{Ber}(\rho_n) \quad i = 1, \dots, n$$

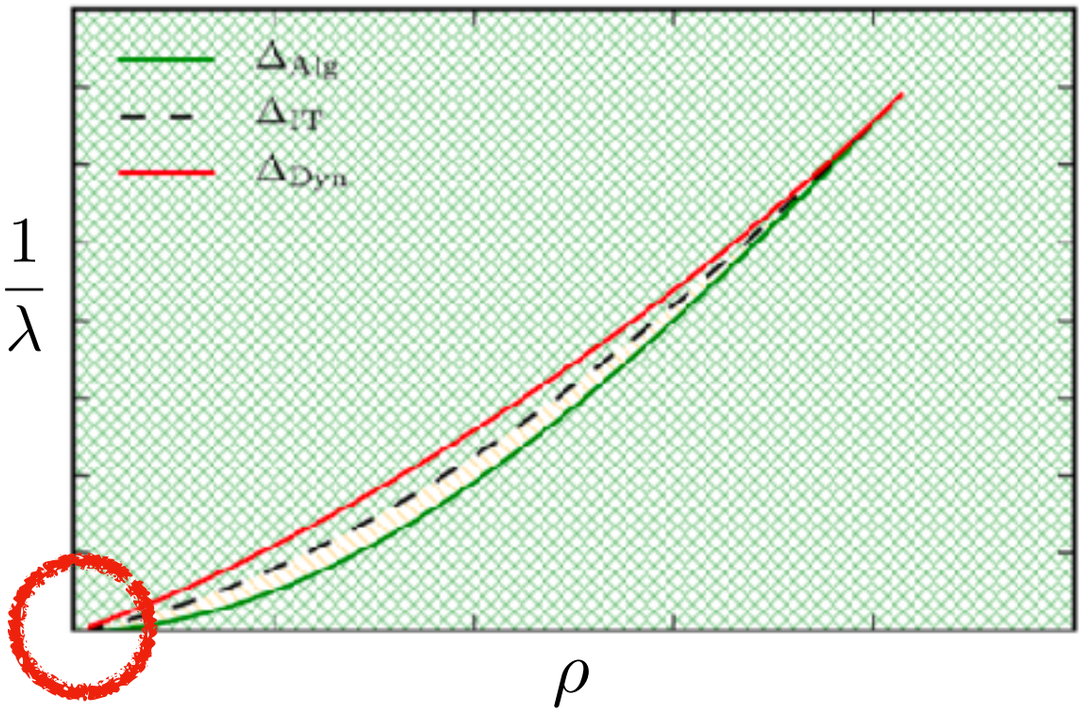
$$\rho_n \to 0_+$$

$$\lambda_n \to +\infty$$

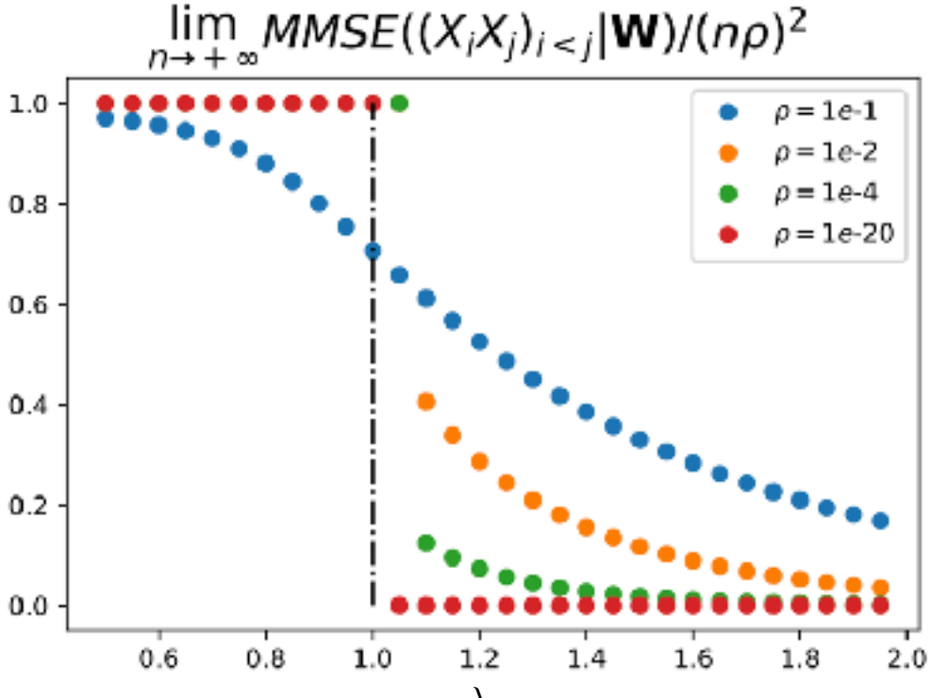
Theorem [Korada Macris 09], [Deshpande Montanari 14], [Barbier Macris Dia Lesieur Krzakala Zdeborová 16], [Miolane Lelarge 17]:

$$\lim_{n\to\infty} \frac{1}{(\rho n)^2}\mathbb{E}\|\mathbf{X}\otimes\mathbf{X} - \mathbb{E}[\mathbf{X}\otimes\mathbf{X}|\mathbf{W}]\|_{\mathrm{F}}^2 = \mathrm{mmse}(q^*)$$

$$\Delta = \frac{1}{\lambda}$$



[Lesieur Krzakala Zdeborová 15]



$$\lambda_{\mathrm{IT}}(\rho) \xrightarrow{\rho\to 0} \frac{4|\ln\rho|}{\rho}$$

Finite size bound for the mutual information

Theorem [Barbier Macris Rush 20]: based on the adaptive interpolation method

Let $P_{0,n} = \rho_n\, p_0 + (1-\rho_n)\delta_0$. Let $\lambda_n = \Theta(|\ln \rho_n|/\rho_n)$ which is the appropriate scaling to observe a phase transition, and $\rho_n = \Theta(n^{-\beta})$ with $\beta \in [0, 1/6)$ Then

$$\frac{1}{\rho_n|\ln\rho_n|}\Big|\frac{1}{n}I(\mathbf{X};\mathbf{W}) - \inf_{q\in[0,1]} i_n^{\rm pot}(q;\lambda_n,\rho_n)\Big| \le C\frac{(\ln n)^{1/3}}{n^{(1-6\beta)/7}}$$

$$\text{with}\quad i_n^{(\rm pot)}(q;\lambda_n,\rho_n) = \frac{\lambda_n\rho_n^2}{4}(q-1)^2 + I_n(X;\sqrt{\lambda_n\rho_n q}\,X+Z)\quad X\sim P_{0,n},\ Z\sim\mathcal{N}(0,1).$$

Corollary [Barbier Macris Rush 20]:

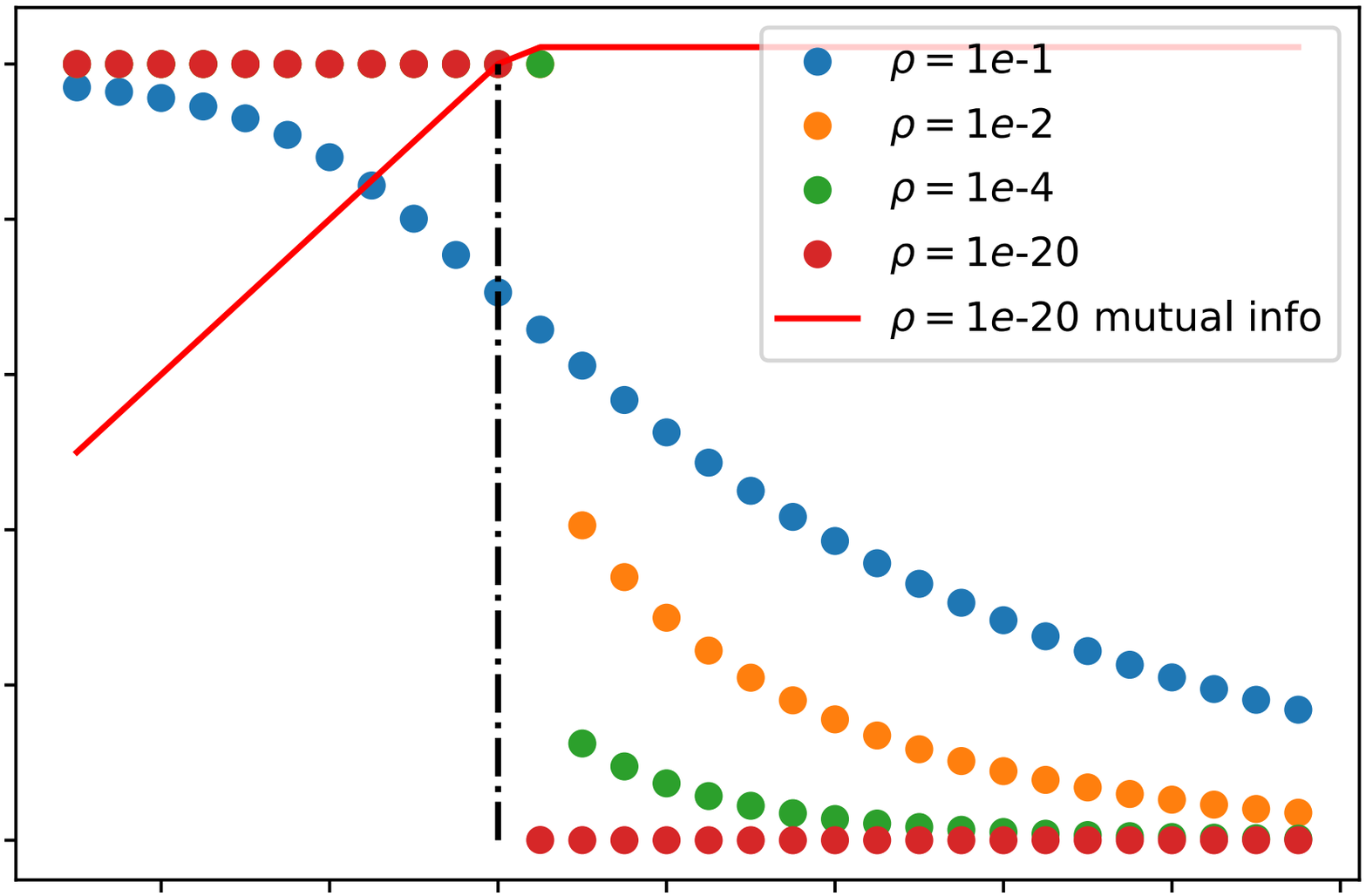
Let $\frac{1}{2}m_n(\lambda,\rho_n) \equiv \rho_n^{-2}\frac{d}{d\lambda}\inf_{q\in[0,\rho_n]} i_n^{\rm pot}(q,\lambda,\rho_n)$. Let $\epsilon>0$ and $\beta\in[0,1/13)$. There exists $C'>0$ independent of n such that

$$m_n(\lambda_n+\epsilon,\rho_n) - \frac{C'}{\epsilon}\frac{(\ln n)^{4/3}}{n^{(1-13\beta)/7}} \le \frac{\mathrm{MMSE}((X_iX_j)_{i<j}|\boldsymbol{W})}{(n\rho_n)^2} \le m_n(\lambda_n-\epsilon,\rho_n) + \frac{C'}{\epsilon}\frac{(\ln n)^{4/3}}{n^{(1-13\beta)/7}}.$$

Information-theoretic all-or-nothing transition in sparse PCA

$$\text{Setting } \lambda_n \equiv \gamma \lambda_{n,\mathrm{IT}}(\rho_n) \text{ where } \lambda_{n,\mathrm{IT}}(\rho_n) \equiv 4\frac{|\ln \rho_n|}{\rho_n}$$

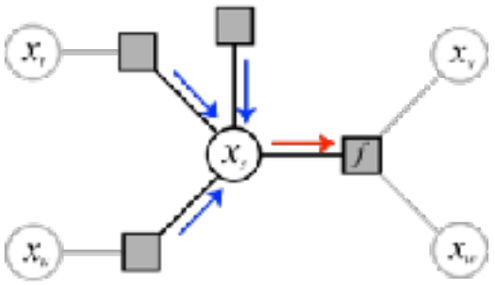
$$\frac{1}{(n\rho_n)^2}\mathbb{E}\|\mathbf{X}\otimes\mathbf{X} - \mathbb{E}[\mathbf{X}\otimes\mathbf{X}|\mathbf{W}]\|_{\mathrm{F}}^2 \xrightarrow[n\to\infty]{} \mathbb{I}(\gamma \le 1)$$



Algorithmic all-or-nothing transition in sparse PCA

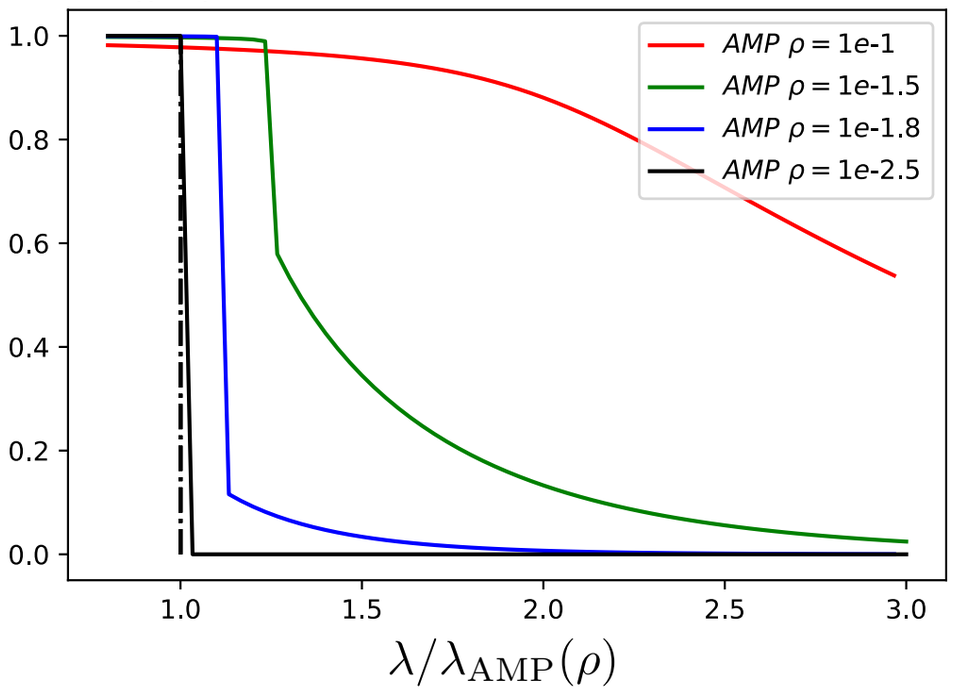
Belief Propagation (BP) / Sum-product algo for DENSE factor graphs:
Approximate Message-Passing algo (AMP), originally introduced for CS

AMP: $$\boldsymbol{x}^{t+1} = \frac{1}{\sqrt{n}}\boldsymbol{W} f_t(\boldsymbol{x}^t) - \mathsf{b}_t f_{t-1}(\boldsymbol{x}^{t-1}), \qquad \mathsf{b}_t = \frac{1}{n}\sum_{i=1}^{n} f_t'(x_i^t), \qquad \hat{\boldsymbol{X}}^t = f_t(\boldsymbol{x}^t)$$

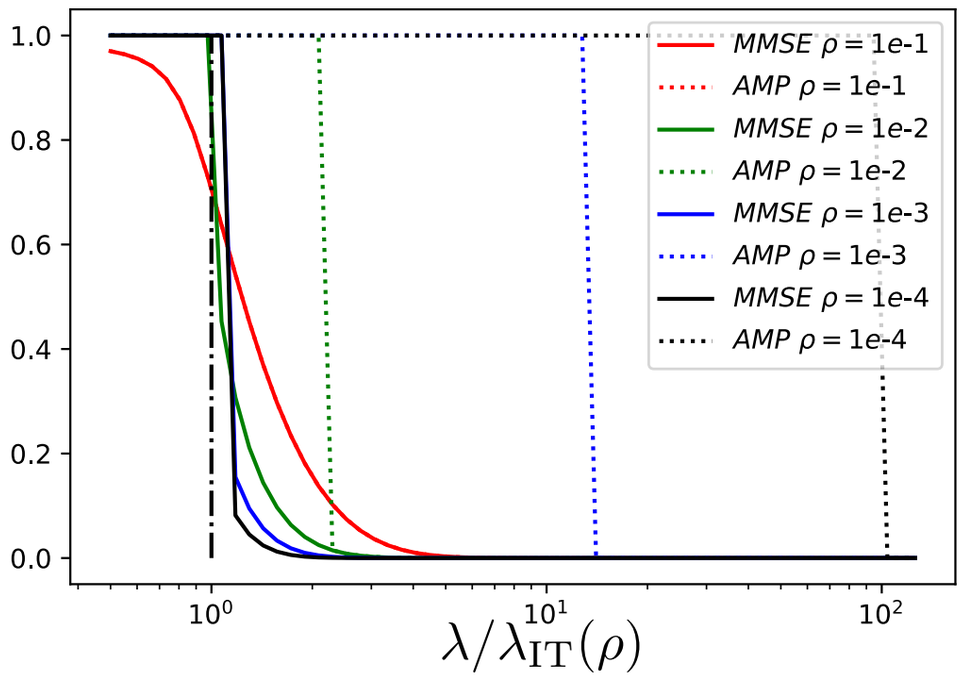


- Exact characterisation in a proper asymptotic limit thanks to **STATE EVOLUTION** [Bayati, Montanari 12']
- **Conjectured to be optimal among all polynomial time algo**

$\lim_{n\to+\infty} MSE_{AMP}((X_iX_j)_{i<j}|\mathbf{W})/(n\rho)^2$



MMSE vs MSE of AMP

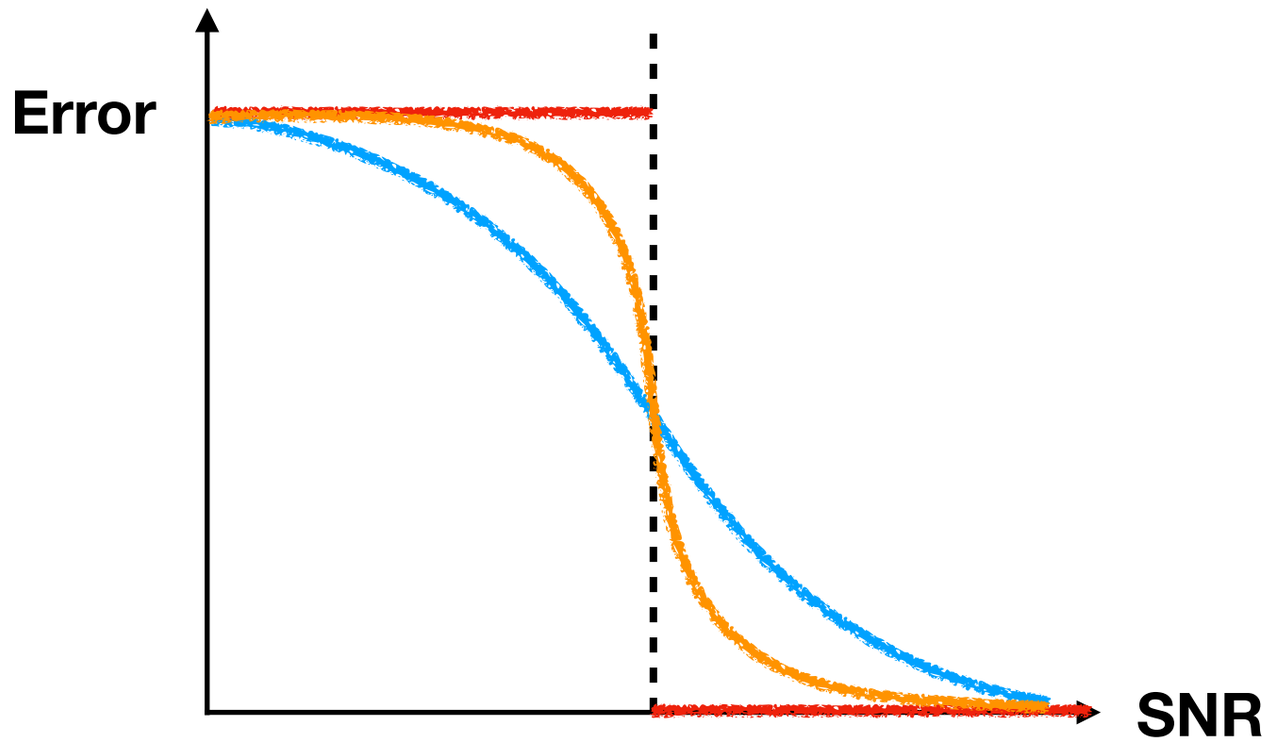


$$\lambda_{\mathrm{AMP}}(\rho) = 1/(e\rho)^2$$

$$\gg \lambda_{\mathrm{IT}}(\rho) = 4\frac{|\ln\rho|}{\rho}$$

(Informal) Theorem [Barbier Macris Rush 20]: AMP finite sample analysis for low sparsity

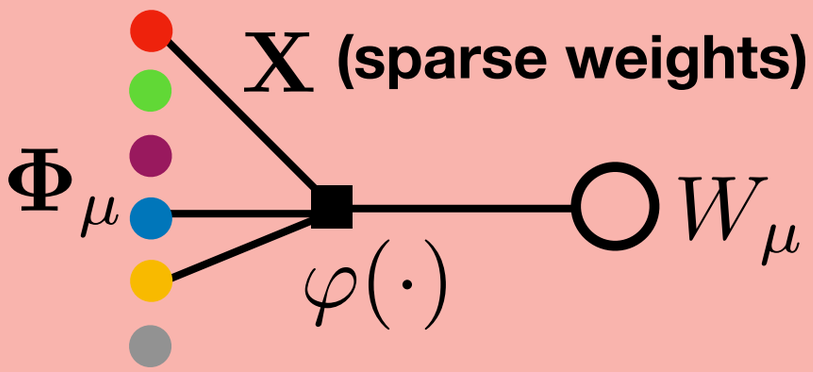
The state evolution recursion allowing to track the performance of AMP, and therefore the curves above, are rigorously valid as $\lambda_n = \Theta(\rho_n^{-2})$ and $\rho_n = \Theta(1/(\ln n)^\alpha)$ for any fixed $\alpha \geq 0$ and n sufficiently large.



All-or-nothing transition: a « universal » phenomenon?

Information-theoretic all-or-nothing transition in GLMs

$$\mathbf{W} = \varphi\left(\sqrt{\frac{1}{n}}\mathbf{\Phi}\mathbf{X}\right)$$



Theorem [Luneau Macris Barbier 20]:

Suppose that the following hypotheses hold:

1. The entries of $\boldsymbol{X}$ are bounded.
2. φ is bounded, and its first and second partial derivatives are bounded and continuous.
3. $(\Phi_{\mu i}) \sim_{\text{iid}} \mathcal{N}(0,1)$.

Let $\rho_n = \Theta(n^{-\lambda})$ with $\lambda \in [0, 1/9)$ and $\alpha_n = \gamma\rho_n|\ln\rho_n|$ with $\gamma > 0$. Then for all n:

$$\left|\frac{I(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{\Phi})}{\alpha_n n} - \inf_{q\geq 0}\sup_{r\geq 0} i_{\text{RS}}(q,r;\alpha_n,\rho_n)\right| \leq \frac{\sqrt{C}\,|\ln n|^{1/6}}{n^{\frac{1}{12}-\frac{3\lambda}{4}}}$$

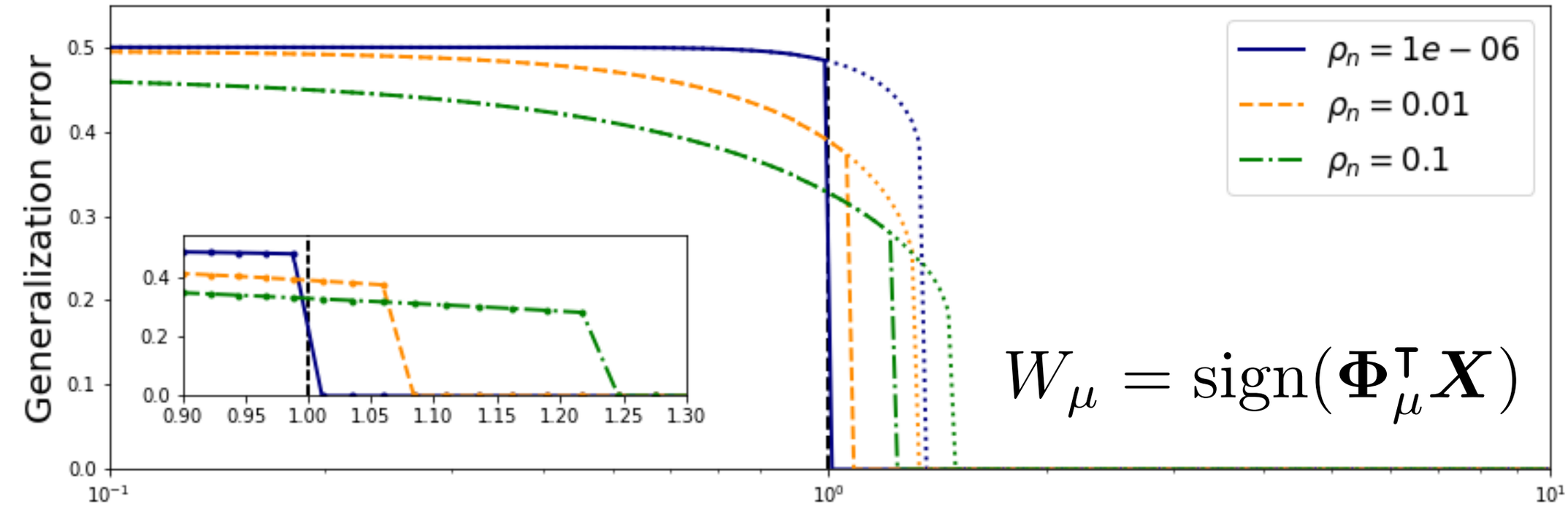
for some constant $C(\varphi) > 0$.

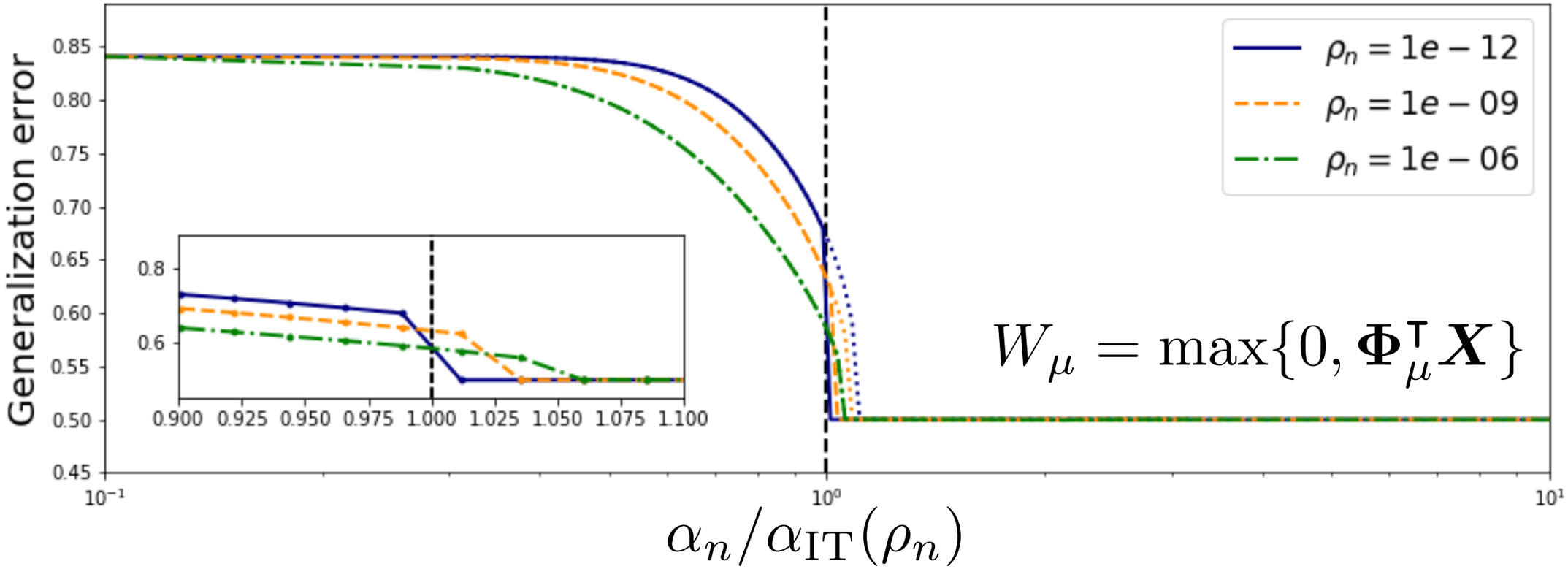
+ analytic exact expression for generalisation: $\lim_{n\to\infty} \mathbb{E}\left[\left(W_{\text{new}} - \mathbb{E}[W_{\text{new}}|\boldsymbol{W},\boldsymbol{\Phi},\boldsymbol{\Phi}_{\text{new}}]\right)^2\right]$

Theorem [Luneau Macris Barbier 20]: In the Bernoulli signal case,

$$\lim_{\rho_n\to 0^+} I(\rho_n,\alpha_n) = \min\left\{I_{P_{\text{out}}}(0,1), \frac{1}{\gamma}\right\}.$$

Information-theoretic all-or-nothing transition in GLMs





$$\alpha_{\mathrm{IT}}(\rho_n) = \gamma_\varphi \big(H_b(\rho_n) + \rho_n H(P_X) \big) \overset{\rho_n \to 0}{\to} \gamma_\varphi \rho_n |\ln \rho_n|$$

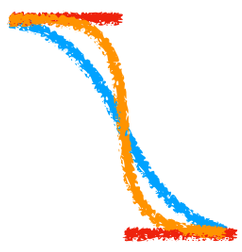
- **Statistical physics predictions rigorously valid in « corners » of phase diagrams (with « control parameters » depending on the system size).**
- **« All-or-nothing » seems quite generic: « Universal » phenomenon in reconstruction of very sparse information (i.e., with « effective low dimension ») from much higher-dimensional data?**

-> [Niles-Weed Zadik 20]: information-theoretic all-or-nothing transition appears for a large class of additive Gaussian models with any sub-linear sparsity:

$$\boldsymbol{W} = \sqrt{\lambda}\boldsymbol{X} + \boldsymbol{Z}$$

- **« All-or-nothing » for AMP algo (at least in sparse PCA). What about other algorithms (Markov chain monte-carlo, gradient-based, etc)?**
- **What about more complex models (multi layers etc)?**

Thank



You