The All-or-Nothing Phenomenon in Sparse Tensor PCA

Jonathan Niles-Weed[1,2] **& Ilias Zadik**[2]

[1]Courant Institute of Mathematical Sciences, NYU & [2]Center for Data Science, NYU

High-dimensional statistical models undergo phase transitions
small changes in the parameters of the model imply *large changes* in the model (e.g. large changes in terms of recovery error).

A canonical example

Principal Component Analysis [BBP'05];
For $\beta \sim \mathrm{Unif}\left(\mathcal{S}^{p-1}\right)$ and $W_{i,j} = W_{j,i} \sim N(0,1)$ (Wigner),

$$Y = \sqrt{\lambda p}\beta\beta^\top + W.$$

$$\lim_{p\to+\infty} \mathbb{E}\left[\|\beta - \mathbb{E}\left[\beta|Y\right]\|_2^2\right] = \begin{cases} < 1 \text{ , if } \lambda > 1 \\ 1 \quad \text{ , if } \lambda < 1\,. \end{cases}$$

Under (sublinear) sparsity: Much sharper transitions

Sparse Linear Regression

For $\beta \sim \mathrm{Unif}(\{v \in \{0, \frac{1}{\sqrt{k}}\}^p, \|v\|_0 = k\})$, $X_{i,j}, W_i \sim N\left(0,1\right)$,

$$Y = X\beta + \sigma W \in \mathbb{R}^n.$$

[Gamarnik, Zadik 17], [Reeves, Xu, Zadik '19] *"The All-or-Nothing Phenomenon"*
For some critical sample size n^* and $k = o(\sqrt{p})$, if $\mathrm{SNR} = k/\sigma^2 \to +\infty$,

$$\lim_{p\to+\infty} \mathbb{E}\left[\|\beta - \mathbb{E}\left[\beta|Y, X\right]\|_2^2\right] = \begin{cases} 0 \text{ , if } n > n^* \\ 1 \text{ , if } n < n^*\,. \end{cases}$$

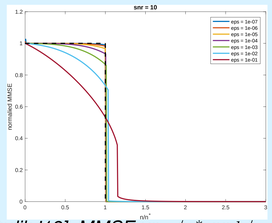


Figure from [Reeves, Xu, Zadik '19], MMSE vs n/n^* as $k/p = \epsilon \to 0$.

Sparse (tensor) PCA

For $\beta \sim \mathrm{Unif}(\{v \in \{0, \frac{1}{\sqrt{k}}\}^p, \|v\|_0 = k\}), W_{i,j} \sim N\left(0,1\right)$ and $d \geq 2$,

$$Y = \sqrt{2\lambda k \log\frac{p}{k}}\beta^{\otimes d} + W.$$

[Barbier, Macris, Rush '20] If $d = 2$ **(the matrix case)**, $p^{\frac{12}{13}} \ll k = o(p)$ then

$$\lim_{p\to+\infty} \mathbb{E}\left[\|\beta - \mathbb{E}\left[\beta|Y\right]\|_2^2\right] = \begin{cases} 0 \text{ , if } \lambda > 1 \\ 1 \text{ , if } \lambda < 1\,. \end{cases}$$

Bernoulli Group Testing

For $\beta \sim \mathrm{Unif}(\{v \in \{0, \frac{1}{\sqrt{k}}\}^p, \|v\|_0 = k\}), X_{i,j} \sim \mathrm{Bern}(\ln 2/k)$

$$Y = \left(\mathbf{1}\left(\langle X, v\rangle \geq 1\right)\right)_{i=1,\ldots,n} \in \{0,1\}^n.$$

[Truong, Aldridge, Scarlett '20] For some critical sample size n^* and $k = p^{o(1)}$,

$$\lim_{p\to+\infty} \mathbb{E}\left[\|\beta - \mathbb{E}\left[\beta|Y, X\right]\|_2^2\right] = \begin{cases} 0 \text{ , if } n > n^* \\ 1 \text{ , if } n < n^*\,. \end{cases}$$

Main Contributions

- Systematic study of a simple but generic inference model (the *Gaussian Additive Model*)
- A *simple necessary and sufficient condition* for all-or-nothing to hold.
- Sparse tensor PCA exhibits all-or-nothing *for all $d \geq 2$ and $k = o(p)$.*

The Gaussian Additive Model

For **arbitrary discrete** $S = S_p \subset \mathcal{S}^{p-1} = \{v \in \mathbb{R}^p : \|v\|_2 = 1\}$, let $\beta \sim \mathrm{P}_p := \mathrm{Unif}(S_p)$ and $W_i \sim N\left(0,1\right)$,

$$Y_\lambda = \sqrt{\lambda}\beta + W.$$

Subcases:

- *Sparse Tensor PCA:* $S = \{v^{\otimes d} : v \in \{0, \frac{1}{\sqrt{k}}\}^p, \|v\|_0 = k\}$.
- *Submatrix Localization:* $S = \{(\mathbf{1}_{\sigma(i)=\sigma(j)} - \frac{1}{k})_{i,j\in[n]} : \sigma : [n] \to [k]$ "balanced"$\}$.
  "Gaussian version of stochastic block model"

Definition 1 *A sequence of priors $(\mathrm{P}_p)_{p\in\mathbb{N}}$ satisfies the all-or-nothing phenomenon holds if for some critical $\lambda_c = \lambda_c(S,p)$:*

$$\lim_{p\to\infty} \mathbb{E}\|\beta - \mathbb{E}[\beta|Y_\lambda]\|_2^2 = \begin{cases} 0 & \text{if } \lambda > \lambda_c \\ 1 & \text{if } \lambda < \lambda_c\,. \end{cases}$$

An equivalent condition

Definition 2 *We denote by D the Kullback-Leibler divergence; given two random variables Z_1, Z_2 with the law of Z_1, P_1, absolutely continuous to the law of Z_2, P_2,*

$$\mathrm{D}(Z_1, Z_2) = \mathbb{E}\left[\frac{\mathrm{dP}_1(Z_2)}{\mathrm{dP}_2(Z_2)}\log\left(\frac{\mathrm{dP}_1(Z_2)}{\mathrm{dP}_2(Z_2)}\right)\right].$$

Theorem 1 *Suppose $|S| = |S_p| \to +\infty$ and S is "sufficiently spread out"* as $p \to +\infty$. A sequence $\{\mathrm{P}_p\}$ satisfies the all-or-nothing phenomenon for some λ_c if and only if*

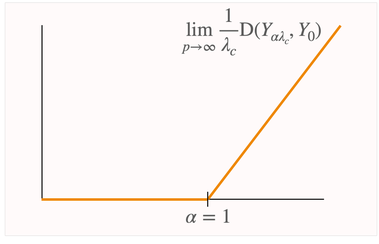
$$\lim_{p\to+\infty} \frac{1}{\log|S_p|}\mathrm{D}(Y_{2\log|S_p|}, Y_0) = 0$$

and $\lambda_c = (1+o(1))2\log|S_p|$.

*Sufficient spread is necessary for "all" recovery to happen.

Key Intuition

- I-MMSE formula:
$$\frac{d}{d\lambda}\mathrm{D}(Y_\lambda, Y_0) = \frac{1}{2} - \frac{1}{2}\mathbb{E}\|\beta - \mathbb{E}[\beta|Y_\lambda]\|_2^2$$
An abrupt change in the slope of KL!
- All-or-nothing phenomenon at λ_c if and only if
$$\lim_{p\to\infty} \frac{1}{\lambda_c}\mathrm{D}(Y_{\alpha\lambda_c}, Y_0) = \frac{1}{2}(\alpha - 1)_+ \qquad \forall \alpha \geq 0\,.$$
- Since limiting KL is Lipschitz, convex, and nonnegative, evaluating the limit at a $\alpha = 1$ and $\alpha \to \infty$ gives information about the whole function.
- For "large" $\alpha > 1$, $\frac{1}{\lambda_c}\mathrm{D}(Y_{\alpha\lambda_c}, Y_0) = \frac{\alpha}{2} - \frac{I(\beta;Y_{\alpha\lambda_c})}{\lambda_c} \approx \frac{\alpha}{2} - \frac{\log|S|}{\lambda_c}$. Hence $\lambda_c \approx 2\log|S|$.



"Near-orthogonality" implies all-or-nothing

Corollary 1 *(Informal) Suppose $S = S_p$ consist of "nearly-orthogonal" vectors, rigorously for any $t > 0$ and large p,*

$$\mathrm{P}_p^{\otimes 2}(\langle X, X'\rangle \geq t) \leq |S|^{-\frac{2t}{t+1}}$$

then all-or-nothing phenomenon always holds at $\lambda_c = 2\log|S_p|$.

Proof by conditional second moment method [Banks et al. '18, Perry et al. '20].

Easy implication: **Any** uniform distribution over discrete subsets of orthogonal vectors satisfies the all-or-nothing phenomenon!

Application: Sparse Tensor PCA

Theorem 2 *For any $d \geq 2$ and $k = o(p)$, the sparse tensor PCA model*

$$Y = \sqrt{2\lambda k \log\left(\frac{p}{k}\right)}\beta^{\otimes d} + W\,, \quad \beta \sim \tilde{\mathrm{P}}_p$$

with $\tilde{\mathrm{P}}_p = \mathrm{Unif}(\{v \in \{0, \frac{1}{\sqrt{k}}\}^p : \|v\|_0 = k\})$ exhibits the all-or-nothing phenomenon:

$$\lim_{p\to\infty} \mathbb{E}\|\beta^{\otimes d} - \mathbb{E}[\beta^{\otimes d}|Y]\|^2 = \begin{cases} 1 & \text{if } \lambda < 1 \\ 0 & \text{if } \lambda > 1\,. \end{cases}$$

Proof idea

- If β and β' are independent draws from $\tilde{\mathrm{P}}_p = \mathrm{Unif}(\{v \in \{0, \frac{1}{\sqrt{k}}\}^p : \|v\|_0 = k\})$, then the overlap $\langle\beta, \beta'\rangle$ is a rescaled Hypergeometric random variable.
- When $k = o(p)$, this distribution satisfies $\tilde{\mathrm{P}}_p^{\otimes 2}(\langle\beta, \beta'\rangle \geq t) \leq |S|^{-t+o(1)}$.
- Therefore, for any $d \geq 2$,
$$\tilde{\mathrm{P}}_p^{\otimes 2}(\langle\beta^{\otimes d}, (\beta')^{\otimes d}\rangle \geq t) \leq |S|^{-t^{1/d}+o(1)} \leq |S|^{-\frac{2t}{t+1}}.$$

Result follows from Corollary 1.

Open questions

- How does Theorem 1 extend to more general models (e.g. GLMs)?
- All-or-nothing for polynomial-time estimators?

References

[BBP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[Barbier, Macris, Rush '20] Jean Barbier, Nicolas Macris, and Cynthia Rush. All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. arXiv:2006.07971, 2020.

[Banks et al. '18] Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *IEEE Trans. Inform. Theory*, 64(7):4872–4994, 2018.

[Gamarnik, Zadik '17] David Gamarnik and Ilias Zadik. High dimensional linear regression with binary coefficients: Mean squared error and a phase transition. *Conference on Learning Theory (COLT)*, 2017.

[Perry et al. '20] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira. Statistical limits of spiked tensor models. *Ann. Inst. Henri Poincaré Probab. Stat.*, 56(1):230–264, 2020.

[Reeves, Xu, Zadik '19] Galen Reeves, Jiaming Xu, and Ilias Zadik. The all-or-nothing phenomenon in sparse linear regression. *Conference on Learning Theory (COLT)*, 2019.

[Truong, Aldridge, Scarlett '20] Lan V. Truong, Matthew Aldridge and Jonathan Scarlett On the All-Or-Nothing Behavior of Bernoulli Group Testing arXiv preprint, 2020.

Advances in Neural Information Processing Systems 34 (NeurIPS 2020)