THE SEQUENTIAL PRICE OF CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential task updates are fundamental to continual learning, but their recency bias can impose a lasting performance cost. We study this cost in an overparameterized linear-regression model with i.i.d. task sampling. We prove that distribution-level forgetting and population loss converge to the same stationary limit. We quantify the additional loss incurred by sequential exact fitting, or the *sequential price*. In more homogeneous task geometries, it equals the intrinsic loss asymptotically attained by joint training, making the total loss twice as large. We further analyze fixed-strength elastic weight consolidation (EWC) under general task curvatures and characterize its stationary sequential price at every regularization strength. Under strong regularization, the price decays inversely with EWC strength while the mean-square coupling horizon grows proportionally. Experiments on Jester and Rotated MNIST support the predicted sequential price and its reduction by EWC, with quantitative agreement on real-world tasks satisfying the theory's assumptions and qualitative agreement under nonlinear finite-step training.

1 INTRODUCTION

Continual learning studies how models learn from a sequence of tasks that arrive over time (Delange et al., 2022; Wang et al., 2024; Mai et al., 2022). A model updates its current state on each new task and then carries the resulting state forward to subsequent tasks, so its capabilities are gradually shaped by a temporally ordered learning process. The most basic learning structure of continual learning is therefore sequential learning: tasks act on the model in sequence and jointly shape its behavior over time. As models and agents are deployed over increasingly long time horizons, understanding the behavior induced by this sequential structure becomes increasingly important.

Sequential learning is inherently asymmetric in time. Each update responds directly to the current task, giving recent tasks a more immediate influence on the current model. Information from earlier tasks is retained and transmitted through a succession of later model states. As tasks accumulate, the model typically places greater emphasis on recent tasks, while performance on past tasks gradually deteriorates, giving rise to forgetting (Parisi et al., 2019; Mirzadeh et al., 2020; Lesort et al., 2023). This recency bias further suggests that sequential learning may impose a persistent cost on the model's overall performance. A basic question is therefore how sequential learning affects the model's stationary population loss over the task distribution.

A natural approach to this question is to view the arriving tasks as independent samples from a latent task distribution. The task distribution provides a stable population-level description of the task sequence and lifts forgetting along a particular training history to a distribution-level quantity. Recent work, *From Order to Distribution* (Xu & Ma, 2026), adopts this perspective to study forgetting in continual learning and connects the dynamics induced by concrete task sequences to properties of the task distribution. This distribution-level formulation provides a natural starting point for characterizing the asymptotic behavior of sequential learning.

That work considers tasks that share a common solution and proves that distribution-level forgetting converges to zero. This setting describes a compatible family of tasks: each task has its own training data, while a single parameter can solve all tasks simultaneously (Evron et al., 2022; Xu & Ma, 2026). More general task families may contain individually realizable tasks with conflicting optima, while the set of parameters that solves all tasks simultaneously is typically empty. This raises a direct follow-up question: how does distribution-level forgetting behave asymptotically when tasks are intrinsically incompatible?

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.

We first connect the historical effects of sequential learning to its overall performance. Distribution-level forgetting measures the current model by its average loss on past tasks, whereas population loss measures its expected loss over the task distribution. For conflicting tasks, both quantities generally remain positive. We prove that, under i.i.d. task sampling and the stability conditions studied here, they converge to the same stationary limit, and the gap between them decays at rate $O(T^{-1})$. The dependence between a remote task and the current model decays with task age, while the recent tasks that remain strongly dependent form a vanishing fraction of the full history. The forgetting plateau therefore also characterizes the sequential learner's stationary population loss over the task distribution.

Identifying the forgetting plateau with stationary population loss allows it to be compared directly with $R^\star$, the minimum population loss attainable within the same model class. This quantity lower-bounds the population loss of any learning procedure that outputs a single set of model parameters (Lemma C.1), and joint training on a growing collection of tasks asymptotically attains it (Lemma C.2). We quantify the additional loss incurred by sequential exact fitting beyond the intrinsic loss $R^\star$ caused by task incompatibility. We call this additional cost the *sequential price*. This separates the unavoidable cost of conflicting tasks from the additional cost of continually adapting to them. In more homogeneous task geometries, the sequential price is as large as the intrinsic loss itself, making the stationary loss twice that attained by joint training.

We next use the sequential price to analyze fixed-strength EWC under general task curvatures (Kirkpatrick et al., 2017; Schwarz et al., 2018; Benzing, 2022), quantifying how regularization changes both stationary performance and convergence to stationarity. For every regularization strength, the stationary sequential price is computable from the distribution's curvatures and targets. In the strong-regularization regime, it decreases as $1/\lambda$, while the mean-square coupling horizon grows as λ.

Contributions. Our main contributions are as follows.

- We prove that distribution-level forgetting converges to population loss with an $O(T^{-1})$ gap, extending shared-solution theory to conflicting tasks.
- Using this result, we split the limiting loss into the unavoidable loss caused by task conflicts and the additional loss caused by learning tasks sequentially, which we call the sequential price, and show that the two are equal in more homogeneous settings.
- We calculate how EWC changes the sequential price at any fixed strength and show that, under strong regularization, the price decreases as $1/\lambda$ while the mean-square coupling horizon grows as λ.
- We apply the theory to Jester, whose naturally conflicting user tasks satisfy our assumptions. Exact predictions match sampled task streams, with stationary loss approximately 2 times the joint-training optimum despite heterogeneous task curvatures. On Rotated MNIST, corresponding loss patterns also emerge under nonlinear models and finite-step training.

2 SETUP

We work with regression tasks $\tau = (X_\tau, y_\tau)$, where $X_\tau \in \mathbb{R}^{n_\tau \times d}$ and $y_\tau \in \mathbb{R}^{n_\tau}$. This is a standard exact-fit linear model for continual-learning theory (Evron et al., 2022; Ding et al., 2024; Goldfarb et al., 2024). Each task is individually realizable, $y_\tau \in \mathrm{range}(X_\tau)$, but different tasks need not share a common solution. Define

$$\ell_\tau(w) := \frac{1}{n_\tau}\|X_\tau w - y_\tau\|_2^2, \qquad C_\tau := \frac{1}{n_\tau}X_\tau^\top X_\tau, \qquad Z_\tau := X_\tau^\dagger y_\tau.$$

Then $C_\tau \succeq 0$, $Z_\tau \in \mathrm{range}(C_\tau)$, and

$$\ell_\tau(w) = \|w - Z_\tau\|_{C_\tau}^2, \qquad \|v\|_A^2 := v^\top A v.$$

Given a task τ and an initialization $u \in \mathbb{R}^d$, the minimum-change exact-fit update is

$$S_\tau(u) := \arg\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - u\|_2^2 \quad \text{subject to} \quad X_\tau w = y_\tau.$$

A standard projection calculation gives

$$S_\tau(u) = u + X_\tau^\dagger(y_\tau - X_\tau u) = P_\tau u + Z_\tau, \qquad P_\tau := I - X_\tau^\dagger X_\tau = I - C_\tau^\dagger C_\tau,$$

where P_τ is the orthogonal projector onto $\ker(X_\tau) = \ker(C_\tau)$. If the tasks admit a common solution w°, then $Z_\tau = (I - P_\tau)w^\circ$ and the update reduces to the homogeneous projection dynamics

$$S_\tau(u) - w^\circ = P_\tau(u - w^\circ).$$

This minimum-change solution is the parameter selected by converged gradient descent on a consistent least-squares task when training starts from u (Evron et al., 2022; Lin et al., 2023; Karpel et al., 2026). Without a common solution, no fixed centering removes the task-dependent shift Z_τ, and the recursion remains affine rather than homogeneous.

Tasks arrive as an i.i.d. stream from a task-generating distribution (Xu & Ma, 2026):

$$\tau_1, \tau_2, \ldots \overset{\text{i.i.d.}}{\sim} \Pi.$$

Starting from an independent W_0 with finite second moment, sequential exact fitting generates

$$W_t = S_{\tau_t}(W_{t-1}) = P_t W_{t-1} + Z_t, \qquad P_t := P_{\tau_t}, \quad Z_t := Z_{\tau_t}.$$

We assume that $\mathbb{E}\|Z_\tau\|_2^2 < \infty$ and that $\|C_\tau\|_{\text{op}} \le L$ almost surely for some finite L.

Let

$$R(w) := \mathbb{E}_{\tau\sim\Pi}[\ell_\tau(w)], \qquad \overline{C} := \mathbb{E}[C_\tau], \qquad \overline{h} := \mathbb{E}[C_\tau Z_\tau].$$

We call $R(w)$ the population loss over the task distribution. Since $C_\tau \succeq 0$ for every task, $\overline{C} \succeq 0$. For a nonzero direction v, equality $v^\top \overline{C} v = 0$ is possible only in the special case $C_\tau v = 0$ almost surely: such a direction lies in $\ker(C_\tau)$ and is invisible to almost every task. We assume that this degeneracy is absent, so $\overline{C} \succ 0$. If invisible directions exist, they do not affect task losses almost surely and can be removed by restricting the analysis to $\text{range}(\overline{C})$ (Appendix C). The population minimizer and minimum population loss are

$$w_\star := \overline{C}^{-1}\overline{h}, \qquad R^\star := R(w_\star),$$

and, since $\overline{C} w_\star = \overline{h}$, direct expansion gives

$$R(w) = \mathbb{E}[Z_\tau^\top C_\tau Z_\tau] + w^\top \overline{C} w - 2w^\top \overline{h} = R^\star + \|w - w_\star\|_{\overline{C}}^2.$$

Consequently, $w_\star$ is the unique population minimizer, and $R^\star$ lower-bounds the expected population loss of any possibly randomized learner whose output is a single parameter vector. Moreover, $R^\star = 0$ if and only if the task distribution admits a common solution almost surely. Thus $R^\star$ is the intrinsic loss induced by task incompatibility within this model class (Lemma C.1). For

$$\widehat{W}_T^{\text{joint}} \in \arg\min_w \frac{1}{T}\sum_{t=1}^{T} \ell_{\tau_t}(w),$$

joint training is empirical risk minimization over the shared parameter w (Vapnik, 1991). Under our assumptions, $\widehat{W}_T^{\text{joint}} \to w_\star$ and $R(\widehat{W}_T^{\text{joint}}) \to R^\star$ almost surely (Lemma C.2). Hence $R^\star$ is also the asymptotic population loss attained by joint training.

Following standard forgetting measures and exact-fit linear analyses (Lopez-Paz & Ranzato, 2017; Evron et al., 2022; Xu & Ma, 2026), we define

$$F_T := \mathbb{E}\left[\frac{1}{T-1}\sum_{s=1}^{T-1} \ell_{\tau_s}(W_T)\right], \qquad G_T := \mathbb{E}[R(W_T)] = \mathbb{E}[\ell_{\widetilde{\tau}}(W_T)],$$

where $\widetilde{\tau} \sim \Pi$ is independent of the training history. The current task is excluded from F_T; under exact fitting, $\ell_{\tau_s}(W_s) = 0$, so this historical loss equals the increase in loss after task acquisition.

Finally, on the space $\mathbb{S}^d$ of symmetric matrices, define

$$\mathcal{S}_\Pi(A) := \mathbb{E}[P_\tau A P_\tau].$$

The collective coverage condition $\overline{C} \succ 0$ implies that $\mathcal{S}_\Pi$ is a strict contraction in Frobenius norm. Consequently, the affine recursion admits a unique invariant distribution ν within the class of distributions with finite second moment, in line with standard contractive random-affine recursions (Brandt, 1986; Diaconis & Freedman, 1999) (Appendix C). For $W_\infty \sim \nu$, define

$$\mu := \mathbb{E}[W_\infty], \qquad \Sigma := \text{Cov}(W_\infty), \qquad G_\infty := \mathbb{E}[R(W_\infty)].$$

3 THE FORGETTING–LOSS EQUIVALENCE

In this section, we prove that distribution-level forgetting approaches population loss at rate $O(T^{-1})$, while both converge to the same stationary limit. This identifies the forgetting plateau with stationary population loss and provides the basis for separating intrinsic loss from sequential price in the next section.

Theorem 3.1 (Asymptotic Equivalence of Forgetting and Population Loss)**.** *Under the assumptions of Section 2, recall that W_T is the parameter obtained after sequentially fitting the first T tasks, F_T is its expected average loss on the first $T-1$ tasks, and $G_T = \mathbb{E}[R(W_T)]$ is its expected population loss. Let ν be the unique invariant distribution established in Section 2. For $W_\infty \sim \nu$, recall that $G_\infty = \mathbb{E}[R(W_\infty)]$ is the stationary population loss. Then*

$$G_T \longrightarrow G_\infty, \qquad |F_T - G_T| = O(T^{-1}).$$

Consequently,

$$F_T \longrightarrow G_\infty.$$

Proof sketch. First, couple the recursion started from W_0 with a second recursion whose initial state is drawn from ν and which is driven by the same task sequence. The strict contraction of $\mathcal{S}_\Pi$ makes the distance between their states converge to zero in L^2, and hence $G_T \to G_\infty$.

We next compare F_T with G_T. Each task τ_s appearing in F_T helped produce W_T, whereas the evaluation task in G_T is independent of W_T. Fix $s < T$ and draw $\tau_s' \sim \Pi$ independently of the original task stream. Run the recursion with τ_s' at time s and with the original tasks at all other times, and denote the resulting final state by $\widetilde{W}_T^{(s)}$. This state has the same distribution as W_T and is independent of τ_s, so

$$\mathbb{E}\Big[\ell_{\tau_s}(\widetilde{W}_T^{(s)})\Big] = G_T.$$

After time s, the original and modified trajectories use the same tasks. Their difference therefore satisfies

$$W_T - \widetilde{W}_T^{(s)} = P_T P_{T-1} \cdots P_{s+1}\big(W_s - \widetilde{W}_s^{(s)}\big).$$

Let $\rho_\Pi := \|\mathcal{S}_\Pi\|_{\mathrm{op},F} < 1$. The contraction of $\mathcal{S}_\Pi$ and the uniform second-moment bounds imply, for a finite constant K independent of s and T,

$$\mathbb{E}\big\|W_T - \widetilde{W}_T^{(s)}\big\|_2^2 \le K\rho_\Pi^{T-s}.$$

Because the task losses are quadratic, $\|C_\tau\|_{\mathrm{op}}$ is bounded, and the relevant second moments are finite, Cauchy–Schwarz then gives

$$|\mathbb{E}[\ell_{\tau_s}(W_T)] - G_T| \le K\rho_\Pi^{(T-s)/2},$$

after enlarging K if necessary. Averaging over the past tasks yields

$$|F_T - G_T| \le \frac{K}{T-1}\sum_{j=1}^{T-1}\rho_\Pi^{j/2} = O(T^{-1}),$$

which proves the claim. A full proof is given in Appendix D.1.

4 INTRINSIC LOSS AND THE SEQUENTIAL PRICE

Having identified the forgetting plateau with stationary population loss, we now separate the two sources of G_∞. The optimum $R^\star$ captures the intrinsic loss induced by task incompatibility, whereas $G_\infty - R^\star$ captures the additional effect of sequential exact fitting. The following theorem defines $G_\infty - R^\star$ as the sequential price and characterizes it through the stationary state.

Theorem 4.1 (Intrinsic Loss and the Sequential Price)**.** *Under the assumptions of Section 2, the stationary population loss satisfies*

$$G_\infty = R^\star + \Delta_{\mathrm{seq}},$$

where the sequential price is

$$\Delta_{\mathrm{seq}} := \|\mu - w_\star\|_{\overline{C}}^2 + \mathrm{tr}(\overline{C}\Sigma), \qquad \mu := \mathbb{E}[W_\infty], \quad \Sigma := \mathrm{Cov}(W_\infty).$$

Here $R^\star$ is the minimum population loss attainable by any learning procedure whose output is a single shared parameter vector, and is also the asymptotic population loss attained by joint training. The stationary moments in the sequential price are characterized by

$$\mu = (I - \mathbb{E}[P_\tau])^{-1}\mathbb{E}[Z_\tau], \qquad \Sigma = (I - \mathcal{S}_\Pi)^{-1}\mathbb{E}[\xi_\tau \xi_\tau^\top],$$

where

$$\xi_\tau := Z_\tau - (I - P_\tau)\mu.$$

Together with Theorem 3.1, this also gives

$$F_T \longrightarrow R^\star + \Delta_{\mathrm{seq}}.$$

Proof sketch. The population-loss identity from Section 2 gives

$$G_\infty - R^\star = \mathbb{E}\|W_\infty - w_\star\|_{\overline{C}}^2.$$

Writing $W_\infty - w_\star = (W_\infty - \mu) + (\mu - w_\star)$ and using $\mathbb{E}[W_\infty - \mu] = 0$ yields

$$G_\infty - R^\star = \|\mu - w_\star\|_{\overline{C}}^2 + \mathrm{tr}(\overline{C}\Sigma).$$

To characterize the two stationary moments, let W'_∞ denote the state after one additional task is applied to W_∞. Stationarity gives

$$W'_\infty = P_\tau W_\infty + Z_\tau, \qquad W'_\infty \overset{d}{=} W_\infty,$$

where W_∞ is independent of the new task τ. Taking expectations gives

$$\mu = \mathbb{E}[P_\tau]\mu + \mathbb{E}[Z_\tau].$$

After subtracting μ and taking covariances, the cross terms vanish and

$$\Sigma = \mathcal{S}_\Pi(\Sigma) + \mathbb{E}[\xi_\tau \xi_\tau^\top].$$

The strict contraction of $\mathcal{S}_\Pi$ then gives the stated formulas for μ and Σ. A full proof is given in Appendix D.2.

The weighting by $\overline{C}$ is inherited directly from the population-loss identity rather than chosen in defining the sequential price. Thus $\|\mu - w_\star\|_{\overline{C}}^2$ is the loss caused by the displacement of the stationary mean from the population minimizer, while $\mathrm{tr}(\overline{C}\Sigma)$ is the loss caused by stationary fluctuations around that mean. The sequential price is therefore the squared bias plus the variance in the geometry induced by the population loss.

In general, μ and $w_\star$ solve different equations,

$$\mu = \mathbb{E}[P_\tau]\mu + \mathbb{E}[Z_\tau], \qquad \overline{C} w_\star = \mathbb{E}[C_\tau Z_\tau],$$

so they need not coincide. For example, consider two equally likely one-dimensional tasks

$$\ell_1(w) = w^2, \qquad \ell_2(w) = 4(w-1)^2.$$

Sequential exact fitting places the parameter at the current task optimum, so in stationarity it is 0 or 1 with equal probability and $\mu = 1/2$. Joint training instead accounts for the different curvatures and gives $w_\star = 4/5$.

The total sequential price remains comparable to the intrinsic loss when the nonzero task curvatures are uniformly comparable, leading to the following corollary.

Corollary 4.2 (Sequential Price under Curvature Variation)**.** *Under the assumptions of Section 2, define $Q_\tau := I - P_\tau$. Suppose that, for some $0 < m \le L < \infty$, almost surely,*

$$mQ_\tau \preceq C_\tau \preceq LQ_\tau, \qquad \kappa := \frac{L}{m}.$$

Then

$$\frac{R^\star}{\kappa} \le \Delta_{\mathrm{seq}} \le \kappa R^\star, \qquad \left(1 + \frac{1}{\kappa}\right) R^\star \le G_\infty \le (1+\kappa)R^\star,$$

and $F_T \to G_\infty$. In particular, if either $C_\tau = cQ_\tau$ almost surely for a fixed $c > 0$ or $C_\tau \equiv C \succ 0$, then

$$\Delta_{\mathrm{seq}} = R^\star, \qquad G_\infty = 2R^\star, \qquad F_T \longrightarrow 2R^\star.$$

Proof sketch. Let $e_t = W_t - w_\star$ and $r_\tau = Q_\tau(Z_\tau - w_\star)$. The centered update is $e_t = P_{\tau_t} e_{t-1} + r_{\tau_t}$, with $P_{\tau_t} r_{\tau_t} = 0$. At stationarity, this orthogonality yields $\mathbb{E}\|W_\infty - w_\star\|_{\overline{Q}}^2 = \mathbb{E}\|r_\tau\|^2$, where $\overline{Q} = \mathbb{E}[Q_\tau]$. The curvature bounds then compare both $R^\star$ and Δ_{seq} to the same projection-space quantity, giving the stated inequalities. The condition controls only the nonzero task curvatures: task ranks and active subspaces may still vary. When $C_\tau = cQ_\tau$, the two comparisons become equalities even though the projections may remain nontrivial. The common positive-definite-curvature case also follows directly because then $P_\tau = 0$. A full proof is given in Appendix D.3.

5 EWC AND THE SEQUENTIAL PRICE

Having separated stationary population loss into intrinsic loss and sequential price, we analyze how fixed-strength EWC changes the sequential price (Kirkpatrick et al., 2017; Schwarz et al., 2018; Heckel, 2022). We use the average task curvature $\overline{C}$ in a fixed population-curvature penalty and study both the resulting stationary population loss and convergence to stationarity under the task distribution of Section 2.

For a fixed $\lambda > 0$, consider the update

$$W_t := \arg\min_w \left\{ \ell_{\tau_t}(w) + \lambda \|w - W_{t-1}\|_{\overline{C}}^2 \right\}.$$

The penalty protects the preceding parameter in directions that are important on average across the task distribution. Its strength remains fixed throughout the task stream. Let G_λ denote the resulting stationary population loss. For $\eta \in (0,1)$, let $T_{\mathrm{mix}}(\eta, \lambda)$ denote the task horizon sufficient for any two trajectories exposed to the same task stream to reduce their expected squared $\overline{C}$-distance to at most an η fraction of its initial value. The formal coupling definition is given in Appendix D.4.

Proposition 5.1 (Exact Finite-Strength Characterization)**.** *Under the assumptions of Section 2, fixed-strength EWC admits a unique stationary distribution within the class of distributions with finite second moment for every $\lambda > 0$. When the task distribution has finite support, its stationary population loss can be computed exactly by solving a finite system of linear equations determined by the task distribution. If the task curvatures are diagonal, these equations further separate into independent scalar equations.*

The corresponding linear system and proof are given in Appendix D.4.

Theorem 5.2 (Sequential Price under Fixed-Strength EWC)**.** *Under the assumptions of Section 2, define*

$$\kappa_\Pi := \frac{1}{2}\mathbb{E}\left[\|C_\tau(Z_\tau - w_\star)\|_{\overline{C}^{-1}}^2 \right].$$

For fixed $\eta \in (0,1)$, as $\lambda \to \infty$,

$$G_\lambda - R^\star = \frac{\kappa_\Pi}{\lambda} + O(\lambda^{-2}), \qquad T_{\mathrm{mix}}(\eta, \lambda) = \frac{\lambda}{2} \log \frac{1}{\eta} + O(1).$$

If $R^\star > 0$, then $\kappa_\Pi > 0$. Thus EWC reduces the stationary sequential price at rate $1/\lambda$, while the horizon for reducing mean-square initialization dependence by a fixed factor grows linearly with λ.

Combining the two asymptotic expressions gives

$$(G_\lambda - R^\star) T_{\mathrm{mix}}(\eta, \lambda) = \frac{\kappa_\Pi}{2} \log \frac{1}{\eta} + O(\lambda^{-1}).$$

For fixed η, reducing stationary sequential price by a constant factor proportionally increases this coupling horizon.

Proof sketch. The first-order condition shows directly how λ changes one task update:

$$W_t - W_{t-1} = -\frac{1}{\lambda} \overline{C}^{-1} C_{\tau_t}(W_{t-1} - Z_{\tau_t}) + O(\lambda^{-2}).$$

Thus, in the large-λ regime, each incoming task changes the parameter by order $1/\lambda$. Because $\mathbb{E}[C_\tau Z_\tau] = \overline{C} w_\star$, averaging over the current task gives

$$\mathbb{E}[W_t - W_{t-1} \mid W_{t-1}] = -\frac{1}{\lambda}(W_{t-1} - w_\star) + O(\lambda^{-2}).$$

Thus the average update pulls the parameter toward $w_\star$. Applying the same expansion to two trajectories exposed to the same tasks shows that their expected squared distance contracts by $1 - 2/\lambda + O(\lambda^{-2})$ per task. After t tasks this factor is approximately $\exp(-2t/\lambda)$, which gives the stated mixing time.

At $w_\star$, a new task injects a random displacement whose expected squared $\overline{C}$-norm is

$$\frac{1}{\lambda^2}\mathbb{E}\|C_\tau(Z_\tau - w_\star)\|^2_{\overline{C}^{-1}} + O(\lambda^{-3}).$$

This is $2\kappa_\Pi/\lambda^2 + O(\lambda^{-3})$. If V_λ denotes the stationary variance contribution to population loss, stationarity balances the error removed by contraction with the noise introduced by the next task:

$$\frac{2}{\lambda}V_\lambda = \frac{2\kappa_\Pi}{\lambda^2} + O(\lambda^{-3}).$$

Hence $V_\lambda = \kappa_\Pi/\lambda + O(\lambda^{-2})$. The squared stationary mean bias is only $O(\lambda^{-2})$, giving the sequential-price formula. A full proof is given in Appendix D.4.

6 Learning from a Stream of User Preferences

We now apply the theory to Jester, a joke-rating dataset with naturally conflicting user preferences. Theorems 3.1 and 4.1 apply directly to its user-level tasks. Proposition 5.1 gives the finite-λ EWC calculation, while Theorem 5.2 describes its price–rate scaling.

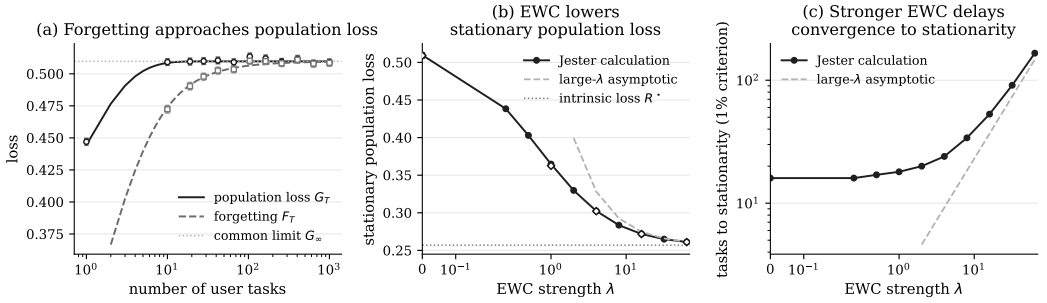


Figure 1: Sequential learning from Jester user preferences. Panel (a) compares the exact forgetting and population-loss curves. Panels (b)–(c) show the stationary population loss and the number of tasks required to reduce the initial-state effect to 1% under EWC. In these two panels, solid curves are exact finite-distribution calculations, dashed curves show the large-λ expressions in Theorem 5.2, and the dotted line marks $R^\star$. Markers show averages over $2{,}000$ independent task streams in panel (a) and $1{,}000$ in panel (b). Error bars denote standard errors and are mostly smaller than the markers.

Jester Dataset 1 (Goldberg et al., 2001) contains ratings by $73{,}421$ users on 100 jokes. One user and all of that user's observed ratings form one task. We retain the original observation pattern, rescale ratings to $[-1, 1]$, and sample tasks independently and uniformly from the empirical user distribution.

Let $\mathcal{J}_i$ be the jokes rated by user i, let $n_i = |\mathcal{J}_i|$, and let Z_{ij} be the corresponding rating. With one-hot joke inputs and a single linear layer $w \in \mathbb{R}^{100}$, task i has loss

$$\ell_i(w) = \frac{1}{n_i}\sum_{j\in\mathcal{J}_i}(w_j - Z_{ij})^2.$$

Its curvature is diagonal, with $(C_i)_{jj} = \mathbf{1}\{j \in \mathcal{J}_i\}/n_i$. Writing $Q_i := \mathrm{diag}(\mathbf{1}\{j \in \mathcal{J}_i\})$, we have $C_i = Q_i/n_i$. Thus task ranks, observed subspaces, and curvature scales vary across users, while $\overline{C} \succ 0$ because every joke has positive coverage. Exact fitting replaces the coordinates rated by the current user and retains the rest. The task distribution falls within the curvature-variation regime of Corollary 4.2 and satisfies the model in Section 2.

Panel (a) shows forgetting and population loss approaching the same value, as predicted by Theorem 3.1. Joint training attains $R^\star = 0.2569$, whereas sequential exact fitting converges to $G_\infty = 0.5097$, approximately 2 times the joint-training loss despite heterogeneous task curvatures. Estimates from independently sampled task streams closely follow these exact curves. Their difference, the sequential price, is 0.2528 and therefore accounts for nearly half of the stationary population loss. Appendix Figure 2 further resolves this sequential price into its stationary-mean and stationary-variance components.

We next evaluate EWC on the same user-task distribution. We use the population mean curvature in the fixed penalty

$$W_t = \arg\min_w \{\ell_{\tau_t}(w) + \lambda \|w - W_{t-1}\|_C^2\},$$

which is the update analyzed in Theorem 5.2. Because the task curvatures are diagonal, the stationary moment equations in Appendix D.4 reduce to independent coordinate-wise calculations.

Panels (b)–(c) show the resulting price–rate trade-off. At $\lambda = 16$, EWC lowers the stationary population loss from 0.5097 to 0.2718, close to $R^\star = 0.2569$, while the number of tasks required by the 1% criterion increases from 16 to 53. The exact finite-distribution curves agree with independently sampled task streams, and the large-λ expressions capture the inverse decrease of the sequential price and the linear increase in the convergence horizon. Full calculations and simulation details are given in Appendix E.

Rotated MNIST exhibits corresponding behavior beyond the exact linear model. With a nonlinear network and finite-step training, historical-task loss approaches population loss, sequential training retains an excess loss over joint training, and moderate functional EWC reduces this empirical sequential price by 95.5%. Full experimental details are provided in Appendix F.

7 Discussion and Scope

Continual-learning sequences often arise from a persistent task-generating source in which the draw of the next task is not determined by the preceding task. Online systems may receive personalized tasks from a stable user population, federated or distributed training may sample participants from a client population in each round, and robots or agents may repeatedly encounter new task instances from a fixed environment distribution. Similarly, an LLM data pipeline may continually generate training data by independently sampling from a prescribed distribution over prompts, topics, or problems. Individual tasks in these settings can have substantially different inputs and targets, while their arrivals follow the same relatively stable generating mechanism. The i.i.d. task distribution in this work abstracts this broad class of settings. It isolates the recency effect induced by sequential access and quantifies its additional cost relative to joint training. When future tasks are instead selected by the current model, its previous performance, or a designed curriculum, the task-generating mechanism evolves with the learning history. Such self-improvement loops and nonstationary task streams constitute important problems beyond the present framework.

Our conclusions depend on the model structure to different degrees. In our analysis, the forgetting–loss equivalence uses i.i.d. task generation and stability of the learning dynamics, while its explicit rate further uses the contractive linear recursion. The sequential price can be defined whenever sequential training has a stable population-loss limit and joint training provides a corresponding optimum. These objects can therefore be formulated in more general continual-learning models. The exact mean–variance expression, the curvature-dependent bounds, and the EWC price–rate law further use the quadratic task geometry and the solvable random updates of our model. The two experiments illustrate these levels. The real Jester user-preference tasks fall exactly within the theoretical framework, and the calculations agree with independently sampled task streams. In the Rotated-MNIST experiment of Appendix F, historical-task loss, population loss, and the empirical sequential price remain consistently measurable under finite-step nonlinear training, and moderate functional EWC substantially reduces this price. The U-shaped effect under strong regularization also shows that the exact EWC asymptotics continue to depend on the dynamics analyzed here.

The analysis begins by asking what forgetting converges to. Forgetting measures the current model's average loss on past tasks, but this value alone mixes two sources: the loss that would remain under joint training and the additional loss caused by sequential updates. Theorem 3.1 shows that forgetting and population loss converge to the same limit. Population loss has an explicit optimal benchmark,

$R^\star$: it is the minimum loss attainable by any shared parameter vector and the loss asymptotically attained by joint training. Subtracting $R^\star$ from the common limit separates task incompatibility from the effect of sequential updates. We define the latter as the sequential price, making the additional effect of sequential access a distinct object of study.

Having isolated the sequential price, we next study two questions. First, which properties of the task distribution determine it? We express the sequential price through the displacement and covariance of the stationary state, and then bound its magnitude through task curvature. Second, how much sequential price can a continual-learning method remove, and how does this affect the decay of initialization dependence? For EWC, we compute both the stationary sequential price and the coupling horizon at fixed relative precision.

The stationary sequential price describes long-run performance, while the coupling horizon describes the decay of initialization dependence at a fixed relative precision. Both quantities are relevant when the task stream has a finite length. Extending this analysis to finite-step nonlinear training, other continual-learning methods such as replay, and nonstationary task streams driven by the learning history are important next steps. Separating what cannot be jointly satisfied by the tasks themselves from what is additionally caused by sequential access provides a basic starting point for understanding and controlling the cost of continual learning.

8 Related Work

Theory of forgetting in linear or overparameterized models analyzes worst-case, random-order, and cyclic projection dynamics for fixed task collections (Evron et al., 2022; Goldfarb & Hand, 2023; Jeong et al., 2023), with extensions addressing task similarity, generalization, classification, and teacher–student dynamics (Lin et al., 2023; Ding et al., 2024; Asanuma et al., 2021). Most directly, *From Order to Distribution* studies a task distribution whose tasks share a solution and characterizes the decay of distribution-level forgetting (Xu & Ma, 2026). We retain this distributional viewpoint while allowing task optima to conflict. The forgetting–loss equivalence then identifies the generally nonzero limit, with the shared-solution result as its zero-loss case.

Task order, access patterns, and optimization paths also affect learning outcomes (Li & Hiratani, 2025; Tsipory et al., 2025; Hess et al., 2024), while multi-task learning describes competition through multi-objective optimization and gradient interference (Sener & Koltun, 2018; Yu et al., 2020; Liu et al., 2021). Within our model, $R^\star$ is the intrinsic loss that remains when one shared model serves the task distribution, whereas the sequential price is the additional loss caused by the access pattern and update rule.

Continual-learning methods preserve historical information through replay or regularization (Lopez-Paz & Ranzato, 2017; Buzzega et al., 2020; Arani et al., 2022). EWC weights parameter changes by their importance to earlier tasks (Kirkpatrick et al., 2017; Schwarz et al., 2018; Benzing, 2022). Our fixed-strength analysis determines the stationary sequential price and the mean-square coupling horizon from the task distribution. On the Jester joke-rating dataset, both quantities are computed exactly for sparse, heterogeneous user tasks and agree with sampled task streams (Goldberg et al., 2001).

AI use statement

Generative AI tools assisted with developing the theoretical formulation, formulating and checking mathematical claims, proof development, literature search, and manuscript drafting and editing. All AI-assisted arguments, citations, and text were manually reviewed. The authors take responsibility for the final content of this work, including all text and claims produced with the aid of generative AI.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

Veronica Alvarez, Santiago Mazuelas, and Jose A. Lozano. Supervised learning with evolving tasks and performance guarantees. *CoRR*, abs/2501.05089, 2025. doi: 10.48550/arXiv.2501.05089. URL https://arxiv.org/abs/2501.05089.

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *ICLR*, 2022. doi: 10.48550/arXiv.2201.12604. URL https://arxiv.org/abs/2201.12604.

Haruka Asanuma, Shiro Takagi, Yoshihiro Nagano, Yuki Yoshida, Yasuhiko Igarashi, and Masato Okada. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, 2021.

Mohammadamin Banayeeanzade, Mahdi Soltanolkotabi, and Mohammad Rostami. Theoretical insights into overparameterized models in multi-task and replay-based continual learning. *CoRR*, abs/2408.16939, 2024. doi: 10.48550/arXiv.2408.16939. URL https://arxiv.org/abs/2408.16939.

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.

Djohan Bonnet, Kellian Cottart, Tifenn Hirtzlin, Tarcisius Januel, Thomas Dalgaty, Elisa Vianello, and Damien Querlioz. Bayesian continual learning and forgetting in neural networks. *Nature Communications*, 16(1):9614, 2025.

Andreas Brandt. The stochastic equation $y_{n+1} = a_n y_n + b_n$ with stationary coefficients. *Advances in Applied Probability*, 18(1):211–220, 1986. doi: 10.2307/1427243.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, 2020. doi: 10.48550/arXiv.2004.07211. URL https://arxiv.org/abs/2004.07211.

Xufeng Cai and Jelena Diakonikolas. Last iterate convergence of incremental methods and applications in continual learning. *CoRR*, abs/2403.06873, 2024. doi: 10.48550/arXiv.2403.06873. URL https://arxiv.org/abs/2403.06873.

Ran Cheng. Context channel capacity: An information-theoretic framework for understanding catastrophic forgetting. *arXiv preprint arXiv:2603.07415*, 2026.

Matthias Delange, 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.

Persi Diaconis and David Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999. doi: 10.1137/S0036144598338446.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

Meng Ding, Kaiyi Ji, Di Wang, and Jinhui Xu. Understanding forgetting in continual learning with linear regression. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10978–11001. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ding24c.html.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 1072–1080. PMLR, 2021.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/evron22a.html.

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *ICML*, pp. 9440–9484, 2023. URL https://proceedings.mlr.press/v202/evron23a.html.

Omer Ginat. The method of alternating projections. *arXiv preprint arXiv:1809.05858*, 2018.

Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001. doi: 10.1023/A:1011419012209.

Daniel Goldfarb and Paul Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *AISTATS*, pp. 2975–2993, 2023. URL https://proceedings.mlr.press/v206/goldfarb23a.html.

Daniel Goldfarb and Paul Hand. Analysis of overparameterization in continual learning under a linear model. *CoRR*, abs/2502.10442, 2025. doi: 10.48550/arXiv.2502.10442. URL https://arxiv.org/abs/2502.10442.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting: An analytical model. In *ICLR*, 2024. doi: 10.48550/arXiv.2401.12617. URL https://arxiv.org/abs/2401.12617.

Reinhard Heckel. Provable continual learning via sketched jacobian approximations. In *AISTATS*, 2022. doi: 10.48550/arXiv.2112.05095. URL https://arxiv.org/abs/2112.05095.

Timm Hess, Tinne Tuytelaars, and Gido M. van de Ven. Two complementary perspectives to continual learning: Ask not only what to optimize, but also how. In *Proceedings of the 1st ContinualAI Unconference*, volume 249 of *Proceedings of Machine Learning Research*, pp. 37–61. PMLR, 2024. URL https://proceedings.mlr.press/v249/hess24a.html.

Naoki Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *Advances in Neural Information Processing Systems 37*, 2024.

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

Hyunji Jung, Hanseul Cho, and Chulhee Yun. Convergence and implicit bias of gradient descent on continual linear classification. In *ICLR*, 2025a. doi: 10.48550/arXiv.2504.12712. URL https://arxiv.org/abs/2504.12712.

Yohan Jung, Hyungi Lee, Wenlong Chen, Thomas Möllenhoff, Yingzhen Li, Juho Lee, and Mohammad Emtiyaz Khan. Compact memory for continual logistic regression. *CoRR*, abs/2511.09167, 2025b. doi: 10.48550/arXiv.2511.09167. URL https://arxiv.org/abs/2511.09167.

Gilad Karpel, Edward Moroshko, Ran Levinstein, Ron Meir, Daniel Soudry, and Itay Evron. Optimal L2 regularization in high-dimensional continual linear regression. *CoRR*, abs/2601.13844, 2026. URL https://arxiv.org/abs/2601.13844.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6109–6119. PMLR, 2021.

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*, pp. 43–65. PMLR, 2023.

Ran Levinstein, Amit Attia, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Itay Evron. Optimal rates in continual linear regression via increasing regularization. *CoRR*, abs/2506.06501, 2025. doi: 10.48550/arXiv.2506.06501. URL https://arxiv.org/abs/2506.06501.

D. Li, Tianqi Wang, Bingrong Xu, Kenji Kawaguchi, Zhigang Zeng, and Ponnuthurai Nagaratnam Suganthan. IF2Net: Innately forgetting-free networks for continual learning. *CoRR*, abs/2306.10480, 2023. doi: 10.48550/arXiv.2306.10480. URL https://arxiv.org/abs/2306.10480.

Haoran Li, Aditya Krishnan, Jingfeng Wu, Soheil Kolouri, Praveen K. Pilly, and Vladimir Braverman. Lifelong learning with sketched structural regularization. In *Proceedings of the 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pp. 985–1000. PMLR, 2021. URL https://proceedings.mlr.press/v157/li21b.html.

Ziyan Li and Naoki Hiratani. Optimal task order for continual learning of multiple tasks. In *ICML*, 2025. doi: 10.48550/arXiv.2502.03350. URL https://arxiv.org/abs/2502.03350.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness B. Shroff. Theory on forgetting and generalization of continual learning. In *ICML*, pp. 21078–21100, 2023. URL https://proceedings.mlr.press/v202/lin23f.html.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6467–6476, 2017.

Yasaman Mahdaviyeh, James Lucas, Mengye Ren, Andreas S. Tolias, Richard S. Zemel, and Toniann Pitassi. Replay can provably increase forgetting. *CoRR*, abs/2506.04377, 2025. doi: 10.48550/arXiv.2506.04377. URL https://arxiv.org/abs/2506.04377.

Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 2022. doi: 10.1016/j.neucom.2021.10.021. URL https://doi.org/10.1016/j.neucom.2021.10.021.

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.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *ICLR*, 2021. URL https://openreview.net/pdf?id=Fmg_fQYUejf.

Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022.

Francesco Mori, Stefano Sarao Mannelli, and Francesca Mignacco. Optimal protocols for continual learning via statistical physics and control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2025. doi: 10.1088/1742-5468/adf296. URL https://doi.org/10.1088/1742-5468/adf296.

Deanna Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010.

Deanna Needell and Joel A. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014.

Peter Oswald and Weiqi Zhou. Convergence analysis for Kaczmarz-type methods in a Hilbert space framework. *Linear Algebra and its Applications*, 478:131–161, 2015.

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.

Liangzu Peng, Juan Elenter, Joshua Agterberg, Alejandro Ribeiro, and René Vidal. LoRanPAC: Low-rank random features and pre-trained models for bridging theory and practice in continual learning. In *ICLR*, 2025. doi: 10.48550/arXiv.2410.00645. URL https://arxiv.org/abs/2410.00645.

Sundar G. Sankaran and A. A. Louis Beex. Convergence behavior of affine projection algorithms. *IEEE Transactions on Signal Processing*, 48(4):1086–1096, 2000.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537. PMLR, 2018. URL https://proceedings.mlr.press/v80/schwarz18a.html.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html.

Brady Steele. Subspace geometry governs catastrophic forgetting in low-rank adaptation. *arXiv preprint arXiv:2603.02224*, 2026.

Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

Matan Tsipory, Ran Levinstein, Itay Evron, Mark Kong, Deanna Needell, and Daniel Soudry. Are greedy task orderings better than random in continual linear regression? *CoRR*, abs/2510.19941, 2025. doi: 10.48550/arXiv.2510.19941. URL https://arxiv.org/abs/2510.19941.

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

Vladimir N. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4, pp. 831–838, 1991.

Chuang Wang, Ameya Agaskar, and Yue M. Lu. Randomized Kaczmarz algorithm for inconsistent linear systems: An exact MSE analysis. *arXiv preprint arXiv:1502.00190*, 2015.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. doi: 10.1109/TPAMI.2024.3367329. URL https://doi.org/10.1109/TPAMI.2024.3367329.

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

Haoming Yang, Ali Hasan, and Vahid Tarokh. Parabolic continual learning. *arXiv preprint arXiv:2503.02117*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

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

Xuyang Zhao, Hui-Yuan Wang, Weiran Huang, and Wei Lin. A statistical theory of regularization-based continual learning. In *ICML*, 2024. doi: 10.48550/arXiv.2406.06213. URL https://arxiv.org/abs/2406.06213.

A Clarifying the Scope and Contribution of the Sequential Price

This work provides a concrete, computable account of the long-run cost of continual adaptation, with distinct roles for the individual results and the established mathematical tools. Theorem 3.1 uses task replacement and contractive coupling to show that the gap between historical forgetting and population loss is $O(T^{-1})$, enabling direct comparison of the forgetting plateau under conflicting tasks with the joint-training optimum. Theorem 4.1 applies the standard quadratic-loss decomposition and moment equations for random affine recursions (Wang et al., 2015) to separate precisely the intrinsic loss caused by task incompatibility from the additional loss incurred by sequential exact fitting. Building on this characterization, Corollary 4.2 translates projection energy balance into a geometric relation between these two losses, identifying conditions for the twofold-loss law and upper and lower bounds under curvature variation; Proposition D.1 further shows that both bounds can be approached arbitrarily closely and provides a task family that departs from the twofold relation. For fixed regularization driven by population curvature, Proposition 5.1 and Theorem 5.2 use stationary-moment analysis and a small-step expansion (Dieuleveut et al., 2017) to characterize how regularization strength changes this additional cost and the time required for initialization effects to decay, with explicit coefficients determined by the task distribution. Together, these results show how established tools for random recursions reveal concrete connections among historical forgetting, task conflict, task geometry, and the cost of continual adaptation.

B Additional Related Work

Catastrophic forgetting has been studied across task-, domain-, and class-incremental continual learning, with broad overviews available in prior surveys (Parisi et al., 2019; Delange et al., 2022; Wang et al., 2024). The distinctions among continual-learning scenarios and evaluation protocols are further developed in prior work (van de Ven et al., 2022; Mai et al., 2022; Lesort et al., 2023). Representative mitigation strategies include replay (Lopez-Paz & Ranzato, 2017; Buzzega et al., 2020; Arani et al., 2022), importance-based regularization (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018), and recent forgetting-resistant architectures, pretrained-model adaptations, or Bayesian updates (Li et al., 2023; Peng et al., 2025; Bonnet et al., 2025). Empirical analyses have also shown that the observed amount of forgetting depends on the training regime, optimization path, model width, and evaluation assumptions (Mirzadeh et al., 2020; 2021; 2022). Our focus is complementary: we quantify the additional stationary population loss induced by a specified sequential learner in a model where that cost can be computed exactly.

The closest theoretical line studies exact-fit or overparameterized linear continual learning through projection dynamics and task geometry. Worst-case, cyclic, and random-order guarantees have been established for this setting (Evron et al., 2022), followed by analyses of random orthogonal tasks, sharper cyclic bounds, and separable classification (Goldfarb & Hand, 2023; Jeong et al., 2023; Evron et al., 2023). Subsequent work treats generalization, task similarity, and linear-regression geometry (Lin et al., 2023; Ding et al., 2024; Goldfarb et al., 2024), as well as the interaction between similarity and overparameterization (Hiratani, 2024; Goldfarb & Hand, 2025; Banayeeanzade et al., 2024). Recent studies further examine task order, replay, and regularization schedules (Li & Hiratani, 2025; Tsipory et al., 2025; Mahdaviyeh et al., 2025), together with last-iterate convergence and optimal linear regularization (Cai & Diakonikolas, 2024; Levinstein et al., 2025; Karpel et al., 2026). Complementary theories use neural-tangent overlap or teacher–student models (Doan et al., 2021; Lee et al., 2021; Asanuma et al., 2021), while other analyses study sketched or compact memory and statistical regularization (Heckel, 2022; Jung et al., 2025a;b). Broader perspectives use statistical physics, parabolic equations, and subspace geometry (Mori et al., 2025; Yang et al., 2025; Steele, 2026), while information-theoretic approaches provide another recent viewpoint (Cheng, 2026). Most

directly, *From Order to Distribution* moves from a fixed task list to i.i.d. draws from a task distribution and characterizes the decay of distribution-level forgetting under a shared solution (Xu & Ma, 2026). We retain that distributional viewpoint while allowing task optima to conflict, which changes the state dynamics from a homogeneous projection recursion to a driven affine recursion and makes the limiting plateau the central object.

Our comparison between sequential and joint training is also related to work on optimization objectives and access patterns. Analyses of evolving tasks, optimization trajectories, and controlled continual-learning protocols show that both the objective and the path used to optimize it can affect retained performance (Alvarez et al., 2025; Hess et al., 2024; Mori et al., 2025). Task ordering provides another mechanism through which sequential access changes learning outcomes (Li & Hiratani, 2025; Tsipory et al., 2025; Goldfarb & Hand, 2023). In multi-task learning, conflicting task objectives have been formalized through multi-objective optimization and gradient interference (Sener & Koltun, 2018; Yu et al., 2020; Liu et al., 2021). Our comparison uses the stationary population loss of sequential exact fitting. We use $R^\star$ for the intrinsic loss, which joint empirical risk minimization asymptotically attains, and define the sequential price as the additional loss of the specified sequential learner. This separates conflict intrinsic to the shared model from the additional cost created by the update rule and access pattern.

Regularization-based continual learning limits movement away from parameters that encode previous tasks. EWC weights this movement by an empirical Fisher matrix, and online EWC consolidates the accumulated information into a running reference state (Kirkpatrick et al., 2017; Schwarz et al., 2018; Benzing, 2022). Related structural regularizers refine or compress the importance geometry (Zenke et al., 2017; Aljundi et al., 2018; Li et al., 2021), and theoretical work studies when quadratic or Jacobian-based regularization can prevent forgetting (Heckel, 2022; Zhao et al., 2024; Lin et al., 2023). Levinstein et al. study vanishing forgetting for jointly realizable tasks under fixed or increasing isotropic regularization (Levinstein et al., 2025), while Karpel et al. study high-dimensional generalization and horizon-dependent isotropic L_2 regularization (Karpel et al., 2026). We instead allow conflicting tasks and characterize stationary sequential price and convergence to stationarity under a fixed population-curvature penalty.

Technically, our recursion is connected to random affine systems and stochastic projection methods. Contractive iterated random functions and affine stochastic equations provide general tools for invariant laws and stationary moments (Brandt, 1986; Diaconis & Freedman, 1999). Randomized Kaczmarz and related projection methods study convergence under successive compatible constraints (Strohmer & Vershynin, 2009; Needell, 2010; Needell & Tropp, 2014), with additional connections to alternating projections and normalized adaptive filtering (Ginat, 2018; Oswald & Zhou, 2015; Sankaran & Beex, 2000). We use these tools to quantify how sequential exact fitting changes the loss induced by task conflict. The resulting analysis connects historical-task loss to the joint-training optimum through stationary population loss, yields curvature-dependent bounds and an exact twofold-loss law in homogeneous settings, and quantifies the EWC price–rate trade-off.

C Auxiliary Results for the Setup

We begin with two elementary facts that formalize the interpretation of $R^\star$ used in the main text.

Lemma C.1 (Population optimum and task compatibility)**.** *Under the assumptions of Section 2,*

$$R(w) = R^\star + \|w - w_\star\|_{\overline{C}}^2 \qquad \text{for every } w \in \mathbb{R}^d.$$

Hence $w_\star$ is the unique minimizer of R. If W is the possibly random output of any learning procedure and $\widetilde{\tau} \sim \Pi$ is an independent evaluation task, then

$$\mathbb{E}[\ell_{\widetilde{\tau}}(W)] = \mathbb{E}[R(W)] \geq R^\star,$$

whenever the expectation is finite, with equality if and only if $W = w_\star$ almost surely. Moreover,

$$R^\star = 0 \quad \Longleftrightarrow \quad \text{there exists } w^\circ \in \mathbb{R}^d \text{ such that } X_\tau w^\circ = y_\tau \text{ for } \Pi\text{-almost every } \tau.$$

Proof. Using $\overline{h} = \overline{C} w_\star$ and expanding the quadratic loss gives

$$\begin{aligned} R(w) &= \mathbb{E}[Z_\tau^\top C_\tau Z_\tau] + w^\top \overline{C} w - 2 w^\top \overline{h} \\ &= R(w_\star) + (w - w_\star)^\top \overline{C} (w - w_\star). \end{aligned}$$

Because $\overline{C} \succ 0$, the second term is nonnegative and vanishes only at $w = w_\star$. Applying this pointwise identity to the random vector W and then taking expectations proves the lower bound and its equality condition. Independence of $\widetilde{\tau}$ gives $\mathbb{E}[\ell_{\widetilde{\tau}}(W)] = \mathbb{E}[R(W)]$.

If a common solution w° exists almost surely, then $R(w^\circ) = 0$, and therefore $R^\star = 0$. Conversely, if $R^\star = 0$, then

$$0 = R(w_\star) = \mathbb{E}[\ell_\tau(w_\star)].$$

The integrand is nonnegative, so $\ell_\tau(w_\star) = 0$ almost surely, which is equivalent to $X_\tau w_\star = y_\tau$ almost surely. Thus $w_\star$ itself is a common solution. □

Lemma C.2 (Consistency of joint training)**.** *Under the assumptions of Section 2, let*

$$\widehat{W}_T^{\mathrm{joint}} \in \arg\min_w \frac{1}{T}\sum_{t=1}^{T} \ell_{\tau_t}(w).$$

Then, almost surely, the empirical objective has the unique minimizer

$$\widehat{W}_T^{\mathrm{joint}} = \overline{C}_T^{-1}\overline{h}_T$$

for all sufficiently large T, where

$$\overline{C}_T := \frac{1}{T}\sum_{t=1}^{T} C_t, \qquad \overline{h}_T := \frac{1}{T}\sum_{t=1}^{T} C_t Z_t.$$

Moreover,

$$\widehat{W}_T^{\mathrm{joint}} \longrightarrow w_\star, \qquad R(\widehat{W}_T^{\mathrm{joint}}) \longrightarrow R^\star \qquad \textit{almost surely}.$$

Proof. The assumptions $\|C_\tau\|_{\mathrm{op}} \le L$ almost surely and $\mathbb{E}\|Z_\tau\|_2^2 < \infty$ imply $\mathbb{E}\|C_\tau Z_\tau\|_2 < \infty$. The strong law of large numbers, applied entrywise in finite dimension, therefore gives

$$\overline{C}_T \longrightarrow \overline{C}, \qquad \overline{h}_T \longrightarrow \overline{h} \qquad \text{almost surely.}$$

Since $\overline{C} \succ 0$, continuity of the smallest eigenvalue implies that $\overline{C}_T \succ 0$ for all sufficiently large T, almost surely. Expanding the empirical objective then shows that its unique minimizer is $\overline{C}_T^{-1}\overline{h}_T$. Continuity of matrix inversion on the positive-definite cone yields

$$\widehat{W}_T^{\mathrm{joint}} = \overline{C}_T^{-1}\overline{h}_T \longrightarrow \overline{C}^{-1}\overline{h} = w_\star \qquad \text{almost surely.}$$

Finally, Lemma C.1 gives

$$R(\widehat{W}_T^{\mathrm{joint}}) - R^\star = \|\widehat{W}_T^{\mathrm{joint}} - w_\star\|_{\overline{C}}^2 \longrightarrow 0.$$

□

We next justify why positive definiteness of the average curvature is only an identifiability condition on the effective parameter space.

Lemma C.3 (Invisible directions)**.** *Let $C_\tau \succeq 0$ and $\overline{C} = \mathbb{E}[C_\tau]$. Then*

$$\ker(\overline{C}) = \{v \in \mathbb{R}^d : C_\tau v = 0 \textit{ almost surely}\}.$$

Consequently, for every $v \in \ker(\overline{C})$, $\ell_\tau(w + v) = \ell_\tau(w)$ and $P_\tau v = v$ almost surely. Thus, if $\overline{C}$ is singular, the analysis may be restricted to $\mathrm{range}(\overline{C}) = \ker(\overline{C})^\perp$ without changing any task loss.

Proof. For any fixed v,

$$v^\top \overline{C} v = \mathbb{E}[v^\top C_\tau v].$$

The integrand is nonnegative. Hence $v^\top \overline{C} v = 0$ if and only if $v^\top C_\tau v = 0$ almost surely, which, because $C_\tau \succeq 0$, is equivalent to $C_\tau v = 0$ almost surely. The two remaining claims follow from $\ell_\tau(w) = \|w - Z_\tau\|_{C_\tau}^2$ and from the fact that P_τ is the orthogonal projector onto $\ker(C_\tau)$. □

Lemma C.4 (Collective coverage implies mean-square contraction)**.** *Suppose that $\overline{C} \succ 0$. On $\mathbb{S}^d$, equipped with the Frobenius norm, define*

$$\mathcal{S}_\Pi(A) := \mathbb{E}[P_\tau A P_\tau].$$

Then

$$\rho_\Pi := \|\mathcal{S}_\Pi\|_{\mathrm{op},F} < 1.$$

In particular, $I - \mathcal{S}_\Pi$ is invertible.

Proof. For each task, the map $\mathcal{Q}_\tau(A) := P_\tau A P_\tau$ is an orthogonal projector on the matrix space $\mathbb{S}^d$ under the Frobenius inner product. Therefore $\mathcal{S}_\Pi = \mathbb{E}[\mathcal{Q}_\tau]$ is self-adjoint, positive semidefinite, and has operator norm at most one.

Suppose for contradiction that $\|\mathcal{S}_\Pi\|_{\mathrm{op},F} = 1$. Since $\mathbb{S}^d$ is finite dimensional, there is a symmetric matrix A with $\|A\|_F = 1$ such that $\mathcal{S}_\Pi(A) = A$. Hence

$$1 = \langle A, \mathcal{S}_\Pi(A)\rangle_F = \mathbb{E}\|P_\tau A P_\tau\|_F^2.$$

Because $\|P_\tau A P_\tau\|_F \le \|A\|_F = 1$, equality implies $P_\tau A P_\tau = A$ almost surely. Choose a nonzero $v \in \mathrm{range}(A)$. Then $P_\tau v = v$ almost surely, so $C_\tau v = 0$ almost surely. It follows that $v^\top \overline{C} v = 0$, contradicting $\overline{C} \succ 0$. Thus $\rho_\Pi < 1$. The Neumann series $(I - \mathcal{S}_\Pi)^{-1} = \sum_{k\ge 0} \mathcal{S}_\Pi^k$ then converges in operator norm. □

Lemma C.5 (Collective coverage implies mean contraction)**.** *Suppose that $\overline{C} \succ 0$ and let $\overline{P} := \mathbb{E}[P_\tau]$. Then*

$$\|\overline{P}\|_{\mathrm{op}} < 1.$$

Consequently, $I - \overline{P}$ is invertible.

Proof. Every P_τ is an orthogonal projector, so $\overline{P}$ is symmetric, positive semidefinite, and has operator norm at most one. Suppose that its operator norm equals one. Then there is a unit vector v such that $\overline{P} v = v$, and hence

$$1 = v^\top \overline{P} v = \mathbb{E}[v^\top P_\tau v] = \mathbb{E}\|P_\tau v\|_2^2.$$

Since $\|P_\tau v\|_2 \le 1$, equality implies $P_\tau v = v$ almost surely. Thus $v \in \ker(C_\tau)$ almost surely and $v^\top \overline{C} v = 0$, contradicting $\overline{C} \succ 0$. Therefore $\|\overline{P}\|_{\mathrm{op}} < 1$, and the inverse is given by the convergent Neumann series $(I - \overline{P})^{-1} = \sum_{k\ge 0} \overline{P}^k$. □

Corollary C.6 (Existence and uniqueness of the stationary law)**.** *Under the assumptions of Section 2, the affine recursion $W_t = P_t W_{t-1} + Z_t$ admits a unique invariant distribution within the class of distributions with finite second moment.*

Proof. Extend the i.i.d. task sequence to integer times and set

$$W_0^{(\infty)} := Z_0 + \sum_{j=1}^{\infty} P_0 P_{-1} \cdots P_{-j+1} Z_{-j}.$$

Let $M_Z := \mathbb{E}[Z_\tau Z_\tau^\top]$ and $\rho_\Pi = \|\mathcal{S}_\Pi\|_{\mathrm{op},F} < 1$. Since Z_{-j} is independent of the later projectors in its summand,

$$\mathbb{E}\big\|P_0 P_{-1} \cdots P_{-j+1} Z_{-j}\big\|_2^2 = \mathrm{tr}\Big(\mathcal{S}_\Pi^j(M_Z)\Big) \le \sqrt{d}\,\rho_\Pi^j \|M_Z\|_F.$$

The series therefore converges in L^2. Shifting all time indices by one shows that its law is invariant under the recursion.

For uniqueness, synchronously couple two copies initialized from any two finite-second-moment invariant laws and use the same future task sequence. Their difference satisfies

$$D_t = P_t P_{t-1} \cdots P_1 D_0,$$

and hence

$$\mathbb{E}\|D_t\|_2^2 = \mathrm{tr}\big(\mathcal{S}_\Pi^t(\mathbb{E}[D_0 D_0^\top])\big) \longrightarrow 0.$$

The two invariant laws must therefore coincide. □

D FULL PROOFS

D.1 PROOF OF THEOREM 3.1

Proof. Write $\rho_\Pi := \|\mathcal{S}_\Pi\|_{\mathrm{op},F} < 1$. We first show that the population loss converges to its stationary value. Let $W_0^{(\nu)} \sim \nu$ be independent of W_0 and of the task stream, and run $W_t^{(\nu)}$ using the same tasks as W_t. Then $W_t^{(\nu)} \sim \nu$ for every t, while

$$D_t := W_t - W_t^{(\nu)} = P_t P_{t-1} \cdots P_1 D_0.$$

Since D_0 is independent of the task stream,

$$\mathbb{E}[D_t D_t^\top] = \mathcal{S}_\Pi^t\big(\mathbb{E}[D_0 D_0^\top]\big)\,,$$

and therefore

$$\mathbb{E}\|D_t\|_2^2 \le \sqrt{d}\,\rho_\Pi^t \left\|\mathbb{E}[D_0 D_0^\top]\right\|_F \longrightarrow 0.$$

In particular, $\sup_t \mathbb{E}\|W_t\|_2^2 < \infty$. Using the population-loss identity from Section 2,

$$\begin{aligned}|G_t - G_\infty| &= \left|\mathbb{E}\|W_t - w_\star\|_{\overline{C}}^2 - \mathbb{E}\|W_t^{(\nu)} - w_\star\|_{\overline{C}}^2\right| \\ &\le \|\overline{C}\|_{\mathrm{op}}\big(\mathbb{E}\|D_t\|_2^2\big)^{1/2}\left(\mathbb{E}\|W_t + W_t^{(\nu)} - 2w_\star\|_2^2\right)^{1/2},\end{aligned}$$

which tends to zero.

We next compare forgetting with population loss. Fix $s < T$ and draw $\tau_s' \sim \Pi$ independently of the original task stream, with associated (P_s', Z_s'). Define a modified trajectory by keeping the original tasks at all times except s and using τ_s' at time s; denote its states by $\widetilde{W}_t^{(s)}$. The final state $\widetilde{W}_T^{(s)}$ has the same distribution as W_T and is independent of the original task τ_s. Consequently,

$$\mathbb{E}\Big[\ell_{\tau_s}(\widetilde{W}_T^{(s)})\Big] = \mathbb{E}[R(\widetilde{W}_T^{(s)})] = G_T.$$

At the replacement time,

$$W_s - \widetilde{W}_s^{(s)} = (P_s - P_s')W_{s-1} + Z_s - Z_s'.$$

Because orthogonal projectors have operator norm at most one, $\sup_t \mathbb{E}\|W_t\|_2^2 < \infty$, and $\mathbb{E}\|Z_\tau\|_2^2 < \infty$, there is a finite constant K_0 such that

$$\sup_{s\ge 1} \mathbb{E}\big\|W_s - \widetilde{W}_s^{(s)}\big\|_2^2 \le K_0.$$

The difference at time s is independent of the future task projectors, and for every $T > s$,

$$W_T - \widetilde{W}_T^{(s)} = P_T P_{T-1} \cdots P_{s+1}\big(W_s - \widetilde{W}_s^{(s)}\big).$$

It follows that, for another finite constant K_1,

$$\mathbb{E}\big\|W_T - \widetilde{W}_T^{(s)}\big\|_2^2 \le K_1 \rho_\Pi^{T-s}.$$

For any $w, v \in \mathbb{R}^d$,

$$\ell_{\tau_s}(w) - \ell_{\tau_s}(v) = (w - v)^\top C_s(w + v - 2Z_s).$$

Since $\|C_s\|_{\mathrm{op}} \le L$ almost surely, Cauchy–Schwarz and the uniform second-moment bounds give a finite constant K_2, independent of s and T, such that

$$\begin{aligned}|\mathbb{E}[\ell_{\tau_s}(W_T)] - G_T| &\le L\big(\mathbb{E}\|W_T - \widetilde{W}_T^{(s)}\|_2^2\big)^{1/2} \\ &\quad \times \left(\mathbb{E}\|W_T + \widetilde{W}_T^{(s)} - 2Z_s\|_2^2\right)^{1/2} \\ &\le K_2 \rho_\Pi^{(T-s)/2}.\end{aligned}$$

Finally,

$$\begin{aligned}|F_T - G_T| &\le \frac{1}{T-1}\sum_{s=1}^{T-1} |\mathbb{E}[\ell_{\tau_s}(W_T)] - G_T| \\ &\le \frac{K_2}{T-1}\sum_{j=1}^{T-1} \rho_\Pi^{j/2} = O(T^{-1}) \longrightarrow 0.\end{aligned}$$

Together with $G_T \to G_\infty$, this proves the theorem. □

D.2 Proof of Theorem 4.1

Proof. Let $W_\infty \sim \nu$. Lemma C.1 gives the pointwise identity

$$R(w) = R^\star + \|w - w_\star\|_{\overline{C}}^2.$$

Taking expectation at W_∞ and writing $W_\infty - w_\star = (W_\infty - \mu) + (\mu - w_\star)$ yields

$$\begin{aligned} G_\infty - R^\star &= \mathbb{E}\|W_\infty - w_\star\|_{\overline{C}}^2 \\ &= \mathbb{E}\|W_\infty - \mu\|_{\overline{C}}^2 + \|\mu - w_\star\|_{\overline{C}}^2 \\ &= \operatorname{tr}(\overline{C}\Sigma) + \|\mu - w_\star\|_{\overline{C}}^2. \end{aligned}$$

The cross term vanishes because $\mathbb{E}[W_\infty - \mu] = 0$. This proves $G_\infty = R^\star + \Delta_{\text{seq}}$ once the stationary moments are identified.

Draw a fresh task $\tau \sim \Pi$ independently of W_∞ and define the one-step update

$$W'_\infty := P_\tau W_\infty + Z_\tau.$$

Stationarity gives $W'_\infty \overset{d}{=} W_\infty$. Taking expectations,

$$\mu = \mathbb{E}[P_\tau]\mu + \mathbb{E}[Z_\tau].$$

Lemma C.5 shows that $I - \mathbb{E}[P_\tau]$ is invertible, and therefore

$$\mu = \big(I - \mathbb{E}[P_\tau]\big)^{-1}\mathbb{E}[Z_\tau].$$

Set $U := W_\infty - \mu$ and

$$\xi_\tau := Z_\tau - (I - P_\tau)\mu.$$

The stationary mean equation implies

$$\mathbb{E}[\xi_\tau] = \mathbb{E}[Z_\tau] - (I - \mathbb{E}[P_\tau])\mu = 0.$$

After centering the one-step recursion,

$$W'_\infty - \mu = P_\tau U + \xi_\tau.$$

The random vector U is independent of the fresh task and has mean zero. Consequently, both cross-covariance terms vanish:

$$\mathbb{E}[P_\tau U \xi_\tau^\top] = \mathbb{E}_\tau\big[P_\tau\, \mathbb{E}[U]\, \xi_\tau^\top\big] = 0,$$

and likewise for its transpose. Taking covariances gives

$$\Sigma = \mathcal{S}_\Pi(\Sigma) + \mathbb{E}[\xi_\tau \xi_\tau^\top].$$

Lemma C.4 makes $I - \mathcal{S}_\Pi$ invertible, so

$$\Sigma = \big(I - \mathcal{S}_\Pi\big)^{-1}\mathbb{E}[\xi_\tau \xi_\tau^\top].$$

The interpretation of $R^\star$ as the optimum attainable by a shared parameter and as the limit of joint training follows from Lemmas C.1 and C.2. Finally, Theorem 3.1 gives $F_T \to G_\infty = R^\star + \Delta_{\text{seq}}$. □

D.3 Proof of Corollary 4.2

Proof. Write $Q_\tau = I - P_\tau$, $\overline{Q} = \mathbb{E}[Q_\tau]$, and

$$r_\tau := Q_\tau(Z_\tau - w_\star).$$

Because $P_\tau Q_\tau = 0$, the centered exact-fitting update is

$$W_t - w_\star = P_{\tau_t}(W_{t-1} - w_\star) + r_{\tau_t}, \qquad P_{\tau_t} r_{\tau_t} = 0.$$

Let $E = W_\infty - w_\star$ and take a fresh task τ independent of E. Stationarity implies that $P_\tau E + r_\tau$ has the same distribution as E. Orthogonality therefore gives

$$\begin{aligned} \mathbb{E}\|E\|^2 &= \mathbb{E}\|P_\tau E\|^2 + \mathbb{E}\|r_\tau\|^2, \\ \mathbb{E}\|E\|_{\overline{Q}}^2 &= \mathbb{E}\|r_\tau\|^2. \end{aligned} \tag{1}$$

Here the second line uses independence and $\|E\|^2 - \mathbb{E}_\tau\|P_\tau E\|^2 = \|E\|^2_{\overline{Q}}$.

Since $C_\tau = C_\tau Q_\tau = Q_\tau C_\tau$, the intrinsic loss is

$$R^\star = \mathbb{E}\|r_\tau\|^2_{C_\tau}.$$

The assumed curvature bounds hence yield

$$m\,\mathbb{E}\|r_\tau\|^2 \le R^\star \le L\,\mathbb{E}\|r_\tau\|^2. \tag{2}$$

Averaging the same bounds gives $m\overline{Q} \preceq \overline{C} \preceq L\overline{Q}$. Together with Theorem 4.1 and Eq. equation 1, this gives

$$m\,\mathbb{E}\|r_\tau\|^2 \le \Delta_{\mathrm{seq}} \le L\,\mathbb{E}\|r_\tau\|^2. \tag{3}$$

Comparing Eqs. equation 2 and equation 3 proves the sequential-price bounds. Adding $R^\star$ gives the bounds on G_∞, and Theorem 3.1 gives $F_T \to G_\infty$.

If $C_\tau = cQ_\tau$ almost surely, then both Eqs. equation 2 and equation 3 are equalities with value $c\,\mathbb{E}\|r_\tau\|^2$. Hence $\Delta_{\mathrm{seq}} = R^\star$, $G_\infty = 2R^\star$, and $F_T \to 2R^\star$. If instead $C_\tau \equiv C \succ 0$, then $P_\tau = 0$ and $W_t = Z_{\tau_t}$, which gives the same twofold conclusion directly. □

Tightness and departures from the twofold law.

Proposition D.1 (Sharpness of the Curvature Bounds)**.** *For each $\kappa \ge 1$, there is a family of task distributions satisfying the assumptions of Corollary 4.2, indexed by $p \in (0,1)$, for which $R^\star > 0$ and*

$$\frac{\Delta_{\mathrm{seq}}}{R^\star} = \frac{1-p}{\kappa} + p\kappa.$$

Consequently, the constants $1/\kappa$ and κ in that corollary are optimal uniformly over the stated class of task distributions.

Proof. Consider the one-dimensional tasks

$$\ell_0(w) = w^2, \qquad \ell_1(w) = \kappa(w-1)^2,$$

sampled with probabilities $1-p$ and p, respectively. Here $Q_\tau = 1$, and the curvature assumption holds with $m = 1$ and $L = \kappa$. Each exact-fit update sets the parameter to the current task optimum, so the stationary parameter W_∞ equals 0 or 1 with probabilities $1-p$ and p. Direct minimization of $R(w) = (1-p)w^2 + p\kappa(w-1)^2$ gives

$$w_\star = \frac{\kappa p}{1-p+\kappa p}, \qquad R^\star = \frac{\kappa p(1-p)}{1-p+\kappa p}.$$

Evaluating the stationary parameter on an independent task gives

$$G_\infty = (1-p)\mathbb{E}[W_\infty^2] + p\kappa\mathbb{E}[(W_\infty - 1)^2] = (1+\kappa)p(1-p).$$

Subtracting $R^\star$ and dividing by it yields the stated ratio. For fixed $\kappa > 1$, the ratio approaches $1/\kappa$ as $p \to 0$ and κ as $p \to 1$, while $R^\star > 0$ for every $p \in (0,1)$. For $\kappa = 1$, the ratio equals 1 for every p. □

For $\kappa = 4$ and $p = 1/2$, this is the two-task example in Section 4, and $G_\infty/R^\star = 3.125$. The mechanism is visible in the weights: exact fitting visits the task optima according to their sampling probabilities, whereas the joint-training optimum also accounts for their curvatures. Changing the frequency of the higher-curvature task can therefore make the sequential price smaller or larger than the intrinsic loss, even with the active subspace held fixed.

D.4 Proofs for Fixed-Strength EWC

Proof of Proposition 5.1. The first-order optimality condition is

$$C_{\tau_t}(W_t - Z_{\tau_t}) + \lambda\overline{C}(W_t - W_{t-1}) = 0.$$

Because $C_{\tau_t} + \lambda\overline{C} \succ 0$, this equation has the unique solution

$$W_t = A_{\tau_t,\lambda}W_{t-1} + B_{\tau_t,\lambda}Z_{\tau_t},$$

where

$$A_{\tau,\lambda} := \lambda(C_\tau + \lambda\overline{C})^{-1}\overline{C}, \qquad B_{\tau,\lambda} := (C_\tau + \lambda\overline{C})^{-1}C_\tau.$$

Their sum is the identity.

We first establish contraction. Introduce the transformed coordinates

$$\widehat{W} := \overline{C}^{1/2}W, \qquad \widehat{Z}_\tau := \overline{C}^{1/2}Z_\tau, \qquad \widehat{C}_\tau := \overline{C}^{-1/2}C_\tau\overline{C}^{-1/2}.$$

Then $\mathbb{E}[\widehat{C}_\tau] = I$, and the transformed old-state matrix is

$$\widehat{A}_{\tau,\lambda} := \overline{C}^{1/2}A_{\tau,\lambda}\overline{C}^{-1/2} = (I + \widehat{C}_\tau/\lambda)^{-1}.$$

It is symmetric and satisfies $0 \preceq \widehat{A}_{\tau,\lambda} \preceq I$. Define

$$\rho_\lambda := \lambda_{\max}\Big(\mathbb{E}[\widehat{A}^2_{\tau,\lambda}]\Big) = \lambda_{\max}\Big(\mathbb{E}[(I + \widehat{C}_\tau/\lambda)^{-2}]\Big).$$

For two chains driven by the same tasks, their transformed difference obeys

$$\widehat{D}_t = \widehat{A}_{\tau_t,\lambda}\widehat{D}_{t-1}.$$

Independence of τ_t and $\widehat{D}_{t-1}$ gives

$$\mathbb{E}\|\widehat{D}_t\|_2^2 \le \rho_\lambda \mathbb{E}\|\widehat{D}_{t-1}\|_2^2.$$

Moreover, $\rho_\lambda < 1$. Otherwise, some unit vector v would satisfy $\mathbb{E}\|\widehat{A}_{\tau,\lambda}v\|_2^2 = 1$. Since every realization is nonexpansive, this would imply $\widehat{C}_\tau v = 0$ almost surely, contradicting $v^\top\mathbb{E}[\widehat{C}_\tau]v = 1$. Iteration proves the stated coupling bound. Together with the finite second moment of the affine forcing, this mean-square contraction gives existence and uniqueness within the class of invariant distributions with finite second moment.

Let W_∞ have this invariant distribution, and define

$$\mu_\lambda := \mathbb{E}[W_\infty], \qquad \Sigma_\lambda := \mathrm{Cov}(W_\infty).$$

Taking expectations in the stationary recursion gives

$$\big(I - \mathbb{E}[A_{\tau,\lambda}]\big)\mu_\lambda = \mathbb{E}[B_{\tau,\lambda}Z_\tau].$$

The inverse exists by the preceding contraction, so

$$\mu_\lambda = \big(I - \mathbb{E}[A_{\tau,\lambda}]\big)^{-1}\mathbb{E}[B_{\tau,\lambda}Z_\tau].$$

Define

$$\xi_{\tau,\lambda} := B_{\tau,\lambda}(Z_\tau - \mu_\lambda), \qquad \mathcal{S}_\lambda(X) := \mathbb{E}[A_{\tau,\lambda}XA_{\tau,\lambda}^\top].$$

Let $\widetilde{W}_\infty$ be an independent stationary state, independent of a fresh task τ, and set

$$W_\infty^+ := A_{\tau,\lambda}\widetilde{W}_\infty + B_{\tau,\lambda}Z_\tau.$$

Then $W_\infty^+ \overset{d}{=} \widetilde{W}_\infty$, and after centering,

$$W_\infty^+ - \mu_\lambda = A_{\tau,\lambda}(\widetilde{W}_\infty - \mu_\lambda) + \xi_{\tau,\lambda},$$

where $\mathbb{E}[\xi_{\tau,\lambda}] = 0$. The cross terms therefore vanish, yielding

$$\Sigma_\lambda = \mathcal{S}_\lambda(\Sigma_\lambda) + \mathbb{E}[\xi_{\tau,\lambda}\xi_{\tau,\lambda}^\top].$$

The same contraction makes $I - \mathcal{S}_\lambda$ invertible, proving the second-moment formula

$$\Sigma_\lambda = \big(I - \mathcal{S}_\lambda\big)^{-1}\mathbb{E}[\xi_{\tau,\lambda}\xi_{\tau,\lambda}^\top].$$

Finally,

$$\mathbb{E}[R(W_\infty)] = R^\star + \mathbb{E}\|W_\infty - w_\star\|_{\overline{C}}^2 = R^\star + \|\mu_\lambda - w_\star\|_{\overline{C}}^2 + \mathrm{tr}(\overline{C}\Sigma_\lambda),$$

so the last two terms give the exact stationary sequential price.

For a finitely supported task distribution, every expectation above is a finite sum. The mean equation is a finite linear system, and vectorizing the covariance equation produces another finite linear system with a unique solution. If the task curvatures are diagonal, then $A_{\tau,\lambda}$ and $B_{\tau,\lambda}$ are diagonal as well, so the moment equations separate coordinate by coordinate. This proves the proposition. □

Proof of Theorem 5.2. Retain the notation established in the preceding proof. Set $\varepsilon := \lambda^{-1}$ and use transformed coordinates throughout. Uniformly over tasks,

$$\widehat{A}_{\tau,\lambda} = I - \varepsilon\widehat{C}_\tau + \varepsilon^2\widehat{C}_\tau^2 + O(\varepsilon^3), \qquad \widehat{B}_{\tau,\lambda} = \varepsilon\widehat{C}_\tau - \varepsilon^2\widehat{C}_\tau^2 + O(\varepsilon^3).$$

The remainders are uniform because $\|C_\tau\|_{\mathrm{op}}$ is bounded and $\overline{C} \succ 0$. Write $\widehat{w}_\star = \overline{C}^{1/2} w_\star$ and $\widehat{d}_\tau = \widehat{Z}_\tau - \widehat{w}_\star$. Population optimality gives

$$\mathbb{E}[\widehat{C}_\tau \widehat{d}_\tau] = 0.$$

The stationary mean equation can thus be written as

$$\mathbb{E}[\widehat{B}_{\tau,\lambda}](\widehat{\mu}_\lambda - \widehat{w}_\star) = \mathbb{E}[\widehat{B}_{\tau,\lambda}\widehat{d}_\tau].$$

Its left matrix is $\varepsilon I + O(\varepsilon^2)$, while the right side is $O(\varepsilon^2)$. Hence $\widehat{\mu}_\lambda - \widehat{w}_\star = O(\varepsilon)$.

The transformed centered forcing satisfies

$$\widehat{\xi}_{\tau,\lambda} = \varepsilon\widehat{C}_\tau\widehat{d}_\tau + O_{L^2}(\varepsilon^2).$$

Define

$$Q := \mathbb{E}\Big[\widehat{C}_\tau \widehat{d}_\tau \widehat{d}_\tau^\top \widehat{C}_\tau\Big].$$

The covariance equation and the expansion of $\widehat{A}_{\tau,\lambda}$ give

$$\widehat{\Sigma}_\lambda = \mathbb{E}[\widehat{A}_{\tau,\lambda}\widehat{\Sigma}_\lambda\widehat{A}_{\tau,\lambda}] + \varepsilon^2 Q + O(\varepsilon^3).$$

More explicitly, the associated operator on symmetric matrices satisfies

$$\mathbb{E}[\widehat{A}_{\tau,\lambda} X \widehat{A}_{\tau,\lambda}] = X - 2\varepsilon X + \mathcal{R}_\varepsilon(X), \qquad \|\mathcal{R}_\varepsilon(X)\|_{\mathrm{F}} \le K\varepsilon^2\|X\|_{\mathrm{F}},$$

because $\mathbb{E}[\widehat{C}_\tau] = I$. Hence its resolvent is $(2\varepsilon)^{-1}I + O(1)$. Applying this resolvent to the centered forcing covariance $\varepsilon^2 Q + O(\varepsilon^3)$ yields

$$\widehat{\Sigma}_\lambda = \frac{\varepsilon}{2} Q + O(\varepsilon^2).$$

Therefore the squared mean bias is $O(\varepsilon^2)$ and

$$\begin{aligned}\mathrm{tr}(\overline{C}\Sigma_\lambda) &= \mathrm{tr}(\widehat{\Sigma}_\lambda) \\ &= \frac{\varepsilon}{2}\mathbb{E}\|\widehat{C}_\tau\widehat{d}_\tau\|_2^2 + O(\varepsilon^2) \\ &= \frac{\kappa_\Pi}{\lambda} + O(\lambda^{-2}).\end{aligned}$$

This proves the expansion of $G_\lambda - R^\star$. If $\kappa_\Pi = 0$, then $C_\tau(Z_\tau - w_\star) = 0$ almost surely, so $\ell_\tau(w_\star) = 0$ almost surely and $R^\star = 0$. Thus $R^\star > 0$ implies $\kappa_\Pi > 0$.

Finally,

$$\mathbb{E}[\widehat{A}_{\tau,\lambda}^2] = I - 2\varepsilon I + O(\varepsilon^2),$$

and hence $\rho_\lambda = 1 - 2/\lambda + O(\lambda^{-2})$. Expanding $\log\rho_\lambda$ and defining

$$T_{\mathrm{mix}}(\eta,\lambda) := \left\lceil \frac{\log\eta}{\log\rho_\lambda} \right\rceil$$

shows that this many tasks are sufficient to reduce the coupled mean-square distance by a factor η. Applying the ceiling changes the expansion by at most a constant, which gives

$$T_{\mathrm{mix}}(\eta,\lambda) = \frac{\lambda}{2}\log\frac{1}{\eta} + O(1).$$

□

E EXPERIMENTAL DETAILS FOR JESTER

E.1 DATA AND TASK DISTRIBUTION

We use Dataset 1 from the Jester release (Goldberg et al., 2001). Its three files contain 73,421 users and 100 joke-rating columns; the first column records the number of ratings made by each user. The value 99 denotes a missing rating. We remove the first column and retain all $N = 73{,}421$ users and all 4,136,360 observed ratings. Dividing observed scores by ten maps the rating scale from $[-10, 10]$ to $[-1, 1]$. A user rates between 15 and 100 jokes, with mean 56.34 and median 52.

User i and the set $\mathcal{J}_i$ of jokes rated by that user define one task. The experimental task distribution is uniform over the N users. Every task arrival is an independent draw with replacement from this distribution, so a user may appear more than once in a task stream. The observation mask and the ratings are kept exactly as recorded in the data.

E.2 MODEL AND EXACT THEORETICAL QUANTITIES

For a one-hot joke input, the single-layer model returns the corresponding coordinate of $w \in \mathbb{R}^{100}$. Write $m_{ij} = \mathbf{1}\{j \in \mathcal{J}_i\}$, $n_i = \sum_j m_{ij}$, and set Z_{ij} to the rescaled rating when $m_{ij} = 1$ and to zero otherwise. Then

$$\ell_i(w) = \sum_{j=1}^{100} c_{ij}(w_j - Z_{ij})^2, \qquad c_{ij} := \frac{m_{ij}}{n_i}, \qquad C_i = \mathrm{diag}(c_{i1}, \ldots, c_{i,100}).$$

Exact fitting overwrites the rated coordinates and retains the unrated ones, so $P_i = I - \mathrm{diag}(m_{i1}, \ldots, m_{i,100})$. The task ranks n_i vary from 15 to 100. Every coordinate has positive empirical coverage, and hence $\overline{C} = N^{-1} \sum_i C_i \succ 0$.

Define $\bar{c}_j = N^{-1} \sum_i c_{ij}$. The population minimizer and intrinsic loss are

$$w_j^\star = \frac{\sum_i c_{ij} Z_{ij}}{N \bar{c}_j}, \qquad R^\star = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{100} c_{ij}(Z_{ij} - w_j^\star)^2.$$

For sequential exact fitting, let

$$p_j := \frac{1}{N} \sum_{i=1}^{N} m_{ij}, \qquad \mu_j := \frac{\sum_i m_{ij} Z_{ij}}{N p_j}, \qquad \sigma_j^2 := \frac{\sum_i m_{ij} Z_{ij}^2}{N p_j} - \mu_j^2.$$

The stationary state at coordinate j is the rating supplied by the most recent task that observed that coordinate. It therefore has mean μ_j and variance σ_j^2. Specializing Theorem 4.1 gives

$$G_\infty = R^\star + \sum_{j=1}^{100} \bar{c}_j \big((\mu_j - w_j^\star)^2 + \sigma_j^2 \big).$$

The three terms are respectively

$$R^\star = 0.2569488, \qquad \sum_j \bar{c}_j (\mu_j - w_j^\star)^2 = 0.0002675, \qquad \sum_j \bar{c}_j \sigma_j^2 = 0.2524891,$$

and hence the sequential price is 0.2527566 and $G_\infty = 0.5097054$.

With $W_0 = 0$, coordinate j has not yet been observed after T tasks with probability $(1 - p_j)^T$. Consequently,

$$\mathbb{E}[W_{Tj}] = (1 - (1 - p_j)^T)\mu_j, \qquad \mathbb{E}[W_{Tj}^2] = (1 - (1 - p_j)^T)(\sigma_j^2 + \mu_j^2),$$

which gives the exact population-loss curve in Figure 1(a). If $g_{\infty,j}$ denotes coordinate j's contribution to G_∞, the corresponding forgetting curve is

$$F_T = \sum_{j=1}^{100} g_{\infty,j} \left(1 - \frac{(1 - p_j)(1 - (1 - p_j)^{T-1})}{(T - 1) p_j} \right), \qquad T \geq 2.$$

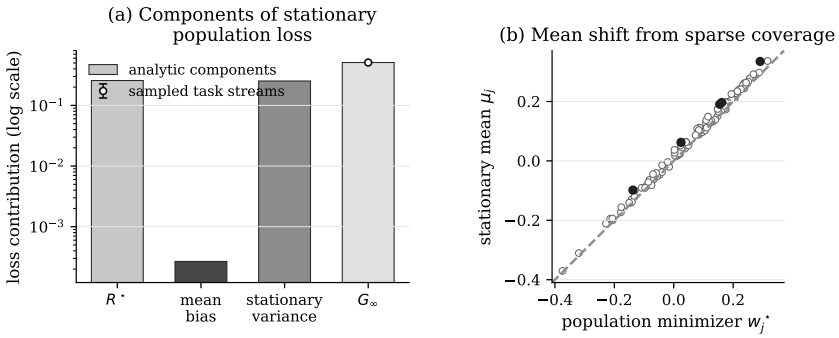


Figure 2: Additional diagnostics for sequential exact fitting on Jester. Panel (a) displays the intrinsic-loss, mean-bias, and stationary-variance components of G_∞, together with an independent task-stream estimate of the total. Panel (b) compares the stationary mean μ_j with the population minimizer $w_j^\star$ for every joke; filled markers identify the five largest coordinate differences.

E.3 Sampled task streams

Panel (a) averages $2{,}000$ independently sampled task streams of length $1{,}000$. At each displayed task index, population loss is evaluated against the complete empirical distribution, while forgetting averages the loss on all earlier tasks in that stream. At task $1{,}000$, the sampled population loss is 0.50899 ± 0.00258 and the sampled forgetting is 0.50854 ± 0.00258, where the reported uncertainties are standard errors.

Appendix Figure 2(a) uses $20{,}000$ independent streams of length 80 to estimate the stationary population loss. The smallest joke-coverage probability is 0.2520, making the probability that any fixed coordinate remains unobserved after 80 tasks less than 10^{-10}. The sampled stationary population loss is 0.50947 ± 0.00081. Panel (b) uses the 100 coordinate-wise population and stationary means computed directly from the empirical distribution.

E.4 EWC calculations

For Figure 1(b)–(c), we use the objective with the fixed population-curvature penalty

$$\ell_i(w) + \lambda \|w - W_{t-1}\|_{\overline{C}}^2.$$

This is the EWC objective in Theorem 5.2. Its coordinate update has the affine form

$$W_{tj} = a_{ij} W_{t-1,j} + b_{ij}, \qquad a_{ij} = \frac{\lambda \overline{c}_j}{c_{ij} + \lambda \overline{c}_j}, \qquad b_{ij} = (1 - a_{ij}) Z_{ij},$$

with $a_{ij} = 1$ and $b_{ij} = 0$ on unrated coordinates. At $\lambda = 0$, this reduces to the exact-fit projection update above.

Let $\alpha_j = \mathbb{E}[a_{ij}]$, $\beta_j = \mathbb{E}[b_{ij}]$, $u_j = \mathbb{E}[a_{ij}^2]$, $v_j = \mathbb{E}[a_{ij} b_{ij}]$, and $s_j = \mathbb{E}[b_{ij}^2]$, where expectations are uniform over users. The stationary coordinate moments are

$$m_j = \frac{\beta_j}{1 - \alpha_j}, \qquad h_j = \frac{2 v_j m_j + s_j}{1 - u_j}.$$

Thus the heterogeneous-curvature curve in panel (b) uses the exact finite-distribution price

$$\Delta_\lambda = \sum_{j=1}^{100} \overline{c}_j \big(h_j - 2 w_j^\star m_j + (w_j^\star)^2 \big).$$

Panel (c) reports the number of tasks required for the worst coordinate's mean-square dependence on initialization to fall below 1%,

$$T_{\mathrm{mix}}(0.01, \lambda) = \left\lceil \frac{\log(0.01)}{\log \rho_\lambda} \right\rceil, \qquad \rho_\lambda := \max_j u_j.$$

For Jester, the distributional constant in Theorem 5.2 is

$$\kappa_\Pi = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{100}\frac{c_{ij}^2(Z_{ij}-w_j^\star)^2}{\bar{c}_j} = 0.2850331.$$

The dashed references show the large-λ expressions $R^\star + \kappa_\Pi/\lambda$ in panel (b) and $(\lambda/2)\log(100)$ in panel (c).

We evaluate $\lambda \in \{0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64\}$. The diamonds in panel (b) independently average 1,000 task streams of length 1,000 at $\lambda \in \{0, 1, 4, 16, 64\}$. Across these five strengths, every sampled stationary population-loss estimate lies within 1.65 standard errors of its exact heterogeneous-curvature value.

F Beyond the Exact Model: Rotated MNIST

We examine how forgetting, population loss, and the sequential price behave in a nonlinear finite-step training problem. This experiment is an out-of-model stress test: the model is a neural network, each task is optimized for a fixed number of steps, and consolidation acts on network outputs. The experiment therefore tests whether forgetting, population loss, the sequential price, and the price–rate trade-off remain informative beyond the setting in which our theorems are exact.

F.1 Task distribution and model

We form four task types by rotating each MNIST image. Their angles are $0°$, $45°$, $90°$, and $135°$. At every task arrival, the angle is sampled independently and uniformly. The task then draws a fresh subset of 5,000 examples from a pool of 50,000 training images. This construction produces an i.i.d. task stream while allowing the input distribution, local network geometry, and learned representation to vary across tasks.

The predictor is a multilayer perceptron with two width-128 ReLU hidden layers and ten outputs. Each task uses mean-squared loss against one-hot class targets and is trained for three epochs with AdamW, learning rate 10^{-3}, batch size 512, and weight decay 0.1. Population quantities are evaluated on all 10,000 test images under each of the four rotations. The empirical joint-training baseline uses the same architecture and optimizer and is trained for 200 passes on the uniform mixture of the four task types.

To align consolidation with the fixed population-curvature penalty in Section 5, we apply the function-space objective

$$\widehat{\ell}_{\tau_t}(\theta) + \lambda\,\mathbb{E}_{x\sim\overline{\mathcal{D}}}\big\|f_\theta(x) - f_{\theta_{t-1}}(x)\big\|_2^2,$$

where $\overline{\mathcal{D}}$ is the uniform mixture of the four rotated input distributions. The expectation is estimated with fresh mixture minibatches, and the preceding network is held fixed when evaluating its outputs. For a linear predictor $f_W(x) = Wx$ with $\overline{C} = \mathbb{E}_{x\sim\overline{\mathcal{D}}}[xx^\top]$, this penalty is exactly

$$\lambda\,\mathrm{tr}\big((W - W_{t-1})\overline{C}(W - W_{t-1})^\top\big),$$

the multi-output form of the population-curvature penalty analyzed in Section 5. We refer to its neural counterpart as functional EWC. Evaluating this penalty uses unlabeled inputs from all four rotations, including those other than the current task. This access mirrors the population curvature used in the theory and aligns the nonlinear experiment with the analyzed regularizer.

We evaluate $\lambda \in \{0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64\}$ over 3,000 tasks and five random seeds. Within each seed, all strengths use the same initialization and task-type sequence. We record every one of the first 20 tasks and then every 20th task.

F.2 Metrics and stationary estimates

At task T, population loss G_T is the uniform average of the current network's test loss over the four rotations. Forgetting F_T averages the same four rotation-specific losses with weights given by their empirical frequencies among the first $T-1$ tasks. Here, F_T measures historical-task loss without subtracting the loss at task acquisition, which need not vanish under finite-step training.

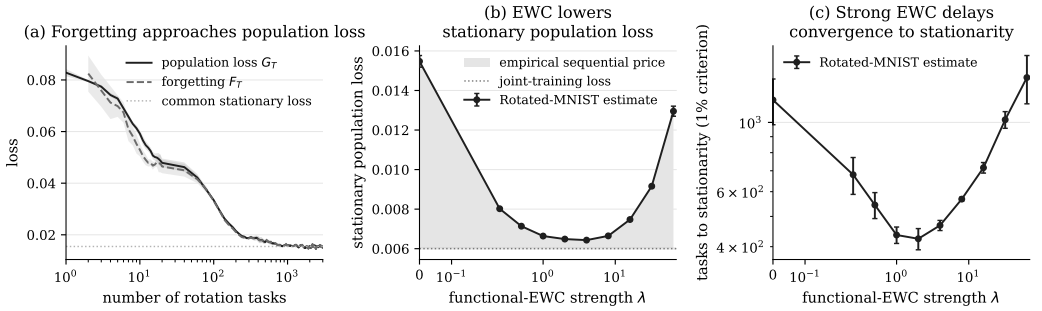


Figure 3: A nonlinear Rotated-MNIST stress test. Panel (a) compares forgetting and population loss under sequential training ($\lambda = 0$). Curves are five-evaluation moving averages and shaded bands denote standard errors across five seeds. Panel (b) reports stationary population loss as a function of functional-EWC strength. The dotted line is the joint-training loss, and the shaded gap is the empirical sequential price. Panel (c) reports the number of tasks required to enter a 1% neighborhood of the stationary population loss. Error bars in panels (b)–(c) denote standard errors across seeds.

For each seed and λ, we estimate stationary population loss by averaging G_T from task $2{,}000$ through task $3{,}000$. The joint-training loss is averaged across the five joint models, and their difference defines the empirical sequential price. Panel (c) first smooths each trajectory over ten recorded evaluations. Its 1% band is one percent of the gap between the task-1 loss and the stationary estimate. The reported relaxation time is the first task beginning ten consecutive recorded evaluations inside this band. All fifty trajectories satisfy this criterion within the experimental horizon; the latest crossing occurs at task $2{,}060$.

F.3 Results beyond the linear dynamics

Panel 3(a) shows forgetting approaching population loss. Over the final $1{,}000$ tasks, their mean absolute difference is 7.95×10^{-5}. Thus the forgetting–loss connection remains visible in this nonlinear task stream. For these four uniformly sampled task types, convergence of the empirical task frequencies to the population weights already implies agreement of the two loss averages when task-wise losses remain bounded. We therefore interpret this observation as qualitative consistency with the forgetting–loss connection.

Panels 3(b)–(c) reveal a U-shaped consolidation trade-off. Without consolidation, stationary population loss is 0.01547 ± 0.00029, compared with the joint-training loss 0.00600 ± 0.00006. At $\lambda = 4$, stationary loss falls to 0.00643 ± 0.00001: the sequential price decreases from 0.00947 to 0.00043, a 95.5% reduction. Moderate consolidation also reaches stationarity fastest; the mean relaxation times are 424 ± 33 tasks at $\lambda = 2$ and 468 ± 17 tasks at $\lambda = 4$.

Stronger consolidation produces a second regime. At $\lambda = 64$, stationary population loss rises to 0.01295 ± 0.00025 and the relaxation time increases to $1{,}392 \pm 245$ tasks. This U-shaped pattern is consistent with moderate functional EWC suppressing variation across successive rotation tasks, while a very large penalty preserves more of the network's earlier approximation error and induces under-adaptation. This bias is absent from the exact linear per-task minimization analyzed in Section 5. The forgetting–loss comparison and sequential-price benchmark therefore remain informative in this nonlinear experiment, whereas the monotone large-λ law remains specific to the exact theoretical dynamics.