EMERGENT PHASES OF SUPERPOSITION: FROM PARTIAL TO FULL REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are thought to represent features by vectors in a hidden space of dimension given by the model's width. Superposition, in which more features are represented than the width by letting representation vectors overlap, is a leading account of how representation vectors are organized. However, how model width and data statistics determine the configuration of representation vectors and the resulting loss when the number of features and the width are large remains less understood. Here we show, in Anthropic's toy model of superposition, that increasing the width drives a continuous phase transition from a partial-representation phase, where only a subset of features receives appreciable representation vectors while the rest vanish, to a full-representation phase, where every feature is represented. Our theory via a partial random projection approximation predicts, and experiments confirm, that the critical width grows linearly with the number of active features up to a logarithmic factor. The loss scaling changes across the transition: below the critical width, the loss grows linearly with the number of active features and depends weakly on the width in a form set by data statistics; above it, the loss grows approximately quadratically with the number of active features and decays inversely with the width. Non-uniform firing probabilities delay the transition and lower the loss, as more frequent features occupy more space. Our results provide an account of how model width and data statistics jointly shape representations and loss, a step toward understanding representation scaling in large models.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable empirical success across a wide range of tasks, yet the internal mechanism behind this success remains unclear. A leading view is that LLMs compute with features. Features are interpretable properties of the input, and serve as the fundamental units of concept representation and computation (Elhage et al., 2022; Bricken et al., 2023). Under the linear representation hypothesis, each feature corresponds to a direction in hidden space (its representation vector), and the hidden state of an input is approximately a linear combination of the representation vectors of its active features (Mikolov et al., 2013; Park et al., 2023). Consistent with this picture, dictionary-learning methods have extracted millions of such feature directions from LLM hidden layers (Bricken et al., 2023; Templeton et al., 2024). How these representation vectors are organized in hidden space is thus a central question for understanding LLMs.

How representation vectors are organized, and how that organization sets performance, has been widely studied. A linear network with hidden dimension m (the model width) can represent at most m features orthogonally, and its optimum is a principal-component projection onto the most important ones; the rest are discarded (Baldi & Hornik, 1989; Elhage et al., 2022). Nonlinearity lifts this ceiling: the network can pack more features than dimensions by letting representation vectors overlap and using the nonlinear layer to filter the resulting interference, a mechanism called superposition (Elhage et al., 2022). Superposition has since been characterized in two regimes. When the numbers of active features and hidden dimensions are both of order one, the representation configuration is determined by the geometry of a few vectors, which settle into specific configurations such as antipodal pairs and polytopes (Elhage et al., 2022; Chen et al., 2023). In the opposite, thermodynamic limit, existing analyses take the picture to be statistical: every feature is assigned a representation vector, the vectors are treated as nearly orthogonal, and their overlaps set a loss that decays inversely

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

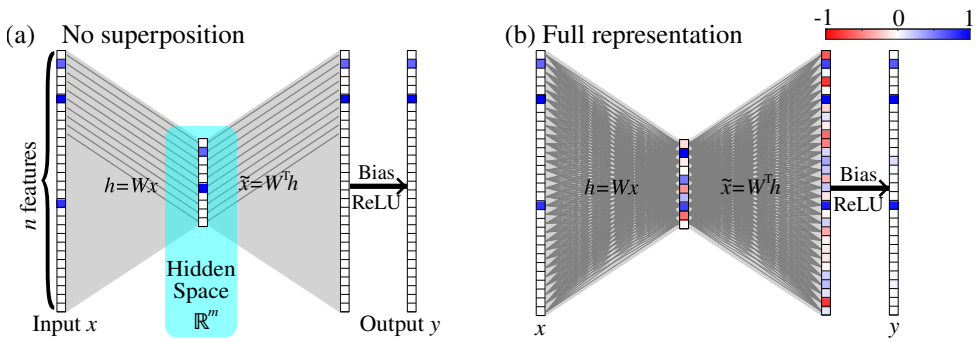


Figure 1: The toy model composed of embedding and unembedding layers exhibits different types of representation configurations. (a) and (b) are schematic illustrations of the toy model. Colored boxes are used to represent the element values of the input, hidden, and output vectors. (a) No superposition. Only m features are represented by vectors without overlap, and the remaining features are discarded. (b) Full representation. All features are represented by vectors with appreciable norms.

with the width (Cowsik et al., 2025; Liu et al., 2025). What has not been established is whether this picture holds across model widths and data statistics: whether a model in this limit can instead leave some features unrepresented, when it does so, and how the loss changes across that boundary. This leads to the question we address: **How do model width and data statistics determine the representation behaviors (which features are represented, with what representation vector) and the resulting loss?**

We address this question in the toy model of superposition introduced by Anthropic (Elhage et al., 2022), varying model width and data statistics. In trained models, we identify two types of configurations of the representation vectors: a partial representation at small model width m, where features split into a group with appreciable norms and a group with vanishing norms (even though all features are uniform), and a full representation at large width, where every feature carries an appreciable vector. As m grows, the number of represented features increases continuously until full representation is reached, and the model undergoes a phase transition from partial representation to full representation. We develop an approximation of the representation behavior and the loss, which reproduces the trends of trained models and captures their asymptotic behavior in the thermodynamic limit. The loss follows the representation, showing different behaviors on two sides of the transition. Finally, we characterize how sparsity and non-uniformity shift the transition and the loss.

2 Toy model

We adopt the toy model of superposition (Elhage et al., 2022), see Fig. 1. In this model, an input $\boldsymbol{x} \in \mathbb{R}^n$ is mapped to a lower-dimensional vector $\boldsymbol{h}$ in the hidden space $\mathbb{R}^m$ through matrix $\boldsymbol{W} \in \mathbb{R}^{m\times n}$, and then unembedded:

$$\boldsymbol{h} \equiv \boldsymbol{W}\boldsymbol{x}, \qquad \boldsymbol{y} \equiv \left[\boldsymbol{W}^T\boldsymbol{h} + \boldsymbol{b}\right]_+, \tag{1}$$

where $[z]_+ \equiv \max(z, 0)$ is the ReLU operator, and $\boldsymbol{b} \in \mathbb{R}^n$ is the bias vector. The hidden space has lower dimensionality than the input space, acting as the bottleneck, hence m is also called the model width. The loss is defined as the mean square error between the output $\boldsymbol{y}$ and the input $\boldsymbol{x}$:

$$\mathcal{L} \equiv \frac{1}{n}\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{2}$$

Unlike Elhage et al. (2022), we assign equal importance to all features. During training, the model aims to minimize $\mathcal{L}$ by optimizing $\boldsymbol{W}$ and $\boldsymbol{b}$. See appendix B for training details.

The input $\boldsymbol{x}$ is interpreted as a list of activations, whose ith element x_i is the activation of feature i. Under this interpretation, n is the total number of features. The activations follow

$$x_i = u_i v_i, \tag{3}$$

where u_i and v_i are all independent. The random variable u_i follows Bernoulli distribution $\mathcal{B}(p_i)$, where p_i is the firing probability of feature i. It controls whether a feature is active (non-zero) or not. Random variables v_i control activation strengths of active features. They are i.i.d. random variables following a distribution $\mathcal{D}$ with mean $\bar{v}$ and variance σ_v^2. In this paper, for simplicity, we let $\mathcal{D}$ be the uniform distribution $\mathcal{U}[\bar{v} - \sqrt{3}\sigma_v, \bar{v} + \sqrt{3}\sigma_v]$ and let $\bar{v} = 1, \sigma_v = \frac{1}{\sqrt{3}}$ (i.e., $v \sim \mathcal{U}[0, 2]$ as in Liu et al. (2025)). The effects of $\bar{v}$ and σ_v are discussed in appendix C.1.2. The input is sparse in the sense that most elements are zero. Denoting the number of active features by $E \equiv \sum_i u_i$, we use its mean $\bar{E}$ to quantify the sparsity of the input ensemble; it satisfies $\bar{E} \ll n$.

Under the feature interpretation, the embedding matrix $\boldsymbol{W}$ can be considered as n vectors:

$$\boldsymbol{W} = (\boldsymbol{w}_1\ \boldsymbol{w}_2\ ...\ \boldsymbol{w}_n), \tag{4}$$

where $\boldsymbol{w}_i$ is the representation vector of feature i in the hidden space $\mathbb{R}^m$. Their norms are denoted by $s_i \equiv \|\boldsymbol{w}_i\|$. Depending on the data properties and the model width m, some features may not be represented at the loss minimum, i.e., the norms of their representation vectors are close to or exactly 0. We define the effective number of represented features as the participation ratio of norms:

$$r \equiv \frac{\left(\sum_i s_i^2\right)^2}{\sum_i s_i^4}. \tag{5}$$

The participation ratio r ranges from 1 to n, counting vectors with appreciable norm while ignoring near-zero ones; if k vectors share the same norm and the remaining vectors vanish, $r = k$. When $r \leq m$, vectors of represented features can be orthogonal and be free of superposition. The corresponding loss is given in appendix A.1. When $m < r < n$, vectors have to overlap and superposition emerges, but some features are sacrificed and the representation is partial. When $r = n$, all features are represented (full representation).

3 Results

In Sec. 3.1, we discuss the effect of model width m on the representation behavior and the loss, and show the phase transition from partial to full representation as m increases. Based on the phenomena, we develop the partial random projection (RP) approximation to estimate the representation behavior and the loss. In Sec. 3.2, we apply the approximation to study the effects of sparsity. In Sec. 3.3, we further introduce non-uniformity to firing probabilities, and study its effects.

3.1 Width drives a transition from partial to full representation

In this section, we focus on how model width m influences the representation behavior and the loss $\mathcal{L}$. For this, we simply let all features have the same firing probability $p \equiv \frac{\bar{E}}{n}$.

Given n and $\bar{E}$, the converged representation of the trained model (for 3×10^4 steps; appendix B) shows different behaviors at different m. At small m, features are split into two groups. In one group, features have nearly vanishing s_i, and are essentially unrepresented. In another group, features have s_i of order 1 and are well represented. Norms s_i are nearly equal within each group. At large m, all features have appreciable representation vectors with nearly equal s_i (Fig. 2a). Biases behave in the same way as norms: At small m, biases are nearly equal within each group; at large m, all biases are nearly equal (Fig. 2b). As m increases, more features are represented, until full representation is reached and a phase transition happens. We define the critical width m^* as the minimal m at which the distribution of s_i becomes unimodal.

Correspondingly, the loss $\mathcal{L}$ also shows different behaviors on two sides of the transition. Before the transition, the loss $\mathcal{L}$ decays slowly with m; after the transition, the loss decays approximately as $1/m$ (Fig. 2c). The $\mathcal{L}$-m curve bends around m^*.

Note that the partial representation before the transition does not mean no superposition. The number r of features represented is always greater than m (Fig. 2d), and r/m peaks around the transition. In what follows we compare r with n (completeness of representation), not with m. In summary:

Key result 1: As the width m increases, the model undergoes a phase transition from partial representation to full representation.

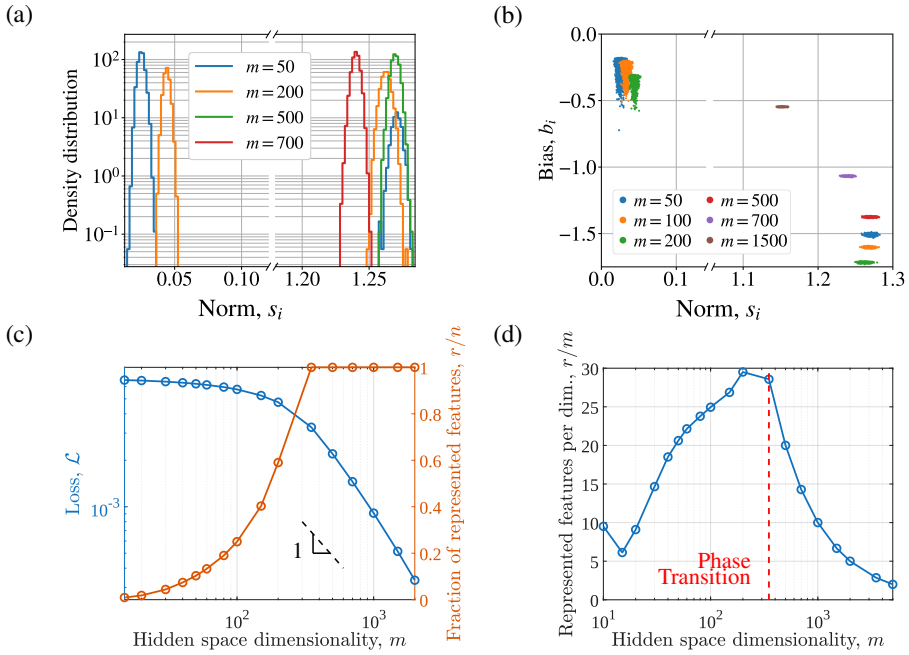


Figure 2: As the width grows, the model undergoes a transition from partial to full representation. (a) Distributions of s_i at different m. At small m the distribution is bimodal while at large m the distribution is unimodal. (b) The distribution of s_i and b_i of each feature. Sample points at small m form two clusters while at large m they form one cluster. (c) The loss $\mathcal{L}$ and the fraction r/n of represented features as functions of m. (d) The ratio r/m as a function of m. It is always greater than 1, indicating superposition. $n = 10000, \bar{E} = 50$.

We use the term phase transition in the following sense. The fraction r/n serves as the order parameter: It is below 1 for $m < m^*$ and equal to 1 for $m > m^*$, and it reaches 1 continuously, so the transition is continuous. The loss is smooth on either side of m^*, but its dependence on m changes at m^*. Strictly, a sharp transition requires the limit $1 \ll \bar{E}, m \ll n$, in which m can be treated as continuous.

To understand these phenomena, we develop an approximation to $\mathcal{L}$, which we call the partial random projection (RP) approximation. We idealize the representation vectors $\boldsymbol{w}_i$ as forming two groups. One group contains vectors of r represented features, each with the same norm $s_i = s$. The other group contains zero vectors of $(n - r)$ unrepresented features (Fig. 3a). In the represented group, biases $b_i = b$ are also assumed equal. We apply this approximation to both partial and full representation phases, as we expect the unrepresented group to vanish spontaneously at the loss minimum in the regime of full representation. The loss is thus composed of the contributions of two groups:

$$\mathcal{L}(\bar{E}, m, n; r) = \frac{1}{n}\left(r\mathcal{L}_{\text{rep}} + (n-r)\mathcal{L}_{\text{unrep}}\right). \tag{6}$$

The loss $\mathcal{L}_{\text{unrep}}$ contributed by unrepresented features is given in appendix A.3:

$$\mathcal{L}_{\text{unrep}} = (\bar{v}^2 + \sigma_v^2)\frac{\bar{E}}{n} - \bar{v}^2\frac{\bar{E}^2}{n^2}. \tag{7}$$

Given an input $\boldsymbol{x}$, the loss $\mathcal{L}_{\text{rep}}$ contributed by represented features can be further divided into two terms:

$$\mathcal{L}_{\text{rep}}|_{\boldsymbol{x}} = \frac{1}{r}\left\{\sum_{i=1}^{E_r}\left[\left[s^2 v_{a_i} + s^2\eta_{a_i} + b\right]_+ - v_{a_i}\right]^2 + \sum_{i\notin\{a_k\}}\left[s^2\eta_i + b\right]_+^2\right\}, \tag{8}$$

where the first term is the loss contribution of active features, and the second term is that of inactive features. The set $\{a_k\}$ contains all active features within the represented group, and E_r is the number

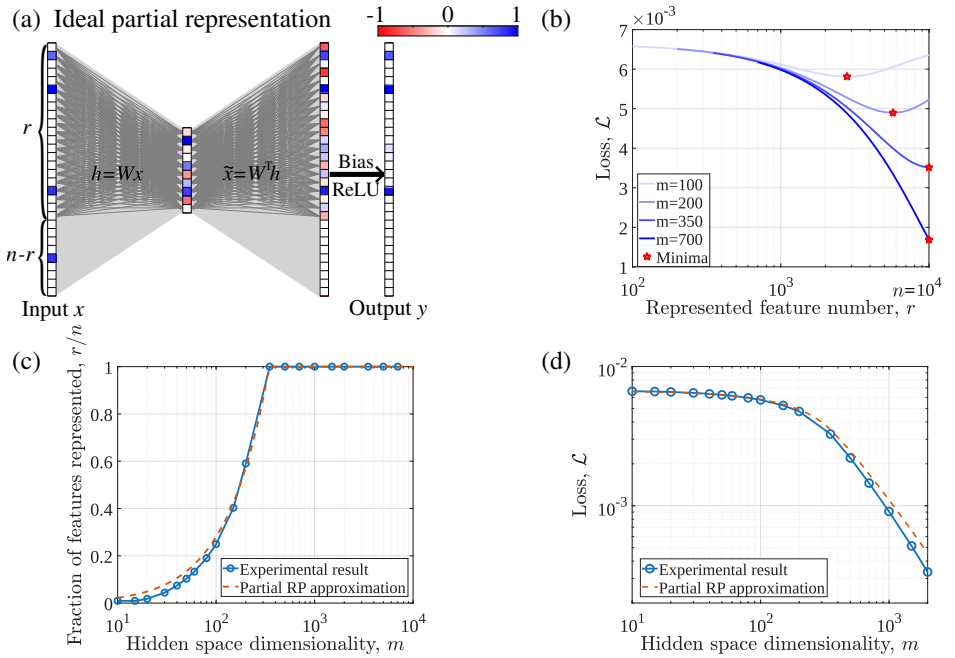


Figure 3: The partial random projection approximation captures the behavior of toy model. (a) Schematic illustration of ideal partial representation. (b) The loss $\mathcal{L}$ predicted by partial RP approximation as a function of r at different m. (c) The predicted r_{opt} shows good consistency with the experimental r. (d) The predicted $\mathcal{L}$ shows good consistency with experimental loss. $n = 10000, \bar{E} = 50$.

of them. The crosstalk noise on feature i is

$$\eta_i \equiv \sum_{j \in \{a_k\} \setminus i} \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_j}{s^2} v_j. \tag{9}$$

We treat η_i as random variables independent of i and of $\boldsymbol{v}$ by mean field approximation, and further use a Gaussian distribution to approximate it:

$$\eta_i \sim \mathcal{N}\left(0, E_r(\bar{v}^2 + \sigma_v^2)\frac{r-m}{m(r-1)}\right). \tag{10}$$

The variance of η_i is factorized into three pieces: the number of active represented features E_r, the activation second moment $(\bar{v}^2 + \sigma_v^2)$, and the minimal mean squared overlap $\frac{r-m}{m(r-1)}$ confined by the Welch bound (Welch, 1974). For $m \ll r$ the overlap approaches $\frac{1}{m}$, the value for random projection vectors, which is the origin of the name. Following Eq. (10), the loss $\mathcal{L}_{\text{rep}}$ averaged over $\boldsymbol{x}$ ensemble can be approximated by

$$\mathcal{L}_{\text{rep}} = \min_{s,b} \mathbb{E}_{\boldsymbol{x}} \mathcal{L}_{\text{rep}}|_{\boldsymbol{x}} \approx \min_{s,b} \frac{1}{r} \mathbb{E}_{E_r} \left\{ E_r \mathbb{E}_{\eta,v} \left[[s^2 v + s^2 \eta + b]_+ - v \right]^2 + (r - E_r) \mathbb{E}_{\eta} [s^2 \eta + b]_+^2 \right\}, \tag{11}$$

where E_r follows a binomial distribution: $p(E_r) = \binom{r}{E_r} p^{E_r} (1-p)^{r-E_r}$. Plugging Eq. (7) and (11) into Eq. (6), we obtain the loss $\mathcal{L}(\bar{E}, m, n; r)$ with r still tunable. The predicted loss is obtained by further optimization with respect to $r \in [m, n]$:

$$\mathcal{L}(\bar{E}, m, n) = \min_r \mathcal{L}(\bar{E}, m, n; r), \quad r_{\text{opt}}(\bar{E}, m, n) \equiv \arg\min_r \mathcal{L}(\bar{E}, m, n; r). \tag{12}$$

The partial RP approximation captures the transition from partial to full representation. The optimal r_{opt} grows with m, while it is always greater than m (vectors are in superposition, see appendix A.1.1). The transition arises because r moves from the interior of (m, n) to n (Fig. 3bc). The loss

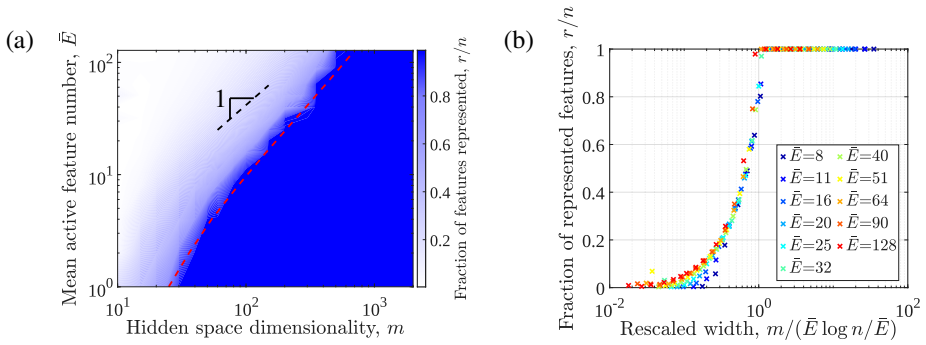


Figure 4: The critical width m^* increases with the active feature number $\bar{E}$, growing asymptotically proportional to it up to a logarithmic correction. (a) Phase diagram of r/n as a function of $\bar{E}$ and m. The red dashed curve is the phase boundary predicted by partial RP approximation. (b) By rescaling m according to Eq. (15), the critical points collapse. $n = 10000$.

trend and the bending around m^* are also predicted (Fig. 3d). At finite $\bar{E}$ and n, the predicted loss in the full-representation phase is slightly higher than the experimental loss due to the finite size effect, which disappears in the limit $1 \ll \bar{E} \ll n$ (appendix C.1.4).

3.2 Sparsity sets the critical width and the loss

We now study how sparsity determines the representation behavior and the loss, using both the converged trained model and the partial RP approximation. The sparsity is measured by the number of active features E in each input (smaller E means sparser input). Here we only focus on its mean value $\bar{E}$, and the discussion of its distribution can be found in appendix A.4.3. To isolate the effect of $\bar{E}$, we keep the uniform setup in Sec. 3.1. The effect of sparsity is summarized as follows (Fig. 4 and 5):

Key result 2: The critical width m^* increases with the mean active feature number $\bar{E}$. In the limit $1 \ll \bar{E} \ll n$, it grows asymptotically as $m^* \sim \bar{E} \log \frac{n}{\bar{E}}$. The loss $\mathcal{L}$ grows with $\bar{E}$. When $m \ll m^*$, the loss $\mathcal{L} \approx \frac{\bar{E}}{n}(\bar{v}^2 + \sigma_v^2)$; when $m \gg m^*$, the loss grows asymptotically as $\mathcal{L} \approx \frac{2\sigma_v^2 \bar{E}^2}{mn} \log \frac{n}{\bar{E}}$.

We use the partial RP approximation to understand the result. Increasing $\bar{E}$ raises $\mathcal{L}$ through both represented and unrepresented features. The loss of unrepresented features (7) is proportional to $\bar{E}$ in the limit $\bar{E} \ll n$. For represented features, the distribution of E_r shifts upwards as $\bar{E}$ increases, hence the crosstalk noise η increases, so $\mathcal{L}_{\text{rep}}$ increases. We can derive the asymptotic behavior of $\mathcal{L}_{\text{rep}}$ in the limit $1 \ll rp \ll m \ll r$ (appendix A.4.1):

$$\mathcal{L}_{\text{rep}} \approx \frac{2\sigma_v^2 p^2 r}{m} \log \frac{1}{p} = \frac{2\sigma_v^2 \bar{E}^2 r}{n^2 m} \log \frac{n}{\bar{E}}. \tag{13}$$

In full representations (i.e., $r = n$), Eq. (13) gives the total loss:

$$\mathcal{L} = \mathcal{L}_{\text{rep}} \approx \frac{2\sigma_v^2 \bar{E}^2}{nm} \log \frac{n}{\bar{E}}. \tag{14}$$

The critical width m^* can now be estimated by comparing the two options of a feature. If it is not represented, it costs $\mathcal{L}_{\text{unrep}}$ in Eq. (7), which is independent of m; if it is represented, it costs $\mathcal{L}_{\text{rep}}$ in Eq. (14), which grows with $\bar{E}$ faster than L_{unrep}, but decays with m. At small m the ratio $\mathcal{L}_{\text{rep}}/\mathcal{L}_{\text{unrep}}$ is large and features are unrepresented. The phase transition happens when $\mathcal{L}_{\text{rep}}/\mathcal{L}_{\text{unrep}}$ becomes of order 1, i.e., at

$$m^* \sim \bar{E} \log \frac{n}{\bar{E}}, \tag{15}$$

with an $O(1)$ prefactor depending on σ_v (appendix A.4.1). The prediction of m^* is consistent with experimental result (Fig. 4b). This scaling coincides with the classical compressed-sensing bound

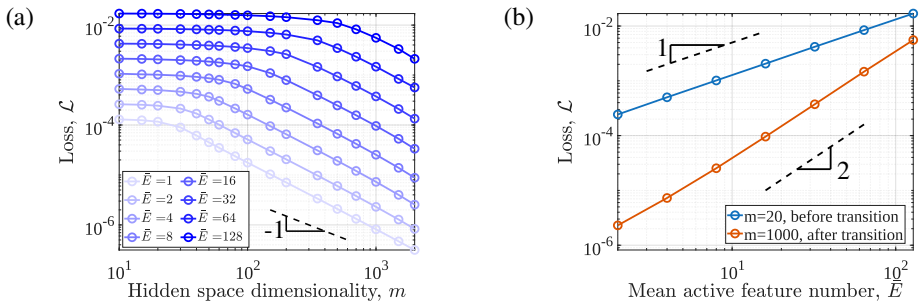


Figure 5: The loss grows with number of active features, approximately proportionally in partial representations and quadratically up to a logarithmic correction in full representations. (a) As $\bar{E}$ increases, the $\mathcal{L}$-m curve shifts upwards. (b) The loss $\mathcal{L}$ as a function of $\bar{E}$ under fixed n and two different m (before and after transition). $n = 10000$.

$m = O(k \log \frac{n}{k})$ for recovering k-sparse signals from m random linear measurements (Donoho, 2006), which Elhage et al. (2022) and Klindt et al. (2026) invoke for superposition. Compressed sensing offers a second reading of Eq. (15): the width at which a fixed dictionary ceases to be decodable by a single thresholding step is also the width below which an optimizing model no longer finds full representation worthwhile (Ivanitskiy et al., 2026).

At $m \ll m^*$, the fraction r/n is small, hence $\mathcal{L}$ is dominated by $\mathcal{L}_{\text{unrep}}$, growing linearly in $\bar{E}$; at $m^* \ll m \ll n$, the loss $\mathcal{L}$ follows Eq. (14), growing quadratically with $\bar{E}$ up to a logarithmic correction. These results match the prediction in Cowsik et al. (2025). The $1/m$ dependence of $\mathcal{L}$ in full-representation phase also matches the prediction in Liu et al. (2025). These asymptotic behaviors all match the experiments well (Fig. 5b).

3.3 Non-uniformity delays the transition and lowers the loss

We now study how the non-uniformity of features' firing probabilities influences the representation behavior and the loss. In this section, the firing probabilities p_i are no longer uniform. Here we use the conventional power-law firing probabilities $p_i \propto i^{-\epsilon_p}$ with the prefactor fixed by $\bar{E}$, hence features are indexed in decreasing order. The data exponent ϵ_p controls the degree of non-uniformity, as p_i becomes more uneven when ϵ_p grows from 0. We only focus on moderate ϵ_p in this section. Discussions on the applicable range of ϵ_p can be found in appendix C.3.

Under the setup of non-uniform independent features, the transition from bimodal s_i distribution (partial representation) to unimodal s_i distribution (full representation) remains valid for moderate ϵ_p. Biases also still behave in the same way as norms (appendix C.3). More specifically, in partial representations, features with higher p_i are more likely to be represented. Unlike the uniform case, vectors of represented features have non-uniform norms depending on p_i (Fig. 6a). Nevertheless, we still use r/n as the order parameter, since its singular behavior around the critical point persists. We have the following result (Fig. 6 and 7):

Key result 3: As the data exponent ϵ_p increases, the critical width m^* increases, and the loss $\mathcal{L}$ decreases.

We adapt the partial RP approximation based on the observations. The represented group is idealized as containing only features with largest p_i. When m increases, features are represented in turn from the first one (largest p_i) to the last one (smallest p_i). Eq. (7) is then replaced by:

$$\mathcal{L}_{\text{unrep}} = \frac{1}{n-r} \sum_{i=r+1}^{n} (\bar{v}^2 + \sigma_v^2) p_i - \bar{v}^2 p_i^2. \tag{16}$$

We still apply the random projection approximation to represented features and formally express $\mathcal{L}_{\text{rep}}$ by Eq. (11), but here E_r follows a Poisson binomial distribution generated by $p_1, p_2, ..., p_r$. The adapted partial RP approximation captures the trend of m^* (Fig. 6b).

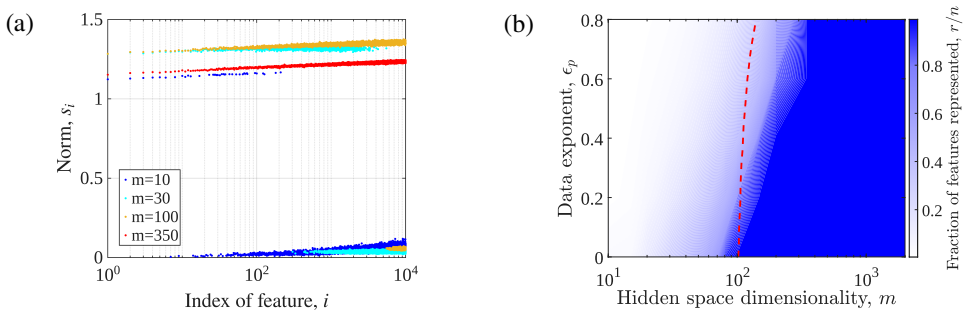


Figure 6: The critical width m^* increases with non-uniformity, i.e., the data exponent ϵ_p. (a) Distributions of s_i at different m. As m increases, two s_i bands change into one band. $\epsilon_p = 0.2$ (b) Phase diagram of r/n as a function of ϵ_p and m. The red dashed curve is the phase boundary predicted by partial RP approximation. $n = 10000, \bar{E} = 10$.

The partial RP approximation treats all represented features isotropically. As a result, in the limit $1 \ll \bar{E} \ll m \ll n$, it still gives the asymptotic expression (14) which only depends on mean firing probability, predicting no dependence on ϵ_p. This conflicts with the experimental result, as $\mathcal{L}$ decays with ϵ_p. To explain it, we need to consider the anisotropic organization of the represented features. For each feature, we define the leverage score for represented features (Scherlis et al., 2025)

$$h_i \equiv \frac{1}{s_i^2} \boldsymbol{w}_i^T (\tilde{\boldsymbol{W}} \tilde{\boldsymbol{W}}^T)^{-1} \boldsymbol{w}_i, \tag{17}$$

where $\tilde{\boldsymbol{W}} = (\boldsymbol{w}_1/s_1 \ ... \ \boldsymbol{w}_r/s_r)$ contains normalized vectors of represented features. We set $h_i = 0$ for unrepresented features by convention. Here h_i acts as the effective number of hidden dimensions allocated to feature i. However, it differs from the feature dimensionality used in Elhage et al. (2022), as h_i upper-bounds their dimensionality (Ivanov et al., 2026). In the limit $1 \ll m \ll r$, the squared overlaps of feature i with other features are controlled by h_i (appendix A.5):

$$\sum_{j:j\neq i} \frac{(\boldsymbol{w}_i \cdot \boldsymbol{w}_j)^2}{s_i^2 s_j^2} \approx \frac{1}{h_i}. \tag{18}$$

The variance of the crosstalk noise on feature i is also proportional to $\frac{1}{h_i}$ accordingly. To see how the representation adapts, we expand $\mathcal{L}_{\text{rep}}$ around the uniform solution $\mathcal{L}_{\text{rep}}^{(0)}$ to first order in $\delta p_i \equiv p_i - \bar{E}/n$, treating s_i, b_i and h_i as independent variables. For a feature with $\delta p_i > 0$, we find $\delta h_i > 0$, $\delta s_i < 0$ and $\delta b_i > 0$ (Fig. 7b). Therefore, a more frequent feature occupies more dimensions, receives less crosstalk noise, hence needs less filtering by its norm and bias. The leading order correction of $\mathcal{L}_{\text{rep}}$ is

$$\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{rep}}^{(0)} - c \sum_{i=1}^{r} \delta p_i^2 \tag{19}$$

with positive c. Therefore, the perturbation analysis captures the trends of s_i and b_i, as well as the downward shift of $\mathcal{L}$, though it is quantitative only for weak non-uniformity.

4 Related Works

The idea that networks represent more features than dimensions arose in studies of word embeddings (Goh, 2016; Arora et al., 2018). Olah et al. (2020) developed it as superposition hypothesis, which was demonstrated in the toy model of Elhage et al. (2022). In small models, the optimal representation vectors form polytopes (Elhage et al., 2022; Chen et al., 2023). Per-feature capacity has been quantified by the feature dimensionality of Elhage et al. (2022), refined into a constrained-allocation framework (Scherlis et al., 2025), and developed into a spectral theory (Ivanov et al., 2026). In the large-system limit, Cowsik et al. (2025) derived the loss by assuming a permutation-symmetric solution.

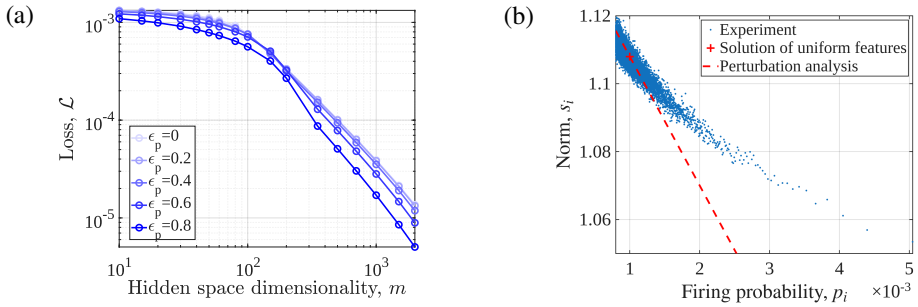


Figure 7: The loss decays with the data exponent ϵ_p, since the representation adapts to the non-uniformity. (a) As ϵ_p increases, the $\mathcal{L}$-m curve shifts downwards. (b) The norm decreases with p_i in the direction predicted. The red cross point is the solution of uniform features given the same $\bar{E}, m$ and n. $\epsilon_p = 0.2, m = 1000$. For (a) and (b), $n = 10000, \bar{E} = 10$.

Loss decreases as a power law in model size across domains (Hestness et al., 2017; Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022).The connection between neural scaling law and superposition was pointed out in Liu et al. (2025), which gave the inverse width scaling supported in LLM experiments (Liu & Gore, 2026; Liu et al., 2026). Our work generalizes their theory.

Sparse coding represents data as sparse combinations of atoms from an overcomplete dictionary (Olshausen & Field, 1996); dictionary learning recovers a ground-truth dictionary from such data (Spielman et al., 2012; Agarwal et al., 2014; Arora et al., 2015), and sparse autoencoders apply it to network activations (Bricken et al., 2023; Templeton et al., 2024), inverting the map that the toy model describes. When the dictionary is given, recovering the code is compressed sensing (Candes et al., 2006; Donoho, 2006), where ℓ_1 recovery fails sharply below $\sim k \log(n/k)$ measurements for k-sparse signals (Donoho & Tanner, 2005; 2008; 2009; Amelunxen et al., 2014). This bound has been used to estimate the capacity of superposition (Elhage et al., 2022; Klindt et al., 2026) and to posit a critical width (Sarkar & Deka, 2026).

5 Discussion

In this paper we focus on the transition from partial to full representation. It has a simple origin: the competition between representing and discarding a feature. Representing a feature results in crosstalk that grows with active feature number $\bar{E}$ and decays with model width m, while discarding it results in an error depending on its variance, independent of m. Their balance sets the critical width $m^* \sim \bar{E} \log(n/\bar{E})$, with the loss dominated by the discarding error below m^* and by crosstalk above it. Under non-uniformity, rare features are represented last, which delays the transition, while the loss is lowered since more frequent features occupy more hidden dimensions.

There are several limitations in this paper. Our work bases fully on the linear representation hypothesis. We focus on overall loss behavior rather than the loss from a finite dataset. We use Gaussian distributions to approximate the crosstalk noise, which discards the fine structure of representation vectors and overestimates the tail that leaks through the ReLU. We treat correlations only in exchangeable form (appendix C.2) and non-uniformity only at moderate strength. Finally, whether our results extend to other architectures and losses remains to be verified.

Parts of our results are consistent with observations in LLMs. Liu et al. (2025) found inverse width scaling in open-source LLMs in strong superposition. Sarkar & Deka (2026) found that the loss of narrow students distilled from an LLM saturates at a floor due to discarded features when the width falls below a critical value scaling as $\bar{E} \log(n/\bar{E})$. Sparse autoencoders would allow more of our predictions to be tested in LLMs, but this is beyond the scope of this paper. Conversely, since m^* and the loss are set by the data statistics, our results open the possibility of measuring the number of features and other statistics of natural language indirectly. Conceptually, since our results rely on generic ingredients: sparse features, a width bottleneck, and nonlinearity, we expect similar transitions in other neural networks. In conclusion, our results provide a step toward understanding the mechanism underlying LLMs.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the development and checking of mathematical derivations. The tools also provided suggestions that informed parts of the theoretical analysis. The authors independently re-derived and verified the resulting mathematical claims.

We also used generative AI tools to generate and modify code for numerical experiments and visualizations, proofread and rewrite manuscript text for readability, critique the manuscript, assist with minor brainstorming, and search for relevant literature. The visualization code generated plots from author-provided data and did not generate the underlying data or results.

All AI-assisted material retained in the paper was reviewed by the authors. The code was checked through inspection, tests, and experiments, and all cited sources identified with AI assistance were independently verified. We take responsibility for the final content of this work, including all text, mathematical claims, analyses, code, figures, and other artifacts produced with the assistance of generative AI.

REFERENCES

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning Sparsely Used Overcomplete Dictionaries. In *Proceedings of The 27th Conference on Learning Theory*, pp. 123–137. PMLR, May 2014. URL https://proceedings.mlr.press/v35/agarwal14a.html.

Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, September 2014. ISSN 2049-8772. doi: 10.1093/imaiai/iau005. URL https://ieeexplore.ieee.org/document/8142448.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, Efficient, and Neural Algorithms for Sparse Coding. In *Proceedings of The 28th Conference on Learning Theory*, pp. 113–149. PMLR, June 2015. URL https://proceedings.mlr.press/v40/Arora15.html.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018. doi: 10.1162/tacl_a_00034. URL https://aclanthology.org/Q18-1034/.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, January 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90014-2. URL https://www.sciencedirect.com/science/article/pii/0893608089900142.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning, October 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52 (2):489–509, February 2006. ISSN 1557-9654. doi: 10.1109/TIT.2005.862083. URL https://ieeexplore.ieee.org/document/1580791.

Zhongtian Chen, Edmund Lau, Jake Mendel, Susan Wei, and Daniel Murfet. Dynamical versus Bayesian Phase Transitions in a Toy Model of Superposition, October 2023. URL http://arxiv.org/abs/2310.06301. arXiv:2310.06301 [cs.LG].

Aditya Cowsik, Kfir Dolev, and Alex Infanger. The persian rug: solving toy models of superposition using large-scale symmetries, 2025. URL https://openreview.net/forum?id=rapXZIfwbX.

David Donoho and Jared Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, July 2008. ISSN 1088-6834, 0894-0347. doi: 10.1090/S0894-0347-08-00600-0. URL https://www.ams.org/jams/2009-22-01/S0894-0347-08-00600-0/.

David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367 (1906):4273–4293, November 2009. ISSN 1364-503X. doi: 10.1098/rsta.2009.0152. URL https://doi.org/10.1098/rsta.2009.0152.

David L. Donoho and Jared Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences*, 102(27):9452–9457, July 2005. doi: 10.1073/pnas.0502258102. URL https://www.pnas.org/doi/10.1073/pnas.0502258102.

D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.871582. URL http://ieeexplore.ieee.org/document/1614066/.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition, September 2022. URL http://arxiv.org/abs/2209.10652. arXiv:2209.10652 [cs].

Gabriel Goh. Decoding the Thought Vector, 2016. URL https://gabgoh.github.io/ThoughtVectors/.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling Laws for Autoregressive Generative Modeling, November 2020. URL http://arxiv.org/abs/2010.14701. arXiv:2010.14701 [cs.LG].

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically, December 2017. URL http://arxiv.org/abs/1712.00409. arXiv:1712.00409 [cs.LG].

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022. URL http://arxiv.org/abs/2203.15556. arXiv:2203.15556 [cs.CL].

Michael I. Ivanitskiy, John Jasper, Emily J. King, and Dustin G. Mixon. Towards a mathematical theory of superposition, August 2026. URL http://arxiv.org/abs/2608.27540. arXiv:2608.27540 [stat.ML].

Georgi Ivanov, Narmeen Oozeer, Shivam Raval, Tasana Pejovic, Shriyash Upadhyay, and Amir Abdullah. Spectral Superposition: A Theory of Feature Geometry, February 2026. URL http://arxiv.org/abs/2602.02224. arXiv:2602.02224 [cs.LG].

Norman Lloyd Johnson and Samuel Kotz. *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. Wiley, 1977. ISBN 978-0-471-44630-9. Google-Books-ID: ZBfvAAAAMAAJ.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs.LG].

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs.LG].

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. A unifying framework from neural superposition to sparse interpretable codes. *Nature Machine Intelligence*, 8(7):1025–1037, July 2026. ISSN 2522-5839. doi: 10.1038/s42256-026-01259-z. URL https://www.nature.com/articles/s42256-026-01259-z.

Yizhou Liu and Jeff Gore. Neural Scaling Universality: If Exponents Are Fixed, Time to Understand Coefficients, June 2026. URL http://arxiv.org/abs/2606.25008. arXiv:2606.25008 [cs.LG].

Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition Yields Robust Neural Scaling. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 159269–159305. Curran Associates, Inc., 2025. doi: 10.52202/085713-5320. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e97ac22927560eb2de6b658498cbc575-Abstract-Conference.html.

Yizhou Liu, Sara Kangaslahti, Ziming Liu, and Jeff Gore. Inverse depth scaling from most layers being similar. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=CEpCpxJqAt.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):10.23915/distill.00024.001, March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996. ISSN 1476-4687. doi: 10.1038/381607a0. URL https://www.nature.com/articles/381607a0.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. October 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Nilesh Sarkar and Dawar Jyoti Deka. Geometric Limits of Knowledge Distillation: A Minimum-Width Theorem via Superposition Theory, April 2026. URL http://arxiv.org/abs/2604.04037. arXiv:2604.04037 [cs.LG].

Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and Capacity in Neural Networks, March 2025. URL http://arxiv.org/abs/2210.01892. arXiv:2210.01892 [cs.NE].

Daniel A. Spielman, Huan Wang, and John Wright. Exact Recovery of Sparsely-Used Dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, pp. 37.1–37.18. JMLR Workshop and Conference Proceedings, June 2012. URL https://proceedings.mlr.press/v23/spielman12.html.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, May 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

L. Welch. Lower bounds on the maximum cross correlation of signals (Corresp.). *IEEE Transactions on Information Theory*, 20(3):397–399, May 1974. ISSN 1557-9654. doi: 10.1109/TIT.1974.1055219. URL https://ieeexplore.ieee.org/document/1055219.

A THEORETICAL ANALYSIS

A.1 NON-SUPERPOSITION LOSS

In this section, we consider the non-superposition configuration. When the model exhibits no superposition, only m features are represented (Fig. 1a). Assume the first m features are represented,

$$\boldsymbol{W} = (\boldsymbol{I}_m \ \boldsymbol{0}_{m\times(n-m)}). \tag{20}$$

Correspondingly, the first m biases $b_1 = b_2 = ... = b_m = 0$, so that the first m features contribute no loss. To minimize the loss, the remaining biases generally do not vanish. We denote them by

$$\tilde{\boldsymbol{b}} = (b_{m+1}, b_{m+2}, ..., b_n)^T. \tag{21}$$

Under this configuration, the output can be analytically given,

$$\begin{aligned}
\boldsymbol{y} &= \left[\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}\right]_+ && (22)\\
&= \left[\begin{pmatrix} \boldsymbol{I}_m & 0_{m\times(n-m)} \\ 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)} \end{pmatrix}\boldsymbol{x} + \begin{pmatrix}\boldsymbol{0}\\ \tilde{\boldsymbol{b}}\end{pmatrix}\right]_+ && (23)\\
&= (x_1\ x_2\ ...\ x_m\ \ [b_{m+1}]_+\ \ [b_{m+2}]_+\ \ ...\ \ [b_n]_+)^T. && (24)
\end{aligned}$$

Therefore, the loss is

$$\begin{aligned}
\mathcal{L}_{\text{NS}} =& \frac{1}{n}\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{y}-\boldsymbol{x}\|^2 && (25)\\
=& \frac{1}{n}\int \mathcal{D}v_{m+1}\mathcal{D}v_{m+2}...\mathcal{D}v_n \sum_{\boldsymbol{u}} p(\boldsymbol{u}) \sum_{i=m+1}^{n} \left([b_i]_+ - u_i v_i\right)^2, && (26)
\end{aligned}$$

where $\mathcal{D}x$ is the shorthand of $\mathrm{d}x\ p(x)$. By defining the marginal firing probabilities of features

$$p_i = \sum_{\boldsymbol{u}} p(\boldsymbol{u}) u_i, \tag{27}$$

Eq. (26) can be simplified as

$$\begin{aligned}
\mathcal{L}_{\text{NS}} =& \frac{1}{n}\sum_{i=m+1}^{n}\int \mathcal{D}v_i \left(p_i([b_i]_+ - v_i)^2 + (1-p_i)[b_i]_+^2\right) && (28)\\
=& \frac{1}{n}\sum_{i=m+1}^{n}\left([b_i]_+^2 - 2p_i[b_i]_+\bar{v} + p_i(\bar{v}^2+\sigma_v^2)\right). && (29)
\end{aligned}$$

The minimum is

$$\mathcal{L}_{\text{NS}} = \frac{1}{n}\sum_{i=m+1}^{n}\left(-p_i^2\bar{v}^2 + p_i(\bar{v}^2+\sigma_v^2)\right). \tag{30}$$

The corresponding optimal bias is

$$b_i = p_i\bar{v}, \quad \text{for } m+1 \le i. \tag{31}$$

A.1.1 SUBOPTIMALITY OF THE NON-SUPERPOSITION CONFIGURATION

Given its loss, it is natural to ask if the non-superposition configuration is the optimal configuration. To answer this question, we construct the following superposition configuration, which outperforms the non-superposition configuration: among n features, $m+1$ of them are represented. Representation vectors of feature 1 to feature m form an orthonormal basis, and that of feature $m+1$ forms

an antipode with feature m (Elhage et al., 2022). In this configuration, the gain of representing one more feature is

$$-p_{m+1}^2\bar{v}^2 + p_{m+1}(\bar{v}^2 + \sigma_v^2), \tag{32}$$

while the increasing loss is

$$p_{m,m+1}\mathbb{E}_{v_m,v_{m+1}}\left(([v_m - v_{m+1}]_+ - v_m)^2 + ([v_{m+1} - v_m]_+ - v_{m+1})^2\right), \tag{33}$$

where $p_{m,m+1}$ is the probability that both feature m and $m+1$ fire.

For the case we consider in the main text, the features are independent with small firing probability $\bar{E}/n \ll 1$, the gain is of order $\bar{E}/n$ while the increasing loss is of order $\bar{E}^2/n^2$. Therefore, the non-superposition configuration is always suboptimal.

A.2 Equal-norm frame and random projection approximation for uniform features

In this section, we use the random projection approximation to estimate the loss for uniform features. By "uniform", we mean all features are exchangeable. They do not need to be independent. Based on the symmetry of features, we consider the representation configuration called equal-norm frame, where all $\boldsymbol{w}_i$ in the weight matrix $\boldsymbol{W}$ have equal norm s. This acts as an ideal full representation, as the participation ratio r defined in Eq. (5) is precisely n. Furthermore, we let $b_i = b$ to be equal for all features as well.

Firstly, we assume that there are precisely E active features in each input, instead of a random number. In this case, the equal-norm loss is

$$\mathcal{L}_{\mathrm{EN}}(E,m,n;s,b) = \frac{1}{n}\mathbb{E}_{\boldsymbol{x}}\|[\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}]_+ - \boldsymbol{x}\|^2 \tag{34}$$

$$= \frac{1}{n\binom{n}{E}} \sum_{a_1,\dots,a_E} \int \mathcal{D}v_{a_1}\mathcal{D}v_{a_2}\dots\mathcal{D}v_{a_E} \sum_{i=1}^{n} \left\{ \left[\sum_{j=1}^{E} (\boldsymbol{W}^T\boldsymbol{W})_{ia_j} v_{a_j} + b_i \right]_+ - x_i \right\}^2 \tag{35}$$

$$= \frac{1}{n\binom{n}{E}} \sum_{a_1,\dots,a_E} \int \mathcal{D}v_{a_1}\dots\mathcal{D}v_{a_E} \left\{ \sum_{i=1}^{E} \left[\left[s^2 v_{a_i} + \sum_{j:j\neq i} s^2 c_{a_i a_j} v_{a_j} + b \right]_+ - v_{a_i} \right]^2 + \sum_{i\notin\{a_k\}} \left[\sum_{j=1}^{E} s^2 c_{i a_j} v_{a_j} + b \right]_+^2 \right\}. \tag{36}$$

The summation $\sum_{a_1,\dots,a_E}$ runs over all combinations of E different active features among n, and

$$c_{ij} \equiv \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_j}{s^2} \tag{37}$$

are shorthands of cosine similarities of representation vectors. $\mathcal{D}v$ indicates integration is weighted by the probability density. The total loss is divided into two parts, one of which is the contribution of all active features, and the other is that of all inactive features. We denote them by $\mathcal{L}_A$ and $\mathcal{L}_I$ respectively. The crosstalk noises in each part are $\sum_{j:j\neq i} s^2 c_{a_i a_j} v_{a_j}$ and $\sum_{j=1}^{E} s^2 c_{i a_j} v_{a_j}$ respectively.

We take the limit $1 \ll E$. By using mean field approximation, crosstalk noises can be regarded as independent random variables drawn from one distribution:

$$\sum_{j:j\neq i} s^2 c_{a_i a_j} v_{a_j} \approx s^2\eta, \quad \sum_{j=1}^{E} s^2 c_{i a_j} v_{a_j} \approx s^2\eta, \tag{38}$$

$$\eta = \sum_{i=1}^{E} c_{(i)} v_{(i)}. \tag{39}$$

Under the mean field approximation, Eq. (36) becomes

$$\mathcal{L}_{\mathrm{EN}}(E,m,n;s,b) \approx \frac{1}{n\binom{n}{E}} \sum_{a_1,\ldots,a_E} \int \mathcal{D}v_{a_1}\ldots\mathcal{D}v_{a_E}\mathcal{D}\eta \Bigg\{ \sum_{i=1}^{E} \left[\left[s^2 v_{a_i} + s^2\eta + b \right]_+ - v_{a_i} \right]^2 + \sum_{i\notin\{a_k\}} \left[s^2\eta + b \right]_+^2 \Bigg\} \tag{40}$$

$$= \frac{1}{n} \left\{ E \int \mathcal{D}v\mathcal{D}\eta \left[\left[s^2 v + s^2\eta + b \right]_+ - v \right]^2 + (n-E) \int \mathcal{D}\eta \left[s^2\eta + b \right]_+^2 \right\} \tag{41}$$

$$= \frac{E}{n}\mathcal{L}_A + \frac{n-E}{n}\mathcal{L}_I. \tag{42}$$

In order to calculate the loss, we need to know the distribution of $c_{ij}|_{i\neq j}$, which is analytically intractable for the trained frame. To proceed, we use Gaussian distribution to approximate $\mathcal{D}_\eta$. The mean value of η cannot contribute to the loss, since we can always replace b by $b - s^2\langle\eta\rangle$ to cancel its effect. Therefore, we simply let $\langle\eta\rangle = 0$. (This is also approximately true in the experiment. The typical experimental result is $\langle c_{ij}|_{i\neq j}\rangle$ reaches the lower bound $-\frac{1}{n-1}$, which is far less than $\frac{1}{m}$ in the limit $m \ll n$. Consequently, $\langle\eta\rangle$ is also very small compared with the standard deviation σ_η. See Sec. C.1.3.) And

$$\sigma_\eta^2 = E\sigma_c^2(\bar{v}^2 + \sigma_v^2) \tag{43}$$

$$\geq E(\bar{v}^2 + \sigma_v^2)\frac{n-m}{m(n-1)}, \tag{44}$$

where σ_c^2 is the variance of $c_{ij}|_{i\neq j}$, which has a lower bound

$$\sigma_c^2 \geq \frac{n-m}{m(n-1)} \tag{45}$$

when $\langle c_{ij}|_{i\neq j}\rangle$ is ignored. This lower bound is achieved when representation vectors form a tight frame. If the distribution of η is confined to be Gaussian, the minimal loss is achieved when the variance is minimal, so we let σ_η^2 be its lower bound. Therefore, we have approximation (10):

$$\eta \sim \mathcal{N}\left(0, E(\bar{v}^2 + \sigma_v^2)\frac{n-m}{m(n-1)}\right). \tag{46}$$

In other words, we assume that the trained representation approximately achieves a tight frame. With Eq. (46), $\mathcal{L}_I$ can be further written as

$$\mathcal{L}_I = F\left(-\frac{b}{\sqrt{2s^4\sigma_\eta^2}}\right) s^4\sigma_\eta^2, \tag{47}$$

where

$$F(\beta) \equiv \frac{2\beta^2+1}{2}(1-\mathrm{erf}\beta) - \frac{\beta}{\sqrt{\pi}}\exp(-\beta^2). \tag{48}$$

Eq. (41) and Eq. (46) give a complete estimation of the loss of the equal-norm frame.

In the limit $m \ll n$, the Gaussian approximation can also be interpreted as using random projection frame to approximate the trained optimal equal-norm frame. This is because in this limit,

$$\sigma_\eta^2 = E(\bar{v}^2 + \sigma_v^2)\frac{n-m}{m(n-1)} \approx \frac{E}{m}(\bar{v}^2 + \sigma_v^2). \tag{49}$$

The Gaussian distribution with variance (49) is the distribution of $c_{ij}|_{i\neq j}$ in the random projection frame (this is also verified experimentally.) So we call mean field + Gaussian approximation **random projection (RP) approximation**.

In most cases, the number of active features in an input is not fixed. For example, in the setup of uniform independent features used in Sec. 3.1 and 3.2, the probability that E active features occur is

$$p_E = \binom{n}{E} \left(\bar{E}/n\right)^E \left(1 - \bar{E}/n\right)^{n-E}. \tag{50}$$

With variable E, the random projection approximation loss is simply the weighted average of losses:

$$\mathcal{L}_{\mathrm{RP}}(\{p_E\}, m, n) = \min_{s,b} \sum_{E=0}^{\infty} p_E \left\{ \frac{n-E}{n} F\left(-\frac{b}{\sqrt{2s^4\sigma_\eta^2(E)}} \right) s^4 \sigma_\eta^2(E) + \frac{E}{n} \mathbb{E}_{v,\eta(E)} \left[[s^2 v + s^2 \eta + b]_+ - v \right]^2 \right\}. \tag{51}$$

Eq. (51) is used numerically to estimate equal-norm frame loss without training. Since $\{p_E\}$ is determined by $\bar{E}$ for uniform independent features, we write $\bar{E}$ in that case.

A.3 Estimating partial representation loss with partial RP approximation

In this section, we still consider uniform features. Based on the experimental result in Sec. 3.1, the partial representation is modeled as a partial equal-norm frame: r features are represented by an equal-norm frame, and other $n - r$ features are just not represented, with $s_i = 0$. Under this approximation, configurations with $r < m$ always have higher loss than that with $r = m$, hence we confine $r \in [m, n]$. The loss can be written as Eq. (6):

$$\mathcal{L}(\{p_E\}, m, n; r) = \frac{r}{n}\mathcal{L}_{\mathrm{rep}} + \frac{n-r}{n}\mathcal{L}_{\mathrm{unrep}}, \tag{52}$$

where $\mathcal{L}_{\mathrm{rep}}$ is the loss contributed by represented features, and $\mathcal{L}_{\mathrm{unrep}}$ is that from features not represented. As in Sec. A.2, we use RP approximation (51) to estimate $\mathcal{L}_{\mathrm{rep}}$:

$$\mathcal{L}_{\mathrm{rep}} = \mathcal{L}_{\mathrm{RP}}\left(\{p_{E_r}\}, m, r\right), \tag{53}$$

where p_{E_r} is the probability of E_r active features occurs among r represented features. Eq. (53) is just Eq. (11) in the main text. For uniform independent features,

$$p_{E_r} = \binom{r}{E_r} \left(\bar{E}/n\right)^{E_r} \left(1 - \bar{E}/n\right)^{r-E_r}; \tag{54}$$

for general uniform features,

$$p_{E_r} = \sum_{E=E_r}^{n} p_E \binom{r}{E_r} \binom{n-r}{E-E_r} \Big/ \binom{n}{E}, \tag{55}$$

where E is the total number of active features among all n features.

$\mathcal{L}_{\mathrm{unrep}}$ can be written out as in Sec. A.1. Still, we assume last $(n - r)$ features are unrepresented:

$$\mathcal{L}_{\mathrm{unrep}} = \frac{1}{n-r}\mathbb{E}_{\boldsymbol{x}} \sum_{i=r+1}^{n} (y_i - x_i)^2 \tag{56}$$

$$= \frac{1}{n-r} \int \mathcal{D}v_{r+1} ... \mathcal{D}v_n \sum_{\boldsymbol{u}} p(\boldsymbol{u}) \sum_{i=r+1}^{n} \left([b_i]_+ - u_i v_i\right)^2 \tag{57}$$

$$= \frac{1}{n-r} \sum_{i=r+1}^{n} \int \mathcal{D}v_i \left(p_i([b_i]_+ - v_i)^2 + (1-p_i)[b_i]_+^2 \right) \tag{58}$$

$$= \frac{1}{n-r} \sum_{i=r+1}^{n} \left([b_i]_+^2 - 2p_i[b_i]_+\bar{v} + p_i(\bar{v}^2 + \sigma_v^2) \right) \tag{59}$$

$$\geq \frac{1}{n-r} \sum_{i=r+1}^{n} \left(-p_i^2\bar{v}^2 + p_i(\bar{v}^2 + \sigma_v^2) \right), \tag{60}$$

where the marginal firing probabilities p_i are defined in Eq. (27). The minimum is achieved with $b_i = p_i\bar{v} = \bar{E}\bar{v}/n$ for $r+1 \le i$. In trained models, the biases of unrepresented features typically settle at negative values rather than at the optimum $p_i\bar{v}$, which is because for $b_i < 0$ the output is zero and the gradient on b_i vanishes, while the resulting excess loss $p_i^2\bar{v}^2$ is negligible.

Plugging Eq. (53) and (60) into Eq. (52), we obtain the **partial RP approximation**

$$\mathcal{L}_{\mathrm{pRP}}(\{p_E\}, m, n; r) = \frac{r}{n}\mathcal{L}_{\mathrm{RP}}\left(\{p_{E_r}\}, m, r\right) + \frac{n-r}{n}\left(-\frac{\bar{E}^2}{n^2}\bar{v}^2 + \frac{\bar{E}}{n}(\bar{v}^2+\sigma_v^2)\right). \tag{61}$$

Since in Eq. (61), r is not given but obtained in optimization, it can be used to calculate the transition point from partial representation to full representation. Following Eq. (12),

$$\mathcal{L}_{\mathrm{pRP}}(\{p_E\}, m, n) = \min_r \mathcal{L}_{\mathrm{pRP}}(\{p_E\}, m, n; r), \tag{62}$$

$$r_{\mathrm{opt}}(\{p_E\}, m, n) = \arg\min_r \mathcal{L}(\{p_E\}, m, n; r). \tag{63}$$

The transition point m^* is defined as

$$m^* \equiv \inf_m \{m : r_{\mathrm{opt}}(\{p_E\}, m, n) = n\}. \tag{64}$$

A.4 Asymptotic behaviors of partial RP approximation

In this section, we analyse the asymptotic behaviors of partial RP approximation (62) in different limits.

A.4.1 Large width limit: $1 \ll \bar{E} \ll m < n$

In this limit, typical crosstalk noise $\sigma_\eta^2 \sim \frac{\bar{E}r}{mn}\frac{r-m}{r-1}(\bar{v}^2+\sigma_v^2) \ll 1$. As a result, the bias to cancel the crosstalk noise on inactive represented features can also be small, typically $b \sim -\sqrt{\sigma_\eta^2}$, and to recover active represented features, the norm s needs to be approximately 1. Therefore, active represented features are nearly not affected by ReLU, and in Eq. (51)

$$\mathbb{E}_{v,\eta(E)}\left[[s^2 v + s^2\eta + b]_+ - v\right]^2 \approx \mathbb{E}_{v,\eta(E)}\left(s^2 v + s^2\eta + b - v\right)^2 \tag{65}$$

$$= (s^2-1)^2\sigma_v^2 + [(s^2-1)\bar{v} + b]^2 + s^4\sigma_\eta^2. \tag{66}$$

Based on the observation, we use the following ansatz for optimal b and s:

$$b = -\beta\sigma_\eta(\bar{E}) + O(\sigma_\eta^2(\bar{E})), \tag{67}$$

$$s = 1 + \Delta s\sigma_\eta(\bar{E}) + O(\sigma_\eta^2(\bar{E})). \tag{68}$$

Under this ansatz, the represented loss becomes

$$\begin{aligned}
\mathcal{L}_{\mathrm{rep}} &= \mathcal{L}_{\mathrm{RP}}(\{p_{E_r}\}, m, r) \\
&\approx \sum_{E_r} p_{E_r}\left\{F\left(\frac{\beta\sigma_\eta(\bar{E})}{\sqrt{2}\sigma_\eta(E_r)}\right)\frac{\sigma_\eta^2(E_r)}{\sigma_\eta^2(\bar{E})} + \frac{E_r}{r}\left[4\Delta s^2\sigma_v^2 + (2\Delta s\bar{v}-\beta)^2 + \frac{\sigma_\eta^2(E_r)}{\sigma_\eta^2(\bar{E})}\right]\right\}\sigma_\eta^2(\bar{E}) \\
&= \sum_{E_r} p_{E_r}\left\{F\left(\frac{\beta}{\sqrt{2}}\sqrt{\bar{E}/E_r}\right)\frac{E_r}{\bar{E}} + \frac{E_r}{r}\left[4\Delta s^2\sigma_v^2 + (2\Delta s\bar{v}-\beta)^2 + \frac{E_r}{\bar{E}}\right]\right\}\sigma_\eta^2(\bar{E}) \\
&= \sum_{E_r} p_{E_r} F\left(\frac{\beta}{\sqrt{2}}\sqrt{\bar{E}/E_r}\right)\frac{E_r}{\bar{E}}\sigma_\eta^2(\bar{E}) + \frac{\bar{E}}{n}\left[4\Delta s^2\sigma_v^2 + (2\Delta s\bar{v}-\beta)^2 + \frac{r}{n} + \frac{r\sigma_{E_r}^2}{n\bar{E}_r^2}\right]\sigma_\eta^2(\bar{E})
\end{aligned} \tag{69}$$

to the leading order. Here we also use $E_r/r \sim \bar{E}/n \ll 1$, and rewrite $\bar{E}_r$ as $\bar{E}r/n$ based on Eq. (55). Eq. (69) is quadratic in Δs, so we can analytically optimize it:

$$\mathcal{L}_{\mathrm{rep}} = \sum_{E_r} p_{E_r} F\left(\frac{\beta}{\sqrt{2}}\sqrt{\bar{E}/E_r}\right)\frac{E_r}{\bar{E}}\sigma_\eta^2(\bar{E}) + \frac{\bar{E}}{n}\left[\frac{\beta^2\sigma_v^2}{\sigma_v^2+\bar{v}^2} + \frac{r}{n} + \frac{r\sigma_{E_r}^2}{n\bar{E}_r^2}\right]\sigma_\eta^2(\bar{E}), \tag{70}$$

with

$$\Delta s_{\rm opt} = \frac{\beta \bar{v}}{2(\sigma_v^2 + \bar{v}^2)}. \tag{71}$$

The parameter β can be optimized numerically. The auxiliary function F decays exponentially (Fig. 8), hence the first term in Eq. (70) is of the same order of the second term at optimal β, even though $\bar{E} \ll n$. Intuitively, the ReLU + bias nonlinearity can filter most of crosstalk noise on inactive represented features. Therefore, at optimum, $\mathcal{L}_{\rm rep} \sim \frac{\bar{E}}{n}\sigma_\eta^2(\bar{E})$, which is far less than $\mathcal{L}_{\rm unrep} \sim \frac{\bar{E}}{n}$ in Eq. (60). In this limit $1 \ll \bar{E} \ll m < n$, all features are represented at optimum, i.e., $r_{\rm opt} = n$. The total loss to the leading order of $\bar{E}/m$ is

$$\mathcal{L}_{\rm pRP}(\{p_E\}, m, n) = \min_\beta \left\{ \sum_E p_E F\left(\frac{\beta}{\sqrt{2}}\sqrt{\bar{E}/E}\right)\frac{E}{\bar{E}} + \frac{\bar{E}}{n}\left[\frac{\beta^2\sigma_v^2}{\sigma_v^2+\bar{v}^2} + 1 + \frac{\sigma_E^2}{\bar{E}^2}\right]\right\}\sigma_\eta^2(\bar{E}). \tag{72}$$

Eq. (72) shows that in this limit, the partial RP loss $\mathcal{L}_{\rm pRP}$ is proportional to $\sigma_\eta^2(\bar{E})$. If we further apply the limit $m \ll n$,

$$\mathcal{L}_{\rm pRP}(\{p_E\}, m, n) \propto \sigma_\eta^2(\bar{E}) \propto \frac{1}{m}. \tag{73}$$

Furthermore, we consider the limit $\sigma_E \ll \bar{E}$ for simplicity, which works for uniform independent features in the limit $1 \ll \bar{E}$, as $\sigma_E = \sqrt{\bar{E}(1-\bar{E}/n)} \ll \bar{E}$. Because of the highly peaked distribution of E, Eq. (72) can be approximated as

$$\mathcal{L}_{\rm pRP}(\{p_E\}, m, n) = \min_\beta \left\{ F\left(\frac{\beta}{\sqrt{2}}\right) + \frac{\bar{E}}{n}\left[\frac{\beta^2\sigma_v^2}{\sigma_v^2+\bar{v}^2} + 1\right]\right\}\sigma_\eta^2(\bar{E}). \tag{74}$$

The optimal β is achieved when

$$\frac{1}{\sqrt{2}}F'(\beta/\sqrt{2}) + \frac{\bar{E}}{n}\frac{2\sigma_v^2\beta}{\sigma_v^2+\bar{v}^2} = 0. \tag{75}$$

The optimal β grows with $n/\bar{E}$. In the large β limit, the following asymptotic equation holds true (Fig. 8):

$$F(x) = F_{\rm asym}(x) \equiv \frac{e^{-x^2}}{2\sqrt{\pi}x^3} \quad \text{as } x \to \infty. \tag{76}$$

The asymptotic solution of Eq. (75) is

$$\beta = \sqrt{2\log\frac{n}{\bar{E}}}. \tag{77}$$

The asymptotic expression of $\mathcal{L}_{\rm pRP}$ becomes

$$\mathcal{L}_{\rm pRP} = 2\sigma_v^2\frac{\bar{E}^2}{nm}\frac{n-m}{n-1}\log\frac{n}{\bar{E}}. \tag{78}$$

In the limit $1 \ll \bar{E} \ll m \ll n$, this expression degenerates to Eq. (14):

$$\mathcal{L}_{\rm pRP} = \frac{2\sigma_v^2\bar{E}^2}{nm}\log\frac{n}{\bar{E}}. \tag{79}$$

For partial representation configurations, the derivation of Eq. (78) holds true for represented features in the limit $1 \ll r\bar{E}/n \ll m < r$ and $\sigma_{E_r} \ll \bar{E}_r = r\bar{E}/n$, and the approximate $\mathcal{L}_{\rm rep}$ is

$$\mathcal{L}_{\rm rep} = 2\sigma_v^2\frac{\bar{E}^2 r}{n^2 m}\frac{r-m}{r-1}\log\frac{n}{\bar{E}}. \tag{80}$$

If we further assume $m \ll r$,

$$\mathcal{L}_{\rm rep} = \frac{2\sigma_v^2\bar{E}^2 r}{n^2 m}\log\frac{n}{\bar{E}}. \tag{81}$$

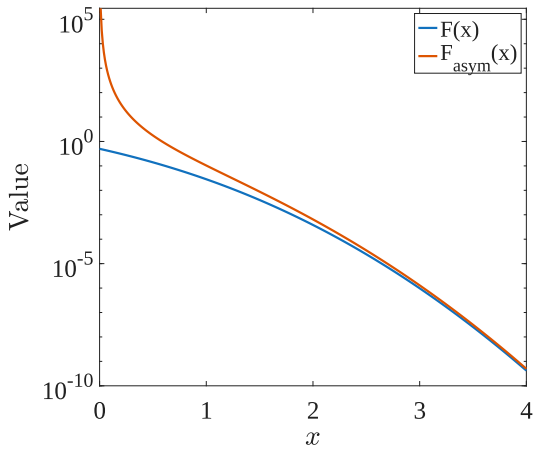


Figure 8: Auxiliary function $F(x)$ and $F_{\text{asym}}(x)$.

Based on Eq. (81), we can estimate the critical width m^*. By plugging Eq. (81) in Eq. (61), the partial RP loss can be written as

$$\mathcal{L}_{\text{pRP}}(\bar{E}, m, n; r) = \frac{2\sigma_v^2 r}{n} \frac{\bar{E}^2 r}{n^2 m} \log \frac{n}{\bar{E}} + \frac{n-r}{n} \frac{\bar{E}}{n} (\bar{v}^2 + \sigma_v^2). \tag{82}$$

According to Eq. (82), the optimal r is

$$r_{\text{opt}} = \min \left(\frac{mn(\bar{v}^2 + \sigma_v^2)}{4\bar{E}\sigma_v^2 \log \frac{n}{\bar{E}}}, n \right). \tag{83}$$

The optimal r_{opt} reaches n at m^*, hence

$$m^* = \frac{4\sigma_v^2 \bar{E}}{(\bar{v}^2 + \sigma_v^2)} \log \frac{n}{\bar{E}}. \tag{84}$$

Note that m^* is twice the width at which representing a feature and sacrificing it cost the same (the heuristic argument in Sec. 3.2). This is because increasing r from $n-1$ to n not only costs $\mathcal{L}_{\text{rep}}$, but also increases the loss of all other represented features, which results in a factor 2. Eq. (83) also predicts a constant $\frac{r_{\text{opt}}}{m}$ in partial representations, which requires $\log \frac{n}{\bar{E}} \gg 1$ and cannot be implemented in experiment. Also, Eq. (84) cannot capture the trend of m^* at different σ_v, see Sec. C.1.2.

A.4.2 Weak superposition limit: $r - m \ll m$

In Sec. A.1.1 we show the non-superposition configuration is suboptimal for uniform independent features with small firing probability. Here we show that the partial RP approximation gives the same result. We maintain the limit condition $1 \ll \bar{E} \ll n$ and $\sigma_E \ll \bar{E}$. For partial RP approximation we additionally require $\bar{E}_r = m\bar{E}/n \gg 1$, i.e., $m \gg n/\bar{E}$. We let

$$r = m + \Delta r, \quad \text{where } \Delta r \ll m. \tag{85}$$

The corresponding crosstalk noise variance is

$$\sigma_\eta^2(\bar{E}_r) = \frac{\bar{E}r}{n}(\bar{v}^2 + \sigma_v^2)\frac{r-m}{m(r-1)} \approx \frac{\bar{E}\Delta r}{nm}(\bar{v}^2 + \sigma_v^2) \ll 1. \tag{86}$$

Since typical crosstalk noise is small, derivation similar to Sec. A.4.1 can be applied here. Here we define

$$b = -\beta\sigma_\eta(\bar{E}_r) + \text{h.o.t.}, \tag{87}$$

$$s = 1 + \Delta s\sigma_\eta(\bar{E}_r) + \text{h.o.t.} \tag{88}$$

The represented loss is

$$\mathcal{L}_{\text{rep}}(r) \approx \min_{\beta,\Delta s} \left\{ F(\beta/\sqrt{2}) + \frac{\bar{E}}{n} \left[4\Delta s^2 \sigma_v^2 + (2\Delta s \bar{v} - \beta)^2 + 1 \right] \right\} \sigma_\eta^2(\bar{E}_r) \tag{89}$$

$$= \min_{\beta} \left\{ F(\beta/\sqrt{2}) + \frac{\bar{E}}{n} \left[\frac{\beta^2 \sigma_v^2}{\sigma_v^2 + \bar{v}^2} + 1 \right] \right\} \sigma_\eta^2(\bar{E}_r). \tag{90}$$

As discussed in Sec. A.4.1, the first factor in Eq. (90) is of order $\frac{\bar{E}}{n}$ at optimal β, hence the loss of representing extra Δr features is

$$r\mathcal{L}_{\text{rep}}(r) \sim \frac{\bar{E}^2 \Delta r}{n^2}. \tag{91}$$

On the other hand, the gain of representing extra Δr features is

$$-\Delta r \mathcal{L}_{\text{unrep}} \approx -\Delta r \frac{\bar{E}}{n}. \tag{92}$$

The gain is much greater than the loss, which indicates that the non-superposition configuration is never the optimal configuration.

A.4.3 Lowest-order correction from the variability of E

In Sec. A.4.1 and A.4.2, we neglect the effect of variable E. Here we derive the lowest-order correction on the partial RP loss of the fluctuation of E, in the limit $\sigma_{E_r} \ll \bar{E}_r$. Since $\mathcal{L}_{\text{unrep}}$ is only determined by the marginal firing probabilities $\bar{E}/n$, we only need to correct $\mathcal{L}_{\text{rep}}$. In the limit $1 \ll \bar{E}_r \ll m < r \leq n$, with

$$b \approx -\beta \sigma_\eta(\bar{E}_r), \tag{93}$$

$$s \approx 1 + \Delta s \sigma_\eta(\bar{E}_r). \tag{94}$$

we have

$$\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{RP}}(\{p_{E_r}\}, m, r)$$

$$= \min_{\beta} \left\{ \sum_{E_r} p_{E_r} F\left(\frac{\beta}{\sqrt{2}} \sqrt{\frac{\bar{E}_r}{E_r}} \right) \frac{E_r}{\bar{E}_r} + \frac{\bar{E}}{n} \left[\frac{\beta^2 \sigma_v^2}{\sigma_v^2 + \bar{v}^2} + 1 + \frac{\sigma_{E_r}^2}{\bar{E}_r^2} \right] \right\} \sigma_\eta^2(\bar{E}_r) \tag{95}$$

By expanding F around $\beta/\sqrt{2}$, the represented loss $\mathcal{L}_{\text{rep}}$ can be further simplified as

$$\mathcal{L}_{\text{rep}} = \min_{\beta} \left\{ F\left(\frac{\beta}{\sqrt{2}}\right) - F'\left(\frac{\beta}{\sqrt{2}}\right) \frac{\beta \sigma_{E_r}^2}{8\sqrt{2}\bar{E}_r^2} + F''\left(\frac{\beta}{\sqrt{2}}\right) \frac{\beta^2 \sigma_{E_r}^2}{16 \bar{E}_r^2} \right.$$

$$\left. + \frac{\bar{E}}{n} \left[\frac{\beta^2 \sigma_v^2}{\sigma_v^2 + \bar{v}^2} + 1 + \frac{\sigma_{E_r}^2}{\bar{E}_r^2} \right] \right\} \sigma_\eta^2(\bar{E}_r). \tag{96}$$

Since F' is negative and F'' is positive, the correction increases $\mathcal{L}_{\text{rep}}$. The optimal β at $\sigma_{E_r}^2 = 0$ is denoted by β_0,

$$\beta = \beta_0 + \Delta\beta \sigma_{E_r}^2 + O(\sigma_{E_r}^4), \tag{97}$$

with

$$\Delta\beta \equiv \frac{F'(\beta_0/\sqrt{2}) \frac{1}{8\sqrt{2}} - F''(\beta_0/\sqrt{2}) \frac{\beta_0}{16} + F'''(\beta_0/\sqrt{2}) \frac{\beta_0^3}{16\sqrt{2}}}{\frac{1}{2} F''(\beta_0/\sqrt{2}) + \frac{2\bar{E}}{n} \frac{\sigma_v^2}{\sigma_v^2 + \bar{v}^2}} \frac{1}{\bar{E}_r^2}. \tag{98}$$

Plugging Eq. (97) into Eq. (96), the loss $\mathcal{L}_{\text{rep}}$ grows linearly with σ_E^2 to the lowest order. Therefore, we expect m^* to increase with σ_E^2, as increased $\mathcal{L}_{\text{rep}}$ requires larger width to outperform $\mathcal{L}_{\text{unrep}}$.

A.5 Perturbation analysis for independent features with unequal firing probabilities

In this section, we consider the independent features with weakly non-uniform firing probabilities:

$$p_i = p_0 + \delta p_i, \quad \text{for } i = 1, ..., n, \tag{99}$$

where p_0 is the mean firing probability, and the deviation δp_i is far less than p_0. Since the non-uniformity is weak, we can treat its effect perturbatively. In Sec. A.5.1, we introduce the leverage score, which we use to quantify the space each feature occupies. Then in Sec. A.5.2, we derive the lowest-order perturbative correction to the partial RP approximation.

A.5.1 THE LEVERAGE SCORE AS AN EFFECTIVE DIMENSIONALITY

We define the leverage score for represented feature and set the leverage score of zero vectors to be 0 by convention. For simplicity, here we assume all features are represented. Let $\tilde{\boldsymbol{W}}$ be the normalized frame

$$\tilde{\boldsymbol{W}} = \left(\frac{\boldsymbol{w}_1}{s_1} \frac{\boldsymbol{w}_2}{s_2} ... \frac{\boldsymbol{w}_n}{s_n} \right). \tag{100}$$

As in Eq. (17), the leverage score h_i of feature i is defined as

$$h_i \equiv \frac{1}{s_i^2} \boldsymbol{w}_i^T (\tilde{\boldsymbol{W}} \tilde{\boldsymbol{W}}^T)^+ \boldsymbol{w}_i. \tag{101}$$

In Eq. (101) the operator $(\cdot)^+$ indicates the Moore-Penrose inverse; when $\tilde{\boldsymbol{W}} \tilde{\boldsymbol{W}}^T$ is invertible, it is simply the inverse as in Eq. (17). The leverage score can be interpreted as an effective dimensionality that a feature occupies. It possesses the following properties:

- Normalization: If vectors span $\mathbb{R}^m$,

$$\sum_i h_i = m. \tag{102}$$

- Invariance under global rotation: If $\boldsymbol{W}$ is replaced by $\boldsymbol{R}\boldsymbol{W}$ (where $\boldsymbol{R} \in O(m)$), the leverage score h_i of each feature remains unchanged.
- Divisibility: if among all vectors, k vectors (subset A) span a d-dimensional subspace, and all other vectors (subset B) are perpendicular to this subspace. Then the leverage score of features in A satisfies $\sum_{i \in A} h_i = d$.
- Reasonable range: $h_i \in [0, 1]$. A vectors at most occupies one dimension, which happens when all other vectors are orthogonal to it. And the leverage score of a zero vector is zero by convention.

We use h_i to estimate the overlap between the representation vector of feature i and other vectors in the limit of weak non-uniformity. The Gram matrix and frame operator of $\tilde{\boldsymbol{W}}$ are denoted respectively by

$$\boldsymbol{G} = \tilde{\boldsymbol{W}}^T \tilde{\boldsymbol{W}} \in \mathbb{R}^{n \times n}, \tag{103}$$

$$\boldsymbol{S} = \tilde{\boldsymbol{W}} \tilde{\boldsymbol{W}}^T \in \mathbb{R}^{m \times m}. \tag{104}$$

The elements of the Gram matrix $G_{ij} = \boldsymbol{w}_i \cdot \boldsymbol{w}_j / s_i s_j$ are cosine similarities of vectors, satisfying

$$\sum_{j:j \neq i} \left(\frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_j}{s_i s_j} \right)^2 = \sum_j G_{ij}^2 - G_{ii}^2 \tag{105}$$

$$= (\boldsymbol{G}\boldsymbol{G}^T)_{ii} - 1. \tag{106}$$

Let

$$\tilde{\boldsymbol{W}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{107}$$

be the SVD decomposition of $\tilde{\boldsymbol{W}}$. Therefore,

$$\sum_{j:j \neq i} \left(\frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_j}{s_i s_j} \right)^2 = (\boldsymbol{G}\boldsymbol{G}^T)_{ii} - 1$$

$$= (\tilde{\boldsymbol{W}}^T \tilde{\boldsymbol{W}} \tilde{\boldsymbol{W}}^T \tilde{\boldsymbol{W}})_{ii} - 1 \tag{108}$$

$$= (\boldsymbol{V}\boldsymbol{\Sigma}^T \boldsymbol{U}^T \, \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \, \boldsymbol{V}\boldsymbol{\Sigma}^T \boldsymbol{U}^T \, \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)_{ii} - 1 \tag{109}$$

$$= \sum_{j:j \leq m} V_{ij}^2 \sigma_j^4 - 1, \tag{110}$$

where σ_j are singular values of $\tilde{\boldsymbol{W}}$. Meanwhile,

$$h_i = [\tilde{\boldsymbol{W}}^T(\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T)^+\tilde{\boldsymbol{W}}]_{ii} \tag{111}$$

$$= \left(\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\ \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T)^+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\right)_{ii} \tag{112}$$

$$= \sum_{j:j\le m} V_{ij}^2. \tag{113}$$

By comparing Eq. (110) and Eq. (113), we have

$$\sum_{j:j\neq i}\left(\frac{\boldsymbol{w}_i\cdot\boldsymbol{w}_j}{s_is_j}\right)^2 = h_i\frac{\sum_{j:j\le m}V_{ij}^2\sigma_j^4}{\sum_{j:j\le m}V_{ij}^2} - 1$$

$$= h_i\left[\left(\mathrm{Mean}_i^V\sigma_j^2\right)^2 + \mathrm{Var}_i^V\sigma_j^2\right] - 1, \tag{114}$$

where

$$\mathrm{Mean}_i^V\sigma_j^2 \equiv \frac{\sum_{j:j\le m}V_{ij}^2\sigma_j^2}{\sum_{j:j\le m}V_{ij}^2} = \frac{G_{ii}}{h_i} = \frac{1}{h_i}, \tag{115}$$

$$\mathrm{Var}_i^V\sigma_j^2 \equiv \frac{\sum_{j:j\le m}V_{ij}^2(\sigma_j^2 - \mathrm{Mean}_i^V\sigma_j^2)^2}{\sum_{j:j\le m}V_{ij}^2}. \tag{116}$$

For tight frames (which we assume the trained representations approximately achieve to minimize σ_η^2, see Sec. A.2), all σ_j are equal and the variance $\mathrm{Var}_i^V\sigma_j^2$ vanishes. Following it, we assume the optimal representation is nearly tight when the non-uniformity is weak, i.e., $\mathrm{Var}_i^V\sigma_j^2 \ll \frac{1}{h_i^2} \sim \frac{n^2}{m^2}$. Furthermore, we assume $m \ll n$. Therefore,

$$\sum_{j:j\neq i}\left(\frac{\boldsymbol{w}_i\cdot\boldsymbol{w}_j}{s_is_j}\right)^2 \approx \frac{1}{h_i} - 1 \approx \frac{1}{h_i}. \tag{117}$$

When $h_i = \frac{m}{n}$ are all equal, Eq. (117) is consistent with Eq. (45).

Based on Eq. (117), the crosstalk noise on feature i is approximated as a Gaussian random variable with mean zero and variance

$$\sigma_{s^2\eta,i}^2 \equiv \mathbb{E}_{\boldsymbol{x}}\left[s_i\sum_{j\in\{a_k\}\setminus i}v_js_j\left(\frac{\boldsymbol{w}_i\cdot\boldsymbol{w}_j}{s_is_j}\right)\right]^2 \approx \frac{s_i^2\langle s^2p\rangle}{h_i}(\bar{v}^2+\sigma_v^2), \tag{118}$$

where $\{a_k\}$ is the set of active features in an input, and $\langle s^2p\rangle \equiv \frac{1}{n}\sum_i s_i^2p_i$. Here we apply the mean field approximation on the crosstalk noise again.

A.5.2 Lowest-order perturbative correction to the representation and the loss

Using the leverage score, we can estimate the loss at weak non-uniformity perturbatively. We first assume that all features are represented, so the random projection approximation (51) acts as the baseline of the perturbation analysis. We work in the limit $1 \ll \bar{E} = p_0n \ll m \ll n$ and ignore the effect of σ_E. Similar to Eq. (69), the loss is approximated by

$$\mathcal{L}(\{p_i\},m,n;\{b_i\},\{s_i\},\{h_i\}) = \frac{1}{n}\sum_{i=1}^n\mathcal{L}_f(s_i,b_i,h_i,p_i,\langle s^2p\rangle) \tag{119}$$

$$= \frac{1}{n}\sum_i\left[(1-p_i)F\left(-\frac{b_i}{\sqrt{2\sigma_{s^2\eta,i}^2}}\right)\sigma_{s^2\eta,i}^2 + p_i\left((s_i^2-1)^2\sigma_v^2 + (s_i^2-1+b_i)^2 + \sigma_{s^2\eta,i}^2\right)\right]. \tag{120}$$

In Eq. (119), $\mathcal{L}_f$ is the loss contribution of a single feature. The predicted loss is obtained by optimizing $\{b_i\}, \{s_i\}$ and $\{h_i\}$ under the constraint (102). Therefore, we use the method of Lagrange multipliers, and let

$$\tilde{\mathcal{L}} \equiv n\mathcal{L} + A\left(\sum_i h_i - m\right). \tag{121}$$

Baseline parameters are

$$h_0 = \frac{m}{n}, \tag{122}$$

$$(s_0, b_0) = \arg\min_{(s,b)} \mathcal{L}_f(s, b, h_0, p_0, \langle s^2 p\rangle)|_{\langle s^2 p\rangle = s^2 p_0}. \tag{123}$$

And we define

$$s_i = s_0 + \delta s_i, \tag{124}$$

$$b_i = b_0 + \delta b_i, \tag{125}$$

$$h_i = h_0 + \delta h_i. \tag{126}$$

We expand $\tilde{\mathcal{L}}$ around the baseline parameters:

$$\begin{aligned}
&\tilde{\mathcal{L}}(\{s_i\}, \{b_i\}, \{h_i\}, \{p_i\}, A) \\
=&n\mathcal{L}|_0 + \sum_i \left(\frac{\partial \mathcal{L}_f}{\partial s}\bigg|_0 \delta s_i + \frac{\partial \mathcal{L}_f}{\partial b}\bigg|_0 \delta b_i + \frac{\partial \mathcal{L}_f}{\partial p}\bigg|_0 \delta p_i + \frac{\partial \mathcal{L}_f}{\partial h}\bigg|_0 \delta h_i \right. \\
&\qquad \left. + A\delta h_i + \frac{\partial \mathcal{L}_f}{\partial \langle s^2 p\rangle}\bigg|_0 \cdot \frac{1}{n}\sum_k \left(2p_0 s_0 \delta s_k + s_0^2 \delta p_k + p_0 \delta s_k^2 + 2s_0 \delta s_k \delta p_k\right) \right) \\
&+ \sum_i \frac{1}{2}(\delta s_i, \delta b_i, \delta h_i; \delta p_i) \cdot \mathcal{H}|_0 \cdot (\delta s_i, \delta b_i, \delta h_i; \delta p_i)^T + \text{h.o.t.},
\end{aligned} \tag{127}$$

where $(\cdot)|_0$ denotes the function evaluated at s_0, b_0, h_0 and p_0, and $\mathcal{H}$ is the Hessian matrix of $\mathcal{L}_f$ at fixed $\langle s^2 p\rangle = s_0^2 p_0$. First order terms vanish because of Eq. (102) and (123). So $\tilde{\mathcal{L}}$ can be simplified as

$$\tilde{\mathcal{L}} = n\mathcal{L}|_0 + \sum_i \left[\frac{1}{2}(\delta s_i, \delta b_i, \delta h_i; \delta p_i) \cdot \mathcal{H}|_0 \cdot (\delta s_i, \delta b_i, \delta h_i; \delta p_i)^T + \frac{\partial \mathcal{L}_f}{\partial \langle s^2 p\rangle}\bigg|_0 \left(p_0 \delta s_i^2 + 2s_0 \delta s_i \delta p_i\right) \right] \tag{128}$$

$$= n\mathcal{L}|_0 + \frac{1}{2}\sum_i \left((\delta s_i, \delta b_i, \delta h_i) \cdot \mathcal{H}'|_0 \cdot (\delta s_i, \delta b_i, \delta h_i)^T + 2(\delta s_i, \delta b_i, \delta h_i) \cdot \mathcal{V}'|_0 \delta p_i + \mathcal{H}_{pp}|_0 \delta p_i^2 \right), \tag{129}$$

where

$$\mathcal{H}' \equiv \begin{pmatrix} \mathcal{H}_{ss} + 2p\frac{\partial \mathcal{L}_f}{\partial \langle s^2 p\rangle} & \mathcal{H}_{sb} & \mathcal{H}_{sh} \\ \mathcal{H}_{bs} & \mathcal{H}_{bb} & \mathcal{H}_{bh} \\ \mathcal{H}_{hs} & \mathcal{H}_{hb} & \mathcal{H}_{hh} \end{pmatrix}, \qquad \mathcal{V}' \equiv \begin{pmatrix} \mathcal{H}_{sp} + 2s\frac{\partial \mathcal{L}_f}{\partial \langle s^2 p\rangle} \\ \mathcal{H}_{bp} \\ \mathcal{H}_{hp} \end{pmatrix}. \tag{130}$$

Therefore, the optimal parameters satisfy

$$\begin{pmatrix} \delta s_i \\ \delta b_i \\ \delta h_i \end{pmatrix} = -(\mathcal{H}'|_0)^{-1} \cdot \mathcal{V}'|_0 \, \delta p_i. \tag{131}$$

Correspondingly, the optimal loss $\mathcal{L}$ is

$$\mathcal{L}(\{p_i\}, m, n) = \mathcal{L}|_0 - c\sum_i \delta p_i^2, \tag{132}$$

with

$$c \equiv \frac{1}{2n}\left[(\mathcal{V}'|_0)^T (\mathcal{H}'|_0)^{-1} \mathcal{V}'|_0 - \mathcal{H}_{pp}|_0\right]. \tag{133}$$

Eq. (131) and (132) give the lowest-order correction to the representation and the loss.

For partial representation, the correction is applied to represented features, with n replaced by r, $\langle s^2 p\rangle$ replaced by the mean over represented features, h_0 replaced by $\frac{m}{r}$, and p_0 taken as the mean firing probability of represented features.

A.5.3 Discussion on the lowest-order perturbative correction

We first discuss the signs of the corrections. In the experiments with uniform features, the norms and biases of all represented features are nearly equal at the loss minimum (Fig. 2ab). Therefore, at $\delta p_i = 0$ the optimum is $\delta s_i = \delta b_i = \delta h_i = 0$, i.e., the uniform solution is a local minimum. This ensures $\mathcal{L}$ in Eq. (129) increases with any fluctuation on $\delta s_i, \delta b_i$ and δh_i, i.e., $\mathcal{H}'|_0$ is positive definite. Moreover, at fixed $\langle s^2 p\rangle$, $\mathcal{L}_f$ in Eq. (119) is affine in p_i, hence $\mathcal{H}_{pp}|_0 = 0$. Following Eq. (133),

$$c = \frac{1}{2n}(\mathcal{V}'|_0)^T(\mathcal{H}'|_0)^{-1}\mathcal{V}'|_0 > 0 \tag{134}$$

for generic parameters, where $\mathcal{V}'|_0 \neq 0$. Therefore, non-uniformity lowers the loss. The signs of the responses δs_i, δb_i and δh_i depend on the values of the Hessian elements, so we verify them numerically. We solve Eq. (131) for all parameters we tested in toy model experiments. For all tested parameters, $\mathcal{H}'|_0$ is positive definite, and

$$\frac{\delta h_i}{\delta p_i} > 0, \quad \frac{\delta b_i}{\delta p_i} > 0, \quad \frac{\delta s_i}{\delta p_i} < 0. \tag{135}$$

Therefore, a more frequent feature occupies more dimensions, receives less crosstalk noise, and hence needs less filtering by its norm and bias.

We emphasize that the perturbation analysis relies on approximations beyond the RP approximation, so it should be regarded as an ansatz. First, we neglect the term $h_i \mathrm{Var}_i^V \sigma_j^2$ in Eq. (114). This term cannot vanish under non-uniformity: if all σ_j are equal, $\tilde{W}\tilde{W}^T = \frac{n}{m}\boldsymbol{I}$, and all $h_i = \frac{m}{n}$. Non-uniform h_i thus requires non-uniform σ_j, and $\mathrm{Var}_i^V \sigma_j^2$ may enter the loss. Its effect is beyond the scope of this paper. Second, the mean field approximation in Eq. (118) assumes that the overlaps of feature i are spread over all other features regardless of their firing probabilities, so that the crosstalk noise depends on other features only through $\langle s^2 p\rangle$. In trained models, the organization of representation vectors is correlated with firing probabilities. For example, Liu et al. (2025) found that the representation vectors of more frequent features tend to be ETF-like, with nearly equal and small overlaps among themselves. The crosstalk noise on a feature then depends on which features it overlaps with, rather than only on the mean.

Finally, the perturbation analysis requires $|\delta p_i| \ll p_0$ for all features. The power-law firing probabilities used in Sec. 3.3 do not satisfy this condition in general. For $p_i \propto i^{-\epsilon_p}$, $\delta p_i/p_0 \approx \epsilon_p(\log \frac{n}{i} - 1)$, so the condition requires $\epsilon_p \log n \ll 1$. For $n = 10000$ and $\epsilon_p = 0.2$ as in Fig. 7, $\epsilon_p \log n \approx 1.8$, and the firing probability of the most frequent feature is about five times p_0. Therefore, we do not expect Eq. (131) and (132) to give correct values in this setup. They only explain the trends: the representation adapts to non-uniformity, and the loss decreases with ϵ_p.

B Toy model training

In this section, we explain how we train the toy model to obtain raw data. We carry out toy model experiments using PyTorch. Hyperparameters we use are given as follow:

- Data dimension n (total number of features): varies from 1000 to 65536. In the main text, we only show results at $n = 10000$.
- Model width m: varied from 10 to 16384. In each case, the model width m is always less than n.
- Batch size: set to be n in each case.
- Total training steps: 30000 (tested up to 50000, which does not change any result; see Sec. C.1.1)
- Optimizer: Adam (Kingma & Ba, 2017) with the default PyTorch settings ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, no weight decay).
- Learning rate: dynamically adjusted using cosine decay scheduling

$$l(t) = \begin{cases} l_{\max}\frac{t}{t_1}, & t < t_1, \\ l_{\mathrm{fnl}} + (l_{\max} - l_{\mathrm{fnl}}) \cdot \frac{1}{2}\left(1 + \cos \pi \frac{t - t_1}{t_2 - t_1}\right), & t_1 \leq t < t_2, \\ l_{\mathrm{fnl}}, & t \geq t_2, \end{cases} \tag{136}$$

where warm-up time $t_1 = 2000$, step number of learning rate decay $t_2 = 20000$, and final learning rate $l_{\text{fnl}} = 0.0002$, the maximal learning rate $l_{\max} = 20 l_{\text{fnl}}$.

- Initialization of $\boldsymbol{W}$: each representation vector $\boldsymbol{w}_i$ is drawn independently and uniformly from the unit sphere in $\mathbb{R}^m$. The initial configuration is thus a random projection frame with $s_i = 1$ for all features.
- Initialization of $\boldsymbol{b}$: $\boldsymbol{b} = \boldsymbol{0}$.

A new batch is randomly sampled from the data distribution at every step, so the model is trained on the input ensemble instead of a fixed dataset. Training is performed using Nvidia L40S, with floating-point precision (FP32).

C Details of experimental results on the toy model

In this section, we show more experimental results on the toy model of superposition. All models are trained as described in Appendix B. The loss $\mathcal{L}$ is averaged over the last 5000 training steps. Representation parameters $\boldsymbol{w}_i$ and b_i are measured after training; parameters s_i, h_i and r are calculated from $\boldsymbol{w}_i$.

C.1 Uniform independent features

C.1.1 Training convergence and experimental errors

Fig. 9 shows an example loss sequence during training. The loss $\mathcal{L}$ reaches a plateau before the learning rate reaches its final value at $t_2 = 20000$, and remains constant within the last 5000 steps, over which we average the loss. Training longer, up to 5×10^4 steps, does not change the results.

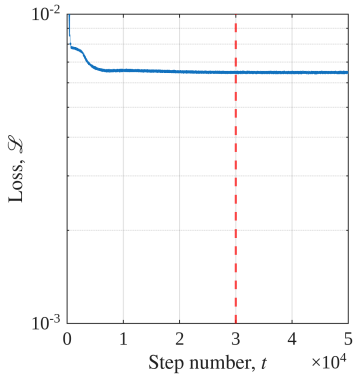


Figure 9: Time sequences of trained loss. The loss converges before $t = 3 \times 10^4$. $\bar{E} = 64, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}, m = 200$.

We repeat several parameter settings with 10 random seeds. In all tested cases, the errors (seed-to-seed standard deviations) of $\mathcal{L}$ and r/n are too small to be visible in the figures (Fig. 10). We therefore use a single seed for most settings and omit error bars in most figures.

Near a continuous phase transition, statistical mechanics suggests that fluctuations of the order parameter r/n (the susceptibility) and of the loss $\mathcal{L}$ (which plays the role of an effective energy) grow. In the thermodynamic limit, these may diverge at the critical point. At our resolution, this effect is weak: the standard deviation of r/n increases only mildly below the transition, and the relative standard deviation of $\mathcal{L}$ shows no peak near it (Fig. 10). It does not affect our results, since we do not locate the critical point precisely; we only extract the trend of m^* from the phase diagrams.

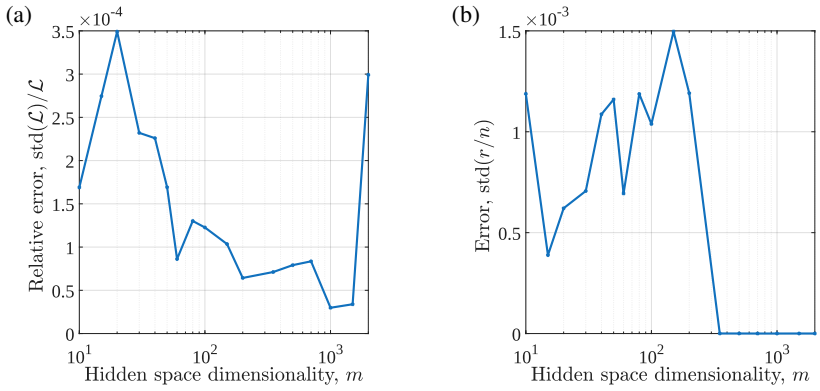


Figure 10: (a) Relative error of the loss $\mathcal{L}$. (b) Error of r/n. The typical value of r/n is of order 1. Since we focus on the transition, where r/n reaches 1, we report the absolute rather than the relative error of r/n. For both panels 10 seeds are used. $\bar{E} = 50, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

C.1.2 Effect of $\bar{v}$ and σ_v

In this section, we discuss the effect of $\bar{v}$ and σ_v. The mean activation $\bar{v}$ trivially acts as a scaling factor of the model. Under the rescaling

$$\bar{v} \to C\bar{v}, \tag{137}$$
$$\sigma_v^2 \to C^2\sigma_v^2, \tag{138}$$
$$\boldsymbol{W} \to \boldsymbol{W}, \tag{139}$$
$$\boldsymbol{b} \to C\boldsymbol{b}, \tag{140}$$

with positive C, the loss $\mathcal{L} \to C^2\mathcal{L}$. So in this paper, without loss of generality, we fix $\bar{v} = 1$.

With $\bar{v} = 1$, the activation strength v is drawn from $U[1-\sqrt{3}\sigma_v,\ 1+\sqrt{3}\sigma_v]$. Nonnegative activations require $\sigma_v \le 1/\sqrt{3}$, so we vary σ_v from 0 to $1/\sqrt{3}$, where $v \sim U[0, 2]$ as in the main text. The activation variance σ_v^2 changes the loss $\mathcal{L}$ and the critical width m^*. Experiments show that $\mathcal{L}$ increases with σ_v^2 (Fig. 11b). This trend is captured by the partial RP approximation, see Eq. (60) and (74). The critical width m^* decreases with σ_v (Fig. 11a). The partial RP approximation numerically captures this trend, but it cannot be described by asymptotic expressions derived in Sec. A.4. This is mainly a consequence of the nonlinearity of the loss.

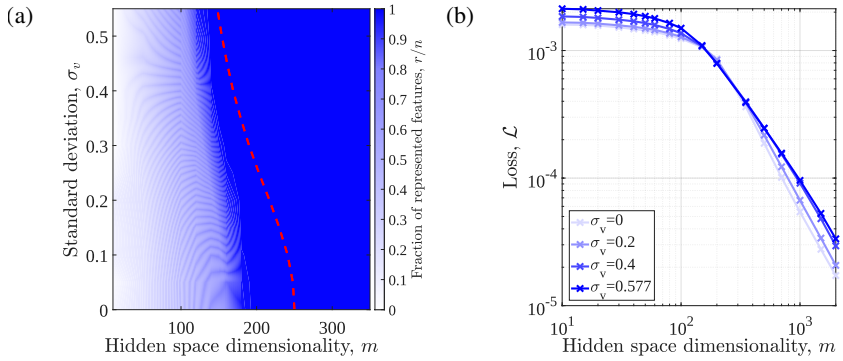


Figure 11: (a) Phase diagram of r/n as a function of m and σ_v. The red dashed curve is the critical width m^* predicted by partial RP approximation. The critical width m^* decreases with σ_v. (b) The $\mathcal{L}$-m at different σ_v. The loss $\mathcal{L}$ increases with σ_v. $\bar{E} = 16, n = 10000$.

C.1.3 TWO CONSTRAINED REPRESENTATIONS: EQUAL-NORM FRAME AND RANDOM PROJECTION FRAME

In experiments and in partial RP approximation, features are divided into two groups: represented and unrepresented groups. To isolate the behavior of the represented group, we introduce two variant models with constrained representations based on the original toy model in Sec. 2:

1. The equal-norm (EN) model: During training, all vectors are forced to have equal norms. So the representation is forced to be an equal-norm frame.
   We parametrize each representation vector as $\boldsymbol{w}_i = s\,\hat{\boldsymbol{w}}_i/\|\hat{\boldsymbol{w}}_i\|$, where $\hat{\boldsymbol{w}}_i \in \mathbb{R}^m$ are unconstrained trainable vectors and s is a single trainable scalar shared by all features. The directions of the representation vectors are thus trained freely, while all norms equal s at every step, so the representation is always an equal-norm frame. The biases b_i are trained independently for each feature. We initialize $\hat{\boldsymbol{w}}_i$ from a standard Gaussian, $s = 1$ and $\boldsymbol{b} = \boldsymbol{0}$, so the initial configuration is the same random projection frame as in the original model (Sec. B).
2. The random projection (RP) model: During training, directions of all representation vectors are not updated and norms of them are forced equal. Furthermore, the bias of all features are forced equal. So there are only two trained parameters: norm s and bias b. The vector configuration remains the initial random projection frame (see Sec. B).

The EN model is the realization of ideal full representation, and the RP model is a further simplified case. Neither of them captures the phase transition from partial to full representation. Experiments show that, in full-representation phase, the loss $\mathcal{L}$ and representation parameters b and s of the EN model are very close to those of the original model (Fig. 12). This result is also consistent with the observation that in full-representation phase, all s_i (and b_i) are highly clustered. Therefore, in this section and the following section C.1.4, we treat the EN model and the original toy model in full-representation phase as equivalent.

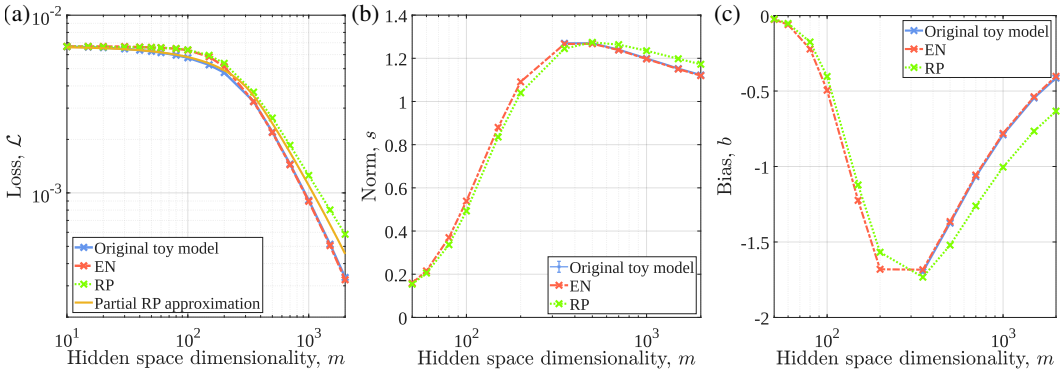


Figure 12: Losses and representation parameters of different models. (a) Losses of different models. All loss curves have the same trend and similar bending. (b) The vector norms of different models. For the original toy model, the mean norm of all features is shown, and only the data in full-representation phase ($350 \lesssim m$) is shown. The error bar is too small to be seen. (c) The biases of different models. For the original toy model, the mean bias of all features is shown; only the data in full-representation phase ($350 \lesssim m$) is shown. $\bar{E} = 50, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

We measure the mean value and the variance of $c_{ij}|_{i\neq j}$ in the trained EN model. For all parameters we tested, the mean value and the variance of $c_{ij}|_{i\neq j}$ both reach the theoretical lower bound (Fig. 13) in the EN model, hence the trained representation is nearly a tight frame.

Using the c_{ij} measured in the experiments, we can also reconstruct the distribution of crosstalk noise η in the EN model: To generate a crosstalk noise value, we sample $\bar{E}$ elements from the pool of $c_{ij}|_{i\neq j}$ and multiply them each by a random v_i drawn from the distribution of activation, and then add them up. (Here we fix $E = \bar{E}$.) The reconstructed crosstalk noise distribution is close to the Gaussian approximation (10) at small m, but deviates from Gaussian distribution as m increases

(Fig. 14). At large m, the distribution of η becomes asymmetric (Fig. 14). Since its variance is fixed at the Welch bound (Fig. 13b), the trained frame can lower the loss only by changing the shape of the distribution. On inactive represented features (which is much more than active represented features), the output is $[s^2\eta + b]_+$ with $b < 0$, so only the positive tail $\eta > -b/s^2$ leaks through the ReLU and bias and contributes to the loss. Correspondingly, the trained distribution is negatively skewed, with a lighter positive tail than the Gaussian distribution of the same variance. This suppresses the leakage, so the Gaussian approximation (10) overestimates the loss. This is the main origin of the slightly higher predicted loss in the full-representation phase (Fig. 3d). We use the skewness to quantify the deviation from Gaussianity. Empirically, it depends approximately on m/n and vanishes as $m/n \to 0$ (Fig. 15a). We therefore regard the loss gap as a finite size effect, which disappears in the limit $1 \ll \bar{E} \ll n$ (Sec. C.1.4). Plugging the reconstructed η distribution into Eq. (41), we obtain a mean-field loss that matches the trained loss of the EN model (Fig. 15b). This indicates that the mean field approximation is not the main cause of approximation error; the error mainly arises in using Gaussian distribution to approximate the crosstalk noise. We expect all these results to hold for the original toy model in full-representation phase too.

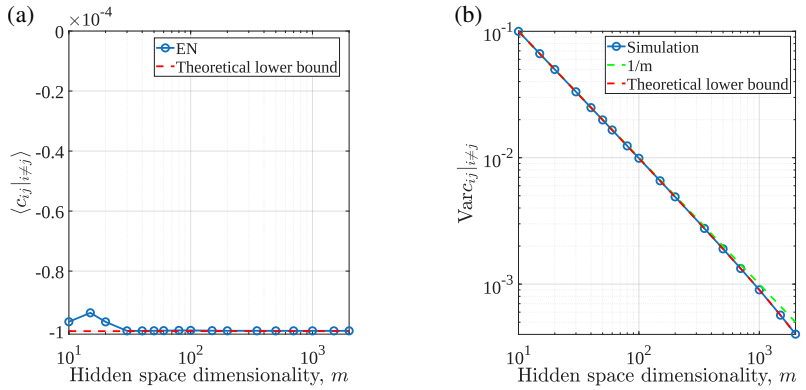


Figure 13: (a) The mean value of $c_{ij}|_{i\neq j}$ in the EN model. It reaches the theoretical lower bound $-\frac{1}{n-1}$. (b) The variance of $c_{ij}|_{i\neq j}$ in the EN model. It reaches the theoretical lower bound $\frac{n-m}{m(n-1)}$. The curve $1/m$ is plotted out as a comparison. $\bar{E} = 10, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

The crosstalk noise distribution of the RP model is highly Gaussian, with a variance $\frac{\bar{E}}{m}(\bar{v}^2 + \sigma_v^2)$ rather than $\bar{E}\frac{n-m}{m(n-1)}(\bar{v}^2+\sigma_v^2)$, as shown in Fig. 16. In the limit $m \ll n$, the two variances converge, and the crosstalk noise distribution of the RP model can be described by Eq. (10). This is why we call our approximation the (partial) random projection approximation.

C.1.4 Approaching the limit $1 \ll \bar{E} \ll n$

Our asymptotic analysis of $\mathcal{L}_{\text{rep}}$ lies in the limit $1 \ll \bar{E} \ll n$. To approach this limit in experiments, we first increase n at fixed $\frac{\bar{E}}{n}$, then decrease $\frac{\bar{E}}{n}$. To isolate the represented features, we use the EN model (Sec. C.1.3) instead of the original toy model, and compare its result with predicted $\mathcal{L}_{\text{rep}}$ given by Eq. (51).

At fixed $\bar{E}/n$, loss curves of different n collapse when n is large enough. For the predicted $\mathcal{L}_{\text{rep}}$, we scale n up and confirm the collapse of $\mathcal{L}_{\text{rep}}$-m/n curves (Fig. 17a). For the EN model, the maximal n we can test is limited by our computational resources, hence we cannot see loss curves collapsing directly. However, by fitting $\mathcal{L}_{\text{EN}}(n)$ at fixed m/n, we find that in a large range of m/n, $\mathcal{L}_{\text{EN}}$ converges to an asymptotic value in the limit $n \to \infty$ (Fig. 17bc).

We then decrease $\bar{E}/n$ to approach the limit $1 \ll \bar{E} \ll n$. Based on the result of the original toy model, we rescale the loss $\mathcal{L}$ to $\mathcal{L}\cdot n/\bar{E}$, and the width m to $m/(\bar{E}\log\frac{n}{\bar{E}})$ or equivalently m/m^*. For $\mathcal{L}_{\text{rep}}$ predicted by Eq. (51), the loss curves collapse well at small m; at large m, due to the finite size effect (as m reaches n, the loss decays toward 0), the loss curves deviate (Fig. 18a). The asymptotic values of $\mathcal{L}_{\text{EN}}$ show similar behaviors (Fig. 18b). As $\bar{E}/n$ approaches 0, the asymptotic values of

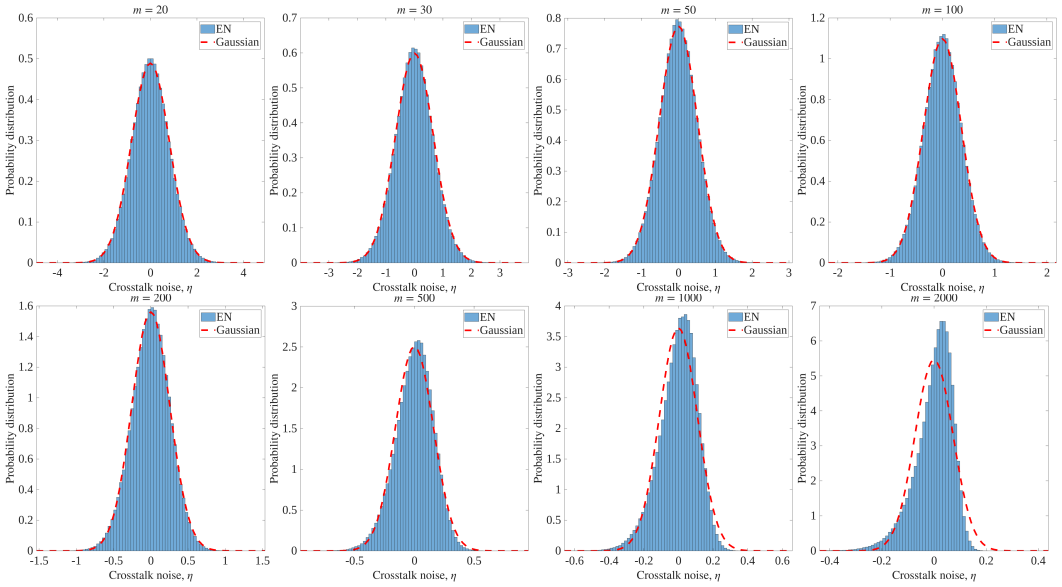


Figure 14: The distribution of the crosstalk noise η reconstructed from c_{ij} of the EN model. The red dashed line is the pdf of Gaussian distribution (10). $\bar{E} = 10, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

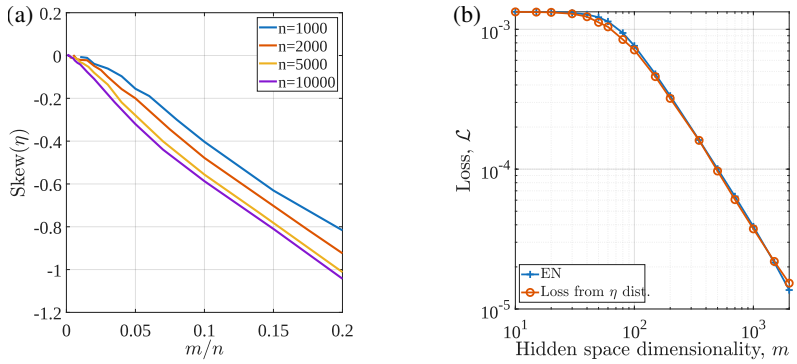


Figure 15: (a) The skewness of η at different n. The skewness decrease with m/n from 0. $\bar{E} = 10, \sigma_v = \frac{1}{\sqrt{3}}$. (b) The trained loss of the EN model and the mean field loss calculated by applying the reconstructed η distribution to Eq. (41). They are highly consistent. $\bar{E} = 10, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

$\mathcal{L}_{\rm EN}$ converge to the predicted limiting value $\lim_{\bar{E}/n\to 0} \lim_{n\to\infty} \mathcal{L}_{\rm rep}$. We therefore conclude that in the limit $1 \ll \bar{E} \ll n$, the EN model loss $\mathcal{L}_{\rm EN}$ converges to the predicted $\mathcal{L}_{\rm rep}$.

All analyses in this section use the EN model. Nevertheless, all results hold for the original toy model in the full-representation phase (see Sec. C.1.3). We also expect these results to be valid for $\mathcal{L}_{\rm rep}$ in partial-representation phase, because s_i and b_i of represented features are highly clustered in that phase. Together with Sec. C.1.3, this shows that the gap between the partial RP prediction and the trained loss in the full-representation phase (Fig. 3) is a finite size effect: it arises from the non-Gaussian crosstalk noise at finite m/n, and vanishes in the limit $1 \ll \bar{E} \ll n$.

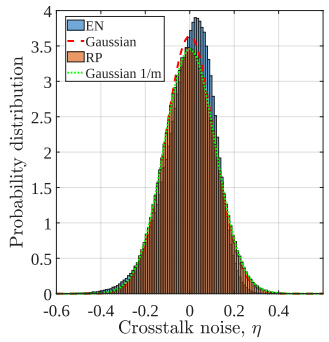


Figure 16: The crosstalk noise distribution of the RP model and the EN model. The red dashed curve is the pdf of Gaussian distribution (10), and the green dashed curve is the pdf of Gaussian distribution with variance $\bar{E}/m \cdot (\bar{v}^2 + \sigma_v^2)$. The crosstalk noise distribution of the RP model is highly consistent with the Gaussian distribution with variance $\bar{E}/m \cdot (\bar{v}^2 + \sigma_v^2)$. $\bar{E} = 10, m = 1000, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

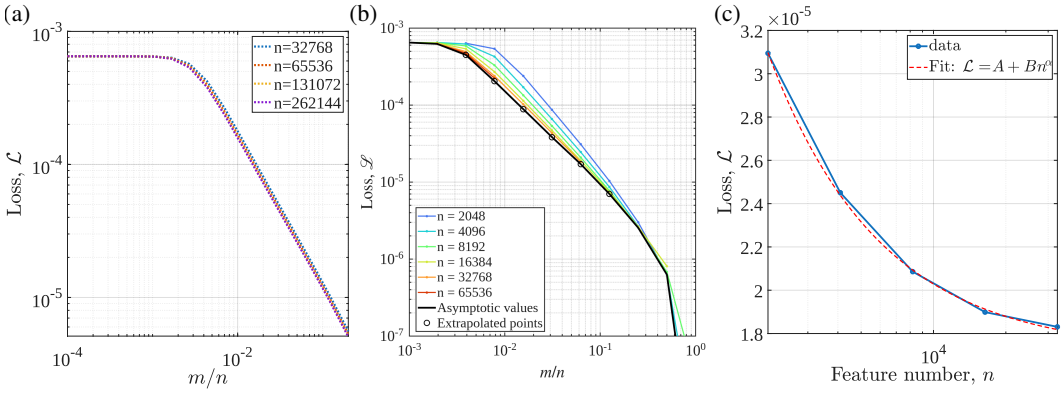


Figure 17: (a) The represented loss $\mathcal{L}_{\text{rep}}$ predicted by Eq. (51) at fixed $\bar{E}/n$. Loss curves collapse after rescaling m to m/n. (b) The EN model loss $\mathcal{L}_{\text{EN}}$ at different n and its asymptotic value in $n \to \infty$. For each m/n, we fit $\mathcal{L}_{\text{EN}}(n)$ by $\mathcal{L}_{\text{EN}} = A + Bn^{\alpha}$. If the fit matches the loss, we use A from fitting as the asymptotic value, otherwise we use the minimal measured loss as the experimental asymptotic value (an upper bound on the real asymptotic value). Circles mark the points based on fit result. (c) $\mathcal{L}_{\text{EN}}$ and the fit result at $m/n = 1/16$. $\bar{E}/n = 1/2048$, $\sigma_v = \frac{1}{\sqrt{3}}$.

C.2 Exchangeable correlated features

C.2.1 Generation of exchangeable correlated features

In this section, we describe the method to generate $\boldsymbol{u}$ for exchangeable correlated features. The generation consists of:

1. Generating E, the number of active features in the input, following a discrete distribution $\{P_E\}$, where P_E is the probability that E active features occur.
2. Randomly selecting E features uniformly at random. Selected features are set active, and the rest are inactive.

The probability array $\{P_E\}$ is tunable, and is not uniquely determined by its mean $\bar{E}$ and variance σ_E^2. We restrict ourselves to the setup of Pólya–Eggenberger urn (Johnson & Kotz, 1977). Under this setup, σ_E has four regimes with different $\{P_E\}$ implementation. At $\sigma_E = 0$, the distribution degenerates to a delta at $\bar{E}$:

$$P_E = \delta_{E,\bar{E}}. \tag{141}$$

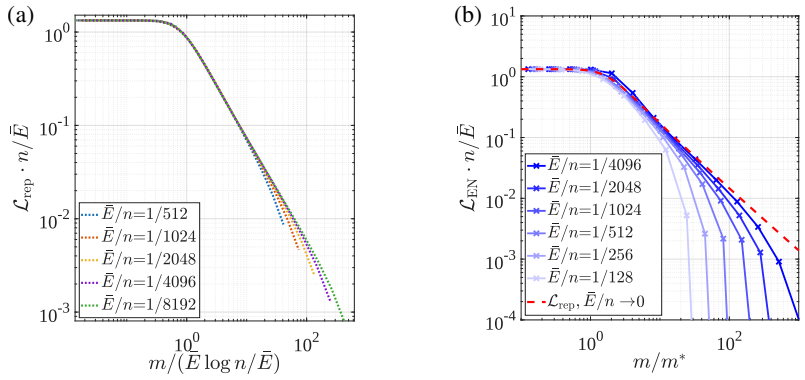


Figure 18: (a) The represented loss $\mathcal{L}_{\rm rep}$ predicted by Eq. (51) at different $\bar{E}/n$. n is large enough so that the large-n collapse happens. (b) The asymptotic values of the EN model loss $\mathcal{L}_{\rm EN}$ at different $\bar{E}/n$. Experimental asymptotic values are obtained by the extrapolation described in Fig. 17. The red dashed curve is the limiting value $\lim_{\bar{E}/n\to 0}\lim_{n\to\infty}\mathcal{L}_{\rm rep}$ computed numerically from Eq. (51). $\sigma_v = \frac{1}{\sqrt{3}}$.

At $\sigma_E = \sigma_E^{\rm bin} = \sqrt{\bar{E}(1-\frac{\bar{E}}{n})}$, the distribution is just binomial distribution:

$$P_E = \binom{n}{E} \cdot \left(\frac{\bar{E}}{n}\right)^E \left(1-\frac{\bar{E}}{n}\right)^{n-E}. \tag{142}$$

Features are activated independently in this case. In the range $0 < \sigma_E < \sigma_E^{\rm bin}$, the activation number E follows hypergeometric distribution:

$$P_E = \frac{\binom{K}{E}\binom{N-K}{n-E}}{\binom{N}{n}}, \quad \text{for } E \in \{\max(0, n+K-N), ..., \min(n, K)\}, \tag{143}$$

where

$$N \equiv \frac{n-(\sigma_E/\sigma_E^{\rm bin})^2}{1-(\sigma_E/\sigma_E^{\rm bin})^2}, \tag{144}$$

$$K \equiv \frac{\bar{E}}{n}N. \tag{145}$$

In the range $0 < \sigma_E^{\rm bin} < \sigma_E$, the activation number E follows beta-binomial distribution:

$$P_E = \binom{n}{E}\frac{B(E+\alpha, n-E+\beta)}{B(\alpha,\beta)}, \tag{146}$$

where

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{147}$$

$$\alpha = \frac{\bar{E}}{n}\frac{(\sigma_E/\sigma_E^{\rm bin})^2 - n}{1-(\sigma_E/\sigma_E^{\rm bin})^2}, \tag{148}$$

$$\beta = \left(1-\frac{\bar{E}}{n}\right)\frac{(\sigma_E/\sigma_E^{\rm bin})^2 - n}{1-(\sigma_E/\sigma_E^{\rm bin})^2}. \tag{149}$$

Following this setup, the possible σ_E ranges from 0 to $\sqrt{\bar{E}(n-\bar{E})}$. We restrict σ_E to be of order $\bar{E}$.

C.2.2 EFFECT OF σ_E

We test various $\bar{E}$ and σ_E, with $\sigma_E^2 \leq 200$, and show $\bar{E} = 20$ as a representative example; other values of $\bar{E}$ give qualitatively the same results. In the parameter range we test, the phase transition from partial representation to full representation remains valid (Fig. 19). Toy model experiment shows that the critical width m^* grows with σ_E. This trend is captured by partial RP approximation (Fig. 20). In the full-representation phase, the loss grows with σ_E^2 approximately linearly at small σ_E^2, as predicted in Sec. A.4.3; at larger σ_E^2, the growth becomes slightly nonlinear (Fig. 21). In the partial-representation phase, since the loss is dominated by unrepresented features, it is insensitive to σ_E.

We expect the representation behavior will be qualitatively different at large enough σ_E. For example, when σ_E reaches its upper bound, all features can only fire together, hence one vector is enough to represent the firing states of all features. If σ_v also vanishes, a model with width 1 can recover all inputs. But the representation behavior at large σ_E lies beyond the scope of this paper.

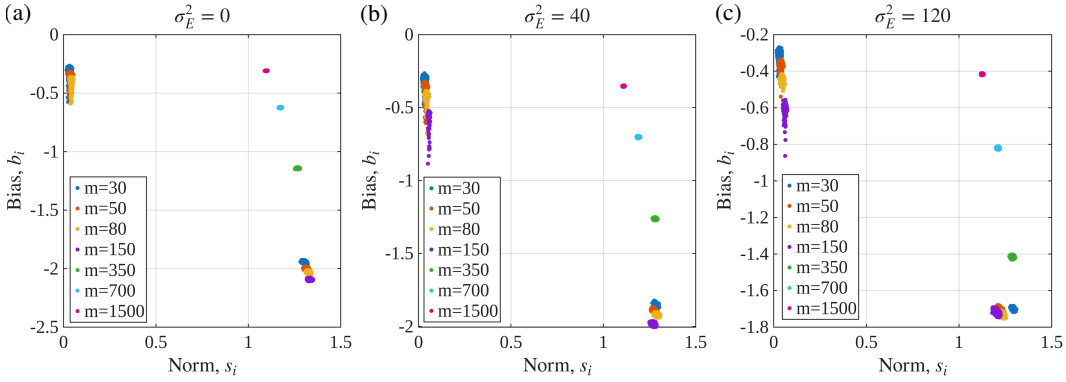


Figure 19: The distribution of s_i and b_i of each feature. Sample points at small m form two clusters while at large m they form one cluster. Different panels are distributions at different σ_E^2. $\bar{E} = 20, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

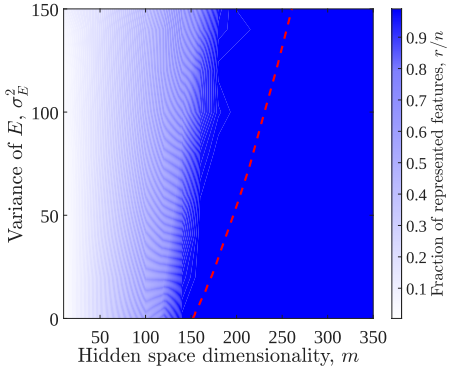


Figure 20: Phase diagram of r/n as a function of m and σ_E^2. The red dashed curve is the critical width m^* predicted by partial RP approximation. The critical width m^* increases with σ_E^2. $\bar{E} = 20, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

C.3 INDEPENDENT FEATURES WITH UNEQUAL FIRING PROBABILITIES

C.3.1 APPLICABLE RANGE OF ϵ_p AND BIAS BEHAVIOR

With power-law firing probabilities $p_i = A\, i^{-\epsilon_p}$, the prefactor is fixed by $\sum_i p_i = \bar{E}$, i.e., $A = \bar{E}/H_n(\epsilon_p)$ with $H_n(\epsilon_p) \equiv \sum_{i=1}^{n} i^{-\epsilon_p}$. Since $p_1 = A$ is a probability, we require $p_1 < 1$, i.e.,

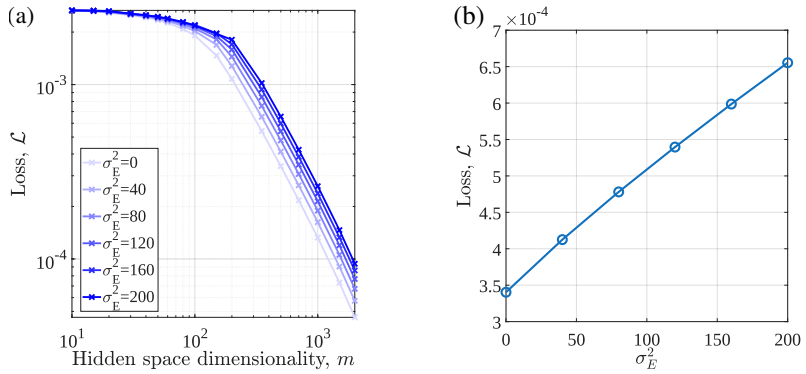


Figure 21: (a) The loss $\mathcal{L}$ as a function of m at different σ_E. In full-representation phase, the loss $\mathcal{L}$ increases with σ_E. (b) The loss $\mathcal{L}$ at $m = 500$, which is greater than the critical width. In full-representation phase, the loss $\mathcal{L}$ is approximately linear in σ_E^2. $\bar{E} = 20, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

$H_n(\epsilon_p) > \bar{E}$. Because $H_n(\epsilon_p)$ decreases with ϵ_p, this condition sets an upper bound on ϵ_p at given $\bar{E}$ and n. For $n = 10000$, the bound is $\epsilon_p < 0.995$ at $\bar{E} = 10$ and $\epsilon_p < 0.70$ at $\bar{E} = 50$, while at $\bar{E} = 1$ any ϵ_p is allowed. We therefore use $\bar{E} = 1$ to probe large ϵ_p (Fig. 22).

The transition from partial to full representation persists only for moderate ϵ_p (Fig. 22). At small ϵ_p, the distribution of s_i at small m is clearly bimodal: frequent features form a band with s_i of order 1, and rare features form a band with nearly vanishing s_i. As ϵ_p increases, the two bands approach each other. At large ϵ_p (e.g., $\epsilon_p = 2$), they merge even at small m: s_i varies continuously with i, and features can no longer be divided into represented and unrepresented groups. In this regime, the partial representation and the partial RP approximation built on it no longer apply. Empirically, the transition remains well defined for $\epsilon_p \lesssim 1$.

Within this range, the biases behave in the same way as the norms (Fig. 23). At small m, the pairs (s_i, b_i) form two clusters, corresponding to represented and unrepresented features; at large m, they merge into one. Unlike the uniform case (Fig. 2b), the represented cluster is elongated rather than point-like, reflecting the p_i-dependent norms and biases (Fig. 6). Along this cluster, b_i decreases as s_i increases, consistent with the perturbative result that more frequent features have smaller norms and larger biases (Eq. (135)). The elongation grows with ϵ_p.

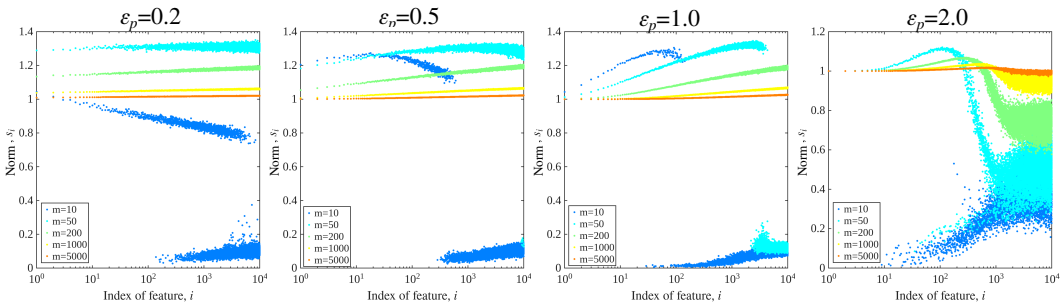


Figure 22: Distributions of s_i at different ϵ_p. At small ϵ_p, the distribution of s_i show clear bimodality at small m; at large ϵ_p, the two bands of s_i merge at small m, and features cannot be divided into represented and unrepresented groups. $\bar{E} = 1, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

C.3.2 Comparison with the perturbation analysis

We compare the perturbation analysis of Sec. A.5 with trained models. We only test the full-representation phase, where the baseline of Sec. A.5.2 (all features represented) applies. The linear-response coefficients in Eq. (131) are evaluated at the uniform solution and depend only on $\bar{E}$, m, n

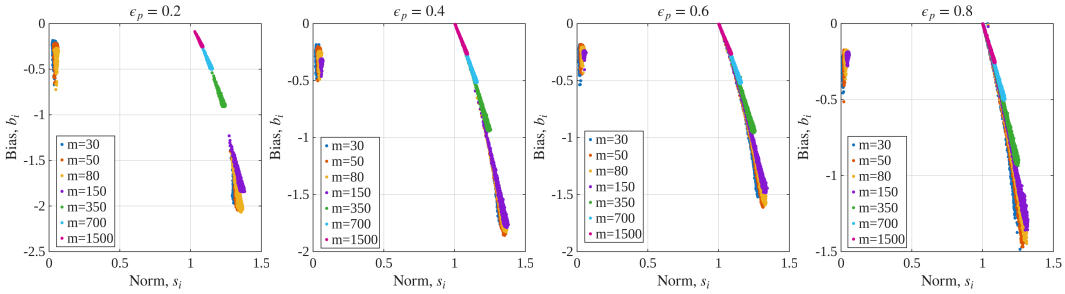


Figure 23: The distribution of s_i and b_i of each feature. Sample points at small m form two clusters while at large m they form one cluster. Different panels are distributions at different ϵ_p. $\bar{E} = 10, n = 10000, \sigma_v = \frac{1}{\sqrt{3}}$.

and σ_v, not on ϵ_p; the data exponent only sets the deviations δp_i. It therefore suffices to test a single ϵ_p. Here we show results at $\epsilon_p = 0.2, \sigma_v = \frac{1}{3}$.

We first test Eq. (117), which relates the overlaps of a feature to its leverage score. Keeping the constant term in Eq. (114) and dropping only the term $h_i \mathrm{Var}_i^V \sigma_j^2$, the sum of squared overlaps of feature i is $1/h_i - 1$. Fig. 24 compares the measured variance of c_{ij} over $j \neq i$ with $(1/h_i - 1)/(n - 1)$ for each feature. The two agree across all features, so the neglected term is small and Eq. (117) holds in trained models under non-uniformity.

We then compare the responses of s_i, b_i and h_i with Eq. (131) (Fig. 25). At all tested widths, the signs of the responses agree with Eq. (135): s_i decreases, while b_i and h_i increase with p_i. The measured h_i are also centered at the baseline value $h_0 = m/n$. The magnitudes, however, are overestimated. The measured responses are also nonlinear in p_i. This is expected, since p_i reaches about $5p_0$, far outside the range $|\delta p_i| \ll p_0$ of the linear expansion (Sec. A.5.3). The perturbation analysis thus captures the direction of the adaptation to non-uniformity, but not its magnitude.

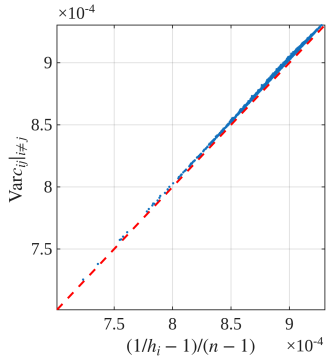


Figure 24: Test of Eq. (117) in the trained model. Each point is one feature. The horizontal axis is $(\frac{1}{h_i} - 1)/(n - 1)$, and the vertical axis is the variance of c_{ij} over $i \neq j$. The red dashed line is $x = y$. $\bar{E} = 10, n = 10000, \sigma_v = \frac{1}{3}, \epsilon_p = 0.2, m = 1000$.

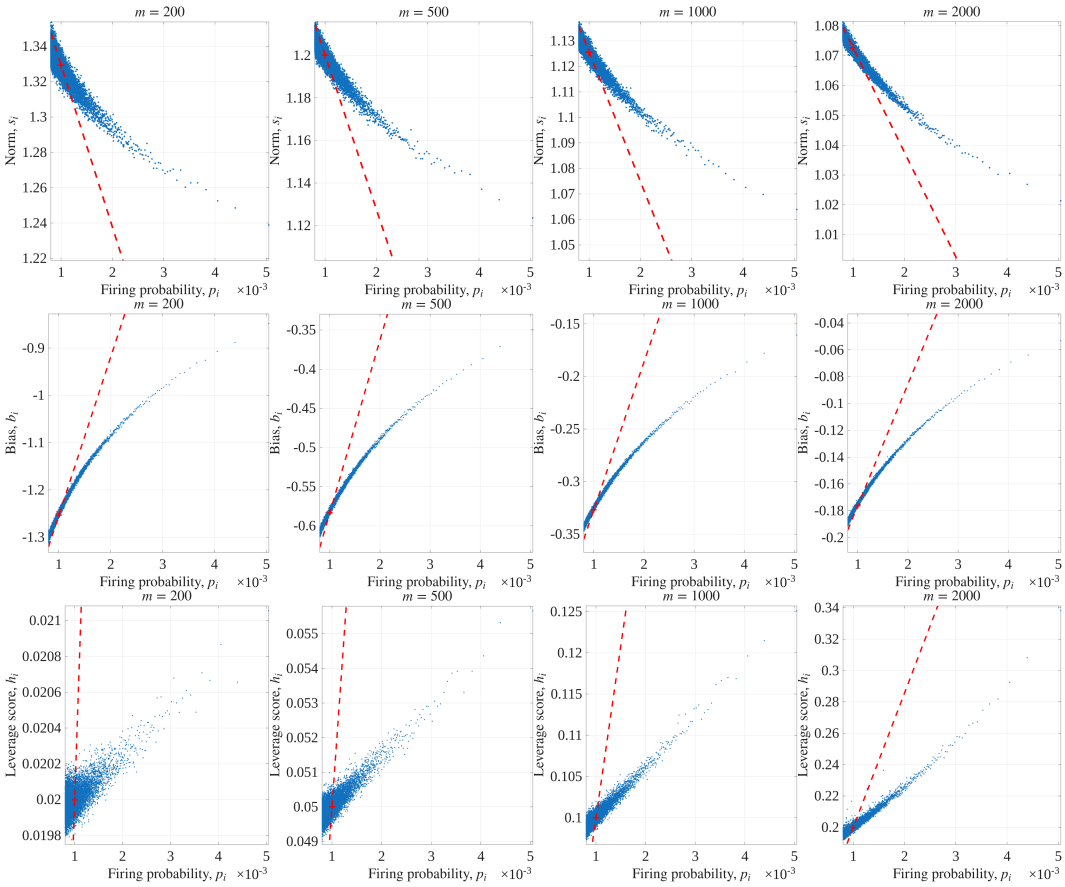


Figure 25: Norms s_i (top row), biases b_i (middle row), and leverage scores h_i (bottom row) as functions of the firing probability p_i at different m. Each point is one feature. The red dashed lines are the linear-response predictions of Eq. (131) around the uniform solution. The predictions capture the signs of the responses but overestimate their magnitudes. $\bar{E} = 10, n = 10000, \sigma_v = \frac{1}{3}, \epsilon_p = 0.2$.