LOW-RANK FRICTION FOR MEMORY-EFFICIENT TRANSFORMER PRETRAINING

PREPRINT

Rajit Rajpal
School of Mathematics
University of Edinburgh
Edinburgh, UK EH9 3FD
s2592586@ed.ac.uk

Benedict Leimkuhler
School of Mathematics
University of Edinburgh
Edinburgh, UK EH9 3FD
b.leimkuhler@ed.ac.uk

September 3, 2026

ABSTRACT

iKFAD is a recently proposed optimizer that closes the Adam–SGD performance gap observed in transformer model pretraining by replacing adaptive learning rates with adaptive friction in the momentum dynamics. Its limitation is that the full friction tensor $\xi \in \mathbb{R}^{m\times n}$ carries the same $\mathcal{O}(nm)$ memory overhead per layer as Adam's second-moment buffer. Here we exploit the rank-1 compression idea that Adafactor applies to Adam's second moment, replacing iKFAD's friction tensor ξ with a rank-1 outer-product factorization built from row and column momentum statistics, resulting in Rank-1 iKFAD (R-iKFAD). This reduces friction memory footprint from $\mathcal{O}(nm)$ to $\mathcal{O}(n+m)$ per layer which approximately halves iKFAD's total optimizer state. Despite this massive decrease in signal, we show that R-iKFAD maintains parity in performance with iKFAD, and consequently Adam on transformer pretraining tasks. Experiments on GPT2-Nano, TinyViT, DistilBERT, and GPT2-S confirm that R-iKFAD matches or exceeds iKFAD while nearly halving the memory footprint and being comparably robust to hyperparameters. We analyze the continuous-time dynamics in two damping regimes. For $\gamma > 0$ (linear damping) we prove exponential convergence under strong convexity. For $\gamma = 0$, the preferred option in our experiments, the friction is generated entirely based on past momentum and it switches off as the momentum vanishes, so geometric convergence cannot be shown. We nonetheless prove unconditional convergence to the minimizer together with explicit algebraic rates, $\mathcal{O}(t^{-1})$ and $\mathcal{O}(t^{-1/2})$ according to the regularization scale. For the discrete scheme that is actually implemented, with stochastic gradients and $\gamma = 0$, we further prove an algebraic rate down to an explicit noise floor. To our knowledge this is the first convergence rate for a rank-1 factored optimizer in continuous time, and the first analysis of one, in either time, that does not require positive damping.

1 Introduction

Adam [10] is the workhorse optimizer in modern deep learning, combining gradient momentum with coordinate-wise second-moment scaling to accelerate convergence on complex loss landscapes. For a model with N scalar parameters, Adam evolves two auxiliary states (the first and second moment estimates) and therefore requires $2N$ auxiliary scalars on top of the N parameters themselves, for a total memory cost of around $3N$. At scale, this $2N$ overhead presents a significant hardware bottleneck during large-model training. To mitigate this, Adafactor [17] introduces a rank-1 factorization of the second-moment matrix. For a matrix weight $W \in \mathbb{R}^{m\times n}$, Adafactor replaces Adam's full second-moment matrix $V \in \mathbb{R}^{m\times n}$ with a factorized approximation $\hat{V}_{ij} = R_i C_j/(\mathbf{1}_m^\top R)$, where $R \in \mathbb{R}^m$ and $C \in \mathbb{R}^n$ track row-wise and column-wise squared gradients, respectively. By eliminating the first-moment buffer, Adafactor reduces per-layer auxiliary state complexity from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$ (achieving memory savings of over 99% for large weight matrices) while maintaining competitive transformer pretraining performance.

An insightful paradigm for analyzing optimizer dynamics is based on continuous-time ordinary differential equations (ODEs) [3, 5], where updates emerge as discrete time-step approximations. For parameter state x, momentum p, and second-moment variable ζ, continuous-time Adam is given in Eqs. (1)-(3). While Adam incorporates parameter-wise adaptation into the *position* equation ($\dot{x}$) via $1/\sqrt{\zeta}$, Karoni et al. [7] demonstrated that coordinate-wise adaptation can alternatively be integrated into the *momentum* equation ($\dot{p}$). This is achieved by replacing the constant damping factor γ with a dynamic, coordinate-wise friction state ξ, resulting in the Individual Kinetic Friction-Adaptive Descent (iKFAD) dynamics (4)-(6)

$$\dot{x} = \frac{p}{\sqrt{\zeta} + \epsilon}, \tag{1}$$

$$\dot{p} = -\nabla f(x) - \gamma p, \tag{2}$$

$$\dot{\zeta} = [\nabla f(x)]^2 - \alpha\zeta. \tag{3}$$

$$\dot{x} = p, \tag{4}$$

$$\dot{p} = -\nabla f(x) - (\gamma + \xi) \odot p, \tag{5}$$

$$\dot{\xi} = \frac{[p]^2}{\mu} - \alpha\,\xi. \tag{6}$$

These dynamics require storing a full adaptive friction tensor ξ alongside momentum p. They incur an auxiliary footprint of $2N$ scalars, identical to Adam's auxiliary storage. Originally introduced for scalar/global friction [6], iKFAD extends friction-based adaptation to individual parameters, matching Adam's performance on standard vision and NLP benchmarks.

2 Methodology: Rank-1 iKFAD

Adafactor compresses Adam's second moment; Rank-1 iKFAD (R-iKFAD) applies the same compression to iKFAD's friction. Writing both as continuous-time systems for a matrix weight $X \in \mathbb{R}^{m\times n}$ makes the correspondence apparent, in the same sense in which the systems above approximate Adam and iKFAD [3, 5]. Most Adafactor implementations omit the first moment, but we enable it throughout for parity, so we retain momentum P and refer to the result as **Adafactor-m** (left). R-iKFAD (right) carries the same momentum and factorizes the friction instead:

$$\dot{X} = \frac{P}{\sqrt{\hat{\zeta}} + \epsilon}, \tag{7}$$

$$\dot{P} = -\nabla f(X) - \gamma P, \tag{8}$$

$$\dot{R} = [\nabla f(X)]^2\,\mathbf{1}_n - \alpha R, \tag{9}$$

$$\dot{C} = \mathbf{1}_m^\top [\nabla f(X)]^2 - \alpha C. \tag{10}$$

$$\dot{X} = P, \tag{11}$$

$$\dot{P} = -\nabla f(X) - \tilde{\xi} \odot P - \gamma P, \tag{12}$$

$$\dot{R} = \frac{1}{\mu}[P]^2\mathbf{1}_n - \alpha R, \tag{13}$$

$$\dot{C} = \frac{1}{\mu}\mathbf{1}_m^\top [P]^2 - \alpha C. \tag{14}$$

Here $\odot$ is the element-wise product and $[P]^2_{ij} = P^2_{ij}$; both systems carry the same linear damping γ. Each builds a rank-1 object from its factors,

$$\hat{\zeta} = \frac{RC^\top}{\mathbf{1}_m^\top R}, \qquad \tilde{\xi}_{ij} = \frac{R_i C_j}{\mathbf{1}_m^\top R + \epsilon_{\text{stab}}}, \tag{15}$$

with $\epsilon, \epsilon_{\text{stab}} > 0$ small regularizers, the latter ensuring numerical stability near $R = 0$. Adafactor-m is thus Adam's system with the full second moment ζ replaced by the rank-1 estimate $\hat{\zeta}$, and R-iKFAD is iKFAD's system with the full friction tensor ξ replaced by the rank-1 estimate $\tilde{\xi}$. The two differ in the same way Adam differs from iKFAD: Adafactor-m's factors aggregate squared *gradients* and act in the position equation, whereas R-iKFAD's aggregate squared *momentum* and act as damping in the momentum equation. In both cases the state per layer falls from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$ while full momentum is retained. We emphasize that R-iKFAD is designed to maintain parity with iKFAD (which matches Adam [7] in performance), rather than outperforming it.

Splitting Discretization. To construct a stable discrete-time algorithm, we rely on principles of geometric numerical integration of Hamiltonian systems [12], which have proved a robust foundation for discretization of dynamics-based optimization schemes. Following [7], we decompose the vector field of (11)–(14) into four sub-operators whose individual flows can be integrated analytically:

$$\begin{pmatrix}\dot{X}\\ \dot{P}\\ \dot{R}\\ \dot{C}\end{pmatrix} = \underbrace{\begin{pmatrix}P\\ 0\\ 0\\ 0\end{pmatrix}}_{\mathrm{A}} + \underbrace{\begin{pmatrix}0\\ -\nabla f(X)\\ 0\\ 0\end{pmatrix}}_{\mathrm{B}} + \underbrace{\begin{pmatrix}0\\ -\dfrac{RC^\top}{\mathbf{1}_m^\top R + \epsilon_{\mathrm{stab}}} \odot P\\ \dfrac{1}{\mu}[P]^2\mathbf{1}_n - \alpha R\\ \dfrac{1}{\mu}\mathbf{1}_m^\top[P]^2 - \alpha C\end{pmatrix}}_{\mathrm{C}} + \underbrace{\begin{pmatrix}0\\ -\gamma P\\ 0\\ 0\end{pmatrix}}_{\mathrm{D}}. \tag{16}$$

Sub-flows A, B, and D integrate analytically: A gives position drift; B applies the gradient kick; D applies linear friction (becoming the identity when $\gamma = 0$, i.e., R-iKFAD0). Sub-flow C couples momentum P and friction factors (R, C) and is integrated semi-analytically via an internal symmetric freeze-then-update scheme. The composition applied in practice is a first-order Lie–Trotter splitting:

$$\Phi_h = \Phi_h^{\mathrm{D}} \circ \Phi_h^{\mathrm{C}} \circ \Phi_h^{\mathrm{A}} \circ \Phi_h^{\mathrm{B}},$$

where sub-steps are executed in the sequence $\mathrm{B} \to \mathrm{A} \to \mathrm{C} \to \mathrm{D}$ (BACD) [7]. Position updates use the gradient-kicked momentum $P^{(B)}$ rather than P_n. Table 3 in Appendix A gives the explicit analytical update rules for each sub-step across the complete BACD sequence over time step h, together with the treatment of the coupling inside sub-flow C.

Memory Complexity. For a network parameter matrix at layer ℓ given by $W_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$, let $N = \sum_\ell m_\ell n_\ell$ denote the total number of scalar model parameters. Table 1 breaks down the auxiliary optimizer memory footprint required on top of the model weights. For non-2D parameters (biases, LayerNorm scales) that are not rank-1 factorizable, R-iKFAD falls back to element-wise iKFAD, contributing a negligible additional scalar per parameter. R-iKFAD requires approximately N auxiliary scalars in total, effectively halving the optimizer auxiliary memory footprint relative to both iKFAD and Adam. Adafactor-m occupies this same $\approx N$ regime and is therefore our primary baseline; the difference is which buffer is compressed, the second moment for Adafactor-m versus the friction tensor for R-iKFAD. Retaining momentum is deliberate: Adafactor is known to suffer training instabilities in certain transformer regimes when momentum is discarded entirely [17].

Table 1: Comparison of auxiliary optimizer memory complexity (excluding parameters).

Optimizer	Auxiliary State Variables	Total Auxiliary Footprint
Adam	First moment M, Second moment V	$2N$ scalars
iKFAD	Momentum P, Friction tensor ξ	$2N$ scalars
Adafactor-m, **R-iKFAD** (ours)	Momentum P, Factors $(R_\ell, C_\ell)_\ell$	$N + \sum_\ell (m_\ell + n_\ell)$ scalars

3 Theoretical Results

We analyze the continuous-time system (11)–(14) in both damping regimes. This section states the results and verifies them numerically; the proofs and the supporting lemmas are in Appendix F.

3.1 Overview

For $\gamma > 0$ a perturbed-energy argument gives exponential convergence under strong convexity (Theorem 1). Every experiment here, however, uses $\gamma = 0$. In R-iKFAD, $\tilde{\xi} \geq 0$ entrywise, and $\dot{\mathcal{E}} = -\langle P, \tilde{\xi} \odot P \rangle_F \leq 0$, so at $\gamma = 0$ the system remains dissipative with no damping imposed at all. What changes is the character of the dissipation: the friction is no longer imposed externally but manufactured by the trajectory out of its own past momentum. Two consequences follow. The dissipation weakens as $P \to 0$, so no exponential rate can hold, and the decay must be algebraic. Moreover $\dot{\mathcal{E}}$ vanishes at every turning point of the oscillation, so estimates must be made over sequential time blocks.

This last point is what makes the $\gamma = 0$ rate technically involved, and it is worth saying where the difficulty lies. Because $\dot{\mathcal{E}} = -\langle P, \tilde{\xi} \odot P \rangle_F$ is zero whenever $P = 0$, there is no differential inequality for $\mathcal{E}$ at any single instant: the usual route, in which one bounds $\dot{\mathcal{E}}$ below by a function of $\mathcal{E}$ and integrates, is unavailable. The argument therefore works on blocks of time and classifies each block three ways. Either the energy has already dropped by a fixed factor across the block, or a scalar moving average of the friction has not yet come down to the energy scale, or the block is *controlled*, meaning the total smoothed friction is trapped in a band proportional to the energy. On a controlled block

a finite-window concentration inequality converts the band into a quantitative lower bound on the dissipation, which is the step that replaces the missing pointwise estimate. Two further ingredients matter. The dissipation on a block is measured over a short observability window whose length is set independently of the block length; tying the two together is not possible, since the available lower bound decays like $\mathrm{e}^{-\alpha L}$ in the block length L while the concentration step simultaneously requires L to be large, and the two demands cannot both be met. Finally, the per-block decrement is multiplicative rather than additive, so it is accumulated through the reciprocal-energy potential $\Phi(e) = \beta/e + \vartheta/(2e^2)$, where $\vartheta = \alpha\mu\,\epsilon_{\mathrm{stab}}$. The two terms of Φ invert the two regimes of the denominator that appears in the decrement, and this is precisely where the two exponents come from: the $1/e$ term produces t^{-1} and the $1/e^2$ term produces $t^{-1/2}$.

3.2 Main theorems

Under linear damping the system converges exponentially, with the rank-1 structure paid for by a weight on the auxiliary block rather than by a restriction on the hyperparameters (Remark 3).

Theorem 1 (Exponential convergence of Rank-1 iKFAD)**.** *Assume $f \in C^2$ satisfies the strong convexity condition: there exist $a, b > 0$ such that for all X,*

$$\langle X - X^*, \nabla f(X)\rangle_F \;\geq\; a\big(f(X) - f(X^*)\big) + b\|X - X^*\|_F^2,$$

where X^* is a global minimizer of f, so that $\nabla f(X^*) = 0$ and $f \geq f(X^*)$ everywhere. Let $\gamma > 0$ and $\alpha, \mu, \epsilon_{\mathrm{stab}} > 0$, and fix an initial condition (X_0, P_0, R_0, C_0) with $R_0, C_0 \geq 0$ entrywise. Then there exist constants $\kappa > 0$ and $C_\star > 0$, explicit in the problem data, such that the solution of (11)–(14) *satisfies for all $t \geq 0$,*

$$f(X(t)) - f(X^*) + \|X(t) - X^*\|_F^2 + \|P(t)\|_F^2 + \|R(t)\|^2 + \|C(t)\|^2 \leq C_\star \mathrm{e}^{-\kappa t}.$$

The hypothesis $\gamma > 0$ is not an artifact of the argument: at $\gamma = 0$ the linearization about the equilibrium is not Hurwitz for any Hessian, so no exponential bound can hold there (Remark 2). What replaces it is a pair of results of quite different character. The first is that the dynamics converge at all, with no restriction on the hyperparameters and no convexity beyond a unique critical point.

Theorem 2 (Convergence at $\gamma = 0$)**.** *Suppose $f \in C^2$ is coercive and has a unique critical point X^*. Let $\alpha, \mu > 0$ and $\epsilon_{\mathrm{stab}} > 0$. Then every solution of* (11)–(14) *with $\gamma = 0$ and $R(0), C(0) \geq 0$ satisfies*

$$\big(X(t), P(t), R(t), C(t)\big) \longrightarrow \big(X^*, 0, 0, 0\big).$$

The same conclusion holds for $\epsilon_{\mathrm{stab}} = 0$ along the trace-matched solution.

The second is a rate, with every constant explicit.

Theorem 3 (Parameter-dependent algebraic rate)**.** *Assume uniform strong convexity and smoothness, $0 < m_f I \preceq \nabla^2 f(X) \preceq M_f I$ for every X, together with $\gamma = 0$ and $R(0) = C(0) = 0$. Let L, $K_\star$, β and $\vartheta = \alpha\mu\,\epsilon_{\mathrm{stab}}$ be the constants constructed in Appendix F.3. There are numerical constants $c, C > 0$ such that*

$$\mathcal{E}(t) \;\leq\; C\left[\mathcal{E}_0\,\mathrm{e}^{-ct/L} + \frac{\beta L}{K_\star(L+t)} + \sqrt{\frac{\vartheta L}{K_\star(L+t)}}\,\right] \qquad (t \geq 0).$$

Consequently $\mathcal{E}(t) = \mathcal{O}\big(t^{-1}\big)$ when $\epsilon_{\mathrm{stab}} = 0$ and $\mathcal{E}(t) = \mathcal{O}\big(t^{-1/2}\big)$ when $\epsilon_{\mathrm{stab}} > 0$. All constants depend only on $m, n, m_f, M_f, \alpha, \mu, \epsilon_{\mathrm{stab}}$ and $\mathcal{E}_0$.

The two exponents are regimes of a single estimate, separated at $\mathcal{E} \approx \alpha\mu\,\epsilon_{\mathrm{stab}}$, so the t^{-1} law is the operative one at any reachable horizon (Remark 4). The gap being closed also resolves an outstanding issue for iKFAD: it is suggested to be deployed at $\gamma = 0$, yet the existing convergence analysis [7] assumes $\gamma > 0$, so the friction-based optimizers have so far lacked a complete convergence theory. The argument is not tied to the rank-1 structure and serves as a template for closing the same gap for full-friction iKFAD. At $m = n = 1$ with $\epsilon_{\mathrm{stab}} = 0$ the reduction is exact: trace matching gives $R \equiv C$, hence $\tilde{\xi} = R^2/R = R$, and the system becomes scalar KFAD [6], so the t^{-1} rate covers that method too.

All three results above concern the continuous-time system. The scheme actually run is the BACD splitting of Section 2, and the perturbed-energy strategy carries over to it, with the additional difficulty that the friction factors are updated between the two momentum half-steps rather than simultaneously with them.

Theorem 4 (Discrete exponential convergence for the BACD scheme)**.** *Assume $f \in C^2$ satisfies $m_f I \preceq \nabla^2 f(X) \preceq MI$ for all X, with $0 < m_f \leq M$. Let $\gamma > 0$ and $\alpha, \mu, \epsilon_{\mathrm{stab}} > 0$, and fix $L > 0$. Then for every initial condition with*

$$\|X_0\|_F + \|P_0\|_F + \|R_0\| + \|C_0\| \leq L \qquad \text{and} \qquad R_0 \geq 0,\; C_0 \geq 0 \;\; \text{entrywise},$$

there exist $h^* > 0$, $\kappa > 0$ and $C_\star > 0$, depending only on $m_f, M, \gamma, \alpha, \mu, \epsilon_{\mathrm{stab}}, L$ and the dimensions m, n, and not *on h, such that the iterates of Table 3 satisfy, for all $h \in (0, h^*)$ and all $n \geq 0$,*

$$f(X_n) - f(X^*) + \|P_n\|_F^2 + \|R_n\|^2 + \|C_n\|^2 \leq C_\star \mathrm{e}^{-\kappa n h}.$$

Two caveats belong with this statement. The result requires $\gamma > 0$, and unlike the continuous case the discrete argument does not reach $\gamma = 0$: every constant in it is built on $\gamma > 0$, and the equilibrium is non-hyperbolic there (Remark 2), so the discrete theory describes a regime adjacent to, rather than identical with, the one the experiments deploy. Second, h^* and κ are existence claims and not usable bounds: the step-size threshold the proof produces is several orders of magnitude below the step sizes at which the scheme is observed to contract, the loss being inherited from the continuous estimate of $\tilde{\xi}$ rather than introduced by the discretization. Two directions remain open, and neither is a routine extension of the argument above. Closing $\gamma = 0$ in discrete time means rebuilding the block construction behind Theorem 3 step by step: at $\gamma = 0$ the only dissipation is $\tilde{\xi} \odot P$, which vanishes whenever P does, so no per-step difference inequality is available, and the observability window becomes a sum whose exact identities must be re-derived for the splitting. Admitting stochastic gradients adds two more obstacles: $\tilde{\xi}$ is quadratic in the momentum, so the dissipation is quartic and the cross terms reach sixth moments of the noise; and the second half-damp depends on the current draw, so conditional expectations do not factor, and at $\gamma = 0$ it cannot simply be discarded because it is part of the only dissipation present. The conclusion would change shape as well, since at a fixed step size with persistent noise the iterates reach a neighbourhood of the minimizer rather than the minimizer itself. Both are left to future work.

3.3 Relation to existing analyses

In the classical inertial system $\ddot{x} + \gamma\dot{x} + \nabla f(x) = 0$ the damping determines the rate. Constant $\gamma > 0$ with strongly convex f gives exponential convergence, the regime of Theorem 1, and the assumption is not cosmetic: at $\gamma = 0$ the system is conservative, the energy is exactly preserved, and no trajectory converges. The standard route to algebraic rates therefore keeps the damping but lets it vanish on a schedule chosen by the analyst, as in the $\ddot{x} + \frac{3}{t}\dot{x} + \nabla f(x) = 0$ limit of Nesterov's method [19], whose $\mathcal{O}(t^{-2})$ rate can be read off the ODE, and its α/t generalization [2], where $\alpha \geq 3$ gives $f - \min f \leq C/t^2$. Our damping is neither constant nor scheduled. It is $\tilde{\xi} \odot P$, generated by the trajectory from its own past momentum: vanishing as $P \to 0$, which precludes an exponential rate, but never identically zero, which is why convergence survives the removal of γ altogether. Being a function of state rather than of time, it lies outside both arguments. Among analyses of factored optimizers, Adafactor's rank-1 state has been given a discrete-time convergence theory in the non-convex smooth setting [4], and closest to the present work, rank-1 parameterized momentum and scaling estimators have been studied in continuous time [14], where convergence to critical points follows from a monotonically decreasing Hamiltonian. Two differences delimit our contribution: that analysis gives descent and asymptotic convergence but no rate, and it requires strictly positive damping. To our knowledge Theorem 3 is the first convergence *rate* for a rank-1 factored optimizer in continuous time, and the first such analysis that does not require positive damping. The block construction does require momentum, so it does not apply to Adafactor as originally proposed, whose continuous limit is first order in X; it may however be useful for Adafactor-m. For full-friction iKFAD at $\gamma = 0$ the same architecture applies with the concentration step removed, since elementwise friction makes the dissipation directly available, and that variant carries no normalizing denominator and hence no ϵ_{stab}, so only the t^{-1} law is expected; we have not carried out that analysis here.

3.4 Numerical validation

The exponents of Theorem 3 and the parameter *relations* predicted by Remark 4 were both checked numerically on quadratic objectives $f(X) = \frac{1}{2}\langle X, \mathcal{A}X\rangle_F$ with $\mathcal{A}$ a random symmetric positive definite operator on $\mathbb{R}^{mn}$, started from $R(0) = C(0) = 0$.

Two independent integrators were used. Parameter sweeps were run with an adaptive Runge–Kutta method (`ode89`, relative tolerance 10^{-10}, absolute tolerance 10^{-12}). The long-horizon runs used the BACD operator splitting of the method itself (Table 3) at fixed step h, which is the natural integrator here: each sub-flow is solved exactly, non-negativity of R and C is preserved by construction, no error control is required, and horizons of 10^7 time units are inexpensive. On their common range the two agree to about 0.01 in every measured slope, and halving h from 0.02 to 0.01 changes every slope by less than 10^{-3}.

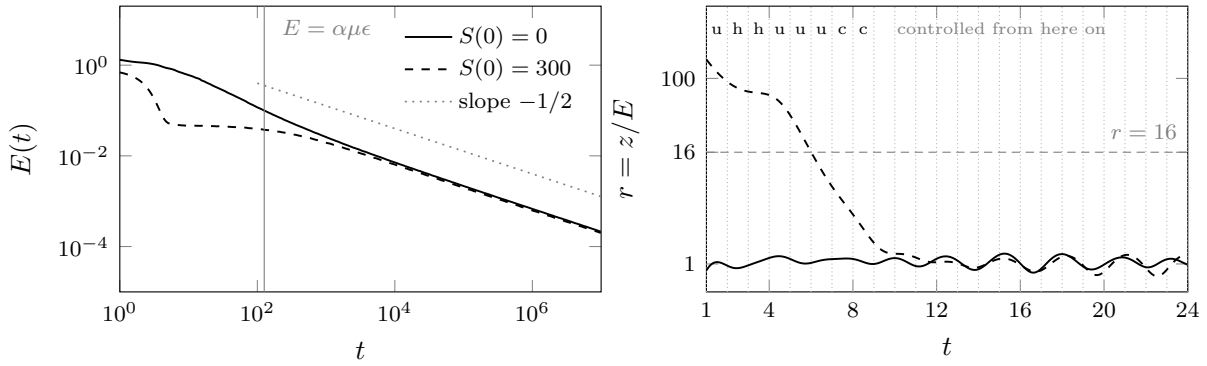

Figure 1: Decay and the block classification of Lemma 8, for $m = 4$, $n = 3$, $\epsilon_{\text{stab}} = 0.1$, $\alpha = \mu = 1$, integrated with the BACD splitting to $t = 10^7$. *Left:* two trajectories, the standard initialization $R(0) = C(0) = 0$ (solid) and one started with a large initial friction, $\mathbf{1}_m^\top R = \mathbf{1}_n^\top C = 300$ (dashed). Both settle onto the predicted $t^{-1/2}$ decay, with the crossover from the first regime occurring near $\mathcal{E} = \alpha\mu\,\epsilon_{\text{stab}}$. *Right:* a linear-time detail of the ratio $r = z/\mathcal{E}$ that decides the classification, with dotted vertical lines marking the block boundaries and the letters above giving the class of each block of the second trajectory: u untracked, h halving, c controlled. The standard trajectory has $r \approx 1$ from the outset, so all of its blocks are controlled. The second begins at $r \approx 160$, spends its first blocks untracked, passes through two halving blocks, and is controlled from the eighth onwards.

Exponents are as predicted, and independent of dimension. Local slopes of $\log \mathcal{E}$ against $\log t$ on successive decades, integrated to $T = 10^7$ at $h = 0.02$ with $\alpha = \mu = 1$:

	$[10^3, 10^4]$	$[10^4, 10^5]$	$[10^5, 10^6]$	$[10^6, 10^7]$
$\epsilon_{\text{stab}} = 0$ (predicted -1)				
$m = n = 2$	-0.955	-0.980	-0.986	-0.989
$m = 4, n = 3$	-0.991	-0.996	-0.996	-0.998
$m = 6, n = 5$	-0.993	-0.997	-0.998	-0.999
$\epsilon_{\text{stab}} = 0.1$ (predicted $-1/2$)				
$m = n = 2$	-0.499	-0.497	-0.496	-0.497
$m = 4, n = 3$	-0.558	-0.518	-0.505	-0.501
$m = 6, n = 5$	-0.593	-0.530	-0.509	-0.503

Both exponents are independent of the matrix dimensions, as Theorem 3 requires. The approach to $-1/2$ is slower for larger mn, and the crossover relation explains why: the second regime begins only once $\mathcal{E} \lesssim \alpha\mu\,\epsilon_{\text{stab}}$, and the constant in the first-regime law $\mathcal{E} \approx c_1\alpha\mu/t$ grows with mn, so the crossing occurs later for larger systems. On a horizon of $2 \cdot 10^4$ the 6×5 system still reads -0.583; by $[10^6, 10^7]$ it has reached -0.503.

The predicted dependence on α, μ and ϵ_{stab} holds. If the asymptotics are $\mathcal{E} \sim c_1\alpha\mu/t$ and $\mathcal{E} \sim c_2\alpha\mu\sqrt{\epsilon_{\text{stab}}/2t}$, the ratios below are constant. Evaluated at $t = 10^4$ with $m = 3$, $n = 2$:

quantity	parameters varied	mean	spread (max/min)
$t\mathcal{E}/(\alpha\mu)$, $\epsilon_{\text{stab}} = 0$	$\alpha, \mu \in \{\frac{1}{2}, 1, 2\}$	4.69	1.29
$\mathcal{E}\sqrt{t}/\big(\alpha\mu\sqrt{\epsilon_{\text{stab}}/2}\big)$, $\epsilon_{\text{stab}} > 0$	$\alpha, \mu \in \{\frac{1}{2}, 1\}$, $\epsilon_{\text{stab}} \in \{0.05, 0.2\}$	2.31	1.14
$\mathcal{E}/(\alpha\mu\,\epsilon_{\text{stab}})$ at the crossing	$\alpha, \mu \in \{\frac{1}{2}, 1\}$, $\epsilon_{\text{stab}} \in \{0.05, 0.2\}$	0.87	1.21

In the first row α and μ each range over a factor of 4, so $\alpha\mu$ ranges over a factor of 16, yet the ratio moves by less than 30%. The third row locates the transition between the two regimes at $\mathcal{E} \approx 0.87\,\alpha\mu\,\epsilon_{\text{stab}}$, confirming the crossover scale. For the scalar system $m = n = 1$ with $\alpha = \mu = 1$, evaluating at $t = 10^5$ gives $t\mathcal{E}/(\alpha\mu) = 0.909$ and $\mathcal{E}\sqrt{t}/(\alpha\mu\sqrt{\epsilon_{\text{stab}}/2}) = 0.884$, both within about 10% of the value 1 predicted by formal averaging; for $m, n > 1$ the same relations hold with a dimension-dependent constant, which the averaging computation does not predict.

Since ϵ_{stab} is small in practice, the crossover at $\mathcal{E} \approx \alpha\mu\,\epsilon_{\text{stab}}$ lies far below any energy reached in a finite run, so the t^{-1} law of the first regime is the operative one at any horizon of practical interest.

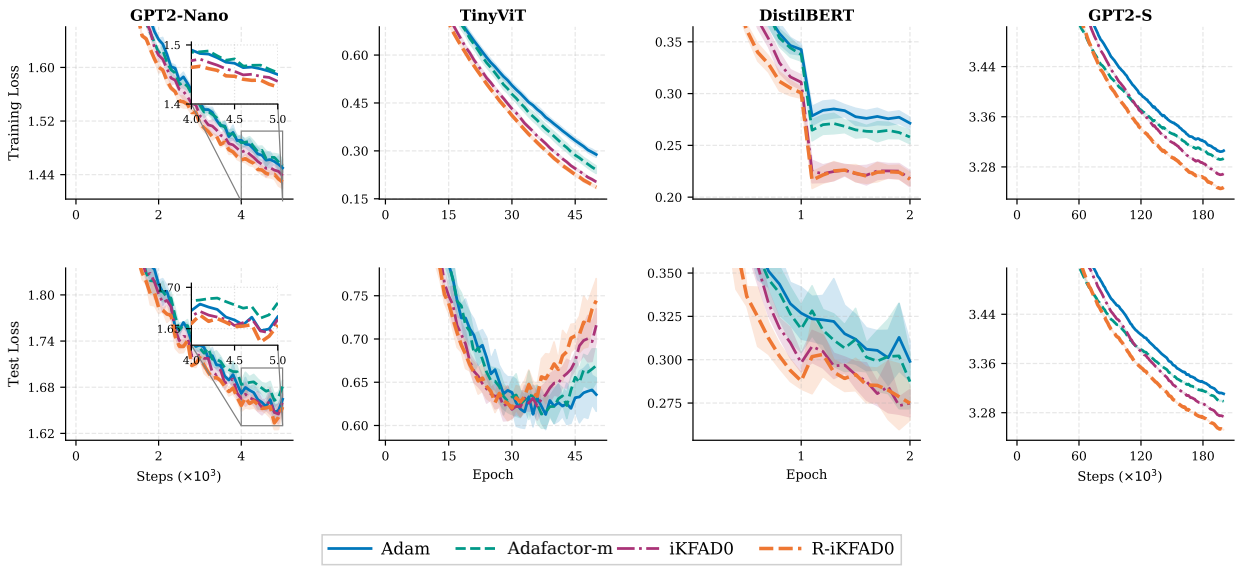

Figure 2: Loss curves for all four benchmarks, averaged over 10 random seeds. R-iKFAD0 matches or improves upon iKFAD0 on all tasks, while using approximately half the optimizer memory. Both methods are competitive with Adam and Adafactor-m, which require comparable or greater memory. Shaded regions denote one standard deviation.

4 Experiments

We set $\gamma = 0$ for both iKFAD and R-iKFAD throughout (denoted **iKFAD0** and **R-iKFAD0** respectively), following [7]; Section 4.2 confirms $\gamma > 0$ yields no benefit. We compare both against Adam and Adafactor-m across four benchmarks spanning image classification and language modeling at scales from 0.85M to 125M parameters:

- **GPT2-Nano (0.85M)** [8] on Shakespeare character-level language modeling.
- **TinyViT (5M)** [20] on CIFAR-10 [11] image classification.
- **DistilBERT (80M)** [16] on SST-2 [18] sentiment classification (pretraining from scratch).
- **GPT2-S (125M)** [15] on OpenWebText pretraining (nanoGPT architecture) [8].

To rigorously isolate the effects of the optimization dynamics and ensure a fair comparison, we use a constant learning rate, disable weight decay, and omit additional regularizers across all runs. Furthermore, we enable Adafactor's optional first moment (β_1 tuned) rather than the conventional momentum-free configuration, that is, we use Adafactor-m, so that it matches R-iKFAD's auxiliary-memory regime and the comparison is apples-to-apples. All hyperparameters are tuned via Optuna [1] (TPE sampler) with equal sweep budgets of approximately 80 trials per optimizer per model; GPT2-Nano and TinyViT received substantially more trials. See Appendix B for full configurations. Just like iKFAD, the optimal scale of μ varies by several orders of magnitude between different tasks due to differing momentum scales. Therefore, we recommend a wide-range initial sweep over μ and accept this as a potential limitation of our method.

4.1 Main Results

Figure 2 shows training and test loss across all benchmarks, and Figure 4 (Appendix C) the corresponding accuracies. R-iKFAD0 matches iKFAD0 in convergence speed and best achieved performance while using about half the optimizer memory, confirming that the rank-1 friction approximation incurs no performance cost. Crucially, R-iKFAD0 is competitive with Adam throughout, consistent with the finding in [7] that adaptive friction achieves parity with adaptive learning rates on LLM pretraining. We do not claim R-iKFAD surpasses Adam. The goal is to preserve iKFAD's performance parity with Adam while reducing memory overhead. Table 2 summarizes the lowest test loss achieved during training for all methods. Note that Adafactor-m requires tuning only two hyperparameters, whereas Adam, iKFAD, and R-iKFAD0 all require three. The fixed trial budget hence favored Adafactor-m's results disproportionately. GPT2-S results use a single seed due to computational constraints; all other benchmarks report mean $\pm$ standard deviation over 10 seeds. The optimizer auxiliary state is the memory an optimizer adds on top of the model parameters. Table 5 in Appendix E reports the absolute values on all four benchmarks in MB (float32).

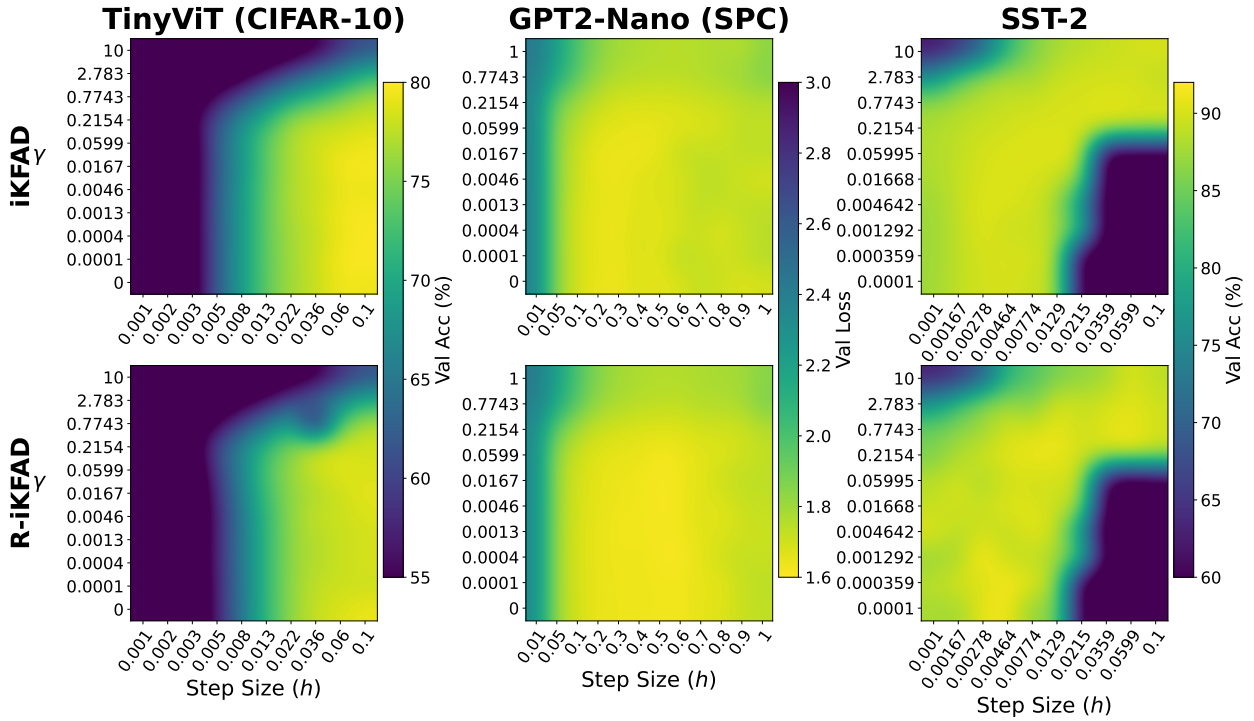

Figure 3: Two-dimensional γ–h sensitivity for R-iKFAD (top) and iKFAD (bottom). Each cell shows best performance at a fixed (h, γ) grid point; brighter = better. R-iKFAD retains comparably broad high-performance plateaus.

4.2 Ablation: $\gamma = 0$ versus $\gamma > 0$

Figure 5 (Appendix D) shows that $\gamma > 0$ yields no consistent improvement over $\gamma = 0$ on any of the three benchmarks tested (GPT2-Nano, TinyViT, DistilBERT). This is consistent with the finding in [7] that the adaptive friction mechanism alone provides sufficient damping. We therefore recommend **R-iKFAD0** ($\gamma = 0$) as the practical default since it has one fewer hyperparameter.

Table 2: Lowest test loss achieved during training (mean±std over 10 seeds; GPT2-S: single seed due to computational limitations). Lower is better. R-iKFAD0 matches or improves upon iKFAD0 across all benchmarks while using ≈2× less optimizer memory.

Dataset	Adam	Adafactor-m	iKFAD0	R-iKFAD0 (ours)
GPT2-Nano	1.643 ± 0.008	1.661 ± 0.007	1.640 ± 0.008	$\mathbf{1.633 \pm 0.011}$
TinyViT	$\mathbf{0.593 \pm 0.015}$	0.596 ± 0.010	0.606 ± 0.011	0.606 ± 0.012
DistilBERT	0.292 ± 0.005	0.280 ± 0.007	0.269 ± 0.005	$\mathbf{0.268 \pm 0.003}$
GPT2-S	3.301	3.287	3.263	**3.242**

4.3 Robustness to Hyperparameters

Figure 3 shows γ–h grid sweeps for R-iKFAD and iKFAD, revealing comparably broad high-performance plateaus. Practically, this allows for coarse hyperparameter tuning. This robustness stems from the adaptive friction mechanism: momentum spikes from large step sizes h automatically scale up the rank-1 factors R and C, increasing $\tilde{\xi}$ to dampen oscillations. This self-stabilizing feedback prevents overshooting, confirming that R-iKFAD inherits iKFAD's tolerance to aggressive hyperparameters [7] without degradation from rank-1 compression.

5 Conclusion

We introduced **R-iKFAD**, an optimizer that applies rank-1 compression to iKFAD's friction tensor to reduce per-layer friction storage from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$. By maintaining a full momentum buffer alongside low-rank friction vectors,

R-iKFAD occupies a sublinear-friction regime derived from provably convergent friction dynamics. The goal is not to surpass Adam, but to preserve iKFAD's established performance parity with Adam while cutting auxiliary optimizer memory in half. Our experiments across GPT2-Nano, TinyViT, DistilBERT, and GPT2-S confirm that R-iKFAD matches iKFAD in convergence speed, best achieved accuracy, and hyperparameter robustness at half the auxiliary memory cost. On the theoretical side, we analyze the continuous-time dynamics in both damping regimes: exponential convergence for $\gamma > 0$ under strong convexity, and for the undamped case $\gamma = 0$ that we actually deploy, unconditional convergence together with algebraic rates $\mathcal{O}(t^{-1})$ and $\mathcal{O}(t^{-1/2})$ separated by the regularization scale. To our knowledge this is the first convergence rate for a rank-1 factored optimizer in continuous time, and the first such analysis that does not require positive damping. For the BACD scheme that is actually implemented we further prove, at $\gamma = 0$ and with stochastic gradients, an algebraic rate down to an explicit noise floor: the regime the experiments deploy, and the one the discrete analysis of iKFAD leaves open. That result is a first foothold rather than a sharp description; removing its localisation, sharpening the discretisation floor, and validating performance at larger scale remain future work.

Acknowledgements

Part of this research was supported by the ProbAI Hub. The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR) [13]. Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/u6ih].

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[2] Hédy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1–2):123–175, 2018.

[3] André Belotto Da Silva and Maxime Gazeau. A general system of differential equations to model first-order adaptive algorithms. *The Journal of Machine Learning Research*, 21(1):5072–5113, 2020.

[4] Yusu Hong and Junhong Lin. Theoretical investigation of adafactor for non-convex smooth optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[5] Aikaterini Karoni. *Higher-order damping mechanisms with applications in optimisation and machine learning*. PhD thesis, The University of Edinburgh, 2024.

[6] Aikaterini Karoni, Benedict Leimkuhler, and Gabriel Stoltz. Friction-adaptive descent: A family of dynamics-based optimization methods. *Journal of Computational Dynamics*, 2023.

[7] Aikaterini Karoni, Rajit Rajpal, Benedict J. Leimkuhler, and Gabriel Stoltz. Adaptive momentum and nonlinear damping for neural network training. In *Forty-third International Conference on Machine Learning*, 2026.

[8] Andrej Karpathy. nanoGPT. `https://github.com/karpathy/nanoGPT`, 2022.

[9] Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 3rd edition, 2002.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

[12] Benedict Leimkuhler and Sebastian Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.

[13] Simon McIntosh-Smith, Sadaf R Alam, and Christopher Woods. Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence, 2024.

[14] Son Nguyen, Lizhang Chen, Bo Liu, and Qiang Liu. Memory-efficient optimization with factorized hamiltonian descent, 2024.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[16] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[17] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[18] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

[19] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[20] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. TinyViT: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

A Update Rules for Splitting Discretization

This appendix gives the explicit sub-step updates for the BACD splitting introduced in Section 2, and the treatment of the one sub-flow that does not admit a closed-form joint solution.

Table 3: Analytical sub-step update rules for the BACD splitting sequence over time step h.

Sub-step	Update Equation
B	$P^{(B)} = P_n - h\,\nabla f(X_n)$
A	$X_{n+1} = X_n + h\,P^{(B)}$
C	$\tilde{\xi}_n = \dfrac{R_n C_n^\top}{\mathbf{1}_m^\top R_n + \epsilon_{\text{stab}}}$
	$P^{(1)} = P^{(B)} \odot \exp\!\Big(-\frac{h}{2}\,\tilde{\xi}_n\Big)$
	$R_{n+1} = \mathrm{e}^{-\alpha h} R_n + \dfrac{1-\mathrm{e}^{-\alpha h}}{\mu\alpha}[P^{(1)}]^2 \mathbf{1}_n$
	$C_{n+1} = \mathrm{e}^{-\alpha h} C_n + \dfrac{1-\mathrm{e}^{-\alpha h}}{\mu\alpha}\mathbf{1}_m^\top[P^{(1)}]^2$
	$\tilde{\xi}_{n+1} = \dfrac{R_{n+1} C_{n+1}^\top}{\mathbf{1}_m^\top R_{n+1} + \epsilon_{\text{stab}}}$
	$P^{(2)} = P^{(1)} \odot \exp\!\Big(-\frac{h}{2}\,\tilde{\xi}_{n+1}\Big)$
D	$P_{n+1} = P^{(2)} \odot \mathrm{e}^{-\gamma h}$ (*identity when* $\gamma = 0$)

Within sub-flow C, no closed-form joint solution exists for coupled P and (R, C). We apply a symmetric half-damp sequence: evaluating $\exp(-\frac{h}{2}\tilde{\xi})$ twice around an exact exponential ODE factor update improves integration accuracy to $\mathcal{O}(h^2)$ within sub-flow C alone. Using the post-first-half-damp momentum $P^{(1)}$ for factor updates reduces forcing magnitude and prevents over-accumulation.

B Optimal Hyperparameters

Across models and tasks, the best-performing scales for μ (iKFAD0, R-iKFAD0) varied by several orders of magnitude. This variation is likely related to the scale of the momentum variables during training, since μ controls the strength of the adaptive friction terms. Normalizing the momentum variables across iterations could potentially reduce this sensitivity, but we leave this for future work.

Hyperparameters were selected by minimum validation loss during the Optuna sweeps; all reported results are computed on held-out test data.

Experimental configurations:

- **TinyViT – CIFAR-10**: batch size 128, 25 epochs, 100 trials. $h \in [10^{-6}, 10^{-1}]$, $\alpha \in [10^{-3}, 1]$, $\mu \in [10^{-8}, 10]$ for iKFAD0 and R-iKFAD0.

- **DistilBERT – SST-2**: batch size 16, 2 epochs, 80 trials. $h \in [10^{-6}, 10^{-1}]$, $\alpha \in [10^{-3}, 1]$, $\mu \in [10^{-8}, 10]$ for iKFAD0 and R-iKFAD0.
- **GPT2-Nano – Shakespeare**: batch size 16, 5000 steps for the hyperparameter sweeps and final reported runs, 500 trials. $h \in [10^{-6}, 5 \times 10^{-1}]$, $\alpha \in [10^{-5}, 10]$, $\mu \in [10^{-8}, 10]$ for iKFAD0 and R-iKFAD0.
- **GPT2-S – OpenWebText**: batch size 16, 75001 steps for the hyperparameter sweeps and 200001 steps for the final reported runs, 80 trials. $h \in [10^{-6}, 10^{-1}]$, $\alpha \in [10^{-3}, 1]$, $\mu \in [10^{-8}, 10]$ for iKFAD0 and $\mu \in [10^{-9}, 10^{-2}]$ for R-iKFAD0.

Table 4: Optimized hyperparameters for all experiments. All iKFAD0 and R-iKFAD0 entries have $\gamma = 0$. Dashes indicate parameters not used by the method.

	h	α	μ	β_1	β_2
TinyViT (CIFAR-10)					
iKFAD0	0.0725	0.0899	1.2643×10^{-5}	–	–
R-iKFAD0	0.0860	0.5327	9.4765×10^{-7}	–	–
Adam	5.5311×10^{-4}	–	–	0.8608	0.8852
Adafactor-m	5.6050×10^{-4}	–	–	0.8907	–
DistilBERT (SST-2)					
iKFAD0	0.0165	0.0574	9.8948×10^{-6}	–	–
R-iKFAD0	0.0423	0.0060	8.1874×10^{-7}	–	–
Adam	3.82×10^{-5}	–	–	0.9187	0.9825
Adafactor-m	2.5638×10^{-5}	–	–	0.9622	–
GPT2-Nano (Shakespeare)					
iKFAD0	0.4941	2.1955	4.5521×10^{-6}	–	–
R-iKFAD0	0.4842	2.8475	1.9407×10^{-6}	–	–
Adam	1.6803×10^{-3}	–	–	0.8876	0.9265
Adafactor-m	2.9923×10^{-3}	–	–	0.8893	–
GPT2-S (OpenWebText)					
iKFAD0	0.0938	0.1933	4.2726×10^{-7}	–	–
R-iKFAD0	0.0918	0.0579	1.3637×10^{-6}	–	–
Adam	1.0478×10^{-4}	–	–	0.9000	0.9971
Adafactor-m	7.2777×10^{-4}	–	–	0.8865	–

For all experiments, Adam was swept over $h \in [10^{-6}, 10^{-2}]$, $\beta_1 \in [0.85, 0.999]$, $\beta_2 \in [0.85, 0.999]$. Adafactor-m was swept over $h \in [10^{-5}, 10^{-2}]$ and $\beta_1 \in [0.85, 0.999]$, with no relative step or warmup. Adafactor-m has no tunable β_2: as originally proposed [17] and in standard practice, the second-moment decay follows the prescribed schedule $\beta_{2,t} = 1 - t^{-0.8}$ (`decay_rate` $= -0.8$), which is why its β_2 entries are left blank. We enable $\beta_1 > 0$ rather than the conventional momentum-free setting so that Adafactor-m occupies the same auxiliary-memory regime as R-iKFAD (Table 5). All sweeps used Optuna's [1] TPE sampler. Table 4 reports the best hyperparameters found. iKFAD0 and R-iKFAD0 are the $\gamma = 0$ variants; all such entries satisfy $\gamma = 0$ by definition. For iKFAD0 and R-iKFAD0, h denotes the step size and μ the adaptive damping coefficient. For Adam and Adafactor-m, h is the learning rate.

C Accuracy Curves

D Ablation: $\gamma = 0$ versus $\gamma > 0$

Figure 5 compares R-iKFAD ($\gamma > 0$, with linear damping) against R-iKFAD0 ($\gamma = 0$, without linear damping) on GPT2-Nano, TinyViT, and DistilBERT (SST-2). Adding the γ term yields no consistent improvement over the $\gamma = 0$ default on any benchmark, confirming that the rank-1 friction encoded in R and C provides sufficient adaptive damping. This is consistent with the analogous finding for iKFAD in [7]. For the $\gamma > 0$ variant, γ was not fixed by hand: we used the same settings as R-iKFAD0 given in Appendix B and additionally swept γ uniformly over $[10^{-6}, 10]$.

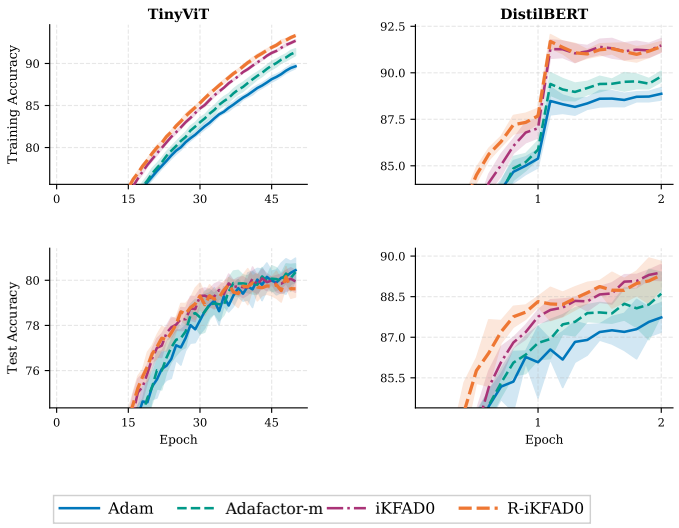

Figure 4: Accuracy averaged over 10 seeds. R-iKFAD0 matches iKFAD0 and Adam. Error bars denote one std.

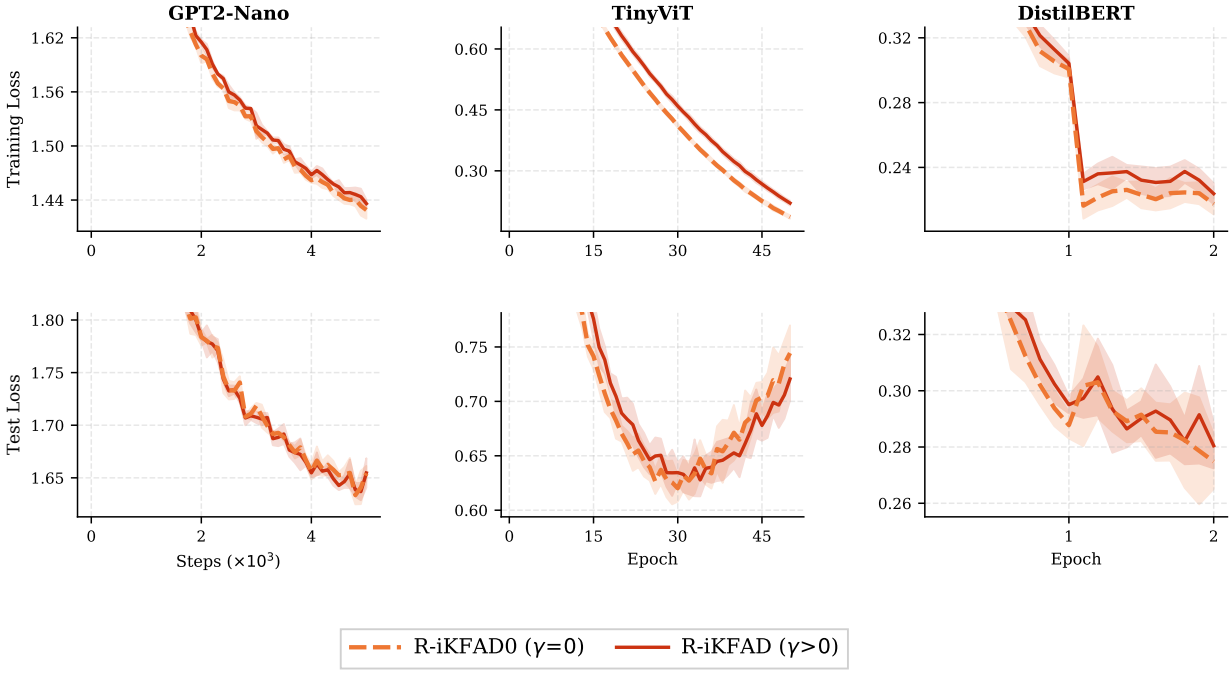

Figure 5: Training and test loss for R-iKFAD ($\gamma > 0$) vs. R-iKFAD0 ($\gamma = 0$) on GPT2-Nano, TinyViT, and DistilBERT (SST-2), averaged over 10 seeds (shaded = std). The two variants are indistinguishable in performance, supporting R-iKFAD0 as the recommended default.

Table 5: Optimizer auxiliary state (MiB, float32), computed over all model parameters. R-iKFAD achieves approximately $2\times$ savings over both iKFAD and Adam because the rank-1 friction factors (R_ℓ, C_ℓ) are negligible relative to the full momentum buffer. The Adafactor-m baseline used throughout enables its optional first moment ($\beta_1 > 0$; Section 4) and therefore carries an auxiliary footprint identical to R-iKFAD's, making the two directly comparable at matched memory. Momentum-free Adafactor is listed for reference.

Model	N	Adam	Adafactor-m	Adafactor	iKFAD	R-iKFAD
TinyViT	5.0M	30.7	15.6	0.2	30.7	15.6 (2.0×)
DistilBERT	80.0M	510.8	256.1	0.7	510.8	256.1 (2.0×)
GPT2-Nano	0.8M	6.1	3.1	< 0.1	6.1	3.1 (2.0×)
GPT2-S	124.4M	948.9	475.4	0.9	948.9	475.4 (2.0×)

E Memory Analysis

F Rank-1 iKFAD Convergence Analysis

We analyze the continuous-time Rank-1 iKFAD dynamics for a matrix parameter $X \in \mathbb{R}^{m\times n}$ (the vector case follows by flattening). The system is given by

$$\dot{X} = P, \tag{17a}$$

$$\dot{P} = -\nabla f(X) - \tilde{\xi} \odot P - \gamma P, \tag{17b}$$

$$\dot{R} = \frac{1}{\mu}[P]^2 \mathbf{1}_n - \alpha R, \tag{17c}$$

$$\dot{C} = \frac{1}{\mu}\mathbf{1}_m^\top [P]^2 - \alpha C, \tag{17d}$$

where $P \in \mathbb{R}^{m\times n}$, $R \in \mathbb{R}^m$, $C \in \mathbb{R}^n$, and $\gamma \geq 0$, $\alpha, \mu > 0$ are hyperparameters. Boundedness (Lemma 5) holds for $\gamma \geq 0$; the exponential convergence results below additionally require $\gamma > 0$ (see Remark 2). The rank-1 friction tensor $\tilde{\xi} \in \mathbb{R}^{m\times n}$ is defined element-wise as

$$\tilde{\xi}_{ij} = \frac{R_i C_j}{\sum_{k=1}^m R_k + \epsilon_{\text{stab}}},$$

with $\epsilon_{\text{stab}} > 0$ a small constant preventing division by zero. The notation $\odot$ denotes element-wise multiplication, $[\,\cdot\,]^2$ element-wise squaring, and $\langle \cdot, \cdot \rangle_F$ the Frobenius inner product.

F.1 Dynamical properties and boundedness

Throughout this appendix X^* denotes a global minimizer of f. Under the hypotheses used below such a point exists, because f is continuous and coercive, and it satisfies

$$\nabla f(X^*) = 0 \qquad \text{and} \qquad f(X) \geq f(X^*) \quad \text{for all } X. \tag{18}$$

Both properties are used repeatedly and neither follows from the growth condition (21) alone: expanding (21) along $X = X^* + tu$ at order t gives $(1-a)\langle u, \nabla f(X^*)\rangle_F \geq 0$ for every u, which forces $\nabla f(X^*) = 0$ only when $a \neq 1$; since $a = 1$ is admissible under strong convexity (Remark 1), the property must be imposed rather than derived.

Lemma 5 (Boundedness of solutions)**.** *Assume $f \in C^2$ with ∇f locally Lipschitz, $f(X) \to +\infty$ as $\|X\|_F \to +\infty$, and let X^* be a global minimizer of f as in* (18)*. Then for any initial condition (X_0, P_0, R_0, C_0) with $R_0, C_0 \geq 0$ entrywise, the solution of* (17a)–(17d) *is well-defined for all $t \geq 0$, and there exists a compact set $\mathcal{K}$ containing the trajectory. In particular, with*

$$P_{\max}^2 := 2\big(f(X_0) - f(X^*) + \tfrac{1}{2}\|P_0\|_F^2\big), \quad R_{\max} := \|R_0\| + \frac{P_{\max}^2}{\alpha\mu}, \quad C_{\max} := \|C_0\| + \frac{P_{\max}^2}{\alpha\mu},$$

all of which are finite and independent of t, we have for all $t \geq 0$

$$0 \leq R_i(t) \leq R_{\max}, \quad 0 \leq C_j(t) \leq C_{\max}, \quad \|P(t)\|_F \leq P_{\max},$$

and moreover the friction tensor obeys the bound

$$0 \leq \tilde{\xi}_{ij}(t) \leq \xi_{\max} := \min\Big\{C_{\max}, \frac{R_{\max}C_{\max}}{\epsilon_{\text{stab}}}\Big\}. \tag{19}$$

Note that $P_{\max}$ depends on neither μ nor γ.

Proof. The ODEs (17c)–(17d) have non-negative forcing and linear decay, so $R_0, C_0 \geq 0$ implies $R_i(t), C_j(t) \geq 0$ for all $t \geq 0$, and hence $\tilde{\xi}_{ij} \geq 0$ throughout.

Consider the reduced energy

$$\mathcal{E}(X, P) = f(X) - f(X^*) + \tfrac{1}{2}\|P\|_F^2,$$

which omits the R, C contributions. Differentiating along trajectories and canceling the gradient terms,

$$\dot{\mathcal{E}} = \langle \nabla f(X), P\rangle_F + \langle P, -\nabla f(X) - \tilde{\xi} \odot P - \gamma P\rangle_F = -\gamma\|P\|_F^2 - \sum_{i,j} \tilde{\xi}_{ij} P_{ij}^2 \leq 0, \tag{20}$$

both terms being non-positive because $\gamma \geq 0$ and $\tilde{\xi} \geq 0$ entrywise. No bound on $\|P\|_F$ is used in deriving (20), so the estimate is not circular. Consequently $\mathcal{E}(t) \leq \mathcal{E}(0)$ for all t, which yields at once

$$\|P(t)\|_F^2 \leq 2\mathcal{E}(0) =: P_{\max}^2,$$

and, since $f(X) - f(X^*) \leq \mathcal{E}(0)$ with f coercive, confinement of $X(t)$ to a compact sublevel set of f.

With $\|P\|_F \leq P_{\max}$ now established independently, the factors follow from the explicit linear filter. Integrating (17c),

$$R(t) = \mathrm{e}^{-\alpha t} R_0 + \frac{1}{\mu} \int_0^t \mathrm{e}^{-\alpha(t-s)} [P(s)]^2 \mathbf{1}_n \, \mathrm{d}s,$$

and using the sharp, dimension-free estimate $\big\|[P]^2\mathbf{1}_n\big\| \leq \|P\|_F^2 \leq P_{\max}^2$ (the entries of $[P]^2\mathbf{1}_n$ are non-negative and sum to $\|P\|_F^2$, so its 2-norm is at most its 1-norm),

$$\|R(t)\| \leq \|R_0\| + \frac{P_{\max}^2}{\alpha\mu} =: R_{\max},$$

with the analogous bound $C_{\max} := \|C_0\| + P_{\max}^2/(\alpha\mu)$. All four variables therefore remain in a compact set $\mathcal{K}$ for all $t \geq 0$, and global existence follows from the standard continuation criterion for locally Lipschitz vector fields.

Finally we establish (19). Two bounds are available. Bounding the denominator below by ϵ_{stab} gives $\tilde{\xi}_{ij} \leq R_{\max} C_{\max}/\epsilon_{\mathrm{stab}}$, which diverges as $\epsilon_{\mathrm{stab}} \downarrow 0$. The second is sharper and uniform in ϵ_{stab}: since every $R_k \geq 0$ we have $R_i \leq \mathbf{1}_m^\top R$, hence

$$\tilde{\xi}_{ij} = \frac{R_i}{\mathbf{1}_m^\top R + \epsilon_{\mathrm{stab}}} C_j \leq C_j \leq C_{\max},$$

and (19) is the minimum of the two. The constant $\epsilon_{\mathrm{stab}} > 0$ remains necessary for well-posedness at $R = 0$, which is exactly how the implementation initializes the factors, but it does not enter the bound on $\tilde{\xi}$, and hence does not enter the convergence rate. □

F.2 Exponential convergence of the continuous dynamics

The rank-1 factorization is not free in this regime. Because $\tilde{\xi}_{ij} \leq C_j$, the factor dynamics inject energy into the Lyapunov function cubically in the state while the available negative terms are quadratic, and the mismatch must be paid for by down-weighting the auxiliary block; in full-friction iKFAD the corresponding terms cancel exactly. Remark 3 makes this precise, and it is the sense in which the condition below is the price of compression rather than an artifact of the argument.

Theorem 1. Assume $f \in C^2$ satisfies the strong convexity condition: there exist $a, b > 0$ such that for all X,

$$\langle X - X^*, \nabla f(X)\rangle_F \geq a\big(f(X) - f(X^*)\big) + b\|X - X^*\|_F^2. \tag{21}$$

where X^* is a global minimizer of f as in (18). Let $\gamma > 0$ and $\alpha, \mu, \epsilon_{\mathrm{stab}} > 0$, and fix an initial condition (X_0, P_0, R_0, C_0) with $R_0, C_0 \geq 0$ entrywise. Then there exist constants $\kappa > 0$ and $C_\star > 0$ such that the solution of (17a)–(17d) satisfies for all $t \geq 0$,

$$f(X(t)) - f(X^*) + \|X(t) - X^*\|_F^2 + \|P(t)\|_F^2 + \|R(t)\|^2 + \|C(t)\|^2 \leq C_\star \mathrm{e}^{-\kappa t}.$$

Remark 1 (Coercivity is implied, not assumed)**.** *Lemma 5 requires f to be coercive, which Theorem 1 does not assume separately: it follows from* (21)*. Since $\nabla f(X^*) = 0$ by* (18)*, expanding* (21) *along $X = X^* + tu$ at second order gives $\big(1 - \frac{a}{2}\big)\langle u, \nabla^2 f(X^*)u\rangle_F \geq b\|u\|_F^2 > 0$, so $a < 2$. Writing $\phi(t) := f(X^* + tu) - f(X^*)$,* (21) *reads $t\phi'(t) \geq a\phi(t) + bt^2\|u\|_F^2$, i.e. $\big(\phi(t)/t^a\big)' \geq b\, t^{1-a}\|u\|_F^2$, which integrates to*

$$f(X) - f(X^*) \;\geq\; \frac{b}{2 - a}\,\|X - X^*\|_F^2.$$

This supplies both the coercivity needed by Lemma 5 and the inequality $f \geq f(X^*)$ of (18)*, and it is what makes $P_{\max}^2 = 2\mathcal{E}(0)$ non-negative and finite.*

Remark 2 (On the restriction $\gamma > 0$)**.** *The hypothesis $\gamma > 0$ is not an artifact of the proof. At $\gamma = 0$ the linearization of* (17a)–(17d) *about $(X^*, 0, 0, 0)$ is*

$$\dot{X} = P, \qquad \dot{P} = -\nabla^2 f(X^*)(X - X^*), \qquad \dot{R} = -\alpha R, \qquad \dot{C} = -\alpha C,$$

since R and C are driven by $[P]^2$ and $\tilde{\xi}$ is a ratio in which they appear, the friction term $\tilde{\xi} \odot P$ vanishes to higher order at the equilibrium in either regime of the denominator. Consequently the (X, P) block reduces to $\left[\begin{smallmatrix} 0 & I \\ -H & 0 \end{smallmatrix}\right]$ with $H = \nabla^2 f(X^*) \succeq 0$, which has zero trace and spectrum symmetric under $\lambda \mapsto -\lambda$, so it is not Hurwitz for any H. For a C^1 vector field a non-Hurwitz linearization precludes local exponential stability [9, Theorem 4.15]; the equilibrium is therefore not exponentially stable at $\gamma = 0$, and no bound of the form asserted in Theorem 1 can hold there.

The Lyapunov argument below also fails at $\gamma = 0$ for a separate and more elementary reason: the term $-\gamma\|P\|_F^2$ is the only negative contribution available to absorb the energy that the factor dynamics inject (Remark 3), and it is absent. What replaces exponential decay is therefore not a defect of this argument but a genuinely different rate, and it is established in Appendix F.3: convergence holds unconditionally (Theorem 2), at the algebraic rate $\mathcal{O}(t^{-1})$ for $\epsilon_{\mathrm{stab}} = 0$ and $\mathcal{O}(t^{-1/2})$ for $\epsilon_{\mathrm{stab}} > 0$ (Theorem 3). Since the experiments in Section 4 use $\gamma = 0$, that is the regime the deployed method occupies.

Remark 3 (The auxiliary channel is a genuine energy source)**.** *It is worth being explicit about why the weight B introduced in* (22) *is needed at all, since this is precisely the price of rank-1 compression. Before any weighting, the friction and factor terms of* (25) *combine to a non-negative quantity: by* (19)*, $\tilde{\xi}_{ij} \le C_j$, hence*

$$-\langle P, \tilde{\xi} \odot P\rangle_F + \langle R, [P]^2 \mathbf{1}_n\rangle + \langle C, \mathbf{1}_m^\top [P]^2\rangle = \sum_{i,j} P_{ij}^2 \big(R_i + C_j - \tilde{\xi}_{ij}\big) \;\ge\; \sum_{i,j} R_i P_{ij}^2 \;\ge\; 0$$

identically: the factor dynamics inject energy into $\mathcal{W}_\epsilon$, cubically in the state, whereas the available negative terms are quadratic. In full-friction iKFAD the corresponding terms cancel exactly, because there the friction obeys $\dot{\xi} = [P]^2/\mu - \alpha\xi$ elementwise and $\langle P, \xi \odot P\rangle_F$ is matched term for term by $\langle \xi, [P]^2\rangle$. Under the rank-1 factorization no such cancellation is available, and the mismatch must be paid for, here by down-weighting the auxiliary block. No sharpening of $P_{\max}$ or of (26) *can remove it.*

Proof of Theorem 1. We employ the perturbed Lyapunov function

$$\mathcal{W}_\epsilon(X, P, R, C) = f(X) - f(X^*) + \frac{1}{2}\|P\|_F^2 + \frac{\mu B}{2}\big(\|R\|^2 + \|C\|^2\big) + \epsilon\langle X - X^*, P\rangle_F + \frac{\epsilon}{2}\|X - X^*\|_F^2, \quad (22)$$

where $\epsilon \in (0, \frac{1}{4}]$ is a small parameter to be chosen later. Applying Young's inequality in the form $\epsilon|\langle X - X^*, P\rangle_F| \le \epsilon\big(\frac{1}{4}\|X - X^*\|_F^2 + \|P\|_F^2\big)$ and using $\epsilon \le 1/4$ gives

$$\mathcal{W}_\epsilon \ge f(X) - f(X^*) + \frac{1}{4}\|P\|_F^2 + \frac{\mu B}{2}(\|R\|^2 + \|C\|^2) + \frac{\epsilon}{4}\|X - X^*\|_F^2, \quad (23)$$

$$\mathcal{W}_\epsilon \le f(X) - f(X^*) + \frac{3}{4}\|P\|_F^2 + \frac{\mu B}{2}(\|R\|^2 + \|C\|^2) + \frac{3\epsilon}{2}\|X - X^*\|_F^2. \quad (24)$$

Differentiating $\mathcal{W}_\epsilon$ along trajectories of (17a)–(17d) yields

$$\begin{aligned}\dot{\mathcal{W}}_\epsilon = {}& \langle \nabla f(X), P\rangle_F + \langle P, -\nabla f(X) - \tilde{\xi} \odot P - \gamma P\rangle_F \\ & + \mu B\langle R, \dot{R}\rangle + \mu B\langle C, \dot{C}\rangle \\ & + \epsilon\|P\|_F^2 - \epsilon\langle X - X^*, \nabla f(X)\rangle_F - \epsilon\langle X - X^*, \tilde{\xi} \odot P\rangle_F + \epsilon(1 - \gamma)\langle X - X^*, P\rangle_F.\end{aligned}$$

Canceling $\langle \nabla f(X), P\rangle_F$ with $-\langle P, \nabla f(X)\rangle_F$ and substituting the auxiliary dynamics gives

$$\begin{aligned}\dot{\mathcal{W}}_\epsilon = {}& -\gamma\|P\|_F^2 - \langle P, \tilde{\xi} \odot P\rangle_F + B\langle R, [P]^2 \mathbf{1}_n\rangle + B\langle C, \mathbf{1}_m^\top [P]^2\rangle - B\alpha\mu(\|R\|^2 + \|C\|^2) \\ & + \epsilon\|P\|_F^2 - \epsilon\langle X - X^*, \nabla f(X)\rangle_F - \epsilon\langle X - X^*, \tilde{\xi} \odot P\rangle_F + \epsilon(1 - \gamma)\langle X - X^*, P\rangle_F. \quad (25)\end{aligned}$$

Define $\mathcal{M} := -\langle P, \tilde{\xi} \odot P\rangle_F + B\langle R, [P]^2 \mathbf{1}_n\rangle + B\langle C, \mathbf{1}_m^\top [P]^2\rangle$, the collection of all terms in (25) generated by the friction and factor dynamics. Expanding element-wise,

$$\mathcal{M} = \sum_{i,j} P_{ij}^2 \Big(BR_i + BC_j - \frac{R_i C_j}{\sum_k R_k + \epsilon_{\mathrm{stab}}}\Big).$$

Since $-\langle P, \tilde{\xi} \odot P\rangle_F \le 0$, we can drop this negative contribution and bound the positive terms. The vector $[P]^2\mathbf{1}_n$ is entrywise non-negative with entries summing to $\|P\|_F^2$, so

$$\big\|[P]^2\mathbf{1}_n\big\|_2 \le \big\|[P]^2\mathbf{1}_n\big\|_1 = \|P\|_F^2,$$

and likewise $\big\|\mathbf{1}_m^\top[P]^2\big\|_2 \le \|P\|_F^2$. This estimate is sharp, with equality when a single row (respectively column) of P is non-zero, and it carries no dimensional factor. Hence by Cauchy–Schwarz followed by Young's inequality with any $\theta > 0$, using $\|P\|_F \le P_{\max}$ from Lemma 5,

$$\langle R, [P]^2\mathbf{1}_n\rangle \le \|R\|\,\|P\|_F^2 \le \frac{\theta}{2}\|R\|^2 + \frac{P_{\max}^2}{2\theta}\|P\|_F^2,$$
$$\langle C, \mathbf{1}_m^\top[P]^2\rangle \le \|C\|\,\|P\|_F^2 \le \frac{\theta}{2}\|C\|^2 + \frac{P_{\max}^2}{2\theta}\|P\|_F^2.$$

Multiplying by B and adding, and discarding $-\langle P, \tilde{\xi} \odot P\rangle_F \le 0$,

$$\mathcal{M} \le \frac{B\theta}{2}\big(\|R\|^2 + \|C\|^2\big) + \frac{B\,P_{\max}^2}{\theta}\|P\|_F^2. \tag{26}$$

For the cross-terms involving $X - X^*$, we apply Young's inequality with constants $\eta, \delta > 0$:

$$\epsilon(1-\gamma)\langle X - X^*, P\rangle_F \le \epsilon|1-\gamma|\Big(\eta\|X - X^*\|_F^2 + \frac{1}{4\eta}\|P\|_F^2\Big), \tag{27}$$

$$-\epsilon\langle X - X^*, \tilde{\xi} \odot P\rangle_F \le \epsilon\Big(\delta\|X - X^*\|_F^2 + \frac{\xi_{\max}^2}{4\delta}\|P\|_F^2\Big), \tag{28}$$

where $\xi_{\max}$ is the entry-wise bound on $\tilde{\xi}$ established in (19). We use it in the form $\xi_{\max} \le C_{\max}$, which is uniform in ϵ_{stab}; the alternative branch $R_{\max}C_{\max}/\epsilon_{\mathrm{stab}}$ is also valid but diverges as $\epsilon_{\mathrm{stab}} \downarrow 0$ and is never needed. Consequently ϵ_{stab} enters the hypotheses only through well-posedness and does not appear in κ.

Substituting (26), (27), (28) and the strong convexity condition (21) into (25) yields

$$\begin{aligned}\dot{\mathcal{W}}_\epsilon \le &-\Big[\gamma - \epsilon - \frac{B\,P_{\max}^2}{\theta} - \epsilon\Big(\frac{|1-\gamma|}{4\eta} + \frac{\xi_{\max}^2}{4\delta}\Big)\Big]\|P\|_F^2 \\ &- B\Big(\alpha\mu - \frac{\theta}{2}\Big)\|R\|^2 - B\Big(\alpha\mu - \frac{\theta}{2}\Big)\|C\|^2 \\ &- a\epsilon\big(f(X) - f(X^*)\big) - \epsilon\Big[b - \eta|1-\gamma| - \delta\Big]\|X - X^*\|_F^2.\end{aligned} \tag{29}$$

We now choose the free parameters to make all coefficients negative. Set $\theta = \alpha\mu$, so that $B\big(\alpha\mu - \frac{\theta}{2}\big) = \frac{B\alpha\mu}{2} > 0$ for every $B > 0$, and likewise for C; the auxiliary coefficients are therefore positive irrespective of B, and the relative decay rate $\frac{B\alpha\mu/2}{\mu B/2} = \alpha$ does not depend on B at all. The leak into the momentum coefficient is $K_1 := BP_{\max}^2/\theta = BP_{\max}^2/(\alpha\mu)$. Next choose

$$\eta = \frac{b}{4\big(1 + |1-\gamma|\big)}, \qquad \delta = \frac{b}{4},$$

which is well defined for every $\gamma \ge 0$, including $\gamma = 1$. Since $|1-\gamma|/\big(1 + |1-\gamma|\big) < 1$ we have $\eta|1-\gamma| < b/4$, hence $b - \eta|1-\gamma| - \delta > b/2 > 0$. Substituting η into (27) gives $|1-\gamma|/(4\eta) = |1-\gamma|\big(1 + |1-\gamma|\big)/b$, so the coefficient of $\|P\|_F^2$ becomes

$$\Gamma(\epsilon) := \gamma - K_1 - \epsilon\,\Lambda, \qquad \Lambda := 1 + \frac{|1-\gamma|\big(1 + |1-\gamma|\big) + \xi_{\max}^2}{b}.$$

Since $K_1 = BP_{\max}^2/(\alpha\mu)$ is proportional to B, and B is a free parameter of the proof which appears nowhere in the statement of Theorem 1, we may simply set

$$B := \min\Big\{1,\ \frac{\gamma\alpha\mu}{2P_{\max}^2}\Big\}, \tag{30}$$

which gives $K_1 \le \gamma/2$ and hence $\gamma - K_1 \ge \gamma/2 > 0$ for *every* $\gamma, \alpha, \mu > 0$. (If $P_{\max} = 0$ then $P_0 = 0$ and $X_0 = X^*$, so the trajectory is already at equilibrium and there is nothing to prove; take $B = 1$.) No lower bound on μ is required.

There are exactly two constraints on ϵ: the bound $\epsilon \le 1/4$ used for (23)–(24), and positivity of $\Gamma(\epsilon)$. We may therefore fix

$$\epsilon := \min\Big\{\frac{1}{4},\ \frac{\gamma - K_1}{2\Lambda}\Big\}, \tag{31}$$

which gives $\Gamma(\epsilon) \ge \frac{\gamma - K_1}{2} \ge \frac{\gamma}{4} > 0$.

Shrinking B does not affect the rate, since by the preceding paragraph the auxiliary decay rate α is B-invariant; it enters only the prefactor. Inverting (23) for each quantity separately gives

$$f(X) - f(X^*) \le \mathcal{W}_\epsilon, \quad \|P\|_F^2 \le 4\mathcal{W}_\epsilon, \quad \|R\|^2 + \|C\|^2 \le \frac{2}{\mu B}\mathcal{W}_\epsilon, \quad \|X - X^*\|_F^2 \le \frac{4}{\epsilon}\mathcal{W}_\epsilon,$$

and since the conclusion of Theorem 1 is the *sum* of these, the four bounds add:

$$C_\star := \Big(5 + \frac{2}{\mu B} + \frac{4}{\epsilon}\Big)\,\mathcal{W}_\epsilon(0). \tag{32}$$

Consequently, there exists $\kappa_0 > 0$ such that

$$\dot{\mathcal{W}}_\epsilon \le -\kappa_0\Big(f(X) - f(X^*) + \|P\|_F^2 + \|R\|^2 + \|C\|^2 + \|X - X^*\|_F^2\Big).$$

Using the upper bound (24), we obtain $\dot{\mathcal{W}}_\epsilon \le -\kappa\mathcal{W}_\epsilon$ for some $\kappa > 0$. Grönwall's inequality then gives $\mathcal{W}_\epsilon(t) \le \mathcal{W}_\epsilon(0)\mathrm{e}^{-\kappa t}$, and the lower bound (23) implies exponential decay of each individual term. □

F.3 Convergence and algebraic rates at $\gamma = 0$

We now treat the undamped case $\gamma = 0$, which is the variant recommended in practice and used in all experiments. Two results are established, of quite different character. The first is that the dynamics converge at all: for any coercive f with a unique critical point, every trajectory tends to the minimizer, with no restriction on α, μ or ϵ_{stab} (Theorem 2). The second is a rate: under uniform strong convexity,

$$\mathcal{E}(t) = \mathcal{O}(t^{-1}) \quad (\epsilon_{\mathrm{stab}} = 0), \qquad \mathcal{E}(t) = \mathcal{O}(t^{-1/2}) \quad (\epsilon_{\mathrm{stab}} > 0),$$

with all constants explicit in $m, n, m_f, M_f, \alpha, \mu, \epsilon_{\mathrm{stab}}$ and $\mathcal{E}(0)$ (Theorem 3). The decay is algebraic, not exponential, and the two exponents correspond to two regimes separated by the size of the smoothed friction relative to ϵ_{stab}.

Why the rate is delicate. At $\gamma = 0$ the only damping is $\tilde{\xi} \odot P$, and $\tilde{\xi}$ is built from *past* momentum. Two consequences follow. First, as $P \to 0$ the friction switches itself off, so no exponential rate can hold and the answer must be algebraic. Second, $\dot{\mathcal{E}} = -\langle P, \tilde{\xi} \odot P\rangle_F$ vanishes at every turning point of the oscillation, where $P = 0$; there is therefore no differential inequality for $\mathcal{E}$ at a single instant, and every estimate must be made over a stretch of time. That is why the argument is organized into blocks, with a three-way alternative on each: either the energy drops by a fixed factor, or a scalar moving average contracts toward the energy scale, or a finite-window concentration inequality forces a definite increase of a reciprocal-energy Lyapunov function.

Notation for this subsection. Throughout, $\gamma = 0$ and $\epsilon_{\mathrm{stab}} \ge 0$ is written ϵ_{stab}; we abbreviate $S = \mathbf{1}_m^\top R$ and use the reduced energy $\mathcal{E}$ of Lemma 5, writing $\mathcal{E}_0 := \mathcal{E}(0)$, so that $P_{\max}^2 = 2\mathcal{E}_0$. Set

$$A := [P]^2, \qquad \sigma := \textstyle\sum_{ij} A_{ij} = \|P\|_F^2, \tag{33}$$

and let W solve the entrywise exponential moving average

$$W' = \alpha\,(A - W), \qquad W(0) = 0. \tag{34}$$

Write

$$\bar{a}_i = \sum_j W_{ij}, \qquad \bar{b}_j = \sum_i W_{ij}, \qquad z = \sum_{ij} W_{ij} = \alpha\mu S, \qquad \vartheta := \alpha\mu\,\epsilon_{\mathrm{stab}}. \tag{35}$$

Solving (17c)–(17d) by variation of constants identifies these with the friction factors,

$$R_i = \frac{\bar{a}_i}{\alpha\mu}, \qquad C_j = \frac{\bar{b}_j}{\alpha\mu}, \qquad S = \frac{z}{\alpha\mu}, \tag{36}$$

so z is the total smoothed friction. When $\epsilon_{\mathrm{stab}} = 0$ we set $\tilde{\xi} = 0$ at $S = 0$; with $R(0) = C(0) = 0$ the identity $\mathbf{1}_m^\top R = \mathbf{1}_n^\top C$ is invariant, so this is the continuous extension along the trajectory. To avoid collision with the symbols of Theorem 1 we write β for the upper band constant, $\nu := mn$, and $K_\star$ for the concentration constant.

F.3.1 Unconditional convergence

Theorem 2. Suppose $f \in C^2$ is coercive and has a unique critical point X^*. Let $\alpha, \mu > 0$ and $\epsilon_{\text{stab}} > 0$. Then every solution of (17a)–(17d) with $\gamma = 0$ and $R(0), C(0) \geq 0$ satisfies

$$\big(X(t), P(t), R(t), C(t)\big) \longrightarrow \big(X^*, 0, 0, 0\big).$$

The same conclusion holds for $\epsilon_{\text{stab}} = 0$ along the trace-matched solution whenever the continuous extension above is used.

Proof of Theorem 2. Lemma 5 makes the trajectory precompact, and (20) with $\gamma = 0$ gives $\dot{\mathcal{E}} = -D \leq 0$ where $D := \langle P, \tilde{\xi} \odot P \rangle_F$, so LaSalle's invariance principle applies. Let (X, P, R, C) be a bounded complete trajectory in the largest invariant subset of $\{D = 0\}$. Bounded completeness removes the backward-growing homogeneous mode from (17c)–(17d), leaving

$$R_i(t) = \frac{1}{\mu} \int_{-\infty}^{t} \mathrm{e}^{-\alpha(t-s)} \sum_j P_{ij}(s)^2 \, \mathrm{d}s, \qquad C_j(t) = \frac{1}{\mu} \int_{-\infty}^{t} \mathrm{e}^{-\alpha(t-s)} \sum_i P_{ij}(s)^2 \, \mathrm{d}s.$$

If $P_{ij}(t_0) \neq 0$ for some i, j, t_0, continuity makes both integrands positive on an interval just before t_0, hence $R_i(t_0) C_j(t_0) > 0$ and therefore $\tilde{\xi}_{ij}(t_0) P_{ij}(t_0)^2 > 0$, contradicting $D = 0$. Thus $P \equiv 0$, and (17b) then gives $\nabla f(X) = 0$, so $X = X^*$ by uniqueness of the critical point. Finally the only bounded complete solutions of $\dot{R} = -\alpha R$ and $\dot{C} = -\alpha C$ are identically zero. □

F.3.2 A quantitative observability window

For the rate we assume uniform strong convexity and smoothness,

$$0 < m_f I \preceq \nabla^2 f(X) \preceq M_f I \qquad \text{for every } X, \tag{37}$$

which implies (21) with $a = m_f / M_f$ and $b = m_f / 2$. Write $Y = X - X^*$ and set

$$K_f = 1 + \frac{1}{m_f}, \quad C_1 = \frac{1}{a} + \frac{1}{2}, \quad C_2 = \frac{2K_f}{a}, \quad C_3 = \frac{\xi_{\max}}{2ab}, \qquad T_0 = 4(C_2 + C_3), \quad c_0 = \frac{1}{2C_1}. \tag{38}$$

Lemma 6 (Window observability)**.** *For every $t \geq 0$, either*

$$\mathcal{E}(t + T_0) \leq \tfrac{1}{2} \, \mathcal{E}(t), \tag{39}$$

or

$$\int_t^{t+T_0} \|P(s)\|_F^2 \, ds \;\geq\; c_0 T_0 \, \mathcal{E}(t + T_0). \tag{40}$$

Proof. For $V = \langle Y, P \rangle_F$ we have $\dot{V} = \|P\|_F^2 - \langle Y, \nabla f(X) \rangle_F - \langle Y, \tilde{\xi} \odot P \rangle_F$. Integrate over $[t, t + T_0]$. By Cauchy–Schwarz against the measure $\tilde{\xi}$, then in time, then Young, using (19) and $\int D = \mathcal{E}(t) - \mathcal{E}(t + T_0)$,

$$\int \big| \langle Y, \tilde{\xi} \odot P \rangle_F \big| \leq \Big(\int D \Big)^{1/2} \Big(\xi_{\max} \int \|Y\|_F^2 \Big)^{1/2} \leq \frac{b}{2} \int \|Y\|_F^2 + \frac{\xi_{\max}}{2b} \big(\mathcal{E}(t) - \mathcal{E}(t + T_0) \big).$$

Also $|V| \leq K_f \mathcal{E}$, since $\|Y\|_F^2 \leq 2f/m_f$ and $f \leq \mathcal{E}$. Using (21), absorbing the $\|Y\|_F^2$ terms and substituting $\int f = \int \mathcal{E} - \frac{1}{2} \int \|P\|_F^2$ gives

$$\int_t^{t+T_0} \mathcal{E} \;\leq\; C_1 \int_t^{t+T_0} \|P\|_F^2 + C_2 \, \mathcal{E}(t) + C_3 \big(\mathcal{E}(t) - \mathcal{E}(t + T_0) \big).$$

If (39) fails then $\mathcal{E}(t) < 2\mathcal{E}(t + T_0)$, so the last two terms are at most $2(C_2 + C_3)\mathcal{E}(t + T_0)$. Since $\mathcal{E}$ is nonincreasing the left side is at least $T_0 \mathcal{E}(t + T_0)$, and the choices (38) give (40). □

F.3.3 Finite-window concentration

Lemma 7 (Tracked block concentration)**.** *Let $I = [u, u + L_1]$ and suppose that for some $e, \ell, \beta > 0$,*

$$\ell e \le z(t) \le \beta e \qquad (t \in I). \tag{41}$$

If

$$L_1 \;\ge\; \max\Big\{\frac{1}{2\alpha},\; \frac{\nu\beta^2}{\alpha\ell^2}\Big\}, \tag{42}$$

then

$$\int_I \sum_{ij} \frac{A_{ij}\,\bar{a}_i\,\bar{b}_j}{z}\; dt \;\ge\; \frac{\ell^4}{8\nu^2\beta^2}\,L_1 e^2. \tag{43}$$

Proof. Set $F = \sum_{ij} W_{ij}^2$ and $H_W = \sum_{ij} A_{ij} W_{ij}$. Since $\bar{a}_i, \bar{b}_j \ge W_{ij}$, two applications of Cauchy–Schwarz give

$$\int_I \sum_{ij} \frac{A_{ij}\bar{a}_i\bar{b}_j}{z} \;\ge\; \int_I \frac{H_W^2}{\sigma z} \;\ge\; \frac{\big(\int_I H_W\big)^2}{\int_I \sigma z}.$$

From (34), $F' = 2\alpha(H_W - F)$ and $(z^2)' = 2\alpha(\sigma z - z^2)$, giving the exact finite-window identities

$$\int_I H_W = \int_I F + \frac{F(u+L_1) - F(u)}{2\alpha}, \qquad \int_I \sigma z = \int_I z^2 + \frac{z(u+L_1)^2 - z(u)^2}{2\alpha}.$$

Because $F \ge z^2/\nu$ and $F(u) \le z(u)^2$, the hypotheses (41)–(42) imply $\int_I H_W \ge L_1\ell^2 e^2/(2\nu)$ and $\int_I \sigma z \le 2L_1\beta^2 e^2$; the first uses $L_1 \ge \nu\beta^2/(\alpha\ell^2)$ and the second $L_1 \ge 1/(2\alpha)$. Substitution proves (43). □

F.3.4 Tracking over blocks

Set

$$\ell = \alpha c_0 T_0\, \mathrm{e}^{-\alpha T_0}, \quad \beta = 36, \quad \nu = mn, \qquad L_1 = \max\Big\{\frac{1}{2\alpha},\; \frac{\nu\beta^2}{\alpha\ell^2},\; \frac{\log 4}{\alpha}\Big\}, \qquad L = T_0 + L_1. \tag{44}$$

Let $t_k = kL$, $\mathcal{E}_k = \mathcal{E}(t_k)$ and $z_k = z(t_k)$. If $\mathcal{E}_K = 0$ for some K then $\mathcal{E}$, being nonnegative and nonincreasing, vanishes on $[t_K, \infty)$ and every bound below is trivial there; we therefore assume $\mathcal{E}_k > 0$ for all k and set $r_k = z_k/\mathcal{E}_k$. A block is *halving* when $\mathcal{E}_{k+1} \le \mathcal{E}_k/2$, *untracked* when it is not halving but $r_k > 16$, and *controlled* otherwise.

Lemma 8 (Tracking alternative)**.** *On every controlled block, with $e = \mathcal{E}_{k+1}$,*

$$\ell e \le z(t) \le \beta e \qquad (t \in [t_k + T_0,\, t_{k+1}]), \tag{45}$$

so Lemma 7 applies on the trimmed interval. Moreover, among the first K blocks the number U_K of untracked blocks satisfies

$$U_K \;\le\; C_{\mathrm{tr}}\Big(1 + \log\frac{\mathcal{E}_0}{\mathcal{E}_K}\Big), \qquad C_{\mathrm{tr}} = \frac{1}{\log\frac{8}{7}}\Big(1 + \frac{2\log\frac{3}{2}}{\log 2} + \log\tfrac{3}{2}\Big). \tag{46}$$

Proof. Suppose block k is not halving, so $\mathcal{E}_{k+1} > \mathcal{E}_k/2$. Every window $[t - T_0, t] \subset [t_k, t_{k+1}]$ also fails to halve, because $\mathcal{E}(t) \ge \mathcal{E}_{k+1} > \mathcal{E}_k/2 \ge \mathcal{E}(t - T_0)/2$. Hence Lemma 6 and $z' = \alpha(\|P\|_F^2 - z)$ give, for $t \ge t_k + T_0$,

$$z(t) \;\ge\; \alpha \mathrm{e}^{-\alpha T_0} \int_{t-T_0}^{t} \|P(s)\|_F^2\, \mathrm{d}s \;\ge\; \ell\,\mathcal{E}(t) \;\ge\; \ell\,\mathcal{E}_{k+1}.$$

If the block is controlled then $z_k \le 16\mathcal{E}_k$, while $\|P\|_F^2 \le 2\mathcal{E}_k$ throughout, so $z(t) \le z_k + 2\mathcal{E}_k \le 18\mathcal{E}_k < 36\mathcal{E}_{k+1}$, proving (45). Note that ℓ depends on the observability window T_0 and not on the block length L_1: had it been obtained from decay across the whole block it would behave like $\mathrm{e}^{-\alpha L_1}$, while (42) simultaneously demands $L_1 \gtrsim 1/\ell^2$, and the two requirements could not both be met. Separating the two lengths removes the conflict.

For the count, put $q = \mathrm{e}^{-\alpha L} \le 1/4$, which holds because $L \ge L_1 \ge (\log 4)/\alpha$. Since $\|P\|_F^2 \le 2\mathcal{E}_k$ on the block, variation of constants gives

$$z_{k+1} \le q z_k + 2(1-q)\mathcal{E}_k, \qquad \text{hence} \qquad r_{k+1} \le d_k\Big(\frac{r_k}{4} + 2\Big), \qquad d_k = \frac{\mathcal{E}_k}{\mathcal{E}_{k+1}} \ge 1. \tag{47}$$

(a) Non-halving blocks. Here $d_k < 2$, so $r_{k+1} \le \frac{1}{2}r_k + 4$. If $r_k > 16$ then $r_{k+1} \le \frac{3}{4}r_k$; if $r_k \le 16$ then $r_{k+1} \le 12$. Thus a controlled block is followed, unless the next block halves, by another controlled block, so a maximal run of untracked blocks can be preceded only by a halving block. Since $r_0 = 0$, no untracked run begins at $k = 0$, and the number of untracked runs is at most $|\mathcal{H}|$, where $\mathcal{H}$ is the set of halving blocks.

(b) Halving blocks. From (47), $1 + \frac{r_{k+1}}{16} \le 1 + \frac{d_k(r_k+8)}{64} \le d_k\big(1 + \frac{r_k}{16}\big)$, the second inequality reducing to $64 \le 56d_k + 3d_k r_k$, which holds since $d_k \ge 2 \ge 8/7$.

(c) Controlled blocks. By (a), $r_{k+1} \le \max\{r_k, 8\}$.

Now set $\psi_k = \log(1 + r_k/16) \ge 0$, so $\psi_0 = 0$. On an untracked block (a) gives $\psi_{k+1} - \psi_k \le \log \frac{3/4 \cdot 16+16}{16+16} = -\log \frac{8}{7}$, since $r \mapsto \log \frac{1+3r/64}{1+r/16}$ is nonincreasing and is evaluated at worst at $r_k = 16$. On a halving block (b) gives $\psi_{k+1} - \psi_k \le \log d_k$. On a controlled block (c) gives $\psi_{k+1} \le \max\{\psi_k, \log \frac{3}{2}\}$, so ψ can increase only up to the level $\log \frac{3}{2}$ and the net increase over any maximal run of controlled blocks is at most $\log \frac{3}{2}$, however long the run; this is the point at which summing the increases block by block would fail, since ψ does increase on individual controlled blocks. The number of maximal controlled runs is at most $1 + |\mathcal{H}| + (\text{untracked runs}) \le 1 + 2|\mathcal{H}|$. Summing over the first K blocks and using $\psi_K \ge 0 = \psi_0$,

$$U_K \log \tfrac{8}{7} \;\le\; \sum_{k \in \mathcal{H}} \log d_k + \big(1 + 2|\mathcal{H}|\big) \log \tfrac{3}{2}.$$

Finally $\prod_{k<K} d_k = \mathcal{E}_0/\mathcal{E}_K$ with every $d_k \ge 1$, so $\sum_{k \in \mathcal{H}} \log d_k \le \log(\mathcal{E}_0/\mathcal{E}_K)$ and $|\mathcal{H}| \le \log(\mathcal{E}_0/\mathcal{E}_K)/\log 2$, which gives (46). □

F.3.5 The algebraic rate

Define

$$K_\star = \frac{\ell^5 L_1}{8\nu^2\beta^2\alpha\mu}. \tag{48}$$

Lemma 9 (Reciprocal-energy increment)**.** *Let* $\Phi(e) = \dfrac{\beta}{e} + \dfrac{\vartheta}{2e^2}$. *On every controlled block,*

$$\Phi(\mathcal{E}_{k+1}) - \Phi(\mathcal{E}_k) \;\ge\; \frac{K_\star}{8}. \tag{49}$$

Proof. On $I_k = [t_k + T_0, t_{k+1}]$, which has length L_1, Lemma 8 supplies the band (41) with $e = \mathcal{E}_{k+1}$, so Lemma 7 applies. From (36),

$$D = \frac{1}{\alpha\mu} \sum_{ij} \frac{A_{ij}\bar{a}_i\bar{b}_j}{z + \vartheta}.$$

Since $z/(z+\vartheta) \ge \ell\mathcal{E}_{k+1}/(\beta\mathcal{E}_{k+1} + \vartheta)$ on I_k, combining with (43) gives

$$\mathcal{E}_k - \mathcal{E}_{k+1} \;\ge\; K_\star \, \frac{\mathcal{E}_{k+1}^3}{\beta\mathcal{E}_{k+1} + \vartheta}. \tag{50}$$

The block is non-halving, so $\mathcal{E}_k < 2\mathcal{E}_{k+1}$, and for $s \in [\mathcal{E}_{k+1}, \mathcal{E}_k]$,

$$-\Phi'(s) = \frac{\beta}{s^2} + \frac{\vartheta}{s^3} \;\ge\; \frac{\beta\mathcal{E}_{k+1} + \vartheta}{8\mathcal{E}_{k+1}^3}.$$

Integrating this over $[\mathcal{E}_{k+1}, \mathcal{E}_k]$ and using (50) proves (49). □

Theorem 3. Assume (37), $\gamma = 0$ and $R(0) = C(0) = 0$, and let the constants be as in (38)–(48). There are numerical constants $c, C > 0$ such that

$$\mathcal{E}(t) \;\le\; C\left[\mathcal{E}_0\, \mathrm{e}^{-ct/L} + \frac{\beta L}{K_\star(L+t)} + \sqrt{\frac{\vartheta L}{K_\star(L+t)}}\,\right] \qquad (t \ge 0). \tag{51}$$

Consequently $\mathcal{E}(t) = \mathcal{O}(t^{-1})$ when $\epsilon_{\mathrm{stab}} = 0$ and $\mathcal{E}(t) = \mathcal{O}(t^{-1/2})$ when $\epsilon_{\mathrm{stab}} > 0$. All constants depend only on $m, n, m_f, M_f, \alpha, \mu, \epsilon_{\mathrm{stab}}$ and $\mathcal{E}_0$ through the displayed formulas.

Proof of Theorem 3. Consider the first K blocks and let H_K, U_K, M_K count the halving, untracked and controlled ones. Every halving block contributes at least $\log 2$ to $\log(\mathcal{E}_0/\mathcal{E}_K)$, and Lemma 8 bounds U_K. If $H_K + U_K \geq K/2$ these two facts force $\log(\mathcal{E}_0/\mathcal{E}_K) \gtrsim K$, hence $\mathcal{E}_K \leq C\mathcal{E}_0 \mathrm{e}^{-cK}$ for numerical $c, C > 0$.

Otherwise $M_K \geq K/2$. Since $\mathcal{E}$ is nonincreasing and Φ is decreasing, $\Phi(\mathcal{E}_k)$ is nondecreasing in k, so summing Lemma 9 over the controlled blocks yields $\Phi(\mathcal{E}_K) \geq K_\star K/16$. At least one of the two terms of Φ is at least half of this, whence

$$\mathcal{E}_K \;\leq\; \max\Big\{\frac{32\beta}{K_\star K},\, 4\sqrt{\frac{\vartheta}{K_\star K}}\Big\}.$$

To pass to continuous time fix $t \geq 0$ and put $K = \lfloor t/L \rfloor$. If $K = 0$ then $t < L$, so $\mathrm{e}^{-ct/L} \geq \mathrm{e}^{-c}$ and $\mathcal{E}(t) \leq \mathcal{E}_0 \leq \mathrm{e}^{c}\mathcal{E}_0 \mathrm{e}^{-ct/L}$, which is absorbed by enlarging C. If $K \geq 1$ then $t_K = KL \leq t$, so $\mathcal{E}(t) \leq \mathcal{E}_K$; moreover $\lfloor x \rfloor \geq x/2$ for $x \geq 1$ gives $K \geq t/(2L)$, and $t \geq L$ gives $L + t \leq 2t$, so $1/K \leq 2L/t \leq 4L/(L+t)$. Substituting turns the three alternatives into the three displayed terms of (51) after renaming numerical constants. □

Remark 4 (Scope)**.** *The estimates are upper bounds and do not preclude faster decay along particular trajectories. Formal averaging predicts the sharper equivalents*

$$\mathcal{E}(t) \sim \frac{\alpha\mu}{t} \quad (\epsilon_{\mathrm{stab}} = 0), \qquad \mathcal{E}(t) \sim \alpha\mu\sqrt{\frac{\epsilon_{\mathrm{stab}}}{2t}} \quad (\epsilon_{\mathrm{stab}} > 0),$$

with the crossover between the two regimes at $\mathcal{E} \approx \alpha\mu\,\epsilon_{\mathrm{stab}}$; proving these equivalents is a separate averaging problem. Section 3.4 confirms both exponents, the crossover scale and the predicted dependence on α, μ and ϵ_{stab} numerically.

F.4 Discrete Convergence Analysis for the BACD Splitting

We analyze the discrete-time scheme defined by the BACD composition with the symmetric half-damp Sub-flow C (Table 3). The one-step update from (X_n, P_n, R_n, C_n) to $(X_{n+1}, P_{n+1}, R_{n+1}, C_{n+1})$ is

$$P^{(B)} = P_n - h\nabla f(X_n), \tag{52}$$

$$X_{n+1} = X_n + hP^{(B)}, \tag{53}$$

$$\tilde{\xi}_n = \frac{R_n C_n^\top}{\mathbf{1}_m^\top R_n + \epsilon_{\mathrm{stab}}},$$

$$P^{(1)} = P^{(B)} \odot \exp\Big(-\frac{h}{2}\,\tilde{\xi}_n\Big), \tag{54}$$

$$R_{n+1} = \mathrm{e}^{-\alpha h} R_n + \frac{1 - \mathrm{e}^{-\alpha h}}{\mu\alpha}\,[P^{(1)}]^2 \mathbf{1}_n, \tag{55}$$

$$C_{n+1} = \mathrm{e}^{-\alpha h} C_n + \frac{1 - \mathrm{e}^{-\alpha h}}{\mu\alpha}\,\mathbf{1}_m^\top [P^{(1)}]^2, \tag{56}$$

$$\tilde{\xi}_{n+1} = \frac{R_{n+1} C_{n+1}^\top}{\mathbf{1}_m^\top R_{n+1} + \epsilon_{\mathrm{stab}}},$$

$$P^{(2)} = P^{(1)} \odot \exp\Big(-\frac{h}{2}\,\tilde{\xi}_{n+1}\Big), \tag{57}$$

$$P_{n+1} = P^{(2)} \odot \mathrm{e}^{-\gamma h}. \tag{58}$$

All exponentials act element-wise.

F.4.1 Equilibria of the discrete dynamics

Proposition 10 (Equilibria preservation)**.** *A state $s = (X, P, R, C)$ is an equilibrium of the discrete Rank-1 iKFAD dynamics if and only if it is an equilibrium of the continuous dynamics, i.e., $P = 0$, $\nabla f(X) = 0$, and $R = C = 0$.*

Proof. Let $s_n = (X_n, P_n, R_n, C_n)$ be an equilibrium ($s_{n+1} = s_n$). From (53), $X_{n+1} = X_n + hP^{(B)} = X_n \implies P^{(B)} = 0$. Since $P_{n+1} = P_n = 0$, step (52) gives $0 = 0 - h\nabla f(X_n) \implies \nabla f(X_n) = 0$. Step (54) gives $P^{(1)} = 0$, and step (55) becomes $R_{n+1} = \mathrm{e}^{-\alpha h} R_n$. Because $R_{n+1} = R_n$, $\mathrm{e}^{-\alpha h} R_n = R_n \implies R_n = 0$ (since $\alpha h > 0$). Similarly $C_n = 0$. The converse is immediate. □

F.4.2 Exponential convergence of the discrete scheme

Theorem 4. Assume $f \in C^2$ satisfies $m_f I \preceq \nabla^2 f(X) \preceq MI$ for all X with $0 < m_f \leq M$; in particular f then satisfies (21) with $a = 1$ and $b = m_f/2$, since m_f-strong convexity gives $\langle X - X^*, \nabla f(X)\rangle_F \geq f(X) - f(X^*) + \frac{m_f}{2}\|X - X^*\|_F^2$. Let $\gamma > 0$ and $\alpha, \mu, \epsilon_{\text{stab}} > 0$, and fix $L > 0$. Then for any initial condition (X_0, P_0, R_0, C_0) with

$$\|X_0\|_F + \|P_0\|_F + \|R_0\| + \|C_0\| \leq L \qquad \text{and} \qquad R_0 \geq 0,\ \ C_0 \geq 0 \ \text{ entrywise},$$

there exist $h^* > 0$, $\kappa > 0$, and $C_\star > 0$ (all depending only on $m_f, M, \gamma, \alpha, \mu, \epsilon_{\text{stab}}, L$ and the dimensions m, n, and *not* on h) such that for all $h \in (0, h^*)$ and all $n \geq 0$,

$$f(X_n) - f(X^*) + \|P_n\|_F^2 + \|R_n\|^2 + \|C_n\|^2 \leq C_\star \mathrm{e}^{-\kappa n h}.$$

Proof of Theorem 4. Translate coordinates so that $X^* = 0$, $f(0) = 0$ and $\nabla f(0) = 0$, and abbreviate

$$g_n := \nabla f(X_n), \qquad \Sigma_n := \|P_n\|_F^2 + \|X_n\|_F^2, \qquad \Lambda := 1 + 2M^2, \qquad r_n := \mathbf{1}_m^\top R_n, \qquad c_n := \mathbf{1}_n^\top C_n.$$

Introduce the perturbed discrete Lyapunov function

$$\mathcal{W}_\epsilon(X, P, R, C) = f(X) + \frac{1}{2}\|P\|_F^2 + \frac{\mu B}{2}\big(\|R\|^2 + \|C\|^2\big) + \epsilon\langle X, P\rangle_F + \frac{\epsilon}{2}\|X\|_F^2, \tag{59}$$

where now

$$\boxed{B \in (0, 1]}$$

and $\epsilon \in (0, \frac{1}{4}]$ are parameters fixed below. (We fix the weight of $\|P\|_F^2$ to 1 to cancel the gradient cross term; this corresponds to $A = 1$ in the generic form. The restriction $B \geq 1$ of the earlier version is replaced by $B \leq 1$: as the coefficient of $\|R_n\|^2$ below shows, nothing in the argument pushes B up, and c_2 pushes it down.) The $\frac{\mu B}{2}(\|R\|^2 + \|C\|^2)$ term is unaffected by the cross term, so for $\epsilon \leq 1/4$ the same Young estimate as in (23) gives

$$\mathcal{W}_\epsilon \geq f(X) + \frac{1}{4}\|P\|_F^2 + \frac{\mu B}{2}(\|R\|^2 + \|C\|^2) + \frac{\epsilon}{4}\|X\|_F^2, \tag{60}$$

$$\mathcal{W}_\epsilon \leq f(X) + \frac{3}{4}\|P\|_F^2 + \frac{\mu B}{2}(\|R\|^2 + \|C\|^2) + \frac{3\epsilon}{2}\|X\|_F^2. \tag{61}$$

Note that both bounds carry the factor B; dropping it from (61) would invalidate the comparison used in the contraction argument. In particular $\mathcal{W}_\epsilon \geq 0$.

We prove by induction on n that

$$\mathcal{W}_\epsilon(X_{n+1}, P_{n+1}, R_{n+1}, C_{n+1}) \leq (1 - \rho h)\, \mathcal{W}_\epsilon(X_n, P_n, R_n, C_n) \tag{62}$$

for a constant $\rho > 0$ and every $h \in (0, h^*)$. The induction hypothesis is the *history* statement

$$\mathcal{W}_\epsilon(X_k, P_k, R_k, C_k) \leq \mathcal{W}_R \quad \text{for all } k \leq n, \mathcal{W}_R := f(X_0) + \tfrac{3}{4}\|P_0\|_F^2 + \tfrac{\mu}{2}\big(\|R_0\|^2 + \|C_0\|^2\big) + \tfrac{3}{8}\|X_0\|_F^2, \tag{63}$$

which holds at $n = 0$ by (61) together with $B \leq 1$ and $\epsilon \leq \frac{1}{4}$; note $\mathcal{W}_R \leq L^2\big(\frac{M}{2} + \frac{\mu}{2} + \frac{9}{8}\big)$, so $\mathcal{W}_R$ is controlled by L alone and, crucially, is *independent of* B, ϵ *and* h. Since (62) with $\rho h \leq 1$ gives $\mathcal{W}_\epsilon(n+1) \leq \mathcal{W}_\epsilon(n)$, establishing (62) at step n propagates (63) to $n + 1$ and closes the induction.

From $m_f I \preceq \nabla^2 f \preceq MI$ and $\nabla f(0) = 0$, $f(0) = 0$,

$$\|g_n\|_F \leq M\|X_n\|_F, \qquad \langle X_n, g_n\rangle_F \geq f(X_n) + \tfrac{m_f}{2}\|X_n\|_F^2 \ \geq 0, \qquad \tfrac{m_f}{2}\|X_n\|_F^2 \leq f(X_n) \leq \tfrac{M}{2}\|X_n\|_F^2. \tag{64}$$

We use repeatedly, for $x \geq 0$,

$$\text{(E1) } 1 - \mathrm{e}^{-x} \leq x, \qquad \text{(E2) } 1 - \mathrm{e}^{-x} \geq x\,\mathrm{e}^{-x}, \qquad \text{(E3) } 0 \leq \mathrm{e}^{-x} - 1 + x \leq \tfrac{x^2}{2}, \tag{65}$$

and the two structural facts

$$\text{(E4) } 0 \leq v \implies \|v\|_2 \leq \|v\|_1 \ \text{ and } \ v_i \leq \mathbf{1}^\top v, \qquad \text{(E5) } \big\|[P]^2\mathbf{1}_n\big\|_2 \leq \big\|[P]^2\mathbf{1}_n\big\|_1 = \|P\|_F^2. \tag{66}$$

(E5) holds because $[P]^2\mathbf{1}_n$ is entrywise nonnegative and its entries sum to $\sum_{ij} P_{ij}^2$; it is sharp, with equality when a single row of P is nonzero. It is the same estimate used in Lemma 5, and it is what keeps every constant below free of any dimension factor.

R_{n+1} in (55) is the sum of $\mathrm{e}^{-\alpha h}R_n$ and a nonnegative vector, so $R_0 \geq 0$ gives $R_n \geq 0$ for all n by induction, and likewise $C_n \geq 0$. Hence $\mathbf{1}_m^\top R_n + \epsilon_{\mathrm{stab}} \geq \epsilon_{\mathrm{stab}} > 0$, $\tilde{\xi}_n \geq 0$ entrywise, every element-wise factor in (54), (57), (58) lies in $(0, 1]$, and therefore

$$\|P^{(1)}\|_F \leq \|P^{(B)}\|_F, \qquad \|P_{n+1}\|_F^2 \leq \mathrm{e}^{-2\gamma h}\|P^{(B)}\|_F^2. \tag{67}$$

Moreover, by (E4), $R_{n,i} \leq \mathbf{1}_m^\top R_n \leq \mathbf{1}_m^\top R_n + \epsilon_{\mathrm{stab}}$, so

$$0 \;\leq\; \tilde{\xi}_{n,ij} \;=\; \frac{R_{n,i}}{\mathbf{1}_m^\top R_n + \epsilon_{\mathrm{stab}}}\, C_{n,j} \;\leq\; C_{n,j} \;\leq\; \|C_n\|_\infty. \tag{68}$$

This gives a bound on $\tilde{\xi}$ that does *not* blow up as $\epsilon_{\mathrm{stab}} \downarrow 0$; both branches are retained in the a priori bounds below.

By (60) and (64), the induction hypothesis (63) yields, for every $k \leq n$,

$$\|P_k\|_F \leq P_{\max} := 2\sqrt{\mathcal{W}_R}, \qquad \|X_k\|_F \leq X_{\max} := \sqrt{2\mathcal{W}_R/m_f}, \tag{69}$$

both independent of ϵ, B and h. Consequently, for $h \leq 1$ and every $k \leq n$,

$$\|P_k^{(1)}\|_F \leq \|P_k^{(B)}\|_F \leq \|P_k\|_F + h\|g_k\|_F \leq P_{\max} + MX_{\max} =: \bar{P}. \tag{70}$$

Contracting (55) with $\mathbf{1}_m$ and (56) with $\mathbf{1}_n$, and using the identity $\mathbf{1}_m^\top\big([P^{(1)}]^2\mathbf{1}_n\big) = \|P^{(1)}\|_F^2 = \big(\mathbf{1}_m^\top[P^{(1)}]^2\big)\mathbf{1}_n$, gives the *exact* scalar recursions

$$r_{k+1} = \mathrm{e}^{-\alpha h}r_k + \big(1 - \mathrm{e}^{-\alpha h}\big)\frac{\|P_k^{(1)}\|_F^2}{\mu\alpha}, \qquad c_{k+1} = \mathrm{e}^{-\alpha h}c_k + \big(1 - \mathrm{e}^{-\alpha h}\big)\frac{\|P_k^{(1)}\|_F^2}{\mu\alpha}. \tag{71}$$

Each right-hand side is a *convex combination* of r_k (resp. c_k) and $\|P_k^{(1)}\|_F^2/(\mu\alpha) \leq \bar{P}^2/(\mu\alpha)$. Hence the interval $[0, \max\{r_0, \bar{P}^2/(\mu\alpha)\}]$ is forward invariant, and by (70) and (E4) we obtain, *for every* $k \leq n+1$,

$$\begin{aligned} \|R_k\|_2 \leq \|R_k\|_1 = r_k \leq R_{\max} &:= \max\Big\{r_0, \frac{\bar{P}^2}{\mu\alpha}\Big\}, \\ \|C_k\|_2 \leq \|C_k\|_1 = c_k \leq C_{\max} &:= \max\Big\{c_0, \frac{\bar{P}^2}{\mu\alpha}\Big\}, \end{aligned} \tag{72}$$

and, combining (68) with the elementary bound $\tilde{\xi}_{k,ij} \leq R_{\max}C_{\max}/\epsilon_{\mathrm{stab}}$,

$$0 \;\leq\; \tilde{\xi}_{k,ij} \;\leq\; \xi_{\max} := \min\Big\{C_{\max},\; \frac{R_{\max}C_{\max}}{\epsilon_{\mathrm{stab}}}\Big\} \qquad \text{for all } k \leq n+1. \tag{73}$$

The index range $k \leq n+1$ in (72)–(73) is essential and is *not* available from a one-step hypothesis: the estimate of the ϵ-terms must bound $\tilde{\xi}_{n+1}$, which depends on R_{n+1}, C_{n+1}, and these are not controlled by $\mathcal{W}_\epsilon(n) \leq \mathcal{W}_R$. The forward-invariance argument supplies exactly that bound, using only (69) at indices $k \leq n$; it is therefore not circular. (Both halves of (69) are used, since $\bar{P}$ involves $X_{\max}$ as well as $P_{\max}$.)

From (52), $\|P^{(B)}\|_F^2 = \|P_n\|_F^2 - 2h\langle P_n, g_n\rangle_F + h^2\|g_n\|_F^2$, while $2|\langle P_n, g_n\rangle_F| \leq \|P_n\|_F^2 + \|g_n\|_F^2 \leq \|P_n\|_F^2 + M^2\|X_n\|_F^2$ and $h^2\|g_n\|_F^2 \leq hM^2\|X_n\|_F^2$ for $h \leq 1$. Hence, for all $h \in (0, 1]$,

$$\|P_n\|_F^2 - h\Lambda\Sigma_n \;\leq\; \|P^{(B)}\|_F^2 \;\leq\; \|P_n\|_F^2 + h\Lambda\Sigma_n \;\leq\; (1+\Lambda)\Sigma_n, \qquad \Lambda = 1 + 2M^2. \tag{74}$$

This is the only device needed to convert every bound stated in terms of $\|P^{(B)}\|_F^2$ into one in terms of $\|P_n\|_F^2$ plus an h^2 remainder carried by Σ_n.

The descent lemma with $X_{n+1} = X_n + hP^{(B)}$, together with $\langle g_n, P^{(B)}\rangle_F = \langle g_n, P_n\rangle_F - h\|g_n\|_F^2$ and (74), gives

$$\begin{aligned} f(X_{n+1}) - f(X_n) \;&\leq\; h\langle g_n, P_n\rangle_F - h^2\|g_n\|_F^2 + \frac{Mh^2}{2}\|P^{(B)}\|_F^2 \\ &\leq\; h\langle g_n, P_n\rangle_F - h^2\|g_n\|_F^2 + \frac{M(1+\Lambda)}{2}h^2\Sigma_n. \end{aligned} \tag{75}$$

No term is hidden in an unnamed $\mathcal{O}(h^2)$: the constant is $M(1+\Lambda)/2 = M(1+M^2)$.

By (67),

$$\begin{aligned}\tfrac{1}{2}\|P_{n+1}\|_F^2 - \tfrac{1}{2}\|P_n\|_F^2 \;\le\; -\frac{1-\mathrm{e}^{-2\gamma h}}{2}\|P^{(B)}\|_F^2 \\ + \underbrace{\Big(\tfrac{1}{2}\|P^{(B)}\|_F^2 - \tfrac{1}{2}\|P_n\|_F^2\Big)}_{=\,-h\langle P_n, g_n\rangle_F + \frac{h^2}{2}\|g_n\|_F^2}.\end{aligned}$$

By (E2) of (65) with $x = 2\gamma h$ one has $\frac{1}{2}(1-\mathrm{e}^{-2\gamma h}) \ge \gamma h\,\mathrm{e}^{-2\gamma h}$; combining with the lower half of (74) and $\mathrm{e}^{-2\gamma h} \le 1$,

$$\begin{aligned}\tfrac{1}{2}\|P_{n+1}\|_F^2 - \tfrac{1}{2}\|P_n\|_F^2 \;\le\; -\gamma\,\mathrm{e}^{-2\gamma h}\,h\,\|P_n\|_F^2 + \gamma\Lambda h^2\Sigma_n \\ - h\langle P_n, g_n\rangle_F + \tfrac{h^2}{2}\|g_n\|_F^2.\end{aligned} \tag{76}$$

Adding (75) and (76), the terms $+h\langle g_n, P_n\rangle_F$ and $-h\langle P_n, g_n\rangle_F$ cancel identically — this is precisely why the weight of $\|P\|_F^2$ in (59) is pinned to 1 — and the surviving gradient terms combine to $-h^2\|g_n\|_F^2 + \frac{h^2}{2}\|g_n\|_F^2 = -\frac{h^2}{2}\|g_n\|_F^2 \le 0$, which we discard. Therefore

$$\Delta f + \tfrac{1}{2}\Delta\|P\|_F^2 \;\le\; -\gamma\,\mathrm{e}^{-2\gamma h}\,h\,\|P_n\|_F^2 + K_{12}\,h^2\Sigma_n, \qquad K_{12} := M(1+M^2) + \gamma(1+2M^2). \tag{77}$$

Write (55) as $R_{n+1} = \mathrm{e}^{-\alpha h}R_n + S_n$ with

$$S_n := \frac{1-\mathrm{e}^{-\alpha h}}{\mu\alpha}\,[P^{(1)}]^2\mathbf{1}_n \;\ge\; 0.$$

Expanding exactly,

$$\|R_{n+1}\|^2 = \mathrm{e}^{-2\alpha h}\|R_n\|^2 + 2\mathrm{e}^{-\alpha h}\langle R_n, S_n\rangle + \|S_n\|^2, \qquad \|S_n\| = \Theta(h), \tag{78}$$

so the cross term is a genuinely positive quantity of order h. Apply Young, $2\langle R_n, S_n\rangle \le \sigma\|R_n\|^2 + \sigma^{-1}\|S_n\|^2$, with the *$h$-proportional* scale

$$\boxed{\sigma := 1 - \mathrm{e}^{-\alpha h}}$$

The two resulting coefficients telescope exactly, with no $\mathcal{O}(h^2)$ slack:

$$\mathrm{e}^{-2\alpha h} + \sigma\,\mathrm{e}^{-\alpha h} = \mathrm{e}^{-2\alpha h} + \mathrm{e}^{-\alpha h} - \mathrm{e}^{-2\alpha h} = \mathrm{e}^{-\alpha h}, \qquad 1 + \frac{\mathrm{e}^{-\alpha h}}{1-\mathrm{e}^{-\alpha h}} = \frac{1}{1-\mathrm{e}^{-\alpha h}}.$$

Hence, using (E5), then (E1), then $\|P^{(1)}\|_F^4 \le \bar{P}^2\|P^{(1)}\|_F^2$ and $\|P^{(1)}\|_F \le \|P^{(B)}\|_F$ from (67)–(70),

$$\begin{aligned}\|R_{n+1}\|^2 \le \mathrm{e}^{-\alpha h}\|R_n\|^2 + \frac{\|S_n\|^2}{1-\mathrm{e}^{-\alpha h}} \;\le\; \mathrm{e}^{-\alpha h}\|R_n\|^2 + \frac{1-\mathrm{e}^{-\alpha h}}{\alpha^2\mu^2}\|P^{(1)}\|_F^4 \\ \le\; \mathrm{e}^{-\alpha h}\|R_n\|^2 + \frac{\bar{P}^2}{\alpha\mu^2}\,h\,\|P^{(B)}\|_F^2,\end{aligned} \tag{79}$$

$$\|C_{n+1}\|^2 \le \mathrm{e}^{-\alpha h}\|C_n\|^2 + \frac{\bar{P}^2}{\alpha\mu^2}\,h\,\|P^{(B)}\|_F^2. \tag{80}$$

The remainder in (79)–(80) *is of order h, not h^2*, and it is displayed as such; no choice of σ can make it $\mathcal{O}(h^2)$. Its decay factor is $\mathrm{e}^{-\alpha h}$, not $2\mathrm{e}^{-2\alpha h}$: it tends to 1 from below, which is what makes contraction possible. Subtracting $\|R_n\|^2$, using (E2) in the form $\mathrm{e}^{-\alpha h} - 1 \le -\alpha h\,\mathrm{e}^{-\alpha h}$, weighting by $\frac{\mu B}{2}$, adding the C analog and applying (74),

$$\begin{aligned}\frac{\mu B}{2}\Delta\big(\|R\|^2 + \|C\|^2\big) \le -\frac{\mu B\alpha\,\mathrm{e}^{-\alpha h}}{2}\,h\big(\|R_n\|^2 + \|C_n\|^2\big) \\ + \frac{B\bar{P}^2}{\alpha\mu}\,h\,\|P_n\|_F^2 + K_3 h^2\Sigma_n,\end{aligned} \tag{81}$$

where $K_3 := B\bar{P}^2\Lambda/(\alpha\mu)$.

Write $P_{n+1} = \mathrm{e}^{-\gamma h}\,Q \odot P^{(B)}$ with $Q = \exp\big(-\frac{h}{2}(\tilde{\xi}_n + \tilde{\xi}_{n+1})\big)$ and $\Delta_Q := Q - \mathbf{1}$. By the invariant cone, (73) and (E1),

$$\Delta_{Q,ij} \in [-h\xi_{\max},\,0], \qquad \text{hence} \qquad \|\Delta_Q \odot P^{(B)}\|_F \le h\xi_{\max}\|P^{(B)}\|_F, \qquad \langle P^{(B)}, \Delta_Q \odot P^{(B)}\rangle_F \le 0. \tag{82}$$

Using $X_{n+1} = X_n + hP^{(B)}$ and $\langle X_n, P^{(B)}\rangle_F = \langle X_n, P_n\rangle_F - h\langle X_n, g_n\rangle_F$, the two ϵ-blocks are, *exactly*,

$$\begin{aligned}\langle X_{n+1}, P_{n+1}\rangle_F = \mathrm{e}^{-\gamma h}\Big[&\langle X_n, P_n\rangle_F - h\langle X_n, g_n\rangle_F \\ &+ \langle X_n, \Delta_Q \odot P^{(B)}\rangle_F + h\langle P^{(B)}, Q\odot P^{(B)}\rangle_F\Big],\\ \tfrac{1}{2}\|X_{n+1}\|_F^2 - \tfrac{1}{2}\|X_n\|_F^2 &= h\langle X_n, P_n\rangle_F - h^2\langle X_n, g_n\rangle_F + \tfrac{h^2}{2}\|P^{(B)}\|_F^2.\end{aligned}$$

We estimate the five resulting groups, each explicitly.

(a) The $\langle X_n, P_n\rangle_F$ terms combine with coefficient $\epsilon(\mathrm{e}^{-\gamma h} - 1 + h) = \epsilon h(1-\gamma) + \epsilon\,\varrho_h$, where $\varrho_h := \mathrm{e}^{-\gamma h} - 1 + \gamma h$ satisfies $0 \le \varrho_h \le \frac{\gamma^2 h^2}{2}$ by (E3). Young with $\eta > 0$ gives $\epsilon h(1-\gamma)\langle X_n, P_n\rangle_F \le \epsilon h|1-\gamma|\big(\eta\|X_n\|_F^2 + \frac{1}{4\eta}\|P_n\|_F^2\big)$, and $|\epsilon\varrho_h\langle X_n, P_n\rangle_F| \le \epsilon\frac{\gamma^2}{4}h^2\Sigma_n$.

(b) By (64) and $\mathrm{e}^{-\gamma h} > 0$, $-\epsilon h\mathrm{e}^{-\gamma h}\langle X_n, g_n\rangle_F \le -\epsilon h\mathrm{e}^{-\gamma h}\big(f(X_n) + \frac{m_f}{2}\|X_n\|_F^2\big)$. Writing $-\mathrm{e}^{-\gamma h} = -1 + (1 - \mathrm{e}^{-\gamma h})$, using (E1) and $f(X_n) + \frac{m_f}{2}\|X_n\|_F^2 \le \frac{M+m_f}{2}\|X_n\|_F^2 \le M\Sigma_n$, this is

$$\le\ -\epsilon h f(X_n) - \epsilon h\tfrac{m_f}{2}\|X_n\|_F^2 + \epsilon\gamma M h^2\Sigma_n.$$

This group is the sole source of c_1 and of the negative $\|X\|_F^2$ coefficient.

(c) By Cauchy–Schwarz, (82), Young with $\delta > 0$ and (74),

$$\begin{aligned}\epsilon\mathrm{e}^{-\gamma h}\langle X_n, \Delta_Q\odot P^{(B)}\rangle_F &\le \epsilon h\xi_{\max}\Big(\delta\|X_n\|_F^2 + \tfrac{1}{4\delta}\|P^{(B)}\|_F^2\Big)\\ &\le \epsilon\xi_{\max}\delta\, h\|X_n\|_F^2 + \tfrac{\epsilon\xi_{\max}}{4\delta}h\|P_n\|_F^2 + \tfrac{\epsilon\xi_{\max}\Lambda}{4\delta}h^2\Sigma_n.\end{aligned}$$

(d) Since $Q_{ij} \in (0,1]$, $0 \le h\langle P^{(B)}, Q\odot P^{(B)}\rangle_F \le h\|P^{(B)}\|_F^2$, giving at most $\epsilon h\|P_n\|_F^2 + \epsilon\Lambda h^2\Sigma_n$ by (74).

(e) $-\epsilon h^2\langle X_n, g_n\rangle_F \le 0$ is discarded by (64), and $\frac{\epsilon h^2}{2}\|P^{(B)}\|_F^2 \le \frac{\epsilon(1+\Lambda)}{2}h^2\Sigma_n$.

Summing (a)–(e),

$$\begin{aligned}\epsilon\Delta\Big[\langle X, P\rangle_F + \tfrac{1}{2}\|X\|_F^2\Big] \le &-\epsilon h f(X_n) - \epsilon h\Big(\tfrac{m_f}{2} - |1-\gamma|\eta - \xi_{\max}\delta\Big)\|X_n\|_F^2\\ &+ \epsilon h\Big(1 + \tfrac{|1-\gamma|}{4\eta} + \tfrac{\xi_{\max}}{4\delta}\Big)\|P_n\|_F^2 + \epsilon K_{45}h^2\Sigma_n,\end{aligned} \tag{83}$$

with $K_{45} := \frac{\gamma^2}{4} + \gamma M + \frac{\xi_{\max}\Lambda}{4\delta} + \Lambda + \frac{1+\Lambda}{2}$.

Adding (77), (81) and (83) — these five blocks exhaust $\mathcal{W}_\epsilon$ — we obtain, for every $h \in (0,1]$,

$$\begin{aligned}\mathcal{W}_\epsilon(n+1) - \mathcal{W}_\epsilon(n) \le &-h\Big(c_1 f(X_n) + c_2\|P_n\|_F^2 + c_3\|R_n\|^2 + c_4\|C_n\|^2 + c_5\|X_n\|_F^2\Big)\\ &+ C_2 h^2\big(\|P_n\|_F^2 + \|X_n\|_F^2\big), \qquad C_2 := K_{12} + K_3 + \epsilon K_{45},\end{aligned} \tag{84}$$

where the coefficients are given *exactly* by

$$\begin{aligned}c_1 &= \epsilon,\\ c_2 &= \underbrace{\gamma\,\mathrm{e}^{-2\gamma h}}_{\text{cross term}} - \underbrace{\frac{B\bar{P}^2}{\alpha\mu}}_{\substack{\text{the } \Theta(h)\\ \text{auxiliary leak}}} - \epsilon\underbrace{\Big(1 + \frac{|1-\gamma|}{4\eta} + \frac{\xi_{\max}}{4\delta}\Big)}_{\text{the } \epsilon\text{-terms}},\\ c_3 &= c_4 = \frac{\mu B\alpha}{2}\,\mathrm{e}^{-\alpha h},\\ c_5 &= \epsilon\Big(\frac{m_f}{2} - |1-\gamma|\eta - \xi_{\max}\delta\Big).\end{aligned} \tag{85}$$

Two structural features of (84) should be noted. First, the h^2 remainder carries only $\Sigma_n = \|P_n\|_F^2 + \|X_n\|_F^2$ — *not* $f(X_n)$, $\|R_n\|^2$ or $\|C_n\|^2$; every h^2 term traces back through (74) to Σ_n. Second, C_2 is a finite constant depending only on $M, \gamma, \alpha, \mu, B, \bar{P}, \xi_{\max}, m_f, \epsilon$, all of which are fixed before h.

Every contribution of $\|R_n\|^2$ to (84) originates in the single summand $\frac{\mu B}{2}\|R\|^2$ of (59): the f-block contains no R; the P-block estimate (67) discards Q altogether; and the ϵ-block sees R, C only through $\tilde{\xi}$, which enters (83) bounded by the *constant* $\xi_{\max}$ and charged entirely to $\|X_n\|_F^2$ and $\|P_n\|_F^2$. Consequently

$$c_3 = c_4 = B\,\vartheta, \qquad \vartheta := \frac{\mu\alpha}{2}\mathrm{e}^{-\alpha h} > 0 \ \text{ independent of } B, \qquad \text{so} \qquad \operatorname{sign} c_3 = \operatorname{sign} B. \tag{86}$$

Hence $c_3 = c_4 > 0$ for *every* $B > 0$: with the corrected estimate (79) the requirement "$c_3, c_4 > 0$" is vacuous and imposes no constraint on B whatsoever. Equivalently, the *relative* decay rate of the auxiliary block,

$$\frac{c_3}{\mu B/2} = \alpha\,\mathrm{e}^{-\alpha h},$$

is independent of B, because rescaling the auxiliary weight rescales the decay and the weight identically; B therefore does not appear in the final rate at all. The only genuine constraint on B is the *upper* bound coming from c_2, since $\partial_B c_2 = -\bar{P}^2/(\alpha\mu) < 0$ and $\bar{P}$ does not depend on B (see the a priori bounds). This settles the matter: B is an upper-constrained free parameter, exactly like ϵ, and the requirements $\{B > 0\}$ and $\{B < \gamma\alpha\mu/\bar{P}^2\}$ are compatible for all $\gamma, \alpha, \mu > 0$. There is no $\|R\|^2$-versus-$\|P\|_F^2$ trade-off to balance.

The constants are fixed in the following strictly acyclic order, each depending only on those already fixed:

$$\mathcal{W}_R \;\to\; (P_{\max}, X_{\max}, \bar{P}) \;\to\; (R_{\max}, C_{\max}, \xi_{\max}) \;\to\; B \;\to\; (\eta, \delta) \;\to\; \epsilon \;\to\; h_1 \;\to\; C_2 \;\to\; h^* \;\to\; \rho.$$

(1) $\mathcal{W}_R$, $P_{\max}$, $X_{\max}$, $\bar{P}$, $R_{\max}$, $C_{\max}$, $\xi_{\max}$ as in (63), (69), (70), (72), (73). These are free of B, ϵ and h, because $\mathcal{W}_R$ was formed using only the a priori restrictions $B \le 1$ and $\epsilon \le \frac{1}{4}$.

(2) *The weight B:*

$$\boxed{B := \min\Big\{1,\ \frac{\gamma\alpha\mu}{2\bar{P}^2}\Big\} \in (0, 1]} \qquad \Longrightarrow \qquad \frac{B\bar{P}^2}{\alpha\mu} \le \frac{\gamma}{2}.$$

This is well defined and strictly positive for all $\gamma, \alpha, \mu > 0$, and is not circular because $\bar{P}$ was fixed in item (1). If $\bar{P} = 0$ the second entry is vacuous; in that case $P_n = 0$ and $X_n = 0$ for all n, so the (X, P) subsystem already sits at equilibrium and only R, C decay — take $B = 1$, for which the $\|P_n\|_F^2$ leak is identically zero.

(3) *The Young parameters:*

$$\boxed{\eta := \frac{m_f}{8|1-\gamma|}\ \ (\gamma \neq 1;\ \text{the term is absent if } \gamma = 1), \qquad \delta := \frac{m_f}{8\xi_{\max}}\ \ (\text{absent if } \xi_{\max} = 0).}$$

Then $|1-\gamma|\eta = \xi_{\max}\delta = \frac{m_f}{8}$, so each consumes exactly one quarter of the strong-convexity budget $\frac{m_f}{2}$, and $\frac{|1-\gamma|}{4\eta} = \frac{2(1-\gamma)^2}{m_f}$, $\frac{\xi_{\max}}{4\delta} = \frac{2\xi_{\max}^2}{m_f}$. Consequently

$$c_5 = \epsilon\Big(\frac{m_f}{2} - \frac{m_f}{8} - \frac{m_f}{8}\Big) = \frac{\epsilon\, m_f}{4}, \qquad c_2 = \gamma\mathrm{e}^{-2\gamma h} - \frac{B\bar{P}^2}{\alpha\mu} - \epsilon\,\Xi, \qquad \Xi := 1 + \frac{2(1-\gamma)^2}{m_f} + \frac{2\xi_{\max}^2}{m_f},$$

and $K_{45} = \frac{\gamma^2}{4} + \gamma M + \frac{2\xi_{\max}^2(1+2M^2)}{m_f} + 2 + 3M^2$.

(4) *The perturbation size:*

$$\boxed{\epsilon := \min\Big\{\frac{1}{4},\ \frac{\gamma}{8\,\Xi}\Big\}} \qquad \Longrightarrow \qquad \epsilon\,\Xi \le \frac{\gamma}{8}.$$

Legitimate and non-circular: Ξ depends on $m_f, \gamma, \xi_{\max}$ only, and $\xi_{\max}$ is ϵ-free by item (1). The bound $\epsilon \le \frac{1}{4}$ is what (60)–(61) require.

(5) *The first step restriction:*

$$\boxed{h_1 := \min\Big\{1,\ \frac{1}{2\gamma}\ln\frac{4}{3}\Big\} > 0} \qquad \Longrightarrow \qquad \mathrm{e}^{-2\gamma h} \ge \tfrac{3}{4} \ \text{ for } h \le h_1.$$

With these choices, for every $h \in (0, h_1]$ all five coefficients admit the *h-independent* positive lower bounds

$$\begin{aligned} &c_1 = \epsilon > 0, \qquad c_2 \ge \frac{3\gamma}{4} - \frac{\gamma}{2} - \frac{\gamma}{8} = \frac{\gamma}{8} > 0, \\ &c_3 = c_4 \ge \frac{\mu B\alpha}{2}\mathrm{e}^{-\alpha} > 0, \qquad c_5 = \frac{\epsilon\, m_f}{4} > 0. \end{aligned} \tag{87}$$

Every c_i depends on h only through the monotone factors $\mathrm{e}^{-2\gamma h}$ and $\mathrm{e}^{-\alpha h}$, which is why replacing them by their infima on $(0, h_1]$ resp. $(0, 1]$ is legitimate.

Every one of $f(X_n)$, $\|P_n\|_F^2$, $\|R_n\|^2$, $\|C_n\|^2$, $\|X_n\|_F^2$ is nonnegative, so comparing the bracket in (84) termwise against the upper bound (61) and factoring out the smallest of the five ratios,

$$\frac{c_1}{1} = \epsilon, \qquad \frac{c_2}{3/4} = \tfrac{4}{3}c_2 \geq \frac{\gamma}{6}, \qquad \frac{c_3}{\mu B/2} = \frac{c_4}{\mu B/2} = \alpha \mathrm{e}^{-\alpha h} \geq \alpha \mathrm{e}^{-\alpha}, \qquad \frac{c_5}{3\epsilon/2} = \frac{m_f}{6},$$

(the factor B canceling between c_3 and (61), as anticipated above), yields

$$c_1 f(X_n) + c_2\|P_n\|_F^2 + c_3\|R_n\|^2 + c_4\|C_n\|^2 + c_5\|X_n\|_F^2 \;\geq\; \underline{c}_{\min}\, \mathcal{W}_\epsilon(n),$$

$$\underline{c}_{\min} := \min\Big\{\epsilon,\ \frac{\gamma}{6},\ \alpha \mathrm{e}^{-\alpha},\ \frac{m_f}{6}\Big\} > 0.$$

For the h^2 term, (60) together with $f(X) \geq \frac{m_f}{2}\|X\|_F^2$ gives $\|P_n\|_F^2 \leq 4\mathcal{W}_\epsilon(n)$ and $\|X_n\|_F^2 \leq \frac{2}{m_f}\mathcal{W}_\epsilon(n)$, hence the conversion

$$\Sigma_n \;\leq\; \Big(4 + \frac{2}{m_f}\Big)\mathcal{W}_\epsilon(n). \tag{88}$$

Substituting both into (84),

$$\mathcal{W}_\epsilon(n+1) \leq \Big[1 - \underline{c}_{\min} h + C_2\Big(4 + \frac{2}{m_f}\Big)h^2\Big]\mathcal{W}_\epsilon(n),$$

so with

$$\boxed{h^* := \min\Big\{1,\ \ h_1,\ \ \frac{\underline{c}_{\min}}{2\,C_2\big(4 + \frac{2}{m_f}\big)}\Big\} > 0, \qquad \rho := \frac{\underline{c}_{\min}}{2},} \tag{89}$$

we obtain (62) for every $h \in (0, h^*)$. All three terms in (89) are needed: $h \leq 1$ normalizes (74) and permits $\mathrm{e}^{-\alpha h} \geq \mathrm{e}^{-\alpha}$; $h \leq h_1$ turns $\gamma \mathrm{e}^{-2\gamma h}$ into the h-free bound $\frac{3\gamma}{4}$; and the third term is exactly the absorption $C_2(4+\frac{2}{m_f})h^2 \leq \frac{\underline{c}_{\min}}{2}h$. The factor $4 + \frac{2}{m_f}$ is the conversion (88) and must not be omitted, since (84) states the remainder in terms of Σ_n rather than $\mathcal{W}_\epsilon(n)$.

h-independence of h^*. Reading the chain of parameter choices backwards: $\mathcal{W}_R$ depends on the initial data; $\bar{P}$ on $\mathcal{W}_R, M, m_f$; $\xi_{\max}$ on $\bar{P}, \mu, \alpha, \epsilon_{\mathrm{stab}}, r_0, c_0$; B on $\gamma, \alpha, \mu, \bar{P}$; η, δ on $m_f, \gamma, \xi_{\max}$; ϵ on $m_f, \gamma, \xi_{\max}$; h_1 on γ; C_2 on $M, \gamma, \alpha, \mu, B, \bar{P}, \xi_{\max}, m_f, \epsilon$; and $\underline{c}_{\min}$ on $\epsilon, \gamma, \alpha, m_f$. None involves h, so h^* in (89) is a fixed positive number determined by $(m_f, M, \gamma, \alpha, \mu, \epsilon_{\mathrm{stab}})$ and L — precisely the dependence the theorem permits.

Finally, $\rho h \leq \frac{\underline{c}_{\min}}{2} \leq \frac{\epsilon}{2} \leq \frac{1}{8} < 1$, so $1 - \rho h \in (0, 1)$; hence (62) gives $\mathcal{W}_\epsilon(n+1) \leq \mathcal{W}_\epsilon(n) \leq \mathcal{W}_R$, which propagates (63) and closes the induction. Iterating,

$$\mathcal{W}_\epsilon(n) \leq (1 - \rho h)^n\, \mathcal{W}_R \leq \mathcal{W}_R\, \mathrm{e}^{-\rho n h},$$

and (60) converts this into the assertion of the theorem with

$$\kappa = -\tfrac{1}{h}\ln(1 - \rho h) \;\geq\; \rho, \qquad C_\star = \Big(5 + \frac{2}{\mu B}\Big)\mathcal{W}_R,$$

using $f \leq \mathcal{W}_\epsilon$, $\|P_n\|_F^2 \leq 4\mathcal{W}_\epsilon$ and $\|R_n\|^2 + \|C_n\|^2 \leq \frac{2}{\mu B}\mathcal{W}_\epsilon$. □