DELTAMOMENTUM: A Key-Value based Anisotropic Momentum Update via Delta Rule

Euijin Hong EHONG@ANDREW.CMU.EDU
Carnegie Mellon University

Guannan Qu GQU@ANDREW.CMU.EDU
Carnegie Mellon University

Abstract

Advances in optimizers have been central to training large-scale neural networks, yet every modern optimizer still forms its momentum as an exponential moving average (EMA) of past gradients, forgetting every direction at one fixed rate regardless of how often it is seen. Recent methods address the anisotropy of training only indirectly, wrapping extra processing around the momentum buffer while leaving the EMA update itself, the true source of this uniform forgetting, unchanged. We propose DeltaMomentum, which instead builds direction-awareness into the update rule. The gradient of a linear layer already splits into an input that acts as a key and an output-side error that acts as a value, and DeltaMomentum updates the momentum buffer by the classical delta rule, the standard update for key-value memories, so each direction is forgotten at a rate set by how often it appears. We prove that DeltaMomentum is a valid momentum for gradient descent, that it performs the input-side curvature correction of curvature-based methods with no matrix inversion, and that it clears stale directions faster than EMA both at a fixed optimum and under a drifting target. It is also a drop-in replacement for the momentum buffer in any optimizer, with a single added setting that transfers across model widths without re-tuning. On FineWeb-Edu language-model pretraining, DeltaMomentum reaches AdamW's validation loss in up to 21% fewer steps at 67M and 21.55% at 370M and continues to a lower final loss, and training-time diagnostics confirm that the predicted mechanism, better gradient tracking and healthier input directions, drives the gain.

1. Introduction

First-order momentum-based optimizers [6, 7, 12] drive modern deep-network training, with the near-universal choice being an exponential moving average (EMA) of stochastic gradients as the first-moment buffer:

$$M_t^{\text{EMA}} \;=\; \beta\, M_{t-1}^{\text{EMA}} \;+\; (1-\beta)\, g_t, \qquad \beta \in [0,1). \tag{1}$$

A central but underexploited structural fact reframes this update. For any linear layer $y = Wx$ with $W \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, and output-side error $\delta = \partial\mathcal{L}/\partial y \in \mathbb{R}^m$, the per-sample gradient factorizes as a rank-1 outer product

$$g_t \;=\; \delta_t\, x_t^\top \;\in\; \mathbb{R}^{m\times n}, \tag{2}$$

. An LLM was used as a drafting and editing aid; all theoretical claims, derivations, and experiments are the authors' own and were human-verified.

© E. Hong & G. Qu.

where x_t is naturally a *key* (*where* signal arrived in input space) and δ_t a *value* (*what* correction to apply there). EMA collapses this factored structure into an unstructured matrix average, imposing the same scalar decay β along every eigendirection.

The cost is misalignment with the anisotropic geometry of deep-network training. Hessian eigenspectra and activation covariances both concentrate a small bulk of large eigenvalues followed by a long tail spanning orders of magnitude [2, 3, 9, 10, 13]. The eigenvalues of the input-feature covariance $\Sigma_x = \mathbb{E}[xx^\top]$ measure the *directional density* of the input distribution: how often the stochastic process queries each direction. EMA's uniform horizon is suboptimal at both ends: along high-density directions, stale gradients are retained at full weight when fresh ones arrive every few steps; along low-density directions, scarce signal is averaged on the same short horizon as dense, fast-changing ones. K-FAC [8], Shampoo [4], SOAP [16], and Muon [5] all address this mismatch through external preconditioning or post-processing of an isotropic EMA buffer; the buffer update itself remains direction-blind.

Contribution. We take a complementary route: rather than preconditioning or post-processing an isotropic momentum buffer, we redesign the momentum update so its dynamics natively respect the rank-1 structure of the layer gradient and the anisotropy of input statistics. We propose DELTAMOMENTUM, derived from the classical delta rule [17], which treats M as an online linear associative memory storing (x_t, δ_t) pairs. A single bridging identity—*anisotropic memory transport* (Lemma 1)—governs the entire dynamics and yields, as corollaries: (i) a Tikhonov-regularized Wiener fixed point, equivalent to implicit input-side natural-gradient preconditioning at no covariance-inversion cost (Theorem 2); (ii) strictly faster tracking-error contraction along every positive-density direction under non-stationary distribution shift (Theorem 4); and (iii) width-invariance of the delta coefficient η under μP. Empirically, on FineWeb-Edu pretraining, DeltaAdamW reaches matched validation-loss levels in up to $21\%/21.55\%$ fewer steps and converges to a lower terminal loss at 67M/370M parameters, and per-layer mechanistic diagnostics confirm the predicted dynamical signatures.

2. DELTAMOMENTUM: An Online Associative Memory of Factored Gradients

Update rule. The proposed update is

$$\boxed{M_t \;=\; \beta\, M_{t-1} \;+\; \eta\,(\delta_t - M_{t-1}x_t)\,x_t^\top,} \qquad \beta \in [0,1),\;\; \eta \in (0,1]. \tag{3}$$

Setting $\beta = 1$ recovers the pure delta rule [17]. Factoring out M_{t-1} yields the form used throughout the dynamical analysis:

$$M_t \;=\; M_{t-1}\big(\beta I_n - \eta\, x_t x_t^\top\big) \;+\; \eta\, \delta_t x_t^\top. \tag{4}$$

The right-multiplying operator $\beta I - \eta x_t x_t^\top$ is a *data-dependent* forgetting factor contracting directions aligned with the current key and leaving orthogonal directions essentially untouched; $\eta\, \delta_t x_t^\top$ injects the new association. Every dynamical property below is a property of this operator.

Rationale. The classical delta rule [14, 17, 19], given a target value v at key k, performs $M \leftarrow M + \eta(v - Mk)k^\top$—exactly one online stochastic-gradient step on the function-space prediction loss $\frac{1}{2}\|v - Mk\|^2$. Identifying $v_t \leftrightarrow \delta_t$, $k_t \leftrightarrow x_t$ and augmented with a uniform decay β gives (3). Thus DELTAMOMENTUM stores associations (x_t, δ_t) in M and refreshes them via the delta rule, with $\beta I - \eta x_t x_t^\top$ erasing stored content along the current key's direction before writing the new

association. The novelty is the *timescale* at which the rule acts: linear-attention and fast-weight programmers [14, 19] run the delta rule as a forward-pass sequence memory updated at inference time, whereas DELTAMOMENTUM runs it on the optimizer's first-moment buffer updated at training time, applying the same associative-memory primitive to gradient transport rather than token mixing.

Modular instantiations and μP width transfer. DELTAMOMENTUM is a drop-in replacement for the first-moment buffer of any base optimizer: *DeltaSGD* applies (3) followed by $\theta_{t+1} = \theta_t - \alpha M_t$, while *DeltaAdamW* applies (3) to the first moment and leaves the second moment, bias correction, and decoupled weight decay of AdamW unchanged. We deploy the *normalized-key* variant $\hat{x}_t = x_t/\|x_t\|$ for input-scale-decoupled stability. Normalization leaves every dynamical result of Section 3 intact, since each holds verbatim with the normalized-key covariance $\Sigma_{\hat{x}} = \mathbb{E}[\hat{x}\hat{x}^\top]$ substituted for Σ_x; its purpose is instead to make the μP transfer width-invariant. Concretely $\|\hat{x}_t\hat{x}_t^\top\|_{\mathrm{op}} = 1$ independent of layer width and $\operatorname{tr}\Sigma_{\hat{x}} = 1$, so under μP [18] the recurrence admits a fixed-point per-coordinate scale of $\Theta(n^{-1/2})$, the EMA scale the base optimizer's μP rule expects, if and only if $\eta = \Theta(1)$ (Appendix A); the base μP learning-rate rule is therefore inherited unchanged. The high-dimensional consequence is zero-shot width transfer of η, verified by coordinate checks at widths $\{128, 256, 512, 1024, 2048\}$ (Appendix A).

3. Dynamics: Anisotropic Memory Transport

A single bridging lemma unifies the steady-state, finite-time, and non-stationary behavior of (3).

The bridging lemma. Iterating (4) from $M_0 = 0$ gives a closed-form expansion of M_t as past error–key outer products transported forward in time by a product of data-dependent contractions:

$$M_t \;=\; \eta \sum_{\ell=1}^{t} \delta_\ell\, x_\ell^\top \underbrace{\prod_{j=\ell+1}^{t} \big(\beta I_n - \eta\, x_j x_j^\top\big)}_{P_{\ell\to t}\,\in\,\mathbb{R}^{n\times n}}, \qquad P_{t\to t} = I_n. \tag{5}$$

Each $\eta\,\delta_\ell x_\ell^\top P_{\ell\to t}$ is the *fading memory* of the association stored at step ℓ. Along directions orthogonal to every subsequent key, $P_{\ell\to t}$ reduces to $\beta^{t-\ell} I$, a uniform geometric decay; along directions aligned with subsequent keys, each factor $-\eta x_j x_j^\top$ adds a direction-selective attenuation that erases content conflicting with the newly arriving key.

Lemma 1 (Anisotropic memory transport) *Treat $\{x_j\}_{j>\ell}$ as drawn independently from a distribution with zero mean and covariance $\Sigma_x = \mathbb{E}[xx^\top] = \sum_i \lambda_i u_i u_i^\top$. Then*

$$\mathbb{E}[P_{\ell\to t}]\, u_i \;=\; (\beta - \eta\lambda_i)^{t-\ell}\, u_i \;=\; \tilde{\beta}_i^{\,t-\ell}\, u_i, \qquad \tilde{\beta}_i := \beta - \eta\lambda_i. \tag{6}$$

Proof By independence, $\mathbb{E}[P_{\ell\to t}] = \prod_{j=\ell+1}^{t} \mathbb{E}[\beta I - \eta x_j x_j^\top] = (\beta I - \eta\Sigma_x)^{t-\ell}$; eigendecomposing Σ_x gives (6). ■

The independence in Lemma 1 is a mean-field idealization of the correlated, θ-dependent key stream, under which the expected memory of the buffer along u_i decays geometrically at rate $\tilde{\beta}_i$, with direction-dependent memory horizon $\tau_i = 1/(1-\tilde{\beta}_i) = 1/((1-\beta) + \eta\lambda_i)$. High-density directions (λ_i large) are forgotten quickly; low-density directions ($\lambda_i \approx 0$) retain an EMA-like horizon $\approx 1/(1-\beta)$. *This direction-dependent horizon is the unifying mechanism behind every result that follows.*

Steady state: a Tikhonov-regularized Wiener predictor. Any momentum buffer is, to first order in the learning rate, an estimator of the population gradient $\bar{G}(\theta_t) = \nabla_\theta \mathcal{L}(\theta_t)$: Taylor-expanding $\mathcal{L}$ along $-\alpha M_t$ gives $\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1}) = \alpha \langle M_t, \bar{G} \rangle_F + O(\alpha^2)$, so any deviation pays a first-order penalty. What distinguishes EMA from DELTAMOMENTUM is *which loss the buffer minimizes online*. EMA (1) performs online SGD on the weight-space MSE $J_w(M) = \frac{1}{2}\mathbb{E}\|M - \delta x^\top\|_F^2$, with fixed point $M_w^\star = \bar{G}$. The pure delta rule ($\beta = 1$ in (3)) performs online SGD on the function-space prediction loss $J_f(M) = \frac{1}{2}\mathbb{E}\|\delta - Mx\|^2$, with fixed point $M_f^\star = \bar{G}\Sigma_x^{-1}$ (Wiener-optimal); the two coincide iff $\Sigma_x = cI$.

Theorem 2 (Tikhonov-regularized Wiener fixed point) *Under the quasi-static approximation ($\bar{G}, \Sigma_x$ constant) and $\rho(\beta I - \eta\Sigma_x) < 1$, the update (3) converges in expectation to*

$$M^\star \;=\; \eta\,\bar{G}\left((1-\beta)I + \eta\Sigma_x\right)^{-1}, \quad M^\star u_i = \phi_i\,\bar{G}u_i, \quad \phi_i = \tfrac{\eta}{(1-\beta)+\eta\lambda_i} = \tfrac{1}{\mu+\lambda_i}, \tag{7}$$

with Tikhonov ridge $\mu := (1-\beta)/\eta$.

Proof [Proof sketch] Taking expectations of (4): $M^\star((1-\beta)I + \eta\Sigma_x) = \eta\bar{G}$. The steady-state ratio of injection to forgetting along u_i is $\eta\tau_i$, a direct reading of Lemma 1. ■

The fixed point (7) interpolates between EMA's direction-uniform target ($\mu \to \infty$, $\phi_i/\phi_j \to 1$) and the unregularized Wiener predictor ($\mu \to 0$, $\phi_i \to 1/\lambda_i$): high-density directions are down-weighted by exactly the directional density that EMA leaves intact—*implicit input-side preconditioning*.

Remark 3 (K-FAC connection) *The K-FAC approximation of the Fisher information of a linear layer is $F \approx \mathbb{E}[\delta\delta^\top] \otimes \mathbb{E}[xx^\top]$, with natural-gradient update $\mathbb{E}[\delta\delta^\top]^{-1}\bar{G}\Sigma_x^{-1}$. The pure delta-rule fixed point $\bar{G}\Sigma_x^{-1}$ is* exactly *the input-side factor;* DELTAMOMENTUM *implements input-side natural-gradient preconditioning without inverting Σ_x, at the cost of a single rank-1 update per step.*

Non-stationary tracking dynamics. In deep-network training $\bar{G}_t$ and $\Sigma_x(\theta_t)$ drift continuously. The same operator that governs the steady state governs the response to distributional shift, with a strictly sharper rate than EMA along every positive-density direction.

Theorem 4 (Direction-selective tracking under distribution shift) *Suppose at step t_0 the population gradient and input-feature covariance shift to new fixed values $(\bar{G}_k, \Sigma_k)$, and let $M^\star$ be the corresponding new fixed point. Define the tracking error $T_{t,i} := (M_t - M^\star)\,u_i$ along eigendirection u_i of Σ_k. Then*

$$\begin{aligned} \textit{EMA:} \quad \mathbb{E}[T_{t,i} \mid T_{t-1,i}] \;&=\; \beta\,T_{t-1,i}, \quad \forall i, \\ \textit{Delta:} \quad \mathbb{E}[T_{t,i} \mid T_{t-1,i}] \;&=\; \tilde{\beta}_i\,T_{t-1,i} \;=\; (\beta - \eta\lambda_i)\,T_{t-1,i}. \end{aligned}$$

EMA contracts uniformly at rate β; DELTAMOMENTUM *contracts along u_i at $\tilde{\beta}_i$, strictly faster for every direction with $\lambda_i > 0$, with no regime restriction.*

Proof [Proof sketch] Take expectations of (4) and use $\eta\bar{G}_k = M^\star[(1-\beta)I + \eta\Sigma_k]$: $\mathbb{E}[M_t - M^\star \mid M_{t-1}] = (M_{t-1} - M^\star)(\beta I - \eta\Sigma_k)$. Projecting on u_i with $\Sigma_k u_i = \lambda_i u_i$ yields $\tilde{\beta}_i$. ■

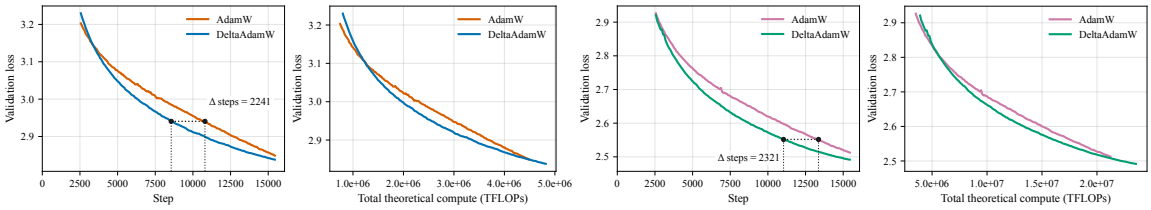
Theorem 4 is the cleanest dynamical expression of DELTAMOMENTUM's mechanism: it follows immediately from Lemma 1 and applies in expectation along every direction. This faster contraction

is a statement about bias (tracking lag); along dense directions it is bought with a shorter averaging window, a bias-variance trade whose net sign is settled empirically in Section 4. *What gets forgotten and what gets preserved is coupled to input-direction density*: high-density directions, where the local gradient changes rapidly with θ, are overwritten on a short horizon; low-density directions, where the gradient is approximately stationary, are preserved on an EMA-like horizon—exactly the data-driven schedule the function-space prediction loss prescribes. The same identity sharpens finite-time behavior in linear regression: per-direction iteration determinants satisfy $\det(A_i^{\mathrm{Delta}}) = \tilde{\beta}_i < \beta = \det(A_i^{\mathrm{EMA}})$, giving $\rho_i^{\mathrm{Delta}} < \rho_i^{\mathrm{EMA}}$ in the underdamped regime (Appendix B).

4. Empirical Validation

We test the three operative predictions of Section 3 by comparing DeltaAdamW vs. AdamW on Llama-2-style [15] language-model pretraining (SwiGLU FFN $\alpha = 8/3$, RoPE, pre-RMSNorm; $d_{\mathrm{model}} \in \{384, 1024\}$, $n_{\mathrm{layers}} = 24$) on FineWeb-Edu 10BT [11], seq. 2048, batch 256, bfloat16, cosine LR with 500-step warmup, $\sim$19k steps. AdamW is the unique baseline differing from DeltaAdamW only in the first-moment update. Hyperparameters tuned on 67M with Optuna [1] and transferred to 370M by μP; 67M is 3-seed averaged, 370M single-seed. Asymptotic per-block FLOP overhead $\approx 15.74\%$ (realized $7.3\%/11.0\%$); zero persistent-state memory overhead.

FLOP accounting. We count per linear layer of shape $m \times n$ over a token batch of size T. Standard training costs $\approx 6Tmn$ there (forward $2Tmn$, input-gradient $2Tmn$, weight-gradient $2Tmn$), against which the EMA update is an $O(mn)$ elementwise step and is negligible. DELTAMOMENTUM adds two token-scaled matmuls of the same order as the weight gradient, forming the prediction $M_{t-1}\hat{x}_t$ ($2Tmn$) and the correction $(M_{t-1}\hat{x}_t)\hat{x}_t^\top$ ($2Tmn$), for $\approx 4Tmn$ extra on each linear layer while reusing the gradient outer product $\eta\, \delta_t \hat{x}_t^\top$ already formed in the backward pass. Aggregating over a transformer block, in which attention scores and value aggregation are token-quadratic and norms and activations are elementwise, all carrying no buffer overhead, dilutes this to the reported $\approx 15.74\%$ theoretical per-block figure. The realized $7.3\%/11.0\%$ is lower still because kernel fusion and the elementwise and data-movement fraction of a step further dilute the matmul-only overhead.

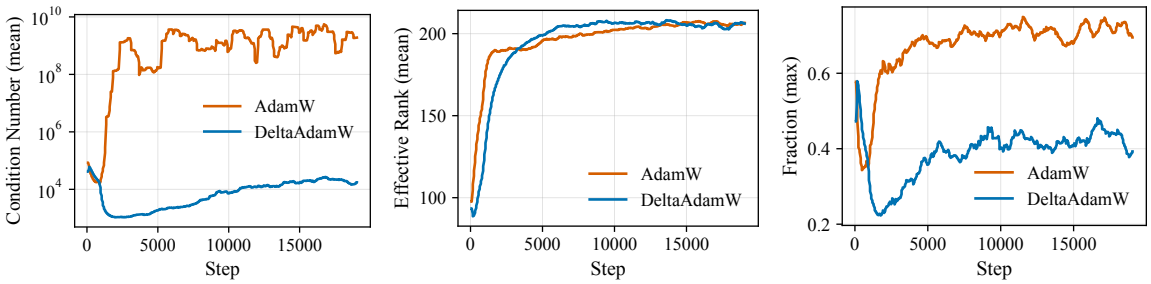


(a) 67M, vs. samples. **(b)** 67M, vs. FLOPs. **(c)** 370M, vs. samples. **(d)** 370M, vs. FLOPs.

Figure 1: Validation loss on FineWeb-Edu (10BT). 67M curves are 3-seed means; 370M single-seed. **(a,c)** Per-sample. **(b,d)** Per theoretical-FLOP: DeltaAdamW is Pareto-superior across the horizon after the crossing ($\sim$2.5k steps).

Findings. At both scales (Figure 1), after a $\sim$2.5k-step crossover DeltaAdamW reaches AdamW's validation loss in fewer steps—max 21% at 67M and 21.55% at 370M across matched loss levels—and continues to a lower terminal value; net of measured per-step overhead this is a FLOP reduction up to $15.26\%/12.94\%$ to reach a fixed target. The advantage opens post-warmup, persists through

cosine decay, and does not degrade with scale—consistent with Theorem 2: the buffer relaxes toward a fixed point that down-weights high-density input directions, scaling $\bar{G}u_i$ by $\phi_i = 1/(\mu + \lambda_i)$ vs. EMA's uniform 1. The downstream dynamical signature is verified directly: feature spectra under DeltaAdamW remain meaningfully healthier, (Figure 2) which is a direct verification that DELTAMOMENTUM minimizes the function-space objective EMA does not.



(a) Condition number (mean). **(b)** Effective rank (mean). **(c)** Top-1 variance fraction (max).

Figure 2: Per-layer input-feature covariance on the 67M run. Layers under DeltaAdamW (blue) maintain healthier spectra than AdamW (orange) throughout training—the downstream dynamical signature of the implicit input-side preconditioning predicted by Theorem 2.

Discussion. DELTAMOMENTUM replaces the EMA accumulator with the classical delta rule, treating the first-moment buffer as an online associative memory of factored gradients $\delta_t x_t^\top$. The single data-conditioned operator $\beta I - \eta x_t x_t^\top$ produces a direction-dependent contraction $\tilde{\beta}_i = \beta - \eta\lambda_i$ that unifies steady-state (Theorem 2: implicit input-side natural-gradient preconditioning), non-stationary (Theorem 4), and width-transfer (zero-shot η under μP) behavior, and the empirical signature agrees with all three predictions. K-FAC, Shampoo, SOAP, and Muon process an isotropic EMA externally; DELTAMOMENTUM addresses anisotropy *inside* the buffer and is complementary—DeltaMuon/Shampoo/SOAP and 1B+ scaling are the immediate follow-ups.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[2] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65, 2019.

[3] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

[4] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[5] Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. https://kellerjordan.github.io/posts/muon, 6(3):4, 2024.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

[9] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

[10] Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. *arXiv preprint arXiv:1901.08244*, 2019.

[11] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

[12] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[13] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[14] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pages 9355–9366. PMLR, 2021.

[15] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[16] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

[17] Bernard Widrow and Marcian E Hoff. Adaptive switching circuits. In *Neurocomputing: foundations of research*, pages 123–134. 1988.

[18] Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34: 17084–17097, 2021.

[19] Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522, 2024.

Appendix A. μP width transfer of the delta coefficient

This appendix proves the "if and only if" invoked in Section 2: with normalized keys, the fixed-point per-coordinate scale of the DELTAMOMENTUM buffer matches that of the EMA buffer the base optimizer's μP rule is calibrated against precisely when $\eta = \Theta(1)$.

Setup. Normalized keys satisfy $\|\hat{x}\| = 1$, so $\operatorname{tr}\Sigma_{\hat{x}} = \mathbb{E}\|\hat{x}\|^2 = 1$ and the eigenvalues $\{\lambda_i\}$ of $\Sigma_{\hat{x}} = \mathbb{E}[\hat{x}\hat{x}^\top] \in \mathbb{R}^{n\times n}$ are nonnegative with $\sum_i \lambda_i = 1$; in particular $\lambda_i \in [0, 1]$, with bulk scale $\lambda_i = \Theta(1/n)$ and at most $O(1)$ directions of scale $\Theta(1)$. Under μP the EMA buffer has fixed point $\bar{G}$ and a per-coordinate scale $\Theta(n^{-1/2})$ that the base learning-rate rule is tuned for. By Theorem 2, $M^\star u_i = \phi_i \bar{G} u_i$ with $\phi_i = \eta/((1-\beta) + \eta\lambda_i)$, so $M^\star$ shares that scale if and only if $\phi_i = \Theta(1)$ uniformly in i.

Proposition 5 (Width-invariant matching) *Fix $\beta \in [0, 1)$ independent of n and use normalized keys. Then $\phi_i = \Theta(1)$ for every i if and only if $\eta = \Theta(1)$. Consequently $M^\star$ inherits the EMA per-coordinate scale $\Theta(n^{-1/2})$, and the base μP learning-rate rule transfers without retuning, exactly when $\eta = \Theta(1)$.*

Proof Write $\eta = \Theta(n^a)$ and recall $(1-\beta) = \Theta(1)$. *Sufficiency* $(a = 0)$. For $\eta = \Theta(1)$ the denominator $(1-\beta)+\eta\lambda_i \in [\,(1-\beta),\ (1-\beta)+\eta\,]$ is $\Theta(1)$ for all i, hence $\phi_i = \Theta(1)/\Theta(1) = \Theta(1)$. *Necessity* $(a \neq 0)$. If $a > 0$, a bulk direction with $\lambda_i = \Theta(1/n)$ gives $\eta\lambda_i = \Theta(n^{a-1})$, so $\phi_i = \Theta(n^a)$ for $0 < a < 1$ and $\phi_i = \Theta(n)$ for $a \geq 1$, in both cases $\omega(1)$, while a top direction with $\lambda_i = \Theta(1)$ keeps $\phi_i = \Theta(1)$; the set $\{\phi_i\}$ is thus not uniformly $\Theta(1)$. If $a < 0$, then $\eta\lambda_i \to 0$ and $\phi_i = \Theta(n^a) = o(1)$ for every i. In either case $\phi_i = \Theta(1)$ fails, so $\eta = \Theta(1)$ is necessary. ■

Coordinate check. We verify Proposition 5 numerically by training the 67M architecture at hidden widths $n \in \{128, 256, 512, 1024, 2048\}$ with η held fixed and tracking the μP coordinate diagnostic, namely the width-normalized per-coordinate scale of the first-moment buffer and of the preactivation update it induces, as a function of n. With $\eta = \Theta(1)$ this profile is flat in n up to the $1 \pm O(1/n)$ correction that ϕ_i predicts, whereas any η scaling with n produces the monotone drift of Proposition 5. The flat profile is what licenses transferring a single η tuned at small width to the 370M model without retuning.

Appendix B. Per-direction iteration determinant on a quadratic

This appendix proves the finite-time claim of Section 3: on a quadratic the per-direction iteration matrices of EMA and DELTAMOMENTUM momentum have determinants β and $\tilde{\beta}_i = \beta - \eta\lambda_i$, so $\det(A_i^{\mathrm{Delta}}) < \det(A_i^{\mathrm{EMA}})$ and the underdamped convergence factor is strictly smaller.

Reduction. Consider scalar linear regression along one eigendirection: predict $y = wx$ with $\mathbb{E}[x^2] = \lambda$, minimize $\frac{1}{2}\mathbb{E}(wx - y)^2$, and let $e_t = w_t - w^\star$ be the parameter error, so the population gradient component is $\mathbb{E}[\delta x] = \lambda e_t$. The buffer recursions are

$$\text{EMA:}\ \ m_t = \beta m_{t-1} + (1-\beta)\lambda\, e_t, \qquad \text{Delta:}\ \ m_t = (\beta - \eta\lambda)\, m_{t-1} + \eta\lambda\, e_t = \tilde{\beta}\, m_{t-1} + \eta\lambda\, e_t,$$

the Delta recursion being the expectation of (3) under $\mathbb{E}[x^2] = \lambda$; both close the loop with $e_{t+1} = e_t - \alpha m_t$.

Iteration matrices. In the state $(e_t,\, m_{t-1})$,

$$A^{\mathrm{EMA}} = \begin{pmatrix} 1 - \alpha(1-\beta)\lambda & -\alpha\beta \\ (1-\beta)\lambda & \beta \end{pmatrix}, \qquad A^{\mathrm{Delta}} = \begin{pmatrix} 1 - \alpha\eta\lambda & -\alpha\tilde{\beta} \\ \eta\lambda & \tilde{\beta} \end{pmatrix}. \tag{8}$$

The off-diagonal contributions cancel in each determinant, leaving

$$\det A^{\mathrm{EMA}} = \beta, \qquad \det A^{\mathrm{Delta}} = \tilde{\beta}(1 - \alpha\eta\lambda) + \alpha\tilde{\beta}\,\eta\lambda = \tilde{\beta} = \beta - \eta\lambda. \tag{9}$$

Underdamped regime. When the discriminant of a real 2×2 iteration matrix is negative its eigenvalues are complex conjugates of modulus $\sqrt{\det}$, so the per-direction convergence factor is $\rho^{\mathrm{EMA}} = \sqrt{\beta}$ and $\rho^{\mathrm{Delta}} = \sqrt{\tilde{\beta}} = \sqrt{\beta - \eta\lambda}$. Hence $\rho^{\mathrm{Delta}} < \rho^{\mathrm{EMA}}$ for every $\lambda > 0$, with the gap widening as the local curvature λ grows. The two matrices have different traces, so their damping regimes differ; the determinant comparison is regime-free, and the convergence-factor comparison applies wherever both modes are underdamped.