The spiked matrix model with generative priors

Benjamin Aubin†, Bruno Loureiro†, Antoine Maillard⋆, Florent Krzakala⋆, Lenka Zdeborová†

† Institut de Physique Théorique, Paris, France, ⋆ Ecole Normale Supérieure, Paris, France

Motivations and outline

- **Sparsity** is a very popular and widely explored kind of dimensionality reduction, that is a generic and powerful way to perform signal processing and statistical inference.
- On the other hand, **generative models based on neural networks**, such as GANs are particularly performant.
- We study spiked matrix models, where a low-rank matrix is observed through a noisy channel. This problem with **sparse structure of the spikes** has attracted broad attention in the past literature [1]. In this work, we replace the **sparsity** by a **generative modelling** prior. Analyzing the Bayes-optimal performance, we investigate the consequences on statistical and algorithmic properties.

"Are generative priors the new sparsity" ?

Summary

1. In contrast with the sparsity assumption, we do not observe regions of parameters where statistical performance is superior to the best known polynomial algorithmic performance.
2. We show that the approximate message passing (AMP) algorithm is able to reach optimal performance.
3. We design a spectral algorithm (L-AMP) and analyze its performance using random matrix theory, and show its superiority to the classical PCA.
4. We illustrate the performance of the spectral algorithm when the spikes come from real datasets.

Spiked matrix model

We observe a matrix $Y \in \mathbb{R}^{p\times p}$ generated by a ground truth vector $\mathbf{v}^\star \in \mathbb{R}^p \sim P_v$

$$Y = \frac{1}{\sqrt{p}}\mathbf{v}^\star(\mathbf{v}^\star)^\intercal + \sqrt{\Delta}\xi \qquad (1)$$

with $\xi \in \mathbb{R}^{p\times p} \sim \mathcal{N}(\mathbf{0}, I_p)$, $p \to \infty$ and $\Delta = \Theta(1)$.

The goal is to reconstruct $\mathbf{v}^\star$, using the prior knowledge P_v.

A very popular flavor of *dimensionality reduction* is *sparsity*. In this case, the model is called **sparse PCA**. But what about replacing it by a **generative prior** to achieve the reconstruction [2] ?

	Sparse PCA	**Generative prior**
Prior	Sparse	Multi-layer
$\mathbf{v}^\star \sim$	$(1-\rho)\delta(\mathbf{v}^\star) + \rho\mathcal{N}(\mathbf{0}, I_p)$	$\varphi^{(L)}\left(..\varphi^{(1)}\left(\frac{1}{\sqrt{k}}W^{(1)}\mathbf{z}^\star\right)\right)$
$k \ll p$	non-zero components	latent variables
$\rho = \frac{1}{\alpha}$	$\rho = \frac{k}{p} = \Theta(1)$	$\alpha = \frac{p}{k} = \Theta(1)$
Critical noise	$\Delta_c = 1$ (BBP)	?
Algorithmic hard phase	Yes, for $\rho \to 0$	?

References

[1] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017.

[2] Soledad Villar. Generative models are the new sparsity? https://solevillar.github.io/2018/03/28/SUNLayer.html, 2018.

[3] Marylou Gabrié and al. Entropy and mutual information in models of deep neural networks. In *Advances in Neural Information Processing Systems 31*, 2018.

Bayesian inference and posterior distribution

The **posterior distribution** of the inference problem

$$P(\mathbf{v}|Y) = \frac{P_v(\mathbf{v})P(Y|\mathbf{v})}{\mathcal{Z}(Y,\Delta)} = \frac{1}{\mathcal{Z}(Y,\Delta)}P_v(\mathbf{v})\prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - \frac{\mathbf{v}_i\mathbf{v}_j}{\sqrt{p}}\right)^2}, \qquad (2)$$

and the **mutual information density**

$$i \equiv \lim_{p\to\infty}\frac{1}{p}I_p(Y,\mathbf{v}^\star) = \lim_{p\to\infty} -\frac{1}{p}\mathbb{E}_Y\left[\log\mathcal{Z}(Y,\Delta)\right] + \frac{\rho_v}{4\Delta} \qquad (3)$$

Theorem: Mutual information and optimal MSE

Assume the spikes $\mathbf{v}^\star$ come from a structured prior P_v on $\mathbb{R}^p$

a) $$i = \inf_{0\le q_v\le\rho_v} i_{\mathrm{RS}}(\Delta, q_v)$$

b) $$\mathrm{MMSE}_v(\Delta) = \rho_v - \mathrm{arginf}_{q_v} i_{\mathrm{RS}}(\Delta, q_v)$$

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

i_{RS} is called the *replica symmetric* potential.

Illustration with a single layer *iid* prior

$$\mathbf{v}^\star \sim \varphi\left(\frac{1}{\sqrt{k}}W\mathbf{z}^\star\right) \Leftrightarrow$$

- non-linearity φ
- *iid* weights $W \in \mathbb{R}^{p\times k}$, $W_{il} \sim \mathcal{N}(0,1)$, $\alpha = \frac{p}{k} = \Theta(1)$
- latent vector $\mathbf{z}^\star \sim P_z$

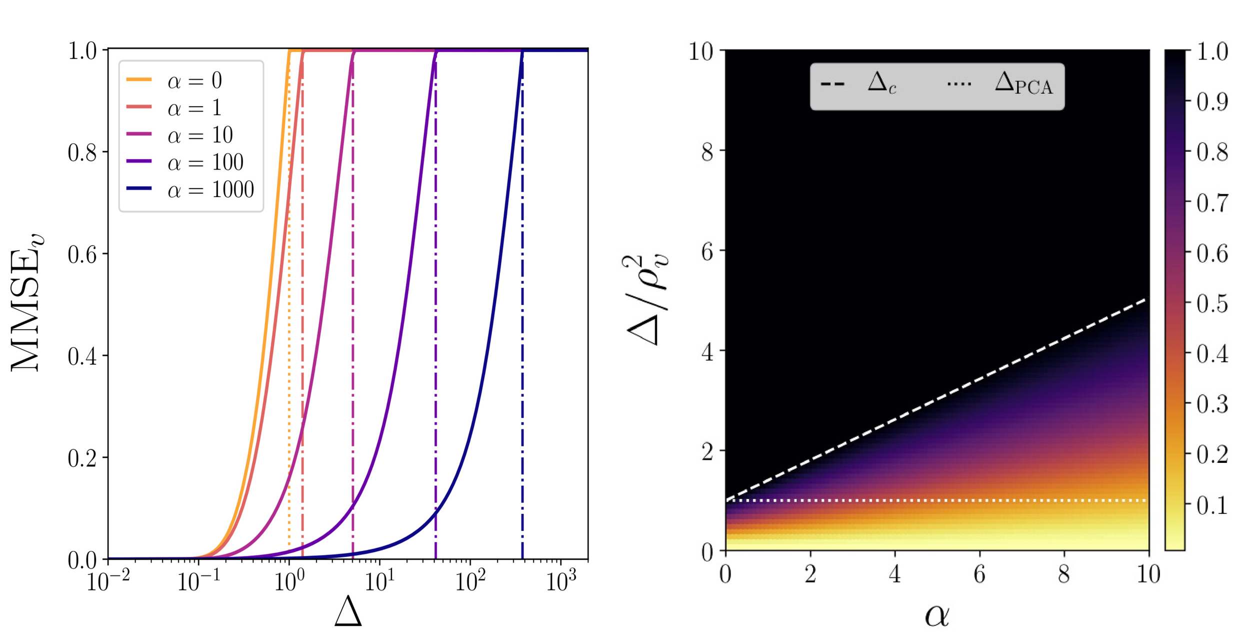
Using mutual information proven in [3]

Corollary: MMSE of the spiked matrix estimation with single layer generative prior

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

$$i_{\mathrm{RS}}(\Delta, q_v) = \frac{\rho_v^2}{4\Delta} + \frac{q_v^2}{4\Delta} + \dots$$
$$\dots + \frac{1}{\alpha}\min_{q_z}\max_{\hat{q}_z}\left[\frac{1}{2}q_z\hat{q}_z - \Psi_z(\hat{q}_z) - \alpha\Psi_{\mathrm{out}}\left(\frac{q_v}{\Delta}, q_z\right)\right]$$

Application: single *sign* layer



We observe a critical noise $\Delta_c(\alpha)$ such that above it, reconstruction is impossible: $\mathrm{MMSE}_v(\Delta > \Delta_c) = 1$.

- Sparse PCA: $\Delta_c = 1$ (dotted white)
- Generative model: $\Delta_c(\alpha) = 1 + \frac{4}{\pi^2}\alpha$ (dashed white)

The gap between sparsity and generative prior increases for large α.

Denoising distributions

$$Q_z(z) \equiv \frac{1}{\mathcal{Z}_z(\gamma,\Lambda)}P_z(z)e^{-\frac{1}{2}\Lambda z^2 + \gamma z},$$

$$Q_{\mathrm{out}}(v,x) \equiv \frac{1}{\mathcal{Z}_{\mathrm{out}}(B,A,\omega,V)}e^{-\frac{1}{2}Av^2 + Bv}P_{\mathrm{out}}(v|x)\,e^{-\frac{1}{2}V^{-1}(x-\omega)^2},$$

$$\Psi_z(x) \equiv \mathbb{E}_\xi\left[\mathcal{Z}_z\left(x^{1/2}\xi, x\right)\log\left(\mathcal{Z}_z\left(x^{1/2}\xi, x\right)\right)\right],$$

$$\Psi_{\mathrm{out}}(x,y) \equiv \mathbb{E}_{\xi,\eta}\left[\mathcal{Z}_{\mathrm{out}}\log\left(\mathcal{Z}_{\mathrm{out}}\left(x^{1/2}\xi, x, y^{1/2}\eta, \rho_z - y\right)\right)\right].$$

Approximate message passing

Is it possible to **algorithmically** achieve the optimal MSE?

AMP algorithm

Input: $Y \in \mathbb{R}^{p\times p}$ and $W \in \mathbb{R}^{p\times k}$:
Initialize with: $\hat{\mathbf{v}}^{t=1} = \mathcal{N}(\mathbf{0}, \sigma^2 I_p)$, $\hat{\mathbf{z}}^{t=1} = \mathcal{N}(\mathbf{0}, \sigma^2 I_k)$, and $\hat{\mathbf{c}}_v^{t=1} = I_p$, $\hat{\mathbf{c}}_z^{t=1} = I_k$, $t = 1$.
repeat
  ***Spiked layer* denoising**:
  $\mathbf{B}_v^t = \frac{1}{\Delta}\frac{Y}{\sqrt{p}}\hat{\mathbf{v}}^t - \frac{1}{\Delta}\frac{\left(I_p^\intercal\hat{\mathbf{c}}_v^t\right)}{p}\hat{\mathbf{v}}^{t-1}$ and $A_v^t = \frac{1}{\Delta p}(\|\hat{\mathbf{v}}^t\|_2)^2 I_p$.
  ***Generative layer* denoising**:
  $V^t = \frac{1}{k}\left(I_k^\intercal\hat{\mathbf{c}}_z^t\right)I_p$, $\boldsymbol{\omega}^t = \frac{1}{\sqrt{k}}W\hat{\mathbf{z}}^t - V^t\mathbf{g}^{t-1}$
  $\mathbf{g}^t = f_{\mathrm{out}}\left(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t\right)$
  $\Lambda^t = \frac{1}{k}\|\mathbf{g}^t\|_2^2 t I_k$ and $\boldsymbol{\gamma}^t = \frac{1}{\sqrt{k}}W^\intercal\mathbf{g}^t + \Lambda^t\hat{\mathbf{z}}^t$.
  Marginals estimation:
  $\hat{\mathbf{v}}^{t+1} = f_v(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t)$ and $\hat{\mathbf{c}}_v^{t+1} = \partial_B f_v(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t)$,
  $\hat{\mathbf{z}}^{t+1} = f_z(\boldsymbol{\gamma}^t, \Lambda^t)$ and $\hat{\mathbf{c}}_z^{t+1} = \partial_\gamma f_z(\boldsymbol{\gamma}^t, \Lambda^t)$,
  $t = t + 1$.
until Convergence.
Output: $\hat{\mathbf{v}}, \hat{\mathbf{z}}$.

$f_z, f_v, f_{\mathrm{out}}$ are respectively the means of z, v and $V^{-1}(x-\omega)$ with respect to Q_z and the joint distribution Q_{out}.

State evolution

Overlaps $q_v = \lim_{p\to\infty}\mathbb{E}\left[\frac{1}{p}\hat{\mathbf{v}}^\intercal\mathbf{v}^\star\right]$, $q_z = \lim_{k\to\infty}\mathbb{E}\left[\frac{1}{k}\hat{\mathbf{z}}^\intercal\mathbf{z}^\star\right]$ measure the reconstruction of the algorithm

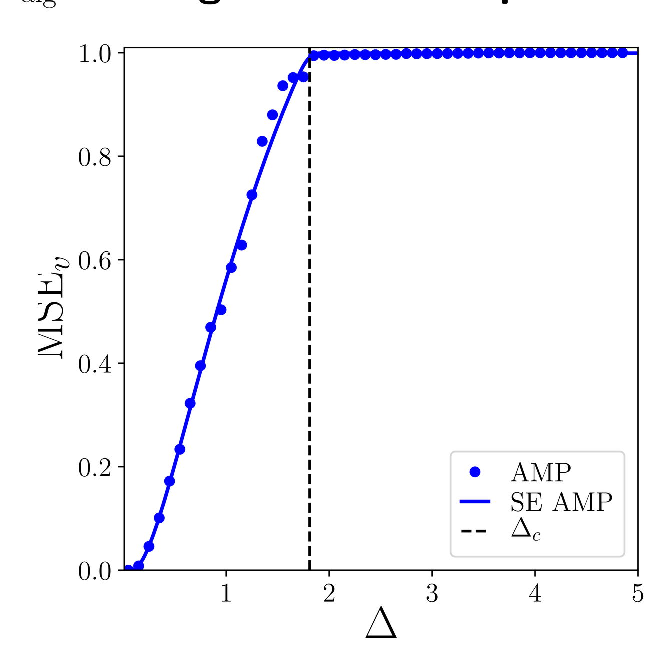
State Evolution equations

For iid weights:

$$q_v^{t+1} = 2\partial_{q_v^t}\Psi_{\mathrm{out}}\left(\frac{q_v^t}{\Delta}, q_z^t\right),\ q_z^{t+1} = 2\partial_{\hat{q}_z}\Psi_z\left(\hat{q}_z^t\right),\ \hat{q}_z^{t+1} = 2\alpha\partial_{q_z}\Psi_{\mathrm{out}}\left(\frac{q_v^t}{\Delta}, q_z^t\right)$$

... exactly the **saddle point equations** of i_{RS} !

- AMP solves the minimization problem of the potential i_{RS}.
- AMP achieves the optimal MSE in the limit $p \to \infty$, as long as i_{RS} has a unique minimizer $q_v^\star$.
- $q_v^\star$ unique in all the models we considered: single/multiple layers with {Linear, Sign, ReLU}.
- $\Delta_c = \Delta_{\mathrm{IT}} = \Delta_{\mathrm{alg}}$: **no algorithmic hard phase!**



L-AMP spectral algorithm

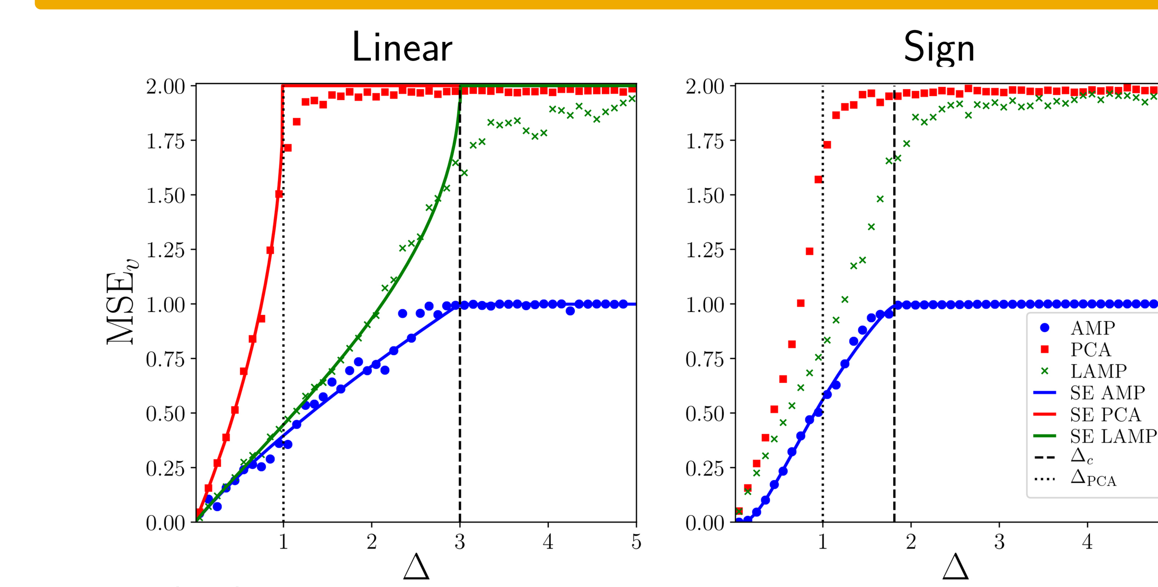
- AMP: correlates with the ground truth signal for $\Delta \le \Delta_c$
- PCA: correlates with the ground truth for $\Delta_{\mathrm{PCA}} \le 1$

Is it possible to design **spectral algorithms**, taking advantage of generative prior ?

L-AMP algorithm

LAMP spectral algorithm: take the leading eigenvector of

$$\Gamma_p = \frac{1}{\Delta}K_p\left[\frac{Y}{\sqrt{p}} - I_p\right] \quad \text{with} \quad K_p \equiv \frac{1}{k}\mathbb{E}[\mathbf{v}\mathbf{v}^\intercal]$$



- PCA (red) works for $\Delta < 1$.
- AMP (blue) achieves the Bayes optimal MSE.
- LAMP (green) has the same statistical threshold than AMP!

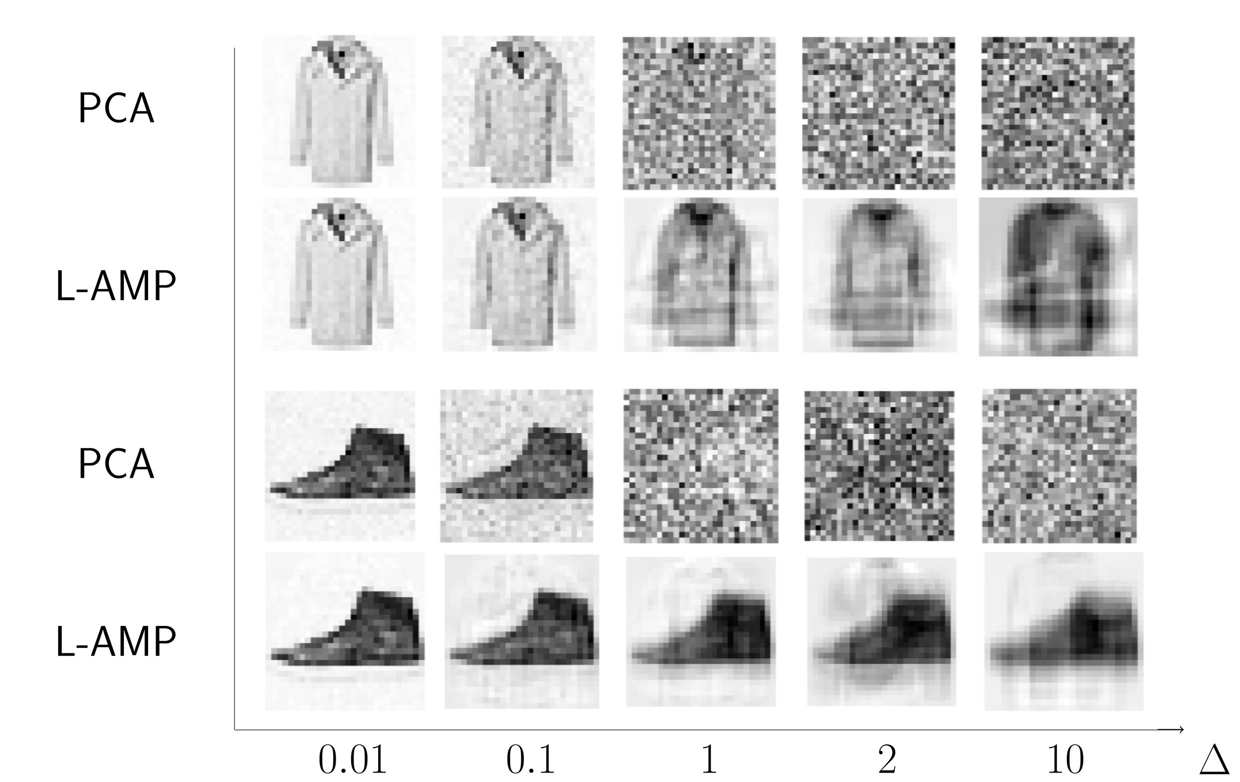
Theorem: RMT & L-AMP (linear)

The leading eigenvector of $\Gamma_p = \frac{1}{\Delta}\frac{WW^\intercal}{k}\left[\frac{Y}{\sqrt{p}} - I_p\right]$ correlates with the ground truth signal for $\Delta < \Delta_c(\alpha) = 1 + \alpha$.

Still an open problem to show the same result for the non-linear case.

Real data generative prior

- P_v: underlying distribution of *Fashion Mnist* dataset $\{\mathbf{v}^\mu\}_{\mu=1}^m$
- Use empirical covariance $K_p = \frac{1}{k}\mathbb{E}[\mathbf{v}\mathbf{v}^\intercal] \simeq \frac{1}{m}\Sigma_{\mu=1}^m\mathbf{v}^\mu(\mathbf{v}^\mu)^\intercal$



Conclusion: Sparsity vs Generative priors

- With **sparse prior**, for low sparsity parameter ρ, large gap between information theoretical and best-known-polynomial algorithm performances: $\Delta_{\mathrm{IT}} < \Delta_{\mathrm{alg}}$. No known algorithm able to beat PCA threshold $\Delta = 1$.
- With **generative prior**, no algorithmic hard phase: $\Delta_{\mathrm{IT}} = \Delta_{\mathrm{alg}}$. Spectral L-AMP algorithm outperforms PCA: reconstruction for larger noise and has the same threshold than AMP (conjectured optimal).

"Generative priors are better than sparsity".

arxiv:1905.12385