RESEARCH ARTICLE

Optimal learning with excitatory and inhibitory synapses

Alessandro Ingrosso¤*

Zuckerman Mind, Brain, Behavior Institute, Columbia University, New York, New York, United States of America

¤ Current address: Quantitative Life Sciences, The Abdus Salam International Centre for Theoretical Physics - ICTP, Trieste, Italy
* ingrosso@ictp.it

Abstract

Characterizing the relation between weight structure and input/output statistics is fundamental for understanding the computational capabilities of neural circuits. In this work, I study the problem of storing associations between analog signals in the presence of correlations, using methods from statistical mechanics. I characterize the typical learning performance in terms of the power spectrum of random input and output processes. I show that optimal synaptic weight configurations reach a capacity of 0.5 for any fraction of excitatory to inhibitory weights and have a peculiar synaptic distribution with a finite fraction of silent synapses. I further provide a link between typical learning performance and principal components analysis in single cases. These results may shed light on the synaptic profile of brain circuits, such as cerebellar structures, that are thought to engage in processing time-dependent signals and performing on-line prediction.

Author summary

A general analysis of learning with biological synaptic constraints in the presence of statistically structured signals is lacking. Here, analytical techniques from statistical mechanics are leveraged to analyze association storage between analog inputs and outputs with excitatory and inhibitory synaptic weights. The linear perceptron performance is characterized and a link is provided between the weight distribution and the correlations of input/output signals. This formalism can be used to predict the typical properties of perceptron solutions for single learning instances in terms of the principal component analysis of input and output data. This study provides a mean-field theory for sign-constrained regression of practical importance in neuroscience as well as in adaptive control applications.

OPEN ACCESS

Citation: Ingrosso A (2020) Optimal learning with excitatory and inhibitory synapses. PLoS Comput Biol 16(12): e1008536. https://doi.org/10.1371/journal.pcbi.1008536

Editor: Abigail Morrison, Research Center Jülich, GERMANY

Received: May 27, 2020

Accepted: November 13, 2020

Published: December 28, 2020

Data Availability Statement: All relevant data are within the manuscript.

Funding: This work was supported by the Gatsby Charitable Foundation (GAT3708) and by NSF NeuroNex Award DBI-1707398. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The author has declared that no competing interests exist.

Introduction

At the most basic level, neuronal circuits are characterized by the subdivision into excitatory and inhibitory populations, a principle called Dale's law. Even though the precise functional role of Dale's law has not yet been understood, the importance of synaptic sign constraints is pivotal in constructing biologically plausible models of synaptic plasticity in the brain [1–5]. The properties of synaptic couplings strongly impact the dynamics and response of neural circuits, thus playing a crucial role in shaping their computational capabilities. It has been argued that the statistics of synaptic weights in neural circuits could reflect a principle of optimality for information storage, both at the level of single-neuron weight distributions [6, 7] and inter-cell synaptic correlations [8] (e.g. the overabundance of reciprocal connections). A number of theoretical studies, stemming from the pioneering Gardner approach [9], have investigated the computational capabilities of stylized classification and memorization tasks in both binary [10–13] and analog perceptrons [14, 15], using synthetic data. With some exceptions mentioned in the following, these studies considered random uncorrelated inputs and outputs, a usual approach in statistical learning theory. One interesting theoretical prediction is that non-negativity constraints imply that a finite fraction of synaptic weights are set to zero at critical capacity [6, 15, 16], a feature which is consistent with experimental synaptic weight distributions observed in some brain areas, e.g. input fibers to Purkinje cells in the cerebellum.

The need to understand how the interaction between excitatory and inhibitory synapses mediates plasticity and dynamic homeostasis [17, 18] calls for the study of heterogeneous multi-population feed-forward and recurrent models. A plethora of mechanisms for excitatory-inhibitory (E-I) balance of input currents onto a neuron have been proposed [19, 20]. At the computational level, it has recently been shown that a peculiar scaling of excitation and inhibition with network size, originally introduced to account for the high variability of neural firing activity [21–27], carries the computational advantage of noise robustness and stability of memory states in associative memory networks [13].

Analyzing training and generalization performance in feed-forward and recurrent networks as a function of statistical and geometrical structure of a task remains an open problem both in computational neuroscience and statistical learning theory [28–32]. This calls for statistical models of the low-dimensional structure of data that are at the same time expressive and amenable to mathematical analyses. A few classical studies investigated the effect of "semantic" (among input patterns) and spatial (among neural units) correlations in random classification and memory retrieval [33–35]. The latter are important in the construction of associative memory networks for place cell formation in the hippocampal complex [36].

For reason of mathematical tractability, the vast majority of analytical studies in binary and analog perceptron models focused on the case where both inputs and outputs are independent and identically distributed. In this work, I relax this assumption and study optimal learning of input/output associations with real-world statistics with a linear perceptron having heterogeneous synaptic weights. I introduce a mean-field theory of an analog perceptron in the presence of weight regularization with sign-constraints, considering two different statistical models for input and output correlations. I derive its critical capacity in a random association task and study the statistical properties of the optimal synaptic weight vector across a diverse range of parameters.

This work is organized as follows. In the first section, I introduce the framework and provide the general definitions for the problem. I first consider a model of temporal (or, equivalently, "semantic") correlations across inputs and output patterns, assuming statistical independence across neurons. I show that optimal solutions are insensitive to the fraction of E and I weights, as long as the external bias is learned. I derive the weight distribution and show that it is characterized by a finite fraction of zero weights also in the general case of E-I constraints and correlated signals. The assumption of independence is subsequently relaxed in order to provide a theory that depends on the spectrum of the sample covariance matrix and the dimensionality of the output signal along the principal components of the input. The implications of these results are discussed in the final section.

Results

Mean-field theory with correlations

Consider the problem of linearly mapping a set of correlated inputs $x_{i\mu}$, with $i \in 1, \ldots, N$ and $\mu = 1, \ldots, P$ from $N_E = f_E N$ excitatory (E) and $N_I = (1 - f_E)$ inhibitory (I) neurons, onto an output y_μ using a synaptic vector $\boldsymbol{w}$, in the presence of a learnable constant bias current b (Fig 1). To account for different statistical properties of E and I input rates, we write the elements of the input matrix as $(X)_{i\mu} \equiv x_{i\mu} = \bar{x}_i + \sigma_i \xi_{i\mu}$ with $\bar{x}_i = \bar{x}_E$ for $i \leq f_E N$ and $\bar{x}_i = \bar{x}_I$ for $i > f_E N$ and the same for σ_i. At this stage, the quantities $\xi_{i\mu}$ have unit variance and are uncorrelated across neurons: $\langle \xi_{i\mu}\, \xi_{i\nu} \rangle = \delta_{ij} C_{\mu\nu}$. In the following, we refer to x and y as signals and μ as a time index, although we consider general "semantic"correlations across the patterns $\boldsymbol{x}_\mu$ [34]. The output signal has average $\langle y_\mu \rangle = \bar{y}$ and variance $\langle (y_\mu - \bar{y})^2 \rangle = \sigma_y^2$. We initially consider output signals y_μ with the same temporal correlations as the input, namely $\langle \delta y_\mu\, \delta y_\nu \rangle = C_{\mu\nu}$, where $y_\mu = \bar{y} + \sigma_y \delta y_\mu$.

For a given input-output set, we are faced with the problem of minimizing the following regression loss (energy) function:

$$E(w; \gamma, x, y) = \frac{1}{2} \sum_{\mu=1}^{P} \left(\sum_{i=1}^{N} w_i x_{i\mu} + b - y_\mu \right)^2 + \frac{N\gamma}{2} \sum_{i=1}^{N} w_i^2 \tag{1}$$

with $w_i > 0$ for $i \leq f_E N$, $w_i < 0$ otherwise. The rationale for using a regularization term lies not

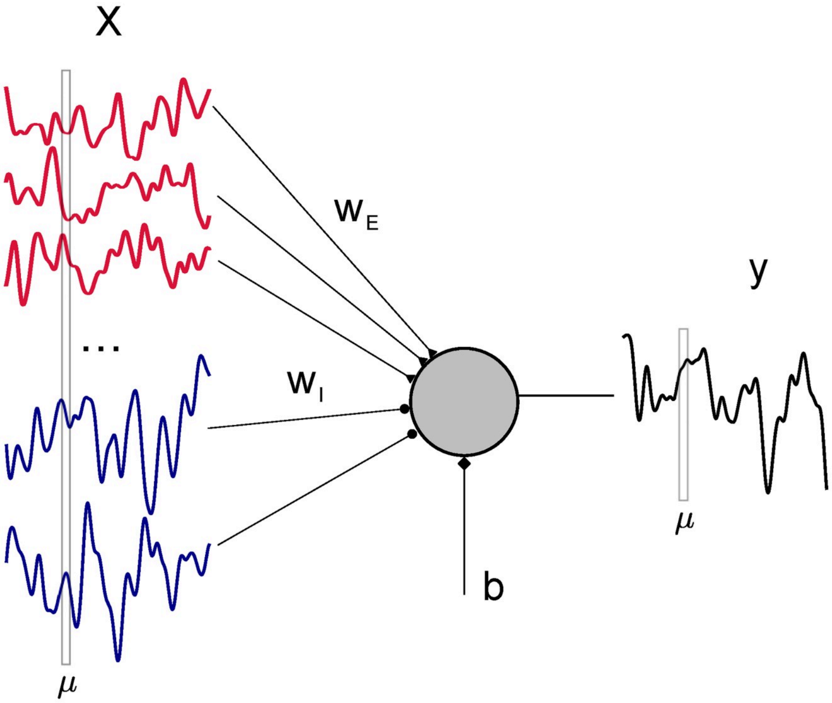

Fig 1. Schematic of the learning problem. A linear perceptron receives N correlated signals (input rates of presynaptic neurons) $x_{i\mu}$ and maps them to the output y_μ through $N_E = f_E N$ excitatory and $N_I = (1 - f_E)N$ plastic inhibitory weights w_i, plus an additional bias current b.

https://doi.org/10.1371/journal.pcbi.1008536.g001

only in alleviating ill-conditioning due to input correlations, but also in controlling the metabolic cost of synaptic plasticity and transmission. Preliminary numerical experiments showed that the typical vector $\boldsymbol{w}$ that solves this sign-constrained least square problem has a squared norm $\sum_{i=1}^{N} w_i^2 = \mathcal{O}(1)$, irrespectively of the L2 regularization, as in the special case of i.i.d input/output and non-negative synaptic weights [15]. Synaptic weights w_i are thus of $\mathcal{O}(1/\sqrt{N})$, hence the scaling of the regularization term $N\gamma$ and the bias current $b = I\sqrt{N}$. In order to consider a well defined $N \to \infty$ limit for E and the spectrum of the matrix C, we take $P = \alpha N$, with α called the *load*, as is costumary in mean-field analysis of perceptron problems [9].

Optimizing with respect to the bias b naturally yields solutions $\boldsymbol{w}$ for which

$$N_{\mathrm{E}}\bar{w}_{\mathrm{E}}\bar{x}_{\mathrm{E}} + N_{\mathrm{I}}\bar{w}_{\mathrm{I}}\bar{x}_{\mathrm{I}} + b = \bar{y} \tag{2}$$

where we call $\bar{w}_c = \frac{1}{N_c}\sum_{i\in c} w_i = \mathcal{O}\left(1/\sqrt{N}\right)$ the average excitatory and inhibitory weight, with $c \in \{\mathrm{E}, \mathrm{I}\}$. We call this property *balance*, in that the same scaling is used in balanced state theory of neural circuits [21, 22, 24].

In order to derive a mean-field description for the typical properties of the learned synaptic vector $\boldsymbol{w}$, we employ a statistical mechanics framework in which the minimizer of E is evaluated after averaging across all possible realizations of the input matrix X and output y. To do so, we compute the free energy density

$$f = -\frac{1}{\beta N}\langle \log Z\rangle_{x,y} \tag{3}$$

where $Z = \int d\mu\,(\boldsymbol{w})e^{-\beta E}$ is the so-called *partition function* and the measure $d\mu(\boldsymbol{w}) = \prod_{i\in \mathrm{E}}\theta(w_i)dw_i\prod_{k\in I}\theta(-w_k)dw_k$ implements the sign-constraints over the synapic weight vector $\boldsymbol{w}$. The brackets in Eq (3) stand for the quenched average over all the quantities $x_{i\mu}$ and y_μ, and the inverse temperature β will allow us to select weight configurations $\boldsymbol{w}$ that minimize the energy E. The free energy density f acts as a generating function from which all the statistical quantities of interest can be calculated by appropriate differentiation and taking the $\beta \to \infty$ limit. In particular, we will be interested in the (normalized) average loss $\epsilon = \frac{\langle E\rangle}{N}$ and the error $\epsilon_{err} = \frac{1}{2N}\langle |X^T\boldsymbol{w} + \boldsymbol{b} - \boldsymbol{y}|^2\rangle$, corresponding to the average value of the first term in Eq (1), where $\boldsymbol{b}$ is a P-dimensional vector containing b in every element. The average in Eq (3) can be computed in the $N \to \infty$ limit with the help of the replica method, an analytical continuation technique that entails the introduction of a number n of *formal* replicas of the vector $\boldsymbol{w}$. A general expression for f can be obtained in the large N limit using the saddle-point method. The crucial quantity in our derivation is the (replicated) cumulant generating function $Z_{\xi,\delta y}$ for the (mean-removed) input x and output y, which can be easily expressed as a function of the eigenvalues λ_μ, $\mu = 1, \ldots, \alpha N$ of the covariance matrix C, plus a set of order parameters to be evaluated self-consistently (Methods).

Critical capacity

The existence of weight vectors $\boldsymbol{w}$'s with a certain value of the regression loss E in the error regime ($\epsilon > 0$) is described by the so-called *overlap* order parameter $\Delta\tilde{q}_w$. In the replica-based derivation of the mean-field theory, overlap parameters are introduced with the purpose of decoupling the w_i's over the i index, and represent the scalar-product of two different configurations of the weights $\boldsymbol{w}$ (Methods: *Replica formalism: ensemble covariance matrix (EC)*). For finite β, the quantity $\Delta q_w = \beta\Delta\tilde{q}_w$ represents the variance of the synaptic weights across different solutions. In the asymptotic limit $\beta \to \infty$ of Eq (3), a simple saddle-point equation for $\Delta\tilde{q}_w$

can be derived when b is chosen to minimize Eq (1):

$$\alpha\Delta\tilde{q}_w\left\langle\frac{\lambda}{1+\Delta\tilde{q}_w\lambda}\right\rangle_{\rho(\lambda)}=\frac{1}{2}-\gamma\Delta\tilde{q}_w \tag{4}$$

where $\rho(\lambda)$ is the distribution of eigenvalues of C.

In the absence of weight regularization ($\gamma = 0$), we define the critical capacity α_c as the maximal load $\alpha = P/N$ for which the patterns $\boldsymbol{x}_\mu$ can be correctly mapped to their outputs y_μ with zero error. When the synaptic weights are not sign-constrained, the critical capacity is obviously $\alpha_c = 1$, since the matrix X is typically full rank. In the sign-constrained case, α_c is found to be the minimal value of α such that Eq (4) is satisfied for $0 < \Delta\tilde{q}_w < +\infty$. Noting that the left-hand side in Eq (4) is a non-decreasing function of $\Delta\tilde{q}_w$ with an asymptote in α, the order parameter $\Delta\tilde{q}_w$ goes to $+\infty$ as the critical capacity is approached from the right. We thus find for $\gamma = 0$ the surpisingly simple result:

$$\alpha_c = 0.5 \tag{5}$$

As shown in Fig 2A in the case of i.i.d. x and y, the loss has a sharp increase at $\alpha = 0.5$. This holds irrespectively of the structure of the covariance matrix C and the ratio of excitatory weights f_E. In Fig 2A, we also show the average minimal loss ϵ for increasing values of the regularization parameter γ.

In [15], the authors showed that, in the case with excitatory synapses only and uncorrelated inputs and outputs, α_c approaches 0.5 in the limit when the quantity $\frac{\sigma_y^2\bar{x}_\mathrm{E}^2}{I^2\sigma_\mathrm{E}^2}$ goes to zero, and analyzed which conditions on inputs and outputs statistics lead to maximize capacity. Here we take a complementary approach, where the x and y statistics are fixed and capacity is optimized within the error regime, so that the optimal bias $I\sqrt{N}$ is well defined in terms of minimizing $\langle E\rangle$ at any load α. The bias optimization leads to a massive simplification of the saddle-point equations and makes results independent of the E/I ratio and the input/output statistics

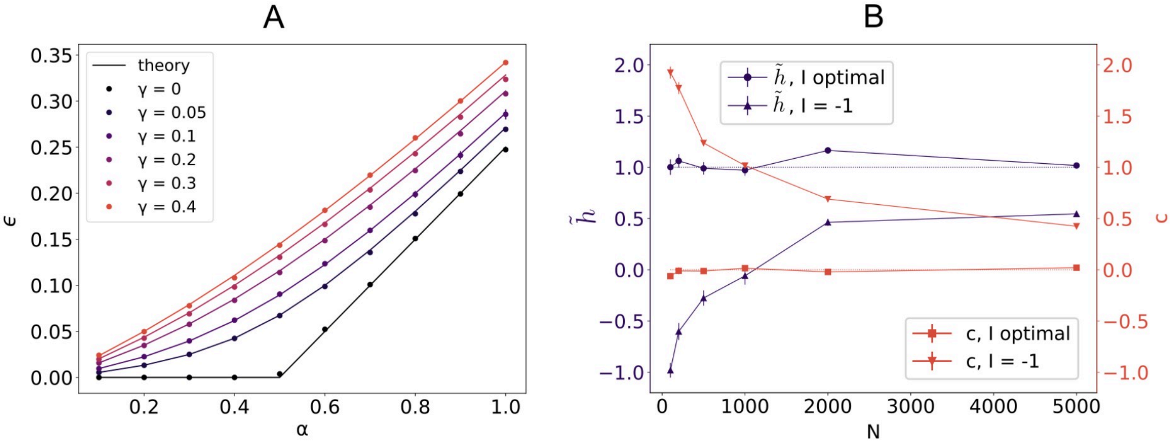

Fig 2. Critical capacity and weight balance. A: Average loss ϵ for a linear perceptron with $f_\mathrm{E} = 0.8$ positive synaptic weights in the case of i.i.d. input X and output y for increasing values of the regularization γ. Parameters: $N = 1000$, $\bar{x}_\mathrm{E}=\bar{x}_I=\sigma_\mathrm{E}=\sigma_I=\bar{y}=\sigma_y=1$. Each point is an average across 50 samples. Full lines show the theoretical results. B: Mean-field component $\tilde{h}$ (left axis, purple) and weight-input correlation c (right axis, red) for increasing dimension N in the case where the bias current $b = I\sqrt{N}$ is either learned (I optimal) or fixed at the outset ($I = -1$) for $f_\mathrm{E} = 1$, $\gamma = 0.1$, $\alpha = 0.8$. Inputs X and output y are time-correlated with un-normalized Gaussian covariance C, $\tau = 10$ (see text). The remaining parameters are as in A. The asymptotic value $\tilde{h}=\bar{y}=1$ is highlighted by the purple dotted line, the value $c = 0$ by the red dotted line as guide for the eye.

https://doi.org/10.1371/journal.pcbi.1008536.g002

(Methods: *EC, Saddle-point equations*). One may observe that, in the particular case studied by [15], α_c is maximal for very large I, due to the divergence of the norm of $\boldsymbol{w}$ at critical capacity for an optimal bias in the absence of regularization.

The independence of our results with respect to the E/I ratio for an optimal bias current signals a *local gauge invariance*, as observed by [37, 38] for a sign-constrained binary perceptron. Indeed, calling g_i = sign w_i, we can write the mean-removed output as $\sum_{i=1}^{N} g_i|w_i|\sigma_i\xi_i^\mu$ and redefine the ξ's as $g_i\xi_i^\mu$, without changing their occurrence probability. This establishes an equivalence to a linear perceptron with non-negative weights (see [37] for more details), once the mean contribution has been removed. Any residual dependence of α_c or ϵ on external parameters must therefore be ascribed to the volume of weights satisfying Eq (2), for a sub-optimal external current b.

For a generic value of the bias current b, there are strong deviations from the condition in Eq (2). In Fig 2B, we compare the value of the average output $\bar{y}$ with $\tilde{h} \equiv \sum_{c\in\{\mathrm{E},I\}} N_c \bar{w}_c \bar{x}_c + b$, and also plot the residual term $c = \frac{1}{NP}\sum_{i\mu}\delta w_i x_{i\mu}$, where we decomposed the weight vector components as $w_i = \bar{w}_c + \delta w_i$ for $c \in \{\mathrm{E}, \mathrm{I}\}$. The quantity c measures weight-rate correlations that are responsible for the cancelation of the $\mathcal{O}(\sqrt{N})$ bias.

The deviation from Eq (2), shown here for a rapidly decaying covariance of the form $C_{\mu\nu} = e^{-\frac{|\mu-\nu|}{2\tau^2}}$, has been previously described in the context of a target-based learning algorithm used to build E-I-separated rate and spiking models of neural circuits capable of solving input/output tasks [3]. In this approach, a randomly initialized recurrent network n_T is driven by a low dimensional signal z. Its currents are then used as targets to train the synaptic couplings of a second (rate or spiking) network n_S, in such a way that the desired output z can later be linearly decoded from the self-sustained activity of n_S. Each neuron of n_S has to independently learn an input/output mapping from firing rates x to currents y, using an on-line sign-constrained least square method. In the presence of an L2 regularization and a constant $b \propto \sqrt{N}$ external current, the on-line learning method typically converges onto a solution for the recurrent synaptic weights for which Eq (2) does not hold. As also shown in [3], in the peculiar case of a self-sustained periodic dynamics (in which case off-diagonal terms of the covariance matrix $C_{\mu\nu}$ do not vanish for large μ or ν) the two contributions $\tilde{h}$ and c scale approximately like $\sqrt{N}$ and cancel each other to produce an $\mathcal{O}(1)$ total average output $\bar{y} = \tilde{h} + c$. In the effort to build heterogeneous functional network models, the emergence of synaptic connectivity compatible with the balanced scaling thus depends on the statistics of incoming currents. Ad-hoc regularization can be avoided by adjusting external currents onto each neuron.

Power spectrum and synaptic distribution

The theory developed thus far applies to a generic covariance matrix C. To connect the spectral properties of C with the signal dynamics, we further assume the $x_{i\mu}$ to be N independent stationary discrete-time processes. In this case, $C_{\mu\nu} = C(\mu - \nu)$ is a matrix of Toeplitz type [39], leading to the following expression for the average minimal loss density in the $N \to \infty$ limit:

$$\epsilon = \frac{\sigma_y^2}{2\pi}\int_0^\pi d\phi \frac{\lambda(\phi)}{1 + \Delta\tilde{q}_w\lambda(\phi)}$$

with $\Delta\tilde{q}_w$ given by Eq (4). The function $\lambda(\phi)$ can be computed exactly in some cases (Methods: *Power spectrum and synaptic distribution*) and corresponds to the average power spectrum of the x and y stochastic processes. Fig 3 shows two representative input signals with Gaussian and exponential covariance matrix C (Fig 3A) and a comparison between the average power

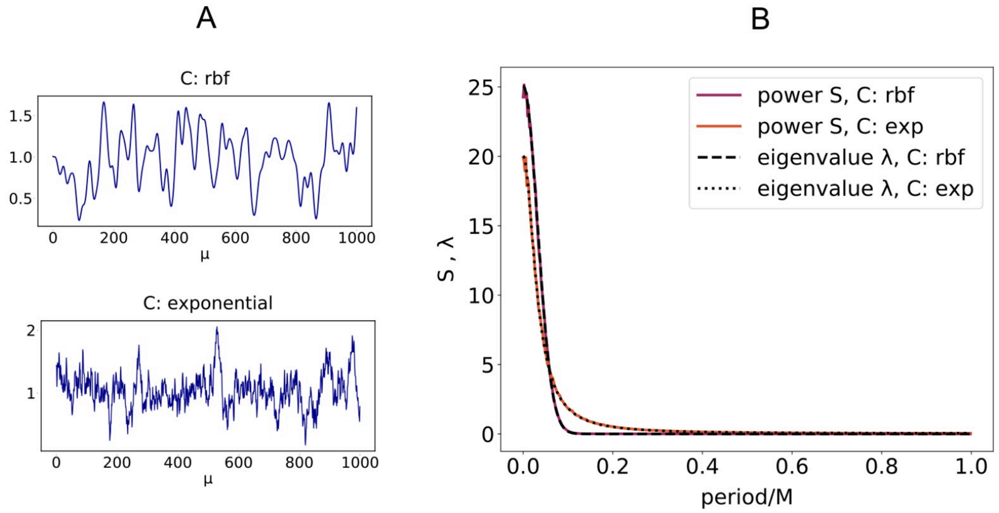

Fig 3. Eigenvalues of *C* and Fourier spectrum. A: Examples of excitatory input signals $x_{i\mu}$ ($i \in$ E) with two different covariance matrices *C*. Top: rbf covariance, $\tau = 10$. Bottom: exponential covariance $C_{\mu\nu} = e^{-\frac{|\mu-\nu|}{\tau}}$, $\tau = 10$. Parameters: $\bar{x}_{\mathrm{E}} = 1$, $\sigma_{\mathrm{E}} = 0.3$. B: Theoretical eigenvalue spectrum of *C* with $\tau = 10$ versus average power spectrum for positive wave numbers across $N = 2000$ independent processes with $P = 1000$ time steps.

https://doi.org/10.1371/journal.pcbi.1008536.g003

spectrum of the input and the analytical results for the eigenvalue spectrum of the matrix *C* (Fig 3B). From now on, we use the terms Gaussian or rfb (radial basis function) indistinguishably to denote the un-normalized Gaussian function $C_{\mu\nu} = e^{-\frac{(\mu-\nu)^2}{2\tau^2}}$.

As shown in Fig 4A in the case of input *x* and output *y* with rbf covariance, the squared norm of the optimal synaptic vector ***w*** (red curve) is in general a non-monotonic function of *α*, its maximum being attained at bigger values of *α* as the time constant *τ* increases. We also show the minimal loss density ϵ and the mean error ϵ_{err} for $\gamma = 0.1$. The curves in Fig 4A are the same for any ratio f_{E}: the use of an optimal bias current *b* cancels any asymmetry between

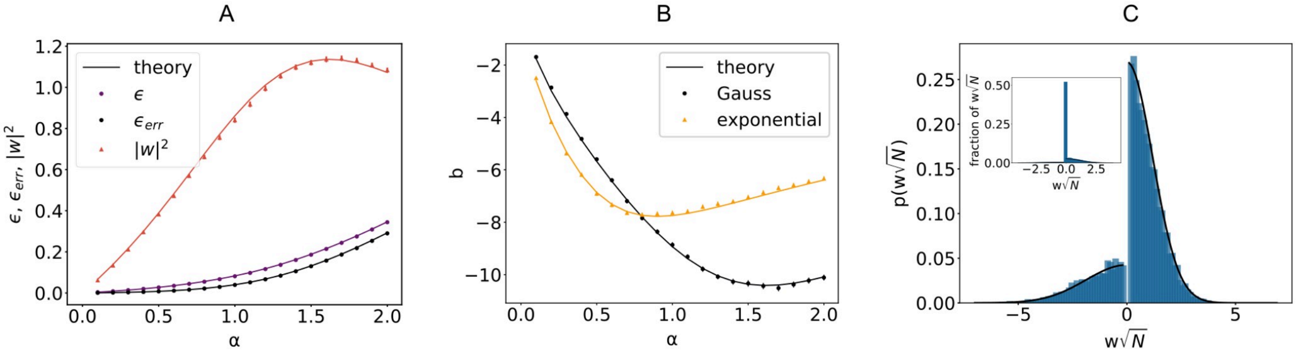

Fig 4. Learning temporally structured signals. A: Minimal loss ϵ, error ϵ_{err} and norm of the weight vector ***w*** as a function of the load *α* for a linear perceptron trained on a time-correlated signal. Covariance matrix *C* is of rbf type with $\tau = 2$. Parameters: $N = 1000$, $f_{\mathrm{E}} = 0.8$, $\gamma = 0.1$, $\bar{x}_{\mathrm{E}} = \bar{x}_I = \sigma_{\mathrm{E}} = \sigma_I = \bar{y} = \sigma_y = 1$. B: Optimal bias *b* for the two sets of signals with rbf (black curve) and exponential (yellow curve) covariance *C*, with $\tau = 2$. Theoretical curves show the value $I\sqrt{N} + \bar{y}$, where *I* has been computed from the saddle-point equations (Methods: *EC, Saddle-point equations*). Parameters as in A. Each point in A and B is an average across 50 samples. C: Probability density of non-zero synaptic weights $w_i\sqrt{N}$ of a linear perceptron with $N = 1000$, a fraction $f_{\mathrm{E}} = 0.8$ of excitatory weights, trained on $P = 600$ exponentially correlated input *x* and output *y*. The *δ* function in zero is omitted for better visualization. Parameters: $\tau = 10$, $\gamma = 0.1$, $\bar{x}_{\mathrm{E}} = \bar{x}_I = 1$, $\sigma_{\mathrm{I}} = 2\sigma_{\mathrm{E}} = 0.4$. The histogram is an average across 50 realizations of input/output signals. Inset: full histogram of synaptic weights $w_i\sqrt{N}$.

https://doi.org/10.1371/journal.pcbi.1008536.g004

E and I populations. For a finite γ, the minimal average loss ϵ for a given f_{E} decreases as either σ_{E} or σ_I increase. For a given set of parameters f_{E} and γ, the optimal bias b will in general depend on the load α and the structure of the covariance matrix C, as shown in Fig 4B.

Using the same analytical machinery employed for the calculation of the free energy Eq (3), the probability distribution of the typical weight w_i can be easily derived. This can be seen by employing a variant of the replica trick (Methods: *Distribution of synaptic weights*) that links the so-called *entropic part* of f to $\langle p(w_i)\rangle$, expressed in terms of the saddle-point values of the same (conjugated) overlap parameters employed thus far. Interestingly, the optimal bias b implies that half of the synapses are zero, irrespectively of f_{E} and the properties of the covariance matrix C. The probability density of the synaptic weights is composed of two truncated Gaussian densities with zero mean for the E and I components, plus a finite fraction $p_0 = 0.5$ of zero weights.

We show in Fig 4C the shape of the optimal weight distribution for a linear perceptron with 80% excitatory synapses, trained on exponentially correlated x and y and with a ratio $\sigma_{\mathrm{I}}/\sigma_{\mathrm{E}} = 2$. It is interesting to note that, in the presence of an optimal external current, both the means of the Gaussian components and the fraction of silent synapses do not depend on the specific properties of input and output signals.

The shape of the synaptic distribution appeared in previous studies both in the binary [8, 11, 13] and linear perceptron [15]. In the linear case with only excitatory synapses [15], for a fixed bias $b = \sqrt{N}$, the fraction of zero E weights is larger than 0.5 at criticality. It generally depends on input parameters and the load in the error region $\alpha \leq \alpha_c$. Let us also mention that a similar property is also apparent in the binary perceptron, where the scale of the typical solutions is set by robustness [13] to input and output noise. For weights $w_i = \mathcal{O}(1/\sqrt{N})$, the sparsity of critical solutions generically depends on properties of E and I inputs. For weights of $\mathcal{O}(1/N)$, robust solutions have a fraction of zero E weights generically larger than 0.5 [6, 11]. When inhibitory synapses are added, their weights are less sparse [11]. Interestingly, in the case without robustness, half of the E and I weights are zero at critical capacity for all $f_{\mathrm{E}} \geq 0.5$.

The dynamic properties of input/output mappings affect the shape of the weight distribution in a computable manner. As an example, in a linear perceptron with non-negative synapses, the explicit dependence of the variance of the weights on the input and output autocorrelation time constant is shown in Fig 5A for various loads α. Previous work considered an analog perceptron with purely excitatory weights as a model for the graded rate response of Purkinje cells in the cerebellum [15]. In the presence of heterogeneity of synaptic properties across cells, a larger variance in their synaptic distribution is expected to be correlated with high frequency temporal fluctuations in input currents. Analogously, the auto-correlation of the typical signals being processed sets the value of the constant external current that a neuron must receive in order to optimize its capacity.

When the input and output have different covariance matrices $C^x \neq C^y$, a joint diagonalization is not possible in general (Methods: *EC, Energetic part*). We can nevertheless write an expression (Eq (23)) that holds when input and output patterns are defined on a ring (with periodic boundary conditions) and use it as an approximation for the general case. Fig 5B shows good agreement between numerical experiment and theoretical predictions for the error ϵ_{err} and the squared norm of the synaptic weight vector $\boldsymbol{w}$, when input and output processes have two different time-constants τ_x and τ_y.

Sample covariance and dimensionality

In the discussion thus far, we assumed independence across the “spatial” index i in the input. It is often the case for input signals to be confined to a manifold of dimension smaller than N,

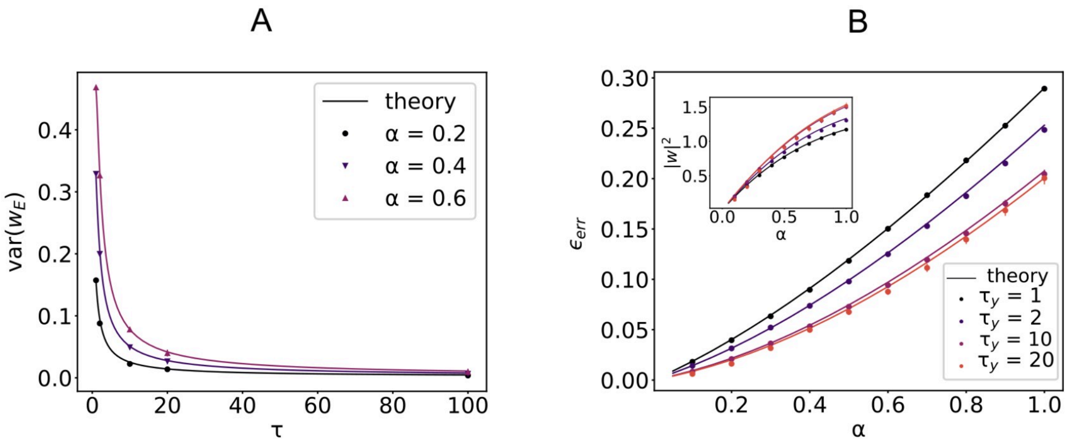

Fig 5. Input/Output time constants and learning performance. A: Variance of synaptic weights (f_E = 1) for a linear perceptron of dimension N = 1000 trained on rbf-correlated signals with increasing time constant τ for three different values of the load α. Parameters: γ = 0.1, $\bar{x}_E = \bar{x}_I = \sigma_E = \sigma_I = \bar{y} = \sigma_y = 1$. B: Average error ϵ_{err} in the case where input and output signals have two different covariance matrices, for increasing time constant τ_y of the output signal y. Parameters: N = 1000, f_E = 0.8, γ = 0.1, $\bar{x}_E = \bar{x}_I = \bar{y} = \sigma_y = 1$, $\sigma_I = 2\sigma_E = 0.6$, C^x rbf with τ_x = 1, C^y rbf with various values of τ_y. Inset: norm of the weight vector $\boldsymbol{w}$. Full lines show analytical results. Points are averages across 50 samples.

https://doi.org/10.1371/journal.pcbi.1008536.g005

a feature that can be described by various dimensionality measures, some of which rely on principal component analysis [40, 41]. In order to relax the independence assumption, we build on a framework originally introduced in the theory of spin glasses with orthogonal couplings [42–44] and further developed in the context of adaptive Thouless-Anderson-Palmer (TAP) equations [45–47]. In the TAP formalism, a set of mean-field equations is derived for a given instance of the random couplings (in our case, for a fixed input/output set). In its adaptive generalization [46], the structure of the TAP equations depends on the specific data distribution, in such a way that averaging the equations over the random couplings yields the same results of the replica approach. Here, following previous work in the context of information theory of linear vector channels and binary perceptrons [48–51], we employ an expression for an ensemble of rectangular random matrices and use the replica method to average over the input X and output y.

Let us write the input matrix $(X)_{i\mu} = \bar{x}_i + \sigma_i \xi_{i\mu}$, with $\xi = USV^T$, S being the matrix of singular values. To analyze the properties of the typical case, we start from a generic singular value distribution S and consider i.i.d. output y_μ. In calculating the cumulant generating function $Z_{\xi,\delta y}$, we perform a homogeneous average across the left and right principal components U and V (Methods: *SC, Energetic part*). Calling $\rho_{\xi\xi^T}(\lambda)$ the eigenvalue distribution of the sample covariance matrix $\xi\xi^T$, we can express $Z_{\xi,\delta y}$ in terms of a function $\mathcal{G}_{\xi,\delta y}$ of an enlarged set of overlap parameters, which depends on the so-called Shannon transform [52] of $\rho_{\xi\xi^T}(\lambda)$, a quantity that measures the capacity of linear vector channels. The resulting self-consistent equations, which describe the statistical properties of the synaptic weights w_i, are expressed in terms of the Stieltjes transform of $\rho_{\xi\xi^T}(\lambda)$, an important tool in random matrix theory [53] (Methods: *SC, Saddle-point equations*).

We show the validity of the mean-field approach by employing two different data models for the input signals. In the first example, valid for $\alpha \leq 1$, all the P vectors $\boldsymbol{\xi}_\mu$ are orthogonal to each other. This yields an eigenvalue distribution of the simple form $\rho(\lambda) = \alpha\delta(\lambda - 1) + (1 - \alpha)\delta(\lambda)$, for which the function $\mathcal{G}_{\xi,\delta y}$ can be computed explicitly [51]. Additionally, we use a synthetic model where we explicitly set the singular value spectrum of ξ to be $s(\alpha) = \chi e^{-\frac{\alpha^2}{2\sigma_x^2}}$, with χ

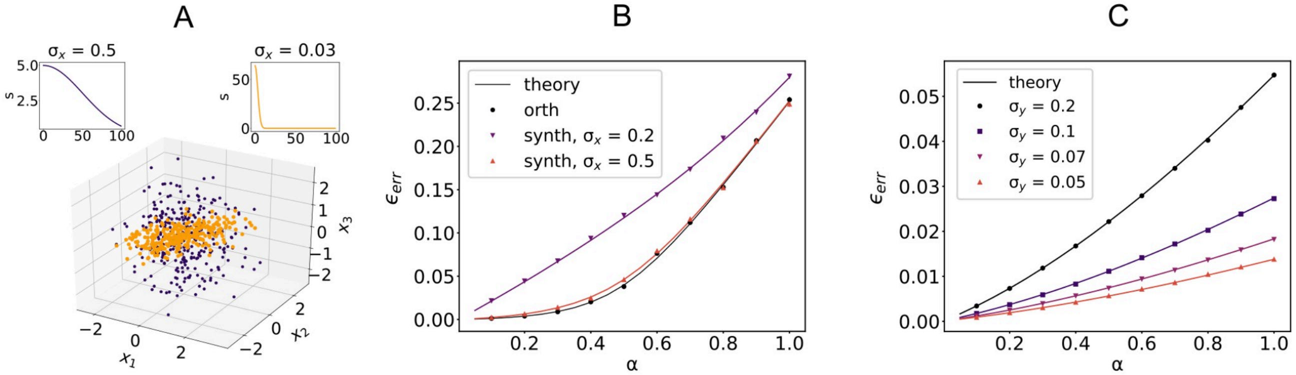

Fig 6. Sample-based PCA and learning performance. A: First three components of inputs $\boldsymbol{\xi}_\mu$ with Gaussian singular value spectrum s for two different values of σ_x (color coded top panels). Parameters: $N = 100$, $P = 300$. B: Average error ϵ_{err} for three different singular value spectra of the input sample covariance matrix: orthogonal model and Gaussian model with increasing σ_x (see main text for definition of σ_x). Outputs are i.i.d Gaussian. Parameters: $N = 1000$, $f_E = 0.8$, $\gamma = 0.1$, $\bar{x}_E = \bar{x}_I = \bar{y} = \sigma_y = 1$, $\sigma_I = 2\sigma_E = 0.6$. B: Average error ϵ_{err} for input with orthogonal-type covariance and output y with rbf-type covariance with decreasing σ_y (see main text for the definition of σ_y). All remaining parameters as in A. Full lines show analytical results. Points are averages across 50 samples.

https://doi.org/10.1371/journal.pcbi.1008536.g006

a normalization factor ensuring matrix ξ has unit variance. The shape of the singular value spectrum s controls the spread of the data points $\boldsymbol{\xi}_\mu$ in the N-dimensional input space, as shown in Fig 6A. As shown in Fig 6B for i.i.d Gaussian output, learning degrades as σ_x decreases, since inputs tend to be confined to a lower dimensional subspace rather than being equally distributed along input dimensions.

For N large enough (in practice, for $N \gtrsim 500$), the statistics of single cases is well captured by the equations for the average case (self-averaging effect). To get a mean-field description for a single case, where a given input matrix X is used, we further assume we have access to the linear expansion c_μ of the output y in the set $\{\boldsymbol{v}_\mu\}$ of the columns of the V matrix, namely $y = \bar{y} + \sigma_y Vc$. The calculation can be carried out in a similar way and yields, for the average regression loss, the following result:

$$\epsilon = \frac{\alpha}{2}\sigma_y^2\tilde{\Lambda}_w\left\langle \frac{\lambda^y}{\lambda^x + \tilde{\Lambda}_w}\right\rangle_{\lambda^x,\lambda^y} \tag{6}$$

The average in Eq (6) is computed over the eigenvalues λ^x of the sample covariance matrix, which correspond to the PCA variances, and $\lambda^y_\mu = c^2_\mu$ (Methods: *SC, Energetic part*). The quantity $\tilde{\Lambda}_w$ can be computed from a set of self-consistent equations that link the order parameter $\Delta\tilde{q}_w$ and the first two moments of the synaptic distribution. To better understand the role of the parameter $\tilde{\Lambda}_w$, it is instructive to compare Eq (6) with the corresponding result for unconstrained weights, which can be derived from the pseudo-inverse solution $w^* = (\xi\xi^T + \gamma)^{-1}\,\xi y$ (Methods: *SC, i.i.d. and unconstrained cases*). The average loss is:

$$\epsilon_{\text{unc}} = \frac{\alpha}{2}\sigma_y^2\gamma\left\langle \frac{\lambda^y}{\lambda^x + \gamma}\right\rangle_{\lambda^x,\lambda^y} \tag{7}$$

Comparing Eqs (7) and (8), we find that $\tilde{\Lambda}_w$ acts as an implicit regularization in the sign-constrained case. The mean-field theory is thus carried out through a diagonalization over independent contributions along the components $\boldsymbol{v}_\mu$, with prescribed input and output variances λ^x and λ^y, respectively. The coupling between different components, induced by the averages $\langle\cdot\rangle_{x,y}$ and the sign-constraints, is incorporated in the effective regularization $\tilde{\Lambda}_w$, acting on each component equally, that depends only on the structure of the input x (see Eqs (56) and (67) in Methods)).

In Fig 6C, we show results when the dimensionality of the output y along the (temporal) components of the input is modulated by taking $c(\alpha) = e^{-\frac{\alpha^2}{2\sigma_y^2}}$. The perceptron performance improves as the output signals spreads out across multiple components $\boldsymbol{v}_\mu$. The case of i.i.d. output is recovered by taking $c_\mu = 1$.

Discussion

In this work, I investigated the properties of optimal solutions of a linear perceptron with sign-constrained synapses and correlated input/output signals, thus providing a general mean-field theory for constrained regression in the presence of correlations. I treated both the case of known ensemble covariances and the case where the sample covariance is given. The latter approach, built on a rotationally invariant assumption, allowed to link the regression performance to the input and output statistical properties expressed by principal component analysis.

I provided the general expression of the weight distribution for regularized regression and found that half of the weights are set to zero, irrespectively of the fraction of excitatory weights, provided the bias is optimized. The shape of the synaptic distribution has been previously described in the binary perceptron with independent input at critical capacity, as well as in the theory of compressed sensing [54]. I elucidated the role of the optimal bias current and its relation to the optimal capacity and the scaling of the solution weights. This analysis also shed light on the structural properties of synaptic matrices that emerge when target-based methods are used for building biologically plausible functional models of rate and spiking networks.

The theory presented in this work is relevant in the effort of establishing quantitative comparisons between the synaptic profile of neural circuits involved in temporal processing of dynamic signals, such as the cerebellum [55–57], and normative theories that take into account the temporal and geometrical complexity of computational tasks. On the other hand, the construction of progressively more biologically plausible models of neural circuits calls for normative theories of learning in heterogeneous networks, which can be coupled to dynamic mean-field analysis of E-I separated circuits [24, 25, 58].

As shown in this work, the interaction between correlational structure of input signals, synaptic metabolic cost and constant external current shapes the distribution of synaptic weights. In this respect, the results presented here offer a first approximation (static linear input-output associations) to account for heterogeneities of the fraction between E and I inputs to single cells in local circuits. Even though a heterogeneous linear neuron is capable of memorizing $N/2$ associations without error for any E/I ratio, the optimal bias does depend on f_E, its minimal value being attained for $f_E = 0.5$. Input current in turn sets the neuron's operating regime and its input/output properties. Moreover, trading memorization accuracy (small output error ϵ_{err}) for smaller weights (small $|\boldsymbol{w}|^2$) could be beneficial when synaptic costs are considered ($\gamma > 0$). It is therefore likely that, for an optimality principle of the 80/20 ratio to emerge from purely representational considerations, dynamical and metabolic effects should be examined all together.

The importance of a theory of constrained regression with realistic input/output statistics goes beyond the realm of neuroscience. Non-negativity is commonly required to provide interpretable results in a wide variety of inference and learning problems. Off-line and on-line least-square estimation methods [59, 60] are also of great practical importance in adaptive control applications, where constraints on the parameter range are usually imposed by physical plausibility.

In this work, I assumed statistical independence between inputs and outputs. For the sake of biological plausibility, it would be interesting to consider more general input-output correlations for regression and binary discrimination tasks. The classical model for such correlations is provided by the so-called teacher-student (TS) approach [61], where the output y is generated by a deterministic parameter-dependent transformation of the input x, with a structure similar to the trained neural architecture. The problem of input/output correlations is deeply related to the issue of optimal random nonlinear expansion both in statistical learning theory [62, 63] and theoretical neuroscience [41, 64], with a history dating back to the Marr-Albus theory of pattern separation in cerebellum [65]. In a recent work, [28] introduced a promising generalization of TS, in which labels are generated via a low-dimensional latent representation, and it was shown that this model captures the training dynamics in deep networks with real world datasets.

A general analysis that fully takes into account spatio-temporal correlations in network models could shed light on the emergence of specific network motifs during training. In networks with non-linear dynamics, the mathematical treatment quickly gets challenging even for simple learning rules. In recent years, interesting work has been done to clarify the relation between learning and network motifs, using a variety of mean-field approaches. Examples are the study of associative learning in spin models [8] and the analysis of motif dynamics for simple learning rules in spiking networks [66]. Incorporating both the temporal aspects of learning and neural cross-correlations in E-I separated models with realistic input/output structure is an interesting topic for future work.

Methods

Replica formalism: Ensemble covariance matrix (EC)

Using the Replica formalism [67], the free energy density is written as:

$$-\beta f = \frac{1}{N}\lim_{n\to 0}\frac{\partial}{\partial n}\log\langle Z^n\rangle_{x,y} \tag{8}$$

The function Z^n can be computed by considering a finite number n of replicas of the vector $\boldsymbol{w}$ and subsequently taking a continuation $n \in \mathbb{R}$. The introduction of n replicas allows to factorize $\langle Z^n\rangle_{x,y}$ over individual weights w_i, at the cost of coupling different replicas after the averages over the x and y are performed. Introducing a small set of *overlap order parameters*, factorization across replicas is restored, so that in the large N limit the replicated partition function takes the form $\langle Z^n\rangle_{x,y} = e^{-\beta N n f}$. In the following, we will usually drop the subscript in the average $\langle\cdot\rangle_{x,y}$.

To simplify the formulas, we introduce the $\mathcal{O}(1)$ weights $J_i = \sigma_i\sqrt{N}w_i$. In terms of these rescaled variables, the loss function in Eq (1) takes the form:

$$E(w;\gamma,\xi,y) = \frac{1}{2}\sum_{\mu=1}^{P}\left(\sum_{i=1}^{N}\frac{J_i}{\sqrt{N}}\xi_{i\mu} + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\bar{x}_i}{\sigma_i}J_i + I\sqrt{N} - y_\mu\right)^2 + \frac{\gamma}{2}\sum_{i=1}^{N}\frac{J_i^2}{\sigma_i^2} \tag{9}$$

by virtue of $x_{i\mu} = \bar{x}_i + \sigma_i\xi_{i\mu}$. We proceed by inserting the definitions $M^a = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\bar{x}_i}{\sigma_i}J_{ia} + I\sqrt{N}$ and $\Delta_{\mu a} = \sum_{i=1}^{N}\xi_{i\mu}\frac{J_{ia}}{\sqrt{N}} - \sigma_y\delta y_\mu$ with the aid of appropriate δ functions. The averaged replicated partition function $\langle Z^n \rangle$ is:

$$\begin{aligned}\langle Z^n \rangle &= \int \prod_a d\mu(J_a) \int \prod_{\mu a} \frac{d\Delta_{\mu a} du_{\mu a}}{2\pi} \prod_a \frac{dM^a d\hat{M}^a}{2\pi/\sqrt{N}} Z_{\xi,\delta y} \\ &\quad e^{\sum_a \hat{M}^a \left(\sqrt{N}M^a - \sum_i \frac{\bar{x}_i}{\sigma_i} J_{ia} - NI\right) - i\sum_{\mu a} u_{\mu a}\Delta_{\mu a} - \frac{\beta}{2}\sum_{\mu a}(\Delta_{\mu a} + M^a - \bar{y})^2 - \frac{\beta\gamma}{2}\sum_{ia}\frac{J_{ia}^2}{\sigma_i^2}}\end{aligned} \tag{10}$$

where:

$$Z_{\xi,\delta y} = \left\langle e^{i\sum_{\mu a} u_{\mu a}\left(\sum_i \xi_{i\mu}\frac{J_{ia}}{\sqrt{N}} - \sigma_y \delta y_\mu\right)} \right\rangle_{\xi,\delta y} \tag{11}$$

In Eq 10, we used a Fourier expansion of the δ functions and introduced the real variables $u_{\mu a}$ as conjugate variables for $\Delta_{\mu a}$. Analogously, we employed the purely imaginary $\hat{M}^a$ for the variables M^a. Once the the average is carried out, second cumulants of ξ and δy get coupled to replica mixing terms of the form $J_{ia} J_{ib}$, which can be dealt with by introducing appropriate overlap order parameters $Nq_w^{ab} = \sum_{i=1}^N J_{ia}J_{ib}$ with the use of $n(n + 1)/2$ additional δ functions, together with their conjugate variables $\hat{q}_w^{ab}$. Cumulants of higher order will not contribute to the expression in the large N limit. Expanding the δ functions for the overlap parameters we get the expression

$$\langle Z^n \rangle = \int \prod_{a\leq b} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi/N} \prod_a \frac{dM^a d\hat{M}^a}{2\pi/\sqrt{N}} e^{\sum_a \hat{M}^a(\sqrt{N}M^a - NI) - N\sum_{a\leq b}\hat{q}^{ab}q^{ab} + N\mathcal{G}_S + \alpha N\mathcal{G}_E} \tag{12}$$

where the two contributions $\mathcal{G}_E$ and $\mathcal{G}_S$, respectively called *energetic* and *entropic* part, will be calculated separately in the following for ease of exposition. Owing to the convexity of the regression problem, we use a Replica Symmetry (RS) [67] ansatz $q_w^{ab} = q_w + \delta_{ab}\Delta q_w$ and $M^a = M$.

EC, Entropic part. The total volume of configurations $\boldsymbol{w}_a$ for fixed values of the overlap parameters is given by the *entropic part*:

$$e^{N\mathcal{G}_S} = \int \prod_a d\mu(J_a) e^{\sum_{a\leq b}\hat{q}_w^{ab}\sum_i J_{ia}J_{ib} - \frac{\beta\gamma}{2}\sum_{ia}\frac{J_{ia}^2}{\sigma_i^2} - \sum_a \hat{M}^a \sum_i \eta_i J_{ia}} \tag{13}$$

where we called $\eta_c = \frac{\bar{x}_c}{\sigma_c}$, with $c \in \{\mathrm{E, I}\}$, and $\eta_i = \eta_{\mathrm{E}}$ ($\eta_i = \eta_I$) if $i \in \mathrm{E}$ ($i \in \mathrm{I}$). Using the RS ansatz $\hat{q}_w^{ab} = \hat{q}_w - \delta_{ab}\frac{\hat{q}_w + \Delta\hat{q}_w}{2}$ and $\hat{M}^a = \hat{M}$, we get:

$$e^{N\mathcal{G}_S} = \int \prod_a d\mu(J_a) e^{-\frac{1}{2}\sum_i \left(\Delta\hat{q}_w + \frac{\beta\gamma}{\sigma_i^2}\right)\sum_a J_{ia}^2 + \frac{\hat{q}_w}{2}\sum_i \sum_{ab} J_{ia}J_{ib} - \hat{M}\sum_{ia}\eta_i J_{ia}} \tag{14}$$

Using the explicit definition of the measure $d\mu(J) \propto \prod_{i\in\mathrm{E}}\theta(J_i)dJ_i \prod_{k\in I}\theta(-J_k)dJ_k$, one has, up to constant terms:

$$\mathcal{G}_S = \sum_{c\in\{\mathrm{E},I\}} f_c \log \int_0^\infty \prod_a dJ_a e^{-\frac{1}{2}\left(\Delta\hat{q}_w + \frac{\beta\gamma}{\sigma_c^2}\right)\sum_a J_a^2 + \frac{\hat{q}_w}{2}\sum_{ab} J_a J_b - s_c \eta_c \hat{M} \sum_a J_a} \tag{15}$$

where we introduced the notations $f_{\mathrm{I}} = 1 - f_{\mathrm{E}}$ and $s_{\mathrm{E}} = -s_{\mathrm{I}} = 1$. In order to disentangle the term $\Sigma_{ab} J_a J_b = (\Sigma_a J_a)^2$, we employ the so-called *Hubbard-Stratonovich* transformation

$e^{\frac{b^2}{2}} = \int Dz e^{bz}$, where $Dz = dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$. Taking the limit $n \to \mathbf{0}$ one gets:

$$\mathcal{G}_S = \sum_{c \in \{\mathrm{E}, I\}} f_c \int Dz \log \int_0^\infty dJ e^{-\frac{J^2}{2}\left(\Delta\hat{q}_w + \frac{\beta\gamma}{\sigma_c^2}\right) + s_c\left(z\sqrt{\hat{q}_w} - \eta_c \hat{M}\right)J} \tag{16}$$

EC, Energetic part. In order to compute the *energetic* part, we first need to evaluate the average with respect to ξ and δy in Eq (11). Performing the two Gaussian integrals we get:

$$Z_{\xi,\delta y} = e^{-\frac{1}{2}\sum_{\mu\nu}\sum_{ab} q_w^{ab} u_{\mu a} u_{\nu b} C^x_{\mu\nu} - \frac{\sigma_y^2}{2}\sum_{\mu\nu}\sum_{ab} u_{\mu a} u_{\nu b} C^y_{\mu\nu}} \tag{17}$$

from which:

$$\begin{aligned} e^{\alpha N \mathcal{G}_E} \quad &= \int \prod_{\mu a} \frac{d\Delta_{\mu a} du_{\mu a}}{2\pi} e^{-\frac{\beta}{2}\sum_{\mu a} \Delta^2_{\mu a} - \frac{1}{2}\sum_{\mu\nu}\sum_{ab} q_w^{ab} u_{\mu a} u_{\nu b} C^x_{\mu\nu}} \\ &\quad e^{-\frac{\sigma_y^2}{2}\sum_{\mu\nu}\sum_{ab} u_{\mu a} u_{\nu b} C^y_{\mu\nu} + i \sum_{\mu a} u_{\mu a}\left(M^a - \bar{y} - \Delta_{\mu a}\right)} \end{aligned} \tag{18}$$

where we performed a translation $\Delta_{\mu a} + M^a - \bar{y} \to \Delta_{\mu a}$. In the special case $C^x = C^y \equiv C$, we can use $C = V \Lambda V^T$ to jointly rotate $\mathbf{\Delta}_a \to V\,\mathbf{\Delta}_a$ and $\boldsymbol{u_a} \to V\boldsymbol{u_a}$, thus leaving scalar products invariant. By doing so, we obtain, within the RS ansatz:

$$\begin{aligned} e^{\alpha N \mathcal{G}_E} \quad &= \int \prod_{\mu a} \frac{d\Delta_{\mu a} du_{\mu a}}{2\pi} e^{-\frac{\beta}{2}\sum_{\mu a} \Delta^2_{\mu a} - \frac{1}{2}\sum_{\mu}\sum_{ab} (q_w + \delta_{ab}\Delta q_w) u_{\mu a} u_{\mu b} \lambda_\mu} \\ &\quad e^{-\frac{\sigma_y^2}{2}\sum_{\mu}\sum_{ab} u_{\mu a} u_{\mu b} \lambda_\mu + i \sum_{\mu a} \zeta_\mu u_{\mu a}\left(M - \bar{y} - \Delta_{\mu a}\right)} \end{aligned} \tag{19}$$

where $\zeta_\mu = \sum_\nu V_{\mu\nu}$. Using a Hubbard-Stratonovich transformation on the term $\sum_{ab} u_{\mu a} u_{\mu b}$, after some algebra, we obtain:

$$\alpha N \mathcal{G}_E = -\frac{1}{2}\sum_\mu \log\left(1 + \beta \Delta q_w \lambda_\mu\right) - \frac{\beta}{2}\sum_\mu \frac{(q_w + \sigma_y^2)\lambda_\mu + \zeta_\mu^2 (M - \bar{y})^2}{1 + \beta\Delta q_w \lambda_\mu} \tag{20}$$

Observing that the free energy only depends on M through the term $(M - \bar{y})^2$ in $\mathcal{G}_E$, we conveniently eliminate the quantities ζ_μ at this stage, using the simple saddle-point relation

$$M = \bar{y} \tag{21}$$

thus getting:

$$\mathcal{G}_E = -\frac{1}{2}\langle \log(1 + \beta\Delta q_w \lambda)\rangle_\lambda - \frac{\beta}{2}\left(q_w + \sigma_y^2\right)\left\langle \frac{\lambda}{1 + \beta\Delta q_w \lambda}\right\rangle_\lambda \tag{22}$$

The brackets $\langle\cdot\rangle_\lambda$ in Eq (22) stand for an average over the eigenvalue distribution $\rho(\lambda)$ of C in the $N \to \infty$ limit, assuming self-averaging. A similar expression for $\mathcal{G}_E$ was previously derived in [34] for spherical weights, i.e. $\sum_{i=1}^N w_i^2 = 1$, in the presence of outputs y_μ generated by a *teacher* linear perceptron. To map Eq (45) in [34] to Eq (22), one substitutes $(1 - q) \to \Delta q_w$ (observing that $q^{aa} = 1$ thanks to the spherical constraint) and sets $R = 0$, since the learning task only involves patterns memorization.

When $C^x \neq C^y$, we can derive a similar expression under the assumption of a ring topology in pattern space (corresponding to periodic boundary conditions in the index μ): in this case, both covariance matrices are circulant and may be jointly diagonalized by discrete Fourier transform [33, 34]. In the main text, we show that the expression

$$\alpha \mathcal{G}_E = -\frac{1}{2N}\sum_{\mu} \log\left(1+\beta\Delta q_w \lambda^x_\mu\right) - \frac{\beta}{2N}\sum_{\mu}\frac{q_w\lambda^x_\mu + \sigma^2_y\lambda^y_\mu}{1+\beta\Delta q_w\lambda^x_\mu} \tag{23}$$

yields good results also when C^x and C^y are covariance matrices of stationary discrete-time processes.

EC, Saddle-point equations. All in all, the free energy density in the saddle-point approximation is:

$$\begin{aligned} -\beta f \quad &= -\hat{M}I + \frac{\Delta\hat{q}_w}{2}(\Delta q_w + q_w) - \frac{\hat{q}_w \Delta q_w}{2} \\ &-\frac{1}{2N}\sum_{\mu}\log\left(1+\beta\Delta q_w\lambda^x_\mu\right) - \frac{\beta}{2N}\sum_{\mu}\frac{q_w\lambda^x_\mu+\sigma^2_y\lambda^y_\mu}{1+\beta\Delta q_w\lambda^x_\mu} + \\ &\sum_{c\in\{\mathrm{E},I\}} f_c \int Dz \log \int_0^\infty dJ e^{-\frac{J^2}{2}\left(\Delta\hat{q}_w + \frac{\beta\gamma}{\sigma^2_c}\right) + s_c\left(z\sqrt{\hat{q}_w} - \eta_c \hat{M}\right)J} \end{aligned} \tag{24}$$

The saddle-point equations stemming from the entropic part can be written as:

$$\Delta q_w = \langle\langle J^2\rangle_J\rangle_z - \langle\langle J\rangle_J^2\rangle_z \tag{25}$$

$$q_w = \langle\langle J\rangle_J^2\rangle_z \tag{26}$$

$$I + \sum_{c\in\{\mathrm{E},I\}} \eta_c \langle\langle J\rangle_J\rangle_z = 0 \tag{27}$$

where the averages $\langle\cdot\rangle_J$ and $\langle\cdot\rangle_z$ in Eqs (25)–(27) are taken with respect to the mean-field distribution of the J weights:

$$p(J;z) \quad \propto \sum_{c\in\{\mathrm{E},I\}} f_c p_c(J;z) \tag{28}$$

$$p_c(J;z) \quad \propto \theta(s_c J) e^{-\frac{J^2}{2}\left(\Delta\hat{q}_w + \frac{\beta\gamma}{\sigma_c}\right) + \left(z\sqrt{\hat{q}_w} - \eta_c\hat{M}\right)J} \tag{29}$$

where z is a standard normal variable and θ is the Heaviside function: $\theta(x) = 1$ when $x > 0$ and 0 otherwise. Eq (25) is obtained by differentiating Eq (24) with respect to $\hat{q}_w$ and then performing an integration by part in z. Eq (26) is easily obtained by subtracting Eq (25) from the saddle-point condition over $\Delta\hat{q}_w$, while Eq (27) originates from the derivative w.r.t. $\hat{M}$.

In the $\beta \to \infty$ limit, the unicity of solution for $\gamma > 0$ implies that $\Delta q_w \to 0$. We therefore use the following scalings for the order parameters:

$$\beta\Delta q_w = \Delta\tilde{q}_w \tag{30}$$

$$\hat{q}_w = \beta^2 C \tag{31}$$

$$\Delta\hat{q}_w = \beta A \tag{32}$$

$$\hat{M} = \beta B\sqrt{C} \tag{33}$$

while $q_w = \mathcal{O}(1)$. In this scaling, Eqs (25)–(27) take the form:

$$\Delta\tilde{q}_w = \sum_{c\in\{\mathrm{E,I}\}} \frac{f_c}{A+\frac{\gamma}{\sigma_c^2}} H(s_c\eta_c B) \tag{34}$$

$$\frac{q_w}{C} = \sum_{c\in\{\mathrm{E,I}\}} \frac{f_c}{\left(A+\frac{\gamma}{\sigma_c^2}\right)^2}\left((1+\eta_c^2B^2)H(s_c\eta_c B) - s_c\eta_c BG(\eta_c B)\right) \tag{35}$$

$$\frac{I}{\sqrt{C}} = \sum_{c\in\{\mathrm{E,I}\}} \frac{f_c}{A+\frac{\gamma}{\sigma_c^2}}\left(\eta_c^2 BH(s_c\eta_c B) - s_c\eta_c G(\eta_c B)\right) \tag{36}$$

where $G(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ and $H(x) = \int_x^\infty Dz$. The two remaining saddle-point equations are:

$$C = \frac{1}{N}\sum_\mu \lambda_\mu^x \frac{q_w\lambda_\mu^x + \sigma_y^2\lambda_\mu^y}{(1+\Delta\tilde{q}_w\lambda_\mu^x)^2} \tag{37}$$

$$A = \frac{1}{N}\sum_\mu \frac{\lambda_\mu^x}{1+\Delta\tilde{q}_w\lambda_\mu^x} \tag{38}$$

Optimizing *f* with respect to the bias $b = I\sqrt{N}$ immediately implies $B = 0$, by virtue of Eq (33), and greatly simplifies the saddle-point equations. Using the scaling assumptions Eqs (30)–(33) together with the saddle-point Eqs (34)–(38), we get Eq (4) in the main text, that is valid for any α for $\gamma > 0$. In the unregularized case ($\gamma = 0$), it describes solutions in the error regime $\alpha > \alpha_c$. The optimal bias b can be computed by $I\sqrt{N}$ using Eq (36), that is valid up to an $\mathcal{O}(1)$ term equal to $\bar{y}$ (Fig 4B). Keeping only the leading terms in the limit $\beta \to \infty$, Eq (24) can be written as:

$$\begin{aligned} -\beta f \quad &= -\beta B\sqrt{C}I + \frac{\beta}{2}Aq_w - \frac{\beta C}{2}\Delta\tilde{q}_w \\ &\quad -\frac{\beta}{2N}\sum_\mu \frac{q_w\lambda_\mu^x + \sigma_y^2\lambda_\mu^y}{1+\Delta\tilde{q}_w\lambda_\mu^x} + \\ &\quad \frac{\beta C}{2}\sum_{c\in\{E,I\}} \frac{f_c}{A+\frac{\gamma}{\sigma_c^2}}\left((1+\eta_c^2B^2)H(s_c\eta_c B) - s_c\eta_c BG(\eta_c B)\right) \end{aligned} \tag{39}$$

From the definition of the free energy density $-\beta N \boldsymbol{f} = \langle \log\int d\mu(w)e^{-\beta E}\rangle$, one has that $\frac{\langle E\rangle}{N} = \partial_\beta(\beta f)$. Using Eq (39) and the relevant saddle-point equations, the expression for the average minimal energy density is then:

$$\epsilon = \frac{\sigma_y^2}{2N}\sum_\mu \frac{\lambda_\mu^y}{1+\Delta\tilde{q}_w\lambda_\mu^x} \tag{40}$$

Also, noting that $\partial_\gamma E = \frac{N}{2}\sum_{i=1}^N w_i^2$, we can compute the average squared norm of the weights $v = \sum_{i=1}^N \langle w_i^2\rangle$ by $v = 2\partial_\gamma f$. We thus obtain:

$$v = C\sum_{c\in\{\mathrm{E,I}\}} \frac{f_c}{\sigma_c^2\left(A+\frac{\gamma}{\sigma_c^2}\right)^2}\left((1+\eta_c^2B^2)H(s_c\eta_c B) - s_c\eta_c BG(\eta_c B)\right) \tag{41}$$

The error $\epsilon_{err} = \frac{1}{2N}\langle |X^T\boldsymbol{w} + \boldsymbol{b} - \boldsymbol{y}|^2\rangle$ can be then computed by $\epsilon_{err} = \epsilon - \frac{\gamma}{2}\nu$.

Distribution of synaptic weights

The synaptic weight distribution appearing in Eqs (28) and (29) can be obtained using a variant of the replica trick [6, 67]. Using the expression $Z^{-1} = \lim_{n \to 0} Z^{n-1}$, the density of excitatory weights can be written as:

$$p(w_{\mathrm{E}}) = \lim_{n\to 0} \int \prod_a d\mu(w_a)\delta(w_{11} - w_{\mathrm{E}})e^{-\beta\sum_a E(w_a)} \tag{42}$$

where we picked the first E weight in the first replica w_{11} without loss of generality. The calculation proceeds along the same lines as for the entropic part above, since the energetic part does not depend on $\boldsymbol{w}_a$ explicitely. Isolating the first replica and taking the limit $n \to 0$, one gets the expression

$$p(J_{\mathrm{E}}) = \theta(J_{\mathrm{E}}) \int Dz \frac{e^{-\frac{J_{\mathrm{E}}^2}{2}\left(\Delta\hat{q}_w + \frac{\beta\gamma}{\sigma_{\mathrm{E}}^2}\right) + \left(z\sqrt{\hat{q}_w} - \eta_{\mathrm{E}}\hat{M}\right)J_{\mathrm{E}}}}{\int_0^\infty dJ e^{-\frac{J^2}{2}\left(\Delta\hat{q}_w + \frac{\beta\gamma}{\sigma_{\mathrm{E}}^2}\right) + \left(z\sqrt{\hat{q}_w} - \eta_{\mathrm{E}}\hat{M}\right)J}} \tag{43}$$

and analogously for the I weights. This expression holds for uncorrelated inputs and outputs and any fixed bias b, as well as for any correlated x and y with optimal bias b, where deviations from Eq (2) do not occur. In the $\beta \to \infty$, using the scaling relations Eqs (30)–(33), it can be easily shown that the mean-field weight probability density of the rescaled weights $\sqrt{N}w_i$ is a superposition of a δ function in zero and two truncated Gaussian densitites:

$$p(\sqrt{N}w) = p_0(B)\delta(w) + \sum_{c\in\{\mathrm{E},I\}} f_c G(\sqrt{N}w; M_c, \Sigma_c)\theta(s_c J) \tag{44}$$

where the mean and standard deviation of the Gaussians $G(\cdot;M, \Sigma)$ are:

$$M_c = -\frac{\eta_c B\sqrt{C}}{\sigma_c A + \frac{\gamma}{\sigma_c}} \tag{45}$$

$$\Sigma_c = \frac{\sqrt{C}}{\sigma_c A + \frac{\gamma}{\sigma_c}} \tag{46}$$

This weight density is valid for $\gamma > 0$ at any α and at critical capacity for $\gamma = 0$. The fraction of zero weights is given by:

$$p_0(B) = f_{\mathrm{E}}H(-\eta_{\mathrm{E}}B) + (1 - f_{\mathrm{E}})H(\eta_I B)$$

Spectrum of exponential and rbf covariance

For the exponential covariance $C_{\mu\nu} = e^{-\frac{|\mu-\nu|}{\tau}}$, one has [33]:

$$\lambda(\phi) = \frac{1-x^2}{1-2x\cos\phi + x^2}$$

with $x = e^{-\frac{1}{\tau}}$. In the rbf case $C_{\mu\nu} = e^{-\frac{|\mu-\nu|^2}{2\tau^2}}$, the spectrum can be computed by Fourier series [39], yielding

$$\lambda(\phi) = \vartheta_3\left(\frac{\phi}{2}, e^{-\frac{1}{2\tau^2}}\right)$$

with $\vartheta_3(z,q) = 1 + 2\sum_{n=1}^{\infty} q^{n^2}\cos(2nz)$ the Jacobi theta function of 3rd type.

Replica formalism: Sample covariance matrix (SC)

Also in the case of a sample covariance matrix, we are interested in statistically structured inputs and output. An independent average across x and y would result in a simple dependence on the variance of y in the energetic part. To capture the geometric dependence between x and y, we thus extend the calculations in [50, 51] to the case where the linear expansion of y_μ on the right singular vectors $V_{\cdot\mu}$ is known, by taking $\delta y_\mu = \Sigma_\nu V_{\mu\nu} c_\nu$.

In order to compute the replicated cumulant generating function Eq (11), we again introduce overlap parameters q_w^{ab}, whose volume is given by the previously computed entropic part $\mathcal{G}_S$. The fact that the entropic part is unchanged in turn implies that the mean-field weight distribution takes the form of Eq (44), with the values of $\{A, B, C\}$ being determined by a new set of saddle-point equations.

SC, Energetic part. Using again the expressions $(X)_{i\mu} = \bar{x}_i + \sigma_i \xi_{i\mu}$ and $\xi = USV^T$, the replicated cumulant generating function for the joint (mean-removed) input and output is:

$$Z_{\xi,\delta y} \quad = \left\langle \exp\left(i\sum_a \tilde{\boldsymbol{J}}_a^T S \tilde{\boldsymbol{u}}_a - i\sigma_y c^T \sum_a \tilde{\boldsymbol{u}}_a \right) \right\rangle_{p(\tilde{\boldsymbol{J}}_a, \tilde{\boldsymbol{u}}_a)} \tag{47}$$

where we used the change of variables $\tilde{J}_{ia} = \sum_k U_{ki} J_{ka}$ and $\tilde{u}_{\mu a} = \sum_k V_{k\mu} u_{ka}$. The average in Eq (47) is taken over the joint distribution $p(\tilde{\boldsymbol{J}}_a, \tilde{\boldsymbol{u}}_a)$ resulting from averaging over the Haar measure on the orthogonal matrices U and V. For a single replica, $Z_{\xi,\delta y}$ will only depend on the squared norms $Q_w = \sum_i \frac{\tilde{J}_i^2}{N}$ and $Q_u = \sum_\mu \frac{\tilde{u}_\mu^2}{P}$ of the two vectors $\tilde{\boldsymbol{J}}$ and $\tilde{\boldsymbol{u}}$. We can therefore write the average in the following way:

$$\langle \exp(i\tilde{\boldsymbol{J}}^T S \tilde{\boldsymbol{u}} - i\sigma_y c^T \tilde{\boldsymbol{u}}) \rangle_{p(\tilde{\boldsymbol{J}}, \tilde{\boldsymbol{u}})} \propto \int \delta(|\tilde{\boldsymbol{J}}|^2 - NQ_w)\delta(|\tilde{\boldsymbol{u}}|^2 - PQ_u) e^{i\tilde{\boldsymbol{J}}^T S\tilde{\boldsymbol{u}} - i\sigma_y c^T \tilde{\boldsymbol{u}}} \tag{48}$$

Introducing Fourier representation for the δ functions, we are left with an expression involving an $N + P$ dimensional Gaussian integral:

$$\begin{aligned} &\int \frac{d\Lambda_w}{4\pi i}\frac{d\Lambda_u}{4\pi i} e^{\frac{N\Lambda_w Q_w}{2} + \frac{P\Lambda_u Q_u}{2}} \int d\tilde{\boldsymbol{J}}\, d\tilde{\boldsymbol{u}}\, e^{-\frac{\Lambda_w}{2}|\tilde{\boldsymbol{J}}|^2 - \frac{\Lambda_u}{2}|\tilde{\boldsymbol{u}}|^2 + i\tilde{\boldsymbol{J}}^T S \tilde{\boldsymbol{u}} - i\sigma_y c^T \tilde{\boldsymbol{u}}} \\ &= \frac{(2\pi)^{\frac{N+p}{2}}}{(4\pi i)^2} \int d\Lambda_w d\Lambda_u e^{\frac{N\Lambda_w Q_w}{2} + \frac{P\Lambda_u Q_u}{2}} \det \mathcal{M}^{-\frac{1}{2}} \exp\left(-\frac{\sigma_y^2}{2} (\,\boldsymbol{0} \quad c\,) \mathcal{M}^{-1} \begin{pmatrix} \boldsymbol{0} \\ c \end{pmatrix} \right) \end{aligned} \tag{49}$$

where

$$\mathcal{M} = \begin{pmatrix} \Lambda_w \mathbb{1}_N & -iS \\ -iS^T & \Lambda_u \mathbb{1}_P \end{pmatrix}$$

and $\mathbb{1}_K$ is the identity matrix of dimension K. Following [51], the determinant can be easily calculated:

$$\begin{aligned} \frac{1}{N}\log\det\mathcal{M} &= \frac{1}{N}\sum_{k=1}^{\min(N,P)}\log\left(\lambda_k^x + \Lambda_w\Lambda_u\right) + \frac{(N-\min(N,P))}{N}\log\Lambda_u \rightarrow \\ &\rightarrow \langle\log(\lambda^x + \Lambda_w\Lambda_u)\rangle_{\lambda^x} + (\alpha-1)\log\Lambda_u \end{aligned} \tag{50}$$

where the limit is taken for $N \to \infty$ and the average is with respect to the eigenvalue distribution $\rho(\lambda^x)$. As for the quadratic portion of the Gaussian integral, calling $\lambda_k^y = c_k^2$, we will use the shorthand

$$\left\langle \frac{\lambda^y}{\lambda^x + \Lambda_w\Lambda_u} \right\rangle_{\lambda^x,\lambda^y} \equiv \frac{1}{P}\sum_{k=1}^{\min(N,P)}\frac{\Lambda_w\lambda_k^y}{\Lambda_w\Lambda_u + \lambda_k} + \frac{1}{P}\sum_{k=\min(N,P)+1}^{P}\frac{\lambda_k^y}{\Lambda_u} \tag{51}$$

Considering now the replicated generating function, all the $n(n + 1)$ cross-product $\boldsymbol{J}_a \cdot \boldsymbol{J}_b = \tilde{\boldsymbol{J}}_a \cdot \tilde{\boldsymbol{J}}_b$ and $\boldsymbol{u}_a \cdot \boldsymbol{u}_b = \tilde{\boldsymbol{u}}_a \cdot \tilde{\boldsymbol{u}}_b$ must be conserved via the multiplication of U and V. Together with the overlap parameters $Nq_w^{ab} = \sum_i J_{ia}J_{ib}$, we additionally introduce the quantities $Pq_u^{ab} = \sum_\mu u_{\mu a}u_{\mu b}$, thus obtaining:

$$e^{\alpha N\mathcal{G}_E} = \int\prod_{\mu a}\frac{d\Delta_{\mu a}du_{\mu a}}{2\pi}Z_{\xi,\delta y}e^{\sum_{a\le b}\hat{q}_u^{ab}\left(Pq_u^{ab}-\sum_\mu u_{\mu a}u_{\mu b}\right)-\frac{\beta}{2}\sum_{\mu a}\Delta_{\mu a}^2-i\sum_{\mu a}u_{\mu a}\left(\Delta_{\mu a}-M^a+\bar{y}\right)} \tag{52}$$

In the RS case, we again take $q_w^{ab} = q_w + \delta_{ab}\Delta q_w$ and, similarly for the u's, $q_u^{ab} = -q_u + \delta_{ab}\Delta q_u$. In the basis where both q_w^{ab} and q_u^{ab} are diagonal, the expression for $Z_{\xi,\delta y}$ becomes

$$Z_{\xi,\delta y} = \left\langle e^{i\tilde{\boldsymbol{J}}_1^T S\tilde{\boldsymbol{u}}_1 - i\sigma_y c^T\sqrt{n}\tilde{\boldsymbol{u}}_1}\prod_{b=2}^{n}e^{i\tilde{\boldsymbol{J}}_b^T S\tilde{\boldsymbol{u}}_b}\right\rangle \tag{53}$$

so, calling $\mathcal{G}_{\xi,\delta y} = \frac{1}{N}\lim_{n\to 0}\log Z_{\xi,\delta y}$, we have:

$$2\mathcal{G}_{\xi,\delta y} = F(\Delta q_w, \Delta q_u) + q_w\frac{\partial F(\Delta q_w,\Delta q_u)}{\partial\Delta q_w} - q_u\frac{\partial F(\Delta q_w,\Delta q_u)}{\partial\Delta q_u} - \alpha\sigma_y^2 K(\Lambda_w,\Lambda_u) \tag{54}$$

with the function F given by:

$$\begin{aligned} F(x,y) &= \mathrm{Extr}_{\Lambda_w,\Lambda_u}\{-\langle\log(\lambda^x + \Lambda_w\Lambda_u)\rangle_{\lambda^x} - (\alpha-1)\log\Lambda_u + \Lambda_w x + \alpha\Lambda_u y\} \\ &-\log x - \alpha\log y - (1+\alpha) \end{aligned} \tag{55}$$

and $K(\Lambda_w,\Lambda_u) = \Lambda_w\langle\frac{\lambda^y}{\lambda^x+\Lambda_w\Lambda_u}\rangle_{\lambda^x,\lambda^y}$. In Eq (54), it is intended that Λ_w and Λ_w are implied by the Legendre Transform conditions:

$$\Delta q_w = \Lambda_u\left\langle\frac{1}{\lambda^x + \Lambda_w\Lambda_u}\right\rangle_{\lambda^x} \tag{56}$$

$$\alpha\Delta q_u = \frac{\alpha-1}{\Lambda_u} + \Lambda_w\left\langle\frac{1}{\lambda^x + \Lambda_w\Lambda_u}\right\rangle_{\lambda^x} \tag{57}$$

The remaining terms in the energetic part $\mathcal{G}_E$ involve the q_u^{ab} overlaps and their conjugated parameters $\hat{q}_u^{ab}$. Introducing the RS ansatz $\hat{q}_u^{ab} = \hat{q}_u + \delta_{ab}\frac{\Delta\hat{q}_u - \hat{q}_u}{2}$, the calculation follows along the same lines of the section *SC, Energetic part*. We get:

$$2\mathcal{G}_E = \frac{2\mathcal{G}_{\xi,\delta y}}{\alpha} + \Delta\hat{q}_u(\Delta q_u - q_u) + \hat{q}_u\Delta q_u - \log(1 + \beta\Delta\hat{q}_u) - \beta\frac{\hat{q}_u + (M - \bar{y})^2}{1 + \beta\Delta\hat{q}_u} \tag{58}$$

Eliminating M, $\hat{q}_u$ and $\Delta\hat{q}_u$ at the saddle-point in Eq (58), $\mathcal{G}_E$ reduces to:

$$\mathcal{G}_E = \frac{\mathcal{G}_{\xi,\delta y}}{\alpha} + \frac{q_u - \Delta q_u}{2\beta} - \frac{q_u}{2\Delta q_u} + \frac{1}{2}\log\Delta q_u \tag{59}$$

SC, Saddle-point equations. The final expression for the free energy density

$$-\beta f = -\hat{M}I + \frac{\Delta\hat{q}_w}{2}(\Delta q_w + q_w) - \frac{\hat{q}_w\Delta q_w}{2} + \mathcal{G}_S + \alpha\mathcal{G}_E \tag{60}$$

implies the following saddle-point equations:

$$\Delta\hat{q}_w + \frac{\partial F}{\partial\Delta q_w} = 0 \tag{61}$$

$$\frac{\alpha}{\Delta q_u} - \frac{\alpha}{\beta} + \frac{\partial F}{\partial\Delta q_u} = 0 \tag{62}$$

$$\hat{q}_w = q_w\frac{\partial^2 F}{\partial\Delta q_w^2} - q_u\frac{\partial^2 F}{\partial\Delta q_w\Delta q_u} - \alpha\sigma_y^2\frac{\partial K}{\partial\Delta q_w} \tag{63}$$

$$\alpha\frac{q_u}{\Delta q_u^2} = q_u\frac{\partial^2 F}{\partial\Delta q_u^2} - q_w\frac{\partial^2 F}{\partial\Delta q_w\Delta q_u} + \alpha\sigma_y^2\frac{\partial K}{\partial\Delta q_u} \tag{64}$$

in addition to the entropic saddle-point Eqs (25)–(27), which are unchanged. The saddle-point values of the conjugate Legendre variables Λ_w, Λ_u greatly simplify the expression for the first and second derivatives of F. Indeed, from Eqs (61) and (62) one has:

$$\Lambda_w = \frac{1}{\Delta q_w} - \Delta\hat{q}_w \tag{65}$$

$$\Lambda_u = \beta^{-1} \tag{66}$$

or, setting $\Lambda_w = \beta\tilde{\Lambda}_w$:

$$\tilde{\Lambda}_w = \frac{1}{\Delta\tilde{q}_w} - A \tag{67}$$

In particular, Eq (56) shows that $\Delta\tilde{q}_w$ is expressed by a Stieltjes transform of $\rho(\lambda^x)$ and the first term in Eq (55) is its Shannon transform. In the limit $\beta \to \infty$, using the following additional scaling relations for the u overlaps:

$$q_u = \beta^2\tilde{q}_u \tag{68}$$

$$\Delta q_u = \beta\Delta\tilde{q}_u \tag{69}$$

we get the expression for the energy density:

$$\epsilon = \frac{\alpha}{2}\sigma_y^2 \tilde{\Lambda}_w \left\langle \frac{\lambda^y}{\lambda^x + \tilde{\Lambda}_w} \right\rangle_{\lambda^x,\lambda^y}$$

SC, i.i.d. and unconstrained cases. Either setting $K = 0$ or $\lambda^y = 0$ reverts back to the i.i.d. output case. In the special case of i.i.d. inputs, the eigenvalue distribution is Marchenko-Pastur

$$\rho(\lambda) = \frac{\sqrt{(\lambda - \lambda_-)(\lambda_+ - \lambda)}}{2\pi\lambda} \tag{70}$$

with $\lambda_{+/-} = (1 \pm \sqrt{\alpha})^2$, from which $F(\Delta q_w, \Delta q_u) = -\frac{\alpha}{2}\Delta q_w \Delta q_u$. The saddle-point equations are essentially the same as the ones in the section *EC, Saddle-point equations* with $C^x_{\mu\nu} = C^y_{\mu\nu} = \delta_{\mu\nu}$.

Let us also note that, in the simple unconstrained case, taking for simplicity $\bar{x}_i = 0$ and $b = 0$, the entropic part can be worked out to be, up to constant terms:

$$2\mathcal{G}_S = \log \Delta q_w + \frac{q_w}{\Delta q_w} - \beta\gamma(\Delta q_w + q_w) \tag{71}$$

which, at the saddle-point, implies $\tilde{\Lambda}_w = \gamma$. The mean-field distribution $p(\sqrt{N}w)$ is a zero-mean Gaussian with variance $v = q_w$. Using the properties of the Hessian of the Legendre Transform, it is easy to show that:

$$q_{w,\text{unc}} = \alpha\frac{\partial K}{\partial \Lambda_w} = \alpha\left\langle \frac{\lambda^x \lambda^y}{(\lambda^x + \gamma)^2} \right\rangle_{\lambda^x,\lambda^y} \tag{72}$$

$$\epsilon_{\text{unc}} = \frac{\alpha}{2}\sigma_y^2 \gamma \left\langle \frac{\lambda^y}{\lambda^x + \gamma} \right\rangle_{\lambda^x,\lambda^y} \tag{73}$$

These expressions can also be derived from the pseudo-inverse solution (we take $\bar{y} = 0$ for simplicity) $w^* = (\xi\xi^T + \gamma)^{-1}\xi y$, by taking an average across ξ and y in the two expressions:

$$v = \langle w^{*T} w^* \rangle = \text{Tr}(\xi y y^T \xi^T (\xi\xi^T + \gamma)^{-2}) \tag{74}$$

$$\langle E \rangle = \frac{1^T}{2}\langle y^T y \rangle - \frac{1}{2}\text{Tr}\big(\xi y y^T \xi^T (\xi\xi^T + \gamma)^{-1}\big) \tag{75}$$

The i.i.d. output case also follows by performing independent averages over y and ξ.

Acknowledgments

The author would like to thank L.F. Abbott and Francesco Fumarola for constructive criticism of an earlier version of the manuscript.

Author Contributions

Conceptualization: Alessandro Ingrosso.

Formal analysis: Alessandro Ingrosso.

Investigation: Alessandro Ingrosso.

Methodology: Alessandro Ingrosso.

Software: Alessandro Ingrosso.

Validation: Alessandro Ingrosso.

Visualization: Alessandro Ingrosso.

Writing – original draft: Alessandro Ingrosso.

Writing – review & editing: Alessandro Ingrosso.

References

1. Song HF, Yang GR, Wang XJ. Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework. PLOS Computational Biology. 2016; 12(2):1–30. https://doi.org/10.1371/journal.pcbi.1004792 PMID: 26928718
2. Nicola W, Clopath C. Supervised learning in spiking neural networks with FORCE training. Nature Communications. 2017; 8(1):2208. https://doi.org/10.1038/s41467-017-01827-3 PMID: 29263361
3. Ingrosso A, Abbott LF. Training dynamically balanced excitatory-inhibitory networks. PLOS ONE. 2019; 14(8):1–18. https://doi.org/10.1371/journal.pone.0220547 PMID: 31393909
4. Kim CM, Chow CC. Learning recurrent dynamics in spiking networks. eLife. 2018; 7:e37124. https://doi.org/10.7554/eLife.37124 PMID: 30234488
5. Brendel W, Bourdoukan R, Vertechi P, Machens CK, Denève S. Learning to represent signals spike by spike. PLOS Computational Biology. 2020; 16(3):1–23. https://doi.org/10.1371/journal.pcbi.1007692 PMID: 32176682
6. Brunel N, Hakim V, Isope P, Nadal JP, Barbour B. Optimal Information Storage and the Distribution of Synaptic Weights: Perceptron versus Purkinje Cell. Neuron. 2004; 43(5):745–757. https://doi.org/10.1016/S0896-6273(04)00528-8 PMID: 15339654
7. Barbour B, Brunel N, Hakim V, Nadal JP. What can we learn from synaptic weight distributions? Trends in Neurosciences. 2007; 30(12):622–629. https://doi.org/10.1016/j.tins.2007.09.005 PMID: 17983670
8. Brunel N. Is cortical connectivity optimized for storing information? Nature Neuroscience. 2016; 19 (5):749–755. https://doi.org/10.1038/nn.4286 PMID: 27065365
9. Gardner E. The space of interactions in neural network models. Journal of Physics A: Mathematical and General. 1988; 21(1):257–270.
10. Clopath C, Nadal JP, Brunel N. Storage of correlated patterns in standard and bistable Purkinje cell models. PLoS computational biology. 2012; 8(4):e1002448–e1002448. https://doi.org/10.1371/journal.pcbi.1002448 PMID: 22570592
11. Chapeton J, Fares T, LaSota D, Stepanyants A. Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons. Proceedings of the National Academy of Sciences. 2012; 109(51): E3614–E3622. https://doi.org/10.1073/pnas.1211467109 PMID: 23213221
12. Zhang D, Zhang C, Stepanyants A. Robust Associative Learning Is Sufficient to Explain the Structural and Dynamical Properties of Local Cortical Circuits. Journal of Neuroscience. 2019; 39(35):6888–6904. https://doi.org/10.1523/JNEUROSCI.3218-18.2019 PMID: 31270161
13. Rubin R, Abbott LF, Sompolinsky H. Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. Proceedings of the National Academy of Sciences. 2017; 114(44): E9366–E9375. https://doi.org/10.1073/pnas.1705841114 PMID: 29042519
14. Seung HS, Sompolinsky H, Tishby N. Statistical mechanics of learning from examples. Phys Rev A. 1992; 45:6056–6091. https://doi.org/10.1103/PhysRevA.45.6056 PMID: 9907706
15. Clopath C, Brunel N. Optimal Properties of Analog Perceptrons with Excitatory Weights. PLOS Computational Biology. 2013; 9(2):1–6. https://doi.org/10.1371/journal.pcbi.1002919 PMID: 23436991
16. Gutfreund H, Stein Y. Capacity of neural networks with discrete synaptic couplings. Journal of Physics A: Mathematical and General. 1990; 23(12):2613–2630. https://doi.org/10.1088/0305-4470/23/12/036
17. Isaacson JS, Scanziani M. How Inhibition Shapes Cortical Activity. Neuron. 2011; 72(2):231–243. https://doi.org/10.1016/j.neuron.2011.09.027 PMID: 22017986
18. Field RE, D'amour JA, Tremblay R, Miehl C, Rudy B, Gjorgjieva J, et al. Heterosynaptic Plasticity Determines the Set Point for Cortical Excitatory-Inhibitory Balance. Neuron. 2020; https://doi.org/10.1016/j.neuron.2020.03.002. PMID: 32213321

19. Hennequin G, Agnes EJ, Vogels TP. Inhibitory Plasticity: Balance, Control, and Codependence. Annual Review of Neuroscience. 2017; 40(1):557–579. https://doi.org/10.1146/annurev-neuro-072116-031005 PMID: 28598717

20. Ahmadian Y, Miller KD. What is the dynamical regime of cerebral cortex? arXiv:190810101. 2019.

21. van Vreeswijk C, Sompolinsky H. Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity. Science. 1996; 274(5293):1724–1726. https://doi.org/10.1126/science.274.5293.1724 PMID: 8939866

22. van Vreeswijk C, Sompolinsky H. Chaotic Balanced State in a Model of Cortical Circuits. Neural Comput. 1998; 10(6):1321–1371. https://doi.org/10.1162/089976698300017214 PMID: 9698348

23. Renart A, de la Rocha J, Bartho P, Hollender L, Parga N, Reyes A, et al. The Asynchronous State in Cortical Circuits. Science. 2010; 327(5965):587–590. https://doi.org/10.1126/science.1179850 PMID: 20110507

24. Kadmon J, Sompolinsky H. Transition to Chaos in Random Neuronal Networks. Phys Rev X. 2015; 5:041030.

25. Harish O, Hansel D. Asynchronous Rate Chaos in Spiking Neuronal Circuits. PLOS Computational Biology. 2015; 11(7):1–38. https://doi.org/10.1371/journal.pcbi.1004266 PMID: 26230679

26. Brunel N. Dynamics of Sparsely Connected Networks of Excitatory and Inhibitory Spiking Neurons. Journal of Computational Neuroscience. 2000; 8(3):183–208. https://doi.org/10.1023/A:1008925309027 PMID: 10809012

27. Tsodyks MV, Sejnowski T. Rapid state switching in balanced cortical network models. Network: Computation in Neural Systems. 1995; 6(2):111–124. https://doi.org/10.1088/0954-898X_6_2_001

28. Goldt S, Mézard M, Krzakala F, Zdeborová L. Modelling the influence of data structure on learning in neural networks: the hidden manifold model. arXiv:190911500. 2019.

29. Chung S, Lee DD, Sompolinsky H. Classification and Geometry of General Perceptual Manifolds. Phys Rev X. 2018; 8:031003.

30. Cohen U, Chung S, Lee DD, Sompolinsky H. Separability and geometry of object manifolds in deep neural networks. Nature Communications. 2020; 11(1):746. https://doi.org/10.1038/s41467-020-14578-5 PMID: 32029727

31. Rotondo P, Lagomarsino MC, Gherardi M. Counting the learnable functions of geometrically structured data. Phys Rev Research. 2020; 2:023169. https://doi.org/10.1103/PhysRevResearch.2.023169

32. Pastore M, Rotondo P, Erba V, Gherardi M. Statistical learning theory of structured data. arXiv:200510002. 2020.

33. Monasson R. Properties of neural networks storing spatially correlated patterns. Journal of Physics A: Mathematical and General. 1992; 25(13):3701–3720. https://doi.org/10.1088/0305-4470/25/13/019

34. Tarkowski W, Lewenstein M. Learning from correlated examples in a perceptron. Journal of Physics A: Mathematical and General. 1993; 26(15):3669–3679. https://doi.org/10.1088/0305-4470/26/15/017

35. Monasson R. Storage of spatially correlated patterns in autoassociative memories. Journal de Physique I. 1993; 3(5):1141–1152. https://doi.org/10.1051/jp1:1993107

36. Battista A, Monasson R. Capacity-Resolution Trade-Off in the Optimal Learning of Multiple Low-Dimensional Manifolds by Attractor Neural Networks. Phys Rev Lett. 2020; 124:048302. https://doi.org/10.1103/PhysRevLett.124.048302 PMID: 32058781

37. Amit DJ, Wong KYM, Campbell C. Perceptron learning with sign-constrained weights. Journal of Physics A: Mathematical and General. 1989; 22(12):2039–2045. https://doi.org/10.1088/0305-4470/22/12/009

38. Amit DJ, Campbell C, Wong KYM. The interaction space of neural networks with sign-constrained synapses. Journal of Physics A: Mathematical and General. 1989; 22(21):4687–4693. https://doi.org/10.1088/0305-4470/22/21/030

39. Gray RM. Toeplitz and Circulant Matrices: A Review. Foundations and Trends in Communications and Information Theory. 2006; 2(3):155–239. https://doi.org/10.1561/0100000006

40. Abbott LF, Rajan K, Sompolinsky H. Interactions between Intrinsic and Stimulus-Evoked Activity in Recurrent Neural Networks. arXiv:09123832. 2009.

41. Litwin-Kumar A, Harris KD, Axel R, Sompolinsky H, Abbott LF. Optimal Degrees of Synaptic Connectivity. Neuron. 2017; 93(5):1153–1164.e7. https://doi.org/10.1016/j.neuron.2017.01.030 PMID: 28215558

42. Marinari E, Parisi G, Ritort F. Replica field theory for deterministic models. II. A non-random spin glass with glassy behaviour. Journal of Physics A: Mathematical and General. 1994; 27(23):7647–7668. https://doi.org/10.1088/0305-4470/27/23/011

43. Parisi G, Potters M. Mean-field equations for spin models with orthogonal interaction matrices. Journal of Physics A: Mathematical and General. 1995; 28(18):5267–5285. https://doi.org/10.1088/0305-4470/28/18/016

44. Cherrier R, Dean DS, Lefèvre A. Role of the interaction matrix in mean-field spin glass models. Phys Rev E. 2003; 67:046112. https://doi.org/10.1103/PhysRevE.67.046112 PMID: 12786441

45. Opper M, Winther O. Tractable Approximations for Probabilistic Models: The Adaptive Thouless-Anderson-Palmer Mean Field Approach. Phys Rev Lett. 2001; 86:3695–3699. https://doi.org/10.1103/PhysRevLett.86.3695 PMID: 11329302

46. Opper M, Winther O. Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling. Phys Rev E. 2001; 64:056131. https://doi.org/10.1103/PhysRevE.64.056131 PMID: 11736038

47. Opper M, Winther O. Expectation Consistent Approximate Inference. Journal of Machine Learning Research. 2005; 6:2177–2204.

48. Takeda K, Uda S, Kabashima Y. Analysis of CDMA systems that are characterized by eigenvalue spectrum. Europhysics Letters (EPL). 2006; 76(6):1193–1199. https://doi.org/10.1209/epl/i2006-10380-5

49. Kabashima Y. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. Journal of Physics: Conference Series. 2008; 95:012001.

50. Shinzato T, Kabashima Y. Learning from correlated patterns by simple perceptrons. Journal of Physics A: Mathematical and Theoretical. 2008; 42(1):015005. https://doi.org/10.1088/1751-8113/42/1/015005

51. Shinzato T, Kabashima Y. Perceptron capacity revisited: classification ability for correlated patterns. Journal of Physics A: Mathematical and Theoretical. 2008; 41(32):324013. https://doi.org/10.1088/1751-8113/41/32/324013

52. Tulino AM, Verdú S. Random Matrix Theory and Wireless Communications. Foundations and Trends in Communications and Information Theory. 2004; 1(1):1–182. https://doi.org/10.1561/0100000001

53. Tao T. Topics in Random Matrix Theory. Graduate studies in mathematics. American Mathematical Soc.;. Available from: https://books.google.com/books?id=Hjq_JHLNPT0C.

54. Ganguli S, Sompolinsky H. Statistical Mechanics of Compressed Sensing. Phys Rev Lett. 2010; 104:188701. https://doi.org/10.1103/PhysRevLett.104.188701 PMID: 20482215

55. Marr D. A theory of cerebellar cortex. The Journal of physiology. 1969; 202(2):437–470. https://doi.org/10.1113/jphysiol.1969.sp008820 PMID: 5784296

56. Wolpert DM, Miall RC, Kawato M. Internal models in the cerebellum. Trends in Cognitive Sciences. 1998; 2(9):338–347. https://doi.org/10.1016/S1364-6613(98)01221-2 PMID: 21227230

57. Herzfeld DJ, Kojima Y, Soetedjo R, Shadmehr R. Encoding of error and learning to correct that error by the Purkinje cells of the cerebellum. Nature Neuroscience. 2018; 21(5):736–743. https://doi.org/10.1038/s41593-018-0136-y PMID: 29662213

58. Mastrogiuseppe F, Ostojic S. Intrinsically-generated fluctuating activity in excitatory-inhibitory networks. PLOS Computational Biology. 2017; 13(4):1–40. https://doi.org/10.1371/journal.pcbi.1005498 PMID: 28437436

59. Chen J, Richard C, Bermudez JM, Honeine P. Variants of Non-Negative Least-Mean-Square Algorithm and Convergence Analysis. IEEE Transactions on Signal Processing. 2014; 62(15):3990–4005. https://doi.org/10.1109/TSP.2014.2332440

60. Nascimento VH, Zakharov YV. RLS Adaptive Filter With Inequality Constraints. IEEE Signal Processing Letters. 2016; 23(5):752–756. https://doi.org/10.1109/LSP.2016.2551468

61. Engel A, Van den Broeck C. Statistical mechanics of learning. Cambridge University Press; 2001.

62. Mei S, Montanari A. The generalization error of random features regression: Precise asymptotics and double descent curve. arXiv:190805355. 2019.

63. Gerace F, Loureiro B, Krzakala F, Mézard M, Zdeborová L. Generalisation error in learning with random features and the hidden manifold model. arXiv:200209339. 2020.

64. Babadi B, Sompolinsky H. Sparseness and Expansion in Sensory Representations. Neuron. 2014; 83(5):1213–1226. https://doi.org/10.1016/j.neuron.2014.07.035 PMID: 25155954

65. Cayco-Gajic NA, Silver RA. Re-evaluating Circuit Mechanisms Underlying Pattern Separation. Neuron. 2019; 101(4):584–602. https://doi.org/10.1016/j.neuron.2019.01.044 PMID: 30790539

66. Ocker GK, Litwin-Kumar A, Doiron B. Self-Organization of Microcircuits in Networks of Spiking Neurons with Plastic Synapses. PLOS Computational Biology. 2015; 11(8):1–40. https://doi.org/10.1371/journal.pcbi.1004458 PMID: 26291697

67. Mézard M, Parisi G, Virasoro M. Spin Glass Theory and Beyond. World Scientific Lecture Notes in Physics; 1987.