# THE RATE MEASURE: HOW FAR LINEARISATION GOVERNS TRAINING AND WHY THAT HORIZON CANNOT BE READ

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The loss curve of a network trained by gradient descent at squared error can be predicted before it trains. The linearised loss is a Laplace transform of a measure carried by the output-space Gauss–Newton operator $G = JJ^\top/n$, and one Krylov pass before training returns it. We measure the horizon, how long real networks follow this forecast, and find it set by how far $G$ moves during training. From $0.11$M to 3B parameters, standard-parameterised transformers follow the forecast through the whole window we score from 16M parameters up. Under maximal-update parameterisation the same networks leave it early. At one size and one initial operator, scaling the targets alone moves the horizon from the whole run to under two hundred steps. At a fixed example count, across size, parameterisation and target scale, runs whose $G$ moves further leave the forecast earlier. A tenth of the way into training, this motion flags the runs whose forecast will fail. On MNIST classifiers its relative size does, and every classifier it certifies holds. On transformers to 3B its growth along the residual does, withholding all but four of the runs that fail. Before the first step, the forecast tracks the measured rate curve of standard-parameterised transformers more closely, in median over log time, than every scaling law we fit. It matches the best predictor that sees none of the run, with no fitted parameter. We prove that in the worst case over networks sharing the linearisation's inputs, no function of them predicts the horizon, and that at cross-entropy the rate is asymptotically model-free.

## 1 INTRODUCTION

A network's loss curve can be predicted before it trains. At squared error under gradient descent, full batch unless stated, the linearised model's loss is a sum of exponentials. Their rates are set by the eigenvalues of the output-space Gauss–Newton operator $G = JJ^\top/n$, and their weights are the initial error's shares along its eigenvectors. One Krylov pass before training, the *read*, returns them. The real run follows that curve while $G$ stays where it was read, and the length of that agreement is the linearisation's *horizon*. How far $G$ changes along the residual the network is still fitting sets the horizon, and to first order that change is the network's excess rate over its linearisation. From $0.11$M to 3B parameters, standard-parameterised transformers follow the curve for most of their loss descent, and longer as they grow. Under maximal-update parameterisation ($\mu$P, Yang & Hu, 2021) the same transformers follow it for less, and on the ladder to $51$M for less as they grow (Figure 1). Scaling only the targets, with $G(0)$ identical bit for bit, moves the horizon from beyond the run to under two hundred steps. Small datasets do not make training lazy by construction, since on the same data the $\mu$P networks leave the read early. With $1{,}024$ examples every $\mu$P run leaves, the larger the earlier, while every standard twin follows the read far longer. Past its horizon each of these networks learns faster than its linearisation. In the worst case no function of the linearisation's inputs can predict the horizon, which depends on a curvature scale the linearisation does not contain. A tenth of the way into training the horizon becomes visible, and the early test reads it there.

**Setting and terms.** The headline runs are full-batch gradient descent at the fixed step $\eta = 0.3/\lambda_{\max}(G(0))$, on sixteen sequences for the transformers trained from scratch, ten or sixteen for released checkpoints. §5.1 carries the read to $16{,}384$ examples. Errors are $|\log_{10}(\hat{q}/q)|$ in decades, so $0.029$ decades is a factor $1.07$. A *ladder* is one architecture at a range of sizes. A **cell** is



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.

one network, seed and data over one window of steps, the unit of every count (Table 1). A cell *holds* when the median rate error over a logarithmic grid of its steps is within tolerance. An early test read at a tenth is scored on the cell's *later error*, that median over the remaining nine tenths. The tolerance is a third of a decade for the drift test and a quarter for the counts over all scored cells. The median is taken *in log time* because the loss falls over several decades and its rate curve is read on log axes. A cell's *horizon* is the first step whose error leaves tolerance, and the forecast *expires* where the error stays above it (§4). A cell can hold past its horizon, and every certificate here is about holding. A *head* cell is a small MNIST network with every weight trained, named for its output layer, a ten-way classifier or an autoencoder, variational, denoising or autoregressive head. The drift test *certifies* a cell when $\delta \leq 0.24$ and otherwise withholds it, where $\delta$ is the relative motion of $G$ and the threshold is empirical (§4.2). Numbered results the body cites are proved in the appendix under the same numbers.

**Contributions.** **A forecast before training, and its horizon** (§3, §5.1). One Krylov pass seeded with the initial residual returns a Gauss rule for the rate measure, and so the whole predicted rate curve $R(k) = -\mathrm{d}\log L/\mathrm{d}k$. Its cost grows as the square root of the steps forecast. On character transformers of $0.11$M to $118$M parameters the read lands within a median factor of $1.07$ of the measured rate in log time. In median it is as close as the best predictor that watches none of the run (Figure 3). Three seeds at each of $12$ of the $16$ sizes give the same median (App. E.8). Standard-parameterised transformers follow the read for a median $88\%$ of the loss descent, all of it from $16$M up, and longer as they grow, at a rank correlation of $0.974$. Under $\mu$P they follow it for $32\%$, shorter as they grow, at $-0.971$. At 1B, on two seeds, and at 3B the standard transformer follows the read for its whole descent and the $\mu$P one for under a fifth (Table 3). Pooled over both parameterisations, the horizon's rank correlation is $-0.891$ with the drift at a run's end and $0.107$ with size. Within each parameterisation it is $-0.914$ and a weak $-0.341$ with the drift at the window's end, while its correlation with size changes sign. From sixteen examples to $16{,}384$ the read's log-time error falls to $0.023$ decades at $3.2$M and $0.013$ at $20$M. From $64$ examples up, both standard networks stay within tolerance for all $3{,}000$ steps.

**An early test of the forecast** (§5.2). A tenth of the way into a run, the drift test reads the operator's relative motion $\delta = \|G(t) - G(0)\|_F / \|G(0)\|_F$. It certified $2{,}581$ of $4{,}923$ MNIST classifier heads, and every one held. It withheld all $312$ that later failed, each within tolerance until then (Figure 2). Across four families it certified $142$ of $192$ heads, all holding, and withheld all $50$ that failed. Scored at a run's end over $555$ full-batch cells in nine families, every cell at or below $0.24$ has held. On transformers the early test is the growth of $G$ along the residual. Out of sample, over $93$ runs, it withholds $49$ of the $53$ that later fail and certifies $39$ of the $40$ that hold. The drift test at a tenth certifies $50$ of those $53$. The early test decides all $13$ runs from $118$M to 3B correctly (Apps. E.5, E.6).

**Two limits of the method** (§4). Runs sharing the initial Jacobian, residual and step size, the read's only inputs, can differ without bound in where the forecast expires. In the worst case over the lazy rescaling of one network on fixed data (Chizat et al., 2019), no function of those three can predict or rank expiry (Theorem 1). Three further passes match the motion that decides expiry in the lazy limit, under the regularity the appendix states. One reaches the point the linearised run arrives at, and two evaluate $G$ there. At cross-entropy the rate is asymptotically model-free, the $1/k$ decay that Soudry et al. (2018) and Lyu & Li (2020) proved. That rate is a mean over examples rather than a spectral quantity, so the forecast declines there by design.

**Related work.** The nearest work reads the same linearised dynamics on the network about to be trained. Zancato et al. (2020) predict a pre-trained network's whole fine-tuning curve in closed form, with accuracy resting on its decay rate and on the time it takes. We score that rate curve itself, on networks trained from scratch, and turn the tangent kernel's motion into a per-run test of a forecast made before training. Fort et al. (2020) found that motion falling together with the advantage of nonlinear over linearised training. Lee et al. (2019) predicted a finite network's linearised dynamics from its tangent kernel at initialisation and compared them with trained loss curves. They tracked the kernel's relative Frobenius change, our $\delta$ at squared error, falling with width. We read the rate on transformers trained from scratch to $118$M by forming $G$ exactly. A $64$-node Krylov read matches it to within $10^{-6}$ decades at nine sizes to $20$M. We read $\delta$ early as a test, the move weather forecasting makes with model drift (Orrell et al., 2001). We find the horizon's opposite trends under two parameterisations to 3B parameters, of which lazy training (Lee et al., 2019) predicts the standard one. We prove that in the worst case over networks sharing the read, nothing computed from it predicts where linearisation fails. Asymptotic rates hold in a regime real runs do not enter (Bordelon et al., 2020; Velikanov & Yarotsky, 2024). Lazy training bounds the linear phase (Chizat et al., 2019;

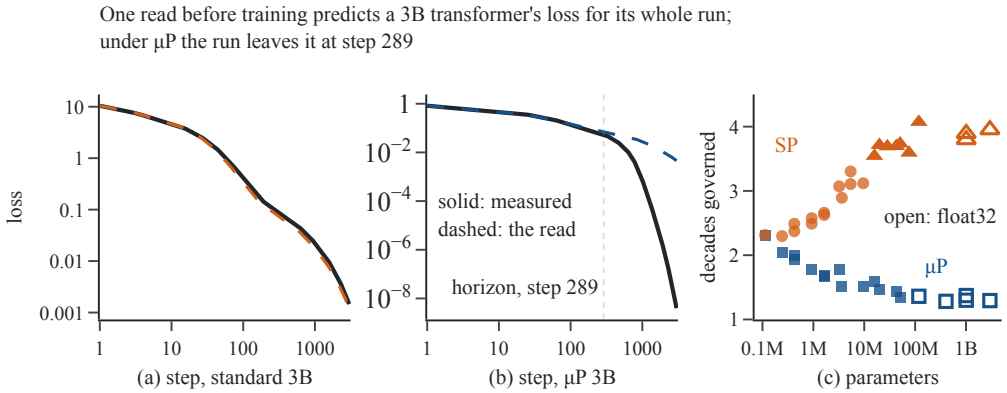



Figure 1: One read taken before the first gradient step forecasts a 3B transformer's loss over its whole run of 3.95 decades (a, measured solid, read dashed). Under $\mu$P the same network leaves its read at step 289 (b). Panel (c) plots the *decades governed*, the loss descent up to the horizon, against size.

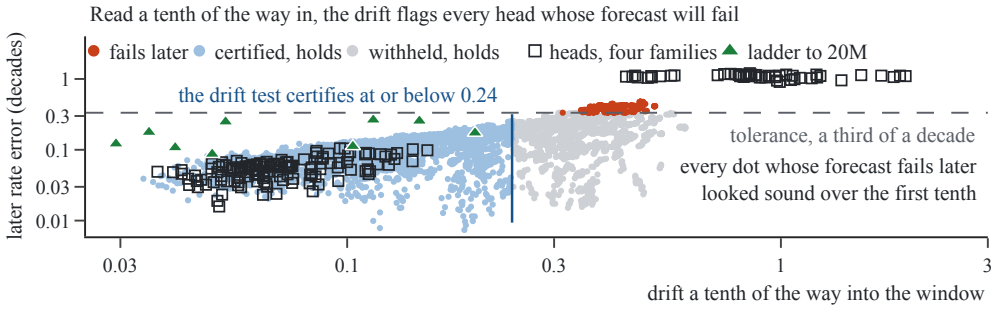



Figure 2: Every MNIST head the drift test certifies a tenth of the way in holds over the remaining nine tenths. The vertical line, 0.24, is an empirical threshold (§4.2), and the dashed one the tolerance. Dots are 4,923 MNIST classifier heads, squares four families and triangles standard transformers.

Liu et al., 2020; Boix-Adserà & Littwin, 2023). Empirical laws fit constants across completed runs (Kaplan et al., 2020; Hoffmann et al., 2022) or to a prefix of the run they predict (Domhan et al., 2015; Adriaensen et al., 2023; Mohr & van Rijn, 2024), and App. B places the rest. Concurrently, Meterez et al. (2026) seed Lanczos in parameter space. The same information sits in a Gram over *outputs*, sized by the data, and our read agrees with the parameter-space one wherever that can be formed (App. E.8).

## 2 The rate measure

Let $n$ be the number of examples, $C$ the outputs per example and $J$ the Jacobian of the outputs in the parameters. The Gauss–Newton operator of the linearised network has the non-zero eigenvalues of the $nC \times nC$ Gram $G = JJ^\top/n$. With step size $\eta$, one step contracts the $i$th eigendirection by $\mu_i = (1 - \eta\lambda_i)^2$, so the loss is a sum of geometric modes,

$$L(k) = \sum_i a_i \mu_i^k, \qquad a_i = \langle r_0, v_i \rangle^2. \tag{1}$$

The residue $a_i$ records how much of the initial residual $r_0 = f(w_0) - Y$ that direction carries, and the prefactor $1/2n$ of the loss cancels in every ratio. Equation (1) is Theorem 4.1 of Arora et al. (2019) and the Landweber filter factorisation of Strand (1974), and App. B gives its other sources. New here is scoring the rate it gives, rather than the level, on networks trained from scratch. The measure is read before the first step from $G$ formed exactly, and a residual-seeded Krylov pass matches it (§5.3). It is exact for the linearised network under full-batch gradient descent at a fixed step on squared error. The nonlinear run follows it as long as $G$ stays where it was read, and §5.2 measures how long that is. The pairs $(\lambda_i, a_i)$ fix the curve, residues as well as eigenvalues, and the method of §3 measures them. They define the *rate measure* $\rho$, which places weight $a_i$ at each $\lambda_i$. It carries more than the curve, since the loss and the weights' squared travel are two integrals against it (Prop. 36). Where (1) is exact, the residues carry the curve to 0.0044 decades against 0.041 for

the best analytic substitute, $a_i \propto \lambda_i$. On the 15 ladder cells whose spectra are stored they carry it to $0.038$ against $0.178$ (App. E.8). That advantage inverts as the operator moves further before the run is scored, and widening the network, which moves it less, restores it (App. E.8). Relabelling the targets leaves every mode and moves only the residues (Figure 12).

The rate the loss is falling at,

$$R(k) = -\frac{\mathrm{d}\log L}{\mathrm{d}k} = \frac{\sum_i a_i \mu_i^k d_i}{\sum_i a_i \mu_i^k}, \qquad d_i = -\log \mu_i, \tag{2}$$

is a Gibbs average of the per-mode rates, a weighted mean whose weights $a_i\mu_i^k$ shift toward the slow modes as training proceeds. Differentiating it,

$$\frac{\mathrm{d}R}{\mathrm{d}k} = -\operatorname{Var}_w(d) \le 0, \tag{3}$$

so $R(k)$ falls monotonically to $R_\infty = \min_{a_i>0} d_i$. Figure 11 is that measure cooling.

The read predicts the rate curve rather than the arrival time at a loss level. Arrival separates into a scale and a shape, and the scale, which carries $97.7\%$ of log arrival's variance, is what survives a nonlinear run (App. E.8). Other losses have a counterpart of (2) in the tangent kernel (Thm. 21).

## 3 Reading the rate measure

The read runs $q$ Lanczos steps on $v \mapsto JJ^\top v/n$ from the residual $r_0 = f(w_0) - Y$ of the $n$ examples at the initial weights $w_0$. The vector has length $m = nC$ in float32 or wider, and each product is one pass over the parameters with $J$ never formed. The Golub–Meurant nodes and weights of the tridiagonal are its pairs $(\lambda_i, a_i)$, and it reports $R_{\mathrm{pred}}(k)$, the rate (2) they give at the step $\eta$. It costs $q$ products of about two gradient steps each, and the nodes a forecast of $k$ steps needs grow as $\sqrt{k}$ (Hochbruck & Lubich, 1997). Our $3{,}000$-step windows sit inside the $64$-node range, and runs past $5{,}000$ steps are scored with $128$ nodes or more (App. E.9). Past that range the registered $64$-node read shows a standard run leaving before its $\mu$P twin in three pairs (App. E.7). App. F gives the exact route for small $m$.

A Krylov identity licenses the move to output space. The subspace built in parameter space from the gradient is the image of the one built in output space from the residual. The Krylov read avoids the cost of *forming* that finite-width kernel, which our from-scratch and head cells paid and Novak et al. (2022) study. We follow Cai et al. (2019) in working with the Gauss–Newton operator in this dual form. A quadrature resolves the spectrum down to its smallest node, so the read is accurate above that node and silent below it. The reference is an exactly formed Gram at $m = 2{,}560$ and condition number above $2 \times 10^6$, on the eight cells whose $m \times m$ Gram can be assembled from products. There $64$ nodes reproduce the predicted loss curve to a median relative error of $1.9 \times 10^{-6}$, worst case $3.4 \times 10^{-5}$, in $40\times$ fewer products (App. E.8).

Every exact check passes. In relative error the matrix-free Gram agrees with an explicitly formed $JJ^\top$ to $1.0 \times 10^{-15}$ on two small cells whose Jacobian can be materialised in full. An explicit Gram matches the analytic one to $2.9 \times 10^{-15}$ (App. E.8). The quadratic model reproduces the measured first-step loss drop to four significant figures in every gradient-descent cell. On $49$ cells the two standard conventions for dating a linearisation place the horizon a median factor of $23$ apart on identical data. Each reports a return to validity that its own definition creates (App. E.8). Ours is dated on the rate, whose first exit is its last on the ladders' stored grid (App. E.3).

## 4 Expiry, and the limit on predicting it

*The read expires* at the depth where its rate error $|\log_{10}(R_{\mathrm{true}}/R_{\mathrm{pred}})|$ first stays above a stated tolerance. No function of the read's inputs $(J_0, r_0, \eta)$ can predict, rank or bound from above that depth in the worst case over the lazy-training family of Theorem 1. Its members share those inputs. The depth turns on a curvature scale those inputs do not carry. It follows the operator's measured motion (§5), and three further passes match that motion in the lazy limit (Corollary 11). At cross-entropy the rate is a loss factor times a Rayleigh quotient of the tangent kernel, and the output Hessian cancels. Past separation, therefore, $kR(k) \to 1$, the $1/k$ decay that Soudry et al. (2018) proved for linear predictors and Lyu & Li (2020) for homogeneous networks. The read is then

left with one constant, which the descent drives to the harmonic-square mean of one number $\kappa_i$ per example. Each $\kappa_i$ is that example's loss decay rate divided by $\eta$ and by its squared output gradient (Thm. 37). On all $42$ cross-entropy cells the parameter-free curve $R = 1/k$ errs $0.052$ decades and the closed equation in that constant $0.055$. The read taken in $G$ errs $0.47$ (Theorem 23, App. D.4).

### 4.1 THE READ'S INPUTS CANNOT PREDICT EXPIRY IN THE WORST CASE

Motion into the read's own residual is what moves its predicted rate. Each term of the residual bound of App. D is attained along that residual, by its own construction (Theorem 31). The bound grows exponentially in depth through the kernel of the read alone. On the range the contraction of the error and the decay of the residual are the same exponential and cancel, leaving an accumulation linear in $k$. Proposition 30 proves this for a fixed kernel with its live spectrum bounded below. Off the residual, motion of any size leaves the rate unchanged. At a guaranteed drift of $2.8$ to $12.3$ the modelled rate moves by under $3 \times 10^{-15}$ nats per step, at float64 precision. So no drift-only bound can be both finite and necessary. Keeping the residual through the same chain gives the three further passes of Corollary 11.

**Theorem 1** (No function of the read's inputs predicts expiry in the worst case, proved in App. D)**.** *Consider the family $f^{(\alpha)}(v) = f(w_0) + \alpha\big[f\big(w_0 + (v - w_0)/\alpha\big) - f(w_0)\big]$, $\alpha > 0$, the lazy-training scaling in the normalisation that holds the tangent kernel fixed (Chizat et al., 2019). Its members are rescalings of one network on fixed data. In the worst case over this family, no function of the read's inputs $(J_0, r_0, \eta)$ can predict, bound from above or rank where the forecast expires. Every member has the same inputs, while the motion deciding expiry is $O(1/\alpha)$, so expiry can be delayed past any step. A predictor needs an ingredient of degree $-1$ in the model's scale, which those inputs lack. Under the regularity stated there, the three further passes of Corollary 11 evaluate second-order information, carry it and match that motion as $\alpha \to \infty$. Under the stability hypothesis $\eta\lambda_{\max}(G(s)) < 1$ at every step $s$, the operator motion the read's residual meets stays below $1/c$ of $\lambda_{\max}$ in operator norm, $c = \eta\lambda_{\max}$. This restates that hypothesis and is sharp over the admissible class. It bounds the drift rather than where the forecast expires.*

The argument is the oracle-indistinguishability device of Nemirovsky & Yudin (1983), applied to the lazy-training rescaling of Chizat et al. (2019). The rescaling leaves the outputs and the Jacobian at initialisation unchanged and shrinks the curvature, and with it their relative scale $\kappa_h$, as $1/\alpha$. Every member therefore shares one read. One network traverses its dynamics, though not its read, under a contraction of its own targets (Prop. 13). By their Theorem 2.3 that $\kappa_h$, of degree $-1$, suffices to control the departure from the linearisation. Theorem 1 adds that the degree is necessary, so no read does better in the worst case, and that three further passes attain it in the lazy limit. On trained runs the ordering is measured, and §5.1 finds that the horizon follows the operator's motion.

### 4.2 A BOUND ON THE RATE ERROR IN THE OPERATOR'S MOTION

Theorem 3 bounds the rate error by the operator's relative motion when the operators share an eigenbasis. Theorem 7 drops that hypothesis, which the scored cells violate. The travel identity below is published as the squared norm of the Landweber-filtered solution (Strand, 1974; Engl et al., 1996). It also appears inside the proof of Lemma 5.3 of Arora et al. (2019) (App. C.1). New is the measurement on real nonlinear networks and its use to scope the rate.

**Lemma 2** (Travel identity, restated from App. C.1)**.** *Let $a_i = \langle r(0), v_i\rangle^2$ and $\nu_i = 1 - \eta\lambda_i$, the residual contraction, so that the loss contraction of* (1) *is $\mu_i = \nu_i^2$. After $K$ steps of gradient descent on a map linear in $w$ over the window, $\|w_K - w_0\|^2 = \frac{1}{n}\sum_{\lambda_i>0} a_i(1 - \nu_i^K)^2/\lambda_i$.*

Predicted travel matches the distance walked to a median $0.47\%$ over 27 classifier cells at nine widths, worst case $1.30\%$. Over six from-scratch transformers it matches to a median $0.55\%$, worst $4.38\%$ (Figure 10, App. C.1).

**Theorem 3** (Rate stability under relative operator motion, proved in App. D)**.** *Let the operators along the window share the initial eigenbasis, with $\eta\lambda_i(s) \in (0, 1)$ and a fixed null space, and let $\bar\theta < 1$ be the largest relative motion of a rate over the window. With* log *natural, let $D(k)$ be the depth descended in nats and $\bar V$ the largest squared coefficient of variation of the Gibbs measure over $[(1 - \bar\theta)k, (1 + \bar\theta)k]$, which is dimensionless and differs from the variance of* (3)*. Then, in the readable form of Corollary 24,*

$$\big|\log R_{\text{true}}(k) - \log R_{\text{pred}}(k)\big| \;\le\; \frac{\bar\theta}{1 - \bar\theta}\Big[1 + 2\big(1 + \bar V\big)D(k)\Big]. \tag{18}$$

Table 1: The populations the body counts over. A cell is one network, seed and data over one window of steps; Table 4 lists every population with its full definition.

| population | cells | one cell is |
|---|---|---|
| from-scratch transformers | 20, 16 sizes to 118M | a character transformer at one size and seed, full batch on 16 sequences; with a control run, 21 curves |
| three seeds | 36, 12 sizes | the later ladder's 12 sizes, at three seeds each |
| with a loss trace | 15, 13 sizes | the 14 with a drift to 118M and the control; the level is scored on these, where the rate errs 0.038 |
| with a drift, to 403M | 19, 17 sizes | the same, with $G$ read at both ends of the window |
| $\mu$P transformers | 14, 12 sizes to 51M | a from-scratch cell with a drift, re-run under $\mu$P |
| transformers in float32 | 13, 118M to 3B | one size, seed and target scale, standard or $\mu$P |
| classifier heads, read early | 4,923, 716 configurations | a small MNIST network trained end to end with a classifier output, at one seed |
| MNIST heads | 256, 5 families | one output head at one seed, width and sample count |
| MNIST perceptrons | 64, 128 scored points | one perceptron over one window |
| released store | 126, 63 with a drift | one reading of a released checkpoint, pretrained or re-initialised |
| all scored cells | 555, 9 families | a cell carrying a drift and a holding verdict, 393 holding and 162 failing |
| each run once | 544 | the same with the 11 runs stored twice counted once, 384 holding and 160 failing |

A search finds the worst error over 906 admissible motions on each of 240 random spectra. The exact form (7) sits a median factor of 2.40 above it, and the displayed form a factor of 4.11 (App. E.8).

**The rule.** The drift test certifies a cell when the drift $\delta = \|G(t) - G(0)\|_F / \|G(0)\|_F$ that the figures plot is 0.24 or less, and otherwise withholds it. This $\delta$ is the relative kernel change that Lee et al. (2019) measure. The threshold is empirical. It was fixed before the 4,923 classifier heads of Figure 2 were run. Under its hypotheses, Theorem 3 gives a value of the same order, found after it. We solve its exact form (7) for the motion at which the bound reaches a third of a decade. The bound is evaluated after as many steps as the run takes to descend 1.8 decades of loss. The solution is $\bar{\theta} = 0.298$, a median over 50 stored spectra of separate cross-entropy runs (App. E.8). $G(0)$ is free of the loss, so they fit this squared-error theorem. That value is 0.235 once read as a drift under the theorem's shared-eigenbasis hypothesis and one more, that the motion is spread evenly over the spectrum. It is the largest of four calibrations, two windows by two tolerances, whose span $[0.06, 0.24]$ is the interval the paper quotes. One window matches the depth the read predicts and one the steps the run measured. Its lower end gates nothing, since the 7 of all 555 scored cells below it all hold, and the counts reported are this one-sided rule's. A value near a tenth had been measured before the bound was derived (App. E). Out of sample the rule has a margin. Of all 555 full-batch cells scored, the nearest that fails sits at a drift of 0.501, twice the threshold. Read early, on the 4,923 classifier heads of §5.2, trained three weeks after the threshold was fixed, the nearest failure sits at 0.314 (App. E).

**The threshold is empirical.** The scored runs lie outside the theorem's hypotheses, since their drift is mostly rotation. On the from-scratch ladder $\bar{\theta}$ is 8.2 to 46.3 times the drift and above 0.298 at every size. Theorem 3 thus covers no cell there, yet the read holds, in median, at every size. Any threshold in $\delta$ alone is empirical, since a motion keeping the read's eigenbasis sends the rate error to infinity far below 0.24 (Prop. 32). App. E.8 gives the full account and the two alternatives that bracket the threshold.

## 5 Results

The read is accurate and cheap where the operator holds still (§4.2), and the drift test finds those runs.

### 5.1 The read is accurate where the operator holds still

On standard-parameterised character transformers trained from scratch the read errs a median 0.029 decades in log time. In median it is as close as the best predictor that watches none of the run. On every cell it is also closer than a power law or the eleven-family ensemble given up to a third of the stored curve. We score 20 character transformers trained from scratch on tiny-shakespeare at squared error on one-hot targets. The data are one corpus and one family, 16 windows of 64 characters over a 65-symbol vocabulary, so 1,040 output dimensions. Training is full-batch gradient descent at the fixed step $\eta = 0.3/\lambda_{\max}(G(0))$ for 3,000 steps in float64. The runs descend 2.78 to 4.07 decades

Table 2: One read before the first gradient step beats a power law and the eleven-family ensemble of Domhan et al. (2015) on all 20 from-scratch cells. The competitors see the first tenth, fifth or third of the stored curve. Errors are in decades, in log time, with the winner in bold. Rows below the line are outside the medians.

| parameters | loss descended (decades) | read ↓ (decades), median **0.029** | eleven families at a third ↓ (decades), median 0.718 | drift |
|---|---|---|---|---|
| 0.11 M | 2.784 | **0.069** | 0.293 | 0.321 |
| 9.6 M | 3.529 | **0.024** | 0.521 | 0.074 |
| 50.6 M | 3.735 | **0.016** | 0.458 | 0.043 |
| 118 M | 4.072 | **0.016** | 0.645 | 0.040 |
| 197 M | 4.022 | **0.024** | 0.535 | 0.038 |
| 227 M | 3.876 | **0.017** | 0.351 | 0.033 |
| 340 M | 3.970 | **0.023** | 2.291 | 0.031 |

Table 3: The standard transformer follows the read for its whole run at each size and example count here; $\mu$P leaves at steps 289 to 2,030. At 1B, one $G(0)$, targets alone move the horizon from the whole run to step 183.

| horizon (decades governed) | | 118M | 403M | 1B | 3B |
|---|---|---|---|---|---|
| 16 sequences | standard | whole run (4.07) | whole run (4.04) | whole run (3.80) | whole run (3.95) |
| | $\mu$P | 364 (1.32) | 348 (1.27) | 364 (1.33) | 289 (1.27) |
| 64 sequences | standard | | | whole run (1.90) | |
| | $\mu$P | | | 749 (0.76) | |
| 256 sequences | standard | whole run (1.65) | whole run (1.65) | whole run (1.51) | |
| | $\mu$P | 2,030 (0.61) | 1,816 (0.52) | 1,813 (0.54) | |
| 16,384, SGD | standard | | | whole run (1.26) | |
| 16 sequences | targets $\times c$ | standard 1B: whole run at $c = 0.25, 0.5, 1, 2$; 660 (1.92) at 4; 183 (1.07) at 8 | | | |

of loss over 3.02 orders of magnitude of parameter count, to 118M (Table 2, Fig. 4). Fourteen carry a measured drift, 13 of them at or below 0.24, the worst erring 0.073 decades, and one above it. Among predictors that also watch nothing, the read is as close as the best. A standard scaling law fitted to every other size is 11.4× worse, and the strongest corpus-fitted law of any form is 1.50× worse at 0.044 decades. The strongest is the neighbouring size's measured curve (Fig. 3), and by held-out parameter count the read is closer on 13 of the 16. Weighted uniformly over steps, that curve is closer on 20 of the 21 (App. E.2). On MNIST the read is closer than the nearest width's measured curve on all 40 held-out cells (App. E). A *competitor* is a predictor that watches the first part of the run and extrapolates. We use two competitors. The first is the better of a power law and a constant at the prefix median. The second is the eleven-family ensemble after Domhan et al. (2015), weighted here by in-sample error rather than by their MCMC posterior. The read *wins* when it errs less than the competitor, so a cell can hold and still lose. In log time the read beats both on all 20 at the first tenth, fifth and third of the stored curve. Its median is 0.029 decades, the ensemble's 0.781 at the first tenth and 0.718 at a third, the fraction of the curve Table 2 prints (App. E.8). A pointwise median of those families errs 0.261 there. A sum of exponentials given a third, 92 steps, is closer on 9 of the 21 (Table 7). Over the 19 cells with a drift, seventeen sizes to 403M, the error tracks the drift at a Spearman 0.842, and the drift falls with size at −0.974.

**The motion of $G$ sets the horizon.** How far $G$ changes along the residual the network is still fitting sets the horizon. To first order that change, of either sign, is the network's relative excess rate over its linearisation (Prop. 38). Standard-parameterised transformers follow the read for a median 88% of a curve's loss descent, 67% to 100% by curve (Fig. 1(c)). On 14 of the 21 curves, one a bit-for-bit repeat, the horizon falls at steps 1,397 to 2,531, a median 1,521. The other 7, every curve from 16M up, stay inside tolerance for the whole window, and so do the standard runs at 1B and 3B (Table 3, App. E.5). Under $\mu$P the same transformers, identical but for the parameterisation, reverse the trend. Over the 20 distinct standard curves the horizon's rank correlation with parameter count is +0.974 in descent governed and +0.913 in steps. Over the 14 $\mu$P runs it is −0.971 and −0.899, and −0.945 in descent governed over 46 runs at three to five seeds a size. Across both ladders the horizon follows how far the operator moves. Over their 28 distinct runs its rank correlation is −0.891 with the drift and −0.953 with the relative change in $\lambda_{\max}$. With $\kappa$, the condition number of $G(0)$, it is −0.281, and with size +0.107. Within the standard ladder and within $\mu$P it is −0.914 and a weak −0.341 with the drift at the window's end, and −0.932 and −0.632 with $\lambda_{\max}$. These correlations are negative in both ladders, while those with size and $\kappa$ change sign. They are read at step 3,000, after the horizon, and the early test of §5.2 gives the same ordering in advance. Target scale moves the horizon with $G(0)$ held fixed. At 3.2M, 20M and 1B, six runs whose targets are scaled by 0.25 to 8 share $G(0)$ bit for bit. Their horizons fall from beyond the window to steps 92, 100 and 183 in the order of the scale, at a rank correlation of

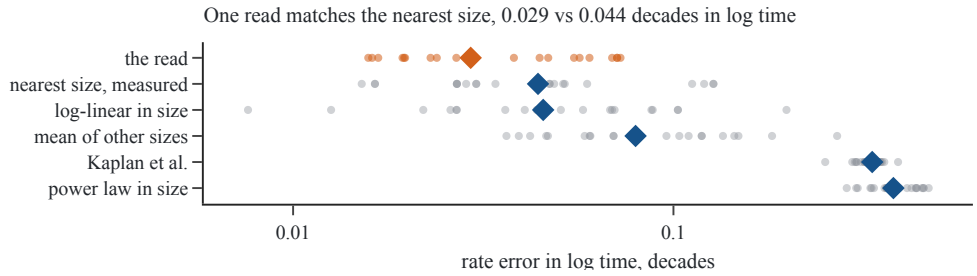



Figure 3: One read before the first gradient step matches the nearest size's measured curve, the best predictor that watches none of the run, and is closer in median than every other. Dots are curves and diamonds medians.

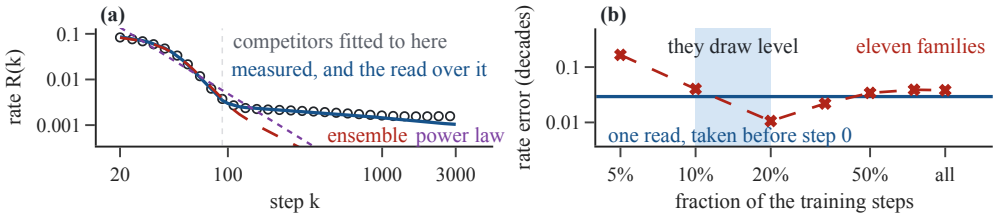



Figure 4: The share of the run a competitor watches decides the comparison. One read before the first gradient step is the better predictor at a twentieth of the steps, and the two draw level between a tenth and a fifth. (a) The 118M model. (b) Median error over 21 curves against the share of the steps the ensemble sees.

$-0.83$ or stronger (Table 3, App. E.3). Past its horizon every run of both ladders that leaves tolerance learns faster than its linearisation. The growth of $G$ along the residual accounts for that speed-up, a median $1.00$ of it over $30$ runs and $0.89$ to $1.23$ on each (App. E.3). Size and parameterisation act through that growth. At a tenth it is non-positive on all $24$ standard runs at unit target scale. Over the last third it falls with size on the $9$ standard ladder runs, at $-0.733$. Under $\mu$P on sixteen sequences it rises with size, at $+0.926$ over $50$ runs, $41$ positive. It rises with target scale at every size, $8$ of $15$ runs positive.

**The example count.** From $16$ to $16{,}384$ examples the read's log-time error falls from $0.044$ to $0.023$ decades at 3.2M and from $0.020$ to $0.013$ at 20M. From $64$ examples up every standard network stays within tolerance for $3{,}000$ steps, and from $1{,}024$ every $\mu$P one does too. The $\mu$P horizon at 20M lengthens from step $458$ at sixteen examples (App. E.7). At $256$ examples, run past the horizon, both trends hold over $24$ runs at four sizes, at $+0.941$ and $-0.789$ in descent governed. At $1{,}024$ and $4{,}096$ examples, at the resolution each window needs, the standard network follows the read longer than its $\mu$P twin in five of the six pairs (App. E.7). Its one earlier exit comes from below, as its shrinking kernel along the residual predicts. App. E.7 also scores the registered $64$-node read. Under minibatch SGD at $16{,}384$ examples the standard 1B network follows the read for its whole window (Table 3).

On released checkpoints to 6.7B parameters in five modalities, the read wins $23$ of $42$ comparisons against a power law given a third of the steps (Table 5, Fig. 6). A sign test puts this at chance. One spectrum, taken once, serves a thirtyfold range of step size (App. E.8).

### 5.2 The drift test, and the theorem that scopes the read

**The drift test taken early.** Every MNIST head certified after a tenth of the window holds over the remaining nine tenths. Over $4{,}923$ classifier heads of one family, in $716$ configurations, it certifies $2{,}581$, and all of them hold (Figure 2). It withholds all $312$ that later fail while looking sound over that first tenth. The certified heads come from $419$ configurations, none with a failing head: a one-sided $95\%$ bound of $0.71\%$ on false certification per configuration (App. E.8). Across four families it certifies $142$ of the $192$ heads, all holding, and none of the $50$ that later fail, each looking sound over that tenth. On the classifier heads the condition number at initialisation ranks the later error at $0.92$ and, cut at its first failure, certifies $4{,}395$ to the drift's $3{,}339$ (App. E.8). App. E.1 gives its boundary.

**The early test on transformers.** It reads $g = r_t^\top G(t)\, r_t / r_t^\top G(0)\, r_t - 1$, the growth of $G$ along the residual, at the same tenth for two backward passes. It measures how much faster the network descends at its current residual than its linearisation would, and past the horizon accounts for its

speed-up (§5.1). Under standard parameterisation it is $\|J_t^\top r_t\|^2/\|J_0^\top r_t\|^2 - 1$. The tenth, step 300 of 3,000, comes before the horizon of 47 of the 53 runs that later fail. Nine tenths of the window remain to score the test on. Its threshold, fixed on the 9 standard runs alone, is their largest growth, $-0.007$. Out of sample, over 93 runs to 3B, it withholds 49 of the 53 that later fail, six of which leave by step 300, where it is read. It certifies 39 of the 40 that hold. The drift test at a tenth certifies 50 of those 53 (App. E.6). The four failing runs the growth certifies are $\mu$P runs at 0.24M and 0.42M, which exceed the tolerance by 0.01 to 0.09 decades. It withholds one run that holds, $\mu$P at 20M on 64 examples, and decides all 13 runs from 118M to 3B correctly (Table 10). Certifying the standard runs at $c \leq 2$ decides them too. Over the 30 runs of the two ladders the growth ranks the horizon at $-0.87$ in descent governed. Within $\mu$P it is $-0.93$, and $-0.55$ given size (App. E.4).

**The drift test at the window's end.** Every certified cell holds, 53 cells at a drift at or below 0.24 over three disjoint populations. They are the 28 head cells and all 12 released cells certified, both drawn in full (Fig. 7), and 13 of the 14 from-scratch transformers carrying a drift. The released cells are scored over the last third of their curves, all their stores keep. The rule is sufficient rather than necessary. Pooled over a wider full-batch set than Figure 7 draws, each run counted once, it certifies 89 of the 384 that hold and withholds all 160 that fail. Holding is scored there at a quarter of a decade. App. E.8 counts the released cells.

**At cross-entropy the drift test withholds every cell.** Over 26 cells at cross-entropy on two populations every drift exceeds 0.99, so the test withholds all 26, and the read exceeds tolerance on 23 of them. Theorem 23 explains this under the hypotheses it states. Past separation the rate is asymptotically model-free (App. E.8).

**Three more passes match the motion that decides expiry.** They close the gap between the residual's drift and the read's bound, which no function of the read closes in the worst case (Thm. 1).

### 5.3 The read's cost, and the training a competitor needs to beat it

On the ladder the 64-node read costs under 5% of a run. At seven of the nine sizes timed, a power law or the ensemble must train past that cost to beat it. The ladder's cells were read by forming the full Gram from 1,040 products, as the drift test did there at both ends, and diagonalising it. On an idle card, at nine sizes from 0.11M to 20M, it costs 1,020 to 1,686 gradient steps. There the read of §3, 64 Lanczos nodes, costs 124 to 147 steps, 4.1 to 4.9% of these runs. Its rate curve matches the exact read's to within $10^{-6}$ decades. The drift test needs no Gram either: by 32 random probes at a tenth it costs as much again as the read, 157 and 161 steps at 16,384 examples (App. E.7). A competitor must instead have trained as far as the last curve point it uses (Fig. 8). Given a fifth or a third of the stored curve, 51 or 92 steps, the competitor is the cheaper. A power law or the ensemble then errs at least 2.8 times as much as the read at every size. A sum of exponentials given a third beats the read at three sizes (App. E.2). The shortest prefix at which a power law or the ensemble beats the read, always the ensemble, ends 20 to 267 steps past the read's cost at seven sizes.

## 6 Conclusion

One Krylov pass before training returns the loss curve a network would follow if it stayed linear. How far $G$ changes along the residual still being fitted sets how long a network follows it, and the drift is that change's coarse certificate. On sixteen sequences, standard-parameterised transformers follow the read for a median 88% of their loss descent, the larger the longer, and from 16M parameters to 3B for the whole window. Under $\mu$P the same transformers follow it for less, the larger the shorter. In the worst case over networks sharing the read's inputs, no function of them can predict the horizon (Thm. 1). The early tests therefore measure the motion that orders it, at about the read's cost. On MNIST classifiers the drift test certifies only heads that hold, and on transformers the growth along the residual withholds 49 of the 53 runs that later fail. At the window's end every full-batch cell certified at or below 0.24, an empirical threshold, has held. On the from-scratch ladder the read errs a median 0.029 decades in log time, as close as the best predictor that watches none of the run. It costs under 5% of a run. For a practitioner this is a forecast before training and an early test flagging failing forecasts. The read applies at squared error under full-batch and minibatch gradient descent, and the drift test under full batch. At cross-entropy the rate is model-free, and App. E.1 gives each claim with the setting it was measured in.

## REPRODUCIBILITY STATEMENT

**Outcomes fixed before measuring.** Each measurement's script states, before it runs, the outcome that would count against the claim it tests, and for a few older results only the result file states it. App. E.1 counts the firings against this paper's claims that the register holds and says what the register does not cover.

Every quantity is stored as JSON beside its log. A verification suite written separately from the experiment scripts re-reads more than eight thousand of the numbers and quoted statements in this paper from the stored JSON and fails on any mismatch; the headline median of $0.029$ decades, the four calibrations behind $[0.06, 0.24]$ and the $53$ cells the drift test certifies are among them. Predictions are scored held out, and holding is scored at both tolerances: over all scored cells, $17$ of the $555$ verdicts move between a quarter and a third of a decade, every one of them above the interval, so no verdict the rule certifies moves (App. E.8). The estimator for $\lambda^{+}_{\min}$ is validated against analytically known spectra before use and passes at every one, worst relative error $4.0 \times 10^{-8}$ at a condition number of $10^{10}$. The released code is a clean package. It implements the read, full-batch training with the compensated (Kahan) update, the scoring and both early tests, and it builds the paper's character transformers under both parameterisations.

## DATA AVAILABILITY

Three kinds of data enter. Our runs were produced by our experiment scripts and are stored as JSON, and the released package rebuilds their method and their character transformers. The corpora are public research datasets, MNIST, CIFAR-10, ImageNet-1k, tiny-shakespeare and WikiText among them. The released checkpoints, Table 5's among them, are third-party models, loaded from their public releases and not redistributed. The code is available at `https://anonymous.4open.science/r/rate-measure-iclr27`.

## CONFLICT OF INTEREST

The authors declare no competing interests.

## ETHICS STATEMENT

This work concerns numerical optimization on standard public benchmarks. We see no direct ethical concerns.

## STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

The research direction and the central objects are the authors': reading the rate curve of a nonlinear run off one eigendecomposition of the output-space Gauss–Newton operator taken before the first gradient step, the drift test that scopes where that reading holds, with its empirical threshold and the bound of Theorem 3, the horizon's opposite trends under two parameterisations, and the worst-case impossibility of predicting that expiry from the read alone with the three further passes that match the deciding motion in the lazy limit (Theorem 1, Corollary 11). The travel identity (Lemma 2) is published; its measurement on nonlinear networks and its use to scope the rate are the authors'. Large language models were used as instruments in the four ways the submission form declares. They aided and polished the writing. They served retrieval and discovery, through literature retrieval and verification of related work. They served research ideation and execution, proposing supporting experiments under the authors' direction, implementing and debugging experiment code, and executing and tabulating runs. They drafted sections of the paper, drafting prose from author-specified content. Every mathematical statement in this paper was derived and verified by the authors, and no proof here rests on a machine-generated argument taken on trust. Errors arising from that use were caught by dedicated verification passes and corrected before submission: a citation misattribution, since withdrawn; a proposed result that proved true by construction and was discarded; a dimensionally inconsistent bound; and figure annotations left stale after a numerical correction.

# REFERENCES


Steven Adriaensen, Herilalaina Rakotoarison, Samuel Müller, and Frank Hutter. Efficient Bayesian learning curve extrapolation using prior-data fitted networks. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2310.20447.

Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. doi: 10.1016/j.neunet.2020.08.022. arXiv:1710.03667.

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378, 2019. URL https://arxiv.org/abs/1810.10082.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, 2019. URL https://arxiv.org/abs/1901.08584.

E. F. Beckenbach. A class of mean value functions. *The American Mathematical Monthly*, 57(1):1–6, 1950.

Enric Boix-Adserà and Etai Littwin. Tight conditions for when the NTK approximation is valid. *Transactions on Machine Learning Research (TMLR)*, 2023. URL https://arxiv.org/abs/2305.13141. arXiv:2305.13141.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2205.09653. arXiv:2205.09653.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *ICML*, 2020. URL https://arxiv.org/abs/2002.02561.

P. S. Bullen. *Handbook of Means and Their Inequalities*. Kluwer Academic Publishers, 2003.

Tianle Cai, Ruiqi Gao, Jikai Hou, Siyu Chen, Dong Wang, Di He, Zhihua Zhang, and Liwei Wang. Gram-Gauss-Newton method: Learning overparameterized neural networks for regression problems. *arXiv preprint arXiv:1905.11675*, 2019.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1812.07956. arXiv:1812.07956.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. arXiv:2103.00065.

Felix Dangel, Lukas Tatzel, and Philipp Hennig. Vivit: Curvature access through the generalized gauss-newton's low-rank structure. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2106.02624. arXiv:2106.02624.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *ICML*, 2017. URL https://arxiv.org/abs/1703.04933.

Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *IJCAI*, 2015.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Springer Netherlands, 1996. doi: 10.1007/978-94-009-1740-8.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2010.15110.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020. doi: 10.1088/1742-5468/abc4de.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2232–2241. PMLR, 2019. URL https://arxiv.org/abs/1901.10159.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Per Christian Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4): 561–580, 1992. doi: 10.1137/1034115.

Marlis Hochbruck and Christian Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 34(5):1911–1925, 1997. doi: 10.1137/S0036142995280572.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv:2203.15556*, 2022. URL https://arxiv.org/abs/2203.15556.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018. URL https://arxiv.org/abs/1806.07572.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Learning curve prediction with Bayesian neural networks. In *ICLR*, 2017. URL https://openreview.net/forum?id=S11KBYclx.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *NeurIPS*, 2019. URL https://arxiv.org/abs/1902.06720.

D. H. Lehmer. On the compounding of certain means. *Journal of Mathematical Analysis and Applications*, 36 (1):183–200, 1971.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv:2003.02218*, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2010.01092.

Kairong Luo, Haodong Wen, Shengding Hu, Zhenbo Sun, Zhiyuan Liu, Maosong Sun, Kaifeng Lyu, and Wenguang Chen. A multi-power law for loss curve prediction across learning rate schedules. In *International Conference on Learning Representations*, pp. 39973–40017, 2025. URL https://openreview.net/forum?id=KnoS9XxIlK.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. arXiv:1906.05890.

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model. *arXiv:2607.21716*, 2026. URL https://arxiv.org/abs/2607.21716.

Felix Mohr and Jan N. van Rijn. Learning curves for decision making in supervised machine learning: a survey. *Machine Learning*, 113:8371–8425, 2024. doi: 10.1007/s10994-024-06619-7. URL https://arxiv.org/abs/2201.12150.

Arkadi S. Nemirovsky and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *ICLR*, 2020. URL https://arxiv.org/abs/1912.02803.

Roman Novak, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Fast finite width neural tangent kernel. In *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2206.08720. arXiv:2206.08720.

David Orrell, Leonard Smith, Jan Barkmeijer, and Tim N. Palmer. Model error in weather forecasting. *Nonlinear Processes in Geophysics*, 8(6):357–371, 2001. doi: 10.5194/npg-8-357-2001.

Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. What can linearized neural networks actually say about generalization? In *Advances in Neural Information Processing Systems*, volume 34, pp. 8998–9010, 2021.

Vardan Papyan. The full spectrum of deepnet Hessians at scale: Dynamics with SGD training and sample size. *arXiv:1811.07062*, 2018. URL https://arxiv.org/abs/1811.07062.

Frank Proschan. Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5(3):375–383, 1963. doi: 10.1080/00401706.2000.10485974. Reprinted in Technometrics 42(1):7–11, 2000.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. arXiv:1710.10345.

Otto Neall Strand. Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind. *SIAM Journal on Numerical Analysis*, 11(4):798–825, 1974. doi: 10.1137/0711066.

Maksim Velikanov and Dmitry Yarotsky. Explicit loss asymptotics in the gradient descent training of neural networks. In *Advances in Neural Information Processing Systems 34*, pp. 2570–2582. Curran Associates, Inc., 2021. arXiv:2105.00507.

Maksim Velikanov and Dmitry Yarotsky. Tight convergence rate bounds for optimization under power law spectral conditions. *Journal of Machine Learning Research*, 25, 2024. URL https://arxiv.org/abs/2202.00992.

Maksim Velikanov, Denis Kuznedelev, and Dmitry Yarotsky. A view of mini-batch sgd via generating functions: conditions of convergence, phase transitions, benefit from negative momenta. In *ICLR*, 2023. URL https://arxiv.org/abs/2206.11124. arXiv:2206.11124.

Martin L. Weitzman. Why the far-distant future should be discounted at its lowest possible rate. *Journal of Environmental Economics and Management*, 36(3):201–208, 1998. doi: 10.1006/jeem.1998.1052.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of the Thirty*

*Third Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673, 2020. arXiv:2002.09277.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737, 2021. arXiv:2011.14522.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. PyHessian: Neural networks through the lens of the Hessian. In *IEEE International Conference on Big Data (BigData)*, 2020. doi: 10.1109/BigData50022.2020.9378171. arXiv:1912.07145.

Zixiong Yu, Songtao Tian, and Guhan Chen. Divergence of empirical neural tangent kernel in classification problems. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2504.11130.

Luca Zancato, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Predicting training time without training. In *NeurIPS*, 2020. URL `https://arxiv.org/abs/2008.12478`.

# A THE TABLES AND FIGURES THE BODY CITES

These are the eleven tables and figures the body's claims rest on, collected here so that a reader of this document can reach every one of them. The populations table, Table 4, names the population behind every count the body gives, and App. F the protocol behind them.

Table 4: The population each count in the body is over, so that no denominator has to be rebuilt from an appendix. A *cell* is one network, seed and corpus over one window; a checkpoint's released and re-initialised reads are two *arms*, and a ladder's sizes are its *rungs*. Rows two to four pool one set of released runs three ways, and the last row pools cells from several rows above and from sweeps they do not list.

| population | size | one member is | where |
|---|---|---|---|
| released ladder | 31 rows, 27 with a drift | one released arm or scratch rung, replicates pooled to one row | Table 5 |
| released cells with a drift | 28 cells, 10 checkpoints | one of the same arms, not pooled | Fig. 7(b) |
| released store | 126 cells, 63 with a drift | every released arm scored anywhere in the paper | §5.2 |
| released cells with curves | 16 cells, 9 checkpoints | an arm whose whole rate curve is stored | – |
| scale ladder | 6 cells, 11 families | one character transformer trained from scratch at one size and seed, six of the next row's twenty, the eleven-family ensemble swept over thirteen budgets against it | Table 6 |
| from-scratch transformers | 20 cells, 14 with a drift | one character transformer trained from scratch at one size and seed, at 16 sizes from 0.11M to 118M; the 6 without a drift come from an earlier ladder that read the operator once. The 0.029 median and the wins at a tenth, fifth and third are over these 20, as is the first timing of the full-Gram route; its idle-card timing and the 64-node read's cost are at nine of their sizes, 0.11M to 20M | §5.1 |
| with its control | 21 curves, 20 distinct | the 20 and the earlier ladder's 0.93M seed-0 cell run again on the later one, repeating its curve bit for bit and adding the drift it lacked, 0.126; the laws fitted to other sizes hold out a size, so the sign test counts the 16 sizes | Fig. 1 |
| with a drift | 15 cells, 13 sizes | the 14 and that control, the cells to 118M storing a loss trace, so the level is scored on these, where the rate errs 0.038; $\bar{\theta}$ is read on the first seed at each of the 12 sizes of the 14 | §5.1 |
| with a drift, to 403M | 19 cells, 17 sizes | those 15 and one cell at each of 197M, 227M, 340M and 403M, over which the error is ranked against the drift; the 1.008B rung is not among them | §5.1 |
| heads read early | 192 cells, 4 families | a built classifier, autoencoder, variational or denoising head, its drift read after a tenth of its window | Fig. 2 |
| classifier heads read early | 4,923 cells, 716 configurations | a small MNIST network trained end to end with a classifier output, at one seed, its drift read after a tenth of its window; of the 732 configurations run at seven seeds, 201 runs reach the scored depth in fewer than the 30 steps a rate curve needs and are not scored | Fig. 2 |
| ladder read early | 9 cells, 0.11M to 20M | a from-scratch rung re-run with its drift read at step 300 | Fig. 2 |
| target-scaled transformers | 18 arms, 6 a size | one network at 3.2M, 20M or 56.9M with its one-hot targets scaled by 0.25 to 8, $G(0)$ identical bit for bit | App. E.3 |
| ladders re-run with the growth | 30 runs, 0.11M to 51M | 9 standard and 21 $\mu$P runs of the two from-scratch ladders, one to three seeds a size, re-run with the growth of $G$ along the residual recorded at every step | App. E.4 |

**Precision of every population.** Every population we trained is descended in float64 except three trained in float32. They are the runs at 256 and 16,384 examples (App. E.7), the 13 runs to 3B (App. E.5) and the runs trained for the two example-count sweeps. Its Gram is formed, diagonalised and, for the drift, differenced in float64. Four of the five rungs above 118M and the fixed-width grid are read in float32. The target-scaled arms of App. E.3 and the ladders re-run in Apps. E.4 and E.6 form their products in float32. The 13 runs to 3B are read by Lanczos from float32 products, their

Table 4: The population each count in the body is over, continued.

| population | size | one member is | where |
|---|---|---|---|
| $\mu$P ladder at three to five seeds | 46 runs, 12 sizes | a run of the maximal-update ladder at one of three to five seeds a size, 0.11M to 51M with the control, the growth of $G$ along the residual recorded at every step; the 21 $\mu$P runs of the row above are among them | App. E.6 |
| early tests out of sample | 93 runs, 0.11M to 3B | a transformer run with both early tests read at a tenth, counted once: 45 of the row above, 10 target-scaled, 19 at 256 examples, 6 $\mu$P at 64 to 4,096 examples and the 13 to 3B | App. E.6 |
| larger example counts | 52 runs, 0.42M to 1B | a rung in float32: at 16,384 examples, 3.2M or 20M, 6 runs, standard full batch or under minibatch SGD and $\mu$P under SGD, the SGD runs carried to 15,000 steps; at 256, standard or $\mu$P, 24 run until the horizon is reached, 3 at 118M and 403M for 6,000 steps, 3 at 403M and 1B for 3,000 steps and 16 more for 3,000 steps | App. E.7 |
| the example count swept | 12 runs, 3.2M and 20M | a standard rung trained full batch for 3,000 steps at one of six example counts, 16 to 16,384, three of them trained for the sweep | App. E.7 |
| the count swept under $\mu$P | 6 runs, 3.2M and 20M | a $\mu$P rung trained as in the row above, at 64, 1,024 or 4,096 examples, the 3.2M run at 4,096 on a second GPU | App. E.7 |
| maximal-update ladder | 15 cells, 13 sizes | the 15 cells with a drift re-run in $\mu$P; §5.1 scores the 14 to 51M, and with the 118M rung, a second copy of the control and two cells of the standard ladder they are 18 cells | §5.1 |
| standard and $\mu$P to 3B | 13 runs, 118M to 3B | a rung in float32, full batch on sixteen sequences for 3,000 steps: standard at 1B and 3B and $\mu$P at 118M, 403M, 1B and 3B, at one seed; at 1B both again at a second seed, and the standard network at five more target scales, $c = 0.25$ to 8 | App. E.5 |
| long-run cells | 10 scored, 14 run | one MNIST or CIFAR-10 perceptron under minibatch momentum for 60,000 to 100,000 steps; 4 CIFAR-10 cells are void because their minibatch loss leaves no rate window to fit, and the 3 that depart do so past 99 per cent of the run | App. E.8 |
| MNIST width sweep | 38 cells, 228 scored points | one perceptron at one of five widths, 16 to 256, scored over its whole curve; the 64 MNIST perceptrons below add wider cells and are scored on the last third of each curve | App. E.8 |
| width-law perceptrons | 123 runs, 10 widths | one perceptron, 104 over MNIST at widths 16 to 8,192 and 19 over CIFAR-10 at 16 to 256 | Fig. 9 |
| built heads | 256 cells, 5 families | one output head at one seed, width and sample count | Fig. 7(a) |
| MNIST perceptrons | 64 cells, 128 scored points | one perceptron over one window, read at several steps of it | §5.1 |
| the bracket, own cells | 39 cells, three ladders | one network we trained, over one window, at one size and seed | App. E.1 |
| released backbones, re-read | 18 arms, 5 backbones | one backbone under one arm on one corpus, at a fixed depth | App. E.1 |
| the ladder above 118M | 5 rungs, 4 bracketed | one character transformer at one size, width and depth moved apart | App. E.1 |
| cross-entropy cells | 42 cells | one network at cross-entropy, 18 built heads and 24 transformers, scored against $R = 1/k$ | §4 |
| with a drift | 26 cells | 8 from-scratch transformers and 18 built heads at cross-entropy | §5.2 |
| cross-entropy spectra | 50 spectra | the stored $G(0)$ of a separate cross-entropy run, where Theorem 3 is calibrated | §4.2 |
| pooled census | 555 cells, 9 families | one cell of eight populations carrying both a drift and a holding verdict, 393 holding and 162 failing | §5.2 |

drift from probes. The long-run cells' stores record no precision, and a float64 re-run reproduces their stored rate errors exactly. Every released arm is descended in float32 and its spectrum read by Lanczos in float32, or in float64 where its store says so: 28 of the released store's 126 cells, the two pooled DINO rows of Table 5 and 4 of the 16 cells with curves, while the 28 drawn cells with a drift, the 12 inside the interval and the 18 re-read arms are all float32. Two of the 126 are precision controls read in bfloat16, below the float32 the read of §3 asks for; both won and neither carries a drift, so without them the store is 124 cells, of which the read wins 79. On released cells the drift is estimated from probes and no Gram is formed, with 24 probes on 49 of the store's 63 cells with a drift, 12 on 13 and 8 on one, the products in float32 on all but 2 and their difference in float64. No producer switched off PyTorch's default that lets cuDNN run convolutions in TF32, and 18 of the store's 20 released checkpoints, all but ESM-2 and GPT-2, have a convolutional input layer. Released reads run $\max(4r, r{+}40)$ Lanczos steps at a rank $r$ chosen per cell, and their prediction uses the top $r$ modes.

## B Related work

**The decomposition and its sources.** Equation (1), $L(k) = \sum_i a_i \mu_i^k$ with $a_i = \langle r_0, v_i \rangle^2$, is Theorem 4.1 of Arora et al. (2019), which states it for a finite-width network with an explicit error term and verifies it against real training; it is the Landweber filter factorisation of Strand (1974); Engl et al. (1996), Theorem 1 of Bordelon & Pehlevan (2022) and Eq. 5 of Zancato et al. (2020), and background in Velikanov et al. (2023); and Velikanov & Yarotsky (2021) read the same residue-weighted measure as a Laplace transform. New here is scoring the rate curve that measure gives, rather than the level, on networks trained from scratch, with the measure read off the finite-width

Table 5: The table lists released checkpoints at the matched depth, with two rungs of the from-scratch ladder §5.1 scores in full, wins and losses together. It is not the whole set: five masked-autoencoder cells at this depth are absent and the read loses every one, as is a supervised backbone it takes three of six on. Counted with those eleven the read wins 23 of 42, and an exact sign test on the 40 decided cells gives $p = 0.43$; on the 29 decided of the 31 listed rows it gives $p = 0.061$, so the table is evidence of breadth of setting and makes no claim of significance. The stores hold 57 further distinct released cells at this depth, among them the re-initialised arms of the three backbones read on several corpora and further seeds and settings, and the read wins 37 of them (App. E.8). Against the eleven-family ensemble, which the table does not score, the read leads on 11 of the 16 released cells whose whole curve is stored at the first tenth of the curve and on 6 at the first fifth. Eleven of the 12 released cells inside the interval are re-initialised arms, pooled into its four rows; the twelfth is a seed of the pretrained AST arm at a drift of 0.233, whose printed row, above the interval, is another seed, at 0.264. Five more are scored outside that depth, three the read wins and two video cells it loses. The row read deeper is the VideoMAE-H/16 cell read again, not a sixth. Backbones are fine-tuned full batch on ten or sixteen examples; the two transformer rows train from scratch on their own corpus, at squared error on one-hot targets, and their drift was not measured at the depths this table scores. Median error factors over the last third of each curve, lower better, winner bold; on the released rows that is the last third of the steps, a median of 4 stored rate points, and the competitor is one power law fitted to the rate up to a third of the steps, while the two from-scratch rows take thirds of their stored points (App. E.8); under 2% is a tie; a factor $f$ is $\log_{10} f$ decades, the unit of Table 2, which takes its median over the whole of its curves, so the unit converts and the measurement does not. Transformer rungs are medians over seeds and backbone rows over corpora, so a second backbone seed is a second row. Rows are ordered by drift, which Theorem 3 bounds the error by; their rank correlation is $+0.50$ over the twenty rows with an unmarked drift and $+0.12$ over the sixteen of them below 5. Head-free and scored at squared error on their own outputs are the masked autoencoders, the protein model and the diffusion families; outside still images the backbones are video, audio and speech. Every released row but the one read deeper is scored at 1.2 decades, and the two from-scratch rows at the full depth their runs reach. Ten cells have since been taken to 2.4 and scored at both depths off one descent, so the shallow column is a prefix of the deep one and no reseeding separates them: the read is closer on 8 of the 10 at 1.2 and on 7 at 2.4, its error falls with depth on 5 and grows on 5, and the competitor's grows on 6. Two further cells reach 1.48 and 1.31 decades inside their budget and are scored at 1.2 alone. Table 2 spans 2.78 to 4.07 decades and is not scored this way. Drift is as in §4.2, blank where the second pass it needs was not run or did not finish. A dagger marks seven drifts, each measured at the depth its row is scored at (App. E.8). On the two pooled DINO rows it is the median of the drift every corpus carries, 1.40 to 20.8. On the other five it is measured by a re-run of the stored cell that reproduces its read to within 2%. All seven lie above the interval, so the drift test certifies none of them. The four blank rows run 4 to 0 against the measured rows' 16 to 9.

| cell | parameters | loss descended (decades) | read ↓ | competitor ↓ | drift |
|---|---|---|---|---|---|
| *inside the interval* (4 rows, 11 cells) | | | | | |
| VideoMAE-L/16, fresh | 304 M | 1.200 | **1.17**× | 1.36× | 0.069 |
| VideoMAE-B/16, fresh | 86 M | 1.200 | **1.44**× | 1.81× | 0.092 |
| TimeSformer-B, fresh | 121 M | 1.200 | **1.07**× | 1.16× | 0.111 |
| Diffusion transformer, fresh | 750 M | 1.200 | **1.71**× | 10.37× | 0.173 |
| *above it* (23 rows, 32 cells) | | | | | |
| WavLM-large, fresh | 316 M | 1.200 | **1.14**× | 1.44× | 0.244 |
| AST, pretrained | 85 M | 1.200 | **1.27**× | 2.11× | 0.264 |
| Masked autoencoder ViT-L, one masked patch | 330 M | 1.201 | **1.12**× | 1.62× | 0.372 |
| AST, fresh, seed 0 | 85 M | 1.200 | **2.20**× | 11.64× | 0.519 |
| Masked autoencoder ViT-B, one masked patch | 112 M | 1.200 | **1.10**× | 1.19× | 0.679 |
| DINOv3 ViT-L/16, fresh | 304 M | 1.201 | 2.36× | **1.11**× | 0.741† |
| VideoMAE-H/16, pretrained | 632 M | 1.206 | 1.47× | **1.04**× | 0.787† |
| Whisper large-v3, fresh | 1,543 M | 1.200 | **1.07**× | 2.13× | 0.80 |
| VideoMAE-H/16, pretrained, read deeper | 632 M | 2.400 | 2.24× | **1.20**× | 0.847† |
| VideoMAE-L/16, pretrained | 304 M | 1.200 | 1.23× | **1.12**× | 0.873 |
| VideoMAE-B/16, pretrained | 86 M | 1.200 | **1.10**× | 1.50× | 0.874 |
| TimeSformer-B, pretrained | 121 M | 1.200 | 1.25× | 1.25× | 0.944 |

output-space Gram before the first gradient step, formed exactly on every from-scratch and head cell and matched by a residual-seeded Krylov pass wherever both exist (§3).

**Predicting loss from other runs, and from a prefix of this one.** Two literatures predict loss and neither reads the run in front of it: laws fitted across completed runs (Kaplan et al., 2020; Hoffmann et al., 2022), extended across learning-rate schedules (Luo et al., 2025), and extrapolation from a prefix of the run (Domhan et al., 2015; Klein et al., 2017; Adriaensen et al., 2023).

**The length of the linear phase.** Lazy training says a wide network stays near its linearisation (Lee et al., 2019; Jacot et al., 2018), and Chizat et al. (2019) trace that to a choice of scaling rather than to width itself; Geiger et al. (2020) and Woodworth et al. (2020) map lazy and feature regimes, large steps leave the linear phase by a catapult (Lewkowycz et al., 2020) or train at the edge of stability (Cohen et al., 2021), and Ortiz-Jiménez et al. (2021) ask what linearisation says of generalisation.

Table 5: Released checkpoints at the matched depth, continued.

| cell | parameters | loss descended (decades) | read ↓ | competitor ↓ | drift |
|---|---|---|---|---|---|
| *above it*, continued | | | | | |
| AST, fresh, seed 1 | 85 M | 1.200 | **3.17**× | 13.92× | 1.112 |
| Whisper large-v3, pretrained | 1,543 M | 1.200 | **1.10**× | 1.19× | 1.69 |
| Protein model | 650 M | 1.204 | **4.69**× | 6.88× | 2.915 |
| DINOv3-B, six corpora | 86 M | 1.200 | **1.30**× | 2.22× | 3.80† |
| WavLM-large, pretrained | 316 M | 1.200 | 3.38× | **2.96**× | 4.33† |
| DINOv3 ViT-L/16, pretrained | 304 M | 1.200 | **1.26**× | 1.76× | 5.74† |
| Rectified-flow diffusion, fresh | 2,243 M | 1.211 | 9.67× | **4.14**× | 6.79 |
| DINOv2-B, five corpora | 86 M | 1.200 | 2.00× | **1.55**× | 13.3† |
| Masked autoencoder ViT-B, own mask ratio | 112 M | 1.200 | 1.59× | 1.61× | 17.8 |
| Rectified-flow diffusion, pretrained | 2,243 M | 1.204 | 4.52× | **1.07**× | 37.7 |
| Diffusion transformer, pretrained | 750 M | 1.200 | 25.28× | **1.76**× | 91.5 |
| *drift unmeasured* (4 rows) | | | | | |
| character transformer | 0.93 M | 3.66 | **2.19**× | 22.4× | |
| character transformer | 75.8 M | 3.59 | **1.17**× | 33.6× | |
| DINOv3 ViT-7B/16, fresh | 6,720 M | 1.201 | **1.01**× | 1.70× | |
| DINOv3 ViT-7B/16, pretrained | 6,720 M | 1.200 | **3.81**× | 4.74× | |

Table 6: The ensemble first errs less than the read within the first fifth of the steps. On these six cells that is at half the measured curve, 7.8% of the steps, and over all 21 curves between a tenth and a fifth (Fig. 4). Median error over six cells in decades of rate, lower is better; a decade error of 0.028 is a factor of 1.07 and one of 0.782 is a factor of 6.05. *The read's own error is* 0.028 *on every row*, because it is taken before the first step and never watches the run; only the competitors' budgets vary, which is the asymmetry the table exists to show. The eleven-family ensemble first beats it at half the measured curve.

| competitor's budget | eleven families ↓ | one power law ↓ | cells the read wins ↑ |
|---|---|---|---|
| 10% of the curve, 1.0% of the steps | 0.782 | 0.629 | 6/6 |
| 20% | 2.608 | 0.374 | 6/6 |
| 33% | 0.632 | 0.548 | 6/6 |
| 40% | 0.147 | 0.657 | 5/6 |
| 50% | **0.011** | 0.384 | 0/6 |
| 90% | 0.040 | 0.379 | 3/6 |

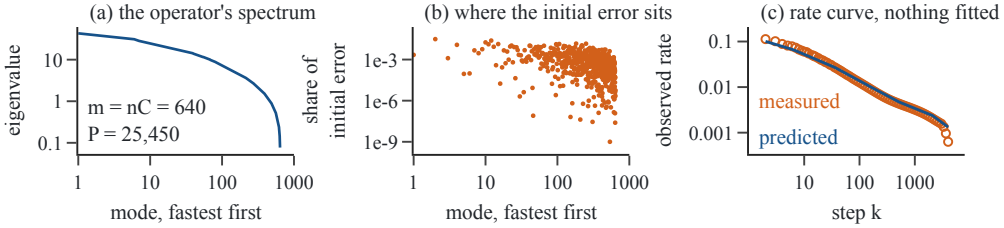



Figure 5: One read of 640 eigenvalues before the first gradient step gives an MNIST perceptron's rate curve with nothing fitted; these panels and the step size are all it uses. (a) The 640 eigenvalues of $G = JJ^\top/n$ span 2.75 decades, 42.9 to 0.077; the operator is $m = nC = 640$ square, set by the data and the class count, not by the 25,450 parameters, so the read takes the same 640 products, each one pass over the parameters, whatever the model's size. (b) The share of the initial error each eigenvector carries, which the prediction averages over. (c) Their curve over the measured rate: the two agree to a median factor of 1.11 over the 73 scored steps.

**The instrument.** Krylov spectral estimation in deep learning is standard and random-seeded (Ghorbani et al., 2019; Yao et al., 2020; Papyan, 2018), and we read Gauss–Newton eigenpairs in output space as Dangel et al. (2023) and Novak et al. (2020) do.

## C THE EXACT STATEMENTS THE INSTRUMENT RESTS ON

Everything in this appendix is derived from the loss decomposition and the operator, with nothing fitted. The measurements that appear alongside are checks on the derivations, and are labelled as such where they could not have come out otherwise.

What follows is the identity that fixes how far the weights travel.

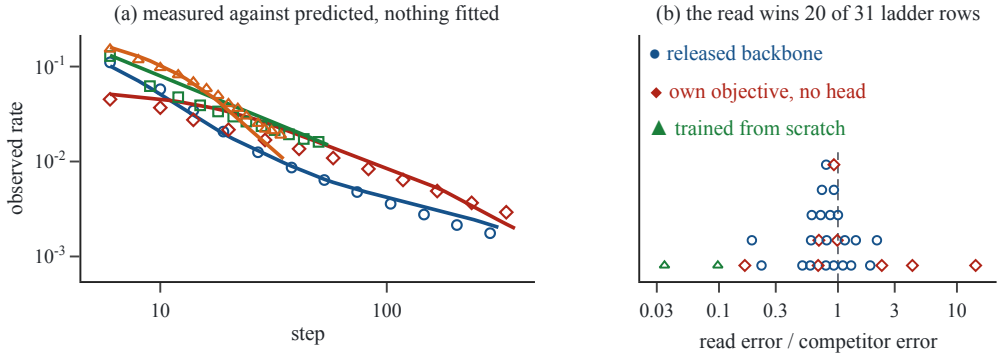



Figure 6: (a) The predicted rate laid over the measured one on four of the 16 released cells whose whole curve is stored, lines predicted and markers measured, subsampled so the line shows, nothing fitted: circles an 85M audio transformer, diamonds a 0.30B video transformer, squares an 86M video transformer, triangles its re-initialised arm. Four cells, not the ladder. A power law fitted to each of these runs recovers only the edge exponent the read already carries, never the edge that ends it. (b) Every row of Table 5 as the ratio of the read's error to the competitor's, one dot per row, stacked where two collide, parity dashed at one, losses plotted with the wins: the read wins 20 of 31, and 23 of 42 counting the eleven same-depth cells the table omits.

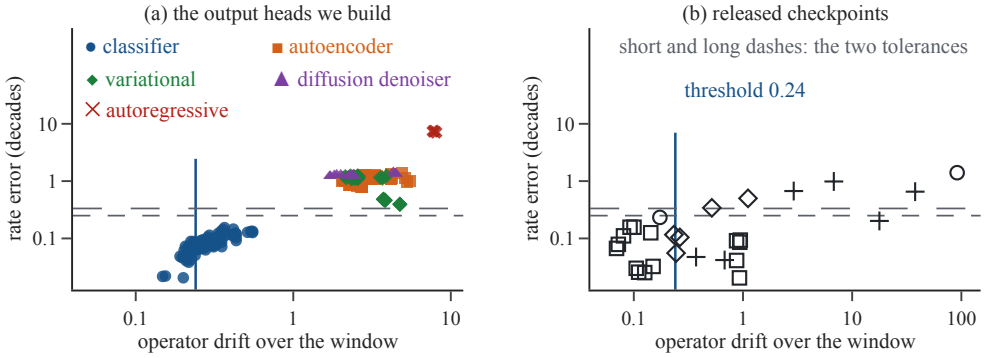



Figure 7: At or below the drift test's threshold, every cell of both panels holds its rate error under either tolerance; above it the threshold covers nothing, and what happens there differs by population: of the 228 head cells above it 111 exceed tolerance, and 6 of the 16 released cells above it do, both counts the same at either tolerance. Short and long dashes are the quarter and third of a decade, the two tolerances on the vertical axis; the solid vertical line is the threshold on the horizontal one, a drift of 0.24, read at $\delta$ under the conditions §4.2 sets out. (a) the 256 head cells this panel draws (Table 4), each seed a cell and each run drawn once within its family, two autoencoder runs recurring among the variational ones at zero regularisation weight; all five families appear above it, no setting of the depth scored or the step size puts a non-classifier family below it, and every head below it is a classifier head, swept across it by the sweep's knobs. Of these cells, 45 are further classifier heads, and seeds were compared on the rest, filed as 74 multi-seeded configurations holding 272 records, 61 of them a run filed in two sweeps: across seeds the win or loss verdict against the competitor is unanimous on 68 of the 74 and the hold or expire verdict drawn here on all 74, the six that split on winning, all near the threshold and five with seeds below it, holding at every seed. (b) the 28 released cells behind the table's 20 rows with an unmarked drift, before pooling: squares the 15 video cells, diamonds the 5 audio, circles the two epsilon-prediction diffusion cells, crosses the six own-objective cells.

### C.1 The initial spectrum fixes the weights' travel

The identity the body restates as Lemma 2 is published, and its sources are listed below, ahead of the statement.

Beyond the bounds of App. D, how far the operator moves is measured rather than derived. The distance the weights move is derived: in the regime the prediction already assumes, it is an identity in the quantities the read has already computed.

It is the squared norm of the Landweber-filtered solution (Strand, 1974; Engl et al., 1996), gradient descent on a least-squares objective being Landweber iteration. The same filtered solution is Eq. (7.40)

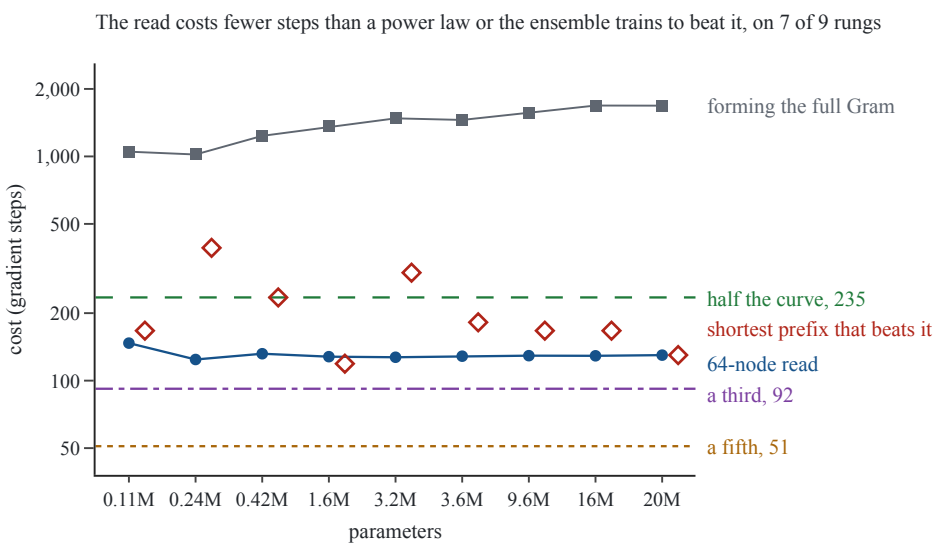



Figure 8: The read of §3 costs fewer gradient steps than a power law or the ensemble must train to beat it on seven of the nine rungs timed. Per rung, on a log scale, in gradient steps timed on the same idle card: circles the 64-node read, 124 to 147 steps; squares the exact route, forming and diagonalising the full Gram, 1,020 to 1,686; lines a competitor given a fifth, a third or half of the stored curve, 51, 92 and 235 steps; diamonds, set beside the circles, the shortest prefix at which a power law or the ensemble errs less than the read, refitted at every prefix and always the ensemble. The first two lines are cheaper than the read and there a power law or the ensemble errs at least 2.8 times as much on every rung. A sum of exponentials, the form of (1), given a third, 92 steps, errs less than the read at 0.42M, 1.6M and 3.6M (App. E.2), so counted with it the read is the cheaper at five of the nine at most. At every rung the 64-node read's rate curve matches the exact read's to within $10^{-6}$ decades. The diamonds lie 20 to 267 steps above the read on seven rungs, level with it at 20M, and 9 steps below it at 1.6M, where at step 119 the ensemble errs 0.052 decades against the read's 0.055 and loses again one prefix later. Keeping one forward pass's graph for every product brings the read to 104 to 115 steps, below every diamond.

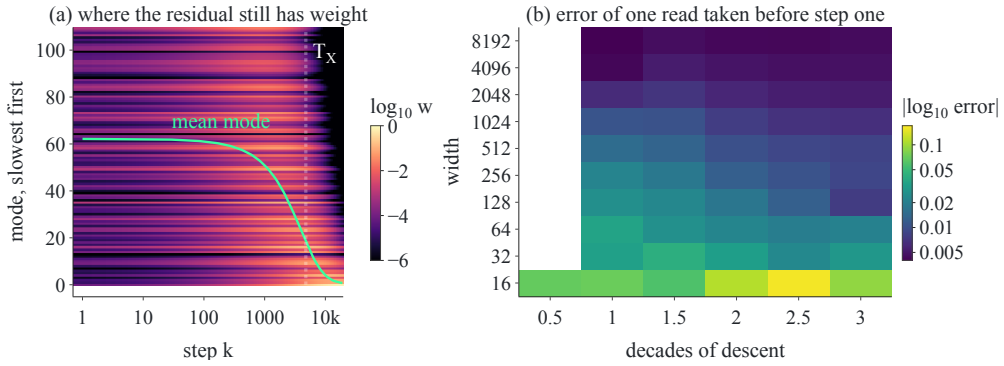



Figure 9: The prediction gets more accurate the wider the network, on perceptrons over MNIST and CIFAR-10 pooled, a trend MNIST alone separates (App. E), and the mechanism is where the residual keeps its weight. (a) The Gibbs weight field $w_i(k)$ over modes and steps: the residual begins spread across the bulk and collapses onto the slow modes, and $T_X$, the step at which the read's rate first comes within a factor $1 + X = 2$ of $R_\infty$, marks where that collapse completes. Colour is $\log w$ over six decades. (b) The error of the prediction against the measured rate, one row per width and one column per half-decade of descent, so widths are compared at matched descent. The band darkens upward, dark being accurate. The leftmost column is blank above the narrowest row because every wider network is already past 0.5 decades at its first scored step.

of Goodfellow et al. (2016), stated in discrete time from $w(0) = 0$, and its continuous-time form in this paper's own $1/n$ normalisation is Lemma 1 of Ali et al. (2019), whose introduction also gives the Landweber provenance quoted above. It is Eq. (38) of Advani et al. (2020) at zero initialisation and in expectation over the draw, that equation being an averaged squared weight norm rather than a displacement; and it appears as an exact intermediate, before the finite-$K$ factor is discarded to reach a $K$-free bound, inside the proof of Lemma 5.3 of Arora et al. (2019). Its vector form is Eq. (8) of

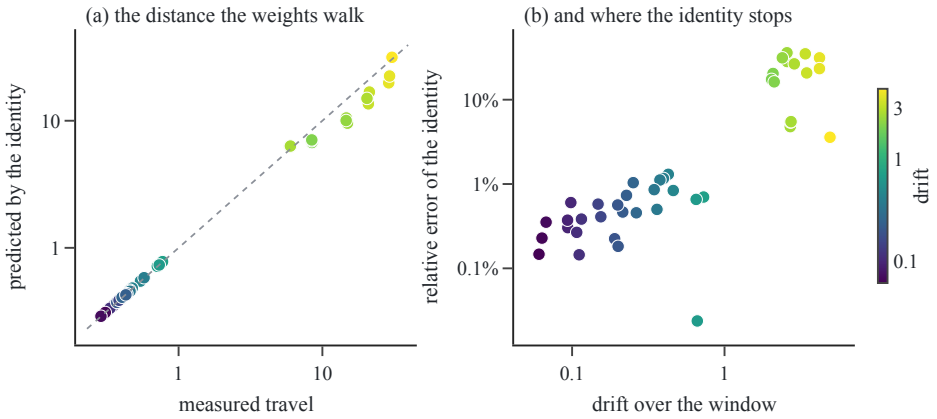



Figure 10: The distance the weights travel is fixed by the initial spectrum and its residues to $0.47\%$ wherever the operator has not moved far, and fails by $22\%$ where it has. Every cell of the width and bottleneck sweeps, coloured by that drift. (a) The distance the weights walk against the distance Lemma 2 predicts from the initial spectrum and residues alone. (b) The identity's relative error against the drift over the same window: $0.47\%$ where drift is below one and $22\%$ above it, a rank correlation of $+0.84$ with interval $[+0.72, +0.91]$, a 95 per cent percentile bootstrap over the 45 cells resampled 20,000 times, its upper end $0.91$ on all 21 bootstrap seeds tried. The two sides of one are the two sweeps, the 27 classifier cells below it, worst $1.30\%$, and the 18 bottleneck cells above it, $3.6\%$ to $36\%$: a drift of one is where the operator has moved by its own norm, and it falls in the empty gap between the sweeps' drifts, $0.73$ and $2.04$, so no cut inside that gap changes the split. The contrast is therefore between sweeps; inside the classifier sweep the error still rises with the drift, at a rank correlation of $+0.57$, and inside the bottleneck sweep, all of it past one, the drift does not order it, at $-0.11$. The identity assumes the map is linear over the window, and it fails exactly where that assumption does.

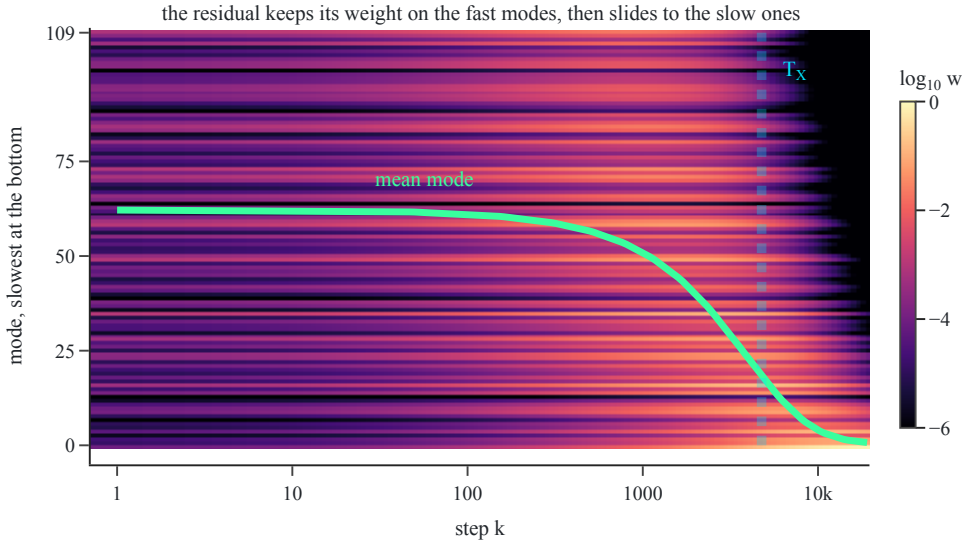



Figure 11: A loss curve is one measure cooling: the weight starts spread over 110 modes and arrives on the slowest, and both ingredients are readable before the first step. The Gibbs weight $w_i(k) \propto a_i \mu_i^k$ over modes and steps on one MNIST perceptron, colour on a log scale over six decades, with the weighted mean mode drawn through it. The residual begins spread across the bulk, and the weight slides toward the slow end until it arrives on the slowest mode, which is where $R(k)$ reaches $R_\infty$ and the transient ends. The rate and the arrival time are readings of this field, and both of its ingredients are fixed before the first step; where the forecast expires is not, since the field does not see the operator move (Theorem 1).

Lee et al. (2019) under the discretisation they state, and paired with the residual norm it is the two coordinates of an L-curve (Hansen, 1992). Two things are ours. The first is measurement. Lee et al. (2019) plot a random sample of one finite-width network's individual weights against the linearised

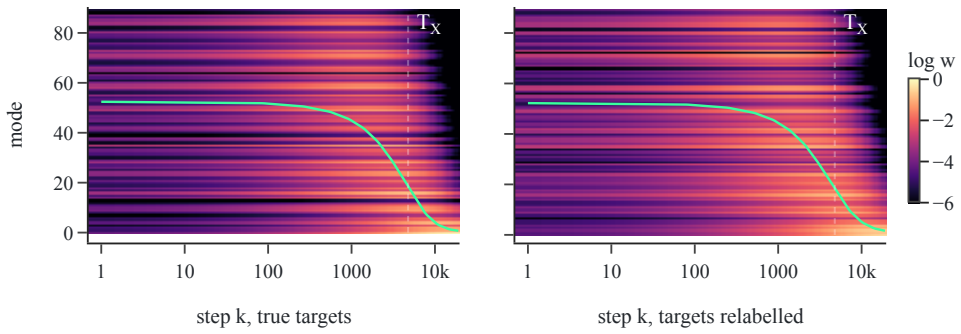



Figure 12: Relabelling the targets leaves the modes and moves only which of them the residual weighs. The Gibbs weight $w_i(k) \propto a_i \mu_i^k$ over modes and steps, centroid in green and arrival $T_X$ dotted, for one network under its true targets and the same network relabelled, on identical axes and one colour scale. The Gram never sees the labels, so the per-mode rates are identical by construction and only the residues differ: the field changes, while arrival, whose scale $1/R_\infty$ the untouched edge fixes, moves only through its shape factor, by five parts in ten thousand, $4{,}765$ steps against $4{,}768$.

dynamics; here the identity itself, the length of the whole displacement, is scored against the distance actually walked on 27 classifier cells and six from-scratch transformers, and holds to the median $0.47\%$ below. The sum runs over the live modes, as the Landweber form does over the nonzero singular values; Arora et al. (2019) and Lee et al. (2019) need no such restriction because they assume the kernel invertible. The second is the use the sum is put to: the factor $(1-\nu_i^K)^2$ is a nuisance to be discarded when the object is a $K$-free ceiling, and is the object of interest when the question is how much of that ceiling a finite run actually spends, which is what makes the second moment a scope condition on the first.

**Lemma 2** (Travel identity)**.** *Let $L = \frac{1}{2n}\|r\|^2$ with $r = f(w) - Y \in \mathbb{R}^m$, let $f$ be linear in $w$ over the window, and let $G = JJ^\top/n$ have eigenpairs $(\lambda_i, v_i)$. Write $a_i = \langle r(0), v_i\rangle^2$ and $\nu_i = 1 - \eta\lambda_i$, the residual contraction, so that the loss contraction $\mu_i$ of* (1) *is $\nu_i^2$. Then after $K$ steps of gradient descent at step size $\eta$,*

$$\|w_K - w_0\|^2 \;=\; \frac{1}{n} \sum_{\lambda_i > 0} \frac{a_i\,(1-\nu_i^K)^2}{\lambda_i},$$

*the sum running over the live modes: a mode in the kernel of $G$ has $J^\top v_i = 0$ and moves no weight.*

*Proof.* The update is $w_{k+1} - w_k = -\frac{\eta}{n} J^\top r_k$, and linearity gives $r_{k+1} = r_k - \frac{\eta}{n} JJ^\top r_k = (I - \eta G) r_k$, so $r_k = (I - \eta G)^k r(0)$ and $\langle r_k, v_i\rangle = \nu_i^k \langle r(0), v_i\rangle$. Summing, $w_K - w_0 = -\frac{\eta}{n} J^\top \sum_{k<K} (I - \eta G)^k r(0) = -\frac{1}{n} J^\top V D V^\top r(0)$ with $D = \operatorname{diag}\big((1-\nu_i^K)/\lambda_i\big)$, using $\sum_{k<K} \nu_i^k = (1-\nu_i^K)/(\eta\lambda_i)$. Then $\|w_K - w_0\|^2 = \frac{1}{n^2} r(0)^\top V D V^\top J J^\top V D V^\top r(0) = \frac{1}{n}\sum_i \lambda_i D_i^2 a_i$, which is the stated sum. □

A mode that has converged contributes $a_i/\lambda_i$; one that has not is suppressed by $(1-\nu_i^K)^2$. *The sum is therefore dominated by the smallest eigenvalue the window has actually driven, near $\eta\lambda_i K \approx 1$,* not by the largest. Writing $\mathcal{S} = \sum_{\lambda_i>0} a_i (1-\nu_i^K)^2/\lambda_i$, travel is $\sqrt{\mathcal{S}/n}$, and if the residue profile is held fixed while the spectrum is stretched then $\mathcal{S} \propto \kappa$ and the motion grows as $\kappa^{1/2}$. Measured, travel scales as $n^{0.50}$ where the identity predicts $n^{0.51}$, cell by cell to within $6.6\%$. The identity is exact for a map linear in $w$ over the window, so what measuring it reads is how far a network departs from that hypothesis. Over 27 classifier cells at nine widths from 16 to 384, scored at a matched $1.8$ decades, predicted travel matches the distance actually walked to a median relative error of $0.47\%$ and a worst case of $1.30\%$, with no width worse than $1.3\%$ and the agreement improving with width. Over six from-scratch character transformers, scored over their whole windows of $2.78$ to $3.57$ decades, it matches to a median $0.55\%$ and a worst case of $4.38\%$. The registered falsifier was $20\%$ and fired on neither population. The six classifier cells run at widths $512$ and $2048$ reach the target depth in fewer than the $30$ steps the rate scorer needs to score a curve and are excluded by it rather than by us, so their travel was not scored and the width trend above is stated over $16$ to $384$.

### C.2 THE EXACT STATEMENTS OF THE NUMBERED RESULTS THE BODY CITES

The results the body cites by number are stated here in full, each under the number the body gives it. Their proofs are in App. D, those of Theorems 21 and 23 in Apps. D.3 and D.4. Lemma 2 is proved in App. C.1 above.

**Theorem 1** (No function of the read's inputs predicts expiry in the worst case, full statement)**.** *In the setting of Theorem 7, with $f \in C^2$ near $w_0$, squared error, full-batch gradient descent at fixed $\eta$, and $G(w) = \frac{1}{n} J(w) J(w)^\top$:*

*(i)* A read-computable ceiling exists, needs no measurement, and is sharp. *Under* (H2′) *alone, $c := \eta\lambda_{\max}(G_0) < 1$ and $\eta\lambda_{\max}(G_s) < 1$ for all $s$ (Theorem 7), for every $s \le K$,*

$$\theta_s^p \;\le\; \theta_s \;=\; \frac{\|\Delta_s\|_{\mathrm{op}}}{\lambda_{\max}} \;<\; \frac{1}{c}.$$

*Over the admissible class this is sharp, and the value depends on whether $G_0$ has a kernel:*

$$\sup_s \theta_s \;=\; \max\Big\{1,\; \frac{1}{c} - \frac{\lambda_{\min}(G_0)}{\lambda_{\max}}\Big\},$$

*with $\lambda_{\min}(G_0)$ the smallest eigenvalue including zero. When $G_0$ is nonsingular the supremum is $\max\{1,\, 1/c - \lambda^+_{\min}/\lambda_{\max}\}$, since the class admits motion in both directions. Upward, $G_s = G_0 + t\, v_{\min} v_{\min}^\top$ as $\eta t \uparrow 1 - \eta\lambda^+_{\min}$ approaches $1/c - \lambda^+_{\min}/\lambda_{\max}$,* (H2′) *capping the moved eigenvalue $\lambda^+_{\min} + t$ rather than $t$; downward, retiring the top mode, $G_s = G_0 - \lambda_{\max} v_{\max} v_{\max}^\top$, stays inside the class and gives $\theta_s = 1$. When $G_0$ has a kernel, which is the case throughout this appendix, the bump goes on a null direction instead: the moved eigenvalue is $t$ itself,* (H2′) *caps it at $\eta t < 1$, and the supremum is $1/c$ exactly. Stating it as $1/c - \lambda^+_{\min}/\lambda_{\max}$ understates the ceiling on every singular operator. It bounds the drift and not the vacuity of* (8)*, which needs the strictly smaller $\theta_k < 1/c - 1$.*

*(ii)* No read-computable floor or price of the drift, and no prediction, upper bound or ordering of expiry. *For $\alpha > 0$ let*

$$f^{(\alpha)}(v) \;:=\; f(w_0) + \alpha\big[f\big(w_0 + (v - w_0)/\alpha\big) - f(w_0)\big], \qquad v_0 = w_0.$$

*This family is the lazy-training scaling of Chizat et al. (2019), written in the renormalised coordinates in which the tangent kernel stays fixed, and it is used here for the opposite purpose: not to send a model into its linearisation, but as a witness along which the read is constant while the motion it must price is not. Dinh et al. (2017) make the neighbouring point for a different statistic and a different family, that the sharpness of a minimum is not invariant under a function-preserving reparametrisation; the family here does not preserve the function, it preserves the read. Each $f^{(\alpha)}$ is a model of the same class with the* same read*: $f^{(\alpha)}(v_0) = f(w_0)$, $J^{(\alpha)}(v_0) = J_0$, $v_0 = w_0$, so every read-computable quantity is identical for every $\alpha$, bit for bit. Assume* (H5) *on a ball $B(w_0, \rho)$, in the one-point form $\|Df(w) - J_0\|_{\mathrm{op}} \le L\|w - w_0\|$ with $L = \Lambda_J \|J_0\|_{\mathrm{op}}$, and not the two-point* (H5′) *of Proposition 35. Fix $K$. Then there are $C_K$, linear in $K$, and $\alpha_0(K)$, polynomial in $K$, both free of $\alpha$, such that for every $\alpha \ge \alpha_0$ the family satisfies* (H2′) *and*

$$\sup_{s \le K} \|w_s - w_0\| \le \frac{C_K}{\alpha}, \qquad \sup_{s \le K} \theta^J(w_s) \le \frac{L C_K}{\alpha \|J_0\|_{\mathrm{op}}}, \qquad \sup_{s \le K} \theta_s^p \le \frac{3 L C_K}{\alpha \|J_0\|_{\mathrm{op}}},$$

*whence by Corollary 28 $\sup_{k \le K} |\log_{10}(R_{\mathrm{true}}(k)/R_{\mathrm{pred}}(k))| \le C'(\mathfrak{r}, K)/\alpha$. Consequently:* (a) *any read-computable $\ell$ with $\ell \le \theta_s^p$ over the family $\{f^{(\alpha)}\}$ has $\ell \le 0$; and whenever the run at $\alpha = 1$ expires at a tolerance $X$, as it does not when $f$ is its own linearisation,* (b) *no read-computable $Q$ satisfies $c_1 Q \le \theta_s^p \le c_2 Q$ over the family and* (c) *for that $X$ and every depth $D_\star$ there is an $\alpha$ whose read has not expired by depth $D_\star$, at a read identical to one that expires, so no read-computable functional predicts the expiry depth, bounds it from above or orders it over the family. The members the bootstrap controls, $\alpha \ge \alpha_0$, only delay expiry, so the family leaves a read-computable lower bound on the expiry depth open.*

*(iii)* Degree $-1$ is the admissible shape, and the probe attains it. *If $f \in C^3$ near $w_0$ then for each fixed $s$,*

$$\alpha\, \theta_s^p(\alpha) \;\longrightarrow\; \gamma_s := \frac{\big\|\frac{1}{n}\big(B[\widehat{\delta w}_s] J_0^\top + J_0 B[\widehat{\delta w}_s]^\top\big) p_s\big\|}{\lambda_{\max}\|p_s\|}, \qquad B[z] = D^2 f(w_0)[z, \cdot\,],$$

*and the probe $\hat{\theta}^p_s$ of Corollary 11 has the* same *limit. Hence any $Q$ with $\alpha Q(\alpha) \to \gamma_s > 0$ must carry an ingredient of degree $-1$ in $\alpha$, and over the family no read-computable quantity is. Its members share one read, so every such quantity is of degree $0$, a ceiling at best, as in (i), and infinitely loose in the lazy limit. When $\gamma_s > 0$, $\theta^p_s/\hat{\theta}^p_s \to 1$: the transfer constant $F$ of Corollary 11 is a finite-$\alpha$ value, not a fitted concession.*

**Theorem 3** (Rate stability under relative operator motion, full statement)**.** *Assume* (H1) *the operators $\{G(s)\}_{s\le k}$ share the eigenbasis $\{v_i\}$ of $G(0)$;* (H2) *$\eta\lambda_i(s) \in (0,1)$ for every live mode $i$ and every $s$, and the relative rate motion over the live modes obeys $\bar{\theta} < 1$ in the sense of* (16)*; a mode in the null space carries $d_i = 0$ and a constant residual, the mass $\rho(\{0\})$ of Theorem 29, and is excluded from the supremum; for that exclusion to be legitimate the null space must not move, so (H2) also asks that $\ker G(s) = \ker G(0)$ for every $s$, which is what makes a mode with $d_i(0) = 0$ carry $d_i(s) = 0$ and a residual that really is constant. Linear stability of the iteration runs to $\eta\lambda < 2$; both theorems are stated on $\eta\lambda < 1$, where $-2\log(1-\eta\lambda)$ is increasing in $\eta\lambda$, and neither says anything above it, which includes the step sizes training settles at. Let $R_{\mathrm{pred}}$ be the rate predicted from the initial read and $R_{\mathrm{true}}$ the rate of the run under the moving operator. Then for every $k > 0$,*

$$\Big|\log R_{\mathrm{true}}(k) - \log R_{\mathrm{pred}}(k)\Big| \;\le\; \underbrace{-\log(1-\bar{\theta})}_{\text{scale}} \;+\; \underbrace{\log \frac{L\big((1-\bar{\theta})k\big)}{L\big((1+\bar{\theta})k\big)}}_{\text{weight, } \mathcal{A}(k)} \;+\; \underbrace{\log \frac{R\big((1-\bar{\theta})k\big)}{R\big((1+\bar{\theta})k\big)}}_{\text{shape, } \mathcal{B}(k)}, \tag{7}$$

*where $L$ and $R$ on the right are the predicted curves, evaluated at real arguments. Every term is computable from the initial spectrum and $\bar{\theta}$ alone, with no fitted constant.*

**Theorem 7** (Rate stability without a shared eigenbasis, direction resolved)**.** *Let $r_{k+1} = (I - \eta G_k)r_k$ be the residual under the moving operator and $p_{k+1} = (I - \eta G_0)p_k$, $p_0 = r_0$, the prediction from the read, with $M = I - \eta G_0$ and $\Delta_k = G_k - G_0$. Assume* (H2′) *$c := \eta\lambda_{\max}(G_0) < 1$ and $\eta\lambda_{\max}(G_k) < 1$ for every $k$. Let $\alpha_k = \eta\|\Delta_k r_k\|/\|Mr_k\|$, $\varepsilon_k = \|r_k - p_k\|/\|p_k\|$, let $\varsigma_k$ be the contraction dispersion of the predicted residual and $\Phi_k = e^{-R_{\mathrm{pred}}(k)}$, and $N_k = 2\varepsilon_k\varsigma_k + c(2-c)\varepsilon_k^2$. Then, whenever $\alpha_k < 1$ and $N_k < \Phi_k$,*

$$\big|R_{\mathrm{true}}(k) - R_M(r_k)\big| \;\le\; -2\log(1-\alpha_k) \;\le\; -2\log\Big(1 - \frac{c\,\theta_k}{1-c}\Big), \tag{8}$$

$$\big|R_M(r_k) - R_{\mathrm{pred}}(k)\big| \;\le\; \frac{N_k}{\Phi_k - N_k}, \tag{9}$$

$$\varepsilon_k \;\le\; c\sum_{s<k} \theta^p_s\, e^{(D(k)-D(s))/2}, \tag{10}$$

*where $\theta_k = \|\Delta_k\|_{\mathrm{op}}/\lambda_{\max}$, $\theta^p_s = \|\Delta_s p_s\|/(\lambda_{\max}\|p_s\|) \le \theta_s$ and $D$ is the predicted depth in nats; $|R_{\mathrm{true}}(k) - R_{\mathrm{pred}}(k)|$ is bounded by the sum of* (8) *and* (9)*, and* (9) *is at most $-2\log(1-c)$ in every case. The outer bound in* (8) *is finite exactly when $c\,\theta_k < 1 - c$ and is read as $+\infty$ otherwise.*

**Corollary 11** (The probe is the same chain with the residual kept)**.** *Let $\hat{\theta}^p_K = \|(G(w_0 + \widehat{\delta w}_K) - G_0)p_K\|/(\lambda_{\max}\|p_K\|)$, two passes at the probe point. Then $\hat{\theta}^p_K \le (1 + \Lambda_J \hat{t}_K)^2 - 1$ by (i) and* (21)*, and $\Upsilon_K \ge 1$ because the run's first step is exactly the linearised one, $w_1 - w_0 = \widehat{\delta w}_1$, so the probe sits below the a-priori profile at every probe depth and differs from it exactly by the residual weighting that (iv) says no operator norm recovers. The transfer to the run is not a hypothesis but a consequence, at the cost of one strengthening, and Proposition 35 states it. The multiplicative form $\theta^p_K \le F\hat{\theta}^p_K$ is not proved, and no step of this paper assumes it; its tightest measured value is $F = 2.01$ up to expiry, over the $47$ probe steps before it on $12$ cells, and the proposition replaces it.*

**Proposition 13** (The witness family is a training configuration)**.** *Let $f$, $w_0$, $Y$ and $\eta$ be as in Theorem 1, let $r_0 = f(w_0) - Y$, and for $\alpha > 0$ put $Y_\alpha := f(w_0) - r_0/\alpha$, the targets contracted toward the network's own initial prediction. Let $(w^\alpha_s)_{s\ge 0}$ be full-batch gradient descent on $\frac{1}{2n}\|f(w) - Y_\alpha\|^2$ from $w^\alpha_0 = w_0$ at step size $\eta$, and let $(v_s)_{s\ge 0}$ be the run of part (ii), gradient descent on $f^{(\alpha)}$ at the same step size from $v_0 = w_0$ against the unchanged targets $Y$. Then for every $s$,*

$$w^\alpha_s = w_0 + \frac{v_s - w_0}{\alpha}, \qquad f^{(\alpha)}(v_s) - Y = \alpha\big(f(w^\alpha_s) - Y_\alpha\big), \qquad G^{(\alpha)}(v_s) = G(w^\alpha_s),$$

*so that $\theta^J(w_s)$, $\theta_s$, $\theta^p_s$, $R_{\mathrm{true}}$, $R_{\mathrm{pred}}$ and the expiry depth which Theorem 1 attributes to $f^{(\alpha)}$ are those of the run $(w^\alpha_s)$. One architecture at one initialisation, one step size and one corpus therefore*

*traverses the family, the only thing varying along it being the scale of the targets about $f(w_0)$. The read is $(J_0, r_0, \eta)$ for every model $f^{(\alpha)}$ and $(J_0, r_0/\alpha, \eta)$ for every run $(w_s^\alpha)$, and $R(\cdot)$, $D(\cdot)$, $w_i(\cdot)$, $\Phi_k$, $\varsigma_k$, $\Sigma_k$ and $\kappa$ are unmoved by the second, being homogeneous of degree 0 in $r_0$. The runs do not share a read, however: along them $\|r_0\|/\alpha$ is read-computable and of degree $-1$. Theorem 1 ranges over models sharing $(J_0, r_0, \eta)$ on fixed data. It does not exclude such a quantity ordering expiry along a scaling of the targets, as c orders the arms of App. E.3.*

**Theorem 21** (The rate is a loss factor times a Rayleigh quotient of the tangent kernel)**.** *Let $z = f(\theta) \in \mathbb{R}^{nC}$ be the outputs, $\ell$ a differentiable loss of $z$, $g = \nabla_z \ell$, $J = \partial z/\partial\theta$, and $\mathcal{K} = \frac{1}{n} J J^\top$ the tangent kernel. Then under $\dot\theta = -\eta \nabla_\theta \ell$,*

$$R \;=\; -\frac{\mathrm{d}\log \ell}{\mathrm{d}k} \;=\; \eta n\, \chi\, \varrho_{\mathcal{K}}, \qquad \chi = \frac{\|g\|^2}{\ell}, \qquad \varrho_{\mathcal{K}} = \frac{\langle g, \mathcal{K} g\rangle}{\|g\|^2}, \tag{11}$$

*and consequently*

$$\lambda_{\min}(\mathcal{K}) \;\le\; \varrho_{\mathcal{K}} \;\le\; \lambda_{\max}(\mathcal{K}) \tag{12}$$

*whatever the loss is and however far it has descended. The loss enters $\varrho_{\mathcal{K}}$ only through the direction of $g$; the output Hessian $H = \nabla_z^2 \ell$ does not enter* (11) *at all.* (11) *is the gradient-flow form and* (2) *the discrete one this paper measures in; the two agree to first order in the step size, and the cells here run at a step size where that difference is not negligible, which §D prices. Moreover $\chi \equiv 2$ identically if and only if $\ell = \frac{1}{2}\operatorname{dist}(z, S)^2$ for a closed set $S$, so $\chi \equiv 2$ at the squared error $\frac{1}{2}\|r\|^2$, and $\chi \equiv 2/n$ under the mean $\frac{1}{2n}\|r\|^2$ every cell here is trained on, the characterisation holding up to that constant; among losses with a unique zero the squared error is the only one. For softmax cross-entropy with label $y$,*

$$(1 + \tfrac{1}{C-1})(1 - p_y)^2 \;\le\; \|g\|^2 \;\le\; 2(1 - p_y)^2, \qquad \ell = \log(1/p_y),$$

*so $\chi$ has the shape $(1 - p_y)^2/\log(1/p_y)$ up to a constant in $[1 + \frac{1}{C-1}, 2]$, and it vanishes as the margins grow.*

**Theorem 23** (The descent selects the harmonic-square mean)**.** *At softmax cross-entropy $\|u^i\|^2 = c_i \varepsilon_i^2$ with $c_i$ in $[1 + \frac{1}{C-1}, 2]$. Suppose the $\kappa_i$ settle and the $c_i$ agree at a common c. Since $\ell_i = -\log(1 - \varepsilon_i)$ the chain rule gives*

$$\mathrm{d}\varepsilon_i/\mathrm{d}k = -\eta c\, \kappa_i \varepsilon_i^2 (1 - \varepsilon_i), \tag{13}$$

*separable with implicit solution $-\varepsilon_i^{-1} + \log\big(\varepsilon_i/(1-\varepsilon_i)\big) = -\eta c \kappa_i k + C_i$. Past separation $1 - \varepsilon_i \to 1$ and* (13) *becomes $\mathrm{d}\varepsilon_i/\mathrm{d}k = -\eta c \kappa_i \varepsilon_i^2$ with $\varepsilon_i(k) = \big(\eta c \kappa_i k + \varepsilon_i(0)^{-1}\big)^{-1}$, which is the regime the limit below is taken in. That closed form holds only in that limit: it drops the factor $(1 - \varepsilon_i)$, which is harmless as $\varepsilon_i \to 0$ and is a factor of $5.3 \times 10^4$ at the $\varepsilon$ a language model carries at initialisation. Integrated from one read to the next on twelve cells,* (13) *is closer than the form it replaces on all twelve, at a median 0.98 decades against 4.20. In either regime the weights are $w_i \propto \kappa_i^{-2}$, and*

$$\varrho_{\mathcal{K}} \;\longrightarrow\; \mathrm{HM}_2(\kappa) \;:=\; \frac{\sum_i \kappa_i^{-1}}{\sum_i \kappa_i^{-2}}. \tag{14}$$

*The shape of* (14) *is classical. A population of independent exponential decays with rates $\kappa_i$, reweighted by its own survivors, has a mixture rate that falls to the slowest member: Proschan (1963) proved it for failure rates and Weitzman (1998) rediscovered it for discount rates, and in both the weights move the same way ours do. Ours is the identification of the limit as a specific Lehmer mean and the strict bound below the harmonic mean it implies.* $\mathrm{HM}_2$ *is the Lehmer mean of order $-1$, $\mathcal{L}_p(x) = \sum_i x_i^p / \sum_i x_i^{p-1}$ at $p = -1$ (Beckenbach, 1950; Lehmer, 1971; Bullen, 2003), and $\mathcal{L}_p$ is increasing in p, so*

$$\mathcal{L}_{-1}(\kappa) \;\le\; \mathrm{HM}(\kappa) \;\le\; \mathrm{AM}(\kappa), \tag{15}$$

*with equality throughout only when the $\kappa_i$ coincide: the descent selects the smallest of the classical means of the per-example numbers, below even their harmonic mean, and the whole travel it asks of $\varrho_{\mathcal{K}}$ spans both gaps of* (15)*, from the arithmetic mean down through the harmonic mean to $\mathcal{L}_{-1}$. At step zero, centred logits make every $\varepsilon_i$ equal and every $w_i$ exactly $1/n$, so $\varrho_{\mathcal{K}}$ begins at the arithmetic mean; measured at initialisation the weights depart from $1/n$ by at most one per cent.* $\mathrm{HM}_2$ *is homogeneous of degree one in $\kappa$, so what is left for an operator to supply is the scale of the per-example constants and nothing else about its shape.*

# D THE DRIFT BOUND AND THE NUMBERED RESULTS, PROVED

This appendix proves the numbered results the body cites. Their statements are in App. C.2; a statement given here for the first time is one the proofs use.

## D.1 THE DRIFT BOUND, PROVED

**The bound is in a relative drift statistic, and under (H1) the paper's is a lower bound on it.** Let $G(s)$, $0 \le s \le k$, be the operator along the window, $G = G(0)$. The paper's drift is $\delta = \|G(k) - G(0)\|_F / \|G(0)\|_F$. The bound is in a different statistic,

$$\bar{\theta} \;=\; \sup_{i,\; s\le k} \Big|\frac{d_i(s)}{d_i(0)} - 1\Big|, \qquad \bar{\sigma} \;=\; \sup_{i,\; s\le k} \Big|\frac{\lambda_i(s)}{\lambda_i(0)} - 1\Big|, \tag{16}$$

the largest *relative* motion of any rate, and of any eigenvalue. The two are related by the elasticity: if $\eta\lambda_{\max}(1+\bar{\sigma}) < 1$, so that no mode is driven across the pole of $\chi$ at $\eta\lambda = 1$, then, since $\chi$ is increasing, integrating $\mathrm{d}\log d/\mathrm{d}\log\lambda = \chi$ over a relative motion of at most $\bar{\sigma}$ gives $\bar{\theta} \le (1+\bar{\sigma})^{\chi(\eta\lambda_{\max}(1+\bar{\sigma}))} - 1$, which is $\bar{\sigma}\chi$ to first order. The relation to $\delta$ is an identity when the motion preserves the eigenbasis: writing $\sigma_i = \lambda_i(k)/\lambda_i(0) - 1$,

$$\delta^2 = \frac{\sum_i \sigma_i^2 \lambda_i^2}{\sum_i \lambda_i^2}, \tag{17}$$

so $\delta$ is the $\lambda^2$-weighted root-mean-square of the same relative motions that $\bar{\sigma}$ takes the supremum of. Hence $\delta \le \bar{\sigma}$, with equality exactly when every live mode ends the window at the largest relative motion any mode reaches in it, $|\sigma_i| = \bar{\sigma}$ for every $i$ with $\lambda_i(0) > 0$: a uniform rescaling $G(k) = (1+\sigma)G(0)$ reached without overshoot is one such motion, and a spectrum whose modes move by $\pm\bar{\sigma}$ with mixed signs is another. *$\delta$ is a lower bound on $\bar{\sigma}$, never an upper one*, and identifying the two is a hypothesis about the shape of the realised motion, (H3): $\bar{\sigma} = \delta$, which we do not prove. The second link of the chain is elementary and holds wherever (H2) does. On $\eta\lambda \in (0,1)$ the map $\lambda \mapsto d = -2\log(1-\eta\lambda)$ is increasing with elasticity $\chi = \mathrm{d}\log d/\mathrm{d}\log\lambda > 1$, so a relative eigenvalue change of size $|\sigma_i|$ moves that mode's rate by at least $|\sigma_i|$ in relative terms; taking suprema over modes and over the window gives $\bar{\sigma} \le \bar{\theta}$. Together with the identity above this is the chain $\delta \le \bar{\sigma} \le \bar{\theta}$ quoted in §4.2, the first inequality holding under (H1) with (H3) as its equality case and the second under (H2). The reverse direction, $\bar{\theta} \le (1+\bar{\sigma})^{\chi} - 1$, is the bound the theorem below uses and is a different statement.

**Corollary 24** (Readable form)**.** *Under the hypotheses of Theorem 3, with* $\bar{V} = \sup_{(1-\bar{\theta})k \le t \le (1+\bar{\theta})k} V(t)$*,*

$$\big|\log R_{\mathrm{true}}(k) - \log R_{\mathrm{pred}}(k)\big| \;\le\; \frac{\bar{\theta}}{1-\bar{\theta}}\Big[\,1 + 2\big(1+\bar{V}\big)D(k)\Big]. \tag{18}$$

*Proof.* $\mathcal{A}(k) = \int_{(1-\bar{\theta})k}^{(1+\bar{\theta})k} R \le 2\bar{\theta}kR((1-\bar{\theta})k)$ and $\mathcal{B}(k) = \int_{(1-\bar{\theta})k}^{(1+\bar{\theta})k} RV \le 2\bar{\theta}kR((1-\bar{\theta})k)\bar{V}$, both by monotonicity of $R$ and $-\mathrm{d}\log R/\mathrm{d}t = RV$. Putting $t = (1-\bar{\theta})k$, $kR(t) = tR(t)/(1-\bar{\theta}) \le D(t)/(1-\bar{\theta}) \le D(k)/(1-\bar{\theta})$. Finally $-\log(1-\bar{\theta}) \le \bar{\theta}/(1-\bar{\theta})$. □

*Theorem 3 in full is stated in App. C.2; its proof follows.*

*Proof.* By (H1) the residual stays diagonal in $\{v_i\}$, so with $a_i = \langle r(0), v_i\rangle^2$ fixed, as in (1), the prefactor $1/2n$ cancelling from every ratio below, and $D_i(k) = \int_0^k d_i(s)\,\mathrm{d}s$,

$$L_{\mathrm{true}}(k) = \sum_i a_i e^{-D_i(k)}, \qquad R_{\mathrm{true}}(k) = \frac{\sum_i a_i d_i(k) e^{-D_i(k)}}{\sum_i a_i e^{-D_i(k)}}.$$

By (H2), $d_i(s) \in [(1-\bar{\theta})d_i, (1+\bar{\theta})d_i]$ for every $s$, hence $D_i(k) \in [(1-\bar{\theta})d_i k, (1+\bar{\theta})d_i k]$. Writing $N(t) = \sum_i a_i d_i e^{-d_i t} = L(t)R(t)$ for the predicted numerator, each of the four extreme substitutions is monotone in the exponent, so

$$N_{\mathrm{true}}(k) \le (1+\bar{\theta})N\big((1-\bar{\theta})k\big), \qquad L_{\mathrm{true}}(k) \ge L\big((1+\bar{\theta})k\big),$$

and symmetrically $N_{\text{true}}(k) \geq (1-\bar{\theta})N((1+\bar{\theta})k)$, $L_{\text{true}}(k) \leq L((1-\bar{\theta})k)$. Dividing,

$$(1-\bar{\theta})\,R\big((1+\bar{\theta})k\big)\frac{L((1+\bar{\theta})k)}{L((1-\bar{\theta})k)} \;\leq\; R_{\text{true}}(k) \;\leq\; (1+\bar{\theta})\,R\big((1-\bar{\theta})k\big)\frac{L((1-\bar{\theta})k)}{L((1+\bar{\theta})k)}.$$

Take logarithms and subtract $\log R(k)$. Since $R$ is non-increasing by (3), $\log R((1+\bar{\theta})k) \leq \log R(k) \leq \log R((1-\bar{\theta})k)$, so both deviations are at most $\mathcal{B}(k)$; the loss ratio contributes $\pm\mathcal{A}(k)$ with $\mathcal{A} \geq 0$; and $|\log(1 \pm \bar{\theta})| \leq -\log(1-\bar{\theta})$. Collecting gives (7). Only the lower bound on $R_{\text{true}}$, and with it the scale term, uses $\bar{\theta} < 1$; the upper bound holds for every $\bar{\theta}$. □

The following statement drops (H1) anyway. It is stated in discrete time, for the per-step rate $-\log\big(L(k+1)/L(k)\big)$ that the runs are scored in, and in a drift statistic that is an operator norm rather than a Frobenius norm, which is what the proof needs and what a few matrix-vector products return. Throughout what follows, $r_{k+1} = (I - \eta G_k)r_k$ is the residual under the moving operator and $p_k = M^k r_0$ the read's prediction, with $M = I - \eta G_0$, $\Delta_k = G_k - G_0$, $c = \eta\lambda_{\max}(G_0)$, $\varepsilon_k = \|r_k - p_k\|/\|p_k\|$ and $\Phi_k = e^{-R_{\text{pred}}(k)}$, and $R_M(x) = -\log\big(\|Mx\|^2/\|x\|^2\big)$ is the per-step rate the read assigns to a residual $x$, so that $R_{\text{pred}}(k) = R_M(p_k)$; Theorem 7 collects this setting.

**Proposition 25** (Operator term at the residual, exact)**.** *In the setting of Theorem 7, for any $r \neq 0$ let $y = Mr$, $z = \eta\Delta_k r$, $q_k(r) = \langle y, z\rangle/\|y\|^2$ and $\alpha_k(r) = \|z\|/\|y\|$. Then*

$$R_{\text{true}}(k) - R_M(r_k) \;=\; -\log\big(1 - 2q_k(r_k) + \alpha_k(r_k)^2\big),$$

*and, with $\theta_k^r = \|\Delta_k r_k\|/(\lambda_{\max}\|r_k\|) \leq \theta_k$,*

$$|q_k| \;\leq\; \alpha_k \;\leq\; \frac{c\,\theta_k^r}{1-c} \;\leq\; \frac{c\,\theta_k}{1-c}, \qquad \big|R_{\text{true}}(k) - R_M(r_k)\big| \;\leq\; -2\log\big(1-\alpha_k\big) \quad (\alpha_k < 1).$$

*Proof.* $(I - \eta G_k)r = y - z$ and $\|y - z\|^2 = \|y\|^2 - 2\langle y, z\rangle + \|z\|^2$; divide by $\|y\|^2$ and take $-\log$. Cauchy–Schwarz gives $|q_k| \leq \alpha_k$, and $\|y\| \geq (1-c)\|r\|$ gives $\alpha_k \leq \eta\|\Delta_k r\|/((1-c)\|r\|)$. Finally $(1-\alpha)^2 \leq 1 - 2q + \alpha^2 \leq (1+\alpha)^2$. □

**Proposition 26** (Residual term through the contraction dispersion)**.** *In the setting of Theorem 7, let $w_i(k) = \langle p_k, v_i\rangle^2/\|p_k\|^2$ be the Gibbs weights of the predicted residual, $\Phi_k = \sum_i w_i(k)\mu_i = e^{-R_{\text{pred}}(k)}$, and*

$$\varsigma_k^2 \;=\; \sum_i w_i(k)\,\mu_i^2 - \Phi_k^2 \;=\; \mathrm{Var}_{w(k)}(\mu),$$

*the dispersion of the loss contraction under those weights, with $\mu_i = 1$ on the kernel of $G_0$. Let $\phi_k$ be the angle between $r_k$ and $p_k$ and $N_k = \varsigma_k \sin 2\phi_k + c(2-c)\sin^2\phi_k$. Then $|e^{-R_M(r_k)} - e^{-R_{\text{pred}}(k)}| \leq N_k$, and whenever $N_k < \Phi_k$,*

$$\big|R_M(r_k) - R_{\text{pred}}(k)\big| \;\leq\; \log\frac{\Phi_k}{\Phi_k - N_k} \;\leq\; \frac{N_k}{\Phi_k - N_k}.$$

*In every case $|R_M(r_k) - R_{\text{pred}}(k)| \leq -2\log(1-c)$. Moreover $\sin\phi_k \leq \varepsilon_k$, so the statement holds with $N_k$ replaced by $2\varepsilon_k\varsigma_k + c(2-c)\varepsilon_k^2$.*

*Proof.* Let $A = M^2$, whose eigenvalues are the $\mu_i$ on the eigenvectors of $G_0$, and $\Phi(x) = x^\top Ax/\|x\|^2$, so $R_M = -\log\Phi$ and $\Phi \in [(1-c)^2, 1]$. Take unit vectors $x = p_k/\|p_k\|$ and $y = r_k/\|r_k\| = \cos\phi_k\, x + \sin\phi_k\, z$ with $z \perp x$ unit. Expanding, $\Phi(y) - \Phi(x) = \sin 2\phi_k\, x^\top Az + \sin^2\phi_k\,(z^\top Az - \Phi(x))$. Since $z \perp x$, $x^\top Az = z^\top(A - \Phi(x)I)x$, and $\|(A - \Phi(x)I)x\|^2 = \sum_i w_i(\mu_i - \Phi_k)^2 = \varsigma_k^2$; and $|z^\top Az - \Phi(x)| \leq \lambda_{\max}(A) - \lambda_{\min}(A) \leq 1 - (1-c)^2$. This gives $|\Phi(y) - \Phi(x)| \leq N_k$. Then $|\log\Phi(y) - \log\Phi(x)| \leq |\Phi(y) - \Phi(x)|/\min(\Phi(x), \Phi(y))$ with $\Phi(y) \geq \Phi_k - N_k$, and $-\log(1-t) \leq t/(1-t)$. The cap is the range of $R_M$. Finally $\sin\phi_k = \mathrm{dist}(p_k, \mathrm{span}\, r_k)/\|p_k\| \leq \|p_k - r_k\|/\|p_k\|$. □

**Proposition 27** (Residual discrepancy, forced at the predicted residual)**.** *Under* (H2′), $e_{k+1} = (I - \eta G_k)e_k - \eta\Delta_k p_k$, *hence*

$$\|e_k\| \;\leq\; \eta\sum_{s<k}\|\Delta_s p_s\|, \qquad \varepsilon_k \;\leq\; c\sum_{s<k}\theta_s^p\, e^{(D(k)-D(s))/2} \;\leq\; c\,\bar{\theta}^p\,\Sigma_k,$$

*with $\theta_s^p = \|\Delta_s p_s\|/(\lambda_{\max}\|p_s\|) \leq \theta_s$ the drift weighted by the predicted residual, $\bar{\theta}^p$ any bound on $\theta_s^p$ for $s < k$, and $\Sigma_k = \sum_{s<k} e^{(D(k)-D(s))/2}$ the amplification functional of the read.*

*Proof.* $e_{k+1} = (I - \eta G_k)(p_k + e_k) - Mp_k = (I - \eta G_k)e_k - \eta \Delta_k p_k$. Under (H2′) the eigenvalues of $I - \eta G_k$ lie in $(0, 1]$, so $\|e_{k+1}\| \le \|e_k\| + \eta\|\Delta_k p_k\|$ and $e_0 = 0$. Then $\eta\|\Delta_s p_s\| = c\,\theta_s^p \|p_s\|$ and $\|p_s\|/\|p_k\| = e^{(D(k)-D(s))/2}$. □

*Proof.* Propositions 25, 26 and 27; the outer bound in (8) is $\theta_k^r \le \theta_k$. □

**Corollary 28** (Shape of the bound in depth and drift)**.** *Let $R(s) = D(s+1) - D(s) = \int_s^{s+1} R(t)\,\mathrm{d}t$ be the predicted per-step rate, with $R(t)$ the continuous rate of* (2)*, so that it is non-increasing by* (3)*, and positive, some live mode being assumed to carry residue. Then*

$$\frac{e^{D(k)/2} - 1}{1 - e^{-R(0)/2}} \;\le\; \Sigma_k = \sum_{s<k} e^{(D(k)-D(s))/2} \;\le\; \frac{e^{D(k)/2} - 1}{1 - e^{-R(k-1)/2}},$$

*so that, under the hypotheses of Theorem 7 with $\theta_s^p \le \bar\theta^p$ for $s < k$ and $c\bar\theta^p \Sigma_k < 1$,*

$$\big|R_{\mathrm{true}}(k) - R_{\mathrm{pred}}(k)\big| \;\le\; \frac{2c\,\theta_k^r}{(1-c) - c\,\theta_k^r} \;+\; \frac{2\varsigma_k\, c\bar\theta^p\, \Sigma_k + c(2-c)\,(c\bar\theta^p \Sigma_k)^2}{\Phi_k - 2\varsigma_k c\bar\theta^p \Sigma_k - c(2-c)(c\bar\theta^p \Sigma_k)^2},$$

*whenever both denominators are positive, the first being so exactly when $c\,\theta_k^r < 1 - c$. At fixed depth the right side is linear in the drift to leading order; at fixed drift the second term grows with depth as $\varsigma_k\, \bar\theta^p\, e^{D(k)/2}/R$, with $R$ between the first and the current predicted rate; and the theorem constrains the error only on the side $c\bar\theta^p \Sigma_k < 1$ of the line*

$$\bar\theta^p_{\mathrm{vac}}(k) \;=\; \frac{1}{c\,\Sigma_k} \;\ge\; \frac{1 - e^{-R(k-1)/2}}{c\,(e^{D(k)/2} - 1)},$$

*which falls exponentially in the depth. Beyond it the residual term is the trivial cap $-2\log(1-c)$.*

*Proof.* Write $u_s = D(k) - D(s)$, so $u_s - u_{s+1} = R(s)$ and $\int_{u_{s+1}}^{u_s} e^{u/2}\,\mathrm{d}u = 2e^{u_s/2}(1 - e^{-R(s)/2})$. Hence $\Sigma_k = \sum_{s<k}\big[2(1 - e^{-R(s)/2})\big]^{-1} \int_{u_{s+1}}^{u_s} e^{u/2}\,\mathrm{d}u$, and since $R(k-1) \le R(s) \le R(0)$ and $\int_0^{D(k)} e^{u/2}\mathrm{d}u = 2(e^{D(k)/2} - 1)$, the two-sided estimate follows. The bound on the error is Proposition 25 with $-2\log(1-x) \le 2x/(1-x)$, Proposition 26 in its $\varepsilon$ form, and Proposition 27. □

**Theorem 29** (Uniform transient bound)**.** *Let $L$ be given by* (1) *with $a_i > 0$, with $\eta\lambda_i \in (0, 1)$ fixed on the live modes, so that $R_i$ is decreasing in $\lambda_i$ and the slowest live mode is $\lambda^+_{\min}$, and with $\mu_i = 1$ on the kernel of $G$, which is what carries $\rho(\{0\})$. Let $\rho$ put positive mass on $(0, \infty)$, and let $XR_\infty < R(0) - R_\infty$ so that $T_X > 0$. Then $\Gamma(k) := \log\big(L(k)/a_1\big) + R_\infty k$ satisfies*

$$R(k) - R_\infty = -\Gamma'(k) \qquad \text{for all } k \ge 0, \tag{19}$$

*$\Gamma$ is convex and non-increasing with $\Gamma(\infty) = \log(1 + \rho(\{0\})) \ge 0$, strictly decreasing where $\rho$ has mass on $(0, \infty)$, and $T_X$ is the unique solution of $-\Gamma'(T) = XR_\infty$. Moreover*

$$T_X \le \frac{\Gamma(0)}{XR_\infty}, \qquad \text{equivalently} \qquad \tau := T_X R_\infty \le \frac{\Gamma(0)}{X}, \qquad \Gamma(0) = \log\frac{L(0)}{a_1}. \tag{20}$$

**Proposition 30** (The depth-exponential lives on the kernel)**.** *In the setting of Corollary 28, suppose in addition that $\ker G_s = \ker G_0$ for $s < k$, so that $\operatorname{ran} G_0$ is invariant along the window, and that the moved operator keeps a floor on its live spectrum, $\lambda^+_{\min}(G_s) \ge \lambda^+_{\min}(G_0)$ for $s < k$, which is what lets one contraction stand for the whole product. Split $\Delta_s p_s$ into its components in $\ker G_0$ and in $\operatorname{ran} G_0$, of relative sizes $\theta_s^{p,\ker}$ and $\theta_s^{p,\mathrm{ran}}$. Write $\nu_{\min} = 1 - \eta\lambda^+_{\min} = e^{-R_{\mathrm{live}}/2}$, where $R_{\mathrm{live}}$ is the rate of the slowest live mode and is not $R_\infty$ when the kernel carries residue. Then for every pair $s < k$,*

$$D(k) - D(s) \;\le\; R_\infty(k - s) + \log\big(1 + S(0)\big),$$

$$\nu_{\min}^{k-s-1} e^{(D(k)-D(s))/2} \;\le\; e^{R_\infty/2}\sqrt{\tfrac{1+S(s)}{1+S(k)}} \;\le\; e^{R_\infty/2}\sqrt{1 + S(0)},$$

*the left inequality an equality exactly when the kernel carries no residue, where $R_\infty = R_{\mathrm{live}}$. Hence the amplification accumulated on the range is bounded by a constant free of $k$ and $s$, and*

$$\varepsilon_k \;\le\; c\sum_{s<k} \theta_s^{p,\ker}\, e^{(D(k)-D(s))/2} \;+\; c\sum_{s<k} \theta_s^{p,\mathrm{ran}}\, \nu_{\min}^{k-s-1} e^{(D(k)-D(s))/2},$$

*whose second sum is at most linear in $k$. The vacuity line of Corollary 28 therefore falls exponentially in depth only through $\ker G_0$; on $\operatorname{ran} G_0$ it falls as $1/k$.*

*Proof.* Both statements are the factorisation of Theorem 29 read twice. From $L(k) = a_1 e^{-R_\infty k}(1 + S(k))$, $D(k) - D(s) = \log(L(s)/L(k)) = R_\infty(k-s) + \log\big((1+S(s))/(1+S(k))\big)$, and $S$ is non-increasing and non-negative, which gives the first display with the ratio at most $1 + S(0)$. Multiplying $e^{(D(k)-D(s))/2}$ by $\nu_{\min}^{k-s-1} = e^{-R_{\text{live}}(k-s-1)/2}$ and using $R_\infty = \min_i R_i \le R_{\text{live}}$ cancels the $e^{R_\infty(k-s)/2}$ and leaves the stated constant: the contraction of the error on the range and the decay of $\|p_k\|$ are the same exponential. On $\ker G_0$ the iteration matrix $I - \eta G_0$ is the identity, nothing contracts, and the factor survives. The split of the sum is then the triangle inequality applied inside Proposition 27. ■ □

**Theorem 31** (The bound is attained along the residual and empty off it)**.** *In the setting of Theorem 7 let $m$ be the dimension of the output space, $\zeta = \|G_0\|_F/\lambda_{\max} \in [1, \sqrt{m}\,]$, $\hat{p}_s = p_s/\|p_s\|$, and $\nu_{\min} = 1 - \eta\lambda^+_{\min}$ the contraction of the slowest live mode.*

*(i)* Motion off the residual path is free. *If $\Delta_s p_s = 0$ for every $s \le k$ then $r_s = p_s$ and $R_{\text{true}}(s) = R_{\text{pred}}(s)$ for every $s \le k$, whatever $\|\Delta_s\|_F$. In particular, with $\Pi_k^\perp$ the projector onto the orthogonal complement of $\operatorname{span}\{p_0, \dots, p_k\}$ and any $t \in [0, \lambda_{\max}(1/c - 1))$, the fixed operator $G_s = G_0 + t\,\Pi_k^\perp$ satisfies* (H2′)*, leaves the rate error at zero through step $k$, and has drift $\delta = t\sqrt{m - \dim\operatorname{span}\{p_0,\dots,p_k\}}\,/\|G_0\|_F$, which as $t \to \lambda_{\max}(1/c-1)$ is at least $(1-c)\sqrt{m-k-1}/(c\,\zeta)$.*

*(ii)* The operator term is attained. *For $r \neq 0$ and $\theta > 0$ with $c(1+\theta) < 1$, let $\hat{y} = Mr/\|Mr\|$. The operator $G_0 + \theta\lambda_{\max}\hat{y}\hat{y}^\top$ satisfies* (H2′)*, has $\theta_k = \theta$, and gives $R_{\text{true}} - R_M(r) = -2\log(1-\alpha)$ exactly, with $\alpha = c\,\theta\,\langle\hat{y}, r\rangle/\|Mr\|$; when $r$ is the top eigenvector of $G_0$, $\alpha = c\theta/(1-c)$ and both inequalities of* (8) *are equalities.*

*(iii)* The dispersion is the exact first-order sensitivity. *For the unit vector $x = \hat{p}_k$, $A = M^2$ and any unit $z \perp x$,*

$$\frac{\mathrm{d}}{\mathrm{d}\phi} R_M(\cos\phi\, x + \sin\phi\, z)\Big|_{\phi=0} \;=\; -\frac{2\langle x, Az\rangle}{\Phi_k}, \qquad \sup_{z\perp x,\ \|z\|=1} |\langle x, Az\rangle| \;=\; \varsigma_k,$$

*attained at $z \propto (A - \Phi_k I)x$. No bound $|R_M(r_k) - R_{\text{pred}}(k)| \le C\sin\phi_k + O(\phi_k^2)$ holds with $C < 2\varsigma_k/\Phi_k$, which is the first-order coefficient of Proposition 26.*

*(iv)* The amplification is attained. *Let $p_s$ lie in the range of $G_0$, let $u$ be a unit vector orthogonal to $p_s$, $\omega_s = \lambda_{\max}\,\hat{p}_s^\top G_0^+ \hat{p}_s \ge 1$, and $0 < t \le 1/\omega_s$ with $c(1+2t) < 1$. The motion that acts at step $s$ alone,*

$$G_s = G_0 - t\lambda_{\max}\,\hat{p}_s\hat{p}_s^\top + t\lambda_{\max}\,(u + \hat{p}_s)(u+\hat{p}_s)^\top, \qquad G_{s'} = G_0 \;\; (s' \neq s),$$

*satisfies* (H2′)*, has $\theta_s^p = t$ and $\theta_{s'}^p = 0$ otherwise, and gives for every $k > s$*

$$\varepsilon_k \;=\; c\,t\,e^{(D(k)-D(s))/2}\,\|M^{k-s-1}u\|, \qquad \|M^{k-s-1}u\| \le 1,$$

*with equality on the right for $k = s+1$, and for $k > s+1$ exactly when $u \in \ker G_0$, and with $\|M^{k-s-1}u\| \le \nu_{\min}^{k-s-1} = e^{-(k-s-1)d_{\min}/2}$ when $u$ lies in the range, attained at the slowest live eigenvector when that eigenvector is orthogonal to $p_s$, as the construction requires of $u$, and otherwise by it orthogonalised against $p_s$, short of the bound by a relative amount of second order in $\langle\hat{p}_s, v_{\min}\rangle$. For $u \in \ker G_0$,* (10) *is an equality and $R_M(r_k) = -\log\big((\Phi_k + \varepsilon_k^2)/(1+\varepsilon_k^2)\big)$, which is $R_{\text{pred}}(k)/(1+\varepsilon_k^2)$ to first order in the rate: one rotation of the residual halves the read's rate at a depth of $2\log(1/(ct))$ nats past the step it entered.*

*Proof in outline.* (i) If $r_s = p_s$ then $r_{s+1} = (I - \eta G_0)p_s - \eta\Delta_s p_s = p_{s+1}$, so the per-step rates agree through step $k$; the fixed operator is $G_0$ plus a non-negative multiple of a projector, positive semidefinite with $\lambda_{\max} \le \lambda_{\max} + t < 1/\eta$, and $\|t\Pi_k^\perp\|_F = t\sqrt{\operatorname{rank}\Pi_k^\perp}$ with $\operatorname{rank}\Pi_k^\perp \ge m - k - 1$. (ii) $z = \eta\Delta r = c\theta\langle\hat{y}, r\rangle\hat{y}$ is a positive multiple of $Mr$, so $q = \alpha$ in Proposition 25 and $1 - 2q + \alpha^2 = (1-\alpha)^2$; for the top eigenvector $Mr = (1-c)r$ and $\alpha = c\theta/(1-c)$. (iii) Along a unit $z \perp x$ the derivative of $\Phi$ at $x$ is $2\langle(A - \Phi_k I)x, z\rangle$, whose supremum is $\varsigma_k$ by Cauchy–Schwarz. (iv) The rank-one motion has $\theta_s^p = t$ and is admissible for $t \le 1/\omega_s$ with $c(1+2t) < 1$; by Proposition 27, $e_{s+1} = -ct\|p_s\|u$ and $e_k = M^{k-s-1}e_{s+1}$, and the eigenvalues of $M$, in $[1-c, \nu_{\min}]$ on the live modes and 1 on the kernel, give the bounds. □

**Proposition 32** (No worst-case bound in $\delta$ alone)**.** *Let $G$ be $m \times m$ with eigenpairs $(\lambda_i, v_i)$ and let $\tau = \delta\lambda_{\max}/\sqrt{m}$. The perturbation $\Delta = -\sum_{i:\lambda_i \le \tau} \lambda_i v_i v_i^\top$ satisfies $\|\Delta\|_F \le \sqrt{m}\,\tau = \delta\lambda_{\max} \le \delta\|G\|_F$, so it is admissible at drift $\delta$, and $G + \Delta$ annihilates every eigenvalue below $\tau$. Writing $\mathcal{I}$ for that set and assuming $\sum_{i\in\mathcal{I}} a_i > 0$,*

$$R_{\text{true}}(k) = \frac{\sum_{i\notin\mathcal{I}} a_i d_i e^{-d_i k}}{\sum_{i\notin\mathcal{I}} a_i e^{-d_i k} + \sum_{i\in\mathcal{I}} a_i} \;\xrightarrow[k\to\infty]{}\; 0,$$

*so $\sup_{\|\Delta\|_F \le \delta\|G\|_F} |\log R_{\text{true}} - \log R_{\text{pred}}| = +\infty$ for every $\delta \ge \lambda^+_{\min}\sqrt{m}/\lambda_{\max}$, and no function $F(\delta, \textit{spectrum}, k)$ can bound the error uniformly at the drifts this paper works at.*

**Lemma 33** (The travel identity fixes the displacement, not only its length)**.** *Let $\widehat{\delta w}_K = -\frac{\eta}{n} J_0^\top \sum_{s<K} p_s$. Then $\widehat{\delta w}_K$ is exactly the displacement the linearised run makes in $K$ steps, and $\|\widehat{\delta w}_K\| = \hat{t}_K$ with $\hat{t}_K^2 = \frac{1}{n}\sum_{\lambda_i>0} a_i(1-\nu_i^K)^2/\lambda_i$, the quantity of Lemma 2. It costs one pass through $J_0^\top$: $\sum_{s<K} p_s$ is a polynomial in $G_0$ applied to $r_0$, which the Lanczos tridiagonal already carries.*

*Proof.* The linearised iteration is $w_{s+1} - w_s = -\frac{\eta}{n} J_0^\top p_s$; summing gives the vector. Then $\|\widehat{\delta w}_K\|^2 = \frac{\eta^2}{n}\big(\sum_{s<K} p_s\big)^\top G_0 \big(\sum_{s<K} p_s\big)$, and in the eigenbasis the $i$th coordinate of $\sum_{s<K} p_s$ is $\langle r_0, v_i\rangle(1-\nu_i^K)/(\eta\lambda_i)$ for $\lambda_i > 0$, which gives $\frac{1}{n}\sum_{\lambda_i>0} a_i(1-\nu_i^K)^2/\lambda_i$; a mode in the kernel contributes $K\langle r_0, v_i\rangle$ to the sum and nothing to the quadratic form. □

Write $\|J_0\|_{\text{op}} = \sqrt{n\lambda_{\max}}$ and, for a weight vector $w$, $\theta^J(w) = \|J(w) - J_0\|_{\text{op}}/\|J_0\|_{\text{op}}$ for the Jacobian's relative motion. The constant the statement below buys with a read is that motion per unit of the travel that produced it, measured by a finite difference of $J$ along the displacement the read predicts:

$$\Lambda_J \;=\; \max_j \; \frac{\theta^J\big(w_0 + \widehat{\delta w}_{k_j}\big)}{\hat{t}_{k_j}} \tag{21}$$

over a ladder of probe depths $k_j$ the read names for itself. It costs one backward pass per depth and a fixed number of power iterations on $(J - J_0)(J - J_0)^\top$, four passes each, independent of the number of steps; no Gram is formed and nothing is trained. Two hypotheses carry it to the run: (H5) $\theta^J(w) \le \Lambda_J\|w - w_0\|$ for every $w$ within $\sup_{s\le K}\|w_s - w_0\|$ of $w_0$, and (H6) the path inflation $\Upsilon_K = \sup_{s\le K}\|w_s - w_0\|/\hat{t}_s$ is finite, which part (ii) below relates to the two quantities the theorem is already written in.

**Theorem 34** (A priori: the read bounds the drift it will meet, and the bound does not order it)**.** *In the setting of Theorem 7, with squared error and $G(w) = \frac{1}{n}J(w)J(w)^\top$:*

- *(i)* From the Jacobian to the drift. *For every $s$, $\theta_s^p \le \|G(w_s) - G_0\|_{\text{op}}/\lambda_{\max} \le \big(1+\theta^J(w_s)\big)^2 - 1$.*
- *(ii)* The run's path leaves the predicted one at second order.

$$\big\|(w_K - w_0) - \widehat{\delta w}_K\big\| \;\le\; \frac{c}{\|J_0\|_{\text{op}}} \sum_{s<K} \Big(\theta^J(w_s)(1+\varepsilon_s) + \varepsilon_s\Big)\|p_s\|,$$

  *so the displacement the read predicts is the one the run makes, to first order in $\theta^J$ and in the residual discrepancy $\varepsilon$ of* (10)*.*

- *(iii)* The a-priori profile. *Under* (H5) *and* (H6)*, with $u = \Lambda_J\Upsilon_K$,*

$$\theta_s^p \;\le\; \hat{b}_s := (1 + u\,\hat{t}_s)^2 - 1, \qquad \varepsilon_K \;\le\; c\sum_{s<K} \hat{b}_s\, e^{(D(K)-D(s))/2},$$

  *which is $2u\,\mathcal{E}_1(K) + u^2\,\mathcal{E}_2(K)$ with $\mathcal{E}_1(K) = c\sum_{s<K}\hat{t}_s e^{(D(K)-D(s))/2}$ and $\mathcal{E}_2(K) = c\sum_{s<K}\hat{t}_s^2 e^{(D(K)-D(s))/2}$ are functionals of the read alone. That bound is increasing in $u$, so at a common $u$ the order in which the cells' bounds reach a fixed level of $\varepsilon$ is fixed by their reads and the constant sets only the depths; by (iv) that order need not be the order in which the runs themselves reach it.*

*(iv)* The relaxation in (i) is unbounded, so the profile cannot price. *For every $k$ and every $\epsilon \in \big(0,(1-c)/2c\big)$ there is a motion inside* (H2′) *with $\theta^p_s = \epsilon$ for every $s \le k$ and $\theta_s \ge 1/c - 1 - 2\epsilon$. Hence $\sup \theta_s/\theta^p_s = \infty$ over the admissible class, and no bound on $\theta^p$ that passes through an operator norm can be tight.*

*Proof.* (i) $\theta^p_s \le \|\Delta_s\|_{\mathrm{op}}/\lambda_{\max}$ by the definition of the operator norm, and $G(w) - G_0 = \frac{1}{n}\big[(J - J_0)J^\top + J_0(J-J_0)^\top\big]$ gives $\|G(w)-G_0\|_{\mathrm{op}} \le \frac{1}{n}\|J-J_0\|_{\mathrm{op}}\big(\|J\|_{\mathrm{op}} + \|J_0\|_{\mathrm{op}}\big) \le \frac{1}{n}\|J_0\|^2_{\mathrm{op}}\,\theta^J(2+\theta^J)$, using $\|J\|_{\mathrm{op}} \le \|J_0\|_{\mathrm{op}} + \|J-J_0\|_{\mathrm{op}}$ and $\|J_0\|^2_{\mathrm{op}}/n = \lambda_{\max}$. (ii) $w_{s+1} - w_s = -\frac{\eta}{n}J_s^\top r_s$ while $\widehat{\delta w}_{s+1} - \widehat{\delta w}_s = -\frac{\eta}{n}J_0^\top p_s$, so their difference is $-\frac{\eta}{n}\big[(J_s - J_0)^\top r_s + J_0^\top e_s\big]$, of norm at most $\frac{\eta}{n}\|J_0\|_{\mathrm{op}}\big(\theta^J(w_s)\|r_s\| + \|e_s\|\big)$; now $\frac{\eta}{n}\|J_0\|_{\mathrm{op}} = c/\|J_0\|_{\mathrm{op}}$, $\|r_s\| \le (1+\varepsilon_s)\|p_s\|$ and $\|e_s\| = \varepsilon_s\|p_s\|$. Sum over $s < K$, both paths starting at $w_0$. (iii) (H5) and (H6) give $\theta^J(w_s) \le \Lambda_J \Upsilon_K \hat{t}_s$; substitute in (i) and in Proposition 27, and expand $(1+u\hat{t}_s)^2 - 1 = 2u\hat{t}_s + u^2\hat{t}_s^2$. (iv) Take $\Delta_s = t\,u_s u_s^\top + \epsilon\lambda_{\max}\,\hat{p}_s\hat{p}_s^\top$ with $u_s$ a unit vector orthogonal to $p_s$, which exists at every $s$ once $m \ge 2$, and $t = \lambda_{\max}(1/c - 1 - 2\epsilon)$. It is positive semidefinite, $\eta\lambda_{\max}(G_0+\Delta_s) \le c + (1-c-2c\epsilon) + c\epsilon < 1$, $u_s^\top p_s = 0$ gives $\Delta_s p_s = \epsilon\lambda_{\max} p_s$, and $\|\Delta_s\|_{\mathrm{op}} \ge t$. A fixed projector off the whole span of $p_0,\dots,p_k$, as in Theorem 31(i), would serve only while that span is a proper subspace, and it is all of $\mathbb{R}^m$ from $k = m-1$ on whenever $r_0$ excites $m$ distinct eigenvalues. □

*Corollary 11 is stated in App. C.2; its proof follows.*

**Proposition 35** (The probe transfers to the run, additively, with a computable correction)**.** *Assume* (H5′)*, that $\|J(w) - J(w')\|_{\mathrm{op}} \le \Lambda_J\|J_0\|_{\mathrm{op}}\|w - w'\|$ for every $w, w'$ in a ball about $w_0$ containing both the run's path and the predicted one. Write $\hat{w}_K = w_0 + \widehat{\delta w}_K$ for the probe point and $\Delta_K = \|(w_K - w_0) - \widehat{\delta w}_K\|$ for the distance between it and the run's. Then*

$$\big|\theta^p_K - \hat{\theta}^p_K\big| \;\le\; \Lambda_J\,\Delta_K\,\big(2 + \theta^J(w_K) + \theta^J(\hat{w}_K)\big). \tag{22}$$

*Theorem 34*(ii) *bounds $\Delta_K$ at second order in $\theta^J$ and the residual discrepancy, and under* (H5) *and* (H6) *every factor on the right is bounded by $\Lambda_J$ and quantities the read computes, so the correction is a probe-computable number and not a fitted one.*

*Proof.* Both drifts weight the same predicted residual $p_K$, so by the triangle inequality, taken in either order, and $\|Ax\| \le \|A\|_{\mathrm{op}}\|x\|$, $\big|\|(G(w_K) - G_0)p_K\| - \|(G(\hat{w}_K) - G_0)p_K\|\big| \le \|G(w_K) - G(\hat{w}_K)\|_{\mathrm{op}}\|p_K\|$; dividing by $\lambda_{\max}\|p_K\|$ gives $|\theta^p_K - \hat{\theta}^p_K| \le \|G(w_K) - G(\hat{w}_K)\|_{\mathrm{op}}/\lambda_{\max}$. Writing $J = J(w_K)$ and $J' = J(\hat{w}_K)$, the identity used in Theorem 34(i) gives $G(w_K) - G(\hat{w}_K) = \frac{1}{n}[(J-J')J^\top + J'(J-J')^\top]$, hence $\|G(w_K) - G(\hat{w}_K)\|_{\mathrm{op}} \le \frac{1}{n}\|J - J'\|_{\mathrm{op}}(\|J\|_{\mathrm{op}} + \|J'\|_{\mathrm{op}})$. By (H5′) the first factor is at most $\Lambda_J\|J_0\|_{\mathrm{op}}\Delta_K$, and $\|J\|_{\mathrm{op}} \le \|J_0\|_{\mathrm{op}}(1 + \theta^J(w_K))$ with the same for $J'$, so the second is at most $\|J_0\|_{\mathrm{op}}(2 + \theta^J(w_K) + \theta^J(\hat{w}_K))$. Multiplying and using $\|J_0\|^2_{\mathrm{op}}/n = \lambda_{\max}$ gives (22). □

*Theorem 1 in full is stated in App. C.2; its proof follows.*

*Proof.* (i) $G_s$ and $G_0$ are positive semidefinite, so for every unit $x$, $x^\top\Delta_s x = x^\top G_s x - x^\top G_0 x \in [-\lambda_{\max}(G_0), \lambda_{\max}(G_s)]$; $\Delta_s$ is symmetric, so $\|\Delta_s\|_{\mathrm{op}} = \sup_x |x^\top\Delta_s x| \le \max\{\lambda_{\max}(G_s), \lambda_{\max}\}$. Under (H2′), $\lambda_{\max}(G_s) < 1/\eta = \lambda_{\max}/c$ and $c < 1$, so $\|\Delta_s\|_{\mathrm{op}} < \lambda_{\max}/c$. The definition of the operator norm gives $\theta^p_s \le \theta_s$. For sharpness, $G_0 + t\,v_{\min}v_{\min}^\top$ is positive semidefinite with $\lambda_{\max}(G_s) = \max\{\lambda_{\max}, \lambda^+_{\min} + t\}$, so (H2′) holds exactly for $t < 1/\eta - \lambda^+_{\min}$, and then $\theta_s = t/\lambda_{\max} \uparrow 1/c - \lambda^+_{\min}/\lambda_{\max}$.

(ii) *Invariance.* $D_v f^{(\alpha)}(v) = Df\big(w_0 + (v - w_0)/\alpha\big)$ by the chain rule, so at $v_0 = w_0$, $J^{(\alpha)}(v_0) = J_0$ and $f^{(\alpha)}(v_0) = f(w_0)$; the read is $(J_0, r_0, \eta)$ and every listed quantity is a functional of it.

*Conjugacy.* Write $u_s = v_s - v_0$ and $w_s = w_0 + u_s/\alpha$. Gradient descent on $f^{(\alpha)}$ reads $u_{s+1} = u_s - \frac{\eta}{n}Df(w_s)^\top r_s$ with $r_s = r_0 + \alpha(f(w_s) - f(w_0))$, and $w_{s+1} - w_s = -\frac{\eta}{n\alpha}Df(w_s)^\top r_s$, so with $\bar{J}_s = \int_0^1 Df(w_s + \tau(w_{s+1} - w_s))\,\mathrm{d}\tau$,

$$r_{s+1} = \big(I - \tfrac{\eta}{n}\bar{J}_s Df(w_s)^\top\big)r_s.$$

The step in $w$ carries the factor $1/\alpha$; the step in $v$ does not.

*Bootstrap.* That the weights stay within $O(1/\alpha)$ of their start over a fixed horizon is the central lazy-training estimate of Chizat et al. (2019); what follows re-derives it in discrete time for the normalisation that holds the tangent kernel fixed, tracking how the constants grow in $K$ so that (H2′) is derived and not assumed. Put $N_0 = \|J_0\|_{\mathrm{op}}$, so $\frac{\eta}{n}N_0^2 = c$, and $A = c\|r_0\|/N_0$. Assume $\|u_j\| \le 2eKA$ and $\|r_j\| \le e\|r_0\|$ for $j \le s$. Then $\|w_j - w_0\| \le 2eKA/\alpha \le \rho$ once $\alpha \ge 2eKA/\rho$, so $\beta_j := \|Df(w_j) - J_0\|_{\mathrm{op}}/N_0 \le \beta_* := 2eKAL/(\alpha N_0)$, and the same bound holds for $\bar J_j$, an average over the segment. Hence $\|u_{j+1}\| \le \|u_j\| + eA(1+\beta_*) \le \|u_j\| + 2eA$ for $\beta_* \le 1$, which closes the first half at $j = K$; and $\|I - \frac{\eta}{n}\bar J_j Df(w_j)^\top\|_{\mathrm{op}} \le \|I - \eta G_0\|_{\mathrm{op}} + c\big((1+\beta_*)^2 - 1\big) \le 1 + 3c\beta_*$, so $\|r_j\| \le e^{3c\beta_* K}\|r_0\| \le e\|r_0\|$ once $3c\beta_* K \le 1$, i.e. $\alpha \ge 6ecLK^2A/N_0$. Both halves close for $\alpha \ge \max\{1,\ 2eKA/\rho,\ 2eKAL/N_0,\ 6ecLK^2A/N_0\}$, with $C_K := 2eKA$ linear in $K$; let $\alpha_0$ be the larger of that and $6cLC_K/((1-c)N_0)$, still quadratic in $K$. Theorem 34(i) then gives $\theta_s^p \le \theta_s \le (1+\theta_s^J)^2 - 1 \le 3\theta_s^J$ for $\theta_s^J \le 1$, and since $\theta_s^J \le \beta_* = LC_K/(\alpha N_0)$, every $\alpha \ge \alpha_0$ has $\eta\lambda_{\max}(G_s^{(\alpha)}) \le c(1+\theta_s) \le (1+c)/2 < 1$, so (H2′) is derived and not assumed. Corollary 28 at $\theta = 3LC_K/(\alpha N_0)$ is $O(1/\alpha)$ in nats, and dividing by $R_{\mathrm{pred}}(K)\ln 10 > 0$, finite because $K$ was fixed first, gives the decade form.

*Consequences.* (a) $\ell \le \inf_\alpha \theta_s^p(\alpha) = 0$. (b) is (a) applied to the lower half of a two-sided price, whose upper half would then force $\theta_s^p \le 0$ on every member; but the run at $\alpha = 1$ expires, so $\theta_s^p > 0$ at some $s$, since with $\theta_s^p = 0$ throughout Proposition 27 gives $r_s = p_s$ and Proposition 25 then gives no rate error. (c) Given $X$ and $D_\star$, choose $K$ with $D(K) > D_\star$ and then $\alpha \ge \alpha_0(K)$ with $C'(\mathfrak{r}, K)/\alpha < X$: no step $k \le K$ crosses the tolerance, while at $\alpha = 1$ the same read expires at $X$ by hypothesis. A predictor, upper bound or orderer of the expiry depth that sees only the read assigns the two the same value, while the second has not expired by $D_\star$, which was arbitrary.

(iii) With $w = w_0 + z/\alpha$ and $f \in C^3$, $Df(w) - J_0 = B[z]/\alpha + O(\alpha^{-2})$, so $G(w) - G_0 = \frac{1}{n}\big(B[z]J_0^\top + J_0 B[z]^\top\big)/\alpha + O(\alpha^{-2})$. At the probe point $z = \widehat{\delta w}_s$ exactly; at the realised point $v_s - v_0 = \widehat{\delta w}_s + O(1/\alpha)$ by Theorem 34(ii) with $\theta^J, \varepsilon = O(1/\alpha)$, so $z = \widehat{\delta w}_s + O(1/\alpha)$ and the two agree at leading order. Multiplying by $\alpha$ and using that $p_s$ is read-computable gives both limits and their equality; the ratio claim needs $\gamma_s > 0$. Read-computable quantities are constant in $\alpha$ by (ii), hence of degree 0. □

*Proposition 13 is stated in App. C.2; its proof follows.*

*Proof.* The chain rule gives $D_v f^{(\alpha)}(v) = Df\big(w_0 + (v - w_0)/\alpha\big)$, as in (ii). Induct, the case $s = 0$ being $w_0^\alpha = v_0 = w_0$ and $f^{(\alpha)}(w_0) - Y = r_0 = \alpha\big(f(w_0) - Y_\alpha\big)$. Assume the first two identities at $s$ and write $w = w_s^\alpha$. Then

$$v_{s+1} = v_s - \frac{\eta}{n} Df(w)^\top\big(f^{(\alpha)}(v_s) - Y\big) = v_s - \frac{\eta\alpha}{n} Df(w)^\top\big(f(w) - Y_\alpha\big),$$

so $w_0 + (v_{s+1} - w_0)/\alpha = w - \frac{\eta}{n}Df(w)^\top\big(f(w) - Y_\alpha\big) = w_{s+1}^\alpha$, which is the first identity at $s+1$; the second follows from it, since $f^{(\alpha)}(v) - Y = \alpha\big(f(w) - f(w_0)\big) + r_0 = \alpha\big(f(w) - Y_\alpha\big)$. The third is $G^{(\alpha)}(v) = \frac{1}{n} D_v f^{(\alpha)}(v) D_v f^{(\alpha)}(v)^\top = \frac{1}{n} Df(w)Df(w)^\top = G(w)$. For the derived quantities, $G_0$, $\lambda_{\max}$, $M$ and each $\Delta_s$ are common to the two pictures; the predicted residual is $p_s = M^s r_0$ in the first and $M^s r_0/\alpha$ in the second, and $\theta_s^p = \|\Delta_s p_s\|/(\lambda_{\max}\|p_s\|)$ is homogeneous of degree 0 in $p_s$; and $\frac{1}{2n}\|f^{(\alpha)}(v_k) - Y\|^2 = \alpha^2 \cdot \frac{1}{2n}\|f(w_k^\alpha) - Y_\alpha\|^2$ makes $R_{\mathrm{true}}(k)$, a difference of logarithms of that quantity at consecutive steps, the same in both, as then is $|\log_{10}(R_{\mathrm{true}}/R_{\mathrm{pred}})|$ and the depth at which it first stays above tolerance. Finally $L(k) = \sum_i a_i(1 - \eta\lambda_i)^{2k}$ with $a_i = \langle r_0, v_i\rangle^2$ is homogeneous of degree 2 in $r_0$, so $R$, $D$, $w_i$, $\Phi$, $\varsigma$, $\Sigma$ and $\kappa$, each a ratio of terms of equal degree, are homogeneous of degree 0 in it. □

**Proposition 36** (Two moments)**.** *Under the hypotheses of Lemma 2, for every $k$ and every window length $K$,*

$$L(k) \;=\; \frac{1}{2n}\int (1 - \eta\lambda)^{2k}\,\mathrm{d}\rho(\lambda), \qquad \|w_K - w_0\|^2 \;=\; \frac{1}{n}\int_{(0,\infty)} \frac{\big(1 - (1-\eta\lambda)^K\big)^2}{\lambda}\,\mathrm{d}\rho(\lambda).$$

*Proof.* The first is $r_i(k) = \nu_i^k r_i(0)$ summed against $L = \|r\|^2/2n$; the second is Lemma 2 rewritten as an integral against $\rho$. □

### D.3 THE HESSIAN CANCELS OUT OF THE RATE

*Theorem 21 is stated in App. C.2; its proof follows.*

*Proof.* $\dot\theta = -\eta J^\top g$ gives $\dot z = J\dot\theta = -\eta J J^\top g$ and hence $\dot\ell = \langle g, \dot z\rangle = -\eta n\langle g, \mathcal{K} g\rangle$, so $R = -\dot\ell/\ell = \eta n\langle g, \mathcal{K} g\rangle/\ell$; multiplying and dividing by $\|g\|^2$ gives (11), and (12) is the Rayleigh bound for the symmetric positive semidefinite $\mathcal{K}$. Under squared error $\ell = \frac{1}{2}\|r\|^2$ and $g = r$, so $\chi = 2$. Conversely $\|\nabla\ell\|^2 = 2\ell$ for all $z$ says that $u = \sqrt{2\ell}$ solves the eikonal equation $\|\nabla u\| = 1$ where $\ell > 0$, whose non-negative solutions, among those differentiable on $\{\ell > 0\}$, are the distance functions $u(z) = \operatorname{dist}(z, S)$ with $S = \{\ell = 0\}$ closed; the squared error is the case $S = \{z^\star\}$, and it is the only case with a unique zero. It is not the only case: $\ell(z) = \frac{1}{2}z_2^2$ on $\mathbb{R}^2$ is smooth, satisfies $\|\nabla\ell\|^2 = 2\ell$ identically and is the squared distance to a line, so the converse needs that hypothesis and not merely the identity. For the softmax, $g = p - e_y$ gives $\|g\|^2 = (1-p_y)^2 + \sum_{j\neq y} p_j^2$, and Cauchy–Schwarz over the $C-1$ tail entries bounds that sum between $(1-p_y)^2/(C-1)$ and $(1-p_y)^2$. □

### D.4 THE CONDITIONS THAT HOLD THE SPECTRAL FACTOR STILL

**Theorem 37** (The spectral factor is a convex combination over examples)**.** *Let $u^i = \nabla_{z^i}\ell_i$ be the per-example output gradient, $w_i = \|u^i\|^2/\|u\|^2$, and*

$$\kappa_i \;=\; \frac{-\,\mathrm{d}\ell_i/\mathrm{d}k}{\eta\,\|u^i\|^2} \;=\; \frac{\langle u^i, \mathcal{K}_{ii}u^i\rangle + \sum_{j\neq i}\langle u^i, \mathcal{K}_{ij}u^j\rangle}{\|u^i\|^2}.$$

*Then under gradient flow, with no hypothesis beyond* (11)*,*

$$\varrho_{\mathcal{K}} \;=\; \sum_{i=1}^{n} w_i\,\kappa_i, \qquad \textit{hence} \qquad \min_i \kappa_i \;\le\; \varrho_{\mathcal{K}} \;\le\; \max_i \kappa_i. \tag{23}$$

*Proof.* $\dot z^i = -\eta\sum_j \mathcal{K}_{ij}u^j$ gives $\dot\ell_i = \langle u^i, \dot z^i\rangle = -\eta[\langle u^i, \mathcal{K}_{ii}u^i\rangle + \sum_{j\neq i}\langle u^i, \mathcal{K}_{ij}u^j\rangle]$, which is the second equality; summing over $i$ and dividing by $\|u\|^2$ gives (23), and the bracket is a convex combination because $w_i \ge 0$ sums to one. □

*Theorem 23 is stated in App. C.2 with its derivation, which uses* (23) *above for the weights $w_i$.*

**The rate's excess over $G(0)$.** At $L = \|r\|^2/2n$ Theorem 21 gives $R = 2\eta\,\varrho_t(r_t)$, $\varrho_t(r) = r^\top G(t)\,r/\|r\|^2$, in continuous time: the rate exceeds that of $G(0)$ at the same residual by exactly the fraction $g_t$ of §5.2.

**Proposition 38** (The rate's excess over the initial operator, in discrete time)**.** *Let $w_{k+1} = w_k - \frac{\eta}{n}J_k^\top r_k$ at squared error $L = \|r\|^2/2n$, $G_s = J_sJ_s^\top/n$, the Jacobian $\beta$-Lipschitz in operator norm on $[w_k, w_{k+1}]$, $\bar c = \eta\max\{\lambda_{\max}(G_0), \lambda_{\max}(G_k)\} < 1$, $r_k \neq 0$, $\varrho_0(r_k) > 0$, $R(k) = -\log(L_{k+1}/L_k)$ and $R_M(r) = -\log(\|Mr\|^2/\|r\|^2)$, $M = I - \eta G_0$. Write $\varrho_s = \varrho_s(r_k)$, $g_k = \varrho_k/\varrho_0 - 1$, $\sigma_s = \|G_sr_k\|^2/\|r_k\|^2$, $B_k = e^{-R_M(r_k)}$, $\tau_k = \beta\eta^2\varrho_k\|r_k\|/2n$, $b_k = \eta^2|\sigma_k - \sigma_0| + 2\tau_k + \tau_k^2$ and $a_k = 2\eta|\varrho_k - \varrho_0| + b_k$. Then $B_k \ge (1-\bar c)^2$ and $R(k) - R_M(r_k) = 2\eta\,\varrho_0\,g_k + E_k$, where, if $a_k \le B_k/2$, $|E_k| \le \big(2\eta|\varrho_k - \varrho_0|(1-B_k) + b_k\big)/B_k + a_k^2/B_k^2$, every term of order $\eta^2$ or higher. So $R(k) - R_{\mathrm{pred}}(k) = 2\eta\,\varrho_0\,g_k + E_k + R_M(r_k) - R_{\mathrm{pred}}(k)$, the last difference bounded by Proposition 26 through the angle between $r_k$ and the read's residual $p_k = (I-\eta G_0)^k r_0$.*

*Proof.* With $\Delta_k = w_{k+1} - w_k$, $r_{k+1} = (I - \eta G_k)r_k + e_k$, $e_k = \int_0^1 (J(w_k + t\Delta_k) - J_k)\Delta_k\,\mathrm{d}t$, so $\|e_k\| \le \frac{\beta}{2}\|\Delta_k\|^2 = \frac{\beta\eta^2}{2n}r_k^\top G_k r_k = \tau_k\|r_k\|$. As $0 \preceq \eta G_k \preceq I$, $\|I - \eta G_k\| \le 1$, so $L_{k+1}/L_k = A_k + T_k$ with $A_k = 1 - 2\eta\varrho_k + \eta^2\sigma_k$ and $|T_k| \le 2\tau_k + \tau_k^2$; and $B_k = 1 - 2\eta\varrho_0 + \eta^2\sigma_0 \ge (1-\bar c)^2$, the eigenvalues of $\eta G_0$ lying in $[0, \bar c]$. So $R(k) - R_M(r_k) = -\log(1 + x_k)$ with $x_k = (A_k - B_k + T_k)/B_k$, where $A_k - B_k = -2\eta\varrho_0 g_k + \eta^2(\sigma_k - \sigma_0)$ and $|x_k| \le a_k/B_k \le \frac{1}{2}$. As $|\log(1+x) - x| \le x^2$ for $|x| \le \frac{1}{2}$ and $-x_k = 2\eta\varrho_0 g_k + \big(2\eta\varrho_0 g_k(1-B_k) - \eta^2(\sigma_k - \sigma_0) - T_k\big)/B_k$, the bound on $E_k$ follows. The last identity adds and subtracts $R_M(r_k)$; Proposition 26 needs only $r_k$, $p_k$ and $G_0$. □

# E THE BOUNDARY OF EACH CLAIM, AND THE EVIDENCE BEHIND THE BODY'S NUMBERS

Where the appendices say *read 1* and *read 2*, they mean the read and the drift test. A threshold near a tenth was known from measurement before the bound was derived. At the window's end no census cell fails below a drift of $0.501$, so every threshold from $0.24$ to $0.50$ certifies the same zero failures and the census does not locate the threshold inside that range. Read early, the classifier heads do: the nearest failure sits at $0.314$, and thresholds of $0.20$, $0.24$, $0.26$ and $0.30$ certify $2{,}091$, $2{,}581$, $2{,}765$ and $3{,}164$ heads with none failing, $0.35$ certifies $3{,}836$ with $9$ failing, $0.40$ certifies $4{,}309$ with $97$ and $0.50$ certifies $4{,}840$ with $304$.

## E.1 LIMITATIONS: EVERY CLAIM, AND THE SETTING IT WAS MEASURED IN

At squared error the results are measured on character transformers from $0.11$M to 3B parameters, on MNIST classifier heads, built heads and perceptrons, and on released checkpoints: the read under full-batch and minibatch gradient descent, the drift test under full batch, and the early growth test on transformers. Each claim is stated below with the setting it was measured in.

**At cross-entropy read 2 withholds every cell, in full.** Over $26$ cells at cross-entropy on two populations, $8$ from-scratch transformers and $18$ built heads, every drift exceeds $0.99$.

**The example count is what read 2 pays for, formed exactly.** Neither read grows with the run, but read 2 formed exactly grows with the example count: the drift needs $G$ at both ends of the window, an $(nC) \times (nC)$ operator formed from $nC$ products, which is why the from-scratch ladder runs full batch on sixteen examples.

**The falsifier register.** Each measurement's script states an outcome that, if it occurred, would count against the claim being tested; for the few whose script is not in the released code the result file's own summary states it, written with the result, so for those few the order of statement and measurement rests on that summary. Across the project $407$ of them carry a recorded outcome: $278$ stayed quiet, $86$ fired, $40$ were evaluated over slices that disagree, and $3$ carry a record whose polarity fixes no verdict.

The entries, by their claims: The read governs to its horizon, and past it each ladder run learns faster than its linearisation; On transformers the early drift test's boundary is the parameterisation and the target's scale; The zero false certifications bound a rate per independent unit; The drift test is claimed for full-batch descent at squared error; Under minibatch SGD the read and its test cost more than the run; In sizing a budget, a trial run charged the read's price wins; On released backbones the read and a power law given a third are level within chance; The populations of §5 and their settings; The decision the read serves; The location of the minimum is predicted by width, not by seed; The minimum is broad, which bounds what a rule built on it can win; Asymptotically-optimal momentum is the wrong constant at realistic budgets; A rank statistic sets the accuracy, and not the estimator's own sensitivity; The band constrains directions, not matrix elements; The corpora, and the linear model they are read through; On shallow perceptron windows the read loses at every batch size, full batch included; The trajectory corpus and its setting; On CIFAR-10 the operator leaves the certified interval early, and above it the read is worse; The setting the bracket is derived for; The basis of the excess relation, and the null beside it; The reach of the released-backbone population; The ladder's reach above $118$M; Read 2's interval is calibrated at sixteen examples; Multiplicity.

## E.2 THE READ'S REACH: FOUR WEIGHTINGS, THE HORIZON, AND A SUM OF EXPONENTIALS

**Four weightings of one error.** The ladder's rate curves are stored on a logarithmic grid of $60$ steps from $20$ to $3{,}000$, so half its points precede step $235$ and a median over it, a median in log time, weights the early descent most. Table 7 rescores the $21$ curves of §5.1 four ways, every competitor under the same weighting as the read. In log time the read is closer than the nearest size's measured curve on $13$ of the $16$ held-out parameter counts; neighbouring sizes share that curve, and on sizes sharing none the best set of seven gives six, $p = 0.125$, so no dependence-aware test separates the two.

**The horizon.** A curve's horizon is its first grid step at which the read's pointwise rate error exceeds a third of a decade.

Table 7: The ladder's 21 from-scratch curves rescored under four weightings of the pointwise rate error: in log time, over the stored logarithmic grid; uniformly over the 3,000 steps; over the last third of the steps, the grid points from 2,136 to 3,000; and at step 3,000. Each entry is the median over curves, and "closer" counts the curves on which that predictor errs strictly less than the read under the same weighting. A predictor given a prefix is scored only on the grid points after it, and the read is scored there too for the count; on those points the read's own median in log time is 0.040, 0.054, 0.049 and 0.099 given a tenth, a fifth, a third and half, and uniformly over steps it is 0.223 given half.

| predictor | log time median | log time closer | uniform in step median | uniform in step closer | last third of steps median | last third of steps closer | step 3,000 median | step 3,000 closer |
|---|---|---|---|---|---|---|---|---|
| the read, which watches nothing | 0.029 | | 0.203 | | 0.384 | | 0.430 | |
| nearest size's measured curve | 0.044 | 7 | 0.035 | 20 | 0.043 | 21 | 0.040 | 21 |
| log-linear in parameters, per step | 0.045 | 6 | 0.051 | 20 | 0.060 | 21 | 0.053 | 21 |
| geometric mean of the other sizes' curves | 0.080 | 2 | 0.053 | 20 | 0.054 | 21 | 0.051 | 21 |
| broken power law, two breaks | 0.081 | 2 | 0.054 | 20 | 0.049 | 21 | 0.044 | 21 |
| power law, given a tenth | 0.734 | 0 | 0.528 | 5 | 0.339 | 12 | 0.315 | 13 |
| eleven families, given a tenth | 0.876 | 0 | 1.064 | 1 | 1.983 | 6 | 2.413 | 6 |
| eleven families, held-out weights, a tenth | 0.812 | 0 | 5.354 | 0 | 8.338 | 1 | 9.557 | 0 |
| the families' pointwise median, a third | 0.400 | 0 | 0.429 | 3 | 0.530 | 5 | 0.557 | 5 |
| power law with a floor, given a third | 4.150 | 0 | 6.922 | 0 | 8.508 | 0 | 9.077 | 0 |
| sum of exponentials, given a tenth | 1.346 | 0 | 1.360 | 0 | 1.299 | 0 | 1.293 | 0 |
| sum of exponentials, given a fifth | 2.780 | 0 | 2.955 | 0 | 3.060 | 0 | 3.077 | 1 |
| sum of exponentials, given a third | 0.063 | 9 | 0.127 | 16 | 0.145 | 20 | 0.144 | 20 |
| sum of exponentials, given half | 0.062 | 18 | 0.104 | 19 | 0.125 | 21 | 0.145 | 21 |

**A sum of exponentials, and its fit.** The competitor is $L(k) = \sum_{j=1}^{K} c_j\, e^{-s_j k}$ with $c_j, s_j > 0$, whose analytic rate is fitted by least squares to the logarithm of the stored rate over a prefix of the curve, the domain it is scored in.

### E.3 The horizon under $\mu$P, its ordering, and the operator held fixed

**Two ladders, opposite trends.** The $\mu$P ladder re-runs the from-scratch cells that carry a drift under maximal-update parameterisation, identical otherwise, so at the base width, 0.11M, the two runs are one run bit for bit.

**The growth of $G$ along the residual carries the departure.** By Proposition 38 a run outpaces its linearisation through the growth of $G$ along its residual or through the residual itself. Over the last third of the window the median of $\log_{10}\big(r_k^\top G(k)\, r_k / r_k^\top G(0)\, r_k\big)$ is a median 1.00 times the run's excess over its linearisation, the median $\log_{10}(R_{\mathrm{meas}}/R_{\mathrm{pred}})$. The ratio is 1.09 over the standard runs and 0.97 over the $\mu$P runs, and 0.89 to 1.23 run by run, so on none of the 30 is it below one half.

**The operator held fixed.** Scaling the one-hot targets by $c$ leaves $G(0)$, $\lambda_{\max}$, $\kappa$ and the step identical by construction, and the captured $G(0)$ is identical bit for bit across the arms. The read also takes the residual, which moves with the targets, so the arms show that $G(0)$ alone does not decide the horizon and are not instances of Theorem 1. At 3.2M, six arms with $c$ from 0.25 to 8 have horizons from beyond step 3,000 to step 92 (Table 8), ordered by $c$ at a Spearman $-1.0$ in the descent governed, exact two-sided $p = 0.003$ over the 720 orderings. At 1B the descent governed falls with $c$ at a rank correlation of $-0.83$, exact one-sided $p = 0.029$ over the 720 orderings; the four arms that never leave enter at their whole descent, which does not fall with $c$ (Table 8, App. E.5).

### E.4 The early test on transformers: the growth of $G$ along the residual

**The test.** On transformers the early test is the growth of $G$ along the current residual, read a tenth of the way into the window, at step $t = 300$ of 3,000: $g = r_t^\top G(t)\, r_t / r_t^\top G(0)\, r_t - 1$. Under standard parameterisation it is $\|J_t^\top r_t\|^2 / \|J_0^\top r_t\|^2 - 1$.

**Its orderings and its certificates.** It is scored on three populations. Over the 30 runs the growth orders the horizon at $-0.871$ in the descent governed and $-0.899$ in steps. Within $\mu$P it orders it at $-0.930$ and $-0.968$, where the drift at the same tenth orders it with the wrong sign, at $+0.88$ in the descent governed, and pooled at $+0.26$. Its Spearman partial correlation with the descent governed is $-0.55$ within $\mu$P, $p = 0.010$ by permutation of the rank residuals, $-0.82$ in steps, and $-0.52$ over both ladders with size held fixed inside each, $p = 0.005$.

Table 8: Six runs each of one 3.2M, one 20M and one 1B network whose one-hot targets are scaled by $c$, trained 3,000 full-batch steps from the ladder's initialisation, in float64 with $G(0)$ identical bit for bit at 3.2M and 20M, and in float32 from weights identical bit for bit at 1B, where $c = 1$ is the standard 1B run of App. E.5: the horizon, the descent governed and the whole descent in decades; at a tenth, the drift, the growth of $G$ along the residual (App. E.4) and the median log-time rate error over the remaining nine tenths; and the drift at the window's end and the median over the whole window. A third of a decade is the tolerance.

| $c$ | horizon, step | governed | descent | drift at 300 | growth at 300 | error after | drift at 3,000 | error, whole |
|---|---|---|---|---|---|---|---|---|
| *3.2M* | | | | | | | | |
| 0.25 | beyond 3,000 | 3.68 | 3.68 | 0.108 | −0.206 | 0.088 | 0.119 | 0.079 |
| 0.5 | beyond 3,000 | 3.35 | 3.35 | 0.108 | −0.195 | 0.060 | 0.099 | 0.042 |
| 1 | 2,531 | 3.07 | 3.34 | 0.103 | −0.134 | 0.088 | 0.136 | 0.044 |
| 2 | 650 | 1.80 | 5.01 | 0.084 | 0.194 | 0.502 | 0.506 | 0.072 |
| 4 | 216 | 1.11 | 10.67 | 0.203 | 3.278 | 1.067 | 1.474 | 0.471 |
| 8 | 92 | 0.79 | 25.05 | 2.342 | 9.817 | 1.425 | 2.367 | 1.272 |
| *20M* | | | | | | | | |
| 0.25 | beyond 3,000 | 4.09 | 4.09 | 0.044 | −0.094 | 0.039 | 0.045 | 0.036 |
| 0.5 | beyond 3,000 | 3.77 | 3.77 | 0.043 | −0.081 | 0.025 | 0.039 | 0.022 |
| 1 | beyond 3,000 | 3.71 | 3.71 | 0.040 | −0.025 | 0.088 | 0.072 | 0.020 |
| 2 | 771 | 2.02 | 5.07 | 0.035 | 0.231 | 0.437 | 0.244 | 0.058 |
| 4 | 256 | 1.20 | 9.67 | 0.125 | 1.943 | 0.922 | 0.641 | 0.328 |
| 8 | 100 | 0.81 | 22.07 | 1.651 | 16.098 | 1.271 | 1.755 | 1.190 |
| *1B, float32* | | | | | | | | |
| 0.25 | beyond 3,000 | 4.20 | 4.20 | 0.051 | −0.210 | 0.062 | 0.055 | 0.054 |
| 0.5 | beyond 3,000 | 4.00 | 4.00 | 0.050 | −0.206 | 0.055 | 0.049 | 0.051 |
| 1 | beyond 3,000 | 3.80 | 3.80 | 0.047 | −0.184 | 0.029 | 0.031 | 0.036 |
| 2 | beyond 3,000 | 4.16 | 4.16 | 0.033 | −0.081 | 0.095 | 0.091 | 0.039 |
| 4 | 660 | 1.92 | 6.36 | 0.052 | 0.417 | 0.477 | 0.312 | 0.104 |
| 8 | 183 | 1.07 | 12.41 | 0.356 | 3.331 | 0.811 | 0.787 | 0.469 |

### E.5 Standard against $\mu$P at 1B and 3B parameters

**The runs.** The standard ladder's top rung, of 1.008B parameters, and a wider and deeper network of 2.946B, called 1B and 3B here, were each trained under both parameterisations, and the $\mu$P ladder was filled in at the standard ladder's 118M and 403M rungs: six runs at one seed.

**The two parameterisations part.** No standard run leaves tolerance within the 3,000 steps, so the read governs each one's whole descent. That is 3.80 and 3.89 decades at 1B, at the two seeds, and 3.95 at 3B.

**The early tests at scale.** The growth of $G$ along the residual is read at a tenth, at step 300, on all 13 runs, against the threshold of $-0.007$ that App. E.4 fixed on the ladders' standard runs, unchanged. It certifies the 6 runs that hold over the remaining nine tenths, the three standard runs and the target-scaled arms with $c \leq 2$, which err 0.019 to 0.095 decades in log time there.

Table 9: The eight float32 runs at $c = 1$, on sixteen sequences, full batch for 3,000 steps, the 1B pair at two seeds. The horizon is scored at every step, and on the ladders' logarithmic grid in the row below; the descent governed by the horizon and the whole descent are in decades of loss; the read's error is its median in log time, over the whole window and over the nine tenths after step 300; the drift and the growth of $G$ along the residual are read at step 300, and the drift again at step 3,000; wall times are on a card shared with other jobs. A third of a decade is the tolerance.

| | standard | | | | | $\mu$P | | |
|---|---|---|---|---|---|---|---|---|
| | 1B | 1B, seed 1 | 3B | 118M | 403M | 1B | 1B, seed 1 | 3B |
| parameters | 1.008B | 1.008B | 2.946B | 118.3M | 403.3M | 1.008B | 1.008B | 2.946B |
| width, depth | 2,048, 20 | 2,048, 20 | 3,072, 26 | 1,280, 6 | 2,048, 8 | 2,048, 20 | 2,048, 20 | 3,072, 26 |
| horizon, step | beyond 3,000 | beyond 3,000 | beyond 3,000 | 364 | 348 | 364 | 343 | 289 |
| on the grid | beyond 3,000 | beyond 3,000 | beyond 3,000 | 391 | 359 | 391 | 359 | 303 |
| governed | 3.80 | 3.89 | 3.95 | 1.32 | 1.27 | 1.33 | 1.27 | 1.27 |
| descent | 3.80 | 3.89 | 3.95 | 7.36 | 8.15 | 7.26 | 7.18 | 8.41 |
| share governed | 100% | 100% | 100% | 18% | 16% | 18% | 18% | 15% |
| error, log time | 0.036 | 0.037 | 0.030 | 0.192 | 0.211 | 0.184 | 0.207 | 0.257 |
| error after step 300 | 0.029 | 0.029 | 0.019 | 0.740 | 0.802 | 0.728 | 0.733 | 0.813 |
| drift at 300 | 0.047 | 0.044 | 0.038 | 0.068 | 0.081 | 0.123 | 0.119 | 0.218 |
| growth at 300 | −0.184 | −0.152 | −0.149 | 0.880 | 0.995 | 0.922 | 1.002 | 1.596 |
| drift at 3,000 | 0.031 | 0.030 | 0.025 | 0.514 | 0.574 | 0.615 | 0.555 | 0.638 |
| wall time, minutes | 34 | 8 | 27 | 2 | 3 | 19 | 9 | 27 |
| peak memory, GiB | 18.8 | 18.8 | 54.9 | 2.4 | 7.6 | 18.8 | 18.8 | 54.9 |

### E.6 The $\mu$P ladder at three to five seeds, and both early tests on every population

**The trend with size, at every seed.** Over all 46 runs the horizon's rank correlation with parameter count is $-0.945$ in the descent governed and $-0.640$ in steps, and none of 200,000 relabellings reaches either.

**The early tests on the seeds.** Read at a tenth, the growth of $G$ along the residual certifies 10 of the 46 runs and withholds the other 36, all of which fail. It certifies all 6 that hold over the remaining nine tenths: the three at the base width, one of them the standard run the threshold was fixed on, and three of the five at 0.24M.

**Both early tests on every population.** Table 10 scores the two tests on every transformer population the paper reads them on, each run counted once. Certifying the standard runs at $c \leq 2$ from the labels alone decides the 13 runs from 118M to 3B as the growth does; over the 93 runs out of sample it certifies 20 of the 40 that hold against the growth's 39, and falsely 2, the standard 3.2M and 20M runs at $c = 2$, against the growth's 4.

Table 10: Both early tests on transformers, each read at a tenth, on every population they are scored on, each run counted once: the runs, those that hold and those that fail over the remaining nine tenths in log time, and the runs each test certifies, with the failing runs among them. The growth's threshold, $-0.007$, is fixed on the first row alone; the drift's is 0.24, the drift exact where $G$ is formed, on sixteen sequences to 51M, and estimated from probes on the other runs. The first $\mu$P row is the $\mu$P ladder at three to five seeds a size (App. E.6) less its base-width run at seed 0, the first row's 0.11M run; the rows across the example count and at 118M and 403M on 256 examples are the runs of App. E.7. Six of the 53 that fail leave tolerance at or before step 300, where both tests are read: the target-scaled runs at $c = 4$ and 8 at 3.2M and 20M and at $c = 8$ at 1B, and the 3B $\mu$P run. A run holds here when its later error is within tolerance, so a run can hold past its horizon. The $\mu$P run at 20M on 64 examples leaves tolerance at step 1,088 and holds, at 0.276. The 16 runs at 256 examples are the 3,000-step runs at seeds 0 and 1. Of these, the $\mu$P pair at 20M leaves tolerance at steps 2,939 and 2,462 and holds, at 0.040 and 0.054. Table 12 carries 24 runs at up to four seeds to the horizon.

| | | | | growth of $G$ along $r_t$ | | drift | |
|---|---|---|---|---|---|---|---|
| population | runs | hold | fail | certified | failing | certified | failing |
| standard ladder, sixteen sequences | 9 | 9 | 0 | 9 | 0 | 9 | 0 |
| $\mu$P ladder, three to five seeds | 45 | 5 | 40 | 9 | 4 | 44 | 40 |
| target-scaled, 3.2M and 20M, $c \neq 1$ | 10 | 4 | 6 | 4 | 0 | 8 | 4 |
| $\mu$P, 64 to 4,096 examples | 6 | 6 | 0 | 5 | 0 | 3 | 0 |
| 256 examples, 0.42M to 20M | 16 | 16 | 0 | 16 | 0 | 8 | 0 |
| 256 examples, 118M and 403M | 3 | 3 | 0 | 3 | 0 | 3 | 0 |
| 118M to 3B, sixteen sequences | 13 | 6 | 7 | 6 | 0 | 12 | 6 |
| out of sample, each run once | 93 | 40 | 53 | 43 | 4 | 78 | 50 |

### E.7 More examples: 16,384 and 256 examples, float32 and minibatch SGD

**Cost.** At full batch the read costs 157 and 161 steps, and the probe drift at a tenth as much again.

**The example count swept.** The standard 3.2M and 20M rungs have been read at six example counts a factor of four apart, from 16 to 16,384, each trained full batch for 3,000 steps at the ladder's step size (Table 11). Three of the twelve runs, both sizes at 64 examples and the 20M rung at 4,096, were trained for this sweep, in float32 with the protocol of the 256-example runs; the other nine are the ladder's two rungs at sixteen examples and runs this appendix describes.

**The example count swept under $\mu$P.** The $\mu$P networks of the same two sizes were trained with the sweep's protocol, full batch in float32 for 3,000 steps at seed 0. At 20M it leaves tolerance at step 458 at sixteen examples and at step 1,088 at 64, where it still holds, its later error 0.276 decades (Table 10).

**At 1,024 and 4,096 examples, past the window.** Six pairs were run past the window by the protocols of the 256-example runs, four at 1,024 examples (3.2M, 20M, 118M, 403M) and two at 4,096 (3.2M, 20M). Each is scored with the finest read valid over its window (Table 13). At 20M the standard run leaves from above at step 29,514, after 1.540 of its 1.542 decades, and the $\mu$P run at 6,775. The standard network stays linear 4.36 times longer. At 3.2M the $\mu$P run leaves at step 17,178, after 0.957 of its 1.038 decades, and the standard run stays within tolerance over all 30,000 steps, more than 1.74 times as long. At 118M the $\mu$P run leaves at step 4,999, after 0.525 of its 0.906 decades, having held over the 3,000-step window. The standard run stays within tolerance over all 15,000 of its steps, more than 3.00 times as long. At 403M the $\mu$P run leaves at step 4,529, after 0.431 of its 0.470 decades, having held over the 3,000-step window. The standard run stays within tolerance over all 12,000 of its steps, more than 2.64 times as long. At 4,096 examples the 20M $\mu$P run leaves at step 16,974. The standard run stays within tolerance over all 30,000 steps, more than 1.76 times as long, its error peaking at 0.118 decades. At 3.2M and 4,096 examples the standard run

leaves first, at step 21,973 or 21,604 at 128 nodes, its error peaking at 0.360 against the tolerance of a third. Its $\mu$P twin stays within tolerance over all 30,000 steps, peaking at 0.222. That standard exit comes from below, as Proposition 38 predicts for a kernel that shrinks along the residual, here to a growth of $-0.886$ by step 30,000. Every other exit at 3.2M and 20M comes from above, with late growth of $G$ along the residual of $+1.25$ to $+4.57$ where it was read. The outcome registered against the contrast is a standard run leaving before its $\mu$P twin. It occurs at the resolution each window needs through one pair, the 3.2M pair at 4,096 examples. On the registered 64-node read the standard run leaves first in three pairs, at 3.2M at step 8,719 against its twin's 20,483. At 4,096 examples it leaves at 9,375 against 10,926 at 3.2M and at 9,062 against 26,362 at 20M. Every one of these exits lies past 6,827 steps, the range the law $q^2/(2\eta\lambda_{\max})$ gives 64 nodes, and the finer reads remove the first and the third. So at 1,024 examples every $\mu$P run leaves, earlier the larger as on sixteen sequences, while every standard run follows the read more than 1.74 times as long.

Table 11: The read across the example count, standard 3.2M and 20M rungs trained full batch for 3,000 steps: its median rate error in log time and uniformly over steps, the descent it governs by its horizon and the whole descent in decades, the horizon, and the drift at a tenth and at the end. Runs are float64 at sixteen and 1,024 examples and for 3.2M at 4,096, float32 elsewhere. The horizon at sixteen is read on the ladder's grid and elsewhere at every step; the drift at sixteen is the exact one and elsewhere the probes'. At 256 examples the float64 runs reproduce the log-time error to within $2 \times 10^{-8}$ decades. Blank where a run's store records no such value.

| size | examples | log time | uniform | governed | descent | horizon | drift, tenth | drift, end |
|---|---|---|---|---|---|---|---|---|
| 3.2M | 16 | 0.0445 | | 3.07 | 3.34 | 2,531 | | 0.136 |
| 3.2M | 64 | 0.0411 | 0.066 | 1.89 | 1.89 | beyond 3,000 | 0.217 | 0.145 |
| 3.2M | 256 | 0.0350 | 0.075 | 1.54 | 1.54 | beyond 3,000 | 0.418 | 0.541 |
| 3.2M | 1,024 | 0.0235 | | 1.45 | 1.45 | beyond 3,000 | | 1.042 |
| 3.2M | 4,096 | 0.0218 | | 1.43 | 1.43 | beyond 3,000 | | 1.114 |
| 3.2M | 16,384 | 0.0227 | 0.097 | 1.42 | 1.42 | beyond 3,000 | 0.615 | 1.132 |
| 20M | 16 | 0.0196 | | 3.71 | 3.71 | beyond 3,000 | | 0.072 |
| 20M | 64 | 0.0190 | 0.040 | 1.99 | 1.99 | beyond 3,000 | 0.074 | 0.061 |
| 20M | 256 | 0.0140 | 0.022 | 1.55 | 1.55 | beyond 3,000 | 0.103 | 0.120 |
| 20M | 1,024 | 0.0137 | | 1.46 | 1.46 | beyond 3,000 | | 0.213 |
| 20M | 4,096 | 0.0136 | 0.036 | 1.43 | 1.43 | beyond 3,000 | 0.137 | 0.250 |
| 20M | 16,384 | 0.0135 | 0.043 | 1.42 | 1.42 | beyond 3,000 | 0.139 | 0.261 |

**At 256 examples, standard against $\mu$P.** Four sizes, 0.42M, 1.6M, 3.2M and 20M, standard and $\mu$P, full batch in float32 on 256 windows, were trained until each had descended three decades or reached 20,000 steps, so that the horizon could be reached, at three seeds a size, 0, 2 and 3, four at 3.2M and two at 20M: 24 runs, 12 standard and 12 $\mu$P (Table 12).

**At 256 examples to 403M.** Three more runs carry the contrast to 118M and 403M at seed 0: the standard network at both sizes and $\mu$P at 118M (Table 12). They were trained in float32 with a compensated weight update, as at 1B (App. E.5), for 6,000 steps, and are scored over the first 3,000 and, in Table 12, over all 6,000.

**At 256 examples to 1B.** Three further runs, at seed 0 by the same procedure and stopped at the 3,000-step window, carry the contrast to 1B: the standard network and $\mu$P at 1B, and $\mu$P at 403M (Table 12). The standard run at 1B stays inside tolerance over the whole window and errs 0.049 decades in log time. The $\mu$P runs leave tolerance at step 1,816 at 403M and 1,813 at 1B, after 0.52 of the 0.59 decades they descend in the window and 0.54 of the 0.61. Two outcomes registered before these runs as counting against the results did not occur: the standard 1B run leaving tolerance before the $\mu$P one, and the $\mu$P run at 403M governing more descent than the one at 118M. A third occurred: the $\mu$P run at 1B governs more descent than the one at 403M, 0.544 decades against 0.519, and the $\mu$P horizon is flat in steps from 403M to 1B, 1,813 against 1,816. Read at a tenth, both early tests certify all three, and all three hold over the remaining nine tenths. Computed after the runs, not registered, and with the 64-node read, the $\mu$P trend over the 15 runs at 256 examples is $-0.86$ in the descent governed, every one leaving tolerance, and the standard trend over 15 runs lies between $+0.34$ and $+0.96$, depending on where the three still inside tolerance leave.

**Standard against $\mu$P under SGD, for 15,000 steps.** At 16,384 examples both parameterisations were trained for 15,000 minibatch steps at 3.2M and 20M. The standard runs continue the SGD runs above from step 3,000; the $\mu$P runs start from initialisation with their own read, taken in the operator the run descends on. Scored with 128 nodes (App. E.9), the read holds over all 15,000 steps at both sizes and under both parameterisations, erring 0.020 to 0.033 decades in log time; the 64-node exits,

Table 12: The horizon at 256 examples, run by run, standard beside $\mu$P, scored at every step, past 5,000 steps with the 128-node read (App. E.9); the descent governed and the whole descent, in decades, and the steps run.

| | | standard | | | | $\mu$P | | | |
|---|---|---|---|---|---|---|---|---|---|
| size | seed | horizon | governed | descent | steps | horizon | governed | descent | steps |
| 0.42M | 0 | 7,759 | 1.61 | 1.96 | 20,000 | 5,490 | 1.31 | 2.04 | 20,000 |
| 0.42M | 2 | 7,610 | 1.57 | 1.92 | 20,000 | 5,389 | 1.28 | 1.98 | 20,000 |
| 0.42M | 3 | 8,717 | 1.57 | 1.83 | 20,000 | 5,983 | 1.28 | 1.90 | 20,000 |
| 1.6M | 0 | 8,704 | 1.62 | 1.91 | 20,000 | 3,723 | 1.05 | 2.47 | 20,000 |
| 1.6M | 2 | 7,031 | 1.61 | 2.03 | 20,000 | 3,097 | 1.04 | 2.60 | 20,000 |
| 1.6M | 3 | 7,326 | 1.64 | 2.06 | 20,000 | 3,125 | 1.06 | 2.69 | 20,000 |
| 3.2M | 0 | 11,983 | 1.64 | 1.82 | 20,000 | 5,120 | 1.05 | 2.21 | 20,000 |
| 3.2M | 1 | 13,392 | 1.65 | 1.78 | 20,000 | 5,845 | 1.06 | 2.09 | 20,000 |
| 3.2M | 2 | 12,348 | 1.63 | 1.79 | 20,000 | 5,207 | 1.04 | 2.15 | 20,000 |
| 3.2M | 3 | 10,553 | 1.66 | 1.93 | 20,000 | 4,620 | 1.07 | 2.39 | 20,000 |
| 20M | 0 | 11,155 | 1.68 | 1.92 | 20,000 | 2,738 | 0.75 | 3.00 | 17,269 |
| 20M | 1 | 10,084 | 1.68 | 1.97 | 20,000 | 2,437 | 0.76 | 3.00 | 16,398 |
| 118M | 0 | beyond 6,000 | 1.65 | 1.65 | 6,000 | 2,030 | 0.61 | 1.11 | 6,000 |
| 403M | 0 | beyond 6,000 | 1.65 | 1.65 | 6,000 | 1,816 | 0.52 | 0.59 | 3,000 |
| 1B | 0 | beyond 3,000 | 1.51 | 1.51 | 3,000 | 1,813 | 0.54 | 0.61 | 3,000 |

Table 13: The runs at 1,024 examples and, below the rule, the 3.2M and 20M pairs at 4,096. Left: the horizon at 64, 128 and 256 nodes (none: no exit; blank: not read) and the decades governed at the finest valid read. Right: the step at which the $q$-node read's rate first leaves a 512-node read's (256 at 20M) by 0.05 decades, beside the law.

| | | horizon at $q =$ | | | | the read holds to, $q =$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| run | steps | 64 | 128 | 256 | governed | 16 | 32 | 64 | 128 | 256 |
| standard, 3.2M | 30,000 | 8,719 | none | none | 1.492 | 325 | 1,213 | 5,828 | 43,695 | 293,126 |
| $\mu$P, 3.2M | 30,000 | 20,483 | 17,182 | 17,178 | 0.957 | 320 | 1,246 | 6,705 | 51,133 | 372,425 |
| standard, 20M | 30,000 | none | none | 29,514 | 1.540 | 306 | 1,164 | 5,316 | 34,391 | |
| $\mu$P, 20M | 15,000 | 7,325 | 6,775 | | 0.672 | | | | | |
| standard, 118M | 15,000 | none | none | | 1.521 | | | | | |
| $\mu$P, 118M | 15,000 | 5,112 | 4,999 | | 0.525 | | | | | |
| standard, 403M | 12,000 | none | none | | 1.516 | | | | | |
| $\mu$P, 403M | 8,000 | 4,561 | 4,529 | | 0.431 | | | | | |
| the law, $q^2/(2\eta\lambda_{\max})$ | | | | | | 427 | 1,707 | 6,827 | 27,307 | 109,227 |
| standard, 3.2M | 30,000 | 9,375 | 21,604 | 21,973 | 1.448 | | | | | |
| $\mu$P, 3.2M | 30,000 | 10,926 | none | none | 0.921 | | | | | |
| standard, 20M | 30,000 | 9,062 | none | none | 1.453 | | | | | |
| $\mu$P, 20M | 30,000 | 26,362 | 17,008 | 16,974 | 0.641 | | | | | |

at steps 9,576 to 11,985, lie past that read's range. Each descends less over the window than the same network governs at 256 examples: the standard runs 1.44 and 1.43 decades against at least 1.63, the $\mu$P runs 0.91 and 0.63 against at least 1.04 and 0.75. Two outcomes registered before these runs as counting against the size contrast at this count occur: the standard 20M run governs its whole 1.434 decades against the 3.2M run's 1.438, a near tie, and neither 20M run leaves tolerance, a tie the registration counts against the contrast. At 1B the standard network under minibatch SGD at batch 1,024 holds over its 1,000-step window, erring 0.045 decades in log time over the 1.26 it descends, so the outcome registered against it did not occur; at a tenth its drift is 0.079 and its growth $-0.29$, and both tests certify it.

### E.8 Further measurements the body quotes

**Section 5's side results, in full.** At 16,384 the read holds in float32 under minibatch SGD too, erring 0.043 to 0.097 decades uniformly over steps; these errors and the drifts below agree with the full-batch runs' to three decimals and differ in the fourth. At its cost, 157 and 161 full-batch steps, every competitor we fit errs more in log time: a sum of exponentials, a power law, one with a floor, one in the rate and the eleven-family ensemble. The drift test at a tenth certifies both 20M runs, at 0.139, and withholds both 3.2M runs, at 0.615, which hold. Released checkpoints built for five modalities, read on image stills, sound clips and protein sequences, up to 6.7B parameters and fine-tuned on ten or sixteen examples as released and re-initialised, give the read 23 of 42 wins against a power law given a third of the steps, at chance by a sign test (App. E.8, Table 5, Fig. 6). The spectrum carries to minibatch momentum, read in its own metric, and on MNIST gains accuracy with width (Figs. 5, 9, App. E.8). On the 4,923 classifier heads read after a tenth, the drift test ranks the later error at 0.64 where the observed error ranks it at 0.17, a difference that resampling whole configurations puts at $[0.40, 0.54]$. Across the four families it ranks the later error at 0.78, against 0.60 for that tenth's own error. At the window's end, above the threshold 111 of the 228 head cells exceed tolerance. At cross-entropy the test withholds a further 15 cells to 118M. Yu et al. (2025) prove that there the empirical tangent kernel diverges, for binary logistic loss under gradient flow; the finite-horizon claim rests on the measured drifts. On every cell where the three passes were run,

the drift the residual meets sits one to two orders of magnitude below the read's bound on it, which restates stability. The passes close that gap; in the worst case over the lazy-scaling family and two more, no quantity computable from the read can (Thm. 1).

**Beyond the theory's setting.** The read carries beyond the theory's setting: under minibatch momentum, read at the effective step the momentum reduction prescribes, it wins 6 of 6 cells at a median $1.47\times$ against $2.95\times$, handed the raw step instead it ties. It withholds at squared error on WikiText-103 too, at 124M and 178M, every drift there being $0.98$; those runs are minibatch. On the WikiText-103 cells the measured descent lies above the read's own bracket by $3.85$ to $5.57$ times.

**Two checks.** Where the parameters enter linearly, in one layer, over $8$ cells, the measured residues carry the curve to $0.0044$ decades against $0.0413$ for the generic profile $a_i \propto \lambda_i$. On the $15$ ladder cells whose spectra are stored, the read's eigenvalues are kept and its residues replaced, and each profile is scored as the read is, by its median rate error in log time. The read errs a median $0.038$ decades; with $a_i \propto \lambda_i$ it errs $0.178$, with uniform residues $0.207$, and with the top mode alone $2.74$. The measured residues are closer than each profile on all $15$ cells, so on the ladder the residues, not the spectrum alone, carry the curve. A count of zero bounds a rate only at a unit that can be treated as independent. The early test's $2{,}581$ certified classifier heads come from $419$ of the $716$ configurations, none of which holds a failing head, so the exact one-sided $95\%$ bound on its rate of false certification per configuration is $0.71\%$, and per head $0.12\%$. The census's $90$ certified cells come from $60$ configurations, which bounds its rate at $4.9\%$. Both bounds assume configurations independent, which shared initialisations and data weaken.

**The released count.** It reproduces from the stored runs: $23$ wins, $17$ losses and $2$ ties, $p = 0.43$.

**Three seeds a size.** Over the $36$ cells of the three seeds the median read error is $0.029$ decades, the published seed-$0$ figure, and no cell leaves tolerance, the worst erring $0.097$.

**The parameter-space read.** On every cell where the operator is full rank the two reads agree to $1.2 \times 10^{-14}$ on the predicted rate and their spectra to $3.4 \times 10^{-12}$.

**The residues and the operator's motion.** Over $27$ cells of a second width sweep, at nine widths from $16$ to $4096$, drift falls from $0.911$ to $0.036$ and the residue advantage rises monotonically from $-0.268$ to $+0.028$ decades, crossing zero between widths $1024$ and $2048$ at a drift of $0.093$.

**Arrival.** On this corpus $-\log R_\infty$ alone carries $97.7\%$ of the variance of $\log T_X$: the intervention that moves every residue while pinning every eigenvalue leaves the shape unable to order arrival, at a concordance of $0.471$ against the edge's $0.880$.

**Exact checks of the read.** Reading both at the same initialisation and evaluating the predicted loss curve at the $254$ distinct steps a log-spaced grid over $1$ to $3{,}000$ produces, the largest relative disagreement on a cell has a median of $1.9 \times 10^{-6}$ and a worst case of $3.4 \times 10^{-5}$, at $40$ times fewer matrix–vector products and $40.6$ times less wall clock. The matrix-free product is checked against an explicitly formed $JJ^\top/n$ on two cells small enough to materialise the Jacobian, one at squared error and one at cross-entropy: in float64 the two Grams agree to a relative $7.4 \times 10^{-16}$ and $1.0 \times 10^{-15}$. A Gram assembled from an explicit Jacobian is checked in turn against the analytic Gram of a one-hidden-layer network of width $64$ at $64$ examples, to a relative $2.9 \times 10^{-15}$.

**Two conventions for the horizon.** Under the criterion of Meterez et al. (2026), a predicted loss decrease within $10\%$ of the measured one, applied two-sided, the first breach falls at a median of $65$ steps over $40$ cells, and at $115.5$ under the one-sided inequality they print. Over the $49$ cells where both conventions are defined, the last step still within tolerance exceeds the first breach by a median factor of $23$; it is also no replacement, scaling one-for-one with how long we watch (log–log slope $+1.11$, $p = 10^{-11}$), and scoring the per-step increment instead re-enters the same way.

**Two alternatives to the threshold.** Every threshold in $\delta$ alone is empirical: a motion built from the read's eigenvectors, which keeps the read's eigenbasis so that nothing rotates, sends the rate error to infinity at drifts far below it (Prop. 32). Two alternatives bracket it. The same motion read at the bottom

does carry a threshold. It certifies 78 of the 90 against the whole-operator drift's 5, and it misclassifies 18 against that statistic's 79, so it is far the more useful of the two. But 6 of the 78 it certifies exceed tolerance, at 0.267 to 0.313 decades, where the whole-operator drift certifies none that fail.

**The bound against its own class.** Searching 906 such motions on each of those 240 spectra: the bound as stated holds on all 240; dropping $\mathcal{A}$ fails on 7, by up to 16.2%; the first-order form fails on 2, by up to 5.4%. Against the same class the stated bound is also far tighter than the constant-motion test suggests, a median slack of 2.40 rather than 4.92, and its readable form 4.11 rather than 7.85.

**The bound at the paper's drifts.** The four calibrations behind the quoted interval are two window conventions by two tolerances: matching the window on predicted depth gives 0.062 and 0.081, on measured steps 0.185 and 0.235, the lower of each at a quarter of a decade and the upper at a third. In $\bar{\theta}$ the same four read 0.076, 0.100, 0.232 and 0.298, medians over the same 50 spectra. Theorem 3 itself gives $\bar{\theta}$, and the four drifts read it under the even spread of §4.2. A second route, through $\theta^p \le \zeta\delta$, gives the interval $[0.024, 0.059]$, which is not these divided by one factor: it takes the predicted-depth pair alone, where $\bar{\theta}^\star$ runs 0.072 to 0.124 across the spectra, and divides through $\zeta$, which runs 2.11 to 3.02, pairing the smallest $\bar{\theta}^\star$ with the largest $\zeta$ at its lower end and the reverse at its upper, so it carries the theorem's threshold into the drift through $\theta^p \le \zeta\delta$, which bounds $\theta^p$ and not $\bar{\theta}$, so membership certifies nothing by proof; it is not a restatement of $[0.06, 0.24]$. The ladder to 118M and the released checkpoints put two cells inside it, holding at 0.016 decades.

**The proved statistic on the ladder.** First, $\bar{\theta}/\delta$ runs 8.18 to 46.3, median 17.6, on every rung: reading at $\delta$ understates the proved statistic here by one to one and a half orders of magnitude, where on classifier heads it overstates it, so the word conservative belongs to that family and not to this one. Second, $\bar{\theta}$ itself falls with the parameter count, 10.18 at 0.11M to 0.371 at 118M, Spearman $-0.986$, with two adjacent inversions, 3.6M above 3.2M and 51M above 43M, and seven of the twelve rungs exceed one: the theorem's own hypothesis is missed at the bottom of the ladder and met at the top, while the read errs 0.0158 to 0.0726 decades across the whole of it with no trend against $\bar{\theta}$. So on the ladder Theorem 3 covers no cell, $\bar{\theta}$ falling with scale to 0.371 at 118M, above 0.298, while the read holds at every size: the theorem's hypotheses are sufficient for the read, which holds beyond them.

**Both tolerances.** The tolerance matters as little: scored at a third of a decade instead of a quarter, 17 of the 555 verdicts change, every one at a drift above the interval, so the 90 certified cells hold at both tolerances and neither certifies a failing cell.

**The nearest width on MNIST.** Each of the 40 cells of the complete MNIST width sweep, the 38 of Table 4 and two width-256 cells that finished after its snapshot, is held out against the nearest width's measured curve on its own ladder of one data draw and seed: the read errs a median 0.024 decades against 0.095 and is closer on all 40, and every maximal set of cells sharing no curve favours it 16 to none, $p = 3.1 \times 10^{-5}$.

**The competitor at a tenth.** The read beats both competitors on 20 of 20 at the first tenth, fifth and third of the curve, at a median 0.029 decades against the eleven-family ensemble's 0.781 and a single power law's 0.683 at the tenth; the ensemble takes 17 of the 20 only once given half.

**One spectrum across a thirtyfold range of step size.** One initialisation spectrum is reused across nine step sizes spanning a factor of 30, from $f = \eta\lambda_{\max} = 0.05$ to 1.5, on 12 cells scored at a matched 2.2 decades of descent and with nothing refitted. The median rate error over those cells varies by 1.4× across the whole range, between 0.021 and 0.031 decades, and its log–log slope in $\eta$ is 0.017.

**The condition number as a certificate.** On this log-time grid the knobs that generate the failures order them better than the drift: the condition number at initialisation ranks the later error at 0.92 and, thresholded at its first failure, certifies 4,395 heads to the drift's 3,339. Nor does $\kappa$ itself, which the exact read carries and the 64-node read does not resolve, stand in for read 2: under the same held-out test on the 332 cells whose $\kappa$ could be recovered, certifying on $\kappa$ alone falsely certifies 19 failing cells and covers 37 of the 257 that hold, where the drift, thresholded the same way on the held-out folds rather than at 0.24, falsely certifies 1 and covers 146.

**A corpus-fitted law at cross-entropy.** Fitted leaving this rung out, a rule that watches nothing of the run it predicts costs $0.033$ decades at $118$M against the read's $0.533$, beating it on all $15$ cells.

### E.9 The read's resolution in time

A $q$-node read follows the linearised curve only while its nodes resolve the part of the spectrum the residual still occupies. For the Lanczos approximation of $e^{-\tau A}v$ that takes $q \gtrsim \sqrt{\tau\lambda_{\max}}$ (Hochbruck & Lubich, 1997), and here $\tau \approx 2\eta k$, so a $q$-node read serves about $q^2/(2\eta\lambda_{\max})$ steps. Measured against reads of $512$ and $256$ nodes on three runs at $1{,}024$ examples (Table 13), the step at which a $q$-node read leaves the reference grows as $q^{2.1}$ to $q^{2.2}$ up to $64$ nodes, close to the law's $q^2$, and faster beyond: a forecast of $k$ steps needs about $\sqrt{0.8k}$ nodes, and $128$ nodes hold past $34{,}000$ steps. On the $35$ stored runs past $5{,}000$ steps the $64$-node read leaves the $128$-node one at steps $6{,}968$ to $9{,}207$ at $256$ examples and $4{,}682$ to $6{,}868$ from $1{,}024$ up. The $3{,}000$-step windows lie inside the $64$-node range, the two reads differing by at most $0.033$ decades over $6{,}000$ steps at $118$M and $403$M. Every run past $5{,}000$ steps is scored with its $128$-node read, the $30{,}000$-step runs with the finest stored. Rescored, every exit at $256$ examples survives, the standard ones $8$ to $31\%$ earlier and the $\mu$P ones at most $3\%$ (Table 12), and no $15{,}000$-step SGD run leaves tolerance (App. E.7).

## F Protocol

**Training and scoring.** Populations are in Table 4, cells, holding and the horizon in §2. Every run is full-batch gradient descent at the fixed step $\eta = 0.3/\lambda_{\max}(G(0))$ unless stated, for a window of $3{,}000$ steps (§5.1); the runs from $118$M to 3B are trained in float32 (Table 4). Before any run, float32 was checked against the ladders' float64 rows on the $3.2$M rung under both parameterisations: the read's log-time error moved by $3.0 \times 10^{-7}$ and $3.1 \times 10^{-7}$ decades, inside a gate of $0.005$. Scoring is in log time on each population's stored grid (App. E.2): of eleven families fitted by weighted least squares, the $7$ to $11$ that converge per cell are combined by inverse in-sample error.

**The read.** The procedure of §3, at $q = 64$ nodes, $128$ or more past $5{,}000$ steps (App. E.9). For small $m$ the nodes and weights are exact, by forming the $m \times m$ Gram from $m$ products and diagonalising it, the route of every from-scratch and head cell; released checkpoints run the Lanczos steps, fully reorthogonalised. Its exact checks are in App. E.8. The two tests are the drift $\delta$ of §4.2, read at a tenth or at the window's end and certified at or below $0.24$, and the growth of $G$ along the residual at step $300$ (App. E.4). The drift is estimated from at least $32$ random probes, until its relative standard error is below $2\%$.