DURABLE REPAIR UNDER STORED-COORDINATE BUDGETS: A RIDGE-TRAINED CHECKPOINT SEPARATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study one-shot factor edits followed by compulsory ridge-regularized minibatch SGD. Capability is the best probability of keeping both task risks safe throughout the continuation window, optimized over the complete legal action class. In a rank-40 linear learner at batch sizes near $2{,}000$, the same stored-coordinate row-write budget gives certain 840-step survival at a young checkpoint but at most $.001$ survival after old training, on a history event of probability at least $.967$. A task subset of prior mass 7.98% is certified repairable at editing time; its contribution to the young-minus-old capability gap exceeds $.0768$. More mobility restores reliable survival. The separation holds on an explicit joint neighborhood in cap, ridge, step size, and risk threshold, including caps on both sides of the exact-representation energy. This neighborhood is small and establishes local openness. The proof combines a full-factor lower bound, feasible witnesses, and stopped finite-batch estimates, with executable arithmetic certificates. Seven-history policy experiments support the ridge mechanism; nonlinear age effects are mixed. The result separates age-generated checkpoint states, without matching their predictors or risks or establishing a representation-invariant limitation.

1 INTRODUCTION

A repair that works at the editing instant may fail after ordinary training resumes. Empirical studies already document this fragility for edited language-model knowledge and text-to-image models (Wen & Zhang, 2025; He et al., 2025; Cheng et al., 2026). We ask a stronger optimization question: among all one-shot edits obeying the same movement, forgetting, and capacity budgets, what is the largest probability of keeping both old and new tasks safe throughout a prescribed continuation window?

At a checkpoint the policy observes the model and the declared task, edits once, and relinquishes control. Fresh minibatch SGD then continues unchanged. This separates immediate repairability, the durability of a selected heuristic, and the best durability attainable by the complete legal action class. The last quantity matters for an impossibility result: failure of several editing methods does not imply that every legal repair must fail.

The mechanism and its scope. Our learner is a ridge-regularized two-layer linear network with a stored factor-row movement budget. Old training changes the predictor, its risk, and retained decoder rows together. Their position changes which edited factorizations the budget can reach, while the resumed optimizer acts on those factors. The resulting chain is

$$\text{old SGD history} \longrightarrow \text{retained stored rows} \longrightarrow \text{best legal edit} \longrightarrow \text{survival under fresh SGD.}$$

The cost is in stored factor coordinates. At the nominal tuple, $K = 192 = 2\,\mathrm{tr}\,B_\theta$ is exactly the minimum factor energy for the noiseless future teacher; it leaves no energy reserve at exact fit. Ridge-induced drift alone is unsurprising. The question is whether the complete budget-constrained action class can counter it at different age-generated checkpoints. The retained-row movement and factor-energy cap drive the separation; the other budgets complete the intervention protocol and are slack on the safety gate.

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.

Results. At batch sizes near 2,000, the same mobility budget gives young capability one and old capability at most .001 on a common history event. On a task subset of mass 7.98%, the old checkpoint is nevertheless immediately repairable; its contribution to the integrated gap exceeds .0768. A small explicit joint parameter box crosses $K = 192$ while retaining the full-action separation. The mathematical evidence covers every legal factor endpoint; the experiments evaluate one declared repair policy. The supplement includes the exact interval cells, checkpoint arrays, 841-state mean replay, raw experimental records, and a single reproduction entry point.

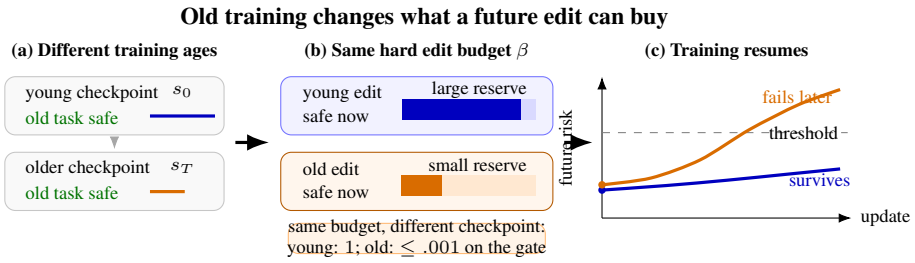

Figure 1: Mechanism overview (schematic, not data). Old SGD moves retained stored rows. Under the same hard movement budget, this changes the best attainable reserve against future SGD. The age theorem compares checkpoints under the same risk limits; their predictors and risks are not matched.

What age means here. Age selects a checkpoint distribution through old SGD. It is not an additional state variable of the future learner: identical full checkpoints have identical capability (Appendix F.9). We do not isolate a causal history effect at matched predictor or risk, or prove monotonicity at intermediate ages. The coordinate cost is an explicit intervention contract, not a measured hardware cost. Its orthogonal covariance and limits are stated separately.

2 Operational Formulation

Before giving the model, we fix four terms used throughout. An *edit* is one checkpoint-dependent parameter change made before future minibatches are seen; here "repair" means this constrained adaptation, not factual knowledge editing. *Legal* means that the edit obeys the same movement, forgetting, and capacity budgets as every competing strategy. *Survival through N steps* means that both task risks stay below threshold at the editing instant and after every one of the next N SGD updates. The *optimal lifetime* is obtained after optimizing over all legal edits, rather than tracking one selected heuristic.

2.1 Learner and Tasks

Let $r = 40$, $k = 16$, and $d = r + k = 56$. The predictor is the two-factor linear network

$$f_{V,W}(x) = V\big(Wx + \delta e_{\text{bias}}\big), \qquad V, W \in \mathbb{R}^{d\times d}, \tag{1}$$

where the fixed column e_{bias} is part of the factor parameterization and cannot be rotated away. Each input is independently uniform on the radius-$\sqrt{d}$ sphere. Both factors receive simultaneous gradients computed from the same ordinary minibatch; no antithetic pairing or empirical centering is imposed. We write m_{o} and m_{f} for the old- and future-phase batch sizes. The theorems use the diagonal specialization $m_{\text{o}} = m_{\text{f}} = m$ unless stated otherwise; the batch diagnostic in Section 4 instead fixes m_{o} and varies m_{f}. Training uses step size $\eta = 1/1{,}680{,}000$, ridge $\lambda = 0.1$, and a projection onto the factor-energy cap

$$\|V\|_F^2 + \|W\|_F^2 \le K, \qquad K = 192. \tag{2}$$

The bias and old-label-noise ranges are $\delta \in [5\times10^{-4}, 10^{-3}]$ and $\nu \in [5\times10^{-3}, 10^{-2}]$. We write $z_t^{(m_{\text{o}})} = (V_t, W_t)$ for the old SGD trajectory and suppress the superscript when the batch size is fixed.

Let $J_0 = I_{r/2} \otimes \mathrm{diag}(1, -1)$, let R_θ apply the same two-dimensional rotation by θ in every active pair, and set $J_\theta = R_\theta J_0 R_\theta^\top$. The old and future teachers are

$$B_{\mathrm{o}} = \mathrm{diag}(2I_r + 0.01J_0, 0_k), \qquad B_\theta = \mathrm{diag}(2I_r + 0.01J_\theta, I_k), \tag{3}$$

so the future task rotates the active spectrum and introduces k previously missing directions. Old labels are $B_{\mathrm{o}}x + \zeta$, with independent coordinates $\zeta_i \sim \mathrm{Unif}[-\nu, \nu]$; future labels are exactly $B_\theta x$. The common angle is uniform on $[0, \pi/2]$ and independent of initialization and data. The task event $\sin^2\theta \le 1/64$ used for certified repairability and the successful old branch has mass $(2/\pi)\arcsin(1/8) > 7/88$. Initialization is uniform in the full factor-pair ball of radius 10^{-3} around

$$\bar{V}_0 = \bar{W}_0 = \mathrm{diag}\Big((2I_r + 0.01J_0)^{1/2}, \sqrt{0.05/k}\, I_k\Big). \tag{4}$$

Old ages satisfy $\eta T \in [1.95, 2]$.

For a teacher B, the population prediction risk uses the half-square convention

$$R_B(V, W) = \tfrac{1}{2}\mathbb{E}\|f_{V,W}(x) - Bx\|^2 = \tfrac{1}{2}\big(\|VW - B\|_F^2 + \delta^2\|Ve_{\mathrm{bias}}\|^2\big). \tag{5}$$

Label noise affects old training, but is not included in this prediction-risk readout. We use R_{o} and R_{f} for the old and future teachers. Success requires

$$R_{\mathrm{f}}(z_j) < \rho k, \qquad R_{\mathrm{o}}(z_j) < 0.51k, \qquad \rho = 2 \times 10^{-6}, \qquad 0 \le j \le N. \tag{6}$$

We denote this whole-window event by Surv_N.

2.2 One-shot stored-coordinate edits and compulsory continuation

At time T, a policy observes the full checkpoint z_T and the declared future task, but not future minibatches. It chooses one edited endpoint $a = (V^a, W^a)$ from the hard action set $\mathcal{A}_\beta(z_T, \theta)$. The scalar β is the allowed movement of the retained old decoder rows, normalized by $\sqrt{r}$; we call it the *mobility budget*. The remaining constraints retain the factorization: encoder and missing-row movements are bounded, incremental old risk is bounded, the distinguished column is respected, and the cap remains satisfied. The complete numerical specification is in Appendix A. Policies may be measurable and privately randomized. After the edit, no further action is allowed; the learner follows fresh same-minibatch SGD.

Why this intervention model? Allowing repeated corrections would mix two questions: whether the checkpoint admits a durable repair and how much future control is continually injected. We isolate the former by charging one checkpoint-relative edit and then restoring the unchanged learner. The budgets act on V and W, rather than only on their product, because SGD also updates these stored factors; a quotient or predictor-space cost would define a different feasible set and continuation problem.

Protected-block interpretation. The intervention model can be read literally as a write budget on named stored parameter blocks: the retained decoder module permits squared edit movement at most $r\beta^2$, the other decoder module at most $k(1.3)^2$, and the encoder at most $36k$. The bias is wired to the fixed hidden coordinate e_1. After this single intervention, the same stored coordinates undergo SGD; the write restriction does not freeze them during training. A factor rescaling changes stored parameters and is charged even when VW is unchanged. This defines an explicit protected-block intervention, but is not a validated hardware-cost law or an optimizer-state model. Balanced-factor, quotient, or predictor-space budgets define different problems; no conclusion for those metrics is claimed.

Definition 1 (Optimal survival capability)**.** For future batch size m_{f}, horizon N, checkpoint z_T, task θ, and mobility β, define

$$V_{N,m_{\mathrm{f}}}(z_T, \theta, \beta) = \sup_{\pi \in \mathcal{A}_\beta} \Pr(\mathsf{Surv}_N \mid z_T). \tag{7}$$

The supremum is over the complete hard-legal action class, not over a prescribed repair family. An endpoint failing either edit-time risk threshold has survival probability zero; an empty action set is assigned value zero.

Definition 2 (Probability-level lifetime)**.** For $p \in (0,1)$ and a truncation $N_{\max} = 1680$, let

$$N_p^\star(z_T, \theta, \beta) = \max\big(\{-1\} \cup \{N \in \{0, \dots, N_{\max}\} : V_{N,m_{\mathrm{f}}}(z_T, \theta, \beta) \geq p\}\big). \tag{8}$$

The sentinel -1 means that even edit-time success fails to reach confidence p; it is not negative physical time. Survival events are nested, so $V_{N,m_{\mathrm{f}}}$ is nonincreasing in N. When $m_{\mathrm{o}} = m_{\mathrm{f}} = m$, we write $V_{N,m}$.

We now define the deterministic object used in the proofs. Let Φ_B be one exact projected population-gradient update for teacher B, with the same step size, ridge, bias column, and cap as the stochastic learner. For an endpoint a, set

$$\bar{z}_0^a = a, \qquad \bar{z}_{j+1}^a = \Phi_{B_\theta}(\bar{z}_j^a), \tag{9}$$

and define its normalized whole-window margin

$$\chi_N(a) = \max_{0 \leq j \leq N} \left\{ \frac{R_{\mathrm{f}}(\bar{z}_j^a)}{k} - \rho, \frac{R_{\mathrm{o}}(\bar{z}_j^a)}{k} - 0.51 \right\}. \tag{10}$$

Let S_β be the closure of hard-legal endpoints satisfying both edit-time risk thresholds strictly. This filters only endpoints whose stochastic survival is already zero. Writing A for the retained old decoder rows, the deterministic comparison value is

$$J_N(A, \beta) = \inf_{a \in S_\beta(A,\theta)} \chi_N(a), \qquad \inf \varnothing = +\infty. \tag{11}$$

The closure gives compact deterministic minima. A strictly negative margin with slack transfers to a nearby initially strict legal endpoint by continuity and the definition of closure; lower bounds remain valid on the larger closed set. On the full history/action gate, encoder movement, missing-row movement, and forgetting are redundant within this safe cap-constrained set. The active restrictions are retained-row mobility and the factor-energy cap. Only on that gate may the checkpoint be abbreviated by A. Old risk stays slack and future risk increases along every such mean path. Thus

$$J_N(A, \beta) = \min_{a \in S_\beta(A,\theta)} \frac{R_{\mathrm{f}}(\bar{z}_N^a)}{k} - \rho. \tag{12}$$

The sign of J_N is converted to stochastic capability by a uniform whole-window estimate. Crucially, J_N is a value function: the optimizer may change with N, and no particular factor lift is frozen.

3 Full-action checkpoint separation and joint robustness

The main results are a moderate-batch separation and its joint parameter neighborhood. Appendix F gives the analytic reductions and their inputs; Appendix F.2 defines the history event and proves every entrance bound. A single current-parameter certificate, described in Appendix C, verifies the numerical implications without executing historical versions.

3.1 A concrete full-action capability crossing

We first establish the operational transition at finite noise. Set $m_{\mathrm{o}} = m_{\mathrm{f}} = m$, let $\Xi = \eta d^2/m$, and consider the box $\Xi \in [8 \times 10^{-7}, 10^{-6}]$, equivalently $m \in \{1867, \dots, 2333\}$.

Theorem 3 (Concrete 840-step capability separation)**.** *For every initializer in the ball defined by equation 4, there is one old-history event $\mathcal{E}_H$, shared by all listed tasks and mobilities, with*

$$\Pr(\mathcal{E}_H \mid u) \geq 1 - 33000\Xi \geq 0.967. \tag{13}$$

On $\mathcal{E}_H$, the low-budget conclusion holds for every task angle; the high-budget conclusion holds for $\sin^2\theta \leq 1/64$*:*

$$\beta \in [0.0093, 0.0098] \Longrightarrow V_{840,m}(z_T, \theta, \beta) \leq 0.001, \tag{14}$$

$$\beta \in [0.0118, 0.0119] \Longrightarrow V_{840,m}(z_T, \theta, \beta) \geq 0.999. \tag{15}$$

The first statement covers every hard-legal strategy; the second is achieved by a measurable factor-level edit satisfying all budgets. After integrating the stated history, initializer, and task laws, the optimized capability gap is greater than 0.0767*.*

The proof pairs a global factor-space lower bound with a constructive full-factor upper bound, leaving a gap below 2×10^{-7} in normalized risk. Initially safe legal endpoints are certified in the low-mobility regime on the stated angle event. The following result compares ages at that same budget, instead of inferring an age effect from two budgets.

Proposition 4 (Same-budget checkpoint-state separation)**.** *For every initializer in equation 4, every task angle, every future batch size, and every $\beta \in [.0093, .0098]$, the young checkpoint at $T = 0$ satisfies $V_{840,m_f}(u, \theta, \beta) = 1$. The balanced teacher edit achieves this for every future sample path. At the old age and batch sizes of Theorem 3, the same budget has capability at most .001 on its common event. Integrating the stated laws, the young-minus-old capability gap exceeds .0768.*

Appendix F.6 proves young survival samplewise. The old failure bound is angle-independent, but immediate old-checkpoint repairability is guaranteed only on the 7.98% task subset. We report its $> .0768$ contribution to distinguish loss of durability from possible edit-time infeasibility elsewhere. The full survival-capability gap exceeds .9660; that larger figure includes tasks without certified immediate repairability. These checkpoints are not matched in predictor, old risk, or edit-time risk frontier. No monotonicity at intermediate ages is asserted.

Proposition 5 (Risk-threshold interval in the rank-40 model)**.** *Keeping the learner and hard budgets fixed, replace the future-risk threshold by any $\rho' \in [1.99, 2.00] \times 10^{-6}$. The conclusions and probabilities of Theorem 3 and Proposition 4 hold simultaneously throughout this interval, with the same budgets and history event.*

The initially safe action class is allowed to change with ρ'. For the lower conclusion it shrinks; the explicit upper construction has initial risk below 1.965333×10^{-6} and retains more than 4×10^{-8} terminal slack at the smallest threshold. This exceeds the stochastic risk buffer 2×10^{-8}. Young risk is below 1.45×10^{-7} on every sample path. This baseline interval is one-sided. The next result permits simultaneous perturbations, including both signs of the cap change.

3.2 A JOINT NEIGHBORHOOD ACROSS THE EXACT-REPRESENTATION CAP

Proposition 6 (Joint rank-40 robustness)**.** *Let $\eta_0 = 1/1680000$. Replace the cap, ridge, step size, and future-risk threshold in both phases by any tuple satisfying*

$$\begin{aligned} |K - 192| \le 10^{-5}, \quad |\lambda - .1| \le 2 \times 10^{-7}, \\ |\eta/\eta_0 - 1| \le 10^{-6}, \quad |\rho - 2 \times 10^{-6}| \le 10^{-8}. \end{aligned} \tag{16}$$

Keep an integer old age with $\eta_0 T \in [1.95, 2]$ and $m \in \{1867, \dots, 2333\}$. For every initializer and each fixed tuple, there is one history event of probability at least .967, shared by all tasks and budgets. On it, the low/high conclusions of Theorem 3 hold with the same budget intervals and $.001/.999$ probabilities. Young capability remains one on every sample path. The low-budget old checkpoint remains immediately repairable on the 7.98% task subset, whose contribution to the same-budget checkpoint gap still exceeds .0768.

The closed box has an open interior containing the nominal tuple, including $K < 192$ and $K > 192$. Its widths are small: cap varies by only 10^{-5}, ridge by 2×10^{-7}, and step size by one part per million. This proves local openness, not wide parameter tolerance or persistence without ridge. The age is a fixed update count, rather than a newly rounded count as η changes. Appendix F.8 proves the joint old-history, all-action lower, cap-scaled upper, and samplewise young bounds. A separate 2048-cell rational cover checks the perturbed lower envelope; no parameter grid or assumed continuity argument replaces the proof.

Additional quantitative results. Appendix B gives the saved-checkpoint mean-threshold bracket $[.01052240335, .010917]$ and the horizon-to-lifetime bridge. Neither computes an optimizer fiber or a stochastic median threshold. Appendix E retains only a short conditional integer-law implication; its full-action Gaussian approximation is an explicit unverified assumption for this instance.

4 EXPERIMENTS: WHAT IS VISIBLE AT ACCESSIBLE SCALE?

The rank-40 simulations test a specified repair and its local noise mechanism. Proposition 7 now bounds the complete-action mean threshold at the same checkpoint used for the scaling experiment; it does not identify the global optimizer. A separate scalar control has a certified global optimum.

We test whether old training moves the row resource and shortens lifetime (Q1), whether mobility shifts that front (Q2), and whether paired ridge removal delays it (Q3).

Protocol. We simulated *seven independent old histories* of the rank-40 learner, each for 3,360,000 updates with $m_o = 2048$. These seven histories are the independent units for history effects. Reusing their checkpoints, we ran 2,568 future trajectories, paired across budgets within each grid. An extended scan and 48 admissible outside-angle continuations give 2,872 baseline trajectories; another 48 are retained as invalid endpoint diagnostics. The primary grid crosses five old ages, seven mobilities, and four future seeds per history at $m_f = 2048$. A paired ridge intervention holds checkpoints, endpoints, and minibatch seeds fixed. Secondary tests vary future batch, angle, precision, and a local-search endpoint. Crossing means the first future-risk violation; old risk and cap activity are recorded separately.

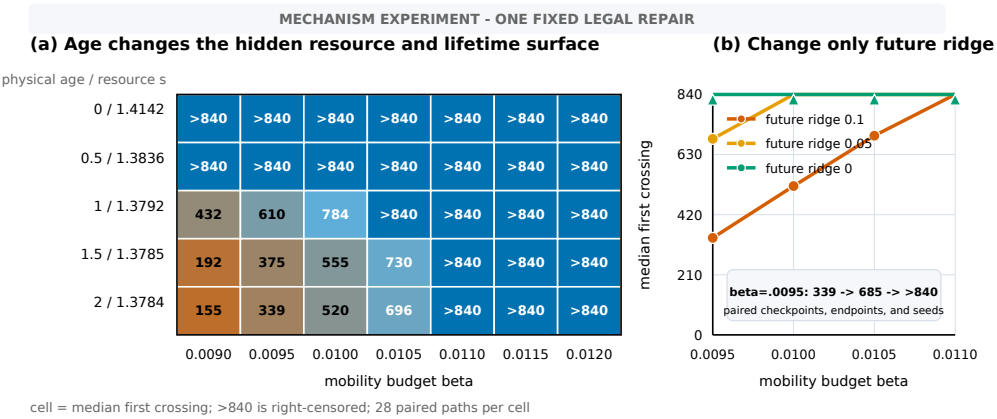

Figure 2: Accessible-scale mechanism tests for one explicit hard-legal repair. **Left:** each cell reports the median first crossing from 28 paired paths; the row label exposes both physical age and retained-row resource s. Older checkpoints move the resource downward and shift the mobility–lifetime front. **Right:** changing only future ridge delays the same front; at $\beta = 0.0095$, the median moves from 339 to 685 to beyond 840 updates.

Q1–Q2: history moves a resource and the lifetime surface with it. The median retained-row coordinate falls from 1.4142 at age 0 to 1.3836, 1.3792, 1.3785, and 1.3784 at physical ages 0.5, 1, 1.5, and 2. At mobility $\beta = 0.0095$, the corresponding median first crossings are beyond 840, beyond 840, 610, 375, and 339 updates; first crossing is nonincreasing with age at each tested mobility. Appendix Figure 5 displays the seven history-level medians rather than pooling them into one curve. Across uncensored history–age–mobility cells, future-seed variability is at most two updates, so the visible uncertainty is chiefly between old histories. On one checkpoint, extending the horizon to 1500 produces crossings from 62 to 1364 over sixteen mobility values; a quadratic describes these points with maximum residual below 0.44 update (Appendix Figure 3).

Q3: deterministic future drift is causally active. At age 2 and mobility $\beta = 0.0095$, future ridge 0.1 gives median crossing 339; halving it gives 685; removing it prevents crossing throughout all 840 updates. The ordering persists across the paired mobility grid, and an independently seeded 0.05 run reproduces the 685-update median. This intervention isolates deterministic ridge drift as one driver of the failure front.

It also narrows the interpretation: at the tested scale, the front is *ridge-driven*, with minibatch noise providing a smaller perturbation. We have not proved an all-action separation for $\lambda = 0$, and the zero-

ridge explicit repair survives the observed window. The experiment therefore supports regularized continued-training dynamics, not a ridge-free plasticity claim.

A direct local scaling test for the explicit repair. We fixed one age-2 checkpoint and numerically bracketed the repair's population 840-step root at $\widehat{\beta}_{\mathrm{exp}} \approx 0.01091621322674$. An independent 64-path calibration fixed a single budget scale $a = 5.152 \times 10^{-8}$. Fresh validation used $\beta = \widehat{\beta}_{\mathrm{exp}} + qa/\sqrt{m_{\mathrm{f}}}$, seven declared offsets $q \in \{-2, -1, -.5, 0, .5, 1, 2\}$, $m_{\mathrm{f}} \in \{256, 1024, 4096\}$, and 128 paths per batch–budget cell. The curves have maximum between-batch probability spread .094 and maximum deviation .076 from the fixed $\Phi(q)$ reference. Every endpoint passes the numerical legality audit, with no old-risk violation or cap projection. Figure 6 reports Wilson intervals; Appendix D.1 gives the preregistered protocol and complete path records. Old batch stays 2048, so this is conditional policy evidence, not a test of the all-action or joint-history limit.

A global bound at the same checkpoint. For the fixed checkpoint above, the certified mean threshold lies in $[.01052240335, .010917]$. The upper endpoint is checked against every hard budget using exact rational arithmetic and against all 841 mean states with explicit rounding allowances. At that budget, its risk is at most 1.999800×10^{-6}; the all-action lower bound exceeds 1.897109×10^{-6}. Thus the policy's normalized-risk gap from the optimum is at most 1.027×10^{-7}. This is a bound, not a claim that the numerical policy root equals the complete-action root or that the same gap holds for nonlinear models.

Task-angle scope. At .15 radians the explicit repair remains legal and has crossings $(340, 783, > 840)$ at budgets $(.0095, .01075, .0115)$. The .25 endpoints exceed the cap by about 4×10^{-5}; their 48 paths are excluded and retained as invalid diagnostics. The .50 construction also violates the cap. The valid .15 check does not extend the certified immediate-repairability or high-mobility angle range. The full-action low-mobility failure bound itself holds for every angle.

A nonlinear control gives mixed evidence. In a small $V \tanh(Wx)$ learner, a separately calibrated threshold and 576 validation continuations show budget and ridge sensitivity (Appendix D.2). With retained decoder rows fixed, median first unsafe observations at old ages $0, 500, 2000$ are $210, 160, 190$; allowing movement $.05\sqrt{8}$ brings all three to 210. All paired zero-ridge continuations remain safe at the sampled times through 500. The oldest checkpoint partially recovers, so this control does not support universal monotone deterioration with age. These are fixed-sample risk readouts every ten updates, not a nonlinear population-survival theorem.

A globally checkable control. For the scalar two-factor model in Appendix G, the full-action optimal 100-step budget is isolated in an interval of width 5.83×10^{-13} about .0366135223364. An independent 90-digit physical-factor replay verifies both bracket signs. At a smaller common budget .005, young capability is one and old capability zero throughout simultaneous perturbations $K \in [1.99, 2.01]$, $\eta_f \in [.00095, .00105]$, $\lambda_f \in [.095, .105]$, and $\epsilon \in [.0245, .0255]$, where the risk threshold is $\epsilon^2/2$. This proves a robust deterministic all-action separation across the exact-representation cap. Scalar sphere gradients are noiseless, so it does not validate an optimized Gaussian limit in rank 40.

Local-search diagnostic. An eight-restart full-factor search finds a legal but dynamically worse endpoint: under 64 paired paths, the explicit repair crosses at 835 and the candidate at 229. The candidate also has worse edit-time risk (1.835 versus 1.424, in 10^{-6} normalized units). This records an unsuccessful search, not an optimality gap or an empirical reversal between edit-time loss and durability.

Appendix Table 3 separates tested axes from parameters that remain calibrated or open.

The collapse concerns one policy at one checkpoint. A linear crossing at j means survival through $j - 1$; the nonlinear control only locates sampled violations.

5 Related work

Edits under subsequent training. Wen & Zhang (2025) study retention of edited knowledge after LLM fine-tuning and find greater fragility than for pretrained knowledge. He et al. (2025) study text-to-image models and observe that later fine-tuning often reverses edits. Cheng et al. (2026), originally circulated as *Quantifying Edits Decay in Fine-tuned LLMs*, compare edit retention after full fine-tuning, LoRA, and DoRA across multiple models and editing methods. These are direct empirical precedents for post-edit durability. Our contribution concerns a complete constrained-action survival value and its analysis under a specified stochastic continuation; we do not claim priority for the phenomenon of edits being lost.

Training age can harm later adaptation in deep, continual, language, and linear models (Dohare et al., 2024; Achille et al., 2019; Kleinman et al., 2024; Springer et al., 2025); repeated resets instead maintain trainability through continuing control (Farias & Jozefiak, 2025). We allow one charged edit, then resume the original learner, and optimize whole-window survival. This differs both from post-edit fine-tuning measurements and from age-dependent adaptation; repeated editing and lifelong edit maintenance are outside our intervention model.

Task overlap, initialization, and imbalance shape transfer dynamics (Tahir et al., 2025; Saxe et al., 2014; Braun et al., 2022; Dominé et al., 2025); factor geometry and balancing are established tools (Tu et al., 2016; Srebro et al., 2004; Lindsey & Menon, 2026). Value sensitivity and path survival also have mature theories (Milgrom & Segal, 2002; Fang & Santos, 2019; Fisac et al., 2015; Sivaramakrishnan et al., 2024). Our model-specific step connects realized finite-step SGD to a stored-coordinate action class and bounds complete-action survival without fixing a common optimizer across horizons. The joint neighborhood establishes local persistence of this construction. Table 2 gives a fuller comparison.

6 Discussion and limitations

The theorem gives a separation between ridge-trained checkpoint states under a stored-coordinate block-write contract. The active constraints are retained-row mobility and factor energy; the other intervention budgets have slack on the main safety gate. An explicit joint neighborhood crosses the nominal exact-fit cap, so the result does not require the equality $K = 192$ exactly. The certified widths are nevertheless small and do not establish broad robustness.

The comparison is not predictor-matched or risk-matched and does not identify history as a cause beyond the complete checkpoint state. Immediate repairability and restored old survival are certified on only 7.98% of the task prior. The all-angle failure bound does not enlarge that repairability claim. Seven-history experiments evaluate a fixed policy, not the optimized capability. The nonlinear control has nonmonotone age effects, and removing ridge prevents observed crossings. Alternative edit metrics, matched-checkpoint separation, broad nonlinear validity, and a verified full-action Gaussian limit remain open. The written proofs, exact arithmetic, and bounded-roundoff replay are auditable evidence, not proof-assistant verification.

AI-USE DISCLOSURE

OpenAI Codex (desktop) and ChatGPT (web) were used as research-support tools. Model identifiers were not consistently available and are therefore not inferred.

We formulated all research questions, candidate proof strategies, and mathematical directions pursued in this work. Codex was used to organize our existing research results, summarize proved statements and unresolved boundaries, and translate our proposed proof approaches into structured prompts and verification checklists. The web interface was used to assess the apparent feasibility of these approaches, identify possible obstacles, and search broadly for relevant prior work or potentially applicable techniques, including work from adjacent research areas. Typical prompts described our current results and one or more proof approaches that we wished to evaluate, and asked the system to locate related literature, compare the approaches, identify missing hypotheses or likely failure points, and suggest primary sources for further examination.

We independently determined whether any retrieved source or suggested connection was mathematically applicable. We carried out all mathematical derivations, proofs, constructions, computations, production of exact certificates, and final research decisions reported in the paper. AI-generated feasibility assessments and literature suggestions were treated only as research aids and not as mathematical evidence. Unsupported or inapplicable suggestions were discarded.

Codex was additionally used to check our arguments for logical gaps, quantifier changes, and omitted hypotheses; to translate proof obligations that we had already specified into machine-checkable verification workflows; and to assist with language and organizational editing, citation checking, LaTeX formatting, pagination, and rendered-layout inspection. Editorial and organizational assistance affected the abstract, the introduction, the presentation of related work, the technical exposition, the appendices, section ordering, this disclosure, and local wording throughout the manuscript. In each case, the starting material consisted of our results, arguments, and source notes; the tools were not used to originate any theorem, proof strategy, construction, computation, certificate, or claimed interpretation. We reviewed every adopted change, verified the cited sources and mathematical claims, and retain responsibility for the entire paper.

REPRODUCIBILITY STATEMENT

Appendix A specifies the model, action class, and quantifiers; Appendix C maps the written proof obligations to the executable certificate. The anonymous supplement includes fixed inputs, checkpoints, seeds, per-trajectory records, environment requirements, and scripts for the current certificate, data audits, and figure reconstruction. The entry point `python scripts/verify_submission.py` verifies input hashes and executes the current numerical chain with Python and NumPy. Appendix D states the experimental recording and sampling limits. Analytic proofs remain separate from finite arithmetic verification.

ETHICS STATEMENT

This work studies mathematical models and synthetic numerical experiments. It does not involve human participants or personal data. The conclusions are limited to the specified edit budgets and continuation dynamics; broader deployment, safety, or fairness benefits are not established.

REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkeStsCcKQ.

Lukas Braun, Clémentine C. J. Dominé, James E. Fitzgerald, and Andrew M. Saxe. Exact learning dynamics of deep linear networks with prior knowledge. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6615–6629, 2022.

Yinjie Cheng, Paul Youssef, Christin Seifert, Jörg Schlötterer, and Zhixue Zhao. Can fine-tuning erase your edits? on the fragile coexistence of knowledge editing and adaptation. *arXiv preprint*

arXiv:2511.05852v4, 2026. URL https://arxiv.org/abs/2511.05852v4. Accepted to KDD 2026; originally circulated as Quantifying Edits Decay in Fine-tuned LLMs.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632: 768–774, 2024. doi: 10.1038/s41586-024-07711-7.

Clémentine C. J. Dominé, Nicolas Anguita, Alexandra M. Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fde1a69a5b6e554b2f1f727197d2651d-Abstract-Conference.html.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412, 2019. doi: 10.1093/restud/rdy049.

Vivek F. Farias and Adam D. Jozefiak. Self-normalized resets for plasticity in continual learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/359c0c4dc53b6bf64041b4c8334f1c11-Abstract-Conference.html.

Jaime F. Fisac, Mo Chen, Claire J. Tomlin, and S. Shankar Sastry. Reach-avoid problems with time-varying dynamics, targets and constraints. In *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, pp. 11–20, 2015. doi: 10.1145/2728606.2728612.

Feng He, Zhenyang Liu, Marco Valentino, and Zhixue Zhao. How robust is model editing after fine-tuning? an empirical study on text-to-image diffusion models. *arXiv preprint arXiv:2506.18428*, 2025. URL https://arxiv.org/abs/2506.18428.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods emerge even in deep linear networks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Aq35gl2c1k.

Kathryn Lindsey and Govind Menon. Regularization implies balancedness in the deep linear network. *arXiv preprint arXiv:2511.01137*, 2026. URL https://arxiv.org/abs/2511.01137v2.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, 2002. doi: 10.1111/1468-0262.00296.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://iclr.cc/archive/2014/conference-proceedings/.

Karthik Sivaramakrishnan, Vignesh Sivaramakrishnan, Rosalyn A. Devonport, and Meeko M. K. Oishi. Stochastic reachability of uncontrolled systems via probability measures: Approximation via deep neural networks. In *2024 IEEE 63rd Conference on Decision and Control*, pp. 7534–7541, 2024. doi: 10.1109/CDC56724.2024.10886521.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56719–56789, 2025. URL https://proceedings.mlr.press/v267/springer25a.html.

Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, volume 17, 2004. URL https://papers.nips.cc/paper_files/paper/2004/hash/e0688d13958a19e087e123148555e4b4-Paper.pdf.

Javan Tahir, Surya Ganguli, and Grant M. Rotskoff. Features are fate: A theory of transfer learning in high-dimensional regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58142–58168, 2025. URL https://proceedings.mlr.press/v267/tahir25a.html.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 416–425, 2016. URL https://proceedings.mlr.press/v48/tu16.html.

Fufang Wen and Shichang Zhang. Retention analysis of edited knowledge after fine-tuning. *arXiv preprint arXiv:2507.14198*, 2025. URL https://arxiv.org/abs/2507.14198.

A Exact model and action contract

This appendix makes the main statements' quantifiers and numerical action contract explicit.

Order of randomness and decisions. First fix the initializer u, task θ, and deterministic age T. Any declared numerical budget is fixed before the old minibatches are drawn. The realized old trajectory produces A_T. An edit may read the complete checkpoint and the declared task, but it may not read future minibatches. Fresh future batches are then sampled and integrated inside $V_{N,m_{\mathrm{f}}}$. Finally, the outer statements integrate over old histories, tasks, and initializers.

Parameter box. Unless a theorem explicitly says otherwise,

$$\begin{gathered} r = 40, \quad k = 16, \quad d = 56, \quad K = 192, \quad \lambda = 0.1, \quad \eta = 1/1{,}680{,}000, \\ \rho = 2 \times 10^{-6}, \quad \delta \in [0.0005, 0.001], \quad \nu \in [0.005, 0.01], \quad \eta T \in [1.95, 2], \quad N \le 1680. \end{gathered} \tag{17}$$

The old and future phases use the same integer batch size $m_{\mathrm{o}} = m_{\mathrm{f}} = m$ in Theorem 3 and Proposition 6. The successful old branch and immediate-repairability guarantee use $\sin^2\theta \le 1/64$.

Stored-coordinate action class. Write A_T for the r retained decoder rows and Z_T for the missing rows. Hard budgets require: (i) $\|A^a - A_T\|_F \le \beta\sqrt{r}$; (ii) missing-decoder movement at most $1.3\sqrt{k}$; (iii) encoder movement at most $6\sqrt{k}$; (iv) incremental old-task risk at most $0.6k$; and (v) factor energy at most K. The bias amplitude is not edited, and $e_{\mathrm{bias}} = e_1$ remains the fixed unit coordinate. Its bias output $c = Ve_1$ moves only through the edited factor; for invertible V, $c^\top(VV^\top)^{-1}c = 1$ holds identically, not as a relaxed inequality. Both edit-time risk thresholds belong to the survival event. Endpoints failing them have zero survival; deterministic proofs use the compact closure S_β of initially strict safe endpoints. All cross blocks, hidden frames, imbalance, measurable checkpoint dependence, and private randomization remain allowed. The successful construction is asserted on its verified gate only.

B Full statements behind the resource–lifetime bridge

B.1 An executable bracket at an experimental checkpoint

Define the generalized mean threshold $\beta_c(z,\theta) = \inf\{\beta \in [.0093, .0119] : J_{840}(z,\theta,\beta) < 0\}$. This definition permits a flat zero interval and does not require a unique optimal factorization or a differentiable value function.

Proposition 7 (A complete-action bracket at a saved checkpoint)**.** *At history* 0 *of the saved age-*2 *experiment, interpreted as exact binary64 coordinates, with* $\theta = 0$ *and* $\delta = .001$*,*

$$.01052240335 \le \beta_c \le .010917, \qquad \textit{width} < 3.946 \times 10^{-4}. \tag{18}$$

At the upper budget, a checked stored-coordinate edit has whole-window mean risk less than ρ *and normalized-risk suboptimality less than* 1.027×10^{-7} *relative to the complete-action optimum.*

The lower side is a bound over all legal factors; the upper side is a feasible policy witness. Appendix F.11 gives the exact input hash, rational budget checks, and propagated rounding allowances. The interval is over five times narrower than $(.0098, .0118)$, but still does not isolate the critical fiber, evaluate optimized Gaussian scales, or supply a stochastic median threshold. It establishes a quantitative bridge between the complete-action theorem and the policy experiment at the same checkpoint.

B.2 A FULL-ACTION RESOURCE–LIFETIME BRIDGE

The first result shows that the optimized margin crosses the risk boundary at a controlled speed even though the best edit is allowed to change with the horizon.

Theorem 8 (Uniform horizon transversality). *On the full history/action gate, with $\|A\|_F/\sqrt{r} < 1.379$ and a nonempty initially safe action set, for $0 < \beta \le 0.0122$ and $0 \le n < n' \le 1680$,*

$$V_{n',m} \le V_{n,m}, \qquad \frac{\eta(n'-n)}{4000} \le J_{n'}(A,\beta) - J_n(A,\beta) \le \frac{\eta(n'-n)}{200}. \tag{19}$$

The margin inequalities cover the complete initially safe action class and do not assume that one optimizer remains active across horizons.

Combined with a uniform future-path buffer $\bar{e}$, Theorem 8 brackets every probability-level lifetime between the last horizon with $J_n < -\bar{e}$ and the first horizon with $J_n > \bar{e}$. This turns a risk-margin certificate into a lifetime certificate.

Let $J_n(A,\beta)$ denote the optimized mean-path risk margin for the complete stored-coordinate action set. The finite stochastic converter supplies a uniform whole-window error $\bar{e} = 3 \times 10^{-8}$ whenever $m \ge 1867$. Define

$$n_- = \max\big(\{-1\} \cup \{n \le 1680 : J_n < -\bar{e}\}\big), \qquad n_+ = \min\big(\{1681\} \cup \{n \le 1680 : J_n > \bar{e}\}\big) - 1. \tag{20}$$

Then, for all $p \in (0.001, 0.999)$,

$$n_- \le N_p^\star \le n_+. \tag{21}$$

When both buffer roots lie in the stated window, Theorem 8 gives $n_+ - n_- \le 404$.

For the finite box $m \in \{1867, \dots, 2333\}$ and $\eta T \in [1.95, 2]$, the same historical event yields

$$\beta \in [0.0093, 0.0098] \Longrightarrow N_p^\star \le 833, \qquad \beta \in [0.0118, 0.0119] \Longrightarrow N_p^\star \ge 846. \tag{22}$$

This coarse result is numerically usable but does not resolve the critical cell. The conditional implication in Appendix E does not provide an instance-specific critical center.

C CURRENT CERTIFICATE AND PROOF DEPENDENCY MAP

The accompanying anonymous ZIP is part of the review artifact: the PDF alone does not contain executable inputs. Run `python scripts/verify_submission.py` for the mathematical chain, or `python scripts/reproduce.py` for that chain together with stored-data audits, scalar checks, and figure reconstruction. The README gives dependency installation and an optional paper-build flag. The package contains fixed SHA-256 input hashes, complete interval cells, the original checkpoint and portable arrays, the feasible endpoint, all 841 checked mean states, and experimental records with their original sampling limits.

Table 1: Analytic obligations and their finite computational dependencies. A program verifies the numerical column; it does not prove the universal scope.

claim and scope	written argument	numerical evidence
all initializers; common old event	Appendix F.2	old decay, mean comparison, joint variance, stopped coefficient
all initially safe endpoints	Appendix F.4	future tube and stochastic risk buffer
global lower / feasible upper	Appendix F.5	2048 rational cells; independent 512-cell enclosure
every young sample path	Appendix F.6	sample Hessian, gradient, movement and risk bounds
all horizon optimizers	Appendix F.7	retained-row induction and per-step risk slopes
joint parameter box; full actions	Appendix F.8	old and future perturbation bounds; second 2048-cell cover
one exact saved checkpoint	Appendix F.11	rational budgets; 841-state bounded-roundoff witness
state sufficiency / covariance	Appendix F.9	analytic coupling; no numerical input
conditional integer implication	Appendix E	no verified instance-specific Gaussian input

Exact parameter contract. The active spectrum is $\{1.99, 2.01\}$, its squared deviation from $2I_r$ is .004, $\operatorname{tr} B_\theta = 96$, and $\|B_\theta\|_F^2 = 176.004$. The nominal parameters and joint box are written as exact fractions in `parameters.json`. The verifier checks the immutable input manifest, the acyclic dependency graph, and the spectrum before executing the programs. The paper and programs are locked together; no archival proof module is used.

What is and is not certified. The arithmetic verifies finite implications of the written full-factor, stopping, and parameter-perturbation arguments. Bounded-roundoff replay verifies the saved endpoint under explicit error allowances; it is not exact real arithmetic execution of its historical training run. The empirical audits recompute stored summaries, not 23.52 million old-training updates. No program establishes LICQ for all closest lifts, a uniform multiplier bound, a full-fiber directional expansion, or a complete-action Gaussian converter. Those claims are not dependencies of the submitted numerical separation.

Why matching the predictor does not determine the next update. For scalar factors $p = vw$, a simultaneous population step for $\frac{1}{2}(p-b)^2 + \frac{\lambda}{2}(v^2+w^2)$ gives

$$p^+ = p - \eta\{(p-b)(v^2+w^2) + 2\lambda p\} + \eta^2\{(p-b)^2 p + \lambda(p-b)(v^2+w^2) + \lambda^2 p\}. \tag{23}$$

Rescaling (v, w) to $(cv, w/c)$ preserves p but generally changes the next prediction. This identity explains why factor state matters; it is not a matched-predictor capability-separation theorem or a hardware cost model.

D Additional experimental details

Table 2: Positioning relative to nearby questions. The defining distinction is what happens after the intervention and what outcome is optimized.

line of work	checkpoint source	control after switch	primary outcome
critical-period studies	models trained to different ages	ordinary target training	adaptation or transfer
plasticity interventions	online learner state	repeated resets or control	maintained trainability
feature-transfer theory	pretrained representation	target-task training	transfer or generalization
long-pretraining studies	realized pretrained model	fine-tuning	later optimization or loss
post-edit retention (Wen & Zhang, 2025; He et al., 2025)	pretrained edited model	later fine-tuning	persistence of selected edits
this paper	one old-SGD checkpoint	one charged edit, then none	best whole-window survival

Table 3: What the experiments do and do not vary. "Open" means that the paper does not claim robustness along that axis.

axis	tested range	conclusion
old age	$\eta T \in \{0, .5, 1, 1.5, 2\}$	resource and explicit-repair lifetime move together
mobility	seven primary budgets; 16-point scan	monotone lifetime front for the tested repair
future ridge	$0, .05, .1$ paired	dominant accessible-scale causal factor; zero ridge is censored at 840
batch	256–8192; local test 256–4096	coarse noise diagnostic and fixed-checkpoint explicit-policy collapse
task angle	0–.15 rad admissible	.25 and .50 endpoints violate the cap; repairability certificate stops at .1253
risk threshold	rank-40: $[1.99, 2.00] \times 10^{-6}$	certified one-sided interval; Proposition 5
step, cap, ridge	rank-40 box in equation 16; scalar box in equation 116	local rank-40 openness; broader scalar control is separate
action metric	stored factor movement	alternatives are open

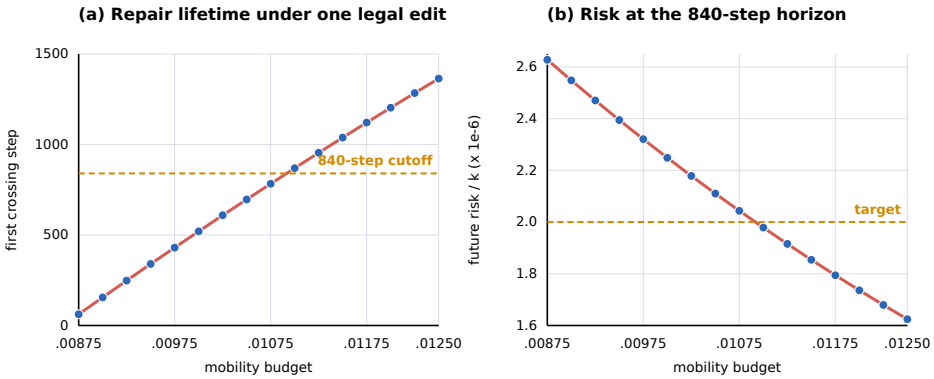

Figure 3: Extended mobility scan on one realized checkpoint. The left panel shows first future-risk crossing; the right panel shows normalized future risk at update 840. Both panels evaluate the same explicit legal repair.

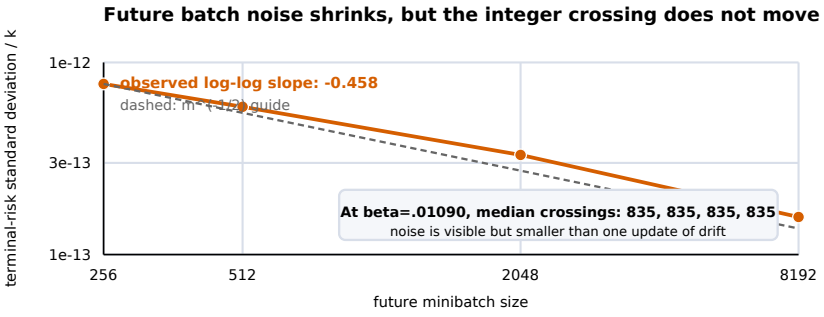

Figure 4: Batch-size diagnostic for the same explicit repair. Terminal-risk variability decreases approximately at the square-root rate, but the integer crossing at $\beta = 0.01090$ remains update 835 for every tested batch. This is exclusionary evidence that accessible-scale sampling noise is smaller than one update of deterministic drift.

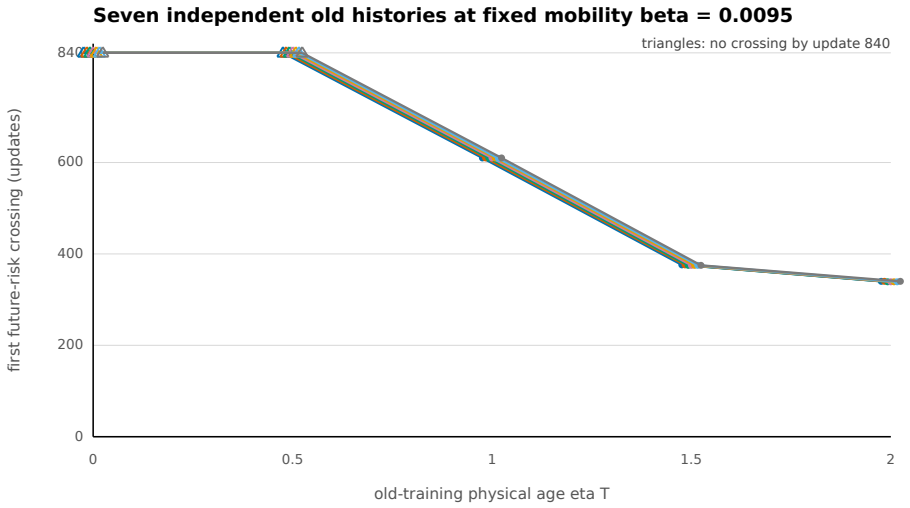

Figure 5: History-level rather than trajectory-pooled results at $\beta = 0.0095$. Each color is one of seven independently trained old histories; each point is the median over four fresh future continuations. Triangles are right-censored at update 840. Future-seed variation is tiny here, so the figure exposes the effective replication level for claims about old training.

Table 4: History-level summary at $\beta = 0.0095$. Each history contributes one median over four fresh future continuations. A value > 840 is right-censored, not treated as the number 840.

physical age ηT	history-median range	histories censored	pooled median
0	> 840	$7/7$	> 840
0.5	> 840	$7/7$	> 840
1	609–610	$0/7$	610
1.5	375–376	$0/7$	375
2	339–340	$0/7$	339

Table 5: Executed linear baseline evidence. "Crossing" means first future-risk violation; all observed crossings preserve the old-risk and cap constraints.

component	configurations	trajectories	main use
old histories	7	23.52M updates	age-conditioned checkpoints
age–mobility grid	5×7	980	resource–lifetime surface
future-ridge ablation	3×4 plus sensitivity	448	paired mechanism intervention
batch grid	4×3	768	noise-scale diagnostic
task-angle grid	6×3 admissible	288	policy sensitivity; another 48 invalid paths are excluded
FP64 replication	1	4	numerical check
local action search	8 restarts	128	negative optimizer diagnostic
single-history extended scan	two 16×8 grids	256	lifetime-curve shape

The baseline artifact stores locked seeds, saved old checkpoints, per-trajectory crossing records, and risk tables at selected steps. It does not store every step of every historical run. Simulation code and per-run configurations are included. Every data figure is regenerated from these records by `scripts/rebuild_figures.py`; no experimental summary array is hard-coded into the figure builders. The new local-collapse experiment additionally stores every future step. The local search compares candidates and provides no all-action optimality certificate.

D.1 A HELD-OUT EXPLICIT-POLICY CRITICAL-WIDTH DIAGNOSTIC

The new experiment uses history index 0 of the saved age-2 checkpoint, with old batch 2048, $\theta = 0$, $\lambda = .1$, and the original explicit endpoint construction. It varies only future batch. The center and scale are fixed before validation trajectories are drawn.

Center and scale. FP64 bisection of the exact population 840-step terminal-risk equation gives

$$\widehat{\beta}_{\mathrm{exp}} \in [0.01091621322669089,\ 0.01091621322678402]. \tag{24}$$

This is a numerical sign bracket, not an outward-interval certificate. The entire root trace is included. A centered finite difference with step 10^{-7} gives terminal normalized-risk derivative -2.5816×10^{-4}. An independent 64-path calibration at future batch 1024 estimates a noise coefficient 1.3300×10^{-11}; their ratio sets $a = 5.1519 \times 10^{-8}$. No validation outcome adjusts either number.

Validation and dependence. For each $m_{\mathrm{f}} \in \{256, 1024, 4096\}$, 128 fresh random streams are reused across seven budgets $\widehat{\beta}_{\mathrm{exp}} + qa/\sqrt{m_{\mathrm{f}}}$. Thus the 2688 validation trajectories use 384 independent future streams, with paired observations across q; they are all conditional on one old checkpoint. The single-cell Wilson intervals in Figure 6 are pointwise, not simultaneous confidence bands. At $q = 0$, the observed survival counts are $59/128$, $71/128$, and $59/128$; their pointwise intervals all contain .5. The maximum between-batch spread is .09375, and the maximum deviation from the fixed normal reference is .07572. Both meet the preregistered descriptive tolerances .20 and .15; these tolerances are not hypothesis-test significance levels.

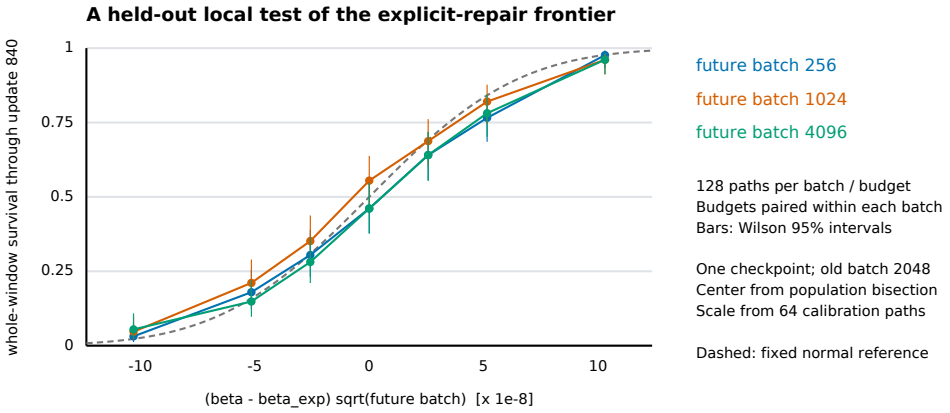

Figure 6: Conditional future-batch collapse for the explicit repair at one fixed checkpoint. The horizontal axis is $(\beta - \widehat{\beta}_{\mathrm{exp}})\sqrt{m_{\mathrm{f}}}$. Center and scale use only population computation and separate calibration paths. Error bars are pointwise Wilson 95% intervals from 128 streams per batch. This tests neither the complete-action root nor old-history fluctuations.

Numerical and data checks. All states, samples, covariance matrices, and means use FP64. The empirical sufficient-statistics update is algebraically the same simultaneous gradient as the direct sample update; their self-check differs by less than 7×10^{-21}. All 21 endpoints pass the full numerical movement, forgetting, risk, and cap audit. Across validation, no old-risk violation or cap projection occurs. Full arrays of future risk, old risk, and energy at every step $0, \ldots, 840$, plus per-trajectory first crossings and censoring, are included in compressed numerical files. The calibration's 64 paths are stored separately. Floating-point legality checks are numerical evidence, not exact global certificates.

D.2 A SMALL NONLINEAR HIDDEN-LAYER CONTROL

We also test a different learner, $f_{V,W}(x) = V \tanh(Wx)$, with input, hidden, and output dimensions 12. There are eight retained output rows and four added outputs. Inputs are uniform on the radius-$\sqrt{12}$ sphere. The future teacher has diagonal factors $\sqrt{2}$ on the first eight coordinates and 1 on

the last four; the old teacher zeros its last four output rows. The experiment uses one separate calibration history and four validation histories. Old SGD has step $.005$, ridge $.05$, batch 128, and ages $0, 500, 2000$ updates. Initialization has diagonal $\sqrt{2}$ on the old block and $\sqrt{.05}$ on the missing block, plus independent Gaussian perturbations of standard deviation $.001$.

A fixed one-shot editor performs 400 projected Adam steps on 1024 editing examples. It minimizes future half-square loss, with retained-row movement at most $\beta\sqrt{8}$, $\beta \in \{0, .05, .15\}$. Missing-row and encoder movement are at most $1.3\sqrt{4}$ and $6\sqrt{4}$, with energy cap 48. Incremental old risk and subsequent old risk are bounded by $.6$ and $.51$ per added output. All risk checks here use a disjoint fixed 2048-example evaluation sample; they are not population certificates.

Locked calibration and validation. The calibration history's age-500, $\beta = .05$ endpoint has initial future risk $.00011855$. After 200 deterministic continuation steps with ridge $.05$ on a separate 4096-example gradient sample, its evaluation risk is $.00358732$; this becomes the fixed future threshold before validation. Thus a crossing near 200 in the calibration setting is by design. Validation crosses four old histories, three ages, three budgets, future ridge $.05$ or 0, and eight future streams per history: 576 continuations, with 32 independent future streams. Minibatches are paired across all settings within a history/replicate. Both risks are evaluated at updates $0, 10, \ldots, 500$.

Table 6: Nonlinear control: median first unsafe *observation*, from 32 paths per cell. Columns β use future ridge $.05$; with zero future ridge every setting remains safe at all observation times. > 500 denotes censoring. The row coordinate is the median retained-row norm divided by $\sqrt{8}$.

old updates	row coordinate	$\beta = 0$	$\beta = .05$	$\beta = .15$	ridge 0
0	1.4142	210	210	210	> 500
500	1.3399	160	210	210	> 500
2000	1.3703	190	210	210	> 500

Mixed evidence and limits. With no retained-row movement, the observed crossing changes from 210 to 160 to 190 as old age increases. The retained-row coordinate likewise falls and partially recovers. Budgets $.05$ and $.15$ remove this observed deficit, but older age is not monotonically worse in this control. All 36 edited endpoints satisfy the numerical budgets and are initially safe; there is no cap projection, and the largest sampled old risk is $.20594$. The manual simultaneous gradient agrees with autograd to 9×10^{-16}. Saved checkpoints, editing/evaluation samples, endpoint audits, and all 51 observations per path are supplied. A first unsafe observation at j does not identify the first violating update between observations. This is evidence of budget and ridge sensitivity in a nonlinear hidden-layer model, with a calibrated threshold and a specified editor; it establishes neither all-action optimality, population whole-window survival, a universal age effect, nor metric invariance.

E A CONDITIONAL INTEGER-LAW IMPLICATION

This appendix isolates what a local Gaussian approximation would imply. It does not establish that approximation for the rank-40 complete action class, and it is not used by the finite-batch separation or joint robustness theorem. In particular, no uniform full-fiber LICQ, multiplier bound, variance floor, or numerical oracle-centered band is asserted here.

Proposition 9 (Conditional two-integer limit)**.** *Fix $1 \le N_0 < N_{\max}$, a reference budget β_0, and $\beta_m = \beta_0 + b/\sqrt{m}$. Suppose that scalar checkpoint fluctuations obey $Z_m \Rightarrow Z \sim \mathcal{N}(0, v)$, $v > 0$, and that the* optimized *capability satisfies*

$$V_{N_0,m}(A_m, \beta_m) = H(b - Z_m) + o_P(1), \qquad H(x) = \begin{cases} \Phi(x/c_+), & x \le 0, \\ \Phi(x/c_-), & x \ge 0, \end{cases} \tag{25}$$

for constants $c_+, c_- > 0$. Also assume $V_{N_0-1,m}(A_m, \beta_m) \to 1$ and $V_{N_0+1,m}(A_m, \beta_m) \to 0$ in probability. Then, for each fixed $p \in (0, 1)$,

$$N_p^\star(A_m, \beta_m) \Rightarrow N_0 - 1 + \text{Bernoulli}\left(\Phi\left(\frac{b - H^{-1}(p)}{\sqrt{v}}\right)\right). \tag{26}$$

Here $H^{-1}(p) = c_+\Phi^{-1}(p)$ for $p < 1/2$, 0 for $p = 1/2$, and $c_-\Phi^{-1}(p)$ for $p > 1/2$.

Proof. Nested survival events and the two neighboring limits place $N_p^\star$ in $\{N_0 - 1, N_0\}$ with probability tending to one. Since H is continuous and strictly increasing and Z has a continuous distribution, $\Pr\{H(b - Z) = p\} = 0$. The assumed optimized approximation therefore gives $\Pr\{V_{N_0,m} \geq p\} \to \Pr\{Z \leq b - H^{-1}(p)\}$, which is the displayed Gaussian probability. The vanishing mass outside the two integers proves convergence in distribution. □

An endpointwise CLT or a successful finite arithmetic run does not establish equation 25: it requires control of the full action supremum and the whole future path under changing checkpoints. We have not computed the relevant center or factor fiber, nor verified these assumptions for the saved rank-40 checkpoint. The explicit-policy scaling experiment has its own calibrated center and cannot substitute for this missing input.

F CURRENT-PARAMETER PROOF CHAIN

This appendix derives the exact updates, old-history event, complete-action gate, stochastic tube, global lower and feasible upper bounds, same-budget young branch, horizon bridge, and joint-parameter extension. The executable programs check finite scalar implications of these analytic reductions. They do not machine-prove universal factor or probability statements.

F.1 EXACT STOCHASTIC AND POPULATION UPDATES

For a sample (x, y), write

$$h = Wx + \delta e_{\mathrm{bias}}, \qquad e = Vh - y. \tag{27}$$

The simultaneous, pre-update gradients for a batch $\{(x_i, y_i)\}_{i=1}^m$ are

$$\begin{aligned} G_V^{(m)}(V, W) &= \frac{1}{m} \sum_{i=1}^m e_i h_i^\top + \lambda V, \\ G_W^{(m)}(V, W) &= \frac{1}{m} \sum_{i=1}^m V^\top e_i x_i^\top + \lambda W. \end{aligned} \tag{28}$$

Let P_K denote radial projection of the pair (V, W) onto $\|V\|_F^2 + \|W\|_F^2 \leq K$. Define the random preprojection map, the actual projected update, and the projected population-gradient map by

$$\begin{aligned} \widetilde{\Psi}_{B,m}(V, W) &= (V, W) - \eta(G_V^{(m)}, G_W^{(m)}), \\ \Psi_{B,m}(V, W) &= \mathsf{P}_K\big(\widetilde{\Psi}_{B,m}(V, W)\big), \\ \Phi_B(V, W) &= \mathsf{P}_K\Big((V, W) - \eta\, \mathbb{E}(G_V^{(m)}, G_W^{(m)})\Big). \end{aligned} \tag{29}$$

Thus Φ_B is not the expectation of a projected stochastic update. On every certified mean tube its preprojection point lies strictly inside the cap, so the final projection is inactive there. In stochastic bounds $\Psi_{B,m}$ retains the projection and uses its nonexpansiveness; inactivity is never assumed before it is proved.

Let $\Psi_{N,m}(a)$ be the probability that the actual trajectory started from endpoint a obeys both inequalities in equation 6 for $0, \ldots, N$. Conditional private randomization only averages these endpoint probabilities. Consequently

$$V_{N,m}(A, \beta) = \sup_{a \in \mathcal{A}_\beta(A,\theta)} \Psi_{N,m}(a), \tag{30}$$

and randomized policies cannot exceed the pointwise supremum.

F.2 OLD-HISTORY EVENT: DEFINITION, PROBABILITY, AND EVERY ENTRANCE BOUND

All norms of factor pairs below are Euclidean Frobenius norms. Let $\bar{z}_j = (D_j, D_j)$ be the exact zero-bias population recursion from the center of equation 4; let ϕ_j^u be the biased population recursion from

the actual initializer u. These are deterministic references, not expectations of projected stochastic updates. Define the single event

$$\mathcal{E}_H = \left\{ \max_{0\le j\le T} \|z_j - \phi_j^u\| < .002 \right\}, \qquad H = \eta T \in [1.95, 2]. \tag{31}$$

It is independent of the subsequently declared task and budget. Its definition does not require the editing policy to observe either reference.

Lemma 10 (Common old-history entrance)**.** *For every initializer in the full ball and every declared integer old age,* $\Pr(\mathcal{E}_H \mid u) \ge 1 - 33000\eta d^2/m$. *On this event,* $\|z_T - \bar{z}_T\| < R = .0033$. *With* $A = V_{T,o}$, $Z = V_{T,m}$, $G = AA^\top$, *and* $s = \|A\|_F/\sqrt{40}$, *all of the following hold:*

$$\|V_T\|_F^2 + \|W_T\|_F^2 \le 192, \qquad \|Z\|_F < .23 < \sqrt{k}/4, \tag{32}$$

$$\sqrt{1.90003} - R/\sqrt{40} \le s \le \sqrt{1.900075} + R/\sqrt{40} < 1.379, \tag{33}$$

$$\bar{G} = \bar{g} I_r + \bar{h} J_0, \quad \bar{g} \in [1.90003, 1.900075], \quad \bar{h} \in [.0099775, .0100225], \tag{34}$$

$$\|G - \bar{G}\|_F \le (2 \cdot 1.383 + R)R = .00913869, \tag{35}$$

$$\lambda_{\min}(G) \ge (\sqrt{1.89003} - R)^2, \quad \lambda_{\max}(G) \le (\sqrt{1.910075} + R)^2. \tag{36}$$

These inequalities, together with $\sin^2\theta \le 1/64$, *are the entire old-history input to the complete-action crossing proof. They imply* $(s+\beta)^2 \in [1.9243, 1.9345]$ *for* $\beta \in [.0093, .0119]$.

Proof. Exact discrete reference. For $b \in \{1.99, 2.01\}$ the corresponding diagonal product satisfies

$$g_{j+1} = g_j[1 - \eta(g_j - b + .1)]^2, \quad g_0 = b. \tag{37}$$

Writing $u_j = g_j - (b - .1)$ gives $u_{j+1} = u_j(1 - 2\eta g_j + \eta^2 g_j u_j)$. Induction gives $b - .1 \le g_j \le b$. The positive factors are at most $e^{-3.7\eta}$ and at least $e^{-4.03\eta}$: for the latter use $-\log(1-x) \le x/(1-x)$ with $x = 4.02\eta$. Consequently

$$.1e^{-4.03\eta j} \le u_j \le .1e^{-3.7\eta j}. \tag{38}$$

At $H \in [1.95, 2]$, both residual products lie in $[.00003, .000075]$. Thus $g_h \in [1.91003, 1.910075]$ and $g_l \in [1.89003, 1.890075]$, which give equation 34. The missing product obeys $p_{j+1} = p_j[1 - \eta(p_j + .1)]^2$, $p_0 = .05/16$; hence $0 < p_j \le p_0$. Throughout the history, the reference factor operator norms are below $a = 1.418$, its energy is at most 160.1, and its residual operator norm is at most $.1$.

Full-pair mean comparison. The zero-bias population Hessian in a direction (U, Z) has quadratic form

$$\|UW + VZ\|_F^2 + 2\langle VW - B_o, UZ\rangle + .1(\|U\|_F^2 + \|Z\|_F^2). \tag{39}$$

At the reference it is nonnegative, since the residual operator norm is at most $.1$. Within pair radius $.005$, differentiating the cubic gradient bounds the change of Hessian by $24(.005) = .12$ and its norm by 5. The bias adds a positive semidefinite Hessian of norm at most 10^{-6}. The squared Lipschitz factor of one mean step is therefore at most $1 + .24\eta + 25\eta^2 \le 1 + .241\eta$, and its norm factor at most $1 + .121\eta$. Comparing the biased and unbiased maps and summing yields

$$\max_{j\le T} \|\phi_j^u - \bar{z}_j\| \le e^{\cdot 121H}(.001 + Ha\delta^2) < .0013. \tag{40}$$

The right side closes the first-exit bootstrap. The reference energy and this radius place every mean candidate strictly inside the cap.

Shared-batch noise and stopping. Put $M = 1.423$. For a single sample residual $e = (VW - B_o)x + \delta V e_1 - \zeta$, the two data gradients together satisfy

$$\|g_{\text{data}}\|^2 \le [(M\sqrt{d} + \delta)^2 + dM^2]\|e\|^2, \quad \mathbb{E}\|e\|^2 = \|VW - B_o\|_F^2 + \delta^2\|Ve_1\|^2 + d\nu^2/3. \tag{41}$$

Here $\|VW - B_o\|_F \le \sqrt{40(.1)^2 + .05^2/16} + \sqrt{2}a(.005) + (.005)^2/2$. Substitution gives $\mathbb{E}\|\widehat{g} - \mathbb{E}\widehat{g}\|^2 \le .05d^2/m$. This is one joint variance bound; no independence of encoder and decoder gradients is assumed.

Stop the actual error relative to ϕ_j^u at its first radius-$.002$ crossing, including the crossing update, and freeze its squared norm thereafter as Y_j. Before crossing, both states lie in the reference $.005$ ball.

Projection is nonexpansive and fixes the next mean state, and centered noise has zero conditional cross term. Thus

$$\mathbb{E}Y_{j+1} \le (1 + .241\eta)\mathbb{E}Y_j + .05\eta^2 d^2/m, \quad Y_0 = 0. \tag{42}$$

On exit, $Y_T \ge (.002)^2$. A geometric sum and Markov's inequality give

$$\Pr(\mathcal{E}_H^c \mid u) \le \frac{.05(e^{.241H} - 1)}{.241(.002)^2} \frac{\eta d^2}{m} < 33000 \frac{\eta d^2}{m}. \tag{43}$$

This pays all $T \le 3{,}360{,}000$ updates at once.

Entrance consequences. Combining the two radii gives $R = .0033$. The norm and singular-value perturbation inequalities give equation 33 and equation 36. The identity $AA^\top - \bar{A}\bar{A}^\top = (A - \bar{A})\bar{A}^\top + \bar{A}(A - \bar{A})^\top + (A - \bar{A})(A - \bar{A})^\top$, with $\|\bar{A}\|_{\mathrm{op}} < 1.383$, gives equation 35. Finally $\|Z\|_F \le \sqrt{.05} + R < .23$, and the actual cap holds by construction. The rational interval program checks every displayed decimal implication, including the full q interval. No previous-version assertion is needed. □

For clarity, the anisotropy allowance used in the upper construction is also a consequence of these explicit inequalities. Put $g = s^2$, $c_z = (2 - \zeta\ell)/g$, and let T be the matrix in equation 70. Centering a Gram error by subtracting its trace is an orthogonal projection in Frobenius norm. Therefore

$$\frac{\|T\|_F}{\sqrt{r}} \le .0025 + |.01 - c_z\bar{h}| + c_z \frac{.00913869}{\sqrt{40}}, \quad \frac{\|T\|_F^2}{r\lambda_{\min}(Q_o)} < 1.115172 \times 10^{-5} < \Omega = 1/80000. \tag{44}$$

The first term uses $\|J_\theta - J_0\|_F/\sqrt{r} = 2|\sin\theta| \le 1/4$; $\zeta\ell < .002$ and the resource interval enclose c_z. This bounds every cross block and physical-column perturbation through the full Gram error.

F.3 Current-parameter gate for the complete action class

Here and below $B = B_\theta$ has eigenvalues $2.01, 1.99, 1$, with multiplicities $20, 20, 16$. Thus

$$\operatorname{tr} B = 96, \quad \|B\|_F^2 = 176.004, \quad K = 2\operatorname{tr} B = 192. \tag{45}$$

The equality for K is structural: it is the least factor energy needed to represent the noiseless teacher exactly. There is no representation-energy reserve at exact fit. The bounds below keep the exact spectrum, the original risk normalization, and the shared-minibatch learner.

Lemma 11 (Elimination of the other checkpoint coordinates)**.** *Let the observed checkpoint obey the cap and $\|Z_T\|_F \le \sqrt{k}/4$. For every edited pair obeying the cap and $R_{\mathrm{f}}(a) \le \rho k$, encoder movement, missing-row movement, incremental forgetting, and old safety hold with strict slack. Consequently the complete initially safe action class depends on the checkpoint only through its retained rows A_T and the closed movement ball about them.*

Proof. Set $P = VW = HO$ (left polar decomposition), $D = V - OW^\top$, and $F = (V + OW^\top)/2$. Initial future safety gives $\|P - B\|_F \le e = .002\sqrt{k}$, so P is invertible. The cap and nuclear-norm duality give

$$g := \|D\|_F^2 = \|V\|_F^2 + \|W\|_F^2 - 2\|P\|_* \le 2\sqrt{d}\, e, \qquad FF^\top = H + DD^\top/4. \tag{46}$$

Write $\gamma = .004\sqrt{3.5} < .007484$. Then

$$\max(\|V\|_F, \|W\|_F)/\sqrt{k} \le \sqrt{6} + \tfrac{1}{2}\sqrt{\gamma} < 2.493. \tag{47}$$

For the missing-row selector J, Cauchy–Schwarz for each diagonal entry of $H = (PP^\top)^{1/2}$ gives $\operatorname{tr}(JHJ^\top) \le \sum_{i\in J} \|P^\top e_i\| \le \sqrt{k}\|JP\|_F \le k + \sqrt{k}e$. Hence

$$\|JV\|_F/\sqrt{k} \le \sqrt{1.002 + \gamma/4} + \tfrac{1}{2}\sqrt{\gamma} < 1.046. \tag{48}$$

The movement bounds are therefore $2.493 + \sqrt{12} < 5.958 < 6$ for W, and strictly below $1.046 + .25 = 1.296 < 1.3$ for the missing rows, in units of $\sqrt{k}$. Apply the triangle inequality to the augmented residual $[P - B, \delta V e_1]$. Since $\|B - B_o\|_F^2/k = 1 + .001\sin^2\theta \le 1.001$,

$$R_{\mathrm{o}}(a)/k \le \tfrac{1}{2}(\sqrt{1.001} + .002)^2 < .503 < .51. \tag{49}$$

Nonnegativity of checkpoint risk makes the incremental forgetting at most $.503k < .6k$. This argument includes every cross block and frame; it does not optimize within a chosen repair template. □

All following future-path estimates quantify over *initially safe* legal endpoints, or their explicitly declared compact closure. An endpoint unsafe at update zero has survival probability zero and is removed before applying any tube or normal approximation. The old-history argument supplies the checkpoint gate in Lemma 11; this lemma itself is deterministic.

F.4 Uniform future tube and full-window conversion

This paragraph supplies the finite-step calculation used in the sequel. Write $G = \nabla(R_f + \lambda\|a\|^2/2)$, $E = VW - B$, $c = Ve_1$, and $\tau = N\eta \le .001$. On the cap with $\|E\|_F \le .003\sqrt{k}$, polar balance as above gives factor operator norms below 3. Differentiating the displayed gradients in equation 28, with the exact spectrum equation 45, yields the following conservative bounds:

$$\begin{gathered} .04I \preceq DG \preceq 19I, \quad \langle G, a\rangle/k > 1.17, \quad \|G\|_F < .624\sqrt{k}, \\ \|(VW)^+ - VW\|_F/\sqrt{k} < .86\eta. \end{gathered} \tag{50}$$

The product increment here is exactly

$$-\eta(EW^\top W + VV^\top E + 2\lambda VW + \delta^2 cw^\top) + \eta^2 G_V G_W, \quad w = W^\top e_1; \tag{51}$$

no continuous-time substitution is used. Radial energy change is $-2\eta\langle G, a\rangle + \eta^2\|G\|^2 < 0$, proving that the mean projection is inactive. Induction from $\|E_0\|/\sqrt{k} \le .002$ gives $\|E_j\|/\sqrt{k} \le \bar{q} = .00286$. Segments within pair radius 10^{-4} stay inside the .003 layer; $I - \eta DG$ is nonexpansive there. The augmented residual bound gives old risk below $.504k$ on this tube.

For completeness, mean risk need not increase at every endpoint. Its discrete convexity is sufficient for a whole-window bound. Let $F = G_V W + VG_W$ and $T = G_V G_W$, so $\Delta P = -\eta F + \eta^2 T$. The same layer bounds give

$$\|DF\| < 82, \quad \|DT\| < 15, \quad \|\Delta P\| > .60\eta\sqrt{k}, \quad \|\Delta^2 P\| < 52\eta^2\sqrt{k}, \quad \|\Delta^2 c\| < 12\eta^2\sqrt{k}. \tag{52}$$

Expanding the two consecutive squared risks, including both bias terms, gives

$$\frac{\Delta^2 R_f}{\eta^2 k} > .60^2 - .003(52) - \frac{\delta^2(3)(12)}{\sqrt{k}} > .2. \tag{53}$$

Thus the mean whole-window maximum is attained at update 0 or N. The stronger increasing-risk statement used for horizon transversality additionally needs the retained-row gate; it is not assumed here.

For the stochastic calculation let $\xi(a, x)$ be the *joint* centered pair gradient. Sphere second and fourth moments applied before separating the layers give

$$\mathbb{E}\|m^{-1}\textstyle\sum_i \xi(a, x_i)\|^2 \le \frac{18(d-1)(.003)^2k + \delta^2[2(.003)^2k + 243d]}{m} < \frac{d^2}{18000m}. \tag{54}$$

Stop at the first distance $R \le 10^{-4}$ from the mean path, including the crossing update. Conditional centering, mean nonexpansiveness, and nonexpansiveness of the actual cap projection yield

$$\Pr\{\text{exit by } N\} \le p_R := \frac{N\eta^2 d^2}{18000mR^2}. \tag{55}$$

On its complement, $D_R = 3\sqrt{2}R + R^2/2$ bounds product displacement, and

$$|R_f(z_j)/k - R_f(\bar{z}_j)/k| \le e(R) := \frac{\bar{q}D_R}{\sqrt{k}} + \frac{D_R^2}{2k} + \frac{\delta^2(3R + R^2/2)}{k}. \tag{56}$$

Accordingly $\chi_N < -e(R)$ implies survival at least $1 - p_R$, whereas $\chi_N > e(R)$ implies survival at most p_R. These are pointwise probability inequalities uniform in the endpoint; they do not assert a common event for uncountably many edits. Taking the supremum is done only afterwards.

F.5 GLOBAL LOWER BOUND AND CONSTRUCTIVE UPPER BOUND

We prove Theorem 3. Put $P_j = V_j W_j$, $E_j = P_j - B_\theta$, and let A_j be the old r rows of V_j. On the common history event $\mathcal{E}_H$, the retained-row norm $s = \|A_T\|_F/\sqrt{r}$ and each admissible budget in Theorem 3 give

$$q = (s+\beta)^2 \in Q = [1.9243, 1.9345]. \tag{57}$$

Lemma 10 gives the full event definition, its probability, and all resource and Gram bounds.

Lemma 12 (All-action terminal lower bound)**.** *For every initially safe legal endpoint (and its declared compact closure) and every $q \in Q$,*

$$\frac{R_{\mathrm{f}}(\Phi_{B_\theta}^{840}(a))}{k} \geq f_-(q) := \frac{\Delta(q)^2}{3.2(\ell^2+.4)}, \qquad \ell = \frac{2}{q}, \tag{58}$$

where, with $b = .00286$, $\tau = 840\eta = 1/2000$, and $h = \ell^2 - 1$,

$$\mathfrak{d}(q) = .1[4.8 - .8\sqrt{3.5\,b}] - (2+h).4b(3.317+b), \tag{59}$$

$$\Delta(q) = 4(\ell-1) - hq + 2\tau\mathfrak{d}(q) - \eta\tau(1+h)(.624)^2(.4). \tag{60}$$

Proof. Let $F = \mathrm{diag}(\ell I_r, I_k)$ and define

$$\mathcal{W}_j = \|V_j\|_F^2 + \|W_j\|_F^2 + h\|A_j\|_F^2. \tag{61}$$

The exact square completion

$$\mathcal{W}_j - 2\langle F, V_j W_j\rangle = \|W_j - V_j^\top F\|_F^2 \geq 0 \tag{62}$$

holds for arbitrary factors, including every cross block and imbalance. Expanding one simultaneous population step, including the η^2 term, gives

$$\begin{aligned}\mathcal{W}_{j+1} - \mathcal{W}_j = &-2\eta\{\lambda\mathcal{W}_j + 2\langle E_j, P_j\rangle + h\langle E_{o,j}, P_{o,j}\rangle + \delta^2(\|V_j e_{\mathrm{bias}}\|^2 + h\|A_j e_{\mathrm{bias}}\|^2)\} \\ &+ \eta^2(\|G_{V,j}\|_F^2 + \|G_{W,j}\|_F^2 + h\|G_{A,j}\|_F^2).\end{aligned} \tag{63}$$

The certified tube gives $\|E_j\|_F \leq b\sqrt{k}$ and $\|(G_{V,j}, G_{W,j})\|_F \leq .624\sqrt{k}$. The bias term is nonnegative. Nuclear-norm duality and equation 62 therefore imply, after summing all 840 exact steps,

$$2\langle F, B_\theta - P_{840}\rangle \geq r\Delta(q). \tag{64}$$

The initial movement constraint is used only as $\|A_0^a\|_F \leq \sqrt{rq}$. Finally Cauchy–Schwarz with $\|F\|_F^2 = r(\ell^2+.4)$ gives equation 58. Equivalently, optimizing $2t\langle F, B_\theta - P_{840}\rangle - t^2\|F\|_F^2$ over $t \geq 0$ yields the same bound. Thus the result covers the full action class rather than a selected factor template. □

Lemma 13 (Measurable hard-legal upper construction)**.** *For every checkpoint on $\mathcal{E}_H$ and $q \in Q$, there is a Borel endpoint $a^\star = (V^\star, W^\star)$ satisfying every hard budget and*

$$J_{840}(A_T, \beta) \leq f_+(q) - \rho. \tag{65}$$

Moreover, over Q,

$$0 \leq J_{840} - (f_-(q) - \rho) \leq f_+(q) - f_-(q) < 2 \times 10^{-7}. \tag{66}$$

Proof. Let $G_0 = A_T A_T^\top$, $\alpha = 1 + \beta/s$, $\ell = 2/q$, and $\Omega = 1/80000$. Define

$$A_q = \ell^2/q, \qquad B_q = 4\ell/q + .8, \qquad C_q = q + 4/q - 4 + \Omega, \tag{67}$$

$$\zeta(q) = \frac{2C_q}{B_q + \sqrt{B_q^2 - 4A_qC_q}}. \tag{68}$$

Choose, by the first-independent-column Gram–Schmidt rule, a Borel matrix N_A with $N_A N_A^\top = I_k$ and $N_A A_T^\top = 0$. Set

$$V^\star = \begin{bmatrix} \alpha A_T \\ \sqrt{1-\zeta}\, N_A \end{bmatrix}, \quad P^\star = B_\theta - \zeta\,\mathrm{diag}(\ell I_r, I_k), \quad W^\star = (V^\star)^{-1}P^\star. \tag{69}$$

Then the old-row movement is exactly $\beta\sqrt{r}$. Writing $Q_o = \alpha^2 G_0 = qI + D$ and $T = .01J_\theta - (2 - \zeta\ell)D/q$, trace cyclicity gives the noncommutative identity

$$\frac{1}{r}\operatorname{tr}[(2I + .01J_\theta - \zeta\ell I)^2 Q_o^{-1}] = \frac{(2-\zeta\ell)^2}{q} + \frac{1}{r}\operatorname{tr}(T^2 Q_o^{-1}). \tag{70}$$

Equation 44 bounds the last term by Ω. The defining quadratic for ζ then gives

$$\|V^\star\|_F^2 + \|W^\star\|_F^2 \le r\{q + (2-\zeta\ell)^2/q + .8(1-\zeta) + \Omega\} = K. \tag{71}$$

The same bounds give edit-time future risk below $1.965332234 \times 10^{-6}k$, encoder movement below $5.958\sqrt{k}$, missing-row movement below $1.296\sqrt{k}$, $R_{\mathrm{o}}/k < .503$, and incremental forgetting below $.6k$. Along its exact mean path, the product update obeys

$$P_{j+1} = P_j - \eta(E_j W_j^\top W_j + V_j V_j^\top E_j + 2\lambda P_j + \delta^2 c_j w_j^\top) + \eta^2 G_{V,j} G_{W,j}. \tag{72}$$

With factor operator norms below 1.5, put $L_0 = 4.7$, $S = \exp(L_0\tau)$, and

$$a_0 = \tfrac{1}{5}\sqrt{11.00025} + \tfrac{9}{16}10^{-6} + .065\eta, \quad q_N^+ = S\zeta\sqrt{2.5\ell^2 + 1} + \frac{a_0}{L_0}(S-1). \tag{73}$$

The finite-step recurrence $q_{j+1} \le (1 + \eta L_0)q_j + \eta a_0$ therefore gives the explicit upper function

$$f_+(q) = \tfrac{1}{2}(q_N^+)^2 + \tfrac{9}{128}10^{-6}. \tag{74}$$

The factor $1/2$ agrees with equation 5. The closed-box certificate checks equation 66 on 2048 rational intervals; a separate 512-cell mean-value enclosure gives 1.676×10^{-7}. These computations certify scalar consequences of the full-factor and history reductions proved here. □

Lemma 14 (Uniform stochastic converter)**.** *In the finite box $m \in \{1867, \dots, 2333\}$, the stopped future path satisfies the following implications, uniformly over initially safe legal endpoints:*

$$J_{840} > 5 \times 10^{-8} \Longrightarrow V_{840,m} \le .001, \qquad \exists a : \chi_{840}(a) < -5 \times 10^{-8} \Longrightarrow V_{840,m} \ge .999. \tag{75}$$

Proof. The stopped envelope in Appendix F.4, including the crossing increment and cap projection, bounds pair-tube exit by

$$\Pr(\text{exit}) \le \frac{N\eta\,\Xi}{18000R^2}. \tag{76}$$

Within radius R, put $D_R = 3\sqrt{2}R + R^2/2$. The half-square risk convention gives a whole-path normalized-risk allowance

$$e(R) = \frac{.00286 D_R}{4} + \frac{D_R^2}{32} + \frac{\delta^2(3R + R^2/2)}{16}. \tag{77}$$

At $N = 840$, $\Xi \le 10^{-6}$, and $R = 1/180000$, these formulas give $\Pr(\text{exit}) \le .0009$ and $e(R) < 2 \times 10^{-8}$, strictly below both stated margins. Old risk remains slack on the tube. The negative implication is pointwise for every endpoint before taking the supremum; the positive implication uses the measurable construction. No single future event over an uncountable action set is asserted. For the longer 1680-step bridge, $R = 1/120000$ gives failure at most .0008 and $e(R) < 3 \times 10^{-8}$, justifying the buffer in Appendix B. □

Proof of Theorem 3. On $\mathcal{E}_H$, the rational enclosures give

$$\beta \le .0098 \Longrightarrow f_-(q) - \rho > 5.15918 \times 10^{-8}, \tag{78}$$

$$\beta \ge .0118 \Longrightarrow f_+(q) - \rho < -5.03044 \times 10^{-8}. \tag{79}$$

Lemmas 12–14 give the conditional $.001/.999$ separation. Mobility nesting makes the capability gap nonnegative off $\mathcal{E}_H$. Integrating its .998 conditional gap over $\Pr(\mathcal{E}_H) \ge .967$ and the task mass $(2/\pi)\arcsin(1/8) > 7/88$ gives a population gap greater than .0767. □

F.6 A SAME-BUDGET COMPARISON WITH THE YOUNG CHECKPOINT

Proof of Proposition 4. At $T = 0$ choose $a_* = (B_\theta^{1/2}, B_\theta^{1/2})$ and retain the actual bias δe_1. Its energy is $2 \operatorname{tr} B_\theta = 192$. Put $d_s = \sqrt{2.01} - \sqrt{1.99} = .02/(\sqrt{2.01} + \sqrt{1.99})$. For every task angle, the retained-row movement from the initializer is at most $\sqrt{r} d_s + .001$, hence less than $.007252\sqrt{r}$. The missing-row movement is at most $4.001 < 5.2$, and encoder movement is below $\sqrt{40} d_s + 4.001 < 4.046 < 24$. These statements use the entire initializer ball, not its center. Also

$$R_{\mathrm{o}}(a_*) \le \tfrac{1}{2}(16.016 + 2.01\delta^2) < 8.009 < .6k. \tag{80}$$

Thus forgetting and both initial safety conditions hold, and a_* is hard legal for every $\beta \ge .0093$.

It remains to bound every future sample path. This argument does not import an old-history or future-noise event. In a pair ball of radius $R = .001$ about a_* use $s = 1.418 > \sqrt{2.01}$, $x_* = 7.484 > \sqrt{56}$, $M = s + R$, and $q_R = s\sqrt{2}R + R^2/2$. For every sphere sample, $\|h\| \le Mx_* + \delta$, $\|e\| \le q_R x_* + \delta M$. The operator norm of the sample-loss Hessian, including ridge, is at most

$$L = (Mx_* + \delta)^2 + M^2 d + (q_R x_* + \delta M)x_* + \lambda < 230. \tag{81}$$

Indeed, its Gauss–Newton part is bounded by the first two terms, and $2\langle e, \Delta V \Delta W x\rangle$ contributes at most $\|e\|\|x\|\|\Delta a\|^2$. At a_* the residual equals $\delta B_\theta^{1/2} e_1$, so the sample-gradient norm is at most

$$\lambda\sqrt{192} + \delta s\sqrt{(s\sqrt{56} + \delta)^2 + s^2 d} < 1.408. \tag{82}$$

Both bounds hold for any minibatch average. The cap projection fixes a_* and is nonexpansive. Consequently, while inside the ball,

$$d_{j+1} \le (1 + 230\eta) d_j + 1.408\eta, \qquad d_0 = 0. \tag{83}$$

For $j \le 840$, $j\eta \le 1/2000$. Summing the geometric series and using $(1 + h)^j \le (1 - jh)^{-1}$ gives

$$d_j \le \frac{1.408/2000}{1 - 230/2000} = \frac{88}{110625} < .0008 < R. \tag{84}$$

Induction closes the ball condition, including a projected update. With $q = s\sqrt{2}(.0008) + (.0008)^2/2$,

$$\frac{R_f(z_j)}{k} \le \frac{q^2 + \delta^2(s + .0008)^2}{32} < 1.45 \times 10^{-7}, \tag{85}$$

$$\frac{R_o(z_j)}{k} \le \frac{(4.002 + q)^2 + \delta^2(s + .0008)^2}{32} < .501. \tag{86}$$

Here $4.002^2 > 16.016$. Thus both barriers hold at every update for every sample path, proving young capability one at every future batch size.

At the old age, Theorem 3 bounds capability by .001 on $\mathcal{E}_H$ and the angle event. Young capability equals one for *all* angles, so its difference from old capability is nonnegative elsewhere. Integration at the *same numerical budget* gives

$$\mathbb{E}[V_{840,m}(u, \theta, \beta) - V_{840,m}(z_T, \theta, \beta)] > .999 \cdot .967 \cdot \frac{7}{88} > .0768. \tag{87}$$

The lower-bound proof is angle-independent, so the unrestricted survival-capability gap also exceeds $.999 \cdot .967 > .9660$. That larger quantity includes angles with no certified immediate old repairability; the $> .0768$ bound isolates the guaranteed-repairable subset. This compares the declared endpoints in age and does not imply monotonicity at intermediate ages. The old branch retains the old-history hypotheses of Theorem 3; the young branch is independent of that proof chain. □

F.7 HORIZON TRANSVERSALITY WITHOUT DIFFERENTIATING THE OPTIMIZER

Proof of Theorem 8. Fix an initially safe legal endpoint a. Throughout $0 \le j \le 1680$, the tube gives factor operator norms below 3, normalized product error below .00286, old risk below $.504k$, and retained-row squared norm below $1.936r$. Exact Taylor expansion of one population step then gives

$$\frac{\eta}{4000} \le \frac{R_{\mathrm{f}}(\bar{z}_{j+1}^a) - R_{\mathrm{f}}(\bar{z}_j^a)}{k} \le \frac{\eta}{200}. \tag{88}$$

For the upper bound, for example,

$$\frac{|f_{j+1} - f_j|}{\eta} \le \frac{.049(1.435) + 10\eta(1.435)^2}{16} < 1/200. \tag{89}$$

The division by $k = 16$ is essential. To see the lower bound, set $\bar{q} = 1.936$, $\ell = 2/\bar{q}$, and $F = \mathrm{diag}(\ell I_r, I_k)$. The weighted square identity and the cap imply

$$\langle F, E_j \rangle \le -\frac{r(2-\bar{q})^2}{2\bar{q}}, \qquad \langle B_\theta, E_j \rangle < -.038. \tag{90}$$

The second inequality uses $\|B_\theta - 2F\|_F < 4.05$ and $\|E_j\|_F \le e_b = .01144$. Exact one-step Taylor expansion then gives

$$\frac{f_{j+1} - f_j}{\eta} \ge \frac{-.049^2 - .2(-.038 + e_b^2) - .9\delta^2 - 10\eta(1.435)^2}{16} > \frac{1}{4000}. \tag{91}$$

The required row bound follows by induction, as follows. Initially $\|A_0^a\|_F^2/r \le (1.379 + .0122)^2 < 1.936$. At every subsequent mean step the tube gives $\|P_{o,j}\|_F \ge \sqrt{160.004} - .01144 > 12.6$, so $\|A_j\|_F > 12.6/3 = 4.2$ and $\|P_{o,j}\|_F < 14$. Writing $D_{A,j} = E_{o,j}W_j^\top + \delta^2 A_j e_1 e_1^\top$, the exact update gives

$$\begin{aligned}\frac{\|A_{j+1}\|_F^2 - \|A_j\|_F^2}{\eta} &= -.2\|A_j\|_F^2 - 2\langle E_{o,j}, P_{o,j}\rangle - 2\delta^2\|A_j e_1\|^2 + \eta\|.1A_j + D_{A,j}\|_F^2 \\ &\le -.2(4.2)^2 + 2(.01144)(14) + \eta\big(.1\sqrt{192} + 3(.01144) + 3\delta^2\big)^2 < -3.\end{aligned} \tag{92}$$

The last square retains the ridge–data cross term. This strict decrease, combined with the initial bound, proves the retained-row condition at every step within the already established future tube. Summing for each a and taking minima gives

$$\frac{\eta(n'-n)}{4000} \le J_{n'} - J_n \le \frac{\eta(n'-n)}{200}. \tag{93}$$

For the left inequality evaluate an n'-minimizer at n; for the right evaluate an n-minimizer at n'. Thus the argument never identifies the two optimizers. Finally, survival through n' is a subset of survival through n for every endpoint, so taking the supremum proves $V_{n',m} \le V_{n,m}$. □

F.8 An explicit joint neighborhood in the rank-40 learner

This proof changes the original rank-40 learner, including both training phases and the cap used by every action and projection. It does not transfer a scalar result. Write $\eta_0 = 1/1680000$, $\lambda_0 = .1$, and $\rho_0 = 2 \times 10^{-6}$. Throughout, let

$$\begin{aligned}|K - 192| \le 10^{-5}, \quad |\lambda - .1| \le 2 \times 10^{-7}, \\ |\eta/\eta_0 - 1| \le 10^{-6}, \quad |\rho - \rho_0| \le 10^{-8}.\end{aligned} \tag{94}$$

The closed box has a nonempty four-dimensional interior containing the nominal tuple. Both $K < 192$ and $K > 192$ are included. Set $\epsilon_K = 10^{-5}$, $\epsilon_\lambda = 2 \times 10^{-7}$, $\epsilon_\eta = 10^{-6}$, $\eta_\pm = \eta_0(1 \pm \epsilon_\eta)$, $\tau_\pm = 840\eta_\pm$, and $\lambda_\pm = .1 \pm \epsilon_\lambda$.

Analytic reduction 1: the old gate survives jointly. Fix an integer T with $\eta_0 T \in [1.95, 2]$, before changing the parameters. Keep the *nominal* zero-bias reference $\bar{z}_j$ from Appendix F.2. Its exact discrete products and Gram bounds are therefore unchanged. Let $\phi_j^{u,p}$ be the actual biased population recursion at the perturbed tuple $p = (K, \lambda, \eta, \rho)$. At $\bar{z}_j$, the nominal old gradient has norm below 1.3; its square is bounded by $.1^2(160) + .103125^2(.1) < 1.3^2$. On the same radius-.005 pair ball the perturbed Hessian has lower bound $-(.12 + \epsilon_\lambda)I$ and norm below 5. Consequently the norm and squared-norm step factors are still bounded by $1 + .121\eta_0$ and $1 + .241\eta_0$, respectively. Comparing one perturbed step to one nominal step introduces forcing at most $\eta_0 F_o$, where

$$F_o = 1.3\epsilon_\eta + (1 + \epsilon_\eta)\{\epsilon_\lambda\sqrt{160.1} + 1.418\delta^2\}. \tag{95}$$

The same geometric-sum bootstrap gives

$$\max_{j\le T}\|\phi_j^{u,p}-\bar z_j\|\le e^{.242}(.001+2F_o)<.0013. \tag{96}$$

The reference energy is at most 160.1, so all these mean candidates are inside even the smallest cap 191.99999.

Define $\mathcal{E}_H(p)=\{\max_{j\le T}\|z_j^p-\phi_j^{u,p}\|<.002\}$. The centered shared-batch variance on the radius-$.005$ layer is unchanged: ridge is deterministic and the feature and label laws are unchanged. It is at most $.05d^2/m$. The crossing-inclusive stopped recurrence now has noise term $.05\eta_+^2d^2/m$ and growth $1+.241\eta_0$. Thus

$$\Pr(\mathcal{E}_H(p)^c\mid u)\le(1+\epsilon_\eta)^2\frac{.05(e^{.482}-1)}{.241(.002)^2}\frac{\eta_0d^2}{m}<33000\frac{\eta_0d^2}{m}\le.033. \tag{97}$$

On this event $\|z_T^p-\bar z_T\|<.0033$, so every resource, Gram, singular-value, and anisotropy bound in Appendix F.2 is retained. The checkpoint energy bound becomes K; this is the only changed gate entry. The event is common to tasks and budgets at each fixed tuple. No event common to an uncountable family of parameter tuples is asserted.

Analytic reduction 2: a uniform 840-step future tube. For any initially safe endpoint at the new cap and threshold, $\|VW-B_\theta\|/\sqrt{k}\le\sqrt{2\rho}\le\sqrt{4.02\times10^{-6}}$. In the larger layer $\|E\|/\sqrt{k}\le.003$, the polar imbalance satisfies

$$\|V-OW^\top\|^2\le\epsilon_K+2\sqrt{d}\|E\|, \tag{98}$$

which still gives both factor operator norms below 3. The exact gradient formulas give

$$\begin{aligned}.04I\preceq DG\preceq19I,\quad \|G\|/\sqrt{k}<.624,\\ \langle G,a\rangle/k>1.17,\quad \|P^+-P\|/\sqrt{k}<.86\eta.\end{aligned} \tag{99}$$

For example the Hessian bounds follow from $\lambda_--.012>.04$ and $18+.012+\lambda_++\delta^2<19$. The gradient bound is $(.012+\lambda_++\delta^2)\sqrt{192+\epsilon_K}/4<.624$. Nuclear-norm duality gives the radial lower bound $\lambda_-(12-2\sqrt{3.5}(.003))-2(.003)(3.317+.003)>1.17$. Finally the exact product increment is bounded by $18(.003)+2\lambda_+(3.317+.003)+9\delta^2/4+2\eta_+(.624)^2<.86$. Radial energy decreases, so mean projection is inactive. Since $\sqrt{4.02\times10^{-6}}+.86\tau_+<.00286$, induction closes the same 840-step mean tube as before. Mean segments within pair radius 10^{-4} remain in the $.003$ layer. Old risk is below $.504k$ there.

The same sphere-moment calculation gives joint centered variance at most $d^2/(18000m)$. At radius $R=1/180000$, the stopped future error obeys

$$p_R\le\frac{840\eta_+^2d^2}{18000mR^2}<.001,\qquad e(R)<2\times10^{-8}, \tag{100}$$

uniformly for $m\ge1867$. Projection remains part of the stochastic update. These are endpointwise uniform bounds, applied before the action supremum.

Analytic reduction 3: the complete-action lower bound. The weighted-square proof uses the cap only to bound $\mathcal{W}_0\le K+hrq$. Therefore the perturbed terminal lower function is bounded below by

$$f_-^{\mathrm{box}}(q)=\frac{\Delta_{\mathrm{box}}(q)^2}{3.2(\ell^2+.4)},\qquad \ell=2/q,\quad h=\ell^2-1, \tag{101}$$

where, for $b=.00286$,

$$d_-(q)=\lambda_-[4.8-.8\sqrt{3.5}\,b]-(2+h).4b(3.317+b), \tag{102}$$

$$\Delta_{\mathrm{box}}(q)=4(\ell-1)-hq-\epsilon_K/40+2\tau_-d_-(q)-\eta_+\tau_+(1+h)(.624)^2(.4). \tag{103}$$

The drift is positive throughout Q. The choice of τ_- is therefore valid; the quadratic-step subtraction uses $\eta_+\tau_+$. A 2048-cell outward interval cover of Q proves positivity of Δ_{box} and strict decrease of f_-^{box}. At $q_{\max}=(\sqrt{1.900075}+.0033/\sqrt{40}+.0098)^2$ it gives

$$f_-^{\mathrm{box}}(q_{\max})-(\rho_0+10^{-8})>4\times10^{-8}. \tag{104}$$

This covers all legal factors, including initially unsafe endpoints whose survival is already zero. It does not use the angle restriction: the weighted trace $\langle F,B_\theta\rangle$ and teacher norm are angle-independent.

Analytic reduction 4: feasible upper and young witnesses. Fix $\sin^2\theta \le 1/64$. At a checkpoint on $\mathcal{E}_H(p)$, construct the nominal upper endpoint $a^\star$ by equation 69, using its actual retained rows. Set

$$t_K = \min\{1, \sqrt{K/192}\}, \qquad a^{\star,p} = t_K a^\star. \tag{105}$$

It obeys the new cap. If $V_o^\star = \alpha A_T$, then $1 < t_K\alpha \le \alpha$ for $\beta \ge .0093$, so this scaling does not increase retained-row movement. The other movement budgets have the strict slack established in Lemma 13. The scaling displacement is at most

$$D_0 = \frac{\epsilon_K}{\sqrt{192} + \sqrt{192 - \epsilon_K}} < 3.62 \times 10^{-7}. \tag{106}$$

Compare its perturbed mean path to the nominal path of $a^\star$. Both have inactive mean projections on the preceding tube. A first-exit bootstrap, the Hessian bound 19, and the parameter forcing give

$$\max_{j\le 840} \|\bar z_j^{\star,p} - \bar z_j^\star\| \le e^{19\tau_+}\left[D_0 + \frac{4(.624)\epsilon_\eta + (1+\epsilon_\eta)\epsilon_\lambda\sqrt{192}}{2000}\right] < 4\times 10^{-7}. \tag{107}$$

On connecting segments normalized future risk has gradient norm below $3\sqrt{2}(.003)/4 + 3\delta^2/16 < .004$, so its change is below 2×10^{-9} at every update. Initial risk is below $1.965332234\times 10^{-6} + 2\times 10^{-9} < \rho_0 - 10^{-8}$. This also certifies edit-time repairability at the low budgets, on the angle event. For $\beta \in [.0118, .0119]$, the nominal upper bound and initial-risk bound consequently imply

$$\max_{j\le 840} R_{\mathrm{f}}(\bar z_j^{\star,p})/k < \rho - 2\times 10^{-8}. \tag{108}$$

In particular the terminal slack exceeds 3.8×10^{-8} at $\beta = .0118$; the whole-window slack also pays the edit-time bound. This proves the high branch with equation 100. All hard budgets remain satisfied, and old safety stays strict. The construction is Borel, including its cap scaling.

For the young checkpoint use $t_K(B_\theta^{1/2}, B_\theta^{1/2})$. The same cap scaling adds at most D_0 to each movement estimate. The samplewise Hessian bound remains 230, and the gradient at this center is at most $1.408 + \epsilon_\lambda\sqrt{192} + 230D_0$. Projection fixes the scaled center. Its entire sample path is therefore within

$$\frac{\tau_+(1.408 + \epsilon_\lambda\sqrt{192} + 230D_0)}{1 - 230\tau_+} + D_0 < .00081 \tag{109}$$

of the unscaled teacher factors. Normalized future risk is below 1.5×10^{-7}, and old risk below .501, for every future sample path. This proves Proposition 6.

Arithmetic evidence and limits. `joint_robustness.py` checks every decimal comparison in these four reductions and saves all 2048 lower-envelope cells. It does not prove the matrix inequalities, the stopped martingale reduction, or the action-class quantifiers. The box is deliberately small: it establishes local openness across the exact-representation cap, not useful tolerance to large cap changes, vanishing ridge, alternative action metrics, or an arbitrary training schedule.

F.9 Checkpoint sufficiency and orthogonal covariance

Proposition 15 (State sufficiency and an exact symmetry)**.** *Fix the task, action contract, and future learner. Two histories ending at the same full pair (V, W) have identical capability at every budget and horizon. Furthermore, for an orthogonal Q with $Qe_1 = e_1$, the simultaneous change*

$$\mathcal{T}_Q(V, W) = (VQ^\top, QW) \tag{110}$$

of checkpoint and action coordinates preserves capability, including every hard budget and the entire future-path law.

Proof. The observed pair fixes the action set and the transition kernel. Coupling the same endpoint and fresh minibatches after identical pairs proves the first claim. No unobserved optimizer state is present in this SGD model.

For the second claim, $Q^\top e_1 = e_1$ gives $VQ^\top(QWx + \delta e_1) = V(Wx + \delta e_1)$. Frobenius norms, named decoder-row movements, encoder movement, and both risks are invariant. This defines a

bijection of the complete legal action sets. Direct substitution into equation 28 gives transformed gradients $(G_V Q^\top, QG_W)$ for the same minibatch, and radial projection commutes with $\mathcal{T}_Q$. Induction couples every subsequent risk value. Taking endpoint suprema proves the claim, also for randomized policies. □

Thus age indexes a process generating different checkpoints; it is not an additional cause after the complete state is fixed. The young and old pairs in Proposition 4 are neither predictor-matched nor risk-matched. The exact symmetry above gives a limited covariance under coherent relabeling. It does not make arbitrary rescaling free, preserve a quotient-space cost, or establish matched-predictor durability separation.

F.10 Finite certificate obligations and reproducibility

The computation checks closed-box inequalities after the analytic reductions and safe-tube arguments in this PDF. The obligations include:

1. the q-box equation 57, positivity of the discriminant and ζ, the anisotropy allowance in equation 70, and all hard-budget slacks of equation 69;
2. positivity of Δ, the two signs at the mobility bands, and the global gap equation 66;
3. the one-step tube bounds used in equation 88; and
4. the joint perturbation inequalities and the second 2048-cell cover in Appendix F.8.

The current envelope module covers the complete scalar interval with 2048 outward rational cells, including a derivative enclosure and bisection of the two scalar roots. The executable dependency graph and fixed input hashes are described in Appendix C; one command rebuilds the current numerical chain. Only the current proof chain is included in the anonymous package. The separate scalar global certificate is in Appendix G.

F.11 A verified rank-40 checkpoint bracket and policy gap

The input is history 0 in the seven-history file at step $3{,}360{,}000$, the same checkpoint used for the policy scaling test. Its source SHA-256 is

`6f9904a717783e47f81c00e4555a7b7b77c9c88d3db01697db4fe51c9eb2ef46.`

The portable certificate exports its binary64 coordinates and a feasible endpoint; every entry is interpreted as its exact dyadic rational value. This is a deterministic certificate for a saved numerical checkpoint, not a claim that floating-point old training equals the real-arithmetic stochastic path. The real-arithmetic history theorem is independent of this calculation.

Global lower endpoint. Compute $s^2 = \sum_{i\le r,j} A_{ij}^2/r$ by exact fractions, enclose its square root, and isolate $f_-(q_L) = \rho$. The certificate gives

$$\begin{gathered} q_L \in (1.9291700961608963586 9222, 1.9291700961608963586 9224), \\ \sqrt{q_L} - s > .01052240335. \end{gathered} \tag{111}$$

The complete-action lower bound rules out strictly safe mean survival for every smaller budget. No local optimization or assumption about an attaining factorization enters this direction. The generic upper envelope independently gives an upper budget below $.011073047$.

Improved feasible upper endpoint. At charged budget $.010917$, construct the explicit edit using an internal budget smaller by 10^{-10}. Exact fractions on the exported factors verify the cap, all three movement budgets, initial old and future risk, and incremental forgetting. The stored endpoint is thus legal despite rounding; legality is not inferred from an approximately saturated constraint.

For the 840-step mean replay, the arithmetic contract is IEEE binary64 round-to-nearest with ordinary length-56 dot products. Let $u = 2^{-53}$ and $\gamma_{128} = 128u/(1-128u)$. On the mean tube, factor Frobenius norms are below 14. Two matrix products per gradient, ridge and bias terms, parameter rounding, and the final updates have pair error at most

$$2\sqrt{2}\,\eta\gamma_{128}(12000) + 28\gamma_{128} < 10^{-12} \tag{112}$$

per step; γ_{128} covers fused and unfused dot products. Every computed raw step lies at least 10^{-8} inside the cap, so no floating-point projection is needed. Propagate $R_0 = 0$, $R_{j+1} = (1 + 20\eta)R_j + 10^{-12}$ by outward rational intervals. It stays below 10^{-8}, closing the tube bootstrap.

The product discrepancy is at most $D_j = 3\sqrt{2}R_j + R_j^2/2$. At all 841 times add the normalized-risk allowance

$$.00286D_j/4 + D_j^2/32 + \delta^2(3R_j + R_j^2/2)/16 + 10^{-13}. \tag{113}$$

The last term pays floating risk evaluation: the matrix-product error is at most $\gamma_{128}(196)$ and the squared-residual sum uses at most 6400 operations with residual norm below .02. Old-risk allowances are also checked. The resulting bounds are

$$\max_{j \le 840} f_j < 1.999800 \times 10^{-6} < \rho, \qquad f_-(q) > 1.897109 \times 10^{-6}. \tag{114}$$

Unrounded interval endpoints give a normalized-risk gap below 1.027×10^{-7}. The lower endpoint and feasible witness prove Proposition 7.

This is a complete-action *enclosure*, not root isolation: the interval can contain a plateau, several optimal lifts, and optimizer switches. The numerical policy root is neither substituted for β_c nor used to establish an optimized Gaussian converter. No nearest-lift regularity, uniform multiplier bound, or full-fiber coefficient is certified by this saved-checkpoint calculation.

G A GLOBALLY SOLVABLE CONTROL WITH A PARAMETER-BOX AGE SEPARATION

This control has a complete two-factor action class and a computed global center. It is a separate, noiseless scalar model: $f(x) = vwx$, $x \in \{-1, 1\}$ equally likely, old teacher .9, and future teacher 1. Thus every minibatch gives the population gradient. It tests the all-action optimization and deterministic age mechanism, not a noisy rank-40 Gaussian limit or nonunique factor fibers.

Start from $v_0 = w_0 = \sqrt{K/2}$. The old phase uses $\eta_o = .001$, $\lambda_o = .1$, and $T = 3000$ simultaneous steps, with the same radial factor cap as the future phase. An edit obeys

$$|v - v_T| \le \beta, \quad |w - w_T| \le .25, \quad v^2 + w^2 \le K, \quad \tfrac{1}{2}(vw - .9)^2 - \tfrac{1}{2}(v_T w_T - .9)^2 \le .001. \tag{115}$$

At every update, old safety is $\frac{1}{2}(vw - .9)^2 < .02$ and future safety is $\frac{1}{2}(vw - 1)^2 < \epsilon^2/2$. All budgets are identical between the young and old comparisons. Consider the continuous box

$$K \in [1.99, 2.01], \quad \eta_f \in [.00095, .00105], \quad \lambda_f \in [.095, .105], \quad \epsilon \in [.0245, .0255]. \tag{116}$$

The cap straddles the exact representation energy 2. The old learning rate and ridge are held fixed; changing K also changes the specified initializer.

Proposition 16 (Robust full-action age separation)**.** *Throughout equation 116, at the same budget $\beta = .005$, $V_{100}(\text{young}) = 1$ and $V_{100}(\text{old}) = 0$. Both checkpoints admit a legal immediately safe edit.*

Proof. Balanced old factors remain balanced, and their product follows $p^+ = p[1 - .001(p - .8)]^2$. The cap is inactive and this map is increasing on $[.8, 1.005]$. Outward endpoint iteration for 3000 steps encloses

$$.80128707439668 < p_T < .80133968670668, \tag{117}$$

$$.89514639830403 < v_T = w_T < .89517578536659. \tag{118}$$

These strict decimal bounds are rounded outwards from the recorded intervals.

At the young checkpoint choose $v = w = \sqrt{\min(K/2, 1)}$. The movement is at most $\sqrt{1.005} - 1 < .005$, energy obeys the cap, and old risk does not increase. Future product obeys $p^+ = p[1 + \eta_f(1 - \lambda_f - p)]^2$. On $p > .905$ this is increasing in p and decreasing in both η_f, λ_f. Iterating from .995 with the largest step and ridge gives $p_{100} > .97629092723416 > 1 - .0245$. The sequence decreases and remains above .905, closing the comparison. Old risk is slack throughout and no cap projection is needed.

For an old edited endpoint that could survive, $v, w > 0$ and $p > 1 - \epsilon$. The movement and cap imply

$$p_0 \le (v_T + .005)\sqrt{K - (v_T + .005)^2} < .98596313989996. \tag{119}$$

Set $a = 1 - \eta_f \lambda_f$, $e = v^2 + w^2$. The exact preprojection product is

$$p^+ = a^2 p + \eta_f a e(1 - p) + \eta_f^2 p(1 - p)^2. \tag{120}$$

While hypothetically safe, $p > 1 - \epsilon_{\max} = .9745$ and $e \le K_{\max}$. Both raw factors stay positive because each update ratio is at least $1 - \eta_{\max}\lambda_{\max} - \eta_{\max}\epsilon_{\max}K_{\max}/(1 - \epsilon_{\max}) > .99983$. Radial projection can only reduce their positive product. Therefore every surviving step decreases p by at least

$$d = \eta_{\min}\left[(2\lambda_{\min} - \eta_{\max}\lambda_{\max}^2)(1 - \epsilon_{\max}) - K_{\max}\epsilon_{\max} - \eta_{\max}(1 + \epsilon_{\max})\epsilon_{\max}^2\right] \tag{121}$$

$$= .000127193617832859687 5. \tag{122}$$

After 100 steps this would force $p_{100} < .97324377811667 < .9745$, a contradiction. The bound holds for the entire action set, including unbalanced factors and projected paths. Finally $v = v_T + .005$, $w = .977/v$ is an immediately safe legal repair: its energy is below $1.988365 < 1.99$, encoder movement below .25, and incremental old risk is negative. Thus the failure is not lack of immediate repairability. □

G.1 Global optimization and a computed critical budget

At nominal parameters $K = 2$, $\eta_f = .001$, $\lambda_f = .1$, $\epsilon = .025$, let $v_T = w_T$ be the old checkpoint above. We now optimize at $\beta \in [.03, .04]$. In addition to equation 120, the exact energy recursion is

$$e^+ = a^2 e + 4\eta a p(1 - p) + \eta^2 e(1 - p)^2. \tag{123}$$

For every initially safe legal pair, $.975 < p_0 \le 1$ and $2p_0 \le e_0 \le 2$. Its product decreases and stays above $.975a^{200} > .955$; its energy decreases and remains at least $2p_j > 1.90$. These claims follow by substituting the extrema into the two increments. They close a 100-step induction and prove cap inactivity on the entire domain. Old risk remains below $.005 < .02$. Whole-window future risk is therefore its terminal value.

For $F(p, e) = p^+$, $G(p, e) = e^+$, the Jacobian is

$$D(F, G) = \begin{pmatrix} a^2 - \eta a e + \eta^2[(1 - p)^2 - 2p(1 - p)] & \eta a(1 - p) \\ 4\eta a(1 - 2p) - 2\eta^2 e(1 - p) & a^2 + \eta^2(1 - p)^2 \end{pmatrix}. \tag{124}$$

Enclose this matrix on the entire tube $p \in [.955, 1]$, $e \in [1.90, 2]$, and multiply interval matrices in the actual time order. The 100 products give

$$.80156 < \frac{\partial p_{100}}{\partial p_0} < .81043, \qquad 0 \le \frac{\partial p_{100}}{\partial e_0} < .004022. \tag{125}$$

All intermediate enclosures are stored. To compare physical initial states, first raise e_0 to 2 at fixed p_0, then raise p_0 along $e_0 = 2$; both paths are feasible product–energy pairs in the above tube. Since $v_T + \beta < 1$, the cap and movement bound give the largest initial product at

$$v_*(\beta) = v_T + \beta, \qquad w_*(\beta) = \sqrt{2 - v_*(\beta)^2}. \tag{126}$$

This pair maximizes both initial product and energy, satisfies encoder and forgetting budgets, and is initially safe. Thus it globally minimizes the whole-window risk over the original full action class. The negative entry in the one-step Jacobian is retained; monotonicity is certified for the ordered product rather than assumed from an entrywise-positive one-step map.

The optimal terminal product is strictly increasing in β. Directed interval bisection of $p_{100}(\beta) = .975$ isolates the unique global critical budget in

$$\beta_* \in [.036613522336119785904884338378906 25, .036613522336701862514019012451171875]. \tag{127}$$

The width is below 5.83×10^{-13}. At equality strict survival fails; above the root it succeeds, and below it every edit fails. This is a scalar all-action center, distinct from both the rank-40 complete-action threshold and the explicit-policy experimental center.

Table 7: Globally certified scalar optimum versus a legal repair with initial product .98. Values are the largest future risk through update 100, divided by the threshold; values below one survive. The optimizer is over the complete physical two-factor action class.

β	global optimum	product-.98 repair
0.03	1.112941	1.861753
0.035	1.026138	1.862489
0.03661	1.000056	1.862712
0.03662	0.999897	1.862713
0.037	0.993871	1.862765
0.04	0.947987	1.863160

Verification. The standalone certificate uses 64-digit directed Decimal interval arithmetic; every operation rounds outwards and square roots are enclosed by adjacent representable numbers. The old-state enclosure, continuous-box drift inequalities, all 100 Jacobian products, and every bisection sign are saved. A separate 90-digit simultaneous-factor implementation verifies both root-bracket signs, records 12 nominal paths, and checks 32 corner paths. Corner checks support implementation consistency; the continuous-box theorem comes from the interval inequalities, not from interpolation of corners. None of these controls demonstrates nontrivial optimizer switching or validates a noisy Gaussian lifetime law.