Fundamental bounds on learning performance in neural circuits

Dhruva Venkita Raman[a,1], Adriana Perez Rotondo[a], and Timothy O'Leary[a,1]

[a]Department of Engineering, University of Cambridge, Cambridge CB21PZ, United Kingdom

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved March 4, 2019 (received for review August 3, 2018)

How does the size of a neural circuit influence its learning performance? Larger brains tend to be found in species with higher cognitive function and learning ability. Intuitively, we expect the learning capacity of a neural circuit to grow with the number of neurons and synapses. We show how adding apparently redundant neurons and connections to a network can make a task more learnable. Consequently, 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. 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. Above the optimal network size, the addition of neurons and synaptic connections starts to impede learning performance. This suggests that the size of brain circuits may be constrained by the need to learn efficiently with unreliable synapses and provides a hypothesis for 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.

learning | neural network | synaptic plasticity | optimization | artificial intelligence

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 (1–3), 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 (4–6). Within and across species, the volume of a brain region often correlates with the importance or complexity of the tasks it performs (7–9). These observations make sense from a theoretical perspective because larger artificial neural networks can solve more challenging computational tasks than smaller networks (10–15). However, we still lack a firm theoretical understanding of how network size improves learning performance.

Biologically it is not clear that there is always a computational advantage to having more neurons and synapses engaged in learning a task. During learning, larger networks face the problem of tuning greater numbers 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 unfavorably 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 (18–21). 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 these questions with a general mathematical analysis of learning performance in neural circuits that is independent of specific learning rules and circuit architectures. For a broad family of learning tasks, we show how the expected learning rate and steady-state performance are related to the size of a network. The analysis reveals how connections can be added to intermediate layers of a multilayer network to reduce the difficulty of learning a task. This gain in overall learning performance is accompanied by slower per-synapse rates 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 (22–25).

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 nonobvious consequence of this result is that the size

Significance

We show how neural circuits can use additional connectivity to achieve faster and more precise learning. Biologically, internal synaptic noise imposes an optimal size of network for learning a given task. Above the optimal size, addition of neurons and synaptic connections starts to impede learning and task performance. Overall brain size may therefore be constrained by pressure to learn effectively with unreliable synapses and may explain why certain neurological learning deficits are associated with hyperconnectivity. Beneath this optimal size, apparently redundant connections are advantageous for learning. Such apparently redundant connections have recently been observed in several species and brain areas.

Author contributions: D.V.R. and T.O. designed research; D.V.R. and A.P.R. performed research; D.V.R., A.P.R., and T.O. analyzed data; D.V.R. and T.O. wrote the paper; and T.O. interpreted results.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Code for figure simulations in this paper is available at https://github.com/olearylab/raman_etal_2018.

[1]To whom correspondence may be addressed. Email: tso24@cam.ac.uk or dvr23@cam.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1813416116/-/DCSupplemental.

Published online May 6, 2019.

of a neural circuit can either reflect the complexity of a fixed task or instead deliver greater learning performance on simpler, arbitrary tasks.

Results

Modeling the Effect of Network Size on Learning. Our goal is to analyze how network size affects learning and steady-state performance in a general setting depicted in Fig. 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 (Fig. 1*A*). 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 signaling 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 (Fig. 1*B*).

Before analyzing the general case, we motivate the analysis with simulations of fully connected, multilayer nonlinear feedforward neural networks that we trained to learn input–output mappings (Fig. 2*A*). We used the so-called student–teacher framework to generate tasks (e.g., refs. 28 and 29). A "teacher" network is initialized with random fixed weights. The task is for a randomly initialized "student" network to learn the input–output mapping of the teacher. This framework models learning of any task that can be performed by a feedforward neural network by setting the teacher as the network optimized to perform the task.

The sizes of the student networks were set by incrementally adding neurons and connections to internal layers of a network with the same initial connection topology as that of the teacher (Fig. 2*B* and *Materials and Methods*). This generated student networks of increasing size with the guarantee that each student can in principle learn the exact input–output mapping of the teacher.

Learning was simulated by modifying synapses with noise-corrupted gradient descent to mimic an imperfect biological learning rule. We emphasize that we do not assume 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 (Eq. **1**). We assume that learning must be performed online; that is, data arrive one sample at a time. We believe this captures a typical biological learning scenario where a learner gets intermittent examples and feedback.

The phenomena we wish to understand are shown in Fig. 2 *C* 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 synaptic noise (Fig. 2*C*). 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 Fig. 2*D*, adding intrinsic noise to the synapses of the student networks results in a nonmonotonic relationship between performance and network size. Beyond a certain size, both learning and asymptotic performance start to worsen.

The simulations in Fig. 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 analyzed how network size and imperfections in feedback learning rules impact learning in a general case.

We note that in machine learning, noise processes such as dropout and stochastic regularization (e.g., refs. 30–32) can be applied to improve generalization from finite training data. Intrinsic synaptic noise is qualitatively different from these regularization processes. In particular, the per-synapse magnitude of intrinsic noise remains constant, independent of network size or training level. Moreover, our simulations use online learning, which is distinct from the common machine-learning paradigm where data are divided into training and test sets. The implications of our paper for this paradigm are considered in *SI Appendix*, *Regularization and Generalization Error* and *Online Learning and Generalization Error*, where we also show that regularization can be incorporated as learning-rule noise.

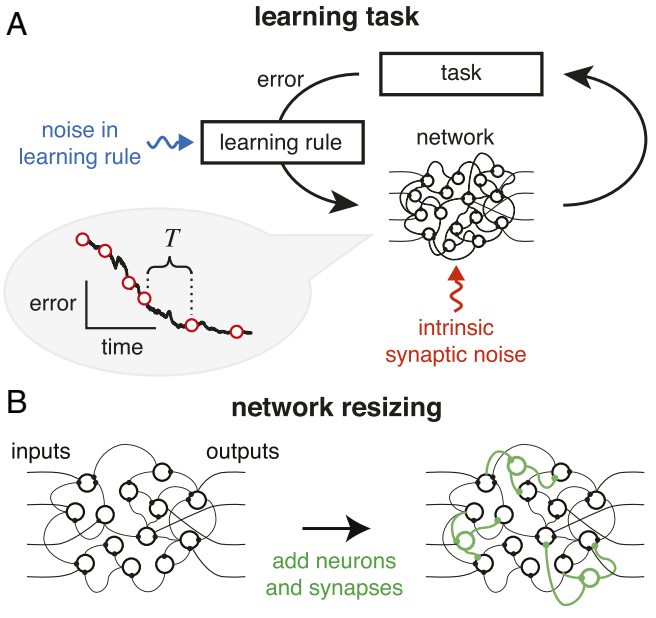

Fig. 1. (*A*) Schematic of learning in a neural network. Information on task error is received by a learning rule which converts this information into synaptic changes that decrease task error. Biologically, the learning rule faces several challenges: It will be subject to noise and perturbations (blue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

Learning Rate and Task Difficulty. We define task error as a smooth function $F[\mathbf{w}]$ 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 time points. Suppose feedback on task error occurs at time points 0 and T, but not in between, for some $T > 0$ (Fig. 3*A*). If the

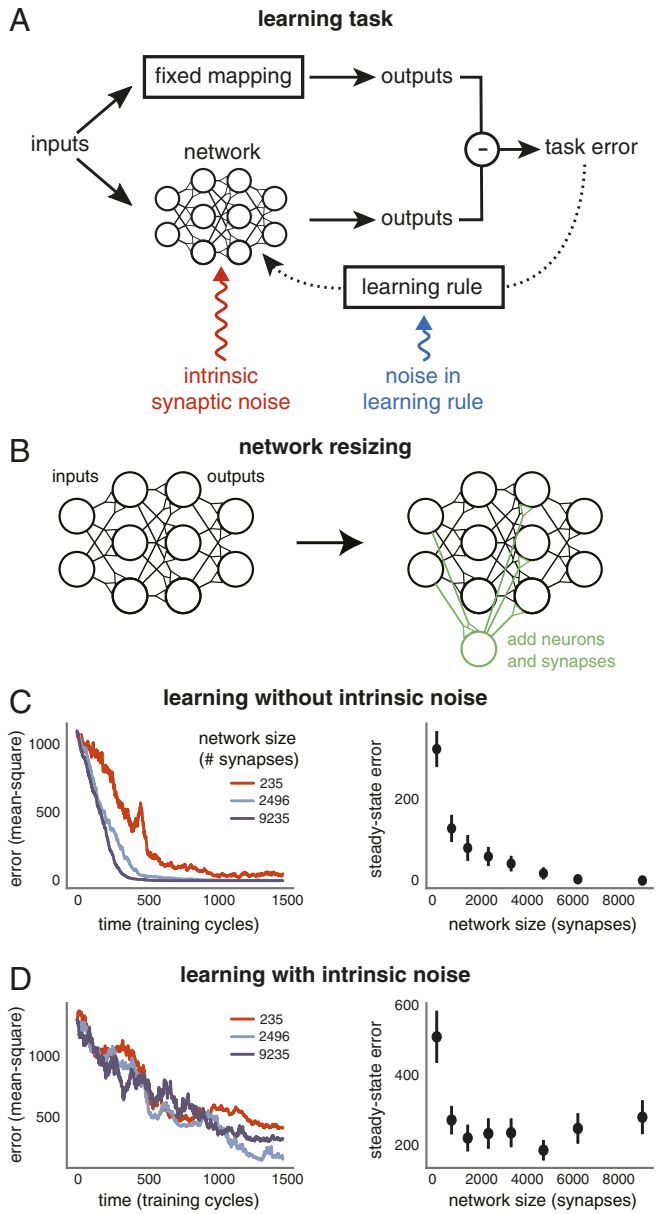

Fig. 2. (*A*) Learning task. Neuronal networks are trained to learn an input–output mapping using feedback error and a gradient-based learning rule to adjust synaptic strengths. The feedback is corrupted with tunable levels of noise (blue), reflecting imperfect sensory feedback, imperfect learning rules, 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. A learning cycle consists of a single input (drawn from a Gaussian distribution) being fed to the network. The gradient of feedback error with respect to this input (i.e., the stochastic gradient) is then calculated and corrupted with noise (blue component in *A*). All networks have five hidden layers of equal size. We vary this size from 5 neurons to 45 neurons across the networks. (*C*, *Right*) Mean task error after 1,500 learning cycles, computed over 12 simulations. Error bars depict ±1 SEM. (*C*, *Left*) Task error over time for a single simulation of each network. (*D*) The same as *C* but each synapse is subject to internal independent noise fluctuations in addition to noise in the learning rule (red component in *A*).

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 (Fig. 3*A*) as

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

where expectation is taken across a uniform distribution of time points in $[0, T]$, dots denote time derivatives, and angle brackets denote the (standard) inner product. Eq. **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 Appendix*, *Learning Rate and Local Task Difficulty*):

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

where

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

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 $[0, T]$, normalized by T:

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

Note that we have made no assumptions on the size of T, so the $\mathcal{O}(T^2)$ term in Eq. **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 Eq. **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 Eq. **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 Fig. 3*A*, where the length of the leaps along the error surface indicates 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

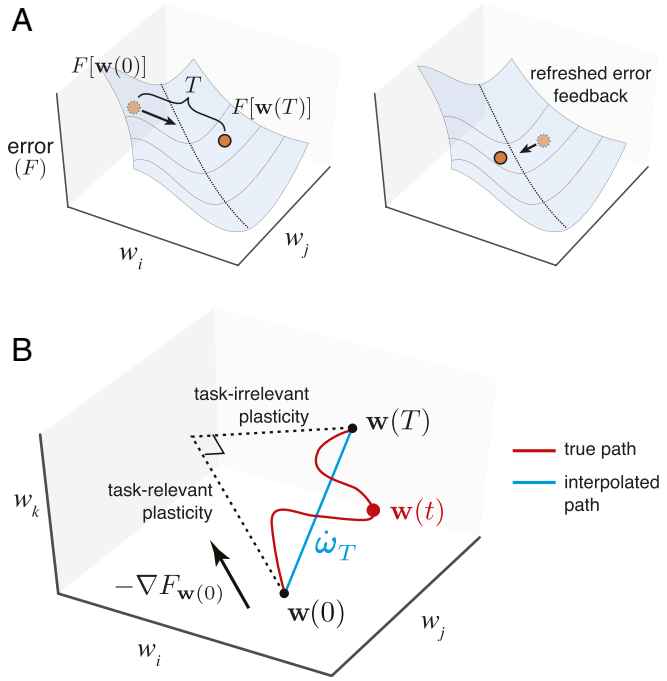

Fig. 3. Geometry of error-based learning in arbitrary networks. (*A*) Schematic of task error as a function of (two) synaptic weights. Learning rule receives and processes task-relevant feedback to provide direction for each synapse to move in weight space. Direction must correlate with steepest-descent 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 summarized 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, which corresponds to the direction $\hat{\mathbf{n}}_2 + \hat{\mathbf{n}}_3$.

$\dot{\boldsymbol{\omega}}_T$ into two components that are parallel and perpendicular to the gradient at time 0, when error information was supplied to the network (Fig. 3*B*),

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

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 this paper can be generalized to eliminate this assumption (*SI Appendix*, *Task-Relevant Plasticity*) but this complicates the presentation without adding insight.

There are several sources of task-irrelevant plasticity. First, there can be inherent imperfections in the learning rule: Information on task error may be imperfectly acquired and transmitted through the nervous system. Second, 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 Fig. 3*A*, where we see that during learning, the local information used to modify synapses leads to a network overshooting local minima in the error surface. Third, in a general biological setting, synapses will be involved in multiple task-irrelevant plasticity processes that contribute to γ_2 (Fig. 1*A*). 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 signaling 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 another expression for synaptic weight change:

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

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:

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

Note that the per-synapse fluctuation due to the γ_3 term is independent of network size. This is because $\|\hat{\mathbf{n}}_3\| = 1$, which implies that the expected magnitude of the ith component of $\hat{\mathbf{n}}_3$ is $\sqrt{\frac{1}{N}}$. If we assume that each component is independent (see *SI Appendix*, *Decomposition of Local Task Difficulty* for justification), we can also write the magnitude of total synaptic rate of change across the network in a convenient form:

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

We see that the γ_3 term, which is a measure of per-synapse fluctuation magnitude, scales with the number of neurons. This stands in contrast to γ_1 and γ_2, which measure fluctuation sources over the network. Eqs. **5** and **6** allow us to rewrite Eq. **2**:

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

For given values of the γ_i and T, we see from Eq. **7** that $\mathbf{G}_F$ controls the learning rate of a network: A higher value of $\mathbf{G}_F[\dot{\boldsymbol{\omega}}_T]$ 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 Eq. **7** to be negative and ceases to occur when this term is zero. This implies

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

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 Eq. **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 Eq. **7** in simulations (Fig. 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 Eq. **8** indeed predicts the steady-state value of $\mathbf{G}_F$. This agreement is confirmed in Fig. 4*A*.

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

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.

Note that in Fig. 4 (as well as in subsequent numerical simulations) we define the entire distribution of input–output pairs to be a finite set or "batch" generated from a fixed, random set of inputs. This a technical necessity that allowed us to numerically calculate true task error (i.e., the error over all inputs) and the true task gradient $\nabla F[\mathbf{w}(0)]$ and thus define specific values of γ_i when applying the update of Eq. **5**. We emphasize that this finite batch is considered to be the entire distribution, not a sample from a true (unknown) distribution. It is not possible to numerically specify the γ_i in the simulations of Fig. 2, because we cannot explicitly calculate the true gradient in this case. Instead, learning occurred online by a single sample from an infinite number of potential inputs, producing a stochastic gradient estimate. Fortunately, conclusions drawn from the both simulation paradigms are interchangeable. Online learning using a stochastic gradient is mathematically equivalent to adding task-irrelevant (γ_2) plasticity to the batch learning setup of Figs. 4 and 6. The amount by which γ_2 is increased depends on the task, but not the network size/architecture (see *SI Appendix*, *Online Learning and Generalization Error* for further details).

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 $\mathbb{E}[\mathbf{n}_2^T\mathbf{n}_3]=0$, which allows us to write an expression for expected local task difficulty:

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

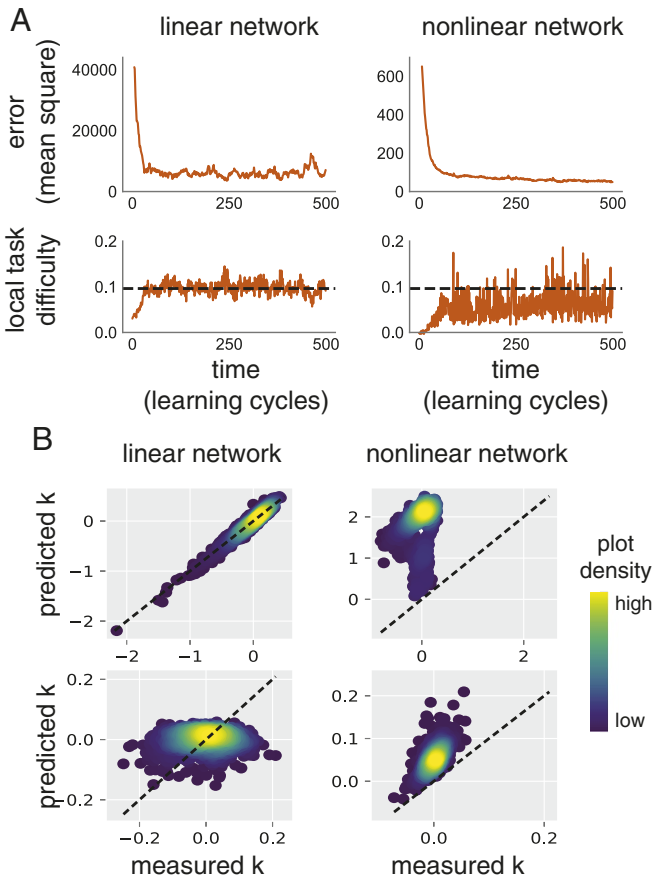

Fig. 4. Numerical validation of learning-rate calculations in simulated neural networks. (*A*) Local task difficulty and mean squared error over time for a linear network (*Left*) with quadratic error function and a nonlinear network (*Right*). Local task difficulty is low when the networks are in an untrained state. As performance improves, it rises, until reaching some steady-state level (black dashed lines). We can predict this steady state a priori, exactly for the quadratic error, and conservatively for the nonlinear error, using Eq. **8**. Both networks are trained using a corrupted learning rule ($\bar{\gamma}=[0.2, 1, 0]$, $T=2$; *Materials and 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 the predicted value of the learning rate k_{pred} (using Eq. **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. Dashed 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 $\bar{\gamma}$ (*Materials and 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 dashed line. In the nonlinear case (*Right*), there is an unknown, nonzero $\mathcal{O}(T^2)$ term (Eq. **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., *A*, *Bottom Right*, dashed line) is consistently an overestimate.

where

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

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

We can gain intuition into how Eq. **9** is derived without going through additional technical details (*SI Appendix, 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

$$F[\mathbf{w}(T)] = F[\mathbf{w}(0)+\mathbf{n}] = F[\mathbf{w}(0)] + \langle \nabla F[\mathbf{w}(0)], \mathbf{n}\rangle + \frac{1}{2}\langle \mathbf{n}, \nabla^2 F[\mathbf{w}(0)]\mathbf{n}\rangle + \mathcal{O}(\|\mathbf{n}\|_2^2). \quad [11]$$

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 Eq. **11** that determines the effect of the perturbation on task error. Its contribution is

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

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. Eq. **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 Fig. 2*C* 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 analyze a linear network and then apply insights from this to a more general nonlinear feedforward case.

Consider a linear network (i.e., a linear map, as shown in Fig. 5*A*) 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 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 semiorthogonal 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

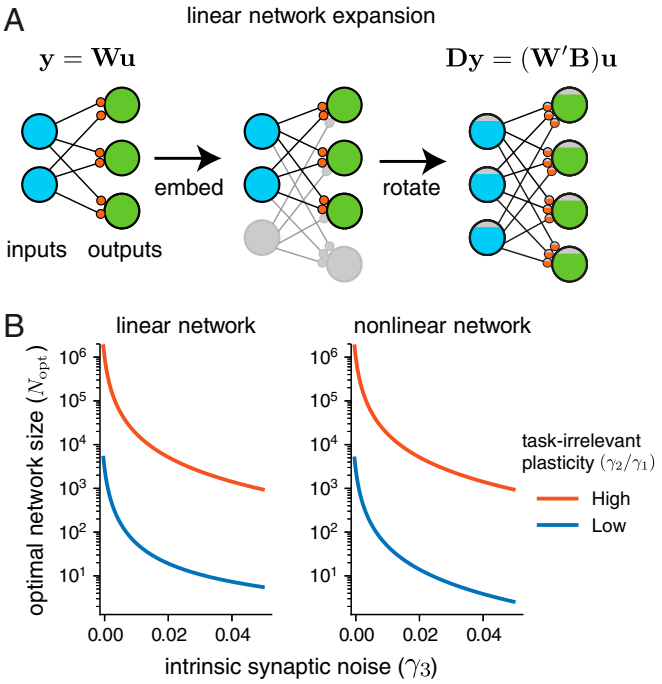

Fig. 5. Optimal network size for linear and nonlinear networks in the presence of intrinsic synaptic noise. (*A*) Network expansion for a linear network, given by an embedding into a larger network, followed by a rotation of the weight matrix. This corresponds to transforming inputs **u** by a projection **B** and outputs **y** by a semiorthogonal mapping **D**. (*B*) Plots show the dependence of N_{opt} in linear and nonlinear networks using Eqs. **17** and **19**. In both cases the learning rule has $\gamma_1 = 0.01$ and $T = 2$. Low task-irrelevant plasticity corresponds to $\gamma_2 = 0.05$, while high task-irrelevant plasticity corresponds to $\gamma_2 = 1$.

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 Fig. 5*A*.

The expanded neural network with weights $W' \in \mathbb{R}^{c_2 o \times c_1 i}$ 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

$$F'[W', u'] = \|y'^*(u') - W'u'\|_2^2 = \|Dy^*(u) - W'Bu\|_2^2 = \|y^*(u) - D^T W'Bu\|_2^2. \quad [12]$$

For some weight configuration W' in an expanded network, Eq. **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

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

Explicit differentiation of $F'[W']$ (see *SI Appendix, Learning in a Linear Network*) yields:

$$\|\nabla F'[W']\|_2 = \|\nabla F[W]\|_2$$
$$Tr(\nabla^2 F'[W']) = 2c_2 o \int_{u\in\mathcal{U}} \|u\|_2^2\ \mathbb{P}(u) = c_2\, Tr(\nabla^2 F[W]). \quad [13]$$

We now rewrite the weight matrices W' and W as vectors $\mathbf{w}' \in \mathbb{R}^{\tilde{N}}$ and $\mathbf{w} \in \mathbb{R}^N$.

We will assume that $\nabla\hat{F}'[\mathbf{w}']$ (which is a normalized 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, Eq. **13** implies

$$\nabla\hat{F}'[\mathbf{w}']^T\nabla^2 F'[\mathbf{w}']\nabla\hat{F}'[\mathbf{w}'] \approx c_2\nabla\hat{F}[\mathbf{w}]^T\nabla^2 F[\mathbf{w}]\nabla\hat{F}[\mathbf{w}]. \quad [14]$$

Bringing together Eqs. **13** and **14** and the formula Eq. **10** for $\mathbf{G}^1_F$, we see that

$$\mathbf{G}'^1_{F'}[\mathbf{w}'] \approx c_2\mathbf{G}^1_F[\mathbf{w}], \quad [15a]$$

and similarly

$$\frac{Tr(\nabla^2 F'[\mathbf{w}'])}{\|\nabla F'[\mathbf{w}']\|_2} \approx c_2\frac{Tr(\nabla^2 F[\mathbf{w}])}{\|\nabla F[\mathbf{w}]\|_2}. \quad [15b]$$

We can therefore write local task difficulty of the expanded network in terms of quantities of the smaller network:

$$\mathbb{E}\left[\mathbf{G}'_{F'}[\dot{\boldsymbol{\omega}}'_T]\right]\|\boldsymbol{\omega}'_T\|_2^2 \approx c_2\mathbf{G}_F[\mathbf{w}] + c_2\frac{Tr(\nabla^2 F[\mathbf{w}])}{2\|\nabla F[\mathbf{w}]\|_2}\left[\frac{\gamma_2^2}{\tilde{N}} + \frac{\gamma_3^2}{T}\right]. \quad [16]$$

As long as the ratio $\frac{c_2}{c_1}$ is fixed, we can rewrite $\tilde{N}$ in terms of c_2 and alter c_2 as an independent parameter in Eq. **16**. Indeed this allows us to optimize the steady-state error of the network by changing N. To see how, recall that

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

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 minimize the expression for $\boldsymbol{\delta}$ in N (or equivalently, c_2). We differentiate $\boldsymbol{\delta}$ in c_2 and note that a single stationary point exists, satisfying the equation

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

Note that C is unknown in general because it depends on the weight configuration of the network. However, we can take the heuristic $C \approx \frac{1}{N^*}$, 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 eigenvectors of the latter, and thus the numerator 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 N^*, the stationary point of $\boldsymbol{\delta}$ in N:

$$N^* \approx \frac{T\gamma_2^2}{\gamma_3^2}\left(1 - \frac{\gamma_1^2}{\gamma_2^2}\right). \quad [17]$$

Since $\lim_{\tilde{N}\to\infty}\boldsymbol{\delta} = \infty$, we have that $\boldsymbol{\delta}(\tilde{N})$ is monotonically increasing in $\tilde{N}$, for $\tilde{N} \geq N^*$. So if $N^* < N$, the size of the original network, then any extra redundancy hurts learning performance. If $N^* > N$, then the optimal network size N_{opt} satisfies $N_{\text{opt}} = N^*$. Our formula is verified numerically in Fig. 6*A*, by evaluating the learning performance of transformed neural networks of different sizes, with different γ_i values.

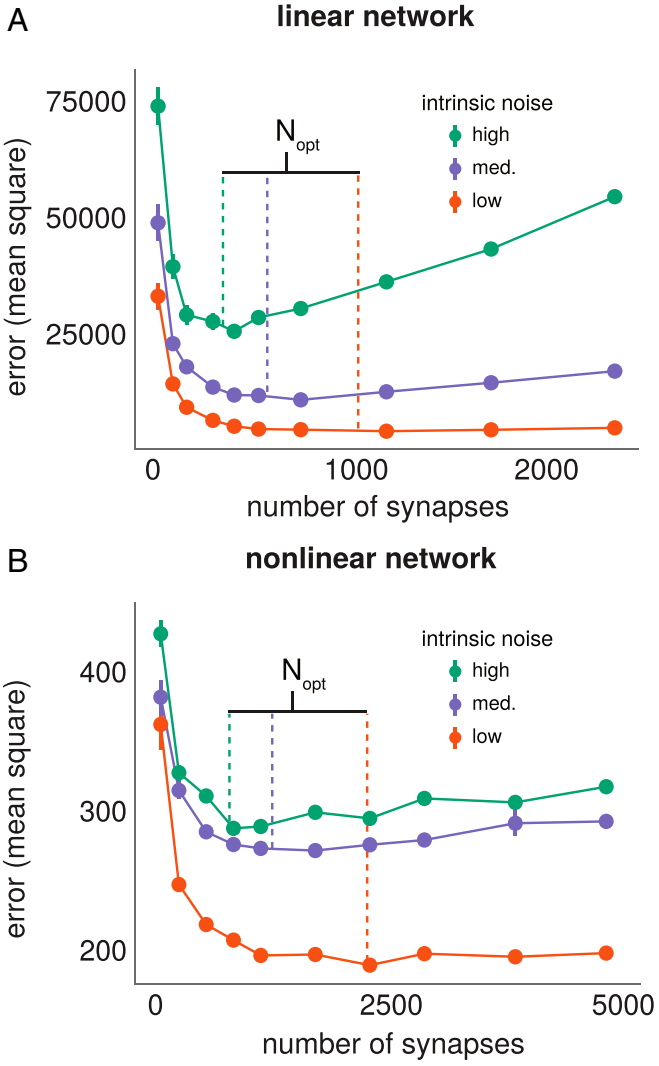

Fig. 6. Testing analytic prediction of optimal network size for linear and nonlinear networks. (*A* and *B*) Linear (*A*) and nonlinear (*B*) networks of different sizes are trained for 1,500 learning cycles of length $T = 1$. Mean steady-state error over 12 repeats is plotted against network size. Error bars denote ± 1 SEM. Colored lines represent a priori predicted optimal network sizes using Eqs. **17** and **19** 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 (detailed in *SI Appendix, Learning in a Linear Network*). The learning rule uses $\bar{\gamma} = [0.07, 1, 0.03]$ (low intrinsic noise), $\bar{\gamma} = [0.06, 1, 0.04]$ (medium intrinsic noise), and $\bar{\gamma} = [0.05, 1, 0.05]$ (high intrinsic noise). (*B*) Nonlinear networks have sigmoidal nonlinearities at each neuron and a single hidden layer (*Materials and Methods*). All networks have 10 input and 10 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.

This estimate of the optimal network size is plotted in Fig. 5*B*, 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 Fig. 6*A* in simulations.

We next consider nonlinear multilayer, feedforward networks. Again, we use the student–teacher framework to generate learning tasks. We consider learning performance of a nominal and an expanded network, both with l layers, and both using the same learning rule. The only difference between the two networks is the larger number of neurons in each hidden layer of the expanded network. We will 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 Fig. 6*B*.

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 kth layer. The concatenation of the synaptic weight matrices across all layers is denoted W and has N elements. We interchangeably denote it as a vector $\mathbf{w} \in \mathbb{R}^N$. The function σ passes its arguments elementwise through some nonlinearity (sigmoidal in simulations; *Materials and Methods*). 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 behavior 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 Appendix, Learning in a Nonlinear, Feedforward Network* for additional details). We get

$$\mathbb{E}[\mathbf{G}'_{F'}[\phi(\dot{\hat{\boldsymbol{\omega}}}_T)]] \|\phi(\dot{\boldsymbol{\omega}}_T)\|_2^2 \approx \sqrt{\frac{N}{\tilde{N}}}\gamma_1^2 \mathbf{G}^1_F[\mathbf{w}(0)] + \sqrt{\frac{\tilde{N}}{N}} \frac{Tr(\nabla^2 F[\mathbf{w}(0)])}{2\|\nabla F[\mathbf{w}(0)]\|_2} \left[\frac{\gamma_2^2}{\tilde{N}} + \frac{\gamma_3^2}{T}\right]. \quad [18]$$

We can use Eq. **18** to minimize

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

in $\tilde{N}$, the number of synapses. There is a single global minimum of $\boldsymbol{\delta}$ in $\tilde{N}$, for $\tilde{N} > 0$. It satisfies

$$N^* \approx \frac{T\gamma_2^2}{\gamma_3^2}\left[\frac{\gamma_1^2}{\gamma_2^2}\frac{N}{N^*} + 1\right]. \quad [19]$$

This gives an optimal size of the transformed network for minimizing steady-state task performance (validated in Fig. 6*B*). Note the dependence of N^* on N, 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 Eq. **19** 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 (33–35). Previous studies focused on the energetic cost of sustaining large numbers of neurons and connecting them efficiently (2, 34–36). 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 volatility of synaptic connections.

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 behaviors (37, 38). It can improve the quality of locally optimal behaviors arrived at after learning (39, 40). 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 analyzing 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 intermittent task error information. Biological learning rules are unlikely to explicitly compute gradients. Regardless, any learning rule that uses error information must effectively approximate a gradient as the network learns.

As an analogy, imagine hiking to the base of a mountain without a map and doing so using intermittent and imperfect estimates of the slope underfoot. An even slope will be easier to descend because slope estimates will remain consistent and random errors in the estimates will average out over time. 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 illustrates why larger neural circuits are able to achieve better learning performance in a given task when 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 hypotheses for the organizing principles in biological circuits which, among other things, predict 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 (26, 27, 41, 42).

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 experience noisy fluctuations due to their molecular makeup (23–25, 43–45). 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 results demonstrate the intuitive fact that insufficient connections impair learning performance. Conversely, and less obviously, excessive numbers of connections impair learning once the optimal network size is exceeded. This provides a hypothesis for why abnormalities in circuit connectivity may lead to learning deficits (18–21).

Our analysis allowed us to predict the optimal size of a network in theoretical learning tasks where we can specify the levels of noise in the learning rule and in synapses. Fig. 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. In particular, optimal network size increases explosively as intrinsic synaptic noise approaches zero. 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.

Materials and Methods

Full details of all simulations are provided in *SI Appendix*. We provide an overview here. Code related to paper is publicly accessible in ref. 46.

Network Architectures. All tested neural networks have fully connected feedforward architectures. Unless otherwise specified networks are nonlinear with multiple hidden layers. Each neuron in these networks passes inputs through a sigmoidal nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ of the form $\sigma(x) = \frac{1}{1+\exp(-x)}$.

Details of the Learning Tasks. All training tasks used in simulation use the student–teacher framework. The basic setup is as follows: We first make a teacher network, which has the same basic network architecture as the student's. We then initialize the teacher weights at fixed values, which are generated as follows (unless otherwise specified): At the kth layer of the teacher network, weights are distributed uniformly on the interval $[-a_k, a_k]$. Here a_k is set so that the SD of weights on the layer is $\frac{4}{\sqrt{i}}$, where i is the number of inputs to the layer. This scaling with i ensures that the magnitude of hidden layer outputs does not increase with the size of hidden layer (47).

We then specify a set $\mathcal{U}$ consisting of 1,000 input vectors. We generate each vector $u \in \mathcal{U}$ componentwise, randomly drawing each component u_i from the distribution $u_i \sim \mathcal{N}(0, 1)$. Note that in Fig. 2 this acts as a test set, which is not used when training the network (details in the next section).

The output vector of the teacher, given an input vector u, is denoted $y^*(u)$. The task of the student network, with weights $\mathbf{w}$, is then to match its output to $y^*(u)$, for all $u \in \mathcal{U}$. Therefore, the task error is

$$F[\mathbf{w}] = \sum_{u \in \mathcal{U}} \|y^*(u) - y(\mathbf{w}, u)\|_2^2, \quad [20]$$

where $y(\mathbf{w}, u)$ denotes the output of the student given input u and weights $\mathbf{w}$.

For the linear networks studied at the beginning of the section Network Expansions That Increase Learning Performance, the dimensionalities of the input and output vectors differ between the teacher and the students. However, fixed matrix transformations lift the teacher inputs/outputs into the appropriate dimensionality (Eq. **12**). For all other networks, the number of network inputs/outputs is shared between the students and the teacher. Teacher and students also have the same number of hidden layers. At the ith hidden layer, each student has at least as many neurons as the teacher. This ensures that the teacher network forms a subset of each student network and therefore that each student network is theoretically capable of exactly recreating the input–output mapping $y^*(u)$ of the teacher.

Network Training. The theoretical analysis in this paper and the simulations in Fig. 2 pertain to online learning, where training data are sampled continuously from an (infinite) distribution defined by the input–output mapping of the teacher networks. However, in cases where we needed to compute a true gradient (specifically Figs. 4 and 6), we needed to define finite distributions to numerically evaluate the true gradient. Having the true gradient allows us to precisely specify values of task-relevant and task-irrelevant components of plasticity.

We emphasize (as we emphasized in the main text) that this differs from treating the finite set as a sample—or batch—from an infinite distribution, which would incur generalization issues because any finite sample will be necessarily biased. The relationship between the results of our paper and the latter scenario is described in *SI Appendix, Regularization and Generalization Error*.

In Figs. 4 and 6 learning is conducted on the finite input set $\mathcal{U}$ described in the previous section. At each learning cycle, we apply the weight update

$$\mathbf{w}_{t+T} = \mathbf{w}_t - T\bar{\gamma}_1 \nabla \hat{F}[\mathbf{w}_t] + T\bar{\gamma}_2 \hat{\mathbf{n}}_2^t + \sqrt{NT}\bar{\gamma}_3 \hat{\mathbf{n}}_3^t. \quad [21]$$

The parameters $\{\bar{\gamma}_i\}_{i=1}^3$ and T specify the quality of the learning rule, the feedback delay, and the intrinsic, per-synapse noise (main text). N is the number of synapses in the network. We calculate the gradient $\nabla F[\mathbf{w}_t]$ by taking the gradient of Eq. **21** using backpropagation. The normalized vectors $\hat{\mathbf{n}}_2^t$ and $\hat{\mathbf{n}}_3^t$ represent sources of task-irrelevant plasticity. The dynamics of the unnormalized vector $\mathbf{n}_2$ satisfy $\mathbf{n}_2^{t+1} = \sqrt{0.1}\mathbf{n}_2^t + \sqrt{0.9}\hat{\nu}$, where ν is a Gaussian random variable, normalized such that $\|\nu\|_2 = 1$. $\hat{\mathbf{n}}_3$, which models intrinsic synaptic noise, is a Gaussian random variable, normalized so that $\|\hat{\mathbf{n}}_3\|_2 = 1$.

The network in Fig. 2 *C* and *D* is conducted online from an infinite distribution. At each learning cycle, we randomly draw a single input vector u of Gaussian components; i.e., $u_i \sim \mathcal{N}(0, 1)$. We replace the term $\nabla \hat{F}[\mathbf{w}_t]$ in the weight update Eq. **21** with $\nabla \hat{F}[\mathbf{w}, u]$, the (normalized) stochastic gradient, where $F[\mathbf{w}, u] := \|y^*(u) - y(\mathbf{w}, u)\|_2^2$. The overall error $F[\mathbf{w}]$ is then the expected error on the next input; i.e.,

$$F[\mathbf{w}] = \int_\Omega F[\mathbf{w}, u]\mathbb{P}[u]\, du,$$

where $\mathbb{P}[u]$ is the componentwise Gaussian probability density function from which u is drawn, with support Ω. We cannot exactly calculate $F[\mathbf{w}]$ in this setting; we therefore use the error function described above, which is constructed from 1,000 inputs, providing an estimate of $F[\mathbf{w}]$.

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

1. Laughlin SB, de Ruyter van Steveninck RR, Anderson JC (1998) The metabolic cost of neural information. *Nat Neurosci* 1:36–41.
2. Tomasi D, Wang G-J, Volkow ND (2013) Energetic cost of brain functional connectivity. *Proc Natl Acad Sci USA* 110:13642–13647.
3. Attwell D, Laughlin SB (2001) An energy budget for signaling in the grey matter of the brain. *J Cereb Blood Flow Metab* 21:1133–1145.
4. Reader SM, Laland KN (2002) Social intelligence, innovation, and enhanced brain size in primates. *Proc Natl Acad Sci USA* 99:4436–4441.
5. Sol D, Duncan RP, Blackburn TM, Cassey P, Lefebvre L (2005) Big brains, enhanced cognition, and response of birds to novel environments. *Proc Natl Acad Sci USA* 102:5460–5465.
6. Joffe TH, Dunbar RIM (1997) Visual and socio–cognitive information processing in primate brain evolution. *Proc R Soc Lond B Biol Sci* 264:1303–1307.
7. Maguire EA, et al. (2000) Navigation-related structural change in the hippocampi of taxi drivers. *Proc Natl Acad Sci USA* 97:4398–4403.
8. Gaser C, Schlaug G (2003) Brain structures differ between musicians and non-musicians. *J Neurosci* 23:9240–9245.
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. *Proc Natl Acad Sci USA* 87:5568–5572.
10. Lawrence S, Giles CL, Tsoi AC (1998) What size neural network gives optimal generalization? Convergence properties of backpropagation (University of Maryland,

Institute for Advanced Computer Studies, College Park, MD), Technical Report UMIACS-TR-96-22 and CS-TR-3617.
11. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, eds Pereira F, Burges CJC, Bottou L, Weinberger KQ (Curran Associates, Inc., Red Hook, NY), pp 1097–1105.
12. Huang G-B (2003) Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Trans Neural Networks* 14:274–281.
13. Takiyama K (2016) Maximization of learning speed due to neuronal redundancy in reinforcement learning. *J Phys Soc Jpn* 85:114801.
14. Takiyama K, Okada M (2012) Maximization of learning speed in the motor cortex due to neuronal redundancy. *PLoS Comput Biol* 8:e1002348.
15. Saxe AM, McClelland JL, Ganguli S (2013) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120. Preprint, posted December 20, 2013.
16. Seung HS (2003) Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron* 40:1063–1073.
17. Werfel J, Xie X, Seung HS (2004) Learning curves for stochastic gradient descent in linear feedforward networks. *Advances in Neural Information Processing Systems*, eds Saul L, Weiss Y, Bottou L (MIT Press, Boston), pp 1197–1204.
18. Contractor A, Klyachko VA, Portera-Cailliau C (2015) Altered neuronal and circuit excitability in fragile X syndrome. *Neuron* 87:699–715.
19. Rinaldi T, Perrodin C, Markram H (2008) Hyper-connectivity and hyper-plasticity in the medial prefrontal cortex in the valproic acid animal model of autism. *Front Neural Circuits* 2:4.
20. Casanova MF, et al. (2006) Minicolumnar abnormalities in autism. *Acta Neuropathol* 112:287–303.
21. Amaral DG, Mills Schumann C, Wu Nordahl C (2008) Neuroanatomy of autism. *Trends Neurosci* 31:137–145.
22. Loewenstein Y, Yanover U, Rumpel S (2015) Predicting the dynamics of network connectivity in the neocortex. *J Neurosci* 35:12535–12544.
23. Ziv NE, Brenner N (2017) Synaptic tenacity or lack thereof: Spontaneous remodeling of synapses. *Trends Neurosci* 41:89–99.
24. Minerbi A, et al. (2009) Long-term relationships between synaptic tenacity, synaptic remodeling, and network activity. *PLoS Biol* 7:e1000136.
25. Puro DG, De Mello FG, Nirenberg M (1977) Synapse turnover: The formation and termination of transient synapses. *Proc Natl Acad Sci USA* 74:4977–4981.
26. Bloss EB, et al. (2018) Single excitatory axons form clustered synapses onto CA1 pyramidal cell dendrites. *Nat Neurosci* 21:353–363.
27. Bartol TM, Jr, et al. (2015) Nanoconnectomic upper bound on the variability of synaptic plasticity. *Elife* 4:e10778.
28. Levin E, Tishby N, Solla SA (1990) A statistical approach to learning and generalization in layered neural networks. *Proc IEEE* 78:1568–1574.
29. Seung HS, Sompolinsky H, Tishby N (1992) Statistical mechanics of learning from examples. *Phys Rev A* 45:6056–6091.
30. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. *J Machine Learn Res* 15:1929–1958.
31. José Hanson S (1990) A stochastic version of the delta rule. *Phys D Nonlinear Phenom* 42:265–272.
32. Frazier-Logue N, José Hanson S (2018) Dropout is a special case of the stochastic delta rule: Faster and more accurate deep learning. arXiv:1808.03578. Preprint, posted August 10, 2018.
33. Chittka L, Niven J (2009) Are bigger brains better? *Curr Biol* 19:R995–R1008.
34. Herculano-Houzel S (2012) The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc Natl Acad Sci USA* 109:10661–10668.
35. Shepherd GMG, Stepanyants A, Bureau I, Chklovskii D, Svoboda K (2005) Geometric and functional organization of cortical circuits. *Nat Neurosci* 8: 782–790.
36. Shulman RG, Rothman DL, Behar KL, Hyder F (2004) Energetic basis of brain activity: Implications for neuroimaging. *Trends Neurosci* 27:489–495.
37. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science* 313:504–507.
38. Tishby N, Zaslavsky N (2015) Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, ed Kschischang FR (Curran Associates, Inc., Red Hook, NY), pp 1–5.
39. Bray AJ, Dean DS (2007) Statistics of critical points of Gaussian fields on large-dimensional spaces. *Phys Rev Lett* 98:150201.
40. Dauphin YN, et al. (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems*, eds Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (Curran Associates, Inc., Red Hook, NY), pp 2933–2941.
41. Druckmann S, et al. (2014) Structured synaptic connectivity between hippocampal regions. *Neuron* 81:629–640.
42. Eichler K, et al. (2017) The complete connectome of a learning and memory centre in an insect brain. *Nature* 548:175–182.
43. Mongillo G, Rumpel S, Loewenstein Y (2017) Intrinsic volatility of synaptic connections—A challenge to the synaptic trace theory of memory. *Curr Opin Neurobiol* 46:7–13.
44. Attardo A, Fitzgerald JE, Schnitzer MJ (2015) Impermanence of dendritic spines in live adult CA1 hippocampus. *Nature* 523:592.
45. Faisal AA, Selen LPJ, Wolpert DM (2008) Noise in the nervous system. *Nat Rev Neurosci* 9:292–303.
46. Raman DV, Perez-Rotondo A, O'Leary TS (2018) Code for figure simulations. Available at https://github.com/olearylab/raman_etal_2018. Deposited December 7, 2018.
47. Bengio Y (2012) Practical recommendations for gradient-based training of deep architectures. *Neural Networks: Tricks of the Trade*, eds Montavon G, Orr GB, Muller KR (Springer, Berlin), pp 437–478.