ON THE STABILITY OF LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning algorithms are usually analyzed under the assumption that informative gradients are continuously available. But this is not always true—learning signals in biological neural networks, for example, may be sparse, local, and only intermittently available at any given synapse. In this work, we show that if gradient information is noisy and only intermittently available, learning may be unstable. In particular, even if average learning dynamics converge to the correct optimum, variance in learning may be asymptotically infinite. We explore this point in a number of analytically tractable settings, including single-layer and multi-layer linear networks, and derive explicit conditions under which learning is stable. These conditions complement, but are conceptually distinct from, existing conditions related to the stability of stochastic gradient descent. Our work demonstrates that unbiased credit assignment is not enough for reliable learning; sufficiently frequent and informative feedback is also required, which yields an additional constraint on models of biologically plausible learning.

1 INTRODUCTION

The claim that the brain does not perform backpropagation (Rumelhart et al., 1986) rests in part on the idea that gradient information may be difficult to communicate. In particular, the gradient of an objective function with respect to any given synaptic weight depends on the synaptic weights of a large number of downstream neurons, and it is not obvious in general how such non-local information could be reliably communicated (Crick, 1989; Lillicrap et al., 2020). This issue has motivated a substantial body of work seeking alternative and more biologically plausible learning rules (Lillicrap et al., 2016; Scellier & Bengio, 2017), as well as schemes for approximately implementing backpropagation in specific contexts (Sacramento et al., 2018; Whittington & Bogacz, 2019; Millidge et al., 2022). It remains unknown if and how this information is communicated, but it is generally agreed that communicating it accurately would be difficult given locality constraints.

In this paper, we consider a related but conceptually distinct issue. If gradient information *can* be communicated, it is not at all obvious that it can be communicated *reliably*. It is plausible that a scheme which delivers the correct updates *on average* only delivers them occasionally, to different neurons at different times, and perhaps somewhat noise-corrupted. How do learning dynamics change when gradients are noisy and only intermittently available? We argue that learning changes substantially, and in fact can become unstable if updates are not delivered sufficiently frequently.

This idea is interesting for a few reasons. First, it implies that credit assignment which is correct *on average* is insufficient, since there remains the additional obstacle that weight updates cannot be too variable. Second, instability is not a generic consequence of noisy weight dynamics, but rather depends crucially on the multiplicative, sparse nature of the noise introduced by assuming intermittent gradient arrivals.

We show that the severity of this issue can be reduced by a few obvious noise mitigation strategies, but that the presence of the issue (i.e., learning becoming unstable if feedback is sufficiently sparse) survives. This problem exists in both linear and nonlinear networks; in shallow and deep networks; and after averaging parameters or recent gradients. The details of the issue change, but the core finding remains: an informative signal must arrive sufficiently often for learning to remain stable.

We proceed as follows. We begin by motivating the simplest version of our model. Next, by analyzing it in detail, we show that there exist gradient-variance-dependent stability issues that are exacerbated as informative feedback becomes less frequent. We then discuss simple examples of

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

this issue near the fixed points of more complex networks, and the extent to which basic noise mitigation strategies address it.

2 SPARSE GRADIENT ARRIVALS CAN AMPLIFY ERRORS

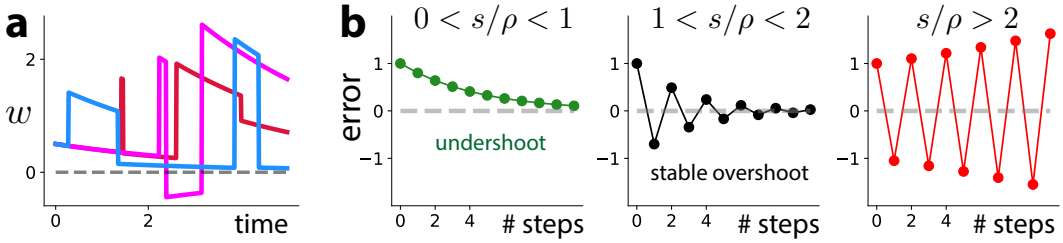


Figure 1: **Sparse gradient arrivals: motivation and intuition.** (**a**) A simple example of sparse gradient arrivals: a weight receives intermittent updates, in between which the regularization term produces relaxation toward zero. Three example stochastic simulations are shown. (**b**) Error as a function of number of arrivals in a toy model. If $s/\rho < 1$, error always decreases (left); if $1 < s/\rho < 2$, error alternates in sign but decreases in magnitude (center); and if $s/\rho > 2$, learning becomes unstable, and error both alternates in sign and increases in magnitude (right).

Suppose we are interested in updating a set of weights $\boldsymbol{w} \in \mathbb{R}^N$ to optimize an objective

$$\tilde{J}(\boldsymbol{w}) := J(\boldsymbol{w}) + \frac{\gamma}{2}\|\boldsymbol{w}\|_2^2 \tag{1}$$

where J is a task-related objective and the second term regularizes $\boldsymbol{w}$. We explicitly separate J and the regularization term because their biological interpretation in what follows is different. In the gradient-flow limit typical of many theoretical treatments of learning (Saxe et al., 2013a; Bordelon et al., 2020; Domine et al., 2025a), gradient descent involves updates of the form

$$\frac{d}{dt}\boldsymbol{w} = \eta\left[-\frac{\partial J}{\partial \boldsymbol{w}} - \gamma \boldsymbol{w}\right] \tag{2}$$

where $\eta > 0$ is the learning rate and $t \in [0, \infty)$ is the amount of elapsed training time. We will investigate the consequences of assuming w_i only has *intermittent* access to the correct gradient. We do this as follows. We model sparse gradient arrivals via a Poisson process (Kingman, 1992): we assume that each synapse has a fixed probability per unit time of receiving a noise-corrupted version of the relevant gradient information. In doing things this way, we ignore potential temporal structure in gradient arrivals and correlations between the information arriving at different synapses; accounting for these factors will likely complexify the resulting model, and more importantly, make the associated stability issues worse rather than better.

Poisson gradient arrivals assumption. To model temporally sparse updates, we assume gradients are coupled to a *Poisson process* with mean arrival rate $\eta\rho$, where $\rho > 0$ is the non-dimensionalized *arrival rate*. As ρ increases, neurons tend to receive more gradient arrivals per learning time scale $1/\eta$. If we denote a realization of this Poisson process by $\boldsymbol{r}(t) \in \mathbb{R}^N$, our model corresponds to the replacement

$$-\frac{\partial J}{\partial \boldsymbol{w}} \to \boldsymbol{r}(t) \odot \frac{1}{\rho}\left[-\frac{\partial J}{\partial \boldsymbol{w}} + \boldsymbol{\xi}\right] \tag{3}$$

where $\odot$ denotes the Hadamard (or *elementwise*) product. The quantity $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\text{obs}}^2 \boldsymbol{I})$ represents *observation noise*. Our main motivation for including it is to illustrate how, unlike sparse gradient arrivals, this kind of noise is comparatively benign.

According to this prescription, weight changes due to gradient arrivals have mean and covariance

$$\mathbb{E}[\Delta \boldsymbol{w}] = -\eta \Delta t \frac{\partial J}{\partial \boldsymbol{w}} \qquad \text{Cov}(\Delta w_i, \Delta w_j) = \frac{\eta \Delta t}{\rho}\,\delta_{ij}\left[\left(\frac{\partial J}{\partial \boldsymbol{w}}\right)^2 + \sigma_{\text{obs}}^2\right]. \tag{4}$$

Useful intuition. We can make a simple heuristic argument to explain why sparse gradient arrivals can destabilize learning. Consider gradient descent (Eq. 2) with only one weight and no regularization ($\gamma = 0$) on a mean-squared error objective

$$J := \frac{s}{2}(w - w^*)^2 \,. \tag{5}$$

For Poisson gradient arrivals, weights do not change until there is an arrival, so learning effectively occurs in discrete time. The error $e_n := w_n - w_n^*$ after n gradient arrivals evolves according to the recurrence relation

$$e_{n+1} = \left(1 - \frac{s}{\rho}\right) e_n \,. \tag{6}$$

These dynamics yield a geometric series with $e_n = (1 - s/\rho)^n e_0$ which has three regimes. For a sufficiently large arrival rate ($s/\rho < 1$), error consistently decreases and asymptotically approaches zero (Fig. 1b, left). For a moderately large arrival rate ($1 < s/\rho < 2$), weight dynamics overshoots its target and the error alternates between positive and negative; it still asymptotically approaches zero (Fig. 1b, center). If the arrival rate is too small ($s/\rho > 2$), learning becomes unstable, and error both alternates in sign and increases in magnitude (Fig. 1b, right). This indicates that, at least for this toy model, the ratio s/ρ between input variance and the gradient arrival rate controls whether learning is stable. Hence, the arrival rate ρ must be larger than

$$\boxed{\rho > \rho_c := \frac{s}{2}} \,. \tag{7}$$

We will see that more complex models exhibit surprisingly similar behavior, and in particular lead to similar expressions for the critical arrival rate ρ_c.

3 Instability in a linear model

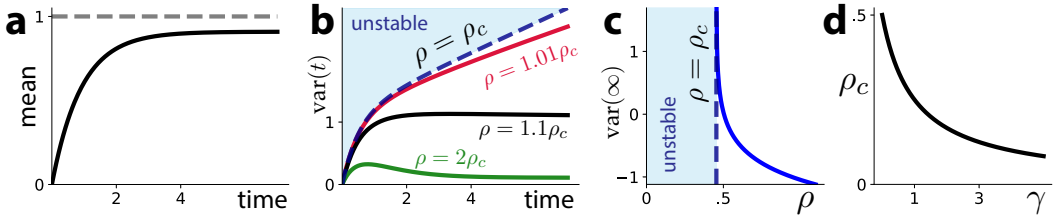


Figure 2: **Instability in a linear model.** (**a**) Average dynamics are unconditionally stable and do not depend on the arrival rate ρ. (**b**) Output variance converges more slowly as ρ decreases and diverges to infinity when $\rho < \rho_c$. (**c**) The steady state output variance diverges like $1/(\rho - \rho_c)$. (**d**) ρ_c decreases as γ increases, so one way to improve stability is to increase the regularization strength.

In this section, we study instability due to sparse gradient arrivals in a tractable linear model, in part because we think it reveals the essence of the phenomenon. Consider an objective

$$\tilde{J} = \mathbb{E}_{\boldsymbol{x}}\left\{\frac{\|\boldsymbol{w}^T\boldsymbol{x} - (\boldsymbol{w}^*)^T\boldsymbol{x}\|_2^2}{2}\right\} + \frac{\gamma}{2}\|\boldsymbol{w}\|_2^2 = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^T\boldsymbol{S}(\boldsymbol{w} - \boldsymbol{w}^*) + \frac{\gamma}{2}\|\boldsymbol{w}\|_2^2 \tag{8}$$

where the input $\boldsymbol{x}$ is assumed to have mean zero and covariance $\boldsymbol{S} \in \mathbb{R}^{N\times N}$. The negative gradient of J equals $\boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w})$, so our prescription yields dynamics

$$\frac{d}{dt}\boldsymbol{w} = \boldsymbol{r} \odot \frac{1}{\rho}[\boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w}) + \boldsymbol{\xi}] - \eta\gamma\boldsymbol{w} \,. \tag{9}$$

The average $\bar{\boldsymbol{w}} := \mathbb{E}[\boldsymbol{w}]$ evolves according to standard-looking linear dynamics

$$\frac{d}{dt}\bar{\boldsymbol{w}} = \eta\,[\boldsymbol{S}(\boldsymbol{w}^* - \bar{\boldsymbol{w}}) - \gamma\bar{\boldsymbol{w}}] \tag{10}$$

which approaches the fixed point at $\boldsymbol{w}' := (\boldsymbol{S} + \gamma \boldsymbol{I})^{-1} \boldsymbol{S} \boldsymbol{w}^*$ exponentially quickly. On the other hand, the dynamics of the covariance matrix $\boldsymbol{C} := \mathbb{E}[(\boldsymbol{w} - \bar{\boldsymbol{w}})(\boldsymbol{w} - \bar{\boldsymbol{w}})^T]$ are substantially more interesting (see Appendix A):

$$\boxed{\frac{1}{\eta}\frac{d}{dt}\boldsymbol{C} = -2\gamma \boldsymbol{C} - \boldsymbol{C}\boldsymbol{S} - \boldsymbol{S}\boldsymbol{C} + \frac{1}{\rho}\sigma_{\text{obs}}^2 \boldsymbol{I} + \frac{1}{\rho}\,\text{diag}(\bar{g}_i^2 + [\boldsymbol{S}\boldsymbol{C}\boldsymbol{S}]_{ii})} \,. \tag{11}$$

We can assess stability by determining the conditions under which $\lim_{t\to\infty} \text{tr}[\boldsymbol{C}(t)]$ is finite. While this is difficult to do analytically in general, it is possible to do if we make at least one additional assumption about $\boldsymbol{S}$: either that (i) $\boldsymbol{S}$ is *diagonal*, or that (ii) $\boldsymbol{S}$ is *circulant*.

3.1 Stability given a diagonal input covariance

If $\boldsymbol{S}$ is diagonal, Eq. 11 simplifies since $[\boldsymbol{S}\boldsymbol{C}\boldsymbol{S}]_{ii} = S_{ii}^2 C_{ii}$. Let us simplify notation by writing $c_i := C_{ii}$ and $s_i := S_{ii}$. The diagonal elements of $\boldsymbol{C}$ then evolve according to

$$\frac{1}{\eta}\frac{d}{dt}c_i = \frac{1}{\rho}(\sigma_{\text{obs}}^2 + \bar{g}_i^2) - \left(2\gamma c_i + 2s_i - \frac{s_i^2}{\rho}\right) c_i \,. \tag{12}$$

These dynamics are linear, and whether c_i converges to a stable value or diverges to infinity depends on the coefficient of c_i—that is, it must be negative. For learning to be stable, the coefficients of all c_i must be negative; hence, it must be true that

$$\boxed{\rho > \rho_c^i := \frac{s_i^2}{2(s_i + \gamma)}} \tag{13}$$

for all i. Equivalently, $\rho > \rho_c := \max_i \rho_c^i$, which depends on the largest input variance s_i. Note that this condition reduces to the critical arrival rate associated with our naive model (Eq. 7) when we take $\gamma \to 0$.

As ρ decreases, c_i converges increasingly slowly, and below ρ_c it becomes unstable (Fig. 2b). The steady state diagonal covariance elements diverge like $1/(\rho - \rho_c^i)$ (Fig. 2c), or more quantitatively,

$$c_i = \frac{1}{\rho - \rho_c^i}\frac{\sigma_{\text{obs}}^2 + \bar{g}_i^2}{2(s_i + \gamma)} \,. \tag{14}$$

Our formula indicates that the critical arrival rate threshold ρ_c can be reduced in two ways: one can either decrease input variance or increase the regularization strength γ (Fig. 2d). The latter method recalls the standard bias-variance tradeoff (Bishop, 2006): we can decrease the (infinite) variance at the cost of increasing bias, since increasing γ makes $\boldsymbol{w}$ more different from the target weight $\boldsymbol{w}^*$.

3.2 Stability given a circulant input covariance

Suppose instead that $\boldsymbol{S}$ is *circulant*. Then $\boldsymbol{S}$ can be diagonalized via a discrete Fourier basis, i.e., $\boldsymbol{S} = \boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^\dagger$ (see Appendix A). The arrival-related term can be written in terms of the covariance matrix in that basis, $\tilde{\boldsymbol{C}} := \boldsymbol{F}^\dagger \boldsymbol{C} \boldsymbol{F}$, since

$$\frac{1}{N}\sum_i [\boldsymbol{S}\boldsymbol{C}\boldsymbol{S}]_{ii} = \frac{1}{N}\sum_q s_q^2 \tilde{C}_{qq} \,. \tag{15}$$

In this case, different $\tilde{C}_{qq}$ do not decouple, but it is still possible to straightforwardly characterize dynamics. After some algebra, we obtain a critical arrival rate

$$\boxed{\rho > \rho_c := \frac{1}{2N}\sum_q \frac{s_q^2}{s_q + \gamma}} \,. \tag{16}$$

Unlike the previous result, this one involves the entire spectrum of $\boldsymbol{S}$ (although it reduces to the previous one when all s_q are the same). Still, the result conveys the same qualitative message: ρ_c increases as the input variance increases, and decreases as the regularization strength γ increases.

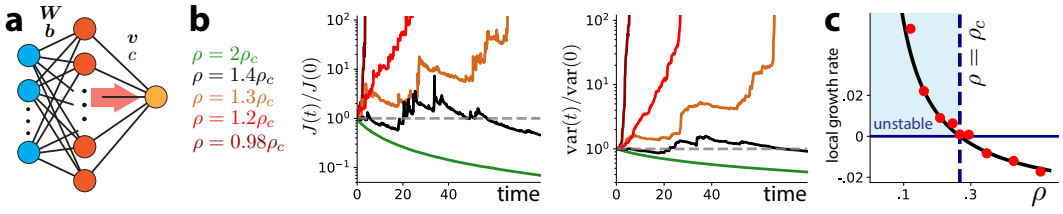


Figure 3: **Instability in a two-layer network.** (**a**) We trained a two-layer network to reproduce the output of an identical teacher network. To ensure our linear model provides a (locally) reasonable description of its dynamics, we initialized it near the teacher's weights. (**b**) Normalized objective and output variance throughout learning. Note that for $\rho < \rho_c$ (dark red), J blows up. For ρ somewhat above the threshold (red, orange), learning still looks unstable on this finite time scale. Learning only appears stable well beyond the threshold (green). (**c**) Instantaneous growth rate of the covariance dynamics near the fixed point. The black line indicates the linear model's prediction and the red dots indicate numerical measurements. Below $\rho = \rho_c$, this local growth rate can be positive, which indicates instability.

4 Instability in a nonlinear network

One virtue of our linear model analysis is that it implies sparse gradient arrivals pose problems for much more general networks. Near a stable fixed point $\boldsymbol{w}^*$ of any well-behaved objective J,

$$J(\boldsymbol{w}) \approx J(\boldsymbol{w}^*) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^T \boldsymbol{H} (\boldsymbol{w} - \boldsymbol{w}^*) \tag{17}$$

where $\boldsymbol{H}$ denotes the (positive definite) Hessian of J at the fixed point $\boldsymbol{w}^*$. Hence, gradient descent dynamics are approximately linear near this fixed point, and if we make the identification $\boldsymbol{S} \to \boldsymbol{H}$, they are well-described by our linear model (Eq. 9).

Example: a two-layer network. Although the above mathematical argument is fairly persuasive, it remains helpful to see an explicit nonlinear example where sparse gradient arrivals destabilize learning. For concreteness' sake, we trained a simple two-layer network to reproduce the output of a teacher network with the same architecture (Fig. 3a). We initialized the student network near the teacher's weights and assessed how changing the arrival rate ρ affected learning. For simplicity and to speed up computation, only the hidden layer's biases $\boldsymbol{b}$ and the output parameters $\boldsymbol{v}$ and c were trained; we found that this did not affect our conclusions, however.

We numerically computed the Hessian around the teacher network's parameters in order to estimate the critical arrival rate ρ_c below which learning ought to be unstable. For $\rho < \rho_c$, we found that the loss function and output variance blew up as learning proceeded (Fig. 3b, c; the dark red line indicates $\rho = 0.98\rho_c$). In fact, we found that ρ marginally above the critical arrival rate ($\rho = 1.2\rho_c$, red; $\rho = 1.3\rho_c$, orange) also exhibited unstable learning dynamics on the finite time scales we examined. Learning only became stable, with a smoothly-decreasing loss, when ρ was much larger than ρ_c ($\rho = 2\rho_c$, green).

We also tested whether the instantaneous change in the output variance (or the "local growth rate"), which is predicted by Eq. 11, empirically describes small-time changes in output variance. We find that this is indeed the case, and that the local growth rate is positive below $\rho = \rho_c$ (Fig. 3c).

5 Instability in a deep linear network

Are deeper networks less stable in the sense we study here? To partially address this question in an analytically tractable setting, we consider an extremely simple deep linear network with $L \geq 1$ layers whose hidden 'layers' are just scalars:

$$y = w_L \cdots w_1 x \tag{18}$$

where w_i is the scalar weight associated with the i-th 'layer', x is a scalar input, and y is a scalar output. Although deep linear networks are admittedly much simpler than their nonlinear counterparts,

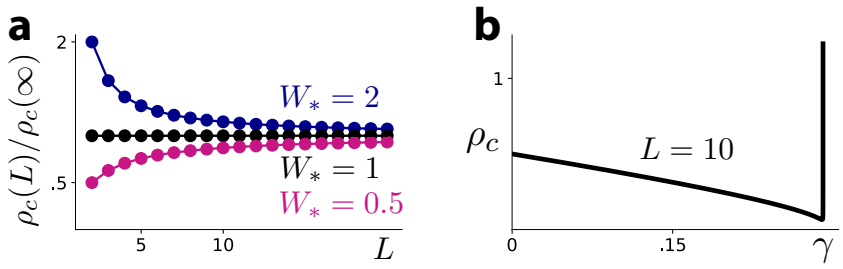


Figure 4: **Depth benignly modulates stability in a deep linear network.** (**a**) ρ_c can increase, decrease, or remain constant with L depending on the value of the target W_*. In all cases, the difference between $\rho_c(L)$ and its asymptotic value $\rho_c(\infty)$ is not substantial. (**b**) ρ_c is non-monotonic as a function of γ ($L = 10$ shown), but this is mainly because sufficiently high regularization destabilizes the nontrivial fixed point.

they have the virtue that they are analytically tractable, and they recapitulate a surprising amount of the phenomenology associated with more realistic networks (Saxe et al., 2013b; Saxe et al., 2019; Domine et al., 2025b).

Consider a teacher-student setting in which our goal is to match y to a target output $y_* := W_* x$:

$$J = \mathbb{E}_x \left\{ \frac{(y - y_*)^2}{2} \right\} = \frac{v}{2}(w_1 \cdots w_L - W_*)^2 \tag{19}$$

where we have assumed that x has mean zero and variance $v > 0$. Although the gradient descent learning dynamics of this network are generally nonlinear and complicated, we can analyze it around the *symmetric fixed point* with $w_1 = \cdots = w_L = q$. We will study the impact of depth on the stability of dynamics near this fixed point. Note that we are studying the fixed point of the regularized objective $\tilde{J}$ (see Eq. 1), not J.

Define the quantity $h := vq^{2L-2}$. The Hessian $\boldsymbol{H}$ of J about the symmetric fixed point has two eigenvalues (see Appendix B):

$$h_\perp = \gamma \quad \text{(with multiplicity } L-1\text{)} \qquad h_\parallel = \gamma + L(h-\gamma) \quad \text{(with multiplicity 1)}\,. \tag{20}$$

Since $\boldsymbol{H}$ is circulant, we can immediately reuse the result from Sec. 3 to find that

$$\boxed{\rho_c(L) = \frac{1}{2L}\left[\frac{h_\parallel^2}{h_\parallel + \gamma} + (L-1)\frac{h_\perp^2}{h_\perp + \gamma}\right] = \frac{1}{2L}\left[\frac{[\gamma + L(h-\gamma)]^2}{L(h-\gamma) + 2\gamma} + \frac{(L-1)}{2}\gamma\right]}\,. \tag{21}$$

In the case of no regularization ($\gamma = 0$), we have the simpler critical arrival rate

$$\boxed{\rho_c(L) = \frac{h}{2} = \frac{v}{2}W_*^{2-2/L}}\,. \tag{22}$$

Interestingly, ρ_c only depends benignly on L, and moreover has a well-behaved infinite-depth ($L \to \infty$) limit (Fig. 4a). There is an apparent instability as γ increases, since ρ_c diverges to infinity at a finite value of γ (Fig. 4b), but this is misleading; it occurs not because covariance dynamics become unstable, but because it is no longer optimal to match W_* when there is too much regularization.

At least in this case, depth does not exacerbate stability issues. In some sense this should be unsurprising, since a result that suggests ρ_c scales poorly with depth would seem inconsistent with biology, whose networks are often deep and highly recurrent; that is, learning is hard, but it should not be impossible.

6 Pooling parameters can improve stability

One approach to improving stability is to average over parameters, since if nearby neurons have similar weights, it may make sense for them to share information (Fig. 5a). Spatially proximal neurons are known to communicate in a variety of non-synaptic ways, including through gap junctions

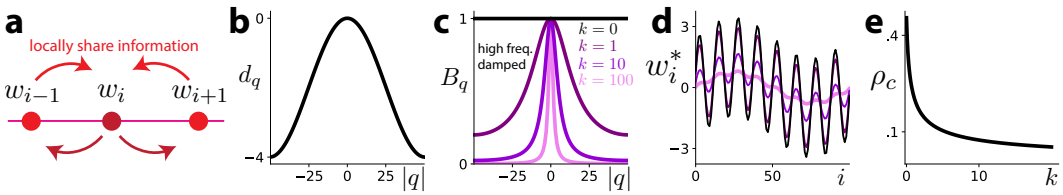


Figure 5: **Pooling parameters can improve stability.** (**a**) If nearby synapses store similar information, neurons may benefit from sharing information with their spatial neighbors. Pictured is a line of synaptic weights; arrows indicate bidirectional information sharing. (**b**) The eigenvalues d_q of $\boldsymbol{D}$ are indexed by spatial frequencies $|q|$, and d_q decreases as $|q|$ increases. (**c**) The q-dependence of d_q causes higher spatial frequencies of the fixed point to be damped, and the effect size increases as the amount of diffusion k increases (more purple $\rightarrow$ higher k). (**d**) One effect of this frequency-dependence is to smooth the fixed point of the average dynamics. In the pictured example, w_i^* is highly oscillatory (black), but dynamics converge to a smoothed version of it (shades of purple). (**e**) ρ_c decreases as k increases, so although parameter-pooling produces bias, it also improves stability.

(Shimizu & Stopfer, 2013) and extracellular vesicles (Nieves Torres & Lee, 2023), and although the computational role of these features is unknown, it is plausible that they share computationally relevant information, including regarding synaptic weights. Parameter-pooling inevitably increases bias, but this tradeoff may be worth it if learning is otherwise unstable.

Assume that weights are arranged along a one-dimensional ring, so that w_i is 'near' w_{i+1} and w_{i-1}. Also assume that $\boldsymbol{S}$ is circulant. Consider a slight modification of Eq. 9 which incorporates parameter sharing via a 'diffusion' matrix $\boldsymbol{D}$ which satisfies $[\boldsymbol{D}\boldsymbol{w}]_i = w_{i-1} - 2w_i + w_{i+1}$:

$$\frac{d}{dt}\boldsymbol{w} = \boldsymbol{r} \odot \frac{1}{\rho}[\boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w}) + \boldsymbol{\xi}] - \eta\gamma\boldsymbol{w} + \eta k\boldsymbol{D}\boldsymbol{w}\,. \tag{23}$$

Here, $k \geq 0$ quantifies the strength of diffusion. This model reduces to the one studied in Sec. 3 when $k = 0$, but otherwise behaves differently. Its average dynamics read

$$\frac{1}{\eta}\frac{d}{dt}\boldsymbol{w} = \boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w}) - \gamma\boldsymbol{w} + k\boldsymbol{D}\boldsymbol{w} \tag{24}$$

and, in the Fourier basis that diagonalizes $\boldsymbol{D}$ and $\boldsymbol{S}$, the components of the fixed point are

$$w_q^{\text{FP}} := \frac{s_q}{s_q + \gamma - kd_q} w_q^* = B_q w_q^*\,. \tag{25}$$

The prefactor B_q, which controls the asymptotic bias in learning, is highly spatial-frequency-dependent: because d_q is more negative for large frequencies $|q|$ (Fig. 5b), B_q is closer to zero for large $|q|$, and the effect size increases as the diffusion strength k increases (Fig. 5c). This implies that what is learned is a spatially smoothed version of the target (Fig. 5d).

Meanwhile, following algebra similar to what we encountered in Sec. 3 (see Appendix C), covariance dynamics are stable provided that a stability condition

$$\boxed{\rho > \rho_c := \frac{1}{2N}\sum_q \frac{s_q^2}{s_q + \gamma - kd_q}} \tag{26}$$

holds. Since ρ_c decreases as k increases, pooling increases stability (Fig. 5e).

7 Averaging over gradient arrivals can improve stability

Another plausible way to improve stability is to effectively reduce the variance of gradient arrivals by not learning *directly* from the available gradient signal, but from a temporal average of recent arrivals (Fig. 6a). We can concretely instantiate this idea by assuming that each parameter w_i has a

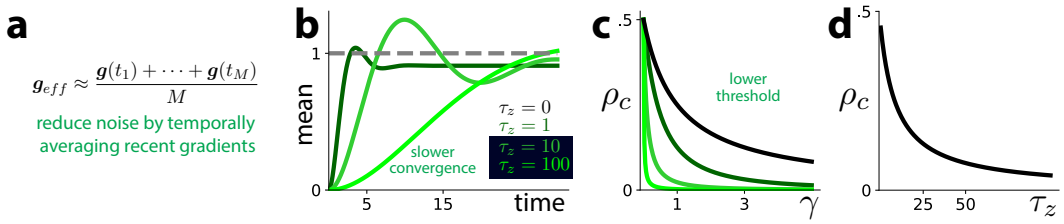


Figure 6: **Averaging over gradient arrivals can improve stability.** (**a**) Intuitively, one can reduce variance by performing a temporal average over recent gradient arrivals. (**b**) Mean dynamics are linear but not monotonic. As the gradient-averaging time scale τ_z increases, they converge more slowly. (black: $\tau_z = 0$; lightest green: $\tau_z = 100$) (**c**, **d**) Increasing τ_z improves stability since ρ_c decreases as long as $\gamma > 0$. But if $\gamma = 0$, there is no stability benefit.

matching accumulator variable z_i which leaky-integrates recent gradients on a time scale $\tau_z > 0$:

$$\begin{aligned} \frac{d}{dt}\boldsymbol{z} &= \boldsymbol{r} \odot \frac{1}{\eta\tau_z\rho}[\boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w}) + \boldsymbol{\xi}] - \frac{1}{\tau_z}\boldsymbol{z} \\ \frac{d}{dt}\boldsymbol{w} &= \eta(\boldsymbol{z} - \gamma\boldsymbol{w}) \, . \end{aligned} \tag{27}$$

By construction, the fixed point of average $\boldsymbol{w}$ dynamics is the same as before, although its transient behavior is slightly different (Fig. 6b). Significantly, though, the addition of the accumulator variable $\boldsymbol{z}$ greatly changes covariance dynamics. The covariances $\boldsymbol{Z} := \mathrm{Cov}(\boldsymbol{z})$, $\boldsymbol{X} := \mathrm{Cov}(\boldsymbol{z}, \boldsymbol{w})$, and $\boldsymbol{W} := \mathrm{Cov}(\boldsymbol{w})$ evolve according to the ODEs (see Appendix D)

$$\begin{aligned} \frac{d}{dt}\boldsymbol{Z} &= -\frac{2}{\tau_z}\boldsymbol{Z} - \frac{1}{\tau_z}(\boldsymbol{X}\boldsymbol{S} + \boldsymbol{S}\boldsymbol{X}^T) + \frac{1}{\eta\rho\tau_z^2}\mathrm{diag}[g_i(\bar{\boldsymbol{w}})^2 + [\boldsymbol{S}\boldsymbol{W}\boldsymbol{S}]_{ii} + \sigma^2_{\text{obs}}] \\ \frac{d}{dt}\boldsymbol{X} &= \eta\boldsymbol{Z} - \left(\eta\gamma + \frac{1}{\tau_z}\right)\boldsymbol{X} - \frac{1}{\tau_z}\boldsymbol{S}\boldsymbol{W} \\ \frac{d}{dt}\boldsymbol{W} &= \eta\left[\boldsymbol{X} + \boldsymbol{X}^T - 2\gamma\boldsymbol{W}\right] \, . \end{aligned} \tag{28}$$

As before, it is much easier to analyze these dynamics if we make specific assumptions about $\boldsymbol{S}$. We will again assume that $\boldsymbol{S}$ is either diagonal or circulant.

7.1 Stability given a diagonal input covariance

If $\boldsymbol{S}$ is diagonal, the dynamics of (Z_{ii}, X_{ii}, W_{ii}) for any fixed i are independent of those for $i' \neq i$, and are characterized by a 3×3 matrix

$$\boldsymbol{M} := \begin{pmatrix} -\frac{2}{\tau_z} & -\frac{2s_i}{\tau_z} & \frac{s_i^2}{\eta\rho\tau_z^2} \\ \eta & -\left(\eta\gamma + \frac{1}{\tau_z}\right) & -\frac{s_i}{\tau_z} \\ 0 & 2\eta & -2\eta\gamma \end{pmatrix} \, . \tag{29}$$

Covariance dynamics are stable provided that $\boldsymbol{M}$ has only negative eigenvalues, and this holds unless ρ is sufficiently small. We can find where the critical value ρ_c^i is by computing where the determinant of $\boldsymbol{M}$ becomes zero. This yields a stability condition

$$\boxed{\rho > \rho_c^i := \frac{s_i^2}{2(1 + \eta\gamma\tau_z)(s_i + \gamma)}} \tag{30}$$

analogous to what we discovered before. It reduces to our earlier result (Eq. 13) in the $\tau_z \to 0$ limit, but otherwise produces a lower critical value of ρ (Fig. 6c, d). Relatedly, we obtain steady state weight variances

$$W_{ii} = \frac{1}{\rho - \rho_c^i}\frac{\sigma^2_{\text{obs}} + \bar{g}_i^2}{2(s_i + \gamma)(1 + \gamma\eta\tau_z)} \tag{31}$$

which again diverge like $1/(\rho - \rho_c^i)$. It is important to emphasize that this gain in stability does not come for free: increases in τ_z can also greatly slow down the convergence of average learning dynamics (Fig. 6b, light green line).

7.2 STABILITY GIVEN A CIRCULANT INPUT COVARIANCE

For circulant $\boldsymbol{S}$, after some algebra, we obtain a result

$$\boxed{\rho > \frac{1}{2N(1+\eta\gamma\tau_z)} \sum_q \frac{s_q^2}{s_q+\gamma}} \tag{32}$$

which is again similar to what we found before (c.f. Eq. 16), and reduces to it in the $\tau_z \to 0$ limit.

8 DISCUSSION

Learning is difficult, and in this work we have identified and analyzed yet another reason why. It is not enough to perform accurate credit assignment *on average* if the relevant learning signal is highly variable, since this high variance can produce unstable learning. The source of this variance, sparse gradient arrivals, is distinct from other types of noise relevant to learning (e.g., noise due to using stochastic rather than full-batch gradient descent), and its effects are apparently more insidious.

Because any well-behaved network is approximately linear near a stable fixed point, one expects this issue to be generic rather than architecture-dependent. We considered three ways of addressing the problem—*regularization*, *parameter-sharing*, and *gradient-averaging*—and each reduce the severity of the issue, but involve tradeoffs. Regularization and parameter-sharing reduce the (infinite) variance at the expense of bias; moreover, parameter-sharing may only be natural when there is specific (e.g., topographic) structure according to which spatial relationships correlate with relationships between synaptic weights. Gradient-averaging does not produce bias, but does slow down learning. It is plausible that all of these mechanisms, and perhaps others, are utilized by biological networks to avoid the kind of instability we study here.

Connection to analyses of stochastic gradient descent. Standard analyses of stochastic gradient descent (Wu et al., 2018; Ahn et al., 2022) involve a mathematical idea similar to the one we study here, whose content is especially similar to the naive model we considered in Sec. 2: if the learning rate or step size is taken to be too large, errors will accumulate and learning will become unstable. In other ways, the issue is qualitatively different. Our model is defined in continuous- rather than discrete-time, and moreover involves a nontrivial interaction between discrete and stochastic gradient arrivals on the one hand, and continuous relaxation (due to regularization) and accumulation (in the model considered in Sec. 7) on the other. Moreover, the biological interpretation of sparse arrivals—which can be viewed as related to error-communication difficulties rather than simply an overall learning rate—is conceptually distinct. Finally, average dynamics are unconditionally stable.

Speculation regarding biological means of variance reduction. If biological systems do indeed confront the problem of potential instability due to sparse gradient (or more generally, learning signal) arrivals, it is plausible that evolution has developed mechanisms to partially combat it. There is biological precedent for each of the three mechanisms—regularization, parameter-pooling, and gradient-averaging—we considered. It is known that the dendritic spines, which are sites of synaptic plasticity, turn over on relatively short time scales, e.g., on the order of one or two weeks in the hippocampus (Attardo et al., 2015). This continual 'forgetting' of synaptic weights can be modeled in the same way we have incorporated regularization. Our gradient-averaging model can be viewed as maintaining a kind of *elibility trace* of learning-related information, and there is mounting evidence that such traces exist and play a major role in biological learning (He et al., 2015; Gerstner et al., 2018). Parameter-pooling has not been directly observed to our knowledge, but as we have mentioned, there are plausible means by which it could occur.

Limitations and future work. For clarity and tractability, we have mainly analyzed simple linear models here. More interesting behavior may be possible in less linear regimes, e.g., away from a stable fixed point. Another major assumption we made is that sparse gradients arrive independently at each synapse according to a Poisson process. In reality, the delivery of gradient information to any given neuron is likely correlated with the delivery of gradients to other neurons (e.g., because the information is coming from roughly the same source). Although details associated with this issue likely differ for more general models, our observations suggest that it is worth further study. Putative methods of credit assignment must produce both good average performance and stable variance.

AI USE STATEMENT

In this work, we used generative AI tools for brainstorming, literature search, to assist with checking the mathematical correctness of our results, and to assist with code generation. The authors have checked all text and results and take full responsibility for them.

REFERENCES

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 247–257. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ahn22a.html.

Alessio Attardo, James E. Fitzgerald, and Mark J. Schnitzer. Impermanence of dendritic spines in live adult CA1 hippocampus. *Nature*, 523(7562):592–596, July 2015. ISSN 1476-4687. doi: 10.1038/nature14467. URL https://doi.org/10.1038/nature14467.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bordelon20a.html.

Francis Crick. The recent excitement about neural networks. *Nature*, 337(6203):129–132, January 1989. ISSN 1476-4687. doi: 10.1038/337129a0. URL https://doi.org/10.1038/337129a0.

Clementine Domine, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro Mediano, and Andrew Saxe. From Lazy to Rich: Exact Learning Dynamics in Deep Linear Networks. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 102485–102536, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fde1a69a5b6e554b2f1f727197d2651d-Paper-Conference.pdf.

Clementine Domine, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro Mediano, and Andrew Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 102485–102536, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fde1a69a5b6e554b2f1f727197d2651d-Paper-Conference.pdf.

Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. *Frontiers in Neural Circuits*, Volume 12 - 2018, 2018. ISSN 1662-5110. doi: 10.3389/fncir.2018.00053. URL https://www.frontiersin.org/journals/neural-circuits/articles/10.3389/fncir.2018.00053.

Kaiwen He, Marco Huertas, Su Z. Hong, XiaoXiu Tie, Johannes W. Hell, Harel Shouval, and Alfredo Kirkwood. Distinct Eligibility Traces for LTP and LTD in Cortical Synapses. *Neuron*, 88(3):528–538, November 2015. ISSN 0896-6273. doi: 10.1016/j.neuron.2015.09.037. URL https://doi.org/10.1016/j.neuron.2015.09.037.

John Frank Charles Kingman. *Poisson processes*, volume 3. Clarendon Press, 1992.

Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7 (1):13276, November 2016. ISSN 2041-1723. doi: 10.1038/ncomms13276. URL https://doi.org/10.1038/ncomms13276.

Timothy P. Lillicrap, Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, June 2020. ISSN 1471-0048. doi: 10.1038/s41583-020-0277-3. URL https://doi.org/10.1038/s41583-020-0277-3.

Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Predictive Coding Approximates Backprop Along Arbitrary Computation Graphs. *Neural Computation*, 34(6):1329–1368, May 2022. ISSN 0899-7667. doi: 10.1162/neco_a_01497. URL https://doi.org/10.1162/neco_a_01497.

Damaris Nieves Torres and Sang H Lee. Inter-neuronal signaling mediated by small extracellular vesicles: wireless communication? *Frontiers in Molecular Neuroscience*, Volume 16 - 2023, 2023. ISSN 1662-5099. doi: 10.3389/fnmol.2023.1187300. URL https://www.frontiersin.org/journals/molecular-neuroscience/articles/10.3389/fnmol.2023.1187300.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL https://doi.org/10.1038/323533a0.

João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1dc3a89d0d440ba31729b0ba74b93a33-Paper.pdf.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv e-prints*, art. arXiv:1312.6120, December 2013a. doi: 10.48550/arXiv.1312.6120.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv e-prints*, art. arXiv:1312.6120, December 2013b. doi: 10.48550/arXiv.1312.6120.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116 (23):11537–11546, 2019. doi: 10.1073/pnas.1820226116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1820226116.

Benjamin Scellier and Yoshua Bengio. Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Frontiers in Computational Neuroscience*, Volume 11 - 2017, 2017. ISSN 1662-5188. doi: 10.3389/fncom.2017.00024. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2017.00024.

Kazumichi Shimizu and Mark Stopfer. Gap junctions. *Current Biology*, 23(23):R1026–R1031, December 2013. ISSN 0960-9822. doi: 10.1016/j.cub.2013.10.067. URL https://doi.org/10.1016/j.cub.2013.10.067.

James C.R. Whittington and Rafal Bogacz. Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, 23(3):235–250, March 2019. ISSN 1364-6613. doi: 10.1016/j.tics.2018.12.005. URL https://doi.org/10.1016/j.tics.2018.12.005.

Lei Wu, Chao Ma, and Weinan E. How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6651526b6fb8f29a00507de6a49ce30f-Paper.pdf.

A Analysis of linear model

In this appendix, we analyze the behavior of the linear model discussed in Sec. 3 of the main text. Recall that it has

$$\frac{d}{dt}\boldsymbol{w} = \boldsymbol{r} \odot \frac{1}{\rho}[\boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w}) + \boldsymbol{\xi}] - \eta\gamma\boldsymbol{w} \tag{33}$$

where $\boldsymbol{r}$ is distributed according to a Poisson arrival process with rate $\eta\rho$ and $\xi \sim \mathcal{N}(0, \sigma^2_{\text{obs}})$ is observation noise associated with a gradient arrival. The master equation that defines how the distribution $P(\boldsymbol{w}|t)$ of weight values evolves in time is

$$\frac{\partial P}{\partial t} = \eta \frac{\partial}{\partial \boldsymbol{w}} \cdot [\gamma \boldsymbol{w} P] + \eta\rho \sum_i \int_{-\infty}^{\infty} d\xi\, \mathcal{N}(\xi; 0, \sigma^2_{\text{obs}})\ [P(\boldsymbol{w} - (1/\rho)[g_i(\boldsymbol{w}) + \xi]\boldsymbol{e}_i|t) - P(\boldsymbol{w}|t)] \tag{34}$$

where $\boldsymbol{e}_i$ denotes the standard basis vector with a 1 as its i-th component and zeros elsewhere. $g_i(\boldsymbol{w})$ is shorthand notation for the i-th component of $\boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w})$, the true gradient.

A.1 Derivation of mean dynamics

Let us first derive the dynamics of the mean weight $\bar{\boldsymbol{w}} := \mathbb{E}[\boldsymbol{w}]$. The master equation (Eq. 34) has the form of a Fokker-Planck equation plus the Poisson arrival term. The regularization term contributes the expected $-\eta\gamma\bar{\boldsymbol{w}}$. We can determine the contribution of the Poisson arrival term by computing its average over $\boldsymbol{w}$. Note that

$$\begin{aligned}
&\eta\rho \int \boldsymbol{w} \left\{ \int_{-\infty}^{\infty} d\xi\, \mathcal{N}(\xi; 0, \sigma^2_{\text{obs}})\ [P(\boldsymbol{w} - (1/\rho)[g_i(\boldsymbol{w}) + \xi]\boldsymbol{e}_i|t) - P(\boldsymbol{w}|t)] \right\} d\boldsymbol{w} \\
&= \eta\rho \int \mathcal{N}(\xi; 0, \sigma^2_{\text{obs}})\ [\boldsymbol{w} P(\boldsymbol{w} - (1/\rho)[g_i(\boldsymbol{w}) + \xi]\boldsymbol{e}_i|t) - \boldsymbol{w} P(\boldsymbol{w}|t)]\ d\boldsymbol{w} \\
&= \eta\rho \int \mathcal{N}(\xi; 0, \sigma^2_{\text{obs}})\ [\boldsymbol{w} + (1/\rho)(g_i(\boldsymbol{w}) + \xi)\boldsymbol{e}_i - \boldsymbol{w}]\, P(\boldsymbol{w}|t)\, d\boldsymbol{w} \\
&= \eta\, \mathbb{E}[g_i(\boldsymbol{w})\boldsymbol{e}_i]\ .
\end{aligned} \tag{35}$$

Summing over i, we find that average dynamics are

$$\boxed{\frac{d}{dt}\bar{\boldsymbol{w}} = \eta\,[\boldsymbol{S}(\boldsymbol{w}^* - \bar{\boldsymbol{w}}) - \gamma\bar{\boldsymbol{w}}]}\ , \tag{36}$$

i.e., they are unbiased.

A.2 Derivation of covariance dynamics

Let us now derive the dynamics of the covariance matrix $\boldsymbol{C} := \text{Cov}(\boldsymbol{w})$. Again, the regularization term contributes to its dynamics in the usual way; the hard part is to determine how the Poisson arrival term contributes.

The average of the Poisson arrival term times $\boldsymbol{w}\boldsymbol{w}^T$ equals

$$\begin{aligned}
&\eta\rho \int \boldsymbol{w}\boldsymbol{w}^T \left\{ \int_{-\infty}^{\infty} d\xi\, \mathcal{N}(\xi; 0, \sigma^2_{\text{obs}})\ [P(\boldsymbol{w} - (1/\rho)[g_i(\boldsymbol{w}) + \xi]\boldsymbol{e}_i|t) - P(\boldsymbol{w}|t)] \right\} d\boldsymbol{w} \\
&= \eta\rho \int \mathcal{N}(\xi; 0, \sigma^2_{\text{obs}})\ \left[\boldsymbol{w}\boldsymbol{w}^T P(\boldsymbol{w} - (1/\rho)[g_i(\boldsymbol{w}) + \xi]\boldsymbol{e}_i|t) - \boldsymbol{w}\boldsymbol{w}^T P(\boldsymbol{w}|t)\right]\ d\boldsymbol{w} \\
&= \eta\rho \int \mathcal{N}(\xi; 0, \sigma^2_{\text{obs}})\ \left\{\ [\boldsymbol{w} + (1/\rho)(g_i(\boldsymbol{w}) + \xi)\boldsymbol{e}_i][\boldsymbol{w} + (1/\rho)(g_i(\boldsymbol{w}) + \xi)\boldsymbol{e}_i]^T - \boldsymbol{w}\boldsymbol{w}^T\ \right\} P(\boldsymbol{w}|t) \\
&= \eta\, \mathbb{E}[(\boldsymbol{w}\boldsymbol{e}_i^T + \boldsymbol{e}_i\boldsymbol{w}^T)g_i(\boldsymbol{w}) + (1/\rho)(g_i(\boldsymbol{w})^2 + \sigma^2_{\text{obs}})\boldsymbol{e}_i\boldsymbol{e}_i^T]\ .
\end{aligned} \tag{37}$$

Summing over i, this becomes

$$\begin{aligned}
&\eta \sum_i \mathbb{E}[(\boldsymbol{w}\boldsymbol{e}_i^T + \boldsymbol{e}_i\boldsymbol{w}^T)g_i(\boldsymbol{w}) + (1/\rho)(g_i(\boldsymbol{w})^2 + \sigma^2_{\text{obs}})\boldsymbol{e}_i\boldsymbol{e}_i^T] \\
&= \eta\, \mathbb{E}[\boldsymbol{w}(\boldsymbol{w}^* - \boldsymbol{w})^T\boldsymbol{S} + \boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w})\boldsymbol{w}^T + (1/\rho)\text{diag}(g_i(\boldsymbol{w})^2 + \sigma^2_{\text{obs}})]\ .
\end{aligned} \tag{38}$$

Meanwhile, the time derivative of the mean product $\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T$ equals

$$\begin{aligned}\frac{d}{dt}(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T) &= \left(\frac{d}{dt}\bar{\boldsymbol{w}}\right)\bar{\boldsymbol{w}}^T + \bar{\boldsymbol{w}}\left(\frac{d}{dt}\bar{\boldsymbol{w}}\right)^T \\ &= \eta\left[\boldsymbol{S}(\boldsymbol{w}^* - \bar{\boldsymbol{w}}) - \gamma\bar{\boldsymbol{w}}\right]\bar{\boldsymbol{w}}^T + \eta\bar{\boldsymbol{w}}\left[\boldsymbol{S}(\boldsymbol{w}^* - \bar{\boldsymbol{w}}) - \gamma\bar{\boldsymbol{w}}\right]^T \ .\end{aligned} \tag{39}$$

Putting these results together, we find that $\boldsymbol{C} = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T$ evolves according to

$$\frac{1}{\eta}\frac{d}{dt}\boldsymbol{C} = -2\gamma\boldsymbol{C} - \boldsymbol{C}\boldsymbol{S} - \boldsymbol{S}\boldsymbol{C} + (1/\rho)\ \mathrm{diag}(\mathbb{E}[g_i(\boldsymbol{w})^2] + \sigma^2_{\text{obs}})\ . \tag{40}$$

This result can be slightly simplified using the fact that

$$\mathbb{E}[g_i(\boldsymbol{w})^2] = \bar{g}_i^2 + [\boldsymbol{S}\boldsymbol{C}\boldsymbol{S}]_{ii}\ , \tag{41}$$

i.e., using the fact that the squared gradient can be written as a sum of 'bias' and 'variance' terms. Here $\bar{g}_i$ is shorthand notation for $g_i(\bar{\boldsymbol{w}})$. We conclude that

$$\boxed{\frac{1}{\eta}\frac{d}{dt}\boldsymbol{C} = -2\gamma\boldsymbol{C} - \boldsymbol{C}\boldsymbol{S} - \boldsymbol{S}\boldsymbol{C} + \frac{1}{\rho}\sigma^2_{\text{obs}}\boldsymbol{I} + \frac{1}{\rho}\ \mathrm{diag}(\bar{g}_i^2 + [\boldsymbol{S}\boldsymbol{C}\boldsymbol{S}]_{ii})}\ . \tag{42}$$

Analysis of Eq. 42 is cumbersome for general $\boldsymbol{S}$, but greatly simplifies for specific choices of $\boldsymbol{S}$. In the next two subsections, we perform a stability analysis of these covariance dynamics assuming $\boldsymbol{S}$ is either diagonal or circulant.

A.3 Stability analysis for diagonal S

Assume $\boldsymbol{S}$ is diagonal. Let $c_i := C_{ii}$ and $s_i := S_{ii}$, and collect both into the vectors $\boldsymbol{c}$ and $\boldsymbol{s}$. Eq. 42 implies that the diagonal entries of $\boldsymbol{C}$ satisfy

$$\begin{aligned}\frac{1}{\eta}\frac{d}{dt}\boldsymbol{c} &= -2\gamma\boldsymbol{c} - 2\boldsymbol{S}\boldsymbol{c} + \frac{1}{\rho}\sigma^2_{\text{obs}}\boldsymbol{1} + \frac{1}{\rho}\left[\boldsymbol{g} + \boldsymbol{S}^2\boldsymbol{c}\right] \\ &= \frac{1}{\rho}[\sigma^2_{\text{obs}}\boldsymbol{1} + \boldsymbol{g}] - [2\gamma\boldsymbol{I} + 2\boldsymbol{S} - \frac{1}{\rho}\boldsymbol{S}^2]\boldsymbol{c}\end{aligned} \tag{43}$$

where we use $\boldsymbol{g}$ to denote the vector whose components are $\bar{g}_i^2$ and $\boldsymbol{1}$ to denote the vector whose components are all ones.

Stability. Because this ODE is linear, it is clear that stability requires that

$$2\gamma + 2s_i - \frac{1}{\rho}s_i^2 > 0 \tag{44}$$

for all i. Rearranging, this implies that we require

$$\boxed{\rho > \rho_c^i := \frac{s_i^2}{2(s_i + \gamma)}} \tag{45}$$

for all i, or equivalently that

$$\boxed{\rho > \rho_c := \frac{s_{\text{max}}^2}{2(s_{\text{max}} + \gamma)}} \tag{46}$$

where s_{max} is the largest input variance s_i.

Steady state covariance. Setting the left-hand side of Eq. 43 to zero, we find that the steady state covariance is

$$\boxed{C_{ii} = \frac{1}{\rho - \rho_c^i}\frac{\sigma^2_{\text{obs}} + g_i(\bar{\boldsymbol{w}}^*)^2}{2(s_i + \gamma)}}\ , \tag{47}$$

where $\bar{\boldsymbol{w}}^*$ denotes the fixed point of mean dynamics, provided that $\rho > \rho_c$.

A.4 STABILITY ANALYSIS FOR CIRCULANT S

Assume $\boldsymbol{S}$ is circulant. Importantly, circulant matrices can be diagonalized via (discrete) Fourier basis functions. Concretely, let $\boldsymbol{F}$ denote the $N \times N$ matrix whose components are

$$F_{jq} = \frac{1}{\sqrt{N}} e^{iqj} \ . \tag{48}$$

Then we can write $\boldsymbol{S} = \boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^\dagger$, where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{S}$. Moreover, if we define the transformed covariance $\tilde{\boldsymbol{C}} := \boldsymbol{F}^\dagger \boldsymbol{C} \boldsymbol{F}$, then

$$\frac{1}{N}\sum_i [\boldsymbol{S}\boldsymbol{C}\boldsymbol{S}]_{ii} = \frac{1}{N}\sum_q s_q^2 \tilde{C}_{qq} \ , \tag{49}$$

where s_q here denotes the q-th eigenvalue of $\boldsymbol{S}$. The virtue of this result, and hence the reason the circulant assumption is useful, is that the above quantity is independent of the index i.

Combining this result with Eq. 42, we find that the diagonal entries $\tilde{C}_{qq}$ of the transformed covariance satisfy

$$\frac{1}{\eta}\frac{d}{dt}\tilde{C}_{qq} = -2\gamma \tilde{C}_{qq} - 2 s_q \tilde{C}_{qq} + \frac{1}{\rho}\sigma^2_{\text{obs}} + \frac{1}{\rho}\left(A_{qq} + \frac{1}{N}\sum_{q'} s_{q'}^2 \tilde{C}_{q'q'}\right) \tag{50}$$

where $\tilde{\boldsymbol{A}} := \boldsymbol{F}^\dagger \text{diag}(\bar{g}_i^2) \boldsymbol{F}$.

Stability. Define $c_q := \tilde{C}_{qq}$ and collect the c_q into the vector $\boldsymbol{c}$. Also, collect the $\tilde{A}_{qq}$ into $\boldsymbol{a}$. Eq. 50 can be written in the form

$$\begin{aligned}
\frac{1}{\eta}\frac{d}{dt}\boldsymbol{c} &= -2\gamma \boldsymbol{c} - 2\boldsymbol{\Lambda}\boldsymbol{c} + \frac{1}{\rho}\sigma^2_{\text{obs}}\mathbf{1} + \frac{1}{\rho}\left(\boldsymbol{a} + \frac{1}{N}\mathbf{1}\mathbf{1}^T\boldsymbol{\Lambda}^2\boldsymbol{c}\right) \\
&= \frac{1}{\rho}(\sigma^2_{\text{obs}}\mathbf{1} + \boldsymbol{a}) - \left[2\gamma \boldsymbol{I} + 2\boldsymbol{\Lambda} - \frac{1}{\rho N}\mathbf{1}\mathbf{1}^T\boldsymbol{\Lambda}^2\right]\boldsymbol{c} \\
&= \frac{1}{\rho}(\sigma^2_{\text{obs}}\mathbf{1} + \boldsymbol{a}) - 2\gamma\left[\boldsymbol{I} + (1/\gamma)\boldsymbol{\Lambda} - \frac{1}{2\rho N\gamma}\mathbf{1}\mathbf{1}^T\boldsymbol{\Lambda}^2\right]\boldsymbol{c} \ .
\end{aligned} \tag{51}$$

From this result, it is clear that the stability of $\boldsymbol{c}$ dynamics depends on the matrix

$$\boldsymbol{M} := \boldsymbol{I} + \frac{1}{\gamma}\boldsymbol{\Lambda} - \frac{1}{2\rho N\gamma}\mathbf{1}\mathbf{1}^T\boldsymbol{\Lambda}^2 \ . \tag{52}$$

This a diagonal matrix plus a rank-one perturbation. We can use the matrix determinant lemma to compute its determinant:

$$\begin{aligned}
\det(\boldsymbol{M}) &= \det(\boldsymbol{I} + \frac{1}{\gamma}\boldsymbol{\Lambda})\left[1 - \frac{1}{2\rho N\gamma}\mathbf{1}^T\boldsymbol{\Lambda}^2(\boldsymbol{I} + \frac{1}{\gamma}\boldsymbol{\Lambda})^{-1}\mathbf{1}\right] \\
&= \det(\boldsymbol{I} + \frac{1}{\gamma}\boldsymbol{\Lambda})\left[1 - \frac{1}{2\rho N\gamma}\sum_q \frac{s_q^2}{s_q/\gamma + 1}\right] \ .
\end{aligned} \tag{53}$$

For sufficiently high ρ, it is clear that this determinant is positive, which indicates stability. We can identify the stability threshold by determining where the determinant is zero; since the eigenvalues of $\boldsymbol{S}$ are all positive, this can only happen when the bracketed factor equals zero.

We conclude that stability requires

$$\boxed{\rho > \rho_c := \frac{1}{2N}\sum_q \frac{s_q^2}{s_q + \gamma}} \tag{54}$$

since beyond this value of ρ, the determinant is negative.

Steady state covariance. The above covariance dynamics can be rewritten as

$$\frac{1}{\eta}\frac{d}{dt}\boldsymbol{c} = \frac{1}{\rho}(\sigma_{\text{obs}}^2\mathbf{1} + \boldsymbol{a}) - 2\gamma\boldsymbol{M}\boldsymbol{c}\,, \tag{55}$$

so its steady state occurs at

$$\boldsymbol{c}^* := \frac{1}{2\rho\gamma}\boldsymbol{M}^{-1}(\sigma_{\text{obs}}^2\mathbf{1} + \boldsymbol{a})\,. \tag{56}$$

This can be simplified if we write out $\boldsymbol{M}^{-1}$ explicitly. By the Sherman-Morrison formula,

$$\begin{aligned}\boldsymbol{M}^{-1} &= (\boldsymbol{I} + \boldsymbol{\Lambda}/\gamma)^{-1} + \frac{1}{2\rho N\gamma}\frac{(\boldsymbol{I} + \boldsymbol{\Lambda}/\gamma)^{-1}\mathbf{1}\mathbf{1}^T\boldsymbol{\Lambda}^2(\boldsymbol{I} + \boldsymbol{\Lambda}/\gamma)^{-1}}{1 - \frac{1}{2\rho N\gamma}\mathbf{1}^T\boldsymbol{\Lambda}^2(\boldsymbol{I} + \boldsymbol{\Lambda}/\gamma)^{-1}\mathbf{1}} \\ &= (\boldsymbol{I} + \boldsymbol{\Lambda}/\gamma)^{-1} + \frac{1}{2N\gamma}\frac{(\boldsymbol{I} + \boldsymbol{\Lambda}/\gamma)^{-1}\mathbf{1}\mathbf{1}^T\boldsymbol{\Lambda}^2(\boldsymbol{I} + \boldsymbol{\Lambda}/\gamma)^{-1}}{\rho - \rho_c}\,.\end{aligned} \tag{57}$$

Although the expression we obtain is more complicated than in the case of diagonal $\boldsymbol{S}$, we still see a divergence that goes like $1/(\rho - \rho_c)$ as ρ approaches the critical arrival rate from above.

B MORE DETAILS ON THE DEEP LINEAR NETWORK

In this appendix, we provide a more thorough account of the deep linear network analysis described in Sec. 5. We consider a scalar deep network with $L \geq 2$ 'layers':

$$y = w_L \cdots w_1 x \; . \tag{58}$$

Here, x is the input, w_i is the weight associated with the i-th 'layer', and y is the output. x is assumed to have mean zero and variance v.

We assume this network is trained via gradient descent to match a target output $y_* := W_* x$ which also depends linearly on its input. The corresponding objective equals

$$\tilde{J} = J + \sum_i \frac{\gamma}{2} w_i^2 = \frac{v}{2}(w_1 \cdots w_L - W_*)^2 + \sum_i \frac{\gamma}{2} w_i^2 \; . \tag{59}$$

Fixed point assumption. Although learning dynamics are generally nonlinear, they are approximately linear near a fixed point. For simplicity, we will consider the symmetric fixed point with

$$w_1 = \cdots = w_L = q \; , \tag{60}$$

where q is chosen to minimize the objective, i.e.,

$$q := \text{argmin}_q \frac{v}{2}(q^L - W_*)^2 + \frac{L\gamma}{2} q^2 \; . \tag{61}$$

Since q minimizes $\tilde{J}$, $d\tilde{J}/dq = 0$, so the condition

$$\boxed{v(W_* - q^L) q^{L-2} = \gamma} \tag{62}$$

holds.

Hessian near fixed point. To characterize linear dynamics near the symmetric fixed point, we need to compute the corresponding $L \times L$ Hessian matrix. Note that the first derivatives at an arbitrary point are

$$\frac{\partial J}{\partial w_i} = v(w_1 \cdots w_L - W_*) \frac{W}{w_i} \tag{63}$$

where we define $W := w_1 \cdots w_L$. The Hessian's components at an arbitrary point are

$$\begin{aligned} H_{ii} &= v \frac{W^2}{w_i^2} \\ H_{ij} &= v \frac{W^2}{w_i w_j} + v(W - W_*) \frac{W}{w_i w_j} \; . \end{aligned} \tag{64}$$

At the symmetric fixed point,

$$\begin{aligned} H_{ii} &= v q^{2L-2} \\ H_{ij} &= v q^{2L-2} + v(q^L - W_*) q^{L-2} \; . \end{aligned} \tag{65}$$

To ease notation, define $h := v q^{2L-2}$. This result implies that we can write the Hessian matrix at the symmetric fixed point as

$$\boldsymbol{H} = \gamma \boldsymbol{I} + (h - \gamma) \mathbf{1}\mathbf{1}^T \; , \tag{66}$$

i.e., it is a diagonal matrix plus a rank-one 'correction'. We immediately learn that $\boldsymbol{H}$ has two eigenvalues: one associated with the correction, and another eigenvalue (with multiplicity $L - 1$) associated with the orthogonal directions:

$$h_\perp = \gamma \qquad\qquad h_\parallel = \gamma + L(h - \gamma) \; . \tag{67}$$

Stability condition near symmetric fixed point. This Hessian is circulant, so we can reuse a linear model result (see Appendix A) to immediately learn that

$$\boxed{\rho_c(L) = \frac{1}{2L}\left[\frac{h_\parallel^2}{h_\parallel + \gamma} + (L-1)\frac{h_\perp^2}{h_\perp + \gamma}\right] = \frac{1}{2L}\left[\frac{[\gamma + L(h-\gamma)]^2}{L(h-\gamma) + 2\gamma} + \frac{(L-1)}{2}\gamma\right]} \; . \tag{68}$$

In the case of no regularization ($\gamma = 0$), we have the simpler critical arrival rate

$$\boxed{\rho_c(L) = \frac{h}{2} = \frac{v}{2} W_*^{2-2/L}} \tag{69}$$

which asymptotically approaches $\rho_c(\infty) := (v/2) W_*^2$ as $L \to \infty$.

C MORE DETAILS ON POOLING PARAMETERS

In this appendix, we derive results related to the parameter-pooling model discussed in Sec. 6. It is defined in almost the same way as our linear model (see Sec. 3 and Appendix A):

$$\frac{d}{dt}\boldsymbol{w} = \boldsymbol{r} \odot \frac{1}{\rho}[\boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w}) + \boldsymbol{\xi}] - \eta\gamma\boldsymbol{w} + \eta k\boldsymbol{D}\boldsymbol{w} \, . \tag{70}$$

The main difference is that there is an additional term $\eta k\boldsymbol{D}\boldsymbol{w}$, where $\boldsymbol{D}$ is a 'diffusion' matrix that implements a sort of spatial parameter pooling. In particular, $\boldsymbol{D}$ is defined such that

$$\sum_j D_{ij} w_j := w_{i+1} - 2w_i + w_{i-1} \tag{71}$$

for all i. The interpretation is that information flows from nodes $i+1$ and $i-1$ to node i, and from node i to nodes $i+1$ and $i-1$.

For simplicity, we assume periodic boundary conditions, so that $\boldsymbol{D}$ is circulant rather than tridiagonal. Under this assumption, its eigenvalues are

$$d_q := -4\sin^2\left(\frac{\pi q}{N}\right) \leq 0 \tag{72}$$

where $q = 0, 1, ..., N-1$, and it can be diagonalized using discrete Fourier basis functions.

The derivation of mean and covariance dynamics for this model is almost the same as it was for the linear model studied in Appendix A; everything is the same up to the identification

$$\gamma\boldsymbol{I} \to \gamma\boldsymbol{I} - k\boldsymbol{D} \, . \tag{73}$$

For this reason, we omit the derivation and directly state the results. As in Appendix A, we assume that $\boldsymbol{S}$ is circulant for analytic tractability. This assumption is particularly important here because it implies that $\boldsymbol{S}$ and $\boldsymbol{D}$ commute.

Mean dynamics. Mean dynamics follow

$$\boxed{\frac{d}{dt}\bar{\boldsymbol{w}} = \eta\left[\boldsymbol{S}(\boldsymbol{w}^* - \bar{\boldsymbol{w}}) - (\gamma\boldsymbol{I} - k\boldsymbol{D})\bar{\boldsymbol{w}}\right]} \, . \tag{74}$$

The corresponding fixed point is

$$\bar{\boldsymbol{w}}^{\text{FP}} := (\boldsymbol{S} + \gamma\boldsymbol{I} - k\boldsymbol{D})^{-1}\boldsymbol{S}\boldsymbol{w}^* \, . \tag{75}$$

In the discrete Fourier basis, this becomes

$$w_q^{\text{FP}} := \frac{s_q}{s_q + \gamma - kd_q} w_q^* = B_q w_q^* \, . \tag{76}$$

where s_q denotes the q-th eigenvalue of $\boldsymbol{S}$ and d_q denotes the q-th eigenvalue of $\boldsymbol{D}$. As noted in the main text, the specific form of d_q makes the overall associated prefactor B_q fairly spatial-frequency-dependent.

Covariance dynamics. Covariance dynamics follow

$$\boxed{\frac{1}{\eta}\frac{d}{dt}\boldsymbol{C} = -2\gamma\boldsymbol{C} - \boldsymbol{C}\boldsymbol{S} + k\boldsymbol{C}\boldsymbol{D} + k\boldsymbol{D}\boldsymbol{C} - \boldsymbol{S}\boldsymbol{C} + \frac{1}{\rho}\sigma_{\text{obs}}^2\boldsymbol{I} + \frac{1}{\rho}\,\text{diag}(\bar{g}_i^2 + [\boldsymbol{S}\boldsymbol{C}\boldsymbol{S}]_{ii})} \, . \tag{77}$$

Moreover, the diagonal elements of the discrete-Fourier-transformed matrix $\tilde{\boldsymbol{C}} := \boldsymbol{F}^\dagger\boldsymbol{C}\boldsymbol{F}$ (see Appendix A for more context) satisfy

$$\frac{1}{\eta}\frac{d}{dt}\tilde{C}_{qq} = -2\gamma\tilde{C}_{qq} + 2kd_q\tilde{C}_{qq} - 2s_q\tilde{C}_{qq} + \frac{1}{\rho}\sigma_{\text{obs}}^2 + \frac{1}{\rho}\left(A_{qq} + \frac{1}{N}\sum_{q'} s_{q'}^2\tilde{C}_{q'q'}\right) . \tag{78}$$

Following a slight modification of the argument from Appendix A, we obtain the stability condition

$$\boxed{\rho > \frac{1}{2N}\sum_q \frac{s_q^2}{s_q + \gamma - kd_q}} \, . \tag{79}$$

D MORE DETAILS ON AVERAGING OVER GRADIENT ARRIVALS

In this appendix, we analyze the mean and covariance dynamics of the gradient-accumulation model discussed in Sec. 7 of the main text. It is defined similarly to the linear model (see Sec. 3 and Appendix A), but with a 'gradient accumulation' variable z_i for each weight w_i:

$$\begin{aligned}\frac{d}{dt}\boldsymbol{z} &= \boldsymbol{r} \odot \frac{1}{\eta\tau_z\rho}[\boldsymbol{S}(\boldsymbol{w}^* - \boldsymbol{w}) + \boldsymbol{\xi}] - \frac{1}{\tau_z}\boldsymbol{z} \\ \frac{d}{dt}\boldsymbol{w} &= \eta(\boldsymbol{z} - \gamma\boldsymbol{w})\ .\end{aligned} \tag{80}$$

This model's master equation has the form

$$\begin{aligned}\frac{\partial P}{\partial t} =\ & -\eta\frac{\partial}{\partial \boldsymbol{w}}\cdot[(\boldsymbol{z} - \gamma\boldsymbol{w})P] + \frac{1}{\tau_z}\frac{\partial}{\partial \boldsymbol{z}}\cdot(\boldsymbol{z}P) \\ & + \eta\rho\sum_i \int_{-\infty}^{\infty} d\xi\, \mathcal{N}(\xi; 0, \sigma^2_{\text{obs}})\ \left[P(\boldsymbol{z} - \frac{1}{\eta\rho\tau_z}[g_i(\boldsymbol{w}) + \xi]\boldsymbol{e}_i, \boldsymbol{w}|t) - P(\boldsymbol{z}, \boldsymbol{w}|t)\right]\ .\end{aligned} \tag{81}$$

In this section, because there are multiple covariance matrices, we will use different notation than elsewhere. Let $\boldsymbol{Z} := \text{Cov}(\boldsymbol{z})$, $\boldsymbol{X} := \text{Cov}(\boldsymbol{z}, \boldsymbol{w})$ denote the cross-covariance, and $\boldsymbol{W} := \text{Cov}(\boldsymbol{w})$.

The calculations required to treat this model largely resemble those from Appendix A, but are lengthier, and hence more tedious. Because they are straightforward, we skip a number of intermediate steps and mostly present the final results.

Mean dynamics. This model's average dynamics are

$$\begin{aligned}\frac{d}{dt}\boldsymbol{z} &= \frac{1}{\tau_z}\boldsymbol{S}(\boldsymbol{w}^* - \bar{\boldsymbol{w}}) - \frac{1}{\tau_z}\bar{\boldsymbol{z}} \\ \frac{d}{dt}\bar{\boldsymbol{w}} &= \eta(\bar{\boldsymbol{z}} - \gamma\bar{\boldsymbol{w}})\ .\end{aligned} \tag{82}$$

In particular, the steady state value of $\bar{\boldsymbol{w}}$ is the same as it is in the linear model; this kind of gradient accumulation does not produce bias, at least asymptotically.

Covariance dynamics. After a tedious calculation, one can find that

$$\begin{aligned}\frac{d}{dt}\boldsymbol{Z} &= -\frac{2}{\tau_z}\boldsymbol{Z} - \frac{1}{\tau_z}(\boldsymbol{X}\boldsymbol{S} + \boldsymbol{S}\boldsymbol{X}^T) + \frac{1}{\eta\rho\tau_z^2}\text{diag}[g_i(\bar{\boldsymbol{w}})^2 + [\boldsymbol{S}\boldsymbol{W}\boldsymbol{S}]_{ii} + \sigma^2_{\text{obs}}] \\ \frac{d}{dt}\boldsymbol{X} &= \eta\boldsymbol{Z} - \left(\eta\gamma + \frac{1}{\tau_z}\right)\boldsymbol{X} - \frac{1}{\tau_z}\boldsymbol{S}\boldsymbol{W} \\ \frac{d}{dt}\boldsymbol{W} &= \eta\left[\boldsymbol{X} + \boldsymbol{X}^T - 2\gamma\boldsymbol{W}\right]\ .\end{aligned} \tag{83}$$

Diagonal S. Assume $\boldsymbol{S}$ is diagonal and let Z_i, X_i, and W_i denote the diagonals of these covariance matrices. The diagonal entries have dynamics

$$\begin{aligned}\frac{d}{dt}Z_i &= -\frac{2}{\tau_z}Z_i - \frac{2}{\tau_z}s_iX_i + \frac{1}{\eta\rho\tau_z^2}[g_i(\bar{\boldsymbol{w}})^2 + s_i^2W_i + \sigma^2_{\text{obs}}] \\ \frac{d}{dt}X_i &= \eta Z_i - \left(\eta\gamma + \frac{1}{\tau_z}\right)X_i - \frac{1}{\tau_z}s_iW_i \\ \frac{d}{dt}W_i &= 2\eta\left[X_i - \gamma W_i\right]\ .\end{aligned} \tag{84}$$

The stability of this three-dimensional system of ODEs is characterized by the 3×3 matrix

$$\boldsymbol{M} := \begin{pmatrix} -\frac{2}{\tau_z} & -\frac{2s_i}{\tau_z} & \frac{s_i^2}{\eta\rho\tau_z^2} \\ \eta & -\left(\eta\gamma + \frac{1}{\tau_z}\right) & -\frac{s_i}{\tau_z} \\ 0 & 2\eta & -2\eta\gamma \end{pmatrix}\ . \tag{85}$$

We can directly compute that its determinant is

$$\begin{aligned} &-\frac{2}{\tau_z}\left[2\eta\gamma(\eta\gamma+1/\tau_z)+2\eta s_i/\tau_z\right]+\frac{2s_i}{\tau_z}\left[-2\eta^2\gamma\right]+\frac{s_i^2}{\eta\rho\tau_z^2}\left[2\eta^2\right] \\ =&-\frac{2}{\tau_z}\left[2\eta\gamma(\eta\gamma+1/\tau_z)+2\eta s_i(1/\tau_z+\eta\gamma)\right]+\frac{s_i^2}{\eta\rho\tau_z^2}\left[2\eta^2\right] \\ =&-\frac{4\eta}{\tau_z}(\eta\gamma+1/\tau_z)\left[\gamma+s_i\right]+\frac{s_i^2}{\eta\rho\tau_z^2}\left[2\eta^2\right]\ . \end{aligned} \tag{86}$$

Stability requires that the determinant is negative. Hence, we need it to be true that

$$2(1+\eta\gamma\tau_z)(\gamma+s_i)-\frac{s_i^2}{\rho}>0\ . \tag{87}$$

Equivalently,

$$\boxed{\rho>\rho_c^i:=\frac{s_i^2}{2(1+\gamma\eta\tau_z)(s_i+\gamma)}}\ . \tag{88}$$

In the infinitely fast integration limit ($\tau_z \to 0$), this reduces to the linear model result (see Appendix A).

The corresponding steady state weight variances are

$$\boxed{W_{ii}=\frac{1}{\rho-\rho_c^i}\frac{\sigma_{\text{obs}}^2+\bar{g}_i^2}{2(s_i+\gamma)(1+\gamma\eta\tau_z)}}\ . \tag{89}$$

Interestingly, this expression is the same as what we found for the linear model, except that ρ_c^i has a different form.

Circulant S. If $\boldsymbol{S}$ is instead circulant, we must go through a lengthier version of the circulant algebra from Appendix A. In the end, we find almost the same results. Most importantly, we find a stability condition

$$\boxed{\rho>\rho_c:=\frac{1}{2N(1+\eta\gamma\tau_z)}\sum_k\frac{s_k^2}{s_k+\gamma}} \tag{90}$$

which reduces to the corresponding linear model condition in the $\tau_z \to 0$ limit.