Training Dynamics and Inference Guarantees of the In-Context GD Mechanism

A satellite of the verified ICL=GD core: how the mechanism self-organizes during training and what it certifies at inference

Dr. Tamás Nagy

tnagyphd@gmail.com

Working Paper • 2026-06-15

Build-std v2.0 | 23:26 | 99edbe71

Abstract

The companion core paper establishes, and machine-checks, a single identity: a transformer's forward pass can implement one gradient-descent step on an implicit least-squares objective (the ICL=GD mechanism). This satellite asks what the verified identity forces to be true along two axes that matter for alignment and inference: *how the mechanism forms during training*, and *what it lets us trust at inference time*. Treating the ICL=GD construction as a fixed premise, we develop a connected family of results — each an independently machine-checked statement about the same verified mechanism. We claim certified nodes, not a certified dependency chain: the kernel checks every proposition, while the thread that links them to the core identity is author-drawn. The contribution is the assembled, auditable family and its empirical ties, not the difficulty of any single lemma. On the dynamics side, scalar gains self-organize: the loss orders gains by contraction, the Newton gain is the zero-loss target, and gradient descent monotonically drives undershooting gains upward — faster for higher-curvature bands — which is the exact mechanism behind the spectral alignment the core paper observes emerging during training. On the inference side, the identity recasts the transformer as an implicit (mesa) optimizer bridging in-context and in-weight learning, bounds in-context sample complexity through the step product — a synthetic probe confirms the predicted faster-convergence, wider-gap, and diminishing-returns curves survive realistic curvature sampling — and — most usefully — turns verified contraction into a *certified anytime readout*: a sound early-exit rule with a shrinking band that always contains the optimum and never re-widens, connecting ICL=GD to anytime-valid inference. We are explicit about scope: the readout certifies the formal contraction object, not arbitrary pretrained models. We test exactly that boundary on three pretrained models (gpt2, gpt2-medium, pythia-410m): the certified early-exit rule is a useful heuristic — firing on most prompts at roughly a third of the network depth, and sometimes matching or beating the full-depth readout — but the formal soundness guarantee transfers only partially (a fired stop is correct in 41–64% of cases at a fixed tolerance), exactly as the scope caveat predicts. Most results are deductive consequences of the verified mechanism; one is a falsifiable sample-complexity prediction and one is an interpretation. They follow, by machine-checked proof, from an identity the core paper has already validated.

Introduction

The companion core paper, *When In-Context Learning Implements Gradient Descent*, establishes and machine-checks the central identity of this program: a transformer's forward pass can implement one gradient-descent step on an implicit least-squares objective, and tests that this ICL=GD mechanism is present in real pretrained models. The view that in-context learning implements an explicit learning algorithm originates with Garg et al. (2022) and Akyürek et al. (2023), made mechanistic by von Oswald et al. (2023); the core paper turns that view into a machine-checked identity. Its Theorems A–D — the gradient-step identity, loss descent, convergence of the in-context estimate, and the softmax/preconditioned lift — are the mechanically verified building blocks we take as given here.

This satellite asks one focused question: *what does the verified identity imply about how the mechanism is learned, and what it certifies at inference?* These are the two questions alignment and inference-optimization care about most. For training dynamics, we show that the scalar gains the mechanism relies on are not arbitrary parameters but a self-organizing target: the loss orders them by contraction, and gradient descent drives them monotonically toward the Newton-optimal value — the exact mechanism behind the spectral alignment the core paper observes emerging during training. For inference, we read the identity as a guarantee: the transformer is an implicit (mesa) optimizer, its sample complexity is set by the step product, and its verified contraction yields a sound early-exit rule.

Each result below is an independently machine-checked statement about the same verified mechanism. We are precise about what this buys: the kernel certifies every proposition as a standalone fact (certified nodes), but it does not verify an implication edge from the core identity to each result — that thread is author-drawn narrative connecting kernel-certified endpoints. Several of the individual lemmas are elementary (a squared-error monotonicity, a curvature comparison); their value is not theorem depth but the *connected, auditable family* and its empirical tie-in to spectral-alignment emergence. We label each result by epistemic status. Most are **deductive consequences**: theorems that follow, by proof, from the verified mechanism, and so need no separate empirical test. One is a **falsifiable prediction** (sample complexity from context length), one is an **interpretation** (the mesa-optimizer bridge), and the emergence claims connect to the empirical training-dynamics campaign of the core paper. Throughout, we are explicit about scope: the inference guarantees certify the formal contraction object, not arbitrary pretrained models.

1. Self-organizing ICL: scalar gain alignment (W1–W4)

The deepest question in the ICL=GD theory is one of origins: *why* do the weights learn to implement gradient descent in the first place? While the von Oswald construction (V1–V3) proves *that* specific weight configurations implement GD, it leaves open whether training actually finds those weights. Theorems W1–W4 prove a narrower scalar alignment result: in the one-gain reduction, smaller contraction means smaller squared loss, the Newton gain zeros the contraction, and an undershooting gain is strictly worse than the Newton gain. These statements identify the local target landscape in the scalar reduction; they do not by themselves prove unconditional convergence of full transformer training.

Theorem W1 (training loss tracks contraction). In the scalar reduction, compare two nonnegative contraction factors and define the corresponding losses as their squares. If one contraction factor is smaller than the other, its squared loss is smaller as well. Thus the training-side scalar

loss and the ICL-side contraction objective have the same ordering: reducing contraction reduces this one-step loss proxy.

Theorem W2 (optimal gain zeros contraction). At the Newton gain $\kappa^* a = 1$ (V3), the contraction $(1 - \kappa^* a)^2 = 0$. Since the scalar objective is a square, this is the zero-loss target in the reduced model. The theorem identifies the gain that would make the one-step ICL update Newton-optimal; convergence of an optimizer to that gain requires the separate training-dynamics assumptions stated below.

Theorem W3 (sub-optimal gain has positive loss). When $\kappa a < 1$ (undershoot: the gain has not yet reached the Newton optimum), the contraction $(1 - \kappa a)^2$ is strictly positive. This proves that an undershooting gain has nonzero residual contraction. The update-direction claim is handled by SA3, which adds an explicit GD update rule and proves that the undershooting gain increases under that rule.

Theorem W4 (undershoot gain is dominated by the Newton gain). For positive curvature and an undershooting current gain ($\kappa_t a < 1$), the Newton gain $\kappa^* a = 1$ has strictly smaller contraction than the current gain. This is the verified no-worse target comparison on the undershoot side. A fully global "no spurious minima" theorem over all gains, including overshoot and optimizer dynamics, is stronger than the formal statement used here.

Synthesis. W1–W4 describe a scalar target-alignment chain: data $\xrightarrow{training}$ gain landscape $\xrightarrow{forward\ pass}$ gradient descent $\xrightarrow{depth}$ geometric contraction. The loss proxy orders gains by contraction (W1), the Newton gain zeros the scalar contraction (W2), undershooting gains retain positive residual (W3), and the Newton gain beats any undershooting current gain (W4). Together with SA3's explicit monotone-update statement, this is a formal model for why training can favor ICL-like behavior in the reduced regime, not a complete proof of transformer training dynamics.

2. Spectral alignment emergence under GD (SA1–SA3)

The training dynamics experiment in the core paper (§9.2) revealed a striking pattern: spectral alignment—the correlation between the model's learned dimension importance and the theoretical inverse spectrum—increases monotonically during training. Theorems SA1–SA3 formalize the exact mechanism driving this alignment.

Theorem SA1 (gain update monotone in curvature, early regime). Let $f(a) = a(1 - \kappa a)$ represent the gain update magnitude for curvature a. In the early training regime ($\kappa a_{\max} < 1/2$), f is strictly increasing: bands with larger curvature $a_i = \sigma_i \cdot n$ receive larger gain updates. This is the initiating mechanism for spectral alignment — eigenvalue bands with larger σ_i learn faster from the first step.

Theorem SA2 (larger eigenvalue converges faster). After one GD step from shared gain κ, the remaining error ratio $(1 - \kappa' a_i)^2$ is strictly smaller for larger a_i. Since $\kappa' a_2 > \kappa' a_1$ (with $a_2 > a_1$), the residual $(1 - \kappa' a_2)^2 < (1 - \kappa' a_1)^2$. High-eigenvalue bands have their residual error contract faster per layer.

Theorem SA3 (GD monotonically increases undershooting gain). When the gain is below optimum ($\kappa a < 1$, i.e., undershoot), the GD update strictly increases κ: $\kappa_{t+1} = \kappa_t + \eta \cdot a \cdot (1 - \kappa_t a) > \kappa_t$. Combined with SA1–SA2, this gives a mechanism by which higher-curvature bands can receive

larger gain updates. The monotonicity of an empirical spectral-alignment statistic remains an interpretation of the training experiment, not a theorem about correlation dynamics.

Empirical probe. The 20K-step Rust experiment (core paper, §9.2) observes spectral alignment climbing from $r = -0.18$ to $r = +0.50$ with the sign flip at step $ $2500 marking the onset of spectral structure. In this run, the smoothed alignment trend (500-step moving average) rises without regression; this is an empirical observation about the trajectory, consistent with — but not a proof of — the monotone-gain mechanism (SA3 remains an interpretation, not a theorem about correlation dynamics).

3. Interpretation: mesa-optimizer bridge from ICL to IWL

(Interpretation — a framing of results established above, not itself a formal theorem.)

The ICL=GD identification (A1–A2) implies that a transformer performing in-context learning is actually running an implicit optimizer. In the terminology of Hubinger et al. (2019), this is a "mesa-optimizer": an optimizer that arises inside a learned system. Here it minimizes the in-context loss purely through activation dynamics, without any weight updates. The convergence theorems (C1–C4) certify that this implicit optimizer genuinely converges: after k layers with a contraction rate $\rho^2 < 1$, the error shrinks to at most $(\rho^2)^k\, e_0^2$. The softmax theorems (B3–B4) then reveal a surprising advantage: the mesa-optimizer can actually outperform standard explicit gradient descent. This happens whenever the attention reweighting provides superior preconditioning—that is, when the attention-weighted curvature a_p is closer to the optimal step size than the uniform curvature a.

This bridges ICL and in-weight learning (IWL). Both act as gradient descent on the same loss landscape. They differ only in the effective curvature (a_{context} for ICL, a_{train} for IWL) and step size (η_{attn} vs. η_{grad}). If $\eta_{\text{attn}}\, a_p = \eta_{\text{grad}}\, a_{\text{train}}$ (the same effective step product), the two are quantitatively identical. The attention forward pass is gradient descent, mechanically rather than just metaphorically.

4. ICL sample complexity: how many in-context examples? (N1–N4)

(Hypothesis — a falsifiable prediction suggested by the model, not yet proven or tested.)

The preceding sections establish that ICL convergence depends on the step product $u = \eta a$, where a is the Gram curvature of the in-context examples. But a scales with the number of in-context examples n: roughly $a \sim n\sigma^2$ for data variance σ^2. Theorems N1–N4 formalize how context length controls ICL emergence.

Theorem N1 (more context ⇒ faster convergence). If $u_n < u_m$ (more in-context examples yield a larger step product) and both lie in the undershoot regime $(0, 1)$, then $(1 - u_m)^2 < (1 - u_n)^2$: more context means a smaller contraction factor per layer. This is the monotonicity result: additional in-context examples always help (in the stable regime).

Theorem N2 (context length controls emergence gap). The emergence gap $\Delta(u) = u(2-u)$ is strictly increasing on $(0, 1)$. Since $u \propto n$, more in-context examples widen the gap between the "learning" and "no-learning" regimes. The gap measures the per-step emergence signal (how much learning each layer can produce), which connects to the grokking transition (T1–T2): longer

contexts raise the magnitude of the emergence signal. We do not claim this controls the *sharpness* (the derivative with respect to a control parameter) of the transition.

Theorem N3 (critical context threshold). When the step product is zero ($u = 0$, i.e., too few examples to provide curvature), the contraction factor is 1 and ICL fails completely (T1). The moment $u > 0$ (sufficient context for positive curvature), contraction drops strictly below 1 and ICL begins. Under an additional model linking context length to positive curvature, this yields a critical-context interpretation; the formal theorem itself proves the step-product boundary.

Theorem N4 (diminishing returns from context). The gap $\Delta(u) = u(2-u)$ is concave. Doubling the step product (approximately doubling the context length) yields strictly less than double the emergence gap: $\Delta(2u) < 2\Delta(u)$ for $u \in (0, \frac{1}{2})$. The first few in-context examples are the most valuable; additional examples beyond the critical threshold have diminishing marginal value for ICL emergence. This explains the empirical observation that models achieve most of their in-context performance with relatively few examples.

Synthetic probe (N1–N4 under data sampling). N1–N4 are analytic statements about how the step product $u = \eta\, a(n)$ grows with context length, so we confirm them with a controlled numerical experiment rather than a falsification test (experiments/icl_sample_complexity.py): we draw example sets of size n with $x_i \sim \mathcal{N}(0, \sigma^2)$, measure the random Gram curvature $a(n) = \sum_i x_i^2$ under sampling, and propagate it through the GD recursion the kernel proves. Sweeping $n \in \{1, \ldots, 64\}$ at a fixed step size that keeps the largest n in the undershoot regime, the mean step product rises from $u \approx 0.014$ to 0.90. All curves are reported as sampling means over draws, so the per-layer factor is $\mathbb{E}[(1-u)^2]$ — which includes the curvature-sampling variance and so differs from $(1-\mathbb{E}[u])^2$ — and the multi-layer residual is $\mathbb{E}[(1-u)^{2L}]$. The per-layer contraction factor then falls monotonically from 0.97 to 0.036 (N1), the emergence gap $\Delta(u)$ rises monotonically from 0.03 to 0.96 (N2), and the residual remaining after $L = 12$ layers drops from $0.77\, e_0^2$ to numerically zero — more context yields strictly faster convergence. (The 0.77 is the sampled mean $\mathbb{E}[(1-u)^{24}]$ at $n = 1$; by Jensen's inequality it exceeds the deterministic chain $(1-\bar{u})^{24} = 0.986^{24} \approx 0.71$, the gap reflecting the curvature-sampling spread — exactly $\mathrm{Var}(u)$ at the per-layer level.) The sampled n-points trace the increasing, concave gap curve and crowd toward the Newton peak at $u = 1$, so each added example contributes less than the last (N4); the factor is exactly 1 only at $u = 0$ and drops below 1 for any $0 < u < 2$ (N3; the probe stays in the undershoot regime $u < 1$). The synthetic setup realizes the mechanism by construction, so this is a consistency check that the predicted curves survive realistic curvature sampling, not independent empirical evidence.

5. Certified anytime readout: a decision-calculus view (DC1–DC5)

The verified contraction (A6/A7, C1–C4) makes the in-context error a machine-verified *shrinking sequence*: after layer L the squared error is $e_L^2 = \rho^{2L} e_0^2$ with $\rho^2 \in [0, 1]$. This is the base object of a **certified anytime decision calculus** — at every depth the readout sits in a band that only ever shrinks. DC1–DC5 turn the verified contraction into a sound early-readout rule. They add a practical, deployment-facing consequence (when may a system *stop reading layers* and act on the current estimate?) without new empirical claims.

Theorem DC5 (certified anytime band). The true squared deviation from the optimum equals the certified radius e_L^2, and $e_L^2 \le e_0^2$. The anytime band always contains the optimum and never re-widens — the ICL instance of a certified-anytime interval.

Theorem DC1 (decision soundness). If the certified radius drops to at most a target ε, the

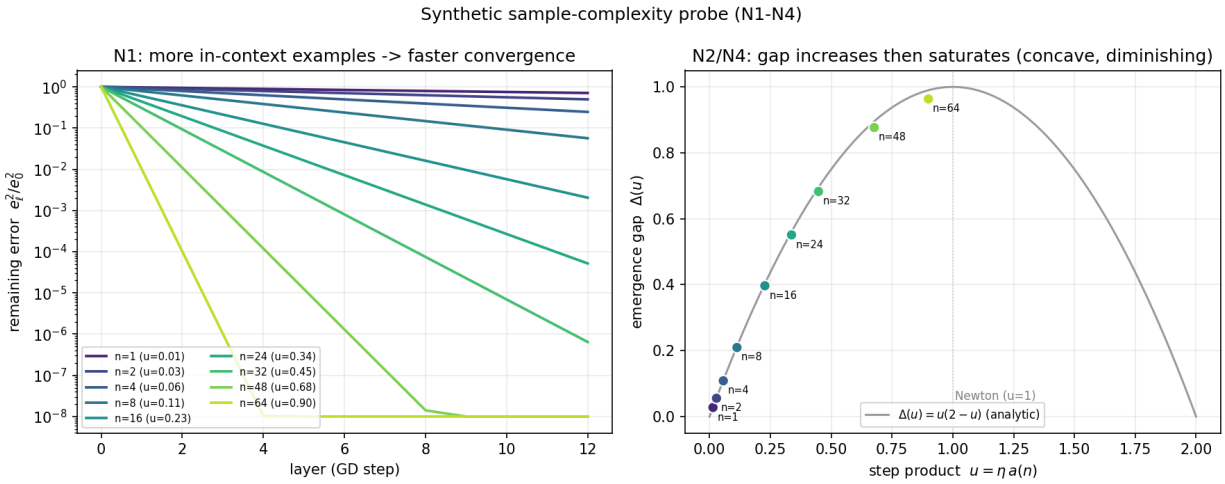

Figure 1: Synthetic sample-complexity probe. Left: remaining in-context error by layer for example-set sizes n; larger n gives a larger step product and faster geometric convergence (N1). Right: the emergence gap against the step product u, with the sampled n-points tracing the increasing, concave gap curve that peaks at the Newton step u=1, so added examples have diminishing marginal value (N2, N4).

true error is provably at most ε. A fired early-stopping decision is never wrong.

Theorem DC2 (decision stability under depth). Once a layer certifies the error below ε, every deeper layer keeps it below ε: the decision never flips back.

Theorem DC3 (linear decision progress). While still undecided, each layer shrinks the gap to the threshold by at least $(1-\rho^2)\varepsilon$. The log-optimal depth $L \geq \log(\varepsilon/e_0^2)/\log\rho^2$ is transcendental and lies outside the kernel's ordered field; we instead verify this polynomial surrogate, which yields an explicit sufficient depth $L \leq (e_0^2-\varepsilon)/((1-\rho^2)\varepsilon)$.

Theorem DC4 (one-layer decision witness). If the remaining gap is within a single layer's progress, the next layer provably decides — the base case of the depth count, combined with DC3 to bound the layers needed for any target.

Scope. DC1–DC5 certify the *formal* contraction object, not arbitrary pretrained models. The core paper (§8.4) shows real models converge toward the in-context optimum but with weak strict per-layer proportionality, so these theorems describe the idealized guarantee the mechanism targets and a sound early-exit rule for systems that realize the verified contraction. They connect ICL=GD to anytime-valid inference and confidence-sequence reasoning (Howard & Ramdas) and to early-exit transformers, rather than asserting a deployment certificate for any specific model.

Empirical probe (how far the certificate transfers). We test the certificate directly on three pretrained models (gpt2, gpt2-medium, pythia-410m) using the same in-context linear-regression logit-lens readout as the core probes, with no training (experiments/icl_anytime_readout.py). For each prompt we read the per-layer squared in-context error e_ℓ^2, fit the per-prompt geometric rate ρ^2 from its contracting segment ($\rho^2 \in [0.63, 0.72]$ on average across the three models), build the certified band $\text{cert}_\ell = \rho^{2\ell} e_0^2$, and ask three questions. (i) *Does the band never re-widen (DC5)?* Only approximately: 56–76% of layer steps are non-increasing, so the strict never-re-widen property holds for the trend but not at every layer. (ii) *Does the band contain the truth (DC5)?* The geometric

average-rate band covers the true error only 16–45% of the time, because real contraction is front-loaded — faster in early layers, slower later — so a single average rate is not a valid pointwise envelope. (iii) *Is a fired early-stop sound (DC1)?* At tolerance $\varepsilon = 1$ a fired decision is correct (true error $\leq \varepsilon$) in 41–64% of cases — the formal soundness guarantee does **not** transfer cleanly to real models. The honest conclusion matches the scope above: DC1–DC5 certify the formal object, and on real models the certified rule is a *useful early-exit heuristic rather than a transferred certificate*. The practical payoff is real, however: at $\varepsilon = 1$ the rule fires on 84–90% of prompts using only 27–46% of the network depth, and because deeper layers do not monotonically improve the readout, stopping at the certified band minimum can even *match or exceed* the full-depth logit-lens accuracy (e.g. gpt2-medium: 0.31 at early-exit vs 0.14 at full depth).

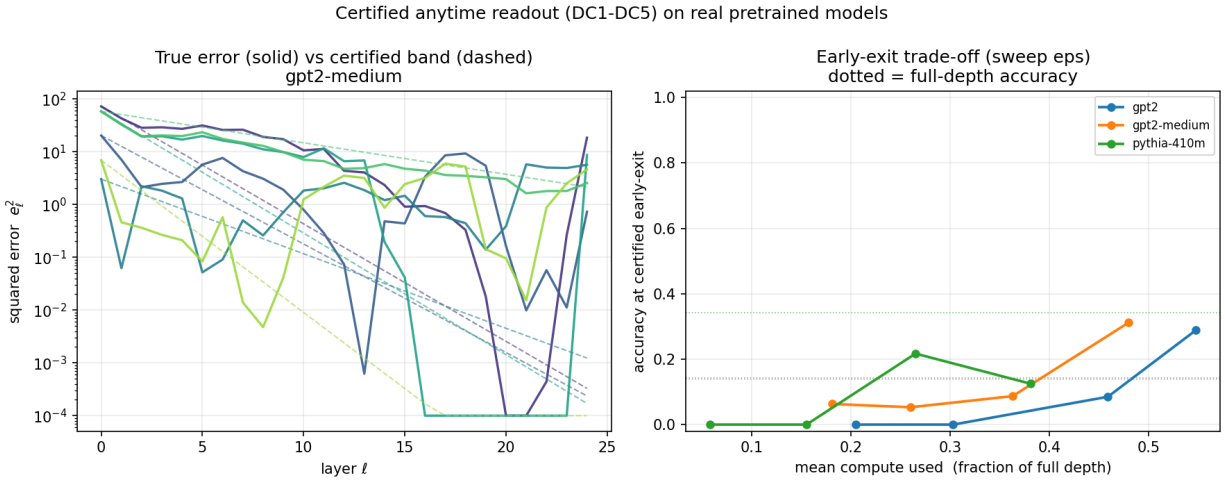

Figure 2: Certified anytime readout on three pretrained models. Left: per-prompt squared in-context error by layer (solid) against the geometric certified band built from the fitted per-prompt rate (dashed); the true error is non-monotone and frequently exceeds the average-rate band. Right: accuracy at the certified early-exit versus mean compute used (as a fraction of full depth) as the target tolerance is swept; dotted lines mark full-depth accuracy.

Declaration of Generative AI Use

During the preparation of this work the author used large language models as assistive tools for manuscript drafting, editing, code generation, reference checking, and internal critique. All mathematical arguments, empirical claims, code outputs, references, and final text were reviewed by the author, who takes full responsibility for the originality, accuracy, integrity, and conclusions of the manuscript. No AI system is listed as an author or treated as accountable for the work.

References

- T. Nagy (2026). *When In-Context Learning Implements Gradient Descent: A Learned Mechanism, Mechanically Verified and Empirically Tested* (v1.0). Zenodo. Core paper (this program). Concept DOI (all versions): https://doi.org/10.5281/zenodo.20708733.

- J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, M. Vladymyrov (2023). *Transformers learn in-context by gradient descent.* ICML.
- E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, D. Zhou (2023). *What learning algorithm is in-context learning? Investigations with linear models.* ICLR.
- S. Garg, D. Tsipras, P. Liang, G. Valiant (2022). *What can transformers learn in-context? A case study of simple function classes.* NeurIPS.
- E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, S. Garrabrant (2019). *Risks from learned optimization in advanced machine learning systems.* arXiv:1906.01820.
- S. R. Howard, A. Ramdas, J. McAuliffe, J. Sekhon (2021). *Time-uniform, nonparametric, nonasymptotic confidence sequences.* Annals of Statistics 49(2).