Fundamental bounds on learning performance in neural circuits

Dhruva V. Raman, Timothy O'Leary

August 2, 2018

Abstract

How does the size of a neural circuit influence its learning performance? Intuitively, we expect the learning capacity of a neural circuit to grow with the number of neurons and synapses. Larger brains tend to be found in species with higher cognitive function and learning ability. Similarly, adding connections and units to artificial neural networks can allow them to solve more complex tasks. However, we show that in a biologically relevant setting where synapses introduce an unavoidable amount of noise, there is an optimal size of network for a given task. Beneath this optimal size, our analysis shows how adding apparently redundant neurons and connections can make tasks more learnable. Therefore large neural circuits can either devote connectivity to generating complex behaviors, or exploit this connectivity to achieve faster and more precise learning of simpler behaviors. Above the optimal network size, the addition of neurons and synaptic connections starts to impede learning performance. This suggests that overall brain size may be constrained by the need to learn efficiently with unreliable synapses, and may explain why some neurological learning deficits are associated with hyperconnectivity. Our analysis is independent of specific learning rules and uncovers fundamental relationships between learning rate, task performance, network size and intrinsic noise in neural circuits.

Introduction

In the brain, computations are distributed across circuits that can include many millions of neurons and synaptic connections. Maintaining a large nervous system is expensive energetically and reproductively [1, 2, 3], suggesting that the cost of additional neurons is balanced by an increased capacity to learn and process information.

Empirically, a 'bigger is better' hypothesis is supported by the correlation of brain size with higher cognitive function and learning capacity across animal species [4, 5, 6]. Within and across species, positive correlations exist between the volume of a specific brain region and the behavioural importance of tasks related to the brain region in question [7, 8, 9]. These observations make sense from a theoretical perspective because larger artificial neural networks can be trained to solve more challenging computational tasks than smaller networks [10, 11, 12, 13, 14, 15].

However, biologically it is not clear that there is always a computational advantage to having more neurons and synapses engaged in learning and solving a task. During learning, larger networks face the problem of tuning a greater number of synapses using limited and potentially corrupted information on task performance [16, 17]. Moreover, no biological component is perfect, so unavoidable noise arising from the molecular machinery in individual synapses might sum unfavourably as the size of a network grows. Intriguingly, a number of well-studied neurodevelopmental disorders exhibit cortical hyperconnectivity at the same time as learning deficits [18, 19, 20, 21]. It is therefore a fundamental question whether learning capacity can grow indefinitely with the number of neurons and synapses in a neural circuit, or whether there is some law of diminishing returns that eventually leads to a decrease in performance beyond a certain network size.

We address this question with a general mathematical analysis of learning performance in neural circuits that is independent of specific learning rules and circuit architectures. Given some fixed task, we show analytically that expected learning rate and steady-state performance is higher in larger networks than smaller networks when there is no intrinsic noise in the connections. The analysis reveals how connections can be added to a network in a principled way to reduce the difficulty of learning a task. This gain in overall learning performance is accompanied by slower per-synapse rate of change, predicting that synaptic turnover rates should vary across brain areas according to the number of connections involved in a task, and the typical task complexity.

If each synaptic connection is intrinsically noisy we show that there is an optimal network size for a given task. Above the optimal network size, adding neurons and connections degrades learning and steady-state performance. This reveals an important disparity between synapses in artificial neural networks, which are not subject to unavoidable intrinsic noise, and those in biology, which are necessarily subject to fluctuations at the molecular level [22, 23, 24, 25].

For networks that are beneath the optimal size, it turns out to be advantageous to add apparently redundant neurons and connections. We show how additional synaptic pathways reduce the impact of imperfections in learning rules and uncertainty in the task error. This provides a potential theoretical explanation for recent, counterintuitive experimental observations in mammalian cortex [26, 27], which show that neurons frequently make multiple, redundant synaptic connections to the same postsynaptic cell. A non-obvious consequence of this result is that the size of a neural circuit can either reflect the complexity of a fixed task, or instead deliver greater learning performance on simpler, arbitrary tasks.

Results

Modelling the effect of network size on learning

Our goal is to analyse how network size affects learning and task performance in a general setting depicted in Figure 1, which is independent of specific tasks, network architectures and learning rules. We assume that there is some error signal that is fed back to the network via a learning rule that adjusts synaptic weights. We also assume that the error signal is limited both by noise and by a finite sampling rate quantified by some time interval T (Figure 1A). In addition to the noise in the learning rule, we also consider noise that is independently distributed across synapses ('intrinsic synaptic noise'). This models molecular noise in signalling and structural apparatus in a biological synapse that is uncorrelated with learning processes, and with changes in other synapses. Network size is adjusted by adding synapses and neurons (Figure 1B).

Before dealing with the general case, we motivate the analysis with simulations of standard nonlinear feedforward neural network models that we trained to compute arbitrary input-output mappings, as shown in Figure 2A. We simulate learning in networks using the so-called student-teacher framework, where we generate a fixed input-output mapping using a nonlinear feedforward 'teacher' network with randomly chosen fixed synaptic weights. Importantly, teacher networks can generate arbitrarily complex input-output mappings with the convenient property of being learnable by a 'student' network with the same underlying connection topology. We added neurons and connections to internal layers to generate student networks of increasing size (Figure 2B). This scales network size while ensuring that all networks can solve the same task.

Learning was simulated by modifying synapses with noise-corrupted gradient descent to mimic an imperfect biological learning rule. We emphasize that do not assume that learning in a biological network occurs by explicit gradient descent. However any error-based learning rule must induce synaptic changes that approximate gradient descent, as we show below (Equation 1).

The phenomenon we wish to understand in detail is shown in Figures 2C and D. We trained networks of varying sizes on the same task, with the same amount of learning rule noise. Larger networks learn more quickly and to a higher steady state performance than smaller networks when there is no intrinsic noise in the synapses Figure 2C. This is surprising because the only difference between the smallest network and larger networks is the addition of redundant synapses and neurons, and the task is designed so that all networks can learn it perfectly in principle. Moreover, as shown in Figure 2D, a small amount of internal noise in each of the connections of the student networks results in a non-monotonic relationship between performance and network size. Beyond a certain size, both learning and asymptotic performance start to worsen.

The simulations in Figure 2 provide evidence of an underlying relationship between learning rate, task performance, network size and intrinsic noise. To understand these observations in a rigorous and general way, we mathematically analysed how network size and imperfections in feedback learning rules impact learning in a general case.

Learning rate and task difficulty

We first establish a general relationship between learning rate and key properties of error-based learning rules. We define task error as a smooth function $F[\mathbf{w}]$ which decreases with increasing levels of performance, and which depends on the vector $\mathbf{w}$ of N synaptic weights in a network. We assume that learning rules use some (potentially imperfect) estimate of this error to adjust synaptic weights.

Biologically, it is reasonable to assume that learning-related synaptic changes occur due to old information. For example, a task-related reward or punishment may be supplied only at the end of a task, which itself takes time to execute. Similarly, even if error feedback is ongoing in time, there will always be some biochemically induced delay between acquisition of this error signal and its integration into plastic changes at each synapse.

Thus, there will be a maximum rate at which task error information can be transmitted to synapses during learning, which for mathematical convenience can be lumped into discrete timepoints. Suppose feedback on task error occurs at timepoints 0 and T, but not in between, for some $T > 0$ (Figure 3A). If the network learned over the interval $[0, T]$ then $F[\mathbf{w}(T)] - F[\mathbf{w}(0)] < 0$ by definition. We quantify learning rate during this interval as the value of k such that

$$F[\mathbf{w}(T)] = (1 - kT)F[\mathbf{w}(0)],$$

with $k < \frac{1}{T}$. A larger positive value of k implies a faster rate of learning. We can write the total change in error over the interval T (see Figure 3A) as:

$$\begin{aligned} F[\mathbf{w}(T)] - F[\mathbf{w}(0)] &= \int_0^T \langle \nabla F[\mathbf{w}(t)], \dot{\mathbf{w}}(t) \rangle \, dt \\ &= T \, \mathbb{E}_t[\langle \nabla F[\mathbf{w}(t)], \dot{\mathbf{w}}(t) \rangle], \end{aligned} \quad (1)$$

where expectation is taken across a uniform distribution of timepoints in $[0, T]$, dots denote time derivatives and angle brackets denote the (standard) inner product. Equation (1) shows that synaptic changes, on average, must anticorrelate with the gradient for learning to occur. We can thus decompose net learning rate during the interval T into contributions as follows (further details in *SI section 'Learning rate and local task difficulty'*):

$$k = \underbrace{\frac{-\|\nabla F[\mathbf{w}(0)]\|_2}{F[\mathbf{w}(0)]}}_{\substack{\text{gradient}\\ \text{strength}}} \left[\overbrace{\langle \dot{\boldsymbol{\omega}}_T, \nabla \hat{F}[\mathbf{w}(0)] \rangle}^{\substack{\text{contribution}\\ \text{from gradient}}} + \underbrace{\mathbf{G}_F[\hat{\dot{\boldsymbol{\omega}}}_T] \|\dot{\boldsymbol{\omega}}_T\|_2^2 T}_{\substack{\text{contribution}\\ \text{from curvature}}} \right] + \mathcal{O}(T^2) \quad (2)$$

where

$$\mathbf{G}_F[\hat{\dot{\boldsymbol{\omega}}}_T] := \frac{1}{2\|\nabla F[\mathbf{w}(0)]\|_2} \left\langle \hat{\dot{\boldsymbol{\omega}}}_T, \nabla^2 F[\mathbf{w}(0)] \hat{\dot{\boldsymbol{\omega}}}_T \right\rangle .$$

Hats indicate unit length normalized vectors (i.e. $\hat{x} = \frac{x}{\|x\|_2}$) and $\dot{\boldsymbol{\omega}}_T$ denotes the average synaptic change over the time interval $[0, T]$, normalised by T:

$$\dot{\boldsymbol{\omega}}_T = \frac{\mathbf{w}(T) - \mathbf{w}(0)}{T}. \quad (3)$$

Note that we have made no assumptions on the size of T, so the $\mathcal{O}(T^2)$ term in (2) is not necessarily small. Nonetheless, we can gain useful insight for how error surface geometry affects learning by examining the other terms on the right hand side of (2). The gradient strength scales the overall learning rate. Inside the brackets, the curvature term (which can change sign and magnitude during learning) can compete with the gradient term to slow down (or reverse) learning.

Informally, the curvature term in (2) therefore controls the learning 'difficulty' at each stage of learning. As we will show, this term can be tuned by changing the number of neurons and synaptic connections in the network.

The learning rate, k, is likely to remain positive during learning if the gradient direction changes gradually as the error surface is traversed (i.e. the error surface is almost linear). In this case a high rate of plasticity - due to a high gain between feedback error and synaptic change - will result in a high learning rate. However, if the descent direction changes rapidly due to the curvature of the error surface (i.e. the surface is crinkled up), then correlation with $-\nabla F[\mathbf{w}(0)]$ becomes a weaker predictor of learning over the entire time interval T. Effective learning therefore involves balancing error surface curvature and per-synapse rate of plasticity. This is illustrated in Figure 3A, where the length of the leaps along the error surface indicate the rate of plasticity.

We next decompose the contributions to the overall synaptic change during a learning increment. First we assume that synapses are perfectly reliable, with no intrinsic noise fluctuations affecting their strengths. In this case, we can decompose $\dot{\boldsymbol{\omega}}_T$ into two components that are parallel and perpendicular to the gradient at time 0, when error information was supplied to the network (see Figure 3B):

$$\dot{\boldsymbol{\omega}}_T = -\gamma_1 \nabla \hat{F}[\mathbf{w}(0)] + \gamma_2 \hat{\mathbf{n}}_2.$$

where γ_1 is the component of synaptic change that projects onto the error gradient direction and γ_2 is the component perpendicular to the gradient direction, with $\hat{\mathbf{n}}_2$ denoting the unit vector in this direction. We call these two components, γ_1 and γ_2, the **task-relevant plasticity** and **task-irrelevant plasticity** respectively.

Note that a learning rule could theoretically induce task-relevant synaptic changes in a direction that is not parallel to the gradient, $\nabla F[\mathbf{w}(0)]$, if information on the Hessian $\nabla^2 F[\mathbf{w}(0)]$ were available. However, as mentioned previously we are assuming that such information is not available biologically and the best the network can do is follow the gradient. In fact, the results of the paper can be generalised so as to eliminate this assumption (see *SI section 'Task-relevant Plasticity'*) but this complicates the presentation without adding insight.

There are several sources of task-irrelevant plasticity. Firstly, there can be inherent imperfections in the learning rule: information on task error may be imperfectly acquired and transmitted through the nervous system. Secondly, as we have emphasized above, the process of integrating feedback error and converting it into synaptic changes takes time. Therefore any learning rule will be using outdated information on task performance, implying that the gradient information will have error in general, unless it is constant for a task. This is illustrated in Figure 3A, where we see that during learning, the local information used to modify synapses leads to a network overshooting local minima in the error surface. Thirdly, in a general biological setting, synapses will be involved in multiple task-irrelevant plasticity processes that contribute to γ_2 (Figure 1A). For instance, the learning of additional, unrelated tasks may induce concurrent synaptic changes; so too could other ongoing cellular processes such as homeostatic plasticity. The common feature of all these components of task-irrelevant plasticity is that they are correlated across the network, but uncorrelated with learning the task.

We now consider the impact of intrinsic noise in the synapses themselves. Synapses are subject to continuous fluctuations due to turnover of receptors and signalling components. Some component of this will be uncorrelated across synapses so we can model these sources of noise as additional, independent white-noise fluctuations at each synapse with some total (per-synapse) variance $\gamma_3 T$ over the interval $[0, T]$. Because this noise is independent across synapses the total variance in the network will scale with the number of synapses. This gives a new expression for synaptic weight change:

$$\begin{aligned}\mathbf{w}(T) - \mathbf{w}(0) &= -\gamma_1 T \nabla \hat{F}[\mathbf{w}(0)] + \gamma_2 T \hat{\mathbf{n}}_2 + \gamma_3 \sqrt{T}\sqrt{N}\hat{\mathbf{n}}_3 \\ &= T\left(-\gamma_1 \nabla \hat{F}[\mathbf{w}(0)] + \gamma_2 \hat{\mathbf{n}}_2 + \gamma_3 \sqrt{\frac{N}{T}}\hat{\mathbf{n}}_3.\right)\end{aligned} \tag{4}$$

Note that γ_3 describes the average degree of intrinsic noise per synapse, whereas γ_1 and γ_2 describe components of synaptic change over the entire network. This in turn gives the following expression for the average weight velocity over the learning interval:

$$\dot{\boldsymbol{\omega}}_T = -\gamma_1 \nabla \hat{F}[\mathbf{w}(0)] + \gamma_2 \hat{\mathbf{n}}_2 + \gamma_3 \sqrt{\frac{N}{T}}\hat{\mathbf{n}}_3. \tag{5}$$

If we assume that each component is independent (see *SI section 'Decomposition of local task difficulty'* for justification) we can also write down the magnitude of total synaptic rate of change in a convenient form:

$$\|\dot{\boldsymbol{\omega}}_T\|_2^2 = \gamma_1^2 + \gamma_2^2 + \gamma_3^2 \frac{N}{T}. \tag{6}$$

Expressions (5) and (6) allow us to rewrite (2):

$$\mathbb{E}(k) = \frac{-\|\nabla F[\mathbf{w}(0)]\|_2}{F[\mathbf{w}(0)]}\Big[-\gamma_1 + \mathbf{G}_F[\dot{\hat{\boldsymbol{\omega}}}_T]\|\dot{\boldsymbol{\omega}}_T\|_2^2 T\Big] + \mathcal{O}(T^2). \tag{7}$$

If we fix values of the γ_i and T, we see from (7) that $\mathbf{G}_F$ controls the learning rate of a network: a higher value of $\mathbf{G}_F[\dot{\hat{\boldsymbol{\omega}}}_T]$ leads to slower or negative learning. For this reason, we refer to $\mathbf{G}_F$ as the **local task difficulty**. Again, the $\mathcal{O}(T^2)$ term may not be small. However, as mentioned, it is reasonable to assume this term cannot be controlled by the learning rule because it depends on higher order derivatives of F that synapses are unlikely to be able to compute. Therefore we can reasonably say that learning requires the first term of (7) to be negative, and ceases to occur when this term is zero. This implies:

$$\mathbf{G}_F[\dot{\hat{\boldsymbol{\omega}}}_T] \leq \frac{\gamma_1}{T\Big(\gamma_1^2 + \gamma_2^2 + \gamma_3^2 \frac{N}{T}\Big)}. \tag{8}$$

This inequality relates the intrinsic 'learnability' of a task (local task difficulty, $\mathbf{G}_F$), the rate of information on task error (T), the quality of the learning rule (relative magnitudes of γ_1 and γ_2), the network size (N), and the intrinsic noisiness of synapses (γ_3).

If inequality (8) is broken, then learning stops entirely. At some point in learning, this breakage is inevitable: as $F[\mathbf{w}]$ approaches a local minimum, the gradient $\nabla F[\mathbf{w}]$ approaches zero, and the Hessian $\nabla^2 F[\mathbf{w}]$ is guaranteed positive semidefinite. At a precise minimum of error, $\mathbf{G}_F$ becomes unbounded. This means that for a nonzero T, cessation of learning is preceded by an increase in local task difficulty, and learning stops just as inequality (8) above is broken.

To validate our analysis we numerically computed the quantities in (7) in simulations (Figure 4). In the case of a linear network with quadratic error the $\mathcal{O}(T^2)$ terms disappear, allowing us to verify that equality in (8) indeed predicts the steady-state value of $\mathbf{G}_F$. This agreement is confirmed in Figure 4A.

For more general error functions, we have observed that (8) is always conservative in numerical simulations: learning stops before local task difficulty reaches the critical value value, implying that the $\mathcal{O}(T^2)$ term of (7) is usually negative. This is demonstrated in Figures 4A and 4B.

In summary, we have shown that local task difficulty $\mathbf{G}_F$ determines the learning rate as well as the steady state learning performance of a network.

Local task difficulty as a function of network size

We next show precisely how network size influences the local task difficulty, and thus learning rate and steady state performance when other factors such as noise and the task itself remain the same.

Recall that $\mathbf{n}_2$ represents the direction in which synapses are perturbed due to error in the learning rule and other task-irrelevant changes that affect all synapses. Meanwhile $\mathbf{n}_3$ represents the direction of weight change due to intrinsic white-noise fluctuations at each synapse. For arbitrary tasks, networks and learning trajectories we can model these terms as coming from mutually independent probability distributions that are independent of task error $F[\mathbf{w}]$ and its derivatives. Thus we assume $\mathbb{E}[\mathbf{n}_2^T \mathbf{n}_3] = 0$, for any i, which allows us to write an expression for expected local task difficulty:

$$\mathbb{E}\big[\mathbf{G}_F[\dot{\hat{\boldsymbol{\omega}}}_T]\big] = \gamma_1^2 \mathbf{G}_F^1[\mathbf{w}(0)] + \frac{Tr(\nabla^2 F[\mathbf{w}(0)])}{2\|\nabla F[\mathbf{w}(0)]\|_2}\left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T}\right], \tag{9}$$

where

$$\mathbf{G}_F^1[\mathbf{w}(0)] = \frac{1}{2\|\nabla F[\mathbf{w}(0)]\|_2}\Big\langle \nabla \hat{F}[\mathbf{w}(0)], \nabla^2 F[\mathbf{w}(0)] \nabla \hat{F}[\mathbf{w}(0)]\Big\rangle. \tag{10}$$

This new expression for the local task difficulty explicitly incorporates N, the number of synaptic weights. So too does the learning rate (7), as we see by substituting into it the expanded forms of $\mathbb{E}[\|\dot{\boldsymbol{\omega}}_T\|_2^2]$ and $\mathbb{E}\big[\mathbf{G}_F[\dot{\hat{\boldsymbol{\omega}}}_T]\big]$.

We can gain intuition into how (9) is derived without going through additional technical details (*SI section 'Decomposition of local task difficulty'*). Suppose that the weights were perturbed by a randomly chosen direction $\mathbf{n}$ over the time interval $[0, T]$. This gives:

$$\begin{aligned} &F[\mathbf{w}(T)] = F[\mathbf{w}(0) + \mathbf{n}] = \\ &F[\mathbf{w}(0)] + \langle \nabla F[\mathbf{w}(0)], \mathbf{n}\rangle + \langle \mathbf{n}, \nabla^2 F[\mathbf{w}(0)]\mathbf{n}\rangle + \mathcal{O}(\|\mathbf{n}\|_2^2) \end{aligned} \tag{11}$$

If the direction $\mathbf{n}$ is drawn independently of task error and its derivatives, then

$$\mathbb{E}[\langle \nabla F[\mathbf{w}(0)], \mathbf{n}\rangle] = 0.$$

Therefore it is the quadratic term of (11) that determines the effect of the perturbation on task error. Its contribution is:

$$\begin{aligned}\mathbb{E}[\langle \mathbf{n}, \nabla^2 F[\mathbf{w}(0)]\mathbf{n}\rangle] &= \|\mathbf{n}\|_2^2 \mathbb{E}[\langle \hat{\mathbf{n}}, \nabla^2 F[\mathbf{w}(0)]\hat{\mathbf{n}}\rangle] \\ &= \|\mathbf{n}\|_2^2 \frac{Tr(\nabla^2 F[\mathbf{w}(0)])}{N}.\end{aligned}$$

So the effect of random perturbations on learning grows with the ratio of $Tr(\nabla^2 F[\mathbf{w}(0)])$ to the number of synapses, N. (9) tells us explicitly how local task difficulty (and thus expected learning rate and steady-state performance) can be modified by changing the size of a network, provided the size change leaves $\mathbf{G}_F^1[\mathbf{w}(0)]$ and $\frac{Tr(\nabla^2 F[\mathbf{w}(0)])}{2\|\nabla F[\mathbf{w}(0)]\|_2}$ unchanged. For different network architectures there are many possible ways of adding neurons and connections while satisfying these constraints. This explains why the naive size increases in Figure 2C generically increased learning performance, and provides a general explanation for enhanced learning performance in larger networks.

Network expansions that increase learning performance

We next give detailed examples of network expansions that increase learning rate and use the theory developed so far to compute the optimal size of a network when intrinsic noise is present. We first provide an analysis in the case of a linear network, which offers useful insight into the more general nonlinear case.

Consider a linear network (i.e. a linear map, as shown in Figure 5A that transforms any input u into an output $y = Wu$ for a matrix $W \in \mathbb{R}^{oi}$ of synaptic weights. The input-dependent error of the network is taken as a simple mean square error:

$$F[W] = \int F'[W', u']\mathbb{P}(u)\ du = \int \|y^*(u) - Wu\|_2^2 \mathbb{P}(u)\ du.$$

where the input vectors are drawn from some distribution $\mathbb{P}(u)$ (e.g. a Gaussian) and $y^*(u)$ is a target output generated by a linear mapping of the same rank and dimension.

We next embed this network in a larger network with $c_1 i$ inputs, $c_2 o$ outputs and a synaptic weight matrix W', for some integers c_1, $c_2 > 1$. We will define the total number of weights as $\tilde{N} = c_1 i c_2 o$. We take the transformation $u' = Bu \in \mathbb{R}^{c_1 i}$, where $B \in \mathbb{R}^{c_1 i \times i}$ is an arbitrary semi-orthogonal matrix. (i.e. it satisfies $B^T B = \mathbb{I}_i$). Geometrically, B therefore represents the composition of a projection into the higher dimensional space $\mathbb{R}^{c_1 i}$ with a rotation. Note that this is an invertible mapping: if $u' = Bu$ then $B^T u' = u$. Similarly, we can take $y'^*(u') = Dy^*(u) \in \mathbb{R}^{c_2 o}$, where $D^T D = \mathbb{I}_o$. This is illustrated in Figure 5A.

The expanded neural network with weights $W' \in \mathbb{R}^{c_1 o \times c_2 i}$ has to learn the same mapping as the original, but with respect to the higher dimensional inputs. So the network receives inputs $u' \in B\mathcal{U}$, and transforms them to outputs $y' = W'u'$, with input-dependent error

$$\begin{aligned}F'[W', u'] &= \|y'^*(u') - W'u'\|_2^2 \\ &= \|Dy^*(u) - W'Bu\|_2^2 \\ &= \|y^*(u) - D^T W'Bu\|_2^2 \qquad (12)\end{aligned}$$

For some weight configuration W' in an expanded network, (12) tells us that if these weights are related to the original network weights by $W = D^T W'B \in \mathbb{R}^{oi}$, then we have:

$$F'[W'] = F[W].$$

If we rewrite the weight matrices W' and W as vectors $\mathbf{w}' \in \mathbb{R}^{\tilde{N}}$ and $\mathbf{w} \in \mathbb{R}^N$, then $\mathbf{w} = H\mathbf{w}'$ for some H with $HH^T = \mathbb{I}_N$. We can then apply the chain rule to relate gradients and hessians in the original and expanded networks:

$$\begin{aligned}H\nabla F'[\mathbf{w}'] &= \nabla F[\mathbf{w}], \\ H\nabla^2 F'[\mathbf{w}']H^T &= \nabla^2 F[\mathbf{w}].\end{aligned}$$

Semi-orthogonality of H implies that it has N singular values with value one, and $\tilde{N} - N$ singular values with value zero. We take the heuristic that $\mathbf{w}'$ should project approximately equally onto each of the associated singular vectors. This is reasonable given that we took a random choice of $\mathbf{w}'$. This implies

$$\frac{\|\nabla F'[\mathbf{w}']\|_2}{\|\nabla F[\mathbf{w}]\|_2} \approx \frac{\tilde{N}}{N} = c_1 c_2 \qquad (13)$$

Meanwhile the quadratic error function implies $\nabla^2 F'[\mathbf{w}]$ is constant. The trace of this matrix can thus be calculated explicitly as

$$Tr(\nabla^2 F'[\mathbf{w}']) = (c_2 o)^2 \int_{u \in \mathcal{U}} \|u\|_2^4\ \ \mathbb{P}(u). \qquad (14)$$

We will also assume that $\nabla \hat{F'}[\mathbf{w}']$ (which is a normalised vector) projects approximately equally onto the different eigenvectors of $\nabla^2 F'[\mathbf{w}']$. The latter is constant, whereas the former is a linear function of the (randomly chosen) W', which justifies this assumption. In this case, (14) implies

$$\nabla\hat{F}'[\mathbf{w}']^T\nabla^2 F'[\mathbf{w}']\nabla\hat{F}'[\mathbf{w}'] \approx (c_2)^2\nabla\hat{F}[\mathbf{w}]^T\nabla^2 F[\mathbf{w}]\nabla\hat{F}[\mathbf{w}]. \qquad (15)$$

Bringing together equations (13), (15), and the formula (10) for $\mathbf{G}_F^1$, we see that

$$\mathbf{G}'^{1}_{F'}[\mathbf{w}'] \approx \frac{(c_2)^2}{c_1 c_2}\mathbf{G}_F^1[\mathbf{w}] = \frac{c_2}{c_1}\mathbf{G}_F^1[\mathbf{w}],$$

similarly

$$\frac{Tr(\nabla^2 F'[\mathbf{w}'])}{\|\nabla F'[\mathbf{w}']\|_2} \approx \frac{c_2}{c_1}\frac{Tr(\nabla^2 F[\mathbf{w}])}{\|\nabla F[\mathbf{w}]\|_2}.$$

Equations (16) imply that we can alter N as an independent parameter in the equations for local task difficulty (9) and speed of weight change (6), as long as the ratio $\frac{c_2}{c_1}$ is preserved. Indeed this allows us to optimise the steady state error of the network by changing N. To see how, recall that

$$\mathbb{E}[k] = \frac{-\|\nabla F[\mathbf{w}(0)]\|_2}{F[\mathbf{w}(0)]}\Big[-\gamma_1 + \boldsymbol{\delta}\Big] + \mathcal{O}(T^2),$$
$$\text{where } \boldsymbol{\delta} = \mathbf{G}_F[\dot{\boldsymbol{\omega}}_T]\|\dot{\boldsymbol{\omega}}_T\|_2^2 T,$$

with $\mathcal{O}(T^2) \equiv 0$ for quadratic error. Suppose the network has reached steady state error, i.e $\mathbb{E}[k] = 0$. If we decreased $\boldsymbol{\delta}$, then $\mathbb{E}[k]$ would also decrease, and the network would learn further. Therefore, to derive the optimal N^*, we should minimise the expression for $\boldsymbol{\delta}$ in N. We differentiate $\boldsymbol{\delta}$ in N, and note that stationary points satisfy the equation:

$$N^2\left[\gamma_1^2 CT + \gamma_3^2\right] = \frac{T^2\gamma_2^2}{\gamma_3^2}\left(\gamma_1^2 + \gamma_2^2\right)$$
$$\text{where } C = \frac{\left\langle \nabla\hat{F}[\mathbf{w}(0)], \nabla^2 F[\mathbf{w}(0)]\nabla\hat{F}[\mathbf{w}(0)]\right\rangle}{Tr(\nabla^2 F[\mathbf{w}(0)])}.$$

For $\gamma_2 \neq 0$, this implies the existence of two stationary points differing only in sign. Since $\lim_{N\to\infty}\boldsymbol{\delta} = \infty$, and $\lim_{N\to-\infty}\boldsymbol{\delta} = -\infty$, the positive stationary point is necessarily a global minimum of $\boldsymbol{\delta}$. So this stationary point defines N^*. We have

$$N_{\text{opt}} = \frac{T\gamma_2}{\gamma_3}\sqrt{\frac{1 + \frac{\gamma_2^2}{\gamma_1^2}}{CT + \frac{\gamma_3^2}{\gamma_1^2}}}. \qquad (17)$$

Note that C is unknown in general because it depends on the weight configuration of the network. However, we can take the heuristic $C \approx \frac{1}{N_{\text{opt}}}$, since if the gradient $\nabla\hat{F}[\mathbf{w}(0)]$ is uncorrelated with the Hessian $\nabla^2\hat{F}[\mathbf{w}(0)]$, then it would project equally onto each of the eigenvalues of the latter, and thus the denominator of C would be the mean eigenvalue, i.e. $\frac{Tr(\nabla^2\hat{F}[\mathbf{w}(0)])}{N}$. This results in an approximate expression for the optimal network size:

$$N_{\text{opt}} \approx \frac{T\gamma_2}{\gamma_3}\sqrt{\frac{1 + \frac{\gamma_2^2}{\gamma_1^2}}{\frac{T}{N_{\text{opt}}} + \frac{\gamma_3^2}{\gamma_1^2}}}, \qquad (18)$$

which can be solved for N_{opt} algebraically. Note that the optimal N_{opt} loses any dependence on the specific value of $\frac{Tr(\nabla^2 F[\mathbf{w}(0)])}{2\|\nabla F[\mathbf{w}(0)]\|_2}$. Our formula is verified numerically in Figure 6A, by evaluating the learning performance of transformed neural networks of different sizes, with different γ_i values.

This estimate of the optimal network size is plotted in Figure 5B, which shows the dependence on intrinsic synaptic noise levels. As noise decreases to zero, we see that the optimal network size grows arbitrarily. In addition, the optimal network size is smaller for a lower amount of task-irrelevant plasticity (i.e. a 'better' learning rule). We validate the optimal network size estimate in Figure 6A with numerical simulations.

We next consider nonlinear multilayer, feedforward networks. Again, we use the student-teacher framework to generate learning tasks. We will consider learning performance of a nominal and expanded network, both with l layers, and both using the same learning rule. The only difference between the two networks will be the larger number of neurons in each hidden layer of the expanded network. Unlike the linear case, this size expansion will modify some factors in the learning rate equation. Nevertheless, we can use our theory to predict an optimal number of synapses (and consequently optimal hidden layer sizes) for the transformed network. As before, this size will depend on the learning rule used by the networks, which is defined by levels of task-relevant plasticity, task-irrelevant plasticity, per-synapse white-noise intensity, and frequency of task error feedback. Our predictions are validated in simulations in Figure 6B.

We first describe the nominal network architecture. Given a vector $h^{(k-1)}$ of neural activities at layer $k-1$, the neural activity at layer $h^{(k)}$ is

$$h^{(k)} = \underline{\sigma}(W^{(k)}h^{(k-1)}).$$

Here, $W^{(k)}$ is the matrix of synaptic weights at the k^{th} layer. The concatenation of the synaptic weight matrices across all layers is denoted W, and has N elements. We will interchangeably denote it as a vector $\mathbf{w} \in \mathbb{R}^N$. The function σ passes its arguments elementwise through some nonlinearity. The first layer of neurons receives an input vector u in place of neural activities $h^{(0)}$. The output $y(W, u)$ is defined as neural activity at the final hidden layer.

For any given state $\mathbf{w}$ of the nominal network, we can construct a state $\phi(\mathbf{w})$ of the expanded network with the same input-output properties, i.e. $y(\mathbf{w}, u) = y'(\phi(\mathbf{w}), u)$, where y' denotes expanded network output. We do this by setting synaptic weights of the added neurons in the expanded network to zero, so the neurons do not contribute at state $\phi(\mathbf{w})$. Nevertheless the extra neurons can affect expanded network behaviour because once they are perturbed by the learning rule they contribute to the error gradient and higher derivatives.

Suppose the nominal network is at state $\mathbf{w}(0)$, and a learning rule picks the direction $\dot{\boldsymbol{\omega}}_T$ for weight change over the time interval $[0, T]$. This direction will have some local task difficulty $\mathbf{G}_F[\dot{\hat{\boldsymbol{\omega}}}_T]$. If we map the state $\mathbf{w}(0)$ and the direction $\dot{\boldsymbol{\omega}}_T$ to the transformed network via the transformation ϕ, then we can estimate local task difficulty $\mathbf{G}'_F[\phi(\dot{\hat{\boldsymbol{\omega}}}_T)]$ of the transformed network (see *SI section 'Learning in a Nonlinear, Feedforward Network'* for additional detail). We get:

$$\mathbb{E}[\mathbf{G}'{}_{F'}[\phi(\dot{\hat{\boldsymbol{\omega}}}_T)]] \approx \frac{N}{\tilde{N}}\gamma_1^2 \mathbf{G}_F^1[\mathbf{w}(0)] + \frac{Tr(\nabla^2 F[\mathbf{w}(0)])}{2\|\nabla F[\mathbf{w}(0)]\|_2}\left[\frac{\gamma_2^2}{\tilde{N}} + \frac{\gamma_3^2}{T}\right] \tag{19}$$

We can use (19) to minimise

$$\boldsymbol{\delta} = \mathbf{G}'_F[\dot{\hat{\boldsymbol{\omega}}}_T]\|\dot{\boldsymbol{\omega}}_T\|_2^2 T$$

in $\tilde{N}$, the number of synapses. This gives an optimal size of the transformed network for minimising steady state task performance (validated in Figure 6B):

$$N_{\text{opt}} \approx \frac{T\gamma_2}{\gamma_3^2}\sqrt{\left(1 + \frac{\gamma_1^2 N}{\gamma_2^2 N_{\text{opt}}}\right)\left(\gamma_1^2 + \gamma_2^2\right)}. \tag{20}$$

Note the dependence of the optimal network size, N_{opt}, on N, which is the number of weights in the nominal and the teacher networks. The teacher networks can generate arbitrary nonlinear mappings whose complexity grows with N. In this way the above formula (Equation 20) reflects the *intrinsic difficulty* of the task.

Discussion

It is difficult to disentangle the physiological and evolutionary factors that determine the size of a brain circuit [28, 29, 30]. Previous studies focused on the energetic cost of sustaining large numbers of neurons and connecting them efficiently [2, 29, 31, 30]. Given the significant costs associated with large circuits [3], it is clear that some benefit must offset these costs, but it is currently unclear whether other inherent tradeoffs constrain network size. We showed under broad assumptions that there is an upper limit to the learning performance of a network which depends on its size and the intrinsic reliability of synapses.

The neural circuits in animals with large brains were presumably shaped on an evolutionary timescale by gradual addition of neurons and connections. Expanding a small neural circuit into a larger one can increase its dynamical repertoire, allowing it to generate more complex behaviours [32, 33]. Less obviously, as we show here, circuit expansion can also allow a network to learn simpler tasks more quickly and to greater precision.

By directly analysing the influence of synaptic weight configurations on task error we derived a quantity we called 'local task difficulty', that determines how easily an arbitrary network can learn. We found that local task difficulty always depends implicitly on the number of neurons, and can therefore be decreased by adding neurons according to relatively unrestrictive constraints. In simple terms, adding redundancy flattens out the mapping between synaptic weights and task error, reducing the local task difficulty on average. This flattening makes learning faster, and steady-state task error lower because the resulting error surface is less contorted and easier to descend using local task error information.

As an analogy, imagine hiking to the base of a mountain without a map, and doing so using intermittent estimates of the slope underfoot. A more even slope will be easier to descend because slope estimates will remain consistent. An undulating slope will be harder to descend because the direction of descent necessarily changes with location. Now consider the same hike in a heavy fog at dusk. The undulating slope will become far harder to descend. However, if it were possible to somehow smooth out the undulations (that is, reduce local task difficulty), the same hike would progress more efficiently. This analogy shows why larger neural circuits are able to achieve better learning performance in a given task where error information is corrupted.

In specific examples we show that adding neurons to intermediate layers of a multilayer, feedforward network, increases the magnitude of the slope (gradient) of the task error function relative to its curvature. From this we provide a template for scaling up network architecture such that both quantities increase approximately equally. This provides novel hypotheses for the organising principles in biological circuits which, among other things, predicts a prevalence of apparently redundant connections in networks that need to learn new tasks quickly and to high accuracy. Recent experimental observations reveal such apparently redundant connections in a number of brain areas across species [34, 35, 27, 26].

Even if neurons are added to a network in a way that obeys the architectural constraints we derive, intrinsic synaptic noise eventually defeats the benefits conferred to learning. All synapses are subject to noisy fluctuations due to their molecular makeup [23, 36, 24, 37, 25, 38]. These sources of noise are distinct from shared noise in a feedback signal that is used in learning. Such independent noise sources accumulate as a network grows in size, outcompeting the benefit of size on learning performance. An immediate consequence is an optimal network size for a given task and level of synaptic noise.

Furthermore, our results show that different noise sources in nervous systems impact learning in qualitatively different ways. Noise in the learning rule as well as external noise in the task error, which may arise from sensory noise or fluctuations in the task, can be overcome in a larger circuit. On the other hand, the impact of intrinsic noise in the synaptic connections only worsens as network size grows.

Our analysis allowed us to predict the optimal size of a network in theoretical learning tasks where where we can specify the levels of noise in the learning rule and in synapses. Figure 5 shows that the optimal network size decreases rapidly as the intrinsic noise in synapses increases. We speculate that the emergence of large neural circuits therefore depended on evolutionary modifications to synapses that reduce intrinsic noise. An intriguing and challenging goal for future work would be to infer noise parameters in synapses across different nervous systems and test whether overall network size obeys the relationships our theory predicts.

Author Affiliations

[a] Department of Engineering, University of Cambridge, Trumpington Street, CB21PZ.

1 Methods

We test our predictions in simulations of neural network models. This section describes the details of the models used. Code is available on request.

1.1 Network architectures

All tested neural networks have fully-connected feedforward architectures, to enable exact gradient computation. The simplest considered network is linear. Given an input $u \in \mathbb{R}^i$, this gives an output of the form

$$y(u) = Wu \in \mathbb{R}^o,$$

where $W \in \mathbb{R}^{oi}$ is a matrix of synaptic weights.

More commonly we consider networks with nonlinearities and hidden layers. In the main text, we refer to these as nonlinear, feedforward networks. Each neuron in these networks passes inputs through a sigmoidal nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ of the form

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

We use the notation $\underline{\sigma} : \mathbb{R}^m \to \mathbb{R}^m$ to represent the elementwise application of σ to a vector of arbitrary length $m \in \mathbb{N}$. Let us denote neural activity at the k^{th} layer via a vector h^k. If $k > 1$, then

$$h^k = \underline{\sigma}(W^k \tilde{h}^{k-1}),$$

where W^k is a matrix of synaptic weights and $\tilde{h}^{k-1}$ is the concatenation of the vector h^{k-1} with the scalar -1. This scalar is known as a bias neuron, and allows for nonzero h^k even when each component of h^{k-1} is zero. Meanwhile for $k = 1$, we have

$$h^1 = \underline{\sigma}(W^1 u),$$

where $u \in \mathbb{R}^i$ is a vector of inputs to the network. Let F be the final hidden layer of the network. Then network output $y \in \mathbb{R}^o$ is taken as

$$y = \underline{\sigma}(W^k \tilde{h}^F),$$

which will subsequently be denoted $y(\mathbf{w}, u)$. Here $\mathbf{w}$ is a vector that concatenates all entries of each weight matrix in the network.

1.2 Details of Learning Tasks

For all training tasks used in simulation, we generate a random fixed mapping $y^*(u)$, along with a set of inputs $\mathcal{U}$, consisting of 1000 elements. The mean square error of the network is given as

$$\sum_{u\in\mathcal{U}} \|y^*(u) - y(\mathbf{w}, u)\|_2^2.$$

This corresponds to the error function $F[\mathbf{w}]$ used in the main text.

We often consider a family of nonlinear neural networks learning the same task (i.e. the same mapping $y^*(u)$, using the same set $\mathcal{U}$ of inputs). Each network in the family has the same number of inputs, outputs, and hidden layers, but the number of hidden layer neurons varies. In this case, we must ensure *a priori* that each neural network in the family is capable of learning the task perfectly, i.e. exactly recreating the mapping $y^*(u)$. To do this, we create a 'teacher' network with the same architecture as the smallest network in the family (i.e. the network with the smallest number of hidden layer neurons). The synaptic weights of the teacher network are initialised randomly and then fixed, to generate a mapping $y^*(u)$ of the form described previously. Specifically, they are generated from a uniform distribution centred at 0. The support of the distribution is set at $[-a, a]$, where a is chosen such that the standard deviation of the weights is $\frac{1}{\sqrt{i}}$, where i is the number of inputs the network receives. The weights of the smallest network in the family are randomly initialised from the same distribution. The weights of larger networks in the family are initialised such that the network's input-output behaviour exactly corresponds to the smallest network in the family. This is achieved by successively adding neurons to the smallest network in such a way that each addition does not change the input-output properties of the network. In this way, the initial task errors $F[\mathbf{w}]$ of all networks in the family are identical.

1.3 Network Training

All training protocols share the same basic template: they combine task relevant plasticity (i.e. gradient information), task-irrelevant plasticity processes, and a white-noise process at each synapse. The proportions of each of these three factors are respectively represented by parameters γ_1, γ_2, and γ_3. Meanwhile, the direction of plasticity changes upon reception of sensory feedback, which happens every T units of time. We refer to each time interval of length T as a single learning cycle. The exact values of the γ components will fluctuate between learning cycles (detailed subsequently). We can set a vector $\bar{\gamma}$ of their expected values. The synaptic weight vector is then updated according to the following formula:

$$\mathbf{w}_{t+T} = \mathbf{w}_t - T\bar{\gamma}_1\nabla\hat{F}[\mathbf{w}_t] + T\bar{\gamma}_2\hat{\mathbf{n}}_2^t + \sqrt{NT}\bar{\gamma}_3\hat{\mathbf{n}}_3^t.$$

N represents the number of synaptic weights, while $\hat{\mathbf{n}}_3^t$ is a Gaussian random variable, normalised to ensure $\|\hat{\mathbf{n}}_3^t\|_2 = 1$. Thus $\sqrt{NT}\bar{\gamma}_3\hat{\mathbf{n}}_3^t$ represents per-synapse white noise. Meanwhile, $\hat{\mathbf{n}}_2^t$ is the state of a normalised random process uncorrelated with the task. The dynamics of the unnormalised random process satisfy $\hat{\mathbf{n}}_2^{t+1} = \sqrt{0.1}\hat{\mathbf{n}}_2^t + \sqrt{0.9}\hat{\nu}$, where ν is a Gaussian random variable, normalised such that $\|\nu\|_2 = 1$. So $T\bar{\gamma}_2\hat{\mathbf{n}}_2^t$ represents the effect of partially systematic, task-irrelevant plasticity processes.

Since

$$\mathbb{E}[\langle\nabla F[\mathbf{w}], \hat{\mathbf{n}}_i^t\rangle = 0, \quad i \in \{1, 2\}$$

we have that

$$\mathbb{E}[\gamma_i] = \bar{\gamma}_i, \quad i \in \{1, 2, 3\}.$$

However, on a given learning cycle, the task-irrelevant components $\hat{\mathbf{n}}_2^t$ and $\hat{\mathbf{n}}_3^t$ will have some nonzero correlation with the gradient of task error. Consequently, the values of γ on this learning cycle will differ from those of $\bar{\gamma}$. For instance, if the task irrelevant components anti-correlate with the gradient, then γ_1 will increase at the expense of γ_2 and γ_3.

We thank Fulvio Forni, Rodrigo Echeveste and Aoife McMahon for careful readings of the manuscript; we thank Rodolphe Sepulchre and Stephen Boyd for helpful discussions. This work is supported by ERC Grant StG2016 - FLEXNEURO (716643).

References

[1] Laughlin SB, van Steveninck RRdR, Anderson JC (1998) The metabolic cost of neural information. *Nature neuroscience* 1(1):36.

[2] Tomasi D, Wang GJ, Volkow ND (2013) Energetic cost of brain functional connectivity. *Proceedings of the National Academy of Sciences* 110(33):13642–13647.

[3] Attwell D, Laughlin SB (2001) An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism* 21(10):1133–1145.

[4] Reader SM, Laland KN (2002) Social intelligence, innovation, and enhanced brain size in primates. *Proceedings of the National Academy of Sciences* 99(7):4436–4441.

[5] Sol D, Duncan RP, Blackburn TM, Cassey P, Lefebvre L (2005) Big brains, enhanced cognition, and response of birds to novel environments. *Proceedings of the National Academy of Sciences* 102(15):5460–5465.

[6] Joffe TH, Dunbar R (1997) Visual and socio–cognitive information processing in primate brain evolution. *Proceedings of the Royal Society of London B: Biological Sciences* 264(1386):1303–1307.

[7] Maguire EA, et al. (2000) Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences* 97(8):4398–4403.

[8] Gaser C, Schlaug G (2003) Brain structures differ between musicians and non-musicians. *Journal of Neuroscience* 23(27):9240–9245.

[9] Black JE, Isaacs KR, Anderson BJ, Alcantara AA, Greenough WT (1990) Learning causes synaptogenesis, whereas motor activity causes angiogenesis, in cerebellar cortex of adult rats. *Proceedings of the National Academy of Sciences* 87(14):5568–5572.

[10] Lawrence S, Giles CL, Tsoi AC (1998) What size neural network gives optimal generalization? convergence properties of backpropagation, (UMIACS), Technical report.

[11] Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks in *Advances in neural information processing systems*. pp. 1097–1105.

[12] Huang GB (2003) Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Transactions on Neural Networks* 14(2):274–281.

[13] Takiyama K (2016) Maximization of learning speed due to neuronal redundancy in reinforcement learning. *Journal of the Physical Society of Japan* 85(11):114801.

[14] Takiyama K, Okada M (2012) Maximization of learning speed in the motor cortex due to neuronal redundancy. *PLoS computational biology* 8(1):e1002348.

[15] Saxe AM, McClelland JL, Ganguli S (2013) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

[16] Seung HS (2003) Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron* 40(6):1063–1073.

[17] Werfel J, Xie X, Seung HS (2004) Learning curves for stochastic gradient descent in linear feedforward networks in *Advances in neural information processing systems*. pp. 1197–1204.

[18] Contractor A, Klyachko VA, Portera-Cailliau C (2015) Altered neuronal and circuit excitability in fragile x syndrome. *Neuron* 87(4):699–715.

[19] Rinaldi T, Perrodin C, Markram H (2008) Hyperconnectivity and hyper-plasticity in the medial prefrontal cortex in the valproic acid animal model of autism. *Frontiers in neural circuits* 2:4.

[20] Casanova MF, et al. (2006) Minicolumnar abnormalities in autism. *Acta neuropathologica* 112(3):287.

[21] Amaral DG, Schumann CM, Nordahl CW (2008) Neuroanatomy of autism. *Trends in neurosciences* 31(3):137–145.

[22] Loewenstein Y, Yanover U, Rumpel S (2015) Predicting the dynamics of network connectivity in the neocortex. *Journal of Neuroscience* 35(36):12535–12544.

[23] Ziv NE, Brenner N (2017) Synaptic tenacity or lack thereof: Spontaneous remodeling of synapses. *Trends in neurosciences*.

[24] Minerbi A, et al. (2009) Long-term relationships between synaptic tenacity, synaptic remodeling, and network activity. *PLoS biology* 7(6):e1000136.

[25] Puro DG, De Mello FG, Nirenberg M (1977) Synapse turnover: the formation and termination of transient synapses. *Proceedings of the National Academy of Sciences* 74(11):4977–4981.

[26] Bloss EB, et al. (2018) Single excitatory axons form clustered synapses onto ca1 pyramidal cell dendrites. *Nature neuroscience* 21(3):353.

[27] Bartol Jr TM, et al. (2015) Nanoconnectomic upper bound on the variability of synaptic plasticity. *Elife* 4:e10778.

[28] Chittka L, Niven J (2009) Are bigger brains better? *Current biology* 19(21):R995–R1008.

[29] Herculano-Houzel S (2012) The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences* 109(Supplement 1):10661–10668.

[30] Shepherd GM, Stepanyants A, Bureau I, Chklovskii D, Svoboda K (2005) Geometric and functional organization of cortical circuits. *Nature neuroscience* 8(6):782.

[31] Shulman RG, Rothman DL, Behar KL, Hyder F (2004) Energetic basis of brain activity: implications for neuroimaging. *Trends in neurosciences* 27(8):489–495.

[32] Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *science* 313(5786):504–507.

[33] Tishby N, Zaslavsky N (2015) Deep learning and the information bottleneck principle in *Information Theory Workshop (ITW), 2015 IEEE.* (IEEE), pp. 1–5.

[34] Druckmann S, et al. (2014) Structured synaptic connectivity between hippocampal regions. *Neuron* 81(3):629–640.

[35] Eichler K, et al. (2017) The complete connectome of a learning and memory centre in an insect brain. *Nature* 548(7666):175.

[36] Mongillo G, Rumpel S, Loewenstein Y (2017) Intrinsic volatility of synaptic connections—a challenge to the synaptic trace theory of memory. *Current opinion in neurobiology* 46:7–13.

[37] Attardo A, Fitzgerald JE, Schnitzer MJ (2015) Impermanence of dendritic spines in live adult ca1 hippocampus. *Nature* 523(7562):592.

[38] Faisal AA, Selen LP, Wolpert DM (2008) Noise in the nervous system. *Nature reviews neuroscience* 9(4):292.

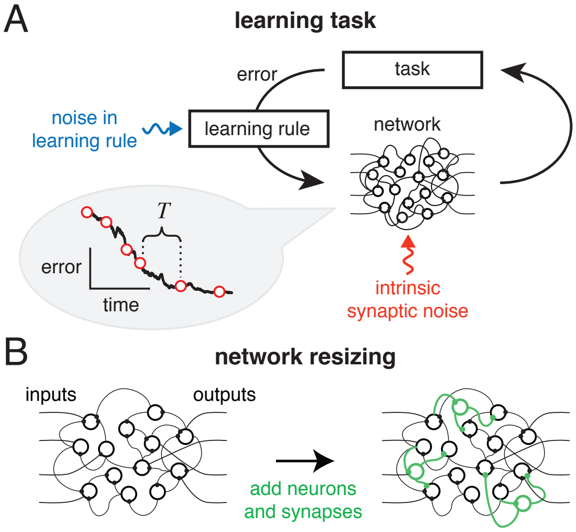

Figure 1: **A**: Schematic of learning in a neural network. Information on task error is received by a learning rule which converts this information into synaptic changes that decrease task error. Biologically, the learning rule faces several challenges: it will be subject to noise and perturbations (blue arrow), and the synapses themselves may suffer from intrinsic noise (red arrow). Error information will only acquired be intermittently, as shown in the learning curve on the left, where T specifies the intermittency of feedback (see main text). **B**: We analyse the effect of network size on learning performance by adding redundant neurons and synapses (green) to an existing network.

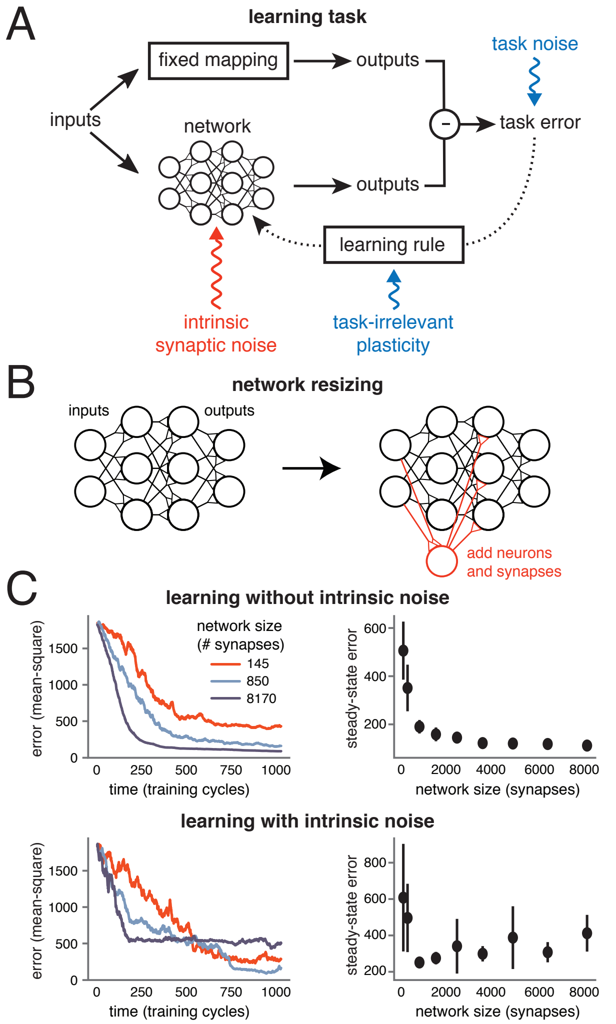

Figure 2: **A** Learning task: neuronal networks are trained to learn an input-output mapping using feedback error and a gradient-based learning rule to adjust synaptic strengths. The feedback is corrupted with tunable levels of noise (blue), reflecting imperfect sensory feedback, imperfect learning rules, and task-irrelevant changes in synaptic strengths. Synaptic strengths are additionally subject to independent internal noise (red), reflecting their inherent unreliability. **B**: Network size is increased by adding neurons and synapses to inner layers. **C**: Three differently sized networks are trained on the same task, with the same noise-corrupted learning rule. (Right) Mean task error after 1000 training cycles, computed over 12 simulations. Error bars depict ±1 standard deviation from the mean. (Left) task error over time for a single simulation of each network. **D**: As for (C) but each synapse is subject to internal independent noise fluctuations in addition to noise in the learning rule.

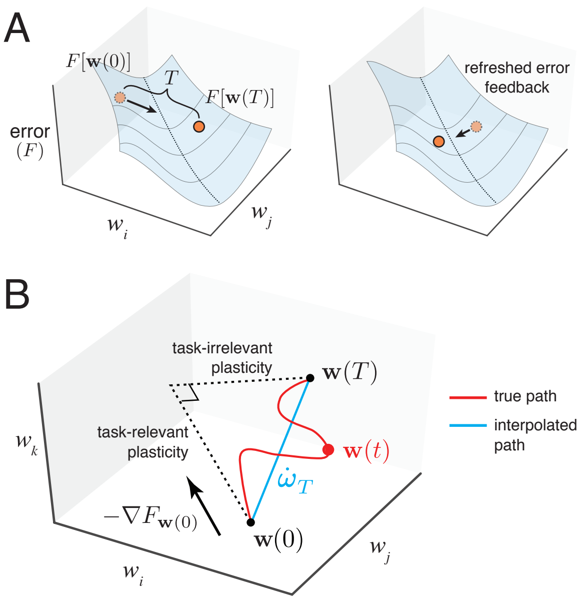

Figure 3: Geometry of error-based learning in arbitrary networks. **A**: Schematic of task error as a function of (two) synaptic weights. Learning rule receives and processes task-relevant feedback to provide direction for each synapse to move in weight space. Direction must correlate with steepest descent direction, resulting in initial improvement of task error. If no new feedback is received over a long time period T, this initially good direction may eventually go uphill, thus becoming bad. Frequent error feedback, a less 'curvy' error surface, and a good correlation with the initial steepest descent direction make learning faster. Local task difficulty captures these factors. **B**: Schematic of changes in three weights over interval $[0, T]$. The true weight trajectory $\mathbf{w}(t)$ over time (red line) is summarised by an interpolated, linear trajectory $\dot{\boldsymbol{\omega}}_T$ (blue line) between $\mathbf{w}(0)$ and $\mathbf{w}(T)$. We can decompose this interpolated trajectory into task-relevant and task-irrelevant plasticity components. The former is the initial steepest descent direction $-\nabla F[\mathbf{w}(0)]$, and the latter is the remaining orthogonal component.

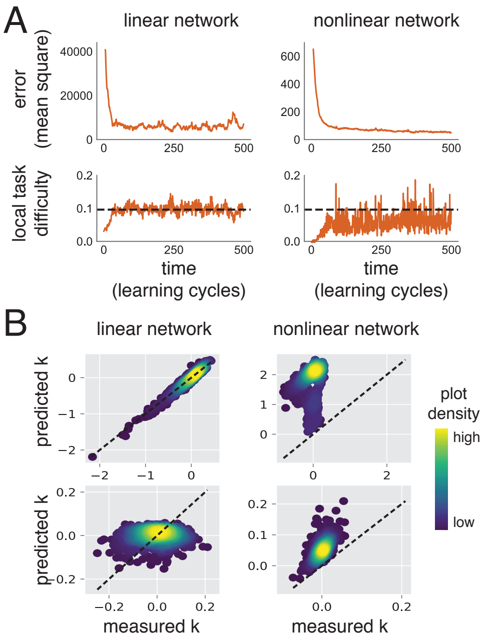

Figure 4: Numerical validation of learning rate calculations in simulated neural networks. **A**: Local task difficulty and MSE over time for a linear network (left) with quadratic error function, and a nonlinear network (right). Local task difficulty is low when the networks are in an untrained state. As performance improves, it rises, until reaching some steady state level (black dotted lines). We can predict this steady state *a priori*, exactly for the quadratic error, and conservatively for the nonlinear error, using (8). Both networks are trained using a corrupted learning rule ($\bar{\gamma} = [0.2, 1, 0]$, $T = 2$, see Methods). Network sizes are 200 synaptic weights (linear) and 220 synaptic weights (nonlinear, one hidden layer). **B**: We use the same linear (left) and nonlinear (right) networks as in A. We compare predicted value of the learning rate k_{pred} (using (7) and $\gamma = \bar{\gamma}$) with the actual value, under low-noise (top , $\bar{\gamma} = [1, 0.05, 0.05]$) and high-noise (bottom, $\bar{\gamma} = [0.2, 0.5, 0.1]$) conditions. Dotted lines represent $k = k_{\text{pred}}$. Density plots of $\{k, k_{\text{pred}}\}$ are shown. Two sources of discrepancy exist. First, k_{pred} is calculated from the mean values $\bar{\gamma}$ (see Methods). Transient correlations between task-irrelevant sources of plasticity and the gradient lead to unbiased fluctuations of γ around $\bar{\gamma}$. This is the only source of discrepancy in the linear case (left). Thus the density distributes equally on either side of the dotted line. In the nonlinear case (right), there is an unknown, nonzero $\mathcal{O}(T^2)$ term ((7)) unaccounted for in calculation of k_{pred}. This term almost always decreases learning rate, as k_{pred} now consistently overestimates k. Thus predicted steady state local task difficulty (e.g. dotted line, bottom-right panel of (A)) is consistently an overestimate.

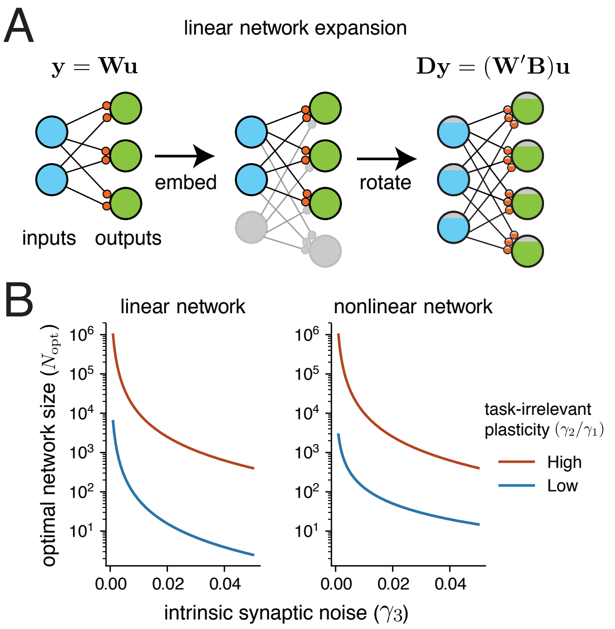

Figure 5: Optimal network size for linear and nonlinear networks in the presence of intrinsic synaptic noise. **A**: network expansion for a linear network, given by a embedding into a larger network, followed by a rotation of the weight matrix. This corresponds to transforming inputs $\mathbf{u}$ by a projection $\mathbf{B}$ and outputs $\mathbf{y}$ by a semiorthogonal mapping $\mathbf{D}$. **B**: Plots show the dependence of N_{opt} in linear and nonlinear networks using equations (18) and (20). In both cases the learning rule has $\gamma_1 = 0.1$. Low task-irrelevant plasticity corresponds to $\gamma_2 = 0.01$, while task-irrelevant plasticity corresponds to $\gamma_2 = 1$.

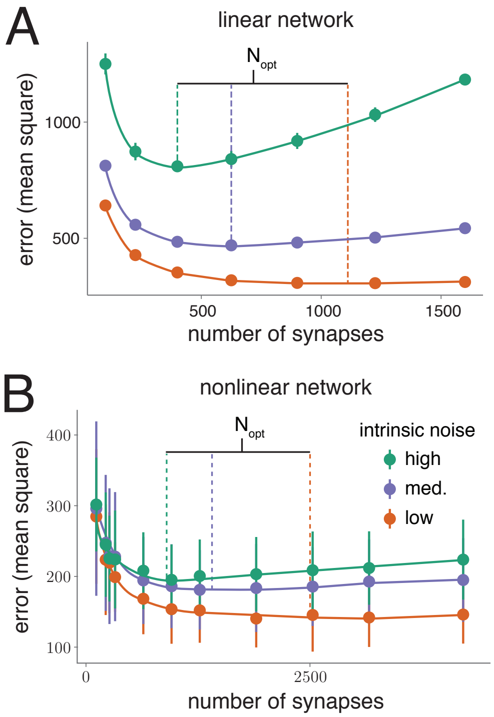

Figure 6: Testing analytic prediction of optimal network size for linear and nonlinear networks. Linear (A) and nonlinear (B) networks of different sizes are trained for 1500 learning cycles of length $T = 1$. Mean steady state error over 12 repeats is plotted against network size. Coloured lines represent *a priori* predicted optimal network sizes using equations (18) and (20) for the linear and nonlinear examples respectively. **A**: Linear networks all have a $2:1$ ratio of inputs to outputs. On each repeat, networks of all considered sizes learn the same mapping, embedded in the appropriate input/output dimension (as detailed in *SI section 'Learning in a Linear Network'*). The learning rule uses $\bar{\gamma} = [0.07, 1, 0.03]$ (high intrinsic noise), $\bar{\gamma} = [0.06, 1, 0.04]$ (medium intrinsic noise) and $\bar{\gamma} = [0.05, 1, 0.05]$ (low intrinsic noise). **B**: Nonlinear networks have sigmoidal nonlinearities at each neuron and a single hidden layer (see Methods). All networks have ten input and ten output neurons, and learn the same task. The number of neurons in the hidden layer is varied from 5 to 120. The learning rules all use $\bar{\gamma}_1 = 0.04$ and $\bar{\gamma}_2 = 1.5$. The value of $\bar{\gamma}_3$ is set respectively at 0.03, 0.04, and 0.05, in the low, medium, and high intrinsic noise cases.