© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. This is the accepted version of: M. E. Nixon, "Sliding Mode Control with Anchor Gradient Correction for Continual Learning in Transformer Language Models," in Proc. 2026 65th IEEE Conference on Decision and Control (CDC), Honolulu, HI, USA, Dec. 2026, to appear.

Sliding Mode Control with Anchor Gradient Correction for Continual Learning in Transformer Language Models

Mason E. Nixon, Ph.D., *Senior Member, IEEE*

***Abstract*— Transformer language models suffer catastrophic forgetting when trained sequentially on shifting data distributions. We formulate continual learning as a discrete-time output-regulation problem in which the parameters are the state, the optimizer is a *normalizing actuator*, and a measured behavioral quantity is the regulated output. The controller combines gradient attenuation with an *anchor gradient correction* computed on a small held-out corpus. We show that under an adaptive-moment optimizer the two gains do not set step magnitude but step *direction*, so that the realized control input is a single mixing ratio between the new-task and anchor descent directions. We prove a bounded-authority lemma for this actuator, and invert the gain schedule so that the efficacy hypothesis can be stated on measurable state rather than on the internal mixing ratio. Under a stated directional-efficacy assumption this yields a quasi-sliding-mode result: the regulated output reaches a band in finite time and that band is forward invariant. Under the weaker stochastic form of the assumption we retain finite mean hitting time, positive recurrence, a bound on the long-run fraction of time spent outside the band, and an exponential bound on excursion overshoot. The result is motion within a band rather than exact sliding. Across 30 seeds on a three-phase sequential character-level task with a 115,328-parameter transformer, the controller holds Task 1 perplexity at 5.30 ± 0.09 against 2193 ± 544 uncontrolled and 282 ± 83 for EWC, and reduces embedding drift $12.9\times$, at a measured plasticity cost of $2.15\times$ on the final task. Two comparators isolate the mechanism: gradient attenuation alone, and A-GEM, the latter with more memory and no plasticity cost. Both leave the error outside tolerance on every step of the final task.**

Keywords: Sliding mode control, machine learning and control, nonlinear systems.

I. Introduction

Catastrophic forgetting, the rapid erasure of previously acquired knowledge during sequential training, remains a central obstacle for deployed transformer language models [1]–[3]. Existing mitigations include regularization (EWC [1]), replay buffers [4], [5], and architectural isolation [6], [7].

None of these treats the training process itself as a controlled dynamical system with formal stability guarantees. Stated plainly, it is a regulation problem of an ordinary kind. A model is trained on an initial task, reaching a parameter vector that meets a behavioral specification, and is deployed; new data then arrives and training resumes, so the parameters move. The question is not whether they move, since they must in order to learn, but whether a measured behavioral quantity stays within tolerance while they do. So stated, the problem has the standard ingredients: a plant whose state is the parameter vector, stochastic new-task gradients as the disturbance pushing it off specification, a measurable output, and an actuator. What is missing is a control law that closes the loop on the measurement.

Recent work treats learning dynamics control-theoretically: drift-plus-penalty formulations give Lyapunov guarantees for continual learning [8], and Lyapunov-stable controllers have been designed for adaptation under concept drift [9]. These regulate hyperparameters of the training procedure. We instead regulate a *measured behavioral output of the model itself*, perplexity on a held-out anchor corpus, and characterize the band the loop attains.

A natural first attempt is a pure gradient-attenuation controller scaling the gradient by $\alpha_t \in (0, 1]$ according to measured drift. Such a *boundary-layer regulator*, recovered by setting $\beta_{\max} = 0$ below, can slow departure from the specification but has no corrective authority: its input is a non-negative multiplier on a direction already leaving the band, so it scales that departure without reversing it. Sec. VI tests this across 30 seeds and finds that attenuation alone never returns the error inside tolerance.

Key insight. A small backward pass on a held-out anchor corpus supplies a corrective direction. Under an adaptive-moment optimizer the two gains then behave differently than a plain gradient-descent reading suggests: the optimizer normalizes gradient scale, so the gains set the update's *direction* rather than its magnitude, and the effective control input is the single mixing ratio $r_t = \beta_t/\alpha_t$. We build the analysis on that fact rather than against it.

Contributions.

1) We formulate transformer continual learning as a discrete-time output-regulation problem in which the parameters are the state, stochastic new-task gradients are bounded disturbances, and a behavioral quantity measured on a held-out anchor corpus is the regulated output (Sec. III).
2) We model the adaptive-moment optimizer as a *normalizing actuator* in Lur'e form and show that the gains set the update's *direction* rather than its magnitude, making the mixing ratio $r_t = \beta_t/\alpha_t$ the realized control input (Sec. IV). The controller adds no parameters, no replay buffer, and no architecture changes, at one anchor backward pass per step.
3) We prove a bounded-authority lemma for that actuator and, under a directional-efficacy assumption, a quasi-sliding-mode result: the behavioral error reaches a band in finite time and that band is forward invariant, with an exponential overshoot bound under the weaker stochastic form (Sec. V). The scheduling signal admits

M. E. Nixon is Lead Consultant at OrbitLink Consulting, LLC, Madison, AL 35758, USA mason.nixon@orbitlink.org

no attractivity claim, yet its parameter-motion term is what keeps the corrective channel open.

4) We validate across 30 seeds, showing that the anchor correction drives Task 1 perplexity *below* its pre-training reference, positive backward transfer rather than mere retention, with $12.9\times$ lower embedding drift, and we report the attained band, the plasticity cost, and a diagnosis of the efficacy assumption that stops short of verifying it (Sec. VI).

II. Background and Related Work

A. Continual Learning

Sequential learning without degrading prior performance is extensively surveyed [2], [10], [11], with the taxonomy identifying regularization (EWC [1], SI [12]), replay [4], and architectural isolation [6]; transformer-specific methods include O-LoRA [13], progressive prompts [14], and self-distillation [15]. No survey identifies a continual learning (CL) method carrying formal stability guarantees on the training dynamics themselves, as opposed to on a static objective.

A parallel line of brain-inspired CL research [16] draws on complementary learning systems, synaptic consolidation, and neurogenesis, and is relevant here because the mechanisms it describes are feedback mechanisms. Parisi et al. [2] observe that homeostatic plasticity acts as a "feedback control signal that regulates the unstable dynamics of Hebbian plasticity," a control statement in all but notation, while [16] notes such mechanisms remain described abstractly. The two gains below give both an explicit form: α_t plays the role of homeostatic regulation, modulating gradient-driven synaptic change from an observed behavioral output, and β_t that of hippocampal replay, reactivating anchor-task gradients in proportion to detected drift [2]. We draw no biological claim from the analogy, but it separates the term supplying corrective authority from the one that does not, a distinction Sec. V makes precise and Sec. VI measures.

Gradient-based memory methods are the closest relatives. GEM [5] projects the new-task gradient into the half-space that does not increase any stored task's loss; A-GEM [17] reduces this to one averaged constraint; DER [18] matches logits on replayed samples. GEM also introduced *backward transfer*, reporting that improving a prior task while learning a new one is attainable; our below-reference result is positive backward transfer in that established sense, and the contribution is the regulation guarantee attached to it.

B. Sliding Mode, and Optimizers as Feedback Systems

SMC [19], [20] drives the state onto a manifold and maintains it there, via a reaching then a sliding phase, under $s\,\Delta s < 0$ [21]. In discrete time exact sliding is generally unattainable: the state attains a *quasi-sliding mode*, remaining within a band around the surface rather than on it [22]. Our results are of this second kind, and the implementation is a smooth, stochastic, normalized-actuator realization rather than a discontinuous switching law; sliding-mode terminology below refers to that lineage, not to the implemented behavior.

First-order optimizers admit control-theoretic readings: as feedback interconnections analyzed by integral quadratic constraints [23], through loop shaping [24], and as PID controllers [25]. We use this literature for *modeling* structure only, and do not invoke the theorems of [23], which assume a convexity a transformer loss does not satisfy. Adam [26] admits a dynamical-system characterization under nonconvexity [27], and its per-coordinate step magnitude is largely decoupled from gradient magnitude [28], the property Sec. IV builds on.

III. Problem Formulation

A model with parameters $\theta \in \mathbb{R}^N$ has been trained on an initial task, reaching θ_0, and training continues on new tasks. Let $\mathcal{D}_A$ be a small fixed *anchor corpus* held out from the initial task, standing in for the specification we must not violate, and let the behavioral measurement be *perplexity* $\mathrm{ppl}(\theta, \mathcal{D})$, the exponentiated mean token-level cross-entropy, lower being better. Write $\mathrm{ppl}^\star = \mathrm{ppl}(\theta_0, \mathcal{D}_A)$.

The quantity we regulate is the signed *behavioral error*

$$\sigma_t = \mathrm{ppl}(\theta_t, \mathcal{D}_A) - \mathrm{ppl}^\star, \tag{1}$$

evaluated on a frozen set. Since perplexity is positive, $\sigma_t > -\,\mathrm{ppl}^\star =: -\sigma_{\min}$. Positive σ_t is forgetting; negative σ_t is backward transfer. Separately, the controller schedules its gains on the nonnegative composite

$$s_t = \max(0, \sigma_t) + w_\theta \left\|\theta_t - \theta_0\right\|_2, \qquad w_\theta > 0. \tag{2}$$

The two signals play different roles. σ_t is the regulated output, while s_t merely selects the gains and, by Corollary 1, cannot be driven to zero by any trajectory that also learns. Every stability statement below concerns σ_t. Given a tolerance band $\mathcal{B} = \{\sigma \le \delta\}$, the objective is to drive σ_t into $\mathcal{B}$ in finite expected time and keep it there for all but a bounded fraction of steps while the new-task loss decreases. We take $\delta = 0$: the band is exactly "no worse than the reference."

IV. Controller and Plant

A. Two-term control law

With g_t^{new} the new-task gradient and $g_t^{\mathrm{anc}} = \nabla_\theta \mathcal{L}(\theta_t, \mathcal{D}_A)$ the anchor gradient, the controller forms

$$h_t = \alpha_t\, g_t^{\mathrm{new}} + \beta_t\, g_t^{\mathrm{anc}}, \tag{3}$$

with boundary-layer gains, $q_t = \min(\max(s_t/\phi, 0), 1)$,

$$\alpha_t = 1 - (1 - \alpha_{\min})\, q_t, \qquad \beta_t = \beta_{\max}\, q_t. \tag{4}$$

Both are continuous, $\alpha_t \in [\alpha_{\min}, 1]$ and $\beta_t \in [0, \beta_{\max}]$, saturating for $s_t \ge \phi$. The overhead is one additional backward pass on an anchor mini-batch.

B. The optimizer as a normalizing actuator

The signal h_t is not applied to the parameters directly. It drives an adaptive-moment optimizer [26], whose moving averages m_t, v_t are first-order linear filters and whose normalization $\hat{m}/(\sqrt{\hat{v}}+\varepsilon)$ is a static elementwise nonlinearity. The actuator is therefore a Lur'e-type interconnection, linear dynamics in feedback with a static nonlinearity [23], [24], followed by the parameter integrator $\theta_{t+1} = \theta_t + u_t$.

That nonlinearity is scale-normalizing. With $\varepsilon = 0$ it is exactly invariant to a constant uniform rescaling of h_t, since $h \mapsto ch$ scales $m \mapsto cm$ and $\sqrt{v} \mapsto c\sqrt{v}$; the invariance is approximate when the gains vary slowly relative to the moment timescales $(1-\beta_1)^{-1}$ and $(1-\beta_2)^{-1}$.

Remark 1 (The gains set direction, not magnitude): Consequently (α_t, β_t) do not attenuate the step. The realized control input is the normalized direction of h_t, parameterized by the single scalar mixing ratio

$$r_t = \frac{\beta_t}{\alpha_t} = \frac{\beta_{\max} q_t}{1-(1-\alpha_{\min})q_t} \in \left[0,\ \frac{\beta_{\max}}{\alpha_{\min}}\right]. \tag{5}$$

Sec. VI confirms this directly: α_t falls by two orders of magnitude while the realized step norm does not fall at all. A plain gradient-descent reading of (3), in which α_t scales the displacement, is not a valid model of this plant.

V. Stability Analysis

A. Bounded control authority

Lemma 1 (Bounded authority): Let the actuator use moment coefficients $\beta_1, \beta_2 \in (0,1)$ with $\rho := \beta_1^2/\beta_2 < 1$. Then for every coordinate i and every $t \geq 1$,

$$|\Delta\theta_{t,i}| \leq c(\beta_1,\beta_2)\,\eta, \qquad c = \frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\rho)}}, \tag{6}$$

independently of the gradient magnitudes, and hence $\|\theta_{t+1}-\theta_t\|_2 \leq c\,\eta\sqrt{N} =: U$.

Proof: Expanding the accumulators, $m_t = (1-\beta_1)\sum_{k\leq t}\beta_1^{t-k} g_k$ and $v_t = (1-\beta_2)\sum_{k\leq t}\beta_2^{t-k} g_k^2$. Splitting each term of m_t and applying Cauchy–Schwarz,

$$\begin{aligned}\sum_k \beta_1^{t-k}\,|g_k| &= \sum_k \frac{\beta_1^{t-k}}{\beta_2^{(t-k)/2}}\cdot \beta_2^{(t-k)/2}\,|g_k| \\ &\leq \Big(\sum_{j\geq 0}\rho^j\Big)^{1/2}\Big(\sum_k \beta_2^{t-k} g_k^2\Big)^{1/2},\end{aligned} \tag{7}$$

so $|m_t|/\sqrt{v_t} \leq (1-\beta_1)\big[(1-\beta_2)(1-\rho)\big]^{-1/2}$. Bias correction contributes the factor $\sqrt{1-\beta_2^t}/(1-\beta_1^t) \leq 1$ for these coefficients, and $\sqrt{\hat{v}}+\varepsilon \geq \sqrt{\hat{v}}$, so the bound carries to the realized update. ∎

Lemma 1 is the formal content of Remark 1: authority is bounded by the learning rate and moment coefficients, not by the gains, so any bound of the form $\eta(\alpha_{\min}G + \beta_{\max}G_a)$, vanishing as $\alpha_t \to \alpha_{\min}$, is invalid for this plant.

Assumption 1 (Local Lipschitz output): σ is L_σ-Lipschitz along the segment joining consecutive visited parameter states.

Proposition 1 (Bounded increment): Under Assumption 1 and Lemma 1, $|\sigma_{t+1} - \sigma_t| \leq L_\sigma U =: \bar{\Delta}$.

Proposition 1 bounds a single increment. On its own it gives no invariance, since a bounded increment repeated indefinitely leaves any bounded set. What converts it into one is a guarantee that the corrective channel is active whenever the error is large.

B. Inverting the gain schedule

Assumption 2 below is stated in terms of the mixing ratio, which is internal to the controller. The analysis needs it in terms of state that can be measured, and inverting the gain schedule (4) supplies the translation.

Lemma 2 (Efficacy threshold): The map $q \mapsto r(q) = \beta_{\max} q/[1-(1-\alpha_{\min})q]$ is strictly increasing on $[0,1]$ with $r(1) = \beta_{\max}/\alpha_{\min}$. Hence for any $\bar{r} \in (0, \beta_{\max}/\alpha_{\min}]$, setting

$$\bar{q} = \frac{\bar{r}}{\beta_{\max} + \bar{r}\,(1-\alpha_{\min})} \in (0,1], \tag{8}$$

we have $r_t \geq \bar{r}$ if and only if $s_t \geq \phi\bar{q}$. Two sufficient conditions follow from (2):

$$\sigma_t \geq \phi\bar{q} \qquad \text{or} \qquad \|\theta_t - \theta_0\|_2 \geq \phi\bar{q}/w_\theta. \tag{9}$$

Proof: $r'(q) = \beta_{\max}\alpha_{\min}[1-(1-\alpha_{\min})q]^{-2} > 0$, and solving $r(\bar{q}) = \bar{r}$ gives (8). Since $q_t = \min(\max(s_t/\phi, 0), 1)$ is nondecreasing in s_t, $r_t \geq \bar{r} \iff q_t \geq \bar{q} \iff s_t \geq \phi\bar{q}$. Both terms of (2) are nonnegative, so $s_t \geq \max(0, \sigma_t)$ and $s_t \geq w_\theta\,\|\theta_t - \theta_0\|$, giving (9). ∎

The derivation is immediate: r is increasing and s_t dominates each of its two nonnegative terms. Its role is to restate Assumption 2's hypothesis on σ_t or on displacement, so Theorem 1's case analysis runs on measurable quantities. The second branch is the one that binds: the term of s_t that Corollary 1 shows cannot be regulated is what keeps the corrective channel open, and a displacement of 6.12 gives $r_t \geq 3$ against a measured 8.72.

C. Reaching and invariance of the band

Assumption 2 (Directional efficacy): There exist $\bar{r} > 0$ and $\gamma > 0$ such that whenever $\sigma_t > \delta$ and $r_t \geq \bar{r}$,

$$\mathbb{E}\big[\sigma_{t+1} - \sigma_t \,\big|\, \mathcal{F}_t\big] \leq -\gamma. \tag{10}$$

It holds in the *strong form* if additionally $\sigma_{t+1} - \sigma_t \leq -\gamma$ almost surely under the same conditions.

Assumption 2 says that when the behavioral error is positive and the mixing ratio large enough, the loop reduces that error, on average in the stated form and pathwise in the strong form. It is stated on σ_t, the signal the anchor gradient can move; the analogous condition on s_t is *false*, since no gradient reduces $w_\theta\,\|\theta_t - \theta_0\|$ while the model learns. It conditions on the mixing ratio rather than β_t alone, per Remark 1. Sec. VI diagnoses but does not verify it, the principal limitation of the analysis. Nor could it be derived here: whether the anchor gradient reduces anchor error against an opposing new-task gradient is a property of the task pair, not of the optimizer. The optimizer-specific content, Lemma 1 and Remark 1, is derived and measured rather than assumed.

Write $\bar{\sigma} := \max(\delta, \phi\bar{q})$, the error level above which Lemma 2 guarantees $r_t \geq \bar{r}$ unconditionally, and define the band

$$\mathcal{A} = \{\sigma : -\sigma_{\min} \leq \sigma \leq \bar{\sigma} + \bar{\Delta}\}. \tag{11}$$

Theorem 1 (Finite-time reaching and forward invariance): Under Lemma 1, Assumption 1, and Assumption 2 in strong form:

(i) $\mathcal{A}$ is forward invariant: $\sigma_t \in \mathcal{A}$ implies $\sigma_{t+1} \in \mathcal{A}$ for every admissible realization;
(ii) from any σ_0, the error enters $\mathcal{A}$ within $\lceil(\sigma_0 - \bar{\sigma})_+/\gamma\rceil$ steps and remains there;
(iii) $|\sigma_t| \leq \bar{\sigma} + \bar{\Delta}$ eventually, so the closed loop is uniformly ultimately bounded.

Proof: (i) Let $\sigma_t \in \mathcal{A}$. If $\sigma_t > \bar{\sigma}$ then $\sigma_t > \delta$ and, by Lemma 2 with $\sigma_t \geq \phi\bar{q}$, also $r_t \geq \bar{r}$; the strong form gives $\sigma_{t+1} \leq \sigma_t - \gamma < \bar{\sigma} + \bar{\Delta}$. If instead $\sigma_t \leq \bar{\sigma}$, Proposition 1 gives $\sigma_{t+1} \leq \sigma_t + \bar{\Delta} \leq \bar{\sigma} + \bar{\Delta}$. The lower bound holds always, perplexity being positive. (ii) While $\sigma_t > \bar{\sigma}$ the same case gives a decrease of at least γ per step, so the error falls into $\mathcal{A}$ within $\lceil(\sigma_0 - \bar{\sigma})_+/\gamma\rceil$ steps; (i) then applies. (iii) is (i) and (ii) with $\sigma \geq -\sigma_{\min}$. ∎

Theorem 1 is the discrete-time quasi-sliding-mode statement in standard form [22]: finite-time reaching of a band and invariance thereafter, in place of the exact sliding motion discrete-time implementations cannot attain. It recovers the two canonical phases without asserting convergence to a surface. In the reported regime we take $\bar{r} = 3$, the largest ratio at which γ is measurable (Sec. VI-D). Then $\bar{q} = 0.917$, the second branch of (9) activates at a displacement of 6.12 against a measured 8.72, and $\bar{\sigma} = \delta = 0$: the certified band is $\sigma \leq \bar{\Delta} = 4.04$ against a measured Task-3 maximum of 1.81. The strong form buys pathwise invariance and is not established: gradients conflict on 45.9% of steps. Read Theorem 1 as the guarantee available *if* efficacy holds pathwise; the run bears on Theorem 2.

D. Practical stability under the stochastic assumption

The strong form of Assumption 2 is a genuine strengthening. Under the conditional-mean form alone, pathwise invariance is unavailable, but the following still holds.

Theorem 2 (Practical stability): Let $\mathcal{B} = \{\sigma \leq \delta\}$ and $T = \inf\{t : \sigma_t \in \mathcal{B}\}$. Under Lemma 1, Assumption 1 and Assumption 2:

(i) $\mathbb{E}[T] \leq (\sigma_0 + \sigma_{\min})/\gamma$;
(ii) after any exit from $\mathcal{B}$ the mean return time is at most $(\delta + \bar{\Delta} + \sigma_{\min})/\gamma$, so $\mathcal{B}$ is visited infinitely often almost surely;
(iii) the long-run fraction of steps outside $\mathcal{B}$ satisfies

$$\limsup_{n\to\infty} \frac{1}{n}\,\mathbb{E}\big[\#\{t < n : \sigma_t > \delta\}\big] \leq \frac{\bar{\Delta}}{\gamma + \bar{\Delta}}; \tag{12}$$

(iv) each excursion above $\mathcal{B}$ overshoots by more than $\lambda > 0$ with probability at most

$$\Pr\Big[\sup\nolimits_{\text{excursion}}\, \sigma > \delta + \bar{\Delta} + \lambda\Big] \leq \exp\big(-\gamma\lambda/\bar{\Delta}^2\big). \tag{13}$$

Proof: (i) By Assumption 2, $M_t := \sigma_{t\wedge T} + \gamma(t \wedge T)$ is a supermartingale, so $\mathbb{E}[\sigma_{t\wedge T}] + \gamma\mathbb{E}[t \wedge T] \leq \sigma_0$. Perplexity is positive, so $\sigma \geq -\sigma_{\min}$ and $\gamma\,\mathbb{E}[t \wedge T] \leq \sigma_0 + \sigma_{\min}$; let $t \to \infty$ by monotone convergence. (ii) By Proposition 1 an exit lands at $\sigma \leq \delta + \bar{\Delta}$; apply (i) from that state. (iii) Sum (10) over n steps, splitting on whether $\sigma_t > \delta$. Outside $\mathcal{B}$ the mean increment is at most $-\gamma$; inside it is at most $\bar{\Delta}$ by Proposition 1. With N_{out} the count outside, $-\sigma_{\min} - \sigma_0 \leq -\gamma\mathbb{E}[N_{\text{out}}] + \bar{\Delta}(n - \mathbb{E}[N_{\text{out}}])$, and rearranging gives the claim. (iv) The increment lies in an interval of width $2\bar{\Delta}$ almost surely, so Hoeffding's lemma and (10) give $\mathbb{E}[e^{\vartheta(\sigma_{t+1}-\sigma_t)} \mid \mathcal{F}_t] \leq \exp(-\vartheta\gamma + \vartheta^2\bar{\Delta}^2/2)$ while $\sigma_t > \delta$. At $\vartheta = \gamma/\bar{\Delta}^2$ that factor is at most $e^{-\gamma^2/2\bar{\Delta}^2} < 1$, so $e^{\vartheta\sigma_t}$ is a nonnegative supermartingale until the excursion ends. Excursions begin at $\sigma \leq \delta + \bar{\Delta}$, and Ville's inequality from that level gives (13). ∎

Theorem 2 is the stochastic counterpart of Theorem 1: the loop attains a band, is recurrent to it, spends a bounded fraction of time outside it, and overshoots only with probability decaying exponentially. Neither asserts convergence to a point; parts (iii) and (iv) are what experiment compares against.

Corollary 1 (No attractivity for the scheduling signal): Both terms of (2) are nonnegative and $w_\theta > 0$, so $s_t = 0$ if and only if $\theta_t = \theta_0$ and $\mathrm{ppl}(\theta_t, \mathcal{D}_A) \leq \mathrm{ppl}^\star$. Hence no trajectory that moves the parameters can approach $s = 0$, and s_t admits no reaching, sliding, or invariance claim of its own.

Remark 2 (Scope): We do not claim exact sliding, finite-time convergence of σ_t to a point, an invariant set for s_t, a chattering bound, or any guarantee over the moment state or a population of models. Assumptions 1 and 2 are assumptions, diagnosed below and not established, and the strong form Theorem 1 needs is strictly stronger than the conditional-mean form Theorem 2 uses.

VI. Experiments

A. Setup

We use a nanoGPT model [29] with 115,328 parameters (2 layers, 2 heads, embedding dim 64, context 64, vocabulary 87) trained on three sequential tasks (2,000 steps each, Adam, $\eta = 3 \times 10^{-4}$, batch 32): Task 1 (Tiny Shakespeare), Task 2 (character-reversed), Task 3 (substitution cipher). Four conditions start from the same Task 1 checkpoint: uncontrolled descent; EWC [1] with $\lambda = 50$; A-GEM [17]; and the controller of (3)–(4) with $\phi = 100$, $\alpha_{\min} = 0.01$, $\beta_{\max} = 0.3$, $w_\theta = 15$, so $r_t \in [0, 30]$ by (5). A-GEM projects the new-task gradient onto the non-conflicting half-space of a reference gradient when the two conflict; its reference batch is 32, its only hyperparameter, left untuned. Its episodic memory covers *all* past tasks, so on Task 3 it draws from Tasks 1 and 2 while the controller anchors on Task 1 alone: the baseline is given strictly more memory. All run across 30 seeds, reported as mean ± one population standard deviation. Three measurement streams are distinguished and never pooled: the *endpoint* stream, the accepted 30-seed run logged every 50 steps, supplying Table I; the *diagnostic*

TABLE I

END-OF-TRAINING METRICS (MEAN ± POPULATION STD, 30 SEEDS).

Metric	Uncontrolled	EWC	Controller
ppl_1 (↓)	2193 ± 544	282 ± 83	$\mathbf{5.30 \pm 0.09}$
ppl_2 (↓)	1393 ± 357	412 ± 106	$\mathbf{61.1 \pm 4.8}$
ppl_3 (↓)	$\mathbf{5.92 \pm 0.09}$	6.48 ± 0.10	12.68 ± 0.44
Δ_{emb} (↓)	1.011 ± 0.077	0.833 ± 0.097	$\mathbf{0.078 \pm 0.009}$
$\|\theta - \theta_0\|$	11.64 ± 0.09	8.59 ± 0.18	8.72 ± 0.11

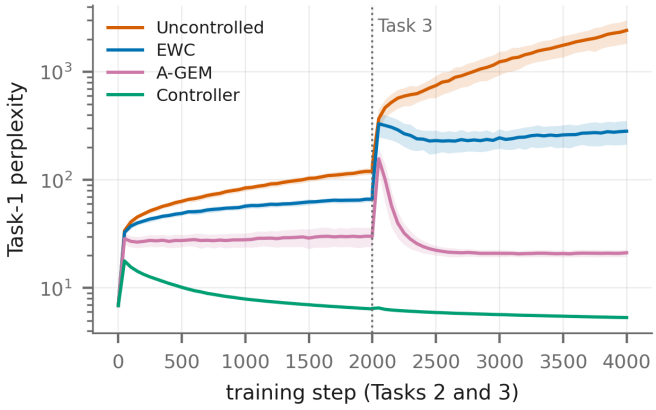


Fig. 1. Task-1 perplexity during Tasks 2 and 3 (mean ± one population std, 30 seeds, log scale). All four arms are drawn from the diagnostic stream so the panel is internally consistent; Table I is the endpoint stream. A-GEM spikes at the Task-3 onset and recovers; only the controller descends.

stream, a per-step instrumented run on a frozen anchor set with controller decisions unchanged, supplying Sec. VI-D and the A-GEM comparison; and the *ablation* stream of Sec. VI-E, the diagnostic run with $\beta_{\max} = 0$. Endpoint and diagnostic estimates of the same quantity differ only in the third significant figure; we name the stream wherever both appear.

B. Retention and plasticity

Fig. 1 traces Task-1 perplexity through both later tasks and Table I gives the endpoints. Uncontrolled and EWC degrade monotonically, while the controller ends at 5.30 ± 0.09, below the reference $\mathrm{ppl}^\star = 6.89$. In the terminology of [5] that is positive backward transfer: the anchor correction fine-tunes on anchor data whenever $\beta_t > 0$, driving the behavioral error to and past zero. The mechanism is ordinary replay; what Theorem 2 adds is the band and the recurrence statement.

The cost is plasticity, a factor of 2.15 on Task 3, and Task 2 ends above its end-of-phase value because the correction prioritizes the anchor task. Embedding drift falls $12.9\times$, so internal representations are preserved and not merely input–output behavior, and parameter displacement is nearly identical for EWC and the controller, so proximity in parameter space does not explain the retention gap.

C. Regulated output

Fig. 2 shows the closed loop in the $(\|\theta_t - \theta_0\|, \sigma_t)$ plane. The behavioral error rises to 10.9 at the onset of Task 2, is driven back through zero at a displacement of 5.46, and ends at -1.59 while the displacement keeps growing to 8.72:

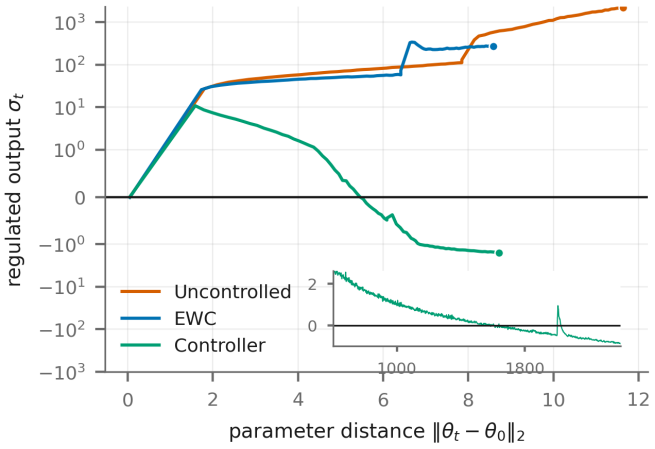


Fig. 2. Regulated output σ_t against parameter distance (30 seeds, endpoint stream, symmetric-log vertical axis). Inset: the crossing region at per-step resolution on the frozen diagnostic set, one seed.

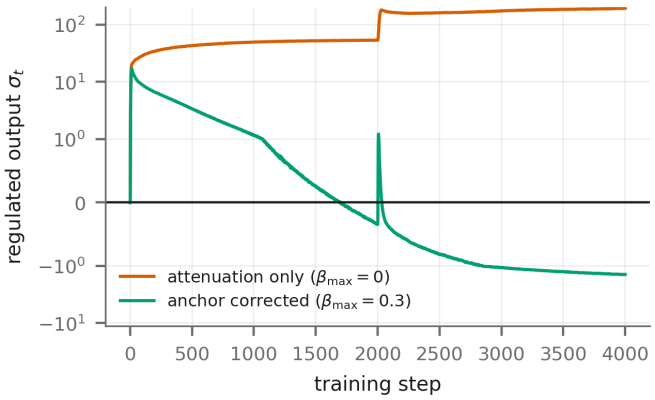


Fig. 3. Regulated output under the full controller and under the degenerate $\beta_{\max} = 0$ case, seed-mean over 30 seeds, symmetric-log vertical axis. Attenuation alone never enters the band.

the model learns and the regulated output returns to band, which is the behavior the formulation targets. The scheduling signal grows with it to 131 ± 2, expected by Corollary 1 and carrying no stability content.

D. Diagnostics and assumption checks

All quantities below come from the 30-seed per-step diagnostic stream.

Bounded authority. With $\beta_1 = 0.9$, $\beta_2 = 0.999$, Lemma 1 gives $c = 7.27$ and $U = 0.741$; the largest observed step is $0.108 = 1.06\,\eta\sqrt{N}$, so the bound holds with roughly sevenfold conservatism. Across the run α_t falls two orders of magnitude, from 0.9995 to 0.0100, while the mean step norm stays at 0.0174 and does not decrease. This is direct evidence for Remark 1 and rules out any magnitude-attenuation reading of the gains. The largest observed $|\Delta\sigma_t|$ is 4.04, so Assumption 1 holds along the visited path with $L_\sigma \le 5.45$ against the bound U.

Mixing ratio and saturation. r_t sweeps the full range $[0, 30]$. The gains saturate at $(\alpha, \beta) = (0.01, 0.30)$ from step 2425, 21% into Task 3, and stay saturated, a saturated fraction of 0.41 over the controlled phase. For the final 79%

of Task 3 the controller is thus a fixed-ratio anchor-replay scheme rather than a proportional regulator, a regime the boundary-layer width ϕ directly sets.

Attained band. σ_t peaks at 17.2 ± 1.1 and changes sign 11.4 times on average during Task 2 as it is driven through zero, then settles: during Task 3 it changes sign twice, stays within $|\sigma_t| \leq 1.81$ across all seeds, and is positive on only 1.55% of steps. Read against Theorem 2(iii) with $\delta = 0$, the measured long-run violation fraction is 0.0155. The reaching and settling behavior is the two-phase structure Theorem 1 describes.

Certified band. Because s_t saturates the boundary layer, Lemma 2 certifies $r_t = 30$ throughout the controlled phase, so $\bar{\sigma} = \delta = 0$ and the band of Theorem 1 is $\sigma \leq \bar{\Delta} = 4.04$. The measured Task-3 maximum of 1.81 lies inside it by a factor of 2.2, the band inheriting Lemma 1's conservatism through $\bar{\Delta} = L_\sigma U$. It is a guarantee, not a prediction, and not a verification of Theorem 1: containment already follows from Theorem 2 under the weaker hypothesis the data support.

Assumption 2, tested directly. The stream logs diagnostic perplexity either side of each update, so the conditional mean the assumption asserts can be formed rather than proxied. On $\{\sigma_t > \delta,\ r_t \geq \bar{r}\}$ it is negative, and γ grows with $\bar{r}$ as the assumption predicts: 0.0022 at $\bar{r} = 1$ over $11{,}197$ steps, 0.0182 at $\bar{r} = 3$ over 721 steps in 29 of 30 seeds, at the conservative end of a seed-clustered bootstrap. Sec. V uses that $\bar{r}$.

The test has two limits. Above $\bar{r} \approx 3.5$ the conditioning set is *empty*, since r_t grows large only once σ_t is already at or below zero, so the assumption is untested at the saturated ratio of 30. And at the measured γ the constants of Theorem 2 do not bind: occupancy $\bar{\Delta}/(\gamma + \bar{\Delta}) = 0.996$ against a measured 0.0155. The qualitative statements survive; their constants are uninformative here. The gradient inner product is also *negative on* 45.9% *of steps* (-0.115 mean), so the correction works against opposition rather than agreement.

E. Is the correction necessary, and is constraining enough?

Setting $\beta_{\max} = 0$ disables the correction while leaving the attenuation schedule untouched, isolating the claim that attenuation alone lacks corrective authority. Over 30 seeds (Fig. 3), attenuation alone ends at $\mathrm{ppl}_1 = 202.0 \pm 33.0$ and $\sigma = +194.1$, *outside the band on* 100% *of Task-3 steps*; the full controller ends at $\mathrm{ppl}_1 = 5.32 \pm 0.05$ and $\sigma = -1.41$, outside on 1.55%. Both arms are read from the per-step streams, so the comparison is like-for-like: the controller's 5.32 ± 0.05 is the diagnostic-stream counterpart of Table I's 5.30 ± 0.09, not a second estimate. The uncontrolled arm is byte-identical across the two runs. Attenuation alone does not reach the band later: it never enters, and σ grows monotonically once Task 3 begins.

The ablation also exposes the tradeoff. Task 3 perplexity is 5.92 ± 0.09 uncontrolled, 9.58 ± 0.17 under attenuation alone, and 12.67 ± 0.45 with the correction (again diagnostic stream, against Table I's 12.68 ± 0.44), against Task 1 perplexities of 2193, 202.0, and 5.32: a stability-plasticity frontier on which the anchor correction buys two orders of magnitude of retention for roughly a further 32% of plasticity.

Attenuation is not the only way to fall short. A-GEM computes the same kind of per-step reference gradient, from more memory, but uses it as a constraint rather than an input: the projection removes the component that would raise anchor loss and never descends it. It spikes at the Task-3 onset and recovers (Fig. 1), ending at $\mathrm{ppl}_1 = 21.17 \pm 1.66$ and $\sigma = +14.3$, again *outside the band on* 100% *of Task-3 steps*, projecting on 39.8% of them. It is much the closest comparator, against $+194.1$ for attenuation alone and $+275.2$ for EWC, and still never enters.

Elsewhere A-GEM is the stronger method, on 30 of 30 seeds each: Task-2 perplexity 20.57 ± 1.49 against 60.97 ± 3.50, and Task-3 5.85 ± 0.08 against 12.67 ± 0.45, indistinguishable from uncontrolled descent, so it pays no measurable plasticity cost. The controller wins Task-1 perplexity and embedding drift, also 30 of 30. At 11.60 ± 0.09 displacement against 11.62 ± 0.10 uncontrolled, A-GEM constrains the update's direction while travelling as far as unconstrained training, where the controller stays at 8.72. Only one of the two regulates an output to a tolerance, and it is not the one with more memory.

VII. Discussion and Limitations

Relation to replay. The anchor correction is controller-driven experience replay: β_t scales with measured drift, so replay is applied when and to the degree forgetting is detected, from a single mini-batch rather than a growing buffer. Cost separates it only from GEM [5], which stores gradients per prior task; A-GEM [17] also runs one extra backward pass per step. What distinguishes the controller is how the reference gradient enters. A projection removes the component that would raise anchor loss, so its ceiling is to do no harm, whereas $\beta_t g_t^{\mathrm{anc}}$ supplies a descent direction and drives σ through zero. Sec. VI-E measures the consequence. We offer no per-task constraint satisfaction, only the band of Theorem 2, and A-GEM is the better method wherever plasticity is what matters.

Limitations. The model is small (115,328 parameters) and the task transformations are synthetic; the comparison is against uncontrolled descent, EWC, A-GEM and a degenerate ablation, and is an initial viability study rather than a benchmark against stronger replay methods such as [18]. Nor is the controller uniformly better than A-GEM. The claim is regulation to a tolerance, not dominance. The stability results are conditional, and the conditioning is now measured rather than asserted: γ is positive where it can be estimated, but the conditioning set empties above $\bar{r} \approx 3.5$, so the assumption is untested at the saturated ratio and the constants it yields leave Theorem 2's bounds uninformative on this run. Theorem 1 needs the strong form, which the 45.9% gradient conflict makes implausible as an exact statement, so the two theorems bracket the guarantee rather than forming one result. L_σ is bounded along the visited path, not a priori. Lemma 1 is roughly sevenfold conservative, so $\bar{\Delta}$ and the certified band are loose in proportion. The scheduling signal is a heuristic:

its parameter-motion term drives the gains to saturation for the last 79% of Task 3, so the band is attained largely under fixed-ratio replay rather than proportional regulation, and that same saturation is what makes Lemma 2 bind so easily; neither should be expected where ϕ is large relative to the parameter motion. Finally, the analysis treats the actuator through its realized displacement; a full treatment of the moment state as part of the closed-loop state remains open.

VIII. Conclusion

We modeled continual learning of a transformer language model as discrete-time output regulation with a normalizing actuator, and showed that the gains act on the direction of the update rather than its magnitude, making the mixing ratio the realized control input. Inverting the gain schedule restates the efficacy hypothesis on measurable state, turning a one-step increment bound into a quasi-sliding-mode result: the regulated output reaches a band in finite time and that band is forward invariant. Under the weaker stochastic form of the efficacy assumption we retain finite mean hitting time, positive recurrence, a bounded long-run violation fraction, and an exponential overshoot bound. Both are band results rather than exact sliding. Across 30 seeds the loop holds Task 1 perplexity at 5.30 ± 0.09, below its reference, with $12.9\times$ lower embedding drift and a plasticity cost of $2.15\times$ on the final task. Future work includes second-order sliding-mode designs [30], adaptive $\beta_{\max}$ scheduling, scaling to larger models, and bringing the moment state inside the closed-loop analysis.

Use of AI Tools

The Claude chatbot (Anthropic) was used to improve syntax and grammar in several paragraphs, to identify related work, and to assist with analysis and figure-generation scripts. All technical content, theoretical derivations, and scientific conclusions are the sole work of the author.

References

[1] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[2] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[3] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychology of Learning and Motivation*, vol. 24, pp. 109–165, 1989.

[4] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. Torr, and M. Ranzato, "On tiny episodic memories in continual learning," in *Workshop on Continual Learning, ICML*, 2019.

[5] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[6] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," in *arXiv preprint arXiv:1606.04671*, 2016.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2022.

[8] N. Shah, G. Arya, B. N. Bharath, and R. Prasad, "Theoretical foundations of continual learning via drift-plus-penalty," *Transactions on Machine Learning Research*, 2026, arXiv:2606.08452.

[9] T. B. Pan, M. Zhu, A. J. Cole, R. Wilson, and D. L. Woodard, "Lyapunov-stable adaptive control for multimodal concept drift," *arXiv preprint arXiv:2510.15944*, 2025.

[10] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, 2024.

[11] H. Shi, Z. Xu, H. Wang, W. Qin, W. Wang, Y. Wang, Z. Wang, S. Ebrahimi, and H. Wang, "Continual learning of large language models: A comprehensive survey," *ACM Computing Surveys*, 2025.

[12] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning (ICML)*, 2017, pp. 3987–3995.

[13] X. Wang, T. Chen, Q. Ge, H. Xia, R. Bao, R. Zheng, Q. Zhang, T. Gui, and X. Huang, "Orthogonal subspace learning for language model continual learning," in *Findings of EMNLP*, 2023.

[14] A. Razdaibiedina, Y. Mao, R. Hou, M. Khabsa, M. Lewis, and A. Almahairi, "Progressive prompts: Continual learning for language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[15] Z. Yang, T. Pang, H. Feng, H. Wang, W. Chen, M. Zhu, and Q. Liu, "Self-distillation bridges distribution gap in language model fine-tuning," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[16] M. Aslam *et al.*, "Continual learning inspired by brain functionality: A comprehensive survey," *International Journal of Intelligent Systems*, 2025.

[17] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," in *International Conference on Learning Representations (ICLR)*, 2019, arXiv:1812.00420.

[18] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: A strong, simple baseline," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, arXiv:2004.07211.

[19] Y. Shtessel, C. Edwards, L. Fridman, and A. Levant, *Sliding Mode Control and Observation*, ser. Control Engineering. New York: Birkhäuser, 2014.

[20] V. I. Utkin, *Sliding Modes in Control and Optimization*. Berlin: Springer-Verlag, 1992.

[21] C. Edwards and S. Spurgeon, *Sliding Mode Control: Theory and Applications*. London: Taylor & Francis, 1998.

[22] W. Gao, Y. Wang, and A. Homaifa, "Discrete-time variable structure control systems," *IEEE Transactions on Industrial Electronics*, vol. 42, no. 2, pp. 117–122, 1995.

[23] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016.

[24] B. Hu and L. Lessard, "Control interpretations for first-order optimization methods," in *American Control Conference (ACC)*, 2017, arXiv:1703.01670.

[25] W. An, H. Wang, Q. Sun, J. Xu, Q. Dai, and L. Zhang, "A PID controller approach for stochastic optimization of deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8522–8531.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] A. Barakat and P. Bianchi, "Convergence and dynamical behavior of the ADAM algorithm for nonconvex stochastic optimization," *SIAM Journal on Optimization*, vol. 31, no. 1, pp. 244–274, 2021.

[28] L. Balles and P. Hennig, "Dissecting Adam: The sign, magnitude and variance of stochastic gradients," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 80, 2018, pp. 413–422.

[29] A. Karpathy, "nanoGPT: The simplest, fastest repository for training/finetuning medium-sized GPTs," https://github.com/karpathy/nanoGPT, 2022.

[30] Y. B. Shtessel, M. Taleb, and F. Plestan, "A novel adaptive-gain supertwisting sliding mode controller: Methodology and application," *Automatica*, vol. 48, no. 5, pp. 759–769, 2012.