Nonlinear dynamics of localization in neural receptive fields

Leon Lufkin
Yale University
leon.lufkin@yale.edu

Andrew Saxe
Gatsby Unit & SWC, UCL
a.saxe@ucl.ac.uk

Erin Grant
Gatsby Unit & SWC, UCL
erin.grant@ucl.ac.uk

Abstract

Localized receptive fields—neurons that are selective for certain contiguous spatiotemporal features of their input—populate early sensory regions of the mammalian brain. Unsupervised learning algorithms that optimize explicit sparsity or independence criteria replicate features of these localized receptive fields, but fail to explain directly how localization arises through learning without efficient coding, as occurs in early layers of deep neural networks and might occur in early sensory regions of biological systems. We consider an alternative model in which localized receptive fields emerge without explicit top-down efficiency constraints—a feedforward neural network trained on a data model inspired by the structure of natural images. Previous work identified the importance of non-Gaussian statistics to localization in this setting but left open questions about the mechanisms driving dynamical emergence. We address these questions by deriving the effective learning dynamics for a single nonlinear neuron, making precise how higher-order statistical properties of the input data drive emergent localization, and we demonstrate that the predictions of these effective dynamics extend to the many-neuron setting. Our analysis provides an alternative explanation for the ubiquity of localization as resulting from the nonlinear dynamics of learning in neural circuits.[1]

1 Introduction

A striking feature of peripheral responses in the animal nervous system is *localization*—that is, the linear receptive fields of simple-cell neurons often respond to contiguous regions much smaller than their full input domain. In vision, retinal ganglion cells approximate localized center-surround filters that tile the input [Dac+00; Doi+12; KK78], and simple cells downstream in primary visual cortex have localized filters that are selective for spatial frequency and orientation [HW59; HW68; RT95; NS08; WMG11; RSH02; Rin02]. In primary somatosensory cortex, neurons respond to stimulation of restricted regions of skin [Cro+11] and in primary auditory cortex, spatiotemporal receptive fields are typically localized in both time and frequency domains [DWZ03; HDZ08]; see Fig. 1 (left).

By contrast, artificial learning systems do not always learn localized filters. Principal component analysis tends to fit weights that span the entire input signal, as do unregularized autoencoder neural network architectures and restricted Boltzmann machines [Sax+11]. This difference has prompted the search for artificial learning models that can learn localized receptive fields from naturalistic stimuli, the most notable of which are sparse coding [OF96; OF97] and independent component analysis [ICA; BS97; vHvdS98]. Sparse coding, ICA, and related compression methods that produce localized receptive fields from naturalistic data share a top-down approach—they find an efficient representation of the input signal by optimizing an explicit sparsity criterion, or an independence criterion that necessitates sparsity in a critically parameterized regime [Fie99; Sax+11].

---

[1] Code to replicate experiments and figures at https://github.com/leonlufkin/localization

38th Conference on Neural Information Processing Systems (NeurIPS 2024).
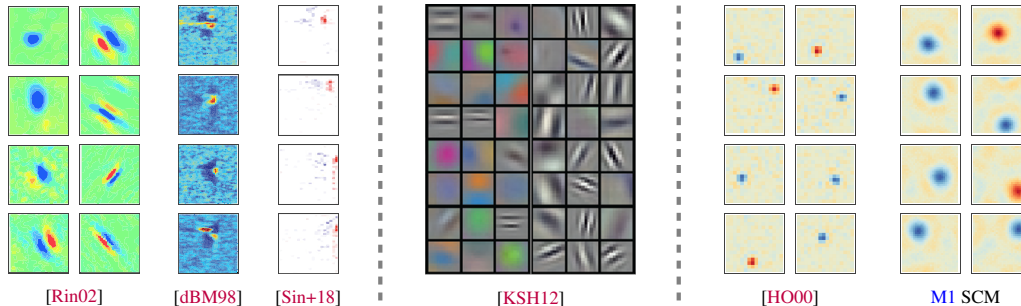
Figure 1: (Left) Localization in spatial receptive fields (RFs) measured from non-human primate (NHP) primary visual cortex [Rin02, Fig. 2] and in spatiotemporal RFs measured from NHP [dBM98, Fig. 2] and ferret [Sin+18, Fig. 2] primary auditory cortex. **(Center)** Half-slice of the localized first-layer kernels of AlexNet trained for ImageNet classification [KSH12]. **(Right)** Localized receptive fields learned from the task of Section 2.3 in 2-D using ICA [HO00] and the soft committee machine (SCM; M1 with fixed second-layer weights) of Section 2.1. *Localization—spatial and/or temporal selectivity—appears across settings, as measured by response maximization in biological systems (left) and by inspecting linear filters in artificial systems (center, right).*

Though sparsity is appealing as a potentially unifying explanation for localization, localization also emerges naturally in networks trained to perform classification tasks without any explicit sparsity regularization [KSH12; ZF13; Yos+15; Sen+18]; see Fig. 1 (center) for an example. Ingrosso and Goldt [IG22] distilled such examples of emergent localization by demonstrating that localized receptive fields emerge in simple feedforward neural networks trained on a data model with properties meant to approximate natural visual input, in particular, locality structure (statistical independence of non-collocated dimensions) and non-Gaussianity (higher-order cumulants are non-null). In simulations, Ingrosso and Goldt [IG22] tie the dynamical emergence of localization to increased tuning to higher-order statistics of the input, and demonstrate that even a single neuron is sufficient to learn a localized receptive field in this setting.

In this work, we build on the demonstration of Ingrosso and Goldt [IG22] with the aim of describing the mechanisms behind the emergence of a localized receptive field in this minimal setting. The higher-order input statistics that drive localization are challenging to analyze with existing tools that exploit implied Gaussianity [Gol+20]. By separating two stages of learning, we are able to derive equations for the effective early-time learning dynamics of the single neuron model that learns a localized receptive field from idealized naturalistic data. Our analytical model identifies a concise description of the higher-order statistics that drive emergence, and we validate both positive and negative predictions of this analytical model via simulations with many neurons; see Fig. 1 (right). These findings suggest an alternative path to account for the ubiquity of localization in early neural responses as resulting from the interaction of the nonlinear dynamics of learning in neural circuits and naturalistic data with higher-order statitistical structure, rather than an explicit efficiency criterion.

2 Modeling approach

We extend the setting of Ingrosso and Goldt [IG22], a minimal example of a neural network that learns localized receptive fields from idealized naturalistic data. We analyze the dynamics of learning in this setting in Section 3 and validate our analytical model with simulations in Section 4.

2.1 Neural network architecture and learning algorithm

We consider a two-layer feedforward neural network with nonlinear activation and scalar output. While simple, this architecture is highly expressive, capable of approximating arbitrary integrable univariate functions with appropriate scaling [Bar93; Pin99], and exhibits rich feature learning dynamics that underlie the performance of models at scale [Woo+20], making this architecture the ongoing subject of theoretical neural network analyses [MMN18; Gol+19; Vei+22]. We denote a two-layer network with N-dimensional input, M hidden units, and one-dimensional scalar output as

2

Model 1 (*many-neuron architecture*).

(M1)
$$\hat{y}(\mathbf{x}) = b^{(2)} + \sum_{m=1}^{M} w_m^{(2)} \sigma \left(b_m^{(1)} + \langle \mathbf{w}_m^{(1)}, \mathbf{x} \rangle \right)$$

---

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a pointwise nonlinearity such as the rectified linear unit (ReLU) or sigmoid function, $\mathbf{w}_m^{(1)} \in \mathbb{R}^N$ and $w_m^{(2)} \in \mathbb{R}$ are learnable weights, $b_m^{(1)}, b^{(2)} \in \mathbb{R}$ are learnable bias terms, and $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner (dot) product on $\mathbb{R}^N$. When the second-layer parameters are fixed, this model is known as a *soft-committee machine* [SCM; SS95], which [IG22] notes learns less noisy receptive fields but exhibits similar localization behavior. The many-neuron architecture in M1 is the focus of our **simulations** (Section 4), but the dynamics of this model are too complex to analyze directly, even for the idealized naturalistic data model considered here. In order to derive **analytical** results (Section 3), we consider the simplest neural network exhibiting the desired localization phenomenon, a single hidden neuron without bias and with rectified linear unit activation, written as

---

Model 2 (*single-neuron architecture*).

(M2)
$$\hat{y}(\mathbf{x}) = \mathrm{ReLU}\left(\langle \mathbf{w}, \mathbf{x} \rangle \right)$$

---

where $\mathrm{ReLU}(x) = \max(x, 0)$, applied pointwise to vectorial input. As Ingrosso and Goldt [IG22] demonstrate, the localized receptive fields learned by the many- and single-neuron models defined in M1 and M2 are qualitatively similar up to spatial translation, which permits us to generalize insights from analyzing the learning dynamics of the single-neuron M2 to the many-neuron M1. For simulations, we initialize the weights and biases as independent draws from an isotropic Gaussian distribution with scaled variance, and train with batch gradient descent with a fixed learning rate on the mean-squared error (MSE) evaluated on input-output pairs from the task; see Section 2.3 for task sampling procedures.

2.2 Stimulus properties

The data model of Ingrosso and Goldt [IG22] can be shown to satisfy three conditions that enable the analysis we give in Section 3. We consider several other data models that share the below properties but differ in generative mechanism in order to probe the effect of these properties on localization. In particular, we consider data $\mathbf{X}$ sampled from distributions p on $\mathbb{R}^N$ satisfying the following:

Stimulus properties 1–3 (*idealization of natural images*).

(S1) (Positional) weak dependence: for any fixed $\rho \in (0, 1)$, as $N \to \infty$,

$$\alpha(N) \triangleq \sup_{A \subseteq \mathbb{R}, B \subseteq \mathbb{R}^{(1-\rho)N}} |\, \mathbb{P}(X_1 \in A, X_{>\rho N} \in B) - \mathbb{P}(X_1 \in A)\, \mathbb{P}(X_{>\rho N} \in B)| \to 0 \,,$$

(S2) Translation invariance: $p(\mathbf{X} = \mathbf{x}) = p(\mathbf{X} = \mathcal{S}\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^N$, where $\mathcal{S}$ is the circular shift operator, and

(S3) Sign symmetry: $p(\mathbf{X} = \mathbf{x}) = p(\mathbf{X} = -\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^N$.

Properties S1 and S2 are defining characteristics of natural image data [HHH09]. Property S3 can also be seen to hold for natural images after centering and is convenient analytically because it implies that $\mathbb{E}[\mathbf{X}] = 0$. Property S1 assumes that p is implicitly parameterized by N in order to state that the statistical dependence between entries of $\mathbf{X}$ vanishes as their separation increases.[2]

We denote the covariance of $\mathbf{X}$ by $\Sigma \triangleq \mathrm{Cov}[\mathbf{X}]$, the square of principal-diagonal entries (the variance of each entry of $\mathbf{X}$) by σ^2, and the i-th row by σ_i. Weak dependence (S1) implies that entries far from the principal diagonal of Σ will be 0, while translation-invariance (S2) implies that Σ is circulant (*i.e.*, entries along each diagonal are equal) and thus identifiable by a single row; see Fig. 2 (center).

---

[2]The weak dependence condition in S1 is based on strong α-mixing, a notion first introduced by [Ros56] to obtain a generalization of the central limit theorem, which we employ later on. We choose α-mixing because it is easy to interpret and verify, but alternative definitions of weak dependence [*e.g.*, Bar+08] can be substituted.

3
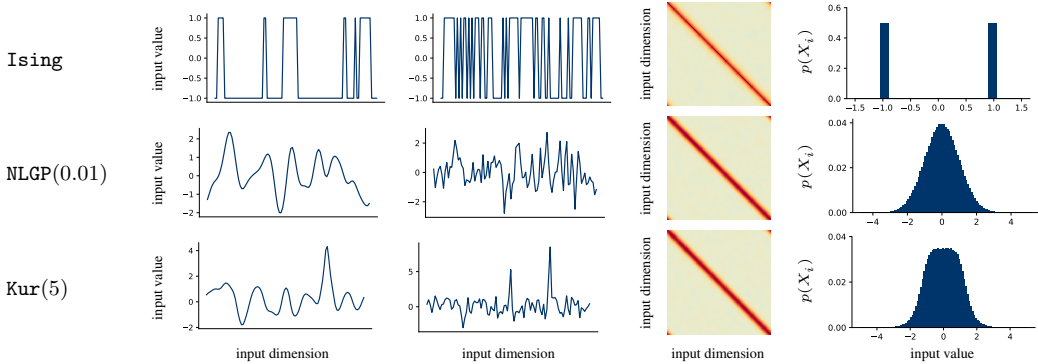
Figure 2: From left: Long- and short-lengthscale samples $\mathbf{x}$, covariances Σ for one lengthscale, and marginals $p(X_i)$ for the data models described in Section 2.3: Ising (with $J = 1.2, 0.3$ for left, right samples), the nonlinear Gaussian process [NLGP; IG22], and the controllable kurtosis model, Kur (with $\xi = 5, 1$ for left, right samples). *Each model generates samples centered about zero and with covariances that can be constrained to be similar, but with differing higher-order statistics, as can be seen from the dimension-wise marginals.*

2.3 Lengthscale discrimination task

Ingrosso and Goldt [IG22] develop a minimal task for which localization emerges in a feedforward neural network: binary discrimination between inputs from two distributions that differ in the lengthscale of the correlations between their entries. This lengthscale discrimination task can be seen as a pretext task for self-supervised learning [KZB19; Che+20] of representations [*cf.* unsupervised: OF96; BS97]. More precisely, we generate data $(\mathbf{X}, Y)$ for supervised training according to

$$\mathbf{X} \mid Y = y \sim p(\mathbf{X}; \Sigma_y) \,, \tag{1}$$

where p is to be defined, Σ_y are distinct covariance matrices for each y, and we sample Y uniformly among a set of increasing *lengthscale correlation classes* $y \in \{0, 1, \ldots\}$, which correspond to the strength of correlation between distant positions. For instance, in the case of two classes ($y = 0, 1$), we take Σ_0 to be closer to $\sigma^2 \mathbb{I}_N$ than Σ_1, where $\mathbb{I}_N$ is the $N \times N$ identity matrix and σ is a fixed value. This construction isolates, via distinct covariance matrices per class, the second-order statistics, which we will see below enter into the learning dynamics separately from other properties of $p(\mathbf{X})$, including, most critically, the implied marginal distributions, $p(X_i)$.

Ising. The first distribution we consider is the one-dimensional Ising model. It is of interest as a distribution that satisfies S1 to S3 with marginals $p(X_i)$ with extreme support on $\{\pm 1\}$, making it the distribution that promotes localization most strongly, as we will see in Section 3. In the absence of an external field, the Ising distribution is

$$p_{\texttt{Ising}}(\mathbf{X} = \mathbf{x}) = p_{\texttt{Ising}}(X_1 = x_1, \ldots, X_N = x_N) = e^{-\sum_{i=1}^N J x_i x_{i+1}}/\mathcal{Z}, \tag{2}$$

where J is a chosen pairwise interaction strength, $\mathcal{Z}$ is the normalizing constant, and we enforce a periodic boundary constraint via $x_{N+1} \equiv x_1$. As J increases, the lengthscale of the correlations in $\mathbf{X}$ also increases. For simulations, we sample from $p_{\texttt{Ising}}$ using a Gibbs sampler [GG84]. Discrimination tasks in the simulations in Section 4 use $J_1 = 0.7$ (for $y = 1$) and $J_0 = 0.3$ (for $y = 0$).

NLGP(g). We also consider the data model used in Ingrosso and Goldt [IG22], the nonlinear Gaussian process (NLGP), which enables one to interpolate between distributions that do and do not yield localization via a single parameter, g. A sample $\mathbf{X} \mid Y = y$ from the NLGP is constructed by first sampling a Gaussian $\mathbf{Z} \mid Y = y \sim \mathcal{N}(0, \tilde{\Sigma}_y)$ and then transforming it via

$$X_i \triangleq \operatorname{erf}(g Z_i)/\mathcal{Z}(g) \qquad 1 \leq i \leq N, \tag{3}$$

where erf is the Gauss error function, $\mathcal{Z}$ is a normalization constant to ensure that the variances of X_i and Z_i are the same, and $\tilde{\Sigma}_y$ is a covariance matrix for $\mathbf{Z}$, where we use $(\tilde{\Sigma}_y)_{ij} = \exp(-(i-j)^2/\xi^2)$ for a lengthscale parameter ξ [IG22]. If $g \approx 0$ (where localization is *not* observed), $g Z_i$ will tend to lie in the linear regime of erf, so Z_i will be untransformed, *i.e.*, $\mathbf{X}$ is Gaussian. However, as $g \to \infty$ (where localization *is* observed), $g Z_i$ will tend to saturate erf, so X_i will have support on $\{\pm 1\}$.

4

Kur(k). The final family we consider is chosen to give us flexibility over the kurtosis κ of the marginals $p(X_i)$. In the Ising model, the *excess* kurtosis ($\kappa - 3$) of the marginals is fixed at -2, while in NLGP(g), it varies from -2 to 0. This family allows us to vary the excess kurtosis from negative through positive values. We sample $\mathbf{X} \mid Y = y$ from this family via inverse transform sampling to vary the marginals while enforcing dependence via Gaussian copulas. More concretely, we sample $\mathbf{Z} \mid Y = y \sim \mathcal{N}(0, \tilde{\Sigma}_y)$ and then transform it via

$$X_i \triangleq f^{-1}(\Phi(Z_i/\tilde{\sigma}))/\mathcal{Z}, \qquad 1 \leq i \leq N, \tag{4}$$

where $\tilde{\sigma}$ is the standard deviation of Z_i, Φ is the standard Gaussian cumulative distribution function (CDF), f is the CDF of the desired marginal distribution for X_i, and $\mathcal{Z}$ is a normalization constant, which we compute numerically. We define $\tilde{\Sigma}_y$ as for NLGP. We choose f to be the generalized *algebraic sigmoid* function (see Appendix A.2) for $k > 0$ to make use of its tractable inverse, simplifying the procedure in Eq. (4). We denote the corresponding distribution by Kur(k). Though we are able to continuously vary excess kurtosis, we lack an explicit form; however, numerical computation shows that for $k \lessapprox 5.8$, excess kurtosis is positive, while for $k \gtrapprox 5.9$, it is negative.

3 Theoretical results

We derive an analytical model for the localization dynamics of the single-neuron architecture in M2. This result establishes necessary and sufficient conditions for localization under Properties S1 to S3 for the minimal case of a binary response, *i.e.*, $y = 0, 1$. The conditions for localization in the single-neuron architecture in M2 are demonstrated in Section 4 to also hold empirically for the many-neuron architecture in M1. Further, we use this model to derive a negative prediction about localization, that the architectures in M1 and M2 fail to learn a localized receptive field on elliptical distributions despite their non-Gaussian—in particular, significantly positive kurtosis—statistics [*cf.* positive kurtosis as an objective or diagnostic for localization, HO00; IG22].

3.1 An analytical model for the dynamics of localization in a single neuron

Previous approaches to obtain analytical dynamics in the architectures in M1 and M2 have studied the gradient flow under the assumption that the preactivation $\langle \mathbf{w}, \mathbf{X} \rangle$ is approximately Gaussian [Gol+20; Ger+20; Gol+22], but this assumption fails to capture the propagation of higher-order statistics through a neural network that promotes localization [IG22]. Happily, the idealized visual input setting set out in S1 to S3 permits us some simplification. In particular, the translation-invariance of the data $\mathbf{X}$ under S2 and the architecture of M2 allow us to work with the marginal distributions of each input dimension, X_i rather than the full joint distribution of $\mathbf{X}$.

We now give the analytical simplifications that allow us to derive an analytical model for the localization dynamics of the single neuron architecture in M2, namely two assumptions on $\mathbf{X} \mid X_i$ for all $i \in \{1, \ldots, N\}$ as a well as a mild condition on the weights that is satisfied at initialization. These are, where σ_i^y to denotes the i-th row of Σ_y:

Analytical simplifications 1–3 (*early-time, limiting dynamics*).

(A1) $\mathbb{E}[\mathbf{X} \mid X_i = x_i, Y = y] = x_i \sigma_i^y$, *i.e.*, the conditional mean scales linearly with x_i.

(A2) $\text{Cov}[\mathbf{X} \mid X_i = x_i, Y = y] = \Sigma_y - \sigma_i^y \sigma_i^{y\top}$, *i.e.*, the conditional covariance is smaller near i, but independent of the exact value of x_i.

(A3) Lindeberg's condition holds for the sequence $w_1 X_1, \ldots, w_N X_N \mid X_i = x_i$ as $N \to \infty$ for all x_i.

Our motivation for Assumptions A1 to A3 is that they replicate the kurtosis of the marginal distributions X_i (discussed further below) of two important and distinct limiting cases where localization does and does not appear, respectively: when $\mathbf{X}$ has support on the vertices of the hypercube $\{\pm 1\}^N$ (satisfied by Ising for any J), and when $\mathbf{X}$ is Gaussian (satisfied by NLGP with $g \approx 0$).

The gradient flow in Lemma 3.1 also relies on Assumption A3 that Lindeberg's condition holds for the sequence $w_i X_i$, which ensures that no single term $w_i X_i$ in the sequence can dominate. If this holds,

5

then we can conclude that $\langle \mathbf{w}, \mathbf{X} \rangle \mid X_i$ is approximately Gaussian. As we discuss in Appendix C.2, this is almost always satisfied for a Gaussian initialization of $\mathbf{w}$, and for slight deviations therefrom, and is satisfied by the settings of Ingrosso and Goldt [IG22]. Using this fact, we obtain an explicit form for the gradient flow early in training, stated in Lemma 3.1.

Lemma 3.1. *Under Assumptions A1 and A2, the gradient flow for the single ReLU neuron in M2 early in training with $y = 0, 1$ trained using MSE loss is*

$$\frac{2}{\tau} \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \varphi \left(\frac{\Sigma_1 \mathbf{w}}{\sqrt{\langle \Sigma_1 \mathbf{w}, \mathbf{w} \rangle}} \right) - (\Sigma_0 + \Sigma_1)\mathbf{w} + o_N(1), \tag{5}$$

where $o_N(1)$ vanishes as $N \to \infty$, and where $\varphi : (-1, 1) \to \mathbb{R}$ is defined as

$$\varphi(a) = \mathbb{E}_{X_1 \mid Y=1} \left[X_1 \operatorname{erf} \left(X_1 \operatorname{alg}^{-1}(a)/\sqrt{2} \right) \right] \tag{6}$$

and $\operatorname{alg}^{-1}(x) = x/\sqrt{1 - x^2}$, the inverse of the algebraic sigmoid function $\operatorname{alg}(x) = x/\sqrt{1 + x^2}$.

Lemma 3.1 reduces the study of higher-order statistics to the marginal distributions, X_1, where, by translation invariance, all marginals have the same distribution, so we refer to X_1 without loss of generality. While Lemma 3.1 technically only holds early in training and breaks down if $\mathbf{w}$ becomes localized due to violation of A3, the gradient flow in Eq. (5) holds sufficiently long to detect the emergence of localization in the weights $\mathbf{w}$. In particular, numerically integrating Eq. (5) yields localized weights $\mathbf{w}$ as $t \to \infty$. Moreover, the location of the peak of final weights from Eq. (5) corresponds closely to the actual peak of the weight, when we observe localization; see Section 4.2 for empirical validation of this fact. The primary difference observed is that the localized bump from Eq. (5) is less peaked than when computed exactly; see Fig. 3 for a comparison between experimentally observed localized receptive fields and theoretical predictions.

3.2 Necessary and sufficient conditions for emergent localization

To establish an exact threshold at which localization emerges requires solving Eq. (5), which is not possible exactly for general nonlinear differential equations.[3] Nevertheless, the form of Eq. (5) reveals that localization is driven solely by the first term. Indeed, the second term depends only on the second-order statistics of the data, and so can be held fixed as $\mathbf{X}$ is varied from a distribution that induces localization to one that does not. Secondly, one can see that the first term in Eq. (5) does not change as $\mathbf{w}$ is scaled, in contrast to the second term. As such, the second term in Eq. (5) serves to constrain the *scale* of $\mathbf{w}$, distinct from localization, while the first is primarily concerned with the *shape* of $\mathbf{w}$, and thus localization. This further motivates the first term, and thus φ, which we will refer to as the *amplifier* and which itself depends on properties of the data distribution $p(\mathbf{X})$, as a focus of study for understanding localization.

We present an analysis of φ in Appendix B.1 that reveals the role of the marginal distribution of the data in driving localization. For each marginal, $\varphi(a) \approx (\sqrt{2/\pi})a$ for $a \approx 0$. For larger a, φ depends more strongly on the data distribution and can be super-linear (sub-linear), *i.e.*, greater (smaller) than $(\sqrt{2/\pi})a$. Super-linear φ encourage entries in $\mathbf{w}$ that are large in some neighborhood to grow faster than those that are smaller, yielding localization. Linear and sub-linear φ are the opposite, encouraging oscillatory or flat weights by suppressing neighborhoods in $\mathbf{w}$. However, super- and sub-linearity may not hold uniformly, as φ can be both over its domain (see Fig. 3, bottom row, black line). As an approximation, we consider a third-order Taylor expansion (red lines in Fig. 3, second column), which reveals that for the canonical setting of $\sigma^2 = 1$, *negative excess kurtosis of the marginals yields super-linearity*, while *positive excess kurtosis yields sub-linearity*; see Appendix B.1. This leads us to the following claim, which is validated by our simulations in Section 4:

Claim 3.2. *For sufficiently large N, if the data $\mathbf{X} \in \mathbb{R}^N$ satisfies conditions S1 to S3 and has marginal distributions with sufficiently negative excess kurtosis, then Model M2 will learn localized receptive fields. Conversely, if the excess kurtosis is sufficiently positive, it will not.*

As a minimal positive example, the distribution with the most negative excess kurtosis is the symmetric Bernoulli, with a value of -2. In our setting, this corresponds to a data vector $\mathbf{X}$ with support on

---

[3] We discuss a partial differential equation limit that faces similar intractabilities in Appendix B.3.

6
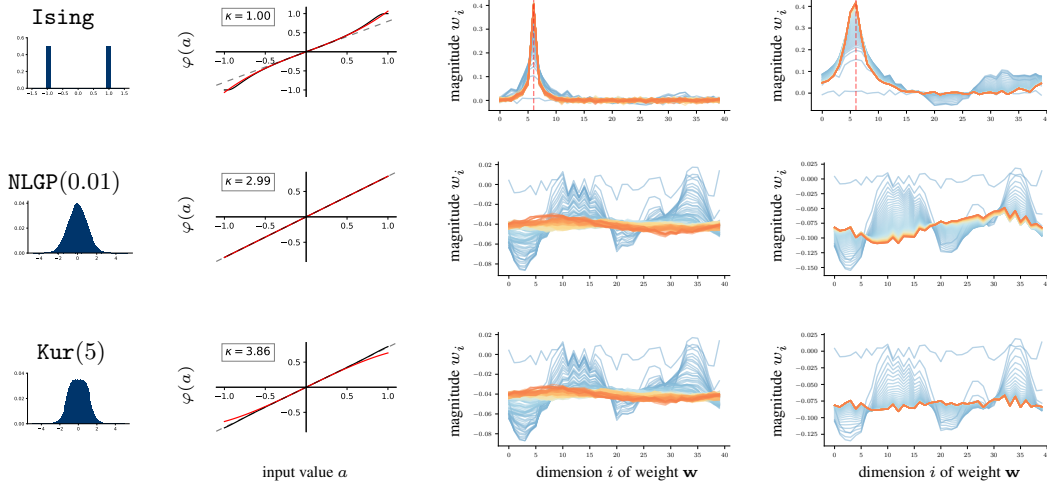
Figure 3: From left and for the same `Ising`, `NLGP`, and `Kur` data models as in Fig. 2: the marginals $p(X_i)$, the amplifier φ defined in Theorem 3.1 and kurtosis κ, and the evolution of simulated receptive fields for the single-neuron model (M2) trained on its data, and lastly the receptive field given by numerically integrating Eq. (5) with φ expanded to a third-order Taylor approximation for the same data; training or evolution time is indicated by line color (blue for early-time; red for late-time). *See Section 4.1 for exposition.*

the vertices of the hypercube, $\{\pm 1\}^N$. As mentioned above, it can be seen from the law of total covariance combined with sign-symmetry that A1 and A2 hold exactly. Note that φ is the same for all such distributions, which leads us to Claim 3.2 that *any* distribution satisfying conditions S1 to S3 whose marginals are maximally concentrated will induce a localized receptive field in M2. Importantly, this claim includes the limiting case of Ingrosso and Goldt [IG22], $g \to \infty$ in `NLGP`. It also includes the Ising model as another example, corroborating an observation for restricted Boltzmann machines [Har+20] that Ising data induces localization in a learning model. These claims are validated for the single-neuron model in Fig. 3 and in Section 4 for the many-neuron model.

3.3   Case study: Elliptical distributions fail to produce localization

Above, we assume weak dependence (S1) as it enables a focus on how the marginals control localization. As a first investigation into departures from this regime, we consider data $\mathbf{X}$ sampled from an elliptical distribution, where weak dependence may not hold. We specialize the definition of an elliptical distribution [Fra04] to our setting of multiple class labels and sign-symmetry:

Def 1. *Samples $\mathbf{X} \in \mathbb{R}^N$ satisfy an* elliptical distribution *if we can write $\mathbf{X} \mid Y = y \stackrel{(d)}{=} R_y \Lambda_y \mathbf{U}_y$ where R_y is a nonnegative random variable, $\Lambda_y \in \mathbb{R}^{N \times D}$ is such that $\Lambda_y \Lambda_y^\top = \Sigma_y$, and $\mathbf{U}_y$ is independent of R_y and uniformly distributed on the D-dimensional sphere.*

The class of elliptical distributions is broad, imposing only the constraint that the contours of the density be ellipses; the multivariate Gaussian and Student-t distributions are examples. As such, they can vary greatly in measures of non-Gaussianity, including kurtosis, while maintaining enough structure for analytical convenience. Proposition 3.3 states that training on elliptical data *prevents* localization in the single ReLU neuron model.

Proposition 3.3. *Assume the data $\mathbf{X}$ are sign-symmetric (S3), translation-invariant (S2), and follow an elliptical distribution such that the MSE on the task in Section 2.3 is always finite. If Σ_0, Σ_1 are such that the ratio of their i-th eigenvalues, $\lambda_i(\Sigma_0)/\lambda_i(\Sigma_1)$, assumes a particular value for at most two distinct i, then the steady states of the weight of M2 are sinusoids, i.e., not localized.*

The condition on the number i such that the ratio of the i-th eigenvalues are the same constrains the number of Fourier components that can be non-zero in the steady state of $\mathbf{w}$. While opaque, this requirement seems to always hold in practice, as even slight increases in length-scale correlation can dramatically change the spectrum of Σ_y.

7
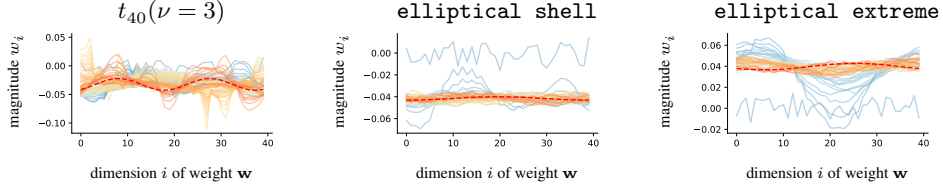
Figure 4: Evolution of receptive fields learned by the single-neuron model (M2), along with sinusoids fit to final states (red dashes) when trained on data from three elliptical distributions: $t_{40}(\nu = 3)$ (**left**), the surface of an ellipse (**middle**), and a custom elliptical distribution that places its mass near the outside of an ellipse (**right**). In all cases, the learned receptive field is oscillatory (a sinusoid), as predicted by Proposition 3.3. The ℓ_2 distances between the fitted oscillatory weights and empirical RFs, as a ratio of the ℓ_2 norm of the empirical RFs, are (left) 9.77%, (center) 3.75%, and (right) 4.14%. *See Section 4.3 for exposition.*

The proposition is surprising because it reveals that the kurtosis of the preactivation is not an appropriate metric for explaining localization. Consider the example of the N-dimensional Student-t distribution with ν degrees of freedom, $t_N(\nu)$. If $\mathbf{X} \sim t_N(\nu)$, then $\langle w, \mathbf{X} \rangle \sim t_1(\nu)$. Note the kurtosis of $t_1(\nu)$ is non-zero, and can be very large or even infinite for small ν. This prediction is validated in Section 4.3. The condition also reveals that not all symmetries in the data (here, elliptical symmetry) induce structure in the trained model weights, if localization is to be seen as a sparsity more structured than oscillatory weights [*cf.* God+23]; indeed, translational symmetry (S2) is more relevant for localization than elliptical symmetry.

4   Experimental results

We describe experiments to validate the generalizability of the analytical results from Section 3. We run all experiments on a single CPU machine locally or on a compute cluster. Since all datasets are procedurally generated, training depends on both the model architecture and the complexity of sampling the data, but is between 10 and 60 minutes for any single simulation run.

4.1   Validating Claim 3.2 with positive and negative predictions

In Fig. 5, we validate Claim 3.2 first via the single-neuron model (M2) with 30 initial conditions trained across a range of excess kurtoses for the NLGP(g) and Kur(k) data models. We use the inverse participation ratio (IPR), defined in Appendix A.3. This measure, also used by Ingrosso and Goldt [IG22], is large when proportionally few weight dimensions "participate" (have large magnitude), and small when weight dimension magnitudes are more uniform. We see that when g and k assume values that yield a negative excess kurtosis, IPR is close to its maximum of 1.0, suggesting the weights are localized; if the excess kurtosis is positive, IPR is nearly zero, suggesting the weights are *not* localized. The IPR is extremely consistent across random initializations, suggesting that localization is determined by data statistics and not initialization. The trend in IPR *vs.* excess kurtosis is very similar between the NLGP(g) and Kur(k) data models, demonstrating that excess kurtosis is a primary driver of localization and localization is largely independent from other properties of the data distribution.
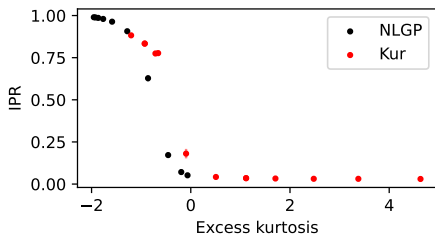
Figure 5: IPR *vs.* excess kurtosis for NLGP and Kur data models, with mean and std. dev. across 30 re-initializations for the single-neuron model (M2); error bars are small and may not be visible.

Figure 3 further validates Claim 3.2 with specific examples. We maintain constant initial conditions for our model and train on the Ising, NLGP($g = 0.01$), and Kur($k = 5$) data models. The marginals of the Ising model have an excess kurtosis of -2, the smallest possible value for any distribution. As a result, we see that the amplifier φ for Ising (top left) is super-linear (the dark line exceeds the dashed light line for larger a), which drives localization via its role in Eq. (5). Integrating Eq. (5) with φ expanded via a third-order Taylor approximation (red line) yields a similar localized receptive field to that from simulation (two right panels), validating this approximation.

8
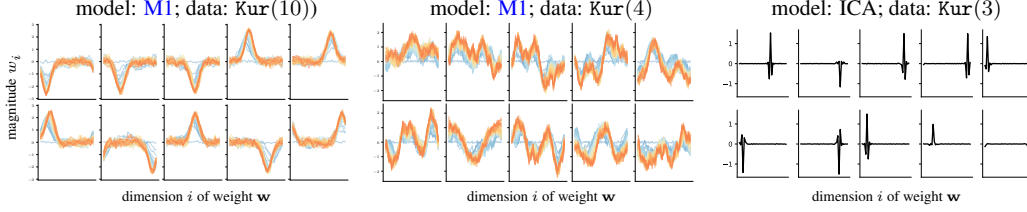
Figure 6: (**Left**, **Center**) Receptive fields learned by many-neuron (M1) soft committee machines (second-layer weights fixed at $\frac{1}{K}$) trained on the Kur(10) and Kur(4) datasets, respectively. The models had $N = 40$ input units, $K = 10$ hidden units, and an initialization variance of 0.1. (**Right**) A random subset of 10 components from the 40 learned by the FastICA algorithm from scikit-learn [Hyv99; Ped+11] on the Kur(3) dataset with length-scale correlation values of $\xi_0 = 1$ and $\xi_1 = 3$. *See Section 4.4 for exposition.*

For the remaining distributions (middle and bottom rows) that elicit oscillatory (sinusoidal) weights, Claim 3.2 is validated due to their positive excess kurtosis. The dynamical steady state (far right) assumes a more negative value than in the simulation (to the left), a difference that is the result of deviations of our *early-time* gradient flow in Eq. (5), but these deviations remain mild enough nevertheless to recover the qualitative structure of the learned receptive field.

4.2 Validating Eq. (5) with localization position prediction

The simulated and integrated receptive fields in Fig. 3 demonstrate that our analytical model is able to meaningfully reproduce localization in receptive fields from neural network training. For the Ising model, we see that the integration even has a peak in the exact same position as the simulation (at index $i = 6$), suggesting precision in our approximation. Indeed, we simulated the condition in Fig. 3 for the Ising model under 28 different initial conditions (weight initializations), and found that in 26 of them (93%), the peaks of the integrated and simulated receptive fields matched exactly. In the two cases where the peaks differed, they did so substantially (see Fig. 8 for an example).

4.3 Validating Proposition 3.3: Elliptical distributions fail to localize

Proposition 3.3 claims that the single-neuron model (M2) trained on elliptical data will yield sinusoidal receptive fields, subject to a condition on the spectra of Σ_0 and Σ_1. We verify this claim in Fig. 4 with three distinct elliptical distributions. The first, $t_{40}(\nu = 3)$, gives preactivations $\langle \mathbf{w}, \mathbf{X} \rangle$ that have *infinite* kurtosis, yet our theory predicts the final receptive field will be sinusoidal. This is confirmed in Fig. 4, where the learned receptive field is indeed a sinusoid with period 1 and intercept at zero.

We also consider data sampled from the surface of an ellipse, which is done by fixing $R_y \equiv 1$ in Definition 1. Here, we observe that the learned receptive field is a near-constant function at -0.04 (note that $\cos(2\pi \cdot 0 \cdot x) \equiv 1$ is a sinusoid, allowing nonzero intercepts and constant functions). Finally, we consider an unconventional elliptical distribution where the density of R is given by $p_R(r) = (4e^{2r+4})/(e^{2r} + e^4)^2 \cdot \mathbb{1}(r \geq 2)$. This particular density places most of its mass near $r = 2$ before rapidly falling off, imposing a minimum norm on $\mathbf{X}$ and pushing support near the surface of an ellipse. This distribution, too, yields an oscillatory steady state, as shown in Fig. 4 (right). We confirm our visual observations by fitting sinusoids to the final receptive fields and see the relative errors are quite low.

4.4 Extensions to many-neuron model and ICA

All of our analysis thus far has concerned single-neuron models with ReLU activation and without hidden-to-output or bias terms, assumptions which were made to make our analysis tractable. Here, we depart from that regime by considering the SCM and the full two-layer network (Model M1). In Fig. 6 (left) and (center), we train a SCM with 10 hidden units and sigmoid activation on the Kur(10) and Kur(4) datasets, which have excess kurtoses of -0.93 and 3.28, respectively. So, based on our single-neuron analysis, we *do* and *do not* expect to see localization for these distributions. Indeed, this is precisely what we observe in Fig. 6, where the receptive fields are sharply localized for the former distribution, while they look like low-frequency oscillations for the latter.

9
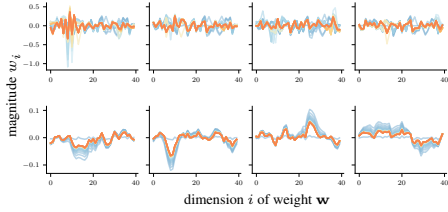
Figure 7: Receptive fields learned by the many-neuron model (M1) with learnable second-layer weights, $N = 40$, $K = 10$. (**Top**) A random subset of 4 receptive fields from a model with sigmoid activation, trained on `Kur(4)` (positive excess kurtosis of 3.28). As predicted by Claim 3.2, the receptive fields are *not* localized, and appear as high-frequency oscillations. (**Bottom**) A random subset of 4 receptive fields from a model with ReLU activation, trained on `Kur(30)` (negative excess kurtosis of -1.17). Receptive fields are localized (**left three**) or exhibit low-frequency oscillations (**right**). *See Section 4.4 for exposition.*

In Fig. 7, we train many-neuron models with $N = 40$ input units and $K = 10$ hidden units, where all weights are learnable. In general, adding flexibility in the second layer leads to more varied structure in the first layer. We train on `Kur(4)` (top), which has an excess kurtosis of 3.28, and `Kur(30)` (bottom), which has an excess kurtosis of -1.17. The receptive fields from the former are not localized, as in the single-neuron model; however, they appear more like high-frequency oscillations than low-frequency sinusoids. For `Kur(30)`, where we expect localization, we see that the first three receptive fields exhibit localization, but less so than for a single neuron. Importantly, not all receptive fields are localized, a result of a variable second-layer weight effectively changing the variance σ^2 in the third-derivative term in Lemma B.1.

We further compare these predictions against ICA, another framework that has been used to model receptive fields in visual cortex. We train on the `Kur(3)` dataset, which has marginals with excess kurtosis 7.66, fitting 10 components using the FastICA implementation from scikit-learn [**sc**; HO00]. We observe in Fig. 6 (right) that we learn localized receptive fields; this contrasts our neural network models, which require negative excess kurtosis. This stems from ICA's objective to maximize non-Gaussianity, regardless of how specifically it is done. The sign of the excess kurtosis is irrelevant, so long as it is nonzero. This deviation between our analytical model and ICA is an interesting avenue for future study, perhaps by validation with natural images.

5   Conclusions

We derive effective learning dynamics for the minimal example of emergent localization in a neural receptive field given by Ingrosso and Goldt [IG22]. The analytical approach we take relies on the assumption that the *conditional* preactivation is Gaussian, a refinement of previous work that assumes Gaussianity of the unconditioned preactivation as asserted by the *Gaussian equivalence property* targeted by Ingrosso and Goldt [IG22]. This approach may prove extensible beyond our specialized setting and may enable further analysis of how statistics of an input task drive emergent structure in neural network learning.

Emergence as an alternative mechanism to top-down constraints like sparsity is in line with recent work that reformulates data-distributional properties as a driver for complex behavior [Cha+22]. Via these analytical effective dynamics, we observe that specific data properties—the covariance structure and the marginals—shape localization in neural receptive fields. Though we cannot capture dynamical interactions between neurons that may shape receptive fields in other settings with the single-neuron analytical model, our empirical validations with many neurons suggest that these interactions do not, in fact, play a significant role in shaping localization [*cf.* Har+20].

The data model we consider is a simplified abstraction of the task faced by early sensory systems, and, as a consequence, we do not yet capture certain features of receptive fields that are observed in early sensory systems. In particular, we do not observe orientation nor phase selectivity, features of simple-cell receptive fields in early sensory cortices and in artificial neural networks that can be seen in a subset of receptive fields in Fig. 1 (left and center, respectively). To capture orientation selectivity, it may be fruitful to follow the approach of Karklin and Simoncelli [KS11], who tie orientation selectivity in a population-based efficient-coding framework to the presence of noise. Furthermore, on-center-off-surround-filtering input data, including the idealized data, gives receptive fields with subfields in our simulations, but is difficult to analyze. Lastly, we do not yet look at the distribution of receptive field shapes and do not validate against other models of receptive field learning beyond a brief comparison with ICA [*cf.* Sax+11], but these are exciting avenues for future work.

10

Acknowledgements

This work was supported by a Schmidt Science Polymath Award to A.S., and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3850). A.S. is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains program.

References

[Bar+08]   J.-M. Bardet, P. Doukhan, G. Lang, and N. Ragache. "Dependent Lindeberg central limit theorem and some applications". In: *ESAIM: Probability and Statistics* 12 (2008), pp. 154–172 (cit. on p. 3).

[Bar93]   A. R. Barron. "Universal approximation bounds for superpositions of a sigmoidal function". In: *IEEE Transactions on Information theory* 39.3 (1993), pp. 930–945 (cit. on p. 2).

[Bra07]   R. C. Bradley. "Introduction to strong mixing conditions". In: *(No Title)* (2007) (cit. on p. 17).

[BS97]   A. J. Bell and T. J. Sejnowski. "The "Independent Components" of Natural Scenes Are Edge Filters". In: *Vision Research* 37.23 (Dec. 1, 1997), pp. 3327–3338. URL: https://www.sciencedirect.com/science/article/pii/S0042698997001211 (cit. on pp. 1, 4).

[Cha+22]   S. C. Y. Chan, A. Santoro, A. K. Lampinen, J. X. Wang, A. Singh, P. H. Richemond, J. McClelland, and F. Hill. "Data Distributional Properties Drive Emergent In-Context Learning in Transformers". In: *Advances in Neural Information Processing Systems*. Conference on Neural Information Processing Systems. Oct. 10, 2022. arXiv: 2205.05055 [cs]. URL: http://arxiv.org/abs/2205.05055 (cit. on p. 10).

[Che+20]   T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. Pmlr, Nov. 21, 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html (cit. on p. 4).

[Cro+11]   S. Crochet, J. F. Poulet, Y. Kremer, and C. C. Petersen. "Synaptic Mechanisms Underlying Sparse Coding of Active Touch". In: *Neuron* 69.6 (Mar. 2011), pp. 1160–1175. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627311001206 (cit. on p. 1).

[Dac+00]   D. Dacey, O. S. Packer, L. Diller, D. Brainard, B. Peterson, and B. Lee. "Center Surround Receptive Field Structure of Cone Bipolar Cells in Primate Retina". In: *Vision Research* 40.14 (June 1, 2000), pp. 1801–1811. URL: https://www.sciencedirect.com/science/article/pii/S0042698900000390 (cit. on p. 1).

[dBM98]   R. C. deCharms, D. T. Blake, and M. M. Merzenich. "Optimizing Sound Features for Cortical Neurons". In: *Science* 280.5368 (1998), pp. 1439–1444. eprint: https://www.science.org/doi/pdf/10.1126/science.280.5368.1439. URL: https://www.science.org/doi/abs/10.1126/science.280.5368.1439 (cit. on p. 2).

[Doi+12]   E. Doi, J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner, T. A. Machado, L. H. Jepson, K. Mathieson, D. E. Gunning, A. M. Litke, L. Paninski, E. J. Chichilnisky, and E. P. Simoncelli. "Efficient Coding of Spatial Information in the Primate Retina". In: *The Journal of Neuroscience* 32.46 (Nov. 14, 2012), pp. 16256–16264. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.4036-12.2012 (cit. on p. 1).

[DWZ03]   M. R. DeWeese, M. Wehr, and A. M. Zador. "Binary Spiking in Auditory Cortex". In: *The Journal of Neuroscience* 23.21 (Aug. 27, 2003), pp. 7940–7949. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.23-21-07940.2003 (cit. on p. 1).

[EC24]   O. Elkabetz and N. Cohen. "Continuous vs. discrete optimization of deep neural networks". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2024 (cit. on p. 17).

11

[Fie99]   D. J. Field. "Wavelets, Vision and the Statistics of Natural Scenes". In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 357.1760 (Sept. 1999). Ed. by B. W. Silverman and J. C. Vassilicos, pp. 2527–2542. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.1999.0446 (cit. on p. 1).

[Fra04]   G. Frahm. "Generalized elliptical distributions: theory and applications". PhD thesis. Universität zu Köln, 2004 (cit. on pp. 7, 19).

[Ger+20]   F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová. "Generalisation error in learning with random features and the hidden manifold model". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3452–3462. URL: http://proceedings.mlr.press/v119/gerace20a.html (cit. on pp. 5, 16).

[GG84]   S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* Pami-6.6 (Nov. 1984), pp. 721–741. URL: https://ieeexplore.ieee.org/document/4767596 (cit. on p. 4).

[God+23]   C. Godfrey, D. Brown, T. Emerson, and H. Kvinge. *On the Symmetries of Deep Learning Models and Their Internal Representations*. Mar. 24, 2023. arXiv: 2205.14258 [cs]. URL: http://arxiv.org/abs/2205.14258. preprint (cit. on p. 8).

[Gol+19]   S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová. "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup". In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 2).

[Gol+20]   S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová. "Modelling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model". In: *Physical Review X* 10.4 (Dec. 3, 2020), p. 041044. arXiv: 1909.11500 [cond-mat, stat]. URL: http://arxiv.org/abs/1909.11500 (cit. on pp. 2, 5, 16).

[Gol+22]   S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová. "The gaussian equivalence of generative models for learning with shallow neural networks". In: *Mathematical and Scientific Machine Learning*. Pmlr. 2022, pp. 426–471 (cit. on pp. 5, 16).

[Har+20]   M. Harsh, J. Tubiana, S. Cocco, and R. Monasson. "'Place-cell' Emergence and Learning of Invariant Data with Restricted Boltzmann Machines: Breaking and Dynamical Restoration of Continuous Symmetries in the Weight Space". In: *Journal of Physics A: Mathematical and Theoretical* 53.17 (Apr. 2020), p. 174002. URL: https://dx.doi.org/10.1088/1751-8121/ab7d00 (cit. on pp. 7, 10).

[HDZ08]   T. Hromádka, M. R. DeWeese, and A. M. Zador. "Sparse Representation of Sounds in the Unanesthetized Auditory Cortex". In: *PLOS Biology* 6.1 (Jan. 29, 2008), e16. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060016 (cit. on p. 1).

[HHH09]   A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics*. Red. by M. Viergever, G. Borgefors, R. Deriche, T. S. Huang, K. Ikeuchi, T. Jiang, R. Klette, A. Leonardis, H.-O. Peitgen, and J. K. Tsotsos. Vol. 39. Computational Imaging and Vision. London: Springer, 2009. URL: http://link.springer.com/10.1007/978-1-84882-491-1 (cit. on p. 3).

[HO00]   A. Hyvärinen and E. Oja. "Independent Component Analysis: Algorithms and Applications". In: *Neural Networks* 13.4 (June 1, 2000), pp. 411–430. URL: https://www.sciencedirect.com/science/article/pii/S0893608000000265 (cit. on pp. 2, 5, 10).

[HW59]   D. H. Hubel and T. N. Wiesel. "Receptive Fields of Single Neurones in the Cat's Striate Cortex". In: *The Journal of Physiology* 148.3 (Oct. 1959), pp. 574–591. pmid: 14403679 (cit. on p. 1).

[HW68]   D. H. Hubel and T. N. Wiesel. "Receptive Fields and Functional Architecture of Monkey Striate Cortex". In: *The Journal of Physiology* 195.1 (Mar. 1968), pp. 215–243. pmid: 4966457 (cit. on p. 1).

12

[Hyv99]   A. Hyvarinen. "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis". In: *IEEE Transactions on Neural Networks* 10.3 (May 1999), pp. 626–634. URL: http://ieeexplore.ieee.org/document/761722/ (cit. on p. 9).

[IG22]   A. Ingrosso and S. Goldt. "Data-driven emergence of convolutional structure in neural networks". In: *Proceedings of the National Academy of Sciences* 119.40 (2022), e2201854119 (cit. on pp. 2–8, 10, 16, 20).

[KK78]   E. I. Knudsen and M. Konishi. "Space and frequency are represented separately in auditory midbrain of the owl". In: *Journal of Neurophysiology* 41.4 (1978), pp. 870–884 (cit. on p. 1).

[KS11]   Y. Karklin and E. Simoncelli. "Efficient Coding of Natural Images with a Population of Noisy Linear-Nonlinear Neurons". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper/2011/hash/12e59a33dea1bf0630f46edfe13d6ea2-Abstract.html (cit. on p. 10).

[KSH12]   A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (cit. on p. 2).

[KZB19]   A. Kolesnikov, X. Zhai, and L. Beyer. *Revisiting Self-Supervised Visual Representation Learning*. Jan. 25, 2019. arXiv: 1901.09005 [cs]. URL: http://arxiv.org/abs/1901.09005. preprint (cit. on p. 4).

[MMN18]   S. Mei, A. Montanari, and P.-M. Nguyen. "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–e7671 (cit. on p. 2).

[NS08]   C. M. Niell and M. P. Stryker. "Highly Selective Receptive Fields in Mouse Visual Cortex". In: *The Journal of Neuroscience* 28.30 (July 23, 2008), pp. 7520–7536. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0623-08.2008 (cit. on p. 1).

[OF96]   B. A. Olshausen and D. J. Field. "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images". In: *Nature* 381.6583 (June 1, 1996), pp. 607–609. URL: https://doi.org/10.1038/381607a0 (cit. on pp. 1, 4).

[OF97]   B. A. Olshausen and D. J. Field. "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?" In: *Vision Research* 37.23 (Dec. 1997), pp. 3311–3325. URL: https://linkinghub.elsevier.com/retrieve/pii/S0042698997001697 (cit. on p. 1).

[Ped+11]   F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 9).

[Pin99]   A. Pinkus. "Approximation theory of the MLP model in neural networks". In: *Acta numerica* 8 (1999), pp. 143–195 (cit. on p. 2).

[Rin02]   D. L. Ringach. "Spatial Structure and Symmetry of Simple-Cell Receptive Fields in Macaque Primary Visual Cortex". In: *Journal of Neurophysiology* 88.1 (July 2002), pp. 455–463. URL: https://journals.physiology.org/doi/full/10.1152/jn.2002.88.1.455 (cit. on pp. 1, 2).

[Ros56]   M. Rosenblatt. "A central limit theorem and a strong mixing condition". In: *Proceedings of the national Academy of Sciences* 42.1 (1956), pp. 43–47 (cit. on p. 3).

[RSH02]   D. L. Ringach, R. M. Shapley, and M. J. Hawken. "Orientation Selectivity in Macaque v1: Diversity and Laminar Dependence". In: *The Journal of Neuroscience* 22.13 (July 1, 2002), pp. 5639–5651. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.22-13-05639.2002 (cit. on p. 1).

[RT95]   E. T. Rolls and M. J. Tovee. "Sparseness of the Neuronal Representation of Stimuli in the Primate Temporal Visual Cortex". In: *Journal of Neurophysiology* 73.2 (Feb. 1, 1995), pp. 713–726. URL: https://www.physiology.org/doi/10.1152/jn.1995.73.2.713 (cit. on p. 1).

13

[Sax+11]   A. Saxe, M. Bhand, R. Mudur, B. Suresh, and A. Ng. "Unsupervised Learning Models of Primary Cortical Receptive Fields and Receptive Field Plasticity". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper/2011/hash/e19347e1c3ca0c0b97de5fb3b690855a-Abstract.html (cit. on pp. 1, 10).

[Sen+18]   A. Sengupta, C. Pehlevan, M. Tepper, A. Genkin, and D. Chklovskii. "Manifold-Tiling Localized Receptive Fields Are Optimal in Similarity-Preserving Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/hash/ee14c41e92ec5c97b54cf9b74e25bd99-Abstract.html (cit. on p. 2).

[Sin+18]   Y. Singer, Y. Teramoto, B. D. Willmore, J. W. Schnupp, A. J. King, and N. S. Harper. "Sensory Cortex Is Optimized for Prediction of Future Input". In: *eLife* 7 (June 18, 2018). Ed. by J. L. Gallant and S. Kastner, e31557. URL: https://doi.org/10.7554/eLife.31557 (cit. on p. 2).

[SS95]   D. Saad and S. A. Solla. "On-line learning in soft committee machines". In: *Physical Review E* 52.4 (1995), p. 4225 (cit. on p. 3).

[Vei+22]   R. Veiga, L. Stephan, B. Loureiro, F. Krzakala, and L. Zdeborová. "Phase Diagram of Stochastic Gradient Descent in High-Dimensional Two-Layer Neural Networks". In: *Advances in Neural Information Processing Systems*. Neural Information Processing Systems. Vol. 35. Dec. 6, 2022, pp. 23244–23255. URL: https://papers.nips.cc/paper%5C%5Ffiles/paper/2022/hash/939bb847ebfd14c6e4d3b5705e562054-Abstract-Conference.html (cit. on p. 2).

[vHvdS98]   J. H. van Hateren and A. van der Schaaf. "Independent Component Filters of Natural Images Compared with Simple Cells in Primary Visual Cortex". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265.1394 (Mar. 7, 1998), pp. 359–366. URL: https://royalsocietypublishing.org/doi/10.1098/rspb.1998.0303 (cit. on p. 1).

[WMG11]   B. D. B. Willmore, J. A. Mazer, and J. L. Gallant. "Sparse Coding in Striate and Extrastriate Visual Cortex". In: *Journal of Neurophysiology* 105.6 (June 2011), pp. 2907–2919. pmid: 21471391. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3118756/ (cit. on p. 1).

[Woo+20]   B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. "Kernel and Rich Regimes in Overparametrized Models". In: *Proceedings of Thirty Third Conference on Learning Theory*. Conference on Learning Theory. Pmlr, July 15, 2020, pp. 3635–3673. URL: https://proceedings.mlr.press/v125/woodworth20a.html (cit. on p. 2).

[Yos+15]   J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. "Understanding Neural Networks through Deep Visualization". In: (June 22, 2015). arXiv: 1506.06579 [cs]. URL: http://arxiv.org/abs/1506.06579. preprint (cit. on p. 2).

[ZF13]   M. D. Zeiler and R. Fergus. "Visualizing and Understanding Convolutional Networks". In: (Nov. 28, 2013). arXiv: 1311.2901 [cs]. URL: http://arxiv.org/abs/1311.2901. preprint (cit. on p. 2).

14

A Definitions and Notation

A.1 Notation

We use $[n]$ to refer to the set $\{i \in \mathbb{N} : 1 \leq i \leq n\}$.

A.2 Algebraic sigmoid

For $k > 0$, the generalized algebraic sigmoid function is defined as

$$\text{alg}_k(x) \triangleq \frac{1}{2}\left(1 + \frac{x}{(1 + |x|^k)^{1/k}}\right) . \tag{7}$$

Following the main text, we drop the subscript when $k = 2$.

A.3 Inverse participation ratio (IPR)

The IPR is defined as:

$$\text{IPR}(\mathbf{w}) \triangleq \left(\sum_{i=1}^{D} w_i^4\right) \Big/ \left(\sum_{i=1}^{D} w_i^2\right)^2 ,$$

where w_i is the magnitude of dimension i of weight $\mathbf{w}$.

B Extensions beyond the scope of the main text

B.1 Analytical properties of the amplifier φ

We present several properties of the amplifying function φ defined in Lemma B.1.

Lemma B.1. *The localization amplifier φ in Eq. (6) satisfies $\varphi(-a) = -\varphi(a)$, for all $a \in (-1, 1)$. Moreover, its derivatives satisfy, where σ^2 and κ denote the variance and kurtosis of X_1, respectively:*

$$\varphi'(0) = \sqrt{\frac{2}{\pi}}\sigma^2 , \qquad and \qquad \varphi'''(0) = -\sqrt{\frac{2}{\pi}}(\kappa^4\sigma^4 - 3\sigma^2) .$$

To gain some understanding of how the marginal distributions of $\mathbf{X}$ impact localization, we use the derivatives in Lemma B.1 to construct a third-order Taylor approximation of φ about 0. The derivatives in Lemma B.1 reveal that every distribution for X_1 with constant variance will look like the same linear function near 0. φ only looks nonlinear once we move sufficiently far away from zero when the third-order term becomes relevant. For the case of $\sigma^2 = 1$ (where the variance of X_1 is equal to the value of the larger target $y = 1$) the third order term suggests that φ is super-linear when $\kappa < 3$, *i.e.*, the excess kurtosis is positive, and sub-linear otherwise.

A super-linear φ will encourage entries where $\Sigma_1\mathbf{w}$ is large to grow at a faster rate than other entries, which are all subject to the same *linear* norm constraint through the second term in Eq. (5). The covariance Σ_1, as a circulant matrix, acts as the convolution operator between some vector and σ_1^1 (the first row in Σ_1). Since $y = 1$ corresponds to the class with a larger length-scale correlation, σ_1^1 will decay relatively slowly and act like a weight local average. Thus, $\Sigma_1\mathbf{w}$ is the weighted local average for each entry in $\mathbf{w}$. So, entries where $\mathbf{w}$ is large within some neighborhood will be encouraged to grow faster than those which are smaller, an effect that compounds as Eq. (5) is integrated. Thus, super-linearity encourages localization.

As we will see in Theorem 3.3 for the setting of elliptical data, if φ is linear, $\mathbf{w}$ learns to be sinusoidal, and thus not localized. In the case of sub-linearity, we expect suppression of larger values, rather than promotion, as in the super-linear setting. Thus, to a first approximation, the sign of the excess kurtosis, $\kappa - 3$ (for $\sigma^2 = 1$), indicates whether $\mathbf{w}$ localizes.

However, simply studying $\varphi'''(0)$ is not sufficient to fully characterize how the marginals impact localization. A function can be sub-linear for small a and super-linear for larger a, making it unclear whether it will yield localization. For marginal distributions where $\kappa \approx 3$ that do not exhibit strict super- or sub-linearity, this condition is no longer precise enough to determine whether we see localization.

15

B.2 PDE limit of Eq. (5)

By taking N to be large and treating w as a continuous function with respect to position, *i.e.*, $w \equiv w(x,t)$, one can treat Eq. (5) as a partial differential equation (PDE). Finding its steady state then amounts to solving

$$\varphi\left(\frac{\sigma^1 \star w}{\sqrt{\langle \sigma^1 \star w, w \rangle}}\right) - \frac{1}{2}(\sigma^0 + \sigma^1) \star w \equiv 0 , \tag{8}$$

where $w : [0,1] \to \mathbb{R}$ is periodic and σ^y is the convolution corresponding to the limiting case of the matrix Σ_y. This equation does not appear to have an explicit solution for non-identity Σ_1, and thus, it may not be possible to solve the steady states of Eq. (5) exactly in this PDE limit or for finite N.

B.3 Assumptions A1 and A2 *vs.* Gaussian equivalence

Assumptions A1 and A2 are equivalent to approximating $\langle \mathbf{w}, \mathbf{X} \rangle \mid X_i$ as Gaussian early in training. Similar ideas have been used to derive gradient flow dynamics for neural networks, including in developing the Gaussian equivalence property of [Gol+20; Ger+20; Gol+22]. However, these works model the unconditional preactivation $\langle \mathbf{w}, \mathbf{X} \rangle$ as a Gaussian, rather than first conditioning on X_i. How this arises is that these previous works assert the Gaussian approximation *prior* to differentiating the loss function for the gradient flow dynamics. However, an approximation at that stage neglects the presence of a multiplicative factor $\mathbf{X}$ that appears as a result of the chain rule applied to $\langle \mathbf{w}, \mathbf{X} \rangle$. Abstractly, this approach assumes that $\mathcal{L}_{\text{exact}} \to \mathcal{L}_{\text{Gauss}}$ implies $\nabla_{\mathbf{w}} \mathcal{L}_{\text{exact}} \to \nabla_{\mathbf{w}} \mathcal{L}_{\text{Gauss}}$, but, in general, this does not follow, and here in particular this assumption does not capture the interplay of learning and higher-order input statistics. This contributes to the failure of Gaussian equivalence in [IG22]. In contrast, we can account for the additional $\mathbf{X}$ term in the derivation of Lemma 3.1 by assuming that $\langle \mathbf{w}, \mathbf{X} \rangle \mid X_i$ rather than $\langle \mathbf{w}, \mathbf{X} \rangle$ is Gaussian. This conditioning approach, along with the translation invariance of the data (Property S2), also motivates considering the marginal distributions X_i as the object of study to obtain gradient flows for neural networks trained on non-Gaussian inputs.

C Proofs of theoretical results

C.1 Gradient flow for mean-squared error (MSE) loss

The loss is given by:

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{\mathbf{X},Y}\left[(Y - \text{ReLU}(\langle \mathbf{w}, \mathbf{X} \rangle))^2\right] \\
&= \mathbb{E}_{\mathbf{X},Y}\left[Y^2\right] - 2\underbrace{\mathbb{E}_{\mathbf{X},Y}\left[Y\,\text{ReLU}(\langle \mathbf{w}, \mathbf{X} \rangle)\right]}_{\triangleq (I)} + \underbrace{\mathbb{E}_{\mathbf{X},Y}\left[\text{ReLU}^2(\langle \mathbf{w}, \mathbf{X} \rangle)\right]}_{\triangleq (II)} .
\end{aligned} \tag{9}$$

The assumption of sign symmetry (S3) gives that $\langle \mathbf{w}, \mathbf{X} \rangle$ is also sign-symmetric. First, this implies that $\mathbb{P}(\langle \mathbf{w}, \mathbf{X} \rangle > 0) = \frac{1}{2}$, so:

$$(II) = \frac{1}{2}\mathbb{E}_{\mathbf{X},Y \mid \langle \mathbf{w}, \mathbf{X} \rangle \geq 0}\left[\text{ReLU}^2(\langle \mathbf{w}, \mathbf{X} \rangle)\right] = \frac{1}{2}\mathbb{E}_{\mathbf{X},Y \mid \langle \mathbf{w}, \mathbf{X} \rangle \geq 0}\left[(\langle \mathbf{w}, \mathbf{X} \rangle)^2\right] .$$

Second, sign-symmetry of $\langle \mathbf{w}, \mathbf{X} \rangle$ implies that we can drop the conditioning on $\langle \mathbf{w}, \mathbf{X} \rangle \geq 0$, since $\langle \mathbf{w}, \mathbf{X} \rangle \overset{d}{=} -\langle \mathbf{w}, \mathbf{X} \rangle$. Thus,

$$\begin{aligned}
(II) &= \frac{1}{2}\mathbb{E}_{\mathbf{X},Y}\left[(\langle \mathbf{w}, \mathbf{X} \rangle)^2\right] \\
&= \frac{1}{2}\mathbf{w}^\top \mathbb{E}_{\mathbf{X},Y}\left[\mathbf{X}\mathbf{X}^\top\right]\mathbf{w} \\
&= \frac{1}{2}\mathbf{w}^\top\left(\frac{1}{K}\sum_{y=0}^{K-1}\Sigma_y\right)\mathbf{w} \\
&= \frac{1}{2K}\sum_{y=0}^{K-1}\mathbf{w}^\top\Sigma_y\mathbf{w} ,
\end{aligned}$$

16

where K is the number of values (classes) of discrete y. Finally, we differentiate $\mathcal{L}$ with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}} \mathcal{L} = 2 \mathbb{E}_{\mathbf{X},Y} \left[Y \mathbb{1}(\langle \mathbf{w}, \mathbf{X} \rangle \geq 0) \mathbf{X} \right] + \frac{1}{K} \sum_{y=0}^{K-1} \Sigma_y \mathbf{w} .$$

The gradient flow [EC24] is given by $\frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = -\tau \nabla_{\mathbf{w}} \mathcal{L}$, where τ is the learning rate. Thus,

$$\frac{1}{\tau} \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = 2 \mathbb{E}_{\mathbf{X},Y} \left[Y \mathbb{1}(\langle \mathbf{w}, \mathbf{X} \rangle \geq 0) \mathbf{X} \right] - \frac{1}{K} \sum_{y=1}^{K} \Sigma_y \mathbf{w} . \tag{10}$$

C.2   Proof of lemma 3.1

Setting $K = 2$ in equation (10), we have

$$\frac{1}{\tau} \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \mathbb{E}_{\mathbf{X}|Y=1}[\mathbb{1}(\langle \mathbf{w}, \mathbf{X} \rangle \geq 0) \mathbf{X}] - \frac{1}{2} \left(\Sigma_0 + \Sigma_1 \right) \mathbf{w} .$$

We wish to express the first term explicitly. Note that the first term is a vector in $\mathbb{R}^N$. We consider each of its entries separately by using the law of total expectation to write the i-th entry as:

$$\mathbb{E}_{\mathbf{X}|Y=1}[\mathbb{1}(\langle \mathbf{w}, \mathbf{X} \rangle \geq 0) X_i] = \mathbb{E}_{X_i|Y=1} \left[X_i \, \mathbb{P}_{\mathbf{X}|X_i=x_i, Y=1} \left[\langle \mathbf{w}, \mathbf{X} \rangle \geq 0 \right] \right] .$$

By Assumption A3, $\{w_i X_i \mid 1 \leq i \leq N\}$ satisfies Lindeberg's condition as $N \to \infty$. This is also known as a *uniform integrability* requirement. Before formally stating it, let us introduce two variables: $S_N \triangleq \sum_{j=1}^N w_j (X_j - \mu_{j|x_i})$, the partial sums, and their variance, $\sigma_N^2 \triangleq \mathbb{E}[S_N^2]$, where $\mu_{j|x_i} \triangleq \mathbb{E}[X_j \mid X_i = x_i]$ is the conditional mean of the j-entry given that i-th entry has value x_i. Then, Lindeberg's condition is formally stated as

$$\sup_N \mathbb{E}_{S_N|X_i=x_i} \left[\left| \frac{S_N^2}{\sigma_N^2} \right| \mathbb{1} \left(\left| \frac{S_N^2}{\sigma_N^2} \right| > x \right) \right] \to 0 \quad \text{as} \quad x \to \infty .$$

This condition effectively states that no term in the partial sum $w_i X_i$ will dominate. Under this condition, along with weak dependence (Property S1), we conclude from [Bra07, Theorems 1.19, 10.2] that $S_N/\sigma_N \xrightarrow{d} \mathcal{N}(0,1)$. Note that

$$S_N = \langle \mathbf{w}, \mathbf{X} \rangle - \langle \mathbf{w}, \mu_{|x_i} \rangle \qquad \text{and} \qquad \sigma_N = \sqrt{\mathbf{w}^\top \Sigma_{|x_i}^1 \mathbf{w}} ,$$

where $\mathbf{w}, \mathbf{X}$ are N-dimensional vectors, $\mu_{|x_i} = \mathbb{E}[\mathbf{X} \mid X_i = x_i]$ is the vector of conditional means, and $\Sigma_{|x_i}^1 \triangleq \mathrm{Cov}[\mathbf{X} \mid X_i = x_i, Y = 1] = \Sigma_1 - \sigma_i^1 \sigma_i^{1\top}$. Since σ_N and $\mu_{|x_i}$ are constant, we can write

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}|X_i=x_i, Y=1} \left[\langle \mathbf{w}, \mathbf{X} \rangle \geq 0 \right] &= \mathbb{P}_{\mathbf{X}|X_i, Y=1} \left[\frac{\langle \mathbf{w}, \mathbf{X} \rangle - \langle \mathbf{w}, \mu_{j|x_i} \rangle}{\sigma_N} \geq -\frac{\langle \mathbf{w}, \mu_{j|x_i} \rangle}{\sigma_N} \right] \\
&= \mathbb{P}_{Z \sim \mathcal{N}(0,1)} \left[Z \geq -\frac{\langle \mathbf{w}, \mu_{j|x_i} \rangle}{\sigma_N} \right] + o_N(1) \\
&= 1 - \frac{1}{2} \left[1 + \mathrm{erf} \left(-\frac{\langle \mathbf{w}, \mu_{j|x_i} \rangle}{\sqrt{2}\sigma_N} \right) \right] + o_N(1) \\
&= \frac{1}{2} + \frac{1}{2} \mathrm{erf} \left(\frac{\langle \mathbf{w}, \mu_{j|x_i} \rangle}{\sqrt{2}\sigma_N} \right) + o_N(1) ,
\end{aligned}$$

where the second step, in which we acquire $o_N(1)$, follows from the definition of convergence in distribution. Under Assumptions A1 and A2, we may express this as

$$\mathbb{P}_{\mathbf{X}|X_i=x_i, Y=1} \left[\langle \mathbf{w}, \mathbf{X} \rangle \geq 0 \right] = \frac{1}{2} + \frac{1}{2} \mathrm{erf} \left(\frac{\langle \mathbf{w}, x_i \sigma_i^y \rangle}{\sqrt{2}\sqrt{\mathbf{w}^\top \Sigma_1 \mathbf{w} - (\langle \mathbf{w}, \sigma_i^y \rangle)^2}} \right) + o_N(1) .$$

17

Therefore,

$$\mathbb{E}_{\mathbf{X}|Y=1}[\mathbb{1}(\langle \mathbf{w}, \mathbf{X}\rangle \geq 0)X_i]$$

$$= \mathbb{E}_{X_i|Y=1}\left[X_i\left(\frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{X_i}{\sqrt{2}}\cdot\frac{\langle \mathbf{w}, \sigma_i^y\rangle}{\sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w} - (\langle \mathbf{w}, \sigma_i^y\rangle)^2}}\right) + o_N(1)\right)\right]$$

$$= \frac{1}{2}\mathbb{E}_{X_i|Y=1}\left[X_i\operatorname{erf}\left(\frac{X_i}{\sqrt{2}}\cdot\frac{\langle \mathbf{w}, \sigma_i^y\rangle/\sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w}}}{\sqrt{1 - (\langle \mathbf{w}, \sigma_i^y\rangle/\sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w}})^2}}\right)\right] + \mathbb{E}_{X_i}[|X_i|]o_N(1)$$

$$= \frac{1}{2}\mathbb{E}_{X_i|Y=1}\left[X_i\operatorname{erf}\left(\frac{X_i}{\sqrt{2}}\cdot\operatorname{alg}^{-1}\left(\frac{\langle \mathbf{w}, \sigma_i^y\rangle}{\sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w}}}\right)\right)\right] + \mathbb{E}_{X_i}[|X_i|]o_N(1)\,.$$

Defining $\varphi_i(a) \triangleq \mathbb{E}_{X_i|Y=1}[X_i\operatorname{erf}(X_i\operatorname{alg}^{-1}(a)/\sqrt{2})]$ we can write

$$\mathbb{E}_{\mathbf{X}|Y=1}[\mathbb{1}(\langle \mathbf{w}, \mathbf{X}\rangle \geq 0)X_i] = \frac{1}{2}\varphi_i\left(\frac{\langle \mathbf{w}, \sigma_i^y\rangle}{\sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w}}}\right) + \mathbb{E}_{X_i}[|X_i|]o_N(1)\,.$$

Note that $\mathbb{E}_{X_i}[|X_i|] \leq \sqrt{\mathbb{E}_{X_i}[X_i^2]} = \sigma$ by Cauchy-Schwarz. So, $\mathbb{E}_{X_i}[|X_i|]o_N(1) = o_N(1)$. Moreover, by translation-invariance, all X_i have the same marginal, so $\varphi_i \equiv \varphi_1 \triangleq \varphi$. Thus,

$$\mathbb{E}_{\mathbf{X}|Y=1}[\mathbb{1}(\langle \mathbf{w}, \mathbf{X}\rangle \geq 0)X_i] = \frac{1}{2}\varphi\left(\frac{\langle \mathbf{w}, \sigma_i^y\rangle}{\sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w}}}\right) + o_N(1)\,.$$

This form holds for all entries i. Concatenating them, we obtain

$$\mathbb{E}_{\mathbf{X}|Y=1}[\mathbb{1}(\langle \mathbf{w}, \mathbf{X}\rangle \geq 0)\mathbf{X}] = \frac{1}{2}\varphi\left(\frac{\Sigma_1\mathbf{w}}{\sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w}}}\right) + o_N(1)\,,$$

which gives the desired result. $\qquad\qquad\square$

C.3 Proof of lemma B.1

Property 1 follows from the fact that alg and erf are odd functions. This implies Property 4 since an odd function must have zero for even Taylor coefficients.

Properties 2 and 3 follow from differentiating Eq. (6). The first derivative is given by

$$\varphi'(a) = \frac{\sqrt{2}}{\sqrt{\pi}(1-a^2)^{\frac{3}{2}}}\mathbb{E}_{X_1}\left[X_1^2 e^{-\frac{X_1^2 a^2}{2(1-a^2)}}\right]\,.$$

Setting $a = 0$, we get

$$\varphi'(0) = \sqrt{\frac{2}{\pi}}\mathbb{E}\left[X_1^2\right] = \sqrt{\frac{2}{\pi}}\sigma^2\,.$$

The third derivative is given by

$$\varphi'''(a) = \frac{\sqrt{2}}{\sqrt{\pi}(1-a^2)^{\frac{7}{2}}(a^2-1)^2}\mathbb{E}_{X_1}\left[X_1^2(12a^6 + (9X_1^2 - 21)a^4 + (X_1^4 - 8X_1^2 + 6)a^2 - X_1^2 + 3)e^{-\frac{X_1^2 a^2}{2(1-a^2)}}\right]\,.$$

Again setting $a = 0$ gives

$$\varphi'''(0) = \sqrt{\frac{2}{\pi}}\mathbb{E}_{X_1}\left[X_1^2(-X_1^2 + 3)\right] = \sqrt{\frac{2}{\pi}}\left(3\mathbb{E}_{X_1}[X_1^2] - \mathbb{E}_{X_1}[X_1^4]\right) = -\sqrt{\frac{2}{\pi}}(\kappa^4\sigma^4 - 3\sigma^2)\,.$$

$$\square$$

18

C.4 Proof of Proposition 3.3

The pdf of $X \sim \mathcal{E}_N(\mu, \Sigma, \phi)$ is

$$p_X(x) = \frac{1}{\sqrt{\det(\Sigma)}} g((x - \mu)^\top \Sigma^{-1} (x - \mu)) \,,$$

for some function $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ [Fra04]. A key property of elliptical distributions is that if $X \sim \mathcal{E}_N(\mu, \Sigma, \phi)$, then its affine transformation is also elliptical: $\langle \mathbf{w}, \mathbf{X} \rangle \sim \mathcal{E}_1(\langle \mathbf{w}, \mu \rangle, \mathbf{w}^\top \Sigma \mathbf{w}, \phi)$ for any $\mathbf{w} \in \mathbb{R}^N$. Thus,

$$p_{\langle \mathbf{w}, \mathbf{X} \rangle}(s) = \frac{1}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}} \tilde{g} \left(\frac{(s - \langle \mathbf{w}, \mu \rangle)^2}{\mathbf{w}^\top \Sigma \mathbf{w}} \right),$$

for some other function $\tilde{g} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

From our assumption of sign-symmetry, we have $\mu = 0$. For brevity, we define $\sigma^2 \triangleq \mathbf{w}^\top \Sigma \mathbf{w}$ and $S \triangleq \langle \mathbf{w}, \mathbf{X} \rangle$. We begin by computing (I) in Eq. (9). Recall that we have $y = 0, 1$ (*i.e.*, $K = 2$). So,

$$2 \times (I) = \mathbb{E}_{S|Y=1}[\text{ReLU}(S)] = \int_0^\infty \frac{1}{\sigma} \tilde{g} \left(\frac{s^2}{\sigma^2} \right) s \, \mathrm{d}s.$$

At this point, we apply a u-substitution with $u = s^2/\sigma^2$, and thus $\mathrm{d}u = 2s \, \mathrm{d}s/\sigma^2 \iff \sigma \mathrm{d}u/2 = s \, \mathrm{d}s/\sigma$. This yields

$$2 \times (I) = \frac{\sigma}{2} \underbrace{\int_0^\infty \tilde{g}(u) \, \mathrm{d}u}_{\triangleq C} = \frac{C}{2} \sqrt{\mathbf{w}^\top \Sigma_1 \mathbf{w}}.$$

Recall that we assume the MSE loss $\mathcal{L}$ is finite. The first term in Eq. (9) is clearly finite for $y = 0, 1$. The term (II) is also easily seen to be finite, since it evaluates to $\mathbf{w}^\top (\Sigma_0 + \Sigma_1) \mathbf{w}/4$. Thus, $\mathcal{L}$ being finite implies (I) in Eq. (9) is finite, which implies $C < \infty$. We computed (II) from Eq. (9) above as

$$\frac{1}{4} \mathbf{w}^\top (\Sigma_0 + \Sigma_1) \mathbf{w}$$

for $K = 2$. Plugging in (I) and (II) and differentiating with respect to $\mathbf{w}$, we get

$$\frac{1}{\tau} \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \frac{\Sigma_1 \mathbf{w}}{\sqrt{\mathbf{w}^\top \Sigma_1 \mathbf{w}}} - \frac{1}{2} (\Sigma_0 + \Sigma_1) \mathbf{w} \,. \tag{11}$$

The steady states of Eq. (11) thus satisfy

$$C \frac{\Sigma_1 \mathbf{w}}{\sqrt{\mathbf{w}^\top \Sigma_1 \mathbf{w}}} = \frac{1}{2} (\Sigma_0 + \Sigma_1) \mathbf{w} \,.$$

Recall that translation-invariance implies that Σ_0 and Σ_1 are circulant, and since they are covariance matrices, they are symmetric. Then, they diagonalize in the basis given by the real and imaginary parts of the first $n/2$ Fourier components in the discrete Fourier transform, which we denote by the $n \times n$ real matrix $\mathcal{F}$. Note that $\mathcal{F}$ is orthogonal. Define $\mathbf{v} = \mathcal{F}^\top \mathbf{w}$ and $\Lambda_y = \mathcal{F}^\top \Sigma_y \mathcal{F}$ for $y = 0, 1$. Thus, the steady states satisfy

$$C \frac{\Lambda_1 \mathbf{v}}{\sqrt{\mathbf{v}^\top \Lambda_1 \mathbf{v}}} = \frac{1}{2} (\Lambda_0 + \Lambda_1) \mathbf{v} \,.$$

This holds iff for all $i \in [n]$,

$$C(\Lambda_1)_{ii} (\mathbf{v}^\top \Lambda_1 \mathbf{v})^{-\frac{1}{2}} v_i = \frac{1}{2} ((\Lambda_0)_{ii} + (\Lambda_1)_{ii}) v_i \,.$$

19

Thus, if $v_i \neq 0$, we must have that

$$\frac{(\Lambda_0)_{ii}}{(\Lambda_1)_{ii}} = 2C(\mathbf{v}^\top \Lambda_1 \mathbf{v})^{-\frac{1}{2}} - 1 \ .$$

That is, the ratio of the i-th eigenvalues of Σ_0 and Σ_1 must be constant for all i s.t. $v_i \neq 0$. The eigenvalues of these matrices always come in pairs because of how we defined $\mathcal{F}$ using both the real and imaginary parts of the discrete Fourier transform. In general, we observe that each pair assumes a unique value. So, since C is finite, the condition above can hold for at most two distinct values of i. Therefore, $v_i = 0$ for all but at most two $i \in [n]$, implying that the steady state w is of the form $a\cos(2\pi kx) + b\sin(2\pi kx)$, *i.e.*, it is oscillatory. As such, it is *not* localized. $\qquad\square$

D   Additional experiments

D.1   Visualizing breakdown of Assumption A3

Fig. 8 demonstrates that our analytical model holds for long enough during training to capture the emergence of localization in the single ReLU neuron (M2). In the first three columns, we visualize the IPR of the weights from our empirical and analytical models, as well as the ℓ_2 difference between these two weights. In the first four rows, we visualize these metrics for four random initializations of the model, training each on NLGP($g = 100$) with $\xi_0 = 0.3$ and $\xi_1 = 0.7$, where we expect from Theorem 3.2 to see localization. We see the error rapidly increase shortly after IPR increases, indicating the formation of localized receptive fields. The last three columns confirm this, as they show snapshots from *before*, *during*, and *after* the divergence between the empirical and analytical weights. We observe the weights are nearly identical *before*, differ only slightly at the most localized point *during*, and are both localized *after*, but possibly with different magnitudes and positions. The difference that emerges *during* is due to a breakdown of Assumption A3 used to create our analytical model, which is violated when the norm of $\mathbf{w}$ is dominated by just a few entries, *i.e.*, it is localized. While [IG22] also observe a breakdown in their analytical model as localization emerges, ours, crucially, holds for long enough to characterize the emergence of localization.

We discuss the individual subplots in more detail. In all but the third row of Fig. 8, the analytical predictions are near-exact; in the third row, we predict localization, but at the wrong position. Focusing again on the first row, we see that at $t = 20$, the weights have not yet become localized (from IPR, left, first, and visually) and analytical and empirical weights match nearlt exactly, as confirmed by the small distance in left, center above. At $t = 30$, a localized peak around $i = 21$ begins to emerge, violating Assumption A3 and weakening analytical precision. The analytical model then underestimates the degree to which the main peak at $i = 21$ dominates, while it overestimates the size of competing peaks at $i = 30, 37$, and 90. Despite this, at $t = 50$, we see that predictions from the analytical model retain a match to the empirical model.

In the last row of Fig. 8, we use the same initialization and setting as in the first row, except that we train on NLGP($g = 0.01$) data instead. From Theorem 3.2, we *do not* expect to see localization. The evolution of IPR confirms this, as it stays low in magnitude. We also see that, because localization never emerges, Assumption A3 is never violated, and so our analytical model accounts for the empirical model nearly perfectly.

20
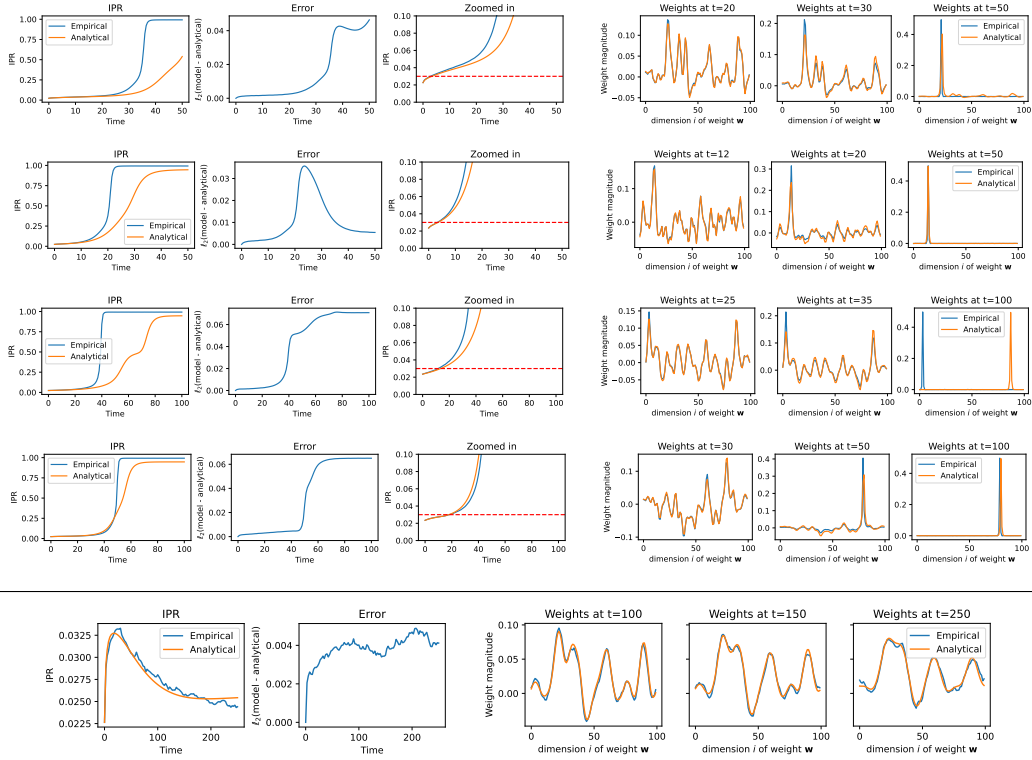
Figure 8: (**Top**) Four initializations trained on NLGP($g = 100$) with $\xi_0 = 0.3$ and $\xi_1 = 0.7$. As expected, weights always localize. In (Left, First) we plot IPR for empirical and analytical receptive fields (RFs) across time (defined as (# of gradient steps) $\times \tau$, the learning rate). In (Left, Second) we plot the time-evolution of ℓ_2 distance between the empirical and analytical RFs. In (Left, Third) we zoom in on (Left, First), restricting the range to $[0, 0.1]$ to more closely see divergence in IPR early in training. In (Right, First) and (Right, Second), we snapshot the empirical and analytical RFs at a time *before* and *just after*, respectively, the analytical model breaks down (according to IPR and ℓ_2 distance) due to localization. Finally, in (Right, Third), we snapshot *at the end* of the training period. (**Bottom**) Same initialization as first row in **top**, but trained on NLGP($g = 0.01$) data, again with $\xi_0 = 0.3$ and $\xi_1 = 0.7$. As expected, weights do not localize. We plot the same quantities as above, but here the predictions of our analytical model hold *throughout* the entire training process as localization never emerges and so assumption (A3) is not violated as above.

21

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, in the abstract and introduction we claim to derive an analytical model for the effective dynamics of localization (Section 3) in the setting described in Section 2 and validate this model with experiments (Section 4).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, in the conclusion we note that our exact analysis is retricted to the single-neuron architecture M2 in the data setting described in Section 2, and that validations beyond that scope (like in Section 4) remain empirical even if they show promise.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, our assumptions are rigourously stated in Section 3 and the proofs are provided in Appendix C.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our setup (Section 4) and experiments (Section 4) provides full details on the data, hyperparameters, and training procedures used to obtain the figures. We also provide code (referenced on the first page).

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, we provide a link to the code repository in the first page of the paper, and the repository contains all the code and data needed to reproduce the experiments.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we provide primary details in Section 4 and all the details in the code.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, for our prediction experiment in Section 4.2 we report a measure of variation over multiple runs.

8. **Experiments Compute Resources**

22

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Yes, we provide this information at the beginning of Section 4.

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Yes, we have reviewed the NeurIPS Code of Ethics and believe that our work conforms to it.

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Since our work is primarily theoretical analysis of an existing phenomenon, we do not consider it to have direct societal impacts.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not involve the release of data or models that have a significant risk for misuse.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We wrote our own code for all experiments to procedurally generate data and train and analyze models. Visualizations from prior work in Fig. 1 are properly credited.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We document our code (the only asset we release) in the accompanying repository.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

23