Architectural Optimizations of the In-Context GD Mechanism

A satellite of the verified ICL=GD core: attention sinks, chain-of-thought, and multi-head structure as optimality conditions of an implicit second-order optimizer

Dr. Tamás Nagy

tnagyphd@gmail.com

Working Paper • 2026-06-15

Build-std v2.0 | 23:27 | 9f25aa4a

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 reads the verified identity as a design principle: it asks which architectural features of real transformers are *optimality conditions* of the implicit optimizer, and which are not. Treating the ICL=GD construction as a fixed premise, we derive a connected family of results — each named deductive theorem stated and checked in the same proof framework as the core identity, by the same independent proof kernel, rather than asserted (the interpretation and empirical claims are labelled as such). The satellite results are independent algebraic consequences inside the shared ICL=GD abstraction, not re-derivations of Theorems A–D. We first recast softmax attention as approximate Newton's method: the error is quadratic in the preconditioning distance and decays geometrically over depth. From this vantage, attention sinks become Newton-optimal regularization (under the null-direction idealization) — allocating weight to a null token reduces effective curvature to exactly the one-step Newton value in the overshoot regime — and chain-of-thought becomes heterogeneous preconditioned multi-step descent, where adaptive steps never do worse than uniform ones. We then turn the same lens on multi-head attention: multiple matched per-channel gains overcome the single-gain Newton barrier as geometric preconditioning, with a certified per-head-doubling improvement that composes into the standard logarithmic head-count argument. Crucially, we report an *honest negative result*: across a confound-controlled panel of pretrained models, head count is not an independent predictor of in-context accuracy once model scale is controlled for — they do not appear to spectrally specialize — separating the genuine architectural consequences from the spurious ones. The complementary forced-specialization ablation is the core paper's theory-guided architecture ablation, our empirical anchor here. Most results are machine-checked algebraic consequences of the optimizer abstraction's definitions; the ICL=GD construction motivates them but is not a logical premise of the arithmetic.

Overview

The companion core paper proves, and machine-checks, that a transformer's forward pass can perform one step of gradient descent on a least-squares problem it builds from its own context.

This paper takes that verified mechanism and asks a designer's question: which familiar features of transformer architecture are the *optimizer's* doing, and which are not?

Reading attention as an implicit optimizer gives a clean account of three much-discussed design features. Attention sinks look like optimal regularization — spending attention on an uninformative token to avoid overshooting. Chain-of-thought looks like multi-step descent, where each reasoning step can carry its own step size. And multi-head attention looks like geometric preconditioning, where different heads cover different parts of the curvature spectrum. A fourth prediction does not survive the data: across a panel of pretrained models, head count stops predicting in-context accuracy once model size is controlled for — pretrained models do not appear to spectrally specialize. (The complementary *forced*-specialization ablation, where specialization is imposed by hand, lives in the core paper, our empirical anchor.) We report this honest negative result alongside the positive ones.

The contribution is as much the discipline as the readings. Every named deductive theorem is checked by an independent proof kernel, and throughout we mark explicitly where an idealized scalar or per-channel result is being *interpreted* as architecture rather than *proven* about real models. The empirical anchor is the core paper's theory-guided architecture ablation and a confound-controlled model panel.

Introduction

With the picture sketched, we make the optimizer reading precise.

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: *if the forward pass is an implicit optimizer, which architectural features are its optimality conditions?* Many empirically discovered "tricks" of transformer design — attention sinks, chain-of-thought prompting, multi-head attention — have been justified post hoc. Read through the verified identity, several of them are not tricks at all but the structure an implicit second-order optimizer must adopt. We make that precise: we recast softmax attention as approximate Newton's method, derive attention sinks as Newton-optimal regularization, derive chain-of-thought as preconditioned multi-step descent, and analyze multi-head attention as geometric preconditioning whose certified per-doubling improvement composes into the standard logarithmic head-count argument.

The discipline of the verified-mechanism view is that it also rules things *out*. We report an honest negative result — across a confound-controlled panel of pretrained models, head count is not an independent predictor of in-context accuracy once model scale is controlled for, so these models do not appear to spectrally specialize — to separate the architectural consequences the mechanism genuinely implies from those it does not. (The complementary forced-specialization ablation, which imposes the specialization by hand, is the core paper's; we use it as our anchor rather than

re-running it here.) Each result below is a machine-checked algebraic consequence of the optimizer abstraction's definitions ($\rho = 1 - \kappa a$ and the per-channel reductions), checked by the same independent proof kernel — the ICL=GD identity motivates those definitions but is not a logical premise of the algebra (the MH inequalities, for instance, are true of the rational functions regardless). We label each by epistemic status: most are **deductive consequences** (theorems that need no separate test); a few are **interpretations** (framings of the verified results), and the negative result is **empirical**, anchored to the core paper's theory-guided architecture ablation.

1. Attention as approximate second-order optimization

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

Theorems B3–B7 and F1–F2 support a reinterpretation of softmax attention as an *approximate Newton's method*. Classical Newton's method achieves one-step exact convergence by using the inverse Hessian as preconditioner ($\eta\, a_p = 1$). Theorem B7 proves that this property holds for softmax attention at the optimal preconditioning. The gap between softmax and exact Newton is controlled by $|v - 1| = |\eta\, a_p - 1|$: the contraction factor $(1 - v)^2 = (v - 1)^2$ is *quadratic* in the distance from optimal preconditioning.

This quadratic error has two consequences. First, small preconditioning errors are cheap — the convergence penalty is second-order in $|v - 1|$. Second, multi-step chains amplify the advantage. By Theorem F2 (mixed preconditioning helps), each CoT token that achieves sub-unit contraction *strictly* reduces the total error. By Theorem F1 (adaptive beats uniform), the heterogeneous chain is at least as good as the best uniform chain.

The combined picture: a transformer with k layers of softmax attention runs k steps of *adaptive, approximately second-order* optimization on the in-context loss. If each layer tunes its preconditioning within ε of optimal ($|v_i - 1| \leq \varepsilon$), the k-layer contraction is at most ε^{2k} — for $\varepsilon < 1$ this is *geometric (contractive) over depth* and *quadratic in the preconditioning error* ε. This provides a mechanistic explanation for why transformers can solve in-context tasks with remarkably few layers.

2. Full softmax mechanism: attention sinks and Newton convergence (SM1–SM5)

B1–B7 prove what softmax preconditioning DOES (bounds, acceleration, amplification). SM1–SM5 show why several observed attention patterns are *optimizer-optimal within the abstraction* — connecting phenomena like attention sinks, induction heads (Olsson et al., 2022), and depth scaling to optimality conditions of the ICL=GD mechanism. They are statements about what the idealized optimizer would do; they do not prove that real transformers develop these patterns by this mechanism. (Scope of verification: SM1–SM5 are verified in the Platonic kernel but, unlike the F- and MH-groups, are not yet exported to Lean — this is the reason lean_verified is partial.)

Theorem SM1 (attention sink achieves Newton step). When the softmax-weighted curvature overshoots ($\eta a_p > 1$), the model can allocate weight s to a null token, reducing effective curvature to $(1-s)a_p$. At the optimal sink fraction ($\eta(1-s)a_p = 1$), the contraction is exactly zero — one-step Newton convergence. The certified content is really an *existence* result: the Newton-zero step follows from $\eta(1 - s)a_p = 1$ alone, and the overshoot hypothesis $\eta a_p > 1$ is what forces the optimal fraction $s \in (0, 1)$ to exist strictly inside the simplex (the genuinely informative part),

rather than the zero step being a separate fact. The reach of this is set by its assumption: *within the ICL=GD abstraction, when sink mass behaves like allocation to a low-curvature / null direction*, it gives a reason attention sinks can be useful in the overshoot regime — they would implement optimal regularization rather than being mere artifacts. We do not claim real sink tokens are shown to be null directions; SM1 is the algebra of that idealization, and the empirical signal below is consistent with, not a proof of, it. Empirically, GPT-2's later-layer ablation-sensitive heads (L5–L6) show 55–77% sink fractions (core paper, §9.3), consistent with this mechanism.

Theorem SM2 (near-Newton quadratic rate). If the preconditioning error satisfies $|1-\eta a_p| \leq \delta$, then contraction $\leq \delta^2$. Small deviation from optimal preconditioning gives quadratically small residual per step. This is the Newton convergence rate and explains why softmax attention (which can tune a_p close to $1/\eta$) gives qualitatively faster ICL than linear attention.

Theorem SM3 (softmax advantage compounds, two-layer case). The verified statement proves the two-layer compounding case: if softmax improves the per-layer contraction ratio by a factor $c < 1$, then two such steps improve by c^2. The displayed c^K law is the standard informal induction under a fixed per-layer ratio, not a separately formalized arbitrary-K theorem.

Theorem SM4 (sink hurts in undershoot). In the undershoot regime ($0 < \eta a_p \leq 1$), any sinking strictly increases contraction — attention should go entirely to informative tokens. Combined with SM1, this gives the local overshoot/undershoot characterization: a properly tuned sink can help only in the overshoot regime, and in undershoot any sink hurts. (This is a one-sided pairing of SM1 and SM4, not a full biconditional for arbitrary sink mass.) Empirically, GPT-2's Layer 0 heads (which see raw input, undershoot regime) show only 8–25% sink fractions, consistent with the undershoot side of the theorem.

Theorem SM5 (multi-head product zeroes when a head reaches Newton). The verified statement is the product-zero algebra behind joint convergence: if total contraction is the product of nonnegative per-head contractions and one head has zero contraction, then the total contraction is zero. Interpreted through SM1, a head can reach zero contraction when its effective softmax-preconditioned curvature satisfies the Newton condition. This connects MH1–MH5 with the softmax mechanism, while stopping short of a separate theorem that every head independently learns or reaches $\eta a_h = 1$.

3. Chain-of-thought as preconditioned multi-step GD

Chain-of-thought (CoT) prompting provides multiple reasoning tokens. Within the ICL=GD abstraction we model CoT as a *sequence of effective optimization steps*, whose preconditioners may vary across positions, layers, and heads. The multi-step composition (C1–C4) then applies with heterogeneous preconditioners per step, so each effective step can carry a different contraction rate. We stress the level of the claim: the formal F1/F2 results prove only the heterogeneous two-step contraction *algebra*. The reading of one reasoning token as one optimization step — rather than the literal architectural fact, where every token position is processed by all heads in a layer — is the interpretation this abstraction licenses, not a theorem that one head is assigned to one CoT token.

Theorem F1 (adaptive beats uniform two-step). If two heterogeneous adaptive steps have contraction rates ρ_a^2, ρ_b^2 each $\leq \rho_{\text{ref}}^2$ (the uniform rate), their product satisfies $\rho_a^2 \rho_b^2 \leq \rho_{\text{ref}}^4$. Adaptive preconditioning across CoT tokens never performs worse than uniform attention.

Theorem F2 (mixed preconditioning helps). If one CoT step achieves a strictly sub-unit contraction ($\rho_b^2 < 1$), the two-step product satisfies $\rho_a^2 \rho_b^2 < \rho_a^2$. In the two-step optimizer abstraction, an additional contracting effective step strictly improves the bound. This establishes the mechanistic basis for why the ICL=GD abstraction predicts a benefit from additional contracting reasoning steps: every effective step that contracts the error makes a non-zero contribution. (This is a prediction of the optimizer model, not a measured claim about real chain-of-thought reasoning.)

4. Multi-head spectral specialization (MH1–MH8)

Theorem M3 proves that a single scalar gain cannot achieve Newton convergence when eigenvalues differ. Theorems MH1–MH5 prove the converse: multiple per-channel gains *can* overcome this limitation. What the kernel certifies is conditional and per-channel — *if* a head's step product is matched to a channel, that channel's contraction is zero — not that real heads learn, or are assigned, eigenvalue bands. Interpreted architecturally, this is the idealized role multi-head specialization *would* play: one head's gain matched to one eigenvalue band. Theorems MH6–MH8 then turn the qualitative "multiple matched gains help" into a *quantitative rate law* in the spectral condition number — the place in this paper where the kernel certifies genuine nonlinear inequalities (MH6–MH7) rather than only degree-≤ 2 identities, with MH8 the optimal-gain identity that grounds the rate.

Theorem MH1 (per-head Newton achievable). If a head's step product satisfies $\kappa_i\, a_i = 1$, the contraction for that channel is exactly zero: $(1 - \kappa_i\, a_i)^2 = 0$. Each head independently attains Newton's-method one-step convergence for its assigned eigenvalue.

Theorem MH2 (two matched heads zero total contraction). When both heads achieve Newton for their respective channels ($\kappa_1\, a_1 = 1$ and $\kappa_2\, a_2 = 1$), the total two-channel contraction is exactly zero: $(1 - \kappa_1 a_1)^2 + (1 - \kappa_2 a_2)^2 = 0$. Compare with M3: a *single* gain cannot zero both channels. Two matched heads *can*. This is the fundamental multi-head theorem.

Theorem MH3 (single-head residual = gain^2 $\times$ spread^2). With a single gain κ at Newton for channel 1 ($\kappa\, a_1 = 1$), the channel-2 contraction decomposes as $(1 - \kappa\, a_2)^2 = \kappa^2\, (a_1 - a_2)^2$. The single-head penalty is the product of the gain scale (κ^2) and the eigenvalue spread ($(a_1 - a_2)^2$). This algebraic identity — which follows from $1 - \kappa a_2 = \kappa(a_1 - a_2)$ when $\kappa a_1 = 1$ — connects the multi-head advantage directly to the spectral geometry.

Theorem MH4 (multi-head advantage is positive). The quantity $\kappa^2\, (a_1 - a_2)^2$ from MH3 is strictly positive when $\kappa > 0$ and $a_1 \neq a_2$. Combined with MH2 (two heads achieve contraction 0) and MH3 (single head leaves residual $\kappa^2(a_1 - a_2)^2$), the multi-head advantage equals $\kappa^2(a_1 - a_2)^2 > 0$. Every unit of eigenvalue spread that a second head covers is strictly valuable.

Theorem MH5 (shared gain has irreducible floor). For *any* single shared gain κ in the bracketing regime ($\kappa\, a_1 \leq 1 \leq \kappa\, a_2$), the total contraction satisfies $(1 - \kappa a_1)^2 + (1 - \kappa a_2)^2 \geq \frac{1}{2}\, \kappa^2\, (a_2 - a_1)^2$. No single-gain choice can reduce total contraction below this eigenvalue-spread floor. Per-channel gains (MH2) achieve 0, while the best single gain is bounded away from 0 by the spectral spread. This proves that multi-head attention's advantage over single-head is *structural*, not merely parametric.

The quantitative rate law (MH6–MH8). MH1–MH5 say multi-head attention helps; MH6–MH8 say *by how much*. Fix a band of eigenvalues with condition ratio $r = \lambda_{\max}/\lambda_{\min} > 1$. The best a single gain can do on that band is the classical steepest-descent worst-case contraction

$\phi(r) = \left(\frac{r-1}{r+1}\right)^2$ (MH8 grounds this ϕ in the actual optimal gain $\kappa^* = 2/(\lambda_{\min} + \lambda_{\max})$). In the idealized geometric-allocation picture, assigning one head per *geometric* sub-band replaces the effective condition ratio r by $r^{1/H}$ for H heads; one doubling replaces s^2 by s. MH6–MH8 certify the per-doubling pieces of this picture — band monotonicity (MH6), the single-doubling inequality (MH7), and the optimal-gain endpoint (MH8) — and, unlike the rest of the formal core, MH6–MH7 are genuine nonlinear inequalities (the kernel certificate is a nonlinear-arithmetic derivation, not a ring identity), while MH8 is the optimal-gain identity that ties the rate ϕ to the GD step product. The full H-head/$\log C$ count is the standard composition of those pieces, not a separate certificate.

Theorem MH6 (rate is monotone in conditioning). For $1 < r_1 < r_2$, the steepest-descent contraction is strictly worse for the worse-conditioned band: $\phi(r_1) < \phi(r_2)$, i.e. $(r_1-1)^2(r_2+1)^2 < (r_2-1)^2(r_1+1)^2$. Worse spectral conditioning means strictly slower in-context convergence — the per-band quantitative content of the condition-number story.

Theorem MH7 (geometric head-doubling strictly improves). For $s > 1$, splitting one head over a band of condition ratio s^2 into two heads over geometric sub-bands of ratio s each strictly lowers the worst-case contraction: $\phi(s^2) > \phi(s)$, i.e. $(s-1)^2(s^2+1)^2 < (s^2-1)^2(s+1)^2$. Each doubling of heads replaces the effective condition number C by $\sqrt{C}$. What the kernel certifies here is the *single* doubling inequality $\phi(s^2) > \phi(s)$ (each doubling replaces the effective condition number C by $\sqrt{C}$). The iterated H-head statement — H geometrically-allocated heads give $\phi\big(C^{1/H}\big)$, with $\phi\big(C^{1/H}\big) \to 0$ as $H \to \infty$, so reaching a target rate needs $H \sim \log C$ heads — is the textbook composition of that certified one-step inequality, *not* a separately machine-checked corollary or limit. With that caveat, multi-head attention reads as *geometric preconditioning*.

Theorem MH8 (the optimal gain attains the rate). The band-optimal shared gain κ^*, characterized by $\kappa^*(a_1 + a_2) = 2$, makes the squared endpoint contraction equal the condition-number rate exactly: $(1 - \kappa^* a_1)^2\,(a_1 + a_2)^2 = (a_2 - a_1)^2$, i.e. $(1 - \kappa^* a_1)^2 = \phi(a_2/a_1)$. (At κ^* the two endpoint residuals are equal and opposite, $1 - \kappa^* a_1 = -(1 - \kappa^* a_2)$, the textbook minimax-gain condition.) This ties the abstract rate ϕ in MH6–MH7 to the concrete GD step product of the mechanism, closing the loop from the von Oswald gain to the convergence rate.

Together MH6–MH8 upgrade the multi-head result from existence (MH2: two heads *can* zero a two-channel contraction) to the *ingredients* for the standard logarithmic scaling argument: the kernel certifies band monotonicity, the single geometric head-doubling inequality $\phi(s^2) > \phi(s)$, and the optimal-gain endpoint κ^*. Composing those doublings — the step that yields "a band of condition number C needs $\sim \log C$ geometrically-tuned heads" — is the textbook argument built on top of the certified inequality, not itself a kernel theorem. This is the quantitative converse of the single-gain Newton barrier (M3 in the core paper; MH5 above) and the deepest single statement in the formal core; we flag honestly that the remaining theorems are mostly elementary identities, and that MH6–MH7 — while genuine inequalities (MH8 being the optimal-gain identity that anchors the rate) — are still about the idealized per-band mechanism, whose contact with real heads is exactly the open L1/L3 link of the core paper's §8.

Empirical status (retained panel, confound-controlled). The panel, the per-model contraction fits, and the partial-correlation procedure are the core paper's logit-lens probe experiments/icl_convergence_probe.py. A model has a *valid geometric fit* when its measured per-layer contraction sequence admits a geometric (constant-rate) decay fit with goodness R^2 above the retained-panel threshold; models without such a fit are excluded so the correlations are not driven by noise. This exclusion is by fit quality, not by head count, so it is orthogonal to the tested variable; still, with $n = 11$ it could in principle bias the partial correlation, and the

head-count finding below should be read as "absent on the fit-reliable subset," with the raw-panel robustness check left to the core paper's released analysis. On the retained panel (11 models with valid geometric fits, spanning 70M–1.1B parameters across GPT-2, Pythia (Biderman et al., 2023), Qwen, and TinyLlama/Qwen-style small causal LMs), the raw correlation between head count and ICL accuracy is moderate ($r = 0.53$). However, **after controlling for model size via partial correlation**, the head-count effect vanishes: $r_{\text{partial}}(H, \text{acc} \mid \log \text{params}) = 0.08$, while model size remains strongly predictive: $r_{\text{partial}}(\log \text{params}, \text{acc} \mid H) = 0.86$.

This means the raw head-count correlation is explained by model size in this panel — larger models have both more heads and better ICL, but head count is not an independent predictor here. Here rank_τ is the *effective curvature rank* — the number of data-Gram eigenvalues above a relative threshold τ, i.e. the count of curvature bands that need covering (as defined in the core paper) — so the *coverage ratio* H/rank_τ is heads per band, and its Newton point $H/\text{rank}_\tau = 1$ (one matched head per band) is where the MH2 per-channel construction first becomes attainable. All models in the panel have $H/\text{rank}_\tau \approx 0.016$, far below 1.0, consistent with current architectures operating deep in the spectrally undercovered regime (MH5).

Interpretation. MH1–MH5 are *normative* results (what multi-head attention *could* achieve if heads specialized spectrally) rather than *descriptive* of current models. The empirical finding is that pretrained transformers achieve ICL primarily through scale (more parameters and layers), not through spectral head specialization. This identifies a concrete architectural inefficiency: models *could* exploit per-head spectral tuning (in the idealized matched-channel model, MH2 guarantees zero contraction), but apparently do not. Whether this is due to training dynamics, initialization, or optimization landscape properties is an open question that W1–W4 partially address (the training landscape is benign for a *single* gain, but the multi-head landscape remains uncharacterized).

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.

- C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, C. Olah (2022). *In-context learning and induction heads.* Transformer Circuits Thread.
- S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, O. van der Wal (2023). *Pythia: A suite for analyzing large language models across training and scaling.* ICML.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever (2019). *Language models are unsupervised multitask learners.* OpenAI Technical Report.
- J. Bai et al. (2023). *Qwen Technical Report.* arXiv:2309.16609.
- P. Zhang, G. Zeng, T. Wang, W. Lu (2024). *TinyLlama: An open-source small language model.* arXiv:2401.02385.