OPTIMAL TRAINING-TIME SCALING IN GRADUAL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In gradual adaptation, how should the training time on each task change as the number of intermediate tasks increases? We study this question for overparameterized linear regression tasks that change smoothly and share a zero-loss solution. With N tasks and training time s_N on each, the final learning progress converges to a continuum curve when $Ns_N \to \tau$. The limiting progress is $\Theta(\tau)$ for small τ and $\Theta(\tau^{-1})$ for large τ, so both very short and very long training produce little progress. It follows that optimal per-task training times scale as $s_N^\star = \Theta(N^{-1})$, equivalently $Ns_N^\star = \Theta(1)$. Experiments on gradually rotated MNIST and a natural Yearbook time shift are consistent with less per-task training as the path is divided more finely.

1 INTRODUCTION

Many learning systems must process sequences of slowly changing training tasks. In gradual domain adaptation, intermediate distributions guide a learner from a source domain to a target domain (Kumar et al., 2020; Chen & Chao, 2021; Wang et al., 2022a). In curriculum learning, related tasks are ordered by difficulty or structure (Bengio et al., 2009; Kumar et al., 2010; Graves et al., 2017; Weinshall et al., 2018; Hacohen & Weinshall, 2019; Klink et al., 2022; Huang et al., 2022). In continual test-time adaptation, a model is repeatedly updated as its deployment data evolve (Sun et al., 2020; Wang et al., 2021; 2022b; Niu et al., 2022; Zhang et al., 2022; Niu et al., 2023; Döbler et al., 2023; Yuan et al., 2023). We use *gradual adaptation* for this setting of sequential learning over nearby tasks, which also falls within the broader scope of continual learning (Parisi et al., 2019; De Lange et al., 2022; van de Ven et al., 2022).

The same gradual adaptation process can be divided into different numbers of tasks. More tasks make adjacent problems closer, but also require the learner to update more frequently. Suppose the sequence contains N tasks and the learner trains for time s_N on each task. The task count N controls the spacing between adjacent tasks, while s_N controls the amount of optimization at each task. We ask: **As the number of tasks N increases, how long should each task be trained to maximize the learning progress after the entire sequence?**

Existing theory studies training time and interference across sequential tasks from different angles. Early-stopping theory treats training time as a statistical regularizer (Yao et al., 2007; Raskutti et al., 2014; Ali et al., 2019), and stability analyses relate training duration to generalization (Hardt et al., 2016). These results explain why stopping before convergence can improve prediction from a finite sample. Our effect is instead present for noiseless population losses and concerns the state passed from one task to the next.

Domain-adaptation theory studies when transfer across distributions is possible, while practical methods align features or predictions between source and target domains (Ben-David et al., 2010; Ganin et al., 2016; Hoffman et al., 2018; Saito et al., 2018). Gradual self-training inserts intermediate distributions and studies how their number and placement affect target error (Kumar et al., 2020; Chen & Chao, 2021; Wang et al., 2022a). These analyses represent each local adaptation by an empirical risk minimization step. They do not treat the amount of optimization at each intermediate task as a variable that changes with the resolution of the path.

Curriculum learning studies the order and pace at which examples or tasks are presented (Bengio et al., 2009; Kumar et al., 2010; Graves et al., 2017; Weinshall et al., 2018; Hacohen & Weinshall,

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.

2019). Continual-learning methods control interference through parameter regularization, stored examples, and gradient projections (Kirkpatrick et al., 2017; Zenke et al., 2017; Ritter et al., 2018; Benzing, 2022; Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019; Farajtabar et al., 2020). Recent theory studies how training regimes and regularized updates affect forgetting (Mirzadeh et al., 2020; Zhao et al., 2024; Levinstein et al., 2025), including intermediate training regimes (Lesort et al., 2023; Graldi et al., 2025; Mori et al., 2025). Exact analyses of linear regression characterize forgetting under fixed and cyclic task sequences (Evron et al., 2022; Jeong et al., 2023), while recent work studies continual models under SGD and task distributions (Evron et al., 2026; Xu & Ma, 2026). These works establish that task order and training regime matter. We isolate a different question by holding the underlying task path fixed and coupling its resolution N to the local training time s_N.

We study this problem in overparameterized linear regression. Task t uses the population squared loss

$$\mathcal{L}_t(w) = \frac{1}{2}\mathbb{E}_t\left[(x^\top w - y)^2\right],$$

and all tasks share a zero-loss solution $w^\star$, so that $y = x^\top w^\star$ almost surely on every task.

To compare different task counts, we index the complete gradual adaptation process by $t \in [0,1]$ and select N tasks at $t_j = j/N$, $j = 1, \dots, N$. Let $H(t) = \mathbb{E}_t[xx^\top]$ be the Hessian of task t. Gradient flow has an exact exponential solution: training task j for time s_N updates the parameter error as

$$v_j = e^{-s_N H(t_j)} v_{j-1}, \qquad v_N = e^{-s_N H(t_N)} \cdots e^{-s_N H(t_1)} v_0.$$

Let

$$\mathcal{M} = \bigcap_{t\in[0,1]} \arg\min_w \mathcal{L}_t(w)$$

be the common solution set, and suppose that the initial model w_0 already fits the first task. With $\bar{w} = \mathrm{Proj}_{\mathcal{M}}(w_0)$ and $v_j = w_j - \bar{w}$, the affine structure gives $\mathrm{dist}(w_j, \mathcal{M}) = \|v_j\|$. We therefore measure the learning progress over the complete sequence by

$$\mathcal{P}_{N,s_N}(v_0) = \mathrm{dist}(w_0, \mathcal{M})^2 - \mathrm{dist}(w_N, \mathcal{M})^2 = \|v_0\|^2 - \|v_N\|^2.$$

The optimal per-task training time $s_N^\star(v_0)$ maximizes this quantity.

Our first theorem gives the continuous limit as the task count grows. If $Ns_N \to \tau < \infty$, the discrete updates converge to

$$\dot{x}_\tau(t) = -\tau H(t) x_\tau(t), \qquad x_\tau(0) = v_0.$$

Here τ is the effective training time spent along the complete task path. The corresponding progress converges to

$$\mathcal{P}_\tau(v_0) = \|v_0\|^2 - \|x_\tau(1)\|^2,$$

which satisfies the exact identity

$$\mathcal{P}_\tau(v_0) = 2\tau \int_0^1 x_\tau(t)^\top H(t) x_\tau(t)\, dt.$$

This gives a common limiting curve $\tau \mapsto \mathcal{P}_\tau(v_0)$ for comparing different choices of N and s_N.

Our second theorem characterizes both ends of this limiting curve. Whenever the initial model is not already in the common solution set,

$$\mathcal{P}_\tau(v_0) = \Theta(\tau) \quad (\tau \downarrow 0), \qquad \mathcal{P}_\tau(v_0) = \Theta(\tau^{-1}) \quad (\tau \to \infty).$$

The relevant resource is the total effective training time Ns_N, whose finite limit is τ. When τ is small, the model barely moves. When τ is large, it remains close to the current task's zero-loss subspace at every point along the path. As these subspaces rotate gradually, the dominant effect is to change the direction of the remaining error rather than its norm. This explains the $\Theta(\tau^{-1})$ decay of the progress produced by overtraining.

Thus, both $\tau \downarrow 0$ and $\tau \to \infty$ yield vanishing progress. Only

$$Ns_N = \Theta(1)$$

can retain a positive limiting amount of progress as the task count grows. It follows that

$$s_N^\star(v_0) = \Theta(N^{-1}).$$

Dividing the same task path more finely therefore requires reducing the training time on each task by the same proportion.

Keeping s_N fixed while increasing N therefore does more than refine the discretization: it also sends the total effective training time Ns_N to infinity, where the progress vanishes. Thus, in the present setting, task resolution and per-task optimization time cannot be treated as independent design choices.

We also characterize the first-order behavior in the regime $Ns_N \to \infty$. For every fixed finite per-task training time, the first-order coefficient is strictly larger than the coefficient obtained by training each task to convergence. If progress is instead measured through any strictly increasing function of squared distance, the ordering of training times and the optimal training time remain unchanged.

Contributions. The paper makes four main contributions.

- We formulate per-task training time as an optimization problem over the task count N and training time s_N, using the reduction in distance to the common solution set after all tasks as the learning objective.
- We derive the limiting learning curve controlled by the effective training time τ and prove its $\Theta(\tau)$ growth for small τ and $\Theta(\tau^{-1})$ decay for large τ.
- We prove that only $Ns_N = \Theta(1)$ can retain positive limiting progress as the task count grows, yielding the optimal scaling $s_N^\star(v_0) = \Theta(N^{-1})$.
- We test the prediction on rotated MNIST and temporal Yearbook data. Across both settings, finer task divisions favor less training per task overall, including under natural temporal shift.

2 Problem Setup

We first specify the overparameterized regression tasks and their smoothly varying zero-loss spaces. We then define the sequential gradient-flow updates, learning progress relative to the common solution set, and the optimal per-task training time.

2.1 A smooth family of regression tasks

Let $t \in [0,1]$ index a gradually changing task. On task t, a feature–label pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ follows a distribution $\mathcal{D}_t$. We assume that there is a parameter $w^\star \in \mathbb{R}^d$ such that $y = x^\top w^\star$ almost surely under every $\mathcal{D}_t$. The population squared loss is therefore

$$\begin{aligned} \mathcal{L}_t(w) &:= \frac{1}{2}\mathbb{E}_{\mathcal{D}_t}\left[(x^\top w - y)^2\right] \\ &= \frac{1}{2}(w - w^\star)^\top H(t)(w - w^\star), \qquad H(t) := \mathbb{E}_{\mathcal{D}_t}[xx^\top]. \end{aligned} \tag{1}$$

Here $H(t)$ is both the population Hessian and the feature second moment. For every $z \in \mathbb{R}^d$,

$$z^\top H(t) z = \mathbb{E}_{\mathcal{D}_t}\left[(x^\top z)^2\right] \geq 0, \tag{2}$$

so $H(t)$ is positive semidefinite. Its kernel consists of parameter perturbations whose linear predictions remain unchanged on task t:

$$V(t) := \ker H(t) = \{z : x^\top z = 0 \text{ almost surely under } \mathcal{D}_t\}.$$

A parameter change along $z \in V(t)$ leaves the predictions on task t unchanged. For a finite regression dataset with design matrix $X(t) \in \mathbb{R}^{n_t \times d}$ and labels $y(t) = X(t)w^\star$, the same representation holds with $H(t) = X(t)^\top X(t)/n_t$. The overparameterized regime corresponds to $\operatorname{rank} H(t) < d$, so each task leaves some parameter directions unconstrained.

We assume that $H : [0,1] \to \mathbb{R}^{d \times d}$ is a C^3 path of symmetric positive-semidefinite matrices with constant rank. Its positive eigenvalues satisfy the uniform bounds

$$0 < \mu \leq \lambda \leq L < \infty. \tag{3}$$

Constant rank ensures that $V(t)$ varies smoothly with t, while Equation (3) keeps every nonzero curvature between μ and L. Let

$$P(t) = \mathrm{Proj}_{V(t)}, \qquad Q(t) = I - P(t), \qquad \mathcal{K} = \bigcap_{t\in[0,1]} V(t).$$

Thus $P(t)$ projects onto the parameter directions unconstrained by task t, while $Q(t)$ projects onto their orthogonal complement. The minimizer set of task t is $\mathcal{M}(t) = w^\star + V(t)$, and the solution set shared by the full path is $w^\star + \mathcal{K}$.

2.2 Sequential training and learning progress

Select tasks on the uniform grid $t_j = j/N$, $j = 0, \dots, N$, and write $v = w - w^\star$. On task t_j, gradient flow for Equation (1) satisfies

$$\frac{d}{du} v(u) = -H(t_j) v(u).$$

Training for time $s \geq 0$ therefore applies the exact update

$$v_j = e^{-sH(t_j)} v_{j-1}. \tag{4}$$

Throughout, s denotes continuous gradient-flow time. A run of k gradient-descent steps with learning rate η approximates Equation (4) with $s \approx \eta k$ in the small-step regime. For task-wise fitting to convergence, we define $e^{-\infty H(t)} := P(t)$.

On the grid above, suppose the learner processes $t_1, \dots, t_N$ in order and runs gradient flow for the same time s on every task. Repeated application of Equation (4) gives

$$\Phi_{N,s} := e^{-sH(t_N)} e^{-sH(t_{N-1})} \cdots e^{-sH(t_1)}, \qquad v_N = \Phi_{N,s} v_0. \tag{5}$$

We begin immediately after the initial task has been fitted, so $v_0 \in V(0)$. This convention focuses the analysis on the learning progress contributed by the N subsequent tasks.

Decompose the initial error orthogonally as

$$v_0 = v_0^{\parallel} + v_0^{\perp}, \qquad v_0^{\parallel} \in \mathcal{K}, \qquad v_0^{\perp} \in \mathcal{K}^{\perp}.$$

Since $\mathcal{K} \subseteq \ker H(t)$ for every t, each update fixes $v_0^{\parallel}$ and preserves $\mathcal{K}^{\perp}$. Hence the distances to the common solution set before and after training are

$$\mathrm{dist}(w^\star + v_0, w^\star + \mathcal{K}) = \|v_0^{\perp}\|, \qquad \mathrm{dist}(w^\star + \Phi_{N,s} v_0, w^\star + \mathcal{K}) = \|\Phi_{N,s} v_0^{\perp}\|.$$

Learning progress therefore depends only on $v_0^{\perp}$. To simplify notation, we now rename $v_0^{\perp}$ as v_0. Together with the assumption that the initial task has been fitted, this gives $v_0 \in V(0) \cap \mathcal{K}^{\perp}$.

The central quantity in our analysis is the reduction in squared distance to the common solution set after the N subsequent tasks:

$$\mathcal{P}_{N,s}(v_0) = \|v_0\|^2 - \|\Phi_{N,s} v_0\|^2 \geq 0, \qquad v_0 \in V(0) \cap \mathcal{K}^{\perp}. \tag{6}$$

This distance has a uniform loss interpretation over the task path. Define the worst population loss of a model w along the path by

$$\mathcal{R}_{\mathrm{path}}(w) := \sup_{t\in[0,1]} \mathcal{L}_t(w).$$

Proposition 1 (Loss interpretation of the learning-progress measure)**.** *Suppose $\mathcal{K}^{\perp} \neq \{0\}$, and let*

$$\bar{H} := \int_0^1 H(t)\, dt, \qquad c_H := \frac{1}{2} \lambda_{\min}\big(\bar{H}\big|_{\mathcal{K}^{\perp}}\big)\,.$$

Then $c_H > 0$ and, for every $w \in \mathbb{R}^d$,

$$c_H \ \mathrm{dist}(w, w^\star + \mathcal{K})^2 \leq \mathcal{R}_{\mathrm{path}}(w) \leq \frac{L}{2} \ \mathrm{dist}(w, w^\star + \mathcal{K})^2. \tag{7}$$

Consequently, $\mathrm{dist}(w_m, w^\star + \mathcal{K}) \to 0$ *for a sequence* (w_m) *if and only if* $\sup_{t\in[0,1]} \mathcal{L}_t(w_m) \to 0$.

Thus squared distance to the common solution set is equivalent, up to path-dependent constants, to the worst population loss along the task path. Equation (6) records the reduction of this state measure along the task sequence. The proof is in Section A.1.

Proposition 2 (Exact decomposition of learning progress)**.** *For every N and $s \in [0, \infty]$,*

$$\mathcal{P}_{N,s}(v_0) = \sum_{j=1}^{N} \left\langle v_{j-1}, \left(I - e^{-2sH(t_j)}\right) v_{j-1} \right\rangle. \tag{8}$$

Proof. Since $H(t_j)$ is symmetric, Equation (4) gives

$$\|v_{j-1}\|^2 - \|v_j\|^2 = \left\langle v_{j-1}, \left(I - e^{-2sH(t_j)}\right) v_{j-1} \right\rangle.$$

Summing over j telescopes to Equation (6). □

Each summand is exactly the reduction in squared error produced on one task. Equation (8) exposes two coupled roles of the per-task training time. A larger s removes more error on the current task, while also changing the state inherited by every later task. Not fitting the current task completely leaves a residual on which subsequent, slightly different tasks can act. Across a dense task sequence, these later reductions can accumulate and produce more progress over the full path than exact fitting. Section 3 quantifies this effect.

All subsequent results characterize how $\mathcal{P}_{N,s}(v_0)$ depends on the task count N and the per-task training time s.

For each task count N, define an optimal per-task training time by

$$s_N^\star(v_0) \in \arg\max_{s \in [0,\infty]} \mathcal{P}_{N,s}(v_0). \tag{9}$$

Because $\Phi_{N,s}$ has a limit as $s \to \infty$, the maximum in Equation (9) is attained.

3 Training-Time Scaling in Gradual Adaptation

We first pass from the discrete task sequence to a continuum learning curve indexed by the effective training time $\tau = Ns_N$. We then characterize its small- and large-budget limits and use them to identify the optimal scaling of the per-task training time.

3.1 The limiting learning curve

For an effective training time $\tau \geq 0$, let $x_\tau : [0, 1] \to \mathbb{R}^d$ solve

$$\dot{x}_\tau(t) = -\tau H(t) x_\tau(t), \qquad x_\tau(0) = v_0, \tag{10}$$

and define

$$\mathcal{P}_\tau(v_0) := \|v_0\|^2 - \|x_\tau(1)\|^2. \tag{11}$$

Theorem 1 (Limit of dense task sequences)**.** *Under the assumptions of Section 2, if $Ns_N \to \tau \in [0, \infty)$, then*

$$\Phi_{N,s_N} v_0 \longrightarrow x_\tau(1), \qquad \mathcal{P}_{N,s_N}(v_0) \longrightarrow \mathcal{P}_\tau(v_0). \tag{12}$$

For every $T < \infty$, the finite-sequence approximation satisfies

$$\sup_{0 \leq \tau \leq T} \left\| \Phi_{N,\tau/N} v_0 - x_\tau(1) \right\| = O_T(N^{-1}), \tag{13}$$

$$\sup_{0 \leq \tau \leq T} \left| \mathcal{P}_{N,\tau/N}(v_0) - \mathcal{P}_\tau(v_0) \right| = O_T(N^{-1}). \tag{14}$$

Moreover,

$$\mathcal{P}_\tau(v_0) = 2\tau \int_0^1 x_\tau(t)^\top H(t) x_\tau(t)\, dt. \tag{15}$$

Proof sketch. Put $b_N = Ns_N$. Since $b_N \to \tau$, each update satisfies

$$e^{-s_N H(t_j)} = I - \frac{b_N}{N} H(t_j) + O(N^{-2})$$

uniformly along the path. The one-step error relative to Equation (10) is $O(N^{-2}) + O(|b_N - \tau|/N)$. Because every update is a contraction, these errors are not amplified by the remaining product. After N steps the endpoint error is therefore $O(N^{-1}) + O(|b_N - \tau|)$, which tends to zero. When $b_N = \tau$ on a compact interval, this also gives Equation (13). The progress rate follows because all trajectories are uniformly bounded. Differentiating $\|x_\tau(t)\|^2$ and integrating from 0 to 1 gives Equation (15). Complete estimates are in Section A.

3.2 Optimal per-task training time

The condition $v_0 \in V(0) \cap \mathcal{K}^\perp$ means that the initial point fits the source task and that components invisible to every task have been removed. Hence a nonzero v_0 is detected by at least one later task.

Theorem 2 (Optimal per-task training time)**.** *Let $v_0 \in V(0) \cap \mathcal{K}^\perp$ be nonzero. Then:*

(i) The limiting progress is strictly positive for every finite $\tau > 0$ and satisfies

$$\mathcal{P}_\tau(v_0) = \Theta(\tau) \quad (\tau \downarrow 0), \qquad \mathcal{P}_\tau(v_0) = \Theta(\tau^{-1}) \quad (\tau \to \infty). \tag{16}$$

(ii) For every sequence of per-task training times s_N,

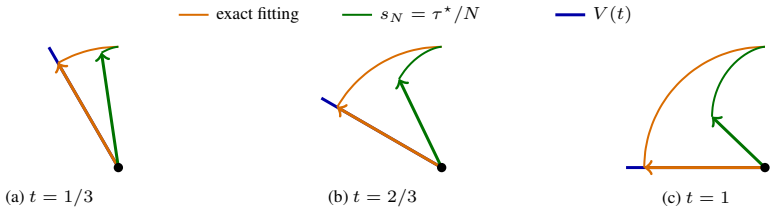
$$\begin{aligned} Ns_N \to 0 &\implies \mathcal{P}_{N,s_N}(v_0) \to 0, \\ Ns_N \to \tau \in (0,\infty) &\implies \mathcal{P}_{N,s_N}(v_0) \to \mathcal{P}_\tau(v_0) > 0, \\ Ns_N \to \infty &\implies \mathcal{P}_{N,s_N}(v_0) \to 0. \end{aligned} \tag{17}$$

(iii) The limiting curve attains its maximum at one or more points in $(0,\infty)$. For all sufficiently large N, every optimizer in Equation (9) is finite and satisfies $s_N^\star(v_0) = \Theta(N^{-1})$, and every accumulation point of $Ns_N^\star(v_0)$ maximizes $\mathcal{P}_\tau(v_0)$. In particular, if the limiting curve has a unique maximizer $\tau^\star(v_0)$, then

$$s_N^\star(v_0) = \frac{\tau^\star(v_0)}{N} + o(N^{-1}). \tag{18}$$

In particular, fitting every task to convergence belongs to the third regime in Equation (17): its progress tends to zero as $N \to \infty$. By contrast, $s_N = \tau/N$ with $0 < \tau < \infty$ retains strictly positive limiting progress.

Proof idea. The distinction is easiest to see by comparing complete trajectories rather than a single update. See Figure 1. Put $h = N^{-1}$. After exact fitting, $v_j \in V(t_j)$. Because adjacent zero-loss spaces differ by only $O(h)$, the next task exposes only an $O(h)$ normal component. Projecting this component away turns the vector by $O(h)$, but changes its squared length by only $O(h^2)$. Across $N = h^{-1}$ tasks, the vector therefore mainly follows the moving zero-loss space along an almost constant-radius path, with only $O(N^{-1})$ total progress.



(a) $t = 1/3$ (b) $t = 2/3$ (c) $t = 1$

Figure 1: Limiting trajectories for the rotating-task example in Section 4. All panels use the same scale. The black dot is the common solution and the blue ray is the current zero-loss space. Exact fitting follows this space along the outer orange arc without approaching the dot. At the optimal scale $s_N = \tau^\star/N$, finite training lags behind the current zero-loss space but moves steadily inward. By the final task it is visibly closer to the common solution.

With $s_N = \tau h$, no task removes its visible component completely:

$$e^{-s_N H(t_j)} v_{j-1} = v_{j-1} - \tau h H(t_j) v_{j-1} + O(h^2).$$

The remaining component need not shrink to $O(h)$ before the next task arrives. Consequently the $O(h)$ decreases in squared length can accumulate over $N = h^{-1}$ tasks, giving the nonzero limit in Equation (15). In the figure this is the difference between merely following the outer arc and moving inward.

The two failures occur on opposite sides of this balance. If $Ns_N \to 0$, the total amount of training vanishes. If $Ns_N \to \infty$, the model is driven increasingly close to each current zero-loss space. The appendix proves the uniform estimate

$$\mathcal{P}_{N,s_N}(v_0) = O\left(\frac{1}{N(1 - e^{-\mu s_N})}\right).$$

Its right-hand side tends to zero whenever $Ns_N \to \infty$. Fix any finite $\bar{\tau} > 0$. The choice $s_N = \bar{\tau}/N$ gives strictly positive limiting progress. The two tail bounds rule out $Ns_N^\star \to 0$ and $Ns_N^\star \to \infty$, so every optimizer satisfies $Ns_N^\star = \Theta(1)$. Uniform convergence on this bounded range then implies that every accumulation point of $Ns_N^\star$ maximizes $\mathcal{P}_\tau(v_0)$. Complete expansions and uniform remainder estimates are given in Sections A.3 and A.3.1.

First-order comparison at fixed per-task time. The main theorem concerns $s_N = \Theta(N^{-1})$, for which progress remains of constant order. A different question is what happens when the per-task time is fixed as N grows. Let $C_s(v_0) = 2\langle v_0, \mathcal{G}_s v_0\rangle$ and $C_H(v_0) = 2\langle v_0, \mathcal{G}_H v_0\rangle$, where the explicit path-dependent operators are defined in Sections A.3 and A.3.1.

Proposition 3 (First-order comparison in the overtraining regime)**.** *For $v_0 \in V(0) \cap \mathcal{K}^\perp$ and every fixed $s \in (0, \infty]$,*

$$\mathcal{P}_{N,s}(v_0) = \frac{C_s(v_0)}{N} + o(N^{-1}). \tag{19}$$

If $s_N \downarrow 0$ and $Ns_N \to \infty$, then

$$\mathcal{P}_{N,s_N}(v_0) = \frac{C_H(v_0)}{Ns_N} + o((Ns_N)^{-1}). \tag{20}$$

Moreover, $C_s(v_0) > C_\infty(v_0)$ for every fixed finite s. Hence $\mathcal{P}_{N,s}(v_0) > \mathcal{P}_{N,\infty}(v_0)$ for all sufficiently large N.

The strict comparison has a direct interpretation. Along a positive-eigenvalue direction of $H(t)$ with eigenvalue λ, finite training leaves the fraction $\rho = e^{-s\lambda}$ of the incoming error. Its contribution to the first-order coefficient can be written as

$$\frac{1}{2}\coth(s\lambda/2) = \frac{1}{2} + \frac{\rho}{1-\rho}. \tag{21}$$

Exact fitting sets $\rho = 0$. Finite fitting adds the positive second term, which records residual error that later, slightly different tasks can still reduce. Thus fixed finite fitting and exact fitting both have progress of order N^{-1}, but the finite-time coefficient is strictly larger. The path-dependent coefficients and proof are in Section A.3.1.

Robustness to monotone distance transformations. This conclusion does not depend on using squared distance itself. For any strictly increasing $f : [0, \infty) \to \mathbb{R}$, define $\mathcal{P}^f_{N,s}(v_0) := f(\|v_0\|^2) - f(\|\Phi_{N,s} v_0\|^2)$. Writing $r = \|v_0\|^2$ gives the exact identity

$$\mathcal{P}^f_{N,s}(v_0) = f(r) - f\big(r - \mathcal{P}_{N,s}(v_0)\big). \tag{22}$$

The right-hand side is strictly increasing in $\mathcal{P}_{N,s}$, so signs, comparisons across training times, and maximizers are unchanged. The conclusion need not hold for anisotropic distances. An explicit reversal is given in Section A.4.

4 Numerical Experiments

Our experiments serve two complementary purposes. The regression experiments verify the theory within its assumptions, while Rotated MNIST and Yearbook deliberately depart from those assumptions to test whether the predicted training-time trend persists under nonlinear models, stochastic optimization, finite samples, cross-entropy losses, and natural distribution shift. In every experiment, the underlying path is held fixed while the number of visited tasks changes.

Rotating regression tasks. We first consider a two-dimensional rank-one problem whose task direction rotates smoothly by $\pi/2$. With $s_N = \tau/N$, the finite-task progress curves approach a common curve with an interior optimum at $\tau^\star \approx 2.20$, and $Ns_N^\star$ approaches the same value. Training every task to convergence instead gives progress of order N^{-1}. The closed-form derivation and a higher-dimensional experiment are in Section B.

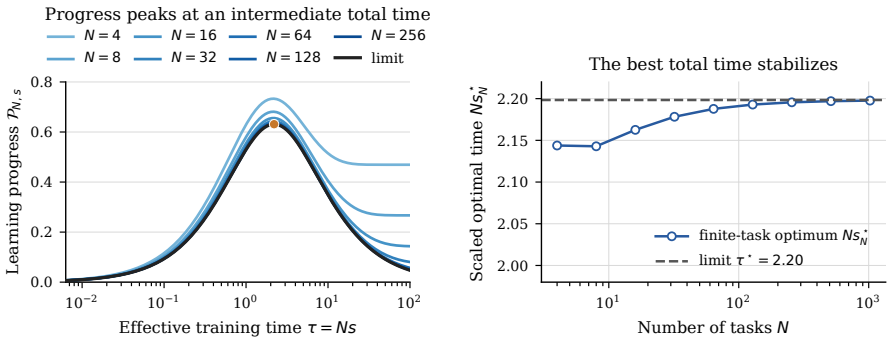


Figure 2: **Training time has an intermediate optimum on a rotating task path. Left:** Finite-task progress curves approach the same rise-and-decay limit. **Right:** The best total training time $Ns_N^\star$ approaches $\tau^\star \approx 2.20$, so the best per-task time is approximately $2.20/N$.

Rotated MNIST. We next train an MLP on upright MNIST and update it along a fixed rotation path from $0°$ to $60°$. Every task count is evaluated on the same 13 angles. Across ten seeds, a log–log fit to the empirical optima gives $k^\star \propto N^{-0.84}$, with bootstrap 95% interval $[0.76, 1.03]$ for the exponent. At $N = 32$, the final-angle loss continues to improve after the mean loss across the rotation path begins to increase. Full details are in Section B.3.

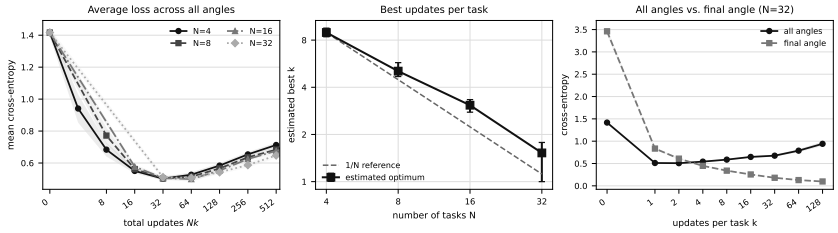


Figure 3: **Finer rotation steps favor fewer updates per task. Left:** Mean loss across the rotation path against total updates Nk. **Middle:** Estimated optimal updates per task and the N^{-1} reference. **Right:** At $N = 32$, final-angle loss keeps decreasing after the mean loss across all angles turns upward. Error bars and bands are 95% intervals over ten seeds.

Yearbook over time. Finally, we use Yearbook photographs spanning 1930–2013 (Ginosar et al., 2017; Yao et al., 2022). Fixed-width temporal smoothing defines the same path for every N. Across five seeds, the estimated optimum falls overall from 79.9 updates per task at $N = 4$ to 5.0 at $N = 32$; the fitted exponent is 1.37, with bootstrap 95% interval $[0.79, 1.66]$. At $N = 32$, increasing k from 4

to 128 lowers target-time cross-entropy from 0.216 to 0.051 but raises the mean across the time path from 0.325 to 0.535. Full details and limitations are in Section B.4.

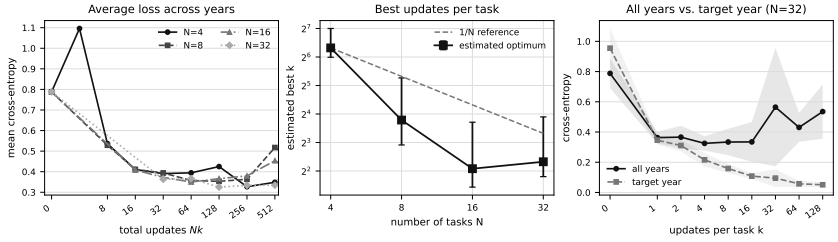


Figure 4: **Finer temporal steps favor less training per task overall. Left:** Mean loss across years against total updates Nk. **Middle:** Estimated optimal updates per task and the N^{-1} reference. **Right:** At $N = 32$, target-year loss keeps decreasing while the mean loss across years eventually increases. Error bars and bands are 95% intervals over five seeds.

Across all settings, increasing N refines a fixed path between unchanged endpoints. Rotating regression recovers $s_N^\star \propto N^{-1}$. On Rotated MNIST, the estimated optimal total update budget remains between 35.8 and 49.2 as N increases from 4 to 32. Yearbook shows an overall decrease in optimal updates per task under natural temporal shift. In both nonlinear experiments, continued improvement on the final task can coexist with increasing average loss along the path.

5 Discussion

Current-task fit and whole-path progress are distinct: longer local training improves the current task but changes the state inherited by later tasks. Hence fitting every task to convergence can reduce whole-path progress even for noiseless population losses with a shared solution. Unlike conventional early stopping, which limits sample-noise fitting to improve generalization (Yao et al., 2007; Raskutti et al., 2014; Ali et al., 2019), our mechanism preserves residual error that later tasks can reduce, connecting local fit to the plasticity–forgetting tension in continual learning (Parisi et al., 2019; De Lange et al., 2022).

Under population gradient flow along a smooth, constant-rank Hessian path with a shared solution and uniform spectral gap, optimal per-task time scales as N^{-1} when N refines a fixed path. Strictly increasing transformations of squared Euclidean distance preserve the ranking of training times, and rotating regression recovers both a finite optimal total time and $s_N^\star \propto N^{-1}$.

Rotated MNIST and Yearbook extend the comparison beyond the theory's assumptions. In both, the fitted-exponent confidence interval contains 1, and continued training can improve the target task while worsening full-path performance. We therefore treat the experiments as evidence for the predicted qualitative trend, not exact verification of the N^{-1} law.

Gradual adaptation is thus a time-scale matching problem: local optimization time must track how quickly the task distribution changes. This matters for gradual self-training, test-time adaptation, and continual updating (Kumar et al., 2020; Wang et al., 2022a; Song et al., 2023; Dorka et al., 2023; Zhang et al., 2025). Recurring tasks, abrupt switches, and model-dependent paths remain open.

AI use statement

The authors conceived the question and developed the theory and methodology. Generative AI assisted with proof development and checking, experiment-code development and validation, literature search, and manuscript drafting and editing. The authors verified all AI-assisted content and take full responsibility for the paper.

Reproducibility statement

The appendix gives complete proofs and experimental details. The supplement provides code, fixed seeds, CSV results, and figure sources for all reported experiments.

References

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 2019. URL https://proceedings.mlr.press/v89/ali19a.html.

J. E. Avron, M. Fraas, G. M. Graf, and P. Grech. Adiabatic theorems for generators of contracting evolutions. *Communications in Mathematical Physics*, 314:163–191, 2012. doi: 10.1007/s00220-012-1504-1.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. doi: 10.1007/s10994-009-5152-4.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380.

Frederik Benzing. Unifying importance based regularisation methods for continual learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2372–2396. PMLR, 2022.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Hong-You Chen and Wei-Lun Chao. Gradual domain adaptation without indexed intermediate domains. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/45017f6511f91be700fda3d118034994-Abstract.html.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

Mario Döbler, Robert A. Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7704–7714, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Dobler_Robust_Mean_Teacher_for_Continual_and_Gradual_Test-Time_Adaptation_CVPR_2023_paper.html.

Nicolai Dorka, Tim Welschehold, and Wolfram Burgard. Dynamic update-to-data ratio: Minimizing world model overfitting. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZIkHSXzd9O7.

Itay Evron, Edward Moroshko, Rachel A. Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 2022. URL https://proceedings.mlr.press/v178/evron22a.html.

Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. From continual learning to SGD and back: Better rates for continual linear models. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–50. PMLR, 2026. URL https://proceedings.mlr.press/v313/evron26a.html.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773. PMLR, 2020.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Shiry Ginosar, Kate Rakelly, Sarah M. Sachs, Brian Yin, Crystal Lee, Philipp Krähenbühl, and Alexei A. Efros. A century of portraits: A visual historical record of american high school yearbooks. *IEEE Transactions on Computational Imaging*, 3(3):421–431, 2017. doi: 10.1109/TCI.2017.2699865.

Jacopo Graldi, Alessandro Breccia, Giulia Lanzillotta, Thomas Hofmann, and Lorenzo Noci. The importance of being lazy: Scaling limits of continual learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20326–20367. PMLR, 2025.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1311–1320. PMLR, 2017. URL https://proceedings.mlr.press/v70/graves17a.html.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2535–2544. PMLR, 2019.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234. PMLR, 2016. URL https://proceedings.mlr.press/v48/hardt16.html.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1989–1998. PMLR, 2018.

Peide Huang, Mengdi Xu, Jiacheng Zhu, Laixi Shi, Fei Fang, and Ding Zhao. Curriculum reinforcement learning using optimal transport via gradual domain adaptation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4556f5398bd2c61bd7500e306b4e560a-Abstract-Conference.html.

Halyun Jeong, Mark Kong, Deanna Needell, William Swartworth, and Rachel Ward. Nearly optimal bounds for cyclic forgetting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d72ae75abaa70a3b19c5d4f436c680d1-Paper-Conference.pdf.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Pascal Klink, Haoyi Yang, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. Curriculum reinforcement learning via constrained optimal transport. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11341–11358. PMLR, 2022. URL https://proceedings.mlr.press/v162/klink22a.html.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5468–5479. PMLR, 2020. URL https://proceedings.mlr.press/v119/kumar20c.html.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Timothée Lesort, Oleksiy Ostapenko, Pau Rodríguez, Diganta Misra, Md Rifat Arefin, Laurent Charlin, and Irina Rish. Challenging common assumptions about catastrophic forgetting and knowledge accumulation. In *Proceedings of the 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 43–65. PMLR, 2023.

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Daniel Soudry, Tomer Koren, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5712575413bb58ae6c7a17aa4f2d64a2-Abstract-Conference.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7308–7320, 2020.

Francesco Mori, Stefano Sarao Mannelli, and Francesca Mignacco. Optimal protocols for continual learning via statistical physics and control theory. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014. URL https://jmlr.org/papers/v15/raskutti14a.html.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

Junha Song, Kwanyong Park, InKyu Shin, Sanghyun Woo, Chaoning Zhang, and In So Kweon. Test-time adaptation in the dynamic world with compound domain knowledge management. *IEEE Robotics and Automation Letters*, 8(11):7583–7590, 2023. URL https://arxiv.org/abs/2212.08356.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 2020.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197, 2022. doi: 10.1038/s42256-022-00568-3.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Haoxiang Wang, Bo Li, and Han Zhao. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22784–22801. PMLR, 2022a. URL https://proceedings.mlr.press/v162/wang22n.html.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5238–5246. PMLR, 2018.

Zonghuan Xu and Xingjun Ma. From order to distribution: A spectral characterization of forgetting in continual learning. *arXiv preprint arXiv:2604.13460*, 2026.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. doi: 10.1007/s00365-006-0663-2.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yuan_Robust_Test-Time_Adaptation_in_Dynamic_Scenarios_CVPR_2023_paper.html.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yunbei Zhang, Akshay Mehra, Shuaicheng Niu, and Jihun Hamm. DPCore: Dynamic prompt coreset for continual test-time adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 75757–75778. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25bf.html.

Xuyang Zhao, Huiyuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61021–61039. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24n.html.

A Proofs and Technical Details

Contribution and relation to existing mathematical tools. Our contribution is to explain how per-task training time determines progress toward a common solution as a fixed task path is refined, and to establish the resulting optimal-time scaling law. The results and the mathematical tools used to obtain them play distinct roles. Theorem 1 uses matrix-exponential expansions and contractive stability to connect discrete task sequences to a continuum learning curve controlled by the total training time $\tau = Ns_N$. The large-budget analysis builds on Kato's parallel transport and the adiabatic expansions for contracting evolutions of Avron et al. (2012), translating tracking of the moving zero-loss space into $\Theta(\tau^{-1})$ decay of learning progress with an explicit path-dependent coefficient. Building on this analysis, Theorem 2 uses a discrete tail bound uniform in the task count and per-task training time to rule out optimal total training times tending to zero or infinity. This turns the two tails of the continuum curve into the discrete optimal-time law $s_N^\star = \Theta(N^{-1})$. Proposition 3 further quantifies residual accumulation in discrete updates through a first-order coefficient with weights $\frac{1}{2}\coth(s\lambda/2)$. For every fixed finite training time, this coefficient is strictly larger than that of exact fitting, yielding greater distance-based progress for sufficiently fine task partitions. Sections A.6 and A.7 clarify, respectively, the connection to average prediction risk and the scope of the result when exact zero curvature is perturbed by small positive curvature. Together, these results explain why, even for noiseless tasks with a common solution, fitting each task fully can mainly make the model track the moving solution space. Preserving residual error that later tasks can reduce enables greater progress toward the common solution, with optimal per-task training time decreasing as the path is refined.

A.1 Pathwise loss and common-solution distance

Proof of Proposition 1. Write $v = w - w^\star = v^\parallel + v^\perp$ with $v^\parallel \in \mathcal{K}$ and $v^\perp \in \mathcal{K}^\perp$. Since $H(t)v^\parallel = 0$ for every t,

$$\mathcal{L}_t(w) = \frac{1}{2}(v^\perp)^\top H(t) v^\perp.$$

The spectral upper bound in Equation (3) gives the right-hand inequality in Equation (7). For the lower bound, first observe that $\ker \bar{H} = \mathcal{K}$. Indeed, if $u^\top \bar{H} u = 0$, then the continuous nonnegative function $t \mapsto u^\top H(t) u$ vanishes everywhere. Positive semidefiniteness then gives $H(t)u = 0$ for all t, so $u \in \mathcal{K}$; the reverse inclusion is immediate. Hence $\bar{H}$ is positive definite on $\mathcal{K}^\perp$ and $c_H > 0$. Finally,

$$\mathcal{R}_{\mathrm{path}}(w) \geq \int_0^1 \mathcal{L}_t(w)\, dt = \frac{1}{2}(v^\perp)^\top \bar{H} v^\perp \geq c_H \|v^\perp\|^2.$$

Because $\|v^\perp\| = \mathrm{dist}(w, w^\star + \mathcal{K})$, this proves Equation (7); the convergence equivalence follows from its two bounds. □

A.2 The dense-task limit

For a total training budget $\tau \geq 0$, let $x_\tau : [0,1] \to \mathbb{R}^d$ solve

$$\dot{x}_\tau(t) = -\tau H(t) x_\tau(t), \qquad x_\tau(0) = v_0, \tag{23}$$

and define its learning progress by

$$\mathcal{P}_\tau(v_0) := \|v_0\|^2 - \|x_\tau(1)\|^2. \tag{24}$$

The following lemma proves Theorem 1, including its uniform $O(N^{-1})$ rates.

Lemma 1 (Dense-task limit)**.** *Under the assumptions of Section 2, if $Ns_N \to \tau \in [0, \infty)$, then*

$$\Phi_{N,s_N} v_0 \longrightarrow x_\tau(1), \qquad \mathcal{P}_{N,s_N}(v_0) \longrightarrow \mathcal{P}_\tau(v_0). \tag{25}$$

For every $T < \infty$,

$$\sup_{0 \leq \tau \leq T} \left\| \Phi_{N,\tau/N} v_0 - x_\tau(1) \right\| = O_T(N^{-1}), \tag{26}$$

$$\sup_{0 \leq \tau \leq T} \left| \mathcal{P}_{N,\tau/N}(v_0) - \mathcal{P}_\tau(v_0) \right| = O_T(N^{-1}). \tag{27}$$

Moreover,

$$\mathcal{P}_\tau(v_0) = 2\tau \int_0^1 x_\tau(t)^\top H(t) x_\tau(t)\, dt. \tag{28}$$

Proof. Set $h = N^{-1}$ and $b_N = Ns_N$, and let x_b denote the solution of Equation (23) with coefficient b. Fix $T < \infty$. Taylor expansion of the matrix exponential and the integral form of the ODE give, uniformly for $b \in [0, T]$ and $j = 1, \dots, N$,

$$\left\| e^{-hbH(t_j)} x_b(t_{j-1}) - x_b(t_j) \right\| \le C_T h^2.$$

Indeed, both terms equal $x_b(t_{j-1}) - hbH(t_j)x_b(t_{j-1})$ up to $O_T(h^2)$; this uses the C^1 regularity of H. If $e_j = v_j - x_b(t_j)$, the exact task update is a contraction, so

$$\|e_j\| \le \|e_{j-1}\| + C_T h^2.$$

Since $e_0 = 0$, summing over the N steps gives $\|\Phi_{N,b/N} v_0 - x_b(1)\| \le C_T h$. This proves Equation (26). Both the discrete updates and the ODE flow are contractions, and hence

$$\left| \mathcal{P}_{N,b/N}(v_0) - \mathcal{P}_b(v_0) \right| \le 2\|v_0\| \, \|\Phi_{N,b/N} v_0 - x_b(1)\|,$$

which proves Equation (27). To compare two coefficients $b, c \ge 0$, subtract their ODEs and apply variation of constants. Since the homogeneous flow is a contraction,

$$\|x_b(1) - x_c(1)\| \le |b - c| \int_0^1 \|H(t)\| \, \|x_c(t)\| \, dt \le L\|v_0\| \, |b - c|.$$

Taking $b = b_N$ and $c = \tau$ proves Equation (25) whenever $b_N \to \tau$.

Finally,

$$\frac{d}{dt}\|x_\tau(t)\|^2 = -2\tau x_\tau(t)^\top H(t) x_\tau(t).$$

Integration from 0 to 1 gives Equation (28). □

A.3 Exact expansions of the limiting curve

The small-budget behavior is described in the ambient parameter coordinates. Define the operator on $V(0)$

$$\mathcal{A}_H := P(0) \left(\int_0^1 H(t) \, dt \right) P(0) \Big|_{V(0)}. \tag{29}$$

For large budgets, the solution directions move with t, so their contributions are compared in a common coordinate system. Let $U(t)$ be Kato transport,

$$\dot{U}(t) = [\dot{P}(t), P(t)]U(t), \qquad U(0) = I. \tag{30}$$

The generator is skew-symmetric, $U(t)$ is orthogonal, and $P(t)U(t) = U(t)P(0)$ (Avron et al., 2012). Write $P_0 = P(0)$ and $Q_0 = I - P_0$. In this fixed frame,

$$U(t)^\top H(t) U(t) = \begin{pmatrix} 0 & 0 \\ 0 & H_\perp(t) \end{pmatrix}, \qquad U(t)^\top \dot{U}(t) = \begin{pmatrix} 0 & -B(t)^\top \\ B(t) & 0 \end{pmatrix}, \tag{31}$$

relative to $V(0) \oplus V(0)^\perp$. Here $H_\perp(t) \succeq \mu I$ and $B(t) = Q_0 U(t)^\top \dot{P}(t) U(t) P_0$. Define

$$\mathcal{G}_H := \int_0^1 B(t)^\top H_\perp(t)^{-1} B(t) \, dt \tag{32}$$

$$= \int_0^1 U(t)^\top \dot{P}(t) H(t)^\dagger \dot{P}(t) U(t) \, dt \Big|_{V(0)}. \tag{33}$$

Lemma 2 (Large-budget tracking expansion)**.** *Let $y_\tau(t) = U(t)^\top x_\tau(t) = (p_\tau(t), q_\tau(t))$ in the decomposition $V(0) \oplus V(0)^\perp$. As $\tau \to \infty$,*

$$\sup_{t \in [0,1]} \|p_\tau(t) - v_0\| = O(\tau^{-1}), \tag{34}$$

$$q_\tau(t) = -\tau^{-1} H_\perp(t)^{-1} B(t) v_0 + r_\tau(t), \tag{35}$$

$$\int_0^1 \|r_\tau(t)\|^2 dt = o(\tau^{-2}). \tag{36}$$

Proof. Using Equations (23) and (31), the transported components satisfy

$$\dot{p}_\tau = B(t)^\top q_\tau, \qquad \dot{q}_\tau = -B(t)p_\tau - \tau H_\perp(t)q_\tau, \qquad (p_\tau(0), q_\tau(0)) = (v_0, 0). \tag{37}$$

Let $E_\tau(t,r)$ denote the evolution operator generated by $-\tau H_\perp(t)$. The spectral gap gives $\|E_\tau(t,r)\| \le e^{-\mu\tau(t-r)}$. Variation of constants in the second equation of Equation (37) gives

$$q_\tau(t) = -\int_0^t E_\tau(t,r)B(r)p_\tau(r)\,dr, \qquad \|q_\tau\|_\infty \le \frac{C}{\tau}\|p_\tau\|_\infty.$$

The first equation of Equation (37) therefore implies

$$\|p_\tau\|_\infty \le \|v_0\| + \frac{C}{\tau}\|p_\tau\|_\infty.$$

For all sufficiently large τ, this bounds p_τ uniformly. Returning to the two estimates above gives $\|q_\tau\|_\infty = O(\tau^{-1})$ and then $\|p_\tau - v_0\|_\infty = O(\tau^{-1})$, proving Equation (34).

Set $a_\tau(t) = H_\perp(t)^{-1}B(t)p_\tau(t)$ and $\widetilde{r}_\tau = q_\tau + \tau^{-1}a_\tau$. Equation (37) gives

$$\dot{\widetilde{r}}_\tau + \tau H_\perp(t)\widetilde{r}_\tau = \tau^{-1}\dot{a}_\tau(t), \qquad \widetilde{r}_\tau(0) = \tau^{-1}a_\tau(0).$$

Smoothness, Equation (34), and the first equation of Equation (37) bound $\dot{a}_\tau$ uniformly. A second use of variation of constants gives

$$\widetilde{r}_\tau(t) = E_\tau(t,0)\frac{a_\tau(0)}{\tau} + \frac{1}{\tau}\int_0^t E_\tau(t,r)\dot{a}_\tau(r)\,dr.$$

The first term has squared L^2 norm $O(\tau^{-3})$. The second is $O(\tau^{-2})$ uniformly in t and hence has squared L^2 norm $O(\tau^{-4})$. Finally, Equation (34) gives

$$\frac{1}{\tau}H_\perp(t)^{-1}B(t)(p_\tau(t) - v_0) = O(\tau^{-2})$$

uniformly. Taking $r_\tau = \widetilde{r}_\tau - \tau^{-1}H_\perp^{-1}B(p_\tau - v_0)$ proves Equations (35) and (36). □

We next give an estimate that is uniform in both task spacing and per-task training time. Its fixed-time expansions prove Proposition 3, and its first bound is used to prove the large-Ns_N regime of Theorem 2.

A.3.1 Dense-task expansion for long per-task training

For $s \in (0,\infty)$, define on the normal block

$$\mathcal{W}_s(t) := \frac{1}{2}\coth\left(\frac{sH_\perp(t)}{2}\right), \qquad \mathcal{W}_\infty(t) := \frac{1}{2}I, \tag{38}$$

and set

$$\mathcal{G}_s := \int_0^1 B(t)^\top \mathcal{W}_s(t)B(t)\,dt. \tag{39}$$

Together with $\mathcal{G}_H$ from Equation (32), these are the operators used in the coefficients of Proposition 3.

Lemma 3 (Finite-time expansion)**.** *Let*

$$\delta_N := 1 - e^{-\mu s_N}.$$

If $N\delta_N \to \infty$, then

$$\mathcal{P}_{N,s_N}(v_0) = O\left(\frac{1}{N\delta_N}\right). \tag{40}$$

For every fixed $s \in (0,\infty]$,

$$P(1)\Phi_{N,s}\big|_{V(0)} = U(1)\left(I - \frac{\mathcal{G}_s}{N}\right) + o(N^{-1}), \tag{41}$$

$$\mathcal{P}_{N,s}(v_0) = \frac{2}{N}\langle v_0, \mathcal{G}_s v_0\rangle + o(N^{-1}). \tag{42}$$

If additionally $s_N \to 0$, then

$$\mathcal{P}_{N,s_N}(v_0) = \frac{2}{Ns_N}\langle v_0, \mathcal{G}_H v_0\rangle + o((Ns_N)^{-1}). \tag{43}$$

Moreover, for every fixed finite s and every nonzero $v_0 \in V(0) \cap \mathcal{K}^\perp$,

$$\langle v_0, (\mathcal{G}_s - \mathcal{G}_\infty)v_0\rangle > 0. \tag{44}$$

Proof. Let $h = N^{-1}$, $U_j = U(t_j)$, and $z_j = U_j^\top v_j = (p_j, q_j)$. Put $B_j = B(t_{j-1})$. Taylor expansion of $U_j^\top U_{j-1}$ in the block frame Equation (31) gives

$$p_j = p_{j-1} + hB_j^\top q_{j-1} - \frac{h^2}{2}B_j^\top B_j p_{j-1} + a_j, \tag{45}$$

$$q_j = R_j(q_{j-1} - hB_j p_{j-1}) + b_j, \tag{46}$$

where $R_j = e^{-s_N H_\perp(t_j)}$ satisfies $\|R_j\| \le 1 - \delta_N$, while

$$\|a_j\| \le C(h^3 + h^2\|q_{j-1}\|), \qquad \|b_j\| \le Ch^2.$$

These bounds are uniform in both j and s_N. Indeed, the block generator $S(t) := U(t)^\top \dot{U}(t)$ in Equation (31) has vanishing diagonal blocks, while $S(t)^2$ has tangent block $-B(t)^\top B(t)$. A second-order Taylor expansion of $U_j^\top U_{j-1}$ therefore gives tangent–normal blocks $hB_j^\top + O(h^2)$ and $-hB_j + O(h^2)$, and tangent block $I - \frac{h^2}{2}B_j^\top B_j + O(h^3)$. The C^3 bounds on U make these remainders uniform; multiplication of the normal block by R_j, with $\|R_j\| \le 1$, preserves the stated estimates. The original updates are contractions, so p_j and q_j remain bounded. Iterating Equation (46) and using $h/\delta_N = (N\delta_N)^{-1} \to 0$ yields

$$\max_{j \le N} \|q_j\| = O\left(\frac{h}{\delta_N}\right).$$

Substitution into Equation (45), followed by discrete Gronwall, gives

$$\max_{j \le N} \|p_j - v_0\| = O\left(\frac{h}{\delta_N}\right).$$

Because U_N is orthogonal,

$$\mathcal{P}_{N,s_N}(v_0) = \|v_0\|^2 - \|p_N\|^2 - \|q_N\|^2.$$

The last two bounds, together with $h/\delta_N \to 0$, prove Equation (40).

Now fix $s \in (0, \infty)$, so that $R_j = e^{-sH_\perp(t_j)}$ is uniformly contractive. Define

$$L_j := -(I - R_j)^{-1}R_j B_j, \qquad W_j := \frac{1}{2}(I + R_j)(I - R_j)^{-1}. \tag{47}$$

Smoothness gives $\|L_j\| \le C$ and $\|L_j - L_{j-1}\| \le Ch$. For $0 \le j \le N-1$, set $e_j = q_j - hL_{j+1}p_j$. The identity $(I - R_j)L_j = -R_j B_j$ and Equations (45) and (46) imply

$$\|e_j\| \le (1 - \delta)\|e_{j-1}\| + Ch^2, \qquad \delta := 1 - e^{-\mu s} > 0.$$

Since $e_0 = O(h)$, the geometric-series bound gives $\|e_j\| \le Ch(1-\delta)^j + Ch^2$ and hence $h\sum_{j=1}^N \|e_{j-1}\| = O(h^2) = o(h)$. Also $q_{j-1} = hL_j p_{j-1} + e_{j-1}$. Substitution into Equation (45) gives

$$p_N = \left[I - h\sum_{j=1}^N hB_j^\top W_j B_j\right] v_0 + o(h), \qquad q_N = O(h). \tag{48}$$

On the normal block,

$$W_j = \frac{1}{2}\coth\left(\frac{sH_\perp(t_j)}{2}\right).$$

The Riemann sum in Equation (48) therefore converges to $\mathcal{G}_s$ from Equation (39). Since $P(1)v_N = U(1)p_N$ and $\|q_N\|^2 = o(h)$, expansion of the endpoint norm proves Equations (41) and (42). The case $s = \infty$ follows by setting $R_j = 0$, equivalently $W_j = I/2$.

For every positive eigenvalue λ of $H_\perp(t)$,

$$\frac{1}{2}\coth(s\lambda/2) - \frac{1}{2} = \frac{e^{-s\lambda}}{1 - e^{-s\lambda}} > 0.$$

Hence $\mathcal{G}_s - \mathcal{G}_\infty$ is positive semidefinite. Equality on v_0 would require $B(t)v_0 = 0$ for every t. To see why this forces $v_0 \in \mathcal{K}$, set $z(t) = U(t)v_0 \in V(t)$. The vector $U(t)^\top \dot{P}(t)z(t)$ lies in $V(0)^\perp$ and equals $B(t)v_0$ there. The Kato equation therefore gives $\dot{z}(t) = \dot{P}(t)z(t) = 0$. Hence $z(t) = v_0 \in V(t)$ for every t, so $v_0 \in \mathcal{K}$. This contradicts nonzero $v_0 \in \mathcal{K}^\perp$ and proves Equation (44).

For the refined expansion, suppose $s_N \to 0$. Then $\delta_N \asymp s_N$. Define

$$L_j := -(I - R_j)^{-1}R_jB_j, \qquad W_j := \frac{1}{2}(I + R_j)(I - R_j)^{-1}.$$

The spectral bounds and smoothness give

$$\|L_j\| \le Cs_N^{-1}, \qquad \|L_j - L_{j-1}\| \le Chs_N^{-1}.$$

Set $e_j = q_j - hL_{j+1}p_j$ for $0 \le j \le N-1$. The identity $(I - R_j)L_j = -R_jB_j$, together with Equations (45) and (46), gives

$$\|e_j\| \le (1 - \delta_N)\|e_{j-1}\| + C\frac{h^2}{s_N} + C\frac{h^3}{s_N^2}.$$

Since $e_0 = O(h/s_N)$, a geometric-series estimate yields

$$\|e_j\| \le C\frac{h}{s_N}(1 - \delta_N)^j + C\frac{h^2}{s_N^2}.$$

Since $\delta_N \asymp s_N$ and the standing assumption gives $h/s_N = (Ns_N)^{-1} \to 0$, summing this bound makes the little-o step explicit:

$$h\sum_{j=1}^{N}\|e_{j-1}\| \le C\frac{h^2}{s_N\delta_N} + C\frac{h^2}{s_N^2} = O\left(\frac{h^2}{s_N^2}\right) = o\left(\frac{h}{s_N}\right).$$

Substituting $q_{j-1} = hL_jp_{j-1} + e_{j-1}$ into Equation (45) and summing gives

$$p_N = \left[I - h\sum_{j=1}^{N}hB_j^\top W_jB_j\right]v_0 + o\left(\frac{h}{s_N}\right).$$

Finally, uniformly on the positive spectrum,

$$W_j = \frac{1}{2}I + (I - e^{-s_NH_\perp(t_j)})^{-1}e^{-s_NH_\perp(t_j)} = s_N^{-1}H_\perp(t_j)^{-1} + O(s_N).$$

The Riemann sum therefore gives $p_N = v_0 - (Ns_N)^{-1}\mathcal{G}_Hv_0 + o((Ns_N)^{-1})$. Since $\|q_N\|^2 = O((Ns_N)^{-2}) = o((Ns_N)^{-1})$, expansion of $\|p_N\|^2 + \|q_N\|^2$ proves Equation (43). □

A.4 An anisotropic distance can reverse the comparison

For a positive-definite matrix M, define the corresponding endpoint progress by

$$\mathcal{P}_{N,s}^M(v) := v^\top Mv - (\Phi_{N,s}v)^\top M(\Phi_{N,s}v).$$

The following closed four-dimensional path shows that the fixed-finite-versus-exact comparison need not survive this change of geometry. Let

$$P_0 = \mathrm{diag}(I_2, 0_2), \qquad \Omega = 2\pi I_2, \qquad J = \begin{pmatrix} 0 & -\Omega^\top \\ \Omega & 0 \end{pmatrix},$$

and define

$$R(t) = e^{tJ}, \qquad P(t) = R(t)P_0R(t)^\top, \qquad H(t) = R(t)H_0R(t)^\top,$$

where

$$H_0 = \operatorname{diag}\left(0, 0, \log 2, \log \frac{5}{4}\right).$$

Take

$$v = \frac{1}{\sqrt{5}}(-2, 1, 0, 0)^\top, \qquad M = \operatorname{diag}(M_0, I_2), \qquad M_0 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

The eigenvalues of M_0 are 0.1 and 1.9, so M is positive definite. The path is closed because $R(1) = I$. Since J is off-diagonal relative to P_0, direct calculation gives $[\dot{P}(t), P(t)] = J$; hence its Kato transport is $U(t) = R(t)$. In the moving frame, $B(t) = \Omega$ and $H_\perp = \operatorname{diag}(\log 2, \log(5/4))$. Therefore, at fitting time $s = 1$,

$$D := \mathcal{G}_1 - \mathcal{G}_\infty = (2\pi)^2 \operatorname{diag}(1, 4). \tag{49}$$

Because $U(1) = I$ and M is block diagonal, the tangent–normal cross term vanishes. The normal endpoint component is $O(N^{-1})$ by Equation (48), so its quadratic contribution is $o(N^{-1})$. The endpoint expansion Equation (41) therefore gives

$$\lim_{N\to\infty} N\big(\mathcal{P}^M_{N,1}(v) - \mathcal{P}^M_{N,\infty}(v)\big) = v^\top(M_0D + DM_0)v.$$

For Euclidean distance, $M_0 = I_2$, and the right-hand side is

$$\frac{16}{5}(2\pi)^2 > 0. \tag{50}$$

For the positive-definite matrix above,

$$M_0 \operatorname{diag}(1, 4) + \operatorname{diag}(1, 4)M_0 = \begin{pmatrix} 2 & 4.5 \\ 4.5 & 8 \end{pmatrix},$$

so the limit is instead

$$-\frac{2}{5}(2\pi)^2 < 0. \tag{51}$$

Thus finite fitting beats exact fitting under Euclidean distance but loses under this anisotropic distance on the same task path and initial direction.

A.5 Proof of the optimal training-time law

The following lemma proves Theorem 2 and records the coefficients of its two tails.

Lemma 4 (Optimal-time law)**.** *Let $v_0 \in V(0) \cap \mathcal{K}^\perp$ be nonzero. Then:*

(i) The limiting progress has the two asymptotic expansions

$$\mathcal{P}_\tau(v_0) = 2\tau\langle v_0, \mathcal{A}_H v_0\rangle + O(\tau^2), \qquad \tau \downarrow 0, \tag{52}$$

$$\mathcal{P}_\tau(v_0) = \frac{2}{\tau}\langle v_0, \mathcal{G}_H v_0\rangle + o(\tau^{-1}), \qquad \tau \to \infty. \tag{53}$$

(ii) The two tail operators satisfy

$$\ker \mathcal{A}_H = \ker \mathcal{G}_H = \mathcal{K} \quad \text{within } V(0), \tag{54}$$

and $\mathcal{P}_\tau(v_0) > 0$ for every $\tau \in (0, \infty)$. Hence

$$\mathcal{T}^\star(v_0) := \arg\max_{\tau>0} \mathcal{P}_\tau(v_0) \tag{55}$$

is a nonempty compact subset of $(0, \infty)$.

(iii) For every sequence s_N,

$$\begin{aligned} Ns_N \to 0 &\implies \mathcal{P}_{N,s_N}(v_0) \to 0, \\ Ns_N \to \tau \in (0,\infty) &\implies \mathcal{P}_{N,s_N}(v_0) \to \mathcal{P}_\tau(v_0) > 0, \\ Ns_N \to \infty &\implies \mathcal{P}_{N,s_N}(v_0) \to 0. \end{aligned} \tag{56}$$

(iv) For all sufficiently large N, every optimizer in Equation (9) is finite. Every sequence of such optimizers satisfies

$$\operatorname{dist}(Ns_N^\star(v_0), \mathcal{T}^\star(v_0)) \longrightarrow 0. \tag{57}$$

Consequently $s_N^\star(v_0) = \Theta(N^{-1})$. If $\mathcal{T}^\star(v_0) = \{\tau^\star(v_0)\}$, then

$$s_N^\star(v_0) = \frac{\tau^\star(v_0)}{N} + o(N^{-1}). \tag{58}$$

Proof. For small τ, the integral form of Equation (23) gives, uniformly on $[0, 1]$,

$$x_\tau(t) = v_0 - \tau \int_0^t H(r)v_0\, dr + O(\tau^2).$$

Since $x_\tau(t) = v_0 + O(\tau)$ uniformly, substitution into Equation (28) proves Equation (52). For large τ, Lemma 2 and Equation (31) give

$$\mathcal{P}_\tau(v_0) = 2\tau \int_0^1 q_\tau(t)^\top H_\perp(t) q_\tau(t)\, dt,$$

Set $g(t) = H_\perp(t)^{-1}B(t)v_0$. By Equations (35) and (36), $q_\tau = -\tau^{-1}g + r_\tau$ with $\|r_\tau\|_{L^2} = o(\tau^{-1})$. Because g and $H_\perp$ are bounded,

$$\int_0^1 q_\tau^\top H_\perp q_\tau\, dt = \frac{1}{\tau^2}\int_0^1 g^\top H_\perp g\, dt + o(\tau^{-2}).$$

The integral equals $\langle v_0, \mathcal{G}_H v_0\rangle$, which proves Equation (53).

For $v \in V(0)$,

$$\langle v, \mathcal{A}_H v\rangle = \int_0^1 \|H(t)^{1/2}v\|^2 dt,$$
$$\langle v, \mathcal{G}_H v\rangle = \int_0^1 \|H_\perp(t)^{-1/2}B(t)v\|^2 dt.$$

By continuity, the first integral vanishes exactly for $v \in \mathcal{K}$. For the second, vanishing is equivalent to $B(t)v = 0$ for every t. Let $z(t) = U(t)v \in V(t)$. Since $\dot{P}(t)$ maps $V(t)$ into $V(t)^\perp$, the normal coordinate of $\dot{P}(t)z(t)$ is exactly $B(t)v$. The Kato equation gives $\dot{z}(t) = \dot{P}(t)z(t)$, so $B(t)v = 0$ makes $z(t)$ constant and places v in every $V(t)$. Conversely, if $v \in \mathcal{K}$, then $P(t)v = v$ and differentiation gives $\dot{P}(t)v = 0$, so the constant path solves the Kato equation and $B(t)v = 0$. This proves Equation (54).

If $\mathcal{P}_\tau(v_0) = 0$ for a finite positive τ, the nonnegative integrand in Equation (28) vanishes for every t. Hence $H(t)x_\tau(t) = 0$ for every t, the ODE gives $\dot{x}_\tau(t) = 0$, and $v_0 = x_\tau(t)$ belongs to every $V(t)$. This contradicts nonzero $v_0 \in \mathcal{K}^\perp$, so $\mathcal{P}_\tau(v_0) > 0$. The curve is continuous, is positive at each finite $\tau > 0$, and tends to zero at both ends by Equations (52) and (53). Its maximum is therefore attained on a compact subinterval of $(0, \infty)$, which proves the claim about Equation (55).

The first two lines of Equation (56) follow from Lemma 1. If $Ns_N \to \infty$, use

$$1 - e^{-\mu s} \geq c_\mu \min\{s, 1\}, \qquad s \geq 0,$$

for a constant $c_\mu > 0$. If $s_N \leq 1$, this gives $N(1 - e^{-\mu s_N}) \geq c_\mu N s_N$; if $s_N > 1$, it gives a lower bound of order N. Thus $N(1 - e^{-\mu s_N}) \to \infty$, and Lemma 3 gives the third line.

It remains to localize the finite-resolution optimizers. Choose any $\bar{\tau} > 0$. By Lemma 1, $\mathcal{P}_{N,\bar{\tau}/N}(v_0) \to \mathcal{P}_{\bar{\tau}}(v_0) > 0$, so the optimal value stays bounded away from zero. If no common positive lower bound existed for $Ns_N^\star$, one could choose task counts and optimizers with $Ns_N^\star \to 0$; the first regime would make their progress tend to zero, a contradiction. If no common upper bound existed, one could similarly choose $Ns_N^\star \to \infty$, including the possibility $s_N^\star = \infty$; the third regime gives the same contradiction. Hence all optimizers are finite and $Ns_N^\star$ lies in one compact subinterval of $(0, \infty)$ for all sufficiently large N.

Consider any convergent subsequence $Ns_N^\star \to \tau$. For each fixed $\gamma > 0$, optimality gives $\mathcal{P}_{N,s_N^\star}(v_0) \geq \mathcal{P}_{N,\gamma/N}(v_0)$. Uniform convergence on the compact interval containing $Ns_N^\star$, together with Lemma 1, yields

$$\mathcal{P}_\tau(v_0) \geq \mathcal{P}_\gamma(v_0) \qquad \text{for every } \gamma > 0.$$

Every accumulation point therefore belongs to $\mathcal{T}^\star(v_0)$. If Equation (57) failed, a subsequence at a fixed positive distance from $\mathcal{T}^\star(v_0)$ would have a further convergent subsequence whose limit belongs to $\mathcal{T}^\star(v_0)$, a contradiction. This proves Equation (57); the final two claims follow. □

A.6 Common-solution distance and prediction risk

We measure learning progress by the reduction in squared Euclidean distance to the common solution set. Following the convention of removing the common-kernel component, define the average path risk by

$$R(v) := \int_0^1 L_t(w^\star + v)\, dt = \frac{1}{2} v^\top \bar{H} v, \qquad \bar{H} := \int_0^1 H(t)\, dt.$$

By Section A.1, $\bar{H}$ is positive definite on $\mathcal{K}^\perp$, so

$$\frac{\lambda_{\min}(\bar{H}|_{\mathcal{K}^\perp})}{2}\|v\|^2 \leq R(v) \leq \frac{\lambda_{\max}(\bar{H}|_{\mathcal{K}^\perp})}{2}\|v\|^2.$$

These bounds give the distance measure a prediction-loss interpretation, but do not ensure that different training times have the same ranking under the two measures. Average path risk also depends on the direction of the error, and training time changes both its length and direction. Section A.4 illustrates how the evaluation geometry can affect comparisons between training times. Thus the general conclusion of Theorem 2 concerns the distance-based progress defined in the main text; extending it to average path risk requires further consideration of the task-path geometry.

The following sufficient condition covers the rotating-regression example in the main text.

Corollary 1 (Average-risk scaling under monotone rotation)**.** *Suppose*

$$H(t) = \lambda(t) n(t) n(t)^\top, \qquad n(t) = (\cos\theta(t), \sin\theta(t))^\top,$$

where $\lambda, \theta \in C^3([0,1])$, $0 < \mu \leq \lambda(t) \leq L$, and θ is monotone with

$$0 < |\theta(1) - \theta(0)| \leq \frac{\pi}{2}.$$

Let $v_0 \in \ker H(0)$ be nonzero. If per-task training time is selected to minimize the final model's average path risk,

$$s_{N,R}^\star \in \arg\min_{s \in [0,\infty]} R(\Phi_{N,s} v_0),$$

then, for all sufficiently large N, every optimizer is finite and satisfies

$$s_{N,R}^\star = \Theta(N^{-1}).$$

Proof. Rescaling the nonzero initial error does not change the optimal training time, so assume $v_0 = k(0)$, where $k(t) = (-\sin\theta(t), \cos\theta(t))^\top$. Write $x_\tau(t) = p(t)k(t) + q(t)n(t)$. The moving-frame equations are

$$p' = -\theta' q, \qquad q' = \theta' p - \tau\lambda q, \qquad (p(0), q(0)) = (1, 0).$$

The expansion used in Section A.3 gives

$$p(1) = 1 - \frac{g}{\tau} + O(\tau^{-2}), \qquad q(1) = \frac{\theta'(1)}{\tau\lambda(1)} + O(\tau^{-2}), \qquad g := \int_0^1 \frac{\theta'(t)^2}{\lambda(t)}\, dt > 0.$$

Let

$$a := k(1)^\top \bar{H} k(1) > 0, \qquad b := k(1)^\top \bar{H} n(1).$$

Monotonicity and the bound on the total rotation imply $b\theta'(1) \leq 0$. Hence, writing $R_\tau := R(x_\tau(1))$ and $R_\infty := R(k(1))$, we obtain

$$R_\tau = R_\infty - \frac{ag - b\theta'(1)/\lambda(1)}{\tau} + O(\tau^{-2}).$$

The numerator is strictly positive, so sufficiently large but finite total training times satisfy $R_\tau < R_\infty$.

At the other endpoint, the small-budget expansion gives

$$R_\tau = R_0 - \tau v_0^\top \bar{H}^2 v_0 + O(\tau^2),$$

with a strictly positive first-order decrease. Comparing the two endpoints, there is always a finite $\tau_0 > 0$ such that

$$R_{\tau_0} < \min\{R_0, R_\infty\}.$$

The endpoint convergence in Theorem 1 and the uniform large-budget estimates in Section A.3.1 now exclude optimal total times tending to zero or infinity, by the optimizer-localization argument in the proof of Lemma 4. This proves the claimed scaling. □

The constant-speed rotating example in the main text satisfies these conditions. Its average-path-risk and common-solution-distance objectives therefore yield the same N^{-1} scaling order, although their optimal proportionality constants need not coincide. This corollary applies to the linear rotating model. The Rotated-MNIST and Yearbook results remain empirical tests of whether a corresponding trend persists beyond the theoretical assumptions.

A.7 Small positive curvature and finite training horizons

The vanishing-progress conclusion at large budgets uses the exactly zero eigenvalues of the task curvatures. Giving these directions small positive curvature can preserve nearby trajectories over finite training horizons while changing the infinite-training behavior. The following exactly solvable perturbation distinguishes these two statements.

Consider

$$H_\varepsilon(t) = H(t) + \varepsilon I, \qquad \varepsilon > 0,$$

with the same common solution $w^\star$ and initial error v_0. Denote the errors after N tasks in the original and perturbed models by $v_N^{(0)}$ and $v_N^{(\varepsilon)}$. Because εI commutes with every $H(t)$,

$$e^{-sH_\varepsilon(t)} = e^{-\varepsilon s} e^{-sH(t)}, \qquad v_N^{(\varepsilon)} = e^{-\varepsilon N s} v_N^{(0)}.$$

Writing $\tau = Ns$, contractivity of the original updates gives

$$\|v_N^{(\varepsilon)} - v_N^{(0)}\| \le (1 - e^{-\varepsilon\tau})\|v_0\| \le \varepsilon\tau\|v_0\|.$$

Thus the final states remain close in absolute error when $\varepsilon\tau \ll 1$.

This perturbation also changes the source-fitting condition. The original initialization satisfies $H(0)v_0 = 0$, whereas its loss in the perturbed model is

$$L_0^{(\varepsilon)}(w^\star + v_0) = \frac{\varepsilon}{2}\|v_0\|^2.$$

Keeping the same initial model therefore corresponds to approximate source fitting. With the positive-definite source curvature $H_\varepsilon(0)$, exact source fitting would already reach the unique common solution.

Closeness of the trajectories and preservation of the original leading-order asymptotics require different comparisons. Define

$$\mathcal{P}_{N,s}^{(\varepsilon)} := \|v_0\|^2 - \|v_N^{(\varepsilon)}\|^2.$$

The exact identity above yields

$$\mathcal{P}_{N,s}^{(\varepsilon)} - \mathcal{P}_{N,s}^{(0)} = (1 - e^{-2\varepsilon\tau})\|v_N^{(0)}\|^2,$$

and hence

$$0 \le \mathcal{P}_{N,s}^{(\varepsilon)} - \mathcal{P}_{N,s}^{(0)} \le 2\varepsilon\tau\|v_0\|^2.$$

In the continuum limit, the same relation gives $x_\tau^{(\varepsilon)}(1) = e^{-\varepsilon\tau} x_\tau^{(0)}(1)$. The original large-budget progress satisfies $\mathcal{P}_\tau^{(0)} \sim C/\tau$ for some $C > 0$. To make the perturbation negligible relative to this

leading term, the stronger scale separation $\varepsilon\tau^2 \ll 1$ is sufficient. More precisely, in the joint limit $\varepsilon \to 0$, $\tau \to \infty$, and $\varepsilon\tau^2 \to 0$,

$$\mathcal{P}_\tau^{(\varepsilon)} \sim \frac{C}{\tau}.$$

This identifies a joint asymptotic regime in which the original large-budget mechanism is retained to leading order.

In contrast, for any fixed $\varepsilon > 0$,

$$\|v_N^{(\varepsilon)}\| \le e^{-\varepsilon\tau}\|v_0\| \longrightarrow 0 \qquad (\tau \to \infty).$$

Sufficiently long training therefore removes the error, and the original conclusion that large-budget progress vanishes no longer holds. This solvable perturbation shows that exactly zero and small positive curvatures can produce nearby finite-horizon trajectories but different infinite-time limits. The global optimal-time theorem relies on the stated nullspace structure; for models with approximately zero curvature, its applicability must be assessed jointly with the curvature magnitude and training horizon.

B Details of the Numerical Experiments

B.1 Rotating rank-one path

The example in Section 4 uses

$$H(t) = u(t)u(t)^\top, \qquad u(t) = \big(\cos(\pi t/2),\, \sin(\pi t/2)\big)^\top, \qquad v_0 = (0,1)^\top.$$

Thus $V(t) = u(t)^\perp$ rotates by $\pi/2$ and $\mathcal{K} = \bigcap_t V(t) = \{0\}$. For finite N, each matrix-exponential update is evaluated as

$$e^{-sH(t_j)}v = v - (1 - e^{-s})u(t_j)u(t_j)^\top v.$$

Write $x_\tau(t) = R(\pi t/2)y_\tau(t)$. Then

$$\dot{y}_\tau(t) = \begin{bmatrix} -\tau & \pi/2 \\ -\pi/2 & 0 \end{bmatrix} y_\tau(t), \qquad y_\tau(0) = v_0.$$

Therefore the limiting endpoint is

$$x_\tau(1) = R(\pi/2)\exp\left(\begin{bmatrix} -\tau & \pi/2 \\ -\pi/2 & 0 \end{bmatrix}\right) v_0,$$

where $R(\theta)$ denotes the planar rotation matrix through angle θ, so the limiting curve can be evaluated without time discretization. Numerical maximization gives $\tau^\star = 2.19836$ and $\mathcal{P}_{\tau^\star}(v_0) = 0.63113$. For exact fitting, consecutive zero-loss directions differ by $\pi/(2N)$. Each projection multiplies the norm by $\cos(\pi/(2N))$, so

$$\mathcal{P}_{N,\infty}(v_0) = 1 - \cos^{2N}\Big(\frac{\pi}{2N}\Big) = \frac{\pi^2}{4N} + O(N^{-2}).$$

Because the single positive eigenvalue of $H(t)$ equals one, Equation (19) specializes to

$$N\mathcal{P}_{N,s}(v_0) \longrightarrow \Big(\frac{\pi}{2}\Big)^2 \coth(s/2) \quad (s < \infty), \qquad N\mathcal{P}_{N,\infty}(v_0) \longrightarrow \Big(\frac{\pi}{2}\Big)^2. \tag{59}$$

Equation (59) makes the new comparison explicit without an additional panel: every finite s has coefficient $(\pi/2)^2 \coth(s/2) > (\pi/2)^2$, the coefficient obtained by training each task to convergence.

B.2 High-dimensional random smooth paths

We also test the scaling law on ten independently generated paths with ambient dimension $d = 20$ and rank $r = 6$. Each path has the form

$$H(t) = Z(t)\Lambda Z(t)^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r), \qquad \lambda_i \in [0.5, 1.5],$$

where $Z(t) \in \mathbb{R}^{d\times r}$ has orthonormal columns and is obtained by applying 80 smooth random Givens rotations to a random orthogonal frame. The rank and positive spectral gap are fixed along the path. The initial error is a random unit vector in $\ker H(0)$. We retain a draw only when the smallest eigenvalue of the average of the projectors onto $\operatorname{im} H(t)$ at 81 equally spaced locations exceeds 10^{-5}; the smallest value among the ten retained paths is 1.48×10^{-4}. This check ensures that the sampled subspaces have trivial common kernel.

For $N \in \{16, 32, 64, 128, 256\}$, we evaluate the exact matrix-exponential updates over a logarithmic grid of effective times $\tau = Ns$ and optimize over τ by bounded scalar minimization. The median scaled optimizer decreases only from 2.82 at $N = 16$ to 2.72 at $N = 256$, while median optimal progress stays positive (0.290 to 0.278). By contrast, median exact-fitting progress decreases from 0.0834 to 0.00568; a log–log fit over $N \geq 32$ has slope -0.981.

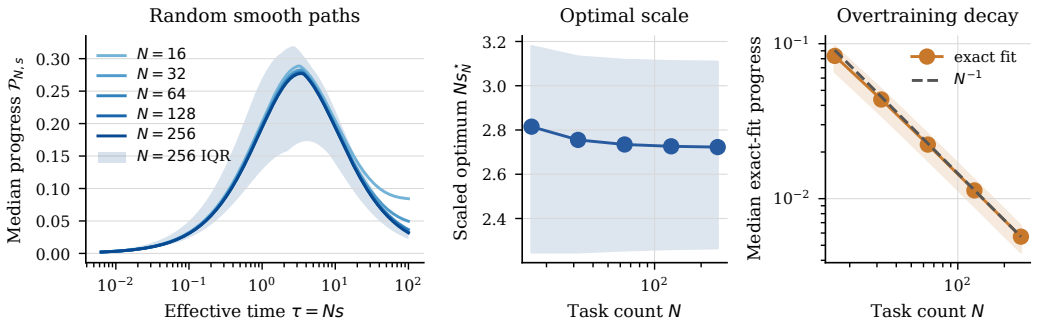


Figure 5: High-dimensional stress test over ten random smooth Hessian paths. Left: median learning-progress curves, with the interquartile range (IQR) shown for $N = 256$. Middle: median scaled optimizer $Ns_N^\star$ and its IQR. Right: median exact-fitting progress and its IQR, compared with an N^{-1} reference line.

This experiment stays within the theorem's assumptions. It tests higher dimension, unequal positive eigenvalues, and noncommuting rotations of $\operatorname{im} H(t)$. All random seeds are fixed in the included experiment code, and the underlying numerical values are provided with the source package.

B.3 Gradually rotated MNIST

The experiment uses the standard MNIST training and test partitions. For each seed, we sample 30,000 images from the training partition for optimization and 3,000 held-out evaluation images from the official test partition. The same held-out sample is used for every (N, k) configuration within that seed. Rotations use bilinear interpolation, zero fill, no canvas expansion, and the standard MNIST normalization. Labels are unchanged along the path. The classifier is an MLP with widths 784–256–128–10 and ReLU activations.

The initial model is trained for 1,000 mini-batch SGD updates on the 0° task. Every adaptation run starts from this same seed-specific checkpoint. For a task count N, the sequential training angles are $60j/N$ degrees for $j = 1, \ldots, N$. We use SGD without momentum, learning rate 0.05, batch size 256, and

$$k \in \{0, 1, 2, 4, 8, 16, 32, 64, 128\}$$

updates per task. Within each (N, seed) pair, different values of k use deterministic prefixes of the same taskwise mini-batch streams. We use the ten consecutive seeds from 20260805 through 20260814.

For evaluation, we rotate each held-out image to the 13 angles $\{0^\circ, 5^\circ, \ldots, 60^\circ\}$. Whole-path cross-entropy averages over all 39,000 image–angle predictions and is therefore independent of the training discretization N. The held-out sample provides a fixed empirical approximation to the whole-path population risk. The reported $k^\star$ is the minimizer of this empirical risk curve, rather than

a hyperparameter chosen to report post-selection predictive performance. No held-out examples are used for gradient updates. Lower values are better.

Table 1: Rotated-MNIST empirical optima over ten seeds. Parentheses in the second column give the number of seeds selecting each grid value. The interpolated columns use a local quadratic fit in $\log_2 k$ around the minimum of each mean evaluation curve. The last column descriptively reports mean $\pm$ standard deviation at the grid value selected by mean whole-path cross-entropy; it is not an independent post-selection test estimate.

N	seedwise best k (count)	interpolated $k^\star$	$Nk^\star$	whole-path CE
4	8 (10)	8.95	35.8	0.502 ± 0.018
8	4 (7), 8 (3)	5.07	40.6	0.503 ± 0.020
16	2 (4), 4 (6)	3.07	49.2	0.499 ± 0.025
32	1 (4), 2 (6)	1.52	48.8	0.507 ± 0.029

The grid-selected best-k sequences are nonincreasing in N for all ten seeds. The interpolated total budgets remain between 35.8 and 49.2, and a log–log fit gives $k^\star \propto N^{-0.84}$. Resampling the seeds and repeating the interpolation gives the 95% confidence interval $[0.76, 1.03]$ for this exponent. This interval includes the theoretical value 1, but the comparison remains qualitative: there are four task counts and an integer update grid. The nonlinear model, cross-entropy loss, finite samples, stochastic updates, and absence of a proved shared zero-loss parameter also differ from the analytical setting.

B.4 Temporal Yearbook

The Yearbook dataset contains aligned American high-school yearbook portraits indexed by calendar year (Ginosar et al., 2017). We use the official Wild-Time preprocessing and retain its provided data split (Yao et al., 2022): 33,431 images are used for training and 3,758 images from the benchmark test partition are held out for evaluation. The images span 1930–2013 and are represented as 32×32 grayscale inputs. We inherit the benchmark's binary labels only to study temporal optimization; the experiment is not intended for demographic inference.

To keep the underlying path fixed as N changes, a task centered at calendar year t reweights example i, with year a_i, according to

$$q_t(i) \propto \exp\left(-\frac{(a_i - t)^2}{2\sigma^2}\right) \mathbf{1}\{|a_i - t| \le 4\sigma\}, \qquad \sigma = 4 \text{ years}.$$

For $N \in \{4, 8, 16, 32\}$, training visits the centers $t_j = 1930 + 83j/N$, $j = 1, \dots, N$. Evaluation uses 17 equally spaced centers from 1930 through 2013 for every N. At each center, held-out cross-entropy is averaged with the corresponding normalized temporal weights; whole-path cross-entropy is the unweighted mean over the 17 centers. The held-out split provides a fixed empirical approximation to the whole-path temporal risk. The reported $k^\star$ is the minimizer of this empirical risk curve, rather than an estimate of post-selection predictive performance, and no held-out examples are used for gradient updates.

The classifier has four 3×3 convolutional blocks with channel widths 32–32–64–64. Each block uses GroupNorm, ReLU, and 2×2 max pooling, followed by global average pooling and a linear two-class head. The source model is trained for 1,500 updates on the source-time distribution. Adaptation uses plain mini-batch SGD with learning rate 0.03, batch size 256, no momentum or weight decay, and

$$k \in \{0, 1, 2, 4, 8, 16, 32, 64, 128\}$$

updates at every visited center. Larger k extends a fixed taskwise mini-batch prefix within each (N, seed) pair. We use five consecutive seeds from 20260805 through 20260809.

A log–log fit to the interpolated mean-curve optima gives $k^\star \propto N^{-1.37}$. Repeating the fit in 10,000 seed-bootstrap replicates gives a median exponent of 1.30 and a 95% interval $[0.79, 1.66]$. The interval includes the theoretical value 1 but does not distinguish it sharply from nearby exponents. Some $N = 4$ replicates select the largest tested value, so the upper endpoint in Table 2 is grid-limited. Across the tested task counts, the estimated optimal total update budgets do not form a stable plateau.

We also perform a necessary endpoint check for approximate joint fit. A model trained on the equally weighted source and target distributions has mean held-out endpoint cross-entropy 0.084, compared

Table 2: Yearbook empirical evaluation-set optima over five seeds. Optima are obtained from local quadratic interpolation in $\log_2 k$ around the minimum of each mean evaluation curve. Confidence intervals resample complete training seeds and repeat the averaging and interpolation; they do not include uncertainty from resampling the fixed evaluation split.

N	grid-selected k	interpolated $k^\star$	bootstrap 95% CI
4	64	79.90	$[63.54, 128.00]$
8	16	13.82	$[7.53, 38.48]$
16	4	4.23	$[2.70, 13.10]$
32	4	5.01	$[3.49, 14.91]$

with 0.057 for separately trained endpoint models; using the unrounded seed means, the mean gap is 0.028 across five seeds. This shows low simultaneous endpoint loss, not a shared zero-loss minimizer along the full path. The experiment further differs from the theorem through finite samples, supervised stochastic updates, cross-entropy loss, temporal smoothing, and the nonlinear classifier. It is therefore a natural-shift consistency check rather than a verification of the assumptions.