POWER-LAW FEATURE STATISTICS EXPLAIN TEST RECONSTRUCTION GAPS IN ASSOCIATIVE MEMORIES

Sergio E.G. Manfrin
Università degli Studi di Milano

Francesco D'Amico
Università di Roma Sapienza

Marco Gherardi
Università degli Studi di Milano

Aurélien Decelle
Universidad Politécnica de Madrid

Beatriz Seoane
Universidad Complutense de Madrid

Matteo Negri
CY Cergy Paris Université

ABSTRACT

Associative memories have been recently shown to be able to generalize, that is to produce attractors near previously unseen examples. While this phenomenon is understood in the synthetic setting of random-features examples as the exploitation of mixed spurious states, it is unclear whether the same explanation extends to real datasets, where the emergent attractors do not coincide perfectly with test examples (a test reconstruction gap is visible). In this work, we introduce a more natural model of data where random features are sampled with power-law occurrence, showing that this change produces many benign effects for generalization, including an implicit form of regularization. Overall, this new data structure provides an interpretation of the test reconstruction gap that is consistent with the known mechanism for generalization based on mixed spurious states.

1 INTRODUCTION

Associative memories (AMs) were originally introduced as content-addressable systems capable of retrieving stored patterns from partial or noisy cues and formalized by Hopfield as energy-based models (Hopfield, 1982). Beyond pure retrieval, recent works have shown that AMs can also *generalize*: unseen examples can be partially memorized, creating new attractors nearby them in place of the original memories (Kalaj et al., 2024; Serricchio et al., 2025; D'Amico et al., 2025a). This *generalization phase* demonstrates the capability of the system to learn a set of patterns shared across examples rather than merely storing individual patterns.

In synthetic random-feature models, this effect admits a compelling explanation Negri et al. (2023); Kalaj et al. (2024). Generalization arises through the formation of *benign spurious states*, corresponding to mixtures of learned features. These mixed states can align with unseen combinations of features, providing a mechanism consistent with the broader idea that learning corresponds to recovering an underlying data manifold (Rahimi & Recht, 2007; Mézard, 2017; Goldt et al., 2020; Gerace et al., 2020).

In this idealized setting, generalization can even coincide with perfect recovery of previously unseen examples (Kalaj et al., 2024). However, when AMs are trained on real datasets, the picture changes. Although unseen examples are attracted to nearby configurations, they do not coincide exactly with them: a clear *test reconstruction gap* is observed, with test examples converging to attractors of overlap strictly smaller than one (Serricchio et al., 2025; D'Amico et al., 2025a). Whether the benign-spurious-state mechanism extends to this more realistic setting remains unclear. Moreover, standard random-feature models assume uniformly sampled components and produce a sharp, finite-rank structure that does not reflect the heterogeneous statistics of natural data.

The main motivation of this work is to introduce a minimal yet realistic deformation of the random-feature model that captures two salient properties of real datasets (Catania et al., 2025): (i) a smooth principal component spectrum and (ii) the presence of a non-trivial test reconstruction gap. We

achieve this by sampling features according to a power-law distribution, motivated by the well-established heavy-tailed statistics observed in many natural datasets Mazzolini et al. (2018a;b) such as language (Zipf's law (Zipf, 2013)), vision (Simoncelli & Olshausen, 2001), and biological signals (Stringer et al., 2019).

In this work, we numerically investigate the generalization transition in random-feature models with power-law feature statistics. We train associative memories by conditional likelihood maximization and show that introducing heavy-tailed feature occurrences qualitatively reshapes their behavior. Learning becomes hierarchical: frequent features are acquired earlier (in training time) than rare ones, and generalization performance depends on how many features have been effectively incorporated. As a consequence, a non-trivial test reconstruction gap naturally emerges.

We support these numerical findings with a signal-to-noise analysis, which clarifies the intuition that only a subset of statistically significant features is learnable at finite sample size.

2 Model definition: associative memory from conditional likelihood

We consider a deterministic recurrent network of N binary neurons $x_i \in \{\pm 1\}$ implemented by a linear layer J_{ij} with activation $\mathrm{sgn}(\cdot)$:

$$x_i^{(t+1)} = \mathrm{sgn}\Big[\sum_{j(\neq i)}^{N} J_{ij}\, x_j^{(t)} \Big], \tag{1}$$

where the sum considers j different from i because we exclude self-couplings, that is $J_{ii} = 0\ \forall i$. The variables $x_i^{(t)}$ describe the dynamical state of the network at time t. The dataset consists of P static configurations $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^P$, where each example $\boldsymbol{\xi}^\mu \in \{\pm 1\}^N$ has the same structure as a network configuration $\boldsymbol{x}$. During training, we interpret each data point $\boldsymbol{\xi}^\mu$ as a target configuration that should become an attractor of the dynamics in Eq. 1.

We train this recurrent network in a self-supervised way, by maximizing the conditional likelihood $p_i(\xi_i^\mu \,|\, \boldsymbol{\xi}_{\setminus i}^{\boldsymbol{\mu}})$ of the component ξ_i^μ of example μ given all the other components $\boldsymbol{\xi}_{\setminus i}^{\boldsymbol{\mu}}$, defined as:

$$p_i(\xi_i^\mu | \boldsymbol{\xi}_{\setminus i}^{\boldsymbol{\mu}}) = \exp\left(\beta \xi_i^\mu \sum_{j(\neq i)}^{N} J_{ij}\, \xi_j^\mu \right) \Big/ 2\cosh\left(\beta \sum_{j(\neq i)}^{N} J_{ij}\, \xi_j^\mu \right). \tag{2}$$

Explicitly, we minimize the loss function $\mathcal{L} = -\frac{1}{P}\sum_{\mu=1}^{P}\sum_{i=1}^{N} \log p_i(\xi_i^\mu | \boldsymbol{\xi}_{\setminus i}^{\boldsymbol{\mu}})$. It was recently shown in D'Amico et al. (2025a) that this minimization produces an associative memory, since cross-entropy loss implicitly maximizes the classification margins (Soudry et al., 2018; Montanari et al., 2024), which in turn produces large basins of attraction around the training examples (Gardner, 1987; Forrest, 1988).

3 Artificial datasets with power-law feature statistics

We construct our dataset of P correlated examples as combinations of random features, as follows:

$$\xi_i^\mu = \mathrm{sgn}\left[\sum_{k=1}^{D} c_k^\mu f_{ki} \right], \tag{3}$$

where $\boldsymbol{f}_k$ are $D = \alpha_D N$ random binary N-dimensional vectors, with distribution $f_{ki} \sim \mathrm{Unif}(\pm 1)$. The coefficient matrix $c_k^\mu \in \{0, \pm 1\}$ is chosen so that, for each example $\boldsymbol{\xi}^\mu$ we select exactly $L = \mathcal{O}(1)$ features, and each feature is chosen with a probability π_k proportional to index k:

$$\pi_k = k^{-\eta}/Z(\eta, D), \tag{4}$$

where the normalization constant $Z(\eta, D)$ ensures that $\sum_{k=1}^{D} \pi_k = 1$. The heavy-tailed feature statistics produce a smooth, slowly decaying principal component spectrum (see Appendix A), in contrast to the sharp rank-D structure generated by uniformly sampled random features. Power-law feature occurrences therefore provide a minimal and realistic deformation of the standard random-feature model (Rahimi & Recht, 2007; Mézard, 2017; Goldt et al., 2020; Gerace et al., 2020; Negri et al., 2023; Kalaj et al., 2024).

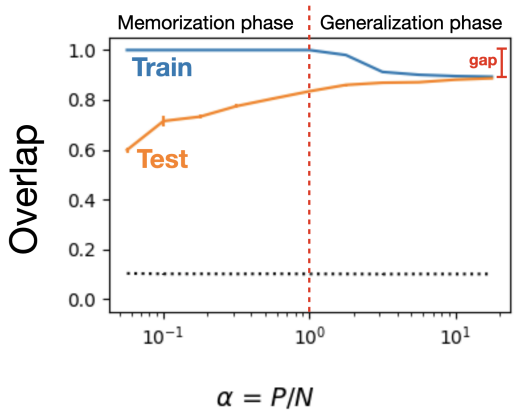

Figure 1: **Power-law feature statistics produce a *test reconstruction gap* in the generalization phase.** Overlap for training and test examples as function of the model load $\alpha = P/N$. Training examples are stable below the red dashed line, which denotes the memory-to-generalization transition. Above the transition, the training overlap decreases and the test overlap increases until they converge for large α, signaling that the model does not distinguish anymore between training and test examples. The horizontal black dashed line represents the typical overlap between examples. $\eta = 1.5$, $L = 3$, $N = 2000$, $\alpha_D = 0.5$. The curves are averages of at least 200 examples. Error bars are mostly too small to be visible.

4 Numerical results

Method. To characterize the generalization properties of the model, we probe the dynamical attractors of the trained network. After training, discussed in Appendix C, we initialize the network on a configuration of the training or test set $\boldsymbol{x}^{(0)}$ and iterate the deterministic dynamics in Eq. 1 until convergence to a fixed point $\boldsymbol{x}^{\star}$. We then measure the overlap $m = \frac{1}{N}\sum_{i=1}^{N} x_i^{(0)} x_i^{\star}$, which quantifies how much of the original configuration is recovered by the attractor. Perfect retrieval corresponds to $m = 1$, while $m < 1$ indicates that the attractor differs from the initialization.

Test reconstruction gap. In Fig. 1 we report the overlaps for training and test examples as function of α. Test examples converge to nearby but distinct attractors, resulting in a *test reconstruction gap*. Importantly, this gap does not signal failure of retrieval. Rather, it reflects the fact that the learned attractors correspond to mixtures of features that only approximately match unseen examples. Unlike the uniform random-feature case, where generalization may coincide with perfect recovery of synthetic mixtures, the power-law structure produces attractors that interpolate between examples but do not exactly reproduce them. This behavior qualitatively matches what is observed on real datasets (Serricchio et al., 2025; D'Amico et al., 2025a).

Incremental feature learning. A central observation is that features are learned gradually, both in training time and as the dataset size increases. By monitoring the effective number of learned features (Fig. 2, green dashed lines), we observe that more frequent features (the ones with a lower index k) are learned first, while rare features are incorporated only later. This hierarchy directly reflects the power-law sampling distribution: frequent components generate a stronger signal during training and thus cross the learning threshold earlier. Crucially, this incremental learning avoids the sharp "memory blackout" transitions typical of uniform random-feature models (Negri et al., 2023; Kalaj et al., 2024). Instead of a sudden collapse due to interference between many equally strong components, the network progressively incorporates features in order of statistical relevance. These observations suggest that the test reconstruction gap is quantitatively controlled by the number of features effectively learned.

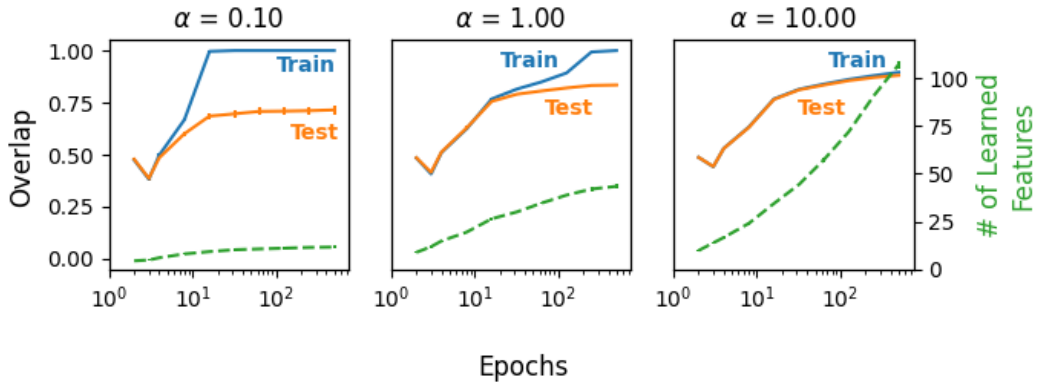

Figure 2: **The number of learned features explains the test reconstruction gap**. Left axis: Overlap for training and test examples during training. Right axis: the green dashed line represent the number of stable features. The higher this number, the higher the test overlap. For all panels: $\eta = 1.5$, $L = 3$, $N = 2000$, $\alpha_D = 0.5$. The curves are averages of at least 200 examples. Error bars are mostly too small to be visible.

5 Signal-to-noise analysis: number of learned features

We are able to explain how many features are learned given $P = \alpha N$ and η by simple signal-to-noise arguments.

We provide a first intuitive argument by making the assumption that a feature is learned if it appears at least P_{crit} times in the examples. Given a dataset of P examples, features with larger k will appear fewer times, and we expect that with $k > k_{\max}$ features will not be learned by the model, where the threshold $k_{\max}$ depends on η.

We can make this more precise using Hebbian learning $J^{\text{Hebb}} = \boldsymbol{\xi}^\top \boldsymbol{\xi}/N$ as a proxy for the true optimizer J^* of $\mathcal{L}$. This is motivated by D'Amico et al. (2025a) and D'Amico et al. (2025b), were they show that J^{Hebb} appears in the early optimization of $\mathcal{L}$, and the full optimization increases the critical capacity only by a factor, without changing the scaling. We verify this numerically in Fig. 3 that the scaling of $k_{\max}$ is the same for J^{Hebb} and J^*.

We proceed by analyzing the local field h_{ik} acting on feature k, namely $h_{ik} = \sum_{j(\neq i)} J^{\text{Hebb}}_{ij} f_{jk}$. Within h_{ik} we can isolate a signal term, proportional to f_{ki}, and a noise term:

$$h_{ik} \simeq \frac{P_k}{L}\left(f_{ki} + O\left(\sqrt{\frac{Z(2\eta, D)}{N\, k^{-2\eta}}}\right)\right) \tag{5}$$

We omitted other subdominant noise terms. As derived in Appendix B, the scaling of $Z(\eta, D))$ with the number of features D depends strongly on η. For the complete expression and the full derivation, see Appendix B. We can find the maximum scaling with N of $k_{\max}(N)$ as the scaling that makes the noise term finite, namely

$$k_{\max}(N) = \mathcal{O}(N^{\frac{1}{2\eta}}) \tag{6}$$

for $\eta > 1/2$, while $k_{\max} = D = \alpha_D N$ if $\eta < 1/2$.

In Fig. 3 we verify that Eq. 6 is the proper scaling (for both J^{Hebb} and J^*). From Eq. 6, we learn that if $\eta > 1/2$ the number of learned features $k_{\max}$ scales slower than $\mathcal{O}(N)$, which means that it is impossible to learn all the $D = \alpha_D N$ features for $\eta > 1/2$.

Therefore, for $\eta > 1/2$, the model remains in a partial-feature regime even as $N \to \infty$, implying a persistent test reconstruction gap in the thermodynamic limit. The power-law deformation is irrelevant for $\eta < 1/2$ in the thermodynamic limit.

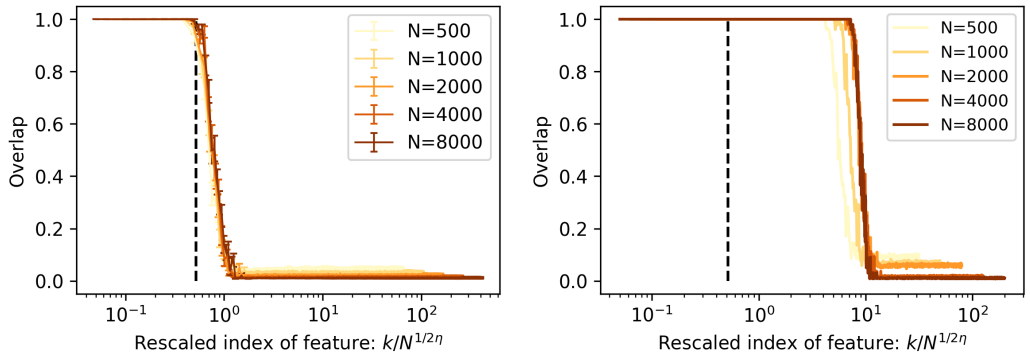

Figure 3: **The number of features learned scales as** $k_{\max} \propto N^{1/2\eta}$. The panels show the feature overlap $m_k = \frac{1}{N}\sum_{i=1}^{N} f_{ki} x_i^\star$ as a function of the (rescaled) feature index k, for increasing values of N. The overlap dropping from 1 to vanishing implies that only a subset of features is learned. The fact that curves for different N collapse when we scale k by $N^{1/2\eta}$ implies that the signal-to-noise analysis described in section 5 is correct. The black vertical line is the prediction of $k_{\max}$ if we assume $P_{\text{crit}} = 0.138N$, that we expect to be close to the transition in Hebbian learning. *Left panel:* Hebbian learning. *Right panel:* conditional likelihood maximization.

6 Conclusion

We investigated how the statistical structure of data influences generalization in associative memories trained by conditional likelihood. By introducing power-law feature statistics, we constructed a minimal deformation of the standard random-feature model that captures two key properties of natural datasets: a smooth principal component spectrum and a non-trivial test reconstruction gap.

Our results show that the test reconstruction gap is directly controlled by the number of features effectively learned by the network. Feature acquisition occurs sequentially: frequent components are incorporated early, while rare components require larger datasets and longer training. When only a subset of features is learned, unseen examples converge to coarse mixtures, leading to imperfect overlap. As additional features are progressively acquired, the attractors better approximate test examples and the gap shrinks monotonically. In this view, the test reconstruction gap is not a failure of the benign-spurious-state mechanism, but a finite-resolution effect arising from partial feature learning.

Overall, our work shows that the natural heavy-tailed structure of data qualitatively reshapes the phase diagram of associative memories. Power-law feature statistics provide a principled explanation for the test reconstruction gap, suggesting that the random feature structure is a good model of data for self-supervised tasks.

References

Giovanni Catania, Aurélien Decelle, Cyril Furtlehner, and Beatriz Seoane. A theoretical framework for overfitting in energy-based modeling. *arXiv preprint arXiv:2501.19158*, 2025.

Francesco D'Amico, Dario Bocchi, Luca Maria Del Bono, Saverio Rossi, and Matteo Negri. Pseudo-likelihood produces associative memories able to generalize, even for asymmetric couplings. *arXiv preprint arXiv:2507.05147*, 2025a.

Francesco D'Amico, Dario Bocchi, and Matteo Negri. Implicit bias produces neural scaling laws in learning curves, from perceptrons to deep networks. *arXiv preprint arXiv:2505.13230*, 2025b.

BM Forrest. Content-addressability and learning in neural networks. *Journal of Physics A: Mathematical and General*, 21(1):245, 1988.

E. Gardner. Maximum Storage Capacity in Neural Networks. *Europhysics Letters*, 4(4):481, August 1987. ISSN 0295-5075. doi: 10.1209/0295-5075/4/4/016. URL https://dx.doi.org/10.1209/0295-5075/4/4/016.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Silvio Kalaj, Clarissa Lauditi, Gabriele Perugini, Carlo Lucibello, Enrico M Malatesta, and Matteo Negri. Random features hopfield networks generalize retrieval to previously unseen examples. *arXiv preprint arXiv:2407.05658*, 2024.

Andrea Mazzolini, Marco Gherardi, Michele Caselle, Marco Cosentino Lagomarsino, and Matteo Osella. Statistics of shared components in complex component systems. *Physical Review X*, 8(2): 021023, 2018a.

Andrea Mazzolini, Jacopo Grilli, Eleonora De Lazzari, Matteo Osella, Marco Cosentino Lagomarsino, and Marco Gherardi. Zipf and heaps laws from dependency structures in component systems. *Physical review E*, 98(1):012315, 2018b.

Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. *Physical Review E*, 95(2):022117, 2017.

Andrea Montanari, Yiqiao Zhong, and Kangjie Zhou. Tractability from overparametrization: The example of the negative perceptron. *Probability Theory and Related Fields*, 188(3–4): 805–910, 2024. doi: 10.1007/s00440-023-01248-y. URL https://doi.org/10.1007/s00440-023-01248-y. arXiv:2110.15824.

Matteo Negri, Clarissa Lauditi, Gabriele Perugini, Carlo Lucibello, and Enrico Malatesta. Storage and learning phase transitions in the random-features hopfield model. *Physical Review Letters*, 131(25):257301, 2023.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 1177–1184, 2007. URL https://papers.nips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Ludovica Serricchio, Dario Bocchi, Claudio Chilin, Raffaele Marino, Matteo Negri, Chiara Cammarota, and Federico Ricci-Tersenghi. Daydreaming hopfield networks and their surprising effectiveness on correlated data. *Neural Networks*, pp. 107216, 2025.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 2013.

A PRINCIPAL COMPONENTS OF RANDOM-FEATURES DATA WITH POWER-LAW FEATURE STATISTICS

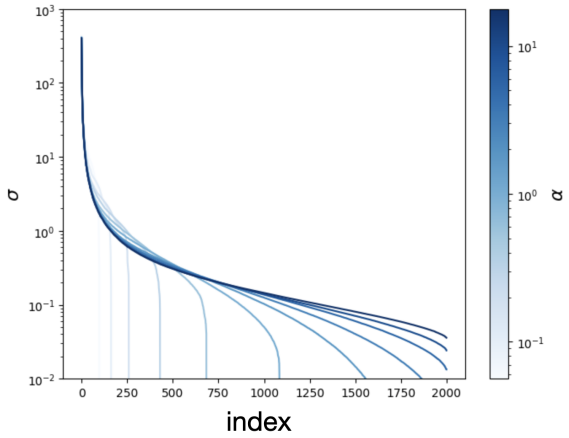

Figure 4: Spectrum of principal components of a random-features dataset with power-law feature statistics. For this plot we chose the same parameters of Fig.1, namely $\eta = 1.5$, $L = 3$, $N = 2000$, $\alpha_D = 0.5$. The color-bar corresponds to different values of the load $\alpha = P/N$.

B DETAILS OF THE SIGNAL-TO-NOISE ANALYSIS

In this section we report the steps for the signal to noise analysis of equation 5. We study the $i-th$ component of the $k-th$ feature local field. To lighten the notation, in the calculation we omit most of the summation boundaries.

B.1 NORMALIZATION

Before evaluating the local field, we need to make some evaluation about the normalization $Z(\eta, D)$, of the power-law feature extracting distribution:

$$p_\eta(k\,;\,D) = \frac{k^{-\eta}}{Z(\eta, D)} \tag{7}$$

It is important to evaluate $Z(\eta, D)$, because, we expect that the normalization of this distribution depends on D and changes behavior with η, which can effect the evaluation of the scalings in the signal-to-noise analysis.

The sum $\sum_k k^{-\eta}$ is not possible to evaluate analytically, therefore, we have to make some approximation. We rewrite the sum:

$$\sum_{k=1}^{D} k^{-\eta} = \int_1^{D+1} \lfloor x \rfloor^{-\eta}\, dx, \tag{8}$$

with $\lfloor x \rfloor$ denoting the floor function, i.e. $\lfloor x \rfloor = \max\{n \in \mathbb{Z}, :, n \leq x\}$. We notice that the continuous function $x^{-\eta}$ is monotonic decreasing, then $\lfloor x \rfloor^{-\eta} \geq x^{-\eta}$ and $\lfloor x \rfloor^{-\eta} \leq (x-1)^{-\eta}$. We observe:

$$\int_1^{D+1} x^{-\eta}\, dx \leq Z(\eta, D) = \int_1^2 \lfloor x \rfloor^{-\eta}\, dx + \int_2^{D+1} \lfloor x \rfloor^{-\eta}\, dx \leq 1 + \int_2^{D+1} (x-1)^{-\eta}\, dx \tag{9}$$

We divided the integral in the second inequality, because $(x-1)^{-\eta}$ diverges for x approaching infinity.

We rewrite this inequality, defining:

$$I(\eta, D) := \int_1^{D+1} x^{-\eta}\, dx, \tag{10}$$

so that,

$$I(\eta, D) \leq Z(\eta, D) \leq 1 + I(\eta, D-1). \tag{11}$$

We calculate the integral in equation 10, to find the scalings:

$$Z(\eta, D) = \begin{cases} O\left(\frac{D^{1-\eta}}{1-\eta}\right) & \text{if } \eta < 1 \\ O\left(\ln D\right) & \text{if } \eta = 1 \\ O\left(\zeta(\eta) + \frac{D^{1-\eta}}{\eta-1}\right) & \text{if } \eta > 1 \end{cases} \tag{12}$$

where $\zeta(\eta)$ is a parameter that does not depends on D.

We observe that the behavior of $Z(\eta, D)$ for $D \to +\infty$ changes at $\eta = 1$. In fact, if $\eta \leq 1$ the normalization diverges with $D^{1-\eta}$ or $\ln D$, indeed, if $\eta > 1$ the normalization converges. From this result we find that the distribution $p_\eta(k)$ is vanishing only if $\eta \leq 1$.

B.2 Extraction without replacement

We notice that the behavior of the probability of selecting the $k-th$ feature in a process of L extractions without replacement $\pi(k)$ is also affected by the value of η. In fact, is we write the probability $\pi(k)$ for $L = 2$:

$$\pi(k) = p_\eta(k; D) + \sum_{k'(\neq k)} \frac{k^{-\eta}}{Z(\eta, D) - (k')^{-\eta}} p_\eta(k'; D) = \tag{13}$$

$$= p_\eta(k; D) \left(1 + \sum_{k'(\neq k)} \frac{(k')^{-\eta}}{Z(\eta, D) - (k')^{-\eta}}\right) \tag{14}$$

We observe that, in case $\eta \leq 1$ the normalization $Z(\eta, D)$ diverges in the limit $D \to \infty$. In this case the correction term $-(k')^{-\eta}$ at the denominator is negligible and we can approximate the term in the parenthesis with 2. We iterate this process for L extraction we find that:

$$\pi(k) \simeq L\, p_\eta(k; D) \tag{15}$$

It is different the case that $Z(\eta, D)$ does not diverges, which corresponds to $\eta > 1$. We notice that:

$$\sum_{k'(\neq k)} \frac{(k')^{-\eta}}{Z(\eta, D) - (k')^{-\eta}} = \sum_{k'(\neq k)} \frac{1}{(k')^{\eta} Z(\eta, D) - 1} < \sum_{k'(\neq k)} \frac{(k')^{-\eta}}{Z(\eta, D)} = 1 - \frac{k^{-\eta}}{Z(\eta, D)} \tag{16}$$

This upper bound approaches 1 only in the case $\eta \leq 1$, because the negative term vanishes as $Z(\eta, D)$ diverges. Indeed, for $\eta > 1$, $Z(\eta, D)$ does not diverges even at the thermodynamic limit, therefore, the upper bound has a dependence by k. We observe that also in this case the negative vanishes if k diverges, meaning that, there are a finite number of feature where extraction without replacement is relevant, but, it is a negligible portion if k diverges. We notice that repeating this process for L extractions increase the effect of the no replacement, but it is not substantial, because, in this case, L is a constant value and does not change when N is increased.

We are now able to define the probability of selecting the $k-th$ feature in a process of L extraction without replacement, as:

$$\pi(k) := \kappa(\eta, L) \frac{k^{-\eta}}{Z(\eta, D)} \tag{17}$$

where $\kappa(\eta, L)$ is an arbitrary value chose to take in account the extraction without replacement.

B.3 FEATURES RETRIEVAL

We study the feature retrieval in case of Hebbian couplings, because is not possible to use an analytical form for the pseudo-likelihood case. We expect that the scaling quantities are not affected by this change, but only their respective load.

Here we expanded the value of the coupling constant J_{ij} and the data ξ_j^μ, with the definitions of power-law features, approximating the sign function $sgn(x) \simeq x$.

$$\begin{aligned} h_{ki} &= \sum_j J_{ij} f_{kj} = \\ &= \frac{1}{N} \sum_j \left(\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \right) f_{kj} = \\ &= \frac{1}{N\,L} \sum_j \sum_\mu \left(\sum_{k'} c_{\mu k'} f_{k'i} \right) \left(\sum_{k''} c_{\mu k''} f_{k''i} \right) f_{kj} = \\ &= \frac{1}{N\,L} \sum_j \sum_\mu \sum_{k'} \sum_{k''} f_{k'i} c_{\mu k'} c_{\mu k''} f_{k''j} f_{kj}. \end{aligned} \tag{18}$$

We now evaluate separately the contributions to the sum, distinguishing the following cases: $k = k' = k''$; $k = k''$ with $k' \neq k''$; and $k' = k''$ with $k \neq k''$.

In particular, when $k = k' = k''$, the corresponding contribution is proportional to f_{ki} and is independent of the specific realization of the other random variables; hence, it contributes deterministically to the signal.

$$\frac{1}{N\,L} \sum_j \sum_\mu f_{ki} c_{\mu k} c_{\mu k} f_{kj} f_{kj}. \tag{19}$$

We notice that the summation over j gives us a term N, berceuse the term f_{kj}^2 is equal to one for every k and j. Otherwise, the term $c_{\mu k}^2$ represents a random variable of a Bernoulli process, which has a probability $\pi(k)$ to be 1 (i.e. we extract the $k-th$ feature to build the $\mu-th$ data) and probability $1-\pi(k)$ to be 0 (i.e. we do not extract it), where $\pi(k)$ is the probability defined in equation 17. In this case the summation over μ is described with a binomial distribution, where the expected value is:

$$P_k := P\,\pi(k).$$

Here we use a law-of-large-numbers argument. The coefficients c_k^μ are independent Bernoulli random variables with mean $\pi(k)$. Since $P \gg 1$, the empirical number of occurrences of feature k, $\sum_{\mu=1}^{P} c_k^\mu$, concentrates around its expectation $P\pi(k)$ with relative fluctuations of order $O(P^{-1/2})$. More precisely, the standard deviation scales as $\sqrt{P\pi(k)(1-\pi(k))}$, which is subleading compared to the signal term $P\pi(k)$ in the thermodynamic limit. In the signal-to-noise analysis we retain only leading-order contributions in N, and therefore replace the empirical frequency by its typical value $P_k = P\pi(k)$. The signal term becomes:

$$\frac{P_k}{L} f_{ki} \tag{20}$$

We now study the terms that are not proportional to f_{ki} or are affected by the randomness of the variables. These are the noise terms which are not contributing to the retrieval of the feature.

Noise 1 : $k' = k''$ and $k \neq k''$

$$\frac{1}{N\,L}\sum_{j(\neq i)}\sum_{\mu}\sum_{k'}\sum_{k''} f_{k'i}c_{\mu k'}c_{\mu k''}f_{k''j}f_{kj}(1-\delta_{kk''})\delta_{k'k''} = \tag{21}$$

$$= \frac{1}{N\,L}\sum_{j(\neq i)}\sum_{\mu}\sum_{k''(\neq k)} f_{k''i}c_{\mu k''}c_{\mu k''}f_{k''j}f_{kj} = \tag{22}$$

$$= \frac{P}{N\,L}\sum_{j(\neq i)}\sum_{k''(\neq k)} \pi(k'')f_{k''i}f_{k''j}f_{kj} \tag{23}$$

$$= \frac{P}{N\,L}\sum_{j(\neq i)}\sum_{k''(\neq k)} \varphi_{ijkk''}\pi(k'') \tag{24}$$

To determine the order of magnitude of this contribution, we introduce the random variable $\varphi_{ijkk''} := f_{k''i}f_{k''j}f_{kj}$. Since the random variables $f_{k''i}$, $f_{k''j}$, and f_{kj} are independently and uniformly distributed over $\{-1, 1\}$, the variable $\varphi_{ijkk''}$ also takes values in $\{-1, 1\}$ with equal probability, i.e.:

$$\begin{cases} p(\varphi_{ijkk''} = +1) & = \frac{1}{2} \\ \\ p(\varphi_{ijkk''} = -1) & = \frac{1}{2} \end{cases} \tag{25}$$

We use the moment-generating function to find the moments of this noise term distribution:

$$g_\eta(t) = \sum_{\boldsymbol{\varphi}} p(\boldsymbol{\varphi}) \exp\left(t\sum_{j(\neq i)}\sum_{k''(\neq k)} \varphi_{ijkk''}\pi(k'')\right) = \tag{26}$$

$$= \sum_{\boldsymbol{\varphi}} p(\boldsymbol{\varphi}) \prod_{j(\neq i)}\prod_{k''(\neq k)} \exp\left(t\varphi_{ijkk''}\pi(k'')\right) = \tag{27}$$

$$= \prod_{j(\neq i)}\prod_{k''(\neq k)}\sum_{\varphi_{ijkk''}} p(\varphi_{ijkk''}) \exp\left(t\varphi_{ijkk''}\pi(k'')\right) = \tag{28}$$

$$= \prod_{j(\neq i)}\prod_{k''(\neq k)} \frac{1}{2} 2\cosh(t\pi(k'')) = \tag{29}$$

$$= \prod_{k''(\neq k)} \left(t\cosh(\pi(k''))\right)^{N-1} \tag{30}$$

This moment-generating function corresponds to a binary variable taking values in $\{-\pi(k''),\, \pi(k'')\}$, repeated $N-1$ times for each $k'' \neq k$. Moreover, it is properly normalized, since $g_\eta(0) = 1$.

We now find the first moment of the distribution, from $\mu_1 = g'(0)$:

$$g'_\eta(t) = (N-1)\sum_{k''(\neq k)} \pi(k'')g_\eta(t)\left(\cosh(t\,\pi(k''))\right)^{-1}\sinh(t\,\pi(k'')). \tag{31}$$

We observe that the average of $\varphi_{ijkk''}$ is zero. Indeed, in the expression of the moment-generating function each term in the summation contains a factor $\sinh\big(t\,\pi(k'')\big)$, which vanishes at $t = 0$.

We now find the second moment of the distribution, from $\mu_2 = g''(0)$:

$$\begin{aligned} g''_\eta(t) =& (N-1)\sum_{k''(\neq k)} \pi(k'')[g'_\eta(t)(\cosh(t\,\pi(k'')))^{-1}\sinh(t\,\pi(k''))+ \\ &+ g_\eta(t)(\cosh(t\,\pi(k'')))^{-2}\pi(k'')\sinh^2(t\,\pi(k''))+ \\ &+ g_\eta(t)(\cosh(t\,\pi(k'')))^{-1}\pi(k'')\cosh(t\,\pi(k''))]. \end{aligned} \tag{32}$$

We notice that, if $t=0$, the first two terms in the square brackets have a term equal to zero ($g'(0)$ or $\sinh(0)$), then the only remaining term is the third one, equal to $\pi(k'')$. The second moment became:

$$\begin{aligned}\mu_2 = g''_\eta(0) &= (N-1)\sum_{k''(\neq k)} \pi^2(k'') = \\ &= (N-1)\left(\frac{\kappa(\eta,L)}{Z(\eta,D)}\right)^2 \sum_{k''(\neq k)} (k'')^{-2\eta} = \\ &= (N-1)\left(\frac{\kappa(\eta,L)}{Z(\eta,D)}\right)^2 (Z(2\eta,D)-k^{-2\eta}) = \\ &= NZ(2\eta,D)\left(\frac{\kappa(\eta,L)}{Z(\eta,D)}\right)^2 \left(1-\frac{k^{-2\eta}}{Z(2\eta,D)}\right)\left(1-\frac{1}{N}\right) = \\ &= \frac{P_k^2}{P^2} NZ(2\eta,D)k^{2\eta}\left(1-\frac{k^{-2\eta}}{Z(2\eta,D)}\right)\left(1-\frac{1}{N}\right).\end{aligned} \tag{33}$$

We estimate the order of magnitude of the noise contribution through its standard deviation, $\sigma_1 = \sqrt{\mu_2 - \mu_1^2}$. In the thermodynamic limit, using that $Z(2\eta,D) \sim D^{1-2\eta}, \ln D$, or $\zeta(\eta)$, depending on the value of η, the terms inside the parentheses approach either 1 or a finite constant. Therefore, they do not affect the scaling with D, and the noise term scales as

$$\longrightarrow \frac{P_k}{L} O\left(\sqrt{\frac{Z(2\eta,D)}{Nk^{-2\eta}}}\right). \tag{34}$$

Where we isolated the signal term, in front of the order of approximation notation term.

We repeat the same process for the other noise terms.

Noise 2 : $k = k''$ and $k' \neq k''$

$$\frac{1}{NL}\sum_{j(\neq i)}\sum_{\mu}\sum_{k'}\sum_{k''} f_{k'i}c_{\mu k'}c_{\mu k''}f_{k''j}f_{kj}\delta_{kk''}(1-\delta_{k'k''}) = \tag{35}$$

$$= \frac{1}{NL}\sum_{j(\neq i)}\sum_{\mu}\sum_{k'(\neq k)} f_{k'i}c_{\mu k'}c_{\mu k''} = \tag{36}$$

$$= \frac{N-1}{NL}\sum_{\mu}\sum_{k'(\neq k)} f_{k'i}c_{\mu k'}c_{\mu k} = \tag{37}$$

$$= \frac{N-1}{NL}\sum_{\mu}\sum_{k'(\neq k)} \varphi_{i\mu kk'}. \tag{38}$$

We define the random variable $\varphi_{i\mu kk'} := f_{k'i}c_{\mu k'}c_{\mu k}$. In this case $\varphi_{i\mu kk'}$ is extracted among $\{-1,0,1\}$, the probability distribution is:

$$\begin{cases} p(\varphi_{i\mu kk'}=+1) & = \frac{1}{2}\pi(k)\pi(k') \\ p(\varphi_{i\mu kk'}=-1) & = \frac{1}{2}\pi(k)\pi(k') \\ p(\varphi_{i\mu kk'}=0) & = 1-\pi(k)\pi(k'). \end{cases} \tag{39}$$

Then, the moment-generating function is now

$$g_\eta(t) = \prod_{\mu}\prod_{k'(\neq k)}\sum_{\varphi_{i\mu kk'}} p(\varphi_{i\mu kk'})\exp\left(t\varphi_{i\mu kk'}\right) = \tag{40}$$

$$= \prod_{\mu}\prod_{k'(\neq k)} \left[\pi(k)\pi(k')\left(\cosh(t)-1\right)+1\right] = \tag{41}$$

$$= \prod_{k'(\neq k)} \left[\pi(k)\pi(k')\left(\cosh(t)-1\right)+1\right]^P. \tag{42}$$

This implies that $g_\eta(0) = 1$. We then compute the first moment from the first derivative,

$$g'_\eta(t) = \sum_{k' \neq k} P\pi(k)\pi(k')\, g_\eta(t)(\pi(k)\pi(k')(\cosh(t) - 1) + 1)^{-1} \sinh(t), \tag{43}$$

from which it follows that $\mu_1 = g'_\eta(0) = 0$.

We then compute the second moment from the second derivative

$$\begin{aligned} g''_\eta(t) = \sum_{k' \neq k} & P\pi(k)\pi(k')[g'_\eta(t)(\cosh(t) - 1) + 1)^{-1} \sinh(t) + \\ & - g_\eta(t)(\cosh(t) - 1) + 1)^{-2} \sinh^2(t)\pi(k)\pi(k') + \\ & + g_\eta(t)(\cosh(t) - 1) + 1)^{-1} \cosh(t)], \end{aligned} \tag{44}$$

thus obtaining

$$\begin{aligned} \mu_2 = g''_\eta(0) &= P\pi(k) \sum_{k'(\neq k)} \pi(k') = \\ &= P\pi(k) \frac{\kappa(\eta)}{Z(\eta, D)} Z(\eta, D) \left(1 - \frac{k^{-\eta}}{Z(\eta, D)}\right) = \\ &= P_k^2 \frac{Z(\eta, D)}{P\, k^{-\eta}} \left(1 - \frac{k^{-\eta}}{Z(\eta, D)}\right). \end{aligned} \tag{45}$$

Concluding that the order of approximation of the second noise term is:

$$\longrightarrow \frac{P_k}{L} O\left(\sqrt{\frac{Z(\eta, D)}{P\, k^{-\eta}}}\right). \tag{46}$$

For the case $k \neq k''$ and $k' \neq k''$, we separate the case $k = k'$ and $k \neq k'$, because, in the first case, the summation over k' is neglected, which gives a different term of noise.

Noise 3a : $k \neq k''$, $k' \neq k''$ and $k = k'$ The term whose noise we want to compute is now

$$\frac{1}{N\, L} \sum_{j(\neq i)} \sum_{\mu} \sum_{k'} \sum_{k''} f_{k'i} c_{\mu k'} c_{\mu k''} f_{k''j} f_{kj} (1 - \delta_{kk''})(1 - \delta_{k'k''}) \delta_{kk'} = \tag{47}$$

$$= \frac{1}{N\, L} f_{ki} \sum_{j(\neq i)} \sum_{\mu} \sum_{k''(\neq k)} c_{\mu k} c_{\mu k''} f_{kj} f_{k''j} = \tag{48}$$

$$= \frac{1}{N\, L} f_{ki} \sum_{j(\neq i)} \sum_{\mu} \sum_{k''(\neq k)} \varphi_{ij\mu kk''}. \tag{49}$$

Again, we define $\varphi_{ij\mu kk''} := c_{\mu k} c_{\mu k''} f_{kj} f_{k''j}$, which has a probability distribution:

$$\begin{cases} p(\varphi_{ij\mu kk''} = +1) & = \frac{1}{2}\pi(k)\pi(k''), \\ p(\varphi_{ij\mu kk''} = -1) & = \frac{1}{2}\pi(k)\pi(k''), \\ p(\varphi_{ij\mu kk''} = 0) & = 1 - \pi(k)\pi(k''). \end{cases} \tag{50}$$

We notice that, in this term of noise, it appears the term f_{ki}, but, because is proportional to a random variable, this term is not part of the signal.

$$g_\eta(t) = \prod_{j(\neq i)} \prod_{\mu} \prod_{k''(\neq k)} \sum_{\varphi_{ij\mu kk''}} p(\varphi_{ij\mu kk''}) \exp\left(t \varphi_{ij\mu kk''}\right) = \tag{51}$$

$$= \prod_{j(\neq i)} \prod_{\mu} \prod_{k''(\neq k)} \left(\pi(k)\pi(k'')\left(\cosh(t) - 1\right) + 1\right) = \tag{52}$$

$$= \prod_{k''(\neq k)} \left(\pi(k)\pi(k'')\left(\cosh(t) - 1\right) + 1\right)^{P(N-1)}. \tag{53}$$

For next computation, we show that $g_\eta(0) = 1$.

$$g'_\eta(t) = \sum_{k'' \neq k} P(N-1)\pi(k)\pi(k'')\, g_\eta(t)(\pi(k)\pi(k'')(\cosh(t)-1)+1)^{-1}\sinh(t). \tag{54}$$

From this result, we find that $\mu_1 = 0$.

$$\begin{aligned} g''_\eta(t) = \sum_{k'' \neq k} & P(N-1)\pi(k)\pi(k'')[g'_\eta(t)(\cosh(t)-1)+1)^{-1}\sinh(t)+ \\ & - g_\eta(t)(\cosh(t)-1)+1)^{-2}\sinh^2(t)\pi(k)\pi(k'')+ \\ & + g_\eta(t)(\cosh(t)-1)+1)^{-1}\cosh(t)]. \end{aligned} \tag{55}$$

$$\begin{aligned} \mu_2 = g''_\eta(0) &= P(N-1)\pi(k)\sum_{k''(\neq k)}\pi(k'') = \\ &= P(N-1)\pi(k)\frac{\kappa(\eta)}{Z(\eta,D)}Z(\eta,D)\left(1-\frac{k^{-\eta}}{Z(\eta,D)}\right) = \\ &= N^2P_k^2\frac{Z(\eta,D)}{N\,P\,k^{-\eta}}\left(1-\frac{k^{-\eta}}{Z(\eta,D)}\right)\left(1-\frac{1}{N}\right). \end{aligned} \tag{56}$$

We find that, this term of noise is:

$$\longrightarrow \frac{P_k}{L}O\left(\sqrt{\frac{Z(\eta,D)}{N\,P\,k^{-\eta}}}\right) \tag{57}$$

Noise 3b : $k \neq k''$, $k' \neq k''$ and $k \neq k'$

$$\frac{1}{N\,L}\sum_{j(\neq i)}\sum_{\mu}\sum_{k'}\sum_{k''} f_{k'i}c_{\mu k'}c_{\mu k''}f_{k''j}f_{kj}(1-\delta_{kk''})(1-\delta_{k'k''})(1-\delta_{kk'}) = \tag{58}$$

$$= \frac{1}{N\,L}\sum_{j(\neq i)}\sum_{\mu}\sum_{k'(\neq k)}\sum_{k''(\neq k,k')} f_{k'i}c_{\mu k'}c_{\mu k''}f_{kj}f_{k''j} = \tag{59}$$

$$= \frac{1}{N\,L}\sum_{j(\neq i)}\sum_{\mu}\sum_{k'(\neq k)}\sum_{k''(\neq k,k')} \varphi_{ij\mu kk'k''}. \tag{60}$$

We define the random variable $\varphi_{ij\mu kk'k''} := f_{k'i}c_{\mu k'}c_{\mu k''}f_{kj}f_{k''j}$, which has the probability distribution:

$$\begin{cases} p(\varphi_{ij\mu kk''} = +1) & = \frac{1}{2}\pi(k')\pi(k''), \\ p(\varphi_{ij\mu kk''} = -1) & = \frac{1}{2}\pi(k')\pi(k''), \\ p(\varphi_{ij\mu kk''} = 0) & = 1-\pi(k')\pi(k''), \end{cases} \tag{61}$$

$$g_\eta(t) = \prod_{j(\neq i)}\prod_{\mu}\prod_{k'(\neq k)}\prod_{k''(\neq k,k')}\sum_{\varphi_{ij\mu kk'k''}} p(\varphi_{ij\mu kk'k''})\exp\left(t\varphi_{ij\mu kk'k''}\right) = \tag{62}$$

$$= \prod_{j(\neq i)}\prod_{\mu}\prod_{k'(\neq k)}\prod_{k''(\neq k,k')}(\pi(k)\pi(k'')\left(\cosh(t)-1\right)+1) = \tag{63}$$

$$= \prod_{k'(\neq k)}\prod_{k''(\neq k,k')}\left(\pi(k')\pi(k'')\left(\cosh(t)-1\right)+1\right)^{P(N-1)}. \tag{64}$$

$$g'_\eta(t) = g_\eta(t)P(N-1)\sum_{k'(\neq k)}\sum_{k''(\neq k,k')}\pi(k')\pi(k'')(\cosh(t)-1)+1)^{-1}\sinh(t). \tag{65}$$

$$\begin{aligned} g''_\eta(t) = & P(N-1)\sum_{k'(\neq k)}\sum_{k''(\neq k,k')}\pi(k')\pi(k'')[g'_\eta(t)(\cosh(t)-1)+1)^{-1}\sinh(t)+ \\ & + \pi(k')\pi(k'')g_\eta(t)(\cosh(t)-1)+1)^{-2}\sinh^2(t)+ \\ & + g_\eta(t)(\cosh(t)-1)+1)^{-1}\cosh(t)]. \end{aligned} \tag{66}$$

$$\begin{aligned}\mu_2 = g_\eta''(0) &= P(N-1)\sum_{k'(\neq k)}\sum_{k''(\neq k,k')}\pi(k')\pi(k'') = \\ &= P(N-1)\left(\frac{\kappa(\eta)}{Z(\eta,D)}\right)^2\sum_{k'(\neq k)}(k')^{-\eta}(Z(\eta,D)-k^{-\eta}-(k')^{-\eta}) = \\ &= N^2P_k^2\frac{k^{2\eta}}{N\,P}\left(1-\frac{1}{N}\right)\left((Z(\eta,D)-k^{-\eta})^2-Z(2\eta,D)+k^{-2\eta}\right).\end{aligned} \tag{67}$$

We find that the last term of noise is:

$$\longrightarrow \frac{P_k}{L}O\left(\sqrt{\frac{(k^\eta Z(\eta,D))^2-k^{2\eta}Z(2\eta,D)}{N\,P}}\right) \tag{68}$$

Collecting all contributions:

Putting everything together, we have:

$$\begin{aligned}h_{ik} = \frac{P_k}{L}\Bigg(f_{ki} &+ O\left(\sqrt{\frac{Z(2\eta,D)}{N\,k^{-2\eta}}}\right) + O\left(\sqrt{\frac{Z(\eta,D)}{Pk^{-\eta}}}\right) + O\left(\sqrt{\frac{Z(\eta,D)}{P\,N\,k^{-\eta}}}\right) + \\ &+ O\left(\sqrt{\frac{(k^\eta Z(\eta,D))^2-k^{2\eta}Z(2\eta,D)+1}{NP}}\right)\Bigg)\end{aligned} \tag{69}$$

As a cross-check, we observe that in the limit $\eta \to 0$, the quantity h_{ik} reduces to the result obtained in the signal-to-noise analysis of random features without a power-law weighting.

We observe that, using the results of Section B.1, the noise exhibits three distinct regimes, due to its dependence on both $Z(\eta, D)$ and $Z(2\eta, D)$. These regimes correspond to $\eta < \frac{1}{2}$, $\frac{1}{2} < \eta < 1$, and $\eta > 1$.

From the signal-to-noise analysis, we can study the scaling of the maximum index k, which represents the number of learnable features. We have to find the possible scaling that make the noise terms at most $O(1)$, i.e. the same order of the signal.

Case $\boldsymbol{\eta < \frac{1}{2}}$:

$$\begin{aligned}h_{ik} = \frac{P_k}{L}\Bigg(f_{ki} &+ O\left(\sqrt{\frac{D}{N}\left(\frac{k}{D}\right)^{2\eta}}\right) + O\left(\sqrt{\frac{D}{P}\left(\frac{k}{D}\right)^{\eta}}\right) + O\left(\sqrt{\frac{D}{P\,N}\left(\frac{k}{D}\right)^{\eta}}\right) + \\ &+ O\left(\sqrt{\frac{D^2}{NP}\left(\frac{k}{D}\right)^{2\eta}}\right)\Bigg)\end{aligned} \tag{70}$$

Assuming that both P and D scale as $O(N)$, we observe that the maximal scaling of k is $O(N)$, as follows from the noise contributions labeled 1, 2, and $3a$. We notice that, in case k and D have the same scaling, the power-law contribution gives a term correction that does not change the scaling of the noise terms, therefore, the behavior of the model is similar to the one of the uniformly distributed random feature.

Case $\boldsymbol{\eta > \frac{1}{2}}$:

First, we notice that, in this case, the first term of noise is not effected by the number of features D. In fact, this term define a maximum scaling of $k = O\left(N^{\frac{1}{2\eta}}\right)$, which does not depends by D. We will see that this is the main term of noise if both P and D are $O(N)$.

Case $\boldsymbol{\frac{1}{2} < \eta < 1}$:

$$h_{ik} = \frac{P_k}{L}\left(f_{ki} + O\left(\sqrt{\frac{k^{2\eta}}{N}}\right) + O\left(\sqrt{\frac{D^{1-\eta}}{Pk^{-\eta}}}\right) + O\left(\sqrt{\frac{D^{1-\eta}}{P\,N\,k^{-\eta}}}\right) + O\left(\sqrt{\frac{(k^\eta D^{1-\eta})^2}{NP}}\right)\right) \tag{71}$$

We notice that, with the assumption which P and D are $O(N)$ and the observation of the scaling of k, the first term of noise is the only one that does not vanish. Also, in this case we are able to increase the scaling of D to a maximum scaling of $O\left(N^{\frac{1}{2(1-\eta)}}\right)$, which is greater than $O(N)$.

Case $\eta > 1$:

$$h_{ik} = \frac{P_k}{L}\left(f_{ki} + O\left(\sqrt{\frac{k^{2\eta}}{N}}\right) + O\left(\sqrt{\frac{k^{\eta}}{P}}\right) + O\left(\sqrt{\frac{k^{\eta}}{P\,N}}\right) + O\left(\sqrt{\frac{k^{2\eta}}{NP}}\right)\right) \tag{72}$$

First, we observe that the noise terms are not affected by D anymore, which has only the constrain of $D \geq k$. Furthermore, we notice that the only noise term which does not vanish is the first one if we assume $P = O(N)$. Also, in this case is we have a minimum scaling for P of $O(\sqrt{N})$.

C Details of the training procedure

Pseudo-likelihood cost is convex, therefore most of complications are not present, and a simple Gradient Descent with full batch is expected to converge to the optimizer. However, if the norm of the weights is not constrained and no regularizer is present, this optimizer would correspond to an infinite norm of the weights. For this reason we start from small weights, we optimize by using gradient descent, and we measure the final overlap of test data. When this overlap stops to increase for $\mathcal{O}(10^2)$ epochs we stop the training, preventing the norm to diverge.