How to Scale Mixture-of-Experts: From μP to the Maximally Scale-Stable Parameterization

Leena Chennuru Vankadara[1,*] **Moritz Haas**[2,*]

Luke Hayward[1,†] **Sebastian Bordt**[3,†] **Alessandro Breccia**[1,†]

[1]Gatsby Computational Neuroscience Unit, University College London

[2]Amazon [3]University of Tübingen, Tübingen AI Center

Abstract

Recent frontier large language models predominantly rely on Mixture-of-Experts (MoE) architectures. Despite empirical progress, there is still no principled understanding of how hyperparameters should scale with network width N, expert width N_e, number of experts M, sparsity K, and depth L to ensure stability and optimal performance at scale. We take a step toward resolving this gap by analyzing three co-scaling regimes: (I) $N \asymp N_e \to \infty$, (II) $N \asymp M \asymp K \to \infty$, and (III) $N \asymp N_e \asymp M \asymp K \to \infty$. For each regime, we derive the parameterization for SGD, Adam, and Muon satisfying all maximal-update (μ) desiderata (μP), and characterize the limiting gradient-flow dynamics through a novel Dynamical Mean Field Theory (DMFT). We then show that the resulting μP prescription does not reliably induce monotonic improvement with scale or robust learning-rate transfer. We trace these pathologies to scale-dependent observables in the dynamics, which motivates a refined set of desiderata that we term *maximal scale stability*. Guided by this principle, we derive a *Maximally Scale-Stable Parameterization* (MSSP) for SGD, Adam, and Muon in all three regimes, and characterize the corresponding limiting dynamics, qualitatively distinct from the μP limit, through a DMFT analysis. We further conjecture an MSSP for arbitrary co-scaling paths with $N \to \infty$, including extremely sparse routing with $K \in o(M)$. Experiments verify that MSSP robustly recovers learning-rate transfer and monotonic improvement with scale across regimes. Combined with existing depth-scaling analyses, these results provide a complete scaling prescription for MoEs as a function of width, depth, expert width, number of experts, and sparsity.

1 Introduction

Scaling the number of parameters consistently improves network quality (Kaplan et al., 2020; Hoffmann et al., 2022), yet the cost of architectural and hyperparameter tuning grows alongside it. At the frontier, where models exceed hundreds of billions of parameters (Brown et al., 2020; Guo et al., 2025), even a single training run makes iterative tuning infeasible. An influential line of work proposes to design parameterizations under which finite-size networks can be regarded as discretizations of a well-defined infinite-size limit (Yang and Hu, 2021; Bordelon and Pehlevan, 2022; Bordelon et al., 2024b) - specifically, parameterizations under which feature and prediction dynamics remain scale invariant and finite networks closely track the limiting object. Under such

*Equal first author. †Equal contributor. Correspondence to: l.vankadara@ucl.ac.uk.
This work does not relate to MH's position at Amazon.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

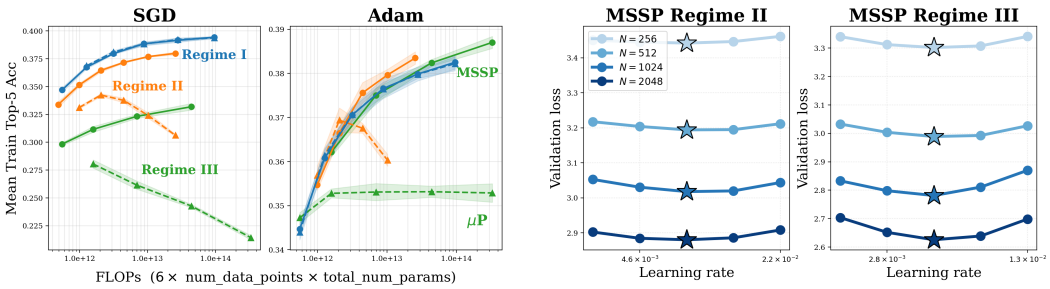


Figure 1: **μP does not reliably improve performance with scale in MoEs. MSSP recovers monotonic improvement and delivers LR transfer across MoE co-scaling regimes.** *Left: Across optimizers and regimes, MSSP (solid lines) outperforms μP (dashed lines) at large scale for MLP MoEs on TinyImageNet. Right:* LR transfer in validation loss for GPT MoEs trained with Adam in MSSP for 2.5B tokens when co-scaling width and number of (active) experts, with many fine-grained experts (Regime II) or many large experts (Regime III). All details in App. E.3.

parameterizations, smaller models serve as faithful proxies for larger ones: optimal hyperparameters, such as learning rates, transfer across scales (Yang et al., 2022; Everett et al., 2024; Dey et al., 2025), and performance improves predictably with scale (Yang et al., 2022; Dey et al., 2025). The Maximal Update Parameterization (μP) is the canonical instance for width (Yang et al., 2022), with analogous prescriptions developed for depth (Bordelon et al., 2024b,a; Dey et al., 2025).

While scale-invariant parametrizations for dense models are well-developed, frontier large language models increasingly employ Mixture-of-Experts (MoE) layers (Jiang et al., 2024; Guo et al., 2025), which decouple parameter count from computational cost via sparse routing (Shazeer et al., 2017). In doing so, they introduce additional scaling axes that must be coordinated jointly with width N and depth L: *the number of experts M, the expert width N_e, and the routing sparsity K/M*. Recent engineering advances render scaling along all of these dimensions practically tractable (NVIDIA, 2026), and a growing body of evidence indicates that increasingly fine-grained experts outperform configurations comprising a few large experts (Clark et al., 2022; Krajewski et al., 2024; Dai et al., 2024). Yet, little is known about how to optimally scale fine-grained, sparse MoE architectures, which motivates the question that this paper addresses:

How should MoE architectures and training hyperparameters be scaled to yield scale-invariant, non-degenerate feature and prediction dynamics in various co-scaling regimes of M, N, N_e, and K?

As we will see in this work, developing scaling theories for MoEs is substantially more involved than for dense networks. One cannot generally assume commutativity across axes (M, N, N_e, K); instead, joint limits governed by their relative rates must be analyzed. Deriving such limits is technically challenging since routing and aggregation in MoEs couple the dynamics of the router, the experts, and update statistics (Shazeer et al., 2017; Fedus et al., 2022). This leads to a combinatorial space of co-scaling regimes, each, as we demonstrate, exhibit qualitatively distinct behavior.

Main Contributions. We study three co-scaling regimes for MoEs: (Regime I, R-I) few large experts, $N \asymp N_e \to \infty$ with M, K fixed; (Regime II, R-II) many fine-grained experts, $N \asymp M \asymp K \to \infty$ with N_e fixed, and (Regime III, R-III) many large experts, $N \asymp N_e \asymp M \asymp K \to \infty$. Across these regimes, we make the following contributions:

- **μP for MoEs across regimes and optimizers.** In each scaling regime, we derive the parameterization satisfying the maximal-update (μ) desiderata (μP) for SGD, Adam, and Muon via signal propagation analysis, and characterize the limiting gradient-flow dynamics through a novel Dynamical Mean Field Theory (DMFT).
- **Scale-dependence of μP in MoEs.** Despite formally satisfying the μP desiderata, we find that $\mu-$parameterizations do not reliably deliver learning-rate transfer or monotonic improvement with scale in MoEs (Fig. 1, G.1 and F.3). We trace this to scale-dependence in the training dynamics.
- **Maximally Scale-Stable Parameterization (MSSP).** We propose *maximal scale stability* as a refined principle generalizing μP desiderata, and derive the Maximally Scale-Stable Parameterization (MSSP) for SGD, Adam, and Muon in each regime (Table 1; Table B.2 for Muon). The

required corrections to μP are structurally distinct across regimes: zero router initialization in R-I, amplified expert-output initialization variance ($1/N_e$ to M/N_e) in R-II, and shared expert weights at initialization in R-III. Based on our signal propagation analysis, we conjecture an MSSP for arbitrary co-scaling paths of (N, N_e, M, K) with $N \to \infty$, including extremely sparse routing with $K \in o(M)$.

- **DMFT under MSSP.** We derive a self-consistent DMFT for the limiting gradient-flow dynamics in each regime under MSSP, including a four-level conditional mean-field hierarchy in R-III induced by the shared expert initialization, qualitatively distinct from the limiting dynamics under μP.
- **Empirical validation.** We verify on MLP and Transformer MoEs that MSSP robustly recovers learning-rate transfer, and restores monotonic improvement with scale across all three regimes. We provide a complete scaling prescription for modern Transformer MoE architectures in Appendix B, along width N, depth L, number of experts M, expert width N_e and number of active experts K.

Overall, this paper takes important steps towards stable, predictable and optimal MoE scaling that preserves reliable hyperparameter transfer and monotonic improvement with increasing network size.

Independent and concurrent work. Jiang et al. (2026) derive a signSGD/Adam parameterization under R-III via DMFT. We derive μP for *SGD, Adam, and Muon across R-I, R-II, and R-III* using a signal propagation analysis complemented by a DMFT for each regime, resulting in a differing router initialization and Adam ε scaling. Crucially, we identify that in R-II and R-III, μP does not reliably yield learning rate transfer or monotonic improvement with scale in MoEs and propose MSSP as a resolution. See Appendix D for an extended discussion.

2 Setting and Approach

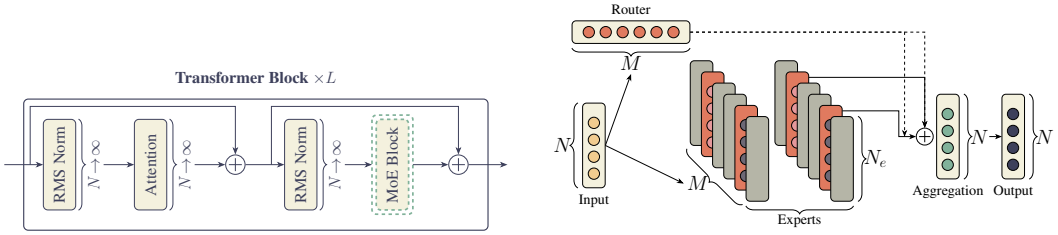


Figure 2: Transformer-MoE architecture (left) and a single MoE block (right).

MoE architecture. Given input $x_t \in \mathbb{R}^D$, the residual stream $h_t^0 = W_t^{\text{in}} x_t$ is transformed as:

$$h_t^l = h_t^{l-1} + K^{-\alpha_{\text{agg}}} \sum_{i \in \text{top-K}} \phi_{t,i}^l \cdot W_t^{l,\text{out},i} \, \varphi(W_t^{l,\text{in},i} h_t^{l-1}), \qquad \phi_{t,i}^l = \sigma\big(\beta \, (Q_t^l h_t^{l-1})_i\big),$$

for $l \in [L]$, and is finally transformed into output logits $f_t = (W_t^{\text{out}})^\top h_t^L$. Here φ is a coordinate-wise nonlinearity, β a tunable inverse temperature, and σ either sigmoid or softmax. Our scaling prescriptions are *local* to the MoE block and apply unchanged when stacked with attention, residual connections, and normalization (with non-MoE components parameterized via μP/mean-field), as validated by our Transformer MoE experiments in Appendix F.

Scaling axes, regimes, and optimizers We study joint scaling in embedding width N, expert width N_e, number of experts M, and active experts K. All $\Theta, \mathcal{O}, \Omega$ are taken as $n \to \infty$ along a trajectory $\mathcal{S}(n) = (N, N_e, M, K)(n)$. We organize results by three regimes: (I) *few large experts*: $M, K = \Theta(1)$ and $N \asymp N_e \to \infty$; (II) *many fine-grained experts*: $N_e = \Theta(1)$ and $N \asymp M \asymp K \to \infty$; (III) *many large experts*: $N \asymp N_e \asymp M \asymp K \to \infty$. We defer depth scaling to the end of Section 4. We consider training with SGD, Adam and Muon. Unless stated otherwise, we treat all other quantities, such as training time, batch size, dataset size, and input dimension, as fixed $\Theta(1)$ quantities.

Parameterizations. A parameterization assigns to *each trainable tensor* W an initialization scale s_W, with $W_{0,ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, s_W^2)$, a learning rate η_W and, for Adam, ϵ_W, and fixes the aggregation exponent $\alpha_{\text{agg}} \in [0, 1]$. Each per-tensor hyperparameter $\theta \in \{s_W, \eta_W, \epsilon_W\}$ is a power law in the scaling axes, $\theta = \bar{\theta} \, N^{r_1} N_e^{r_2} M^{r_3} K^{r_4} L^{r_5}$, with exponents $r = r(\theta) \in (\mathbb{R} \cup \{-\infty\})^5$, using the convention $x^{-\infty} := 0$, and a scale-independent multiplier $\bar{\theta} > 0$.

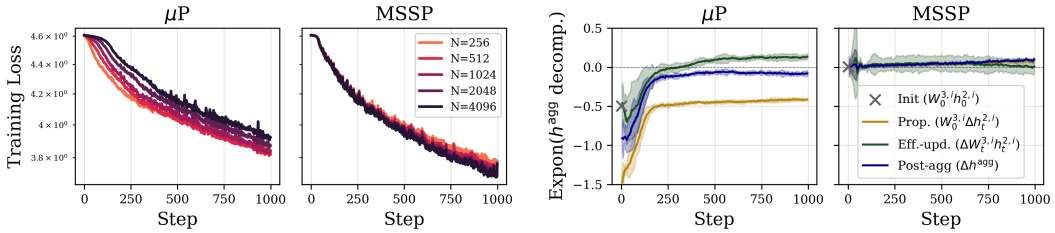


Figure 3: **Non-monotonic improvement with scale and scale dependent dynamics of μP resolved by MSSP (SGD, Regime II).** Training loss (left) is worse at large scale in μP, but monotonically improves in MSSP. Scaling exponents of sub-terms of the aggregated MoE activations h_t^l (right) are approximately 0 in all time steps in MSSP, signaling scale-independent training dynamics. In μP, initially vanishing sub-terms cause cascading time-dependent scale dependence of all terms. Even after 1000 updates, post-aggregation propagating updates still vanish in μP, as predicted.

Our approach. We seek a scaling of hyperparameters that yields stable, performant dynamics at scale, and formalize this goal as desiderata in Sections 3 and 4. We derive and validate the scaling exponents with complementary tools of varying formal status. We derive them through a *signal propagation analysis* for SGD: we track every forward and backward term through the first updates to leading order in n, including the correlations and alignments induced by training, and impose the desiderata (App. J). Under these exponents, we formally derive the limiting dynamics of full-batch gradient flow via a *DMFT*, showing that the derived exponents yield a well-defined scaling limit (App. K–O). We extend the analysis heuristically to Adam and Muon (App. H.2.3).

Experiment Setup. Our experiments cover both MLP and Transformer MoEs. We utilize MLP MoEs to isolate scaling degeneracies at a manageable scale; this allows for the exhaustive hyperparameter tuning that would be computationally prohibitive with Transformers. Specifically, we perform MLP MoE experiments with a single MoE block without residual connection but nonlinearity applied to h_t^0. This setup is more sensitive to miss-scaling than Transformers with RMSNorms and residual connections trained with Adam. We train single pass using 100 classes of TinyImageNet (Le et al., 2015) with batch size 50. This amounts to 1000 steps over 50000 images. For each scaling configuration, we separately tune a global initialization multiplier and a learning rate multiplier for each weight matrix. As these multipliers interact, we sweep all six jointly at small size $N = 128$, amounting to at least $5^6 = 15\,625$ runs per parameterization and regime.

In our transformer MoE experiments, we train GPTs where dense layers are replaced with MoE layers (Muennighoff et al., 2024). We train with warmup and cosine learning rate decay on Dolma3 (Olmo et al., 2025) up to width 2048 or 2.5B parameters using AdamW (Loshchilov and Hutter, 2019). We use RMS pre-norm and qk-norm. For load balancing, quantile balanced (QB) expert biases outperform the auxiliary loss by Fedus et al. (2022) both in validation loss and scale stability across model size and token horizon. We discuss load balancing in detail in Appendix I and provide empirical evidence in Appendix F.4. Open source code to fully reproduce our experiments is publicly available. Our setup is detailed in Appendix E. Additional empirical insights are provided in Appendix F for Transformer MoEs and Appendix G for MLP MoEs, for example showing the importance of multiplier tuning (Appendix G.5) and of layerwise Adam ϵ-scaling (Figure F.9).

3 Maximal Update Desiderata for MoEs and their Shortcomings

The maximal-update (μ) desiderata have served as the fundamental guiding principle for deriving principled scaling rules such as μP for dense networks that deliver stable and predictable training dynamics, monotonic improvement in performance with scale, and learning-rate transfer across model sizes (Yang and Hu, 2021; Bordelon et al., 2024b; Vyas et al., 2024; Dey et al., 2025). A natural starting point for MoE scaling is therefore to seek a parameterization that satisfies these desiderata in each of our scaling regimes. We now recapitulate the μ-desiderata and then derive the parameterizations satisfying them in each regime summarized in Table 1. We will then see that, despite satisfying the μ-criteria by construction, μP can induce degeneracies in MoEs, motivating the refined desiderata we propose in the next section.

Norms. For a vector, matrix or tensor T with n entries, $\|T\|_{\mathrm{RMS}} := \left(\frac{1}{n}\sum_i T_i^2\right)^{1/2}$. For a matrix W, $\|W\|_{\mathrm{op}} := \sup_{x\neq 0} \|Wx\|_{\mathrm{RMS}}/\|x\|_{\mathrm{RMS}}$ is the operator norm induced by $\|\cdot\|_{\mathrm{RMS}}$.

Group	Qty.	Regime I ($N, N_e \asymp n \to \infty$) SGD	Regime I ($N, N_e \asymp n \to \infty$) Adam	Regime II ($N, M, K \asymp n \to \infty$) SGD	Regime II ($N, M, K \asymp n \to \infty$) Adam	Regime III ($N, N_e, M, K \asymp n \to \infty$) SGD	Regime III ($N, N_e, M, K \asymp n \to \infty$) Adam
Router Q	Init σ	N^{-1} | $\mathbf{0}$	N^{-1} | $\mathbf{0}$	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$
	LR η	N^{-1}	N^{-1}	MN^{-1}	N^{-1}	1	N^{-1}
	Adam ϵ	–	1	–	M^{-1}	–	M^{-1}
Expert in	Init σ	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$ | $(\mathbf{N^{-1/2}})^{\texttt{tied}}$	$N^{-1/2}$ | $(\mathbf{N^{-1/2}})^{\texttt{tied}}$
	LR η	1	N^{-1}	MN^{-1}	N^{-1}	M	N^{-1}
	Adam ϵ	–	N^{-1}	–	M^{-1}	–	$M^{-1}N^{-1}$
Expert out	Init σ	$N_e^{-1/2}$	$N_e^{-1/2}$	$N_e^{-1/2}$ | $\mathbf{M^{1/2}N_e^{-1/2}}$	$N_e^{-1/2}$ | $\mathbf{M^{1/2}N_e^{-1/2}}$	$N_e^{-1/2}$ | $(\mathbf{N_e^{-1/2}})^{\texttt{tied}}$	$N_e^{-1/2}$ | $(\mathbf{N_e^{-1/2}})^{\texttt{tied}}$
	LR η	1	N_e^{-1}	MN	N_e^{-1}	M	N_e^{-1}
	Adam ϵ	–	N^{-1}	–	$M^{-1}N^{-1}$	–	$M^{-1}N^{-1}$

Table 1: **μP and MSSP for SGD and Adam across scaling regimes**. Scaling prescriptions as a function of (N, N_e, M, K) for initialization standard deviation σ, learning rate η, and Adam ε for each MoE weight matrix, using expert aggregation factor $\alpha^{\text{agg}} = 1$ for sigmoid routers. Entries where μP and MSSP differ are highlighted as μP | MSSP. All other weight matrices should be scaled as in μP. In addition, decoupled weight decay (Loshchilov and Hutter, 2019) and Adam's β_1, β_2 should be scale-independent. `tied`: weights should be shared across experts *at initialization*. The full MSSP scaling prescription including load balancing and depth dependence is summarized in Table B.1.

Recapitulating the Maximal Update Desiderata. The μ-desiderata for non-degenerate training dynamics under model scaling are captured in the following three criteria (Yang and Littwin, 2023). Intuitively, they require that parameter updates induce $\Theta(1)$ feature updates that propagate forward through all linear layers without vanishing or diverging.

Desideratum μ-1: Stability and feature learning. For every layer (including router and expert sub-layers), activations and their updates neither vanish nor diverge with model scale: $\|h_t\|_{\text{RMS}} = O(1)$ for all $t \geq 0$ and $\|\Delta h_{t+1}\|_{\text{RMS}} = \Theta(1)$ for some $t \geq 0$.

Desideratum μ-2: Maximal effective and propagating updates. For any trainable layer with weights W_t and input activations x_t, write $h_t := W_t x_t$. The pre-activation update decomposes into an *effective update* (from updates to the current layer's parameters) and a *propagating update* (from updates to upstream layers): $\Delta h_t = \Delta W_t\, x_t + W_0\, \Delta x_t$. Both contributions are $\Theta(1)$ at some $t \in O(1)$.

Desideratum μ-3: Faithfulness. Adam's ϵ scales as the layerwise gradient RMS.

Adam's ϵ acts as a crucial scale-dependent hyperparameter: values that are too large relative to gradients lead to suboptimal performance, while values that are too small are numerically unstable (Everett et al., 2024). Figure F.9 shows that gradient entries in the expert layers follow our predicted exponents and cross the standard global $\epsilon = 10^{-8}$ at moderate width 4500.

We derive the parameterizations that satisfy Desiderata μ-(1-3) in each scaling regime R-I - R-III.

Result 1 (μ-parameterization for MoEs). For each scaling regime (I–III) and each optimizer (SGD, Adam), the μ-parameterization (μP) in Table 1 satisfies Desiderata μ-(1-3).

Empirical degeneracies of μP in MoEs. Having established the μ-parameterization for each regime, we now examine whether it robustly delivers the scaling benefits expected in dense networks such as learning-rate transfer and monotonic improvement with increasing scale.

Regime I. Under μP, training in R-I broadly behaves as expected: loss improves monotonically with scale, and learning-rate transfer holds across scales (Section G.1.1). One empirical anomaly stands out: propagating updates of the router typically vanish with scale (Figure G.18), even though our derivation predicts them to be $\Theta(1)$. We will return to this discrepancy in Section 4.

Regimes II and III. Interestingly, however, we find that μP *can fail* empirically in R-II and R-III: (i) Effective and propagating updates do not remain $\Theta(1)$ at initialization, and can take many steps to drift toward their predicted asymptotic values (Fig. 4 and E.3). (ii) Performance often degrades with scale (Fig. 1, 3, and E.2; App. G.2.2 and G.2.3) and (iii) Learning rates that are optimal at small scales do not remain optimal at larger scales (Fig. G.1 and F.3, App. G.1.2 and G.1.3). *In short, μP does not reliably deliver the scaling benefits in MoEs that motivate it in dense networks.*

HP tuning. These findings are robust to extensive layerwise HP tuning conducted at $N = 128$ and transferred to larger scales (Appendix E). We resolve the underlying mechanisms in the next section.

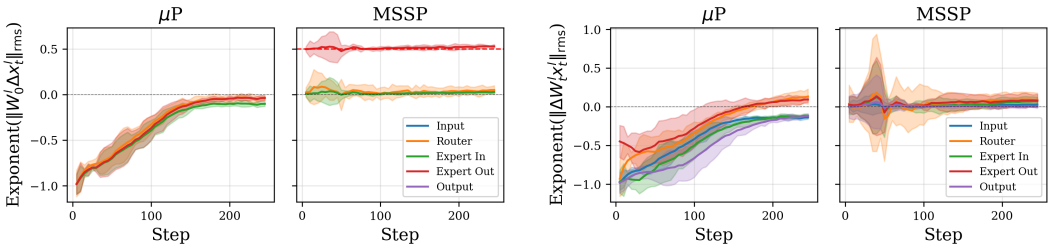


Figure 4: **Consistent exponents in MSSP, but not μP (SGD, Regime II).** Corresponds to Figure 3. Initially vanishing post-aggregation terms in μP induce width-dependence to cascade into all layers. Over time, exponents partially self-correct in μP and almost recover width-independent propagating updates (left) as well as effective updates (right) in all layers. By allowing benign divergence of pre-aggregation propagating updates (red dashed line), MSSP recovers approximately width-independent effective updates at all time steps.

Diagnosing the mechanism for failure in Regimes II and III. To understand these empirical failures under μP, examine the aggregation over experts unique to MoEs. The post-aggregation activations h_t^{l+1} can be written as $h_t^l + h_t^{\text{agg}}$, where h_t^{agg} admits the following decomposition:

$$h_t^{\text{agg}} \;=\; \frac{1}{M}\sum_{i=1}^{M} \phi_i^l \cdot \Bigg(\underbrace{W_0^{l,\text{out},i}\, h_{0,i}^{l,\text{in}}}_{\text{init.}} + \underbrace{\Delta W_t^{l,\text{out},i}\, h_{t,i}^{l,\text{in}}}_{\text{effective}} + \underbrace{W_0^{l,\text{out},i}\, \Delta h_{t,i}^{l,\text{in}}}_{\text{propagating}} \Bigg). \qquad \text{(MoE)}$$

A single principle organizes the scaling properties of each of these quantities. The cross-expert average exhibits LLN-like behaviour and contributes at $\Theta(1)$ whenever the per-expert summands share a *coherent direction across experts*; when the per-expert summands have independent random directions, the corresponding aggregate exhibits CLT-like behaviour which suppresses a factor of $\sqrt{M}$. The underlying mechanism that produces or destroys this coherence varies across these aggregate terms and across regimes, so they can occupy distinct orders in N, and the order of h_t^{l+1} at any finite scale is set by whichever sub-term dominates. We now provide intuition for how each of these aggregates behaves. We recommend non-technical readers to skip to '*Empirical consequences*'.

Init aggregate. Each per-expert summand $\phi_i^l \cdot W_0^{l,\text{out},i}\, h_{0,i}^{l,\text{in}}$ is a chain $W_0^{l,\text{out},i}\, W_0^{l,\text{in},i}$ of two independent Gaussian matrices acting on the shared input x_0^l. By rotation invariance, the cross-expert dependence between summands is mediated by its scaled norm, $\|x_0^l\|^2/M$, which as $M \to \infty$ converges to a deterministic scalar. Since summands are asymptotically i.i.d., the init aggregate exhibits CLT-like behaviour and is $\Theta(1/\sqrt{M})$ in both R-II and R-III.

Propagating aggregate. The propagating aggregate admits a further decomposition. At leading order, $\Delta h_t^{l,\text{in},i} \approx W_0^{l,\text{in},i}\, \Delta x_t^l + \Delta W_t^{l,\text{in},i}\, x_t^l$, splitting the propagating aggregate into two contributions: $\frac{1}{M}\sum_i \phi_i^l\, W_0^{l,\text{out},i}\, W_0^{l,\text{in},i}\, \Delta x_t^l$ and $\frac{1}{M}\sum_i \phi_i^l\, W_0^{l,\text{out},i}\, \Delta W_t^{l,\text{in},i}\, x_t^l$. The former is governed by the same CLT mechanism as the init aggregate, with one caveat: Δx_t^l depends on each $W_0^{l,\text{in},i}$ through backpropagation. However, a single vector in $\mathbb{R}^N$ cannot align nontrivially with $M \to \infty$ independent random Gaussian chains simultaneously. The per-expert summands therefore remain approximately independent across experts, and the same CLT argument applies. This contribution is therefore also of order $\Theta(1/\sqrt{M})$ in both R-II and R-III. The second contribution behaves differently in different scaling regimes. Substituting the rank-one gradient $\Delta W_t^{l,\text{in},i} \propto (W_0^{l,\text{out},i})^\top \delta_t^{l+1} (x_t^l)^\top$ (where $\delta_t^{l+1} = \nabla_{h_t^{l+1}} \mathcal{L}$ represents the backpropagated error signal from the subsequent layer) yields the form $\frac{1}{M}\sum_{i=1}^M \phi_i^l\, G_i\, u$, where $G_i := W_0^{l,\text{out},i}(W_0^{l,\text{out},i})^\top$ and $u \propto \delta_t^{l+1}$ is shared across experts.

(a) *Regime II.* When N_e is held fixed, G_i has rank $N_e = \Theta(1)$ and after rescaling, it is the orthogonal projector onto a $O(1)$ dimensional subspace of $\mathbb{R}^N$ that is Haar-distributed on the Grassmannian $\text{Gr}(N_e, N)$ (i.e., uniformly random among N_e-dimensional subspaces), and these subspaces are independent across experts. Therefore, the per-expert vectors $\{G_i u\}_{i=1}^M$ are approximately i.i.d. across experts with no coherent direction, and the cross-expert average exhibits (CLT-like behaviour). Both contributions to the propagating aggregate are therefore $\Theta(1/\sqrt{M})$ under μP.

(b) *Regime III.* When $N_e \asymp N$, $W_0^{l,\text{out},i}$ is generically full rank and the empirical spectrum of $G_i = W_0^{l,\text{out},i}(W_0^{l,\text{out},i})^\top$ concentrates around its mean (Marchenko and Pastur, 1967). Consequently $G_i \approx c\, I_N$ for a deterministic scalar $c = \Theta(1)$, so $G_i u \approx c\, u$ for every i. The per-expert summands therefore share a coherent direction u, and this contribution exhibits LLN-like behaviour at $\Theta(1)$ scale. Together with the $\Theta(1/\sqrt{M})$ first contribution, the propagating aggregate contains a strictly subleading contribution and exhibits a milder version of the same CLT/LLN imbalance. Moreover, the init term in both regimes exhibits CLT behaviour so remains subleading.

Effective aggregate. Substituting the rank-one gradient $\Delta W_t^{l,\text{out},i} \propto \delta_t^{l+1}\big(h_{t-1}^{l,\text{in},i}\big)^\top$ reveals that the summands of the effective aggregate share a coherent direction δ_t^{l+1}. The effective aggregate therefore exhibits LLN-like behaviour at $\Theta(1)$ along δ_t^{l+1} in both R-II and R-III.

Empirical consequences. In summary, h_t^{agg} *is composed of unbalanced contributions*: the init term is of order $\Theta(1/\sqrt{M})$ in both regimes; the effective term is of order $\Theta(1)$ in both; the propagating term decomposes into a first contribution of $\Theta(1/\sqrt{M})$ in both regimes, and a second contribution of $\Theta(1/\sqrt{M})$ in R-II but $\Theta(1)$ in R-III. The training-induced update Δh_t^{agg}, formed from the effective and propagating terms, is therefore unbalanced in both regimes, with the imbalance more pronounced in R-II. This is confirmed empirically in R-II under μP (Fig. 3): at finite scale Δh_t^{agg} is dominated by the scale-suppressed init and propagating terms. Feature learning is therefore vanishing early in training and takes many optimization steps to become $\Theta(1)$ and performance consequently degrades with increasing scale (Fig. 1 and 3).

4 A More Fundamental Desideratum Beyond Maximal Updates: Scaling MoEs Requires Maximal Scale Stability

In Section 3, we traced the failure modes of μP to scale-dependent contributions arising in the training dynamics. This motivates a more general and fundamental desideratum, which we call *maximal scale stability*: *every primitive interaction obtained by decomposing each weight as $W = W_0 + \Delta W$ and propagating this split through the network should have a $\Theta(1)$ impact on the forward and (appropriately normalized) backward dynamics.* For most settings of interest, this reduces to a compact set of operational conditions; we instantiate the desideratum for MoEs as follows:

Maximal Scale Stability Desiderata for MoEs.

1. *Forward.* The maximal update Desiderata μ-1, μ-2, μ-3 hold.
2. *Backward.* Along any linear map $h_t^\ell = W_t^\ell\, x_t^{\ell-1}$, analogous to activations, gradients admit the transpose recursion $\bar{\delta}_t^{\ell-1} = (W_t^\ell)^\top\, \delta_t^\ell$ and consequently the analogous effective/propagating decomposition $\bar{\delta}_t^{\ell-1} = (W_0^\ell)^\top\, \delta_t^\ell + (\Delta W_t^\ell)^\top\, \delta_t^\ell$, where $\bar{\delta}_t^\ell := \nabla_{x_t^\ell}\mathcal{L}$ and $\delta_t^\ell := \nabla_{h_t^\ell}\mathcal{L}$. Both contributions are balanced and remain $\Theta(1)$ under appropriate normalization.
3. *Forward and Backward Aggregation.* Each of the init, propagating, and effective contributions of the cross-expert aggregation (MoE) is $\Theta(1)$. The analogous requirement applies to the expert-aggregated gradient at the shared input h^l (see Eq. (H.6)).

We refer to a parameterization that satisfies these conditions as the *Maximally Scale-Stable Parameterization* (MSSP). For linear layers with diverging input and output dimensions (`fan-in` $\asymp$ `fan-out`), forward and backward desiderata coincide by the symmetry of the Gaussian initialization. MSSP desiderata reduce to μ-desiderata in dense architectures such as MLPs and Transformers. But in general, forward scale stability does not imply backward scale stability as we will see below.

MSSP for MoEs. We derive parameterizations that satisfy the above desiderata in R-I - R-III.

Result 2 (MSSP for MoEs). The maximally scale-stable parameterization (MSSP) in Table 1 satisfies all MSSP desiderata in R-I and R-III. In R-II, it satisfies Desiderata 1–3 except that the propagating update of $h^{l,\text{out},i}$ is of order $\Theta(\sqrt{M})$; however, its impact on the aggregate dynamics remains $\Theta(1)$.

Regime II ($N, M, K \to \infty, N_e \in \Theta(1)$). Under μP in R-II, gradients at *every layer* except $h^{l,\text{out},i}$ have scale-dependent contributions. In addition, we saw in Sec. 3 that both the forward and backward

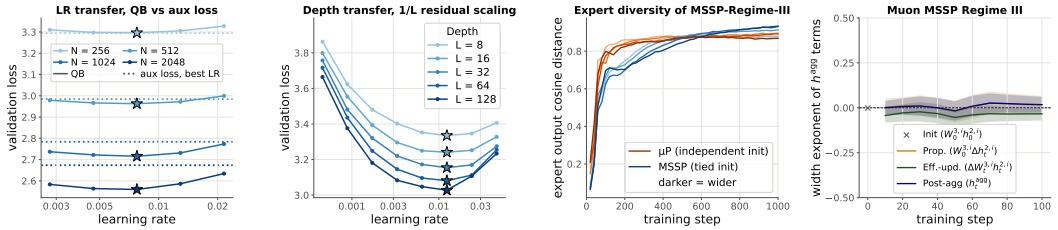


Figure 5: **MSSP performs well in Transformer MoEs.** 1. MSSP using QB expert biases preserves learning rate transfer and outperforms MSSP using aux loss. 2. Transfer under pure depth scaling. 3. Despite tied expert initialization, MSSP in R-III outperforms μP in R-III in terms of expert diversity after only 600 steps. 4. MSSP desiderata hold empirically for Muon using Table B.2.

aggregations receive scale-dependent contributions. MSSP rectifies all these imbalances with a single intervention: **the initialization variance of the expert output is amplified from $1/\mathbf{N_e}$ to $\mathbf{M/N_e}$.** Although failure modes in μP arise through structurally distinct mechanisms, this variance amplification simultaneously rebalances every forward and backward cross-expert sub-aggregate, every previously sub-leading contribution to per-expert gradients as well as to the gradient flowing through the router (cf. tables in App. J.3.2 for μP versus App. J.3.4 for MSSP).

The only quantity that does not remain $\Theta(1)$ under MSSP is the per-expert propagating update $W_0^{l,\text{out},i} h_t^{l,\text{in},i} = \Theta(\sqrt{M})$, which is impossible to avoid. We argue that this divergence is benign since its impact on the training dynamics is $\Theta(1)$, and no training observable inherits this divergence. The dynamics admits a well-defined DMFT limit under MSSP (App. L). We empirically verify that all relevant terms remain scale-stable over time (Figure 4, App. G.2.2). Consequently, MSSP in R-II restores both monotonic performance improvement with scale and learning rate transfer from small to large scale for both SGD and Adam (Fig. 1, G.1 and F.3, App. G.1.2).

Regime III $(N, N_e, M, K, \to \infty)$. When $N_e \asymp N$, amplifying the expert output initialization variance induces divergence in the backward pass. The viable fix in R-III is instead structural: **share the expert initialization, $\mathbf{W_0^{l,out,i} = W_0^{l,out}}$ and $\mathbf{W_0^{l,in,i} = W_0^{l,in}}$ for all** i and allow the routing mechanism to diversify experts over time. No parameterization with independent initialization as defined in Section 2 satisfies all desiderata in R-III. Under this intervention, complete scale stability is restored (see theory App. J.4.5 and experiment App. G.2.3). Figure 1 shows that this intervention can have impactful practical consequences: performance of both SGD and Adam on TinyImageNet reliably improves with scale only under MSSP but not μP, all relevant quantities remain $\Theta(1)$ through training (see App. G.2.3), and learning rate reliably transfers across scales (Fig. G.1 and F.3). In Transformers, even though expert weights are shared at initialization, MSSP matches or even exceeds the expert diversity obtained under independently initialized experts after only 600 steps (Figure 5).

Hence, both regimes call for structurally distinct interventions: increased expert output initialization in R-II, and shared initialization in R-III. Neither solution transfers to the other regime. A detailed mechanism-by-mechanism account of how MSSP rectifies each μP failure is provided in App. H.

Regime I $(M, K \in \Theta(1))$. μP satisfies our operational MSSP desiderata for MoEs in R-I. However, splitting Δh^{l-1} into the contributions of the router-gradient and the expert-gradient pathways reveals that the former is correlated with Q_0 and contributes coherently at $\Theta(1)$, while the latter is weakly correlated with Q_0 and is CLT-suppressed at $\Theta(1/\sqrt{N})$. To recover residual balance in the router's propagating update $Q_0^l \Delta h^{l-1}$, **MSSP initializes the router to zero**. The empirical impacts of this adjustment are marginal (Figure 1, App. G.2.1).

Arbitrary co-scaling paths. Based on our signal propagation analysis for R-II and R-III, we conjecture that, within our class of parameterizations including shared initialization within groups of experts, an MSSP along a co-scaling path with $N \to \infty$ exists if and only if $K \cdot N_e = \Omega(N)$. Otherwise, if $K \cdot N_e = o(N)$, no initialization balances the coherent and fluctuating contributions to the propagating aggregate. For $K \cdot N_e = \Omega(N)$, our conjectured MSSP shares initial expert weights within $G \asymp N/N_e$ groups and amplifies the expert-output initialization variance to $G/N_e \asymp N/N_e^2$. This interpolates between independent experts with amplified variance M/N_e in R-II ($G \asymp M$) and fully shared experts with variance $1/N_e$ in R-III ($G = 1$); in R-I it yields an alternative MSSP with tied experts. The full prescription, which covers extremely sparse routing with $K \in o(M)$, is in

the last column of Table B.1 for Adam and Table B.2 for Muon. Refined coordinate checks on an extremely sparse path with fixed K, for MLPs with Adam and for Transformers with Muon, provide preliminary validation (App. G.3, F.6).

Scaling depth L. In residual MoEs, including Transformer MoEs, depth composes naturally with our analysis: once the per-layer MoE block is parameterized according to MSSP, with $\Theta(1/L)$ scaling of the residual MoE blocks (Dey et al., 2025), the analysis of Bordelon et al. (2024a) extends with minimal modification to give a well-defined infinite-depth limit. Figure 5 shows clean learning rate transfer under pure depth scaling. Table B.1 gives a scaling prescription for modern transformer MoE architectures as a function of width, depth, expert width, experts, and sparsity.

Load balancing. We provide a heuristic derivation of the scale-preserving auxiliary loss scaling in Appendix I and verifying Transformer experiments in Appendix F.4. The standard auxiliary loss scaling by Fedus et al. (2022) induces increasingly small router logits with scale. But even under correctly scaled auxiliary loss, router logits develop increasingly negative entries over time resulting in small post-sigmoid routing weights. Quantile-balanced expert biases preserve the router logit distribution across model scales, and non-vanishing post-sigmoid routing weights across time (Figure F.10), resulting in improved validation loss without requiring any HP tuning (Figure 5).

Regime Comparison. As a preliminary evaluation of which regime and parameterization scale most favorably in practice, Figure F.1 shows the scaling ladder of 8-layer all-layer MoEs after 10B tokens and multiplier tuning as discussed in Section E.2.2. As expected, all MoEs outperform their dense equivalent. While different regimes perform similarly well at the same number of active non-embedding parameters, all-scaling R-III is more scalable than fine-grained R-II as the expert communication overhead quickly grows infeasible in R-II. MSSP increasingly outperforms μP with increasing model scale in R-III. On a 2.5B token horizon, both μP and MSSP outperform standard scaling in R-II and R-III where the margin also increases with scale (Figure F.5). We also validate the correctness of our MSSP prescriptions for Muon and for sparse scaling keeping K fixed in Figures F.20 and F.21 and Appendix G.3, but a full performance comparison across optimizers and sparsity is left to future work.

5 Self-consistent DMFT for MoE training dynamics

For each regime, we derive self-consistent equations describing the limiting MoE training dynamics using a path-integral formulation based on the Martin–Siggia–Rose–De Dominicis–Janssen (MSRDJ) framework (Martin et al., 1973; De Dominicis, 1976; Janssen, 1976). This framework underlies pioneering DMFT analyses of training dynamics in infinite-width and infinite-depth deep neural networks (Bordelon and Pehlevan, 2022; Bordelon et al., 2024b,a). We derive DMFT limits for MoE architectures in each of the three scaling regimes, under both μP and MSSP; each admits a well-defined scaling limit. In each regime, the limiting MoE dynamics under MSSP enjoy several analytical properties described below. The full set of DMFT equations is provided in App.K-N. We defer the comparison to μP limits to App. O.

Regime I $(M, K \in \Theta(1))$. The DMFT has a finite-expert mean-field factorization. Residual-stream neurons and their backward signals are i.i.d. draws from a global single-site distribution; within each expert, hidden-layer neurons are i.i.d. draws from an expert-local single-site distribution. Both distributions are characterized self-consistently by a finite set of macroscopic order parameters. Because the number of experts is finite, the expert index is not replaced by a population mean field: the experts and router variables remain explicitly represented. Averages over the global distribution define the global feature and gradient kernels; Together with router variables, these order parameters close the limiting dynamics. Under soft routing, the stochastic router field vanishes in this limit, inducing deterministic routing. Without an additional symmetry-breaking mechanism (Top-K, routing noise, or random router biases), the router remains uniform and experts stay identical throughout training. This is empirically verified in Figure G.42.

Regime II. The simultaneous $N, M, K \asymp n \to \infty$ limit with $N_e \in \Theta(1)$ reveals a finite expert width mean-field factorization in MSSP. Residual-stream neurons and their backward signals are i.i.d. draws from an effective global single-site distribution; experts together with their router variables are i.i.d. draws from an effective expert/router distribution. Both are self-consistently characterized by a finite set of order parameters. Since per-expert hidden width is fixed, the hidden layer of a typical expert is a finite-dimensional component of the expert/router process. Averages over the global distribution

define the global feature and gradient kernels; averages over the expert/router distribution define the routed expert-level kernels. These order parameters close the limiting dynamics.

Regime III. The simultaneous $N, N_e, M, K \asymp n \to \infty$ limit reveals a *four-level, conditional, mean-field hierarchy* in MSSP. Schematically, $\mathcal{X}_{\text{glob}} \sim P_{\text{glob}}(\cdot; \mathcal{K})$, $\mathcal{F}_{\text{sh}} \sim P_{\text{sh}}(\cdot; \mathcal{K})$, $\mathcal{E} \mid \mathcal{F}_{\text{sh}} \sim P_{\text{ex/r}}(\cdot \mid \mathcal{F}_{\text{sh}}; \mathcal{K})$ and $\mathcal{U} \mid (\mathcal{F}_{\text{sh}}, \mathcal{E}) \sim P_{\text{loc}}(\cdot \mid \mathcal{F}_{\text{sh}}, \mathcal{E}; \mathcal{K})$, where $\mathcal{X}_{\text{glob}}$ denotes the global single-site process, $\mathcal{F}_{\text{sh}}$ the shared expert-hidden single-site process, $\mathcal{E}$ the expert/router single-site process, and $\mathcal{U}$ the within-expert hidden-neuron single-site process. *Global single-site process.* Residual-stream coordinates and their backward signals are i.i.d. draws from an effective distribution characterized self-consistently by a finite set of order parameters; averages over this process define the global feature and gradient kernels. *Shared expert-hidden single-site process.* Coordinates of *expert-averaged* hidden-layer fields, arising from expert-averaged hidden activations and gradients, are i.i.d. draws from an effective shared distribution. Averages define the shared expert-hidden kernels, which summarize the correlation structure of these fields induced by the shared initial expert weights. *Expert/router single-site process.* Conditional on the shared expert-hidden fields, expert/router sites are i.i.d. draws from an effective expert/router distribution; averages define expert-level kernels. *Within-expert hidden-neuron single-site process.* Conditional on the shared expert-hidden fields and the corresponding router state, neurons in the expert hidden layers are i.i.d. draws from an effective distribution; averages define the expert-local forward and backward kernels. The four processes form a closed self-consistent system: kernels defined at each level enter as parameters of the single-site distributions at the others, and the limiting dynamics are determined by solving all four levels simultaneously.

Top-K routing. Selecting the top K experts is equivalent to thresholding the router scores so that a fraction $\rho = K/M$ of the experts is active. In R-II and R-III, where $M \to \infty$, this threshold becomes the $(1-\rho)$-quantile of the limiting router-score distribution (App. M, O). In R-I, M is finite and top-K remains an explicit selection among the M experts. At initialization, the router scores converge to a deterministic limit as $N \to \infty$ under both μP and MSSP, which is identical across experts by symmetry, so the selection would be decided by finite-width fluctuations that the DMFT does not capture. Under additive routing noise or a symmetry-breaking bias, the $N, N_e \to \infty$ limit yields a DMFT conditional on its realization.

6 Discussion and Future Work

Scaling laws and hyperparameter transfer have become the backbone of modern model development. Yet, for MoE architectures there remained a critical gap of how to scale the optimization hyperparameters jointly with MoE model dimensions M, N, N_e, K, L. The Maximally Scale-Stable Parameterization (MSSP) introduced here closes this gap. We believe generalizing the μ-desiderata to maximally scale-stable desiderata provides a framework for scale-invariant dynamics across a broader set of architectures, optimizers, and scaling dimensions, with practical benefits including predictable performance gains and hyperparameter transfer. MSSP enables the first principled compute-optimal comparison of the three co-scaling regimes. Earlier work on optimal expert granularity (Clark et al., 2022; Krajewski et al., 2024; Dai et al., 2024) was conducted under suboptimal model scaling and merits reevaluation. The DMFT limits are of independent interest, opening routes to studying finite-size corrections, expert specialization, and the role of routing. They also provide a framework for analyzing auxiliary load-balancing and z-losses, which we verify empirically do not alter the scaling exponents but may shed light on router collapse through their interaction with the limiting dynamics.

Limitations. In R-II, MSSP causes the per-expert pre-aggregation $W_0^{l,\text{out},i}\, h_t^{l,\text{in},i}$ to scale as $\Theta(\sqrt{M})$. However, its dynamical impact remains $\Theta(1)$, and a fused kernel can avoid materializing the divergent activations (App. B). In R-III, shared initialization relies on asymmetric routing to differentiate the experts. We observe rapid differentiation in our settings (Figure 5; App. F.2), but its robustness beyond them remains to be established. Our signal propagation analysis is a leading-order scaling argument, and our DMFT covers full-batch gradient flow with P and T fixed; both are at the level of rigor of physics, and fully rigorous proofs remain open. The Adam and Muon scalings are empirically verified heuristic extensions. Our MSSP for arbitrary co-scaling paths, including sparse routing with $K \in o(M)$, remains a conjecture with preliminary empirical validation. We leave a complete signal propagation analysis, a thorough empirical validation, and the characterization of the corresponding limiting dynamics to future work. More fundamentally, why hyperparameters transfer under scale-invariant parameterizations, in MoEs, dense networks, or beyond, remains an open theoretical problem.

Acknowledgments

This work has been supported by the Gatsby Charitable Foundation (GAT4058). SB has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning - New Perspectives for Science" (EXC 2064/1 number 390727645).

References

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2010.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, 2 edition, 2010.

Charlie Blake, Constantin Eichenberg, Josef Dean, Lukas Balles, Luke Yuri Prince, Björn Deiseroth, Andres Felipe Cruz-Salinas, Carlo Luschi, Samuel Weinbach, and Douglas Orr. u-μp: The unit-scaled maximal update parametrization. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Enric Boix-Adsera and Philippe Rigollet. The power of fine-grained experts: Granularity boosts expressivity in mixture of experts. *arXiv preprint arXiv:2505.06839*, 2025.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 32240–32256, 2022.

Blake Bordelon, Hamza Tahir Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024a.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

Louis-Pierre Chaintron, Lénaïc Chizat, and Javier Maas. Resnets of all shapes and sizes: Convergence of training dynamics in the large-scale limit. *arXiv preprint arXiv:2603.18168*, 2026.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Aidan Clark, Diego De Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George Bm Van Den Driessche, Eliza Rutherford, Tom Hennigan, Matthew J Johnson, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Marc'Aurelio Ranzato, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

C. De Dominicis. Techniques de renormalisation de la théorie des champs et dynamique des phénomènes critiques. *Journal de Physique Colloques*, 37(C1):C1–247–C1–253, 1976.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers. *arXiv:2505.01618*, 2025.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A Alemi, Roman Novak, Peter J Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, et al. Scaling exponents across parameterizations and optimizers. *arXiv:2407.05872*, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Jacob Fein-Ashley, Dhruv Parikh, Rajgopal Kannan, and Viktor Prasanna. Mixture of thoughts: Learning to aggregate what experts think, not just what they say. *arXiv preprint arXiv:2509.21164*, 2025.

Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer. *arXiv preprint arXiv:2512.22768*, 2025.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Moritz Haas, Sebastian Bordt, Ulrike von Luxburg, and Leena Chennuru Vankadara. On the surprising effectiveness of large learning rates under standard width scaling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In *International Conference on Machine Learning (ICML)*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE international conference on computer vision (ICCV)*, 2015.

Xu Owen He. Mixture of a million experts. *arXiv preprint arXiv:2407.04153*, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:30016–30030, 2022.

J. Hubbard. Calculation of partition functions. *Physical Review Letters*, 3(2):77–78, 1959.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Hans-Karl Janssen. On a lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties. *Zeitschrift für Physik B Condensed Matter*, 23(4):377–380, 1976.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Tianze Jiang, Blake Bordelon, Cengiz Pehlevan, and Boris Hanin. Hyperparameter transfer with mixture-of-expert layers. *arXiv preprint arXiv:2601.20205*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An Optimizer for Hidden Layers in Neural Networks. https://kellerjordan.github.io/posts/muon/, 2024.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, 2019.

Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, et al. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Jan Małaśnicki, Kamil Ciebiera, Mateusz Boruń, Maciej Pióro, Jan Ludziejewski, Maciej Stefaniak, Michał Krutul, Sebastian Jaszczur, Marek Cygan, Kamil Adamczewski, et al. mu-parametrization for mixture of experts. *arXiv preprint arXiv:2508.09752*, 2025.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sbornik (N.S.)*, 72(114)(4):507–536, 1967.

Paul Cecil Martin, E. D. Siggia, and H. A. Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423, 1973.

Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Mansour Zoubeirou Mayaki. Generalization and scaling laws for mixture-of-experts transformers, 2026. Manuscript.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. Olmoe: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

Radford M. Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*. Springer New York, 1996.

Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024a.

Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024b.

NVIDIA. Emerging-Optimizers. https://github.com/NVIDIA-NeMo/Emerging-Optimizers, 2025.

NVIDIA. Scalable training of mixture-of-experts models with megatron core. *arXiv preprint arXiv:2603.07685*, 2026.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

Babak Shahbaba and Radford Neal. Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, 10(8), 2009.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Jianlin Su. Travels in MoE: 6. promoting load balance via optimal assignment, 2026. Blog post, Scientific Spaces (in Chinese).

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*, 2024.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. *arXiv preprint arXiv:2502.05795*, 2025.

Volker Tresp. Mixtures of gaussian processes. *Advances in neural information processing systems*, 13, 2000.

Leena Chennuru Vankadara, Jin Xu, Moritz Haas, and Volkan Cevher. On feature learning in structured state space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Nikhil Vyas, Alexander Atanasov, Blake Bordelon, Depen Morwani, Sabarish Sainathan, and Cengiz Pehlevan. Feature-learning networks are consistent across widths at realistic scales. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024.

William Held. Make QB Fast Enough to Use. GitHub issue #3972, marin-community/marin, https://github.com/marin-community/marin/issues/3972, 2026.

Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie E Everett, Alexander A Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Greg Yang and Etai Littwin. Tensor programs ivb: Adaptive optimization in the infinite-width limit. *arXiv:2308.01814*, 2023.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv:2203.03466*, 2022.

Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

Appendices

Appendix Contents.

I Quick Reference **19**

A Appendix Roadmap and Suggested Reading Paths **19**

B Practitioner's Guide to Scaling Transformer MoEs with MSSP **21**

II Background and Extended Related Work **25**

C Background on Width-scaling Parameterizations **25**

D Detailed Related Work **26**

III Experiments **28**

E Experimental Setup **28**

E.1 MLP MoE experiments on TinyImagenet 28

E.2 Transformer MoE experiments 29

E.3 Figure details 31

F Additional Transformer MoE Experiments **33**

F.1 Learning rate sweeps and performance evaluations for Transformer MoEs 33

F.2 Expert diversity despite weight sharing 36

F.3 Layerwise gradient norms crossing Adam ε 37

F.4 Corrected load balance scaling 37

F.5 Refined coordinate checks for Transformer MoEs 39

F.6 MSSP for Muon 42

G Additional MLP MoE Experiments **45**

G.1 Learning rate sweeps for MLP MoEs 45

G.2 Fine-grained scaling evaluations in MLP MoEs 48

G.3 MSSP for sparse scaling with fixed K 62

G.4 Soft softmax routing collapses to uniform for μP in Regime I 63

G.5 MoEs require layerwise learning rate tuning 64

G.6 Random search and 2D multiplier tuning do not suffice 66

G.7 Global Adam ε induces width dependence at sufficient scale 66

IV Theory **68**

H Intuitive Explanations of the Shortcomings of μP for MoEs and How MSSP Fixes Them **68**

H.1 Overview . . . 68
H.2 Per-layer alignment: scaling of linear (forward) maps $W \cdot x$ . . . 69
H.3 MSSP Desiderata . . . 71
H.4 Backward scaling . . . 72
H.5 Aggregation layers . . . 72
H.6 Width dependence in μP and how MSSP fixes it . . . 72
H.7 Cross-expert aggregation: a taxonomy of mechanisms . . . 73
H.8 Summary: regime imbalances under μP and how MSSP resolves them . . . 77
H.9 Justification of MSSP in Regime II . . . 78
H.10 Approaches We Tried That Do Not Restore Scale Stability . . . 79

I Load balancing Approaches and their Scaling **80**

J Signal Propagation Analysis **81**

J.1 Preliminaries . . . 81
J.2 Standing assumptions and notation . . . 82
J.3 Scaling derivation for Regime II . . . 82
J.4 Scaling derivation for Regime III . . . 103

K DMFT Analysis for Regime I **118**

K.1 Setup . . . 118
K.2 Dynamics . . . 119
K.3 Mean field theory . . . 122
K.4 Averages of dual variables . . . 129
K.5 The coupling kernels A and B . . . 132
K.6 Final DMFT . . . 133

L DMFT Analysis for Regime II (MSSP) **135**

L.1 Architectural definitions . . . 135
L.2 Learning rates and initialization . . . 135
L.3 Gradient definitions . . . 135
L.4 DMFT kernels and order parameters . . . 136
L.5 Learning Dynamics and the Neural Tangent Kernel . . . 136
L.6 Decomposition of the MoE NTK . . . 137
L.7 Evolution of weights, preactivations, and pregradients . . . 138
L.8 Stochastic initial fields . . . 140
L.9 Deriving the DMFT Action . . . 141
L.10 Softmax . . . 143

L.11 Partition function . . . 144
L.12 Saddle point approximation . . . 146
L.13 Saddle-Point Equations . . . 147
L.14 Expert-local order parameters . . . 148
L.15 Hubbard-Stratonovich transformation . . . 148
L.16 DMFT Dynamics . . . 149

M DMFT Analysis for Regime II (μP) 150

N DMFT Analysis for Regime III (MSSP) 151
N.1 Architectural Definitions . . . 151
N.2 Learning Rates and Initialization . . . 151
N.3 Gradient Definitions . . . 151
N.4 DMFT Kernels and Order Parameters . . . 152
N.5 Learning Dynamics and the Neural Tangent Kernel . . . 152
N.6 Decomposition of the MoE NTK . . . 153
N.7 Evolution of weights, preactivations, and pregradients . . . 154
N.8 Stochastic Initial Fields . . . 155
N.9 Deriving the DMFT Action . . . 157
N.10 Softmax . . . 159
N.11 Partition function . . . 159
N.12 Saddle Point Approximation . . . 162
N.13 Saddle-Point Equations . . . 162
N.14 Expert-Local Order Parameters . . . 164
N.15 Hubbard-Stratonovich transformation . . . 164
N.16 DMFT Dynamics . . . 164

O DMFT Analysis for Regime III (μP) 166

Part I
Quick Reference

This part serves as a concise entrypoint to the appendix that enables the reader to orient themselves quickly and to navigate the subsequent parts in a structured way. Appendix A provides a roadmap and suggested reading paths depending on the reader's background and interests.

For practitioners, Appendix B concisely provides our *complete MSSP scaling prescription for modern transformer MoE architectures*.

A Appendix Roadmap and Suggested Reading Paths

The remaining parts of this appendix are self-contained and can be summarized follows:

- **Part II (Background and Extended Related Work)** provides background on the established infinite width literature, in particular the most common abc-parameterizations for standard dense networks, necessary to understand our corrected parameterizations for MoEs (Appendix C) and a detailed account of related work (Appendix D).
- **Part III (Experiments)** contains the experimental setup for our MLP MoE and Transformer MoE experiments (Appendix E), followed by extensive empirical evidence for our claims about learning rate transfer and scaling properties of μP and MSSP for both SGD and Adam in all 3 scaling regimes for Transformer MoEs (Appendix F) and MLP MoEs (Appendix G). Appendix F.4 evaluates the scaling properties of load balancing approaches. Appendix F.6 contains a verification of MSSP for Muon even under extremely sparse scaling.
- **Part IV (Theory)** provides three complementary levels of formalism: an intuitive scaling explanation (Appendix H), followed by a full forward and backward signal propagation analysis for deriving the μP and MSSP hyperparameter scaling rules in each scaling regime (Appendix J) followed by a self-consistent DMFT for μP and MSSP in each of the three scaling regimes (Appendices K to O). Note that Regimes II and III provide differing challenges that require their own resolutions, both in terms of scaling and analysis technique. Whenever the DMFT equations for a differing parameterization our routing mechanism follows from an analogous derivation, we omit the duplication and just provide the final set of self-consistent DMFT equations.

If you only have 30 minutes: read (i) Appendix B for our complete practical MSSP recipe, and (ii) Appendix H.8 for the regime-by-regime summary of why μP breaks and how MSSP fixes it.

For conceptual understanding: We recommend familiarizing oneself with the background on width scaling (Appendix C), then reading Appendix H for a comprehensive introduction into the scaling arguments necessary for understanding the signal propagation pathologies of MoEs and how to fix them. These arguments will provide the reader with tools that are more broadly applicable to similar scaling derivations for other non-standard architectures and optimizers. We then recommend reading the complete MSSP recipe for modern MoEs (Appendix B), and ending with a quick pass over the empirical results in Appendix F. In particular, we recommend skipping Appendices J to O during this first read.

Deep dive: For the ambitious reader, we recommend starting with the above path for developing sufficient conceptual understanding, before diving into the detailed signal propagation analyses in Appendix J, and the DMFT analyses in Appendices K to O. The signal propagation analysis decomposes all terms in the training dynamics and determines their scaling with arguments from random matrix theory. This makes the order of contributions of all observables explicit. The DMFT provides complementary value by clarifying how different variables interact via evolution of their kernels.

Practitioners: For readers that are primarily interested in the practical takeaways, we recommend starting with our complete practical MSSP recipe in Appendix B. Our Transformer setup is introduced in Appendix E.2 and our empirical findings are provided in Appendix F.

Notation mapping. The notation in the appendix differs slightly from the main paper. For readability, we write out the theory for a single block, but, as is common in the related DMFT and μP literature

(Yang et al., 2022; Jiang et al., 2026), every derivation follows in the same way for the architecture that stacks multiple blocks that is considered in the main paper.

Recall the following main paper notation.

MoE architecture. Given input $x_t \in \mathbb{R}^D$, the residual stream $h_t^0 = W_t^{\text{in}} x_t$ is transformed as:

$$h_t^l = h_t^{l-1} + K^{-\alpha_{\text{agg}}} \sum_{i \in \text{top-K}} \phi_{t,i}^l \cdot W_t^{l,\text{out},i} \, \varphi(W_t^{l,\text{in},i} h_t^{l-1}), \qquad \phi_{t,i}^l = \sigma\big(\beta \, (Q_t^l h_t^{l-1})_i\big),$$

for $l \in [L]$, and is finally transformed into output logits $f_t = (W_t^{\text{out}})^\top h_t^L$. Here φ is a coordinatewise nonlinearity, β a tunable inverse temperature, and σ either sigmoid or softmax.

For theoretical purposes, it suffices to analyze the signal propagation through the following single MoE block without residual connection.

Minimal MoE architecture. For readability, we work with the minimal MoE architecture

$$h^1 = W^1 x, \tag{A.1}$$
$$\psi = Q h^1, \tag{A.2}$$
$$\phi = \sigma(\psi), \tag{A.3}$$
$$h^{2,i} = W^{2,i} h^1, \qquad i = 1, \dots, M, \tag{A.4}$$
$$h^{3,i} = W^{3,i} h^{2,i}, \tag{A.5}$$
$$h^3 = \frac{1}{M} \sum_{i=1}^{M} \phi_i \, h^{3,i}, \tag{A.6}$$
$$f = W^4 h^3, \tag{A.7}$$

where $x \in \mathbb{R}^D$ is the input, $f \in \mathbb{R}$ the scalar output, and the trainable parameters are

$$\theta \;=\; \big(W^1,\; Q,\; \{W^{2,i},\, W^{3,i}\}_{i=1}^{M},\; W^4\big)$$

with shapes $W^1 \in \mathbb{R}^{N \times D}$, $Q \in \mathbb{R}^{M \times N}$, $W^{2,i} \in \mathbb{R}^{N_e \times N}$, $W^{3,i} \in \mathbb{R}^{N \times N_e}$, $W^4 \in \mathbb{R}^{1 \times N}$. The same scaling exponents hold under sigmoid gating with aggregation factor $\alpha_{\text{agg}} = 1$ and under softmax gating without an explicit aggregation multiplier.

In particular, here we write h^1 instead of h^l, and $h^3 = \frac{1}{M} \sum_{i=1}^{M} \phi_i \, h^{3,i}$ instead of $h^{l+1} = h^l + h^{\text{agg}}$.

B Practitioner's Guide to Scaling Transformer MoEs with MSSP

This section collects what is needed to scale a Transformer MoE with MSSP. Table B.1 gives the complete prescription for AdamW and Table B.2 for Muon. The scaling regimes are illustrated in Figure B.1.

Practical impact.

- **Tune once at small width, then scale.** Under MSSP, the learning rate tuned at base width $N = 256$ transfers to large widths and depths (Figure 5). Under μP the optimal learning rate slightly shifts in Regimes II and III. Under standard parameterization it shifts as well, and the best loss much worse than in MSSP at large scale (Figure F.5).
- **Better models at no extra cost.** MSSP changes only the initialization relative to μP. In Regime III its advantage over μP grows with width, from 0.002 at $N = 256$ to 0.036 at $N = 2048$ after 2.5B tokens.
- **Reliable architecture comparisons.** Expert granularity (number, width and active count of experts) is usually chosen by extrapolating from small models. Without scale stability, such comparisons are confounded by width-dependent training dynamics. MSSP removes this confound in all co-scaling regimes.

Recommended recipe for AdamW.

1. **Scale all MoE dimensions with width (Regime III).** Grow M and N_e proportionally to N, and ablate the optimal scaling of K. This was the best of the scaling paths we compared at matched active parameters (Figure F.1). MSSP exists whenever the active expert width KN_e grows at least proportionally to N, which includes fixed K with $N_e \asymp N$ (Appendix G.3). If $KN_e = o(N)$, as for fixed K and fixed N_e, no parameterization in our class is maximally scale stable.
2. **Tie the expert initialization in groups.** Partition the experts into $G \asymp N/N_e$ groups and give all experts of a group the same initial weights, drawn independently across groups. In Regime III ($G = 1$) all experts start identical. Tying makes the experts' propagating updates add up coherently in the aggregation instead of cancelling, so that every term of the aggregated MoE output stays $\Theta(1)$. Experts are at least as specialized as under independent initialization after only 600 steps (Figure 5).
3. **Initialize the router to zero on sparse paths and route stochastically.** For $K = o(N)$, the standard router initialization $N^{-1/2}$ unbalances the contributions of the different paths to the router gradient, so the router is initialized to zero (Table B.1). Zero router logits then require randomness in the routing mechanism. With deterministic top-K selection and tied experts, all tokens select the same experts, which receive identical gradients and never separate. We recommend sampling the active experts from the router distribution with Gumbel sampling (Kool et al., 2019; Sukhbaatar et al., 2024; Fein-Ashley et al., 2025) over additive routing noise. Gumbel sampling is uniform at zero initialization, learns the router smoothly instead of through sharp deterministic routing flips, and regularizes expert specialization. For $K \asymp N$, as in Regimes II and III, the router can be initialized with standard deviation $N^{-1/2}$.
4. **Aggregate with K^{-1}.** Multiply the sum of the K gate-weighted expert outputs by K^{-1}. Sigmoid gates are $\Theta(1)$ each and therefore do not provide this factor, unlike top-K-renormalized softmax gates.
5. **Scale learning rates and ϵ per tensor.** Use the tensorwise scaling rules for Adam learning rates and Adam ε specified in Table B.1. For most layers, Adam learning rates scale as $\propto (\text{fan-in})^{-1}$, since this is the length of the inner product between the weight update and the incoming activations. Each expert receives a fraction K/M of the tokens, each weighted by K^{-1}, so router and expert gradients carry a factor M^{-1} which affects the ε but not the Adam learning rate scaling. This holds for every implementation that backpropagates the token-averaged loss. With a global $\epsilon = 10^{-8}$, expert output updates start to vanish at width about 4500 (Appendix F.3).
6. **Load balancing with QB expert biases.** QB expert biases do not require hyperparameter tuning and keep routing weights stable over training and across widths (Figure F.10). The auxiliary loss needs a width-dependent prefactor (Appendix I), still drives router logits down over training, and is worse than QB by 0.02, 0.07 and 0.11 at $N = 512$, 1024 and 2048 (Figure 5). We compute

router logits in FP32 with sigmoid gates, QB biases enter only the top-K selection, and no tokens are dropped.

7. **Scale residual branches by** L^{-1} when scaling depth, and average rather than sum the per-layer auxiliary and z-losses.
8. **Width-independent weight decay.** PyTorch's AdamW decays each tensor by $\eta_t \, s_W \lambda$ per step, where s_W is the tensor's scale dependent learning rate multiplier. Dividing each group's weight decay by s_W makes every tensor decay by $\lambda \eta_t$ per step, which follows the learning rate schedule but not the width. Dividing additionally by the peak learning rate η recovers the decoupled weight decay of Loshchilov and Hutter (2019), $\lambda \eta_t / \eta$, which also does not change when η is retuned.
9. **Tune multipliers once at base width.** For optimal transfer and performance, we tune one learning rate multiplier per weight type, a global initialization multiplier, the attention logit multiplier and the global learning rate η at small base size (Section E.2.2 and Table E.1). They are constant factors on top of the scaling rules.

Essential rules, safeguards and calibration. The *scaling exponents* of Table B.1 are essential. They comprise tied group initialization, router initialization, the K^{-1} aggregation, layerwise learning rates and ϵ, and L^{-1} residual scaling, and omitting any of them reintroduces width-dependent dynamics, as the μP and standard parameterization baselines show. Layerwise ϵ matters at frontier scale. Expert output gradient entries decay as N^{-2} in Regime III and cross the common global $\epsilon = 10^{-8}$ at width about 4500, and at even smaller scale late in training as gradients shrink, beyond which expert updates vanish (Appendices F.3 and G.7). To avoid numerical underflows, the gradients or Adam moments can be scaled up rather than shrinking ϵ to tiny values. The *multipliers* are a finite-size calibration and essential for clean transfer. They are tuned once at base width and improved the loss by more than 0.025 nats over all-one multipliers in our tuning runs.

MSSP initializes the expert output weights $\sqrt{M}$ times larger, so per-expert outputs are $\Theta(\sqrt{M})$ while their aggregate is $\Theta(1)$. With $W^{\mathrm{out},i} = \sqrt{M}\, \widetilde{W}^{\mathrm{out},i}$, a fused kernel that computes $\widetilde{W}^{\mathrm{out},i} h^{\mathrm{in},i}$, scales it by $\phi_i / \sqrt{M}$ and accumulates directly into the aggregate never materializes the large tensors. We did not need this up to $M = 512$. All our experiments train stably in BF16 mixed precision with standard kernels and without loss scaling or unit scaling.

Muon Optimizer. Muon (Jordan et al., 2024) replaces the update of a weight matrix by the orthogonalized momentum. We follow NVIDIA Emerging-Optimizers (NVIDIA, 2025) exactly: momentum $B_t = \mu B_{t-1} + (1-\mu) G_t$ with $\mu = 0.9$ and no Nesterov, orthogonalization $O_t = \mathrm{orth}(B_t) \approx U V^\top$ by 8 Newton–Schulz steps with Polar Express coefficients (Amsel et al., 2025) in bfloat16, and the update

$$W_{t+1} = (1 - \lambda \, \eta_t / \eta_{\max}) \, W_t - \eta_t \sqrt{\texttt{fan_out}/\texttt{fan_in}} \, O_t .$$

Since $\|O_t\|_{\mathrm{op}} = 1$, the factor $\sqrt{\texttt{fan_out}/\texttt{fan_in}}$ makes every update scale-stable in RMS to RMS operator norm, following the spectral condition for feature learning (Yang et al., 2023; Pethick et al., 2025). Every expert is orthogonalized as its own matrix, and so are the query, key and value blocks. Orthogonalizing the stacked expert tensor instead would couple the experts, and for similar expert gradients, as after tied initialization, it shrinks each expert's update roughly as $M^{-1/2}$. Muon updates the router, the expert and the attention matrices. Embedding, unembedding and normalization gains keep AdamW with the MSSP scaling of Table B.1.

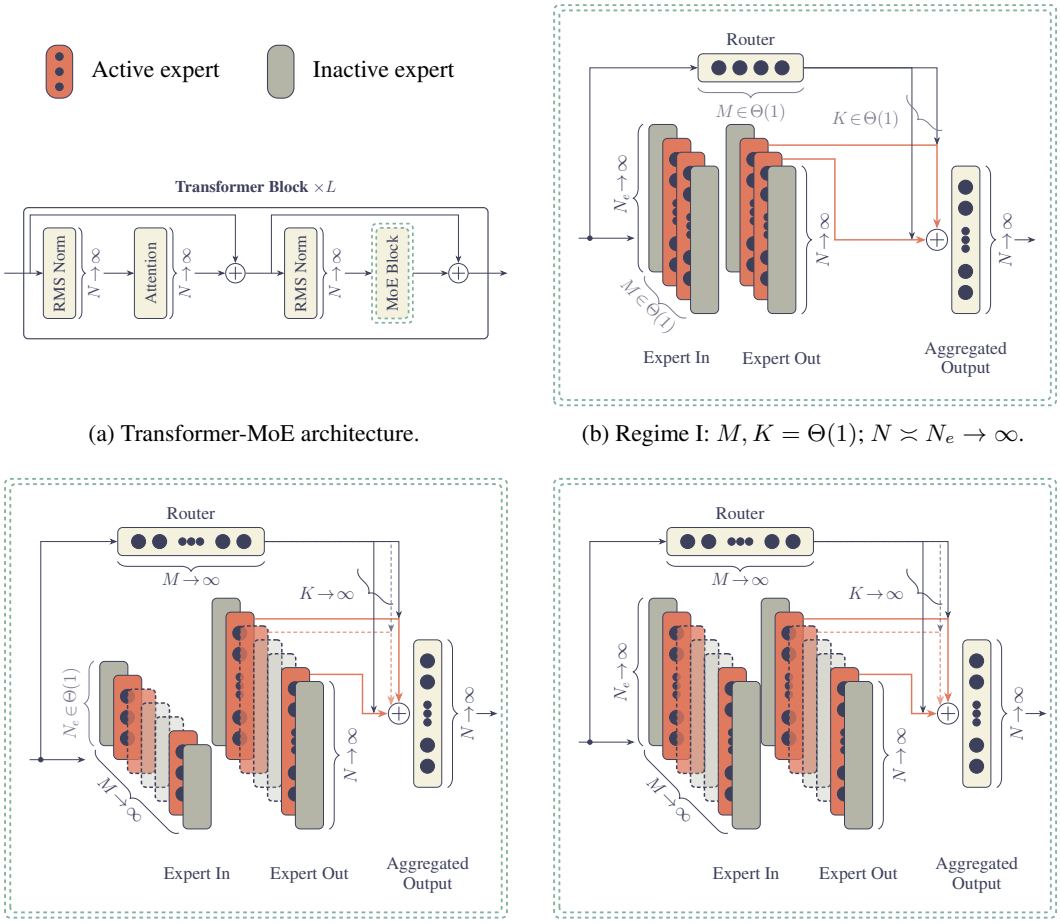
Scaling Muon. Muon normalizes away the gradient magnitude, so the K- and M-dependent gradient scales that the Adam prescription corrects for with ϵ do not enter. The scalings of initialization, propagating updates and expert aggregation remain unaffected. This results in MSSP for Muon as summarized in Table B.2. In Figures F.20 and F.21, we empirically verify scale stability of effective and propagating updates and of all terms of the h^{agg} decomposition.

Table B.1: **MoE AdamW hyperparameter scaling.** Columns correspond to different MoE scaling regimes of width N, expert width N_e, number of experts M and top-K. The last column states our conjectured MSSP whenever $K \cdot N_e \in \Omega(N)$, and Regimes I–III are special cases of it. In the case $K \cdot N_e \in o(N)$, no MSSP exists in the considered space of parameterizations. Width and depth control terms are highlighted. Dimensions that scale proportionally could be replaced by a single width multiplier $m_N = N/N_0$ w.r.t. some base width N_0. The experts are partitioned into $G \asymp N/N_e$ groups of $M/G \asymp MN_e/N$ experts each. $\texttt{tied}_G$: expert weights are shared within each group and independent across groups at initialization. This gives independent experts in Regime II ($G \asymp M$) and fully tied experts in Regime III ($G = 1$). In Regime I it gives fully tied experts, an alternative to the independent initialization listed in its column. Adam ϵ scales as the batch-averaged gradient entries $\epsilon \asymp |g|$ (a fraction K/M of the tokens reaches each expert, each with weight K^{-1}). In AdamW as proposed by Loshchilov and Hutter (2019), weight decay should stay fixed across model scales. In the PyTorch implementation of AdamW, the weight decay multiplier should always be the inverse of the learning rate multiplier so that the effective weight decay $\eta \cdot \text{wd}$ stays scale-independent. The minimal fine-graining of multiplier tuning at small model size recommended includes a global initialization multiplier and learning rate multipliers based on layer type (embedding, readout, gain, router, expert in, expert out, other).
$*$: Router zero init. requires initial randomness in the routing mechanism or top-k tie breaking.
†: For the DeepSeek-normalized load-balancing loss, which is scale stable. The standard Fedus et al. (2022) loss additionally needs the factor $(MK)^{-1}$ (Appendix I). The multiplier L^{-1} averages instead of sums the per-layer terms. We recommend using HP-free QB expert biases instead.

Parameterization	**Regime I** $N, N_e \to \infty$, M, K fixed	**Regime II** $N, M, K \to \infty$, N_e fixed	**Regime III** $N, M, K, N_e \to \infty$	**General path** $N \to \infty$, $K \cdot N_e \in \Omega(N)$
Emb. Init. Std.			1	
Emb. Adam LR			1	
Emb. Adam ϵ			N^{-1}	
Pre-LN Init. Std.			-	
Pre-LN Adam LR			1	
Pre-LN Adam ϵ			$N^{-1}L^{-1}$	
Hidden Init. Std.			$(\text{fan-in})^{-1/2}$	
Hidden Adam LR			$(\text{fan-in})^{-1}$	
Hidden Bias Adam LR			1	
Hidden Adam ϵ			$N^{-1}L^{-1}$	
MoE routing & experts				
Expert init. groups G	M	M	1	NN_e^{-1}
Router Init. Std.	0^*	$N^{-1/2}$	$N^{-1/2}$	0^* if $K = o(N)$ $N^{-1/2}$ if $K \asymp N$
Expert In Init. Std.	$N^{-1/2}$	$N^{-1/2}$	$(N^{-1/2})^{\texttt{tied}}$	$(N^{-1/2})^{\texttt{tied}_G}$
Expert Out Init. Std.	$N_e^{-1/2}$	$M^{1/2}N_e^{-1/2}$	$(N_e^{-1/2})^{\texttt{tied}}$	$(N^{1/2}N_e^{-1})^{\texttt{tied}_G}$
Router Adam LR			N^{-1}	
Expert In Adam LR			N^{-1}	
Expert Out Adam LR			N_e^{-1}	
Router Adam ϵ	L^{-1}	$M^{-1}L^{-1}$	$M^{-1}L^{-1}$	$M^{-1}L^{-1}$
Expert In Adam ϵ	$N^{-1}L^{-1}$	$M^{-1}L^{-1}$	$N^{-1}M^{-1}L^{-1}$	$M^{-1}N_e^{-1}L^{-1}$
Expert Out Adam ϵ	$N^{-1}L^{-1}$	$N^{-1}M^{-1}L^{-1}$	$N^{-1}M^{-1}L^{-1}$	$M^{-1}N^{-1}L^{-1}$
Aggregation multiplier			K^{-1}	
Aux. load-balancing loss†			L^{-1}	
Router z-loss			L^{-1}	
QB Expert Biases			-	
MHA / MoE Residual	$X^l + L^{-1} \cdot F(LN(X^l))$, for any block $F \in \{\text{MHA}, \text{MoE}\}$			
Final-LN Init. Std.			-	
Final-LN Adam LR			1	
Final-LN Adam ϵ			N^{-1}	
Unemb. Init. Std.			N^{-1}	
Unemb. Adam LR			N^{-1}	
Unemb. AdamW ϵ			1	
Unemb. Fwd.			1	

Table B.2: **MSSP for Muon.** For the general case $N \to \infty$ with $K \cdot N_e = \Omega(N)$. Learning rates are relative to the single tuned rate η, and Muon applies them to the orthogonalized momentum of each matrix separately. Experts share their initial weights within each of $G \asymp N/N_e$ groups ($\texttt{tied}_G$). Everything else as in Table B.1.

Weights	**Optimizer**	**Init. std**	**Learning rate**
Attention W^q, W^k, W^v, W^o	Muon, per block	$N^{-1/2}$	1
Router Q	Muon	0 if $K/N \to 0$, else $N^{-1/2}$	$(M/N)^{1/2}$
Expert in $W^{\mathrm{in},i}$	Muon, per expert	$(N^{-1/2})^{\texttt{tied}_G}$	$(N_e/N)^{1/2}$
Expert out $W^{\mathrm{out},i}$	Muon, per expert	$(N^{1/2}N_e^{-1})^{\texttt{tied}_G}$	$(N/N_e)^{1/2}$
Embedding, unembedding, gains	AdamW	as in Table B.1	as in Table B.1



(a) Transformer-MoE architecture.

(b) Regime I: $M, K = \Theta(1)$; $N \asymp N_e \to \infty$.

(c) Regime II: $N \asymp M \asymp K \to \infty$; $N_e = \Theta(1)$.

(d) Regime III: $N \asymp N_e \asymp M \asymp K \to \infty$.

Figure B.1: Transformer-MoE architecture and the MoE block under the three scaling regimes.

Part II

Background and Extended Related Work

This part collects background on existing width-scaling parameterizations and a detailed account of related-work.

C Background on Width-scaling Parameterizations

Table C.1 gives an overview over the most common abc-parameterizations for training dense networks with SGD. Yang and Hu (2021) use width-dependent weight multipliers $n^{-a} \cdot W$ to scale gradients and avoid layerwise learning rates. This extra degree of freedom introduces an equivalence class of valid μP abc-parameterizations. Mean field parameterization (Mei et al., 2018; Chizat and Bach, 2018; Bordelon and Pehlevan, 2022) is equivalent to μP, using the multipliers $\texttt{fan-in}^{-1/2}$ except N^{-1} in the output layer. For practical readability, we remove this extra degree of freedom in the main paper by using naive weight multipliers $a = 0$ for all trainable weights.

			Weight-multiplier version			**Weight-multiplier-free version**		
			Input-like	Hidden-like	Output-like	Input-like	Hidden-like	Output-like
SP	$\alpha_l \cdot W^l$,	$\alpha_l \propto$		-		1	1	1
	$\mathcal{N}(0, \sigma_l^2)$,	$\sigma_l \propto$				1	$N^{-1/2}$	$N^{-1/2}$
	$\eta_l \cdot \nabla_{W^l} \mathcal{L}$,	$\eta_l \propto$				N^{-c}	N^{-c}	N^{-c}
NTP	$\alpha_l \cdot W^l$,	$\alpha_l \propto$	1	$N^{-1/2}$	$N^{-1/2}$	1	1	1
	$\mathcal{N}(0, \sigma_l^2)$,	$\sigma_l \propto$	1	1	1	1	$N^{-1/2}$	$N^{-1/2}$
	$\eta_l \cdot \nabla_{W^l} \mathcal{L}$,	$\eta_l \propto$	1	1	1	1	N^{-1}	N^{-1}
μP	$\alpha_l \cdot W^l$,	$\alpha_l \propto$	$N^{1/2}$	1	$N^{-1/2}$	1	1	1
	$\mathcal{N}(0, \sigma_l^2)$,	$\sigma_l \propto$	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$	1	$N^{-1/2}$	N^{-1}
	$\eta_l \cdot \nabla_{W^l} \mathcal{L}$,	$\eta_l \propto$	1	1	1	n	1	N^{-1}

Table C.1: **(Common abc-parameterizations)** Standard parameterization (SP), neural tangent parameterization (NTP) and the maximal update parameterization (μP) for MLPs trained with SGD in their multiplier version which purely adapts the architecture and allows width-independent global learning rates (*left*) and in their weight multiplier-free version (*right*). Parameterizations differ in their layerwise choice of width-dependent weight multipliers α_l, initialization variances σ_l and learning rates η_l. Weight multiplier-free representatives of an abc-equivalence class purely adapt the optimization algorithm highlighting the fact that parameterizations effectively only induce layerwise learning rates. Knowing that μP correctly scales the updates in all layers, observe that the input- and hidden-layer learning rates in NTP induce vanishing updates. The same holds in SP when choosing $c \geq 1$ as is necessary for avoiding logit blowup in the infinite-width limit. SP with large learning rates $c = 1/2$ recovers stable hidden-layer feature learning despite logit divergence (Haas et al., 2025).

μP for Adam in standard architectures. Adam requires the same layerwise initialization variances as SGD. For clarity, consider the weight-multiplier-free version of μP. Since Adam normalizes its update, but the update direction remains correlated with the incoming activations, Adam's layerwise learning rate scaling simplifies to $\eta(W) = 1/\texttt{fan-in}(W)$ for any linear weight matrix $W : \mathbb{R}^{\texttt{fan-in}} \to \mathbb{R}^{\texttt{fan-out}}$. For faithfulness, Adam $\varepsilon(W)$ should always scale as the gradient RMS norm of W, that is $\Theta(1)$ for the readout layer and $\Theta(N^{-1})$ otherwise.

Table C.2: **μP-heuristic** $(b, c_{\text{SGD}}, c_{\text{Adam}}, d_{\text{Adam}})$ by layer type. Both SGD and Adam share the same initialization $\mathcal{N}(0, N^{-b})$. SGD requires learning rates $\eta_N = N^{-c_{\text{SGD}}}$, and Adam requires learning rates $\eta_N = N^{-c_{\text{Adam}}}$ and gradient scaling $N^{-d_{\text{Adam}}} \cdot \nabla_W \mathcal{L}$.

	b	c_{SGD}	c_{Adam}	d_{Adam}
Input-like	0	-1	0	1
Matrix-like	0.5	0	1	1
Output-like	1	1	1	0

Baseline Parameterizations: SP and the μP heuristic. Standard parametrization (SP) uses Xavier/Glorot-type variance scaling (Glorot and Bengio, 2010; He et al., 2015), i.e. typical dense layers take $b_W = \frac{1}{2}$ (with common exceptions for input embeddings), and uses the same optimizer hyperparameters such as learning rate and Adam ε for all layers.

In contrast, μP is designed so that maximal-update behavior (and, for adaptive methods, faithfulness) hold in the infinite-width limit for dense networks (Yang and Hu, 2021; Yang and Littwin, 2023). Practitioners often use the following *μP heuristic*: classify a weight tensor by whether it maps fixed$\to n$ (*input-like*), $n \to n$ (*matrix-like*), or $n \to$fixed (*output-like*), and assign (b, c, d) accordingly as summarized in Table C.2. For MoEs in Regime I, this heuristic treats expert MLP weights matrix-like, while router projections are output-like due to the fixed number of experts M. In Regime I, μP heuristic indeed satisfies all μ desiderata, but it does not suffice in Regimes II and III, when co-scaling the number of experts $M \to \infty$.

D Detailed Related Work

Mixture of Experts models. Mixture of experts models have a long history in machine learning (Jacobs et al., 1991), with instantiations such as hierarchical models (Jordan and Jacobs, 1994), Gaussian processes (Tresp, 2000) or Dirichlet processes (Shahbaba and Neal, 2009). After the first application in deep networks by Eigen et al. (2013), Shazeer et al. (2017) sparked the use of MoEs in modern architectures, using the top-k operation for sparsification and trainable noise for load balancing. Fedus et al. (2022) lowered communication and computational cost for improved scalability, using Kaiming initialization and simplifying the load balancing auxiliary loss. Most MoE architectures still use a width-independent initialization of 0.006 (Dai et al., 2024). The trend in recent frontier models goes toward increased fine-graining, from 2 out of 8 in Mixtral 8x7B and 8x22B (Jiang et al., 2024) to 8 out of 256 in Deepseek-R1 (Guo et al., 2025). Zoph et al. (2022) introduce the router z-loss for router logit regularization; Haas et al. (2025) suggest that miss-scaling causes logit divergence, hence the additional regularization might be unnecessary in MSSP. Insightful empirical investigations can be found in the fully open OlMoE report (Muennighoff et al., 2024). They identify removing the need for the load balancing loss as an important future direction, since it significantly constrains the model flexibility and may prevent experts from sufficiently specializing over the course of training. To date, the theory of MoEs is very limited. Chen et al. (2022) show that MoEs are able to learn data with cluster structure using two-layer nonlinear convolutional neural networks as experts.

Scaling theory. Infinite-width theory dates back to Neal (1996); Matthews et al. (2018); Jacot et al. (2018). The Tensor Program series (Yang, 2019; Yang and Hu, 2021; Yang and Littwin, 2023) has served as a crucial tool for developing flexible and general infinite-width theory allowing to study non-vanishing feature learning limits (Everett et al., 2024; Vankadara et al., 2024; Haas et al., 2025). This theory gave rise to the maximal update parameterization (μP) (Yang et al., 2022) under which the optimal learning rate and further dynamical properties (Noci et al., 2024b) have been observed to transfer to larger model width. The mean-field parameterization is equivalent to μP. DMFT has proven useful as it allows to simulate from the limit (Chizat and Bach, 2018; Mei et al., 2018; Bordelon and Pehlevan, 2022; Bordelon et al., 2024a). Similar techniques allow infinite depth limits (Bordelon et al., 2024b; Hayou and Yang, 2023; Noci et al., 2024a; Chaintron et al., 2026). These works apply depth-dependent scaling factors like $L^{-1/2}$ or L^{-1} (Dey et al., 2025) to the residual stream; it remains open whether these interventions suffice. Dey et al. (2025) show that L^{-1} outperforms $L^{-1/2}$ in Transformer training, but Sun et al. (2025) argue that scaling should depend on the layer index.

Scaling mixture of experts models. Clark et al. (2022); Krajewski et al. (2024) and Dai et al. (2024, Figure 3) all suggest that increasingly fine-grained experts outperform fewer larger ones in terms of learning performance, with saturating gains. Boix-Adsera and Rigollet (2025) prove an exponentially improved expressivity due to fine graining. Mayaki (2026) provides theory showing that performance can improve either by scaling active capacity or by increasing the number of experts, depending on the dominant bottleneck. In practice, eventually the increased communication cost of routing dominates compute-optimality considerations, but increasingly efficient implementations are being developed. NVIDIA (2026) provides a highly optimized distributed training framework across scaling regimes, that also allows efficient training with many fine-grained experts. He (2024) proposes PEER layers that take fine graining to the extreme. However, all existing works use suboptimal model scaling, hence it will be crucial to evaluate the impact of granularity on scaling in MSSP.

Independent and concurrent work. Jiang et al. (2026) derive a parameterization for *Sign SGD* in Regime III via a DMFT analysis, and present Transformer MoE experiments showing approximate learning rate transfer when scaling one axis at a time with all others fixed. We derive μP for *SGD and Adam across all three Regimes I, II, and III* using a signal propagation analysis complemented by

a DMFT for each regime, and our experiments scale jointly along the axes each regime prescribes. Our μP in Regime III coincides with theirs up to a larger router initialization. More importantly, we identify that in Regimes II and III, μP principles do not suffice to robustly yield learning rate transfer or monotonic improvement with scale in MoEs, due to scale-dependent dynamics, and we propose MSSP as a resolution. In Regime III, our DMFT for MSSP is qualitatively distinct: it exhibits a four-layer mean-field hierarchy (versus a three-layer hierarchy under μP), and we additionally derive novel DMFT limits for Regimes I and II under MSSP.

Under Regimes II and III ($M, K \to \infty$), μP differs from shape-based μP-heuristic. For dense networks, it is common to scale the layerwise initialization variance and learning rate of a weight matrix purely based on its shape (Yang et al., 2022, 2023), distinguishing input-like (fixed $\to \infty$), hidden-like ($\infty \to \infty$) and output-like ($\infty \to$ fixed) layers. Małaśnicki et al. (2025) apply this heuristic to MoEs, treating the router output-like in Regime I ($M, K \in \Theta(1)$). Table B.1 shows that μP-heuristic coincides with μP in Regime I, but fails to induce all μ-desiderata in Regimes II and III.

Part III
Experiments

This part covers the experimental setup for our MLP MoE as well as Transformer MoE experiments in Appendix E, provides more details concerning the main figures in Appendix E.3, followed by detailed empirical evidence for our claims about learning rate transfer and scaling properties of μP and MSSP for both SGD and Adam in all 3 scaling regimes for Transformer MoEs (Appendix F) and MLP MoEs (Appendix G). We close with further empirical scaling insights that practitioners should be aware of in Sections G.4 to G.7.

E Experimental Setup

Open source code to fully reproduce our experiments can be found at:

https://github.com/vankadara-lab/mssp-moe

E.1 MLP MoE experiments on TinyImagenet

MoE Architecture. We train the same embedding-MoE-readout architecture with 2-layer fully connected experts that we analyse theoretically (without scale-dependent weight multipliers) using PyTorch (Paszke et al., 2019). It consists of a linear input layer with GeLU activation function, a linear router layer, 2-layer expert MLPs with GeLU activation function, followed by $M^{-1}\cdot$ sigmoid aggregation, followed by a linear output layer. In this architecture without normalization layers and residual connections, instabilities in the scaling procedure become apparent at more moderate scale.

Data. We train single pass over all 50000 available images from 100 classes of TinyImageNet (Le et al., 2015) for 1000 update steps with batch size 50. Our initial experiments used CIFAR-10 (Krizhevsky et al., 2009), but robustly benefitting from expert specialization in MoEs requires more diverse data.

We vary the optimizer (SGD and Adam), the scaling regime (fixed M, fine-grained and all-scaling), the routing mechanism (soft routing and top-k routing) and the parameterization (μP and MSSP).

E.1.1 Scaling configurations

Unless otherwise specified, we initialize the last layer to zero.

MSSP. Depending on the optimizer and scaling regime, Table B.1 summarizes our proposed parameterization MSSP as a function of width N, expert width N_e and number of experts M. Scaling of non-MoE trainable weights remains in μP.

μP. Table B.1 specifies μP for SGD and Adam in each scaling regime. The provided HP scaling rules achieve stability as well as scale-independent effective and propagating updates after sufficiently many update steps, as verified in Appendix G.2. However their initial vanishing signal propagation induces more scale dependence in the dynamics, delayed learning and/or no monotonic improvement with scale.

For Regime I, we compare the baseline router init std $1/N$ (standard muP) versus zero router initialization for improved scale independence. Both variants barely differ.

E.1.2 Multiplier tuning

For MLPs, we use base width 1, which uncovers scaling properties of each parameterization at more moderate scale. Before running extended evaluations or learning rate sweeps, we tune multipliers at small model scale $N = 128$. The network at size $N = 128$ has dimensions ($N = N_e = 128,\ M = 8$) in Regime I, ($N = 128,\ M = N/16,\ N_e = 16$) in Regime II and ($N = N_e = 128,\ M = N/16$) in Regime III, from which we scale up proportionally $N \to \infty$.

We observe that the optimal multipliers can be as far as 10^6 from all ones (since the gradient RMS norm of the expert output layer decays as fast as N^{-2} in Regimes II and III (see Sections G.2.2

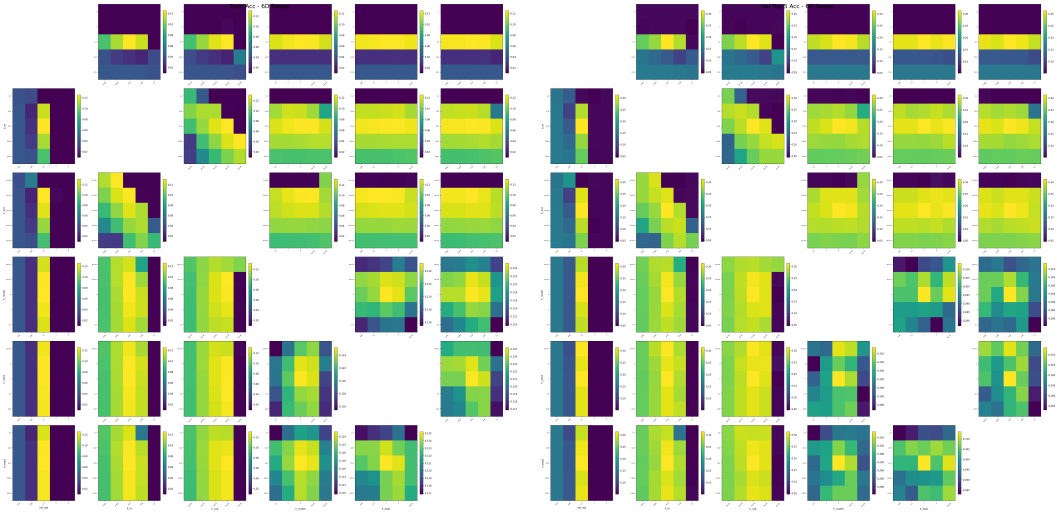


Figure E.1: **6D multiplier sweeps at small scale $N = 128$ (MSSP, SGD, fine-grained).** 2D heatmaps showing training accuracy (left) and top-5 validation accuracy (right) of all HP pairs, while fixing all remaining HPs at the optimum. The optimum in each multiplier remains consistent across HP pairs. Train and validation optimum can slightly differ.

and G.2.3)), and some update terms can be negligibly small without tuning, resulting in non-balanced learning (Appendix G.5).

For each configuration of optimizer, parameterization, routing mechanism and scaling regime, we tune the following set of multipliers:

- global initialization variance,
- global learning rate,
- layerwise learning rate multiplier (input, router, expert input, expert output, output layer)

Our tuning algorithm proceeds as follows. We first tune the learning rate at width $N = 128$, then run a broad 6D grid of the above init and layerwise learning rate multipliers, then run a full 6D grid with 5 grid points per dimension centered at the optimum from the previous stage at multiplicative resolution 4. This is possible since several small model training runs in parallel take less than 40 seconds on one A100 GPU.

Figure E.1 shows the example of the last step for the case of SGD MSSP in Regime II. Top-5 validation accuracy optima tend to me more localized than training accuracy optima. Hence we choose the optimum based on top-5 validation accuracy. The optimal initialization variance multiplier tends to be very clearly localized and essential in all combinations of optimizer, parameterization and scaling regime.

Appendix G.6 shows that a cheaper multiplier tuning procedure, which starts with a random search over the grid, followed by extensive 2D sweeps over all combinations of multipliers does not suffice to reliably provide a near-optimal combination of multipliers.

E.1.3 Coordinate checks and learning rate sweeps

For each scaling configuration, we repeat the same experiment for 4 independent random seeds, affecting the random weight initialization as well as data shuffling. Uncertainty bands denote 2σ-confidence bands. If not mentioned otherwise, MLP experiments use soft routing, and the same compute budget for multiplier tuning was invested for μP and MSSP.

E.2 Transformer MoE experiments

We adapt the nanoGPT and nanoMoE codebases to train transformer models with a simple but modern architecture.

General Architecture and Training Details. By default, models have 8 blocks with mixture-of-expert layers (no dense layers). We scale the number of attention heads proportionally with width, while the head dimension remains fixed at $\texttt{d_head} = 64$. We use standard $\texttt{d_head}^{-1/2}$ attention scaling, which can be viewed as a tunable multiplier since $\texttt{d_head}$ remains fixed. We use pre-attention and qk-RMSNorm (Wortsman et al., 2024). We train for 4768 steps (2.5B tokens) using AdamW with a single, tuned maximal learning rate, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$ with layerwise scaling according to the respective parameterization, sequence length 1024, batch size 512, 1400 steps of warmup followed by cosine learning rate decay to 10% of the maximal learning rate, weight decay 0.1 divided by each tensor's LR scale so that the effective weight decay is width-independent, and gradient clipping at 1.

MoE layers. Mixture-of-expert layers are scaled according to the respective scaling regime, using a sigmoid activation function, sigmoid routing and quantile balanced (QB) expert biases as explained in Appendix I. We also ablate auxiliary loss with weight 0.01 for load balancing (Fedus et al., 2022), but use QB expert biases if not stated otherwise. We use token-choice routing and the number of active experts per token is half the total number of experts. We do not drop tokens.

Architecture for Regime III. In all all-scaling Regime III experiments, we use the following architecture. At base width $N = 256$ each of the 8 MoE layers has $n_{\text{exp}} = 8$ experts of hidden width $N_e = N/2 = 128$; at $N = 2048$ the same layer has $n_{\text{exp}} = 64$ experts of hidden width $N_e = 1024$.

Architecture for Regime I. We use the same base size as Regime III, and scale keeping the number of experts fixed to $n_{\text{exp}} = 8$.

Architecture for Coordinate Checks in Regime II. For the coordinate checks in fine-grained Regime II, we use 8 MoE layers with a fixed expert hidden width of $N_e = 16$. For keeping a total expansion ratio of 4, the expert count is $M = 64$ at $N = 256$ and grows to $M = 512$ at $N = 2048$.

Architecture for Learning Rate Sweeps for Regime II. For the learning rate sweeps in fine-grained Regime II, we use a more computationally efficient fine-grained architecture due to compute constraints. We use 4 blocks alternating between dense and MoE layers (Dense, MoE, Dense, MoE). At width $N = 256$, each MoE block holds $M = 32$ experts of fixed hidden width $N_e = 32$, scaling to $M = 256$ experts at $N = 2048$. This amounts to 408M parameters at $N = 2048$.

Data. Models are trained on 2.5B tokens of dolma3_mix-150B-1025 (Olmo et al., 2025). Validation loss is computed on 210M tokens.

E.2.1 Scaling configurations

As in the MLP experiments, models are trained with the respective scaling configuration. We initialize the last layer to zero.

E.2.2 Multiplier tuning

All multipliers are tuned at base width $N = 256$ and act on top of the respective parameterization, so they transfer to larger widths unchanged. We tune one learning-rate multiplier per weight type (input embedding, attention $W^{q,k,v}$, attention output W^o, expert in, expert out, router, unembedding), a single global initialization multiplier applied to the initialization standard deviation of all weights, the attention logit multiplier and the global learning rate η. Normalization gains are not tuned.

Layerwise learning rates. An exhaustive grid search, as described for MLPs in Section E.1.2, was not computationally feasible. Instead, we tune with round-robin coordinate descent at 1B tokens. Each round sweeps one multiplier over the factors $\{2^{-4}, 2^{-3}, \ldots, 2^4\}$ relative to the current best vector with one seed, validates the best factor with three further seeds and promotes it if the 4-seed mean validation loss improves on the current best. After the one-dimensional rounds have converged, we sweep coupled pairs jointly over $\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}^2$ (embedding and unembedding, expert in and expert out, and learning rate and initialization of each weight type). In Regime II, the round-robin tuning ran with soft routing, and a final one-dimensional scan with sparse routing at 2.5B tokens moved the expert-out multiplier from 512 to 8192. We also ran a random search over 9 multipliers (100 vectors drawn from $\{2^{-7}, \ldots, 2^7\}$ per setting, top 5 validated over 4 seeds). It improved over all-one multipliers by 0.025 to 0.054 nats but remained worse than the round-robin vectors in both regimes (by 0.024 to 0.036 nats in Regime III), so we do not use it.

Shared across parameterizations. μP and MSSP use the same layerwise learning-rate multipliers (Table E.1). Only the global initialization multiplier, the attention logit multiplier and η are tuned separately for each parameterization, the first two by one-dimensional scans over $\{2^{-7}, \ldots, 2^{7}\}$ at 2.5B tokens. The multipliers were tuned with the auxiliary loss. After switching to QB load balancing we kept them and only retuned η on the grid $\{10^{-3+2k/9}\}_{k=0}^{9}$. Regime I coincides with μP-Regime-III at base width ($M = 8$, $K = 4$, $N_e = N/2$) and uses the Regime III multipliers.

Table E.1: **Tuned multipliers for Transformer MoEs** at base width $N = 256$. Learning-rate multipliers are shared by μP and MSSP. Regime I uses the Regime III values.

	Regime II		**Regime III**	
	μP	MSSP	μP	MSSP
LR mult. input embedding	64		16	
LR mult. attention $W^{q,k,v}$	2		2	
LR mult. attention W^{o}	2		2	
LR mult. expert in	2		2	
LR mult. expert out	8192		128	
LR mult. router	4		8	
LR mult. unembedding	4		8	
Global init. mult.	1	$1/2$	$1/4$	1
Attention logit mult.	1	1	$1/4$	1
Learning rate η (QB)	$7.74 \cdot 10^{-3}$	$7.74 \cdot 10^{-3}$	$4.64 \cdot 10^{-3}$	$7.74 \cdot 10^{-3}$

E.2.3 Coordinate checks and learning rate sweeps

We use torch-module-monitor to monitor the training dynamics and perform coordinate checks.

E.3 Figure details

Figure 1: *Right:* Learning rate sweeps for GPT MoEs trained with Adam using auxiliary load balancing loss in Regime II and III, as described in Appendix E.2. For Regime II, the figure shows the mean over 4 seeds. Appendix F.1 shows uncertainty bands and the corresponding training loss.

Left: We compare the top-5 training accuracy averaged over the last 50 steps of μP (dashed lines) and MSSP (solid lines) for the soft routing MLP MoE Adam scaling runs from each scaling regime in App. G.2.1 for Regime I, G.2.2 for Regime II and G.2.3 for Regime III. Where both variants are available, we use the 0 last-layer initialization variant for the μP baseline for direct comparability with MSSP. These experiments use the optimal multipliers and learning rate from $N = 128$ and transfer them to large model scales, as is common for μP in dense networks. Both μP and MSSP use the same compute budget for multiplier tuning at small size $N = 128$. In Regime I, μP and MSSP barely differ, so that the lines overlap heavily for both SGD and Adam.

Figure 3: Soft-routing MLP MoEs trained with SGD in Regime II, using zero last-layer initialization in both μP and MSSP. The equivalent figure for Regime III (Figure E.2) also shows delayed learning and initially vanishing MoE outputs in μP but not MSSP. Under maximal stable last-layer initialization $\sigma \asymp 1/N$, exponents are even more unstable and learning even more delayed (Figure G.23). More detailed evaluations are provided in Section G.2.2 for Regime II and Section G.2.3 for Regime III. All width-scaling exponents Expon(v_N) are computed as OLS linear regression in log-log-space, hence fitting $\alpha \in \mathbb{R}$ in a model $v_N = C \cdot N^{\alpha}$ based on all available widths.

The right subplots show, at each step, the width-scaling exponents of the individual terms of Equation (MoE) and the overall updates to the post-aggregation activations $\|\Delta h_t^l\|_{RMS}$, which are the sum of effective and propagating updates, where the M^{-1} aggregation scaling is included in all terms.

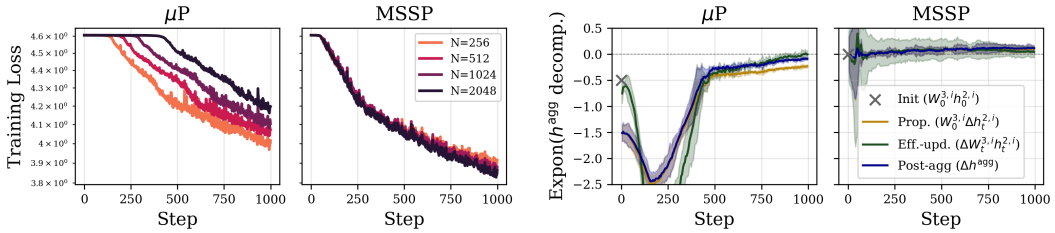


Figure E.2: **Non-monotonic improvement with scale and scale dependent dynamics of μP resolved by MSSP (SGD, R-III).** Training loss (left) is worse at large scale in μP, but monotonically improves in MSSP. Scaling exponents of sub-terms of the aggregated MoE activations h_t^l (right) are approximately 0 in all time steps in MSSP, signaling scale-independent training dynamics. In μP, initially vanishing sub-terms cause cascading time-dependent scale dependence of all terms. Even after 1000 updates, post-aggregation propagating updates still vanish in μP, as predicted.

Figure 4: Width-scaling exponents of the propagating and effective updates of each linear weight matrix for the same setting as in Figure 3. The expert output weights are also evaluated pre-aggregation. Missing lines indicate that the respective term is exactly 0 (such as propagating updates in the input layer). The equivalent figure for Regime III (Figure E.3) also shows initially vanishing propagating and effective updates in all layer types in μP but not MSSP. Sections G.2.2 and G.2.3 show that for Adam, while less severe, analogous width dependence in μP is resolved by MSSP, and monotonic improvement with scale is recovered.

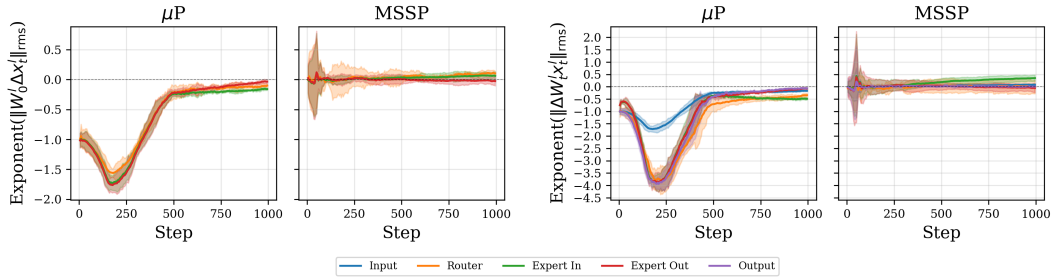


Figure E.3: **Consistent exponents in MSSP, but not μP (SGD, R-III).** Corresponds to Figure E.2. Initially vanishing post-aggregation terms in μP induce width-dependence to cascade into all layers. Over time, exponents partially self-correct in μP and almost recover width-independent propagating updates (left) as well as effective updates (right) in all layers. By allowing benign divergence of pre-aggregation propagating updates (red dashed line), MSSP recovers approximately width-independent effective updates at all time steps.

Figure 5: All four panels use Regime-III Transformer MoEs with an MoE layer in every block, $M = N/32$ experts, $K = M/2$ active experts and expert width $N_e = N/2$, as described in Appendix E.2, and $L = 8$ layers unless stated otherwise. Panels 1 and 3 use the learning rate multipliers from Section E.2.2, panels 2 and 4 set all base multipliers to 1.

1. Learning rate transfer under QB load balancing. Adam, $N \in \{256, 512, 1024, 2048\}$, 2.5B tokens (4768 steps of 512 sequences of 1024 tokens), 1400 warmup steps followed by cosine decay, one seed. We report the final validation loss on 210M held-out tokens. Solid lines use QB expert biases. Stars mark the best learning rate $7.74 \cdot 10^{-3}$ at every width. The dotted line of the same color marks the best validation loss of an otherwise identical sweep with the auxiliary load balancing loss (weight 0.01). QB matches the auxiliary loss at $N = 256$ and improves on it by 0.02, 0.07 and 0.11 at $N = 512$, 1024 and 2048.

2. Depth transfer. Adam, starting at Regime III base size $N = 256$, $M = 8$, $K = 4$ and scaling $L \in \{8, 16, 32, 64, 128\}$, with residual branches scaled by L^{-1} (Table B.1), auxiliary load balancing loss, 2.5B tokens, 1400 warmup steps, one seed. The best learning rate is $1.23 \cdot 10^{-2}$ at every depth.

3. Expert diversity. Mean pairwise cosine distance between the outputs of all experts of a layer, where every expert is evaluated on the same held-out tokens without routing, averaged over the MoE layers. We compare μP-Regime-III, which initializes the experts independently, with MSSP-Regime-III,

which ties them at initialization, so that their cosine distance starts at 0 in MSSP. Both use QB load balancing and learning rate $4.64 \cdot 10^{-3}$ for 2.5B tokens.

4. Refined coordinate check for Muon. MSSP with Muon as in Table B.2, learning rate $10^{-13/9} \approx 3.6 \cdot 10^{-2}$, decoupled weight decay $1.29 \cdot 10^{-3}$, QB load balancing, 100 steps, with 4 seeds each. At every step, we fit the width exponent of each term of the h^{agg} decomposition (Equation (MoE)) in the first MoE layer by OLS in log-log space over all widths and seeds, and the bands show ± 2 standard errors of the fitted slope. The effective term includes the cross term. All Muon coordinate checks are in Figures F.20 and F.21.

F Additional Transformer MoE Experiments

F.1 Learning rate sweeps and performance evaluations for Transformer MoEs

For a preliminary evaluation, which regime and parameterization scales most favorably in practice, Figure F.1 shows the scaling ladder of 8-layer all-layer MoEs after 10B tokens (19073 steps, 1400 warmup steps) against the number of active non-embedding parameters, which counts the attention weights, the router and the K active experts of all 8 blocks. Layerwise multipliers are tuned as discussed in Section E.2.2. As expected, all MoEs outperform their dense equivalent. While different regimes perform similarly well at the same number of active non-embedding parameters, all-scaling R-III is more scalable than fine-grained R-II as the expert communication overhead quickly grows infeasible in R-II. MSSP increasingly outperforms μP with increasing model scale in R-III. Every model uses the learning rate tuned at base width for 2.5B tokens, halved, which we found to be optimal at 10B tokens, and one seed.

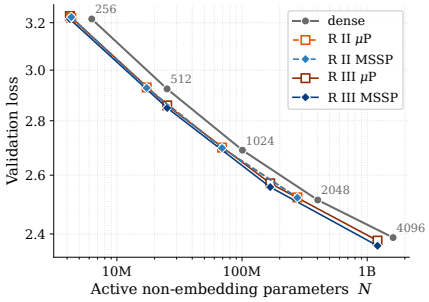


Figure F.1: Training on 10B tokens in μP versus MSSP under QB load balancing.

All learning rate sweeps below are run for 2.5B tokens.

Figure F.2 shows that the optimal learning rate transfers from small to large widths in both μP and MSSP in Regimes II and III, except for a small saturating drift in μP-Regime-III. Sensitivity to suboptimal learning rates is low overall, but increases with scale.

Figure F.3 shows that learning rate transfer also holds under auxloss in alternate layer MoEs instead of QB expert biases in all-layer MoEs. μP and MSSP perform similarly well with cleaner transfer in MSSP.

Figure F.4 shows clean learning rate transfer of both μP and MSSP in Regime I. As expected, μP and MSSP do not differ significantly here.

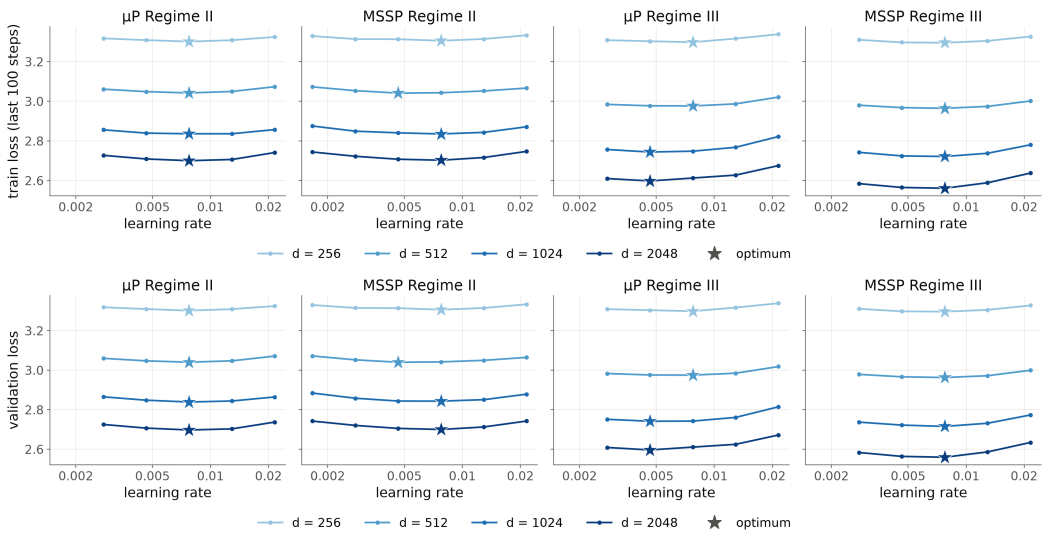


Figure F.2: **Learning rate sweeps for μP and MSSP (Adam, all-layer MoE, QB expert biases).** Training loss (top) and validation loss (bottom) of an all-layer MoE with QB expert biases.

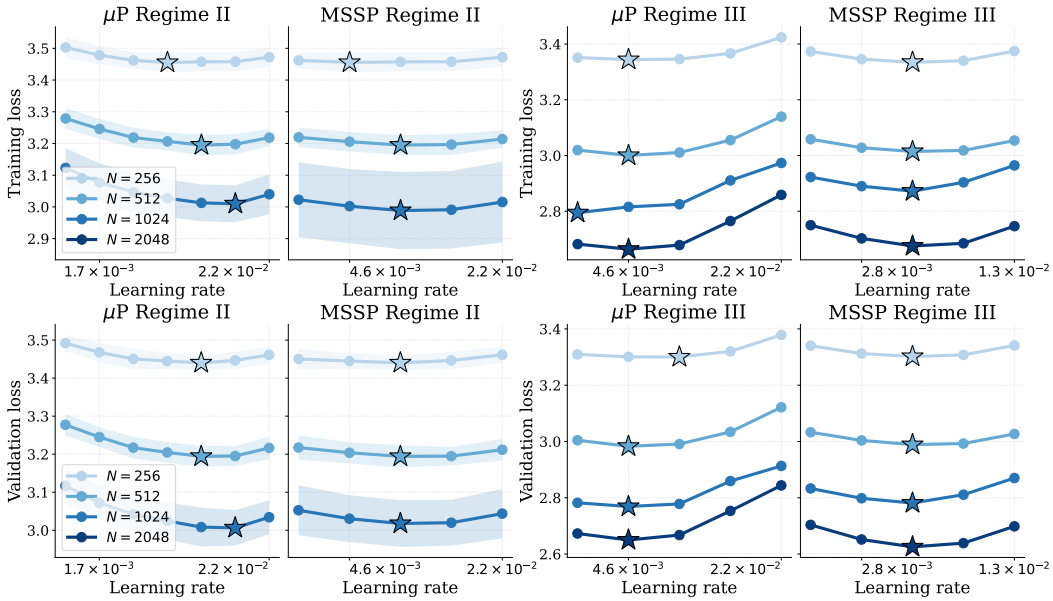


Figure F.3: **Learning rate sweeps for μP and MSSP (Adam, aux loss, alternate-layer MoE).** Training loss (top) and validation loss (bottom) of an alternate-layer MoE with aux loss instead of QB expert biases, and with σ-confidence bands across 4 seeds for fine-grained Regime II.

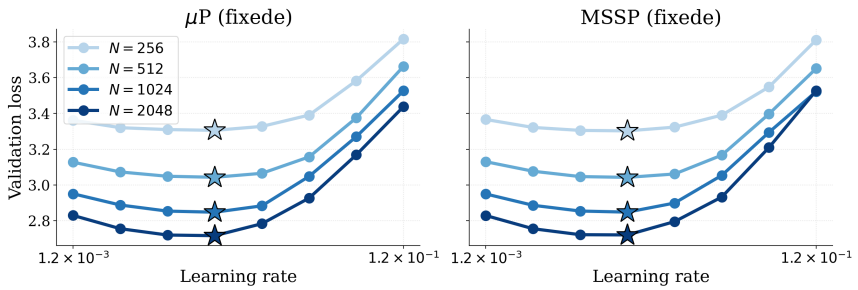


Figure F.4: **LR Transfer in Regime I.** Same as Figure F.3, but for Regime I. As expected, we get LR transfer and μP and MSSP do not significantly differ.

As expected standard scaling does not transfer the optimal learning rate. Figure F.5 shows that it is even outperformed by MSSP at all scales, and the difference is increasing with scale.

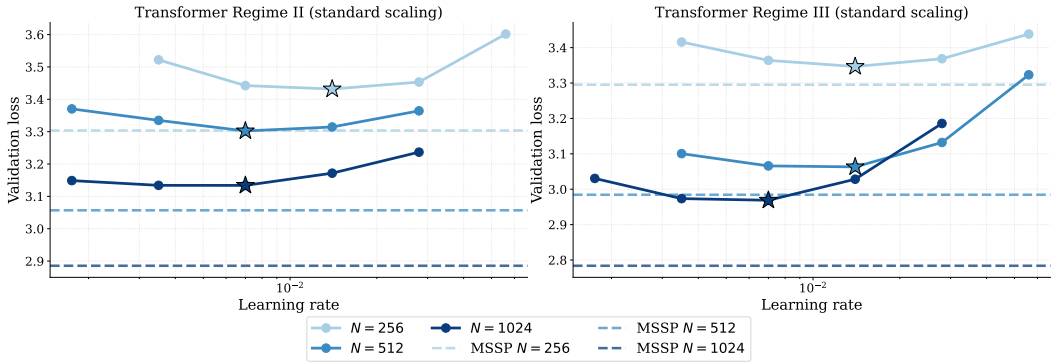


Figure F.5: **Standard scaling is outperformed and does not transfer.** A learning rate sweep for standard scaling without layerwise multiplier tuning compared to the equivalent auxloss MSSP performance at 2.5B token budget.

Figure F.6 is a sanity check that increasing the total number of parameters while keeping active parameters fixed by increasing number of experts M monotonically improves performance in MSSP-Regime-III over the dense baseline.

The optimal learning rate also transfers cleanly under pure depth scaling in Transformer MoEs. Surprisingly, the $1/L$ residual multiplier does not significantly affect the results. This likely stems from the approximate scale-invariance induced by the final RMSNorm layer as long as the residual stream RMS norm is significantly larger than RMSNorm's ϵ. Note however that increasing residual stream norm under naive scaling decreases the gradient passed backward from the final RMSNorm layer, which will cause vanishing updates at sufficient depth. Observe that without residual scaling, the $L = 128$ already crosses the smaller depths at larger than optimal LR.

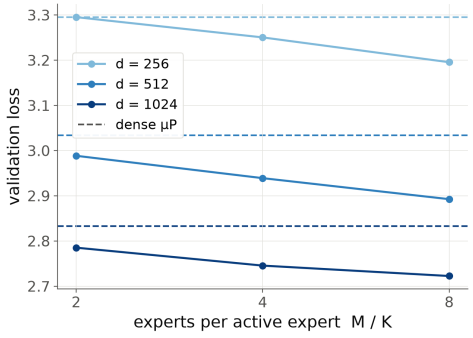


Figure F.6: **Sparsity improves performance.** Increasing total number of parameters while keeping active parameters fixed by increasing number of experts M in the auxloss MSSP-Regime-III setup improves performance as expected.

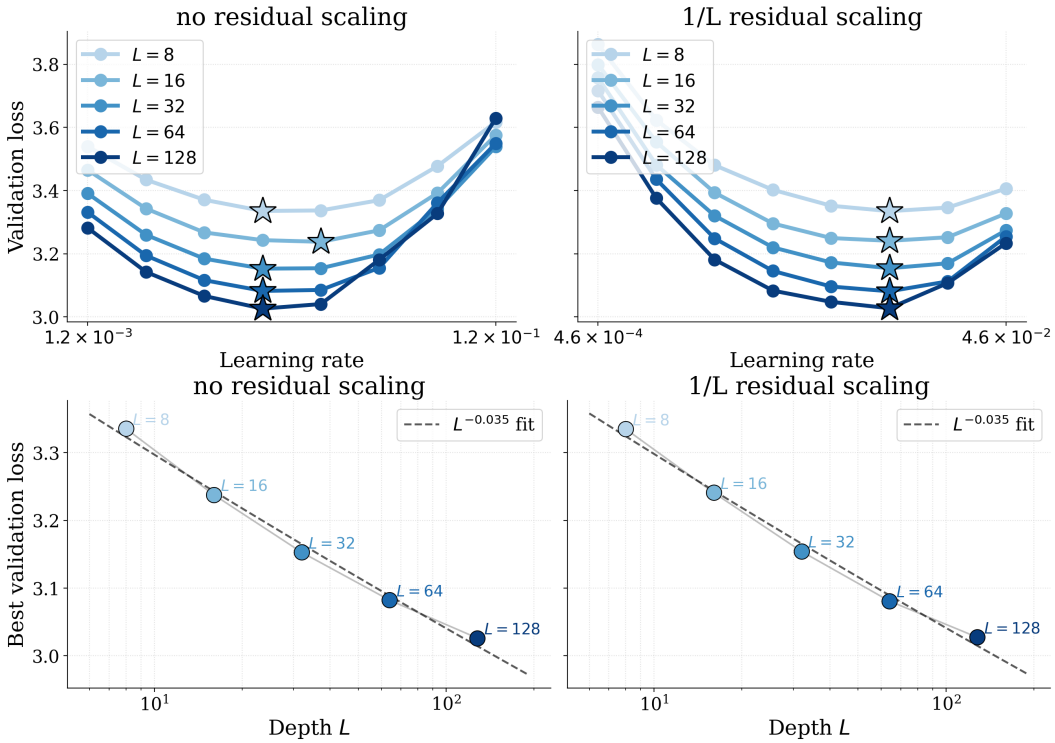


Figure F.7: **LR Transfer across depths.** Regime I and III base model, pure depth scaling. Naive versus CompleteP residual scaling barely differ at moderate depth due to approximate scale invariance induced by final RMSNorm. Loss has depth exponent -0.035 at fixed token budget.

F.2 Expert diversity despite weight sharing

By definition of MSSP in Regime III, the cosine distance between experts is zero at initialization, which raises the question whether this harms specialization. But due to top-k selection, different experts still train on different subsets of the data. *Remarkably, after only 600 steps, experts are more diverse in MSSP than under independent initialization in μP (Figure F.8).* The larger the model, the more MSSP eventually outperforms μP in terms of expert diversity. Later in training, the difference tends to vanish.

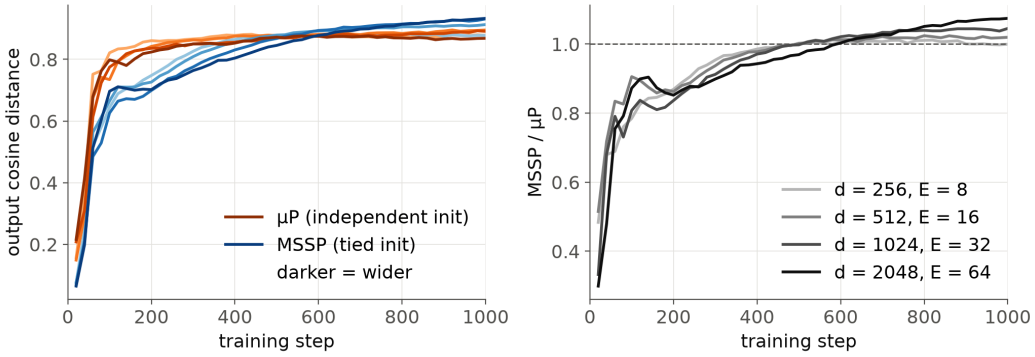


Figure F.8: **Experts are more diverse in MSSP than μP after only 600 steps (Regime III).** Cosine distance between expert outputs on the same fixed validation batch over the course of training on 2.5B tokens using QB load balancing.

F.3 Layerwise gradient norms crossing Adam ε

Gradient entries scale with the predicted exponents. The expert output gradient entries therefore cross a typical constant Adam $\varepsilon = 10^{-8}$ already at moderate width 4500, leading to vanishing expert updates beyond that scale. Gradients also shrink over the course of training, so that this issue likely arises at even smaller scale. In MSSP, the layerwise gradient RMS norms and Adam ε remain approximately equidistant in log space across model sizes, preserving approximately scale-invariant update dynamics. For numerical stability, rather scale up the gradients or vanishing moments m and v instead of scaling down Adam ε.

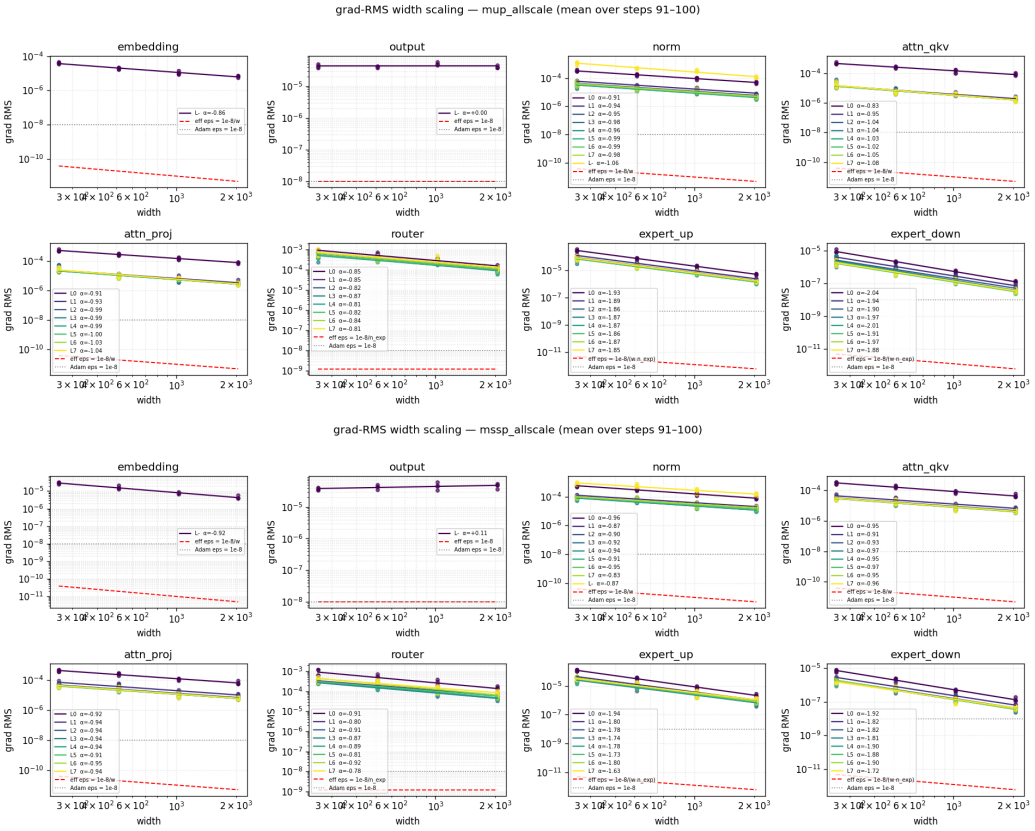


Figure F.9: **Layerwise gradient RMS norms cross global Adam ε.** μP with global Adam ε (top) and MSSP with layerwise Adam ε scaling (bottom) using QB load balancing in Regime III. The dotted gray line denotes the fixed global Adam $\epsilon = 10^{-8}$. The dashed red line denotes our layerwise, scale dependent ϵ scaling. Different colors denote layers 1 to 8. Under our ϵ scaling, the relative scale of ϵ to the tensor-wise gradient RMS norms is approximately preserved across model scales. Expert gradient entries decay as fast as predicted $\propto N^{-2}$ and cross constant Adam $\varepsilon = 10^{-8}$ at moderate width 4500, inducing vanishing expert output layer updates at larger scale.

F.4 Corrected load balance scaling

The load balancing approaches and heuristic scaling derivation is provided in Appendix I.

Figure F.11 and Figure F.10 shows router logit drift towards increasingly negative logit distributions in large MoEs under standard auxiliary loss scaling following (Fedus et al., 2022), both pre and post sigmoid. A corrected pre-factor of $M^{-1}L^{-1}$ recovers a scale-invariant auxiliary loss which results in a scale-invariant router logit distribution, both pre and post sigmoid. Still, the auxiliary loss term is reduced by driving the router logits more negative which results in vanishing post-sigmoid routing weights over time. Quantile balanced expert biases quickly achieve scale-invariant load balancing and stable post-sigmoid routing weights, without needing to tune any hyperparameters.

QB expert biases also remain scale stable without logit drift over time in μP (Figure F.12) and in Regime III (**??**). In Regime I, neither K nor M increase, so that load balancing is not a scaling issue (Figure F.13). Still, across all regimes, the router logit distribution drifts downward with time under auxloss but remains stable under QB expert biases (Figure F.10).

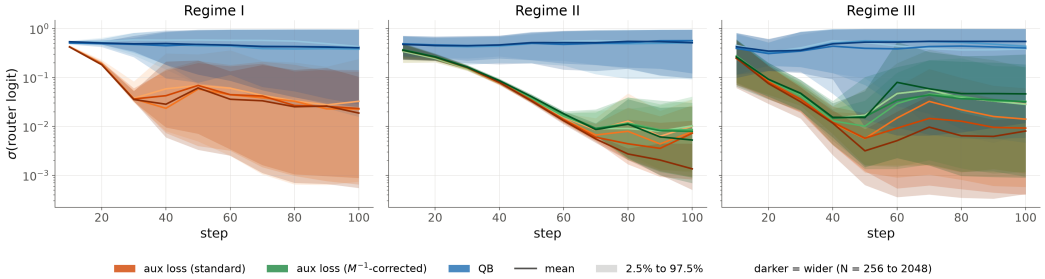


Figure F.10: **Stable router logits under QB expert biases.** Mean post-sigmoid router logits with 2.5% to 97.5% bands for MSSP across regimes (left to right). Post-sigmoid routing weights decay over time under aux loss, and decay with model scale under standard aux loss scaling. QB expert biases (blue) preserve stable routing weights over time and model scales.

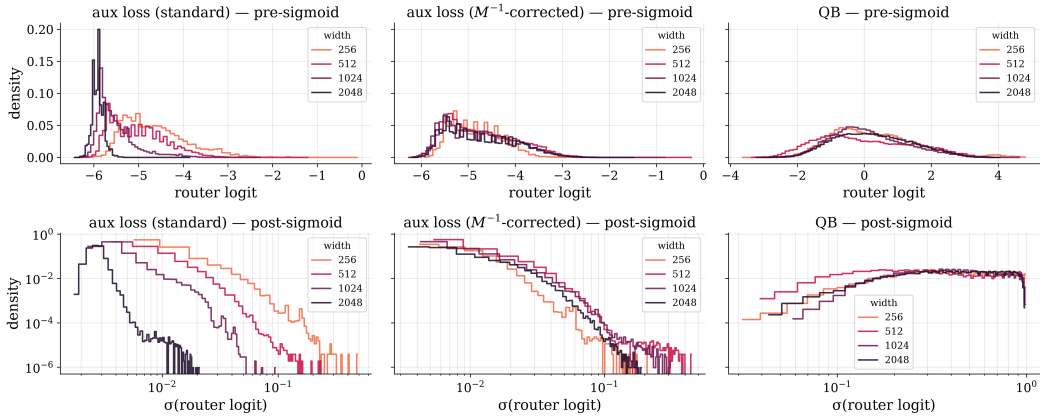


Figure F.11: **Quantile balanced expert biases preserve non-vanishing routing weights. (MSSP, Regime II)** First router in 8-layer GPT MoE after 80 steps under MSSP in Regime II.

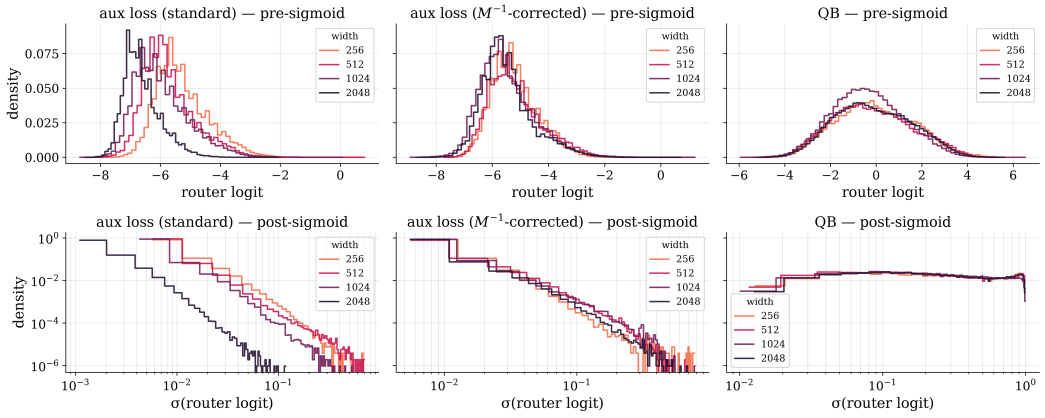


Figure F.12: **Stable router logits under QB expert biases (μP, Regime II).** Same as Figure F.11 but for μP in Regime II. Quantile balanced expert biases remain scale stable without logit drift over time.

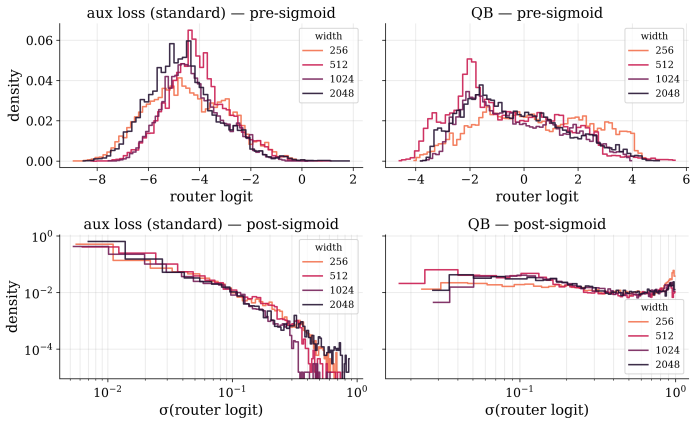


Figure F.13: **Stable router logits under QB expert biases (Regime I).** Same as Figure F.11 but for MSSP in Regime I. There is only one scale stable aux loss variant as K and M both stay fixed. The router logit distribution still drifts downward with time under auxloss but remains stable under QB expert biases.

F.5 Refined coordinate checks for Transformer MoEs

We use `torch-module-monitor` to measure refined coordinate checks as a sanity check for correct implementation of the MSSP scaling rules in each scaling regime. Propagating update exponents 0 verify correct initialization variance of the respective layer, given correctly scaled inputs. Effective update exponents 0 verify correct learning rate scaling of the respective layer, given correctly scaled inputs. One could isolate the effect of the layer's learning rate scaling by normalizing the input activations $x_{\text{in}}/\|x_{\text{in}}\|_{RMS}$. We instead verify that $\|x_{\text{in}}\|_{RMS} = \Theta(1)$ approximately holds in all layers (not shown).

As a unit test for correctly implemented μP and MSSP, Figure F.14 shows clean effective and propagating updates across layer types over the first 100 steps. In μP, observe vanishing propagating updates in the router.

Figure F.15 validates that propagating terms vanish after the expert aggregation under μP in Regimes II and III. MSSP preserves approximately scale invariant post aggregation dynamics in all subterms of the expert aggregation.

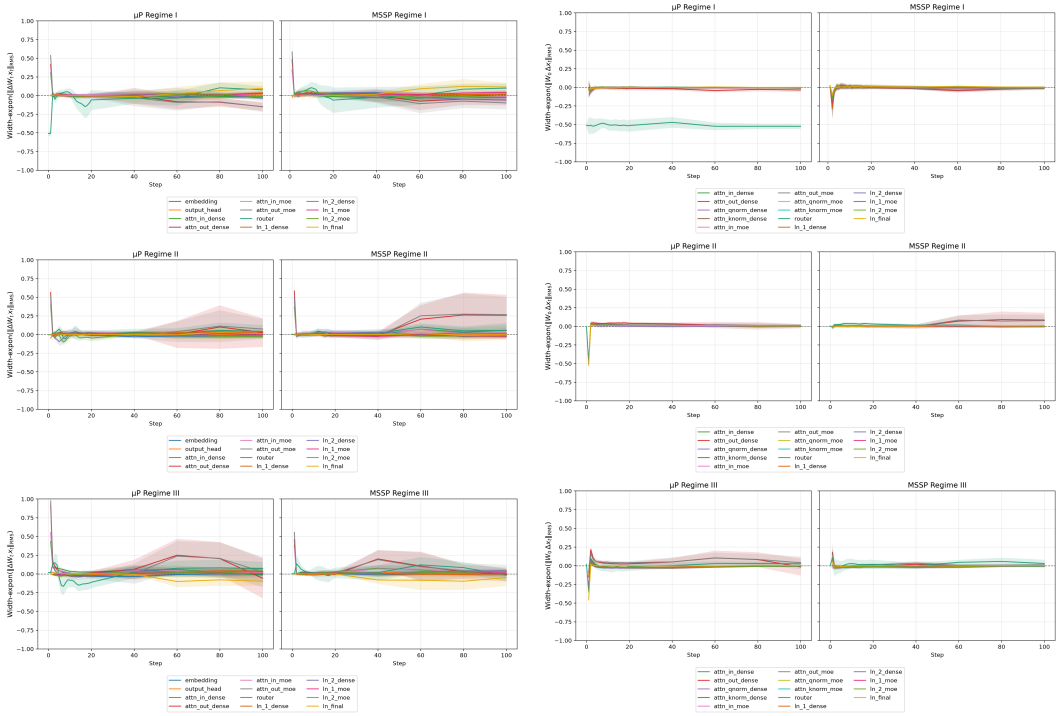


Figure F.14: **Refined coordinate checks under QB load balancing.** Scaling exponents of effective updates (left) and propagating updates (right) across scaling regimes (rows). The refined coordinate check acts as a unit test for scale stable training dynamics. As expected, all plotted exponents are close to 0, except router propagating updates in μP-Regime-I is at -0.5.

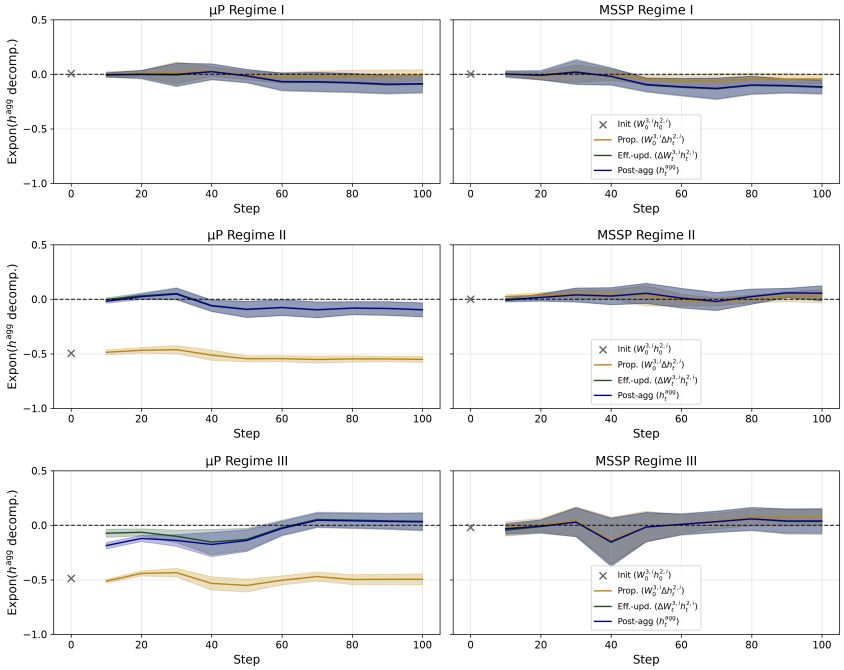


Figure F.15: **Expert aggregation decomposition.** Forward expert aggregation decomposition (MoE) in the first of 8 MoE blocks. Initialization and propagating updates are clearly too small in μP Regimes II and III. μP and MSSP barely differ in Regime I.

Figure F.16 to Figure F.19 below verify scale stable scaling in all layer types for the aux loss setup.

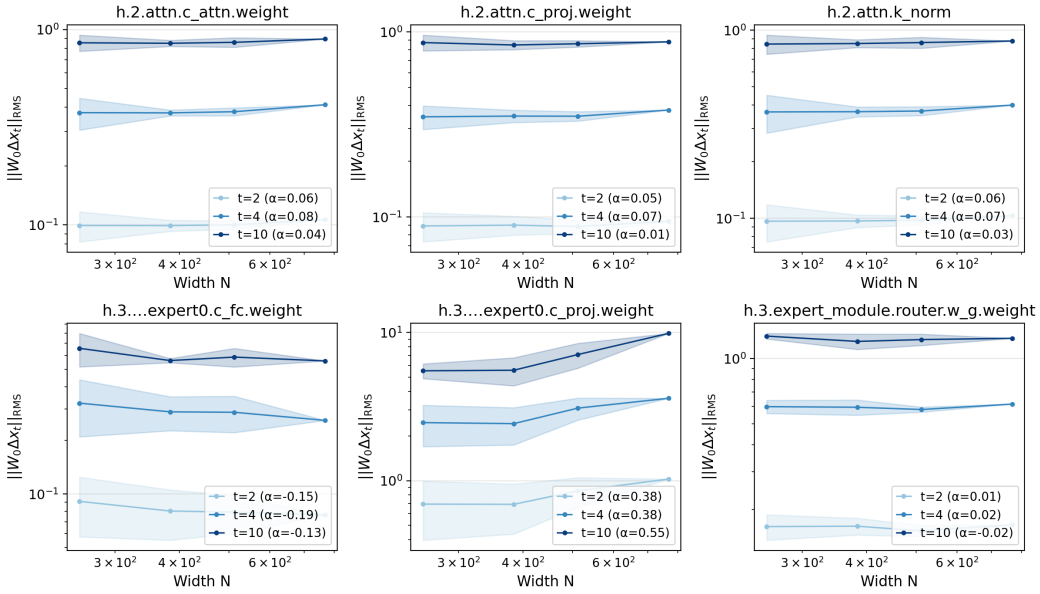


Figure F.16: **Propagating updates in MSSP (Adam, Regime II).** Propagating updates of some example layers approximately approximately follow the desired scaling exponents. First and second expert layer stem from a single expert. Measuring such small experts is particularly noisy, depending heavily on how many tokens are routed to them in the respective step. Recall that the desired propagating update exponent in the expert output layer `expert0.c_proj` is 0.5 in Regime II.

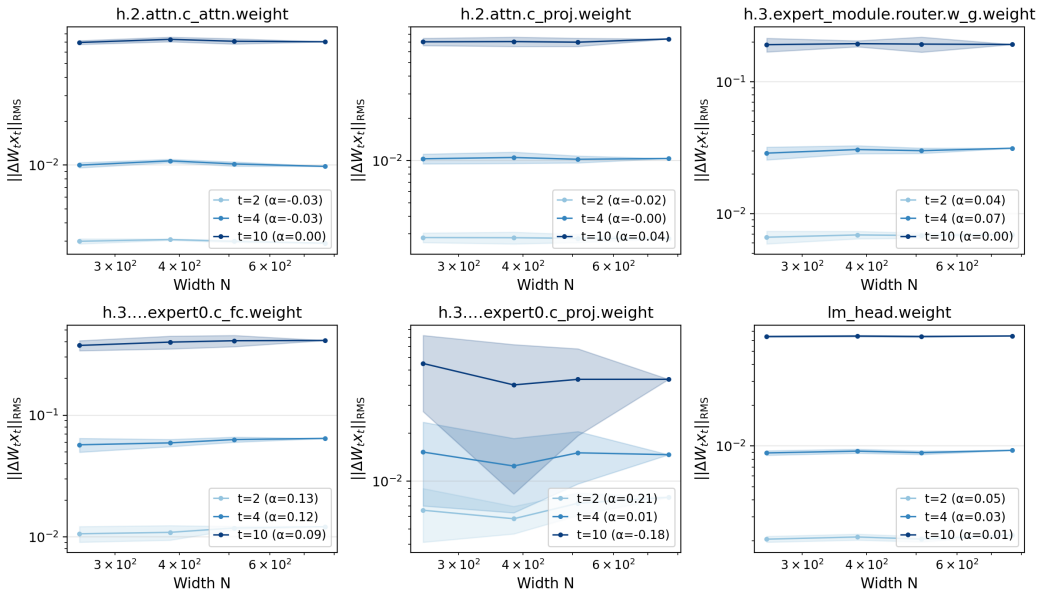


Figure F.17: **Effective updates in MSSP (Adam, Regime II).** Effective updates of some example layers are approximately scale-preserving at all times. First and second expert layer stem from a single expert. Measuring such small experts is particularly noisy, depending heavily on how many tokens are routed to them in the respective step.

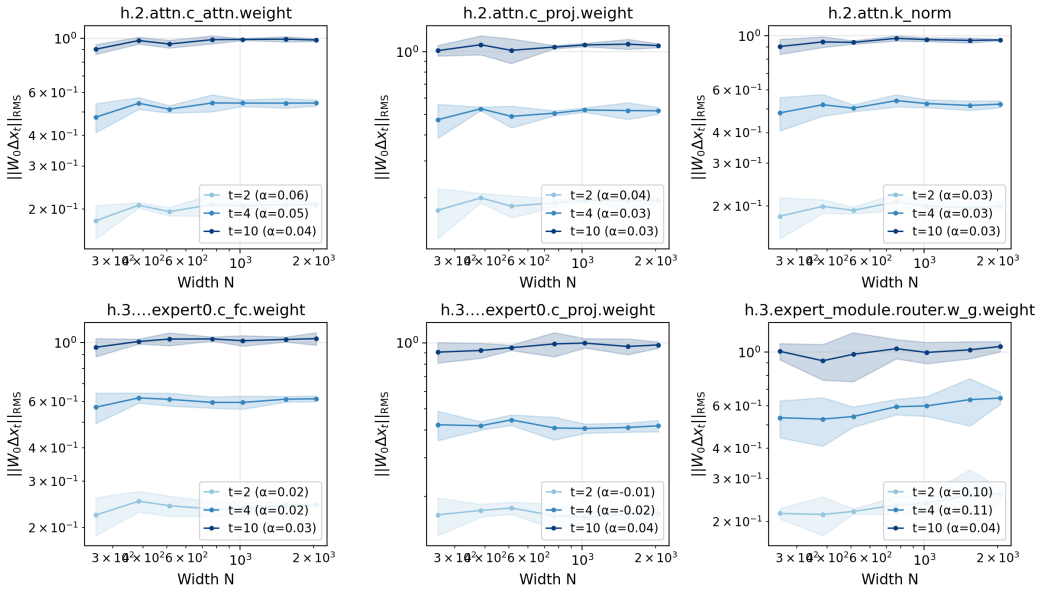


Figure F.18: **Propagating updates in MSSP (Adam, Regime III).** Propagating updates of some example layers are approximately scale-preserving at all times. First and second expert layer stem from a single expert.

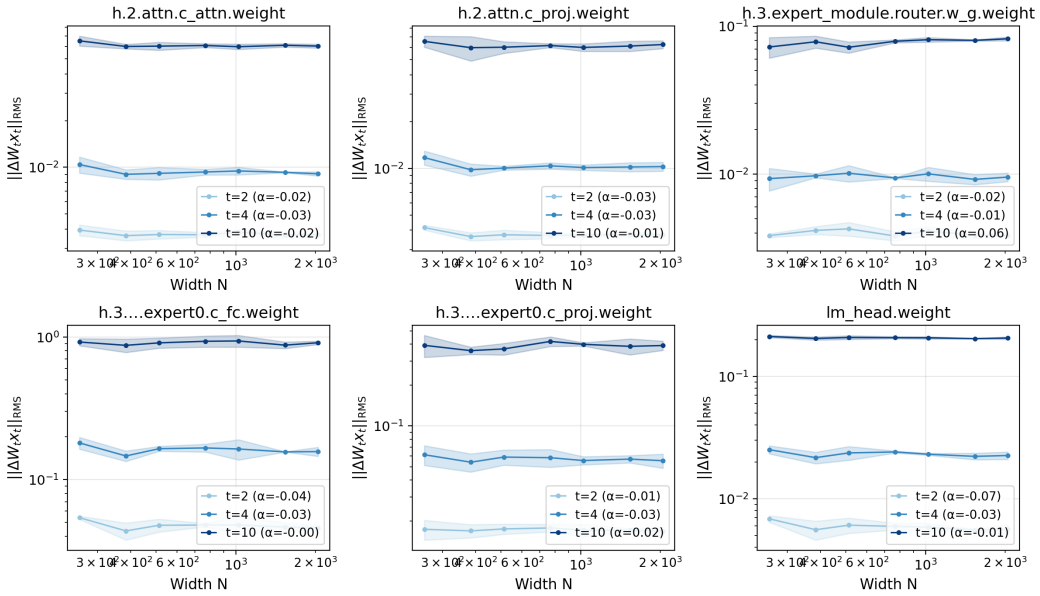


Figure F.19: **Effective updates in MSSP (Adam, Regime III).** Effective updates of some example layers are approximately scale-preserving at all times. First and second expert layer stem from a single expert.

F.6 MSSP for Muon

Figures F.20 and F.21 repeat the refined coordinate checks of Appendix F.5 with Muon in Regimes I, II and III and on the extremely sparse path with $K = 4$ fixed while N, N_e and M grow (router initialized to zero, experts tied, QB load balancing). We tune learning rate at base width, and fix all layerwise multipliers at 1. Regimes I and III use $\eta = 10^{-13/9} \approx 3.6 \cdot 10^{-2}$ and Regime II uses $\eta = 10^{-15/9} \approx 2.2 \cdot 10^{-2}$. Weight decay is decoupled and independent of the absolute learning

rate, $\lambda = 1.29 \cdot 10^{-3}$ per step at peak learning rate. After a transient in the first step, the width exponents of the effective and propagating updates of all monitored layers and of all terms of the h^{agg} decomposition stay within ± 0.17 of 0 for $N = 256$ to 2048 in all four settings, empirically suggesting that the prescription of Table B.2 is maximally scale stable for Muon as well.

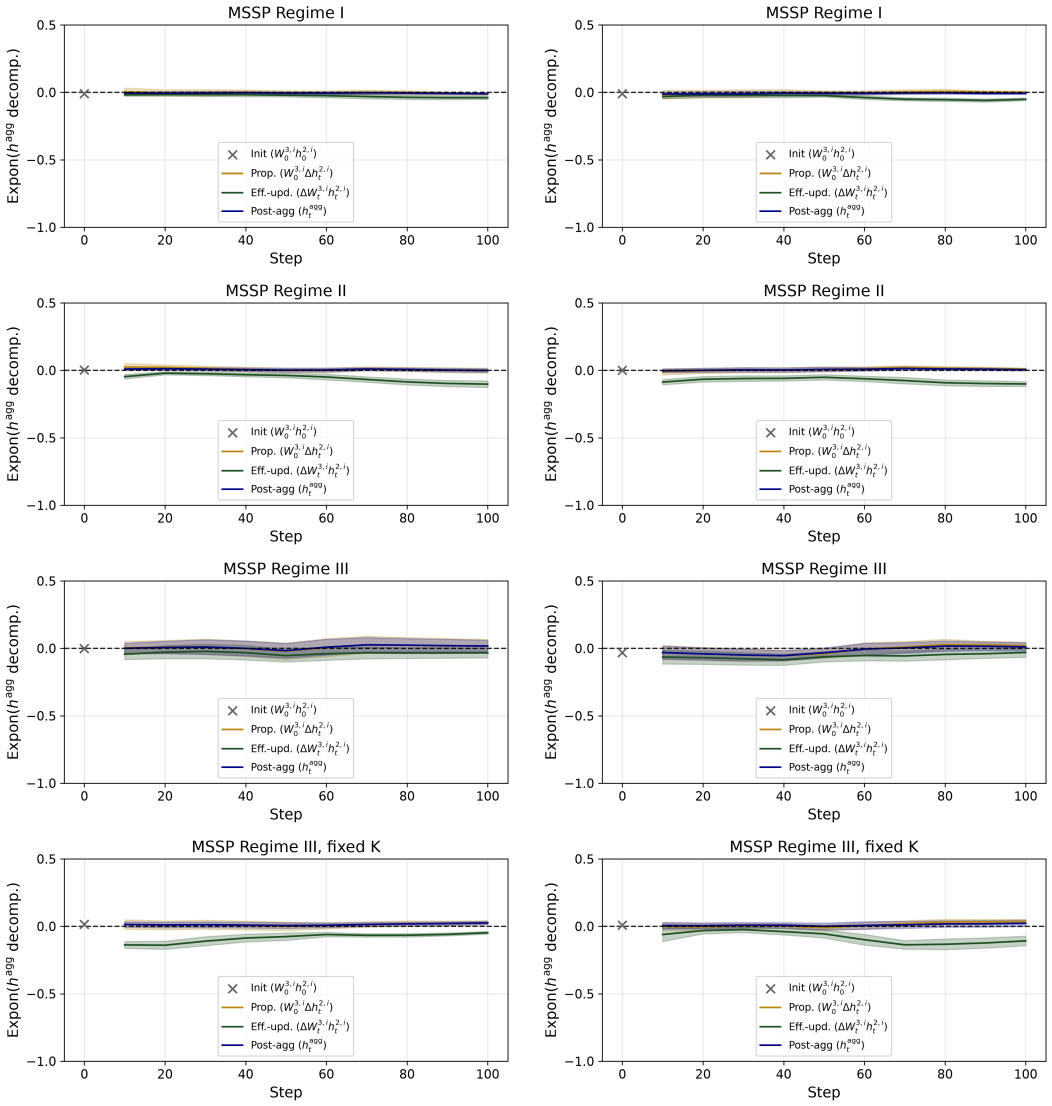


Figure F.20: h^{agg} **decomposition for Muon with MSSP.** Width exponents of the initial, propagating and effective-update terms of the expert output and of the aggregated output h^{agg} in the first (left) and last (right) MoE layer, in Regime I, Regime II, Regime III and Regime III with fixed $K = 4$ (top to bottom). $N = 256$ to 2048, 4 seeds, 100 steps.

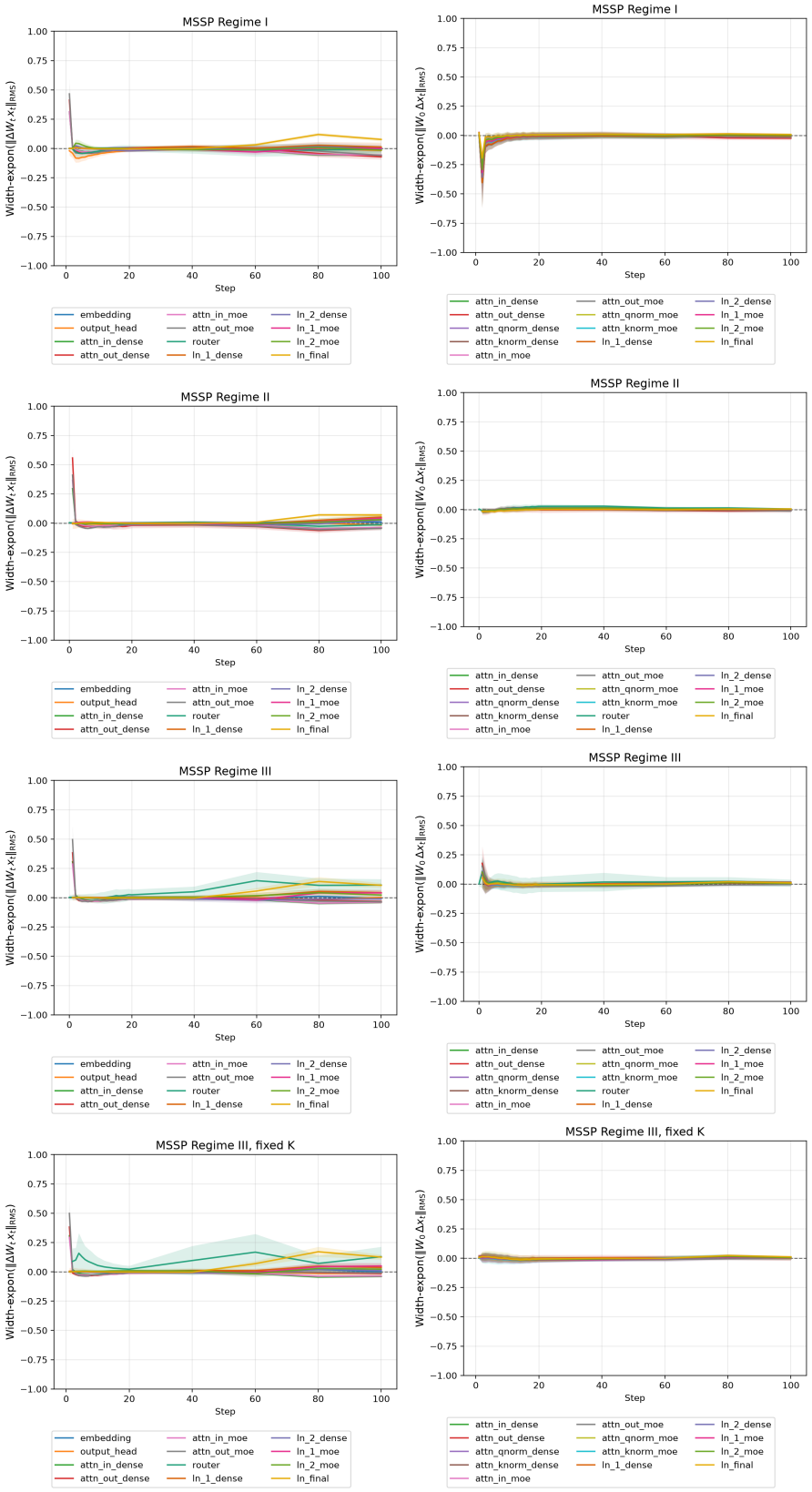


Figure F.21: **Refined coordinate checks for Muon with MSSP.** Width exponents of the effective updates (left) and the propagating updates (right) of all monitored layers in Regime I, Regime II, Regime III and Regime III with fixed $K = 4$ (top to bottom). $N = 256$ to 2048, 4 seeds, 100 steps. Exponents stay near 0 after the first step.

G Additional MLP MoE Experiments

G.1 Learning rate sweeps for MLP MoEs

Generally observe cleaner learning rate transfer across model sizes in MSSP than in μP (Figure G.1). Also observe more robust monotonic improvement with scale in MSSP across optimizers and scaling regimes.

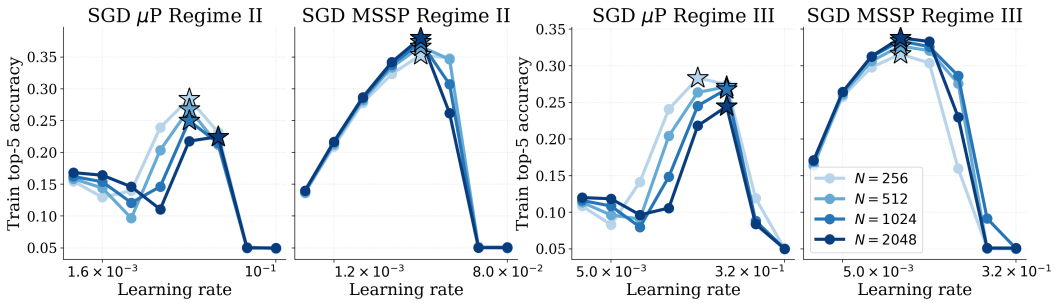


Figure G.1: **Robust LR transfer in MSSP, but not μP.** Top-5 training accuracy of MLP MoEs trained with SGD on TinyImageNet. The optimal learning rate often grows in μP saturating at the maximal stable learning rate with degrading performance at large scale. MSSP recovers learning rate transfer and monotonic improvements with increasing scale.

G.1.1 Regime I: Fixed number of experts

While the optimal learning rate transfers and performance monotonically improves with scale, the performance saturates at large scales. The differing router initialization has negligible impact.

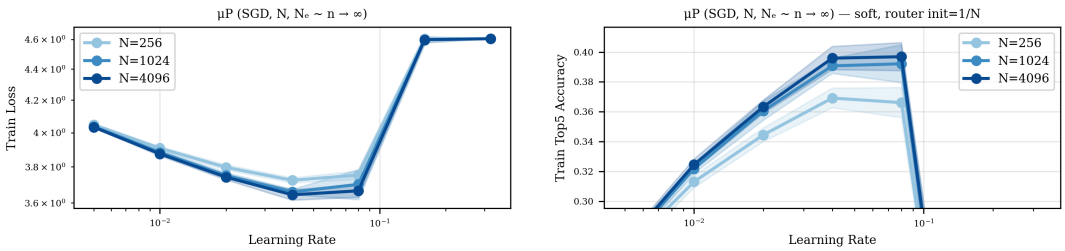


Figure G.2: **LR sweep, μP with $1/N$ router init (SGD, soft, Regime I).**

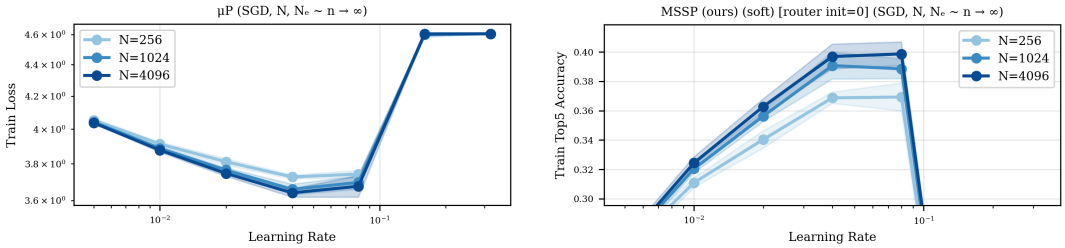


Figure G.3: **LR sweep, MSSP with zero router init (SGD, soft, Regime I).**

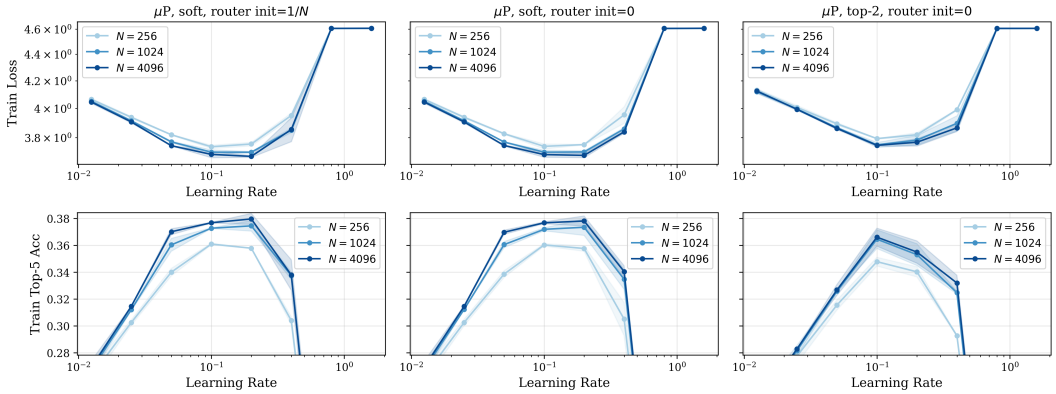


Figure G.4: **LR sweep comparison (Adam, Regime I).**

G.1.2 Regime II: Fixed expert width

Observe greatly improved learning rate transfer of MSSP over μP in both SGD and Adam in Regime II. Larger performance gains from small to large model width result in improved absolute performance at scale across optimizers.

The optimal learning rate in μP tends to grow toward the instability threshold since subcomponents of the expert aggregation dynamics are vanishing at fixed learning rate. Since the correctly scaled subcomponents of the expert aggregation would diverge when increasing the learning rate, the stability threshold does not grow with width. These conflicting objectives of stability of some terms versus effective learning in others results in worse performance at scale. By balancing all subcomponents, MSSP recovers width independence of both the maximal stable and most effective learning rate scaling for all terms in the training dynamics, so that monotonic improvement with scale is preserved. Slight saturating growth of the optimal learning rate can occur in MSSP. Analogous slight shifts have often been observed in dense architectures in μP (Yang et al., 2022; Blake et al., 2025; Ghosh et al., 2025).

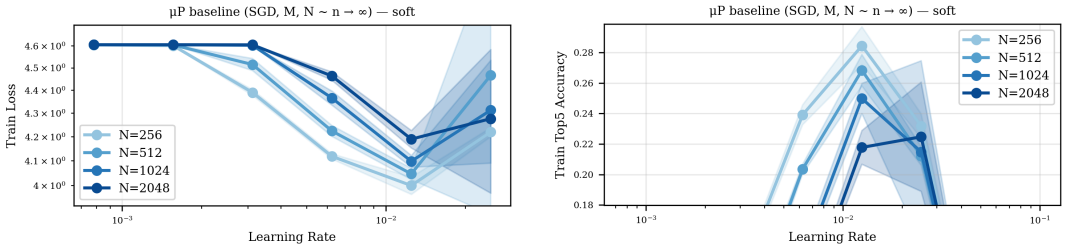


Figure G.5: **LR sweep, μP baseline (SGD, soft, Regime II).** Ending lines denote divergence of the training run. The optimal learning rate drifts toward the maximal stable threshold, which does not increase with width. Observe delayed learning in the corresponding coordinate checks, which results in performance getting worse with model scale.

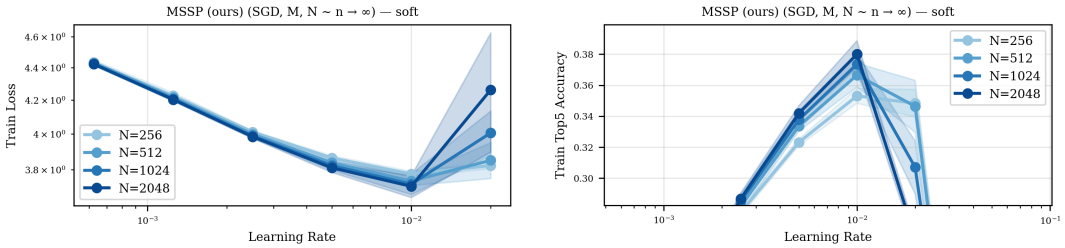


Figure G.6: **LR sweep, MSSP (SGD, soft, Regime II).**

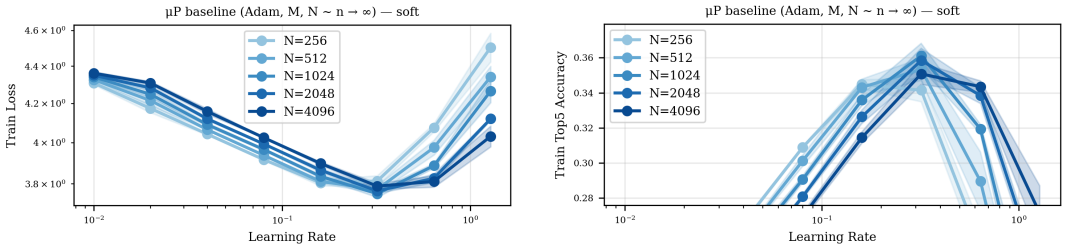


Figure G.7: **LR sweep, μP baseline (Adam, soft, Regime II).**

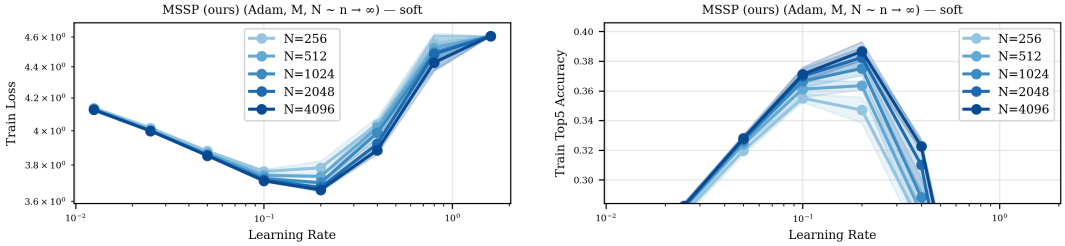


Figure G.8: **LR sweep, MSSP (Adam, soft, Regime II).**

G.1.3 Regime III: Joint proportional scaling

Similar to Regime II, the optimal learning rate in μP saturates at the maximal stable learning rate. The optimal learning rate in MSSP approximately transfers from small to large scale, and achieves better performance than μP at large scale by consistently recovering monotonic improvement with scale.

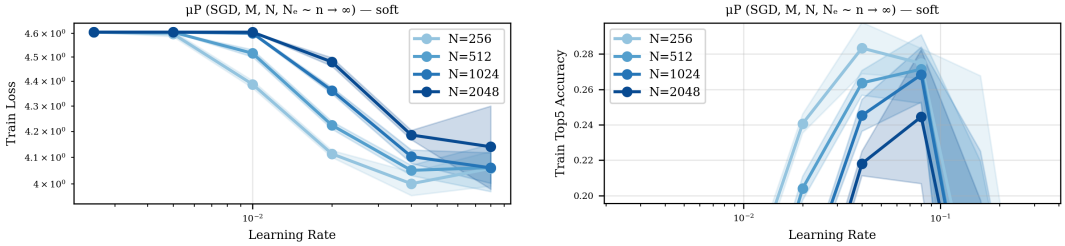


Figure G.9: **LR sweep, μP without shared experts (SGD, soft, Regime III).**

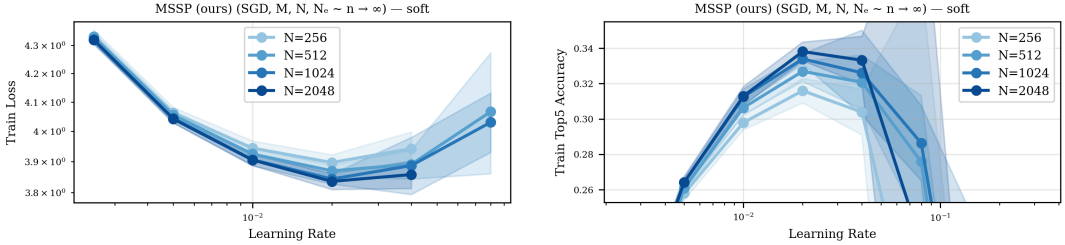


Figure G.10: **LR sweep, MSSP with shared experts (SGD, soft, Regime III).**

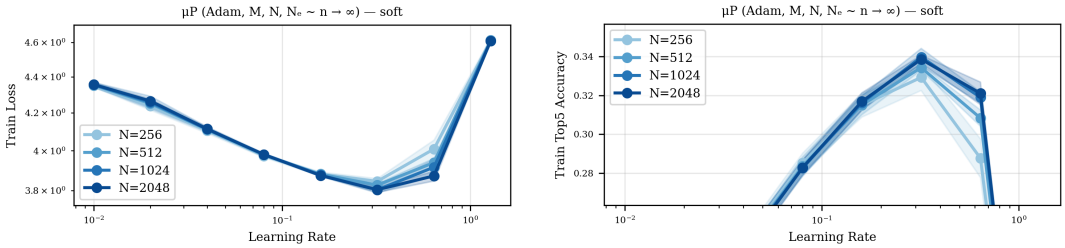


Figure G.11: **LR sweep, μP without shared experts (Adam, soft, Regime III).**

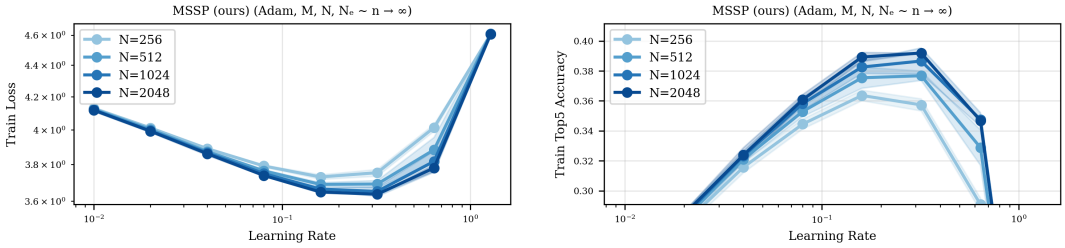


Figure G.12: **LR sweep, MSSP with shared experts (Adam, soft, Regime III).**

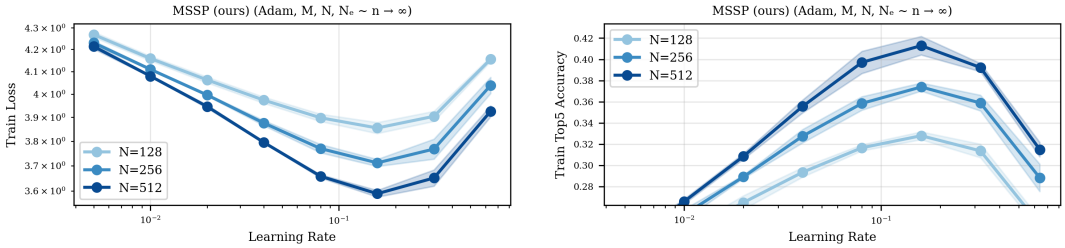


Figure G.13: **LR sweep, MSSP with shared experts (Adam, top-k, Regime III).**

G.2 Fine-grained scaling evaluations in MLP MoEs

Here we show standardized evaluation figures for training the optimal configuration of many combination of optimizer, parameterization, routing mechanism and scaling regime. Across the board, we observe more desirable scaling properties in MSSP than in μP.

The first row shows training loss, training accuracy, the cumulative feature learning of the entire MoE $\|\Delta h_t^L\|_{RMS}$, as well as the routing logit norm $\|\psi_t\|_{RMS}$ before the sigmoid. These allow to verify delayed learning versus monotonic improvement with scale, as well as the scale dependence in feature and router learning.

The second and third rows show refined coordinate check exponents as a function of width N. The general desired exponent for propagating and effective updates is zero, signaling scale independence of the respective component of the training dynamics (with the exception of expert output propagating update exponent 0.5 in MSSP in Regime II). The expected layerwise gradient scaling exponents vary between -2 and 0, depending on the scaling regime. Dashed horizontal lines indicate where expected exponents are non-zero. Decomposition exponents (third row first and second subplots) should generally be balanced, otherwise subcomponents of the dynamics strictly dominate others, which causes strong finite-width effects. Exponents should ideally also be independent of the step, otherwise non-trivial initial dynamical effects such as delayed learning are to be expected.

Missing lines in the propagating update and decomposition plots denote 0 values due to zero initialization of the relevant layer. For example under zero last-layer initialization, the initial gradient is zero except in the output layer. This helps to reduce width dependence from vanishing initial terms.

We show both raw effective updates (second row, center) and a variant with normalized incoming activation norm (second row, right) to distinguish width dependence from the combination of the current and previous layers versus isolated width dependence in the current layer.

In all regimes, the predicted exponents hold surprisingly well across training, suggesting that our theory is predictive of practical training dynamics far beyond the first few iterations. The router gradient exponents are most noisy, but the closest clean exponent of all layers still follows our prediction in MSSP remarkably well throughout training.

In μP, width dependence in individual subterms of the training dynamics cascades into the entire dynamics such that exponents become much more width and time dependent across layers, and often converge to intermediate values between the clean exponents $\{-0.5, 0, 0.5\}$.

Extensive layerwise multiplier tuning for each scaling config, optimizer and routing type is paramount for achieving stable training at all.

G.2.1 Regime I: Fixed number of experts

Under maximal stable router initialization $\sigma = 1/N$, the propagating updates in the router are too small. Setting the router initialization to 0 removes this source of width dependence and results in cleaner scaling exponents. The effect of this intervention on the final performance is negligible.

The top-k selection mechanism with $k \asymp M$ does not change any expected scaling exponents in this paper. Indeed, Figures G.16, G.17, G.20 and G.21 verify that all exponents remain unaltered, albeit slightly more noisy.

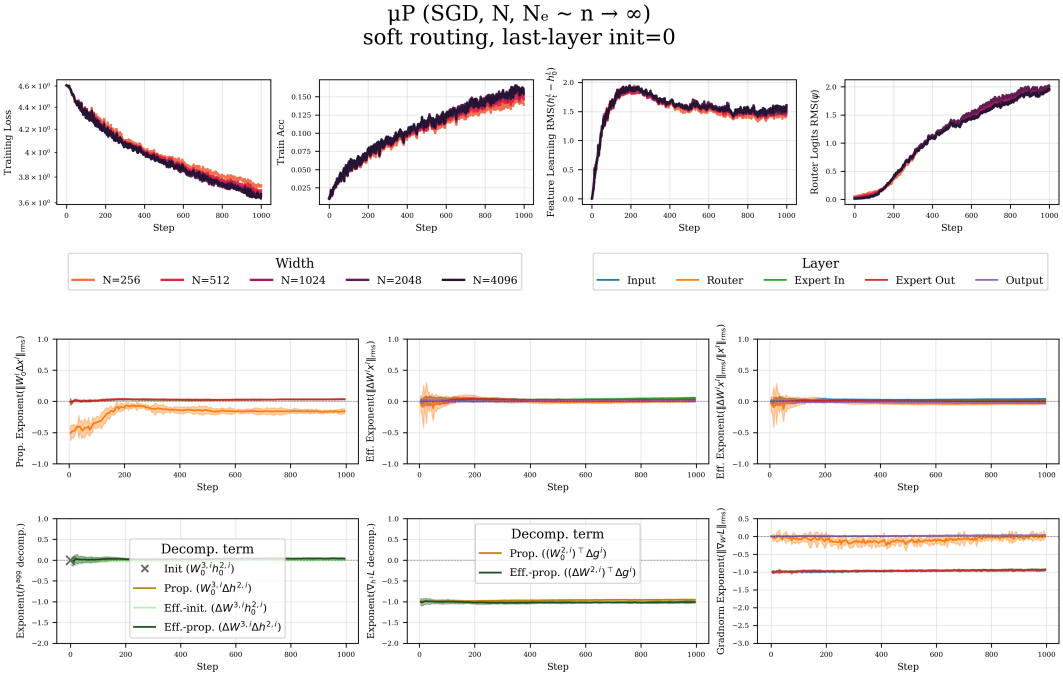


Figure G.14: **μP with 1/N router init (SGD, Regime I).**

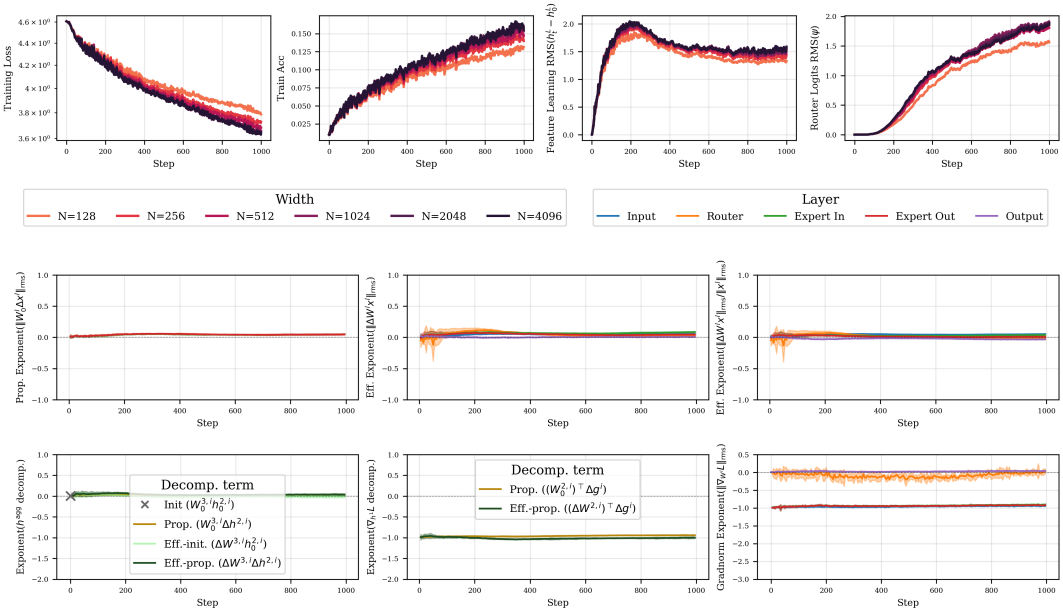


Figure G.15: μ**P with zero router init (SGD, Regime I).**

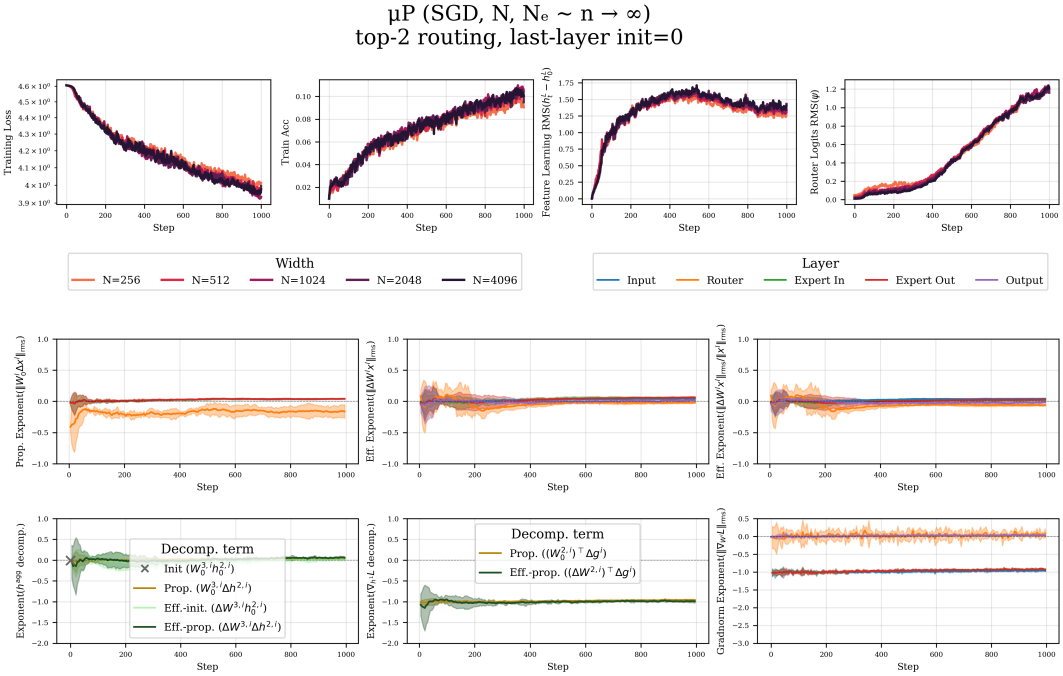


Figure G.16: μ**P with 1/N router init (SGD, Regime I, top-k).**

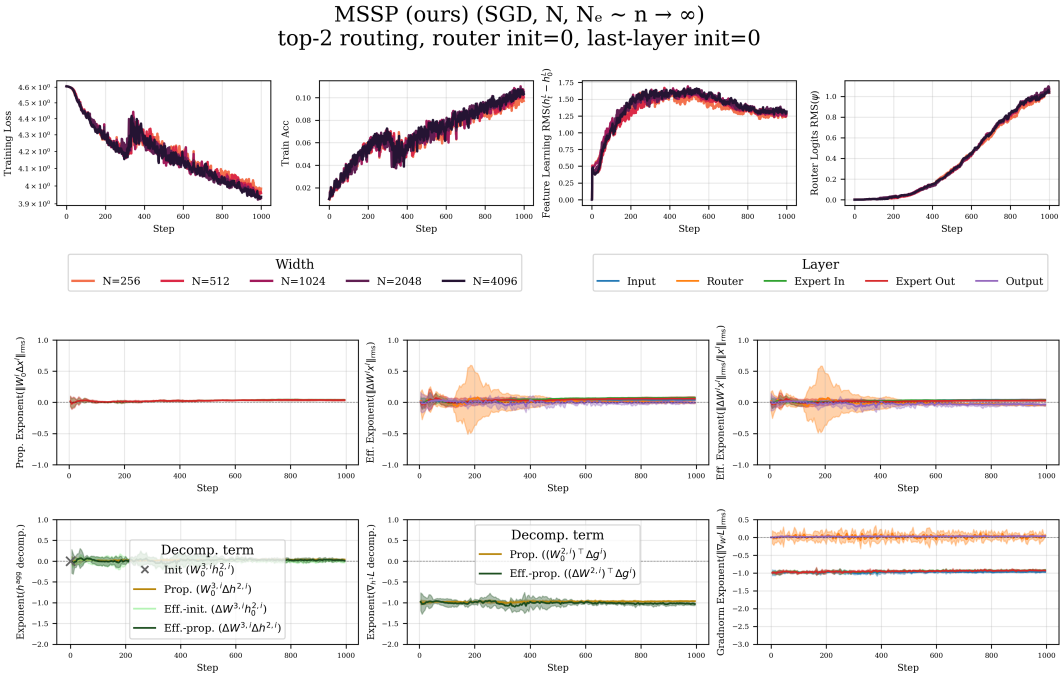


Figure G.17: **MSSP with zero router init (SGD, Regime I, top-k).**

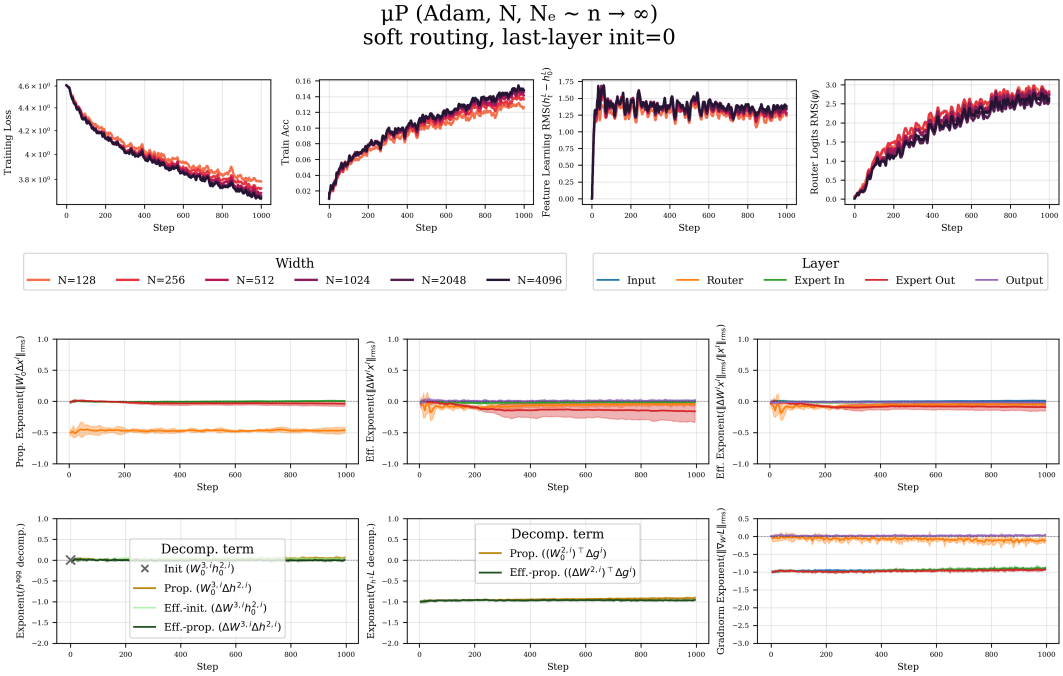


Figure G.18: **μP with 1/N router init (Adam, Regime I).**

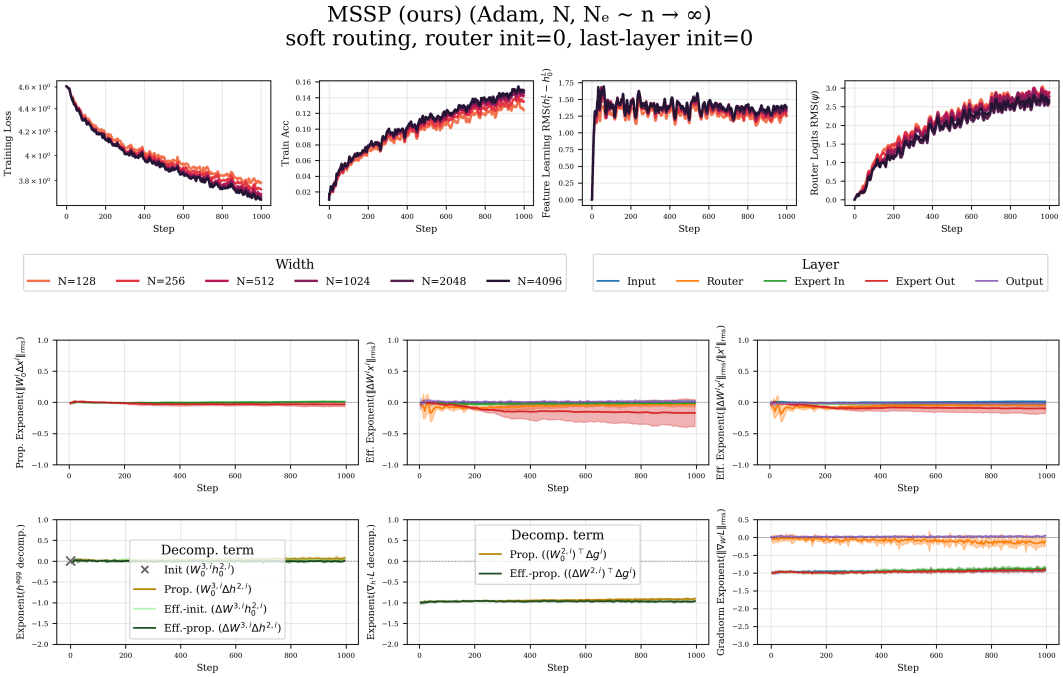


Figure G.19: **MSSP with zero router init (Adam, Regime I).**

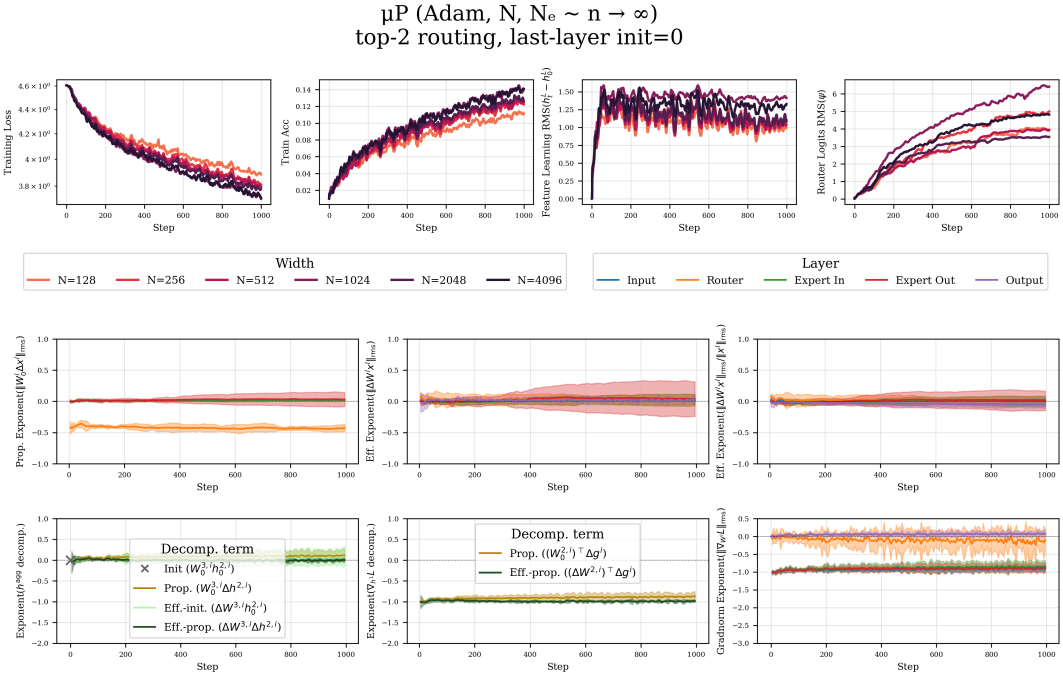


Figure G.20: μ**P with 1/N router init (Adam, Regime I, top-k).**

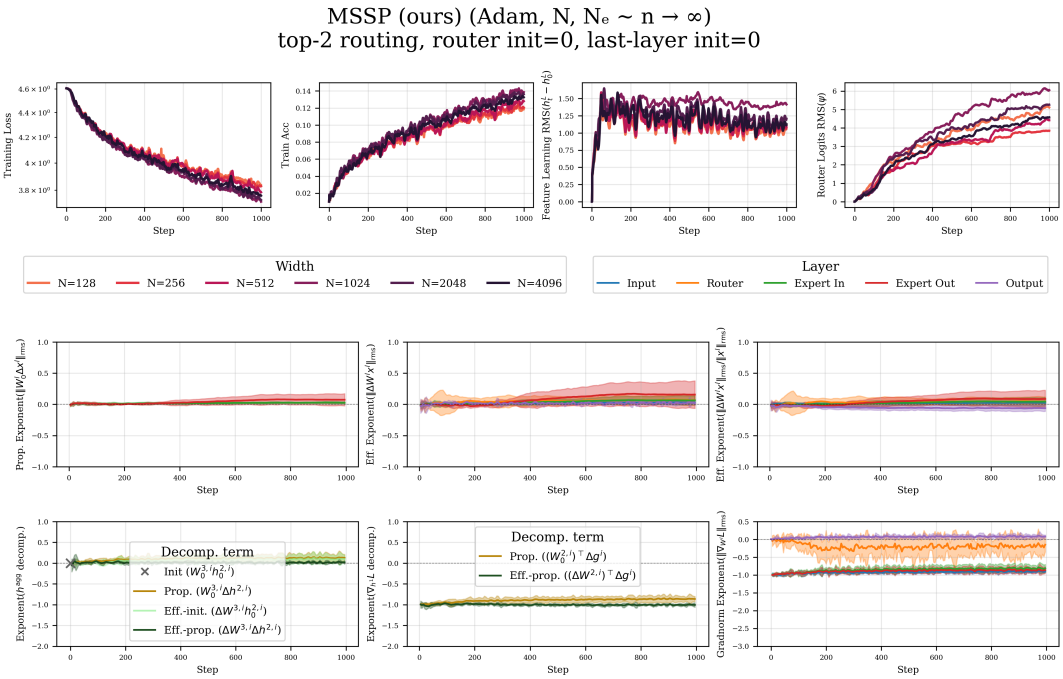


Figure G.21: **MSSP with zero router init (Adam, Regime I, top-k).**

G.2.2 Regime II: Fixed expert width

To ensure a stable baseline, we provide μP both with maximal stable and with 0 last-layer initialization. Both suffer from the same scaling degeneracies for SGD and Adam.

Starting with SGD, we observe strong delayed learning under maximal stable last-layer initialization $\sigma = 1/N$. This delay is visible not only in training loss, but also feature and router logit learning. Initial vanishing terms in the expert aggregation cascade even stronger into all layers than under last-layer 0 initialization presented in the main paper, resulting in final exponents that do not follow any clean scaling exponent from $\{-0.5, 0, 0.5\}$. μP with last-layer zero initialization starts out too small but partially self-corrects, which verifies that this parameterization indeed satisfies the μP desiderata, as any larger initialization or learning rate scaling would induce divergence in at least one layer after sufficiently many steps.

Adam in μP stabilizes much faster than SGD, and, while exponents are still not clean, they differ less from 0. Still, performance does not monotonically improve with scale and reduced feature learning at large scales is visible. Again, MSSP resolves these issues and shows monotonic improvement with scale as well as clean and balanced scaling exponents.

Propagating update scaling in μP remains vanishing throughout training for both SGD and Adam, as our theory predicts.

Note that, as predicted, expert output layer gradient entries decay extremely fast as $\Theta(N^{-2})$ in both SGD and Adam. If ignored, this can cause numerical precision issues in practice at moderate model sizes. We recommend layerwise gradient scaling or equivalent Adam moment scaling to prevent numerical underflows.

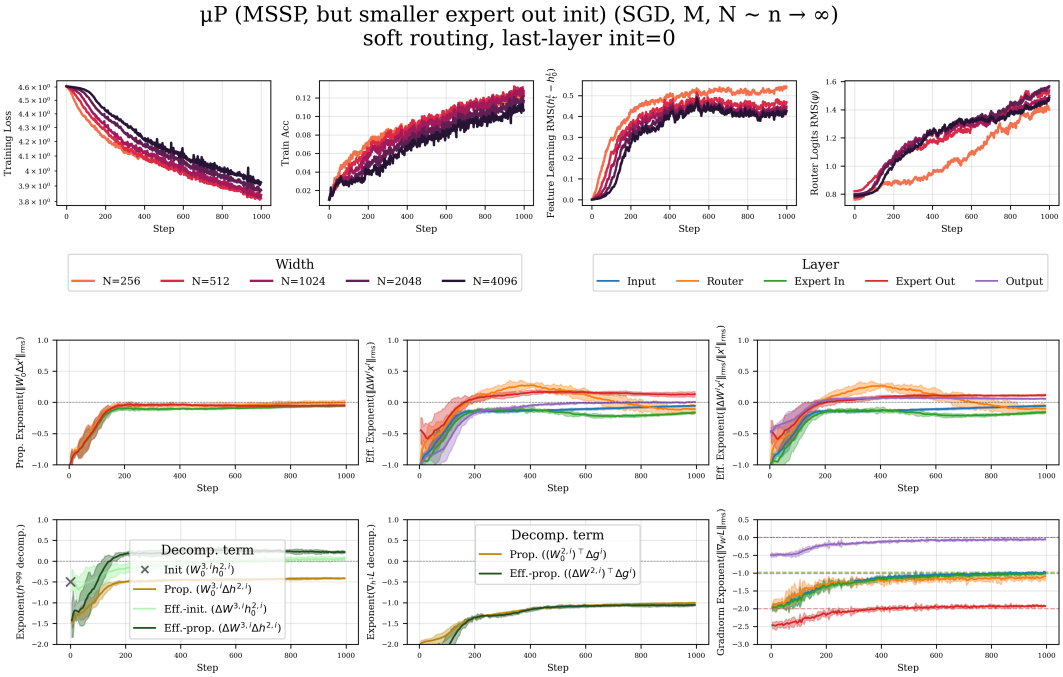


Figure G.22: **μP baseline (SGD, Regime II).**

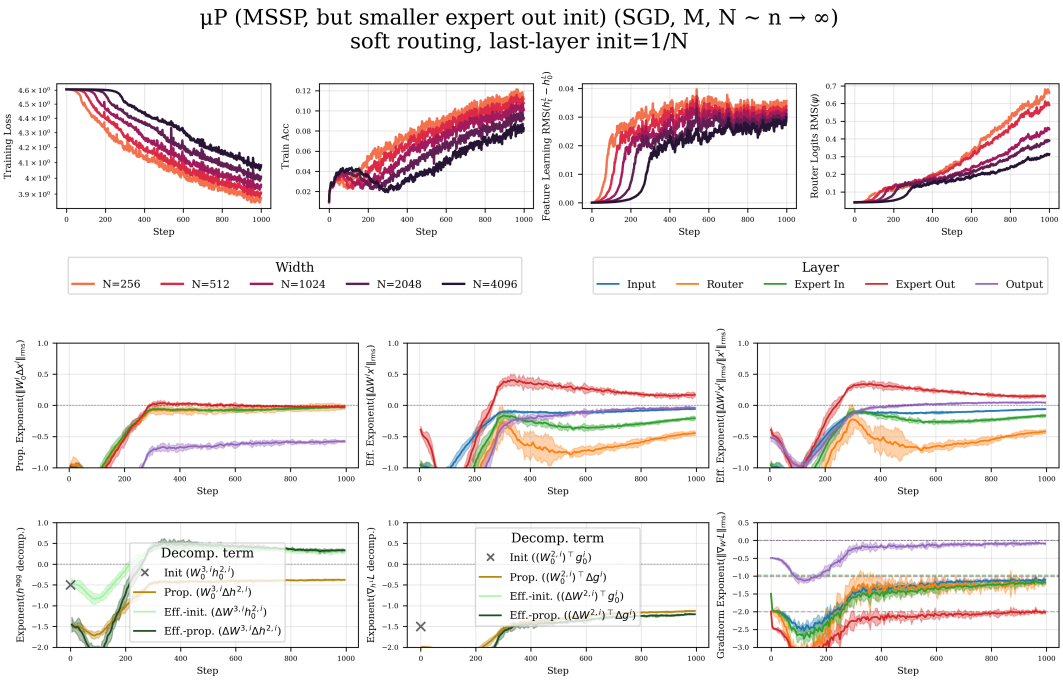


Figure G.23: **μP baseline with 1/N last-layer init (SGD, Regime II).** Interesting initial growth in training accuracy even though training loss and features are barely moving. Learning is still delayed with increasing width.

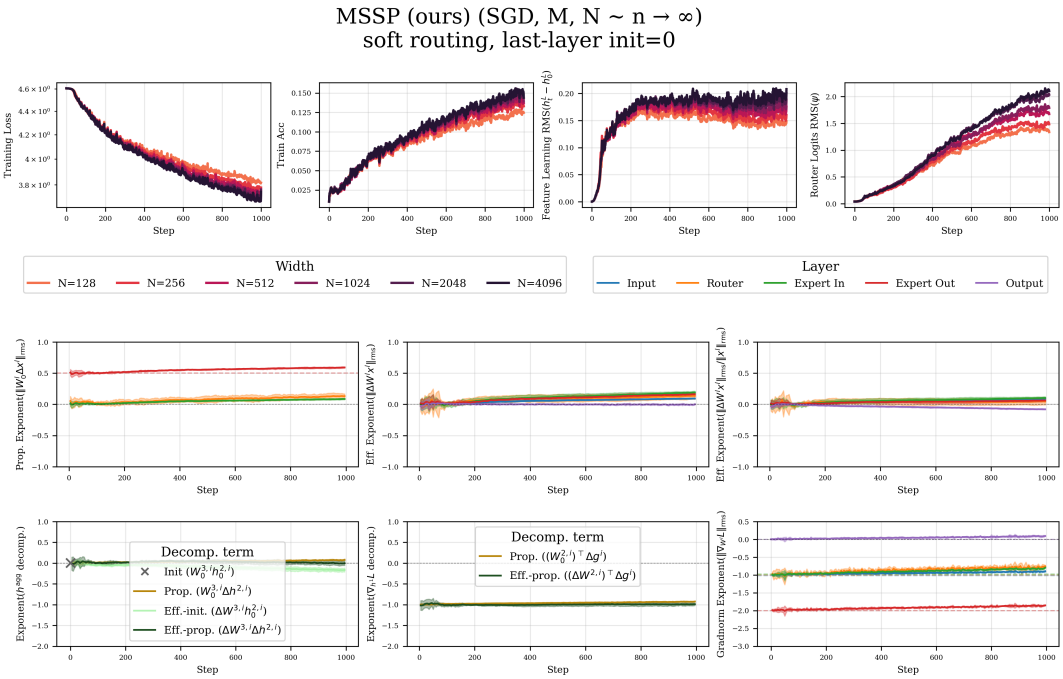


Figure G.24: **MSSP (SGD, Regime II).** Note that the propagating update exponent 0.5 of the expert output layer is desired here, since it is cancelled out by a clean CLT effect in the expert aggregation. The width independence of all other quantities throughout training verifies that the divergence of the individual expert outputs $h^{2,\text{out}}$ through this propagating update term results in approximately width-independent training dynamics.

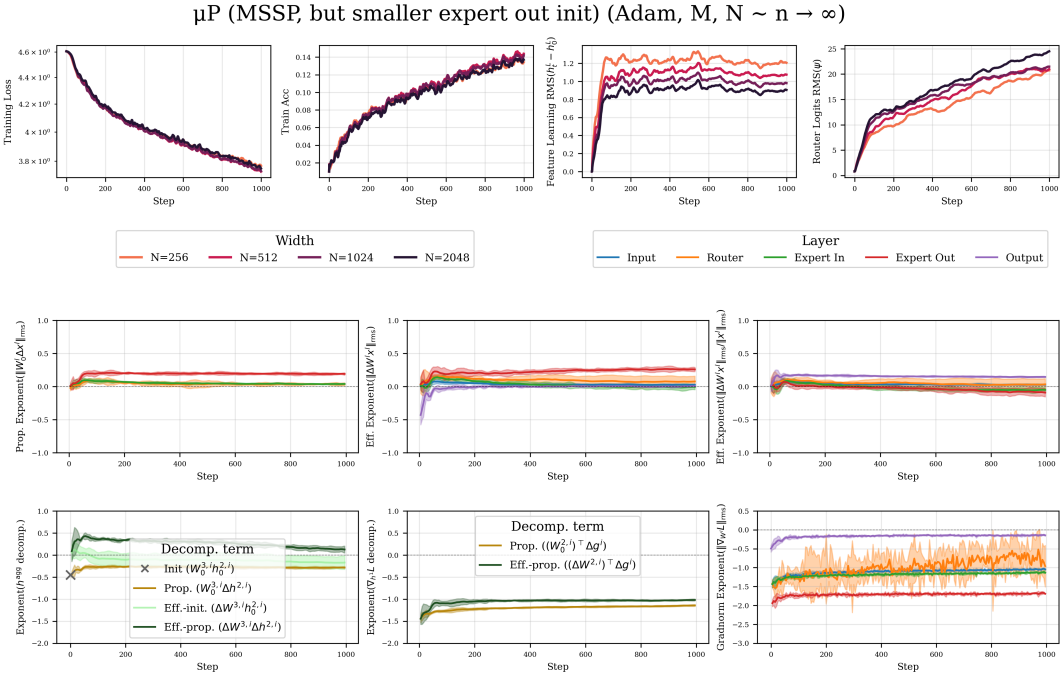


Figure G.25: **μP baseline (Adam, Regime II).**

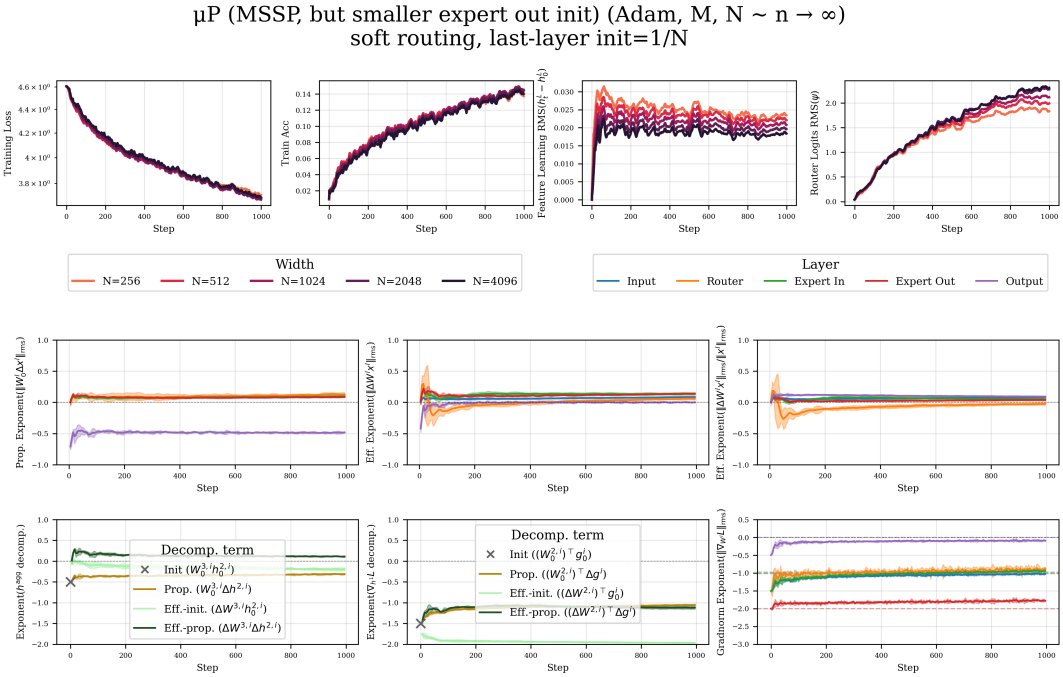


Figure G.26: **μP baseline with 1/N last-layer init (Adam, Regime II).**

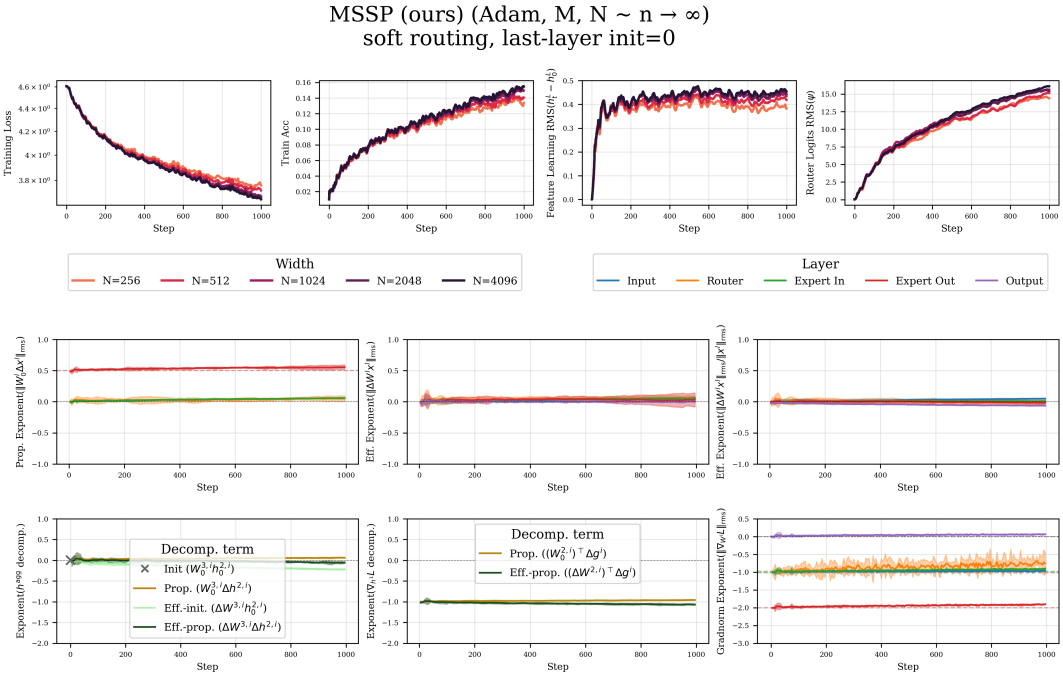


Figure G.27: **MSSP (Adam, Regime II).** Note that the propagating update exponent 0.5 of the expert output layer is desired here, since it is cancelled out by a clean CLT effect in the expert aggregation. The width independence of all other quantities throughout training verifies that the divergence of the individual expert outputs $h^{2,\text{out}}$ through this propagating update term results in approximately width-independent training dynamics.

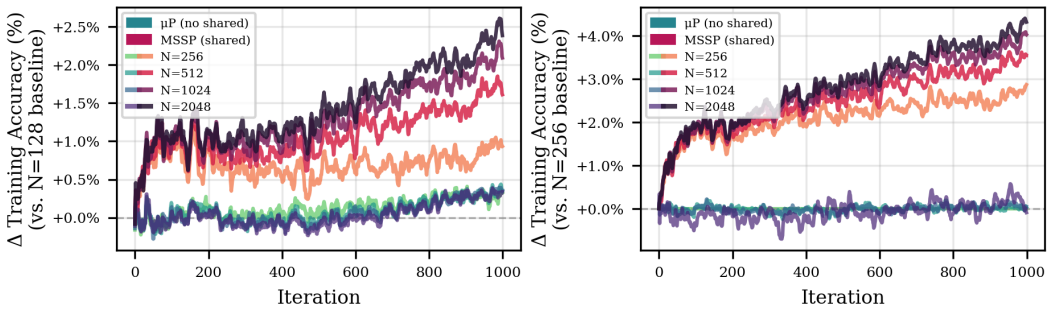


Figure G.29: **Monotonic improvement with scale only in MSSP, not in μP (Adam, Regime III).** Training accuracy difference compared to μP with $1/N$ (left) and 0 (right) last-layer initialization at $N = 128$, with separately tuned multipliers. MSSP outperforms both versions of μP.

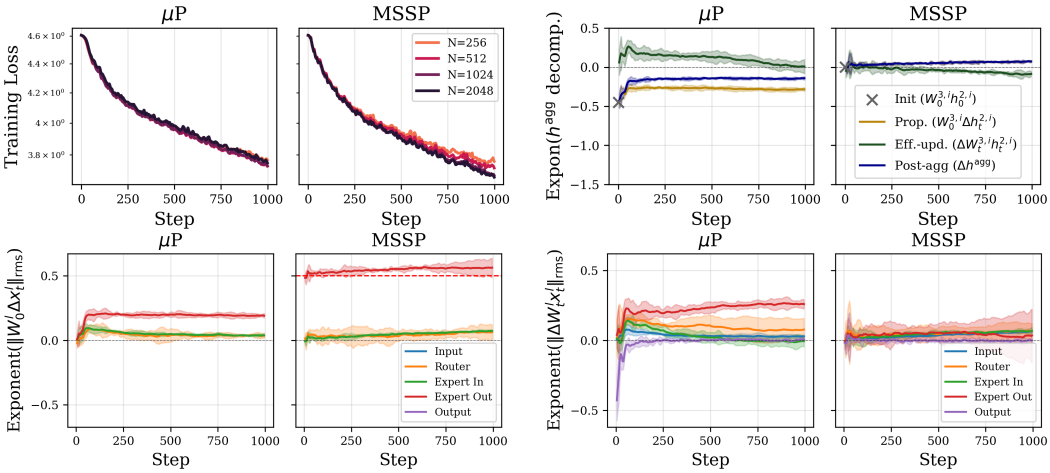


Figure G.28: **Monotonic improvement with scale only in MSSP, not in μP (Adam, Regime II).** Same as Figure E.2, but for Adam. While Adam's stability is less severely impacted at moderate model scale, expert output exponents are not clean and performance does not improve with model scale in μP.

G.2.3 Regime III: Joint proportional scaling

For SGD, observe delayed learning in μP similar to Regime II. In both μP and MSSP, the router gradient norm is very noisy, but this does not affect the width-independent signal propagation in MSSP.

Adam in μP is surprisingly stable in the all-scaling Regime III. Still, Adam's performance only improves robustly and significantly with scale in MSSP, both under soft and top-k routing.

top-k selection does not affect qualitative scaling properties as theoretically predicted.

As in Regime II, note that expert input and output layer gradients are decaying extremely fast with exponent -2 as predicted in both SGD and Adam, requiring layerwise gradient or Adam moment scaling to prevent numerical underflows at moderate scale.

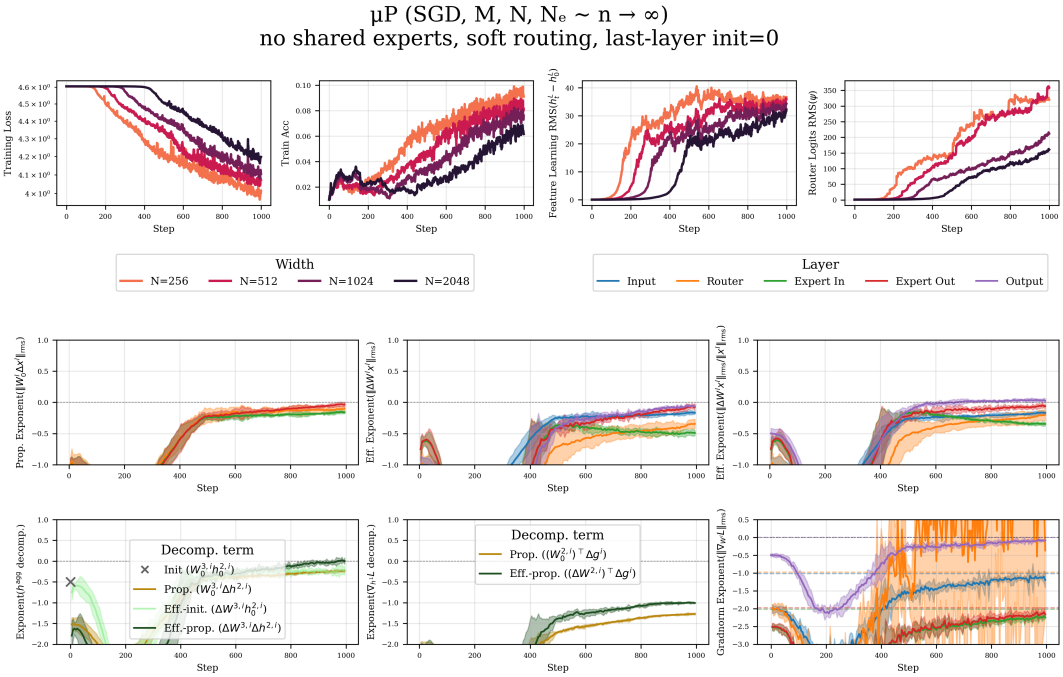


Figure G.30: μ**P without shared experts (SGD, Regime III).**

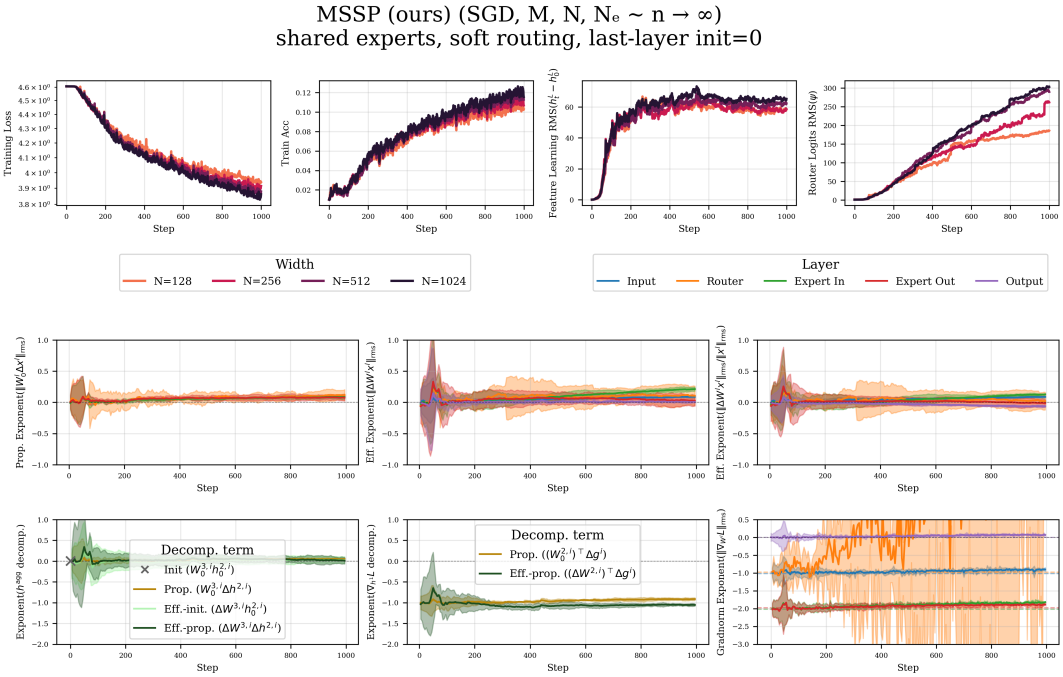


Figure G.31: **MSSP with shared experts (SGD, Regime III).**

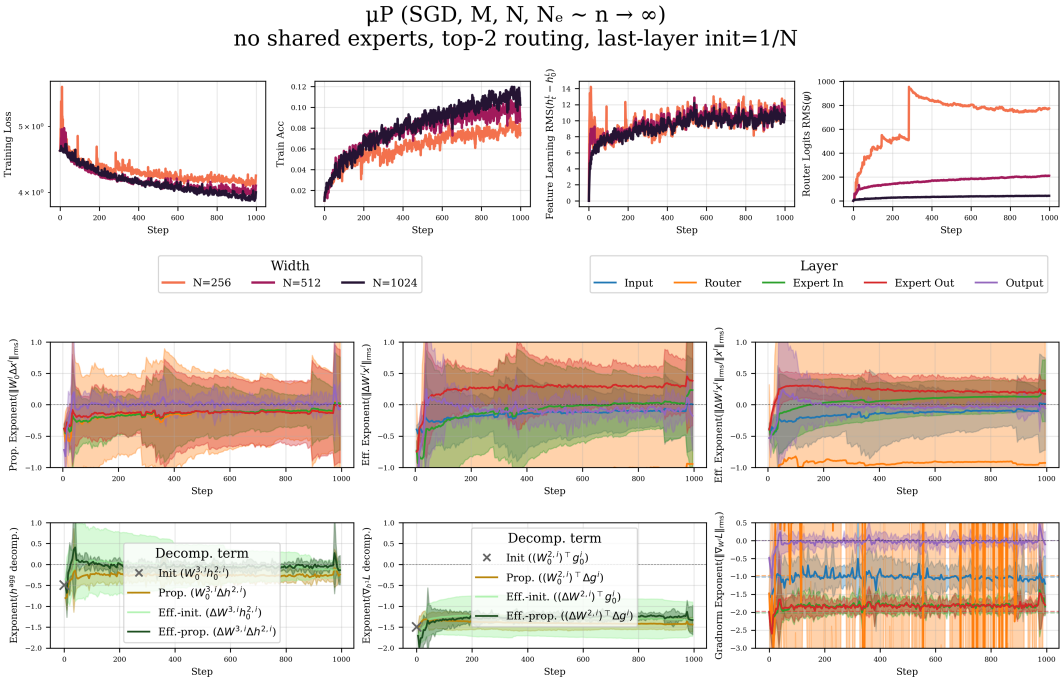


Figure G.32: **μP without shared experts with 1/N last-layer init (SGD, Regime III, top-k).**

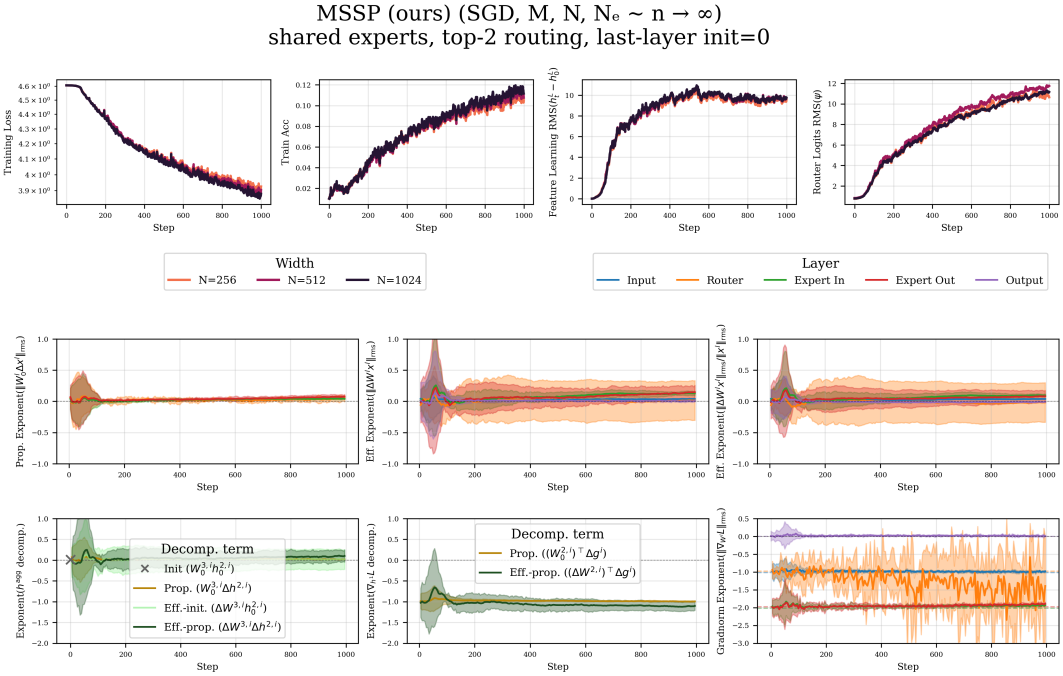


Figure G.33: **MSSP with shared experts (SGD, Regime III, top-k).**

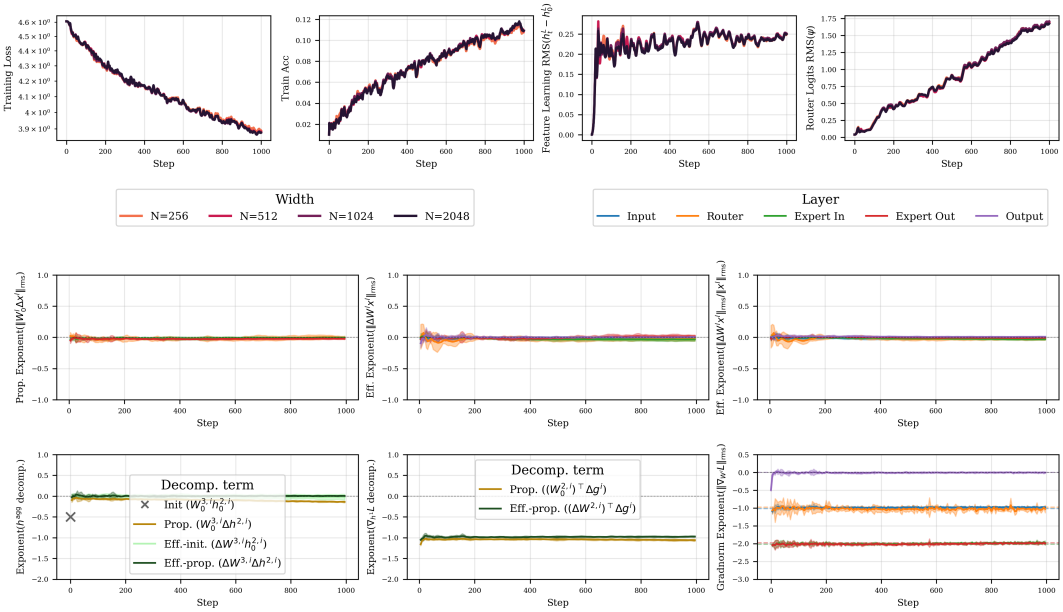


Figure G.34: μ**P without shared experts (Adam, Regime III).**

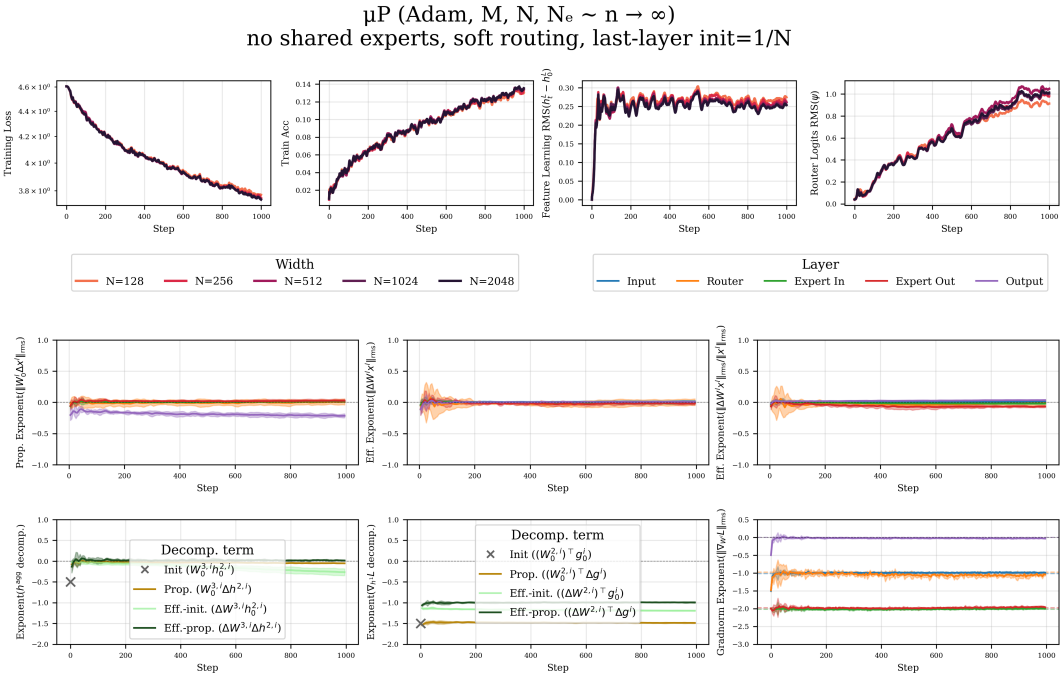


Figure G.35: μ**P with 1/N last-layer init, no shared experts (Adam, Regime III).**

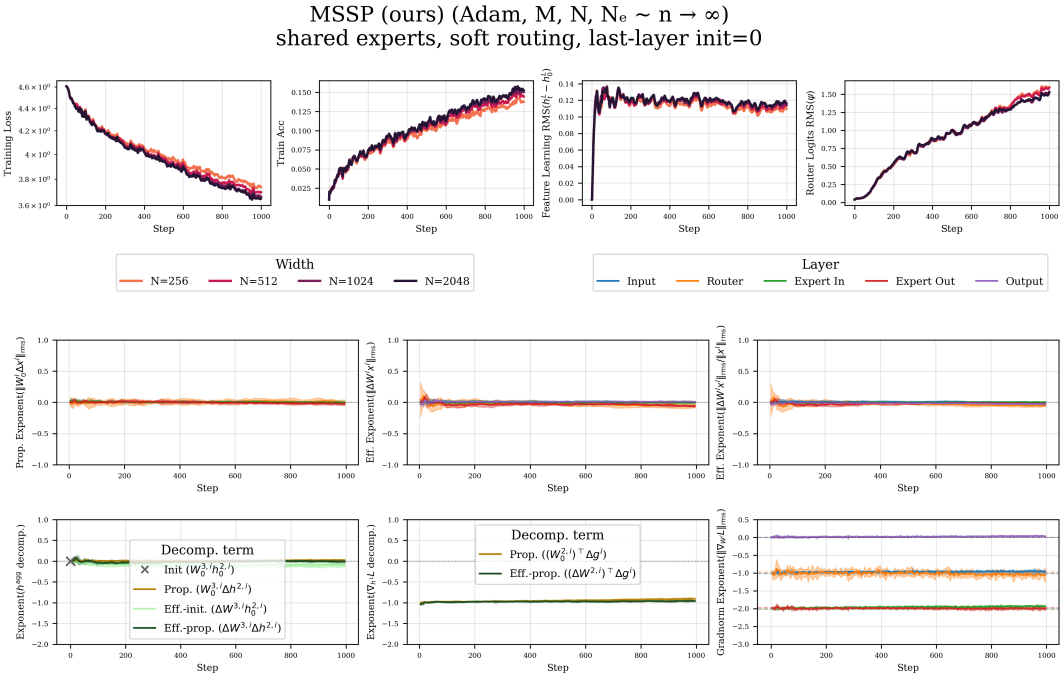


Figure G.36: **MSSP with shared experts (Adam, Regime III).**

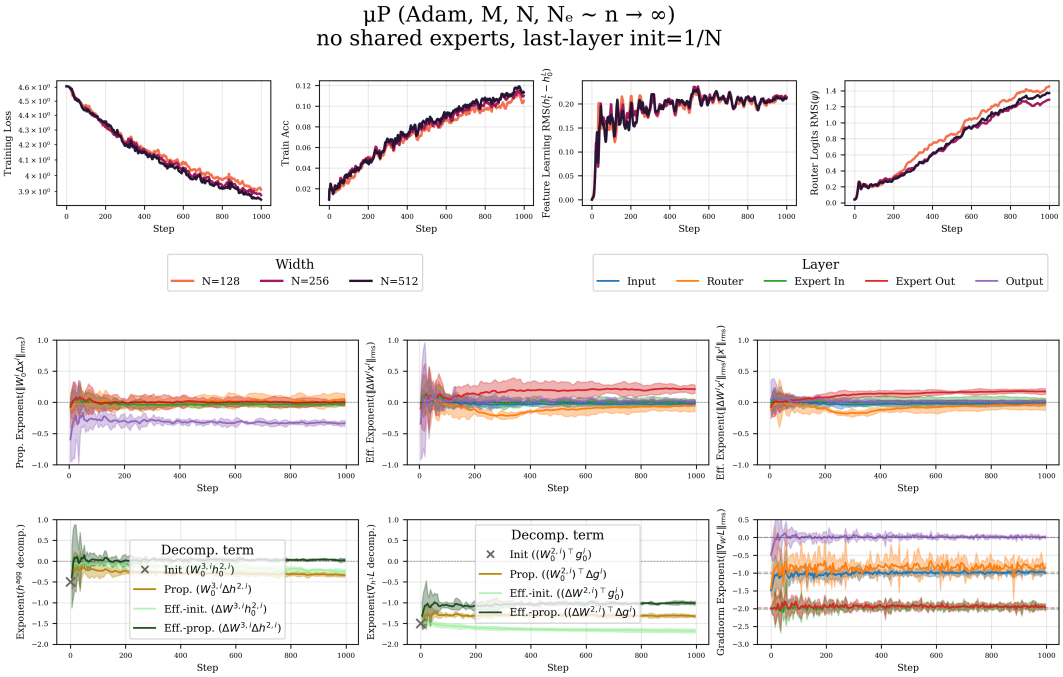


Figure G.37: **μP without shared experts with 1/N last-layer init (Adam, Regime III, top-k).**

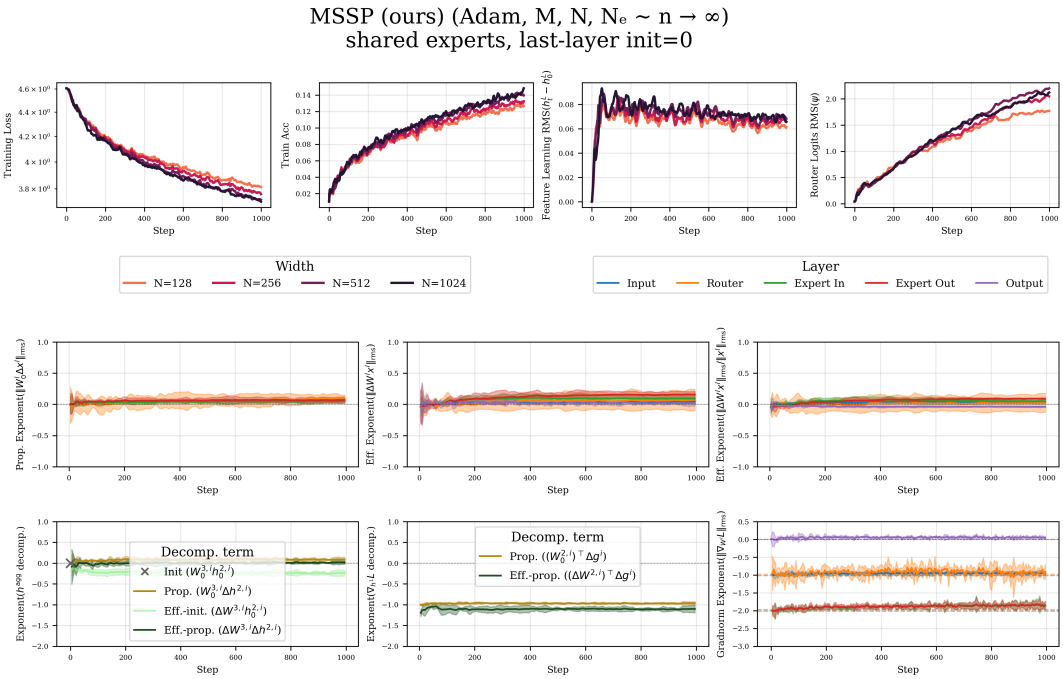


Figure G.38: **MSSP with shared experts (Adam, Regime III, top-k).**

G.3 MSSP for sparse scaling with fixed K

Our extension of MSSP to extremely sparse scaling preserves scale stability. Figure G.39 shows monotonic improvement with scale, and clean exponents. With tied experts, a zero-initialized router never breaks the symmetry under deterministic top-K tie-breaking, so we use the standard $N^{-1/2}$ router initialization here. Instead, one could also use router 0 initialization with Gumbel sampling top-k. The slightly growing router effective updates are a selection effect of top-K routing, not a failure of the parameterization. The refined coordinate check measures the router update $\Delta Q\, h$ only on the K active experts, and with fixed K these are the top 4 of $M = N/16 \in \{16, \ldots, 128\}$ experts. Over all experts, the router logits stay width-independent (with RMS 1.40, 1.59, 1.50 and 1.30 for $N = 256$ to 2048 at the end of training). Since the logits are small at initialization (RMS 0.04), they essentially equal the accumulated router update. The selected logits, however, are the largest K of M values, and they grow with M even when the distribution of the logits does not change. For i.i.d. Gaussian logits, the RMS of the top 4 of M grows from 1.33 at $M = 16$ to 2.24 at $M = 128$, a fitted width exponent of 0.25. This matches the measured exponent of the total router effective update, 0.26 ± 0.02 over 4 seeds. The growth follows the extreme-value scaling $\sqrt{2 \log M}$, so it is logarithmic rather than polynomial in width. No width-dependent multiplier can remove it, and the saturating sigmoid gate further damps its effect on the MoE output.

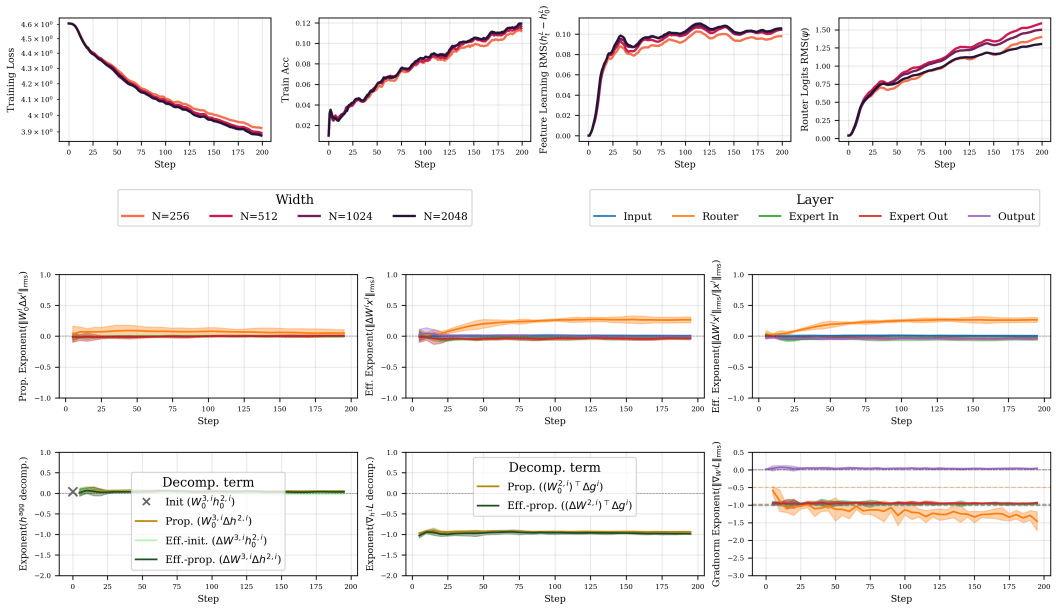


Figure G.39: **MSSP for sparse scaling K fixed, but $M \to \infty$ (Adam, Regime III, but fixed K).**

G.4 Soft softmax routing collapses to uniform for μP in Regime I

Here we train with single pass SGD with the optimal learning rate under MSE loss using soft softmax routing over binary classification from CIFAR-10 using all data points from class 'airplane' $+1$ and class 'automobile' -1 with batch size 64.

Figure G.40 shows learning rate sweeps scaling with μP versus standard parameterization (SP) under MSE loss. Lines ending on the right indicates divergence with NaNs under larger learning rates. Validation performance gets worse with scale in SP. Due to router collapse, performance does not improve with scale under soft routing in μP either. The optimal and maximal stable learning rate shrinks with width in SP, but transfers in μP.

As theoretically predicted, the router gradient vanishes with increased width (Figure G.41), which results in vanishing router updates inducing vanishing router logits inducing uniform routing. Figure G.42 shows that this collapse prevails over the entire course of training.

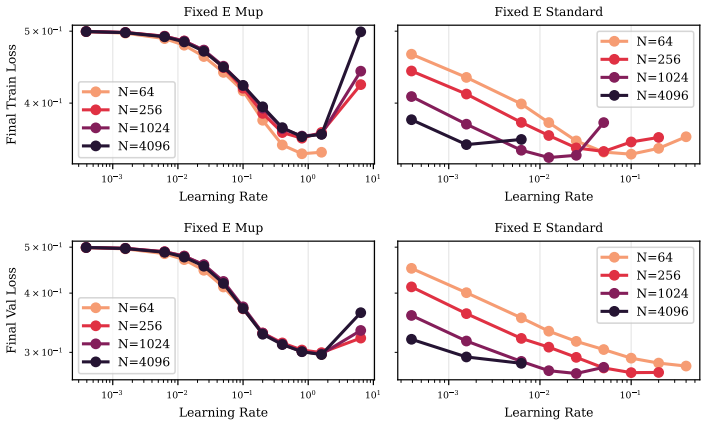


Figure G.40: **Learning rate transfer in Regime I (soft softmax routing, MSE loss, CIFAR-10).** Top: training loss. Bottom: validation loss. Left: μP. Right: SP. Ending lines denote divergence. Observe the maximal stable and optimal learning rate shrinking in SP, but staying consistent in μP. Performance does not monotonically improve due to router and expert collapse.

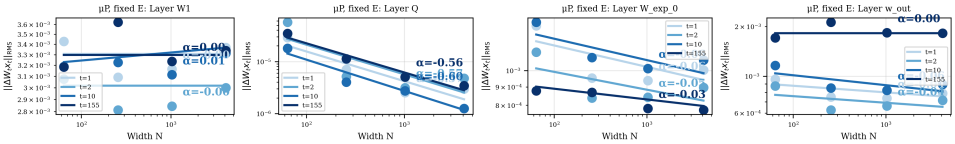


Figure G.41: **Effective updates across layers (μP, soft, Regime I).** Columns (left to right): input layer (W1), router layer (Q), first expert layer (W_exp_0), output layer (w_out). Approximate width-independence in μP, except in the router, which collapses under soft softmax routing.

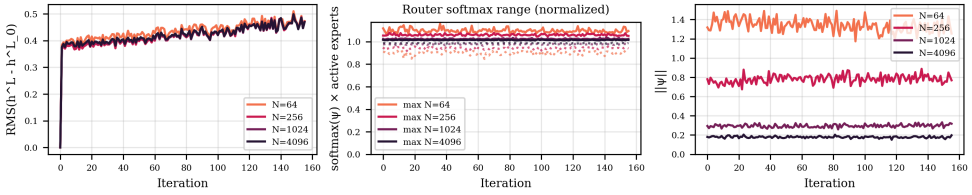


Figure G.42: **Soft softmax routing collapses to uniform in μP in Regime I.** Here, we train single pass SGD in μP over the CIFAR-10 subset described in this section. Columns (left to right): feature learning $\|\Delta h^L\|_{RMS}$, normalized post-softmax routing weights, router logit norm $\|\psi\|_{RMS}$. Routing collapses as routing logits converge to 0 with scale over the entire course of training. Feature learning is still approximately width-independent.

Across several settings, we found that expert specialization does not necessarily improve performance on CIFAR-10, as the dataset is too small and not diverse enough. Hence we use TinyImageNet for all other MLP experiments.

G.5 MoEs require layerwise learning rate tuning

Without layerwise learning rate multiplier tuning, effective updates in router and expert input layer can initially vanish even in MSSP (Figure G.43).

Figure G.44 shows that, while all exponents are approximately 0 as theoretically predicted, the propagating updates dominate the effective updates in absolute scale in the first steps by a factor of more than 10^6 in both the forward and backward pass aggregation operations.

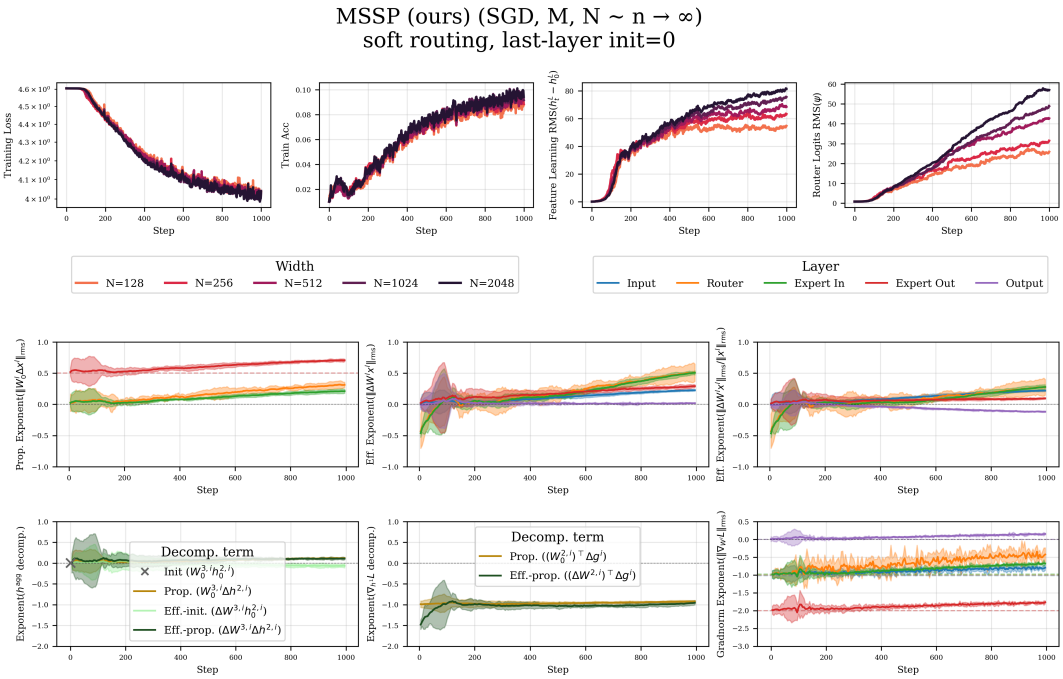


Figure G.43: **MSSP without tuned multipliers (SGD, Regime II).** Here we set all initialization variance and layerwise learning rate multipliers to 1.0 and only tune the global learning rate.

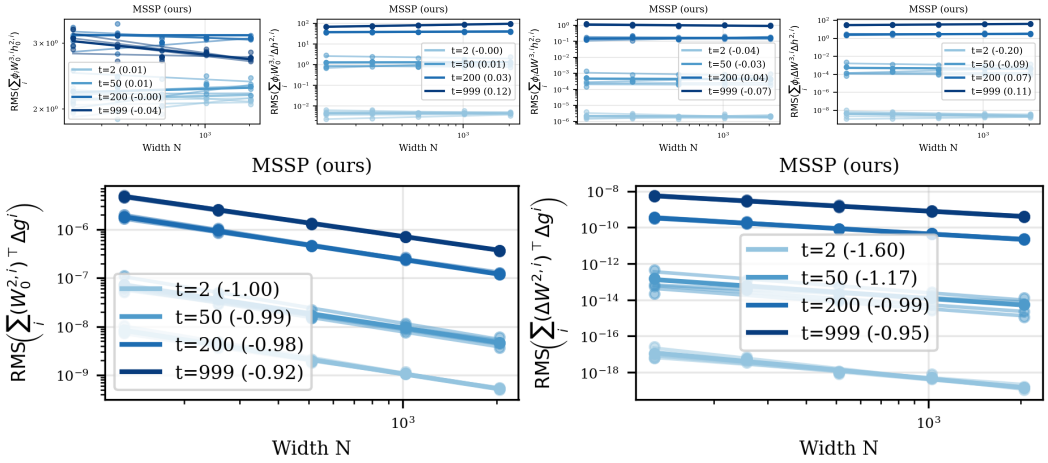


Figure G.44: **Without tuned multipliers, sub-term contributions differ by orders of magnitude (MSSP, SGD, Regime II).** The forward pass expert aggregation components (top) and the backward pass components to $\nabla_{h^1}\mathcal{L}_t$. Propagating updates dominate the effective updates in absolute scale for several hundred steps. In the forward pass, propagating updates start out at around 10^0 and effective updates around 10^{-6}. In the backward pass, the propagating update term starts around 10^{-9} at $N = 1024$ and $t = 2$, whereas the effective update term starts around much smaller 10^{-18}. Hence the weight updates contribute vanishingly to the overall first-layer gradient $\nabla_{h^1}\mathcal{L}_t$.

Hence, in parameterizations where correctly scaled effective updates should dominate vanishing propagating updates, the propagating updates can still dominate in absolute value for several hundred steps at realistic scales. Hence it can seem that the empirical exponents do not follow the predicted ones, when not measured in sufficient granularity.

After layerwise learning rate tuning, which can consequently require grids with ranges beyond 10^6, we observe that propagating and effective updates generally end up having a similar order of

magnitude in the first 1000 steps, and hence we accurately measure the predicted limiting exponents at moderate model scales.

Overall this highlights that extensively tuning layerwise learning rate multipliers is even more essential in MoEs than in dense models.

G.6 Random search and 2D multiplier tuning do not suffice

Instead of full 6D sweeps, one could also tune multipliers more efficiently by running a random search, and, starting from the HPs found in the first stage, doing 2D sweeps of all HP pairs around the optimum. This approach finds near-optimal HPs if the suboptimality of the random stage is predominantly restricted to a subspace allowing 2 interacting multipliers.

However, Figure G.45 shows that 2D sweeps after a random sweep do not suffice. For example in SGD μP Regime III (left), all 2D sweeps containing the expert input or output layer lr multiplier suggest that smaller values would perform better, but the joint change of (lr router, lr out) dominates and does not allow a third change. Similarly in SGD μP Regime II (center) the pair (lr router, lr expert2) marginally dominates other improvements such as even smaller expert output learning rate. Hence, after such a 2D sweep, it remains unclear whether the new expert output learning rate, and all other HPs, are indeed robustly near-optimal, or whether higher-order interactions continue to induce an indefinite 2D update without ever converging to a local optimum. In Adam MSSP Regime I (right), (lr router, lr in) are updated and it remains unclear whether the global optimum would also include a reduced expert output lr multiplier.

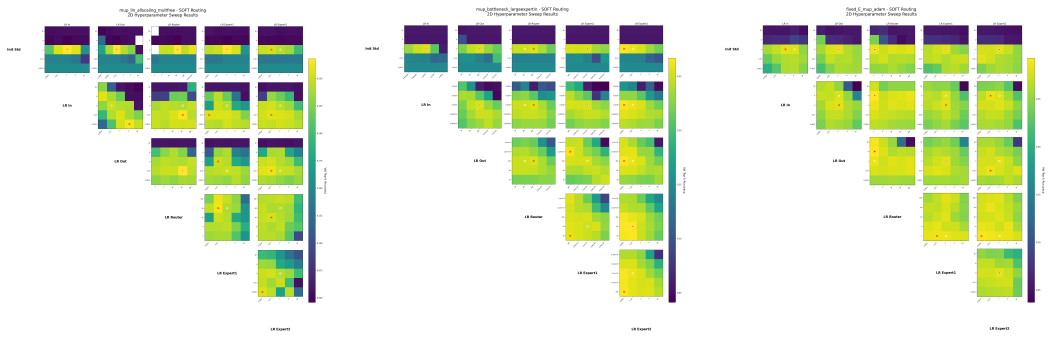


Figure G.45: **2D multiplier sweeps at small scale $N = 128$ from random search stage.** 2D heatmaps showing top-5 validation accuracy of all HP pairs, while fixing all other multipliers at the optimum found from a random sweep over at least 1024 runs. From left to right: SGD μP Regime III, SGD μP Regime II, Adam MSSP Regime I. White circles denote the optimum after this 2D stage, red 'x' denote the optimum within the respective 2D sweep. Often these do not align, suggesting that higher dimensional interactions make this hyperparameter tuning strategy insufficient.

G.7 Global Adam ε induces width dependence at sufficient scale

Here, we mimick the effects of further scaling – which we cannot run due to compute constraints – by increasing Adam ε, since gradient RMS norms decay with model scale. We compare naive constant Adam ε versus our layerwise Adam ε scaling under allscaling and layerwise LR multipliers tuned at width 128. Figure G.46 shows more time-dependent and less clean effective and propagating update exponents under global ε, especially of expert effective updates. Figure G.47 shows feature learning and router learning is reduced with scale under global epsilon, as updates are beginning to vanish.

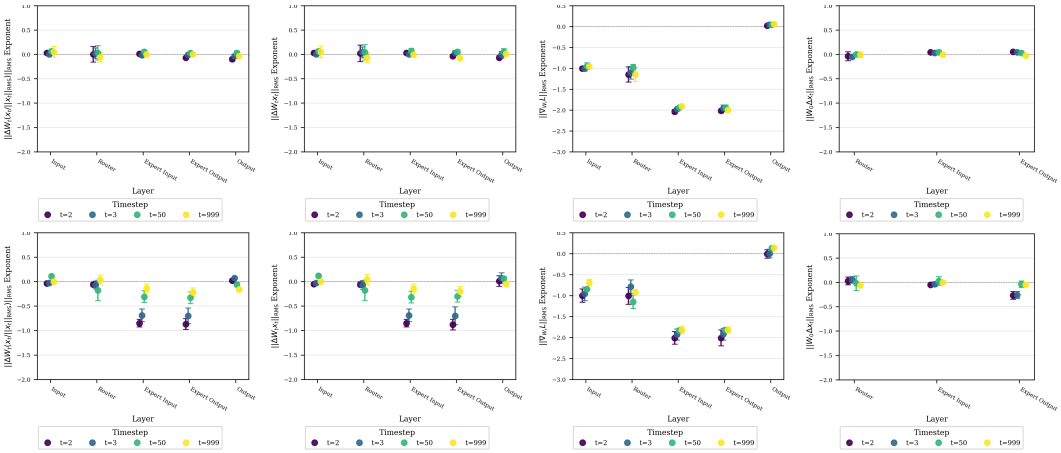


Figure G.46: **Global Adam ε induces entangled exponents.** RCC exponents of Adam MSSP in Regime III with layerwise ε scaling (Ours, top), the same scaling rule with constant global ε (bottom). Columns (left to right): Effective Updates (Normalized), Effective Updates (Raw), Gradient Norms (Raw), Propagating Updates (Raw).

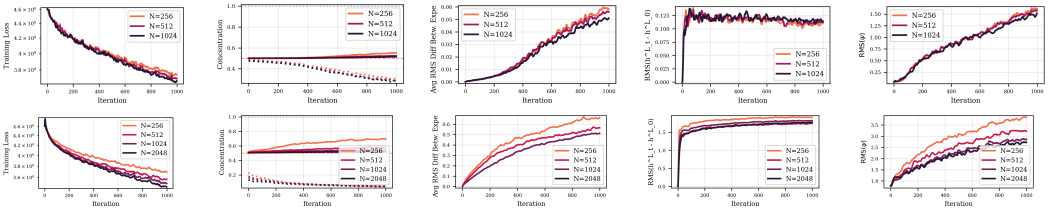


Figure G.47: **Reduced feature learning and router learning under global Adam ε.** Adam MSSP in Regime III with layerwise ε scaling (top) versus the same scaling rule with constant global Adam ε (bottom). Columns (left to right): Training loss, router concentration (0.5 indicates uniform routing, solid lines denote the maximal and dotted lines the minimal routing weight in relation to uniform routing), average RMS norm difference between experts, accumulated feature learning ($\|\Delta h^L\|_{RMS}$), router logit norm $\|\psi\|_{RMS}$.

Part IV
Theory

This part of the appendix presents our scaling analysis at three levels of formalism, each progressively more precise. We start with an intuitive explanation in Appendix H, then provide the detailed signal propagation analysis in Appendix J, and finally provide a full DMFT analysis for each scaling regime in Appendices K to O.

H Intuitive Explanations of the Shortcomings of μP for MoEs and How MSSP Fixes Them

To set the stage for our formal results, we provide intuition for how to derive scaling rules for MoE blocks that satisfy Desiderata 1 and 2. Table B.1 reports μP and MSSP in each scaling regime. As described in Section 2, we derive the SGD scaling in each regime by signal propagation and validate it with the corresponding DMFT; the Adam and Muon scalings are heuristic extensions. The conditions under which these parameterizations satisfy Desideratum 3 are discussed in Section J. Here and throughout, asymptotic notation (e.g., $\Theta(\cdot)$) is understood with respect to the joint scaling trajectory $\mathcal{S}(n)$ as $n \to \infty$, specialized to the scaling regime (I–III) under consideration. For the intuitive explanation, we focus mainly on Regimes II and III. Regime I is more straightforward and is omitted for brevity.

H.1 Overview

Let's start by recalling the Maximal update desiderata[2].

H.1.1 Maximal Update Desiderata

The Maximal Update Desiderata require that in each layer, the effective and propagating updates remain $\Theta(1)$ in RMS at some $t > 0$:

$$\|\mathcal{T}^{\ell}_{\text{eff}}(t)\|_{\text{RMS}} = \|\Delta W^{\ell}_t\, x^{\ell-1}_t\|_{\text{RMS}} = \Theta(1), \qquad \|\mathcal{T}^{\ell}_{\text{prop}}(t)\|_{\text{RMS}} = \|W^{\ell}_0\, \Delta x^{\ell-1}_t\|_{\text{RMS}} = \Theta(1). \tag{H.1}$$

For the MoE aggregation layer, we additionally require that the updates to the aggregated representation remains non-vanishing and non-diverging:

$$\|\Delta h^{\text{agg}}_t(t)\|_{\text{RMS}} = \Theta(1). \tag{H.2}$$

H.1.2 Effective and propagating updates

Alignment exponents for effective and propagating updates. In all regimes, the asymptotic scales of forward and backward effective and propagating terms are governed by the training-induced correlation structure between the two factors appearing in each product. Following Yang et al. (2023); Everett et al. (2024), we summarise this effect via *alignment exponents* p_ℓ and q_ℓ (p^{∇}_ℓ and q^{∇}_ℓ analogously for backward), defined by

$$\begin{aligned} \|\Delta W^{\ell}_t x^{\ell-1}_t\|_{\text{RMS}} &= \Theta\big((n^{\ell}_{\text{in}})^{p_\ell}\, \|\Delta W^{\ell}_t\|_{\text{RMS}}\, \|x^{\ell-1}_t\|_{\text{RMS}}\big)\,, \\ \|W^{\ell}_0 \Delta x^{\ell-1}_t\|_{\text{RMS}} &= \Theta\big((n^{\ell}_{\text{in}})^{q_\ell}\, \|W^{\ell}_0\|_{\text{RMS}}\, \|\Delta x^{\ell-1}_t\|_{\text{RMS}}\big)\,, \end{aligned} \tag{H.3}$$

where $n^{\ell}_{\text{in}} := \dim(x^{\ell-1})$ denotes the fan-in of layer ℓ. Informally, p_ℓ quantifies the degree of alignment between ΔW^{ℓ}_t and the current features $x^{\ell-1}_t$, while q_ℓ quantifies the degree of alignment between the initial weights W^{ℓ}_0 and the upstream feature change $\Delta x^{\ell-1}_t$.

From (p_ℓ, q_ℓ) to scaling rules for optimizers. Fix a layer ℓ and suppose the alignment exponents (p_ℓ, q_ℓ) are known. Assume inductively that at some time $t > 0$ the previous-layer quantities satisfy

[2]We omit an explicit discussion of the optimizer faithfulness desideratum here for brevity but we will briefly discuss ϵ scaling.

$\|x_t^{\ell-1}\|_{\mathrm{RMS}} = \Theta(1)$ and $\|\Delta x_t^{\ell-1}\|_{\mathrm{RMS}} = \Theta(1)$. Enforcing Desideratum 1 at layer ℓ then reduces, via (H.3), to choosing scales for $\|W_0^\ell\|_{\mathrm{RMS}}$ and $\|\Delta W_t^\ell\|_{\mathrm{RMS}}$ so that the effective and propagating updates remain $\Theta(1)$. These marginal scales are directly controlled by optimizer hyperparameters: initialization variances set the scaling of $\|W_0^\ell\|_{\mathrm{RMS}}$, while learning-rate scaling sets the scaling of $\|\Delta W_t^\ell\|_{\mathrm{RMS}}$ once the scale of the layerwise gradients is known. For Adam, optimizer faithfulness further requires ϵ to remain $\Theta(1)$ relative to the typical gradient scale; accordingly, we choose ϵ to scale with the layerwise gradient RMS norm. The same reduction applies to the backward pass, with backward alignment exponents $(p_\ell^\nabla, q_\ell^\nabla)$ playing the role of (p_ℓ, q_ℓ) in setting the leading-order scales of $(W_0^\ell)^\top \delta_t^\ell$ and $(\Delta W_t^\ell)^\top \delta_t^\ell$, where $\delta_t^\ell := \nabla_{h_t^\ell} \mathcal{L}$ denotes the pre-activation gradient passed backward.

H.1.3 DMFT predictions.

Our DMFT analysis provides the regime-specific asymptotic inputs needed to instantiate the above program. It predicts the forward alignment behaviour (p_ℓ, q_ℓ), the backward alignment behaviour $(p_\ell^\nabla, q_\ell^\nabla)$, and the leading-order scaling of expert-indexed aggregates that enter both $\Delta h^{\mathrm{agg}}(t)$ and the gradients at shared representations when M and K scale. Together, these predictions determine the regime-appropriate scaling of initialization variances, learning rates, and (for Adam) ϵ, yielding the canonical parameterizations reported in Table **??**.

The remainder of this section follows the programme described above in two stages. Section H.2 treats single linear maps $W \cdot x$ and the alignment exponents (p_ℓ, q_ℓ) that govern them. Section H.7 treats the MoE-specific cross-expert aggregations, organized as a small taxonomy of structural mechanisms that determine whether each aggregation term survives at leading scale or is suppressed by the cross-expert CLT effect. The same taxonomy makes the imbalances under μP in Regimes II and III transparent, and identifies, for each mechanism, the structural change that MSSP introduces to restore balance.

H.2 Per-layer alignment: scaling of linear (forward) maps $W \cdot x$

Each of the products entering the alignment-exponent decomposition (H.3) is a single linear map $W\,x$, and the prefactors in $\|Wx\|_{\mathrm{RMS}}$ are governed by random-matrix and rank-1-alignment arguments below. We focus mainly on the forward pass; the backward arguments are analogous.

H.2.1 Random-matrix preliminaries

Let $A \in \mathbb{R}^{n\times n}$ have i.i.d. entries $A_{ij} \sim \mathcal{N}(0, \sigma^2)$. Then A acts as a near-isometry rescaled by $\sigma\sqrt{n}$ along almost every direction in $\mathbb{R}^n$: its singular values are $\Theta(\sigma\sqrt{n})$ throughout the Marchenko–Pastur bulk (Marchenko and Pastur, 1967; Bai and Silverstein, 2010), and the corresponding singular vectors are uniformly distributed on S^{n-1} (Anderson et al., 2010, Ch. 2).

The directions along which A fails to act as such an isometry, namely those aligned with its smallest singular values, span only a vanishing fraction of $\mathbb{R}^n$ and form an exponentially thin set on the sphere (e.g. Vershynin, 2018, §3.1). Therefore, for generic $x \in \mathbb{R}^n$ (even when x depends on A, provided it is not adversarially chosen to lie in this subspace),

$$\|Ax\|_2 \asymp \sigma\sqrt{n}\,\|x\|_2.$$

Equivalently, in RMS norm,

$$\|Ax\|_{\mathrm{RMS}} \asymp \sqrt{n}\,\|A\|_{\mathrm{RMS}}\,\|x\|_{\mathrm{RMS}}.$$

We refer to this as *CLT-like scaling*: each entry of Ax is a sum of n uncorrelated mean-zero terms, so its typical magnitude grows only as $\sqrt{n}$ relative to the per-term scale, in analogy with the Central Limit Theorem. Later, we will contrast this with *LLN-like scaling*, where the prefactor becomes n rather than $\sqrt{n}$.

H.2.2 Propagating updates

Standard MLP hidden layer: CLT-like scaling. Applying the arguments above to a standard MLP hidden layer, with initial weights (W_0) an i.i.d. Gaussian matrix whose fan-in and fan-out both

diverge at rate n, a propagating perturbation Δx inherits this CLT-like scaling:

$$\|W_0\,\Delta x\|_{\mathrm{RMS}} \asymp \sqrt{n}\,\|W_0\|_{\mathrm{RMS}}\,\|\Delta x\|_{\mathrm{RMS}}.$$

Non-trivial feature learning requires the previous layer's updates to satisfy $\|\Delta x\|_{\mathrm{RMS}} = \Theta(1)$, and for the propagating update to remain $\Theta(1)$ after passing through the weight matrix we therefore need

$$\|W_0\|_{\mathrm{RMS}} \asymp \tfrac{1}{\sqrt{n}}, \qquad \text{equivalently,} \qquad \sigma \asymp \tfrac{1}{\sqrt{n}}.$$

The CLT-like scaling above hinges on A mapping between two diverging dimensions: for an $n \times n$ i.i.d. Gaussian, the operator norm is $\Theta(\sigma\sqrt{n})$, which caps the prefactor at $\sqrt{n}$ regardless of how x depends on A. No alignment, however structured, can lift the scaling to n in the $\infty \to \infty$ case. When at least one dimension of A is finite or A has an extremely skewed aspect ratio, this argument no longer applies and alignment-induced LLN scaling becomes possible, although whether it actually occurs depends on the surrounding network dynamics; we shall give some examples below to illustrate this. In the $\infty \to \infty$ case, x has too few degrees of freedom to align coherently with all n rows of A at once: maximal alignment with any single row boosts only the corresponding entry of Ax to LLN order, while the remaining $n-1$ entries stay at CLT order and dominate the norm. There is no way for one x to align with n independent rows simultaneously.

Remark. In all three regimes, propagating updates (both forward and backward) of all the layers that have diverging input and output dimensions admit CLT-like scaling. For instance, in Regime III all the layers (including the experts and the router), except the input and the output layer, fall into this category.

Readout layer: the canonical LLN-like scaling. For the readout layer of a standard network, $W_0 \in \mathbb{R}^{k\times n}$ maps the diverging width n to a fixed output dimension k. The perturbation $\Delta x \in \mathbb{R}^n$ entering the layer has been shaped by backpropagation through $W_0^\top$, so its entries acquire systematic alignment with the columns of W_0. The inner products $\langle (W_0)_{j\cdot}, \Delta x\rangle$ no longer cancel, and the LLN-like prefactor is realized in the diverging dimension being summed over (in the large-width regime $n \gg k$ (Haas et al., 2025)):

$$\|W_0\,\Delta x\|_{\mathrm{RMS}} \asymp n\,\|W_0\|_{\mathrm{RMS}}\,\|\Delta x\|_{\mathrm{RMS}}.$$

MoE first expert layer (Regime II): alignment broken by expert independence. The MoE bottleneck regime (Regime II) is an instructive intermediate case. The first expert layer $W_0^{(i)} \in \mathbb{R}^{N_e\times n}$ maps the diverging width n to the finite expert hidden width N_e, geometrically the same situation as a readout layer, so alignment-induced LLN scaling is in principle permitted. Yet the alignment mechanism is broken here: the M expert first-layers $\{W_0^{(i)}\}_{i=1}^M$ are independent at initialization, and a single shared input perturbation Δx cannot simultaneously align with the column structure of every $W_0^{(i)}$. CLT-like scaling therefore survives,

$$\|W_0^{(i)}\,\Delta x\|_{\mathrm{RMS}} \asymp \sqrt{n}\,\|W_0^{(i)}\|_{\mathrm{RMS}}\,\|\Delta x\|_{\mathrm{RMS}}.$$

MoE second expert layer (Regime II): alignment is restored at finite dimension. The second expert layer $W_0^{(i,2)} \in \mathbb{R}^{n\times N_e}$ maps the finite hidden width N_e back to the diverging width n. Here the relevant sum runs over only N_e terms, so the LLN prefactor cannot exceed N_e regardless of alignment. The perturbation entering this layer is private to expert i and is propagated by training through its own $W_0^{(i,2)\top}$, so the alignment is realized:

$$\|W_0^{(i,2)}\,\Delta x\|_{\mathrm{RMS}} \asymp N_e\,\|W_0^{(i,2)}\|_{\mathrm{RMS}}\,\|\Delta x\|_{\mathrm{RMS}} \asymp \|W_0^{(i,2)}\|_{\mathrm{RMS}}\,\|\Delta x\|_{\mathrm{RMS}} \quad \text{since } N_e \in \Theta(1).$$

LLN-like in the fixed dimension being summed over.

H.2.3 Effective updates: rank-1 alignment

The SGD update $\Delta W = -\eta\,(\nabla_{x'} L)\,x^\top$ is rank-1, so applying it to x saturates Cauchy-Schwarz: $\|\Delta W \cdot x\|_2 = \|\Delta W\|_F\,\|x\|_2$, equivalently

$$\|\Delta W \cdot x\|_{\mathrm{RMS}} \asymp n\,\|\Delta W\|_{\mathrm{RMS}}\,\|x\|_{\mathrm{RMS}},$$

where n is the layer's fan-in. The prefactor n (vs. the $\sqrt{n}$ of CLT) reflects the rank-1 alignment between ΔW and x due to gradient descent, which precludes any mean-zero cancellation. When the fan-in is fixed (as for the second expert layer in Regime II, with $N_e = \Theta(1)$), the prefactor is $\Theta(1)$ rather than diverging. By transpose, the backward effective piece $(\Delta W)^\top \delta$ satisfies the same alignment relation with fan-out replacing fan-in. Together with the gradient scale at the layer, the rank-1 relation places a constraint on the SGD learning rate at every layer in any regime; the gradient scale is the only missing input to instantiate it.

In contrast, for Adam, entrywise normalization sets ΔW entries to $\Theta(\eta_{\mathrm{Adam}})$ regardless of the gradient scale, so the alignment argument alone constrains η_{Adam}: the input-side LLN factor gives $\eta_{\mathrm{Adam}} = 1/n$ when the input dimension diverges and $\eta_{\mathrm{Adam}} = 1$ when it is fixed. Setting Adam's ε, however, still requires the gradient scale: ε must remain Θ of the typical per-coordinate gradient magnitude (else either ε dominates and the update becomes trivial, or the gradients dominate and the normalization becomes vacuous), so it scales with the gradient at the layer.

Extension to Adam and Muon. Initialization and aggregation exponents carry over from SGD: they are fixed by forward stability at initialization and by the propagating-update and aggregation conditions, which involve the updates only through their $\Theta(1)$ effect on the features and their rank-one alignment with the layer input, not through the update rule. For Adam, this leaves the learning rate $\eta \propto 1/\text{fan-in}$ and ϵ at the layerwise gradient RMS derived above, so that ϵ's effect on the update is independent of scale (Desideratum μ-3); the gradient-RMS exponents are those of our SGD signal propagation analysis (App. J). Muon replaces each matrix gradient by its orthogonalization, rescaled by $\sqrt{\text{fan-out}/\text{fan-in}}$, so every update satisfies the spectral condition for feature learning independently of the gradient scale (Yang et al., 2023; Pethick et al., 2025). The M^{-1} factor that expert and router gradients inherit from the aggregation (App. B) therefore does not enter, and no ϵ is needed. What remains MoE-specific is the initialization and the aggregation, which we take from μP or MSSP (Table 1; Table B.2 for MSSP).

Together with the propagating analysis above, this yields per-layer constraints relating σ, η (and ε for Adam) to the layerwise gradient scale, which is in turn determined by the cross-expert aggregations of Section H.7. Our goal is a parameterization that simultaneously satisfies these per-layer constraints together with the cross-expert balance constraints below; trade-offs between competing constraints can in principle require expanding the hyperparameter space (e.g., layer-specific scalings) to admit a feasible solution.

Crucially, the μP desideratum (that the layerwise effective and propagating terms in the *forward* pass remain $\Theta(1)$) already yields a parameterization: the per-layer initialization variances, learning rates, and (for Adam) ε. We refer to this parameterization as the *μP baseline*. The alignment-exponent framework (H.3) furnishes the analytical machinery for solving the corresponding $\Theta(1)$ conditions. The μP desideratum, however, imposes no explicit constraints on the backward pass or on the cross-expert aggregations (H.4), (H.6); as discussed in §H.1.1, balance across the terms of these decompositions constitutes the additional desiderata required of an MoE parameterization. In what follows, we proceed case by case through the μP baseline, identifying where it produces imbalances under our extended desiderata and showing how MSSP resolves each.

H.3 MSSP Desiderata

For MoE blocks, the analogous requirement applies to each term in the effective/propagating decomposition of the aggregated representation. Writing $W^{3,i}(t) = W^{3,i}(0) + \Delta W^{3,i}(t)$ and $h^{2,i}(t) = h^{2,i}(0) + \Delta h^{2,i}(t)$, the aggregation $h^{\mathrm{agg}}(t) = \sum_{i=1}^{M} \phi_i(t)\, W^{3,i}(t)\, h^{2,i}(t)$ decomposes as

$$h^{\mathrm{agg}}(t) = \underbrace{\sum_{i=1}^{M} \phi_i\, W^{3,i}(0)\, h^{2,i}(0)}_{\text{A1: init.}} + \underbrace{\sum_{i=1}^{M} \phi_i\, W^{3,i}(0)\, \Delta h^{2,i}(t)}_{\text{A2: propagating}} + \underbrace{\sum_{i=1}^{M} \phi_i\, \Delta W^{3,i}(t)\, h^{2,i}(t)}_{\text{A3: effective}}. \tag{H.4}$$

The MoE-specific extension of Desideratum 1 requires each of these three contributions (init., propagating, effective) to remain $\Theta(1)$ in RMS at some $t > 0$. Expanding the propagating term one step further (substituting $\Delta h^{2,i}(t) \approx W^{2,i}(0)\, \Delta h^1 + \Delta W^{2,i}(t)\, h^1(0)$) yields the four-term decomposition $h^{\mathrm{agg}}(t) = A_1 + A_{2,1} + A'_{2,2} + A_3$ which will be used in the case studies below.

Backward desiderata. Recall that the same structure carries over to the backward pass. Along any linear map $h_t^\ell = W_t^\ell x_t^{\ell-1}$, writing $\bar{\delta}_t^\ell := \nabla_{x_t^\ell}\mathcal{L}$ and $\delta_t^\ell := \nabla_{h_t^\ell}\mathcal{L}$, the transpose recursion $\bar{\delta}_t^{\ell-1} = (W_t^\ell)^\top \delta_t^\ell$ admits the analogous effective/propagating split

$$\bar{\delta}_t^{\ell-1} \;=\; \underbrace{(W_0^\ell)^\top \delta_t^\ell}_{\text{propagating}} \;+\; \underbrace{(\Delta W_t^\ell)^\top \delta_t^\ell}_{\text{effective}}. \tag{H.5}$$

When backpropagation crosses a representation that fans out into M expert branches, the gradient at the shared node aggregates expert contributions in the same form as the forward aggregation. Writing $g^{2,i}(t) := \nabla_{h^{2,i}(t)}\mathcal{L}$ for the per-expert gradient at the bottleneck representation, the expert-pathway contribution to the gradient at the input shared across experts is $\bar{\delta}^{1,\exp}(t) = \sum_{i=1}^M \phi_i(t)\,(W^{2,i}(t))^\top g^{2,i}(t)$, and expanding $W^{2,i}(t) = W^{2,i}(0) + \Delta W^{2,i}(t)$ and $g^{2,i}(t) = g^{2,i}(0) + \Delta g^{2,i}(t)$ gives

$$\bar{\delta}^{1,\exp}(t) \;=\; \underbrace{\sum_{i=1}^M (W^{2,i}(0))^\top g^{2,i}(0)}_{\text{A4: init.}} + \underbrace{\sum_{i=1}^M (W^{2,i}(0))^\top \Delta g^{2,i}(t)}_{\text{A5: propagating}} + \underbrace{\sum_{i=1}^M (\Delta W^{2,i}(t))^\top g^{2,i}(t)}_{\text{A6: effective}} \tag{H.6}$$

Desideratum 1 extends naturally to the backward pass: each piece of the backward decomposition (the layerwise effective/propagating split (H.5) and each term in the expert-aggregation decomposition (H.6)) is required to be *balanced*, i.e. to share a common scale within each decomposition. (Under an appropriate per-layer normalization that places backward signals on a scale comparable to forward signals, the common scale is $\Theta(1)$; since this normalization depends on architecture, algorithm, and layer type, we work with the normalization-free *balance* statement throughout, and omit the explicit normalization.)

H.4 Backward scaling

The forward arguments above transpose directly to the backward pass at every layer with diverging input and output dimensions: $(W_0^\ell)^\top \delta$ inherits the same CLT-like scaling as $W_0^\ell\,\Delta x$, and $(\Delta W^\ell)^\top \delta$ inherits the same rank-1 alignment as $\Delta W^\ell \cdot x$ (with fan-out replacing fan-in). We do not repeat these case studies.

H.5 Aggregation layers

In MoEs, aggregations across experts arise in both the forward pass (Eq. (H.4)) and the backward pass (Eq. (H.6)). The aggregation operator introduces sums over *expert-indexed* contributions (weighted by routing), and the leading-order scale of these sums depends on the evolving cross-expert correlation structure. In particular, expert sums can exhibit CLT-type scaling (weak cross-expert correlations), LLN-type scaling (strongly correlated contributions), or mixed behaviour across different contributions; our analysis makes these distinctions regime-by-regime.

H.6 Width dependence in μP and how MSSP fixes it

Below, we provide some case studies on the key imbalances in MoE dynamics under μP and follow this by how MSSP fixes them. All computations use $\alpha = 1$.

H.6.1 Expert hidden layer gradient $\partial f/\partial h^{2,i}$

Consider the four-piece decomposition

$$\partial f/\partial h^{2,i} \;\propto\; (W_0^{3,i} + \Delta W^{3,i})^\top (W_0^4 + \Delta W^4)^\top.$$

In Regime III, $(W_0^{3,i})^\top (\Delta W^4)^\top$ admits CLT-like scaling, since $W_0^{3,i}$ has diverging input and output dimensions. In Regime II, $W_0^{3,i}$ is a low-rank matrix and ΔW^4 is shaped by $W_0^{3,i}$ through backpropagation, so LLN-like scaling is in principle possible. However, ΔW^4 is built from sums over all $M \to \infty$ experts' $W_0^{3,i}$, and hence cannot simultaneously align with any single $W_0^{3,i}$; the term therefore also admits CLT-like scaling.

$$\|(W_0^{3,i})^\top(\Delta W^4)^\top\|_{\mathrm{RMS}} = \sqrt{N}\,\|(W_0^{3,i})^\top\|_{\mathrm{RMS}}\,\|(\Delta W^4)^\top\|_{\mathrm{RMS}} \asymp \sqrt{N}\cdot 1/\sqrt{N_e}\cdot 1/N \quad \text{under } \mu\text{P}.$$

Now consider the term $(\Delta W^{3,i})^\top(W_0^4)^\top$. As discussed in the effective-updates section, since the updates to the second expert-layer weights are rank-1 in structure and all experts share the same gradient direction, the rank-1 alignment argument applies, yielding LLN-like scaling:

$$\|(\Delta W^{3,i})^\top(W_0^4)^\top\|_{\mathrm{RMS}} = N\,\|(\Delta W^{3,i})^\top\|_{\mathrm{RMS}}\,\|(W_0^4)^\top\|_{\mathrm{RMS}} \asymp N\cdot 1/N_e\cdot 1/N \quad \text{under } \mu\text{P}.$$

The remaining two contractions in the four-piece expansion follow the same scaling laws: $(W_0^{3,i})^\top(W_0^4)^\top$ behaves identically to the CLT-like term analysed above, and $(\Delta W^{3,i})^\top(\Delta W^4)^\top$ behaves identically to the rank-1 term.

In Regime III we have $N_e \asymp N$, so both terms are of order $1/N$ and are therefore balanced (and, under the appropriate normalization, both terms are of order 1). In Regime II, $N_e \in \Theta(1)$, and the two terms have scales $1/\sqrt{N}$ and 1 respectively.

In summary, in Regime III, $|W_0^{3,i}| = \Theta(1/\sqrt{N})$ and $|\Delta W^{3,i}| = \Theta(1/N)$, leading to an entry-size gap of $\sqrt{N}$. This $\sqrt{N}$ exactly compensates the CLT-vs.-rank-1 alignment-type gap, so both contractions land at $\Theta(1/N)$: *balanced*. In Regime II, $|W_0^{3,i}| = \Theta(1)$ and $|\Delta W^{3,i}| = \Theta(1)$, so the entry sizes are equal (since $\sigma_3^2 = 1/N_e = \Theta(1)$). The CLT-vs.-rank-1 alignment-type gap is no longer cancelled, so the two pieces split by $\sqrt{N}$: *imbalanced*.

MSSP correction in Regime II. The MSSP-Regime-II boost $\sigma_3^2 = M/N_e$ increases the scale of $|W_0^{3,i}|$ from $\Theta(1)$ to $\Theta(\sqrt{M}) = \Theta(\sqrt{N})$, while the update $|\Delta W^{3,i}| = \Theta(1)$ is unchanged (the SGD learning rate $\eta_3 = \eta MN$ is identical to its μP-Regime-II value). The entry-size ratio $|\Delta W^{3,i}|/|W_0^{3,i}|$ thus moves from $\Theta(1)$ under μP to $\Theta(1/\sqrt{N})$ under MSSP, restoring the buffer to the same $\sqrt{N}$ value as in Regime III. The CLT-vs-rank-1 alignment-type gap is then cancelled exactly, and the four pieces of $\partial f/\partial h^{2,i}$ all sit at the leading $\Theta(1/N)$ scale.

H.7 Cross-expert aggregation: a taxonomy of mechanisms

We now turn to the MoE-specific cross-expert aggregations. Two such aggregations arise: the forward aggregation $h^{\mathrm{agg}}(t)$ of (H.4), with expanded four-term form $h^{\mathrm{agg}}(t) = A_1 + A_{2,1} + A'_{2,2} + A_3$; and the backward aggregation $\bar{\delta}^{1,\mathrm{exp}}(t)$ of (H.6), which collects the expert-pathway contributions to the gradient with respect to the input h^1 shared across experts, with an analogous decomposition. Each summand depends on the per-expert weights $W_0^{2,i}, W_0^{3,i}$ and their updates; the leading-order scale of the aggregate is determined by the cross-i correlation structure of these summands. The scaling of these aggregation terms is determined by the crude CLT–LLN dichotomy: summands sharing a coherent direction aggregate by LLN (no $1/\sqrt{M}$ suppression), whereas summands with i.i.d. random directions aggregate by CLT ($1/\sqrt{M}$ suppression). The structurally interesting question, however, is: *What generates a coherent direction in the first place, or prevents one from emerging?* The four mechanisms identified below (A–D) classify the case studies that follow according to the algebraic feature that controls this question, and we examine each in turn under the μP baseline and under MSSP. For each mechanism we (i) state the algebraic feature and case-study term it governs, (ii) state its scaling under the μP baseline, and (iii) close with a paragraph describing the corresponding behaviour under MSSP.

H.7.1 Mechanism A: cross-expert CLT under independent or weakly correlated summands

Mechanism A governs aggregates $A := \frac{1}{M}\sum_i \phi_i\, W_0^{3,i}\, W_0^{2,i}\, v$ with v shared across experts. The init and init-propagating forward aggregation terms

$$A_1 = \sum\nolimits_{i=1}^{M} \phi_i\, W^{3,i}(0)\, h^{2,i}(0), \qquad A_{2,1} = \sum\nolimits_{i=1}^{M} \phi_i\, W^{3,i}(0) W^{2,i}(0)\, \Delta h^1,$$

are of this form with $v = h^1$ and $v = \Delta h^1$ respectively. In both cases a cross-expert central limit theorem applies, suppressing the aggregate by $1/\sqrt{M}$ relative to the per-summand scale.

Independent case (A_1). Take $v \in \mathbb{R}^N$ independent of all expert weights. With $W_0^{2,i} \in \mathbb{R}^{N_e \times N}$ Gaussian (entries $\mathcal{N}(0, \sigma_2^2)$),

$$W_0^{2,i}\, v \;\sim\; \mathcal{N}\big(0,\, \sigma_2^2\, \|v\|^2\, I_{N_e}\big),$$

isotropic on $\mathbb{R}^{N_e}$ and depending on v only through $\|v\|^2$. Conditional on $W_0^{2,i}\, v$,

$$u_i := W_0^{3,i}\, W_0^{2,i}\, v \;\big|\; W_0^{2,i}\, v \;\sim\; \mathcal{N}\big(0,\, \sigma_3^2\, \|W_0^{2,i}\, v\|^2\, I_N\big).$$

Marginalising over $W_0^{2,i}$ preserves rotation invariance: the distribution of u_i depends on v only through $\|v\|^2$. Across experts, $(W_0^{2,i}, W_0^{3,i})_{i=1}^M$ are i.i.d., so the only source of cross-expert dependence among $\{u_i\}$ is the shared $\|v\|^2$.

The scalar $\|v\|^2 = \Theta(N)$ does not itself converge. Combined with the $1/M$ aggregation prefactor and $M \asymp N$, it appears in the aggregate variance $\mathbb{E}\|A\|^2 \asymp M^{-1}\, \mathbb{E}\|u_i\|^2$ as $\|v\|^2/M \asymp \|v\|^2/N = \|v\|^2_{\mathrm{RMS}}$, which converges deterministically under standard initialisation. The summands are asymptotically independent at the aggregate level, and a multivariate CLT applies.

Weakly dependent case ($A_{2,1}$). Δh^1 is shaped by gradient backflow through every expert and depends on each $W_0^{2,i}$, so strict independence fails. A single vector in $\mathbb{R}^N$ cannot, however, align nontrivially with $M \to \infty$ independent Gaussian matrices simultaneously: the dependence of Δh^1 on any one $W_0^{2,i}$ enters through a single summand of the cross-expert loss gradient and is of relative order $1/M$. Decompose $\Delta h^1 = \Delta h^{1,(\neq i)} + \Delta h^{1,(i)}$ with $\Delta h^{1,(\neq i)}$ independent of $W_0^{2,i}$ and $\|\Delta h^{1,(i)}\|/\|\Delta h^1\| = O(1/M)$; the dominant term $W_0^{3,i}\, W_0^{2,i}\, \Delta h^{1,(\neq i)}$ is governed by the previous paragraph, the residual by a $1/\sqrt{M}$-suppressed correction.

Per-summand and aggregate magnitudes. For init and propagating aggregation terms $A \in \{A_1, A_{2,1}\}$ in both Regime II (N_e fixed) and Regime III ($N_e \asymp N$), we arrive at the per-expert and aggregate scalings,

$$\mathbb{E}\|u_i\|^2 \;=\; N\sigma_3^2 \cdot N_e\sigma_2^2\, \|v\|^2 \;=\; \sigma_2^2\, \sigma_3^2\, N\, N_e\, \|v\|^2,$$

$$\|u_i\|_{\mathrm{RMS}} \;\asymp\; (\sigma_2\sqrt{N})(\sigma_3\sqrt{N_e})\, \|v\|_{\mathrm{RMS}}, \qquad \|A\|_{\mathrm{RMS}} \;\asymp\; \frac{\sigma_2\, \sigma_3\, \sqrt{N\, N_e}}{\sqrt{M}}\, \|v\|_{\mathrm{RMS}}.$$

Under μP, $\sigma_2 = 1/\sqrt{N}$ and $\sigma_3 = 1/\sqrt{N_e}$, so $\|A_1\|_{\mathrm{RMS}}, \|A_{2,1}\|_{\mathrm{RMS}}$ are of order $1/\sqrt{M}$, strictly subleading to $A'_{2,2}$ and A_3 at $\Theta(1)$.

MSSP correction. The two MSSP fixes resolve this in regime-specific ways. In Regime III, MSSP shares expert weights at initialization ($W_0^{2,i} = W_0^2$ and $W_0^{3,i} = W_0^3$ for all i); the per-expert summands then become strongly correlated across i rather than independent, the cross-expert CLT collapses to an LLN, and both terms are lifted to $\Theta(1)$ along the shared direction $W_0^3 W_0^2\, h^1$ (or its Δh^1 analogue). In Regime II, MSSP boosts σ_3 from $1/\sqrt{N_e}$ to $\sqrt{M/N_e}$; the per-summand RMS scale is amplified by $\sqrt{M}$, exactly compensating the $1/\sqrt{M}$ cross-expert CLT suppression and lifting both terms to $\Theta(1)$ in expert-specific directions.

H.7.2 Mechanism B: Gram operators on a shared vector

We will use the forward propagating-update aggregation term

$$A'_{2,2} \;=\; \frac{1}{M} \sum_i \phi_{i,2}\, W_0^{3,i}\, \Delta_t W^{2,i}\, h_0^1$$

to illustrate this mechanism. Each summand has the form $G_i u$, where $G_i = W_0^{3,i}(W_0^{3,i})^\top$ acts on a vector u that is shared across experts (no i index). The cross-expert direction of $G_i u$ is then governed by whether G_i is full-rank (so $G_i \approx N_e\sigma_3^2 \cdot I$ and $G_i u$ inherits the shared direction u) or rank-deficient (so G_i is, after rescaling, a projection onto a random N_e-dimensional subspace of $\mathbb{R}^N$, distinct across experts). The same rank threshold separates Regime III ($N_e \asymp N$, isotropic, LLN) from Regime II ($N_e = \Theta(1)$, anisotropic, CLT).

Intuitive scaling of $A'_{2,2}$ across regimes. Substituting the rank-1 update $\Delta_1 W^{2,i} \propto (W_0^{3,i})^\top (W_0^4)^\top (h_0^1)^\top$ and using $\Delta_1 W^{2,i} h_0^1 \propto \|h_0^1\|^2 (W_0^{3,i})^\top (W_0^4)^\top$, each summand of $A'_{2,2}$ reduces to

$$W_0^{3,i} \Delta_1 W^{2,i} h_0^1 \;\propto\; \|h_0^1\|^2 \underbrace{W_0^{3,i} (W_0^{3,i})^\top}_{=: G_i} (W_0^4)^\top.$$

Writing $u := (W_0^4)^\top \in \mathbb{R}^N$, the analysis splits into a within-expert step (the action of G_i on the fixed vector u) and a cross-expert step (the aggregation $(1/M)\sum_i$).

Within-expert structure (fix i). The matrix $W_0^{3,i} \in \mathbb{R}^{N \times N_e}$ has rank $\min(N, N_e)$ generically, and so does G_i. The image of G_i is the column space of $W_0^{3,i}$ in $\mathbb{R}^N$; let Π_i denote the orthogonal projection onto it. By rotational invariance of the entries of $W_0^{3,i}$, the random subspace $\mathrm{col}(W_0^{3,i})$ is Haar-distributed on the Grassmannian $\mathrm{Gr}(\min(N, N_e),\, N)$.

The non-zero eigenvalues of G_i are the squared singular values of $W_0^{3,i}$, which concentrate around $N\sigma_3^2$ by Marchenko–Pastur (Marchenko and Pastur, 1967). Consequently G_i admits the operator-level approximation

$$G_i \;\approx\; (N\sigma_3^2)\, \Pi_i,$$

giving $G_i u \approx (N\sigma_3^2)\, \Pi_i u$. Since u is independent of $W_0^{3,i}$ and $\mathrm{col}(W_0^{3,i})$ is Haar-distributed[3], each projection retains

$$\|\Pi_i u\|_2^2 \;\asymp\; \frac{\min(N, N_e)}{N} \|u\|_2^2.$$

Cross-expert structure. The matrices $\{W_0^{3,i}\}_{i=1}^M$ are i.i.d., so the projections $\{\Pi_i u\}_{i=1}^M$ are independent across experts, each marginally Haar-distributed in direction within $\mathbb{R}^N$.

Regime III ($N_e \asymp N$): isotropic Gram matrix preserves LLN scaling. For each fixed i, since $W_0^{3,i}$ is generically full rank at leading order, so $\Pi_i \approx I_N$ and

$$G_i\, u \;\approx\; (N_e\sigma_3^2)\, u \;=\; \Theta(1) \cdot u$$

(taking $\sigma_3^2 = 1/N_e$). The Gram acts as a near-scalar multiple of the identity, and the result lies along $u = (W_0^4)^\top$ (shared direction across all experts). Cross-expert aggregation admits a law of large numbers like scaling:

$$\|A'_{2,2}\|_{\mathrm{RMS}} \;\asymp\; \|h_0^1\|^2 \cdot N_e\sigma_3^2 \cdot \|(W_0^4)^\top\|_{\mathrm{RMS}} \;=\; \Theta(1) \text{ along } (W_0^4)^\top \quad \text{under } \mu\mathrm{P}$$

Regime II (N_e fixed, $N \to \infty$): anisotropic Gram yields CLT scaling. For each fixed i, the column span $\mathrm{col}(W_0^{3,i})$ is a fixed-dimensional random subspace within a diverging ambient space, on which G_i has eigenvalue $N\sigma_3^2 = N/N_e$ under μP. By rotational invariance, the Haar distribution of this subspace implies $\|\Pi_i u\|_2^2 \asymp (N_e/N)\, \|u\|_2^2$: each orthonormal basis vector of $\mathrm{col}(W_0^{3,i})$ is uniform on S^{N-1} by rotation invariance, contributing $\|u\|^2/N$ in expectation, and there are N_e such basis vectors. Therefore,

$$\|G_i u\|_2 \;\asymp\; (N\sigma_3^2)\sqrt{N_e/N}\, \|u\|_2 \;=\; \sigma_3^2 \sqrt{N N_e}\, \|u\|_2,$$

with $G_i u$ lying entirely in the expert-specific subspace $\mathrm{col}(W_0^{3,i})$. Across experts, these subspaces are independent random elements of the Grassmannian, so the directions $\{G_i u / \|G_i u\|_2\}_{i=1}^M$ are independent and there is no shared coherent direction along which $\{G_i u\}_i$ can accumulate. Cross-expert aggregation therefore proceeds by the central limit theorem rather than the law of large numbers: $A'_{2,2}$ is suppressed by an additional $1/\sqrt{M}$ relative to the per-summand scale, and is not aligned with $(W_0^4)^\top$.

[3]The same argument holds under weak correlations.

MSSP correction. The Regime II imbalance is removed under MSSP-Regime-II by the amplification $\sigma_3^2 = M/N_e$, which raises the eigenvalue of G_i on its rank-N_e column span from $N\sigma_3^2 = N/N_e$ under μP to $N\sigma_3^2 = NM/N_e$ under MSSP. The per-summand scale $\|G_i u\|_2 \asymp \sigma_3^2\sqrt{NN_e}\,\|u\|_2$ is thereby amplified by exactly $\sqrt{M}$, and the cross-expert $1/\sqrt{M}$ CLT cancellation is exactly absorbed; $A'_{2,2}$ then sits at $\Theta(1)$, in expert-specific directions rather than along $(W_0^4)^\top$. In Regime III, $A'_{2,2}$ is already $\Theta(1)$ under μP (the Gram is isotropic and aligned along u); MSSP-Regime-III's shared experts collapse the marginal randomness of $\{\Pi_i\}$ entirely, but do not change the Θ-class.

H.7.3 Mechanism C: Expert sums of effective contributions

To illustrate this mechanism, we use the forward effective update aggregation term

$$A_3 = \sum\nolimits_{i=1}^{M} \phi_i\, \Delta W^{3,i}(t)\, h^{2,i}(t)$$

Each summand has the form $\Delta W^{3,i}\, x_i$ with $\Delta W^{3,i}$ rank-1 of the form $u_{\text{shared}}\, v_i^\top$, where u_{shared} is a backpropagated factor shared across experts and v_i is expert-specific. The product $\Delta W^{3,i}\, x_i = \langle v_i, x_i\rangle\, u_{\text{shared}}$ then writes the summand as a (per-expert) scalar times a shared direction. Cross-expert aggregation reduces to a scalar mean, and the LLN survives whenever that scalar mean is dominated by a strictly positive contribution. Unlike Mechanism B, the shared direction is forced by the rank-1 structure of the gradient itself rather than recovered from a Gram acting on a fixed vector, so this mechanism is regime-insensitive.

Scaling A_3. Recall that

$$A_3 \;=\; \frac{1}{M}\sum_i \phi_{i,t}\, \Delta_t W^{3,i}\, h_t^{2,i},$$

where $\Delta_t W = \sum_{t=1}^{T} \delta_t W$ and under the SGD update rule,

$$\delta_t W^{3,i}\, h_t^{2,i} \;=\; -\frac{\eta_3\, \chi\, \phi_{i,t-1}}{M}\, \langle h_{t-1}^{2,i},\, h_t^{2,i}\rangle\, (W_{t-1}^4)^\top.$$

Therefore, the per-expert summand is a single inner product times the shared readout direction. Decomposing the inner product,

$$\langle h_{t-1}^{2,i},\, h_t^{2,i}\rangle \;=\; \|h_{t-1}^{2,i}\|_2^2 \;+\; \langle h_{t-1}^{2,i},\, \Delta_t h^{2,i}\rangle.$$

The first term, $\|h_{t-1}^{2,i}\|_2^2$, is strictly positive for every expert (due to expert activation stability). The second term is a smaller fluctuation around it: at $t = 1$ it is random across i, while at $t = 2$ it becomes coherent via the alignment chain. In either case the sum is dominated by the strictly positive $\|h_{t-1}^{2,i}\|_2^2$, so the cross-expert scalar average is LLN-preserved:

$$\frac{1}{M}\sum_i \phi_{i,t}\, \phi_{i,t-1}\, \langle h_{t-1}^{2,i},\, h_t^{2,i}\rangle \;\asymp\; \|h^{2,i}\|_2^2.$$

Substituting back, using that $\|h^{2,i}\|_2^2 = \Theta(\|h^{2,1}\|_2^2)$,

$$A_3 \;\asymp\; -\frac{\eta_3\, \chi}{M}\, \|h^{2,1}\|_2^2\, (W_{t-1}^4)^\top$$

along the shared readout direction and in $\Theta(1)$ in Regime II under both μP and MSSP. In Regime III, the same argument shows that these terms are $\Theta(1)$ under μP. Under MSSP (which uses shared experts), similar arguments show that these terms are $\Theta(1)$ as well.

H.7.4 Mechanism D: Scalar self-pairing structural invariance

Certain quantities in the aggregation are built from summands that comprise scalar self-pairings $u^\top G_i u$ rather than a vector $G_i u$. The expectation $\mathbb{E}[u^\top G_i u] = \|u\|_2^2$ depends only on the diagonal mean $\mathbb{E}[(G_i)_{aa}] = N_e\sigma_3^2$, which is $\Theta(1)$ in both Regimes II and III under μP (and is therefore regime-symmetric). The per-layer parameterization, however, is calibrated against the (regime-asymmetric) vector form, so when the structure happens to be scalar the calibration finds nothing to act on and passes through unattenuated, leading to width dependence.

Mechanism: The structural mechanism underlying the imbalance of A_6. The following term A_6,

$$A_6 = (1/M) \sum_i \phi_{i,t} \, (\Delta W^{2,i})^\top \, (W_0^{3,i})^\top \, (W_t^4)^\top$$

arises in the expansion of the backward expert-pathway gradient $\bar{\delta}^{1,\exp}(t)$ of (H.6). A_6 captures the contribution in which $\Delta W^{2,i}$ is the perturbed expert factor while $W^{3,i}$ remains at initialization. Substituting $\Delta W^{2,i}$ into A_6 produces the scalar self-pairing

$$S_i \;:=\; u^\top G_i \, u \;=\; \big\| (W_0^{3,i})^\top u \big\|_2^2, \qquad u := (W_0^4)^\top, \;\; G_i := W_0^{3,i} (W_0^{3,i})^\top.$$

Since $(G_i)_{ab} = \sum_{k=1}^{N_e} (W_0^{3,i})_{ak} (W_0^{3,i})_{bk}$ with entries i.i.d. $\mathcal{N}(0, \sigma_3^2)$,

$$\mathbb{E}[(G_i)_{ab}] \;=\; N_e \sigma_3^2 \, \delta_{ab} \;=\; \delta_{ab} \qquad (\sigma_3^2 = 1/N_e),$$

hence

$$\mathbb{E}[S_i] \;=\; \|u\|_2^2 \;=\; \Theta(1/N) \quad \text{in both regimes.}$$

The remaining terms of the decomposition of the aggregated gradient depend on $W_0^{3,i}$ only through linear, vector-form quantities, which are regime-asymmetric by a factor of $\sqrt{N}$ in per-coordinate magnitude and which are accounted for by the parameterization. Due to the regime symmetry of A_6, the parameterization therefore passes through it unattenuated, yielding *imbalanced* scaling in Regime II.

$$A_6 \;\in\; \Theta(1/N^2) \;\text{ in Regime II,} \qquad A_6 \;\in\; \Theta(1/N) \;\text{ in Regime III.}$$

MSSP correction. The MSSP-Regime-II boost $\sigma_3^2 = M/N_e$ promotes the diagonal mean to $\mathbb{E}[(G_i)_{aa}] = N_e \sigma_3^2 = M = \Theta(N)$, so the scalar self-pairing satisfies $\mathbb{E}[S_i] = M\|u\|_2^2 = \Theta(N) \cdot \Theta(1/N) = \Theta(1)$. The $\Theta(N)$ amplification of the scalar exactly fills the $\sqrt{N} \cdot \sqrt{N}$ deficit that μP-Regime-II's vector-form-calibrated parameterization had been unable to address, and A_6 moves from $\Theta(1/N^2)$ to $\Theta(1/N)$, the leading scale in Regime II. The structural symmetry of $\mathbb{E}[G_i] \propto I$ that made A_6 inert under μP-Regime-II is exploited in reverse by MSSP: the boost is precisely tuned so that the regime-symmetric part of G_i becomes regime-asymmetric in the way the parameterization expects. In Regime III, A_6 is already at the leading scale under μP, and MSSP-Regime-III's shared experts do not change its Θ-class.

H.8 Summary: regime imbalances under μP and how MSSP resolves them

Regime II. In Regime II, MSSP introduces a single change to the μP baseline: the initialization variance of the second expert layer is increased to $\sigma_3^2 = M/N_e$. Although the change operates through a single parameter, it restores width independence in all four sources of imbalance present under μP in Regime II. The cross-expert CLT under independent or weakly correlated summands (Mechanism A) is corrected because the increased variance amplifies each per-summand by $\sqrt{M}$, exactly offsetting the $1/\sqrt{M}$ suppression of the cross-expert sum. The Gram operators on a shared vector (Mechanism B) and the scalar self-pairing structural invariance (Mechanism D) are corrected by the corresponding amplification of the Gram's eigenvalues and diagonal entries. The gradient at the expert hidden layer (§H.6.1) is corrected because the increased σ_3 raises $\|W_0^{3,i}\|$ to a scale that allows the four contractions of $\partial f / \partial h^{2,i}$ to balance. The same change to the parameterization therefore corrects all four structurally distinct sources of width dependence in Regime II simultaneously.

Regime III. In Regime III, MSSP introduces a different single change: expert weights are shared at initialization, so that all experts begin from a common pair (W_0^2, W_0^3). Under μP, the only mechanism that produces a width dependence in Regime III is the cross-expert CLT under independent or weakly correlated summands (Mechanism A). With shared experts, the per-expert summands are identical across i and accumulate by an LLN rather than cancelling by a CLT, and the affected terms are lifted to the leading scale. The remaining mechanisms (the Gram operators on a shared vector, the expert sums of effective contributions, the scalar self-pairing structural invariance, and the gradient at the expert hidden layer) are already balanced under μP in Regime III, and shared experts leave their leading behaviour unchanged.

The two fixes cannot be interchanged. The two fixes are matched to the two distinct sources of width dependence in their respective regimes, and cannot be interchanged. Applying the Regime-II

fix in Regime III would amplify the second expert layer's initial weights too aggressively, causing the backward gradient at the expert hidden layer to scale as $\sqrt{N}$ above its target; this is the very imbalance that the Regime-II fix was introduced to repair in Regime II. Conversely, sharing experts in Regime II would correct only Mechanism A: the Gram operators on a shared vector, the scalar self-pairing structural invariance, and the gradient at the expert hidden layer would all remain imbalanced, since shared experts do not affect the σ_3-dependent scales through which those mechanisms operate. Regime II requires the boosted variance; Regime III requires shared experts; each is the appropriate fix for its regime and would cause width dependence in the other.

H.9 Justification of MSSP in Regime II

We close the Regime II analysis by addressing a structural feature of the MSSP parameterization (Definition J.2) that distinguishes it from conventional feature-learning parameterizations: the per-expert hidden state $h^{3,i}$ (specifically the initial and propagating contributions $W_0^{3,i} h_0^{2,i}$ and $W_0^{3,i} \Delta h^{2,i}$) has coordinate scale $\Theta(\sqrt{N})$ vs the standard $\Theta(1)$ scale. We argue in this subsection that this is admissible because $h^{3,i}$ never functions as a primitive variable in the dynamics, and that MSSP-Regime-II achieves the correct $\Theta(1)$ scale for the decomposition of the aggregated activation h^3 in Regime II while confining the resulting amplification to a quantity that the dynamics avoid as a primitive.

The dynamics never realize $h^{3,i}$ as a primitive variable. We claim that, throughout both forward and backward passes, every place at which $h^{3,i}$ would arise in a natural decomposition of the dynamics admits a contraction-order rewriting that bypasses it. The forward aggregate is computable directly:

$$h^3 \;=\; \frac{1}{M} \sum_{i=1}^{M} \phi_i \, W^{3,i} \, h^{2,i}, \tag{H.7}$$

without forming any $h^{3,i}$ explicitly. The backward gradients factor through small-dimensional contractions:

$$\frac{\partial f}{\partial h^{2,i}} \;=\; \frac{\phi_i}{M} \, (W^{3,i})^\top (W^4)^\top, \tag{H.8}$$

$$\frac{\partial L}{\partial W^{3,i}} \;=\; \frac{\chi \, \phi_i}{M} \, (W^4)^\top (h^{2,i})^\top, \tag{H.9}$$

$$\frac{\partial f}{\partial \phi_i} \;=\; \frac{1}{M} \, (W^4 \, W^{3,i}) \, h^{2,i}. \tag{H.10}$$

The gating gradient (H.10) is the only quantity in which $h^{3,i}$ would appear under the natural decomposition $\partial f / \partial \phi_i = (1/M) \, (h^{3,i})^\top (W^4)^\top$. The contraction reordering on the right-hand side replaces the formation of $h^{3,i} \in \mathbb{R}^N$ with the inner-product chain $W^4 \, W^{3,i} \in \mathbb{R}^{1 \times N_e}$, an N_e-dimensional row vector. Because $N_e = \Theta(1)$, this object is small-dimensional and may be formed at numerical magnitude $\Theta(\sqrt{N})$ without storage or precision concerns.

The full forward and backward dynamics therefore close on the variable set

$$\mathcal{V} \;=\; \big\{ \, h^1, \; \psi, \; \phi, \; \{h^{2,i}\}_{i=1}^M, \; h^3, \; f, \; \{W^{2,i}\}_{i=1}^M, \; \{W^{3,i}\}_{i=1}^M, \; W^4, \; Q \, \big\}, \tag{H.11}$$

none of whose elements simultaneously carry both large coordinate magnitude ($\Theta(\sqrt{N})$) and large dimension ($\Theta(N)$). The amplified per-expert hidden state $h^{3,i}$ — the unique object in the architecture that combines both — is structurally absent from $\mathcal{V}$ as a primitive variable.

Mean-field interpretation. The structure above is the signature of a mean-field theory: per-particle quantities are amplified, the aggregate is the natural order parameter, and the closed dynamics involve only aggregated moments of per-particle quantities together with low-dimensional inputs. The DMFT effective fields associated with MSSP-Regime-II can be expressed entirely in terms of aggregated weight-product moments such as

$$\bar{G}_t \;=\; \frac{1}{M} \sum_{i=1}^{M} \phi_{i,t} \, W_t^{3,i} \, W_t^{2,i}, \qquad \bar{K}_t \;=\; \frac{1}{M} \sum_{i=1}^{M} \phi_{i,t}^2 \, (W_t^{2,i})^\top \, W_t^{2,i},$$

together with the per-expert layer-1 activations $\{h^{2,i}\}$, all of which sit at the standard $\Theta(1)$ coordinate scale. The per-expert hidden state $h^{3,i}$ does not appear in the effective theory.

H.10 Approaches We Tried That Do Not Restore Scale Stability

Before settling on the parameterization-level prescription that defines MSSP, we explored a number of architectural and parameterization-level interventions intended to remove the delayed-learning phenomenon and the underlying width-dependent subterms in MoE training under μP. None of the interventions enumerated below, applied on its own, restored clean refined-coordinate-check exponents while avoiding delayed learning at scale. We document them here because each is a natural intervention that practitioners may consider, and the reasons they fail clarify why the underlying problem is structural and why a parameterization-level fix is necessary.

H.10.1 RMSNorm (or LayerNorm) after the aggregation

A natural first response to the observation that $\|h^{\text{agg}}\|_{\text{RMS}}$ vanishes with width under μP-Regime-II is to insert a normalization layer after the aggregation, rescaling h^{agg} back to $\Theta(1)$. We tested this intervention with both RMSNorm and post-aggregation sigmoid and observed that neither restores width-independent dynamics. The reasons fall into four distinct points.

Norm layers preserve ratios, not balance. The post-aggregation activation h^{agg} decomposes into multiple subterms with distinct width scalings under μP-Regime-II — most notably the LLN-aligned effective contribution $A_3 \in \Theta(1)$ and the CLT-suppressed propagating contribution $A_{2,1} \in \Theta(1/\sqrt{N})$ (§J.3.2). RMSNorm divides the entire aggregate by its own RMS norm, rescaling every subterm by the same factor. The relative magnitudes of the subterms are preserved: after normalization, the post-norm activation is dominated by the effective contribution at the leading scale, while the propagating contribution sits at $\Theta(1/\sqrt{N})$ of the signal — a vanishing fraction. The propagating channel that should carry information from upstream-layer updates into the current block has been effectively erased.

The norm's ϵ has to be width-scaled, and the right scaling is exactly the imbalanced-subterm structure we are trying to avoid having to track. Practical RMSNorm computes $\tilde{h} = h/\sqrt{\|h\|^2 + \epsilon^2}$, where ϵ prevents over- and under-flow. With $\|h_0^{\text{agg}}\|^2 \in \Theta(1/N)$ in μP-Regime-II, there is a width-dependent crossover between two regimes: when $\|h\|^2 \ll \epsilon^2$ the norm acts as a constant rescaling h/ϵ (no normalization in the limit), while when $\|h\|^2 \gg \epsilon^2$ the norm acts as a full RMS normalization. A fixed ϵ therefore behaves qualitatively differently at small versus large widths, and the threshold between the two behaviors moves with N. To recover width-independent behavior, ϵ would need to be scaled in proportion to $\|h^{\text{agg}}\|$; but $\|h^{\text{agg}}\|$ in μP-Regime-II is itself a sum of subterms with mixed scaling exponents, and the dominant subterm changes during training. There is no single power-of-N schedule for ϵ that works across widths without first having solved the underlying scaling problem. The norm therefore moves the scaling burden into a different hyperparameter rather than eliminating it.

The backward pass through the norm introduces its own width-dependent rescaling. The Jacobian of RMSNorm scales as $1/\|h^{\text{agg}}\|$, so gradients flowing backward through the norm layer are inflated by that factor. At initialization in μP-Regime-II this factor is $\sqrt{N}$, and gradients upstream of the norm are amplified by a width-dependent multiplier whose magnitude also evolves over training as $\|h^{\text{agg}}\|$ climbs. The effective layerwise learning rate on every parameter upstream of the norm therefore becomes a function of both width and step number — the time-varying width-dependence that scale stability is designed to remove. So even granting the loss of the propagating channel discussed above and the hyperparameter burden of scaling ϵ, the backward pass through a normalization layer is itself a fresh source of instability rather than a fix for the existing one.

A post-aggregation norm cannot reach imbalances that live elsewhere in the dynamics. The imbalanced subterms in μP-Regime-II are not localized to h^{agg}. The four-piece expert-hidden-gradient buffer $\partial f/\partial h^{2,i}$ (§H.6.1) and the scalar self-pairing imbalance in A_6 (Mechanism D, §H.7.4) sit on the backward pathway and never pass through any post-h^{agg} operation. Even granting that a post-aggregation norm somehow fixed the forward aggregation, three of the four mechanisms identified in §H.7 would remain broken. A single normalization layer placed at one point in the architecture cannot address imbalances that arise across multiple structurally distinct points in the dynamics; this is one of the empirical reasons we converged on a parameterization-level fix rather than an architectural patch.

I Load balancing Approaches and their Scaling

We fix a layer l and a step t and drop both indices. For a batch $\mathcal{B}$ of tokens x, $\psi_i(x) = (Qh(x))_i$ denotes the router logits and $\phi_i(x) = \sigma_{\text{sig}}(\beta\,\psi_i(x))$ the routing scores.

Standard auxiliary loss. GShard/Switch Transformers (Fedus et al., 2022) add the auxiliary loss term

$$\mathcal{L}_{\text{aux}} = \lambda\, M \sum_{i=1}^{M} f_i P_i, \qquad f_i = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \mathbf{1}\{i \in \text{top-K}(\psi(x))\}, \qquad P_i = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \phi_i(x),$$

where f_i is the fraction of tokens routed to expert i and P_i its average routing score. They use softmax routing with $K = 1$, so $f_i = P_i = 1/M$ at perfect balance, and the factor M makes $\mathcal{L}_{\text{aux}} = \lambda$ independent of M.

The standard scaling is wrong when K grows. Every token selects exactly K experts, so $\sum_i f_i = K$ for every batch and $f_i = K/M$ under balance. Assuming $0 < c \le P_i \le C < 1$ under sigmoid routing, which holds as long as the router logits remain $\Theta(1)$ (Desideratum μ-1), we obtain $cK \le \sum_i f_i P_i \le CK$ and hence $\mathcal{L}_{\text{aux}} = \Theta(\lambda M K)$ along any path $M \asymp N^{\nu}$, $K \asymp N^{\kappa}$. Replacing the factor M by K^{-1} yields $\mathcal{L}_{\text{aux}} \in [\lambda c,\, \lambda C]$, which is scale independent. The gradients lead to the same factor. The cross-entropy gradient reaches the router row Q_i only through the fraction K/M of tokens routed to expert i, each with weight K^{-1} from the aggregation, so it is $\Theta(M^{-1})$. With the factor K^{-1}, the auxiliary gradient $\lambda K^{-1} f_i\, \partial P_i / \partial Q_i$ is $\Theta(\lambda M^{-1})$ as well, so the ratio of the two stays $\Theta(\lambda)$. Both arguments assume that every expert receives many tokens per batch, $|\mathcal{B}|K/M \gg 1$.

DeepSeek variant. DeepSeekMoE (Dai et al., 2024) and DeepSeek-V3 (DeepSeek-AI, 2024) normalize both factors,

$$\mathcal{L}_{\text{bal}} = \lambda \sum_{i=1}^{M} f'_i P'_i, \qquad f'_i = \frac{M}{K} f_i, \qquad P'_i = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \frac{\phi_i(x)}{\sum_{j=1}^{M} \phi_j(x)}.$$

Since $f'_i = 1$ under balance and $\sum_i P'_i = 1$, the sum $\sum_i f'_i P'_i$ is a P'-weighted average of the f'_i. It equals 1 at balance for all M and K, without any assumption on the routing scores. Under the bounds above $\sum_j \phi_j(x) = \Theta(M)$, so $\mathcal{L}_{\text{bal}} = \Theta\big(\lambda \frac{M}{K} M^{-1} \sum_i f_i P_i\big) = \Theta\big(\lambda K^{-1} \sum_i f_i P_i\big)$, which is exactly the K^{-1} scaling on every path.

Depth scaling. The auxiliary loss is summed over the MoE layers, so for fixed λ it grows as $\Theta(L)$, while the cross-entropy loss stays $\Theta(1)$. Under $\Theta(1/L)$ residual MoE blocks (Dey et al., 2025), the cross-entropy gradient reaches Q^l only through ϕ^l and is therefore $\Theta(1/L)$, whereas the l-th auxiliary term acts on the router logits directly and its gradient does not depend on depth. The weight should therefore be $\lambda M^{-1} L^{-1}$ for the standard loss and λL^{-1} for the DeepSeek loss. In other words, the auxiliary loss is averaged rather than summed over layers. The same applies to the router z-loss.

Quantile balancing (QB). QB (Su, 2026) replaces the auxiliary loss by per-expert biases b_i that enter only the expert selection (Wang et al., 2024; DeepSeek-AI, 2024). Token x selects top-K$(\psi(x) + b)$, while the routing scores $\phi_i(x)$ use the unbiased logits. The biases come from the dual of the balanced assignment problem $\max_{z \in \{0,1\}^{\mathcal{B} \times M}} \sum_{x,i} z_{x,i}\, \psi_i(x)$ subject to $\sum_i z_{x,i} = K$ and $\sum_x z_{x,i} = |\mathcal{B}|K/M$. After each step, one coordinate-descent step on the dual computes

$$\tau_x = (K{+}1)\text{-th largest entry of } \psi(x){+}b, \qquad \mu_i = \left\lfloor \tfrac{|\mathcal{B}|K}{M} \right\rfloor\text{-th largest entry of } \{\psi_i(x){+}b_i{-}\tau_x\}_{x \in \mathcal{B}},$$

and updates $b \leftarrow b - (\mu - \bar{\mu}\,\mathbf{1})$ with $\bar{\mu} = \frac{1}{M} \sum_i \mu_i$. An overloaded expert has $\mu_i > 0$ and is shifted down by the amount that would leave it exactly balanced. We compute μ per micro-batch and average it over micro-batches and ranks (William Held, 2026). Unlike the sign update of loss-free balancing (Wang et al., 2024), QB has neither a step size nor a loss weight, so no hyperparameter has to be rescaled with M or K. Empirically, we observe that QB reduces the number of starved experts fast, recovers a stable post-sigmoid router logit distribution across both model scales and token horizon, and improves performance over auxiliary loss setups (Appendix F.4).

J Signal Propagation Analysis

Here we provide a self-consistent forward and backward signal propagation analysis through an MoE block for the two more interesting Regimes II and III.

This analysis complements the DMFT treatment in the subsequent sections. We use the minimal MoE architecture defined below, which places the MoE block between an input and an output layer so that its incoming and outgoing signals carry the same asymptotic scaling as in a standard dense architecture. The arguments therefore extend to MoE blocks embedded in larger networks, such as a Transformer MoEs, whenever additional modules preserve this scaling. This is precisely what μP's layer-wise parametrization guarantees (Yang and Hu, 2021; Yang et al., 2022). This compositionality is verified by our Transformer MoE coordinate checks in Appendix F.5. For brevity, we omit the analysis for Regime I, which follows the known arguments for standard architectures.

J.1 Preliminaries

Minimal MoE architecture. We work with the minimal MoE architecture

$$h^1 = W^1 x, \tag{J.1}$$
$$\psi = Q h^1, \tag{J.2}$$
$$\phi = \sigma(\psi), \tag{J.3}$$
$$h^{2,i} = W^{2,i} h^1, \qquad i = 1, \dots, M, \tag{J.4}$$
$$h^{3,i} = W^{3,i} h^{2,i}, \tag{J.5}$$
$$h^3 = \frac{1}{M} \sum_{i=1}^{M} \phi_i \, h^{3,i}, \tag{J.6}$$
$$f = W^4 h^3, \tag{J.7}$$

where $x \in \mathbb{R}^D$ is the input, $f \in \mathbb{R}$ the scalar output, and the trainable parameters are

$$\theta \;=\; \big(W^1,\; Q,\; \{W^{2,i},\, W^{3,i}\}_{i=1}^{M},\; W^4\big)$$

with shapes $W^1 \in \mathbb{R}^{N \times D}$, $Q \in \mathbb{R}^{M \times N}$, $W^{2,i} \in \mathbb{R}^{N_e \times N}$, $W^{3,i} \in \mathbb{R}^{N \times N_e}$, $W^4 \in \mathbb{R}^{1 \times N}$. Here N is the hidden width, N_e the expert hidden width, M the number of experts, and D the input dimension. The derivations below are written for *sigmoid* gating, $\sigma(\psi)_i = (1 + e^{-\psi_i})^{-1}$, paired with the explicit $1/M$ aggregation factor in (J.1); the same scaling exponents hold under softmax gating without an explicit aggregation multiplier.

Scaling regimes. As described in the main paper, we analyze three asymptotic regimes, distinguished by which of the width parameters N, N_e, M diverge:

- **Regime I:** ($N, N_e \asymp n \to \infty$ with $M, K = \Theta(1)$),
- **Regime II:** ($N, M, K \asymp n \to \infty$ with $N_e = \Theta(1)$),
- **Regime III:** ($N, N_e, M, K \asymp n \to \infty$).

The heuristic derivation for Regimes II and III is provided in Sections J.3, and J.4 respectively.

In each of these regimes, we want to understand the scale of all quantities that arise in the forward and backward computations in the network, both at initialization and during training. While the initial random parameters are independent of the data distribution, correlations develop over time since updates carry information about the data; this is the case for both SGD and Adam (Yang et al., 2022; Yang and Littwin, 2023; Everett et al., 2024). To derive the correct scaling it is crucial to consider the correlations that arise during training and how these correlations propagate forward and backward through the network. In general, after one step of SGD the full set of post-update correlations is already present(Yang and Hu, 2021; Yang et al., 2022; Bordelon and Pehlevan, 2022; Everett et al., 2024), so analysing the first two forward and backward passes suffices to determine the asymptotic width scaling of every quantity arising in the computation[4].

[4]When $W_{L+1}^{(0)} = 0$, only the readout updates in the first SGD step, since the gradient into every non-readout parameter carries a factor of $W_{L+1}^{\top}$ that vanishes at initialization. A third forward and backward pass at $\theta^{(2)}$ is then needed, since the non-readout parameters first move at $t = 2$.

J.2 Standing assumptions and notation

We track leading-order coordinate-scale exponents only, ignoring $\Theta(1)$ prefactors.

Standing assumptions.

(B1) $\chi_t = \partial L / \partial f_t \in \Theta(1)$ at every step t; absorbed into $\Theta(1)$ scalar prefactors throughout.

(B2) Nonlinearities preserve leading scaling exponents. For sigmoid gating with $1/M$ aggregation, $\phi_{i,t},\ \dot{\phi}_{i,t} \in \Theta(1)$.

Notation. $X \in \Theta(\alpha)$ denotes entry-wise (coordinate) scale. Cumulative SGD updates: $W_t^\ell = W_0^\ell + \Delta_t W^\ell$, $\Delta_t X = X_t - X_{t-1}$, with init values subscripted by 0.

Standard tools. We invoke the following standard probabilistic facts as needed, without per-instance citation:

- Central limit theorem (CLT) for sums of approximately independent terms;
- Law of large numbers (LLN) for empirical averages;
- Operator-norm concentration for iid (sub-)Gaussian matrices: for $W \in \mathbb{R}^{n\times m}$ with iid $\mathcal{N}(0,\sigma^2)$ entries, $\|W\|_{op} = \sigma(\sqrt{n}+\sqrt{m})\,(1+o(1))$ with high probability (Vershynin, 2018).

Feature update decomposition. For $h^\ell = W^\ell h^{\ell-1}$,

$$\Delta h^\ell \;=\; \underbrace{W_0^\ell\,\Delta h^{\ell-1}}_{\text{propagating}} + \underbrace{\Delta W^\ell\, h_0^{\ell-1}}_{\text{effective}} + \underbrace{\Delta W^\ell\,\Delta h^{\ell-1}}_{\text{cross term}}.$$

J.3 Scaling derivation for Regime II

In this section, we provide the heuristic scaling derivation for Regime II.

Definition J.1 (μP baseline, Regime II)**.**

- **Initialization.** All parameters are drawn *independently* according to:

$$W_0^1 \sim \mathcal{N}(0, D^{-1}),\ Q_0 \sim \mathcal{N}(0, N^{-1}),\ W_0^{2,i} \sim \mathcal{N}(0, N^{-1}),$$
$$W_0^{3,i} \sim \mathcal{N}(0, N_e^{-1}),\ W_0^4 \sim \mathcal{N}(0, N^{-2}),\ \text{for } i \in [M].$$

- **SGD learning rates.** $\eta_1 = \eta\, N,\ \eta_Q = \eta(M/N),\ \eta_2 = (M/N)\eta,\ \eta_3 = (MN)\eta,\ \eta_4 = \eta\, N^{-1}$.
- **Adam learning rates.** $\eta_1 = \eta,\ \eta_Q = \eta_2 = \eta_4 = \eta\, N^{-1},\ \eta_3 = 1/N_e\eta$.
- **Adam epsilon.** $\epsilon_1 = \epsilon\, N^{-1},\ \epsilon_Q = \epsilon\, M^{-1},\ \epsilon_2 = \epsilon\, M^{-1},\ \epsilon_3 = \epsilon\, N^{-1}M^{-1},\ \epsilon_4 = \epsilon$.

Definition J.2 (SSP, Regime II)**.**

- **Initialization.** All parameters are drawn *independently* according to:

$$W_0^1 \sim \mathcal{N}(0, D^{-1}),\ Q_0 \sim \mathcal{N}(0, N^{-1}),\ W_0^{2,i} \sim \mathcal{N}(0, N^{-1}),$$
$$W_0^{3,i} \sim \mathcal{N}(0, MN_e^{-1}),\ W_0^4 \sim \mathcal{N}(0, N^{-2}),\ \text{for } i \in [M].$$

- **SGD learning rates.** $\eta_1 = \eta\, N,\ \eta_Q = \eta(M/N),\ \eta_2 = (M/N)\eta,\ \eta_3 = (MN)\eta,\ \eta_4 = \eta\, N^{-1}$.
- **Adam learning rates.** $\eta_1 = \eta,\ \eta_Q = \eta_2 = \eta_4 = \eta\, N^{-1},\ \eta_3 = 1/N_e\eta$.
- **Adam epsilon.** $\epsilon_1 = \epsilon\, N^{-1},\ \epsilon_Q = \epsilon\, M^{-1},\ \epsilon_2 = \epsilon\, M^{-1},\ \epsilon_3 = \epsilon\, N^{-1}M^{-1},\ \epsilon_4 = \epsilon$.

J.3.1 Deriving μP in Regime II

This subsection presents a self-contained derivation for the μP baseline (Definition J.1) in Regime II.

Setup and standing assumptions

Architecture. As in §J.1: minimal MoE with sigmoid gating and explicit $1/M$ aggregation $h^3 = (1/M)\sum_i \phi_i\, h^{3,i}$.

Initialization (μP baseline). Per Definition J.1, with all parameters drawn independently:

$$(W_0^1)_{ab} \sim \mathcal{N}(0, 1/D), \qquad (Q_0)_{ia} \sim \mathcal{N}(0, 1/N),$$
$$(W_0^{2,i})_{ab} \sim \mathcal{N}(0, 1/N), \qquad (W_0^{3,i})_{ab} \sim \mathcal{N}(0, 1/N_e),$$
$$(W_0^4)_a \sim \mathcal{N}(0, 1/N^2).$$

Resulting entry scales: W_0^4 entries $\Theta(1/N)$, $W_0^{2,i}$ entries $\Theta(1/\sqrt{N})$, $W_0^{3,i}$ entries $\Theta(1)$ (since $\sigma_3^2 = 1/N_e = \Theta(1)$).

Learning rates (SGD). $\eta_1 = \eta N$, $\eta_Q = \eta M/N$, $\eta_2 = \eta M/N$, $\eta_3 = \eta MN$, $\eta_4 = \eta/N$, $\eta \in \Theta(1)$.

Gating. Sigmoid: $\phi_{i,t}, \dot{\phi}_{i,t} \in \Theta(1)$.

Auxiliary scaling lemmas We reuse Lemma J.7 from §J.4.4.

Specialization of Lemma J.7 to the rank-deficient initialization Gram matrices in Regime II. For the random initialization $W_0^{3,i} \in \mathbb{R}^{N\times N_e}$ with $\sigma_W^2 = 1/N_e$ and $N_e = \Theta(1)$, the Gram matrix $W_0^{3,i}(W_0^{3,i})^\top \in \mathbb{R}^{N\times N}$ has rank at most $N_e = \Theta(1)$: its diagonal mean is 1 and its off-diagonal entries have coordinate scale $\Theta(1)$ (variance $1/N_e = \Theta(1)$). Consequently, acting on a vector $v = (W_0^4)^\top$ with entries of coordinate scale $\Theta(1/N)$, the product $W_0^{3,i}(W_0^{3,i})^\top v$ has entries of coordinate scale $\Theta(1/\sqrt{N})$ in mostly-random directions, not aligned along v.

Analogously, for the random initialization $W_0^{2,i} \in \mathbb{R}^{N_e\times N}$ with $\sigma_W^2 = 1/N$, the Gram matrix $(W_0^{2,i})^\top W_0^{2,i} \in \mathbb{R}^{N\times N}$ has rank at most $N_e = \Theta(1)$ and entries of coordinate scale $\Theta(1/N)$. Acting on h_0^1, it produces entries of coordinate scale $\Theta(1/\sqrt{N})$ in mostly-random directions, not aligned along h_0^1.

We refer to $W_0^{3,i}(W_0^{3,i})^\top$ and $(W_0^{2,i})^\top W_0^{2,i}$ collectively as the *rank-deficient initialization Gram matrices* of Regime II.

First forward pass

$$h_0^1 = W_0^1\, x \in \Theta(1). \quad (\text{CLT};\ \sigma_1^2 = 1/D,\ \|x\|^2 \in \Theta(D)) \tag{R2μ-F1.1}$$
$$\psi_0 = Q_0\, h_0^1 \in \Theta(1). \quad (\text{CLT};\ \sigma_Q^2 = 1/N,\ \|h_0^1\|^2 \in \Theta(N)) \tag{R2μ-F1.2}$$
$$\phi_{i,0} = \sigma(\psi_{i,0}) \in \Theta(1). \quad (\text{sigmoid bounded; gating assumption}) \tag{R2μ-F1.3}$$
$$h_0^{2,i} = W_0^{2,i}\, h_0^1 \in \Theta(1), \quad \text{independent across } i$$
$$(\text{CLT};\ \sigma_2^2 = 1/N,\ \|h_0^1\|^2 \in \Theta(N);$$
$$\|h_0^{2,i}\|^2 = N_e \cdot \Theta(1) = \Theta(1) \text{ since } N_e = \Theta(1),\ \text{not } \Theta(N_e) \to \infty). \tag{R2μ-F1.4}$$
$$h_0^{3,i} = W_0^{3,i}\, h_0^{2,i} \in \Theta(1)$$
$$(\text{variance per entry } \sigma_3^2\|h_0^{2,i}\|^2 = (1/N_e)\cdot\Theta(1) = \Theta(1);$$
$$\text{direct variance, not asymptotic CLT, since } N_e = O(1)). \tag{R2μ-F1.5}$$
$$h_0^3 = (1/M)\textstyle\sum_i \phi_{i,0}\, h_0^{3,i} \in \Theta(1/\sqrt{M}) = \Theta(1/\sqrt{N})$$
$$(\text{cross-}i \text{ CLT on entries of coordinate scale } \Theta(1);$$
$$\{h_0^{3,i}\} \text{ independent across } i). \tag{R2μ-F1.6}$$
$$f_0 = W_0^4\, h_0^3 = \langle W_0^4, h_0^3\rangle \in \Theta(1/N)$$
$$(\text{variance } \textstyle\sum_a \Theta(1/N^2)\cdot\Theta(1/N) = \Theta(1/N^2);$$
$$h_0^3 \text{ entry scale } \Theta(1/\sqrt{N}) \text{ via (R2}\mu\text{-F1.6)}). \tag{R2μ-F1.7}$$

First backward pass and step-1 updates

New intermediate scaling. The intermediate quantity $(W_0^{3,i})^\top (W_0^4)^\top$ has entries of variance $\sigma_3^2 \, \|W_0^4\|^2 = \Theta(1) \cdot \Theta(1/N) = \Theta(1/N)$, hence coordinate scale $\Theta(1/\sqrt{N})$.

Per-layer gradients (with $W_0^4/\Delta W^4$ split).

$$\partial f_0/\partial h_0^3 = (W_0^4)^\top \in \Theta(1/N). \quad (W_0^4 \text{ entries } \Theta(1/N);\ \Delta_t W^4 = 0 \text{ at } t = 0) \tag{R2μ-B1.1}$$

$$\partial f_0/\partial h_0^{3,i} = (\phi_{i,0}/M)(W_0^4)^\top \in \Theta(1/(MN)) = \Theta(1/N^2). \quad (\phi_{i,0} \in \Theta(1)) \tag{R2μ-B1.2}$$

$$\begin{aligned}\partial f_0/\partial h_0^{2,i} = (\phi_{i,0}/M)(W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/(M\sqrt{N})) = \Theta(1/N^{3/2}) \\ ((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/\sqrt{N}) \text{ from new intermediate scaling above}).\end{aligned} \tag{R2μ-B1.3}$$

$$\begin{aligned}\partial f_0/\partial \phi_{i,0} = (1/M)\langle h_0^{3,i}, W_0^4\rangle \in \Theta(1/N^{3/2}) \\ (\langle h_0^{3,i}, W_0^4\rangle \text{ variance } \textstyle\sum_a \Theta(1) \cdot \Theta(1/N^2) = \Theta(1/N), \text{ coord } \Theta(1/\sqrt{N}); \\ h_0^{3,i} \text{ entry scale } \Theta(1) \text{ via (R2}\mu\text{-F1.5})).\end{aligned} \tag{R2μ-B1.4}$$

$$\begin{aligned}(\partial f_0/\partial h_0^1)_{\exp} = (1/M) \textstyle\sum_i \phi_{i,0} (W_0^{2,i})^\top (W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/N^{3/2}) \\ (\text{at } t = 0 \text{ only the all-init } A_{4.1} \text{ piece is non-zero;} \\ \text{per summand } (W_0^{2,i})^\top \text{ acts on } (W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/\sqrt{N}) \\ \text{with } \|v\|^2 = N_e\, \Theta(1/N) = \Theta(1/N); \\ \text{variance } (1/N)\, \Theta(1/N) = \Theta(1/N^2), \text{ per-}i \text{ coord } \Theta(1/N); \\ \text{cross-}i \text{ CLT}).\end{aligned} \tag{R2μ-B1.5}$$

$$\begin{aligned}(\partial f_0/\partial h_0^1)_{\text{router}} = (1/M) Q_0^\top v, \quad v_i := \dot\phi_{i,0} \langle h_0^{3,i}, W_0^4\rangle \\ \in \Theta(1/N^{3/2}) \\ (v_i \in \Theta(1/\sqrt{N}) \text{ via (R2}\mu\text{-B1.4)};\ \|v\|^2 = M\, \Theta(1/N) = \Theta(1); \\ \text{cross-layer CLT giving } Q_0^\top v \text{ coord } \Theta(1/\sqrt{N}); \\ \text{W}^{3,i}\text{-split } v_i = v_i^I + v_i^U \text{ where} \\ v_i^I := \dot\phi_{i,0} \langle (W_0^{3,i}) h_0^{2,i}, W_0^4\rangle \\ v_i^U := \dot\phi_{i,0} \langle (\Delta_0 W^{3,i}) h_0^{2,i}, W_0^4\rangle = 0 \text{ since } \Delta_0 W^{3,i} = 0; \\ \text{so } (1/M) Q_0^\top v^I = \Theta(1/N^{3/2}) \text{ and } (1/M) Q_0^\top v^U = 0).\end{aligned} \tag{R2μ-B1.6}$$

$$\partial f_0/\partial h_0^1 = (\partial f_0/\partial h_0^1)_{\exp} + (\partial f_0/\partial h_0^1)_{\text{router}} \in \Theta(1/N^{3/2}). \tag{R2μ-B1.7}$$

Step-1 parameter updates.

$$\begin{aligned}\Delta_1 W^4 = -(\eta/N) \chi_0 (h_0^3)^\top \in \Theta(1/N^{3/2}) \\ (h_0^3 \text{ entry scale } \Theta(1/\sqrt{N}) \text{ via (R2}\mu\text{-F1.6)};\ \text{sub-leading vs } W_0^4 \in \Theta(1/N)).\end{aligned} \tag{R2μ-U1.4}$$

$$\begin{aligned}\Delta_1 W^{3,i} = -\eta_3 \chi_0 (\phi_{i,0}/M)(W_0^4)^\top (h_0^{2,i})^\top = -\eta N \chi_0 \phi_{i,0}\, (W_0^4)^\top (h_0^{2,i})^\top \\ \in \Theta(1) \text{ rank-1 along } (W_0^4)^\top \otimes h_0^{2,i} \\ (\eta_3 = \eta M N;\ h_0^{2,i} \in \Theta(1) \text{ via (R2}\mu\text{-F1.4)}; \\ \text{same scale as } W_0^{3,i} \in \Theta(1)).\end{aligned} \tag{R2μ-U1.3}$$

$$\Delta_1 W^{2,i} = -(\eta\chi_0\phi_{i,0}/N)(W_0^{3,i})^\top (W_0^4)^\top (h_0^1)^\top \in \Theta(1/N^{3/2}) \text{ rank-1}$$
$$(\eta_2 = \eta M/N;\ (W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/\sqrt{N}) \text{ from new intermediate scaling;}$$
$$\text{sub-leading vs } W_0^{2,i} \in \Theta(1/\sqrt{N})). \qquad \text{(R2}\mu\text{-U1.2)}$$

$$\Delta_1 W^1 = -\eta_1\chi_0(\partial f_0/\partial h_0^1)\, x^\top \in \Theta(1/\sqrt{N})$$
$$(\eta_1 = \eta N;\ \partial f_0/\partial h_0^1 \in \Theta(1/N^{3/2}) \text{ via (R2}\mu\text{-B1.7)}). \qquad \text{(R2}\mu\text{-U1.1a)}$$

$$\Delta_1 h^1 = \Delta_1 W^1\, x \in \Theta(1/\sqrt{N}) \text{ aligned along } \partial f_0/\partial h_0^1. \qquad \text{(R2}\mu\text{-U1.1b)}$$

$$\Delta_1 Q = -\eta_Q\chi_0(\partial f_0/\partial\phi)(h_0^1)^\top \in \Theta(1/N^{3/2}). \quad (\partial f_0/\partial\phi \in \Theta(1/N^{3/2}) \text{ via (R2}\mu\text{-B1.4)}) \qquad \text{(R2}\mu\text{-U1.Q)}$$

Second forward pass

$$h_1^1 = h_0^1 + \Delta_1 h^1 \in \Theta(1). \quad (h_0^1 \in \Theta(1) \text{ dominates } \Delta_1 h^1 \in \Theta(1/\sqrt{N}) \text{ via (R2}\mu\text{-U1.1b)}) \qquad \text{(R2}\mu\text{-F2.1)}$$

$$\psi_1 = \psi_0 + \Delta_1\psi \in \Theta(1), \quad \phi_1 \in \Theta(1) \qquad \text{(R2}\mu\text{-F2.2–R2}\mu\text{-F2.3)}$$
$$(\Delta_1\psi \in \Theta(1/\sqrt{N}) \text{ from } Q_0\Delta_1 h^1 \text{ and } \Delta_1 Q\, h_0^1;$$
$$\text{scales by (R2}\mu\text{-U1.1b), (R2}\mu\text{-U1.Q);}$$
$$\text{cross } \Delta_1 Q\, \Delta_1 h^1 \in \Theta(1/N^{3/2}) \text{ sub-leading via } \partial f_0/\partial h_0^1 \text{ via (R2}\mu\text{-B1.7))}$$

(R2μ-F2.4) $h_1^{2,i} = W_1^{2,i}\, h_1^1$, four-piece decomposition:

$$h_1^{2,i} = h_0^{2,i} + W_0^{2,i}\Delta_1 h^1 + \Delta_1 W^{2,i}\, h_0^1 + \Delta_1 W^{2,i}\, \Delta_1 h^1. \qquad \text{(R2}\mu\text{-F2.4)}$$

$$\text{init: } W_0^{2,i} h_0^1 = h_0^{2,i} \in \Theta(1). \quad \text{(R2}\mu\text{-F1.4)} \qquad \text{(R2}\mu\text{-F2.4a)}$$

$$\text{prop: } W_0^{2,i}\Delta_1 h^1 \in \Theta(1/\sqrt{N}) \qquad \text{(R2}\mu\text{-F2.4b)}$$
$$(\sigma_2^2 \|\Delta_1 h^1\|^2 = (1/N)\,\Theta(1) = \Theta(1/N);$$
$$\|\Delta_1 h^1\|^2 \in \Theta(1) \text{ via (R2}\mu\text{-U1.1b))}$$

$$\text{eff: } \Delta_1 W^{2,i}\, h_0^1 = -(\eta\chi_0\phi_{i,0}/N)\, \|h_0^1\|^2\, (W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/\sqrt{N}) \qquad \text{(R2}\mu\text{-F2.4c)}$$
$$(\text{(R2}\mu\text{-U1.2)substitution;}$$
$$(W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/\sqrt{N}) \text{ from new intermediate scaling;}$$
$$\|h_0^1\|^2 \in \Theta(N))$$

$$\text{cross: } \Delta_1 W^{2,i}\, \Delta_1 h^1 = -(\eta\chi_0\phi_{i,0}/N)\, \langle h_0^1, \Delta_1 h^1\rangle\, (W_0^{3,i})^\top (W_0^4)^\top$$
$$\in \Theta(1/N^{3/2}) \qquad \text{(R2}\mu\text{-F2.4d)}$$
$$((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/\sqrt{N}) \text{ from new intermediate scaling;}$$
$$\langle h_0^1, \Delta_1 h^1\rangle \in \Theta(1) \text{ random, derived via}$$
$$\partial f_0/\partial h_0^1 \in \Theta(1/N^{3/2}) \text{ via (R2}\mu\text{-B1.7);}$$
$$\text{sub-leading vs propagating and effective)}$$

(R2μ-F2.5) $h_1^{3,i} = W_1^{3,i}\, h_1^{2,i}$, four-piece decomposition:

$$h_1^{3,i} = h_0^{3,i} + W_0^{3,i}\Delta_1 h^{2,i} + \Delta_1 W^{3,i}\, h_0^{2,i} + \Delta_1 W^{3,i}\, \Delta_1 h^{2,i}. \qquad \text{(R2}\mu\text{-F2.5)}$$

$$\text{init: } h_0^{3,i} \in \Theta(1). \quad \text{(R2}\mu\text{-F1.5)} \qquad \text{(R2}\mu\text{-F2.5a)}$$

$$\text{prop: } W_0^{3,i}\Delta_1 h^{2,i} \in \Theta(1/\sqrt{N}) \tag{R2μ-F2.5b}$$
$$(\sigma_3^2\|\Delta_1 h^{2,i}\|^2 = \Theta(1)\cdot\Theta(1/N) = \Theta(1/N);$$
$$\|\Delta_1 h^{2,i}\|^2 \in \Theta(1) \text{ via (R2}\mu\text{-F2.4))}$$

$$\text{eff: } \Delta_1 W^{3,i}\, h_0^{2,i} = -\eta N \chi_0 \phi_{i,0}\, \|h_0^{2,i}\|^2\, (W_0^4)^\top \in \Theta(1) \text{ along } (W_0^4)^\top \tag{R2μ-F2.5c}$$
$$(\text{(R2}\mu\text{-U1.3)substitution; } \|h_0^{2,i}\|^2 \in \Theta(1) \text{ via (R2}\mu\text{-F1.4);}$$
$$(W_0^4)^\top \in \Theta(1/N))$$

$$\text{cross: } \Delta_1 W^{3,i}\, \Delta_1 h^{2,i} = -\eta N \chi_0 \phi_{i,0}\, \langle h_0^{2,i}, \Delta_1 h^{2,i}\rangle\, (W_0^4)^\top \in \Theta(1/\sqrt{N}) \text{ along } (W_0^4)^\top \tag{R2μ-F2.5d}$$
$$(\text{(R2}\mu\text{-U1.3)substitution;}$$
$$\text{dominant piece of } \Delta_1 h^{2,i} \propto (W_0^{3,i})^\top (W_0^4)^\top \text{ via (R2}\mu\text{-F2.4c),}$$
$$\text{so } \langle h_0^{2,i}, \Delta_1 h^{2,i}\rangle \sim -(\eta\chi_0\phi/N)\|h_0^1\|^2 \langle h_0^{3,i}, W_0^4\rangle \in \Theta(1/\sqrt{N})$$
$$\text{random across } i, \text{ via (R2}\mu\text{-B1.4);}$$
$$\text{random sign across } i, \text{ sub-leading vs effective)}$$

Effective dominates: $\Delta_1 h^{3,i} \in \Theta(1)$ aligned along $(W_0^4)^\top$ — the engine of feature learning at $t = 1$.

(R2μ-F2.6) $h_1^3 = A_1 + A_{2,1} + A_{2,2} + A_3 + D$, where

$$A_1 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\, h_0^{3,i}, \qquad A_{2,1} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\, W_0^{3,i} W_0^{2,i}\, \Delta_1 h^1,$$
$$A_{2,2} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\, W_0^{3,i}\, \Delta_1 W^{2,i}\, h_0^1, \qquad A_3 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\, \Delta_1 W^{3,i}\, h_0^{2,i},$$
$$D := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\, \Delta_1 W^{3,i}\, \Delta_1 h^{2,i}.$$

Here and below, $A_{2,2}$ is evaluated at the current input $h_t^1 = h_0^1 + \Delta h^1$. The additional contribution $\frac{1}{M}\sum_i \phi_i W_0^{3,i}\Delta W^{2,i}\Delta h^1$ equals $\langle h_0^1, \Delta h^1\rangle/\|h_0^1\|^2$ times the h_0^1 term at the first hidden update, since $\Delta W^{2,i}$ is rank one. It therefore has the same scale and does not change the conclusions.

$$A_1 \in \Theta(1/\sqrt{N}). \quad \text{(cross-}i\text{ CLT on } \Theta(1)\text{-entry independent vectors)} \tag{R2μ-F2.6a}$$

$$A_{2,1} \in \Theta(1/(\sqrt{M}\sqrt{N})) = \Theta(1/N) \tag{R2μ-F2.6b}$$
$$(\text{per summand chain CLT entries } \Theta(1/\sqrt{N}) \text{ random;}$$
$$\|\Delta_1 h^1\|^2 \in \Theta(1) \text{ via (R2}\mu\text{-U1.1b); cross-}i\text{ CLT)}$$

$$A_{2,2} = -(\eta\chi_0/N)\, \|h_0^1\|^2\, \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\, W_0^{3,i}(W_0^{3,i})^\top (W_0^4)^\top$$
$$\in \Theta(1/N) \text{ in mostly-random directions} \tag{R2μ-F2.6c}$$
$$(\text{(R2}\mu\text{-U1.2)substitution;}$$
$$W_0^{3,i}(W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/\sqrt{N}) \text{ in mostly-random directions}$$
$$\text{by rank-deficient initialization Gram (§J.3.1);}$$
$$\|h_0^1\|^2 \in \Theta(N); \text{ per summand entries } \Theta(1/\sqrt{N}), \text{ cross-}i\text{ CLT)}$$

$$A_3 = -\eta N \chi_0\, (W_0^4)^\top \left(\tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\, \|h_0^{2,i}\|^2\right) \in \Theta(1) \text{ along } (W_0^4)^\top \tag{R2μ-F2.6d}$$
$$(\text{(R2}\mu\text{-U1.3)substitution; } \|h_0^{2,i}\|^2 \in \Theta(1) \text{ via (R2}\mu\text{-F1.4); LLN;}$$
$$(W_0^4)^\top \in \Theta(1/N); \text{ entry } \eta N \cdot (1/N) \cdot \Theta(1) = \Theta(1))$$

$$D = -\eta N \chi_0 \, (W_0^4)^\top \, \tfrac{1}{M} \textstyle\sum_i \phi_{i,1}\phi_{i,0} \, \langle h_0^{2,i}, \Delta_1 h^{2,i} \rangle$$
$$\in \Theta(1/N) \text{ along } (W_0^4)^\top \qquad \text{(R2}\mu\text{-F2.6e)}$$
$$(\text{(R2}\mu\text{-U1.3)substitution;}$$
$$\langle h_0^{2,i}, \Delta_1 h^{2,i} \rangle \in \Theta(1/\sqrt{N}) \text{ random across } i$$
$$\text{via } \langle h_0^{3,i}, W_0^4 \rangle \in \Theta(1/\sqrt{N}) \text{ via (R2}\mu\text{-B1.4);}$$
$$\text{cross-}i \text{ CLT in the random factor;}$$
$$\text{sub-leading vs } A_3 \in \Theta(1))$$

$$h_1^3 \in \Theta(1) \text{ along } (W_0^4)^\top, \text{ from } A_3 \text{ alone.} \qquad \text{(R2}\mu\text{-F2.6)}$$

$$f_1 = W_1^4 \, h_1^3 \approx \langle W_0^4, A_3 \rangle = -\eta N \chi_0 \, \|W_0^4\|^2 \left(\tfrac{1}{M} \textstyle\sum_i \phi_{i,1}\phi_{i,0} \, \|h_0^{2,i}\|^2 \right) \in \Theta(1) \qquad \text{(R2}\mu\text{-F2.7)}$$
$$(A_3 \text{ aligned along } (W_0^4)^\top \text{ via (R2}\mu\text{-F2.6d);}$$
$$\|W_0^4\|^2 = N \cdot (1/N^2) = 1/N; \; \|h_0^{2,i}\|^2 \in \Theta(1) \text{ via (R2}\mu\text{-F1.4);}$$
$$\text{net } \eta N \cdot \Theta(1) \cdot (1/N) = \Theta(1))$$

Second backward pass and step-2 updates

New intermediate at $t = 1$.

$$(\Delta_1 W^{3,i})^\top (W_1^4)^\top \approx -\eta N \chi_0 \phi_{i,0} \, \|W_0^4\|^2 \, h_0^{2,i} \in \Theta(1) \text{ aligned } \| \; h_0^{2,i}$$
$$(\text{(R2}\mu\text{-U1.3)rank-1 substitution; } \|W_0^4\|^2 = 1/N;$$
$$h_0^{2,i} \in \Theta(1) \text{ via (R2}\mu\text{-F1.4); net } \eta N \cdot (1/N) = \Theta(1))$$

This $\Theta(1)$ aligned-along-$h_0^{2,i}$ contribution from the rank-1 update is what drives $\partial f_1 / \partial h_1^{2,i}$ to climb from $\Theta(1/N^{3/2})$ at $t = 0$ to $\Theta(1/N)$ at $t = 1$.

Per-layer gradients (with $W_0^4/\Delta W^4$ binary split).

$$\partial f_1 / \partial h_1^3 = (W_0^4)^\top + (\Delta_1 W^4)^\top \in \Theta(1/N) \qquad \text{(R2}\mu\text{-B2.1)}$$
$$(W_0^4 \in \Theta(1/N) \text{ dominates } \Delta_1 W^4 \in \Theta(1/N^{3/2}) \text{ via (R2}\mu\text{-U1.4))}$$

$$\partial f_1 / \partial h_1^{3,i} = (\phi_{i,1}/M)(W_1^4)^\top \in \Theta(1/N^2) \qquad \text{(R2}\mu\text{-B2.2)}$$
$$(\text{init piece } \Theta(1/N^2); \; \Delta_1 W^4 \text{ piece } \Theta(1/N^{5/2}) \text{ via (R2}\mu\text{-U1.4))}$$

(R2μ-B2.3) $\partial f_1 / \partial h_1^{2,i} = (\phi_{i,1}/M)(W_1^{3,i})^\top (W_1^4)^\top$, four-piece decomposition:

$$(\phi_{i,1}/M)(W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/N^{3/2}) \qquad \text{(R2}\mu\text{-B2.3a)}$$
$$((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/\sqrt{N}) \text{ from new intermediate scaling at } t = 0)$$

$$(\phi_{i,1}/M)(W_0^{3,i})^\top (\Delta_1 W^4)^\top = -(\eta \chi_0 \phi_{i,0}/(MN))(W_0^{3,i})^\top h_0^3$$
$$\in \Theta(1/N^2) \qquad \text{(R2}\mu\text{-B2.3b)}$$
$$(\text{(R2}\mu\text{-U1.4)substitution;}$$
$$(W_0^{3,i})^\top h_0^3 \in \Theta(1)$$
$$\text{from } j = i \text{ coherent contribution to the average})$$

$$(\phi_{i,1}/M)(\Delta_1 W^{3,i})^\top (W_0^4)^\top \in \Theta(1/N) \text{ along } h_0^{2,i} \qquad \text{(R2}\mu\text{-B2.3c)}$$
$$((\Delta_1 W^{3,i})^\top (W_0^4)^\top \in \Theta(1) \text{ along } h_0^{2,i}$$
$$\text{from new intermediate at } t = 1 \text{ above; } \textbf{dominant})$$

$$(\phi_{i,1}/M)(\Delta_1 W^{3,i})^\top(\Delta_1 W^4)^\top \in \Theta(1/N^2) \qquad \text{(R2}\mu\text{-B2.3d)}$$
(same alignment as (R2μ-B2.3c);
$\Delta_1 W^4$ smaller by $1/\sqrt{N}$ via (R2μ-U1.4))

$$\partial f_1/\partial h_1^{2,i} \in \Theta(1/N) \text{ along } h_0^{2,i}, \text{ from (R2}\mu\text{-B2.3c).} \qquad \text{(R2}\mu\text{-B2.3)}$$

$$\partial f_1/\partial \phi_{i,1} = (1/M)\langle h_1^{3,i}, W_1^4\rangle \in \Theta(1/N) \qquad \text{(R2}\mu\text{-B2.4)}$$
(Decompose $h_1^{3,i} = (W_0^{3,i})\, h_1^{2,i} + (\Delta_1 W^{3,i})\, h_1^{2,i}$ and $W_1^4 = W_0^4 + \Delta_1 W^4$;
four-piece grid:
(init · init) $\langle (W_0^{3,i})\, h_1^{2,i}, W_0^4\rangle \in \Theta(1/\sqrt{N})$
via random IP $\langle h_0^{3,i}, W_0^4\rangle \in \Theta(1/\sqrt{N})$ (R2μ-B1.4);
(init · update) $\langle (W_0^{3,i})\, h_1^{2,i}, \Delta_1 W^4\rangle = -(\chi_0/N)\langle h_0^{3,i}, h_0^3\rangle \in \Theta(1/N)$
via $\langle h_0^{3,i}, h_0^3\rangle \in \Theta(1)$ ($j = i$ diagonal $(\phi_{i,0}/M)\|h_0^{3,i}\|^2 = \Theta(1)$ in μP);
(update · init) $\langle (\Delta_1 W^{3,i})\, h_1^{2,i}, W_0^4\rangle = -\eta N \chi_0 \phi_{i,0} \|h_0^{2,i}\|^2 \|W_0^4\|^2 \in \Theta(1)$
via rank-1 alignment of $\Delta_1 W^{3,i} h_0^{2,i}$ with $(W_0^4)^\top$; $\eta N \cdot \Theta(1) \cdot \Theta(1/N) = \Theta(1)$;
(update · update) $\langle (\Delta_1 W^{3,i})\, h_1^{2,i}, \Delta_1 W^4\rangle = \kappa \langle W_0^4, \Delta_1 W^4\rangle \in \Theta(1/N)$
via $\kappa \in \Theta(N)$ (rank-1 prefactor) and $\langle W_0^4, \Delta_1 W^4\rangle \in \Theta(1/N^2)$;
$/M$: $(\Theta(1/N^{3/2}), \Theta(1/N^2), \Theta(1/N), \Theta(1/N^2))$ respectively;
(update · init) dominates; total $\Theta(1/N)$)

Expert pathway: $(\partial f_1/\partial h_1^1)_{\text{exp}} = A_4 + A_5 + A_6 + E$, full 8-piece expansion. Each of A_4, A_5, A_6, E splits into a .1 piece (using W_0^4) and a .2 piece (using $\Delta_1 W^4$):

$$A_{4.1} \in \Theta(1/N^{3/2}). \quad \text{(R2}\mu\text{-B1.5)} \qquad \text{(R2}\mu\text{-B2.5a.1)}$$
$$A_{4.2} \in \Theta(1/N^2). \quad (\Delta_1 W^4 \text{ smaller by } 1/\sqrt{N} \text{ via (R2}\mu\text{-U1.4)}) \qquad \text{(R2}\mu\text{-B2.5a.2)}$$

$$A_{5.1} = \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}(W_0^{2,i})^\top(\Delta_1 W^{3,i})^\top(W_0^4)^\top$$
$$\in \Theta(1/N) \text{ in mostly-random directions} \qquad \text{(R2}\mu\text{-B2.5b.1)}$$
$((\Delta_1 W^{3,i})^\top(W_0^4)^\top \in \Theta(1)$ along $h_0^{2,i}$ from new intermediate at $t = 1$;
$(W_0^{2,i})^\top h_0^{2,i} = (W_0^{2,i})^\top W_0^{2,i} h_0^1 \in \Theta(1/\sqrt{N})$ in mostly-random directions
by rank-deficient initialization Gram (§J.3.1);
coherent $(N_e/N)\, h_0^1$ piece is $\Theta(1/N)$ sub-leading;
cross-i CLT: $\Theta(1/\sqrt{N})/\sqrt{M} = \Theta(1/N)$)

$$A_{5.2} \in \Theta(1/N^2). \quad (\text{same structure as (R2}\mu\text{-B2.5b.1)};\ \Delta_1 W^4 \text{ smaller by } 1/\sqrt{N} \text{ via (R2}\mu\text{-U1.4)}) \qquad \text{(R2}\mu\text{-B2.5b.2)}$$

$$A_{6.1} = \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}(\Delta_1 W^{2,i})^\top(W_0^{3,i})^\top(W_0^4)^\top$$
$$= -(\eta\chi_0/N)\, h_0^1\, \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\, S_i \in \Theta(1/N^2) \text{ along } h_0^1 \qquad \text{(R2}\mu\text{-B2.5c.1)}$$
((R2μ-U1.2)substitution; $S_i := W_0^4 W_0^{3,i}(W_0^{3,i})^\top(W_0^4)^\top \in \Theta(1/N)$
quadratic form in W_0^4 across rank-N_e Gram, leading piece $\|W_0^4\|^2$;
per summand entries $(1/N)\cdot 1 \cdot (1/N) = \Theta(1/N^2)$ aligned along h_0^1)

$$A_{6.2} \in \Theta(1/N^3). \quad (\text{same structure as (R2}\mu\text{-B2.5c.1)};\ \Delta_1 W^4 \text{ smaller by } 1/\sqrt{N} \text{ via (R2}\mu\text{-U1.4)}) \qquad \text{(R2}\mu\text{-B2.5c.2)}$$

$$E_1 = \tfrac{1}{M}\sum_i \phi_{i,1}(\Delta_1 W^{2,i})^\top(\Delta_1 W^{3,i})^\top(W_0^4)^\top \in \Theta(1/N^2) \quad \text{(R2}\mu\text{-B2.5d.1)}$$

(combining (R2μ-U1.2), (R2μ-U1.3) substitutions;
per summand involves $\langle W_0^4, h_0^{3,i}\rangle \in \Theta(1/\sqrt{N})$ random via (R2μ-B1.4);
per summand entries $\Theta(1/N^{3/2})$; cross-i CLT)

$$E_2 \in \Theta(1/N^{5/2}). \quad \text{(same as (R2}\mu\text{-B2.5d.1); } \Delta_1 W^4 \text{ smaller by } 1/\sqrt{N} \text{ via (R2}\mu\text{-U1.4))} \quad \text{(R2}\mu\text{-B2.5d.2)}$$

$$(\partial f_1/\partial h_1^1)_{\exp} \in \Theta(1/N) \text{ in mostly-random directions, from } A_{5.1} \text{ alone.} \quad \text{(R2}\mu\text{-B2.5)}$$

Router pathway. Decomposing $\mathbf{v} = \mathbf{v}^{(0)} + \mathbf{v}^{(\Delta)}$ where $\mathbf{v}_i^{(0)} = \dot\phi_{i,1}\langle h_1^{3,i}, W_0^4\rangle \in \Theta(1)$ (coherent via the alignment in $h_1^{3,i}$ from (R2μ-F2.5c) / (R2μ-B2.4)) and $\mathbf{v}_i^{(\Delta)} \in \Theta(1/N)$ (via (R2μ-U1.4)):

$$(1/M)Q_0^\top \mathbf{v}^{(0)} \in \Theta(1/N) \text{ random direction} \quad \text{(R2}\mu\text{-B2.6a)}$$

(Q_0 indep. of $\mathbf{v}^{(0)}$; $\sigma_Q^2 = 1/N$; $\|\mathbf{v}^{(0)}\|^2 \in \Theta(M)$)

$$(1/M)Q_0^\top \mathbf{v}^{(\Delta)} \in \Theta(1/N^2). \quad \text{(R2}\mu\text{-B2.6b)}$$

$$(1/M)\Delta_1 Q^\top \mathbf{v}^{(0)} \in \Theta(1/N^2) \text{ along } h_0^1 \quad \text{(R2}\mu\text{-B2.6c)}$$

($\Delta_1 Q$ rank-1 along $(h_0^1)^\top$ via (R2μ-U1.Q);
$\langle w_0, \mathbf{v}^{(0)}\rangle = (\dot\phi^2/M)\sum_i [r_i c_i + r_i^2] \in \Theta(1/N)$
(both the cross-i CLT piece $\sum r_i c_i \sim 1$ and the LLN self-pairing $\sum r_i^2 \sim M/N$ contribute $\Theta(1)$ before the $(1/M)$ prefactor);
$\Delta_1 Q^\top \mathbf{v}^{(0)} = h_0^1 \eta_Q \chi_0 \langle w_0, \mathbf{v}^{(0)}\rangle \sim 1/N$ entries;
after $(1/M)$: $\Theta(1/N^2)$ — sub-leading vs (R2μ-B2.6a))

$$(1/M)\Delta_1 Q^\top \mathbf{v}^{(\Delta)} \in \Theta(1/N^3). \quad \text{(R2}\mu\text{-B2.6d)}$$

$$(\partial f_1/\partial h_1^1)_{\text{router}} \in \Theta(1/N), \text{ from (R2}\mu\text{-B2.6a) alone (other three pieces sub-leading).} \quad \text{(R2}\mu\text{-B2.6)}$$

$W^{3,i}$-imbalance propagates into the router pathway. The $W^{3,i}$-split of (R2μ-B2.4), $\mathbf{v} = \mathbf{v}^I + \mathbf{v}^U$ with $\mathbf{v}_i^I = \dot\phi_{i,1}\langle (W_0^{3,i})h_1^{2,i}, W_1^4\rangle$ and $\mathbf{v}_i^U = \dot\phi_{i,1}\langle (\Delta_1 W^{3,i})h_1^{2,i}, W_1^4\rangle$, transmits as: $\mathbf{v}_i^I \in \Theta(1/\sqrt{N})$ random across i (from $\langle h_0^{3,i}, W_0^4\rangle$), giving $\|\mathbf{v}^I\|^2 \in \Theta(1)$ and $(1/M)Q_t^\top \mathbf{v}^I \in \Theta(1/N^{3/2})$ via cross-layer CLT; whereas $\mathbf{v}_i^U \in \Theta(1)$ coherent across i (rank-1 alignment with $(W_0^4)^\top$), giving $\|\mathbf{v}^U\|^2 \in \Theta(M)$ and $(1/M)Q_t^\top \mathbf{v}^U \in \Theta(1/N)$. The same $\sqrt{N}$ deficit between $W_0^{3,i}$- and $\Delta W^{3,i}$-pieces is transmitted into $\partial f_1/\partial h_1^1$ via the router pathway.

$$\partial f_1/\partial h_1^1 = (\partial f_1/\partial h_1^1)_{\exp} + (\partial f_1/\partial h_1^1)_{\text{router}} \in \Theta(1/N). \quad \text{(R2}\mu\text{-B2.7)}$$

Step-2 parameter updates.

$$\Delta_2 W^4 = -(\chi_1/N)(h_1^3)^\top \in \Theta(1/N). \quad (h_1^3 \in \Theta(1) \text{ via (R2}\mu\text{-F2.6); now comparable to } W_0^4) \quad \text{(R2}\mu\text{-U2.4)}$$

$$\Delta_2 W^{3,i} = -\eta N \chi_1 \phi_{i,1}(W_1^4)^\top (h_1^{2,i})^\top \in \Theta(1) \text{ rank-1 along } (W_1^4)^\top \otimes h_1^{2,i} \quad \text{(R2}\mu\text{-U2.3)}$$

(same structure as (R2μ-U1.3);
$(W_1^4)^\top \in \Theta(1/N)$; $h_1^{2,i} \in \Theta(1)$ via (R2μ-F2.4))

$$\Delta_2 W^{2,i} = -(\eta\chi_1\phi_{i,1}/N)(W_1^{3,i})^\top (W_1^4)^\top (h_1^1)^\top \in \Theta(1/N) \text{ rank-1 along } h_0^{2,i} \otimes h_1^1 \quad \text{(R2}\mu\text{-U2.2)}$$

($(W_1^{3,i})^\top (W_1^4)^\top \in \Theta(1)$ along $h_0^{2,i}$ via new intermediate at $t = 1$;
$\sqrt{N}$ larger than $\Delta_1 W^{2,i} \in \Theta(1/N^{3/2})$)

$$\Delta_2 W^1 = -\eta_1 \chi_1 (\partial f_1/\partial h_1^1)\, x^\top \in \Theta(1) \qquad \text{(R2}\mu\text{-U2.1a)}$$
$$(\eta_1 = \eta N;\ \partial f_1/\partial h_1^1 \in \Theta(1/N) \text{ via (R2}\mu\text{-B2.7))}$$

$$\Delta_2 h^1 = \Delta_2 W^1\, x \in \Theta(1) \text{ aligned along } h_0^1. \quad \text{(embedding feature-learns at } t=2) \qquad \text{(R2}\mu\text{-U2.1b)}$$

$$\Delta_2 Q \in \Theta(1/N). \quad (\partial f_1/\partial \phi \in \Theta(1/N) \text{ via (R2}\mu\text{-B2.4))} \qquad \text{(R2}\mu\text{-U2.Q)}$$

Third forward pass We compute activations at $\theta^{(2)}$. The cumulative embedding change $\Delta h^1 := h_2^1 - h_0^1 = \Delta_1 h^1 + \Delta_2 h^1$, with $\Delta_1 h^1 \in \Theta(1/\sqrt{N})$ via (R2μ-U1.1b) and $\Delta_2 h^1 \in \Theta(1)$ aligned along h_0^1 via (R2μ-U2.1b); the latter dominates.

$$h_2^1 = h_0^1 + \Delta h^1 \in \Theta(1). \qquad \text{(R2}\mu\text{-F3.1)}$$

(R2μ-F3.2) $h_2^{2,i} = W_2^{2,i} h_2^1$, four-piece decomposition with cumulative $\Delta W^{2,i}$:

$$\text{init: } W_0^{2,i} h_0^1 = h_0^{2,i} \in \Theta(1). \quad \text{(R2}\mu\text{-F1.4)} \qquad \text{(R2}\mu\text{-F3.2a)}$$

$$\text{prop: } W_0^{2,i} \Delta h^1 \in \Theta(1) \qquad \text{(R2}\mu\text{-F3.2b)}$$
$$(\sigma_2^2 \|\Delta h^1\|^2 = (1/N)\,\Theta(N) = \Theta(1);$$
$$\|\Delta h^1\|^2 \in \Theta(N) \text{ via } \Delta_2 h^1 \in \Theta(1) \text{ via (R2}\mu\text{-U2.1b))}$$

$$\text{eff: } \Delta_2 W^{2,i}\, h_0^1 = -(\eta\chi_1\phi_{i,1}/N)(W_1^{3,i})^\top (W_1^4)^\top \langle h_1^1, h_0^1\rangle \in \Theta(1) \text{ along } h_0^{2,i} \qquad \text{(R2}\mu\text{-F3.2c)}$$
$$(\text{(R2}\mu\text{-U2.2)substitution;}$$
$$(W_1^{3,i})^\top (W_1^4)^\top \in \Theta(1) \text{ along } h_0^{2,i} \text{ via new intermediate at } t=1;$$
$$\langle h_1^1, h_0^1\rangle \in \Theta(N) \text{ coherent)}$$

$$\text{cross: } \Delta_2 W^{2,i}\, \Delta h^1 = -(\eta\chi_1\phi_{i,1}/N)(W_1^{3,i})^\top (W_1^4)^\top \langle h_1^1, \Delta h^1\rangle \in \Theta(1) \text{ along } h_0^{2,i} \qquad \text{(R2}\mu\text{-F3.2d)}$$
$$(\text{(R2}\mu\text{-U2.2)substitution;}$$
$$(W_1^{3,i})^\top (W_1^4)^\top \in \Theta(1) \text{ along } h_0^{2,i} \text{ via new intermediate at } t=1;$$
$$\langle h_1^1, \Delta h^1\rangle \in \Theta(N) \text{ coherent via } \Delta h^1 \parallel h_0^1 \text{ from (R2}\mu\text{-U2.1b);}$$
$$\text{leading-scale contribution)}$$

$\Delta h^{2,i} \in \Theta(1)$ entry-wise, with the coherent $\Theta(1)$-along-$h_0^{2,i}$ piece emerging from (R2μ-F3.2c) + (R2μ-F3.2d).

(R2μ-F3.3) $h_2^{3,i} = W_2^{3,i} h_2^{2,i}$, four-piece decomposition with cumulative $\Delta W^{3,i}$:

$$\text{init: } h_0^{3,i} \in \Theta(1). \quad \text{(R2}\mu\text{-F1.5)} \qquad \text{(R2}\mu\text{-F3.3a)}$$

$$\text{prop: } W_0^{3,i} \Delta h^{2,i} \in \Theta(1) \qquad \text{(R2}\mu\text{-F3.3b)}$$
$$(\sigma_3^2 \|\Delta h^{2,i}\|^2 = \Theta(1)\cdot\Theta(1) = \Theta(1);$$
$$\|\Delta h^{2,i}\|^2 \in \Theta(1) \text{ via (R2}\mu\text{-F3.2);}\ N_e = \Theta(1))$$

$$\text{eff: } \Delta_2 W^{3,i}\, h_0^{2,i} = -\eta N \chi_1 \phi_{i,1}\, \|h_0^{2,i}\|^2\, (W_1^4)^\top \in \Theta(1) \text{ along } (W_0^4)^\top \qquad \text{(R2}\mu\text{-F3.3c)}$$
$$(\text{(R2}\mu\text{-U2.3)substitution;}\ \|h_0^{2,i}\|^2 \in \Theta(1) \text{ via (R2}\mu\text{-F1.4);}$$
$$(W_1^4)^\top \in \Theta(1/N) \text{ dominated by } (W_0^4)^\top)$$

$$\text{cross: } \Delta_2 W^{3,i}\, \Delta h^{2,i} \text{ tracked as part of } D' \text{ in (R2}\mu\text{-F3.4).} \qquad \text{(R2}\mu\text{-F3.3d)}$$

$\Delta h^{3,i} \in \Theta(1)$ aligned along $(W_0^4)^\top$.

(R2μ-F3.4) $h_2^3 = A'_1 + A'_{2,1} + A'_{2,2} + A'_3 + D'$, where

$$A'_1 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, h_0^{3,i}, \qquad A'_{2,1} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, W_0^{3,i} W_0^{2,i}\, \Delta h^1,$$
$$A'_{2,2} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, W_0^{3,i}\, \Delta_2 W^{2,i}\, h_0^1, \qquad A'_3 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, \Delta_2 W^{3,i}\, h_0^{2,i},$$
$$D' := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, \Delta_2 W^{3,i}\, \Delta h^{2,i}.$$

$A'_1 \in \Theta(1/\sqrt{N})$. (cross-$i$ CLT on $\Theta(1)$-entry independent vectors; same as (R2μ-F2.6a)) (R2μ-F3.4a)

$A'_{2,1} \in \Theta(1/\sqrt{N})$ (R2μ-F3.4b)
(climbs from $\Theta(1/N)$ at $t=1$ because $\|\Delta h^1\|^2$ grows from $\Theta(1)$ to $\Theta(N)$ via (R2μ-U2.1b))

$A'_{2,2} \in \Theta(1/\sqrt{N})$ in mostly-random directions (R2μ-F3.4c)
(same mechanism as $A'_{2,1}$; $\|\Delta h^1\|^2 \in \Theta(N)$ via (R2μ-U2.1b); rank-deficient initialization Gram (§J.3.1) still does not collapse, so sub-leading)

$A'_3 = -\eta N \chi_1 (W_0^4)^\top \left(\tfrac{1}{M}\sum_i \phi_{i,2}\phi_{i,1} \|h_0^{2,i}\|^2\right) \in \Theta(1)$ along $(W_0^4)^\top$ (R2μ-F3.4d)
(same structure as (R2μ-F2.6d) with cumulative $\Delta_2 W^{3,i}$)

$D' = -\eta N \chi_1 (W_0^4)^\top \tfrac{1}{M}\sum_i \phi_{i,2}\phi_{i,1} \langle h_0^{2,i}, \Delta h^{2,i}\rangle$
$\in \Theta(1)$ along $(W_0^4)^\top$ (R2μ-F3.4e)
(**climbs from** $\Theta(1/N)$ at $t=1$ to $\Theta(1)$ at $t=2$; $\langle h_0^{2,i}, \Delta h^{2,i}\rangle \in \Theta(1)$ coherent across i via (R2μ-F3.2c)+(R2μ-F3.2d), replacing the $t=1$ random factor $\langle h_0^{3,i}, W_0^4\rangle \in \Theta(1/\sqrt{N})$ via (R2μ-B1.4); LLN gives empirical average $\Theta(1)$ rather than $\Theta(1/\sqrt{M})$)

$h_2^3 \in \Theta(1)$ along $(W_0^4)^\top$, from $A'_3 + D'$. (R2μ-F3.4)

$f_2 = W_2^4 h_2^3 \in \Theta(1)$. (R2$\mu$-F3.5)
($A'_3 + D'$ aligned along $(W_0^4)^\top$; $\|W_0^4\|^2 = 1/N$; net $\eta N \cdot \Theta(1) \cdot (1/N) = \Theta(1)$)

Third backward pass The structure parallels (R2μ-B2.1)–(R2μ-B2.7) with cumulative ΔW's. The key change at $t=2$ vs $t=1$:

$(\Delta W^4)^\top \in \Theta(1/N)$, climbs from $\Theta(1/N^{3/2})$ at $t=1$ (R2μ-B3.1)
($h_1^3 \in \Theta(1)$ via (R2μ-F2.6); $(\Delta_2 W^4)^\top \in \Theta(1/N)$ via (R2μ-U2.4); now comparable to $(W_0^4)^\top \in \Theta(1/N)$)

Combined with the alignment mechanism (driven by $\langle W_1^4, h_1^{3,i}\rangle \in \Theta(1)$ coherent across i via (R2μ-B2.4)), several pieces climb at $t=2$:

Router pathway $\partial f_2/\partial \phi_{i,2}$.

$$\partial f_2/\partial \phi_{i,2} = (1/M)\langle h_2^{3,i}, W_2^4\rangle \in \Theta(1/N), \tag{R2μ-B3.4}$$

four-piece grid (same decomposition as (R2μ-B2.4) at $t=2$) :

(init $\cdot$ init) $\langle (W_0^{3,i})\, h_2^{2,i}, W_0^4\rangle \in \Theta(1/\sqrt{N})$ random, unchanged from $t=1$;

(init $\cdot$ update) $\langle (W_0^{3,i})\, h_2^{2,i}, \Delta_2 W^4\rangle \in \Theta(1/\sqrt{N})$, climbs from $\Theta(1/N)$ at $t=1$

via $\langle h_0^{3,i}, A_3\rangle \in \Theta(\sqrt{N})$ ($A_3 \parallel (W_0^4)^\top$ from R2μ-F2.6d)

followed by the $1/N$ prefactor of $\Delta_2 W^4$;

(update $\cdot$ init) $\langle (\Delta_2 W^{3,i})\, h_2^{2,i}, W_0^4\rangle \in \Theta(1)$ via rank-1 alignment, unchanged;

(update $\cdot$ update) $\langle (\Delta_2 W^{3,i})\, h_2^{2,i}, \Delta_2 W^4\rangle \in \Theta(1)$, climbs from $\Theta(1/N)$ at $t=1$

via $\langle W_0^4, h_1^3\rangle \in \Theta(1)$ from A_3

($\kappa \in \Theta(N)$ rank-1 prefactor times $\langle W_0^4, \Delta_2 W^4\rangle \in \Theta(1/N)$);

$/M$: $(\Theta(1/N^{3/2}), \Theta(1/N^{3/2}), \Theta(1/N), \Theta(1/N))$ respectively;

both update $\cdot *$ pieces dominate; total $\Theta(1/N)$;

the entire $W_0^{3,i}$ row stays at $\Theta(1/N^{3/2})$: $\sqrt{N}$ deficit preserved.

Climbing pieces in expert pathway $(\partial f_2/\partial h_2^1)_{\exp}$.

$$A_{4.2} \in \Theta(1/N^{3/2}), \quad \text{climbs from } \Theta(1/N^2) \text{ at } t=1 \tag{R2μ-B3.5a.2}$$

($\sqrt{N}$ lift via (R2μ-B3.1); still sub-leading vs $A_{4.1}$)

$$A_{5.2} \in \Theta(1/N), \quad \text{climbs from } \Theta(1/N^2) \text{ at } t=1 \tag{R2μ-B3.5b.2}$$

(combined $\sqrt{N}$ lift via (R2μ-B3.1) and alignment-driven mechanism;

$\langle W_1^4, h_1^{3,i}\rangle \in \Theta(1)$ coherent via alignment of $h_1^{3,i}$ along $(W_0^4)^\top$

from (R2μ-F2.5c); joins $A_{5.1}$ at leading scale)

$$A_{6.1} \in \Theta(1/N^2) \text{ at every step}; \quad A_{6.2} \in \Theta(1/N^3) \text{ at } t=1,\ \Theta(1/N^2) \text{ at } t=2 \tag{R2μ-B3.5c}$$

($A_{6.1} : S_i := W_0^4 \cdot M_W \cdot (W_0^4)^\top$ has trace-induced coherent piece

$\sigma_4^2 \cdot \mathrm{tr}(M_W) = \Theta(1/N)$ at every step, LLN preserves $\Theta(1/N^2)$;

$A_{6.2}$ at $t=1 : S_i$ random ($\Delta_1 W^4$ indep of W_0^4), cross-i CLT gives $1/\sqrt{N}$

combined with $\Delta_1 W^4$ smaller by $1/\sqrt{N}$ via (R2μ-U1.4) $\Rightarrow \Theta(1/N^3)$;

$A_{6.2}$ at $t=2 : \Delta W^4$ acquires alignment with W_0^4 via h_1^3

(FL piece in (R2μ-F2.5c)), restoring trace-induced coherent piece in S_i;

combined with ΔW^4 climb to $\Theta(1/N)$ via (R2μ-B3.1) gives $\Theta(1/N^2)$;

both stay sub-leading vs $A_{5.x}, E_x \in \Theta(1/N)$)

$$E_1 \in \Theta(1/N), \quad \text{climbs from } \Theta(1/N^2) \text{ at } t=1 \tag{R2μ-B3.5d.1}$$

(alignment-driven climb;

inner product structure inside E no longer CLT-cancels but LLN-accumulates

once $\langle W_1^4, h_1^{3,i}\rangle \in \Theta(1)$ is coherent via (R2μ-B2.4);

joins leading scale)

$$E_2 \in \Theta(1/N), \quad \text{climbs from } \Theta(1/N^{5/2}) \text{ at } t=1 \tag{R2μ-B3.5d.2}$$

(combined alignment-driven climb and $\sqrt{N}$ lift via (R2μ-B3.1);

joins leading scale)

$$(\partial f_2/\partial h_2^1)_{\exp} \in \Theta(1/N), \text{ from } A_{5.1}, A_{5.2}, E_1, E_2. \tag{R2μ-B3.5}$$

Router pathway.

$$(\partial f_2/\partial h_2^1)_{\text{router}} \in \Theta(1/N), \text{ all four pieces at leading scale.} \tag{R2μ-B3.6}$$

The $\mathbf{v}^{(\Delta)}$-pieces lift via (R2μ-B3.1); the $\Delta_t Q^\top \mathbf{v}^{(0)}$ piece remains coherent via the rank-1 alignment of $\Delta_t Q$ with $\mathbf{v}^{(0)}$ (extending (R2μ-B2.6c) to cumulative $\Delta_t Q$).

$W^{3,i}$-imbalance preserved at $t = 2$. Define $v_i := \dot{\phi}_{i,2}\langle h_2^{3,i}, W_2^4\rangle$ and split $v_i = v_i^I + v_i^U$ where

$$v_i^I := \dot{\phi}_{i,2}\langle (W_0^{3,i}) h_2^{2,i}, W_2^4\rangle,$$
$$v_i^U := \dot{\phi}_{i,2}\langle (\Delta_2 W^{3,i}) h_2^{2,i}, W_2^4\rangle.$$

$v_i^I \in \Theta(1/\sqrt{N})$ random across i (both $(W_0^{3,i}, W_0^4)$ and $(W_0^{3,i}, \Delta_2 W^4)$ entries are $\Theta(1/\sqrt{N})$, the latter via $\langle h_0^{3,i}, A_3\rangle \in \Theta(\sqrt{N})$ followed by the $1/N$ prefactor of $\Delta_2 W^4$). $v_i^U \in \Theta(1)$ coherent. Cross-layer CLT then gives $(1/M) Q_2^\top v^I \in \Theta(1/N^{3/2})$ (sub-leading) and $(1/M) Q_2^\top v^U \in \Theta(1/N)$ (leading). The entire $W_0^{3,i}$-row of the router pathway sits a factor $\sqrt{N}$ below the $\Delta_2 W^{3,i}$-row.

$$\partial f_2/\partial h_2^1 \in \Theta(1/N), \tag{R2μ-B3.7}$$

with leading-scale contributions from $A_{5.1}, A_{5.2}, E_1, E_2$ in the expert pathway and all four pieces in the router pathway.

J.3.2 Summary tables of signal propagation for μP in Regime II

Notation: $\Delta_t W^\ell$ denotes the cumulative update $W_t^\ell - W_0^\ell$. Green = at the proper feature-learning scale at $t = 2$; red = sub-leading throughout.

Forward.

Quantity	$t=0$	$t=1$	$t=2$
$h_t^1 = h_0^1 + \Delta_t h^1$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $h_0^1 = W_0^1 x$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
effective: $\Delta_t h^1 = \Delta_t W^1\, x$	0	$\Theta(1/\sqrt{N})$	$\Theta(1)$
ψ_t, ϕ_t	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$h_t^{2,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $h_0^{2,i} = W_0^{2,i} h_0^1$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
propagating: $W_0^{2,i} \Delta_t h^1$	0	$\Theta(1/\sqrt{N})$	$\Theta(1)$
effective: $\Delta_t W^{2,i}\, h_0^1$	0	$\Theta(1/\sqrt{N})$	$\Theta(1)$
cross: $\Delta_t W^{2,i}\, \Delta_t h^1$	0	$\Theta(1/N^{3/2})$	$\Theta(1)$
$h_t^{3,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $h_0^{3,i} = W_0^{3,i} h_0^{2,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
propagating: $W_0^{3,i} \Delta_t h^{2,i}$	0	$\Theta(1/\sqrt{N})$	$\Theta(1)$
effective: $\Delta_t W^{3,i}\, h_0^{2,i}$	0	$\Theta(1)$	$\Theta(1)$
cross: $\Delta_t W^{3,i}\, \Delta_t h^{2,i}$	0	$\Theta(1/\sqrt{N})$	$\Theta(1)$
$h_t^3 = A_1 + A_{2,1} + A_{2,2} + A_3 + D$	$\Theta(1/\sqrt{N})$	$\Theta(1)$	$\Theta(1)$
$A_1 = (1/M)\sum_i \phi_{i,t}\, h_0^{3,i}$	$\Theta(1/\sqrt{N})$	$\Theta(1/\sqrt{N})$	$\Theta(1/\sqrt{N})$
$A_{2,1} = (1/M)\sum_i \phi_{i,t}\, W_0^{3,i} W_0^{2,i}\, \Delta_t h^1$	0	$\Theta(1/N)$	$\Theta(1/\sqrt{N})$
$A_{2,2} = (1/M)\sum_i \phi_{i,t}\, W_0^{3,i}\, \Delta_t W^{2,i}\, h_0^1$	0	$\Theta(1/N)$	$\Theta(1/\sqrt{N})$
$A_3 = (1/M)\sum_i \phi_{i,t}\, \Delta_t W^{3,i}\, h_0^{2,i}$	0	$\Theta(1)$	$\Theta(1)$
$D = (1/M)\sum_i \phi_{i,t}\, \Delta_t W^{3,i}\, \Delta_t h^{2,i}$	0	$\Theta(1/N)$	$\Theta(1)$
$f_t = W_t^4\, h_t^3$	$\Theta(1/N)$	$\Theta(1)$	$\Theta(1)$

Backward.

Quantity	$t=0$	$t=1$	$t=2$
$\partial f_t/\partial h_t^3 = (W_0^4)^\top + (\Delta_t W^4)^\top$	$\Theta(1/N)$	$\Theta(1/N)$	$\Theta(1/N)$
init: $(W_0^4)^\top$	$\Theta(1/N)$	$\Theta(1/N)$	$\Theta(1/N)$
update: $(\Delta_t W^4)^\top$	0	$\Theta(1/N^{3/2})$	$\Theta(1/N)$
$\partial f_t/\partial h_t^{3,i} = (\phi_{i,t}/M)(W_t^4)^\top$	$\Theta(1/N^2)$	$\Theta(1/N^2)$	$\Theta(1/N^2)$
init: $(\phi_{i,t}/M)(W_0^4)^\top$	$\Theta(1/N^2)$	$\Theta(1/N^2)$	$\Theta(1/N^2)$
update: $(\phi_{i,t}/M)(\Delta_t W^4)^\top$	0	$\Theta(1/N^{5/2})$	$\Theta(1/N^2)$
$\partial f_t/\partial h_t^{2,i} = (\phi_{i,t}/M)(W_t^{3,i})^\top(W_t^4)^\top$	$\Theta(1/N^{3/2})$	$\Theta(1/N)$	$\Theta(1/N)$
init·init: $(\phi_{i,t}/M)(W_0^{3,i})^\top(W_0^4)^\top$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$
init·update: $(\phi_{i,t}/M)(W_0^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N^{3/2})$
update·init: $(\phi_{i,t}/M)(\Delta_t W^{3,i})^\top(W_0^4)^\top$	0	$\Theta(1/N)$	$\Theta(1/N)$
update·update: $(\phi_{i,t}/M)(\Delta_t W^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$\partial f_t/\partial \phi_{i,t} = (1/M)\langle h_t^{3,i}, W_t^4\rangle$	$\Theta(1/N^{3/2})$	$\Theta(1/N)$	$\Theta(1/N)$
init·init: $(1/M)\langle (W_0^{3,i})\, h_t^{2,i}, W_0^4\rangle$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$
init·update: $(1/M)\langle (W_0^{3,i})\, h_t^{2,i}, \Delta_t W^4\rangle$	0	$\Theta(1/N^2)$	$\Theta(1/N^{3/2})$
update·init: $(1/M)\langle (\Delta_t W^{3,i})\, h_t^{2,i}, W_0^4\rangle$	0	$\Theta(1/N)$	$\Theta(1/N)$
update·update: $(1/M)\langle (\Delta_t W^{3,i})\, h_t^{2,i}, \Delta_t W^4\rangle$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$(\partial f_t/\partial h_t^1)_{\text{exp}}$	$\Theta(1/N^{3/2})$	$\Theta(1/N)$	$\Theta(1/N)$
$A_{4.1} = (1/M)\sum_i \phi_{i,t}\,(W_0^{2,i})^\top(W_0^{3,i})^\top(W_0^4)^\top$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$
$A_{4.2} = (1/M)\sum_i \phi_{i,t}\,(W_0^{2,i})^\top(W_0^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N^{3/2})$
$A_{5.1} = (1/M)\sum_i \phi_{i,t}\,(W_0^{2,i})^\top(\Delta_t W^{3,i})^\top(W_0^4)^\top$	0	$\Theta(1/N)$	$\Theta(1/N)$
$A_{5.2} = (1/M)\sum_i \phi_{i,t}\,(W_0^{2,i})^\top(\Delta_t W^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$A_{6.1} = (1/M)\sum_i \phi_{i,t}\,(\Delta_t W^{2,i})^\top(W_0^{3,i})^\top(W_0^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N^2)$
$A_{6.2} = (1/M)\sum_i \phi_{i,t}\,(\Delta_t W^{2,i})^\top(W_0^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^3)$	$\Theta(1/N^2)$
$E_1 = (1/M)\sum_i \phi_{i,t}\,(\Delta_t W^{2,i})^\top(\Delta_t W^{3,i})^\top(W_0^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$E_2 = (1/M)\sum_i \phi_{i,t}\,(\Delta_t W^{2,i})^\top(\Delta_t W^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^{5/2})$	$\Theta(1/N)$
$(\partial f_t/\partial h_t^1)_{\text{router}} = (1/M)Q_t^\top[\dot\phi_{i,t}\langle h_t^{3,i}, W_t^4\rangle]_i$	$\Theta(1/N^{3/2})$	$\Theta(1/N)$	$\Theta(1/N)$
$(1/M)Q_0^\top[\dot\phi_{i,t}\langle (W_0^{3,i})h_t^{2,i}, W_0^4\rangle]_i$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$
$(1/M)Q_0^\top[\dot\phi_{i,t}\langle (W_0^{3,i})h_t^{2,i}, \Delta_t W^4\rangle]_i$	0	$\Theta(1/N^2)$	$\Theta(1/N^{3/2})$
$(1/M)Q_0^\top[\dot\phi_{i,t}\langle (\Delta_t W^{3,i})h_t^{2,i}, W_0^4\rangle]_i$	0	$\Theta(1/N)$	$\Theta(1/N)$
$(1/M)Q_0^\top[\dot\phi_{i,t}\langle (\Delta_t W^{3,i})h_t^{2,i}, \Delta_t W^4\rangle]_i$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$(1/M)\Delta_t Q^\top[\dot\phi_{i,t}\langle (W_0^{3,i})h_t^{2,i}, W_0^4\rangle]_i$	0	$\Theta(1/N^2)$	$\Theta(1/N^2)$
$(1/M)\Delta_t Q^\top[\dot\phi_{i,t}\langle (W_0^{3,i})h_t^{2,i}, \Delta_t W^4\rangle]_i$	0	$\Theta(1/N^3)$	$\Theta(1/N^2)$
$(1/M)\Delta_t Q^\top[\dot\phi_{i,t}\langle (\Delta_t W^{3,i})h_t^{2,i}, W_0^4\rangle]_i$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$(1/M)\Delta_t Q^\top[\dot\phi_{i,t}\langle (\Delta_t W^{3,i})h_t^{2,i}, \Delta_t W^4\rangle]_i$	0	$\Theta(1/N^3)$	$\Theta(1/N)$
$\partial f_t/\partial h_t^1 = (\partial f_t/\partial h_t^1)_{\text{exp}} + (\partial f_t/\partial h_t^1)_{\text{router}}$	$\Theta(1/N^{3/2})$	$\Theta(1/N)$	$\Theta(1/N)$

J.3.3 Deriving MSSP in Regime II

This subsection presents the MSSP-Regime-II derivation per Definition J.2, with $W_0^4 = 0$. MSSP-Regime-II differs from μP-Regime-II (Definition J.1) in exactly two places: the variance of $W_0^{3,i}$ is M/N_e (entries $\Theta(\sqrt{N})$) rather than $1/N_e$ (entries $\Theta(1)$); and $W_0^4 = 0$ rather than variance $1/N^2$. All other init variances, learning rates, gating, and standing assumptions are identical to μP-Regime-II.

Two structural consequences propagate through the derivation:

1. The boosted $W_0^{3,i}$ variance gives $h_0^{3,i}$ entries of scale $\Theta(\sqrt{N})$ (vs $\Theta(1)$ in μP), which after $1/M$-aggregation with the cross-i CLT gives h_0^3 entries $\Theta(1)$. So the aggregated activation is at the leading scale at init, even though the prediction $f_0 = 0$ vanishes.
2. Since $W_0^4 = 0$, every gradient at $t = 0$ that carries a W_0^4 factor is zero, so only W^4 updates at $t = 1$. Hidden weights first move at $t = 2$. Three forward + three backward passes are needed to capture all post-update correlations.

Compared to μP-Regime-II: MSSP-Regime-II's larger $W^{3,i}$ init means the bottleneck-layer expert weights start big and the per-expert hidden state $h^{3,i}$ inherits that scale; updates to $W^{3,i}$ are then sub-leading relative to its init (vs comparable in μP-Regime-II). The aggregated h^3 still feature-learns at $\Theta(1)$, just through a different mechanism: the prediction reaches $f_1 \in \Theta(1)$ at $t = 1$ purely from the $W^4 \cdot h_0^3$ contraction (since $h_0^3 \in \Theta(1)$ entries, $\|h_0^3\|^2 \in \Theta(N)$), with no hidden updates required.

Setup and standing assumptions

Architecture. Same as in §J.3.1.

Initialization. Per Definition J.2 with $W_0^4 = 0$:

$$(W_0^1)_{ab} \sim \mathcal{N}(0, 1/D), \qquad (Q_0)_{ia} \sim \mathcal{N}(0, 1/N),$$
$$(W_0^{2,i})_{ab} \sim \mathcal{N}(0, 1/N), \qquad (W_0^{3,i})_{ab} \sim \mathcal{N}(0, M/N_e),$$
$$W_0^4 = 0.$$

All parameters drawn independently across i. Resulting entry scales: $W_0^4 = 0$, $W_0^{2,i} \in \Theta(1/\sqrt{N})$, $W_0^{3,i} \in \Theta(\sqrt{M}) = \Theta(\sqrt{N})$, $W_0^1 \in \Theta(1)$, $Q_0 \in \Theta(1/\sqrt{N})$.

Learning rates (SGD). Same as μP-Regime-II: $\eta_1 = \eta N$, $\eta_Q = \eta M/N$, $\eta_2 = \eta M/N$, $\eta_3 = \eta MN$, $\eta_4 = \eta/N$.

Gating. Sigmoid + $1/M$ aggregation: $\phi_{i,t}, \dot{\phi}_{i,t} \in \Theta(1)$.

Auxiliary scaling lemmas We reuse CLT, LLN, and Lemma J.7.

Specialization of Lemma J.7 to MSSP-Regime-II. For $W = W_0^{3,i} \in \mathbb{R}^{N \times N_e}$ with $\sigma_W^2 = M/N_e$ and $N_e = \Theta(1)$:

- $W_0^{3,i}(W_0^{3,i})^\top \in \mathbb{R}^{N \times N}$ has rank $\le N_e = \Theta(1)$. Diagonal mean $n\sigma^2 = M = \Theta(N)$, off-diagonal entries variance $n\sigma^4 = M^2/N_e = \Theta(N^2)$, coordinate scale $\Theta(N)$. **Every entry is** $\Theta(N)$ — much larger than μP-Regime-II's $\Theta(1)$ entries.
- $(W_0^{3,i})^\top W_0^{3,i} \in \mathbb{R}^{N_e \times N_e}$, rank-$N_e$ full. Diagonal mean $m\sigma^2 = NM/N_e = \Theta(N^2)$, off-diagonal coord scale $\Theta(\sigma^2\sqrt{m}) = \Theta(N^{3/2})$. So acting on a $\Theta(1)$-coordinate N_e-vector $h_0^{2,i}$ gives entries of coordinate scale $\Theta(N) \cdot \Theta(1) + \Theta(N^{3/2}) \cdot \Theta(1) = \Theta(N^2)$ at leading.

The other Gram products ($W_0^{2,i}$ versions) are unchanged from μP-Regime-II.

First forward pass

$$h_0^1 = W_0^1 x \in \Theta(1). \quad (\text{CLT};\ \sigma_1^2 = 1/D,\ \|x\|^2 \in \Theta(D)) \tag{R2s-F1.1}$$

$$\psi_0 = Q_0 h_0^1 \in \Theta(1). \quad (\text{CLT};\ \sigma_Q^2 = 1/N,\ \|h_0^1\|^2 \in \Theta(N)) \tag{R2s-F1.2}$$

$$\phi_{i,0} = \sigma(\psi_{i,0}) \in \Theta(1). \quad (\text{sigmoid bounded; gating assumption}) \tag{R2s-F1.3}$$

$$h_0^{2,i} = W_0^{2,i} h_0^1 \in \Theta(1), \quad \text{independent across } i$$
$$(\text{CLT};\ \sigma_2^2 = 1/N,\ \|h_0^1\|^2 \in \Theta(N);$$
$$\|h_0^{2,i}\|^2 = N_e \cdot \Theta(1) = \Theta(1) \text{ since } N_e = \Theta(1)). \tag{R2s-F1.4}$$

$$h_0^{3,i} = W_0^{3,i} h_0^{2,i} \in \Theta(\sqrt{N}), \quad \|h_0^{3,i}\|^2 \in \Theta(N^2) \quad (\textbf{boosted by MSSP variance}) \tag{R2s-F1.5}$$
$$(\text{CLT};\ \sigma_3^2 = M/N_e = \Theta(N);\ \|h_0^{2,i}\|^2 \in \Theta(1) \text{ via (R2}s\text{-F1.4)};$$
$$\text{variance per entry } \Theta(N) \cdot \Theta(1) = \Theta(N))$$

$$h_0^3 = (1/M)\sum_i \phi_{i,0}\, h_0^{3,i} \in \Theta(1) \qquad \text{(R2}s\text{-F1.6)}$$

(cross-i CLT on $h_0^{3,i}$ entries $\Theta(\sqrt{N})$ via (R2s-F1.5);
$\{h_0^{3,i}\}$ independent across i;
variance per entry $(1/M^2)\sum_i \phi_{i,0}^2\,\Theta(N) = \Theta(N/M) = \Theta(1)$;
$\sqrt{M}$ cross-i CLT cancellation balances the $\sqrt{M}$ amplification of $h_0^{3,i}$)

$$f_0 = W_0^4\, h_0^3 = 0. \quad (W_0^4 = 0 \text{ by MSSP-Regime-II standing assumption}) \qquad \text{(R2}s\text{-F1.7)}$$

First backward pass and step-1 updates By the chain rule, every hidden gradient at $t = 0$ ($\partial f_0/\partial h_0^3$, $\partial f_0/\partial h_0^{3,i}$, $\partial f_0/\partial h_0^{2,i}$, $\partial f_0/\partial \phi_{i,0}$, and both expert and router contributions to $\partial f_0/\partial h_0^1$) carries a $W_0^4 = 0$ factor and is therefore zero. Hence $W_1^\ell = W_0^\ell$ for $\ell \in \{1, Q, (2,i), (3,i)\}$; only W^4 is updated at $t = 1$.

Step-1 parameter update.

$$\Delta_1 W^4 = -(\eta\chi_0/N)\,(h_0^3)^\top \in \Theta(1/N). \quad (h_0^3 \in \Theta(1) \text{ via (R2}s\text{-F1.6)}) \qquad \text{(R2}s\text{-U1.4)}$$

Second forward pass Since the non-readout weights are unchanged at $t = 1$, every hidden activation is unchanged: $h_1^\ell = h_0^\ell$ for $\ell \in \{1, (2,i), (3,i), 3\}$, $\psi_1 = \psi_0$, $\phi_1 = \phi_0$.

$$f_1 = W_1^4\, h_0^3 = -\tfrac{\eta\chi_0}{N}\,\|h_0^3\|^2 \in \Theta(1). \quad (\|h_0^3\|^2 \in \Theta(N) \text{ via (R2}s\text{-F1.6)}) \qquad \text{(R2}s\text{-F2.7)}$$

Second backward pass and step-2 updates

New intermediate at $t = 1$.

$(W_0^{3,i})^\top h_0^3 \in \Theta(N)$ entry-wise
(using (R2s-F1.6) to expand $h_0^3 = (1/M)\sum_j \phi_{j,0} h_0^{3,j}$;
$j = i$ piece $(\phi_{i,0}/M)(W_0^{3,i})^\top W_0^{3,i} h_0^{2,i}$ with
$(W_0^{3,i})^\top W_0^{3,i} h_0^{2,i} \in \Theta(N^2)$ from Lemma J.7 specialization,
giving $(\phi_{i,0}/M)\cdot\Theta(N^2) = \Theta(N)$;
$j \neq i$ pieces give cross-j noise of comparable $\Theta(N)$ scale)

$\Rightarrow\ (W_0^{3,i})^\top (W_1^4)^\top = -(\eta\chi_0/N)(W_0^{3,i})^\top h_0^3 \in \Theta(1)$ entry-wise
(W_1^4 via (R2s-U1.4))

Per-layer gradients.

$$\partial f_1/\partial h_1^3 = (W_1^4)^\top \in \Theta(1/N). \quad (\text{via (R2}s\text{-U1.4)}) \qquad \text{(R2}s\text{-B2.1)}$$

$$\partial f_1/\partial h_1^{3,i} = (\phi_{i,0}/M)(W_1^4)^\top \in \Theta(1/N^2). \quad (\phi_{i,0} \in \Theta(1);\ M = \Theta(N)) \qquad \text{(R2}s\text{-B2.2)}$$

$$\partial f_1/\partial h_1^{2,i} = (\phi_{i,0}/M)(W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1/N) \qquad \text{(R2}s\text{-B2.3)}$$
$((W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1)$ via the new intermediate above)

$$\partial f_1/\partial \phi_{i,0} = (1/M)\langle h_0^{3,i}, W_1^4\rangle = -(\eta\chi_0/(MN))\langle h_0^{3,i}, h_0^3\rangle \in \Theta(1/N) \tag{R2s-B2.4}$$

(Four-piece grid: only (init · update) is non-zero at $t = 1$,
since $W_0^4 = 0$ kills the * · init column
and $\Delta_1 W^{3,i} = 0$ (hidden weights frozen) kills the update · * row;
(init · update) $\langle (W_0^{3,1})\, h_0^{2,i}, \Delta_1 W^4\rangle \in \Theta(1)$:
expanding $h_0^3 = (1/M)\sum_j \phi_{j,0} h_0^{3,j}$ gives
$\langle h_0^{3,i}, h_0^3\rangle = (\phi_{i,0}/M)\|h_0^{3,i}\|^2 + (1/M)\sum_{j\neq i}\phi_{j,0}\langle h_0^{3,i}, h_0^{3,j}\rangle$;
$j = i$ diagonal term: $(\phi_{i,0}/M)\|h_0^{3,i}\|^2 = (1/M)\Theta(N^2) = \Theta(N)$
using $\|h_0^{3,i}\|^2 \in \Theta(N^2)$ via (R2s-F1.5);
$j \neq i$ off-diagonal terms give comparable $\Theta(N)$ via cross-j CLT;
so $\langle h_0^{3,i}, h_0^3\rangle \in \Theta(N)$;
substituting: $\partial f_1/\partial \phi_{i,0} = -(\eta\chi_0/(MN)) \cdot \Theta(N) = \Theta(1/M) = \Theta(1/N)$;
lifted from μP-R2's (init · update) $= \Theta(1/N^2)$ entry
by the boost $\sigma_3^2 = M/N_e$)

$$(\partial f_1/\partial h_1^1)_{\text{exp}} = (1/M)\textstyle\sum_i \phi_{i,0}(W_0^{2,i})^\top (W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1/N) \tag{R2s-B2.5}$$

($(W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1)$ via the new intermediate above;
$(W_0^{2,i})^\top$ acts on a $\Theta(1)$ vector with $\|v\|^2 = N_e\Theta(1) = \Theta(1)$;
variance per entry $(1/N)\Theta(1) = \Theta(1/N)$, coord $\Theta(1/\sqrt{N})$ random per i;
cross-i CLT: $\Theta(1/N)$)

$$(\partial f_1/\partial h_1^1)_{\text{router}} = (1/M)Q_0^\top v, \quad v_i := \dot\phi_{i,0}\langle h_0^{3,i}, W_1^4\rangle \in \Theta(1/N) \tag{R2s-B2.6}$$

($v_i \in \Theta(1)$ via (R2s-B2.4); $\|v\|^2 = M\Theta(1) = \Theta(M)$;
$Q_0^\top v$ entries variance $(1/N)\Theta(M) = \Theta(1)$ via cross-layer CLT;
$(1/M)\Theta(1) = \Theta(1/N)$;
W^{3,i}-split $v_i = v_i^I + v_i^U$ where
$v_i^I := \dot\phi_{i,0}\langle (W_0^{3,1}) h_0^{2,i}, W_1^4\rangle$
$v_i^U := \dot\phi_{i,0}\langle (\Delta_1 W^{3,i}) h_0^{2,i}, W_1^4\rangle = 0$ since $\Delta_1 W^{3,i} = 0$;
$v_i^I \in \Theta(1)$ coherent via boost (cf. R2s-B2.4);
$(1/M)Q_0^\top v^I = \Theta(1/N)$ and $(1/M)Q_0^\top v^U = 0$)

$$\partial f_1/\partial h_1^1 = (\partial f_1/\partial h_1^1)_{\text{exp}} + (\partial f_1/\partial h_1^1)_{\text{router}} \in \Theta(1/N). \tag{R2s-B2.7}$$

Step-2 parameter updates.

$$\Delta_2 W^4 = -(\eta\chi_1/N)(h_0^3)^\top \in \Theta(1/N). \quad (h_0^3 \in \Theta(1) \text{ via (R2}s\text{-F1.6)}) \tag{R2s-U2.4}$$

$$\Delta_2 W^{3,i} = -\eta N \chi_1 \phi_{i,0}\, (W_1^4)^\top (h_0^{2,i})^\top \in \Theta(1) \text{ rank-1} \tag{R2s-U2.3}$$

($(W_1^4)^\top \in \Theta(1/N)$ via (R2s-U1.4); $h_0^{2,i} \in \Theta(1)$ via (R2s-F1.4);
sub-leading vs $W_0^{3,i} \in \Theta(\sqrt{N})$ via (R2s-F1.5))

$$\Delta_2 W^{2,i} = -(\eta\chi_1\phi_{i,0}/N)(W_0^{3,i})^\top (W_1^4)^\top (h_0^1)^\top \in \Theta(1/N) \text{ rank-1} \tag{R2s-U2.2}$$

($(W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1)$ via the new intermediate;
sub-leading vs $W_0^{2,i} \in \Theta(1/\sqrt{N})$)

$$\Delta_2 W^1 = -\eta_1 \chi_1 (\partial f_1 / \partial h_1^1)\, x^\top \in \Theta(1) \tag{R2s-U2.1a}$$
$$(\eta_1 = \eta N;\ \partial f_1/\partial h_1^1 \in \Theta(1/N) \text{ via (R2}s\text{-B2.7)})$$
$$\Delta_2 h^1 = \Delta_2 W^1\, x \in \Theta(1),\ \text{aligned along } h_0^1. \tag{R2s-U2.1b}$$
$$\Delta_2 Q \in \Theta(1/N). \quad (\partial f_1/\partial \phi \in \Theta(1/N) \text{ via (R2}s\text{-B2.4)}) \tag{R2s-U2.Q}$$

Third forward pass

$$h_2^1 = h_0^1 + \Delta_2 h^1 \in \Theta(1). \quad (\Delta_2 h^1 \in \Theta(1) \text{ via (R2}s\text{-U2.1b)};\ h_0^1 \in \Theta(1)) \tag{R2s-F3.1}$$

(R2s-F3.2) $h_2^{2,i} = W_2^{2,i} h_2^1$, four-piece decomposition:

$$h_2^{2,i} = h_0^{2,i} + W_0^{2,i}\Delta_2 h^1 + \Delta_2 W^{2,i} h_0^1 + \Delta_2 W^{2,i}\Delta_2 h^1. \tag{R2s-F3.2}$$
$$\text{init: } h_0^{2,i} \in \Theta(1). \quad (\text{R2}s\text{-F1.4}) \tag{R2s-F3.2a}$$

$$\text{prop: } W_0^{2,i}\Delta_2 h^1 \in \Theta(1)$$
$$(\sigma_2^2 \|\Delta_2 h^1\|^2 = (1/N)\Theta(N) = \Theta(1);$$
$$\|\Delta_2 h^1\|^2 \in \Theta(N) \text{ via (R2}s\text{-U2.1b)}). \tag{R2s-F3.2b}$$

$$\text{eff: } \Delta_2 W^{2,i} h_0^1 = -(\eta\chi_1\phi_{i,1}/N)\|h_0^1\|^2 (W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1)$$
$$(\text{(R2}s\text{-U2.2)substitution;}$$
$$(W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1) \text{ via the new intermediate at } t=1;$$
$$\|h_0^1\|^2 \in \Theta(N)). \tag{R2s-F3.2c}$$

$$\text{cross: } \Delta_2 W^{2,i}\Delta_2 h^1 = -(\eta\chi_1\phi_{i,1}/N)\langle h_0^1, \Delta_2 h^1\rangle (W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1)$$
$$((W_0^{3,i})^\top (W_1^4)^\top \in \Theta(1) \text{ via the new intermediate;}$$
$$\langle h_0^1, \Delta_2 h^1\rangle \in \Theta(N) \text{ coherent via (R2}s\text{-U2.1b)}). \tag{R2s-F3.2d}$$

$\Delta_2 h^{2,i} \in \Theta(1)$ entry-wise.

(R2s-F3.3) $h_2^{3,i} = W_2^{3,i} h_2^{2,i}$, four-piece decomposition:

$$h_2^{3,i} = h_0^{3,i} + W_0^{3,i}\Delta_2 h^{2,i} + \Delta_2 W^{3,i} h_0^{2,i} + \Delta_2 W^{3,i}\Delta_2 h^{2,i}. \tag{R2s-F3.3}$$
$$\text{init: } h_0^{3,i} \in \Theta(\sqrt{N}). \quad (\text{R2}s\text{-F1.5}) \tag{R2s-F3.3a}$$

$$\text{prop: } W_0^{3,i}\Delta_2 h^{2,i} \in \Theta(\sqrt{N})$$
$$(\sigma_3^2 \|\Delta_2 h^{2,i}\|^2 = \Theta(N)\cdot\Theta(1) = \Theta(N);$$
$$\|\Delta_2 h^{2,i}\|^2 = N_e\Theta(1) = \Theta(1) \text{ via (R2}s\text{-F3.2)};\ N_e = \Theta(1)). \tag{R2s-F3.3b}$$

$$\text{eff: } \Delta_2 W^{3,i} h_0^{2,i} = -\eta N \chi_1 \phi_{i,1} \|h_0^{2,i}\|^2 (W_1^4)^\top \in \Theta(1)$$
$$(\text{(R2}s\text{-U2.3)substitution;}$$
$$\|h_0^{2,i}\|^2 \in \Theta(1) \text{ via (R2}s\text{-F1.4);}$$
$$(W_1^4)^\top \in \Theta(1/N) \text{ via (R2}s\text{-U1.4)}). \tag{R2s-F3.3c}$$

$$\text{cross: } \Delta_2 W^{3,i}\Delta_2 h^{2,i} \text{ tracked as part of } D \text{ in (R2}s\text{-F3.4)}. \tag{R2s-F3.3d}$$

Propagating dominates: $\Delta_2 h^{3,i} \in \Theta(\sqrt{N})$ — same scale as $h_0^{3,i}$ at init. The effective piece $\Theta(1)$ is sub-leading per-expert but contributes coherently to the aggregation in (R2s-F3.4d).

(R2s-F3.4) $h_2^3 = A_1 + A_{2,1} + A_{2,2} + A_3 + D$, where

$$A_1 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2} h_0^{3,i}, \qquad A_{2,1} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2} W_0^{3,i} W_0^{2,i}\Delta_2 h^1,$$
$$A_{2,2} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2} W_0^{3,i}\Delta_2 W^{2,i} h_0^1, \qquad A_3 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\Delta_2 W^{3,i} h_0^{2,i},$$
$$D := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\Delta_2 W^{3,i}\Delta_2 h^{2,i}.$$

$$A_1 \in \Theta(1). \quad \text{(cross-}i\text{ CLT, same calculation as (R2}s\text{-F1.6))} \tag{R2s-F3.4a}$$

$$\begin{aligned} A_{2,1} &\in \Theta(1) \\ &\quad \text{(per summand } W_0^{3,i} W_0^{2,i} \Delta_2 h^1 \in \Theta(\sqrt{N}) \text{ random;} \\ &\qquad \|\Delta_2 h^1\|^2 \in \Theta(N) \text{ via (R2}s\text{-U2.1b);} \\ &\qquad \text{cross-}i\text{ CLT: } \Theta(\sqrt{N}/\sqrt{M}) = \Theta(1)). \end{aligned} \tag{R2s-F3.4b}$$

$$\begin{aligned} A_{2,2} &\in \Theta(1) \\ &\quad \text{(per summand } W_0^{3,i}[\Delta_2 W^{2,i} h_0^1] \in \Theta(\sqrt{N}) \\ &\qquad \text{since } \Delta_2 W^{2,i} h_0^1 \in \Theta(1) \text{ via (R2}s\text{-F3.2c);} \\ &\qquad \text{cross-}i\text{ CLT: } \Theta(\sqrt{N}/\sqrt{M}) = \Theta(1)). \end{aligned} \tag{R2s-F3.4c}$$

$$\begin{aligned} A_3 &= -\eta N \chi_1 (W_1^4)^\top \left(\tfrac{1}{M} \textstyle\sum_i \phi_{i,2} \phi_{i,1} \|h_0^{2,i}\|^2 \right) \in \Theta(1) \\ &\quad \text{((R2}s\text{-U2.3)substitution;} \\ &\qquad \|h_0^{2,i}\|^2 \in \Theta(1) \text{ via (R2}s\text{-F1.4); LLN;} \\ &\qquad (W_1^4)^\top \in \Theta(1/N)). \end{aligned} \tag{R2s-F3.4d}$$

$$\begin{aligned} D &= -\eta N \chi_1 (W_1^4)^\top \tfrac{1}{M} \textstyle\sum_i \phi_{i,2} \phi_{i,1} \langle h_0^{2,i}, \Delta_2 h^{2,i} \rangle \in \Theta(1) \\ &\quad \text{((R2}s\text{-U2.3)substitution;} \\ &\qquad \langle h_0^{2,i}, \Delta_2 h^{2,i} \rangle \in \Theta(1) \text{ coherent across } i, \\ &\qquad \text{from two pieces: } \Delta_2 h^1 \parallel h_0^1 \text{ via (R2}s\text{-U2.1b),} \\ &\qquad \text{and the } j = i \text{ Gram self-overlap in the effective piece (R2}s\text{-F3.2c);} \\ &\qquad \text{LLN gives empirical average } \Theta(1)). \end{aligned} \tag{R2s-F3.4e}$$

$$h_2^3 \in \Theta(1) \text{ entry-wise.} \tag{R2s-F3.4}$$

$$\begin{aligned} f_2 &= W_2^4 \, h_2^3 = -(\eta(\chi_0 + \chi_1)/N) \langle h_0^3, h_2^3 \rangle \in \Theta(1) \\ &\quad (W_2^4 = W_1^4 + \Delta_2 W^4 \text{ via (R2}s\text{-U2.4);} \\ &\qquad \langle h_0^3, h_2^3 \rangle \in \Theta(N) \text{ coherent, dominated by } A_3 + D \text{ via (R2}s\text{-F3.4);} \\ &\qquad \text{net } \eta \cdot (1/N) \cdot \Theta(N) = \Theta(1)). \end{aligned} \tag{R2s-F3.5}$$

Third backward pass The structure parallels (R2s-B2.1)–(R2s-B2.7) with cumulative weights at step 2. Since $W_0^4 = 0$ in MSSP, the four ".1" pieces $A_{4.1}, A_{5.1}, A_{6.1}, E_1$ — using W_0^4 rather than ΔW^4 — are identically zero at every step; only the four ".2" pieces contribute. With $W_2^4 = \Delta W^4$ now non-zero (via (R2s-U2.4)), all gradients are non-zero.

Per-layer gradients.

$$\partial f_2 / \partial h_2^3 = (W_2^4)^\top \in \Theta(1/N). \quad \text{((R2}s\text{-U2.4))} \tag{R2s-B3.1}$$

$$\partial f_2 / \partial h_2^{3,i} = (\phi_{i,2}/M)(W_2^4)^\top \in \Theta(1/N^2). \quad (\phi_{i,2} \in \Theta(1);\ M = \Theta(N)) \tag{R2s-B3.2}$$

$$\begin{aligned} \partial f_2 / \partial h_2^{2,i} &= (\phi_{i,2}/M)(W_2^{3,i})^\top (W_2^4)^\top \in \Theta(1/N) \\ &\quad ((W_0^{3,i})^\top (W_2^4)^\top \in \Theta(1) \text{ via the cumulative analog of the new intermediate at } t = 1; \\ &\qquad (\Delta_2 W^{3,i})^\top (W_2^4)^\top \in \Theta(1) \text{ via (R2}s\text{-U2.3)).} \end{aligned} \tag{R2s-B3.3}$$

Router pathway $\partial f_2/\partial \phi_{i,2}$.

$$\partial f_2/\partial \phi_{i,2} = (1/M)\langle h_2^{3,i}, W_2^4\rangle \in \Theta(1/N), \tag{R2s-B3.4}$$

Four-piece grid: $*\cdot$init column is identically zero ($W_0^4 = 0$);
(init $\cdot$ update) $\langle (W_0^{3,1})\, h_2^{2,i}, \Delta_2 W^4\rangle \in \Theta(1)$:
$\langle h_0^{3,i}, W_2^4\rangle = -\frac{\chi_0+\chi_1}{N}\langle h_0^{3,i}, h_0^3\rangle$,
expanding $h_0^3 = (1/M)\sum_j \phi_{j,0} h_0^{3,j}$ gives the diagonal $j = i$ term
$(\phi_{i,0}/M)\|h_0^{3,i}\|^2 = (1/M)\Theta(N^2) = \Theta(N)$ via (R2s-F1.5);
$j \neq i$ off-diagonal terms give comparable $\Theta(N)$;
so $\langle h_0^{3,i}, W_2^4\rangle \in \Theta(1)$;
(update $\cdot$ update) $\langle (\Delta_2 W^{3,i})\, h_2^{2,i}, \Delta_2 W^4\rangle \in \Theta(1)$:
$\langle (\Delta_2 W^{3,i})\, h_2^{2,i}, W_2^4\rangle = \kappa \langle (W_1^4)^\top, W_2^4\rangle$
with $\kappa = -\eta N \chi_1 \phi_{i,1}\|h^{2,i}\|^2 \in \Theta(N)$
and $\langle W_1^4, W_2^4\rangle = (\chi_0(\chi_0+\chi_1)/N^2)\|h_0^3\|^2 \in \Theta(1/N)$;
net $\Theta(N)\cdot\Theta(1/N) = \Theta(1)$;
substituting: $\partial f_2/\partial \phi_{i,2} = (1/M)\big(\Theta(1) + \Theta(1)\big) = \Theta(1/M) = \Theta(1/N)$;
lifted from μP-R2's grid $(\Theta(1/N^{3/2}), \Theta(1/N^{3/2}), \Theta(1/N), \Theta(1/N))$
to MSSP-R2's $(0, \Theta(1/N), 0, \Theta(1/N))$ via $W_0^4 = 0$ and $\sigma_3^2 = M/N_e$.

Expert pathway: $(\partial f_2/\partial h_2^1)_{\exp} = A_{4.2} + A_{5.2} + A_{6.2} + E_2$.

$$A_{4.2} = \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}(W_0^{2,i})^\top (W_0^{3,i})^\top (W_2^4)^\top \in \Theta(1/N)$$

$((W_0^{3,i})^\top (W_2^4)^\top \in \Theta(1)$ via the cumulative analog of the new intermediate;
$(W_0^{2,i})^\top$ acts on $\Theta(1)$ vector with $\|v\|^2 = \Theta(1)$;
coord $\Theta(1/\sqrt{N})$ random per i, cross-i CLT: $\Theta(1/N)$). (R2s-B3.5a)

$$A_{5.2} = \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}(W_0^{2,i})^\top (\Delta_2 W^{3,i})^\top (W_2^4)^\top \in \Theta(1/N)$$

((R2s-U2.3)substitution gives $(\Delta_2 W^{3,i})^\top (W_2^4)^\top \in \Theta(1)$ along $h_0^{2,i}$;
$(W_0^{2,i})^\top h_0^{2,i} \in \Theta(1/\sqrt{N})$ random per i;
cross-i CLT: $\Theta(1/N)$). (R2s-B3.5b)

$$A_{6.2} = \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}(\Delta_2 W^{2,i})^\top (W_0^{3,i})^\top (W_2^4)^\top \in \Theta(1/N)$$

((R2s-U2.2)substitution;
per summand $-(\eta\chi_1\phi_{i,1}/N)h_0^1 \cdot S_i$ where
$S_i := W_1^4\, W_0^{3,i}(W_0^{3,i})^\top (W_2^4)^\top \in \Theta(1)$
(Lemma J.7 specialization: $W_0^{3,i}(W_0^{3,i})^\top$ entries $\Theta(N)$,
acting on $(W_2^4)^\top \in \Theta(1/N)$ gives $\Theta(\sqrt{N})$;
W_1^4 contraction yields $\Theta(1)$);
per summand entries $\Theta(1/N)$, cross-i avg: $\Theta(1/N)$). (R2s-B3.5c)

$$E_2 = \tfrac{1}{M}\sum_i \phi_{i,2}(\Delta_2 W^{2,i})^\top(\Delta_2 W^{3,i})^\top(W_2^4)^\top \in \Theta(1/N)$$
(combining (R2s-U2.2), (R2s-U2.3) substitutions;
$(\Delta_2 W^{3,i})^\top(W_2^4)^\top \in \Theta(1)$ along $h_0^{2,i}$;
$\langle W_1^4, h_0^{3,i}\rangle = -(\eta\chi_0/N)\langle h_0^3, h_0^{3,i}\rangle \in \Theta(1)$
(coherent $j=i$ piece via $\|h_0^{3,i}\|^2 \in \Theta(N^2)$ from (R2s-F1.5);
plus $j \neq i$ random of comparable scale);
per summand entries $\Theta(1/N)$;
cross-i LLN on coherent piece: $\Theta(1/N)$). (R2s-B3.5d)

$$(\partial f_2/\partial h_2^1)_{\text{exp}} \in \Theta(1/N), \text{ all four non-zero pieces at leading scale.} \quad \text{(R2}s\text{-B3.5)}$$

Router pathway.

$$(\partial f_2/\partial h_2^1)_{\text{router}} = (1/M)Q_2^\top v, \quad v_i := \dot\phi_{i,2}\langle h_2^{3,i}, W_2^4\rangle \in \Theta(1/N)$$
($v_i \in \Theta(1)$ via (R2s-B3.4);
by the analysis of (R2s-B2.6) extended to cumulative $Q_2 = Q_0 + \Delta_2 Q$;
Q_0 piece via cross-layer CLT;
$\Delta_2 Q$ piece coherent via (R2s-U2.Q);
both contribute at $\Theta(1/N)$). (R2s-B3.6)

$W^{3,i}$-balance transmits through the router pathway. The W^{3,i}-split of (R2s-B3.4), $\mathbf{v} = \mathbf{v}^I + \mathbf{v}^U$ with $\mathbf{v}_i^I = \dot\phi_{i,2}\langle (W_0^{3,1})h_2^{2,i}, \Delta_2 W^4\rangle$ and $\mathbf{v}_i^U = \dot\phi_{i,2}\langle (\Delta_2 W^{3,i})h_2^{2,i}, \Delta_2 W^4\rangle$, transmits as: both $\mathbf{v}_i^I \in \Theta(1)$ and $\mathbf{v}_i^U \in \Theta(1)$ coherent across i — the boost $\sigma_3^2 = M/N_e$ promotes the $W_0^{3,1}$-piece's diagonal $j=i$ contribution to $\Theta(1)$ (cf. R2s-B3.4), matching the rank-1-aligned $\Delta W^{3,i}$-piece. So $(1/M)Q_t^\top \mathbf{v}^I \in \Theta(1/N)$ and $(1/M)Q_t^\top \mathbf{v}^U \in \Theta(1/N)$: balanced, in contrast to the $\sqrt{N}$ deficit under μP-R2.

$$\partial f_2/\partial h_2^1 = (\partial f_2/\partial h_2^1)_{\text{exp}} + (\partial f_2/\partial h_2^1)_{\text{router}} \in \Theta(1/N). \quad \text{(R2}s\text{-B3.7)}$$

J.3.4 Summary tables of signal propagation for MSSP in Regime II

Forward features.

Quantity	$t=0$	$t=1$	$t=2$
h_t^1 (init)	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$\Delta_t h^1$	—	0 (no update)	$\Theta(1)$
effective $\Delta W^1\, x$	—	0	$\Theta(1)$
ψ_t	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$\Delta_t \psi$	—	0	$\Theta(1)$
effective $\Delta Q\, h_0^1$	—	0	$\Theta(1)$
propagating $Q_0\, \Delta h^1$	—	0	$\Theta(1)$
ϕ_t	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$\Delta_t \phi$	—	0	$\Theta(1)$
$h_t^{2,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$\Delta_t h^{2,i}$	—	0	$\Theta(1)$
effective $\Delta W^{2,i} h_0^1$	—	0	$\Theta(1)$
propagating $W_0^{2,i}\Delta h^1$	—	0	$\Theta(1)$
$h_t^{3,i}$	$\Theta(\sqrt{N})$	$\Theta(\sqrt{N})$	$\Theta(\sqrt{N})$
$\Delta_t h^{3,i}$	—	0	$\Theta(\sqrt{N})$
effective $\Delta W^{3,i} h_0^{2,i}$	—	0	$\Theta(1)$
propagating $W_0^{3,i}\Delta h^{2,i}$	—	0	$\Theta(\sqrt{N})$
h_t^3	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
f_t	0	$\Theta(1)$	$\Theta(1)$

Forward aggregation: $h_t^3 = A_1 + A_{2,1} + A_{2,2} + A_3 + D$**.**

Piece	$t=0$	$t=1$	$t=2$
$A_1 = (1/M)\sum \phi h_0^{3,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$A_{2,1} = (1/M)\sum \phi W_0^{3,i} W_0^{2,i} \Delta h^1$	0	0	$\Theta(1)$
$A_{2,2} = (1/M)\sum \phi W_0^{3,i} \Delta W^{2,i} h_0^1$	0	0	$\Theta(1)$
$A_3 = (1/M)\sum \phi \Delta W^{3,i} h_0^{2,i}$	0	0	$\Theta(1)$
$D = (1/M)\sum \phi \Delta W^{3,i} \Delta h^{2,i}$	0	0	$\Theta(1)$
Total h_t^3	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$

Backward gradients (per-layer with $W_0^4/\Delta W^4$ split).

Piece	$t=0$	$t=1$	$t=2$
$(W_0^4)^\top$	0	0	0
$(\Delta W^4)^\top$	0	$\Theta(1/N)$	$\Theta(1/N)$
$\partial f_t/\partial h_t^3$ **total**	0	$\Theta(1/N)$	$\Theta(1/N)$
$\partial f_t/\partial h_t^{3,i}$ **total**	0	$\Theta(1/N^2)$	$\Theta(1/N^2)$
$(\phi/M)(W_0^{3,i})^\top (W_0^4)^\top$	0	0	0
$(\phi/M)(W_0^{3,i})^\top (\Delta W^4)^\top$	0	$\Theta(1/N)$	$\Theta(1/N)$
$(\phi/M)(\Delta W^{3,i})^\top (W_0^4)^\top$	0	0	0
$(\phi/M)(\Delta W^{3,i})^\top (\Delta W^4)^\top$	0	0	$\Theta(1/N)$
$\partial f_t/\partial h_t^{2,i}$ **total**	0	$\Theta(1/N)$	$\Theta(1/N)$
init·init: $(1/M)\langle (W_0^{3,i})\, h_t^{2,i}, W_0^4\rangle$	0	0	0
init·update: $(1/M)\langle (W_0^{3,i})\, h_t^{2,i}, \Delta_t W^4\rangle$	0	$\Theta(1/N)$	$\Theta(1/N)$
update·init: $(1/M)\langle (\Delta_t W^{3,i})\, h_t^{2,i}, W_0^4\rangle$	0	0	0
update·update: $(1/M)\langle (\Delta_t W^{3,i})\, h_t^{2,i}, \Delta_t W^4\rangle$	0	0	$\Theta(1/N)$
$\partial f_t/\partial \phi_{i,t}$ **total**	0	$\Theta(1/N)$	$\Theta(1/N)$

Expert pathway: 8-piece decomposition.

Piece	Δ **factors**	$t=0$	$t=1$	$t=2$
$A_{4.1}$ (no ΔW^4)	none	0	0	0
$A_{4.2}$ (with ΔW^4)	W^4	0	$\Theta(1/N)$	$\Theta(1/N)$
$A_{5.1}$ (no ΔW^4)	$W^{3,i}$	0	0	0
$A_{5.2}$ (with ΔW^4)	$W^{3,i}, W^4$	0	0	$\Theta(1/N)$
$A_{6.1}$ (no ΔW^4)	$W^{2,i}$	0	0	0
$A_{6.2}$ (with ΔW^4)	$W^{2,i}, W^4$	0	0	$\Theta(1/N)$
E_1 (no ΔW^4)	$W^{2,i}, W^{3,i}$	0	0	0
E_2 (with ΔW^4)	all three	0	0	$\Theta(1/N)$
Expert total	—	0	$\Theta(1/N)$	$\Theta(1/N)$

Router pathway.

Piece	$t=0$	$t=1$	$t=2$
$(1/M)Q_t^\top[\dot\phi_{i,t}\langle h_t^{3,i}, \Delta_t W^4\rangle]_i$ **(total)**	0	$\Theta(1/N)$	$\Theta(1/N)$
$(1/M)Q_0^\top[\dot\phi_{i,t}\langle (W_0^{3,1}) h_t^{2,i}, \Delta_t W^4\rangle]_i$	0	$\Theta(1/N)$	$\Theta(1/N)$
$(1/M)Q_0^\top[\dot\phi_{i,t}\langle (\Delta_t W^{3,i}) h_t^{2,i}, \Delta_t W^4\rangle]_i$	0	0	$\Theta(1/N)$
$(1/M)\Delta Q^\top[\dot\phi_{i,t}\langle (W_0^{3,1}) h_t^{2,i}, \Delta_t W^4\rangle]_i$	0	0	$\Theta(1/N)$
$(1/M)\Delta Q^\top[\dot\phi_{i,t}\langle (\Delta_t W^{3,i}) h_t^{2,i}, \Delta_t W^4\rangle]_i$	0	0	$\Theta(1/N)$

Total backward.

Quantity	$t=0$	$t=1$	$t=2$
$\partial f_t/\partial h_t^1$	0	$\Theta(1/N)$	$\Theta(1/N)$

J.4 Scaling derivation for Regime III

In this section, we provide the heuristic scaling derivation for Regime III. Throughout we work under the proportional limit $N \asymp N_e \asymp M \asymp K \to \infty$.

Definition J.3 (μP baseline, Regime III)**.**

- **Initialization.** All parameters are drawn *independently* according to:

$$W_0^1 \sim \mathcal{N}(0, D^{-1}),\ Q_0 \sim \mathcal{N}(0, N^{-1}),\ W_0^{2,i} \sim \mathcal{N}(0, N^{-1}),$$
$$W_0^{3,i} \sim \mathcal{N}(0, N_e^{-1}),\ W_0^4 \sim \mathcal{N}(0, N^{-2}),\ \text{for } i \in [M].$$

- **SGD learning rates.** $\eta_1 = \eta\, N,\ \eta_Q = \eta,\ \eta_2 = \eta_3 = \eta\, M,\ \eta_4 = \eta\, N^{-1}$.

- **Adam learning rates.** $\eta_1 = \eta,\ \eta_Q = \eta_2 = \eta_3 = \eta_4 = \eta\, N^{-1}$.

- **Adam epsilon.** $\epsilon_1 = \epsilon\, N^{-1},\ \epsilon_Q = \epsilon\, M^{-1},\ \epsilon_2 = \epsilon_3 = \epsilon\, N^{-1}M^{-1},\ \epsilon_4 = \epsilon$.

Definition J.4 (Maximally Scale-Stable Parameterization (MSSP), Regime III)**.** MSSP adopts the same per-layer initialization variances and learning rate scalings as the μP baseline (Definition J.3), with the sole exception that expert weights are **shared across experts at initialization**: a single pair (W_0^2, W_0^3) is sampled and assigned to every expert, such that

$$W_0^{2,i} = W_0^2 \quad \text{and} \quad W_0^{3,i} = W_0^3 \quad \text{for all } i = 1, \dots, M.$$

J.4.1 Compound aggregate lemmas in Regime III

The analyses in this section invoke three compound aggregate-level results stating how sums across M independent experts of products and Gram products of the per-expert weight matrices behave under proportional scaling. All three are exact moment computations under Gaussian initialization, with independence across experts and from the input vector v as the only inputs.

Lemma J.5 (Cross-layer product sum across experts)**.** *Let $\{W^{2,i}\}_{i=1}^M$ have i.i.d. centered Gaussian entries of variance $\sigma_2^2 = 1/N$ in $\mathbb{R}^{N_e \times N}$, and let $\{W^{3,i}\}_{i=1}^M$ have i.i.d. centered Gaussian entries of variance $\sigma_3^2 = 1/N_e$ in $\mathbb{R}^{N \times N_e}$, with all matrices mutually independent across i and across the two layers. Let $c_i \in \Theta(1)$ be approximately independent of these weights, and let $v \in \mathbb{R}^N$ have entries of coordinate scale $\Theta(\alpha)$, independent of all $W^{2,i}, W^{3,i}$. Then*

$$G \;:=\; \frac{1}{M}\sum_{i=1}^{M} c_i\,(W^{2,i})^\top (W^{3,i})^\top \;\in\; \mathbb{R}^{N\times N}$$

has zero mean, entries of coordinate scale $\Theta(1/\sqrt{NM})$, and Gv has entries of coordinate scale $\Theta(\alpha/\sqrt{M}) = \Theta(\alpha/\sqrt{N})$.

Proof. $\mathbb{E}[G_{a,b}] = (1/M)\sum_i c_i\,\mathbb{E}[\sum_k W^{2,i}_{k,a} W^{3,i}_{b,k}] = 0$ by independence across the two layers. Variance:

$$\mathrm{Var}(G_{a,b}) = \frac{1}{M^2}\sum_i c_i^2\,\mathrm{Var}\Big(\sum_k W^{2,i}_{k,a} W^{3,i}_{b,k}\Big) = \frac{1}{M^2}\sum_i c_i^2 \cdot N_e\sigma_2^2\sigma_3^2 = \Theta\Big(\tfrac{N_e\sigma_2^2\sigma_3^2}{M}\Big) = \Theta\Big(\tfrac{1}{MN}\Big).$$

Cross-entry covariances $\mathbb{E}[G_{a,b_1}G_{a,b_2}]$ for $b_1 \neq b_2$ vanish by independence, so $\mathrm{Var}((Gv)_a) = \sum_b \mathrm{Var}(G_{a,b})\,v_b^2 = N \cdot \Theta(1/(MN)) \cdot \alpha^2 = \Theta(\alpha^2/M)$. □

Lemma J.6 (Gram concentration across experts)**.** *Let $\{W^{a,i}\}_{i=1}^M$ have i.i.d. centered Gaussian entries of variance σ^2 in $\mathbb{R}^{m\times n}$, independent across i. Let $c_i \in \Theta(1)$ be approximately independent of the weights, with empirical mean $\bar{c}$. Let $G^{a,i}$ denote either Gram $(W^{a,i})^\top W^{a,i} \in \mathbb{R}^{n\times n}$ or $W^{a,i}(W^{a,i})^\top \in \mathbb{R}^{m\times m}$, and let v be a vector of compatible dimension at coordinate scale $\Theta(\alpha)$, independent of all $W^{a,i}$. Then*

$$\frac{1}{M}\sum_{i=1}^{M} c_i\,G^{a,i}\,v = \bar{c}\,(\text{tr-dim}\cdot\sigma^2)\,v + r, \qquad r \text{ at coordinate scale } \Theta(\sigma^2\sqrt{mn/M}\,\alpha),$$

where tr-dim *is m for the inner Gram $((W^{a,i})^\top W^{a,i})$ and n for the outer Gram $(W^{a,i}(W^{a,i})^\top)$.*

For the specific instances used in this section:

- *Inner Gram:* $W^{2,i} \in \mathbb{R}^{N_e \times N}$ *with* $\sigma_2^2 = 1/N$ *gives leading constant* $m\sigma^2 = N_e/N$*;* $W^{3,i} \in \mathbb{R}^{N \times N_e}$ *with* $\sigma_3^2 = 1/N_e$ *gives leading constant* $m\sigma^2 = N/N_e$*.*
- *Outer Gram: both layers give* $n\sigma^2 = 1$*.*

All leading constants are $\Theta(1)$ *under proportional scaling, with random correction* $\Theta(\alpha/\sqrt{N})$*.*

Proof. Inner Gram case: $\mathbb{E}[G_{j,k}] = (1/M)\sum_i c_i\, \mathbb{E}[\sum_a W_{a,j}^{a,i} W_{a,k}^{a,i}] = \bar{c}\, m\sigma^2\, \delta_{j,k}$, with off-diagonal variance $\Theta(m\sigma^4/M)$ giving coordinate scale $\Theta(\sigma^2\sqrt{m/M})$ and off-diagonal contribution to $(Gv)_j - \bar{c}m\sigma^2 v_j$ at coordinate scale $\Theta(\sigma^2\sqrt{mn/M}\,\alpha)$. Outer Gram case is the same statement with the roles of m and n swapped, equivalently obtained by applying the inner-Gram result to $W^\top$ (whose entries are i.i.d. Gaussian of the same variance with the dimension pair swapped). □

J.4.2 Deriving μP in Regime III

This subsection provides a self-contained derivation of the heuristic width scaling for the μP baseline (Definition J.3) in Regime III. We invoke the compound aggregate lemmas of §J.4.1 for sums across independent experts of weight-product structures, and use CLT/LLN inline for single matrix-vector products and gating averages.

The principal structural consequence is the following. Although hidden weights move at $t = 1$ (since $W_0^4 \neq 0$ in μP), the embedding update $\Delta_1 h^1$ comes out at $\Theta(1/\sqrt{N})$, not $\Theta(1)$, because both the expert and router contributions to $\partial f_0/\partial h_0^1$ cancel across independent experts via CLT. Full $\Theta(1)$ feature learning at the embedding emerges at $t = 2$, when the rank-one updates $\Delta_1 W^{2,i}, \Delta_1 W^{3,i}$ inject coherent contributions (A_5, A_6) into the expert pathway gradient that align across experts along $(W_0^4)^\top$ and so escape the cross-i cancellation. We therefore analyse three forward and three backward passes.

First forward pass

$$h_0^1 = W_0^1 x \in \Theta(1). \quad (\text{CLT};\ \sigma_1^2 = 1/D,\ \|x\|^2 \in \Theta(D)) \tag{R3μ-F1.1}$$

$$\psi_0 = Q_0 h_0^1 \in \Theta(1). \quad (\text{CLT};\ \sigma_Q^2 = 1/N,\ \|h_0^1\|^2 \in \Theta(N)) \tag{R3μ-F1.2}$$

$$\phi_0 = \sigma(\psi_0) \in \Theta(1). \quad (\text{sigmoid bounded; gating assumption}) \tag{R3μ-F1.3}$$

$$h_0^{2,i} = W_0^{2,i} h_0^1 \in \Theta(1),\ \text{indep. across } i. \quad (\text{CLT};\ \sigma_2^2 = 1/N,\ \|h_0^1\|^2 \in \Theta(N)) \tag{R3μ-F1.4}$$

$$h_0^{3,i} = W_0^{3,i} h_0^{2,i} \in \Theta(1),\ \text{indep. across } i. \quad (\text{CLT};\ \sigma_3^2 = 1/N_e,\ \|h_0^{2,i}\|^2 \in \Theta(N_e)) \tag{R3μ-F1.5}$$

$$h_0^3 = (1/M)\sum_i \phi_{i,0} h_0^{3,i} \in \Theta(1/\sqrt{M}) = \Theta(1/\sqrt{N}). \quad (\text{cross-}i\ \text{CLT}) \tag{R3μ-F1.6}$$

$$f_0 = W_0^4 h_0^3 = \langle W_0^4, h_0^3 \rangle \in \Theta(1/N). \quad (\text{CLT};\ W_0^4 \text{ entries } \Theta(1/N) \text{ indep. of } h_0^3) \tag{R3μ-F1.7}$$

First backward pass and step-1 updates

$$\partial f_0/\partial h_0^3 = (W_0^4)^\top \in \Theta(1/N). \tag{R3μ-B1.1}$$

$$\partial f_0/\partial h_0^{3,i} = (\phi_{i,0}/M)(W_0^4)^\top \in \Theta(1/(MN)). \tag{R3μ-B1.2}$$

$$\partial f_0/\partial h_0^{2,i} = (\phi_{i,0}/M)(W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/(MN)). \tag{R3μ-B1.3}$$
$$((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/N) \text{ by CLT};\ \sigma_3^2 = 1/N_e)$$

$$\partial f_0/\partial \phi_{i,0} = (1/M)\langle h_0^{3,i}, W_0^4 \rangle \in \Theta(1/(M\sqrt{N})) = \Theta(1/N^{3/2}). \tag{R3μ-B1.4}$$
$$(\langle h_0^{3,i}, W_0^4 \rangle \in \Theta(1/\sqrt{N}) \text{ by CLT})$$

$$\left(\tfrac{\partial f_0}{\partial h_0^1}\right)_{\text{exp}} = \tfrac{1}{M}\sum_i \phi_{i,0}\,(W_0^{2,i})^\top (W_0^{3,i})^\top (W_0^4)^\top$$
$$\in \Theta\big(\sigma_2\sigma_3\sqrt{N_e}\,\|(W_0^4)^\top\|/\sqrt{M}\big) \text{ random direction} \quad (\text{Lem. J.5; cross-}i\text{ CLT})$$
$$= \Theta(1/\sqrt{N}\cdot 1/\sqrt{N_e}\cdot\sqrt{N_e}\cdot 1/\sqrt{N}\cdot 1/\sqrt{M}) = \Theta(1/N^{3/2}). \qquad (\text{R3}\mu\text{-B1.5})$$

$$\left(\tfrac{\partial f_0}{\partial h_0^1}\right)_{\text{router}} = \tfrac{1}{M} Q_0^\top v, \qquad v_i := \dot\phi_{i,0}\langle W_0^4, h_0^{3,i}\rangle$$
$$\in \Theta\big(\sigma_Q\|v\|/M\big) \text{ random direction} \quad (Q_0 \text{ indep. of } v;\ \text{cross-layer CLT})$$
$$= \Theta(1/\sqrt{N}\cdot 1\cdot 1/N) = \Theta(1/N^{3/2}) \qquad (\text{R3}\mu\text{-B1.6})$$
$$(\sigma_Q^2 = 1/N;\ v_i \in \Theta(1/\sqrt{N}) \text{ random across } i \Rightarrow \|v\|^2 \in \Theta(1)).$$

$$\partial f_0/\partial h_0^1 = (\partial f_0/\partial h_0^1)_{\text{exp}} + (\partial f_0/\partial h_0^1)_{\text{router}} \in \Theta(1/N^{3/2}). \qquad (\text{R3}\mu\text{-B1.7})$$

Step-1 parameter updates.

$$\Delta_1 W^4 = -\tfrac{\chi_0}{N}(h_0^3)^\top \in \Theta(1/N^{3/2}). \quad (h_0^3 \in \Theta(1/\sqrt{N}) \text{ by (R3}\mu\text{-F1.6)}) \qquad (\text{R3}\mu\text{-U1.4})$$

$$\Delta_1 W^{3,i} = -\eta\chi_0\phi_{i,0}\,(W_0^4)^\top (h_0^{2,i})^\top \in \Theta(1/N). \quad (\eta_3 = \eta M) \qquad (\text{R3}\mu\text{-U1.3})$$

$$\Delta_1 W^{2,i} = -\eta\chi_0\phi_{i,0}\,(W_0^{3,i})^\top (W_0^4)^\top (h_0^1)^\top \in \Theta(1/N). \qquad (\text{R3}\mu\text{-U1.2})$$
$$((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/N) \text{ by (R3}\mu\text{-B1.3))}$$

$$\Delta_1 W^1 = -\eta N\chi_0\,(\partial f_0/\partial h_0^1)\,x^\top \in \Theta(1/\sqrt{N}). \qquad (\text{R3}\mu\text{-U1.1a})$$
$$(\eta_1 = \eta N,\ \partial f_0/\partial h_0^1 \in \Theta(1/N^{3/2}) \text{ by (R3}\mu\text{-B1.7))}$$

$$\Delta_1 h^1 = \Delta_1 W^1\, x \in \Theta(1/\sqrt{N}),\ \text{sub-leading vs } h_0^1 \in \Theta(1). \qquad (\text{R3}\mu\text{-U1.1b})$$

$$\Delta_1 Q = -\eta_Q\chi_0\,\mathrm{diag}(\dot\phi_0)\,(\partial f_0/\partial\phi)\,(h_0^1)^\top \in \Theta(1/N^{3/2}). \qquad (\text{R3}\mu\text{-U1.Q})$$
$$(\partial f_0/\partial\phi \in \Theta(1/N^{3/2}) \text{ by (R3}\mu\text{-B1.4);}\ \eta_Q = \eta)$$

Second forward pass

$$h_1^1 = h_0^1 + \Delta_1 h^1 \in \Theta(1). \quad (h_0^1 \in \Theta(1) \text{ dominates } \Delta_1 h^1 \in \Theta(1/\sqrt{N}) \text{ by (R3}\mu\text{-U1.1b))} \qquad (\text{R3}\mu\text{-F2.1})$$

$$\psi_1 = \psi_0 + \Delta_1\psi \in \Theta(1), \quad \phi_1 \in \Theta(1). \qquad (\text{R3}\mu\text{-F2.2–R3}\mu\text{-F2.3})$$
$$(\Delta_1\psi \in \Theta(1/\sqrt{N}) \text{ from } Q_0\Delta_1 h^1 \text{ and } \Delta_1 Q\, h_0^1\text{; scales by (R3}\mu\text{-U1.1b), (R3}\mu\text{-U1.Q))}$$

(R3μ-F2.4) $h_1^{2,i} = W_1^{2,i}\, h_1^1$, four-piece decomposition:

$$h_1^{2,i} = h_0^{2,i} + W_0^{2,i}\Delta_1 h^1 + \Delta_1 W^{2,i}\, h_0^1 + \Delta_1 W^{2,i}\,\Delta_1 h^1. \qquad (\text{R3}\mu\text{-F2.4})$$

$$\text{init: } W_0^{2,i} h_0^1 = h_0^{2,i} \in \Theta(1). \quad (\text{R3}\mu\text{-F1.4}) \qquad (\text{R3}\mu\text{-F2.4a})$$

Op-norm of iid Gaussian $W_0^{2,i}$ $\Delta_1 h^1 \in \Theta(1/\sqrt{N})$ by (R3μ-U1.1b):

$$\text{prop: } W_0^{2,i}\Delta_1 h^1 \in \Theta(1/\sqrt{N}). \qquad (\text{R3}\mu\text{-F2.4b})$$

$$\text{eff: } \Delta_1 W^{2,i}\, h_0^1 = -\eta\chi_0\phi_{i,0}(W_0^{3,i})^\top (W_0^4)^\top \|h_0^1\|^2 \in \Theta(1). \qquad (\text{R3}\mu\text{-F2.4c})$$
$$((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/N) \text{ by (R3}\mu\text{-B1.3);}\ \|h_0^1\|^2 \in \Theta(N))$$

$$\text{cross: } \Delta_1 W^{2,i}\,\Delta_1 h^1 = -\eta\chi_0\phi_{i,0}(W_0^{3,i})^\top (W_0^4)^\top \langle h_0^1, \Delta_1 h^1\rangle \in \Theta(1/N). \qquad (\text{R3}\mu\text{-F2.4d})$$
$$((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/N) \text{ by (R3}\mu\text{-B1.3);}\ \langle h_0^1, \Delta_1 h^1\rangle \in \Theta(1) \text{ by CLT)}$$

(R3μ-F2.5) $h_1^{3,i} = W_1^{3,i}\, h_1^{2,i}$, four-piece decomposition:

$$h_1^{3,i} = h_0^{3,i} + W_0^{3,i}\Delta_1 h^{2,i} + \Delta_1 W^{3,i}\, h_0^{2,i} + \Delta_1 W^{3,i}\,\Delta_1 h^{2,i}. \tag{R3μ-F2.5}$$

$$\text{init: } h_0^{3,i} \in \Theta(1). \quad \text{(R3}\mu\text{-F1.5)} \tag{R3μ-F2.5a}$$

$$\text{prop: } W_0^{3,i}\Delta_1 h^{2,i} \in \Theta(1). \quad \text{(op-norm of iid Gaussian } W_0^{3,i};\ \Delta_1 h^{2,i} \in \Theta(1)) \tag{R3μ-F2.5b}$$

$$\text{eff: } \Delta_1 W^{3,i}\, h_0^{2,i} = -\eta\chi_0\phi_{i,0}(W_0^4)^\top \|h_0^{2,i}\|^2 \in \Theta(1). \tag{R3μ-F2.5c}$$
$$(\|h_0^{2,i}\|^2 \in \Theta(N_e);\ (W_0^4)^\top \in \Theta(1/N))$$

$$\text{cross: } \Delta_1 W^{3,i}\,\Delta_1 h^{2,i} \text{ tracked as part of } D \text{ in (R3}\mu\text{-F2.6).} \tag{R3μ-F2.5d}$$

(R3μ-F2.6) $h_1^3 = A_1 + A_{2,1} + A_{2,2} + A_3 + D$, where

$$A_1 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\, h_0^{3,i}, \qquad A_{2,1} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\, W_0^{3,i}W_0^{2,i}\,\Delta_1 h^1,$$
$$A_{2,2} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\, W_0^{3,i}\,\Delta_1 W^{2,i}\, h_0^1, \qquad A_3 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\,\Delta_1 W^{3,i}\, h_0^{2,i},$$
$$D := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\,\Delta_1 W^{3,i}\,\Delta_1 h^{2,i}.$$

$$A_1 \in \Theta(1/\sqrt{N}). \quad \text{(cross-}i\text{ CLT on } \Theta(1)\text{-entry independent vectors)} \tag{R3μ-F2.6a}$$

$$A_{2,1} \in \Theta(1/(\sqrt{N}\sqrt{M})) = \Theta(1/N). \tag{R3μ-F2.6b}$$
$$\text{(Lemma J.5;}\ \Delta_1 h^1 \text{ entries } \Theta(1/\sqrt{N}) \text{ by (R3}\mu\text{-U1.1b))}$$

$$\begin{aligned} A_{2,2} &= -\eta\chi_0\,\|h_0^1\|^2\,\tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\, W_0^{3,i}(W_0^{3,i})^\top(W_0^4)^\top \\ &\in \Theta\big(\|h_0^1\|^2\,\tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\big)\cdot(W_0^4)^\top \text{ along } (W_0^4)^\top \quad \text{(Lem. J.6)} \\ &= \Theta(N\cdot 1\cdot 1/N) = \Theta(1) \text{ along } (W_0^4)^\top \\ &\quad (\|h_0^1\|^2 \in \Theta(N);\ \text{LLN: } \tfrac{1}{M}\textstyle\sum \phi\phi \in \Theta(1);\ (W_0^4)^\top \in \Theta(1/N)) \end{aligned} \tag{R3μ-F2.6c}$$

$$A_3 = -\eta\chi_0\,(W_0^4)^\top\Big(\tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\,\|h_0^{2,i}\|^2\Big) \in \Theta(1) \text{ along } (W_0^4)^\top. \tag{R3μ-F2.6d}$$
$$(\|h_0^{2,i}\|^2 \in \Theta(N_e);\ \text{LLN})$$

$$\begin{aligned} D &= \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\,\Delta_1 W^{3,i}\,\Delta_1 h^{2,i} = -\eta\chi_0\,(W_0^4)^\top\,\tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\,\langle h_0^{2,i}, \Delta_1 h^{2,i}\rangle \\ &\in \Theta\big(\tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\,\langle h_0^{2,i}, \Delta_1 h^{2,i}\rangle\big)\cdot(W_0^4)^\top \text{ along } (W_0^4)^\top \\ &\quad (\text{cross-}i\text{ CLT: } \langle\cdot,\cdot\rangle \text{ random across } i,\ \text{mean } 0) \\ &= \Theta(\sqrt{N}/\sqrt{M}\cdot 1/N) = \Theta(1/N) \text{ along } (W_0^4)^\top \\ &\quad (\langle h_0^{2,i}, \Delta_1 h^{2,i}\rangle \in \Theta(\sqrt{N});\ (W_0^4)^\top \in \Theta(1/N);\ M \in \Theta(N)) \end{aligned} \tag{R3μ-F2.6e}$$

$$h_1^3 \in \Theta(1) \text{ along } (W_0^4)^\top, \text{ from } A_{2,2} + A_3. \tag{R3μ-F2.6}$$

$$W_0^4\, h_1^3 = \langle W_0^4, A_{2,2} + A_3\rangle + \cdots \in \Theta(1). \tag{R3μ-F2.7a}$$
$$(A_{2,2}, A_3 \text{ aligned along } (W_0^4)^\top;\ \langle W_0^4, A_3\rangle = -\eta\chi_0\|W_0^4\|^2\bar{c}\,N_e \in \Theta(1))$$

$$\Delta_1 W^4\, h_1^3 = -(\chi_0/N)\langle h_0^3, h_1^3\rangle \in \Theta(1/N). \tag{R3μ-F2.7b}$$

$$f_1 = W_1^4\, h_1^3 \in \Theta(1). \tag{R3μ-F2.7}$$

Second backward pass and step-2 updates

$$\partial f_1/\partial h_1^3 = (W_1^4)^\top = (W_0^4)^\top + (\Delta_1 W^4)^\top \in \Theta(1/N). \tag{R3μ-B2.1}$$

$$(W_0^4 \in \Theta(1/N) \text{ dominates } \Delta_1 W^4 \in \Theta(1/N^{3/2}) \text{ by (R3}\mu\text{-U1.4)})$$

$$\partial f_1/\partial h_1^{3,i} = (\phi_{i,1}/M)(W_1^4)^\top \in \Theta(1/(MN)). \tag{R3μ-B2.2}$$

$$\partial f_1/\partial h_1^{2,i} = (\phi_{i,1}/M)(W_1^{3,i})^\top (W_1^4)^\top \in \Theta(1/(MN)). \tag{R3μ-B2.3}$$

$$((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/N) \text{ by (R3}\mu\text{-B1.3)};\ (\Delta_1 W^{3,i})^\top W_0^4 \in \Theta(1/N);$$
$$\phi_{i,1} \in \Theta(1);\ M \in \Theta(N))$$

$$\partial f_1/\partial \phi_{i,1} = (1/M)\langle h_1^{3,i}, W_1^4\rangle \in \Theta(1/M). \tag{R3μ-B2.4}$$

$$(\langle h_1^{3,i}, W_0^4\rangle = -\eta\chi_0\phi_{i,0}\|h_0^{2,i}\|^2\|W_0^4\|^2 \in \Theta(1) \text{ via eff. term of (R3}\mu\text{-F2.5)};$$
$$\|h_0^{2,i}\|^2 \in \Theta(N_e);\ \|W_0^4\|^2 \in \Theta(1/N))$$

(R3μ-B2.5) $(\partial f_1/\partial h_1^1)_{\exp} = A_4 + A_5 + A_6 + E$, where

$$A_4 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}(W_0^{2,i})^\top (W_0^{3,i})^\top (W_1^4)^\top,$$
$$A_5 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}(W_0^{2,i})^\top (\Delta_1 W^{3,i})^\top (W_1^4)^\top,$$
$$A_6 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}(\Delta_1 W^{2,i})^\top (W_0^{3,i})^\top (W_1^4)^\top,$$
$$E := \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}(\Delta_1 W^{2,i})^\top (\Delta_1 W^{3,i})^\top (W_1^4)^\top.$$

$$A_4 \in \Theta(1/N^{3/2}). \quad (\text{Lemma J.5};\ v = (W_1^4)^\top \in \Theta(1/N)) \tag{R3μ-B2.5a}$$

$$A_5 = -\eta\chi_0 \|W_0^4\|^2 \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\,(W_0^{2,i})^\top W_0^{2,i}\, h_0^1$$
$$\in \Theta\big(\|W_0^4\|^2 \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\big)\cdot h_0^1 \text{ along } h_0^1 \quad (\text{Lem. J.6})$$
$$= \Theta(1/N\cdot 1\cdot 1) = \Theta(1/N) \text{ along } h_0^1 \tag{R3μ-B2.5b}$$
$$(\|W_0^4\|^2 \in \Theta(1/N);\ \text{LLN: } \tfrac{1}{M}\textstyle\sum \phi\phi \in \Theta(1);\ h_0^1 \in \Theta(1))$$

$$A_6 = -\eta\chi_0\, h_0^1 \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\, W_0^4 W_0^{3,i}(W_0^{3,i})^\top (W_1^4)^\top$$
$$\in \Theta\big(\tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}\,\|W_0^4\|^2\big)\cdot h_0^1 \text{ along } h_0^1$$
$$(\text{Lem. J.6: } W_0^{3,i}(W_0^{3,i})^\top (W_1^4)^\top \to (W_1^4)^\top;\ \langle W_0^4, (W_1^4)^\top\rangle \to \|W_0^4\|^2)$$
$$= \Theta(1\cdot 1/N) = \Theta(1/N) \text{ along } h_0^1 \tag{R3μ-B2.5c}$$
$$(\|W_0^4\|^2 \in \Theta(1/N);\ \text{LLN};\ h_0^1 \in \Theta(1))$$

$$E = \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}(\Delta_1 W^{2,i})^\top (\Delta_1 W^{3,i})^\top (W_1^4)^\top$$
$$= \eta^2\chi_0^2\, h_0^1 \tfrac{1}{M}\textstyle\sum_i \phi_{i,1}\phi_{i,0}^2\,\langle W_0^4, h_0^{3,i}\rangle\,(h_0^3)^\top (W_1^4)^\top \quad (\text{substituting (R3}\mu\text{-U1.2),(R3}\mu\text{-U1.3)})$$
$$\in \Theta\big(\tfrac{1}{M}\textstyle\sum_i \langle W_0^4, h_0^{3,i}\rangle\big)\cdot h_0^1 \text{ along } h_0^1 \quad (\text{cross-}i\text{ CLT: } \langle W_0^4, h_0^{3,i}\rangle \text{ random across } i,\ \text{mean } 0)$$
$$= \Theta(1/\sqrt{N}\cdot 1/\sqrt{M}\cdot \|h_0^3\|^2/N) = \Theta(1/N^2) \text{ along } h_0^1 \tag{R3μ-B2.5d}$$
$$(\langle W_0^4, h_0^{3,i}\rangle \in \Theta(1/\sqrt{N}) \text{ by (R3}\mu\text{-B1.4)};\ \|h_0^3\|^2 \in \Theta(1);\ M \in \Theta(N))$$
$$(\partial f_1/\partial h_1^1)_{\exp} \in \Theta(1/N) \text{ along } h_0^1. \tag{R3μ-B2.5}$$

$$(\partial f_1/\partial h_1^1)_{\text{router}} = \tfrac{1}{M} Q_0^\top v, \qquad v_i := \dot\phi_{i,1}\langle h_1^{3,i}, W_1^4\rangle$$
$$\in \Theta\big(\sigma_Q\|v\|/M\big) \text{ random direction}$$
$$(Q_0 \text{ indep. of } v;\ \text{cross-layer CLT};\ \Delta_1 Q \text{ sub-leading by (R3}\mu\text{-U1.Q)})$$
$$= \Theta(1/\sqrt{N}\cdot\sqrt{M}\cdot 1/M) = \Theta(1/\sqrt{MN}) = \Theta(1/N) \tag{R3μ-B2.6}$$
$$(\sigma_Q^2 = 1/N;\ v_i \in \Theta(1) \text{ via (R3}\mu\text{-B2.4)} \Rightarrow \|v\|^2 \in \Theta(M);\ M \in \Theta(N))$$

$$\partial f_1/\partial h_1^1 = (\partial f_1/\partial h_1^1)_{\exp} + (\partial f_1/\partial h_1^1)_{\text{router}} \in \Theta(1/N), \tag{R3μ-B2.7}$$

the proper feature-learning rate, an order of magnitude larger than $\partial f_0/\partial h_0^1$.

Step-2 parameter updates.

$$\Delta_2 W^4 = -\tfrac{\chi_1}{N}(h_1^3)^\top \in \Theta(1/N). \quad (h_1^3 \in \Theta(1) \text{ by (R3}\mu\text{-F2.6)}) \tag{R3μ-U2.4}$$

$$\Delta_2 W^{3,i},\ \Delta_2 W^{2,i} \in \Theta(1/N). \tag{R3μ-U2.3, R3μ-U2.2}$$
$$(\text{same structure as (R3}\mu\text{-U1.3), (R3}\mu\text{-U1.2) with cumulative weights})$$

$$\Delta_2 W^1 = -\eta_1 \chi_1 (\partial f_1/\partial h_1^1)\, x^\top \in \Theta(1). \tag{R3μ-U2.1a}$$
$$(\eta_1 = N,\ \partial f_1/\partial h_1^1 \in \Theta(1/N) \text{ by (R3}\mu\text{-B2.7)})$$

$$\Delta_2 h^1 = \Delta_2 W^1\, x \in \Theta(1) \text{ aligned along } h_0^1. \quad (\text{embedding feature-learns at } t=2) \tag{R3μ-U2.1b}$$

$$\Delta_2 Q \in \Theta(1/N). \quad (\partial f_1/\partial \phi \in \Theta(1/N) \text{ by (R3}\mu\text{-B2.4)}) \tag{R3μ-U2.Q}$$

Third forward pass We compute activations at $\theta^{(2)}$. The cumulative embedding change $\Delta h^1 := h_2^1 - h_0^1 = \Delta_1 h^1 + \Delta_2 h^1$, with $\Delta_1 h^1 \in \Theta(1/\sqrt{N})$ by (R3μ-U1.1b) and $\Delta_2 h^1 \in \Theta(1)$ aligned along h_0^1 by (R3μ-U2.1b).

$$h_2^1 = h_0^1 + \Delta h^1 \in \Theta(1). \tag{R3μ-F3.1}$$

(R3μ-F3.2) $h_2^{2,i} = W_2^{2,i} h_2^1$, four-piece decomposition with cumulative $\Delta W^{2,i}$:

$$\text{init: } W_0^{2,i} h_0^1 = h_0^{2,i} \in \Theta(1). \quad (\text{R3}\mu\text{-F1.4}) \tag{R3μ-F3.2a}$$

$$\text{prop: } W_0^{2,i} \Delta h^1 \in \Theta(1). \quad (\text{op-norm of iid Gaussian } W_0^{2,i};\ \Delta h^1 \in \Theta(1) \text{ by (R3}\mu\text{-U2.1b)}) \tag{R3μ-F3.2b}$$

$$\text{eff: } \Delta_t W^{2,i} h_0^1 = -\eta\chi\phi(W_{t-1}^{3,i})^\top (W_{t-1}^4)^\top \langle h_{t-1}^1, h_0^1\rangle \in \Theta(1). \tag{R3μ-F3.2c}$$
$$((W_{t-1}^{3,i})^\top (W_{t-1}^4)^\top \in \Theta(1/N) \text{ by (R3}\mu\text{-B1.3)/(R3}\mu\text{-B2.3);}$$
$$\langle h_{t-1}^1, h_0^1\rangle \in \Theta(N) \text{ coherent via (R3}\mu\text{-U2.1b))}$$

$$\text{cross: } \Delta_t W^{2,i} \Delta h^1 = -\eta\chi\phi(W_{t-1}^{3,i})^\top (W_{t-1}^4)^\top \langle h_{t-1}^1, \Delta h^1\rangle \in \Theta(1). \tag{R3μ-F3.2d}$$
$$((W_{t-1}^{3,i})^\top (W_{t-1}^4)^\top \in \Theta(1/N) \text{ by (R3}\mu\text{-B1.3)/(R3}\mu\text{-B2.3);}$$
$$\langle h_{t-1}^1, \Delta h^1\rangle \in \Theta(N) \text{ coherent via (R3}\mu\text{-U2.1b))}$$

$$h_2^{2,i} \in \Theta(1). \tag{R3μ-F3.2}$$

(R3μ-F3.3) $h_2^{3,i} = W_2^{3,i} h_2^{2,i}$, four-piece decomposition:

$$\text{init: } W_0^{3,i} h_0^{2,i} = h_0^{3,i} \in \Theta(1). \quad (\text{R3}\mu\text{-F1.5}) \tag{R3μ-F3.3a}$$

$$\text{prop: } W_0^{3,i} \Delta h^{2,i} \in \Theta(1). \quad (\text{op-norm of iid Gaussian } W_0^{3,i};\ \Delta h^{2,i} \in \Theta(1)) \tag{R3μ-F3.3b}$$

$$\text{eff: } \Delta_t W^{3,i} h_0^{2,i} = -\eta\chi\phi(W_{t-1}^4)^\top \langle h_{t-1}^{2,i}, h_0^{2,i}\rangle \in \Theta(1). \tag{R3μ-F3.3c}$$
$$((W_{t-1}^4)^\top \in \Theta(1/N);\ \langle h_{t-1}^{2,i}, h_0^{2,i}\rangle \in \Theta(N_e) \text{ coherent})$$

$$\text{cross: } \Delta W^{3,i} \Delta h^{2,i} \text{ tracked as part of } D' \text{ in (R3}\mu\text{-F3.4).} \tag{R3μ-F3.3d}$$

$$h_2^{3,i} \in \Theta(1). \tag{R3μ-F3.3}$$

(R3μ-F3.4) $h_2^3 = A_1' + A_{2,1}' + A_{2,2}' + A_3' + D'$:

$$A_1' = (1/M)\textstyle\sum_i \phi_{i,2} h_0^{3,i} \in \Theta(1/\sqrt{N}). \quad (\text{cross-}i \text{ CLT}) \tag{R3μ-F3.4a}$$

$$A_{2,1}' \in \Theta(1/\sqrt{M}) = \Theta(1/\sqrt{N}). \quad (\text{Lemma J.5};\ \Delta h^1 \in \Theta(1) \text{ by (R3}\mu\text{-U2.1b)}) \tag{R3μ-F3.4b}$$

$$A'_{2,2} = -\eta\chi_1 \langle h_1^1, h_0^1\rangle \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\phi_{i,1} W_0^{3,i}(W_1^{3,i})^\top (W_1^4)^\top$$
$$\in \Theta\big(\langle h_1^1, h_0^1\rangle \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\phi_{i,1}\big)\cdot (W_0^4)^\top \text{ along } (W_0^4)^\top \quad \text{(Lem. J.6)}$$
$$= \Theta(N\cdot 1\cdot 1/N) = \Theta(1) \text{ along } (W_0^4)^\top$$
$$(\langle h_1^1, h_0^1\rangle \in \Theta(N) \text{ coherent; LLN; } (W_0^4)^\top \in \Theta(1/N)). \qquad \text{(R3}\mu\text{-F3.4c)}$$

$$A'_3 \in \Theta(1) \text{ along } (W_0^4)^\top. \quad (\Delta_2 W^{3,i} h_0^{2,i} \in \Theta(1) \text{ along } (W_1^4)^\top; \text{ LLN}) \qquad \text{(R3}\mu\text{-F3.4d)}$$

$$D' \in \Theta(1) \text{ along } (W_0^4)^\top. \qquad \text{(R3}\mu\text{-F3.4e)}$$
$$(\langle h_1^{2,i}, \Delta h^{2,i}\rangle \in \Theta(N_e) \text{ coherent via alignment chain through (R3}\mu\text{-F3.2); LLN})$$
$$h_2^3 \in \Theta(1) \text{ along } (W_0^4)^\top. \qquad \text{(R3}\mu\text{-F3.4)}$$

(R3μ-F3.5) $f_2 = W_2^4 h_2^3$, with $W_2^4 = W_0^4 + \Delta_1 W^4 + \Delta_2 W^4$:

$$W_0^4\, h_2^3 \in \Theta(1). \quad (A'_{2,2} + A'_3 + D' = \kappa (W_0^4)^\top \text{ with } \kappa \in \Theta(N);\ \kappa \|W_0^4\|^2 \in \Theta(1)) \qquad \text{(R3}\mu\text{-F3.5a)}$$

$$\Delta_2 W^4\, h_2^3 = -(\chi_1/N)\langle h_1^3, h_2^3\rangle \in \Theta(1). \qquad \text{(R3}\mu\text{-F3.5b)}$$
$$(\langle h_1^3, h_2^3\rangle \in \Theta(N) \text{ via aligned components along } (W_0^4)^\top;$$
$$h_1^3 \in \Theta(1) \text{ along } (W_0^4)^\top \text{ by (R3}\mu\text{-F2.6))}$$

$$\Delta_1 W^4\, h_2^3 \in \Theta(1/\sqrt{N}). \quad (\Delta_1 W^4 \in \Theta(1/N^{3/2}) \text{ sub-leading by (R3}\mu\text{-U1.4)}) \qquad \text{(R3}\mu\text{-F3.5c)}$$
$$f_2 \in \Theta(1). \qquad \text{(R3}\mu\text{-F3.5)}$$

Third backward pass We compute gradients at $\theta^{(2)}$, with $W_2^4 = W_0^4 + \Delta_1 W^4 + \Delta_2 W^4$ where $W_0^4 \in \Theta(1/N)$, $\Delta_1 W^4 \in \Theta(1/N^{3/2})$ by (R3μ-U1.4), and $\Delta_2 W^4 = -(\chi_1/N)(h_1^3)^\top \in \Theta(1/N)$ aligned along $(h_1^3)^\top$ by (R3μ-U2.4).

$$\partial f_2/\partial h_2^3 = (W_2^4)^\top \in \Theta(1/N). \qquad \text{(R3}\mu\text{-B3.1)}$$
$$(W_0^4 \in \Theta(1/N) \text{ and } \Delta_2 W^4 \in \Theta(1/N) \text{ by (R3}\mu\text{-U2.4) both contribute;}$$
$$\Delta_1 W^4 \in \Theta(1/N^{3/2}) \text{ by (R3}\mu\text{-U1.4) sub-leading)}$$

$$\partial f_2/\partial h_2^{3,i} = (\phi_{i,2}/M)(W_2^4)^\top \in \Theta(1/(MN)). \quad (\phi_{i,2} \in \Theta(1)) \qquad \text{(R3}\mu\text{-B3.2)}$$

$$\partial f_2/\partial h_2^{2,i} = (\phi_{i,2}/M)(W_2^{3,i})^\top (W_2^4)^\top \in \Theta(1/(MN)). \qquad \text{(R3}\mu\text{-B3.3)}$$
$$((W_0^{3,i})^\top (W_0^4)^\top \in \Theta(1/N) \text{ via (R3}\mu\text{-B1.3);}$$
$$(\Delta W^{3,i})^\top (W_2^4)^\top \in \Theta(1/N) \text{ via Lem. J.6 on aligned components)}$$

$$\partial f_2/\partial \phi_{i,2} = (1/M)\langle h_2^{3,i}, W_2^4\rangle \in \Theta(1/M). \qquad \text{(R3}\mu\text{-B3.4)}$$
$$(\langle h_2^{3,i}, W_2^4\rangle \in \Theta(1) \text{ via the } (W_0^4)^\top\text{-aligned effective}$$
$$\text{and propagating components of (R3}\mu\text{-F3.3))}$$

(R3μ-B3.5) Expert contribution. The chain through the experts gives

$$\Big(\frac{\partial f_2}{\partial h_2^1}\Big)_{\text{exp}} = \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}(W_2^{2,i})^\top (W_2^{3,i})^\top (W_2^4)^\top.$$

Expand $W_2^{a,i} = W_0^{a,i} + \Delta W^{a,i}$ with cumulative update $\Delta W^{a,i} = \Delta_1 W^{a,i} + \Delta_2 W^{a,i}$. The four-term decomposition is:

$$A''_4 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}(W_0^{2,i})^\top (W_0^{3,i})^\top (W_2^4)^\top,$$
$$A''_5 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}(W_0^{2,i})^\top (\Delta W^{3,i})^\top (W_2^4)^\top,$$
$$A''_6 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}(\Delta W^{2,i})^\top (W_0^{3,i})^\top (W_2^4)^\top,$$
$$E'' := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}(\Delta W^{2,i})^\top (\Delta W^{3,i})^\top (W_2^4)^\top.$$

$$A_4'' \in \Theta(1/N^{3/2}). \quad \text{(Lemma J.5; } v = (W_2^4)^\top \in \Theta(1/N)) \tag{R3μ-B3.5a}$$

$$\begin{aligned} A_5'' &= -\eta\chi \, \|W_0^4\|^2 \, \tfrac{1}{M} \textstyle\sum_i \phi_{i,2}\phi \, (W_0^{2,i})^\top W_0^{2,i} \, h_0^1 + \cdots \\ &\in \Theta\big(\|W_0^4\|^2 \, \tfrac{1}{M} \textstyle\sum_i \phi_{i,2}\phi\big) \cdot h_0^1 \text{ along } h_0^1 \\ &\quad \text{(Lem. J.6; } \Delta_1 \text{ and } \Delta_2 \text{ contributions share leading direction)} \\ &= \Theta(1/N \cdot 1 \cdot 1) = \Theta(1/N) \text{ along } h_0^1 \\ &\quad (\|W_0^4\|^2 \in \Theta(1/N); \text{ LLN; } h_0^1 \in \Theta(1)) \end{aligned} \tag{R3μ-B3.5b}$$

$$A_6'' \in \Theta(1/N) \text{ along } h_0^1. \tag{R3μ-B3.5c}$$
$$\text{(Lemma J.6; LLN; } (\Delta W^{2,i})^\top \text{ rank-1 with leading direction } h_0^1 \otimes (W_0^4 W_0^{3,i}))$$

$$\begin{aligned} E'' &= \eta^2\chi^2 \, h_0^1 \, \tfrac{1}{M} \textstyle\sum_i \phi_{i,2}\phi^2 \, \langle W_t^4, h_t^{3,i}\rangle \, (h_t^3)^\top (W_2^4)^\top + \cdots \\ &\in \Theta\big(\tfrac{1}{M} \textstyle\sum_i \langle W_t^4, h_t^{3,i}\rangle \cdot \langle h_t^3, W_2^4\rangle\big) \cdot h_0^1 \text{ along } h_0^1 \\ &\quad \text{(at } t = 2\text{: } \langle W_t^4, h_t^{3,i}\rangle \in \Theta(1) \text{ coherent via eff. term of (R3}\mu\text{-F2.5); LLN)} \\ &= \Theta(1 \cdot 1) = \Theta(1/N) \text{ along } h_0^1 \\ &\quad \text{(after multiplying by } \langle h_t^3, W_2^4\rangle \in \Theta(1) \text{ by (R3}\mu\text{-F3.5); prefactor } \eta^2\|W_0^4\|^2 \in \Theta(1/N)) \end{aligned} \tag{R3μ-B3.5d}$$

$$(\partial f_2/\partial h_2^1)_{\text{exp}} = A_4'' + A_5'' + A_6'' + E'' \in \Theta(1/N) \text{ along } h_0^1. \tag{R3μ-B3.5}$$

$$\begin{aligned} (\partial f_2/\partial h_2^1)_{\text{router}} &= \tfrac{1}{M} Q_0^\top v + \tfrac{1}{M} (\Delta_2 Q)^\top v, \qquad v_i := \dot\phi_{i,2} \langle h_2^{3,i}, W_2^4\rangle \\ &\in \Theta\big(\sigma_Q \|v\|/M\big) \text{ random direction} + \Theta\big(\langle \partial f_1/\partial\phi, v\rangle/M\big) \, h_1^1 \text{ along } h_1^1 \\ &\quad (Q_0 \text{ piece via cross-layer CLT; } \Delta_2 Q \text{ piece coherent}) \\ &= \Theta(1/\sqrt{N} \cdot \sqrt{M} \cdot 1/M) = \Theta(1/\sqrt{MN}) = \Theta(1/N) \\ &\quad (\sigma_Q^2 = 1/N; \; v_i \in \Theta(1) \text{ via (R3}\mu\text{-B2.4)} \Rightarrow \|v\|^2 \in \Theta(M); \; M \in \Theta(N)) \end{aligned} \tag{R3μ-B3.6}$$

$$\partial f_2/\partial h_2^1 = (\partial f_2/\partial h_2^1)_{\text{exp}} + (\partial f_2/\partial h_2^1)_{\text{router}} \in \Theta(1/N), \tag{R3μ-B3.7}$$

dominated by $A_5'' + A_6'' + E''$ along h_0^1; same scale as $\partial f_1/\partial h_1^1$ at $t = 1$, with E'' now matching A_5'', A_6'' via the alignment-coherence shift.

J.4.3 Summary table of signal propagation for μP in Regime III

Notation: $\Delta_t W^\ell$ denotes the cumulative update $W_t^\ell - W_0^\ell$.

Forward.

Quantity	$t=0$	$t=1$	$t=2$
$h_t^1 = h_0^1 + \Delta_t h^1$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $h_0^1 = W_0^1 x$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
effective: $\Delta_t h^1 = \Delta_t W^1 x$	0	$\Theta(1/\sqrt{N})$	$\Theta(1)$
ψ_t, ϕ_t	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$h_t^{2,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $h_0^{2,i} = W_0^{2,i} h_0^1$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
propagating: $W_0^{2,i} \Delta_t h^1$	0	$\Theta(1/\sqrt{N})$	$\Theta(1)$
effective: $\Delta_t W^{2,i}\, h_0^1$	0	$\Theta(1)$	$\Theta(1)$
cross: $\Delta_t W^{2,i}\, \Delta_t h^1$	0	$\Theta(1/N)$	$\Theta(1)$
$h_t^{3,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $h_0^{3,i} = W_0^{3,i} h_0^{2,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
propagating: $W_0^{3,i} \Delta_t h^{2,i}$	0	$\Theta(1)$	$\Theta(1)$
effective: $\Delta_t W^{3,i}\, h_0^{2,i}$	0	$\Theta(1)$	$\Theta(1)$
cross: $\Delta_t W^{3,i}\, \Delta_t h^{2,i}$	0	$\Theta(1/N^{3/2})$	$\Theta(1)$
$h_t^3 = A_1 + A_{2,1} + A_{2,2} + A_3 + D$	$\Theta(1/\sqrt{N})$	$\Theta(1)$	$\Theta(1)$
$A_1 = (1/M) \sum_i \phi_{i,t}\, h_0^{3,i}$	$\Theta(1/\sqrt{N})$	$\Theta(1/\sqrt{N})$	$\Theta(1/\sqrt{N})$
$A_{2,1} = (1/M) \sum_i \phi_{i,t}\, W_0^{3,i} W_0^{2,i}\, \Delta_t h^1$	0	$\Theta(1/N)$	$\Theta(1/\sqrt{N})$
$A_{2,2} = (1/M) \sum_i \phi_{i,t}\, W_0^{3,i}\, \Delta_t W^{2,i}\, h_0^1$	0	$\Theta(1)$	$\Theta(1)$
$A_3 = (1/M) \sum_i \phi_{i,t}\, \Delta_t W^{3,i}\, h_0^{2,i}$	0	$\Theta(1)$	$\Theta(1)$
$D = (1/M) \sum_i \phi_{i,t}\, \Delta_t W^{3,i}\, \Delta_t h^{2,i}$	0	$\Theta(1/N)$	$\Theta(1)$
$f_t = W_t^4 h_t^3$	$\Theta(1/N)$	$\Theta(1)$	$\Theta(1)$
init: $W_0^4\, h_0^3$	$\Theta(1/N)$	$\Theta(1/N)$	$\Theta(1/N)$
propagating: $W_0^4\, \Delta_t h^3$	0	$\Theta(1)$	$\Theta(1)$
effective: $\Delta_t W^4\, h_0^3$	0	$\Theta(1/N)$	$\Theta(1/N)$
cross: $\Delta_t W^4\, \Delta_t h^3$	0	$\Theta(1/N)$	$\Theta(1)$

Backward.

Quantity	$t=0$	$t=1$	$t=2$
$\partial f_t/\partial h_t^3 = (W_0^4)^\top + (\Delta_t W^4)^\top$	$\Theta(1/N)$	$\Theta(1/N)$	$\Theta(1/N)$
init: $(W_0^4)^\top$	$\Theta(1/N)$	$\Theta(1/N)$	$\Theta(1/N)$
update: $(\Delta_t W^4)^\top$	0	$\Theta(1/N^{3/2})$	$\Theta(1/N)$
$\partial f_t/\partial h_t^{3,i} = (\phi_{i,t}/M)(W_t^4)^\top$	$\Theta(1/(MN))$	$\Theta(1/(MN))$	$\Theta(1/(MN))$
init: $(\phi_{i,t}/M)(W_0^4)^\top$	$\Theta(1/(MN))$	$\Theta(1/(MN))$	$\Theta(1/(MN))$
update: $(\phi_{i,t}/M)(\Delta_t W^4)^\top$	0	$\Theta(1/(MN^{3/2}))$	$\Theta(1/(MN))$
$\partial f_t/\partial h_t^{2,i} = (\phi_{i,t}/M)(W_t^{3,i})^\top(W_t^4)^\top$	$\Theta(1/(MN))$	$\Theta(1/(MN))$	$\Theta(1/(MN))$
init·init: $(\phi_{i,t}/M)(W_0^{3,i})^\top(W_0^4)^\top$	$\Theta(1/(MN))$	$\Theta(1/(MN))$	$\Theta(1/(MN))$
init·update: $(\phi_{i,t}/M)(W_0^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/(MN^{3/2}))$	$\Theta(1/(MN))$
update·init: $(\phi_{i,t}/M)(\Delta_t W^{3,i})^\top(W_0^4)^\top$	0	$\Theta(1/(MN))$	$\Theta(1/(MN))$
update·update: $(\phi_{i,t}/M)(\Delta_t W^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/(MN^2))$	$\Theta(1/(MN))$
$\partial f_t/\partial \phi_{i,t} = (1/M)\langle h_t^{3,i}, W_t^4\rangle$	$\Theta(1/N^{3/2})$	$\Theta(1/M)$	$\Theta(1/M)$
init: $(1/M)\langle h_t^{3,i}, W_0^4\rangle$	$\Theta(1/N^{3/2})$	$\Theta(1/M)$	$\Theta(1/M)$
update: $(1/M)\langle h_t^{3,i}, \Delta_t W^4\rangle$	0	$\Theta(1/N^{3/2})$	$\Theta(1/M)$
$(\partial f_t/\partial h_t^1)_{\text{exp}}$	$\Theta(1/N^{3/2})$	$\Theta(1/N)$	$\Theta(1/N)$
$A_{4,1} = (1/M)\sum_i \phi_{i,t}\,(W_0^{2,i})^\top(W_0^{3,i})^\top(W_0^4)^\top$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$	$\Theta(1/N^{3/2})$
$A_{4,2} = (1/M)\sum_i \phi_{i,t}\,(W_0^{2,i})^\top(W_0^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N^{3/2})$
$A_{5,1} = (1/M)\sum_i \phi_{i,t}\,(W_0^{2,i})^\top(\Delta_t W^{3,i})^\top(W_0^4)^\top$	0	$\Theta(1/N)$	$\Theta(1/N)$
$A_{5,2} = (1/M)\sum_i \phi_{i,t}\,(W_0^{2,i})^\top(\Delta_t W^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$A_{6,1} = (1/M)\sum_i \phi_{i,t}\,(\Delta_t W^{2,i})^\top(W_0^{3,i})^\top(W_0^4)^\top$	0	$\Theta(1/N)$	$\Theta(1/N)$
$A_{6,2} = (1/M)\sum_i \phi_{i,t}\,(\Delta_t W^{2,i})^\top(W_0^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$E_1 = (1/M)\sum_i \phi_{i,t}\,(\Delta_t W^{2,i})^\top(\Delta_t W^{3,i})^\top(W_0^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$E_2 = (1/M)\sum_i \phi_{i,t}\,(\Delta_t W^{2,i})^\top(\Delta_t W^{3,i})^\top(\Delta_t W^4)^\top$	0	$\Theta(1/N^2)$	$\Theta(1/N)$
$(\partial f_t/\partial h_t^1)_{\text{router}} = (1/M)Q_t^\top v$	$\Theta(1/N^{3/2})$	$\Theta(1/\sqrt{MN})$	$\Theta(1/\sqrt{MN})$
init: $(1/M)Q_0^\top v$	$\Theta(1/N^{3/2})$	$\Theta(1/\sqrt{MN})$	$\Theta(1/\sqrt{MN})$
update: $(1/M)(\Delta_t Q)^\top v$	0	$\Theta(1/(MN))$	$\Theta(1/\sqrt{MN})$

J.4.4 Deriving MSSP in Regime III

This subsection provides a self-contained, heuristic scaling analysis for the Maximally Scale-Stable Parameterization (MSSP, Definition J.4) in Regime III.

Auxiliary scaling lemma

Lemma J.7 (Gram concentration). *Let $W \in \mathbb{R}^{m\times n}$ have i.i.d. centered entries of variance σ_W^2. Then*

$$\mathbb{E}\big[(W^\top W)_{jk}\big] = m\sigma_W^2\,\delta_{jk}, \qquad \mathbb{E}\big[(WW^\top)_{ij}\big] = n\sigma_W^2\,\delta_{ij},$$

with off-diagonal fluctuations of order $\sigma_W^2\sqrt{m}$ (resp. $\sigma_W^2\sqrt{n}$). For any fixed $v \in \mathbb{R}^n$ with $\|v\|^2 \in \Theta(n)$ that is independent of W, $W^\top W\,v = m\sigma_W^2\,v + r$ with r of coordinate scale $\Theta(\sigma_W^2\sqrt{mn})$. The analogous statement holds for $WW^\top$ acting on $v' \in \mathbb{R}^m$ with $\|v'\|^2 \in \Theta(m)$.

Sketch. $(W^\top W)_{jk} = \sum_a W_{aj}W_{ak}$ is a sum of m centered i.i.d. products with mean $\sigma_W^2\delta_{jk}$ and variance σ_W^4 off-diagonal. The residual $r = (W^\top W - m\sigma_W^2 I)v$ has entry-wise variance $\sum_k m\sigma_W^4\,v_k^2 = m\sigma_W^4\,\|v\|^2$, hence coordinate scale $\sigma_W^2\sqrt{mn}$. □

Specializations of Lemma J.7 under proportional scaling. Both Gram matrices of $W_0^{2,1}$ and of $W_0^{3,1}$ produce $\Theta(1)$-coordinate output when applied to $\Theta(1)$-coordinate input vectors (independent of or weakly correlated with the corresponding W):

$$(W_0^{2,1})^\top W_0^{2,1}\,v = (N_e/N)\,v + r, \qquad W_0^{2,1}(W_0^{2,1})^\top\,v' = v' + r',$$
$$(W_0^{3,1})^\top W_0^{3,1}\,u = (N/N_e)\,u + r'', \qquad W_0^{3,1}(W_0^{3,1})^\top\,u' = u' + r''',$$

where each residual is of the same $\Theta(1)$ entry-wise order as the leading deterministic part. The Gram approximation is order-of-magnitude correct, with comparable random fluctuations.

First forward pass

$$h_0^1 = W_0^1\, x \in \Theta(1). \quad (\text{CLT};\ \sigma_1^2 = 1/D,\ \|x\|^2 \in \Theta(D)) \tag{R3s-F1.1}$$

$$\psi_0 = Q_0\, h_0^1 \in \Theta(1). \quad (\text{CLT};\ \sigma_Q^2 = 1/N,\ \|h_0^1\|^2 \in \Theta(N)) \tag{R3s-F1.2}$$

$$\phi_0 = \sigma(\psi_0) \in \Theta(1). \quad (\text{sigmoid bounded; gating assumption}) \tag{R3s-F1.3}$$

$$h_0^{2,i} = h_0^{2,1} = W_0^{2,1}\, h_0^1 \in \Theta(1)$$
$$(\text{CLT};\ \sigma_2^2 = 1/N,\ \|h_0^1\|^2 \in \Theta(N);\ h_0^{2,i} = h_0^{2,1} \text{ by MSSP}). \tag{R3s-F1.4}$$

$$h_0^{3,i} = h_0^{3,1} = W_0^{3,1}\, h_0^{2,1} \in \Theta(1)$$
$$(\text{CLT};\ \sigma_3^2 = 1/N_e,\ \|h_0^{2,1}\|^2 \in \Theta(N_e);\ h_0^{3,i} = h_0^{3,1} \text{ by MSSP}). \tag{R3s-F1.5}$$

$$h_0^3 = (1/M)\textstyle\sum_i \phi_{i,0}\, h_0^{3,1} = (\sum_k \phi_{k,0}/M)\, h_0^{3,1} \in \Theta(1)$$
$$((\textstyle\sum_k \phi_{k,0}/M) \in \Theta(1) \text{ by LLN}). \tag{R3s-F1.6}$$

$$f_0 = W_0^4\, h_0^3 = 0. \quad (\text{standing assumption } W_0^4 = 0) \tag{R3s-F1.7}$$

First backward pass and step-1 updates By the chain rule, every hidden gradient ($\partial f_0/\partial h_0^3$, $\partial f_0/\partial h_0^{3,i}$, $\partial f_0/\partial h_0^{2,i}$, $\partial f_0/\partial \phi_{i,0}$, $\partial f_0/\partial h_0^1$ via either the expert or router path) carries a $W_0^4 = 0$ factor and is therefore zero. Hence every non-readout weight is unchanged at $t = 1$: $W_1^\ell = W_0^\ell$ for $\ell \in \{1, Q, 2, 3\}$.

Step-1 parameter updates.

$$\Delta_1 W^4 = -\tfrac{\chi_0}{N}\,(h_0^3)^\top \in \Theta(1/N). \quad (\text{aligned along } (h_0^3)^\top) \tag{R3s-U1.4}$$

Second forward pass Since the non-readout weights are unchanged at $t = 1$, every hidden activation is unchanged: $h_1^\ell = h_0^\ell$ for $\ell \in \{1, (2,i), (3,i), 3\}$, $\psi_1 = \psi_0$, $\phi_1 = \phi_0$.

$$f_1 = W_1^4\, h_0^3 = -\tfrac{\chi_0}{N}\|h_0^3\|^2 \in \Theta(1). \quad (\|h_0^3\|^2 \in \Theta(N) \text{ via (R3}s\text{-F1.6)}) \tag{R3s-F2.7}$$

Second backward pass and step-2 updates

$$\partial f_1/\partial h_1^3 = (W_1^4)^\top = (\Delta_1 W^4)^\top \in \Theta(1/N). \quad (\text{(R3}s\text{-U1.4)};\ W_0^4 = 0) \tag{R3s-B2.1}$$

$$\partial f_1/\partial h_1^{3,i} = (\phi_{i,0}/M)(W_1^4)^\top \in \Theta(1/(MN)). \quad (\phi_{i,0} \in \Theta(1)) \tag{R3s-B2.2}$$

$$\partial f_1/\partial h_1^{2,i} = (\phi_{i,0}/M)(W_0^{3,1})^\top (W_1^4)^\top \in \Theta(1/(MN))$$
$$((W_0^{3,1})^\top (W_1^4)^\top = -\tfrac{\chi_0}{MN}(\textstyle\sum \phi)(W_0^{3,1})^\top W_0^{3,1} h_0^{2,1} \in \Theta(1/N) \text{ by Lem. J.7};$$
$$\phi_{i,0} \in \Theta(1)). \tag{R3s-B2.3}$$

$$\partial f_1/\partial \phi_{i,0} = (1/M)(h_0^{3,1})^\top (W_1^4)^\top \in \Theta(1/M)$$
$$((h_0^{3,1})^\top (W_1^4)^\top = -\tfrac{\chi_0}{MN}(\textstyle\sum \phi)\|h_0^{3,1}\|^2 \in \Theta(1) \text{ by LLN};$$
$$\|h_0^{3,1}\|^2 \in \Theta(N_e) \text{ via (R3}s\text{-F1.5)};\ N_e/N \in \Theta(1) \text{ in Regime III}). \tag{R3s-B2.4}$$

$$\Big(\tfrac{\partial f_1}{\partial h_1^1}\Big)_{\text{exp}} = \big(\tfrac{1}{M}\textstyle\sum \phi\big)(W_0^{2,1})^\top (W_0^{3,1})^\top (W_1^4)^\top$$
$$\in \Theta\big(\tfrac{1}{M}\textstyle\sum \phi\big) \cdot (W_0^{2,1})^\top h_0^{2,1}\,/N \text{ along } (W_0^{2,1})^\top h_0^{2,1}$$
$$(\text{Lem. J.7 on } (W_0^{3,1})^\top h_0^3)$$
$$= \Theta(1 \cdot 1/N) = \Theta(1/N)$$
$$(\text{LLN};\ (W_0^{2,1})^\top h_0^{2,1} \in \Theta(1) \text{ by Lem. J.7}). \tag{R3s-B2.5}$$

$$\Big(\tfrac{\partial f_1}{\partial h_1^1}\Big)_{\text{router}} = \tfrac{1}{M} Q_0^\top v, \qquad v_i := \dot\phi_{i,0}(h_0^{3,1})^\top (W_1^4)^\top$$
$$\in \Theta\big(\sigma_Q \|v\|/M\big) \text{ random direction} \quad (Q_0 \text{ indep. of } v;\ \text{cross-layer CLT})$$
$$= \Theta(1/\sqrt{N} \cdot \sqrt{M} \cdot 1/M) = \Theta(1/\sqrt{MN})$$
$$(\sigma_Q^2 = 1/N;\ v_i \in \Theta(1) \text{ via (R3}s\text{-B2.4)} \Rightarrow \|v\|^2 \in \Theta(M)). \tag{R3s-B2.6}$$

$$\partial f_1/\partial h_1^1 = (\partial f_1/\partial h_1^1)_{\text{exp}} + (\partial f_1/\partial h_1^1)_{\text{router}} \in \Theta(1/N). \tag{R3s-B2.7}$$

Step-2 parameter updates.

$$\Delta_2 W^4 = -\tfrac{\chi_1}{N}(h_1^3)^\top \in \Theta(1/N), \quad W_2^4 = -\tfrac{\chi_0+\chi_1}{N}(h_0^3)^\top \\ (h_1^3 = h_0^3). \tag{R3s-U2.4}$$

$$\Delta_2 W^{3,i} = -\tfrac{\chi_1 \phi_{i,0}}{N}\, h_0^3\, (h_0^{2,1})^\top \in \Theta(1/N). \quad (\eta_3 = M;\ \text{(R3}s\text{-B2.2))} \tag{R3s-U2.3}$$

$$\Delta_2 W^{2,i} = -\tfrac{\chi_1 \phi_{i,0}}{N}\, (W_0^{3,1})^\top h_0^3\, (h_0^1)^\top \in \Theta(1/N). \quad ((W_0^{3,1})^\top h_0^3 \in \Theta(1) \text{ via (R3}s\text{-B2.3))} \tag{R3s-U2.2}$$

$$\Delta_2 W^1 = -\eta_1 \chi_1 (\partial f_1/\partial h_1^1)\, x^\top \in \Theta(1) \\ (\eta_1 = N;\ \partial f_1/\partial h_1^1 \in \Theta(1/N) \text{ via (R3}s\text{-B2.7)}). \tag{R3s-U2.1a}$$

$$\Delta_2 h^1 = \Delta_2 W^1\, x \in \Theta(1) \text{ aligned along } \partial f_1/\partial h_1^1. \tag{R3s-U2.1b}$$

$$\Delta_2 Q = -\eta_Q \chi_1\, (\partial f_1/\partial \phi)\, (h_1^1)^\top \in \Theta(1/M) = \Theta(1/N) \\ (\eta_Q = 1;\ \partial f_1/\partial \phi \in \Theta(1/M) \text{ via (R3}s\text{-B2.4)}). \tag{R3s-U2.Q}$$

Third forward pass We compute activations at $\theta^{(2)}$, expanding each updated weight as $W_2^\ell = W_0^\ell + \Delta_2 W^\ell$.

$$h_2^1 = h_0^1 + \Delta_2 h^1 \in \Theta(1). \quad (\Delta_2 h^1 \in \Theta(1) \text{ via (R3}s\text{-U2.1b))} \tag{R3s-F3.1}$$

(R3s-F3.2) $h_2^{2,i} = W_2^{2,i}\, h_2^1$, four-piece decomposition:

$$h_2^{2,i} = h_0^{2,1} + W_0^{2,1}\, \Delta_2 h^1 + \Delta_2 W^{2,i}\, h_0^1 + \Delta_2 W^{2,i}\, \Delta_2 h^1. \tag{R3s-F3.2}$$

$$\text{init: } W_0^{2,1} h_0^1 = h_0^{2,1} \in \Theta(1). \quad \text{(R3}s\text{-F1.4)} \tag{R3s-F3.2a}$$

$$\text{prop: } W_0^{2,1}\, \Delta_2 h^1 \in \Theta(1) \\ (\text{op-norm of iid Gaussian } W_0^{2,1};\ \Delta_2 h^1 \in \Theta(1) \text{ via (R3}s\text{-U2.1b)}). \tag{R3s-F3.2b}$$

$$\text{eff: } \Delta_2 W^{2,i}\, h_0^1 = -\tfrac{\chi_1 \phi_{i,0}}{N}(W_0^{3,1})^\top h_0^3\, \|h_0^1\|^2 \in \Theta(1) \\ ((W_0^{3,1})^\top h_0^3 \in \Theta(1) \text{ via (R3}s\text{-B2.3)};\ \|h_0^1\|^2 \in \Theta(N)). \tag{R3s-F3.2c}$$

$$\text{cross: } \Delta_2 W^{2,i}\, \Delta_2 h^1 = -\tfrac{\chi_1 \phi_{i,0}}{N}(W_0^{3,1})^\top h_0^3\, \langle h_0^1, \Delta_2 h^1 \rangle \in \Theta(1) \\ ((W_0^{3,1})^\top h_0^3 \in \Theta(1) \text{ via (R3}s\text{-B2.3)}; \\ \langle h_0^1, \Delta_2 h^1 \rangle \in \Theta(N) \text{ coherent via (R3}s\text{-U2.1b)}). \tag{R3s-F3.2d}$$

(R3s-F3.3) $h_2^{3,i} = W_2^{3,i}\, h_2^{2,i}$, four-piece decomposition:

$$h_2^{3,i} = h_0^{3,1} + W_0^{3,1}\, \Delta_2 h^{2,i} + \Delta_2 W^{3,i}\, h_0^{2,1} + \Delta_2 W^{3,i}\, \Delta_2 h^{2,i}. \tag{R3s-F3.3}$$

$$\text{init: } h_0^{3,1} \in \Theta(1). \quad \text{(R3}s\text{-F1.5)} \tag{R3s-F3.3a}$$

$$\text{prop: } W_0^{3,1}\, \Delta_2 h^{2,i} \in \Theta(1) \\ (\text{op-norm of iid Gaussian } W_0^{3,1};\ \Delta_2 h^{2,i} \in \Theta(1) \text{ via (R3}s\text{-F3.2)}). \tag{R3s-F3.3b}$$

$$\text{eff: } \Delta_2 W^{3,i}\, h_0^{2,1} = -\tfrac{\chi_1 \phi_{i,0}}{N}\, h_0^3\, \|h_0^{2,1}\|^2 \in \Theta(1) \\ (\|h_0^{2,1}\|^2 \in \Theta(N_e);\ h_0^3 \in \Theta(1) \text{ via (R3}s\text{-F1.6)}). \tag{R3s-F3.3c}$$

$$\text{cross: } \Delta_2 W^{3,i}\, \Delta_2 h^{2,i} \text{ tracked as part of } D \text{ in (R3}s\text{-F3.4)}. \tag{R3s-F3.3d}$$

(R3s-F3.4) $h_2^3 = A_1 + A_2 + A_3 + D$, where

$$A_1 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, W_0^{3,1} h_0^{2,1}, \qquad A_2 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, \Delta_2 W^{3,i}\, h_0^{2,1},$$
$$A_3 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, W_0^{3,1}\, \Delta_2 h^{2,i}, \qquad D := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, \Delta_2 W^{3,i}\, \Delta_2 h^{2,i}.$$

$$A_1 = \big(\textstyle\sum \phi_{\cdot,2}/M\big)\, h_0^{3,1} \in \Theta(1). \quad (i\text{-independent matrix; LLN}) \tag{R3s-F3.4a}$$

$$A_2 = -\tfrac{\chi_1 \|h_0^{2,1}\|^2}{N}\Big(\tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\phi_{i,0}\Big)\, h_0^3 \in \Theta(1) \text{ along } h_0^3$$
$$(\|h_0^{2,1}\|^2/N \in \Theta(1);\ \text{LLN}). \tag{R3s-F3.4b}$$

$$A_3 \in \Theta(1)$$
(three-piece expansion of $\Delta_2 h^{2,i}$ from (R3s-F3.2);
Lemma J.7 on h_0^3, h_0^1; $\langle h_0^1, \Delta_2 h^1\rangle \in \Theta(N)$; LLN). (R3$s$-F3.4c)

$$D = -\tfrac{\chi_1}{N}\, h_0^3\, \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\phi_{i,0}\, \langle h_0^{2,1}, \Delta_2 h^{2,i}\rangle$$
$$\in \Theta\big(\tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\phi_{i,0}\, \langle h_0^{2,1}, \Delta_2 h^{2,i}\rangle/N\big)\cdot h_0^3 \text{ along } h_0^3$$
(Lem. J.7; LLN)
$$= \Theta(N/N) = \Theta(1) \text{ along } h_0^3$$
$(\langle h_0^{2,1}, \Delta_2 h^{2,i}\rangle \in \Theta(N)$ coherent via (R3s-U2.1b); $h_0^3 \in \Theta(1))$. (R3s-F3.4d)

$$h_2^3 = A_1 + A_2 + A_3 + D \in \Theta(1). \tag{R3s-F3.4}$$

$$f_2 = W_2^4\, h_2^3 = -\tfrac{\chi_0+\chi_1}{N}\langle h_0^3, h_2^3\rangle \in \Theta(1)$$
$(\langle h_0^3, A_1\rangle = (\textstyle\sum \phi_{\cdot,2}/M)\langle h_0^3, h_0^{3,1}\rangle \in \Theta(N)$ coherent via (R3s-F3.4a);
$\langle h_0^3, A_2 + A_3 + D\rangle \in \Theta(N)$ via aligned components (R3s-F3.4b)–(R3s-F3.4d)).
(R3s-F3.5)

Third backward pass

$$\partial f_2/\partial h_2^3 = (W_2^4)^\top = -\tfrac{\chi_0+\chi_1}{N}\, h_0^3 \in \Theta(1/N). \quad ((\text{R3}s\text{-U2.4});\ h_0^3 \in \Theta(1)) \tag{R3s-B3.1}$$

$$\partial f_2/\partial h_2^{3,i} = (\phi_{i,2}/M)(W_2^4)^\top \in \Theta(1/(MN)). \quad (\phi_{i,2} \in \Theta(1)) \tag{R3s-B3.2}$$

$$\partial f_2/\partial h_2^{2,i} = (\phi_{i,2}/M)(W_2^{3,i})^\top (W_2^4)^\top \in \Theta(1/(MN))$$
$((W_0^{3,1})^\top (W_2^4)^\top \in \Theta(1/N)$ via (R3s-B2.3);
$(\Delta_2 W^{3,i})^\top (W_2^4)^\top \in \Theta(1/N)$ along $h_0^{2,1}$ using $\|h_0^3\|^2 \in \Theta(N)$ via (R3s-F1.6);
$\phi_{i,2} \in \Theta(1))$. (R3$s$-B3.3)

$$\partial f_2/\partial \phi_{i,2} = (1/M)(h_2^{3,i})^\top (W_2^4)^\top \in \Theta(1/M)$$
$((h_2^{3,i})^\top (W_2^4)^\top = -\tfrac{\chi_0+\chi_1}{N}\langle h_2^{3,i}, h_0^3\rangle \in \Theta(1)$;
$\langle h_2^{3,i}, h_0^3\rangle \in \Theta(N)$ coherent via (R3s-F3.3)). (R3s-B3.4)

(R3s-B3.5) $(\partial f_2/\partial h_2^1)_{\text{exp}} = \widetilde{A}_4 + \widetilde{A}_5 + \widetilde{A}_6 + \widetilde{E}$, where

$$\widetilde{A}_4 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, (W_0^{2,1})^\top (W_0^{3,1})^\top (W_2^4)^\top,$$
$$\widetilde{A}_5 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, (W_0^{2,1})^\top (\Delta_2 W^{3,i})^\top (W_2^4)^\top,$$
$$\widetilde{A}_6 := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, (\Delta_2 W^{2,i})^\top (W_0^{3,1})^\top (W_2^4)^\top,$$
$$\widetilde{E} := \tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\, (\Delta_2 W^{2,i})^\top (\Delta_2 W^{3,i})^\top (W_2^4)^\top.$$

$$\begin{aligned}\widetilde{A}_4 &= -\tfrac{\chi_0+\chi_1}{N}\big(\textstyle\sum \phi_{\cdot,2}/M\big)(W_0^{2,1})^\top (W_0^{3,1})^\top h_0^3 \\ &\in \Theta\big((W_0^{2,1})^\top (W_0^{3,1})^\top h_0^3/N\big) \\ &\quad (i\text{-indep. matrix under MSSP};\ \text{Lem. J.7};\ \text{LLN}) \\ &= \Theta(1\cdot 1/N) = \Theta(1/N) \\ &\quad ((W_0^{2,1})^\top (W_0^{3,1})^\top h_0^3 \in \Theta(1) \text{ via (R3}s\text{-B2.3)}).\end{aligned} \tag{R3s-B3.5a}$$

$$\begin{aligned}\widetilde{A}_5 &= \tfrac{\chi_1(\chi_0+\chi_1)\,\|h_0^3\|^2}{N^2}\,\tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\phi_{i,0}\,(W_0^{2,1})^\top h_0^{2,1} \\ &\in \Theta\big(\|h_0^3\|^2\,\tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\phi_{i,0}/N^2\big)\cdot h_0^1 \text{ along } h_0^1 \\ &\quad (\text{Lem. J.7: } (W_0^{2,1})^\top h_0^{2,1} \to h_0^1) \\ &= \Theta(N\cdot 1\cdot 1/N^2) = \Theta(1/N) \text{ along } h_0^1 \\ &\quad (\|h_0^3\|^2 \in \Theta(N) \text{ via (R3}s\text{-F1.6)};\ \text{LLN};\ h_0^1 \in \Theta(1)).\end{aligned} \tag{R3s-B3.5b}$$

$$\begin{aligned}\widetilde{A}_6 &= -\tfrac{\chi_1}{N}\,h_0^1\,\tfrac{1}{M}\textstyle\sum_i \phi_{i,0}\,S_i, \qquad S_i := (h_0^3)^\top W_0^{3,1}(W_0^{3,1})^\top (W_2^4)^\top \\ &\in \Theta\big(\tfrac{1}{M}\textstyle\sum_i \phi_{i,0}\,S_i/N\big)\cdot h_0^1 \text{ along } h_0^1 \\ &\quad (\text{Lem. J.7: } S_i \in \Theta(1);\ \text{LLN}) \\ &= \Theta(1\cdot 1/N) = \Theta(1/N) \text{ along } h_0^1 \\ &\quad (S_i \in \Theta(1);\ h_0^1 \in \Theta(1)).\end{aligned} \tag{R3s-B3.5c}$$

$$\begin{aligned}\widetilde{E} &= -\tfrac{\chi_1^2(\chi_0+\chi_1)\,\|h_0^3\|^2}{N^3}\,h_0^1\,\tfrac{1}{M}\textstyle\sum_i \phi_{i,2}\phi_{i,0}^2\,\langle h_0^3, h_0^{3,1}\rangle \\ &\in \Theta\big(\langle h_0^3, h_0^{3,1}\rangle\,\|h_0^3\|^2/N^3\big)\cdot h_0^1 \text{ along } h_0^1 \\ &\quad (\text{LLN};\ \langle h_0^3, h_0^{3,1}\rangle \text{ coherent under MSSP}) \\ &= \Theta(N\cdot N/N^3) = \Theta(1/N) \text{ along } h_0^1 \\ &\quad (\langle h_0^3, h_0^{3,1}\rangle \in \Theta(N) \text{ coherent via (R3}s\text{-F1.6)};\ \|h_0^3\|^2 \in \Theta(N) \text{ via (R3}s\text{-F1.6)}).\end{aligned} \tag{R3s-B3.5d}$$

$$(\partial f_2/\partial h_2^1)_{\text{exp}} \in \Theta(1/N) \text{ along } h_0^1. \tag{R3s-B3.5}$$

$$\begin{aligned}(\partial f_2/\partial h_2^1)_{\text{router}} &= \tfrac{1}{M}Q_0^\top v + \tfrac{1}{M}(\Delta_2 Q)^\top v, \qquad v_i := \dot\phi_{i,2}(h_2^{3,i})^\top (W_2^4)^\top \\ &\in \Theta\big(\sigma_Q\|v\|/M\big) \text{ random direction} + \Theta\big(\langle \partial f_1/\partial\phi, v\rangle/M\big)\,h_1^1 \text{ along } h_1^1 \\ &\quad (Q_0 \text{ piece via cross-layer CLT};\ \Delta_2 Q \text{ piece coherent under MSSP}) \\ &= \Theta(1/\sqrt{N}\cdot\sqrt{M}\cdot 1/M) = \Theta(1/\sqrt{MN}) \\ &\quad (\sigma_Q^2 = 1/N;\ \|v\|^2 \in \Theta(M) \text{ via } v_i \in \Theta(1) \text{ from (R3}s\text{-B3.4)}; \\ &\qquad \langle \partial f_1/\partial\phi, v\rangle \in \Theta(1) \text{ via MSSP coherence}).\end{aligned} \tag{R3s-B3.6}$$

$$\partial f_2/\partial h_2^1 = (\partial f_2/\partial h_2^1)_{\text{exp}} + (\partial f_2/\partial h_2^1)_{\text{router}} \in \Theta(1/N). \tag{R3s-B3.7}$$

J.4.5 Summary table of signal propagation for MSSP in Regime III

Notation: $\Delta_t W^\ell$ denotes the cumulative update $W_t^\ell - W_0^\ell$.

Forward.

Quantity	$t=0$	$t=1$	$t=2$
$h_t^1 = h_0^1 + \Delta_t h^1$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $h_0^1 = W_0^1 x$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
effective: $\Delta_t h^1 = \Delta_t W^1 x$	0	0	$\Theta(1)$
ψ_t, ϕ_t	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$h_t^{2,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $W_0^{2,1} h_0^1$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
propagating: $W_0^{2,1} \Delta_t h^1$	0	0	$\Theta(1)$
effective: $\Delta_t W^{2,i}\, h_0^1$	0	0	$\Theta(1)$
cross: $\Delta_t W^{2,i}\, \Delta_t h^1$	0	0	$\Theta(1)$
$h_t^{3,i}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
init: $W_0^{3,1} h_0^{2,1}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
propagating: $W_0^{3,1} \Delta_t h^{2,i}$	0	0	$\Theta(1)$
effective: $\Delta_t W^{3,i}\, h_0^{2,1}$	0	0	$\Theta(1)$
cross: $\Delta_t W^{3,i}\, \Delta_t h^{2,i}$	0	0	$\Theta(1)$
$h_t^3 = A_1 + A_2 + A_3 + D$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$A_1 = (1/M) \sum_i \phi_{i,t}\, W_0^{3,1} h_0^{2,1}$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$A_2 = (1/M) \sum_i \phi_{i,t}\, \Delta_t W^{3,i}\, h_0^{2,1}$	0	0	$\Theta(1)$
$A_3 = (1/M) \sum_i \phi_{i,t}\, W_0^{3,1}\, \Delta_t h^{2,i}$	0	0	$\Theta(1)$
$D = (1/M) \sum_i \phi_{i,t}\, \Delta_t W^{3,i}\, \Delta_t h^{2,i}$	0	0	$\Theta(1)$
$f_t = W_t^4 h_t^3$	0	$\Theta(1)$	$\Theta(1)$
effective: $\Delta_t W^4\, h_0^3$	0	$\Theta(1)$	$\Theta(1)$
cross: $\Delta_t W^4\, \Delta_t h^3$	0	0	$\Theta(1)$

Backward.

Quantity	$t=0$	$t=1$	$t=2$
$\partial f_t / \partial h_t^3 = (\Delta_t W^4)^\top$	0	$\Theta(1/N)$	$\Theta(1/N)$
$\partial f_t / \partial h_t^{3,i} = (\phi_{i,t}/M)(\Delta_t W^4)^\top$	0	$\Theta(1/(MN))$	$\Theta(1/(MN))$
$\partial f_t / \partial h_t^{2,i} = (\phi_{i,t}/M)(W_t^{3,i})^\top (\Delta_t W^4)^\top$	0	$\Theta(1/(MN))$	$\Theta(1/(MN))$
init: $(\phi_{i,t}/M)(W_0^{3,1})^\top (\Delta_t W^4)^\top$	0	$\Theta(1/(MN))$	$\Theta(1/(MN))$
update: $(\phi_{i,t}/M)(\Delta_t W^{3,i})^\top (\Delta_t W^4)^\top$	0	0	$\Theta(1/(MN))$
$\partial f_t / \partial \phi_{i,t} = (1/M) \langle h_t^{3,i}, \Delta_t W^4 \rangle$	0	$\Theta(1/M)$	$\Theta(1/M)$
$(\partial f_t / \partial h_t^1)_{\text{exp}}$	0	$\Theta(1/N)$	$\Theta(1/N)$
$A_4 = (1/M) \sum_i \phi_{i,t}\, (W_0^{2,1})^\top (W_0^{3,1})^\top (\Delta_t W^4)^\top$	0	$\Theta(1/N)$	$\Theta(1/N)$
$A_5 = (1/M) \sum_i \phi_{i,t}\, (W_0^{2,1})^\top (\Delta_t W^{3,i})^\top (\Delta_t W^4)^\top$	0	0	$\Theta(1/N)$
$A_6 = (1/M) \sum_i \phi_{i,t}\, (\Delta_t W^{2,i})^\top (W_0^{3,1})^\top (\Delta_t W^4)^\top$	0	0	$\Theta(1/N)$
$E = (1/M) \sum_i \phi_{i,t}\, (\Delta_t W^{2,i})^\top (\Delta_t W^{3,i})^\top (\Delta_t W^4)^\top$	0	0	$\Theta(1/N)$
$(\partial f_t / \partial h_t^1)_{\text{router}} = (1/M) Q_t^\top v$	0	$\Theta(1/\sqrt{MN})$	$\Theta(1/\sqrt{MN})$
init: $(1/M) Q_0^\top v$	0	$\Theta(1/\sqrt{MN})$	$\Theta(1/\sqrt{MN})$
update: $(1/M)(\Delta_t Q)^\top v$	0	0	$\Theta(1/\sqrt{MN})$

K DMFT Analysis for Regime I

K.1 Setup

We use the following architecture:

$$
\begin{aligned}
h^1_\mu &= W^1 x_\mu \\
\psi_\mu &= \frac{1}{N} Q \sigma(h^1_\mu) \\
\phi_\mu &= \text{softmax}(\psi_\mu) \\
h^2_{\mu,i} &= \frac{1}{\sqrt{N}} W^2_i \sigma(h^1_\mu), \\
h^{\ell+1}_{\mu,i} &= \frac{1}{\sqrt{N}} W^{\ell+1}_i \sigma(h^\ell_{\mu,i}), \quad \ell \in \{2,3,...,L-1\} \\
h^L_\mu &= \sum_{i=1}^{M} \phi^i_\mu \cdot h^L_{\mu,i} \\
h^{L+1}_\mu &= \frac{1}{\sqrt{N}} w^{L+1\mathsf{T}} h^L_\mu \\
f_\mu &= \frac{1}{\gamma} h^{L+1}_\mu \\
w^{L+1}_\alpha(0) &\sim \mathcal{N}(0,1) \\
W^\ell_{\alpha\beta,i}(0) &\sim \mathcal{N}(0,1) \text{ for } \ell \in \{2,3,...,L\}, i \in \{1,2,...,M\} \\
W^1_{\alpha\beta}(0) &\sim \mathcal{N}(0,1) \\
Q_{\alpha\beta}(0) &\sim \mathcal{N}(0,1) \\
\eta &= \eta_0 \gamma^2 \\
\gamma &= \gamma_0 \sqrt{N} \\
\eta_0, \gamma_0 &\sim O(1)
\end{aligned}
\tag{K.1}
$$

$x_\mu \in \mathbb{R}^D$ is the embedding vector of input $\mu \in \{1,2,...,P\}$, $\psi_\mu \in \mathbb{R}^M$ gives the router's preferences for each of the M experts, ϕ^i_μ denotes the ith component of ϕ_μ, for $i \in \{1,2,...,M\}$, σ is an element-wise nonlinearity, $Q \in \mathbb{R}^{M\times N}$, $w^4 \in \mathbb{R}^N$, $W^2_i, W^3_i \in \mathbb{R}^{N\times N}$, $W^1 \in \mathbb{R}^{N\times D}$, the preactivations h are in $\mathbb{R}^N$, except for h^4_μ, which is a scalar. We define softmax : $\mathbb{R}^p \to (0,1)^p$, $p > 1$ as the function which takes a tuple $\boldsymbol{z} = (z_1, z_2, ..., z_p) \in \mathbb{R}^p$, and computes each component of the vector $\text{softmax}(\boldsymbol{z}) \in (0,1)^p$ with $[\text{softmax}(\boldsymbol{z})]_i = \frac{e^{\beta z_i}}{\sum_{j=1}^p e^{\beta z_j}}$, where $T = \frac{1}{\beta}$ is the temperature of the softmax. We will be interested in the low temperature limit $\beta \to \infty$, where softmax implements Top-1.

We denote for brevity $\boldsymbol{\theta} = \text{Vec}\{W^1, W^\ell_i, w^{L+1}, Q\}_{i,\ell}$, for $i \in \{1,2,...,M\}$ and $\ell \in \{2,3,...,L\}$, and define the gradients

$$
\begin{aligned}
g^l_\mu &= \sqrt{N} \frac{\partial h^{L+1}_\mu}{\partial h^l_\mu}, \; l \in \text{Vec}\{1, (2,i), (3,i), ..., (L,i), L, L+1\}_i \\
g^\phi_\mu &= \frac{1}{\sqrt{N}} \frac{\partial h^{L+1}_\mu}{\partial \phi_\mu}
\end{aligned}
\tag{K.2}
$$

g^1 has two terms, as we must track the gradient flowing both through the router, and through the experts. Using K.4, the components of g^ϕ_μ are

$$
(g^\phi_\mu)_i = \frac{1}{\sqrt{N}} \frac{\partial h^{L+1}_\mu}{\partial \phi^i_\mu} = \frac{1}{\sqrt{N}} \frac{\partial h^{L+1}_\mu}{\partial \boldsymbol{h}^L_\mu} \cdot \frac{\partial \boldsymbol{h}^L_\mu}{\partial \phi^i_\mu} = \frac{1}{N} \boldsymbol{g}^L \cdot \boldsymbol{h}^L_{\mu,i} = O(1). \tag{K.3}
$$

In addition, we define the following *pre-gradient fields* which are defined as the gradients with respect to the activations instead of the pre-activations. For linear layers, they clearly reduce to the gradient fields.

$$\begin{aligned}
g_\mu^L &= z_\mu^L = w^{L+1} \\
g_\mu^{L,i} &= z_{\mu,i}^L = g_\mu^L \phi_\mu^i \\
g_{\mu,i}^\ell &= \dot\sigma(h_{\mu,i}^\ell) \odot z_{\mu,i}^\ell, \quad z_{\mu,i}^\ell = \frac{1}{\sqrt{N}} W_i^{\ell+1\mathsf{T}} g_{\mu,i}^{\ell+1}, \quad \ell \in \{2,3,4,...,L-1\}, i \in \{1,2,...,M\} \\
(g_\mu^\phi)_i &= (z_\mu^\phi)_i = \frac{1}{N} g_\mu^L \cdot h_{\mu,i}^L, \quad i \in \{1,2,...,M\} \\
g_\mu^1 &= \dot\sigma(h_\mu^1) \odot z_\mu^1, \quad z_\mu^1 = \sum_{i=1}^M \frac{1}{\sqrt{N}} W_i^{2\mathsf{T}} g_{\mu,i}^2 + Q^\top g_\mu^\phi
\end{aligned} \tag{K.4}$$

For convenience, we define $\tilde z_\mu^\phi := Q^\top g_\mu^\phi$ and $\tilde z_{\mu,i}^1 := \frac{1}{\sqrt{N}} W_i^{2\mathsf{T}} g_{\mu,i}^2$. Then we have $z_\mu^1 = \sum_{i=1}^M \tilde z_{\mu,i}^1$ and $z_\mu^\phi = \tilde z_\mu^\phi$.

We define also the following feature and gradient kernels (The gradient kernels may be considered to be defined in terms of the z-objects via K.4) :

$$\begin{aligned}
\Phi_{\mu\nu}^0 &:= x_\mu \cdot x_\nu, \\
\Phi_{\mu\nu}^1(s,t) &:= \frac{1}{N} \sigma(h_\mu^1(s)) \cdot \sigma(h_\nu^1(t)), \\
\Phi_{\mu\nu,i}^\ell(s,t) &:= \frac{1}{N} \sigma\big(h_{\mu,i}^\ell(s)\big) \cdot \sigma\big(h_{\nu,i}^\ell(t)\big) \qquad \ell \in \{2,3,...,L-1\}, \\
\Phi_{\mu\nu}^L(s,t) &:= \frac{1}{N} h_\mu^L(s) \cdot h_\nu^L(t) \\
G_{\mu\nu}^1(s,t) &:= \frac{1}{N} g_\mu^1(s) \cdot g_\nu^1(t), \\
G_{\mu\nu,i}^\ell(s,t) &:= \frac{1}{N} g_{\mu,i}^\ell(s) \cdot g_{\nu,i}^\ell(t) \qquad \ell \in \{2,3,...,L\} \\
G_{\mu\nu}^\phi(s,t) &:= g_\mu^\phi(s) g_\nu^\phi(t)
\end{aligned} \tag{K.5}$$

All of these kernels are normalised to be order unity in N.

K.2 Dynamics

Throughout, we take $N \to \infty$ with the number of training samples P and the training time t held fixed. We train using full-batch gradient flow with learning rate η on the loss function

$$\mathcal{L} = \frac{1}{P} \sum_{\mu=1}^P \ell(f_\mu, y_\mu) \tag{K.6}$$

This induces the following dynamics:

$$\begin{aligned}
\frac{d\theta}{dt} &= \frac{\eta_0 \gamma}{P} \sum_\mu \Delta_\mu \frac{\partial h_\mu^{L+1}}{\partial \theta} \\
\Delta_\mu &= -\frac{\partial \mathcal{L}}{\partial f_\mu}
\end{aligned} \tag{K.7}$$

In particular, for a MSE loss $\mathcal{L} = \frac{1}{P} \sum_\nu (y_\nu - f_\nu)^2$, we have $\Delta_\nu = 2(y_\nu - f_\nu)$.

The logits update as

$$\frac{df_\mu(t)}{dt} = \frac{\partial f_\mu(t)}{\partial \theta} \cdot \frac{d\theta}{dt} = \frac{\eta}{P} \sum_\alpha \Delta_\alpha \, K^{\mathrm{NTK}}_{\mu\alpha}(t,t) \tag{K.8}$$

$$\text{where} \quad K^{\mathrm{NTK}}_{\mu\alpha}(t,s) \equiv \frac{\partial f_\mu(t)}{\partial \theta} \cdot \frac{\partial f_\alpha(s)}{\partial \theta} \quad \text{is the neural tangent kernel.} \tag{K.9}$$

We can derive from this definition and the repeated use of chain rule that

$$\gamma^2 K^{NTK}_{\mu\nu}(t,t) = G^1_{\mu\nu}\Phi^0_{\mu\nu} + \sum_{i=1}^{M}[G^2_{\mu\nu,i}\Phi^1_{\mu\nu} + \sum_{\ell=3}^{L} G^\ell_{\mu\nu,i}\Phi^{\ell-1}_{\mu\nu,i}] + \Phi^L_{\mu\nu} + G^\phi_{\mu\nu}\Phi^1_{\mu\nu} \tag{K.10}$$

Hence

$$\begin{aligned}\frac{\mathrm{d}f_\mu(t)}{\mathrm{d}t} = \frac{\eta_0}{P}\sum_{\nu=1}^{P}\Big[&G^1_{\mu\nu}(t,t)\Phi^0_{\mu\nu} + \sum_{i=1}^{M}[G^2_{\mu\nu,i}(t,t)\Phi^1_{\mu\nu}(t,t) + \sum_{\ell=3}^{L} G^\ell_{\mu\nu,i}(t,t)\Phi^{\ell-1}_{\mu\nu,i}(t,t)] + \Phi^L_{\mu\nu}(t,t) \\ &+ G^\phi_{\mu\nu}(t,t)\Phi^1_{\mu\nu}(t,t)\Big]\Delta_\nu(t)\end{aligned} \tag{K.11}$$

Using the learning dynamics K.7, we find for the weight matrices

$$\frac{dW^1(t)}{dt} = \frac{\eta_0\gamma_0}{P}\sum_\mu \Delta_\mu \, g^1_\mu(t) x^\top_\mu \tag{K.12}$$

$$\frac{d\,W^2_i(t)}{dt} = \frac{\eta_0\gamma_0}{\sqrt{N}\,P}\sum_\mu \Delta_\mu \, g^2_{\mu,i}\sigma\big(h^1_\mu\big)^\top \tag{K.13}$$

$$\frac{d\,W^\ell_i(t)}{dt} = \frac{\eta_0\gamma_0}{\sqrt{N}\,P}\sum_\mu \Delta_\mu \, g^\ell_{\mu,i}\sigma\big(h^{\ell-1}_{\mu,i}\big)^\top, \quad \ell \in \{3,4,...,L\} \tag{K.14}$$

$$\frac{d\,w^{L+1}(t)}{dt} = \frac{\eta_0\gamma_0\sqrt{N}}{P}\sum_\mu \Delta_\mu \frac{\partial h^{L+1}_\mu}{\partial w^{L+1}} = \frac{\eta_0\gamma_0}{P}\sqrt{N}\sum_\mu \Delta_\mu \frac{1}{\sqrt{N}} h^L_\mu = \frac{\gamma_0\eta_0}{P}\sum_\mu \Delta_\mu h^L_\mu \tag{K.15}$$

$$\underbrace{\frac{dQ(t)}{dt}}_{\in\mathbb{R}^{M\times N}} = \frac{\eta_0\gamma_0}{P}\sum_{\mu=1}^{P}\Delta_\mu(t)\,\underbrace{g^\phi_\mu(t)}_{\in\mathbb{R}^{M\times 1}}\,\underbrace{[\sigma(h^1_\mu(t))]^\top}_{\in\mathbb{R}^{1\times N}} \tag{K.16}$$

We can now integrate these expressions and use the definitions of the pre-activations from K.1 to obtain:

$$
\begin{aligned}
W^1(t) &= W^1(0) + \int_0^t ds\, \frac{\eta_0\gamma_0}{P} \sum_\mu \Delta_\mu\, g^1_\mu(s) \cdot x_\mu \\
\implies h^1_\mu(t) &= W^1(0)x_\mu + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu\, g^1_\nu(s)\, \Phi^0_{\mu\nu} \\
W^2_i(t) &= W^2_i(0) + \int_0^t ds\, \frac{\eta_0\gamma_0}{\sqrt{N}P} \sum_\mu \Delta_\mu\, g^2_{\mu,i}(s) \sigma(h^1_\mu(s))^\top \\
\implies h^2_{\mu,i}(t) &= \frac{1}{\sqrt{N}} W^2_i(0)\sigma(h^1_\mu(t)) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu\, g^2_{\nu,i}(s)\, \Phi^1_{\mu\nu}(s,t) \\
W^\ell_i(t) &= W^\ell_i(0) + \frac{\eta_0\gamma_0}{\sqrt{N}P} \int_0^t ds \sum_\mu \Delta_\mu\, g^\ell_{\mu,i}(s)\, \sigma\big(h^{\ell-1}_{\mu,i}\big)^T \\
\implies h^\ell_{\mu,i}(t) &= \frac{1}{\sqrt{N}} W^\ell_i(0)\, \sigma\big(h^{\ell-1}_{\mu,i}(t)\big) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu\, g^\ell_{\nu,i}(s)\, \Phi^{\ell-1}_{\mu\nu,i}(s,t), \quad \ell \in \{3,4,...,L\} \\
Q(t) &= Q(0) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_{\mu=1}^P \Delta_\mu(s)\, g^\phi_\mu(s)\, \big(h^1_\mu(s)\big)^\top \\
\implies \psi_\mu(t) &= \frac{1}{N}\, Q(0)\, \sigma(h^1_\mu(t)) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_{\nu=1}^P \Delta_\nu(s)\, g^\phi_\nu(s)\, \Phi^1_{\mu\nu}(s,t)
\end{aligned}
\tag{K.17}
$$

This is useful, because we hope to average over the weights at initialization to obtain expressions in terms of only features. These expressions K.17 motivate the definition of the following fields:

$$
\begin{aligned}
\chi^1_\mu &= W^1(0)x_\mu \in \mathbb{R}^N \\
\chi^2_{\mu,i}(t) &= \frac{1}{\sqrt{N}} W^2_i(0)\sigma(h^1_\mu(t)) \in \mathbb{R}^N \\
\chi^\ell_{\mu,i}(t) &= \frac{1}{\sqrt{N}} W^\ell_i(0)\sigma(h^{\ell-1}_{\mu,i}(t)) \in \mathbb{R}^N \quad \ell \in \{3,4,...,L\} \\
\chi^\phi_\mu(t) &= \frac{1}{N} Q(0)\sigma(h^1_\mu(t)) \in \mathbb{R}^M
\end{aligned}
\tag{K.18}
$$

We repeat the analysis of K.17 for the gradient fields which are defined as $z^\ell_\mu = \frac{1}{\sqrt{N}} W^{\ell+1}(t)^\top g^{\ell+1}_\mu(t)$, where ℓ indexes the layer, and may involve an expert index too, and $z^\phi_\mu(t) = Q(t)^\top g^\phi_\mu(t)$: In particular, we have $\forall i \in \{1,2,...,M\}$

$$
\begin{aligned}
z^\phi_\mu(t) &= \xi^\phi_\mu(t) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_{\nu=1}^P \Delta_\nu(s)\sigma(h^1_\nu(s))G^\phi_{\mu\nu}(t,s) \\
z^1_{\mu,i}(t) &= \xi^{1,i}_\mu(t) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu\, \sigma(h^1_\nu(s))\, G^2_{\mu\nu,i}(s,t) \\
z^\ell_{\mu,i}(t) &= \xi^\ell_{\mu,i}(t) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu\, \sigma(h^\ell_{\nu,i}(s))\, G^{\ell+1}_{\mu\nu,i}(s,t) \quad \ell \in \{2,3,...,L-1\} \\
z^L_{\mu,i} &= \phi^i_\mu z^L_\mu \\
z^L_\mu(t) &= w^{L+1}_\mu(t) = \xi^L_\mu(t) + \frac{\gamma_0\eta_0}{P} \int_0^t ds \sum_\nu \Delta_\nu h^L_\nu,
\end{aligned}
\tag{K.19}
$$

where we define the following fields:

$$
\begin{aligned}
\xi^{\phi}_{\mu}(t) &= Q(0)^{\top} g^{\phi}_{\mu}(t) \\
\xi^{\ell}_{\mu,i}(t) &= \frac{1}{\sqrt{N}} W^{\ell+1}_{i}(0)^{\intercal} g^{\ell+1}_{\mu,i}(t) \quad \ell \in \{1, 2, ..., L-1\} \\
\xi^{L}_{\mu} &= w^{L+1}(0) \sim \mathcal{N}(0, 1)
\end{aligned}
\tag{K.20}
$$

Note that ξ^{L}_{μ} is in fact time-independent, and equal to the initialization of w^{L+1}. We know that this object has entries distributed as $\mathcal{N}(0, 1)$, and we will also recover this from the DMFT for completeness.

Note also that the final gradient is $g^{4}_{\mu} = \sqrt{N}$, so we can think of the final equation in K.19 as involving the identity kernel.

Finally, we write

$$
\begin{aligned}
\big[g^{\phi}\big]_{i} &= \frac{1}{N} \sum_{n=1}^{N} g^{L}_{n} \left[h^{L}_{i}\right]_{n} = \frac{1}{N^{3/2}} \sum_{n,m=1}^{N} g^{L}_{n} \big[W^{L}_{i}\big]_{nm} \sigma\big(\big[h^{L-1}_{i}\big]_{m}\big) \\
\big[g^{\phi}_{\mu}(t)\big]_{i} &= \xi^{g^{\phi}}_{\mu i}(t) \\
&\quad + \frac{1}{N^{2}} \frac{\eta_0 \gamma_0}{P} \int_{0}^{t} ds \sum_{\nu} \Delta_{\nu}(s) \sum_{n,m=1}^{N} g^{L}_{\mu n}(t) g^{L}_{\nu n}(s) \phi^{i}_{\nu}(s) \sigma\big(\big[h^{L-1}_{\mu i}(t)\big]_{m}\big) \, \sigma\big(\big[h^{L-1}_{\nu i}(s)\big]_{m}\big) \\
&= \xi^{g^{\phi}}_{\mu i}(t) + \frac{\eta_0 \gamma_0}{P} \int_{0}^{t} ds \sum_{\nu} \Delta_{\nu}(s) \, G^{L}_{\mu\nu}(t, s) \, \Phi^{L-1}_{\mu\nu}(t, s) \, \phi^{i}_{\nu}(s)
\end{aligned}
\tag{K.21}
$$

Where we define the field

$$
\xi^{g^{\phi}}_{\mu i}(t) = \frac{1}{N^{3/2}} \, g^{L\,\top}_{\mu}(t) \, W^{L}_{i}(0) \, \sigma\big(h^{L-1}_{\mu i}(t)\big). \tag{K.22}
$$

K.3 Mean field theory

If we can show that the kernels defined above concentrate, then our DMFT will be determined by the stochastic fields $\mathcal{F} = \{\chi^{1}_{\mu}, \chi^{2}_{\mu,i}, \chi^{3}_{\mu,i}, ..., \chi^{L}_{\mu,i} \chi^{\phi}_{\mu}, \xi^{1}_{\mu,i}, \xi^{2}_{\mu,i}, ..., \xi^{L-1}_{\mu,i} \xi^{L}_{\mu}, \xi^{\phi}_{\mu}\}_{i \in \{1,2,...,M\}}$. Therefore to construct the DMFT, we must compute the moment generating functional of these stochastic processes.

For brevity, we define, alongside the set $\mathcal{F}$ of fields, the set $\mathcal{S}$ of sources, which contains a corresponding source j (or v) for every χ (or ξ) in $\mathcal{F}$. So $\mathcal{S} = \{j^{1}_{\mu}, j^{2}_{\mu,i}, j^{3}_{\mu,i}, ..., j^{L}_{\mu,i} j^{\phi}_{\mu}, v^{1}_{\mu,i}, v^{2}_{\mu,i}, ..., v^{L-1}_{\mu,i} v^{L}_{\mu}, v^{\phi}_{\mu}\}_{i \in \{1,2,...,M\}}$. This notation is constructed so that we can write $\mathcal{F} \cdot \mathcal{S}$ to denote $j^{1}_{\mu}(t) \cdot \chi^{1}_{\mu}(t) + j^{2}_{\mu,i}(t) \cdot \chi^{2}_{\mu,i}(t) + ... + v^{\phi}_{\mu}(t) \cdot \xi^{\phi}_{\mu}(t)$. Note that $v^{\phi}_{\mu} \in \mathbb{R}^{M}$, but all the other dummy variables are in $\mathbb{R}^{N}$. We also define $\theta_0 = \mathrm{Vec}\big\{W^{1}(0), W^{2}_{i}(0), ..., W^{L}_{i}(0), w^{L+1}(0), Q(0)\big\}_{i \in \{1,...,M\}}$.

The MGF with which we are concerned is therefore

$$
Z[\{j, v\}] \left\langle \exp\left(\sum_{\mu} \int_{0}^{\infty} dt \, \mathcal{F} \cdot \mathcal{S} \right) \right\rangle_{\theta_0} \tag{K.23}
$$

$v^{\phi}_{\mu} \in \mathbb{R}^{M}$, but all the other dummy variables are in $\mathbb{R}^{N}$.

In order to integrate, we enforce definitions of our random fields using resolutions of the identity to obtain:

$$
\begin{aligned}
Z =& \Bigg\langle \frac{1}{(2\pi)^{8N}} \prod_{\mu=1}^{P} \prod_{i=1}^{M} \int_0^\infty dt \left\{ \exp\left(\sum_\mu \int_0^\infty dt\, \mathcal{F} \cdot \mathcal{S} \right) + \right. \\
& \int d\mathcal{F} \exp \left[i \left(\hat{\chi}^1_\mu \cdot \left(\chi^1_\mu - W^1(0) x_\mu \right) + \hat{\chi}^2_{\mu,i} \cdot \left(\chi^2_{\mu,i} - \frac{1}{\sqrt{N}} W^2_i(0) \sigma(h^1_\mu(t)) \right) \right. \right. \\
& \quad + \sum_{\ell=3}^{L} \hat{\chi}^\ell_{\mu,i} \cdot \left(\chi^\ell_{\mu,i} - \frac{1}{\sqrt{N}} W^\ell_i(0) \sigma(h^{\ell-1}_{\mu,i}(t)) \right) + \sum_{\ell=1}^{L-1} \hat{\xi}^\ell_{\mu,i} \cdot \left(\xi^\ell_{\mu,i} - \frac{1}{\sqrt{N}} W^{\ell+1}_i(0)^\top g^{\ell+1}_{\mu,i}(t) \right) \\
& \quad + \hat{\xi}^L_\mu \cdot \left(\xi^L_\mu - w^{L+1}(0)^\top \right) + \hat{\chi}^\phi_\mu \cdot \left(\chi^\phi_\mu - \frac{1}{N} Q(0) \sigma(h^1_\mu(t)) \right) + \hat{\xi}^\phi_\mu \cdot \left(\xi^\phi_\mu - Q(0)^\top g^\phi_\mu(t) \right) \\
& \quad \left. \left. + \hat{\xi}^{g^\phi} \cdot \left(\xi^{g^\phi}_\mu - \frac{1}{N^{3/2}} g^{L\top}_\mu W^L_i(0) \sigma(h^{L-1}_{\mu,i}) \right) \right) \right] \Bigg\} \Bigg\rangle_{\theta_0}
\end{aligned}
\tag{K.24}
$$

We then integrate using the fact that for a Gaussian variable $z \sim \mathcal{N}(0, \sigma^2)$, $\langle e^{-ia \cdot z} \rangle_z \propto e^{-\frac{1}{2}\sigma^2 |a|^2}$ to obtain

$$
\begin{aligned}
Z \propto & \prod_{\mu=1}^{P} \prod_{\nu=1}^{P} \prod_{i=1}^{M} \int_0^\infty dt \int dt \int d\mathcal{F} \\
& \times \exp \Bigg\{ -\frac{1}{2} \sum_{\mu,\nu=1}^{P} \int_0^\infty dt \int_0^\infty ds \Big[\hat{\chi}^1_\mu(t) \cdot \hat{\chi}^1_\nu(s)\, \Phi^0_{\mu\nu}(t,s) \\
& + (\hat{\chi}^2_{\mu,i}(t) \cdot \hat{\chi}^2_{\nu,i}(s) + \frac{1}{N} \hat{\chi}^\phi_\mu(t) \cdot \hat{\chi}^\phi_\nu(s))\, \Phi^1_{\mu\nu}(t,s) + \sum_{\ell=3}^{L} \hat{\chi}^\ell_{\mu,i}(s) \cdot \hat{\chi}^\ell_{\nu,i}(t)\, \Phi^{\ell-1}_{\mu\nu,i}(s,t) \\
& + \sum_{\ell=2}^{L} \hat{\xi}^\ell_{\mu,i}(s) \cdot \hat{\xi}^\ell_{\nu,i}(t)\, G^\ell_{\mu\nu,i}(s,t) + \hat{\xi}^{L+1}_\mu(s) \cdot \hat{\xi}^{L+1}_\nu(t) + \hat{\xi}^\phi_\mu(s) \cdot \hat{\xi}^\phi_\nu(t) G^\phi_{\mu\nu}(t,s) \Big] \\
& - i \sum_{\mu,\nu=1}^{P} \int_0^\infty dt \int_0^\infty ds \left[\sum_{\ell=2}^{L} \hat{\chi}^\ell_{\mu,i}(t) \cdot g^\ell_{\nu,i}(t) A^\ell_{\mu\nu,i}(t,t) + \hat{\chi}^\phi_\mu(t) \cdot g^\phi_\nu(t) A^\phi_{\mu\nu}(t,t) \right] \\
& + \sum_{\mu=1}^{P} \int_0^\infty dt \int_0^\infty ds \left[\sum_{\ell=2}^{L-1} \left((j^\ell_{\mu,i}(t) + i\hat{\chi}^\ell_{\mu,i}(t)) \cdot \chi^\ell_{\mu,i}(t) + (v^\ell_{\mu,i}(t) + i\hat{\xi}^\ell_{\mu,i}) \cdot \xi^\ell_{\mu,i}(t) \right) \right. \\
& + (j^1_\mu(t) + i\hat{\chi}^1_\mu(t)) \cdot \chi^1_\mu(t) \quad + (v^L_\mu + i\hat{\xi}^L_\mu(t)) \cdot \xi^L_\mu(t) \\
& + (j^L_{\mu,i}(t) + i\hat{\chi}^L_{\mu,i}(t)) \cdot \chi^L_{\mu,i}(t) + (v^1_{\mu,i}(t) + i\hat{\xi}^1_{\mu,i}(t)) \cdot \xi^1_{\mu,i}(t) + (v^\phi_\mu + i\hat{\xi}^\phi_\mu(t)) \cdot \xi^\phi_\mu(t) \\
& \left. + (j^\phi_\mu + i\hat{\chi}^\phi_\mu(t)) \cdot \chi^\phi_\mu(t) + (v^{g^\phi}_\mu + i\hat{\xi}^{g^\phi}_\mu(t)) \cdot \xi^{g^\phi}_\mu(t) \right] \Bigg\}.
\end{aligned}
\tag{K.25}
$$

Where we have defined the kernels

$$
A^\ell_{\mu\nu,i}(s,t) = \begin{cases} -\frac{i}{N} \hat{\xi}^1_{\mu,i}(s) \cdot \sigma(h^1_\nu(t)) & \ell = 2 \\ -\frac{i}{N} \hat{\xi}^{\ell-1}_{\mu,i}(s) \cdot \sigma(h^{\ell-1}_{\nu,i}(t)) & \ell \in \{3, 4, ..., L\} \end{cases} \tag{K.26}
$$

$$
A^\phi_{\mu\nu}(s,t) = -\frac{i}{N} \hat{\xi}^\phi_\mu(s) \cdot \sigma(h^1_\nu(t)) \tag{K.27}
$$

There are also three terms arising from the coupling of ξ^{g^ϕ} to itself and to the forwards and backwards fields at layer L. These vanish as $\frac{1}{N}$, so have been excluded from K.25 for brevity.

The Φ and G kernels are required for computing the evolution of the function output, via the NTK. The kernels A are not involved in the NTK, but rather arise from the coupling of the fields across a single layer's initial weight matrix. We now enforce the definitions of all three types of kernels using integral representations of Dirac delta-functions. In particular, for each pair μ, ν of samples, each expert i and each pair t, s of times, we multiply by

$$1 = \int \frac{d\Phi^0_{\mu\nu}(t,s)\, d\hat{\Phi}^0_{\mu\nu}(t,s)}{2\pi i} \exp\Big[\hat{\Phi}^0_{\mu\nu}(t,s)\left(\Phi^0_{\mu\nu}(t,s) - x_\mu \cdot x_\nu\right)\Big] \tag{K.28}$$

$$1 = \int \frac{d\Phi^1_{\mu\nu}(t,s)\, d\hat{\Phi}^1_{\mu\nu}(t,s)}{2\pi i\, N^{-1}} \exp\Big[\hat{\Phi}^1_{\mu\nu}(t,s)\left(N\Phi^1_{\mu\nu}(t,s) - \sigma(h^1_\mu(t)) \cdot \sigma\big(h^1_\nu(s)\big)\right)\Big] \tag{K.29}$$

$$1 = \int \frac{d\Phi^\ell_{\mu\nu,i}(t,s)\, d\hat{\Phi}^\ell_{\mu\nu,i}(t,s)}{2\pi i\, N^{-1}} \exp\Big[\hat{\Phi}^\ell_{\mu\nu,i}(t,s)\big(N\Phi^\ell_{\mu\nu,i}(t,s) - \sigma\big(h^\ell_{\mu,i}(t)\big) \cdot \sigma\big(h^\ell_{\nu,i}(s)\big)\big)\Big] \tag{K.30}$$
$$\ell \in \{2, 3, ..., L-1\}$$

$$1 = \int \frac{d\Phi^L_{\mu\nu}(t,s)\, d\hat{\Phi}^L_{\mu\nu}(t,s)}{2\pi i\, N^{-1}} \exp\Big[\hat{\Phi}^L_{\mu\nu}(t,s)\left(N\Phi^L_{\mu\nu}(t,s) - h^L_\mu(t) \cdot h^L_\nu(s)\right)\Big] \tag{K.31}$$

$$1 = \int \frac{dG^1_{\mu\nu}(t,s)\, d\hat{G}^1_{\mu\nu}(t,s)}{2\pi i\, N^{-1}} \exp\Big[\hat{G}^1_{\mu\nu}(t,s)\left(NG^1_{\mu\nu}(t,s) - g^1_\mu(t) \cdot g^1_\nu(s)\right)\Big] \tag{K.32}$$

$$1 = \int \frac{dG^\ell_{\mu\nu,i}(t,s)\, d\hat{G}^\ell_{\mu\nu,i}(t,s)}{2\pi i\, N^{-1}} \exp\Big[\hat{G}^\ell_{\mu\nu,i}(t,s)\big(NG^\ell_{\mu\nu,i}(t,s) - g^\ell_{\mu,i}(t) \cdot g^\ell_{\nu,i}(s)\big)\Big] \tag{K.33}$$
$$\ell \in \{2, 3, ..., L\}$$

$$1 = \int \frac{dA^2_{\mu\nu,i}(t,s)\, dB^2_{\mu\nu,i}(t,s)}{2\pi i\, N^{-1}} \exp\Big[-B^2_{\mu\nu,i}(t,s)\left(NA^2_{\mu\nu,i}(t,s) + i\hat{\xi}^1_i(t) \cdot \sigma(h^1_\nu(s))\right)\Big] \tag{K.34}$$

$$1 = \int \frac{dA^\phi_{\mu\nu}(t,s)\, dB^\phi_{\mu\nu}(t,s)}{2\pi i\, N^{-1}} \exp\Big[-B^\phi_{\mu\nu}(t,s)\left(NA^\phi_{\mu\nu}(t,s) + i\hat{\xi}^\phi_\mu(t) \cdot \sigma(h^1_\nu(s))\right)\Big] \tag{K.35}$$

$$1 = \int \frac{dA^\ell_{\mu\nu,i}(t,s)\, dB^\ell_{\mu\nu,i}(t,s)}{2\pi i\, N^{-1}} \exp\Big[-B^\ell_{\mu\nu,i}(t,s)\big(NA^\ell_{\mu\nu,i}(t,s) + i\hat{\xi}^{\ell-1}_{\mu,i}(t) \cdot \sigma(h^{\ell-1}_{\nu,i}(s))\big)\Big] \tag{K.36}$$
$$\ell \in \{3, 4, ..., L\}$$

$$1 = \int \frac{dG^\phi_{\mu\nu}(t,s)\, d\hat{G}^\phi_{\mu\nu}(t,s)}{2\pi i} \exp\Big[\hat{G}^\phi_{\mu\nu}(t,s)\left(G^\phi_{\mu\nu}(t,s) + g^\phi_\mu(t) \cdot g^\phi_\nu(t)\right)\Big] \tag{K.37}$$

We call the conjugate kernel of A B, rather than $\hat{A}$. This is because it will turn out to be equal to $\hat{\chi} \cdot g$, which appears in the gradient stream update equation in the final DMFT.

Multiplying in all these factors of unity, we arrive at a partition function which factorises over each of the N sites in each hidden layer. There are two things to note here:

1. We want to treat g,h as scalars, so each of the resolutions of the identity K.28 to K.37 above will be duplicated N times leading to an extensive factor N, which pulls through to the front of the exponential, since all the neurons in a given layer are indistinguishable in the limit, so have the same Φ or G. That is, we can write for instance

$$1 = \int \frac{dG^1_{\mu\nu}(t,s)\, d\hat{G}^1_{\mu\nu}(t,s)}{2\pi i\, N^{-1}} \exp\Big[N\hat{G}^1_{\mu\nu}(t,s)\left(G^1_{\mu\nu}(t,s) - g^1_\mu(t) g^1_\nu(s)\right)\Big] \tag{K.38}$$

   Where $g^1_\mu \in \mathbb{R}$ are the components of the former vector of the same name, which are considered statistically indistinguishable in the limit. The exception to this however is K.37, since $g^\phi_\mu \in \mathbb{R}^M$.

M is not being taken to the infinite limit, so we cannot consider the M components of g^ϕ_μ to be indistinguishable. Rather we must sum over them:

$$1 = \int \frac{dG^\phi_{\mu\nu}(t,s)\, d\hat{G}^\phi_{\mu\nu}(t,s)}{2\pi i} \exp\left[\sum_{\beta=1}^{M} \hat{G}^\phi_{\mu\nu}(t,s)\left(G^\phi_{\mu\nu}(t,s) + g^\phi_{\mu,\beta}(t) g^\phi_{\nu,\beta}(t)\right)\right]. \tag{K.39}$$

2. Each term in the argument of the exponential in K.25 is a dot product of two vectors in $\mathbb{R}^N$, so contains N terms which are identical in the limit. We can therefore bring out a factor N here too. Once again, the exception is the term $(j^\phi_\mu(t) + i\hat{\chi}^\phi_\mu(t))\chi^\phi_\mu(t)$, which is repeated M times. Therefore when we bring a factor N out front of the action, we must include a prefactor $\frac{M}{N}$ with this term.

Define the set $\mathcal{K}$ of all order parameters. This includes all the kernels defined above .

$$Z \propto \int \prod_{\mu,\nu,t,s,i} d\mathcal{K}\, \exp\Big(N S[\{\hat{\Phi}, \Phi, \hat{G}, G, A, B\}]\Big) \tag{K.40}$$

Where the DMFT action $S$is $O(1)$, and has the form

$$\begin{aligned} S[\{\hat{\Phi}, \Phi, \hat{G}, G, A, B\}] = \sum_{\mu,\nu} \int_0^\infty dt \int_0^\infty ds \Big[& \hat{G}^1_{\mu\nu}(t,s) G^1_{\mu\nu}(t,s) + \sum_{\ell=2}^{L} \hat{G}^\ell_{\mu\nu,i}(t,s) G^\ell_{\mu\nu,i}(t,s) \\ & + \hat{\Phi}^1_{\mu\nu}(t,s)\Phi^1_{\mu\nu}(t,s) + \sum_{\ell=2}^{L-1} \hat{\Phi}^2_{\mu\nu,i}(t,s)\Phi^2_{\mu\nu,i}(t,s) + \hat{\Phi}^L_{\mu\nu}(t,s)\Phi^L_{\mu\nu}(t,s) \\ & - \sum_{L=2}^{L} B^\ell_{\mu\nu,i}(t,s) A^\ell_{\mu\nu,i}(t,s) + \frac{1}{N}\hat{G}^\phi_{\mu\nu}(t,s) G^\phi_{\mu\nu}(t,s) - B^\phi_{\mu\nu}(t,s) A^\phi_{\mu\nu}(t,s) \Big] \\ & + \frac{1}{N}\sum_{\alpha=1}^{N} \ln \mathcal{Z}_N[\Phi, \hat{\Phi}, G, \hat{G}, A, B, j_\alpha, v_\alpha] + \frac{1}{N}\sum_{\beta=1}^{M} \ln \mathcal{Z}_M[\Phi^1, \hat{G}^\phi, j^\phi_\beta] \end{aligned} \tag{K.41}$$

such that it consists of inner products of the order parameters $\{\Phi, G, A\}$ and their duals $\{\hat{\Phi}, \hat{G}, B\}$, in addition to the single-site MGFs

$$
\begin{aligned}
&\mathcal{Z}_N[\{\Phi,\hat{\Phi},G,\hat{G},A,B,j,v\}]\\
&\quad=\prod_{\mu,\nu,i,t}\int_0^\infty dt\int d\mathcal{F}\\
&\quad\times\exp\Bigg(+\sum_{\mu=1}^{P}\int_0^\infty dt\int_0^\infty ds\Bigg[\sum_{\ell=2}^{L}\Big((j^\ell_{\mu,i}(t)+i\hat{\chi}^\ell_{\mu,i}(t))\cdot\chi^\ell_{\mu,i}(t)+(v^\ell_{\mu,i}(t)+i\hat{\xi}^\ell_{\mu,i})\cdot\xi^\ell_{\mu,i}(t)\Big)\\
&\quad+(j^1_\mu(t)+i\hat{\chi}^1_\mu(t))\cdot\chi^1_\mu(t)+(v^{L+1}_\mu+i\hat{\xi}^{L+1}_\mu(t))\cdot\xi^{L+1}_\mu(t)+(j^\phi_\mu(t)+i\hat{\chi}^\phi_\mu(t))\cdot\chi^\phi_\mu(t)\\
&\quad+(v^\phi_\mu(t)+i\hat{\xi}^\phi_\mu(t))\cdot\xi^\phi_\mu(t)\Bigg]\Bigg)\\
&\quad\times\exp\Bigg(-\frac{1}{2}\sum_{\mu,\nu,i}\int_0^\infty dt\int_0^\infty ds\Big[\hat{\chi}^1_\mu(t)\cdot\hat{\chi}^1_\nu(s)\,\Phi^0_{\mu\nu}(t,s)+(\hat{\chi}^2_{\mu,i}(t)\cdot\hat{\chi}^2_{\nu,i}(s)\\
&\quad+\frac{1}{N}\hat{\chi}^\phi_\mu(t)\cdot\hat{\chi}^\phi_\mu(s))\Phi^1_{\mu\nu}(t,s)+\sum_{\ell=3}^{L}\hat{\chi}^\ell_{\mu,i}(s)\cdot\hat{\chi}^\ell_{\nu,i}(t)\,\Phi^{\ell-1}_{\mu\nu,i}(s,t)+\sum_{\ell=2}^{L}\hat{\xi}^\ell_{\mu,i}(s)\cdot\hat{\xi}^\ell_{\nu,i}(t)\,G^\ell_{\mu\nu,i}(s,t)\\
&\quad+\hat{\xi}^{L+1}_\mu(s)\cdot\hat{\xi}^{L+1}_\nu(t)+\hat{\xi}^\phi_\mu(t)\hat{\xi}^\phi_\nu(s)G^\phi_{\mu\nu}(t,s)\Big]\\
&\qquad-i\sum_{\mu,\nu=1}^{P}\int_0^\infty dt\int_0^\infty ds\sum_{\ell=2}^{L}\hat{\chi}^\ell_{\mu,i}(t)\cdot g^\ell_{\nu,i}(t)A^\ell_{\mu\nu,i}(t,t)\\
&\quad+i\sum_{\mu}\int_0^\infty dt\left[\hat{h}^3_\mu(t)h^3_\mu(t)-\sum_{i=1}^{M}\phi^i_\mu(t)\hat{h}^3_\mu(t)h^3_\mu(t)\right]\\
&\quad\times\exp\Bigg(-\sum_{\mu,\nu,i}\int_0^\infty dt\int_0^\infty dt\Big[x_\mu x_\nu\hat{\Phi}^0_{\mu\nu}+\sigma(h^1_\mu(t))\sigma(h^1_\nu(s))\hat{\Phi}^1_{\mu\nu}(t,s)\\
&\quad+\sum_{\ell=2}^{L-1}\sigma(h^\ell_{\mu,i}(t))\sigma(h^\ell_{\nu,i}(s))\hat{\Phi}^\ell_{\mu\nu,i}(t,s)+h^L_\mu(t)h^L_\nu(s)\hat{\Phi}^L_{\mu\nu}(t,s)+g^1_\mu(t)g^1_\nu(s)\hat{G}^1_{\mu\nu}(t,s)\\
&\quad+\sum_{\ell=2}^{L}g^\ell_{\mu,i}(t)g^\ell_{\nu,i}(s)\hat{G}^\ell_{\mu\nu,i}(t,s)+\hat{\xi}^2_{\mu,i}(t)\sigma(h^1_\nu(s))B^2_{\mu\nu,i}(t,s)\\
&\quad+i\sum_{\ell=3}^{L}\hat{\xi}^\ell_{\mu,i}(t)\sigma(h^{\ell-1}_{\nu,i}(s))B^\ell_{\mu\nu,i}(t,s)+i\hat{\xi}^\phi_\mu(t)\sigma(h^1_\nu(s))B^\phi_{\mu\nu}(t,s)\Big]\Bigg)
\end{aligned}
\tag{K.42}
$$

And

$$
\begin{aligned}
&\mathcal{Z}_M[\{\Phi^1,\hat{G}^\phi,j^\phi_\beta\}]\\
&\quad=\prod_{\mu,\nu,t}\int_0^\infty dt\int d\hat{\chi}^\phi_{\mu,\beta}(t)\,d\chi^\phi_{\mu,\beta}(t)\exp\Big((j^\phi_{\mu,\beta}(t)+i\hat{\chi}_{\mu,\beta}(t))\chi^\phi_\mu(t)+(v^{g^\phi}_\mu+i\hat{\xi}^{g^\phi}_\mu(t))\cdot\xi^{g^\phi}_\mu(t)\\
&\qquad-\frac{1}{2}\frac{1}{N}\hat{\chi}^\phi_{\mu,\beta}(t)\hat{\chi}^\phi_{\nu,\beta}(s)\Phi^1_{\mu\nu}(t,s)-g^\phi_{\mu,\beta}g^\phi_{\nu,\beta}\hat{G}^\phi_{\mu\nu}(t,s)-i\hat{\chi}^\phi_{\mu,\beta}(t)g^\phi_{\nu,\beta}(t)A^\phi_{\mu\nu}(t,t)\Big)
\end{aligned}
\tag{K.43}
$$

The sum over β is an average over sites (rows of vectors in $\mathbb{R}^M$). We regard $h_\mu(t)$ and $g_\mu(t)$ as functions of χ and ξ.

It is manifest in the form of K.40 that the actual state of the system in the limit $N\to\infty$ will be a saddle point of the action. This is the point where $\delta S=0$ for *any* variation of our many order parameters. That is, we impose the following $\forall t,s,\mu,\nu,i$:

$$
\begin{aligned}
&\frac{\delta \mathcal{S}}{\delta G^1_{\mu\nu}(t,s)} = \hat{G}^1_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta G^1_{\mu\nu}(t,s)} = \hat{G}^1_{\mu\nu}(t,s) - 0 = 0 \\
&\frac{\delta \mathcal{S}}{\delta G^\ell_{\mu\nu,i}(t,s)} = \hat{G}^\ell_{\mu\nu,i}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta G^\ell_{\mu\nu,i}(t,s)} = \hat{G}^\ell_{\mu\nu,i}(t,s) - \frac{1}{2}\frac{1}{N}\sum_{\alpha=1}^{N}\langle \hat{\xi}^\ell_{\mu,i}(t)\hat{\xi}^\ell_{\nu,i}(s)\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta \Phi^1_{\mu\nu}(t,s)} = \hat{\Phi}^1_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta \Phi^1_{\mu\nu}(t,s)} = \hat{\Phi}^1_{\mu\nu}(t,s) - \frac{1}{2}(\frac{1}{N}\sum_{\alpha=1}^{N}\langle \hat{\chi}^2_{\mu,i}(t)\hat{\chi}^2_{\nu,i}(s)\rangle_\alpha \\
&\qquad\qquad + \frac{1}{N}\frac{1}{N}\sum_{\beta=1}^{M}\langle \hat{\chi}^\phi_{\mu}(t)\hat{\chi}^\phi_{\nu}(s)\rangle_\beta) = 0 \\
&\frac{\delta \mathcal{S}}{\delta \Phi^\ell_{\mu\nu,i}(t,s)} = \hat{\Phi}^\ell_{\mu\nu,i}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta \Phi^\ell_{\mu\nu,i}(t,s)} = \hat{\Phi}^\ell_{\mu\nu,i}(t,s) - \frac{1}{2}\frac{1}{N}\sum_{\alpha=1}^{N}\langle \hat{\chi}^\ell_{\mu,i}(t)\hat{\chi}^\ell_{\nu,i}(s)\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta \Phi^L_{\mu\nu}(t,s)} = \hat{\Phi}^L_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta \Phi^L_{\mu\nu}(t,s)} = \hat{\Phi}^L_{\mu\nu}(t,s) - 0 = 0 \\
&\frac{\delta \mathcal{S}}{\delta A^\ell_{\mu\nu,i}(t,s)} = -B^\ell_{\mu\nu,i}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta A^\ell_{\mu\nu,i}(t,s)} = -B^\ell_{\mu\nu,i}(t,s) - i\frac{1}{N}\sum_{\alpha=1}^{N}\langle \hat{\chi}^\ell_{\mu,i}(t) g^\ell_{\nu,i}(s)\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta A^\phi_{\mu\nu}(t,s)} = -B^\phi_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta A^\phi_{\mu\nu}(t,s)} = -B^\phi_{\mu\nu}(t,s) - i\frac{1}{N}\sum_{\beta=1}^{M}\langle \hat{\chi}^\phi_{\mu}(t) g^\phi_{\nu}(s)\rangle_\beta = 0 \\
&\frac{\delta \mathcal{S}}{\delta G^\phi_{\mu\nu}(t,s)} = \frac{1}{N}\hat{G}^\phi_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta G^\phi_{\mu\nu}(t,s)} = \frac{1}{N}\hat{G}^\phi_{\mu}(t,s) - \frac{1}{2}\frac{1}{N}\sum_{\alpha=1}^{N}\langle \hat{\xi}^\phi_{\mu}(t)\hat{\xi}^\phi_{\nu}(s)\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta \hat{\Phi}^1_{\mu\nu}(t,s)} = \Phi^1_{\mu\nu}(t,s) - \frac{1}{N}\sum_{\alpha=1}^{N}\langle \sigma(h^1_\mu(t))\sigma(h^1_\nu(s))\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta \hat{\Phi}^\ell_{\mu\nu,i}(t,s)} = \Phi^\ell_{\mu\nu,i}(t,s) - \frac{1}{N}\sum_{\alpha=1}^{N}\langle \sigma(h^\ell_{\mu,i}(t))\,\sigma(h^\ell_{\nu,i}(s))\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta \hat{\Phi}^L_{\mu\nu}(t,s)} = \Phi^L_{\mu\nu}(t,s) - \frac{1}{N}\sum_{\alpha=1}^{N}\langle h^L_\mu(t) h^L_\nu(s)\rangle_\alpha = 0
\end{aligned}
\tag{K.44}
$$

$$
\begin{aligned}
&\frac{\delta \mathcal{S}}{\delta \hat{G}^1_{\mu\nu}(t,s)} = G^1_{\mu\nu}(t,s) - \frac{1}{N}\sum_{\alpha=1}^{N}\langle g^1_\mu(t) g^1_\nu(s)\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta \hat{G}^\ell_{\mu\nu,i}(t,s)} = G^\ell_{\mu\nu,i}(t,s) - \frac{1}{N}\sum_{\alpha=1}^{N}\langle g^\ell_{\mu,i}(t) g^\ell_{\nu,i}(s)\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta B^2_{\mu\nu,i}(t,s)} = -A^2_{\mu\nu,i}(t,s) - i\frac{1}{N}\sum_{\alpha=1}^{N}\langle \hat{\xi}^2_{\mu,i}(t)\sigma(h^1_\nu(s))\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta B^\ell_{\mu\nu,i}(t,s)} = -A^\ell_{\mu\nu,i}(t,s) - i\frac{1}{N}\sum_{\alpha=1}^{N}\langle \hat{\xi}^\ell_{\mu,i}(t)\sigma(h^{\ell-1}_{\nu,i}(s))\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta B^\phi_{\mu\nu}(t,s)} = -A^\phi_{\mu\nu}(t,s) - i\frac{1}{N}\sum_{\alpha=1}^{N}\langle \hat{\xi}^\phi_{\mu}(t)\sigma(h^1_\nu(s))\rangle_\alpha = 0 \\
&\frac{\delta \mathcal{S}}{\delta \hat{G}^\phi_{\mu\nu}(t,s)} = \frac{1}{N}G^\phi_{\mu\nu}(t,s) - \frac{1}{N}\sum_{\beta=1}^{M}\langle g^\phi_\mu(t) g^\phi_\nu(s)\rangle_\beta = 0
\end{aligned}
\tag{K.45}
$$

Here the average $\langle\rangle_\alpha$ denotes the αth single-site average of an observable $\mathcal{O}(\{\chi,\xi,u\})$, defined as:

$$\langle\mathcal{O}(\{\chi,\xi\})\rangle_\alpha = \frac{1}{\mathcal{Z}[j_\alpha, v_\alpha]}\int\prod_\mu\prod_{\text{layers}} d\chi_\mu^\ell d\xi_\mu^\ell \exp(-\mathcal{H}[\{\chi,\xi\},\{j,v\}])\mathcal{O}(\{\chi,\xi\}) \tag{K.46}$$

Where $\mathcal{H}$ is the logarithm of the integrand in K.42 or K.43.

At zero source $j, v \to 0$, all single-site averages over N terms are equivalent, and we can write (for instance) $\Phi^1_{\mu\nu} = \langle\sigma(h^1_\mu)\,\sigma(h^1_\nu)\,\rangle$, where $\langle\cdot\rangle$ is the average over single-site distributions for $j, v \to 0$. The sum $\frac{1}{N}\frac{1}{N}\sum_{\beta=1}^M\langle\hat{\chi}^\phi_\mu(t)\hat{\chi}^\phi_\nu(s)\rangle_\beta \to 0$ in the limit, but notably, $\frac{1}{N^2}\sum_{\beta=1}^M\langle g^\phi_\mu(t)g^\phi_\nu(s)\rangle_\beta$ remains

$O(1)$ but doesn't concentrate. We have

$$
\begin{aligned}
\frac{\delta \mathcal{S}}{\delta G^1_{\mu\nu}(t,s)} &= \hat{G}^1_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta G^1_{\mu\nu}(t,s)} = \hat{G}^1_{\mu\nu}(t,s) - 0 = 0 \\
\frac{\delta \mathcal{S}}{\delta G^\ell_{\mu\nu,i}(t,s)} &= \hat{G}^\ell_{\mu\nu,i}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta G^\ell_{\mu\nu,i}(t,s)} = \hat{G}^\ell_{\mu\nu,i}(t,s) - \frac{1}{2}\langle \hat{\xi}^\ell_{\mu,i}(t)\hat{\xi}^\ell_{\nu,i}(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \Phi^0_{\mu\nu}} &= \hat{\Phi}^0_{\mu\nu} - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta \Phi^0_{\mu\nu}} = \hat{\Phi}^0_{\mu\nu} - \frac{1}{2N}\langle \hat{\chi}^1_\mu(t)\hat{\chi}^1_\nu(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \Phi^1_{\mu\nu}(t,s)} &= \hat{\Phi}^1_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta \Phi^1_{\mu\nu}(t,s)} = \hat{\Phi}^1_{\mu\nu}(t,s) - \frac{1}{2}(\langle \hat{\chi}^2_{\mu,i}(t)\hat{\chi}^2_{\nu,i}(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \Phi^\ell_{\mu\nu,i}(t,s)} &= \hat{\Phi}^\ell_{\mu\nu,i}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta \Phi^\ell_{\mu\nu,i}(t,s)} = \hat{\Phi}^\ell_{\mu\nu,i}(t,s) - \frac{1}{2}\langle \hat{\chi}^\ell_{\mu,i}(t)\hat{\chi}^\ell_{\nu,i}(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \Phi^L_{\mu\nu}(t,s)} &= \hat{\Phi}^L_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta \Phi^L_{\mu\nu}(t,s)} = \hat{\Phi}^L_{\mu\nu}(t,s) - 0 = 0 \\
\frac{\delta \mathcal{S}}{\delta A^\ell_{\mu\nu,i}(t,s)} &= -B^\ell_{\mu\nu,i}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta A^\ell_{\mu\nu,i}(t,s)} = -B^\ell_{\mu\nu,i}(t,s) - i\langle \hat{\chi}^\ell_{\mu,i}(t) g^\ell_{\nu,i}(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta A^\phi_{\mu\nu}(t,s)} &= -B^\phi_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta A^\phi_{\mu\nu}(t,s)} = -B^\phi_{\mu\nu}(t,s) = 0 \\
\frac{\delta \mathcal{S}}{\delta G^\phi_{\mu\nu}(t,s)} &= \frac{1}{N}\hat{G}^\phi_{\mu\nu}(t,s) - \frac{1}{\mathcal{Z}}\frac{\delta \mathcal{Z}}{\delta G^\phi_{\mu\nu}(t,s)} = \frac{1}{N}\hat{G}^\phi_{\mu}(t,s) - \frac{1}{2}\langle \hat{\xi}^\phi_\mu(t)\hat{\xi}^\phi_\nu(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \hat{\Phi}^1_{\mu\nu}(t,s)} &= \Phi^1_{\mu\nu}(t,s) - \langle \sigma(h^1_\mu(t))\sigma(h^1_\nu(s))\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \hat{\Phi}^\ell_{\mu\nu,i}(t,s)} &= \Phi^\ell_{\mu\nu,i}(t,s) - \langle \sigma(h^\ell_{\mu,i}(t))\,\sigma(h^\ell_{\nu,i}(s))\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \hat{\Phi}^L_{\mu\nu}(t,s)} &= \Phi^L_{\mu\nu}(t,s) - \langle h^L_\mu(t) h^L_\nu(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \hat{G}^1_{\mu\nu}(t,s)} &= G^1_{\mu\nu}(t,s) - \langle g^1_\mu(t) g^1_\nu(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \hat{G}^\ell_{\mu\nu,i}(t,s)} &= G^\ell_{\mu\nu,i}(t,s) - \langle g^\ell_{\mu,i}(t) g^\ell_{\nu,i}(s)\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta B^2_{\mu\nu,i}(t,s)} &= -A^2_{\mu\nu,i}(t,s) - i\langle \hat{\xi}^2_{\mu,i}(t)\sigma(h^1_\nu(s))\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta B^\ell_{\mu\nu,i}(t,s)} &= -A^\ell_{\mu\nu,i}(t,s) - i\langle \hat{\xi}^\ell_{\mu,i}(t)\sigma(h^{\ell-1}_{\nu,i}(s))\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta B^\phi_{\mu\nu}(t,s)} &= -A^\phi_{\mu\nu}(t,s) - i\langle \hat{\xi}^\phi_{\mu}(t)\sigma(h^1_\nu(s))\rangle = 0 \\
\frac{\delta \mathcal{S}}{\delta \hat{G}^\phi_{\mu\nu}(t,s)} &= G^\phi_{\mu\nu}(t,s) - \sum_{\beta=1}^{M}\langle g^\phi_\mu(t) g^\phi_\nu(s)\rangle_\beta = 0
\end{aligned}
\tag{K.47}
$$

K.4 Averages of dual variables

We see that the dual variables $\{\hat{\Phi}, \hat{G}, B\}$ are given in terms of such correlators as $\langle \hat{\chi}\hat{\chi}\rangle$. We will now show that these vanish, finding along the way expressions for the fields $\{\chi, \xi\}$, as well as the kernels $\{A, B\}$.

It will serve brevity to work with a vectorised notation. To wit: let $\boldsymbol{\chi}^\ell_i = \mathrm{Vec}\{\chi^\ell_{\mu,i}(t)\}_{\mu\in\{1,\dots,P\},t\in\mathbb{R}^+}$ denote the vectorization of the stochastic field over different samples and times, with analogous objects defined for the other fields. We denote the dot product between such ob-

jects $\boldsymbol{a}\cdot\boldsymbol{b}=\sum_{\mu=1}^{P}\int_0^\infty a_\mu(t)b_\mu(t)$. A similar procedure is applied to matrices by defining $\boldsymbol{\Phi}=\text{Mat}\{\Phi_{\mu\nu}\}_{\mu\nu\in\{1,\dots,P\},ts\in\mathbb{R}^+}$, with the appropriate matrix product defined as $[\boldsymbol{Ab}]_{\mu,t}=\int_0^t ds\frac{1}{P}\sum_{\nu=1}^{P}A_{\mu\nu}(t,s)b_\nu(s)$.

With this notation in place, we can write $\langle\hat{\boldsymbol{\chi}}_i^\ell\hat{\boldsymbol{\chi}}_i^\ell\rangle$ for $\ell\in\{3,...,L-1\}$ in terms of the moment generating function for $\hat{\boldsymbol{\chi}}_i^\ell$, and the dummy field $\boldsymbol{w}$

$$\begin{aligned}
\langle\hat{\boldsymbol{\chi}}_i^\ell\hat{\boldsymbol{\chi}}_i^\ell\rangle &= -\frac{\partial^2}{\partial\boldsymbol{w}\,\partial\boldsymbol{w}^\top}\left\langle\exp\big(i\,\boldsymbol{w}\cdot\hat{\boldsymbol{\chi}}_i^\ell\big)\right\rangle\Big|_{\boldsymbol{w}=0}\\
&= -\frac{\partial^2}{\partial\boldsymbol{w}\,\partial\boldsymbol{w}^\top}\left[\frac{1}{\mathcal{Z}}\int d\hat{\boldsymbol{\chi}}_i^\ell\,d\boldsymbol{\chi}_i^\ell\;\exp\big(i\,\boldsymbol{w}\cdot\hat{\boldsymbol{\chi}}_i^\ell\big)\,\exp\big(-\mathcal{H}\big)\right]\Bigg|_{\boldsymbol{w}=0}\\
&= -\frac{1}{\mathcal{Z}}\frac{\partial^2}{\partial\boldsymbol{w}\,\partial\boldsymbol{w}^\top}\int d\hat{\boldsymbol{\chi}}_i^\ell\,d\boldsymbol{\chi}_i^\ell\;\exp\big(-\tfrac{1}{2}\,\hat{\boldsymbol{\chi}}_i^{\ell\top}\boldsymbol{\Phi}_i^{\ell-1}\hat{\boldsymbol{\chi}}_i^\ell+i\big(-\boldsymbol{\chi}_i^\ell+\boldsymbol{w}+\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\cdot\hat{\boldsymbol{\chi}}_i^\ell\big)\Big|_{\boldsymbol{w}=0}\\
&= -\frac{1}{\mathcal{Z}}\frac{\partial^2}{\partial\boldsymbol{w}\,\partial\boldsymbol{w}^\top}\int d\boldsymbol{\chi}_i^\ell\;\exp\Big(-\tfrac{1}{2}\,\big(-\boldsymbol{\chi}_i^\ell+\boldsymbol{w}+\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)^\top[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\big(-\boldsymbol{\chi}_i^\ell+\boldsymbol{w}+\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\Big)\Big|_{\boldsymbol{w}=0}\\
&= \frac{1}{\mathcal{Z}}\int d\boldsymbol{\chi}_i^\ell\;\exp\Big(-\tfrac{1}{2}\,\big(-\boldsymbol{\chi}_i^\ell+\boldsymbol{w}+\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)^\top[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\big(-\boldsymbol{\chi}_i^\ell+\boldsymbol{w}+\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\Big)\\
&\qquad\times\Big([\boldsymbol{\Phi}_i^{\ell-1}]^{-1}-[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\big(-\boldsymbol{\chi}_i^\ell+\boldsymbol{w}+\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\big(-\boldsymbol{\chi}_i^\ell+\boldsymbol{w}+\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)^\top[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\Big)\Big|_{\boldsymbol{w}=0}\\
&= [\boldsymbol{\Phi}_i^{\ell-1}]^{-1}-[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\Big\langle\big(\boldsymbol{\chi}_i^\ell-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\big(\boldsymbol{\chi}_i^\ell-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)^\top\Big\rangle[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}
\end{aligned}\tag{K.48}$$

Where we use the definition K.46 of the average, and retain only the relevant parts of the integral. Through similar derivations, we obtain also the following:

$$\langle\hat{\boldsymbol{\chi}}_i^2\hat{\boldsymbol{\chi}}_i^2\rangle=[\boldsymbol{\Phi}^1]^{-1}-[\boldsymbol{\Phi}^1]^{-1}\Big\langle\big(\boldsymbol{\chi}_i^2-\mathbf{A}_i^2\mathbf{g}_i^2\big)\big(\boldsymbol{\chi}_i^2-\mathbf{A}_i^2\mathbf{g}_i^2\big)^\top\Big\rangle[\boldsymbol{\Phi}^1]^{-1}\tag{K.49}$$

$$\langle\hat{\boldsymbol{\chi}}^1\hat{\boldsymbol{\chi}}^1\rangle=\big[\boldsymbol{\Phi}^0\big]^{-1}-\big[\boldsymbol{\Phi}^0\big]^{-1}\Big\langle\boldsymbol{\chi}^1\boldsymbol{\chi}^{1\top}\Big\rangle\big[\boldsymbol{\Phi}^0\big]^{-1}\tag{K.50}$$

$$\langle\hat{\boldsymbol{\xi}}_i^\ell\hat{\boldsymbol{\xi}}_i^\ell\rangle=[\boldsymbol{G}_i^\ell]^{-1}-[\boldsymbol{G}_i^\ell]^{-1}\Big\langle\big(\boldsymbol{\xi}_i^\ell-\mathbf{B}_i^\ell\mathbf{g}_i^\ell\big)\big(\boldsymbol{\xi}_i^\ell-\mathbf{B}_i^\ell\mathbf{g}_i^\ell\big)^\top\Big\rangle[\boldsymbol{G}_i^\ell]^{-1}\quad\ell\in\{2,3,...,L\}\tag{K.51}$$

$$\langle\hat{\boldsymbol{\xi}}^\phi\hat{\boldsymbol{\xi}}^\phi\rangle=[\boldsymbol{G}^\phi]^{-1}-[\boldsymbol{G}^\phi]^{-1}\Big\langle\big(\boldsymbol{\xi}^\phi-\mathbf{B}^\phi\mathbf{g}^\phi\big)\big(\boldsymbol{\xi}^\phi-\mathbf{B}^\phi\mathbf{g}^\phi\big)^\top\Big\rangle[\boldsymbol{G}^\phi]^{-1}\tag{K.52}$$

$$\begin{aligned}
-i\,\big\langle\hat{\boldsymbol{\chi}}_i^\ell\,\mathbf{g}_i^\ell\big\rangle &= \frac{\partial}{\partial\mathbf{w}}\Big\langle\exp\big(-i\,\mathbf{w}\cdot\hat{\boldsymbol{\chi}}_i^\ell\big)\,\mathbf{g}_i^{\ell\top}\Big\rangle\Big|_{\mathbf{w}=0}\\
&= \frac{1}{\mathcal{Z}}\frac{\partial}{\partial\mathbf{w}}\int d\hat{\boldsymbol{\chi}}_i^\ell\,d\boldsymbol{\chi}_i^\ell\;\exp\big(-\tfrac{1}{2}\,\hat{\boldsymbol{\chi}}_i^{\ell\top}\boldsymbol{\Phi}_i^{\ell-1}\hat{\boldsymbol{\chi}}_i^\ell+i\big(\boldsymbol{\chi}_i^\ell+\mathbf{w}-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\cdot\hat{\boldsymbol{\chi}}_i^\ell\big)\,\mathbf{g}_i^{\ell\top}\Big|_{\mathbf{w}=0}\\
&= \frac{1}{\mathcal{Z}}\frac{\partial}{\partial\mathbf{w}}\int d\boldsymbol{\chi}_i^\ell\;\exp\Big(-\tfrac{1}{2}\,\big(\boldsymbol{\chi}_i^\ell+\mathbf{w}-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)^\top[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\big(\boldsymbol{\chi}_i^\ell+\mathbf{w}-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\Big)\,\mathbf{g}_i^{\ell\top}\Big|_{\mathbf{w}=0}\\
&= -\frac{1}{\mathcal{Z}}\int d\boldsymbol{\chi}_i^\ell\;\exp\Big(-\tfrac{1}{2}\,\big(\boldsymbol{\chi}_i^\ell+\mathbf{w}-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)^\top[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\big(\boldsymbol{\chi}_i^\ell+\mathbf{w}-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\Big)\\
&\qquad\qquad\times[\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\big(\boldsymbol{\chi}_i^\ell+\mathbf{w}-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\mathbf{g}_i^{\ell\top}\Big|_{\mathbf{w}=0}\\
&= [\boldsymbol{\Phi}_i^{\ell-1}]^{-1}\Big\langle\big(\boldsymbol{\chi}_i^\ell-\mathbf{A}_i^\ell\mathbf{g}_i^\ell\big)\,\mathbf{g}_i^{\ell\top}\Big\rangle\quad\ell\in\{2,3,...,L\}
\end{aligned}\tag{K.53}$$

$$-i\,\Big\langle\hat{\boldsymbol{\xi}}_i^\ell\,\boldsymbol{\sigma}\big(\mathbf{h}_i^{\ell-1}\big)\Big\rangle=\big[\mathbf{G}_i^\ell\big]^{-1}\Big\langle\big(\boldsymbol{\xi}_i^\ell-\mathbf{B}_i^\ell\boldsymbol{\sigma}\big(\mathbf{h}_i^{\ell-1}\big)\big)\,\boldsymbol{\sigma}\big(\mathbf{h}_i^{\ell-1}\big)\Big\rangle\quad\ell\in\{3,4,...,L\}\tag{K.54}$$

$$-i\left\langle \hat{\boldsymbol{\xi}}^{\phi}\,\sigma(\mathbf{h}^1)\right\rangle = \left[\mathbf{G}^{\phi}\right]^{-1}\left\langle \left(\boldsymbol{\xi}^{\phi}-\mathbf{B}^{\phi}\sigma(\mathbf{h}^1)\right)\sigma(\mathbf{h}^1)\right\rangle \tag{K.55}$$

$$-i\left\langle \hat{\boldsymbol{\xi}}^{\,2,i}\,\sigma\big(\mathbf{h}^1\big)\right\rangle = \left[\mathbf{G}^{\,2,i}\right]^{-1}\left\langle \left(\boldsymbol{\xi}^{\,2,i}-\mathbf{B}_i^2\sigma\big(\mathbf{h}^1\big)\right)\sigma\big(\mathbf{h}^1\big)\right\rangle \tag{K.56}$$

Now the Hubbard Trick Hubbard (1959) states that

$$\exp\!\big(-\tfrac{1}{2}\,\mathbf{x}^{\top}A\,\mathbf{x}\big)=\textstyle\int_{\mathbb{R}^d}\frac{d\mathbf{u}}{(2\pi)^{d/2}\sqrt{\det A}}\,\exp\!\big(-\tfrac{1}{2}\,\mathbf{u}^{\top}A^{-1}\mathbf{u}-i\,\mathbf{u}\cdot\mathbf{x}\big)=\big\langle\,\exp(-i\,\mathbf{u}\cdot\mathbf{x})\big\rangle_{\mathbf{u}\sim\mathcal{N}(0,A)}. \tag{K.57}$$

This allows us to rewrite the quadratic terms in our single-site MGF as follows:

$$\begin{aligned}\exp\!\left(-\frac{1}{2}\sum_{\mu\nu}\int_0^{\infty}dt\int_0^{\infty}ds\;\hat{\chi}^1_{\mu}(t)\cdot\hat{\chi}^1_{\nu}(s)\;\Phi^0_{\mu\nu}(t,s)\right)=\\ =\left\langle\exp\!\left(-i\sum_{\mu}\int_0^{\infty}dt\;\alpha^1_{\mu}(t)\,\hat{\chi}^1_{\mu}(t)\right)\right\rangle_{\alpha^1\sim\mathcal{GP}\,(0,\Phi^{\,0})}\end{aligned} \tag{K.58}$$

$$\begin{aligned}\exp\left(-\frac{1}{2}\sum_{\mu\nu i}\int_0^{\infty}dt\int_0^{\infty}ds\;\hat{\chi}^2_{\mu,i}(t)\cdot\hat{\chi}^2_{\nu,i}(s)\,\Phi^1_{\mu\nu}(t,s)\right)=\\ =\left\langle\exp\left(-i\sum_{\mu i}\int_0^{\infty}dt\;\alpha^2_{\mu,i}(t)\,\hat{\chi}^2_{\mu,i}(t)\right)\right\rangle_{\{\alpha^2_i\}\sim\mathcal{GP}\,(0,\Phi^{\,1})}\end{aligned} \tag{K.59}$$

$$\begin{aligned}\exp\!\left(-\frac{1}{2}\sum_{\mu\nu i}\int_0^{\infty}dt\int_0^{\infty}ds\;\hat{\chi}^{\ell}_{\mu,i}(t)\cdot\hat{\chi}^{\ell}_{\nu,i}(s)\,\Phi^{\ell-1}_{\mu\nu,i}(t,s)\right)=\\ =\left\langle\exp\!\left(-i\sum_{\mu i}\int_0^{\infty}dt\;\alpha^{\ell}_{\mu,i}(t)\,\hat{\chi}^{\ell}_{\mu,i}(t)\right)\right\rangle_{\{\alpha^{\ell}_i\}\sim\mathcal{GP}\left(0,\Phi^{\ell-1}_i\right)}\qquad\ell\in\{3,4,...,L\}\end{aligned} \tag{K.60}$$

$$\begin{aligned}\exp\!\left(-\frac{1}{2}\sum_{\mu\nu i}\int_0^{\infty}dt\int_0^{\infty}ds\;\hat{\xi}^{\ell}_{\mu,i}(t)\cdot\hat{\xi}^{\ell}_{\nu,i}(s)\,G^{\ell}_{\mu\nu,i}(t,s)\right)=\\ =\left\langle\exp\!\left(-i\sum_{\mu i}\int_0^{\infty}dt\;\beta^{\ell}_{\mu,i}(t)\,\hat{\xi}^{\ell}_{\mu,i}(t)\right)\right\rangle_{\{\beta^{\ell}_i\}\sim\mathcal{GP}\left(0,G^{\ell}_i\right)}\qquad\ell\in\{2,3,4,...,L\}\end{aligned} \tag{K.61}$$

$$\begin{aligned}\exp\!\left(-\frac{1}{2}\sum_{\mu\nu i}\int_0^{\infty}dt\int_0^{\infty}ds\;\hat{\xi}^{L+1}_{\mu}(t)\cdot\hat{\xi}^{L+1}_{\nu}(s)\right)=\\ =\left\langle\exp\!\left(-i\sum_{\mu i}\int_0^{\infty}dt\;\beta^{L+1}_{\mu}(t)\,\hat{\xi}^{L+1}_{\mu}(t)\right)\right\rangle_{\{\beta^{L+1}\}\sim\mathcal{GP}\,(0,1)}\end{aligned} \tag{K.62}$$

$$\exp\!\left(-\frac{1}{2}\sum_{\mu\nu i}\int_0^{\infty}dt\int_0^{\infty}ds\;\hat{\xi}^{\phi}_{\mu}(t)\cdot\hat{\xi}^{\phi}_{\nu}(s)\right)=\left\langle\exp\!\left(-i\sum_{\mu i}\int_0^{\infty}dt\;\beta^{\phi}_{\mu}(t)\,\hat{\xi}^{\phi}_{\mu}(t)\right)\right\rangle_{\{\beta^{\phi}\}\sim\mathcal{GP}\,(0,G^{\phi})} \tag{K.63}$$

We can now easily integrate over all the $\{\hat{\chi}, \hat{\xi}\}$ variables, since the argument of the exponential in $\mathcal{Z}$ has been linearised with respect to them all. Doing so yields the following delta functions:

$$\int \prod_{\mu,t} \frac{d\hat{\chi}^1_\mu(t)}{2\pi} \, \exp\big(i\,\hat{\boldsymbol{\chi}}^1 \cdot (\boldsymbol{\chi}^1 - \boldsymbol{\alpha}^1)\big) \;=\; \delta\big(\boldsymbol{\chi}^1 - \boldsymbol{\alpha}^1\big) \tag{K.64}$$

$$\int \prod_{\mu,t} \frac{d\hat{\chi}^{\,\ell}_{\mu,i}(t)}{2\pi} \, \exp\big[i\,\hat{\boldsymbol{\chi}}^{\ell}_i \cdot (\boldsymbol{\chi}^{\ell}_i - \boldsymbol{\alpha}^{\ell}_i - \mathbf{A}^{\ell}_i\,\mathbf{g}^{\ell}_i)\big] = \delta\big(\boldsymbol{\chi}^{\ell}_i - \boldsymbol{\alpha}^{\ell}_i - \mathbf{A}^{\ell}_i\,\mathbf{g}^{\ell}_i\big) \quad \ell \in \{2,3,...,L\} \tag{K.65}$$

$$\int \prod_{\mu,t} \frac{d\hat{\xi}^{2,i}_\mu(t)}{2\pi} \, \exp\Big(i\,\hat{\boldsymbol{\xi}}^{\ell}_i \cdot \Big(\boldsymbol{\xi}^2_i - \boldsymbol{\beta}^2_i - \mathbf{B}^{2\top}_i\,\sigma\big(\mathbf{h}^1\big)\Big)\Big) = \delta\big(\boldsymbol{\xi}^2_i - \boldsymbol{\beta}^2_i - \mathbf{B}^{2\top}_i\,\sigma\big(\mathbf{h}^1\big)\big) \tag{K.66}$$

$$\int \prod_{\mu,t} \frac{d\hat{\xi}^{\,\ell}_{\mu,i}(t)}{2\pi} \, \exp\Big(i\,\hat{\boldsymbol{\xi}}^{\ell}_i \cdot \Big(\boldsymbol{\xi}^{\ell}_i - \boldsymbol{\beta}^{\ell}_i - \mathbf{B}^{\ell\top}_i\,\sigma\big(\mathbf{h}^{\ell-1}_i\big)\Big)\Big) = \delta\big(\boldsymbol{\xi}^{\ell}_i - \boldsymbol{\beta}^{\ell}_i - \mathbf{B}^{\ell\top}_i\,\sigma\big(\mathbf{h}^{\ell-1}_i\big)\big) \quad \ell \in \{3,4,...,L\} \tag{K.67}$$

$$\int \prod_{\mu,t} \frac{d\hat{\xi}^{\phi}_\mu(t)}{2\pi} \, \exp\Big(i\,\hat{\boldsymbol{\xi}}^{\phi} \cdot \Big(\boldsymbol{\xi}^{\phi} - \boldsymbol{\beta}^{\phi} - \mathbf{B}^{\phi\top}\,\sigma(\mathbf{h}^1)\Big)\Big) = \delta\big(\boldsymbol{\xi}^{\phi} - \boldsymbol{\beta}^{\phi} - \mathbf{B}^{\phi\top}\,\sigma(\mathbf{h}^1)\big) \tag{K.68}$$

$$\int \prod_{\mu,t} \frac{d\hat{\xi}^{L+1}_\mu(t)}{2\pi} \, \exp\Big(i\,\hat{\boldsymbol{\xi}}^{L+1} \cdot \Big(\boldsymbol{\xi}^{L+1} - \boldsymbol{\beta}^{L+1}\Big)\Big) = \delta\big(\boldsymbol{\xi}^{L+1} - \boldsymbol{\beta}^{L+1}\big) \tag{K.69}$$

Using these in the expressions K.49 to K.56, we note that all four classes of correlators vanish, since the average term collapses to the covariance matrix of the respective Gaussian process $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$, which cancels with its inverse, giving zero overall. We summarise this as follows (where the expression is implied to hold for all allowed arguments, superscripts and subscripts):

$$\hat{G} = \hat{\Phi} = \langle \hat{\chi}\hat{\chi} \rangle = \langle \hat{\xi}\hat{\xi} \rangle = 0 \tag{K.70}$$

K.5 The coupling kernels A and B

What of A and its dual B? Well inserting K.66 into K.54, and inserting this in turn into the relevant saddle point equation, we obtain for $\ell \in \{3,4,...,L\}$

$$\begin{aligned} A^{\ell}_{\mu\nu,i}(t,s) &= -\,i\,\big\langle \hat{\xi}^{\ell}_{\nu,i}(t)\,\sigma\big(h^{\ell-1}_{\mu,i}(s)\big)\big\rangle \\ &= \big[G^{\ell}_i\big]^{-1}\Big\langle \big(\xi^{\ell}_i - B^{\ell}_i\,\sigma\big(h^{\ell-1}_i\big)\big)\,\sigma\big(h^{\ell-1}_i\big)\Big\rangle \\ &= \left\langle \frac{\partial\sigma(h^{\ell-1}_{\mu,i}(t))}{\partial\beta^{\ell\top}_{\nu,i}(s)} \right\rangle \end{aligned} \tag{K.71}$$

In the final line, we use Stein's lemma Stein (1981) which states that for a normally distributed random variable X with expectation μ and variance σ^2, and differentiable function g, $\mathbb{E}[g(X)(X-\mu)] = \sigma^2\mathbb{E}[g'(X)]$.

Similarly:

$$A^2_{\mu\nu,i}(t,s) = \left\langle \frac{\partial \sigma(h^1_\mu(t))}{\partial \beta^{2\top}_{\nu,i}(s)} \right\rangle \tag{K.72}$$

$$A^\phi_{\mu\nu}(t,s) = \left\langle \frac{\partial \sigma(h^1_\mu(t))}{\partial \beta^{\phi\top}_{\nu}(s)} \right\rangle \tag{K.73}$$

$$B^\ell_{\mu\nu,i}(t,s) = \left\langle \frac{\partial g^\ell_{\mu,i}(t)}{\partial \alpha^{\ell\top}_{\nu,i}(s)} \right\rangle \tag{K.74}$$

Note that the saddle equations tell us that $B^\phi_{\mu\nu}(t,s) = 0$.

K.6 Final DMFT

To remedy the asymmetry manifest in equations K.66 and K.67, we redefine B as its transpose. We also rescale A to $\gamma_0\eta_0 A$, and B to $\gamma_0\eta_0 B$. This makes it clear that the non-Gaussian corrections to h and z are $O(\gamma_0\eta_0)$. We also replace instances of g with their corresponding z using K.4.

We then have the following complete self-consistent DMFT equations:

$$\begin{aligned}
\alpha^1(t) &\sim \mathcal{GP}(0, \Phi^0), \qquad \alpha^\ell_i(t) \sim \begin{cases} \mathcal{GP}(0, \Phi^1(t,s)) & \ell = 2 \\ \mathcal{GP}(0, \Phi^{\ell-1}_i(t,s)) & \ell \in \{3,4,...,L\} \end{cases} \\
\beta^\ell_i(t) &\sim \mathcal{GP}(0, G^{\ell+1}_i(t,s)) \quad \ell \in \{1,2,...,L-1\}, \qquad \beta^L(t) \sim \mathcal{N}(0,1), \\
\beta^\phi(t) &\sim \mathcal{GP}(0, G^\phi(t,s)) \qquad \Phi^0_{\mu\nu} = \langle x_\mu x_\nu \rangle, \qquad \Phi^1_{\mu\nu}(t,s) = \langle \sigma(h^1_\mu(t)) \sigma(h^1_\nu(s)) \rangle, \\
\Phi^\ell_{\mu\nu,i}(t,s) &= \langle \sigma(h^\ell_{\mu,i}(t))\, \sigma(h^\ell_{\nu,i}(s)) \rangle \quad \ell \in \{2,3,...,L-1\}, \\
\Phi^L_{\mu\nu}(t,s) &= \left\langle \sum_{i,j=1}^M \phi^i_\mu(t)\, \phi^j_\nu(s)\, h^L_{\mu,i}(t)\, h^L_{\nu,i}(s) \right\rangle, \\
G^1_{\mu\nu}(t,s) &= \langle [\dot\sigma(h^1_\mu(t)) \odot z^1_\mu(t)][\dot\sigma(h^1_\nu(s)) \odot z^1_\nu(s)] \rangle \\
G^\ell_{\mu\nu,i}(t,s) &= \langle [\dot\sigma(h^\ell_{\mu,i}(t)) \odot z^\ell_{\mu,i}(t)][\dot\sigma(h^\ell_{\nu,i}(s)) \odot z^\ell_{\nu,i}(s)] \rangle \quad \ell \in \{2,3,...,L-1\}, \\
G^L_{\mu\nu,i}(t,s) &= \langle z^L_{\mu,i}(t) z^L_{\nu,i}(s) \rangle \\
G^\phi_{\mu\nu}(t,s) &= \sum_{i=1}^M \langle h^L_{\mu,i}(t) z^L_\mu(t) \rangle_i \langle h^L_{\nu,i}(s) z^L_\nu(s) \rangle_i \\
A^2_{\mu\nu,i}(t,s) &= \left\langle \frac{\partial \sigma(h^1_\mu(t))}{\partial \beta^{2\top}_{\nu,i}(s)} \right\rangle, \qquad A^\ell_{\mu\nu,i}(t,s) = \left\langle \frac{\partial \sigma(h^{\ell-1}_{\mu,i}(t))}{\partial \beta^{\ell\top}_{\nu,i}(s)} \right\rangle \quad \ell \in \{3,4,...,L\}, \\
A^\phi_{\mu\nu}(t,s) &= \left\langle \frac{\partial \sigma(h^1_\mu(t))}{\partial \beta^{\phi\top}_{\nu}(s)} \right\rangle \\
B^\ell_{\mu\nu,i}(t,s) &= \left\langle \frac{\partial (\dot\sigma(h^\ell_{\mu,i}(t)) z^\ell_{\mu,i}(t))}{\partial \alpha^{\ell\top}_{\nu,i}(s)} \right\rangle \quad \ell \in \{2,3,...,L-1\}, \qquad B^L_{\mu\nu,i}(t,s) = \left\langle \frac{\partial z^L_{\mu,i}(t)}{\partial \alpha^{L\top}_{\nu,i}(s)} \right\rangle
\end{aligned} \tag{K.75}$$

$$
\begin{aligned}
\tilde{z}^1_{\mu,i}(t) &= \beta^1_{\mu,i}(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu [B^2_{\mu\nu,i}(s,t) + \Delta_\nu(s) G^2_{\mu\nu,i}(s,t)]\, \sigma(h^1_\nu(s)) \\
\tilde{z}^\phi_\mu(t) &= \beta^\phi_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu [\Delta_\nu(s) G^\phi_{\mu\nu}(s,t)]\, \sigma(h^1_\nu(s)), \\
z^1_\mu(t) &= \sum_{i=1}^M \tilde{z}^1_{\mu,i}(t) + \tilde{z}^\phi_\mu \\
z^\ell_{\mu,i}(t) &= \beta^\ell_{\mu,i}(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu [B^{\ell+1}_{\mu\nu,i}(s,t) + \Delta_\nu(s) G^{\ell+1}_{\mu\nu,i}(s,t)]\, \sigma(h^\ell_{\nu,i}(s)) \\
&\qquad\qquad \ell \in \{2,3,...,L-1\} \\
z^L_\mu(t) &= \beta^L_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu \Delta_\nu(s)\, h^L_\nu(s), \qquad\qquad z^L_{\mu,i}(t) = \phi^i(t) z^L_\mu(t)
\end{aligned}
\tag{K.76}
$$

$$
\begin{aligned}
h^1_\mu(t) &= \alpha^1_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu \Delta_\nu\, \Phi^0_{\mu\nu}\, (\dot\sigma(h^1_\mu) \odot z^1_\mu(s)), \\
h^2_{\mu,i}(t) &= \alpha^2_{\mu,i}(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu [A^2_{\mu\nu,i}(s,t) + \Delta_\nu(s)\Phi^1_{\mu\nu}(s,t)]\, (\dot\sigma(h^2_{\nu,i}(s)) \odot z^2_{\nu,i}(s)), \\
h^\ell_{\mu,i}(t) &= \alpha^\ell_{\mu,i}(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu [A^\ell_{\mu\nu,i}(s,t) + \Delta_\nu(s)\Phi^{\ell-1}_{\mu\nu,i}(s,t)]\, (\dot\sigma(h^\ell_{\nu,i}(s)) \odot z^\ell_{\nu,i}(s)) \\
&\qquad\qquad \ell \in \{3,4,...,L-1\}, \\
h^L_{\mu,i}(t) &= \alpha^L_{\mu,i}(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu [A^L_{\mu\nu,i}(s,t) + \Delta_\nu(s)\Phi^{L-1}_{\mu\nu,i}(s,t)]\, z^L_{\nu,i}(s)\,, \\
h^L_\mu &= \sum_{i=1}^M \phi^i_\mu \cdot h^L_{\mu,i}, \qquad\qquad \phi_\mu(t) = \text{softmax}(\psi_\mu(t)) \\
\frac{\mathrm{d}f_\mu(t)}{\mathrm{d}t} &= \frac{\eta_0}{P}\sum_{\nu=1}^P \Big[G^1_{\mu\nu}(t,t)\Phi^0_{\mu\nu} + \sum_{i=1}^M [G^2_{\mu\nu,i}(t,t)\Phi^1_{\mu\nu}(t,t) + \sum_{\ell=3}^L G^\ell_{\mu\nu,i}(t,t)\Phi^{\ell-1}_{\mu\nu,i}(t,t)] \\
&\quad + \Phi^L_{\mu\nu}(t,t) + G^\phi_{\mu\nu}(t,t)\Phi^1_{\mu\nu}(t,t)\Big]\Delta_\nu(t) \\
\psi^i_\mu(t) &= \frac{\gamma_0\eta_0}{P}\int_0^t ds \sum_{\nu=1}^P [A^\phi_{\nu\mu}(s,t) + \Delta_\nu(s)\Phi^1_{\nu\mu}(s,t)]\, \left[g^\phi_\nu(s)\right]_i\,, \\
\left[g^\phi_\mu(t)\right]_i &= \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu \Delta_\nu(s)\, G^L_{\mu\nu}(t,s)\, \Phi^{L-1}_{\mu\nu}(t,s)\, \phi^i_\nu(s)
\end{aligned}
\tag{K.77}
$$

L DMFT Analysis for Regime II (MSSP)

L.1 Architectural definitions

Here we distinguish the external width N and the internal width N_e, maintaining a standard continuous MoE architecture, but now with $Q \in \mathbb{R}^{M\times N}$ and $W_i^{2,\text{in}} \in \mathbb{R}^{N_e\times N}$. All other architectural objects live in the standard spaces and the total number of experts is M.

We are concerned with the thermodynamic limit where $M, N \to \infty$ at a fixed ratio. Specifically, we define $\kappa = \frac{M}{N}$, where κ is a constant of order unity in N. The internal expert width N_e will also remain strictly order unity in this setting. The forward pass, following the ***MSSP***, is defined as:

$$\begin{aligned}
\mathbb{R}^N \ni h_\mu^1 &= \frac{1}{\sqrt{D}} W^1 x_\mu & \mathbb{R}^M \ni \psi_\mu &= \frac{1}{\sqrt{N}} Q\sigma(h_\mu^1) \\
\mathbb{R}^{N_e} \ni h_{\mu,i}^{2,\text{in}} &= \frac{1}{\sqrt{N}} W_i^{2,\text{i}} \sigma(h_\mu^1) & \mathbb{R}^M \ni \phi_\mu &= \text{softmax}(\psi_\mu) \\
\mathbb{R}^N \ni h_{\mu,i}^{2,\text{o}} &= \sqrt{\frac{\kappa N}{N_e}} W_i^{2,\text{o}} \sigma(h_{\mu,i}^{2,\text{in}}) & \mathbb{R}^N \ni h_\mu^3 &= \sum_{i=1}^{M} \phi_\mu^i \, h_{\mu,i}^{2,\text{o}} \\
\mathbb{R} \ni h_\mu^4 &= \frac{1}{\sqrt{N}} w^{4\intercal} h_\mu^3 & f_\mu &= \frac{1}{\gamma} h_\mu^4
\end{aligned} \tag{L.1}$$

L.2 Learning rates and initialization

To ensure that the network evolves dynamically in the infinite-width limit without gradients vanishing or exploding, we apply specific learning rate scalings:

$$\eta = \eta_0 \gamma^2 \qquad \eta_Q = \eta_0 \gamma^2 \kappa \qquad \gamma = \gamma_0 \sqrt{N} \qquad \eta_0, \gamma_0 \sim O(1) \tag{L.2}$$

The network weights are initialized from standard Gaussian distributions:

$$w_\alpha^4(0), \left[W_i^{2,\text{out}}(0)\right]_{\alpha\beta}, \left[W_i^{2,\text{in}}(0)\right]_{\alpha\beta}, W_{\alpha\beta}^1(0), Q_{\alpha\beta}(0) \sim \mathcal{N}(0,1) \tag{L.3}$$

For brevity in our derivations, we denote the collection of all network parameters as $\boldsymbol{\theta} = \text{Vec}\{W^1, W_i^{2,\text{in}}, W_i^{2,\text{out}}, w^4, Q\}$.

Due to the normalization constraint of the softmax operator over M experts, the routing probabilities scale as $\phi^i \sim O(\frac{1}{\kappa N})$. It is mathematically tidier to formulate the DMFT using variables which remain $O(1)$ in the limit. We therefore define and track the rescaled routing variables:

$$\tilde{\phi}_\mu^i := \kappa N \phi_\mu^i \tag{L.4}$$

L.3 Gradient definitions

We mathematically *define* the pre-activation gradients, ensuring they contain the correct scaling factors to remain finite in the $N \to \infty$ limit:

$$
\begin{aligned}
g^1_\mu &:= \sqrt{N}\frac{\partial h^4_\mu}{\partial h^1_\mu} & g^{2,\text{in}}_{\mu,i} &:= \kappa\sqrt{N}\frac{\partial h^4_\mu}{\partial h^{2,\text{in}}_{\mu,i}} \\
g^{2,\text{out}}_{\mu,i} &:= \kappa N^{\frac{3}{2}}\frac{\partial h^4_\mu}{\partial h^{2,\text{o}}_{\mu,i}} & g^3_\mu &:= \sqrt{N}\frac{\partial h^4_\mu}{\partial h^3_\mu} = w^4 = z^3 \\
g^\phi_\mu &:= \frac{1}{\sqrt{N}}\frac{\partial h^4_\mu}{\partial \phi_\mu} & z^{2,\text{out}}_{\mu,i} &:= \kappa N \phi^i_\mu g^3_\mu = \tilde{\phi}^i_\mu g^3_\mu \\
z^{2,\text{in}}_{\mu,i} &:= \sqrt{\frac{\kappa}{N_e N}}\, W^{2,\text{out}\,\top}_i g^{2,\text{out}}_{\mu,i} & \tilde{z}^\phi_\mu &:= \frac{1}{\kappa\sqrt{N}} Q^\top g^{1,\phi} \\
\tilde{z}^1_{\mu,i} &:= \frac{1}{\kappa\sqrt{N}} W^{2,\text{in}\,\top}_i g^{2,\text{in}}_{\mu,i} & z^1 &:= \frac{1}{\kappa\sqrt{N}} Q^\top g^{1,\phi} + \frac{1}{\kappa\sqrt{N}}\sum_{i=1}^M W^{2,\text{in}\,\top}_i g^{2,\text{in}}_{\mu,i} \\
& & &= \frac{1}{\kappa\sqrt{N}}\sum_{i=1}^M \tilde{z}^1_{\mu,i} + \tilde{z}^{1,\phi}_\mu
\end{aligned}
\tag{L.5}
$$

Note that the components of the router gradient g^ϕ_μ naturally scale as $O(1)$:

$$
(g^\phi_\mu)_i = \frac{1}{\sqrt{N}}\frac{\partial h^4_\mu}{\partial \phi^i_\mu} = \frac{1}{\sqrt{N}}\frac{\partial h^4_\mu}{\partial \boldsymbol{h}^3_\mu}\cdot\frac{\partial \boldsymbol{h}^3_\mu}{\partial \phi^i_\mu} = \frac{1}{N}\boldsymbol{g}^3\cdot \boldsymbol{h}^{2,\text{out}}_{\mu,i} = O(1). \tag{L.6}
$$

The backward pass through the router dictates that the gradient at layer 1 involves the quantity $g^{1,\phi}$, defined component-wise to handle the Jacobian of the softmax:

$$
g^{1,\phi}_k := \kappa N \sum_{i=1}^{\kappa N} g^\phi_i \phi^i(\delta_{ik} - \phi^k) = \sum_{i=1}^{\kappa N} g^\phi_i \tilde{\phi}^i(\delta_{ik} - \frac{\tilde{\phi}^k}{\kappa N}) \tag{L.7}
$$

L.4 DMFT kernels and order parameters

We define the following macroscopic *kernels*, all of which are $O_N(1)$ and will serve as the fundamental order parameters in the dynamics:

$$
\begin{aligned}
\Phi^0_{\mu\nu} &= \frac{1}{D}\, x_\mu \cdot x_\nu & \Phi^1_{\mu\nu}(s,t) &= \frac{1}{N}\,\sigma\big(h^1_\mu(s)\big)\cdot\sigma\big(h^1_\nu(t)\big) \\
\Phi^{2,\text{in}}_{\mu\nu,i}(s,t) &= \frac{1}{N_e}\,\sigma\big(h^{2,\text{in}}_{\mu,i}(s)\big)\cdot\sigma\big(h^{2,\text{in}}_{\nu,i}(t)\big) & \Phi^{2,\text{o}}_{\mu\nu,i}(s,t) &= \frac{1}{N}\,\sigma\big(h^{2,\text{out}}_{\mu,i}(s)\big)\cdot\sigma\big(h^{2,\text{out}}_{\nu,i}(t)\big) \\
\Phi^3_{\mu\nu}(s,t) &= \frac{1}{N}\, h^3_\mu(s)\cdot h^3_\nu(t) & G^1_{\mu\nu}(s,t) &= \frac{1}{N}\, g^1_\mu(s)\cdot g^1_\nu(t) \\
G^{2,\text{in}}_{\mu\nu,i}(s,t) &= \frac{1}{N_e}\, g^{2,\text{in}}_{\mu,i}(s)\cdot g^{2,\text{in}}_{\nu,i}(t) & G^{2,\text{out}}_{\mu\nu,i}(s,t) &= \frac{1}{N}\, g^{2,\text{out}}_{\mu,i}(s)\cdot g^{2,\text{out}}_{\nu,i}(t) \\
G^{1,\phi}_{\mu\nu}(s,t) &= \frac{1}{\kappa N}\, g^{1,\phi}_\mu(s)\cdot g^{1,\phi}_\nu(t) & G^3_{\mu\nu}(s,t) &= \frac{1}{N}\, g^3_\mu(s)\cdot g^3_\nu(t)
\end{aligned}
\tag{L.8}
$$

With the exception of the input data Gram matrix Φ^0, all of these kernels are dynamically evolving macroscopic variables.

L.5 Learning Dynamics and the Neural Tangent Kernel

We train the network using full-batch continuous-time gradient flow on the P training samples, and take $N, M \to \infty$ with P and the training time t held fixed. The variable learning rate scheme defined previously is necessary so that the activation updates within the experts and for the router scores do not vanish or explode as $N, M \to \infty$.

Taking a standard empirical risk minimization loss of the form:

$$\mathcal{L} = \frac{1}{P}\sum_{\mu=1}^{P} \ell(f_\mu, y_\mu) \tag{L.9}$$

The network parameters evolve according to:

$$\begin{aligned} \frac{d\theta}{dt} &= \frac{\eta}{P\gamma}\sum_{\mu} \Delta_\mu \frac{\partial h^4_\mu}{\partial \theta} \\ \Delta_\mu &= -\frac{\partial \mathcal{L}}{\partial f_\mu} \end{aligned} \tag{L.10}$$

For a Mean Squared Error (MSE) loss where $\mathcal{L} = \frac{1}{P}\sum_\nu (y_\nu - f_\nu)^2$, the error signal (or residual) simplifies to $\Delta_\nu = 2(y_\nu - f_\nu)$.

The logits update dynamically via the chain rule, driven by the Neural Tangent Kernel (NTK):

$$\frac{df_\mu(t)}{dt} = \frac{\partial f_\mu(t)}{\partial \theta}\cdot\frac{d\theta}{dt} = \frac{\partial f_\mu(t)}{\partial \theta}\cdot\frac{\eta_\theta}{P}\sum_{\alpha}^{P}\Delta_\alpha \frac{\partial f_\alpha(t)}{\partial \theta} = \frac{\eta}{P}\sum_{\alpha}\Delta_\alpha\, K^{\mathrm{NTK}}_{\mu\alpha}(t,t) \tag{L.11}$$

Where the infinite-width NTK is defined as:

$$K^{\mathrm{NTK}}_{\mu\alpha}(t,s) \equiv \frac{\partial f_\mu(t)}{\partial \theta}\cdot\frac{\partial f_\alpha(s)}{\partial \theta} \tag{L.12}$$

L.6 Decomposition of the MoE NTK

To explicitly compute the NTK, we require that the magnitude $\eta_\theta \frac{\partial f_\mu(t)}{\partial \theta_i}\cdot\frac{\partial f_\alpha(s)}{\partial \theta_i}$ remains order 1 for each parameter block θ_i. We evaluate this constraint block by block.

For the expert input layer $W^{2,\mathrm{in}}$:

$$\begin{aligned} &\eta_0 \sum_{i=1}^{\kappa N}\sum_{l,m,j=1}^{N_e}\sum_{k=1}^{N} \frac{\partial h^4}{\partial\left[h_i^{2,\mathrm{in}}\right]_l}\frac{\partial\left[h_i^{2,\mathrm{in}}\right]_l}{\partial W^2_{jk,i}}\frac{\partial h^4}{\partial\left[h_i^{2,\mathrm{in}}\right]_m}\frac{\partial\left[h_i^{2,\mathrm{in}}\right]_m}{\partial W^2_{jk,i}} = \\ &= \eta_0 N \sum_{i=1}^{\kappa N}\sum_{l,m,j=1}^{N_e}\sum_{k=1}^{N} \frac{\left[g_i^{2,\mathrm{in}}\right]_l}{\kappa\sqrt{N}}\frac{\left[g_i^{2,\mathrm{in}}\right]_m}{\kappa\sqrt{N}}\frac{\sigma(h^1_k)}{\sqrt{N}}\frac{\sigma(h^1_k)}{\sqrt{N}}\delta_{mj}\delta_{lj} \\ &= \eta_0 \frac{N_e}{\kappa}\frac{1}{\kappa N}\sum_{i=1}^{\kappa N}\left[\frac{1}{N_e}\sum_{j=1}^{N_e}\left[g_i^{2,\mathrm{in}}\right]_j\left[g_i^{2,\mathrm{in}}\right]_j\right]\left[\frac{1}{N}\sum_{k=1}^{N}\sigma(h^1_k)^2\right] \\ &= \eta_0 \frac{N_e}{\kappa}\frac{1}{\kappa N}\sum_{i=1}^{\kappa N} G_i^{2,\mathrm{in}}\Phi^1 = \eta_0\frac{N_e}{\kappa}\bar{G}^2\Phi^1 \end{aligned} \tag{L.13}$$

For the expert output layer $W^{2,\mathrm{out}}$:

$$\begin{aligned} &\eta_0 \sum_{i=1}^{\kappa N}\sum_{l,m,k=1}^{N_e}\sum_{j=1}^{N} \frac{\partial h^4}{\partial\left[h_i^{2,\mathrm{out}}\right]_l}\frac{\partial\left[h_i^{2,\mathrm{out}}\right]_l}{\partial W^{2,\mathrm{out}}_{jk}}\frac{\partial h^4}{\partial\left[h_i^{2,\mathrm{out}}\right]_m}\frac{\partial\left[h_i^{2,\mathrm{out}}\right]_m}{\partial W^{2,\mathrm{out}}_{jk}} \\ &= \eta_0 \sum_{i=1}^{\kappa N}\sum_{l,m,k=1}^{N_e}\sum_{j=1}^{N} \frac{\left[g_i^{2,\mathrm{out}}\right]_l}{\kappa N^{3/2}}\frac{\left[g_i^{2,\mathrm{out}}\right]_m}{\kappa N^{3/2}}\frac{\sqrt{\kappa N}}{\sqrt{N_e}}\sigma(\left[h_i^{2,\mathrm{in}}\right]_k)\frac{\sqrt{\kappa N}}{\sqrt{N_e}}\sigma(\left[h_i^{2,\mathrm{in}}\right]_k)\delta_{mj}\delta_{lj} \\ &= \eta_0 \frac{1}{\kappa N}\sum_{i=1}^{\kappa N} G_i^{2,\mathrm{out}}\Phi_i^{2,\mathrm{in}} \end{aligned} \tag{L.14}$$

For the shared base layer W^1:

$$\eta_0 \sum_{k=1}^{D} \sum_{j,m=1}^{N} \frac{\partial h^4}{\partial h_n^1} \frac{\partial h_n^1}{\partial W_{jk}^1} \frac{\partial h^4}{\partial h_m^1} \frac{\partial h_m^1}{\partial W_{jk}^1} = \eta_0 \sum_{k=1}^{D} \sum_{j,m,n=1}^{N} \frac{g_n^1}{\sqrt{N}} \frac{g_m^1}{\sqrt{N}} \delta_{nj} \, \delta_{mj} \, \frac{x_k}{\sqrt{D}} \frac{x_k}{\sqrt{D}} = \eta_0 \, G^1 \, \Phi^0 \tag{L.15}$$

The router weight matrix Q involves the complex Jacobian of the softmax operator. Integrating this out, we find:

$$\eta_0 \kappa \sum_{j=1}^{\kappa N} \sum_{k=1}^{N} \frac{\partial h^4}{\partial Q_{jk}} \frac{\partial h^4}{\partial Q_{jk}} = \eta_0 \kappa N \sum_{j=1}^{\kappa N} \sum_{i,i'=1}^{\kappa N} \Phi^1 \, g_i^\phi \, g_{i'}^\phi \, \phi^i \, \phi^{i'} (\delta_{ij} - \phi^j)(\delta_{i'j} - \phi^j) = \eta_0 \Phi^1 G^{1,\phi} \tag{L.16}$$

Aggregating these contributions yields the full macroscopic evolution equation:

$$\begin{aligned} \frac{df_\mu(t)}{dt} = \frac{\eta_0}{P} \sum_\nu \Delta_\nu \Bigg[& G_{\mu\nu}^1(t,t) \, \Phi_{\mu\nu}^0(t,t) + \frac{N_e}{\kappa} \frac{1}{\kappa N} \sum_{i=1}^{\kappa N} G_{\mu\nu,i}^{2,\text{in}}(t,t) \Phi_{\mu\nu}^1(t,t) \\ & + \frac{1}{\kappa N} \sum_{i=1}^{\kappa N} G_{\mu\nu,i}^{2,\text{out}}(t,t) \Phi_{\mu\nu,i}^{2,\text{in}}(t,t) \; + \Phi_{\mu\nu}^3(t,t) + \Phi_{\mu\nu}^1(t,t) G_{\mu\nu}^{1,\phi}(t,t) \Bigg] \end{aligned} \tag{L.17}$$

L.7 Evolution of weights, preactivations, and pregradients

The updates to the router weights are:

$$\begin{aligned} \frac{dQ_{jk}}{dt} &= \frac{\eta}{P} \sum_\mu \Delta_\mu \frac{\partial f_\mu}{\partial Q_{jk}} = \frac{\eta_0 \gamma \kappa \sqrt{N}}{P} \sum_\mu \Delta_\mu \frac{\partial h_\mu^4}{\partial Q_{jk}} = \frac{\eta_0 \gamma_0 \kappa N}{P} \sum_\mu \Delta_\mu \frac{\partial h_\mu^4}{\partial Q_{jk}} \\ &= \frac{\eta_0 \gamma_0}{P\sqrt{N}} \sum_\mu \Delta_\mu \left[\kappa N \sum_{a=1}^{M} g_a^\phi \, \phi_a (\delta_{aj} - \phi_j) \right] \sigma(h_k^1) \\ &= \frac{\eta_0 \gamma_0}{P\sqrt{N}} \sum_\mu \Delta_\mu g_j^{1,\phi} \sigma(h_k^1) \end{aligned} \tag{L.18}$$

We obtain thus:

$$\begin{aligned} \psi_\mu(t) &= \frac{1}{\sqrt{N}} \, Q(t) \, \sigma\big(h_\mu^1(t)\big) \\ &= \chi_\mu^\phi(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu(s) \, \Phi_{\mu\nu}^1(s,t) g_\nu^{1,\phi}(t) \\ \tilde{z}_\mu^{1,\phi} &= \tilde{\xi}_\mu^{1,\phi} + \frac{\eta_0 \gamma_0}{P} \int_0^\infty dt \sum_\nu \Delta_\nu G_{\nu\mu}^{1,\phi} \, \sigma(h_{\nu j}^1) \end{aligned} \tag{L.19}$$

The other preactivation and pregradient updates follow similarly:

$$
\begin{aligned}
\frac{dW^1(t)}{dt} &= \frac{\eta_0\gamma_0\sqrt{N}}{P}\sum_\mu \Delta_\mu(t)\,\frac{g^1_\mu(t)}{\sqrt{N}}\,\frac{1}{\sqrt{D}}\,x_\mu \\
\Rightarrow \quad h^1_\mu(t) &= \frac{1}{\sqrt{D}}\,W^1(0)\,x_\mu + \int_0^t ds\,\frac{\eta_0\gamma_0}{P}\sum_\nu \Delta_\nu(s)\,g^1_\nu(s)\,\Phi^0_{\mu\nu}(t,s) \\
\frac{dW^{2,\text{in}}_i(t)}{dt} &= \frac{\eta_0\gamma_0\sqrt{N}}{P}\sum_\mu \Delta_\mu(t)\,\frac{g^{2,\text{in}}_{\mu,i}(t)}{\kappa\sqrt{N}}\,\frac{1}{\sqrt{N}}\,\sigma\big(h^1_\mu(t)\big) \\
\Rightarrow \quad h^{2,\text{in}}_{\mu,i}(t) &= \frac{1}{\sqrt{N}}W^{2,\text{in}}_i(0)\,\sigma(h^1_\mu(t)) + \frac{\eta_0\gamma_0}{P}\int_0^t ds\,\sum_\nu \Delta_\nu(s)\,g^{2,\text{in}}_{\nu_i}(s)\,\Phi^1_{\mu\nu}(t,s) \\
\frac{dW^{2,\text{out}}_i(t)}{dt} &= \frac{\eta_0\gamma_0\sqrt{N}}{P}\sum_\mu \Delta_\mu(t)\,\frac{g^{2,\text{out}}_{\mu,i}(t)}{N^{3/2}\kappa}\,\frac{\sqrt{\kappa N}}{\sqrt{N_e}}\,\sigma\big(h^{2,\text{in}}_{\mu,i}(t)\big) \\
&= \frac{\eta_0\gamma_0}{P\sqrt{\kappa}}\frac{1}{\sqrt{NN_e}}\sum_\mu \Delta_\mu(t) g^{2,\text{out}}_i(t)\sigma(h^{2,\text{in}}_i(t)) \\
\Rightarrow \quad h^{2,\text{out}}_{\mu,i}(t) &= \frac{\sqrt{\kappa N}}{\sqrt{N_e}}\,W^{2,\text{out}}_i(0)\,\sigma\big(h^{2,\text{in}}_{\mu,i}(t)\big) + \frac{\eta_0\gamma_0}{P}\int_0^t ds\,\sum_\nu \Delta_\nu(s)\,g^{2,\text{out}}_{\nu,i}(s)\,\Phi^{2,\text{in}}_{\mu\nu,i}(t,s) \\
\frac{dw^4(t)}{dt} &= \frac{\eta_0\gamma_0\sqrt{N}}{P}\sum_\mu \Delta_\mu(t)\,\frac{1}{\sqrt{N}}\,h^3_\mu(t) = \frac{\eta_0\gamma_0}{P}\sum_\mu \Delta_\mu(t)\,h^3_\mu(t)
\end{aligned}
\tag{L.20}
$$

Using the same expressions for the evolution of the weights to derive expressions for the pregradients:

$$
\begin{aligned}
z^3_\mu(t) = g^3_\mu(t) = w^4_\mu(t) &= \xi^3_\mu(0) + \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\nu \Delta_\nu(s)\,h^3_\nu(s) \\
z^{2,\text{out}}_{\mu,i}(t) = \kappa g^{2,\text{out}}_{\mu,i}(t) &= \tilde\phi^i_\mu(t)\,g^3_\nu(t) = \frac{\eta_0\gamma_0}{P}\tilde\phi^i_\mu(t)\int_0^t ds\sum_\nu \Delta_\nu(s)\,h^3_\nu(s) \\
z^{2,\text{in}}_{\mu,i}(t) &= \sqrt{\frac{\kappa}{NN_e}}\,W^{2,\text{out}\,\top}_i(t)\,g^{2,\text{out}}_{\mu,i}(t) \\
&= \xi^{2,\text{in}}_{\mu,i}(t) + \frac{\eta_0\gamma_0}{PN_e}\int_0^t ds\sum_\nu \Delta_\nu(s)\,G^{2,\text{out}}_{\nu\mu,i}(s,t)\,\sigma\big(h^{2,\text{in}}_{\nu,i}(s)\big) \\
\tilde z^1_{\mu,i}(t) &= W^{2,\text{in}\,\top}_i(t)\,g^{2,\text{in}}_i(t) \\
&= \tilde\xi^1_{\mu,i}(t) + \frac{\eta_0\gamma_0 N_e}{\sqrt{N}P}\int_0^t ds\sum_\nu \Delta_\nu(s)\,G^{2,\text{in}}_{\nu\mu,i}(s,t)\,\sigma\big(h^1_\nu(s)\big)
\end{aligned}
\tag{L.21}
$$

Finally for the router gradient:

$$
\begin{aligned}
g^{1,\phi}_i &= \frac{1}{\kappa N}\sum_{m=1}^{\kappa N} g^\phi_m\tilde\phi^m(\kappa N\delta_{mi} - \tilde\phi^i) \\
&= \xi^{g^{1,\phi}}_i + \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\nu \Delta_\nu G^3\Phi^{2,in}_i\tilde\phi^i\tilde\phi^i - \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\nu \Delta_\nu G^3\bar\Phi^{2,in}\tilde\phi^i.
\end{aligned}
\tag{L.22}
$$

L.8 Stochastic initial fields

To isolate the purely deterministic part of the trajectory from the random initialization, we define a set of initial stochastic fields $\mathcal{F}$. These fields encapsulate the randomness of the initial weights:

$$\mathcal{F} = \{\chi^1_\mu(t), \chi^{2,\text{in}}_{\mu,i}(t), \chi^{2,\text{out}}_{\mu,i}(t), \chi^\phi_\mu(t), \tilde{\xi}^1_{\mu,i}(t), \xi^{2,\text{in}}_{\mu,i}(t), \xi^3_\mu(t), \xi^\phi_\mu(t), \xi^{g^\phi}_{\mu i}(t)\}_{i\in\{1,\dots,M\},\mu\in\{1,\dots,P\}} \tag{L.23}$$

$$\begin{aligned}
\chi^1_\mu &= \frac{1}{\sqrt{D}}\, W^1(0)\, x_\mu \\
\chi^{2,\text{in}}_{\mu,i}(t) &= \frac{1}{\sqrt{N}}\, W^{2,\text{in}}_i(0)\, \sigma\big(h^1_\mu(t)\big) \\
\chi^{2,\text{out}}_{\mu,i}(t) &= \sqrt{\frac{\kappa N}{N_e}} W^{2,\text{out}}_i(0)\, \sigma\big(h^{2,\text{in}}_{\mu,i}(t)\big) \\
\chi^\phi_\mu(t) &= \frac{1}{\sqrt{N}}\, Q(0)\, \sigma(h^1_\mu(t)) \\
\tilde{\xi}^1_{\mu,i}(t) &= \big(W^{2,\text{in}}_i(0)\big)^\top g^{2,\text{in}}_{\mu,i}(t) \\
\xi^{2,\text{in}}_{\mu,i}(t) &= \sqrt{\frac{\kappa}{NN_e}}\, \big(W^{2,\text{out}}_i(0)\big)^\top g^{2,\text{out}}_{\mu,i}(t) \\
\xi^{1,\phi}_\mu(t) &= \frac{1}{\kappa\sqrt{N}} Q(0)^\top g^{1,\phi}_\mu(t) \\
\xi^{g^\phi}_{\mu i}(t) &= \frac{1}{\sqrt{N_e N}} g^{3\top}_\mu(t)\, W^{2,\text{out}}_i(0)\, \sigma(h^{2,\text{in}}_{\mu,i}(t))
\end{aligned} \tag{L.24}$$

A critical step in rendering the MoE partition function computationally tractable is distinguishing between expert-local fields (which are specific to an expert i) and global fields (which impact the shared base layer or the final aggregated output). This distinction allows the massive partition function to factorize over the experts.

We define the router-averaged stochastic fields to capture the macroscopic effect of the local processes:

$$\bar{\chi}^3_\mu(t) = \frac{1}{\kappa N} \sum_{i=1}^{\kappa N} \tilde{\phi}^i_\mu(t) \chi^{2,\text{out}}_{\mu,i}(t) \tag{L.25}$$

$$\bar{\xi}^1_\mu(t) = \frac{1}{\kappa\sqrt{N}} \sum_{i=1}^{\kappa N} \tilde{\xi}^1_{\mu,i}(t) \tag{L.26}$$

Substituting these, the global aggregated output h^3_μ and the gradient arriving at the base layer z^1_μ can be expressed entirely in terms of order parameters that are smooth over the ensemble of experts:

$$h^3_\mu(t) = \bar{\chi}^3_\mu(t) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu(s) \bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) z^3_\nu(s) \tag{L.27}$$

$$z^1_\mu(t) = \bar{\xi}^1_\mu(t) + \xi^{1,\phi}_\mu(t) + \frac{\eta_0\gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu(s) \left[\frac{1}{\sqrt{N}} \bar{G}^{2,\text{in}}_{\mu\nu}(s,t) + G^{1,\phi}_{\mu\nu}(s,t)\right] \sigma(h^1_\nu(s)) \tag{L.28}$$

L.9 Deriving the DMFT Action

We formulate the DMFT by writing the moment-generating function for the system's trajectories and performing a disorder average over the initial weights θ_0.

$$
\begin{aligned}
Z \propto \Bigg\langle \int d\mathcal{F} \exp\Bigg(\sum_{\mu} \int_0^{\infty} dt\, i \Bigg[& \hat{\chi}^1_\mu \cdot \big(\chi^1_\mu - \frac{1}{\sqrt{D}} W^1(0) x_\mu\big) \\
& + \sum_{i=1}^{\kappa N} \hat{\chi}^{2,\text{in}}_{\mu,i} \cdot \Big(\chi^{2,\text{in}}_{\mu,i} - \frac{1}{\sqrt{N}} W^{2,\text{in}}_i(0) \sigma(h^1_\mu(t))\Big) \\
& + \hat{\bar{\chi}}^3_\mu \cdot \Big(\bar{\chi}^3_\mu - \frac{1}{\kappa N} \sum_{i=1}^{\kappa N} \tilde{\phi}^i_\mu(t) \sqrt{\frac{\kappa N}{N_e}} W^{2,\text{out}}_i(0) \sigma(h^{2,\text{in}}_{\mu,i}(t))\Big) \\
& + \hat{\bar{\xi}}^1_\mu \cdot \Big(\bar{\xi}^1_\mu - \frac{1}{\kappa\sqrt{N}} \sum_{i=1}^{\kappa N} W^{2,\text{in}}_i(0)^\top g^{2,\text{in}}_{\mu,i}(t)\Big) \\
& + \sum_{i=1}^{\kappa N} \hat{\xi}^{2,\text{in}}_{\mu,i} \cdot \Big(\xi^{2,\text{in}}_{\mu,i} - \sqrt{\frac{\kappa}{N N_e}} W^{2,\text{out}}_i(0)^\top g^{2,\text{out}}_{\mu,i}(t)\Big) + \hat{\xi}^3_\mu \cdot \Big(\xi^3_\mu - w^4(0)^\top\Big) \\
& + \hat{\chi}^\phi_\mu \cdot \Big(\chi^\phi_\mu - \frac{1}{\sqrt{N}} Q(0) \sigma(h^1_\mu(t))\Big) + \hat{\xi}^{1,\phi}_\mu \cdot \Big(\xi^{1,\phi}_\mu - \frac{1}{\kappa\sqrt{N}} Q(0)^\top g^{1,\phi}_\mu(t)\Big) \\
& + \sum_{m=1}^{\kappa N} \hat{\xi}^{g^{1,\phi}}_{\mu,m}(t) \Big(\xi^{g^{1,\phi}}_{\mu,m}(t) - \frac{1}{\kappa N} \sum_{i=1}^{\kappa N} \frac{1}{\sqrt{N_e}\, N} g^{3\,T}_\mu(t) W^{2,\text{out}}_i \sigma\big(h^{2,\text{in}}_{\mu,i}(t)\big) \tilde{\phi}^i_\mu(t) \left(\kappa N\, \delta_{im} - \tilde{\phi}^m_\mu(t)\right)\Big) \Bigg] \Bigg) \\
& \times \exp\Bigg(\sum_{\mu} \int_0^{\infty} dt \Bigg[j^1_\mu(t) \cdot \chi^1_\mu(t) + \sum_{i=1}^{M} j^{2,\text{in}}_{\mu,i}(t) \cdot \chi^{2,\text{in}}_{\mu,i}(t) + \bar{j}^3_\mu(t) \cdot \bar{\chi}^3_\mu(t) + \sum_{i=1}^{M} v^{2,\text{in}}_{\mu,i}(t) \cdot \xi^{2,\text{in}}_{\mu,i}(t) \\
& + \bar{v}^1_\mu(t) \cdot \bar{\xi}^1_\mu(t) + v^3_\mu \cdot \xi^3_\mu(t) + j^\phi_\mu(t) \cdot \chi^\phi_\mu(t) + v^{1,\phi}_\mu(t) \cdot \xi^{1,\phi}_\mu(t) + v^{g^{1,\phi}}_\mu(t) \cdot \xi^{g^{1,\phi}}_\mu(t) \Bigg] \Bigg) \Bigg\rangle_{\theta_0} .
\end{aligned}
\tag{L.29}
$$

The resulting partition function obtained after the integration over the initial weights θ_0 is:

$$
\begin{aligned}
Z \propto \int d\mathcal{F} \exp\Bigg\{ & -\frac{1}{2}\sum_{\mu\nu}\int_0^\infty \mathrm{d}t \int_0^\infty \mathrm{d}s \Bigg[\hat{\chi}^1_\mu(t)\cdot\hat{\chi}^1_\nu(s)\,\Phi^0_{\mu\nu}(t,s) + (\sum_{i=1}^{M} \hat{\chi}^{2,\text{in}}_{\mu,i}(t)\cdot\hat{\chi}^{2,\text{in}}_{\nu,i}(s) \\
& + \hat{\chi}^\phi_\mu(t)\cdot\hat{\chi}^\phi_\nu(s))\,\Phi^1_{\mu\nu}(t,s) + \frac{N_e}{\kappa}\hat{\bar{\xi}}^1_\mu(s)\cdot\hat{\bar{\xi}}^1_\nu(t)\,\bar{G}^{2,\text{in}}_{\mu\nu}(s,t) \\
& + \frac{\kappa}{N_e}\sum_{i=1}^{M}\hat{\xi}^{2,\text{in}}_{\mu,i}(s)\cdot\hat{\xi}^{2,\text{in}}_{\nu,i}(t)\,G^3_{\mu\nu}(s,t)\tilde{\phi}^i_\mu(s)\tilde{\phi}^i_\nu(t) \\
& + \hat{\xi}^3_\mu(s)\cdot\hat{\xi}^3_\nu(t) + \frac{1}{\kappa}\hat{\xi}^{1,\phi}_\mu(s)\cdot\hat{\xi}^{1,\phi}_\nu(t)G^{1,\phi}_{\mu\nu}(t,s) \\
& + \kappa\sum_{i=1}^{\kappa N}\hat{\xi}^{g^{1,\phi}}_{\mu,i}(s)\,\hat{\xi}^{g^{1,\phi}}_{\nu,i}(t)\,\Phi^{2,\text{in}}_{\mu\nu,i}(s,t)\,G^3_{\mu\nu}(s,t)\,\tilde{\phi}^i_\mu(s)\,\tilde{\phi}^i_\nu(t) \\
& + \bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t)\,\left(\hat{\bar{\chi}}^3_\mu(s)\cdot\hat{\bar{\chi}}^3_\nu(t)\right)\Bigg] - i\sum_{\mu,\nu}\int_0^\infty dt\int_0^\infty ds\,\Big[\bar{A}^1_{\mu\nu}(s,t)\frac{1}{\kappa}\sum_{i=1}^{\kappa N}\hat{\chi}^{2,\text{in}}_{\mu,i}(s)\cdot g^{2,\text{in}}_{\nu,i}(t) \\
& + \frac{1}{\kappa}\hat{\chi}^\phi_\mu(s)\cdot g^{1,\phi}_\nu(t)A^\phi_{\mu\nu}(s,t) + \frac{1}{N_e}\sum_{i=1}^{\kappa N}\left(\hat{\xi}^{2,\text{in}}_{\mu,i}(s)\;\cdot\;\sigma(h^{2,\text{in}}_{\nu,i}(t))\right)\tilde{\phi}^i_\mu(s)\tilde{\phi}^i_\nu(t)\,\bar{B}^{2,\text{in}}_{\mu\nu}(s,t) \\
& + \sum_{i=1}^{\kappa N}\hat{\xi}^{g^{1,\phi}}_{\mu,\,i}(t)\tilde{\phi}^i_\nu(s)\,\tilde{A}^{\tilde{\phi}}_{\mu\nu,\,i}(t,s)\Big] + \sum_\mu\int_0^\infty dt\,\Bigg[j^1_\mu(t)\cdot\chi^1_\mu(t) + \sum_{i=1}^{M} j^{2,\text{in}}_{\mu,i}(t)\cdot\chi^{2,\text{in}}_{\mu,i}(t) + \bar{j}^3_\mu(t)\cdot\bar{\chi}^3_\mu(t) \\
& + \sum_{i=1}^{M} v^{2,\text{in}}_{\mu,i}(t)\cdot\xi^{2,\text{in}}_{\mu,i}(t) + \bar{v}^1_\mu(t)\cdot\bar{\xi}^1_\mu(t) + v^3_\mu\cdot\xi^3_\mu(t) + j^\phi_\mu(t)\cdot\chi^\phi_\mu(t) + v^{1,\phi}_\mu(t)\cdot\xi^{1,\phi}_\mu(t)\Bigg]\Bigg\}
\end{aligned}
\tag{L.30}
$$

Where we have introduced the $O(1)$ kernels (also called response functions)

$$
\begin{aligned}
\bar{B}^{2,\text{in}}_{\mu\nu}(s,t) &= -\frac{i}{N}\hat{\bar{\chi}}^3_\mu(s)\cdot g^3_\nu(t) \\
\bar{A}^1_{\mu\nu}(s,t) &= -\frac{i}{N}\hat{\bar{\xi}}^1_\mu(s)\cdot\sigma(h^1_\nu(t)) \\
A^{1,\phi}_{\mu\nu}(s,t) &= -\frac{i}{N}\,\hat{\xi}^{1,\phi}_\mu(s)\cdot\sigma(h^1_\nu(t)) \\
\tilde{A}^{\tilde{\phi}}_{\mu\nu,\,i}(t,s) &= \kappa\Phi^{2,\text{in}}_{\mu\nu,\,i}(t,s)\,G^3_{\mu\nu}(t,s)\,\tilde{\phi}^i_\mu(t)\,A^{g^{1,\phi}}_{\mu\nu}(t,s) + \kappa A^{2,\text{in}}_{\mu\nu,\,i}(t,s)\,G^3_{\mu\nu}(t,s)\,\tilde{\phi}^i_\mu(t) \\
&\quad + \Phi^{2,\text{in}}_{\mu\nu,\,i}(t,s)\,\tilde{\phi}^i_\mu(t)\,\bar{B}^{2,\text{in}}_{\mu\nu}(t,s) \quad + \frac{1}{2}\kappa\bar{\Phi}^{2,\text{in}}_{\mu\nu}(t,s)\,G^3_{\mu\nu}(t,s)\,A^{g^{1,\phi}}_{\mu\nu}(t,s) \\
&\quad + \bar{\Phi}^{2,\text{in}}_{\mu\nu}(t,s)\,\bar{B}^{2,\text{in}}_{\mu\nu}(t,s) + \kappa\bar{A}^{2,\text{in}}_{\mu\nu}(t,s)\,G^3_{\mu\nu}(t,s)
\end{aligned}
\tag{L.31}
$$

We must constrain the solutions to be physical, so for each i,μ,ν,s,t we multiply in the following resolutions of the identity to enforce the definitions of the order parameters.

$$
1 = \int \frac{\mathrm{d}\Phi^1_{\mu\nu}(s,t)\,\mathrm{d}\hat{\Phi}^1_{\mu\nu}(s,t)}{2\pi i N^{-1}}\exp\left[\hat{\Phi}^1_{\mu\nu}(s,t)\Big(N\Phi^1_{\mu\nu}(s,t) - \sigma(h^1_\mu(s))\cdot\sigma(h^1_\nu(t))\Big)\right] \tag{L.32}
$$

$$
1 = \int \frac{\mathrm{d}\Phi^{2,\text{in}}_{\mu\nu,i}(s,t)\,\mathrm{d}\hat{\Phi}^{2,\text{in}}_{\mu\nu,i}(s,t)}{2\pi i {N_e}^{-1}}\exp\left[\hat{\Phi}^{2,\text{in}}_{\mu\nu,i}(s,t)\Big(N_e\Phi^{2,\text{in}}_{\mu\nu,i}(s,t) - \sigma(h^{2,\text{in}}_{\mu,i}(s))\cdot\sigma(h^{2,\text{in}}_{\nu,i}(t))\Big)\right] \tag{L.33}
$$

$$
1 = \int \frac{\mathrm{d}G^1_{\mu\nu}(s,t)\,\mathrm{d}\hat{G}^1_{\mu\nu}(s,t)}{2\pi i N^{-1}}\exp\left[\hat{G}^1_{\mu\nu}(s,t)\Big(NG^1_{\mu\nu}(s,t) - g^1_\mu(s)\cdot g^1_\nu(t)\Big)\right] \tag{L.34}
$$

$$1 = \int \frac{\mathrm{d}G^{2,\text{in}}_{\mu\nu,i}(s,t)\,\mathrm{d}\hat{G}^{2,\text{in}}_{\mu\nu,i}(s,t)}{2\pi i {N_e}^{-1}} \exp\Big[\hat{G}^{2,\text{in}}_{\mu\nu,i}(s,t)\Big(N_e G^{2,\text{in}}_{\mu\nu,i}(s,t) - g^2_{\mu,i}(s)\cdot g^2_{\nu,i}(t)\Big)\Big] \tag{L.35}$$

$$1 = \int \frac{\mathrm{d}G^{3}_{\mu\nu}(s,t)\,\mathrm{d}\hat{G}^{3}_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[\hat{G}^{3}_{\mu\nu}(s,t)\Big(N G^{3}_{\mu\nu}(s,t) - g^3_{\mu}(s)\cdot g^3_{\nu}(t)\Big)\Big] \tag{L.36}$$

$$1 = \int \frac{\mathrm{d}G^{1,\phi}_{\mu\nu}(s,t)\,\mathrm{d}\hat{G}^{1,\phi}_{\mu\nu}(s,t)}{2\pi i \kappa N^{-1}} \exp\Big[\hat{G}^{1,\phi}_{\mu\nu}(s,t)\Big(\kappa N G^{1,\phi}_{\mu\nu}(s,t) - g^{1,\phi}_{\mu}(s)\cdot g^{1,\phi}_{\nu}(t)\Big)\Big] \tag{L.37}$$

$$1 = \int \frac{\mathrm{d}\bar{B}^{2,\text{in}}_{\mu\nu}(s,t)\,\mathrm{d}\bar{A}^{2,\text{in}}_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[-\bar{A}^{2,\text{in}}_{\mu\nu}(s,t)\Big(N \bar{B}^{2,\text{in}}_{\mu\nu}(s,t) + i\hat{\tilde{\chi}}^3_{\mu}(s)\cdot g^3_{\nu}(t)\Big)\Big] \tag{L.38}$$

$$1 = \int \frac{\mathrm{d}\bar{A}^{1}_{\mu\nu}(s,t)\,\mathrm{d}\bar{B}^{1}_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[-\bar{B}^{1}_{\mu\nu}(s,t)\Big(N \bar{A}^{1}_{\mu\nu}(s,t) + i\hat{\tilde{\xi}}^1_{\mu}(s)\cdot \sigma(h^1_{\nu}(t))\Big)\Big] \tag{L.39}$$

$$1 = \int \frac{\mathrm{d}A^{1,\phi}_{\mu\nu}(s,t)\,\mathrm{d}B^{1,\phi}_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[-B^{1,\phi}_{\mu\nu}(s,t)\Big(N A^{1,\phi}_{\mu\nu}(s,t) + i\hat{\xi}^{1,\phi}_{\mu}(s)\cdot \sigma(h^1_{\nu}(t))\Big)\Big] \tag{L.40}$$

$$1 = \int \frac{\mathrm{d}A^{g^{1,\phi}}_{\mu\nu}(s,t)\,\mathrm{d}B^{g^{1,\phi}}_{\mu\nu}(s,t)}{2\pi i (\kappa N)^{-1}} \exp\Big[-B^{g^{1,\phi}}_{\mu\nu}(s,t)\Big(\kappa N A^{g^{1,\phi}}_{\mu\nu}(s,t) + i\hat{\xi}^{g^{1,\phi}}_{\mu}(s)\cdot \tilde{\phi}_{\nu}(t)\Big)\Big] \tag{L.41}$$

$$1 = \int \frac{\mathrm{d}A^{2,\text{in}}_{\mu\nu,i}(s,t)\,\mathrm{d}B^{2,\text{in}}_{\mu\nu,i}(s,t)}{2\pi i N_e^{-1}} \exp\Big[B^{2,\text{in}}_{\mu\nu,i}(s,t)\Big(N_e A^{2,\text{in}}_{\mu\nu,i}(s,t) + i\hat{\xi}^{2,\text{in}}_{\mu,i}(s)\cdot \sigma(h^{2,\text{in}}_{\nu,i}(t))\Big)\Big] \tag{L.42}$$

$$1 = \int \frac{d\Phi^{3}_{\mu\nu}(t,s)\,d\hat{\Phi}^{3}_{\mu\nu}(t,s)}{2\pi i\, N^{-1}} \exp\Big[\hat{\Phi}^{3}_{\mu\nu}(t,s)\,\big(N\Phi^{3}_{\mu\nu}(t,s) - h^3_{\mu}(t)\cdot h^3_{\nu}(s)\big)\Big] \tag{L.43}$$

$$1 = \int \frac{d\bar{\Phi}^{2,\text{in}}_{\mu\nu}(t,s)\,d\hat{\bar{\Phi}}^{2,\text{in}}_{\mu\nu}(t,s)}{2\pi i} \exp\left[\hat{\bar{\Phi}}^{2,\text{in}}_{\mu\nu}(t,s)\left(\kappa N \bar{\Phi}^{2,\text{in}}_{\mu\nu}(t,s) - \sum_{i=1}^{\kappa N} \tilde{\phi}^i_{\mu}(s)\tilde{\phi}^i_{\nu}(t)\Phi^{2,\text{in}}_{\mu\nu,i}(s,t)\right)\right] \tag{L.44}$$

$$1 = \int \frac{d\bar{G}^{2,\text{in}}_{\mu\nu}(t,s)\,d\hat{\bar{G}}^{2,\text{in}}_{\mu\nu}(t,s)}{2\pi i\,(\kappa N)^{-1}} \exp\left[\hat{\bar{G}}^{2,\text{in}}_{\mu\nu}(t,s)\left(\kappa N \bar{G}^{2,\text{in}}_{\mu\nu}(t,s) - \sum_{i=1}^{\kappa N} G^{2,\text{in}}_{\mu\nu,i}(s,t)\right)\right] \tag{L.45}$$

L.10 Softmax

We need to encode the behaviour of softmax in the DMFT. Recall that $\phi^i = \text{softmax}(\psi^i)$, so $\tilde{\phi}^i = \kappa N \phi^i = \kappa N\,\text{softmax}(\psi^i)$. That is,

$$\tilde{\phi}^i_{\mu}(t) = \kappa N \phi^i_{\mu}(t) = \frac{e^{\psi^i_{\mu}(t)}}{\frac{1}{\kappa N}\sum_{j=1}^{\kappa N} e^{\psi^j_{\mu}(t)}} \tag{L.46}$$

This invites the definition for each input μ of the *global* order parameter

$$\mathcal{S}_\mu(t) := \frac{1}{\kappa N} \sum_{j=1}^{\kappa N} e^{\psi_\mu^j(t)}, \tag{L.47}$$

which we enforce via the Fourier-transformed delta function

$$1 = \int \frac{d\mathcal{S}_\mu(t)\, d\hat{\mathcal{S}}_\mu(t)}{2\pi i\, (\kappa N)^{-1}} \exp\left[\hat{\mathcal{S}}_\mu(t) \left(\kappa N \mathcal{S}_\mu(t) - \sum_{i=1}^{\kappa N} e^{\psi_\mu^i(t)}\right)\right] \tag{L.48}$$

L.11 Partition function

Define the set of all kernels and conjugates (indexed for time and feature, although this is omitted for brevity in L.49):

$$\begin{aligned}\mathcal{K} = \{&\Phi^1, \hat{\Phi}^1, \Phi_i^{2,\text{in}}, \hat{\Phi}_i^{2,\text{in}}, \bar{\Phi}^{2,\text{in}}, \hat{\bar{\Phi}}^{2,\text{in}}, \Phi^3, \hat{\Phi}^3, G^1, \hat{G}^1, \bar{G}^{2,\text{in}}, \hat{\bar{G}}^{2,\text{in}}, G_i^{2,\text{in}}, \hat{G}_i^{2,\text{in}}, \\ &G^3, \hat{G}^3, G^{1,\phi}, \hat{G}^{1,\phi}, \bar{A}^1, \bar{B}^1, \bar{B}^{2,\text{in}}, \bar{A}^{2,\text{in}}, A^{1,\phi}, B^{1,\phi}, A^{g^{1,\phi}}, B^{g^{1,\phi}}, B_i^{2,\text{in}}, A_i^{2,\text{in}}\}\end{aligned} \tag{L.49}$$

We can partition these order parameters into the set $\mathcal{K}_{global}$ of global order parameters, and the set $\mathcal{K}_{exp-loc}$ of expert-local order parameters.

$$\begin{aligned}\mathcal{K}_{global} = \{&\Phi^1, \hat{\Phi}^1, \bar{\Phi}^{2,\text{in}}, \hat{\bar{\Phi}}^{2,\text{in}}, \Phi^3, \hat{\Phi}^3, G^1, \hat{G}^1, \bar{G}^{2,\text{in}}, \hat{\bar{G}}^{2,\text{in}}, G^{1,\phi}, \hat{G}^{1,\phi}, G^3, \hat{G}^3, \\ &A^1, \bar{B}^1, \bar{B}^{2,in}, \bar{A}^{2,in}, A^{1,\phi}, B^{1,\phi}, A^{g^{1,\phi}}, B^{g^{1,\phi}}\}\end{aligned} \tag{L.50}$$

$$\mathcal{K}_{exp-loc} = \{\Phi_i^{2,\text{in}}, \hat{\Phi}_i^{2,\text{in}}, \Phi_i^{2,\text{out}}, \hat{\Phi}_i^{2,\text{out}}, G_i^{2,\text{in}}, \hat{G}_i^{2,\text{in}}, B_i^{2,\text{in}}, A_i^{2,\text{in}}\} \tag{L.51}$$

$$Z \;\propto\; \int \Big(\prod_{\mu,\nu} \prod_{t,s} d\mathcal{K}_{global}\Big)\; \exp\Big(N\, S[\mathcal{K}]\Big), \quad \text{where} \tag{L.52}$$

$$\begin{aligned}S[\mathcal{K}] = \sum_{\mu,\nu} \int dt \int ds \Bigg[&\hat{\Phi}_{\mu\nu}^1(t,s)\, \Phi_{\mu\nu}^1(t,s) + \hat{G}_{\mu\nu}^1(t,s)\, G_{\mu\nu}^1(t,s) + \kappa\, \hat{G}_{\mu\nu}^{1,\phi}(t,s)\, G_{\mu\nu}^{1,\phi}(t,s) \\ &- \bar{A}_{\mu\nu}^{2,\text{in}}(t,s)\, \bar{B}_{\mu\nu}^{2,\text{in}}(t,s) - \bar{B}_{\mu\nu}^1(t,s)\, \bar{A}_{\mu\nu}^1(t,s) - B_{\mu\nu}^{1,\phi}(t,s)\, A_{\mu\nu}^{1,\phi}(t,s) - \kappa B_{\mu\nu}^{g^{1,\phi}}(t,s)\, A_{\mu\nu}^{g^{1,\phi}}(t,s) \\ &+ \hat{\Phi}_{\mu\nu}^3(t,s)\, \Phi_{\mu\nu}^3(t,s) + \kappa \hat{\bar{\Phi}}_{\mu\nu}^{2,\text{in}}(t,s) \bar{\Phi}_{\mu\nu}^{2,\text{in}}(t,s) + \kappa \hat{\bar{G}}_{\mu\nu}^{2,\text{in}}(t,s) \bar{G}_{\mu\nu}^{2,\text{in}}(t,s) \\ &+ \kappa \hat{\mathcal{S}}_\mu(t)\, \mathcal{S}_\nu(s) + \hat{G}_{\mu\nu}^3(t,s)\, G_{\mu\nu}^3(t,s)\Bigg] \\ &+ \frac{1}{N} \sum_{n=1}^{N} \ln \mathcal{Z}_N^{global}[j_n^1, \bar{j}_n^3, \bar{v}_n^1, v_n^{1,\phi}, v_n^3] + \frac{\kappa}{\kappa N} \sum_{i=1}^{\kappa N} \ln Z^{local}[j_i^{2,\text{in}}, v_i^{2,\text{in}}, j_i^{3,\text{out}}, j_i^\phi]\end{aligned} \tag{L.53}$$

$$\ln \mathcal{Z}_N^{global}[j_n^1, \bar{j}_n^3, \bar{v}_n^1, v_n^{1,\phi}, v_n^3] = -\frac{1}{2}\sum_{\mu,\nu}\int dt \int ds \Big[\hat{\chi}_\mu^1(t)\hat{\chi}_\nu^1(s)\,\Phi_{\mu\nu}^0(t,s) + \hat{\xi}_\mu^3(t)\hat{\xi}_\nu^3(s)$$
$$+ \frac{1}{\kappa}\hat{\xi}_\mu^{1,\phi}(t)\hat{\xi}_\nu^{1,\phi}(s)\,G_{\mu\nu}^{1,\phi}(t,s) + \frac{N_e}{\kappa}\hat{\bar{\xi}}_\mu^1(t)\hat{\bar{\xi}}_\nu^1(s)\bar{G}_{\mu\nu}^{2,\text{in}}(t,s) + \hat{\bar{\chi}}_\mu^3(t)\,\hat{\bar{\chi}}_\mu^3(t)\,\bar{\Phi}^{2,\text{in}}\Big]$$
$$- \sum_{\mu,\nu}\int dt \int ds \Big[\hat{\Phi}_{\mu\nu}^1(t,s)\,\sigma(h_\mu^1(t))\,\sigma(h_\nu^1(s))$$
$$+ \hat{G}_{\mu\nu}^1(t,s)\,g_\mu^1(t)\,g_\nu^1(s) + \hat{\Phi}_{\mu\nu}^3(t,s)\,h_\mu^3(t)h_\nu^3(s) + \hat{G}_{\mu\nu}^3(t,s)\,g_\mu^3(t)\,g_\nu^3(s)\Big]$$
$$- i\sum_{\mu,\nu}\int dt \int ds \Big[\, B_{\mu\nu}^{1,\phi}(t,s)\hat{\xi}_\mu^{1,\phi}(t)\sigma(h_\nu^1(s)) + \bar{B}_{\mu\nu}^1(t,s)\hat{\bar{\xi}}_\mu^1(t)\sigma(h_\nu^1(s))$$
$$+ \bar{A}_{\mu\nu}^{2,\text{in}}(t,s)\hat{\bar{\chi}}_\mu^3(t)g_\nu^3(s)\Big]$$
$$+ \sum_{\mu}\int dt\Big[(v_{\mu,n}^3 + i\hat{\xi}_{\mu,n}^3(t))\xi_{\mu,n}^3(t) + (v_{\mu,n}^{1,\phi}(t) + i\hat{\xi}_{\mu,n}^{1,\phi}(t))\xi_{\mu,n}^{1,\phi}(t) + (j_{\mu,n}^1(t)$$
$$+ i\hat{\chi}_{\mu,n}^1(t))\chi_{\mu,n}^1(t) + (\bar{j}_{\mu,n}^3(t) + i\hat{\bar{\chi}}_{\mu,n}^3(t))\bar{\chi}_{\mu,n}^3(t) + (\bar{v}_{\mu,n}^1(t) + i\hat{\bar{\xi}}_{\mu,n}^1(t))\bar{\xi}_{\mu,n}^1(t)\Big] \tag{L.54}$$

$$\mathcal{Z}^{local}[j_i^2, v_i^2, j^{3,i}, j_i^\phi] = \int \Big(\prod_{\mu\nu}\prod_{t,s} d\mathcal{K}_{exp-loc}\Big) \times \exp\left(\sum_{\mu\nu}\int_0^\infty dt \int_0^\infty ds\, \Big[N_e\,\hat{\Phi}_{\mu\nu,i}^{2,\text{in}}(t,s)\,\Phi_{\mu\nu,i}^{2,\text{in}}(t,s)\right.$$
$$\left. + N_e\,\hat{G}_{\mu\nu,i}^{2,\text{in}}(t,s)\,G_{\mu\nu,i}^{2,\text{in}}(t,s) + N_e\,B_{\mu\nu,i}^{2,\text{in}}(t,s)\,A_{\mu\nu,i}^{2,\text{in}}(t,s)\Big]\right)$$
$$\times \frac{1}{N}\prod_{j=1}^{N_e}\mathcal{Z}_{N_e}\left[\left[j_i^{2,\text{in}}\right]_j, \left[v_i^{2,\text{in}}\right]_j\right] \times \mathcal{Z}_M\left[j_i^\phi, v_i^{g^{1,\phi}}\right] \tag{L.55}$$

$$\mathcal{Z}_{N_e\, j}\left[\left[j_i^{2,\text{in}}\right]_j,\left[v_i^{2,\text{in}}\right]_j\right] := \int d\mathcal{F}\ \exp\Bigg\{ -\frac{1}{2}\sum_{\mu,\nu}\int dt\int ds\Big[\ \left[\hat{\chi}_{\mu,i}^{2,\text{in}}(t)\right]_j\ \left[\hat{\chi}_{\nu,i}^{2,\text{in}}(s)\right]_j\ \Phi_{\mu\nu}^1(t,s)$$
$$+\frac{\kappa}{N_e}\ \left[\hat{\xi}_{\mu,i}^{2,\text{in}}(t)\right]_j\ \left[\hat{\xi}_{\nu,i}^{2,\text{in}}(s)\right]_j\ G_{\mu\nu}^3(s,t)\tilde{\phi}_\mu^i(s)\tilde{\phi}_\nu^i(t)$$
$$+\frac{1}{N_e}G_{\mu\nu}^3(s,t)\sigma(\left[h_{\mu,i}^{2,\text{in}}(s)\right]_j)\sigma(\left[h_{\nu,i}^{2,\text{in}}(t)\right]_j)\hat{\xi}_{\mu,i}^{g^\phi}(t)\hat{\xi}_{\nu,i}^{g^\phi}(s)\Big]$$
$$-i\sum_{\mu,\nu}\int dt\int ds\Bigg[\frac{1}{\kappa}\ \left[\hat{\chi}_{\mu,i}^{2,\text{in}}(t)\right]_j\ \left[g_{\nu,i}^{2,\text{in}}(s)\right]_j\ \bar{A}_{\mu\nu}^1(t,s)$$
$$-B_{\mu\nu,i}^{2,\text{in}}(t,s)\ \left[\hat{\xi}_{\mu,i}^{2,\text{in}}(t)\right]\sigma(\left[h_{\nu,i}^{2,\text{in}}(s)\right])\Bigg]$$
$$-\sum_{\mu,\nu}\int dt\int ds\Big[\hat{\Phi}_{\mu\nu,i}^{2,\text{in}}(t,s)\,\sigma(\left[h_{\mu,i}^{2,\text{in}}(t)\right]_j)\,\sigma(\left[h_{\nu,i}^{2,\text{in}}(s)\right]_j)$$
$$+\hat{G}_{\mu\nu,i}^{2,\text{in}}(t,s)\ \left[g_{\mu,i}^{2,\text{in}}(t)\right]_j\ \left[g_{\nu,i}^{2,\text{in}}(s)\right]_j\Big]$$
$$+\sum_{\mu}\int dt\Big[\big(\left[j_{\mu,i}^{2,\text{in}}(t)\right]_j+i\left[\hat{\chi}_{\mu,i}^{2,\text{in}}(t)\right]_j\big)\ \left[\chi_{\mu,i}^{2,\text{in}}(t)\right]_j$$
$$+\big(\left[v_{\mu,i}^{2,\text{in}}(t)\right]_j+i\left[\hat{\xi}_{\mu,i}^{2,\text{in}}(t)\right]_j\big)\ \left[\xi_{\mu,i}^{2,\text{in}}(t)\right]_j\Big]\Bigg\}. \tag{L.56}$$

$$\mathcal{Z}_M[j_m^\phi,v_m^{g^{1,\phi}}] := \int d\mathcal{F}\ \exp\Bigg\{ -\frac{1}{2}\sum_{\mu,\nu}\int dt\int ds\Big[\hat{\chi}_{\mu m}^\phi(t)\hat{\chi}_{\nu m}^\phi(s)\,\Phi_{\mu\nu}^1(t,s)$$
$$+\kappa\hat{\xi}_{\mu,m}^{g^{1,\phi}}(s)\,\hat{\xi}_{\nu,m}^{g^{1,\phi}}(t)\,\Phi_{\mu\nu,m}^{2,\text{in}}(s,t)\,G_{\mu\nu}^3(s,t)\,\tilde{\phi}_\mu^m(s)\,\tilde{\phi}_\nu^m(t)\Big]$$
$$-i\sum_{\mu,\nu}\int dt\int ds\,\Big[\frac{1}{\kappa}\hat{\chi}_{\mu m}^\phi(t)g_\nu^{1,\phi}(s)A_{\mu\nu}^{1,\phi}(t,s)+\hat{\xi}_{\mu m}^{g^{1,\phi}}(t)\tilde{\phi}_\nu^m(s)\left[B_{\mu\nu}^{g^{1,\phi}}(t,s)+\tilde{A}_{\mu\nu,\,m}^{\tilde{\phi}}(t,s)\right]\Big]$$
$$+\sum_{\mu,\nu}\int_0^\infty dt\int_0^\infty ds\ \Big[A_{\mu\nu,m}^{2,\text{in}}(s,t)\tilde{\phi}_\mu^m(s)\tilde{\phi}_\nu^m(t)\,\bar{B}_{\mu\nu}^{2,\text{in}}(s,t)\Big]$$
$$-\sum_{\mu,\nu}\int dt\int ds\Big[\hat{G}_{\mu\nu}^{1,\phi}(t,s)\,g_{\mu m}^{1,\phi}(t)\,g_{\nu m}^{1,\phi}(s)+\hat{\bar{\Phi}}_{\mu\nu}^{2,\text{in}}(t,s)\,\tilde{\phi}_\mu^m(t)\,\tilde{\phi}_\nu^m(s)\Phi_{\mu\nu,m}^{2,\text{in}}(t,s)$$
$$+\hat{\bar{G}}_{\mu\nu}^{2,\text{in}}(t,s)\,G_{\mu\nu,m}^{2,\text{in}}(t,s)\Big]$$
$$-\sum_\mu\int dt\,\hat{\mathcal{S}}_\mu(t)\,e^{\psi_\mu^m(t)}+\sum_\mu\int dt(j_{\mu m}^\phi(t)+i\hat{\chi}_{\mu m}^\phi(t))\,\chi_{\mu m}^\phi(t)+(v_{\mu m}^{g^\phi}(t)+i\hat{\xi}_{\mu m}^{g^{1,\phi}}(t))\,\xi_{\mu m}^{g^{1,\phi}}(t)\Bigg\}. \tag{L.57}$$

L.12 Saddle point approximation

To write down the saddle point equations, we define first the single-site distributions.

We can write

$$\mathcal{Z}_N^{global}[j_n^1,\bar{j}_n^3,\bar{v}_n^1,v_n^{1,\phi},v_n^3] = \int d\mathcal{F}\ \exp\Big\{ -\mathcal{H}_N^{global}[\{\chi_{\mu n}^1,\bar{\chi}_{\mu n}^3,\bar{\xi}_{\mu n}^1,\xi_{\mu n}^{1,\phi},\xi_{\mu n}^3,j_n^1,\bar{j}_n^3,\bar{v}_n^1,v_n^{1,\phi},v_n^3\}_\mu]\Big\} \tag{L.58}$$

$$\mathcal{Z}_M[j_m^\phi] = \int d\mathcal{F}\ \exp\left\{ -\mathcal{H}_M[\{\chi_{\mu,m}^\phi, j_{\mu,m}^\phi\}_\mu] \right\} \tag{L.59}$$

Where $\mathcal{H}$ in each case is the logarithm of the integrand of the corresponding $\mathcal{Z}$. We can then define for each $\mathcal{Z} \in \{\mathcal{Z}_N^{global}, \mathcal{Z}_{N_e\, j}, \mathcal{Z}_M\}$ and the corresponding $\mathcal{H}$ the average

$$\langle \mathcal{O}(\{\chi,\xi\})\rangle_{\mathcal{Z}} = \frac{1}{\mathcal{Z}} \int \prod_\mu d\mathcal{F} \exp(-\mathcal{H}[\{\chi,\xi\},\{j,v\}])\mathcal{O}(\{\chi,\xi\}) \tag{L.60}$$

With this apparatus in place, we can take saddle equations. We treat expert-local kernels as microvariables which are implicitly defined in terms of χ, ξ, and so do not take saddle equations of them at this point.

L.13 Saddle-Point Equations

Since the integrand in eqn.L.52 has the form $e^{NS[\mathcal{K}]}$, in the $\lim_N \to \infty$ we can use the saddle point approximation, which consist in finding the set of equations that lead to a stationary action $S[\mathcal{K}]$.

We note the following:

- At zero source, all single site averages $\langle\rangle_{\mathcal{Z}_N^{global}[j_n^1, \bar{j}_n^3, \bar{v}_n^1, v_n^{1,\phi}, v_n^3]}$ are equivalent, so we may write $\langle\rangle_{\mathcal{Z}_N^{global}}$ for the average over the single-site distributions when $\boldsymbol{j}^1, \bar{\boldsymbol{j}}^3, \bar{\boldsymbol{v}}^1, \boldsymbol{v}^{1,\phi}, \boldsymbol{v}^3 \to 0$.
- *Conditional on the set* $\mathcal{K}_{global}$ of global kernels, $\mathcal{Z}^{local}$ factorises over experts, and so we can write $\langle\rangle_{\mathcal{Z}_{N_e j}|\mathcal{K}_{global}}$ for the conditional average over the distribution defined by $\mathcal{Z}_{N_e\, j}\left[\left[j_i^{2,\text{in}}\right]_j, \left[v_i^{2,\text{in}}\right]_j\right]$ as $\boldsymbol{j}^{2,\text{in}}, \boldsymbol{v}^{2,\text{in}} \to 0$, and $\langle\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}$ for the conditional average over the distribution defined by $\mathcal{Z}_M[j_m^\phi, v_m^{g^\phi}]$ as $\boldsymbol{j}^\phi, \boldsymbol{v}^{g^{1,\phi}} \to 0$.
- Expert-local variables follow single-site processes, so we can drop expert indices
- By a derivation similar to that in sec.K, we can prove that conjugate kernels defined as covariances between conjugate fields vanish, since they have no physical meaning and can not influence the dynamics in addition to imposing constraint when introducing kernel definitions.

The global kernels are given by their expectation values over the single-site global distribution $\mathcal{Z}_N^{\text{global}}$:

$$\begin{aligned}
\Phi^1_{\mu\nu}(s,t) &= \left\langle \sigma(h^1_\mu(s))\, \sigma(h^1_\nu(t)) \right\rangle_{\mathcal{Z}_N^{\text{global}}} \\
\Phi^3_{\mu\nu}(s,t) &= \left\langle h^3_\mu(s)\, h^3_\nu(t) \right\rangle_{\mathcal{Z}_N^{\text{global}}} \\
G^1_{\mu\nu}(s,t) &= \left\langle g^1_\mu(s)\, g^1_\nu(t) \right\rangle_{\mathcal{Z}_N^{\text{global}}} \\
G^3_{\mu\nu}(s,t) &= \left\langle g^3_\mu(s)\, g^3_\nu(t) \right\rangle_{\mathcal{Z}_N^{global}} \\
A^{1,\phi}_{\mu\nu}(s,t) &= -i\left\langle \hat{\xi}^{1,\phi}_\mu(s)\, \sigma(h^1_\nu(t))(t) \right\rangle_{\mathcal{Z}_N^{global}} \\
\bar{A}^1_{\mu\nu}(s,t) &= -i\left\langle \hat{\bar{\xi}}^1_\mu(t) \sigma(h^1_\nu(s)) \right\rangle_{\mathcal{Z}_N^{global}} \\
\bar{B}^{2,\text{in}}_{\mu\nu}(s,t) &= -i\left\langle \hat{\bar{\chi}}^3_\mu(t) g^3_\nu(s) \right\rangle_{\mathcal{Z}_N^{global}}
\end{aligned} \tag{L.61}$$

The router kernels require expectations over the expert ensemble distribution $\mathcal{Z}_M$, conditioned on the global fields:

$$\begin{aligned}
\bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) &= \big\langle \tilde{\phi}_\mu(s)\, \tilde{\phi}_\nu(t) \Phi^{2,\text{in}}_{\mu\nu}(s,t) \big\rangle_{\mathcal{Z}_M | \mathcal{K}_{global}} \\
\bar{G}^{2,\text{in}}_{\mu\nu}(s,t) &= \big\langle G^{2,\text{in}}_{\mu\nu,}(s,t) \big\rangle_{\mathcal{Z}_M | \mathcal{K}_{global}} \\
G^{1,\phi}_{\mu\nu}(s,t) &= \big\langle g^{1,\phi}_\mu(s)\, g^{1,\phi}_\nu(t) \big\rangle_{\mathcal{Z}_M | \mathcal{K}_{global}} \\
\mathcal{S}_\mu(t) &= \big\langle e^{\psi_\mu(t)} \big\rangle_{\mathcal{Z}_M | \mathcal{K}_{global}}
\end{aligned} \tag{L.62}$$

L.14 Expert-local order parameters

The local kernels are determined by independent processes within each expert's internal dimensions:

$$\begin{aligned}
\Phi^{2,\text{in}}_{\mu\nu}(s,t) &= \frac{1}{N_e} \sum_{j=1}^{N_e} \sigma\big(\big[h^{2,\text{in}}_\mu(t)\big]_j\big)\, \sigma\big(\big[h^{2,\text{in}}_\nu(s)\big]_j\big) \\
G^{2,\text{in}}_{\mu\nu}(s,t) &= \frac{1}{N_e} \sum_{j=1}^{N_e} \big[g^{2,\text{in}}_\mu(t)\big]_j\, \big[g^{2,\text{in}}_\nu(s)\big]_j \\
A^{2,\text{in}}_{\mu\nu}(s,t) &= i\,\kappa \sum_{j=1}^{N_e} \Big[\hat{\xi}^{2,\text{in}}_\mu(s)\Big]_j\, \sigma\big(\big[h^{2,\text{in}}_\nu\big]_j\big) \\
B^{2,\text{in}}_{\mu\nu}(s,t) &= -\frac{1}{N_e} \langle \tilde{\phi}_\mu(s) \tilde{\phi}_\nu(t) \,\rangle_{\mathcal{Z}_M | \mathcal{K}_{global}} \bar{B}^{2,\text{in}}_{\mu\nu}(s,t)
\end{aligned} \tag{L.63}$$

All non-physical conjugate kernels identically vanish at the saddle point: $\hat{\Phi}^3 = \hat{\Phi}^1 = \hat{G}^1 = \hat{\bar{\Phi}}^{2,\text{in}} = \cdots = 0$.

L.15 Hubbard-Stratonovich transformation

The Hubbard-Stratonovich trick allows us to rewrite the quadratic terms over conjugate fields as Gaussian integrals over linear conjugate fields:

$$\exp\big(-\tfrac{1}{2}\,\hat{\mathbf{x}}^\top A\,\hat{\mathbf{x}}\big) = \textstyle\int_{\mathbb{R}^d} \frac{d\mathbf{u}}{(2\pi)^{d/2}\sqrt{\det A}}\, \exp\big(-\tfrac{1}{2}\,\mathbf{u}^\top A^{-1}\mathbf{u} - i\,\mathbf{u}\cdot\hat{\mathbf{x}}\big) = \big\langle \exp(-i\,\mathbf{u}\cdot\hat{\mathbf{x}})\big\rangle_{\mathbf{u}\sim\mathcal{N}(0,A)}. \tag{L.64}$$

where $\hat{\mathbf{x}}$ is a generic conjugate fields in the partition function above. Using Stein's lemma, (integration by parts on a Gaussian variable), we can reformulate the definitions of each of the kernels A and B as response functions:

$$\begin{aligned}
A^{\ell}_{\mu\nu,i}(t,s) &= -\,i\,\big\langle \hat{\xi}^{\ell}_{\nu,i}(t)\, \sigma\big(h^{\ell-1}_{\mu,i}(s)\big)\big\rangle_{\beta^{\ell}_{\nu,i}(s)} \\
&= \big[G^{\ell}_i\big]^{-1} \Big\langle \big(\xi^{\ell}_i - B^{\ell}_i\, \sigma\big(h^{\ell-1}_i\big)\big)\, \sigma\big(h^{\ell-1}_i\big)\Big\rangle_{\beta^{\ell}_{\nu,i}(s)} \\
&= \left\langle \frac{\partial \sigma(h^{\ell-1}_{\mu,i}(t))}{\partial \beta^{\ell\top}_{\nu,i}(s)} \right\rangle_{\beta^{\ell}_{\nu,i}(s)}
\end{aligned} \tag{L.65}$$

It easier now integrate over all the $\hat{\mathbf{x}}$'s, since the argument of the exponential in $\mathcal{Z}$ has been linearised with respect to them all. Doing so yields delta functions that give us the final DMFT dynamics.

L.16 DMFT Dynamics

Piecing together the self-consistent order parameters and the exact integrations over the Hubbard-Stratonovich fields, the complete DMFT description of the infinite-width MoE is summarized below.

$$
\begin{aligned}
&\alpha^1(t) \sim \mathcal{N}(0, \Phi^0) \qquad \alpha^\phi(t) \sim \mathcal{GP}(0, \Phi^1) \quad \left[\alpha^{2,\text{in}}\right]_j \sim \mathcal{GP}(0, \Phi^1) \\
&\bar{\alpha}^3(t) \sim \mathcal{GP}(0, \bar{\Phi}^{2,\text{in}}) \quad \beta^{g^{1,\phi}} \sim \mathcal{GP}(0, \kappa \Phi^{2,\text{in}} \tilde{\phi}\tilde{\phi} G^3) \\
&\bar{\beta}^1(t) \sim \mathcal{GP}(0, \frac{N_e}{\kappa} \bar{G}^{2,\text{in}}) \quad \left[\beta^{2,\text{in}}\right]_j \sim \mathcal{GP}(0, \frac{\kappa}{N_e} G^3 \tilde{\phi}\tilde{\phi}) \\
&\beta^3(t) \sim \mathcal{N}(0, \mathbb{I}) \quad \tilde{\beta}^{1,\phi}(t) \sim \mathcal{GP}(0, \frac{1}{\kappa} G^{1,\phi})
\end{aligned}
\tag{L.66}
$$

$$
\begin{aligned}
g_\mu^{1,\phi}(t) &= \beta_\mu^{g^{1,\phi}}(t) + \frac{\eta_0 \gamma_0}{\kappa P} \int_0^t ds \sum_\nu \big[\Delta_\nu(s) G^3_{\mu\nu}(s,t) \left[\Phi^{2,in}_{\mu\nu}(s,t) \tilde{\phi}_\nu(t) - \bar{\Phi}^{2,in}_{\mu\nu}(s,t) \right] \\
&\quad + \tilde{A}^{\tilde{\phi}} \big] \tilde{\phi}_\mu(s) \\
z_\mu^1(t) &= \bar{\beta}_\mu^1(t) + \tilde{\beta}_\mu^{1,\phi}(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \left[B^{1,\phi}_{\nu\mu}(s,t) + \bar{B}^1_{\mu\nu}(s,t) \right. \\
&\quad \left. + \Delta_\nu(s) G^{1,\phi}_{\mu\nu}(s,t) \right] \sigma(h^1_\nu(s)) \\
\left[z_\mu^{2,\text{in}}(t)\right]_j &= \left[\beta_\mu^{2,\text{in}}(t)\right]_j + \frac{\eta_0 \gamma_0}{P N_e} \int_0^t ds \sum_\nu \left[\Delta_\nu(s) G^3_{\mu\nu}(s,t)\, \tilde{\phi}_\mu(s)\, \tilde{\phi}_\nu(t) + B^{2,\text{in}}_{\mu\nu}(t,s) \right] \sigma\left(\left[h_\nu^{2,\text{in}}(s)\right]_j \right) \\
z_\mu^3(t) &= \beta_\mu^3(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu(s)\, h^3_\nu(s) \\
h_\mu^1(t) &= \alpha_\mu^1(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu\, \Phi^0_{\mu\nu}\, (z^1_\mu(s)\, \dot{\sigma}(h^1_\mu(s))) \\
\left[h_\mu^{2,\text{in}}(t)\right]_j &= \alpha_\mu^{2,\text{in}}(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \left[\Delta_\nu(s) \Phi^1_{\mu\nu}(t,s) \right. \\
&\qquad \left. + \frac{1}{\kappa} \bar{A}^1_{\mu\nu}(s,t) \right] \left(\dot{\sigma}\left(\left[h_\mu^{2,\text{in}}(t)\right]_j \right) \odot \left[z_\nu^{2,\text{in}}(s)\right]_j \right) \\
h_\mu^3(t) &= \bar{\alpha}_\mu^3(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \left[\bar{A}^{2,\text{in}}_{\mu\nu}(t,s) + \Delta_\nu(s) \bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) \right] z^3_\nu(s)
\end{aligned}
\tag{L.67}
$$

M DMFT Analysis for Regime II (μP)

For μP in Regime II, an analogous derivation as in the previous section, which we omit for conciseness, yields the following set of self-consistent DMFT equations:

$$\alpha^1(t) \sim \mathcal{N}(0, \Phi^0) \qquad \alpha^\phi(t) \sim \mathcal{GP}(0, \Phi^1) \quad \left[\alpha^{2,\text{in}}\right]_j \sim \mathcal{GP}(0, \Phi^1)$$

$$\bar{\beta}^1(t) \sim \mathcal{GP}(0, \frac{N_e}{\kappa}\bar{G}^{2,\text{in}}) \quad \beta^3(t) \sim \mathcal{N}(0, \mathbb{I}) \quad \tilde{\beta}^{1,\phi}(t) \sim \mathcal{GP}(0, \frac{1}{\kappa}G^{1,\phi})$$

$$\Phi^0_{\mu\nu} = \frac{1}{D}x_\mu x_\nu^\top, \Phi^1_{\mu\nu}(t,s) = \langle \sigma(h^1_\mu(t))\sigma(h^1_\nu(s))\rangle_{\mathcal{Z}_N^{global}},$$

$$\Phi^{2,\text{in}}_{\mu\nu}(s,t) = \frac{1}{N_e}\sum_{j=1}^{N_e}\left\langle \sigma(\left[h^{2,\text{in}}_\mu(t)\right]_j)\,\sigma(\left[h^{2,\text{in}}_\nu(s)\right]_j)\right\rangle_{\mathcal{Z}_{N_e\,j}|\mathcal{K}_{global}}$$

$$\Phi^3_{\mu\nu}(s,t) = \left\langle h^3_\mu(t)h^3_\nu(s)\right\rangle_{\mathcal{Z}_N^{global}}, \quad G^1_{\mu\nu}(t,s) = \left\langle[\dot\sigma(h^1_\mu(t)) \odot z^1_\mu(t)][\dot\sigma(h^1_\nu(s)) \odot z^1_\nu(s)]\right\rangle_{\mathcal{Z}_N^{global}}$$

$$\bar\Phi^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle \tilde\phi_\mu(s)\,\tilde\phi_\nu(t)\Phi^{2,\text{in}}_{\mu\nu}(s,t)\right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}} \qquad \bar G^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle G^{2,\text{in}}_{\mu\nu}(s,t)\right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}$$

$$G^{2,\text{in}}_{\mu\nu}(s,t) = \frac{1}{N_e}\sum_{j=1}^{N_e}\left\langle \left[\dot\sigma(\left[h^{2,\text{in}}_\mu(t)\right]_j) \odot \left[z^{2,\text{in}}_\mu(t)\right]_j\right] \left[\dot\sigma(\left[h^{2,\text{in}}_\nu(s)\right]_j) \odot \left[z^{2,\text{in}}_\nu(s)\right]_j\right] \right\rangle_{\mathcal{Z}_{N_e\,j}|\mathcal{K}_{global}}$$

$$G^{1,\phi}_{\mu\nu}(s,t) = \left\langle g^{1,\phi}_\mu(s)\, g^{1,\phi}_\nu(t)\right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}, \quad G^3_{\mu\nu}(s,t) = \left\langle z^3_\mu(s)\, z^3_\nu(t)\right\rangle_{\mathcal{Z}_N^{global}}$$

$$\bar A^1_{\mu\nu}(t,s) = \left\langle \frac{\partial\sigma(h^1_\mu(t))}{\partial\bar\beta^{1\top}_\nu(s)}\right\rangle_{\mathcal{Z}_N^{global}}, \quad \bar B^1_{\mu\nu}(t,s) = \sum_{j=1}^{N_e}\left\langle \frac{\partial\left[g^{2,\text{in}}_\mu(t))\right]_j}{\partial\left[\alpha^{2,\text{in}\top}_\nu(s)\right]_j}\right\rangle_{\mathcal{Z}_{N_e\,j}|\mathcal{K}_{global}}$$

$$A^{1,\phi}_{\mu\nu}(t,s) = \left\langle \frac{\partial\sigma(h^1_\mu(t))}{\partial\tilde\beta^{1,\phi\top}_\nu(s)}\right\rangle_{\mathcal{Z}_N^{global}}, \quad B^{1,\phi}_{\mu\nu}(t,s) = \left\langle \frac{\partial g^{1,\phi}_\mu(t))}{\partial\alpha^{\phi\top}_\nu(s)}\right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}$$

$$g^{1,\phi}_\mu(t) = \frac{\eta_0\gamma_0}{\kappa P}\int_0^t ds\sum_\nu \Delta_\nu(s)G^3_{\mu\nu}(s,t)\left[\Phi^{2,in}_{\mu\nu}(s,t)\tilde\phi_\nu(t) - \bar\Phi^{2,in}_{\mu\nu}(s,t)\right]\tilde\phi_\mu(s)$$

$$z^1_\mu(t) = \bar\beta^1_\mu(t) + \tilde\beta^{1,\phi}_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\nu\left\{B^{1,\phi}_{\nu\mu}(s,t) + \bar B^1_{\mu\nu}(s,t) + \Delta_\nu(s)\left[\bar G^{2,\text{in}}_{\mu\nu}(s,t) + G^{1,\phi}_{\mu\nu}(s,t)\right]\right\}\sigma(h^1_\nu(s))$$

$$\left[z^{2,\text{in}}_\mu(t)\right]_j = \frac{\eta_0\gamma_0}{P\sqrt{N_e}}\int_0^t ds\sum_\nu \Delta_\nu(s)G^3_{\mu\nu}(s,t)\,\tilde\phi_\mu(s)\,\tilde\phi_\nu(t)\,\sigma\left(\left[h^{2,\text{in}}_\nu(s)\right]_j\right), \quad j\in\{1,2,...,N_e\}$$

$$z^3_\mu(t) = \beta^3_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\nu \Delta_\nu(s)\,h^3_\nu(s)$$

$$h^1_\mu(t) = \alpha^1_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\nu \Delta_\nu\,\Phi^0_{\mu\nu}\,(z^1_\mu(s)\,\dot\sigma(h^1_\mu(s)))$$

$$\left[h^{2,\text{in}}_\mu(t)\right]_j = \alpha^{2,\text{in}}_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\nu\left[\Delta_\nu(s)\Phi^1_{\mu\nu}(t,s) + \frac{1}{\kappa}\bar A^1_{\mu\nu}(s,t)\right]\left(\dot\sigma\left(\left[h^{2,\text{in}}_\mu(t)\right]_j\right) \odot \left[z^{2,\text{in}}_\nu(s)\right]_j\right), \quad \forall j$$

$$h^3_\mu(t) = \frac{\eta_0\gamma_0}{\kappa P}\int_0^t ds\sum_\nu \Delta_\nu(s)\bar\Phi^{2,\text{in}}_{\mu\nu}(s,t)z^3_\nu(s)$$

$$\tilde\phi_\mu(t) = \frac{e^{\hat\psi_\mu(t)}}{\mathcal{S}_\mu(t)}, \qquad \mathcal{S}_\mu(t) = \left\langle e^{\hat\psi_\mu(t)}\right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}, \qquad \hat\psi_\mu(t) = m_\mu(t)\psi_\mu(t),$$

$$m_\mu(t) = \mathbf{1}(\psi_\mu(t) - \tau_\mu(t)) \qquad \text{where } \tau_\mu(t) \text{ is chosen such that } \rho = \langle m_\mu(t)\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}$$

$$\frac{df_\mu(t)}{dt} = \frac{\eta_0}{P}\sum_\nu\Delta_\nu(t)\Bigg[G^1_{\mu\nu}(t,t)\,\Phi^0_{\mu\nu}(t,t) + \left[\frac{N_e}{\kappa}\bar G^{2,\text{in}}_{\mu\nu}(t,t) + G^{1,\phi}_{\mu\nu}(t,t)\right]\Phi^1_{\mu\nu}(t,t)$$
$$+ \kappa G^3_{\mu\nu}(t,t)\bar\Phi^{2,\text{in}}_{\mu\nu}(t,t) + \Phi^3_{\mu\nu}(t,t)\Bigg]$$

N DMFT Analysis for Regime III (MSSP)

N.1 Architectural Definitions

In this section we are considering the case in which $M, N, N_e \to \infty$ at a fixed ratio. Specifically, we define $\kappa = \frac{M}{N}, \iota = \frac{N_e}{N}$, where κ,ι are order one in N. The forward pass, following ***MSSP***, is defined as:

$$\begin{aligned}
\mathbb{R}^N \ni h^1_\mu &= \frac{1}{\sqrt{D}} W^1 x_\mu & \mathbb{R}^M \ni \psi_\mu &= \frac{1}{\sqrt{N}} Q\sigma(h^1_\mu) \\
\mathbb{R}^{N_e} \ni h^{2,\text{in}}_{\mu,i} &= \frac{1}{\sqrt{N}} W^{2,\text{i}}_i \sigma(h^1_\mu) & \mathbb{R}^M \ni \phi_\mu &= \text{softmax}(\psi_\mu) \\
\mathbb{R}^N \ni h^{2,\text{out}}_{\mu,i} &= \frac{1}{\sqrt{N}} W^{2,\text{out}}_i \sigma(h^{2,\text{in}}_{\mu,i}) & \mathbb{R}^N \ni h^3_\mu &= \sum_{i=1}^{M} \phi^i_\mu \, h^{2,\text{o}}_{\mu,i} \\
\mathbb{R} \ni h^4_\mu &= \frac{1}{\sqrt{N}} w^{4\intercal} h^3_\mu & f_\mu &= \frac{1}{\gamma} h^4_\mu
\end{aligned} \tag{N.1}$$

N.2 Learning Rates and Initialization

To ensure that the network evolves dynamically in the infinite-width limit without gradients vanishing or exploding, we apply specific learning rate scalings:

$$\eta = \eta_0 \gamma^2 \qquad \eta_Q = \eta_0 \gamma^2 \kappa \qquad \eta_E = \eta_0 \gamma^2 N \qquad \gamma = \gamma_0 \sqrt{N} \qquad \eta_0, \gamma_0 \sim O(1) \tag{N.2}$$

The network weights are initialized from standard Gaussian distributions:

$$w^4_\alpha(0), \left[W^{2,\text{out}}_i(0)\right]_{\alpha\beta}, \left[W^{2,\text{in}}_i(0)\right]_{\alpha\beta}, W^1_{\alpha\beta}(0), Q_{\alpha\beta}(0) \sim \mathcal{N}(0,1) \tag{N.3}$$

but noticing that for each expert's matrices $W^{2,\text{in}}_i, W^{2,\text{out}}_i$ we share the same initial weights $W^{2,\text{in}}_i(0) = W^{2,\text{in}}(0), W^{2,\text{out}}_i(0) = W^{2,\text{out}}(0)$ For brevity in our derivations, we denote the collection of all network parameters as $\boldsymbol{\theta} = \text{Vec}\{W^1, W^{2,\text{in}}_i, W^{2,\text{out}}_i, w^4, Q\}$.

Due to the normalization constraint of the softmax operator over M experts, the routing probabilities scale as $\phi^i \sim O(\frac{1}{\kappa N})$. It is mathematically tidier to formulate the DMFT using variables which remain $O(1)$ in the limit. We therefore define and track the rescaled routing variables:

$$\tilde{\phi}^i_\mu := \kappa N \phi^i_\mu \tag{N.4}$$

N.3 Gradient Definitions

We mathematically *define* the pre-activation gradients, ensuring they contain the correct scaling factors to remain finite in the $N \to \infty$ limit:

$$
\begin{aligned}
g^1_\mu &:= \sqrt{N}\frac{\partial h^4_\mu}{\partial h^1_\mu} & g^{2,\text{in}}_{\mu,i} &:= \kappa N^{\frac{3}{2}}\frac{\partial h^4_\mu}{\partial h^{2,\text{in}}_{\mu,i}} \\
g^{2,\text{out}}_{\mu,i} &:= \kappa N^{\frac{3}{2}}\frac{\partial h^4_\mu}{\partial h^{2,\text{o}}_{\mu,i}} & g^3_\mu &:= \sqrt{N}\frac{\partial h^4_\mu}{\partial h^3_\mu} = w^4 = z^3 \\
g^\phi_\mu &:= \frac{1}{\sqrt{N}}\frac{\partial h^4_\mu}{\partial \phi_\mu} & z^{2,\text{out}}_{\mu,i} &:= \kappa N \phi^i_\mu g^3_\mu = \tilde{\phi}^i_\mu g^3_\mu \\
z^{2,\text{in}}_{\mu,i} &:= \frac{1}{\sqrt{N}} W^{2,\text{out}\,\top}_i g^{2,\text{out}}_{\mu,i} & \tilde{z}^\phi_\mu &:= \frac{1}{\kappa\sqrt{N}} Q^\top g^{1,\phi} \\
\tilde{z}^1_{\mu,i} &:= \frac{1}{\kappa\sqrt{N}} W^{2,\text{in}\,\top}_i g^{2,\text{in}}_{\mu,i} & z^1 &:= \frac{1}{\kappa\sqrt{N}} Q^\top g^{1,\phi} + \frac{1}{\kappa\sqrt{N}}\sum_{i=1}^M W^{2,\text{in}\,\top}_i g^{2,\text{in}}_{\mu,i} \\
& & &= \frac{1}{\kappa\sqrt{N}}\sum_{i=1}^M \tilde{z}^1_{\mu,i} + \tilde{z}^{1,\phi}_\mu
\end{aligned}
\tag{N.5}
$$

Note that the components of the router gradient g^ϕ_μ naturally scale as $O(1)$:

$$
(g^\phi_\mu)_i = \frac{1}{\sqrt{N}}\frac{\partial h^4_\mu}{\partial \phi^i_\mu} = \frac{1}{\sqrt{N}}\frac{\partial h^4_\mu}{\partial \boldsymbol{h}^3_\mu}\cdot\frac{\partial \boldsymbol{h}^3_\mu}{\partial \phi^i_\mu} = \frac{1}{N}\boldsymbol{g}^3\cdot\boldsymbol{h}^{2,\text{out}}_{\mu,i} = O(1). \tag{N.6}
$$

The backward pass through the router dictates that the gradient at layer 1 involves the quantity $g^{1,\phi}$, defined component-wise to handle the Jacobian of the softmax:

$$
g^{1,\phi}_k := \kappa N \sum_{i=1}^{\kappa N} g^\phi_i \phi^i(\delta_{ik} - \phi^k) = \sum_{i=1}^{\kappa N} g^\phi_i \tilde{\phi}^i(\delta_{ik} - \frac{\tilde{\phi}^k}{\kappa N}) \tag{N.7}
$$

N.4 DMFT Kernels and Order Parameters

We define the following macroscopic *kernels*, all of which are $O_N(1)$ and will serve as the fundamental order parameters in the dynamics:

$$
\begin{aligned}
\Phi^0_{\mu\nu} &= \frac{1}{D}\, x_\mu \cdot x_\nu & \Phi^1_{\mu\nu}(s,t) &= \frac{1}{N}\,\sigma\big(h^1_\mu(s)\big)\cdot\sigma\big(h^1_\nu(t)\big) \\
\Phi^{2,\text{in}}_{\mu\nu,i}(s,t) &= \frac{1}{\iota N}\,\sigma\big(h^{2,\text{in}}_{\mu,i}(s)\big)\cdot\sigma\big(h^{2,\text{in}}_{\nu,i}(t)\big) & \Phi^{2,\text{o}}_{\mu\nu,i}(s,t) &= \frac{1}{N}\,\sigma\big(h^{2,\text{out}}_{\mu,i}(s)\big)\cdot\sigma\big(h^{2,\text{out}}_{\nu,i}(t)\big) \\
\Phi^3_{\mu\nu}(s,t) &= \frac{1}{N}\, h^3_\mu(s)\cdot h^3_\nu(t) & G^1_{\mu\nu}(s,t) &= \frac{1}{N}\, g^1_\mu(s)\cdot g^1_\nu(t) \\
G^{2,\text{in}}_{\mu\nu,i}(s,t) &= \frac{1}{\iota N}\, g^{2,\text{in}}_{\mu,i}(s)\cdot g^{2,\text{in}}_{\nu,i}(t) & G^{2,\text{out}}_{\mu\nu,i}(s,t) &= \frac{1}{N}\, g^{2,\text{out}}_{\mu,i}(s)\cdot g^{2,\text{out}}_{\nu,i}(t) \\
G^{1,\phi}_{\mu\nu}(s,t) &= \frac{1}{\kappa N}\, g^{1,\phi}_\mu(s)\cdot g^{1,\phi}_\nu(t) & G^3_{\mu\nu}(s,t) &= \frac{1}{N}\, g^3_\mu(s)\cdot g^3_\nu(t)
\end{aligned}
\tag{N.8}
$$

With the exception of the input data Gram matrix Φ^0, all of these kernels are dynamically evolving macroscopic variables.

N.5 Learning Dynamics and the Neural Tangent Kernel

We train the network using full-batch continuous-time gradient flow on the P training samples, and take $N, N_e, M \to \infty$ with P and the training time t held fixed. The variable learning rate scheme defined previously is necessary so that the activation updates within the experts and for the router scores do not vanish or explode as $N, M \to \infty$.

Taking a standard empirical risk minimization loss of the form:

$$\mathcal{L} = \frac{1}{P}\sum_{\mu=1}^{P} \ell(f_\mu, y_\mu) \tag{N.9}$$

The network parameters evolve according to:

$$\begin{aligned} \frac{d\theta}{dt} &= \frac{\eta}{P\gamma}\sum_{\mu} \Delta_\mu \frac{\partial h^4_\mu}{\partial \theta} \\ \Delta_\mu &= -\frac{\partial \mathcal{L}}{\partial f_\mu} \end{aligned} \tag{N.10}$$

For a Mean Squared Error (MSE) loss where $\mathcal{L} = \frac{1}{P}\sum_\nu (y_\nu - f_\nu)^2$, the error signal (or residual) simplifies to $\Delta_\nu = 2(y_\nu - f_\nu)$.

The logits update dynamically via the chain rule, driven by the Neural Tangent Kernel (NTK):

$$\frac{df_\mu(t)}{dt} = \frac{\partial f_\mu(t)}{\partial \theta}\cdot\frac{d\theta}{dt} = \frac{\partial f_\mu(t)}{\partial \theta}\cdot\frac{\eta_\theta}{P}\sum_{\alpha}^{P}\Delta_\alpha \frac{\partial f_\alpha(t)}{\partial \theta} = \frac{\eta}{P}\sum_\alpha \Delta_\alpha\, K^{\mathrm{NTK}}_{\mu\alpha}(t,t) \tag{N.11}$$

Where the infinite-width NTK is defined as:

$$K^{\mathrm{NTK}}_{\mu\alpha}(t,s) \equiv \frac{\partial f_\mu(t)}{\partial \theta}\cdot\frac{\partial f_\alpha(s)}{\partial \theta} \tag{N.12}$$

N.6 Decomposition of the MoE NTK

To explicitly compute the NTK, we require that the magnitude $\eta_\theta \frac{\partial f_\mu(t)}{\partial \theta_i}\cdot\frac{\partial f_\alpha(s)}{\partial \theta_i}$ remains order 1 for each parameter block θ_i. We evaluate this constraint block by block.

For the expert input layer $W^{2,\text{in}}$:

$$\begin{aligned} &\eta_0 \sum_{i=1}^{\kappa N}\sum_{l,m,j=1}^{N_e}\sum_{k=1}^{N} \frac{\partial h^4}{\partial\left[h_i^{2,\text{in}}\right]_l}\frac{\partial\left[h_i^{2,\text{in}}\right]_l}{\partial W^2_{jk,i}}\frac{\partial h^4}{\partial\left[h_i^{2,\text{in}}\right]_m}\frac{\partial\left[h_i^{2,\text{in}}\right]_m}{\partial W^2_{jk,i}} = \\ &= \eta_0 N \sum_{i=1}^{\kappa N}\sum_{l,m,j=1}^{N_e}\sum_{k=1}^{N} \frac{\left[g_i^{2,\text{in}}\right]_l}{\kappa N^{3/2}}\frac{\left[g_i^{2,\text{in}}\right]_m}{\kappa N^{3/2}}\frac{\sigma(h^1_k)}{\sqrt{N}}\frac{\sigma(h^1_k)}{\sqrt{N}}\delta_{mj}\delta_{lj} \\ &= \eta_0 \frac{\iota}{\kappa}\frac{1}{\kappa N}\sum_{i=1}^{\kappa N}\left[\frac{1}{N_e}\sum_{j=1}^{N_e}\left[g_i^{2,\text{in}}\right]_j\left[g_i^{2,\text{in}}\right]_j\right]\left[\frac{1}{N}\sum_{k=1}^{N}\sigma(h^1_k)^2\right] \\ &= \eta_0 \frac{\iota}{\kappa}\frac{1}{\kappa N}\sum_{i=1}^{\kappa N} G_i^{2,\text{in}}\Phi^1 = \eta_0\frac{\iota}{\kappa}\bar{G}^2\Phi^1 \end{aligned} \tag{N.13}$$

For the expert output layer $W^{2,\text{out}}$:

$$\begin{aligned} &\eta_0 \sum_{i=1}^{\kappa N}\sum_{l,m,k=1}^{N_e}\sum_{j=1}^{N} \frac{\partial h^4}{\partial\left[h_i^{2,\text{out}}\right]_l}\frac{\partial\left[h_i^{2,\text{out}}\right]_l}{\partial W^{2,\text{out}}_{jk}}\frac{\partial h^4}{\partial\left[h_i^{2,\text{out}}\right]_m}\frac{\partial\left[h_i^{2,\text{out}}\right]_m}{\partial W^{2,\text{out}}_{jk}} \\ &= \eta_0 \sum_{i=1}^{\kappa N}\sum_{l,m,k=1}^{N_e}\sum_{j=1}^{N} \frac{\left[g_i^{2,\text{out}}\right]_l}{\kappa N^{3/2}}\frac{\left[g_i^{2,\text{out}}\right]_m}{\kappa N^{3/2}}\frac{\sqrt{\kappa N}}{\sqrt{N_e}}\sigma(\left[h_i^{2,\text{in}}\right]_k)\frac{\sqrt{\kappa N}}{\sqrt{N_e}}\sigma(\left[h_i^{2,\text{in}}\right]_k)\delta_{mj}\delta_{lj} \\ &= \eta_0 \frac{1}{\kappa N}\sum_{i=1}^{\kappa N} G_i^{2,\text{out}}\Phi_i^{2,\text{in}} \end{aligned} \tag{N.14}$$

For the shared base layer W^1:

$$\eta_0 \sum_{k=1}^{D} \sum_{j,m=1}^{N} \frac{\partial h^4}{\partial h_n^1} \frac{\partial h_n^1}{\partial W_{jk}^1} \frac{\partial h^4}{\partial h_m^1} \frac{\partial h_m^1}{\partial W_{jk}^1} = \eta_0 \sum_{k=1}^{D} \sum_{j,m,n=1}^{N} \frac{g_n^1}{\sqrt{N}} \frac{g_m^1}{\sqrt{N}} \delta_{nj} \, \delta_{mj} \, \frac{x_k}{\sqrt{D}} \frac{x_k}{\sqrt{D}} = \eta_0 \, G^1 \, \Phi^0 \tag{N.15}$$

The router weight matrix Q involves the complex Jacobian of the softmax operator. Integrating this out, we find:

$$\eta_0 \kappa \sum_{j=1}^{\kappa N} \sum_{k=1}^{N} \frac{\partial h^4}{\partial Q_{jk}} \frac{\partial h^4}{\partial Q_{jk}} = \eta_0 \kappa N \sum_{j=1}^{\kappa N} \sum_{i,i'=1}^{\kappa N} \Phi^1 \, g_i^\phi \, g_{i'}^\phi \, \phi^i \, \phi^{i'} (\delta_{ij} - \phi^j)(\delta_{i'j} - \phi^j) = \eta_0 \Phi^1 G^{1,\phi} \tag{N.16}$$

Aggregating these contributions yields the full macroscopic evolution equation:

$$\begin{aligned} \frac{df_\mu(t)}{dt} = \frac{\eta_0}{P} \sum_\nu \Delta_\nu \Bigg[& G_{\mu\nu}^1(t,t) \, \Phi_{\mu\nu}^0(t,t) + \frac{\iota}{\kappa} \Big[G_{\mu\nu,i}^{2,\text{in}} + G_{\mu\nu}^{1,\phi}(t,t) \Big](t,t) \Phi_{\mu\nu}^1(t,t) \\ & + \frac{\iota}{\kappa} G_{\mu\nu}^3(t,t) \bar{\Phi}_{\mu\nu}^{2,\text{in}}(t,t) + \Phi_{\mu\nu}^3(t,t) \Bigg] \end{aligned} \tag{N.17}$$

N.7 Evolution of weights, preactivations, and pregradients

The updates to the router weights are:

$$\begin{aligned} \frac{dQ_{jk}}{dt} &= \frac{\eta}{P} \sum_\mu \Delta_\mu \frac{\partial f_\mu}{\partial Q_{jk}} = \frac{\eta_0 \gamma \kappa \sqrt{N}}{P} \sum_\mu \Delta_\mu \frac{\partial h_\mu^4}{\partial Q_{jk}} = \frac{\eta_0 \gamma_0 \kappa N}{P} \sum_\mu \Delta_\mu \frac{\partial h_\mu^4}{\partial Q_{jk}} \\ &= \frac{\eta_0 \gamma_0}{P\sqrt{N}} \sum_\mu \Delta_\mu \left[\kappa N \sum_{a=1}^{M} g_a^\phi \, \phi_a (\delta_{aj} - \phi_j) \right] \sigma(h_k^1) \\ &= \frac{\eta_0 \gamma_0}{P\sqrt{N}} \sum_\mu \Delta_\mu g_j^{1,\phi} \sigma(h_k^1) \end{aligned} \tag{N.18}$$

We obtain thus:

$$\begin{aligned} \psi_\mu(t) &= \frac{1}{\sqrt{N}} \, Q(t) \, \sigma\big(h_\mu^1(t)\big) \\ &= \chi_\mu^\phi(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu(s) \, \Phi_{\mu\nu}^1(s,t) g_\nu^{1,\phi}(t) \\ \tilde{z}_\mu^{1,\phi} &= \tilde{\xi}_\mu^{1,\phi} + \frac{\eta_0 \gamma_0}{P} \int_0^\infty dt \sum_\nu \Delta_\nu G_{\nu\mu}^{1,\phi} \, \sigma(h_{\nu j}^1) \end{aligned} \tag{N.19}$$

The other preactivation and pregradient updates follow similarly:

$$
\begin{aligned}
\frac{dW^1(t)}{dt} &= \frac{\eta_0\gamma_0\sqrt{N}}{P}\sum_\mu \Delta_\mu(t)\,\frac{g^1_\mu(t)}{\sqrt{N}}\,\frac{1}{\sqrt{D}}\,x_\mu \\
\Rightarrow \quad h^1_\mu(t) &= \frac{1}{\sqrt{D}}\,W^1(0)\,x_\mu + \int_0^t ds\,\frac{\eta_0\gamma_0}{P}\sum_\nu \Delta_\nu(s)\,g^1_\nu(s)\,\Phi^0_{\mu\nu}(t,s) \\
\frac{dW^{2,\text{in}}_i(t)}{dt} &= \frac{\eta_0\gamma_0\sqrt{N}N}{P}\sum_\mu \Delta_\mu(t)\,\frac{g^{2,\text{in}}_{\mu,i}(t)}{\kappa\sqrt{N}}\,\frac{1}{N^{3/2}}\,\sigma\big(h^1_\mu(t)\big) \\
\Rightarrow \quad h^{2,\text{in}}_{\mu,i}(t) &= \frac{1}{\sqrt{N}}W^{2,\text{in}}_i(0)\,\sigma(h^1_\mu(t)) + \frac{\eta_0\gamma_0}{\kappa P}\int_0^t ds\,\sum_\nu \Delta_\nu(s)\,g^{2,\text{in}}_{\nu_i}(s)\,\Phi^1_{\mu\nu}(t,s) \\
\frac{dW^{2,\text{out}}_i(t)}{dt} &= \frac{\eta_0\gamma_0\sqrt{N}N}{P}\sum_\mu \Delta_\mu(t)\,\frac{g^{2,\text{out}}_{\mu,i}(t)}{N^{3/2}\kappa}\,\frac{1}{\sqrt{N}}\,\sigma\big(h^{2,\text{in}}_{\mu,i}(t)\big) \\
\Rightarrow \quad h^{2,\text{out}}_{\mu,i}(t) &= \frac{1}{\sqrt{N}}\,W^{2,\text{out}}_i(0)\,\sigma\big(h^{2,\text{in}}_{\mu,i}(t)\big) + \frac{\eta_0\gamma_0\iota}{P\kappa}\int_0^t ds\,\sum_\nu \Delta_\nu(s)\,g^{2,\text{out}}_{\nu,i}(s)\,\Phi^{2,\text{in}}_{\mu\nu,i}(t,s) \\
\frac{dw^4(t)}{dt} &= \frac{\eta_0\gamma_0\sqrt{N}}{P}\sum_\mu \Delta_\mu(t)\,\frac{1}{\sqrt{N}}\,h^3_\mu(t) = \frac{\eta_0\gamma_0}{P}\sum_\mu \Delta_\mu(t)\,h^3_\mu(t)
\end{aligned}
\tag{N.20}
$$

Using the same expressions for the evolution of the weights to derive expressions for the pregradients:

$$
\begin{aligned}
z^3_\mu(t) &= g^3_\mu(t) = w^4_\mu(t) = \xi^3_\mu(0) + \frac{\eta_0\gamma_0}{P}\int_0^t ds\sum_\nu \Delta_\nu(s)\,h^3_\nu(s) \\
z^{2,\text{out}}_{\mu,i}(t) &= \kappa g^{2,\text{out}}_{\mu,i}(t) = \tilde\phi^i_\mu(t)\,g^3_\nu(t) = \frac{\eta_0\gamma_0}{P}\tilde\phi^i_\mu(t)\int_0^t ds\sum_\nu \Delta_\nu(s)\,h^3_\nu(s) \\
z^{2,\text{in}}_{\mu,i}(t) &= \frac{1}{\sqrt{N}}\,W^{2,\text{out}\,\top}_i(t)\,g^{2,\text{out}}_{\mu,i}(t) \\
&= \xi^{2,\text{in}}_{\mu,i}(t) + \frac{\eta_0\gamma_0}{P\kappa}\int_0^t ds\sum_\nu \Delta_\nu(s)\,G^{2,\text{out}}_{\nu\mu,i}(s,t)\,\sigma\big(h^{2,\text{in}}_{\nu,i}(s)\big) \\
\tilde z^1_{\mu,i}(t) &= W^{2,\text{in}\,\top}_i(t)\,g^{2,\text{in}}_i(t) \\
&= \tilde\xi^1_{\mu,i}(t) + \frac{\eta_0\gamma_0\iota}{P\kappa}\int_0^t ds\sum_\nu \Delta_\nu(s)\,G^{2,\text{in}}_{\nu\mu,i}(s,t)\,\sigma\big(h^1_\nu(s)\big)
\end{aligned}
\tag{N.21}
$$

Finally for the router gradient:

$$
\begin{aligned}
g^{1,\phi}_i &= \frac{1}{\kappa N}\sum_{m=1}^{\kappa N} g^\phi_m\tilde\phi^m(\kappa N\delta_{mi} - \tilde\phi^i) \\
&= \xi^{g^{1,\phi}}_i + \frac{\eta_0\gamma_0\iota}{P\kappa}\int_0^t ds\sum_\nu \Delta_\nu G^3\Phi^{2,in}_i\tilde\phi^i\tilde\phi^i - \frac{\eta_0\gamma_0\iota}{P\kappa}\int_0^t ds\sum_\nu \Delta_\nu G^3\bar\Phi^{2,in}\tilde\phi^i.
\end{aligned}
\tag{N.22}
$$

N.8 Stochastic Initial Fields

To isolate the purely deterministic part of the trajectory from the random initialization, we define a set of initial stochastic fields $\mathcal{F}$. These fields encapsulate the randomness of the initial weights:

$$\mathcal{F} = \{\chi^1_\mu(t), \chi^{2,\text{in}}_{\mu,i}(t), \chi^{2,\text{out}}_{\mu,i}(t), \chi^\phi_\mu(t), \tilde{\xi}^1_{\mu,i}(t), \xi^{2,\text{in}}_{\mu,i}(t), \xi^3_\mu(t), \xi^\phi_\mu(t), \xi^{g^\phi}_{\mu i}(t)\}_{i\in\{1,\dots,M\},\mu\in\{1,\dots,P\}} \tag{N.23}$$

$$\begin{aligned}
\chi^1_\mu &= \frac{1}{\sqrt{D}}\, W^1(0)\, x_\mu \\
\chi^{2,\text{in}}_{\mu,i}(t) &= \frac{1}{\sqrt{N}}\, W^{2,\text{in}}_i(0)\, \sigma\big(h^1_\mu(t)\big) \\
\chi^{2,\text{out}}_{\mu,i}(t) &= \frac{1}{\sqrt{N}} W^{2,\text{out}}_i(0)\, \sigma\big(h^{2,\text{in}}_{\mu,i}(t)\big) \\
\chi^\phi_\mu(t) &= \frac{1}{\sqrt{N}}\, Q(0)\, \sigma(h^1_\mu(t)) \\
\tilde{\xi}^1_{\mu,i}(t) &= \frac{1}{\sqrt{N}} \big(W^{2,\text{in}}_i(0)\big)^\top\, g^{2,\text{in}}_{\mu,i}(t) \\
\xi^{2,\text{in}}_{\mu,i}(t) &= \frac{1}{\sqrt{N}}\, \big(W^{2,\text{out}}_i(0)\big)^\top\, g^{2,\text{out}}_{\mu,i}(t) \\
\xi^{1,\phi}_\mu(t) &= \frac{1}{\kappa\sqrt{N}} Q(0)^\top g^{1,\phi}_\mu(t) \\
\xi^{g^\phi}_{\mu i}(t) &= \frac{1}{N^{3/2}} g^{3\top}_\mu(t)\, W^{2,\text{out}}_i(0)\, \sigma(h^{2,\text{in}}_{\mu,i}(t))
\end{aligned} \tag{N.24}$$

A critical step in rendering the MoE partition function computationally tractable is distinguishing between expert-local fields (which are specific to an expert i) and global fields (which impact the shared base layer or the final aggregated output). This distinction allows the massive partition function to factorize over the experts.

We define the router-averaged stochastic fields to capture the macroscopic effect of the local processes:

$$\bar{\chi}^3_\mu(t) = \frac{1}{\kappa N} \sum_{i=1}^{\kappa N} \tilde{\phi}^i_\mu(t) \chi^{2,\text{out}}_{\mu,i}(t) \tag{N.25}$$

$$\bar{\xi}^1_\mu(t) = \frac{1}{\kappa\sqrt{N}} \sum_{i=1}^{\kappa N} \tilde{\xi}^1_{\mu,i}(t) \tag{N.26}$$

Substituting these, the global aggregated output h^3_μ and the gradient arriving at the base layer z^1_μ can be expressed entirely in terms of order parameters that are smooth over the ensemble of experts:

$$h^3_\mu(t) = \bar{\chi}^3_\mu(t) + \frac{\eta_0\gamma_0\iota}{\kappa P} \int_0^t ds \sum_\nu \Delta_\nu(s) \bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) z^3_\nu(s) \tag{N.27}$$

$$z^1_\mu(t) = \bar{\xi}^1_\mu(t) + \xi^{1,\phi}_\mu(t) + \frac{\eta_0\gamma_0\iota}{\kappa P} \int_0^t ds \sum_\nu \Delta_\nu(s) \left[\frac{1}{\sqrt{N}} \bar{G}^{2,\text{in}}_{\mu\nu}(s,t) + G^{1,\phi}_{\mu\nu}(s,t)\right] \sigma(h^1_\nu(s)) \tag{N.28}$$

N.9 Deriving the DMFT Action

We formulate the DMFT by writing the moment-generating function for the system's trajectories and performing a disorder average over the initial weights θ_0.

$$
\begin{aligned}
Z \propto \Bigg\langle \int d\mathcal{F} \exp\Bigg(\sum_{\mu} \int_0^{\infty} dt\, i \Bigg[& \hat{\chi}^1_\mu \cdot \big(\chi^1_\mu - \frac{1}{\sqrt{D}} W^1(0) x_\mu\big) + \hat{\chi}^{2,\text{in}}_\mu \cdot \Big(\chi^{2,\text{in}}_\mu - \frac{1}{\sqrt{N}} W^{2,\text{in}}(0)\sigma(h^1_\mu(t))\Big) \\
& + \hat{\bar{\chi}}^3_\mu \cdot \Big(\bar{\chi}^3_\mu - \frac{1}{\kappa N}\sum_{i=1}^{\kappa N} \tilde{\phi}^i_\mu(t) \frac{1}{\sqrt{N}} W^{2,\text{out}}(0)\sigma(h^{2,\text{in}}_{\mu,i}(t))\Big) + \hat{\bar{\xi}}^1_\mu \cdot \Big(\bar{\xi}^1_\mu - \frac{1}{\kappa N}\sum_{i=1}^{\kappa N} \frac{1}{\sqrt{N}} W^{2,\text{in}}(0)^\top g^{2,\text{in}}_{\mu,i}(t)\Big) \\
& + \frac{1}{\kappa N}\sum_{i=1}^{\kappa N} \hat{\xi}^{2,\text{in}}_{\mu,i} \cdot \Big(\xi^{2,\text{in}}_{\mu,i} - \frac{1}{\sqrt{N}} W^{2,\text{out}}(0)^\top g^{2,\text{out}}_{\mu,i}(t)\Big) + \hat{\xi}^3_\mu \cdot \Big(\xi^3_\mu - w^4(0)^\top\Big) \\
& + \hat{\chi}^\phi_\mu \cdot \Big(\chi^\phi_\mu - \frac{1}{\sqrt{N}} Q(0)\sigma(h^1_\mu(t))\Big) + \hat{\xi}^{1,\phi}_\mu \cdot \Big(\xi^{1,\phi}_\mu - \frac{1}{\kappa\sqrt{N}} Q(0)^\top g^{1,\phi}_\mu(t)\Big) \\
& + \sum_{m=1}^{\kappa N} \hat{\xi}^{g^{1,\phi}}_{\mu,m}(t)\Big(\xi^{g^{1,\phi}}_{\mu,m}(t) - \frac{1}{\kappa N}\sum_{i=1}^{\kappa N} \frac{1}{N^{3/2}} g^{3\,T}_\mu(t) W^{2,\text{out}} \sigma\big(h^{2,\text{in}}_{\mu,i}(t)\big)\tilde{\phi}^i_\mu(t)\left(\kappa N\, \delta_{im} - \tilde{\phi}^m_\mu(t)\right)\Big)\Bigg]\Bigg) \\
\times \exp\Bigg(\sum_{\mu} \int_0^{\infty} dt \Bigg[& j^1_\mu(t)\cdot\chi^1_\mu(t) + j^{2,\text{in}}_\mu(t)\cdot\chi^{2,\text{in}}_\mu(t) + \bar{j}^3_\mu(t)\cdot\bar{\chi}^3_\mu(t) + \frac{1}{\kappa N}\sum_{i=1}^{M} v^{2,\text{in}}_{\mu,i}(t)\cdot\xi^{2,\text{in}}_{\mu,i}(t) \\
& + \bar{v}^1_\mu(t)\cdot\bar{\xi}^1_\mu(t) + v^3_\mu\cdot\xi^3_\mu(t) + j^\phi_\mu(t)\cdot\chi^\phi_\mu(t) + v^{1,\phi}_\mu(t)\cdot\xi^{1,\phi}_\mu(t) + v^{g^{1,\phi}}_\mu(t)\cdot\xi^{g^{1,\phi}}_\mu(t)\Bigg]\Bigg)\Bigg\rangle_{\theta_0}.
\end{aligned}
\tag{N.29}
$$

Given the shared initialisation across experts we now have terms coupling different experts (e.g. $\sigma(h^{2,\text{in}}_\mu)_i\,\sigma(h^{2,\text{in}}_\nu)_j\,\tilde{\phi}^i_\mu\tilde{\phi}^j_\nu$). In order to obtain a stable set of order parameters that scale coherently with N we define the following set of "averaged fields":

$$
\begin{aligned}
1 &= \int \frac{d\hat{\bar{g}}^{2,\text{in}}_\mu(t)}{2\pi}\ \exp\left[\hat{\bar{g}}^{2,\text{in}}_\mu(t)\left(\kappa N \bar{g}^{2,\text{in}}_\mu(t) - \sum_{m=1}^{\kappa N} g^{2,\text{in}}_{\mu,m}(t)\right)\right] \\
1 &= \int \frac{d\hat{\bar{\sigma}}(h_\mu{}^{2,\text{in}}(t))}{2\pi}\ \exp\left[\hat{\bar{\sigma}}(h_\mu{}^{2,\text{in}}(t))\left(\kappa N \bar{\sigma}(h_\mu{}^{2,\text{in}}(t)) - \sum_{i=1}^{\kappa N}\sigma(h^{2,\text{in}}_{\mu,i}(t))\tilde{\phi}^i(t)\right)\right] \\
1 &= \int \frac{d\hat{\bar{\sigma}}^\phi(h_\mu{}^{2,\text{in}}(t))}{2\pi}\ \exp\left[\hat{\bar{\sigma}}^\phi(h_\mu{}^{2,\text{in}}(t))\left(\kappa N \bar{\sigma}^\phi(h_\mu{}^{2,\text{in}}(t)) - \sum_{i=1}^{\kappa N}\sigma(h^{2,\text{in}}_{\mu,i}(t))\tilde{\phi}^i(t)\hat{\xi}^{g^{1,\phi}}_i(t)\right)\right] \\
1 &= \int \frac{d\hat{\bar{\xi}}^{2,\text{in}}_\mu(t)}{2\pi}\ \exp\left[\hat{\bar{\xi}}^{2,\text{in}}_\mu(t)\left(\kappa N \bar{\xi}^{2,\text{in}}_\mu(t) - \sum_{i=1}^{\kappa N}\hat{\xi}^{2,\text{in}}_{\mu,i}(t)\tilde{\phi}^i(t)\right)\right]
\end{aligned}
\tag{N.30}
$$

We must constrain the solutions to be physical, so for each i,μ,ν,s,t we multiply in the following resolutions of the identity:

$$
1 = \int \frac{d\Phi^1_{\mu\nu}(s,t)d\hat{\Phi}^1_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[\hat{\Phi}^1_{\mu\nu}(s,t)\Big(N\Phi^1_{\mu\nu}(s,t) - \sigma(h^1_\mu(s))\cdot\sigma(h^1_\nu(t))\Big)\Big] \tag{N.31}
$$

$$
1 = \int \frac{dG^1_{\mu\nu}(s,t)d\hat{G}^1_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[\hat{G}^1_{\mu\nu}(s,t)\Big(NG^1_{\mu\nu}(s,t) - g^1_\mu(s)\cdot g^1_\nu(t)\Big)\Big] \tag{N.32}
$$

$$
1 = \int \frac{dG^3_{\mu\nu}(s,t)d\hat{G}^3_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[\hat{G}^3_{\mu\nu}(s,t)\Big(NG^3_{\mu\nu}(s,t) - g^3_\mu(s)\cdot g^3_\nu(t)\Big)\Big] \tag{N.33}
$$

$$
1 = \int \frac{dG^{1,\phi}_{\mu\nu}(s,t)d\hat{G}^{1,\phi}_{\mu\nu}(s,t)}{2\pi i\kappa N^{-1}} \exp\Big[\hat{G}^{1,\phi}_{\mu\nu}(s,t)\Big(\kappa NG^{1,\phi}_{\mu\nu}(s,t) - g^{1,\phi}_\mu(s)\cdot g^{1,\phi}_\nu(t)\Big)\Big] \tag{N.34}
$$

$$1 = \int \frac{d\Phi^{2,\text{in}}_{\mu\nu,i}(s,t) d\hat{\Phi}^{2,\text{in}}_{\mu\nu,i}(s,t)}{2\pi i (\iota N)^{-1}} \exp\Big[\hat{\Phi}^{2,\text{in}}_{\mu\nu,i}(s,t)\Big(\iota N \Phi^{2,\text{in}}_{\mu\nu,i}(s,t) - \sigma(h^{2,\text{in}}_{\mu,i}(s)) \cdot \sigma(h^{2,\text{in}}_{\nu,i}(t))\Big)\Big] \tag{N.35}$$

$$1 = \int \frac{dG^{2,\text{in}}_{\mu\nu,i}(s,t) d\hat{G}^{2,\text{in}}_{\mu\nu,i}(s,t)}{2\pi i (\iota N)^{-1}} \exp\Big[\hat{G}^{2,\text{in}}_{\mu\nu,i}(s,t)\Big(\iota N G^{2,\text{in}}_{\mu\nu,i}(s,t) - g^2_{\mu,i}(s) \cdot g^2_{\nu,i}(t)\Big)\Big] \tag{N.36}$$

$$1 = \int \frac{d\bar{A}^1_{\mu\nu}(s,t) d\bar{B}^1_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[-\bar{B}^1_{\mu\nu}(s,t)\Big(N\bar{A}^1_{\mu\nu}(s,t) + i\hat{\bar{\xi}}^1_{\mu}(s) \cdot \sigma(h^1_{\nu}(t))\Big)\Big] \tag{N.37}$$

$$1 = \int \frac{dA^{1,\phi}_{\mu\nu}(s,t) dB^{1,\phi}_{\mu\nu}(s,t)}{2\pi i N^{-1}} \exp\Big[-B^{1,\phi}_{\mu\nu}(s,t)\Big(NA^{1,\phi}_{\mu\nu}(s,t) + i\hat{\xi}^{1,\phi}_{\mu}(s) \cdot \sigma(h^1_{\nu}(t))\Big)\Big] \tag{N.38}$$

$$1 = \int \frac{dA^{g^{1,\phi}}_{\mu\nu}(s,t) dB^{g^{1,\phi}}_{\mu\nu}(s,t)}{2\pi i (\kappa N)^{-1}} \exp\Big[-B^{g^{1,\phi}}_{\mu\nu}(s,t)\Big(\kappa N A^{g^{1,\phi}}_{\mu\nu}(s,t) + i\hat{\xi}^{g^{1,\phi}}_{\mu}(s) \cdot \tilde{\phi}_{\nu}(t)\Big)\Big] \tag{N.39}$$

$$1 = \int \frac{d\Phi^3_{\mu\nu}(t,s)\, d\hat{\Phi}^3_{\mu\nu}(t,s)}{2\pi i\, N^{-1}} \exp\Big[\hat{\Phi}^3_{\mu\nu}(t,s)\,\big(N\Phi^3_{\mu\nu}(t,s) - h^3_{\mu}(t) \cdot h^3_{\nu}(s)\big)\Big] \tag{N.40}$$

$$1 = \int \frac{d\bar{\Phi}^{2,\text{in}}_{\mu\nu}(t,s)\, d\hat{\bar{\Phi}}^{2,\text{in}}_{\mu\nu}(t,s)}{2\pi i} \exp\left[\hat{\bar{\Phi}}^{2,\text{in}}_{\mu\nu}(t,s)\,\left(\kappa N \bar{\Phi}^{2,\text{in}}_{\mu\nu}(t,s) - \sum_{i=1}^{\kappa N} \tilde{\phi}^i_{\mu}(s) \Phi^{2,\text{in}}_{\mu\nu,i}(s,t)\right)\right] \tag{N.41}$$

$$1 = \int \frac{d\bar{G}^{2,\text{in}}_{\mu\nu}(t,s)\, d\hat{\bar{G}}^{2,\text{in}}_{\mu\nu}(t,s)}{2\pi i\,(\kappa N)^{-1}} \exp\left[\hat{\bar{G}}^{2,\text{in}}_{\mu\nu}(t,s)\,\left(\kappa N \bar{G}^{2,\text{in}}_{\mu\nu}(t,s) - \sum_{i=1}^{\kappa N} G^{2,\text{in}}_{\mu\nu,i}(s,t)\right)\right] \tag{N.42}$$

$$1 = \int \frac{d\tilde{G}^{2,\text{in}}_{\mu\nu}(s,t)\, d\hat{\tilde{G}}^{2,\text{in}}_{\mu\nu}(s,t)}{2\pi i\,(\iota N)^{-1}} \exp\Big[\hat{\tilde{G}}^{2,\text{in}}_{\mu\nu}(s,t)\,\Big(\iota N\, \tilde{G}^{2,\text{in}}_{\mu\nu}(s,t) - \bar{g}^{2,\text{in}}_{\mu}(s) \cdot \bar{g}^{2,\text{in}}_{\nu}(t)\Big)\Big] \tag{N.43}$$

$$1 = \int \frac{d\tilde{\Phi}^{2,\text{in}}_{\mu\nu}(s,t)\, d\hat{\tilde{\Phi}}^{2,\text{in}}_{\mu\nu}(s,t)}{2\pi i\,(\iota N)^{-1}} \exp\Big[\hat{\tilde{\Phi}}^{2,\text{in}}_{\mu\nu}(s,t)\,\Big(\iota N\, \tilde{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) - \bar{\sigma}^{2,\text{in}}_{\mu}(s) \cdot \bar{\sigma}^{2,\text{in}}_{\nu}(t)\Big)\Big] \tag{N.44}$$

$$1 = \int \frac{d\tilde{\Phi}^{2,\text{in},\phi}_{\mu\nu}(s,t)\, d\hat{\tilde{\Phi}}^{2,\text{in},\phi}_{\mu\nu}(s,t)}{2\pi i\,(\iota N)^{-1}} \exp\Big[\hat{\tilde{\Phi}}^{2,\text{in},\phi}_{\mu\nu}(s,t)\,\Big(\iota N\, \tilde{\Phi}^{2,\text{in},\phi}_{\mu\nu}(s,t) - \bar{\sigma}^{2,\text{in},\phi}_{\mu}(s) \cdot \bar{\sigma}^{2,\text{in}}_{\nu}(t)\Big)\Big] \tag{N.45}$$

$$1 = \int \frac{d\hat{\bar{\Xi}}^{2,\text{in}}_{\mu\nu}(s,t)\, d\hat{\tilde{\bar{\Xi}}}^{2,\text{in}}_{\mu\nu}(s,t)}{2\pi i\,(\iota N)^{-1}} \exp\left[\hat{\tilde{\bar{\Xi}}}^{2,\text{in}}_{\mu\nu}(s,t)\,\left(\iota N\, \hat{\bar{\Xi}}^{2,\text{in}}_{\mu\nu}(s,t) - \hat{\bar{\xi}}^{2,\text{in}}_{\mu}(s) \cdot \hat{\bar{\xi}}^{2,\text{in}}_{\nu}(t)\right)\right] \tag{N.46}$$

$$1 = \int \frac{d\Phi^{\phi\phi,2,\text{in}}_{\mu\nu}(s,t)\, d\hat{\Phi}^{\phi\phi,2,\text{in}}_{\mu\nu}(s,t)}{2\pi i\,(\iota N)^{-1}} \exp\Big[\hat{\Phi}^{\phi\phi,2,\text{in}}_{\mu\nu}(s,t)\,\big(\iota N\, \Phi^{\phi\phi,2,\text{in}}_{\mu\nu}(s,t) - \bar{\sigma}^{\phi,2,\text{in}}_{\mu}(s) \cdot \bar{\sigma}^{\phi,2,\text{in}}_{\nu}(t)\big)\Big] \tag{N.47}$$

$$1 = \int \frac{d\tilde{A}^{2,\text{in}}_{\mu\nu}(s,t)\, d\tilde{B}^{2,\text{in}}_{\mu\nu}(s,t)}{2\pi i\,(\iota N)^{-1}} \exp\Big[-\tilde{B}^{2,\text{in}}_{\mu\nu}(s,t)\,\Big(\iota N\, \tilde{A}^{2,\text{in}}_{\mu\nu}(s,t) - \bar{\sigma}^{2,\text{in}}_{\mu}(s) \cdot \bar{\xi}^{2,\text{in}}_{\nu}(t)\Big)\Big] \tag{N.48}$$

$$1 = \int \frac{d\tilde{A}^{\phi,2,\text{in}}_{\mu\nu}(s,t)\, d\tilde{B}^{\phi,2,\text{in}}_{\mu\nu}(s,t)}{2\pi i\,(\iota N)^{-1}} \exp\Big[-\tilde{B}^{2,\text{in}}_{\mu\nu}(s,t)\,\Big(\iota N\, \tilde{A}^{2,\text{in}}_{\mu\nu}(s,t) - \bar{\sigma}^{\phi,2,\text{in}}_{\mu}(s) \cdot \bar{\xi}^{2,\text{in}}_{\nu}(t)\Big)\Big] \tag{N.49}$$

N.10 Softmax

As for the previous regimes we enforce the softmax layer order parameter via the Fourier-transformed delta function

$$1 = \int \frac{d\mathcal{S}_\mu(t)\, d\hat{\mathcal{S}}_\mu(t)}{2\pi i\, (\kappa N)^{-1}} \exp\left[\hat{\mathcal{S}}_\mu(t)\left(\kappa N \mathcal{S}_\mu(t) - \sum_{i=1}^{\kappa N} e^{\psi^i_\mu(t)}\right)\right] \tag{N.50}$$

N.11 Partition function

Define the set of all kernels and conjugates (indexed for time and feature, although this is omitted for brevity in N.51):

$$\begin{aligned}\mathcal{K} = \{&\Phi^1, \hat{\Phi}^1, , \bar{\Phi}^{2,\text{in}}, \hat{\bar{\Phi}}^{2,\text{in}}, \Phi^3, \hat{\Phi}^3, G^1, \hat{G}^1, \bar{G}^{2,\text{in}}, \hat{\bar{G}}^{2,\text{in}}, \Phi_i^{2,\text{in}}, \hat{\Phi}_i^{2,\text{in}}, G_i^{2,\text{in}}, \hat{G}_i^{2,\text{in}},\\ &G^3, \hat{G}^3, G^{1,\phi}, \hat{G}^{1,\phi}, \bar{A}^1, \bar{B}^1, \bar{B}^{2,\text{in}}, A^{1,\phi}, B^{1,\phi}, A^{g^{1,\phi}}, B^{g^{1,\phi}}, B^3, A^3,\\ &\tilde{\bar{G}}^{2,\text{in}}, \tilde{G}^{2,\text{in}}\hat{\tilde{\Phi}}^{2,\text{in}}, \tilde{\Phi}^{2,\text{in}}, \hat{\tilde{\Phi}}^{2,\text{in},\phi}, \tilde{\Phi}^{2,\text{in},\phi}, \tilde{B}^{2,\text{in}}, \tilde{A}^{2,\text{in}}, \tilde{B}^{2,\text{in},\phi}, \tilde{A}^{2,\text{in},\phi}, \hat{\tilde{\Xi}}^{2,\text{in}}, \hat{\tilde{\Xi}}^{2,\text{in}}, \hat{\Phi}^{\phi\phi,2,\text{in}}, \Phi^{\phi\phi,2,\text{in}}\}\end{aligned} \tag{N.51}$$

We can partition these order parameters into the set $\mathcal{K}_{global}$ of global order parameters, the set $\mathcal{K}_{exp-loc}$ of expert-local ones, together with the shared fields $\mathcal{F}_j^{2,\text{in,sh}}$, the router fields $\mathcal{F}_i^\phi$, the experts fields $\mathcal{F}_{i,j}^{2,\text{in}}$,$\mathcal{F}_n^{\text{global-site}}$:

$$\begin{aligned}\mathcal{K}_{\text{global}} = \{&\Phi^1, \hat{\Phi}^1, \bar{\Phi}^{2,\text{in}}, \hat{\bar{\Phi}}^{2,\text{in}}, \Phi^3, \hat{\Phi}^3, G^1, \hat{G}^1, \bar{G}^{2,\text{in}}, \hat{\bar{G}}^{2,\text{in}}, G^{1,\phi}, \hat{G}^{1,\phi}, G^3, \hat{G}^3, \bar{A}^1, \bar{B}^1, \bar{A}^{2,\text{in}}, A^{1,\phi},\\ &B^{1,\phi}, A^{g1,\phi}, B^{g1,\phi}, \widetilde{G}^{2,\text{in}}, \hat{\widetilde{G}}^{2,\text{in}}, \widetilde{\Phi}^{2,\text{in}}, \hat{\widetilde{\Phi}}^{2,\text{in}}, \widetilde{\Phi}^{2,\text{in},\phi}, \hat{\widetilde{\Phi}}^{2,\text{in},\phi}, \widetilde{B}^{2,\text{in}},\\ &\widetilde{A}^{2,\text{in}}, \widetilde{B}^{2,\text{in},\phi}, \widetilde{A}^{2,\text{in},\phi}, \hat{\widetilde{\Xi}}^{2,\text{in}}, \hat{\widetilde{\Xi}}^{2,\text{in}}, \Phi^{\phi\phi,2,\text{in}}, \hat{\Phi}^{\phi\phi,2,\text{in}}, S, \hat{S}\}.\end{aligned} \tag{N.52}$$

$$\mathcal{K}_{\text{exp}-\text{loc},i} = \{\Phi_i^{2,\text{in}}, \hat{\Phi}_i^{2,\text{in}}, G_i^{2,\text{in}}, \hat{G}_i^{2,\text{in}}\}. \tag{N.53}$$

$$\mathcal{F}_j^{2,\text{in,sh}} = \{\hat{\chi}^{2,\text{in}}_{\mu,j}(t), \bar{g}^{2,\text{in}}_{\mu,j}(t), \hat{\bar{g}}^{2,\text{in}}_{\mu,j}(t),\ \bar{\sigma}^{2,\text{in}}_{\mu,j}(t), \hat{\bar{\sigma}}^{2,\text{in}}_{\mu,j}(t), \bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(t), \hat{\bar{\sigma}}^{\phi,2,\text{in}}_{\mu,j}(t), \bar{\xi}^{2,\text{in}}_{\mu,j}(t), \hat{\bar{\xi}}^{2,\text{in}}_{\mu,j}(t)\}_{\mu,t}.$$

$$d\mathcal{F}_j^{2,\text{in,sh}} := \textstyle\prod_{\mu,t} d\hat{\chi}^{2,\text{in}}_{\mu,j}(t) d\bar{g}^{2,\text{in}}_{\mu,j}(t) d\hat{\bar{g}}^{2,\text{in}}_{\mu,j}(t) \times d\bar{\sigma}^{2,\text{in}}_{\mu,j}(t) d\hat{\bar{\sigma}}^{2,\text{in}}_{\mu,j}(t) d\bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(t) d\hat{\bar{\sigma}}^{\phi,2,\text{in}}_{\mu,j}(t) \times d\bar{\xi}^{2,\text{in}}_{\mu,j}(t) d\hat{\bar{\xi}}^{2,\text{in}}_{\mu,j}(t).$$ (N.54)

$$\mathcal{F}_i^\phi = \{\chi^\phi_{\mu i}(t), \hat{\chi}^\phi_{\mu i}(t), g^{1,\phi}_{\mu i}(t), \xi^{g1,\phi}_{\mu i}(t), \hat{\xi}^{g1,\phi}_{\mu i}(t), \tilde{\phi}^i_\mu(t), \psi^i_\mu(t)\}_{\mu,t}. \tag{N.55}$$

$$\mathcal{F}_{i,j}^{2,\text{in}} = \{\chi^{2,\text{in}}_{\mu,i,j}(t), h^{2,\text{in}}_{\mu,i,j}(t), g^{2,\text{in}}_{\mu,i,j}(t), \xi^{2,\text{in}}_{\mu,i,j}(t), \hat{\xi}^{2,\text{in}}_{\mu,i,j}(t)\}_{\mu,t}. \tag{N.56}$$

$$\begin{aligned}\mathcal{F}_n^{\text{global-site}} = \{&h^1_{\mu,n}(t), g^1_{\mu,n}(t), \chi^1_{\mu,n}(t), \hat{\chi}^1_{\mu,n}(t), h^3_{\mu,n}(t), g^3_{\mu,n}(t), \xi^3_{\mu,n}(t), \hat{\xi}^3_{\mu,n}(t),\\ &\xi^{1,\phi}_{\mu,n}(t), \hat{\xi}^{1,\phi}_{\mu,n}(t), \bar{\chi}^3_{\mu,n}(t), \hat{\bar{\chi}}^3_{\mu,n}(t), \bar{\xi}^1_{\mu,n}(t), \hat{\bar{\xi}}^1_{\mu,n}(t)\}_{\mu,t}.\end{aligned}$$

Then the partition function can be written as:

$$
\begin{aligned}
Z \propto & \int \left(\prod_{\mu,\nu} \prod_{t,s} d\mathcal{K}_{\text{global}} \right) \exp\left\{ N \mathcal{S}_{\text{global}}[\mathcal{K}_{\text{global}}] \right\} \times \prod_{n=1}^{N} \mathcal{Z}_N^{\text{global}}[j_n^1, \bar{j}_n^3, \bar{v}_n^1, v_n^{1,\phi}, v_n^3] \\
& \times \int \left[\prod_{j=1}^{N_e} d\mathcal{F}_j^{2,\text{in,sh}} \right] \times \left[\exp\left\{ \sum_{j=1}^{N_e} \mathcal{S}_j^{2,\text{in,sh}} \right\} \right. \\
& \left. \times \prod_{i=1}^{\kappa N} \mathcal{Z}_i^{\text{local}} \left[j_i^{2,\text{in}}, v_i^{2,\text{in}}, j_i^{3,\text{out}}, j_i^{\phi}, v_i^{g1,\phi} \; ; \; \{\mathcal{F}_j^{2,\text{in,sh}}\}_{j=1}^{N_e} \right] \right].
\end{aligned}
$$

with the global action defined as:

$$
\begin{aligned}
\mathcal{S}_{\text{global}} = & \sum_{\mu,\nu} \int dt \int ds \Big[\hat{\Phi}_{\mu\nu}^1(t,s) \Phi_{\mu\nu}^1(t,s) + \hat{G}_{\mu\nu}^1(t,s) G_{\mu\nu}^1(t,s) + \kappa \hat{G}_{\mu\nu}^{1,\phi}(t,s) G_{\mu\nu}^{1,\phi}(t,s) \\
& - \bar{B}_{\mu\nu}^1(t,s) \bar{A}_{\mu\nu}^1(t,s) - B_{\mu\nu}^{1,\phi}(t,s) A_{\mu\nu}^{1,\phi}(t,s) + \hat{\Phi}_{\mu\nu}^3(t,s) \Phi_{\mu\nu}^3(t,s) \\
& + \kappa \hat{\bar{\Phi}}_{\mu\nu}^{2,\text{in}}(t,s) \bar{\Phi}_{\mu\nu}^{2,\text{in}}(t,s) \\
& + \kappa \hat{\bar{G}}_{\mu\nu}^{2,\text{in}}(t,s) \bar{G}_{\mu\nu}^{2,\text{in}}(t,s) + \kappa \hat{S}_{\mu}(t) S_{\nu}(s) + \hat{G}_{\mu\nu}^3(t,s) G_{\mu\nu}^3(t,s) \\
& - \kappa B_{\mu\nu}^{g^{1,\phi}}(t,s) A_{\mu\nu}^{g^{1,\phi}}(t,s) + \iota \hat{\widetilde{G}}_{\mu\nu}^{2,\text{in}}(t,s) \widetilde{G}_{\mu\nu}^{2,\text{in}}(t,s) \\
& + \iota \hat{\widetilde{\Phi}}_{\mu\nu}^{2,\text{in}}(t,s) \widetilde{\Phi}_{\mu\nu}^{2,\text{in}}(t,s) + \iota \hat{\widetilde{\Phi}}_{\mu\nu}^{2,\text{in},\phi}(t,s) \widetilde{\Phi}_{\mu\nu}^{2,\text{in},\phi}(t,s) \\
& - \iota \widetilde{B}_{\mu\nu}^{2,\text{in}}(t,s) \widetilde{A}_{\mu\nu}^{2,\text{in}}(t,s) - \iota \widetilde{B}_{\mu\nu}^{2,\text{in},\phi}(t,s) \widetilde{A}_{\mu\nu}^{2,\text{in},\phi}(t,s) \\
& + \iota \hat{\widetilde{\Xi}}_{\mu\nu}^{2,\text{in}}(t,s) \widetilde{\Xi}_{\mu\nu}^{2,\text{in}}(t,s) + \iota \hat{\Phi}_{\mu\nu}^{\phi\phi,2,\text{in}}(t,s) \Phi_{\mu\nu}^{\phi\phi,2,\text{in}}(t,s) \\
& + i\kappa \bar{A}_{\mu\nu}^{2,\text{in}}(t,s) G_{\mu\nu}^3(t,s) A_{\mu\nu}^{g^{1,\phi}}(t,s) - \iota \frac{1}{2} \hat{\widetilde{\Xi}}_{\mu\nu}^{2,\text{in}} G^3 \\
& + \kappa A_{\mu\nu}^{g^{1,\phi}}(t,s) A_{\mu\nu}^{g^{1,\phi}}(t,s) G_{\mu\nu}^3(t,s) \widetilde{\Phi}_{\mu\nu}^{2,\text{in}}(t,s) \\
& + i\kappa\iota \widetilde{\Phi}_{\mu\nu}^{2,\text{in},\phi}(t,s) G_{\mu\nu}^3(t,s) A_{\mu\nu}^{g1,\phi}(t,s) - \frac{\iota}{2} \Phi_{\mu\nu}^{\phi\phi,2,\text{in}}(t,s) G_{\mu\nu}^3(t,s) \\
& - \iota\kappa \widetilde{A}_{\mu\nu}^{\phi,2,\text{in}}(t,s) G_{\mu\nu}^3(t,s) + i\kappa \widetilde{A}_{\mu\nu}^{2,\text{in}}(t,s) G_{\mu\nu}^3(t,s) A_{\mu\nu}^{g1,\phi}(t,s) \Big].
\end{aligned}
$$

and the single-site global action as:

$$
\mathcal{Z}_N^{\text{global}}[j_n^1, \bar{j}_n^3, \bar{v}_n^1, v_n^{1,\phi}, v_n^3] := \int d\mathcal{K}_n^{\text{global-site}} \; \exp\left\{ \mathcal{S}_n^{\text{global-site}} \right\}.
$$

$$
\begin{aligned}
\mathcal{S}_n^{\text{global-site}} = &-\frac{1}{2}\sum_{\mu,\nu}\int dt\int ds\Big[\hat{\chi}^1_{\mu,n}(t)\hat{\chi}^1_{\nu,n}(s)\Phi^0_{\mu\nu}(t,s)+\hat{\xi}^3_{\mu,n}(t)\hat{\xi}^3_{\nu,n}(s)\\
&\quad+\hat{\xi}^{1,\phi}_{\mu,n}(t)\hat{\xi}^{1,\phi}_{\nu,n}(s)G^{1,\phi}_{\mu\nu}(t,s)\Big]\\
&-\sum_{\mu,\nu}\int dt\int ds\Big[\hat{\Phi}^1_{\mu\nu}(t,s)\sigma(h^1_{\mu,n}(t))\sigma(h^1_{\nu,n}(s))+\hat{G}^1_{\mu\nu}(t,s)g^1_{\mu,n}(t)g^1_{\nu,n}(s)\\
&\quad+\hat{\Phi}^3_{\mu\nu}(t,s)h^3_{\mu,n}(t)h^3_{\nu,n}(s)+\hat{G}^3_{\mu\nu}(t,s)g^3_{\mu,n}(t)g^3_{\nu,n}(s)\\
&\quad-\frac{1}{2}\widetilde{\Phi}^{2,\text{in}}_{\mu\nu}(t,s)\hat{\bar{\chi}}^3_{\mu,n}(t)\hat{\bar{\chi}}^3_{\nu,n}(s)+\frac{1}{2}\widetilde{G}^{2,\text{in}}_{\mu\nu}(t,s)\hat{\bar{\xi}}^1_{\mu,n}(t)\hat{\bar{\xi}}^1_{\nu,n}(s)\\
&\quad-\iota\kappa\tilde{\Phi}^{2,\text{in}}_{\mu\nu}\hat{\bar{\chi}}^3g^3A^{g^{1,\phi}}+\iota\kappa\bar{A}^{2,\text{in}}_{\mu\nu}\hat{\bar{\chi}}^3g^3\Big]\\
&-i\sum_{\mu,\nu}\int dt\int ds\Big[B^{1,\phi}_{\mu\nu}(t,s)\hat{\xi}^{1,\phi}_{\mu,n}(t)\sigma(h^1_{\nu,n}(s))+\bar{B}^1_{\mu\nu}(t,s)\hat{\bar{\xi}}^1_{\mu,n}(t)\sigma(h^1_{\nu,n}(s))\\
&+\iota\kappa\tilde{\Phi}^{2,\text{in},\phi}_{\mu\nu}\hat{\bar{\chi}}^3g^3\Big]\\
&+\sum_{\mu}\int dt\Big[\big(v^3_{\mu,n}(t)+i\hat{\xi}^3_{\mu,n}(t)\big)\xi^3_{\mu,n}(t)+\big(v^{1,\phi}_{\mu,n}(t)+i\hat{\xi}^{1,\phi}_{\mu,n}(t)\big)\xi^{1,\phi}_{\mu,n}(t)\\
&\quad+\big(j^1_{\mu,n}(t)+i\hat{\chi}^1_{\mu,n}(t)\big)\chi^1_{\mu,n}(t)+\big(\bar{j}^3_{\mu,n}(t)+i\hat{\bar{\chi}}^3_{\mu,n}(t)\big)\bar{\chi}^3_{\mu,n}(t)\\
&\quad+\big(\bar{v}^1_{\mu,n}(t)+i\hat{\bar{\xi}}^1_{\mu,n}(t)\big)\bar{\xi}^1_{\mu,n}(t)\Big].
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{S}_j^{2,\text{in,sh}} = &-\frac{1}{2}\sum_{\mu,\nu}\int dt\int ds\ N\Big[\hat{\chi}^{2,\text{in}}_{\mu,j}(t)\hat{\chi}^{2,\text{in}}_{\nu,j}(s)\Phi^1_{\mu\nu}(t,s)-\hat{\tilde{G}}^{2,\text{in}}_{\mu\nu}(t,s)\bar{g}^{2,\text{in}}_{\mu,j}(t)\bar{g}^{2,\text{in}}_{\nu,j}(s)\\
&-\hat{\tilde{\Phi}}^{2,\text{in}}_{\mu\nu}(t,s)\bar{\sigma}^{2,\text{in}}_{\mu,j}(t)\bar{\sigma}^{2,\text{in}}_{\nu,j}(s)-\hat{\tilde{\Phi}}^{2,\text{in},\phi}_{\mu\nu}(t,s)\bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(t)\bar{\sigma}^{2,\text{in}}_{\nu,j}(s)\\
&-\hat{\tilde{\Xi}}^{2,\text{in}}_{\mu\nu}(t,s)\hat{\bar{\xi}}^{2,\text{in}}_{\mu,j}(t)\hat{\bar{\xi}}^{2,\text{in}}_{\nu,j}(s)\\
&-\hat{\Phi}^{\phi\phi,2,\text{in}}_{\mu\nu}(t,s)\bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(t)\bar{\sigma}^{\phi,2,\text{in}}_{\nu,j}(s)-\widetilde{B}^{2,\text{in}}_{\mu\nu}(t,s)\bar{\sigma}^{2,\text{in}}_{\mu,j}(t)\hat{\bar{\xi}}^{2,\text{in}}_{\nu,j}(s)\\
&-\widetilde{B}^{\phi,2,\text{in}}_{\mu\nu}(t,s)\bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(t)\hat{\bar{\xi}}^{2,\text{in}}_{\nu,j}(s)\Big]\\
&-i\sum_{\mu}\int dt\Big[\hat{\bar{g}}^{2,\text{in}}_{\mu,j}(t)\bar{g}^{2,\text{in}}_{\mu,j}(t)+\hat{\bar{\sigma}}^{2,\text{in}}_{\mu,j}(t)\bar{\sigma}^{2,\text{in}}_{\mu,j}(t)\\
&+\hat{\bar{\sigma}}^{\phi,2,\text{in}}_{\mu,j}(t)\bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(t)+\hat{\bar{\bar{\xi}}}^{2,\text{in}}_{\mu,j}(t)\hat{\bar{\xi}}^{2,\text{in}}_{\mu,j}(t)\Big]\\
&-i\sum_{\mu,\nu}\int dt\int ds\Big[-\frac{1}{2}\Phi^1_{\mu\nu}(t,s)\hat{\chi}^{2,\text{in}}_{\mu,j}(t)\hat{\chi}^{2,\text{in}}_{\nu,j}(s)\Big].
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{Z}_i^{\text{local}}\Big[j_i^{2,\text{in}},v_i^{2,\text{in}},j_i^{3,\text{out}},j_i^{\phi},v_i^{g1,\phi}\ ;\ \{\mathcal{K}_j^{2,\text{in,sh}}\}_{j=1}^{N_e}\Big] := &\int d\mathcal{K}_{\text{exp}-\text{loc},i}\\
&\int d\mathcal{K}_i^{\phi}\times\exp\Big\{\mathcal{S}_i^{\text{exp}-\text{loc}}+\mathcal{S}_i^{\phi}\Big\}\times\frac{\iota}{\iota N}\prod_{j=1}^{N_e}\Big[\int d\mathcal{K}_{i,j}^{2,\text{in}}\ \exp\Big\{\mathcal{S}_{i,j}^{2,\text{in}}\Big\}\Big].
\end{aligned}
$$

$$
\mathcal{S}_i^{\text{exp}-\text{loc}}=\sum_{\mu,\nu}\int_0^{\infty}dt\int_0^{\infty}ds\Big[\iota\hat{\Phi}^{2,\text{in}}_{\mu\nu,i}(t,s)\Phi^{2,\text{in}}_{\mu\nu,i}(t,s)+\iota\hat{G}^{2,\text{in}}_{\mu\nu,i}(t,s)G^{2,\text{in}}_{\mu\nu,i}(t,s)\Big]. \tag{N.57}
$$

$$
\begin{aligned}
\mathcal{S}_i^{\phi} =& \sum_{\mu} \int_0^{\infty} dt\; g^{1,\phi}_{\mu i}(t) g^{1,\phi}_{\mu i}(t) \hat{G}^{1,\phi}_{\mu\nu} - \frac{1}{2} \sum_{\mu,\nu} \int dt \int ds \Big[\hat{\chi}^{\phi}_{\mu i}(t) \hat{\chi}^{\phi}_{\nu i}(s) \Phi^{1}_{\mu\nu}(t,s) \Big] \\
& - i \sum_{\mu,\nu} \int dt \int ds \Big[\hat{\chi}^{\phi}_{\mu i}(t) g^{1,\phi}_{\nu i}(s) A^{1,\phi}_{\mu\nu}(t,s) - B^{g1,\phi}_{\mu\nu}(t,s) \hat{\xi}^{g1,\phi}_{\mu i}(t) \tilde{\phi}^{i}_{\nu}(s) \Big] \\
& - \sum_{\mu,\nu} \int dt \int ds \Big[\hat{G}^{1,\phi}_{\mu\nu}(t,s) g^{1,\phi}_{\mu i}(t) g^{1,\phi}_{\nu i}(s) + \hat{\bar{\Phi}}^{2,\mathrm{in}}_{\mu\nu}(t,s) \tilde{\phi}^{i}_{\mu}(t) \tilde{\phi}^{i}_{\nu}(s) \Phi^{2,\mathrm{in}}_{\mu\nu,i}(t,s) \\
& + \hat{\bar{G}}^{2,\mathrm{in}}_{\mu\nu}(t,s) G^{2,\mathrm{in}}_{\mu\nu,i}(t,s) \Big] \\
& - \sum_{\mu} \int dt\; \hat{S}_{\mu}(t) e^{\psi^{i}_{\mu}(t)} + \sum_{\mu} \int dt \Big[\big(j^{\phi}_{\mu i}(t) + i \hat{\chi}^{\phi}_{\mu i}(t) \big) \chi^{\phi}_{\mu i}(t) + \big(v^{g1,\phi}_{\mu i}(t) + i \hat{\xi}^{g1,\phi}_{\mu i}(t) \big) \xi^{g1,\phi}_{\mu i}(t) \Big].
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{S}^{2,\mathrm{in}}_{i,j} =& - i \sum_{\mu,\nu} \int dt \int ds \Big[\frac{1}{\kappa} \hat{\chi}^{2,\mathrm{in}}_{\mu,j}(t) g^{2,\mathrm{in}}_{\nu,i,j}(s) \bar{A}^{1}_{\mu\nu}(t,s) \Big] \\
& - i \sum_{\mu} \int dt \Big[\hat{\bar{g}}^{2,\mathrm{in}}_{\mu,j}(t) g^{2,\mathrm{in}}_{\mu,i,j}(t)\; \hat{\bar{\sigma}}^{2,\mathrm{in}}_{\mu,j}(t) \sigma\Big(h^{2,\mathrm{in}}_{\mu,i,j}(t) \Big)\; \tilde{\phi}^{i}_{\mu}(t) \\
& \quad - \hat{\bar{\sigma}}^{\phi,2,\mathrm{in}}_{\mu,j}(t) \sigma\Big(h^{2,\mathrm{in}}_{\mu,i,j}(t) \Big)\; \tilde{\phi}^{i}_{\mu}(t) \hat{\xi}^{g1,\phi}_{\mu i}(t) - \hat{\bar{\xi}}^{2,\mathrm{in}}_{\mu,j}(t) \hat{\xi}^{2,\mathrm{in}}_{\mu,i,j}(t) \tilde{\phi}^{i}_{\mu}(t) \Big] \\
& - \sum_{\mu,\nu} \int dt \int ds \Big[\hat{\Phi}^{2,\mathrm{in}}_{\mu\nu,i}(t,s) \sigma\Big(h^{2,\mathrm{in}}_{\mu,i,j}(t) \Big)\; \sigma\Big(h^{2,\mathrm{in}}_{\nu,i,j}(s) \Big) + \hat{G}^{2,\mathrm{in}}_{\mu\nu,i}(t,s) g^{2,\mathrm{in}}_{\mu,i,j}(t) g^{2,\mathrm{in}}_{\nu,i,j}(s) \Big] \\
& + \sum_{\mu} \int dt\; N \Big[\big(j^{2,\mathrm{in}}_{\mu i}(t) + i \hat{\chi}^{2,\mathrm{in}}_{\mu,j}(t) \big) \chi^{2,\mathrm{in}}_{\mu,i,j}(t) + \big(\frac{1}{\kappa N} v^{2,\mathrm{in}}_{\mu,i}(t) + i \hat{\xi}^{2,\mathrm{in}}_{\mu,i,j}(t) \big) \xi^{2,\mathrm{in}}_{\mu,i,j}(t) \Big].
\end{aligned}
$$

One way to look at it is the following: there are four conditional factorization levels in the full partition function. The first is the standard global N-site DMFT factorization over n. The remaining three form a nested local/shared sector: a shared N_e-component factorization over j, an expert factorization over i conditional on the shared j-fields, and an intra-expert N_e-component factorization over (i,j)-local fields. The important point is that these are *conditional* factorizations.

N.12 Saddle Point Approximation

To write down the saddle point equations, we first define the single-site distributions, where $\mathcal{H}$ in each case is the logarithm of the integrand of the corresponding $\mathcal{Z}$. We can then define for each $\mathcal{Z} \in \{\mathcal{Z}_N^{global}, Z_{\mathrm{sh}}, \mathcal{Z}_{N_e\, j}, \mathcal{Z}_M\}$ and the corresponding $\mathcal{H}$ the average

$$
\langle \mathcal{O}(\{\chi,\xi\}) \rangle_{\mathcal{Z}} = \frac{1}{\mathcal{Z}} \int \prod_{\mu} d\mathcal{F} \exp(-\mathcal{H}[\{\chi,\xi\},\{j,v\}]) \mathcal{O}(\{\chi,\xi\}) \tag{N.58}
$$

With this apparatus in place, we can take saddle equations. We treat expert-local kernels as microvariables which are implicitly defined in terms of χ, ξ, and so do not take saddle equations of them at this point.

N.13 Saddle-Point Equations

As in Regime II, we can use the saddle point approximation, which consist in finding the set of equations that lead to a stationary action $\bar{S}[\mathcal{K}]$.

We note the following:

- At zero source, all single site averages $\langle \rangle_{\mathcal{Z}_N^{global}}, \langle \rangle_{\mathcal{Z}_{sh}}$ are equivalent.

- *Conditional on the set $\mathcal{K}_{global}, \mathcal{F}_{sh}$* of global and shared kernels, $\mathcal{Z}^{local}$ factorises over experts, and so we can write $\langle\rangle_{\mathcal{Z}_{N_e j}}$ for the conditional average over the distribution over the experts neurons and $\langle\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}$ for the conditional average over the distribution of the router's neurons.
- Expert-local variables follow single-site processes, so we can drop expert indices
- By a derivation similar to that in K, we can prove that conjugate kernels defined as covariances between conjugate fields vanish, since they have no physical meaning and can not influence the dynamics in addition to imposing constraint when introducing kernel definitions.

The global kernels are given by their expectation values over the single-site global distribution $\mathcal{Z}_N^{\text{global}}$:

$$\begin{aligned}
\Phi^1_{\mu\nu}(s,t) &= \left\langle \sigma(h^1_\mu(s))\, \sigma(h^1_\nu(t)) \right\rangle_{\mathcal{Z}_N^{\text{global}}|\mathcal{K}_{global}} \\
\Phi^3_{\mu\nu}(s,t) &= \left\langle h^3_\mu(s)\, h^3_\nu(t) \right\rangle_{\mathcal{Z}_N^{\text{global}}|\mathcal{K}_{global}} \\
G^1_{\mu\nu}(s,t) &= \left\langle g^1_\mu(s)\, g^1_\nu(t) \right\rangle_{\mathcal{Z}_N^{\text{global}}|\mathcal{K}_{global}} \\
G^3_{\mu\nu}(s,t) &= \left\langle g^3_\mu(s)\, g^3_\nu(t) \right\rangle_{\mathcal{Z}_N^{global}|\mathcal{K}_{global}} \\
A^{1,\phi}_{\mu\nu}(s,t) &= -i\left\langle \hat{\xi}^{1,\phi}_\mu(s)\, \sigma(h^1_\nu(t))(t) \right\rangle_{\mathcal{Z}_N^{global}|\mathcal{K}_{global}} \\
\bar{A}^1_{\mu\nu}(s,t) &= -i\left\langle \hat{\bar{\xi}}^1_\mu(t)\sigma(h^1_\nu(s)) \right\rangle_{\mathcal{Z}_N^{global}|\mathcal{K}_{global}} \\
\bar{B}^{2,\text{in}}_{\mu\nu}(s,t) &= -i\left\langle \hat{\bar{\chi}}^3_\mu(t) g^3_\nu(s) \right\rangle_{\mathcal{Z}_N^{global}|\mathcal{K}_{global}}
\end{aligned} \tag{N.59}$$

The shared kernels are given by their expectation values over the single-site shared distribution $\mathcal{Z}_{sh}$:

$$\widetilde{G}^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle \bar{g}^{2,\text{in}}_{\mu,j}(s) \bar{g}^{2,\text{in}}_{\nu,j}(t) \right\rangle_{\mathcal{Z}_{sh}|\mathcal{K}_{\text{global}}}, \tag{N.60}$$

$$\widetilde{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle \bar{\sigma}^{2,\text{in}}_{\mu,j}(s) \bar{\sigma}^{2,\text{in}}_{\nu,j}(t) \right\rangle_{\mathcal{Z}_{sh}|\mathcal{K}_{\text{global}}}, \tag{N.61}$$

$$\widetilde{\Phi}^{2,\text{in},\phi}_{\mu\nu}(s,t) = \left\langle \bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(s) \bar{\sigma}^{2,\text{in}}_{\nu,j}(t) \right\rangle_{\mathcal{Z}_{sh}|\mathcal{K}_{\text{global}}}, \tag{N.62}$$

$$\hat{\widetilde{\Xi}}^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle \hat{\bar{\xi}}^{2,\text{in}}_{\mu,j}(s) \hat{\bar{\xi}}^{2,\text{in}}_{\nu,j}(t) \right\rangle_{\mathcal{Z}_{sh}|\mathcal{K}_{\text{global}}} \tag{N.63}$$

$$\tilde{\Phi}^{\phi\phi,2,\text{in}}_{\mu\nu}(s,t) = \left\langle \bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(s) \bar{\sigma}^{\phi,2,\text{in}}_{\nu,j}(t) \right\rangle_{\mathcal{Z}_{sh}|\mathcal{K}_{\text{global}}} \tag{N.64}$$

$$\widetilde{A}^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle \bar{\sigma}^{2,\text{in}}_{\mu,j}(s) \bar{\xi}^{2,\text{in}}_{\nu,j}(t) \right\rangle_{\mathcal{Z}_{sh}|\mathcal{K}_{\text{global}}} \tag{N.65}$$

$$\widetilde{A}^{\phi,2,\text{in}}_{\mu\nu}(s,t) = \left\langle \bar{\sigma}^{\phi,2,\text{in}}_{\mu,j}(s) \bar{\xi}^{2,\text{in}}_{\nu,j}(t) \right\rangle_{\mathcal{Z}_{sh}|\mathcal{K}_{\text{global}}} \tag{N.66}$$

$$\widetilde{B}^{2,\text{in}}_{\mu\nu}(s,t) = i\kappa G^3_{\mu\nu}(s,t) A^{g^{1,\phi}}_{\mu\nu}(s,t) \tag{N.67}$$

$$\widetilde{B}^{\phi,2,\text{in}}_{\mu\nu}(s,t) = -i\kappa G^3_{\mu\nu}(s,t) \tag{N.68}$$

The router kernels require expectations over the expert ensemble distribution $\mathcal{Z}_M$, conditioned on the global kernels and shared fields:

$$\begin{aligned}
\bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) &= \left\langle \tilde{\phi}_\mu(s)\, \tilde{\phi}_\nu(t) \Phi^{2,\text{in}}_{\mu\nu}(s,t) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
\bar{G}^{2,\text{in}}_{\mu\nu}(s,t) &= \left\langle G^{2,\text{in}}_{\mu\nu,}(s,t) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
G^{1,\phi}_{\mu\nu}(s,t) &= \left\langle g^{1,\phi}_\mu(s)\, g^{1,\phi}_\nu(t) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
\mathcal{S}_\mu(t) &= \left\langle e^{\psi_\mu(t)} \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
A^{g^{1,\phi}}_{\mu\nu}(s,t) &= -i \left\langle \hat{\xi}^{g^{1,\phi}}_\mu(s) \tilde{\phi}_\nu(t) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
B^{1,\phi}_{\mu\nu}(s,t) &= -i\kappa \left\langle \hat{\chi}^\phi_\mu(s) g^{1,\phi}_\nu(t) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
B^{g^{1,\phi}}_{\mu\nu}(s,t) &= +\, i\bar{A}^{2,\text{in}}_{\mu\nu}(t,s) G^3_{\mu\nu}(t,s) + A^{g^{1,\phi}}_{\mu\nu}(t,s) G^3_{\mu\nu}(t,s) \widetilde{\Phi}^{2,\text{in}}_{\mu\nu}(t,s) \\
&\quad + i\iota \widetilde{\Phi}^{2,\text{in},\phi}_{\mu\nu}(t,s) G^3_{\mu\nu}(t,s) + i\widetilde{A}^{2,\text{in}}_{\mu\nu}(t,s) G^3_{\mu\nu}(t,s) - \iota \tilde{\Phi}^{2,\text{in}}_{\mu\nu} \langle \hat{\tilde{\chi}}^3 g^3 \rangle_{\mathcal{Z}_M|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}}
\end{aligned} \tag{N.69}$$

N.14 Expert-Local Order Parameters

The local kernels are determined by independent processes within each expert's internal dimensions:

$$\begin{aligned}
\Phi^{2,\text{in}}_{\mu\nu}(s,t) &= \left\langle \sigma(\left[h^{2,\text{in}}_\mu(t)\right]_j)\, \sigma(\left[h^{2,\text{in}}_\nu(s)\right]_j) \right\rangle_{\mathcal{Z}_{N_e\, j}|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
G^{2,\text{in}}_{\mu\nu}(s,t) &= \left\langle \left[g^{2,\text{in}}_\mu(t)\right]_j \left[g^{2,\text{in}}_\nu(s)\right]_j \right\rangle_{\mathcal{Z}_{N_e\, j}|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
A^{2,\text{in}}_{\mu\nu}(s,t) &= \left\langle \left[\hat{\xi}^{2,\text{in}}_\mu(s)\right]_j \sigma(\left[h^{2,\text{in}}_\nu\right]_j) \right\rangle_{\mathcal{Z}_{N_e\, j}|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}} \\
\bar{B}^1_{\mu\nu}(s,t) &= -\frac{i}{\kappa} \left\langle \hat{\chi}^{2,\text{in}}_\mu(s) g^{2,\text{in}}_\nu(t) \right\rangle_{\mathcal{Z}_{N_e\, j}|\mathcal{K}_{\text{global}},\mathcal{F}_{\text{shared}}}
\end{aligned} \tag{N.70}$$

All non-physical conjugate kernels identically vanish at the saddle point: $\hat{\Phi}^3 = \hat{\Phi}^1 = \hat{G}^1 = \hat{\bar{\Phi}}^{2,\text{in}} = \cdots = 0$.

N.15 Hubbard-Stratonovich transformation

The Hubbard-Stratonovich trick allows us to rewrite the quadratic terms over conjugate fields as Gaussian integrals over linear conjugate fields:

$$\exp\left(-\tfrac{1}{2}\, \hat{\mathbf{x}}^\top A\, \hat{\mathbf{x}}\right) = \int_{\mathbb{R}^d} \frac{d\mathbf{u}}{(2\pi)^{d/2}\sqrt{\det A}}\, \exp\left(-\tfrac{1}{2}\, \mathbf{u}^\top A^{-1}\mathbf{u} - i\, \mathbf{u}\cdot\hat{\mathbf{x}}\right) = \left\langle \exp(-i\, \mathbf{u}\cdot\hat{\mathbf{x}}) \right\rangle_{\mathbf{u}\sim\mathcal{N}(0,A)}. \tag{N.71}$$

where $\hat{\mathbf{x}}$ is a generic conjugate fields in the partition function above. Using Stein's lemma, we can reformulate the definitions of each of the kernels A and B as response functions:

$$\begin{aligned}
A^\ell_{\mu\nu,i}(t,s) &= -\, i \left\langle \hat{\xi}^\ell_{\nu,i}(t)\, \sigma\big(h^{\ell-1}_{\mu,i}(s)\big) \right\rangle_{\beta^\ell_{\nu,i}(s)} \\
&= \left[G^\ell_i\right]^{-1} \left\langle \left(\xi^\ell_i - B^\ell_i\, \sigma\big(h^{\ell-1}_i\big)\right) \sigma\big(h^{\ell-1}_i\big) \right\rangle_{\beta^\ell_{\nu,i}(s)} \\
&= \left\langle \frac{\partial \sigma(h^{\ell-1}_{\mu,i}(t))}{\partial \beta^{\ell\top}_{\nu,i}(s)} \right\rangle_{\beta^\ell_{\nu,i}(s)}
\end{aligned} \tag{N.72}$$

It easier now integrate over all the $\hat{\mathbf{x}}$'s, since the argument of the exponential in $\mathcal{Z}$ has been linearised with respect to them all. Doing so yields delta functions that give us the final DMFT dynamics.

N.16 DMFT Dynamics

Piecing together the self-consistent order parameters and the exact integrations over the Hubbard-Stratonovich fields, the DMFT description of the infinite-width MoE is summarized below.

$$
\begin{aligned}
h^1_\mu(t) &= \alpha^1_\mu + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu \Delta_\nu\, g^1_\nu(s)\,\Phi^0_{\mu\nu} \qquad \alpha^1 \sim \mathcal{N}(0, K^x)\\
h^{2,\text{in}}_{\mu,i}(t) &= \alpha^{2,\text{in}}_\mu(t) + \frac{\eta_0\gamma_0}{\kappa P}\int_0^t ds \sum_\nu \left[\Delta_\nu \Phi^1_{\mu\nu}(s,t) + \bar{A}^1_{\nu\mu}(s,t)\right] g^{2,\text{in}}_{\nu,i}(s) \qquad \alpha^{2,\text{in}} \sim \mathcal{N}(0, \Phi^1(t))\\
h^3_\mu(t) &= \bar{\alpha}^3_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu \left[\Delta_\nu \bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) + \mathcal{R}^3_{\nu\mu}(s,t)\right] g^3_\nu(s) \qquad \bar{\alpha}^3 \sim \mathcal{N}(0, \tilde{\Phi}^{2,\text{in}}(t))\\
&\quad \text{with } \mathcal{R}^3_{\nu\mu}(s,t) = \kappa \bar{A}^{2,\text{in}}_{\nu\mu}(s,t) + \kappa\widetilde{\Phi}^{2,\text{in},\phi}_{\nu\mu}(s,t) - \kappa\widetilde{\Phi}^{2,\text{in}}_{\nu\mu}(s,t) A^{g^{1,\phi}}_{\nu\mu}(s,t)\\
\psi_\mu(t) &= \alpha^\phi_\mu + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_{\nu=1}^P \left[\Delta_\nu(s)\Phi^1_{\mu\nu}(s,t) + A^{1,\phi}_{\mu\nu}\right] g^{1,\phi}_\nu(s) \qquad \alpha^\phi \sim \mathcal{N}(0, \frac{1}{\kappa}\Phi^1(t))\\
\tilde{\phi}_\mu(t) &= \frac{e^{\psi_\mu(t)}}{\mathcal{S}_\mu(t)}, \qquad \hat{\psi}_\mu(t) = m_\mu(t)\psi_\mu(t), \qquad m_\mu(t) = \mathbf{1}(\psi_\mu(t) - \tau_\mu(t))\\
&\quad \text{where } \tau_\mu(t) \text{ is chosen such that } \rho = \langle m_\mu(t)\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}
\end{aligned}
\tag{N.73}
$$

$$
\begin{aligned}
g^\phi_\mu(t) &= \frac{\eta_0\gamma_0\iota}{P\kappa}\int_0^t ds \sum_\nu G^3_{\mu\nu}(t,s)\tilde{\phi}_\mu\left[\Phi^{2,\text{in}}_{\mu\nu,i}(t,s)\tilde{\phi}_\nu - \bar{\Phi}^2_{\mu\nu}(t,s)\right]\Delta_\nu\\
z^1_\mu(t) &= \bar{\beta}^1_\mu(t) + \beta^{1,\phi}_\mu(t) + \frac{\eta_0\gamma_0}{P}\int_0^t ds \sum_\nu \Big[\bar{B}^1_{\mu\nu} + B^{1,\phi}_{\mu\nu}(t,s) + \Big[\Delta_\nu \frac{\iota}{\kappa}\bar{G}^2_{\mu\nu}(s,t)\\
&\quad + \Delta_\nu G^{1,\phi}_{\mu\nu}(s,t)\Big]\Big]\,\sigma(h^1_\nu(s))\\
&\quad \bar{\beta}^1_\mu(t) \sim \mathcal{N}(0, \tilde{G}^{2,\text{in}}(t,t)) \qquad \beta^{1,\phi}_\mu(t) \sim \mathcal{N}(0, G^{1,\phi}(t,t))\\
z^{2,\text{in}}_\mu(t) &= \beta^{2,\text{in}}_\mu(t)\tilde{\phi}^i_\mu + \frac{\eta_0\gamma_0}{\kappa P}\int_0^t ds \sum_\nu \Delta_\nu\,\sigma(h^{2,\text{in}}_{\nu,i}(s))\,G^3_{\mu\nu}(s,t)\tilde{\phi}_\mu\tilde{\phi}_\nu \qquad \beta^{2,\text{in}}(t) \sim \mathcal{N}(0, G^3(t,t))\\
z^3_\mu(t) &= \beta^3(t) + \frac{\gamma_0\eta_0}{P}\int_0^t ds \sum_\nu \Delta_\nu h^3_\nu(s) \qquad \beta^3 \sim \mathcal{N}(0,1),
\end{aligned}
\tag{N.74}
$$

O DMFT Analysis for Regime III (μP)

In this section we are considering the case in which $M, N, N_e \to \infty$ at fixed ratios. Specifically, we define $\kappa = \frac{M}{N}, \iota = \frac{N_e}{N}$, where κ,ι are order one in N. The forward pass, following μP is defined as:

$$
\begin{aligned}
\mathbb{R}^N \ni h^1_\mu &= \frac{1}{\sqrt{D}} W^1 x_\mu \\
\mathbb{R}^M \ni \psi_\mu &= \frac{1}{\sqrt{N}} Q \sigma(h^1_\mu) \\
\mathbb{R}^M \ni \phi_\mu &= \text{softmax}(\psi_\mu) \\
\mathbb{R}^{N_e} \ni h^{2,\text{in}}_{\mu,i} &= \frac{1}{\sqrt{N}} W^{2,\text{in}}_i \sigma(h^1_\mu), \\
\mathbb{R}^N \ni h^{2,\text{out}}_{\mu,i} &= \frac{1}{\sqrt{N}} W^{2,\text{out}}_i \sigma(h^{2,\text{in}}_{\mu,i}), \\
\mathbb{R}^N \ni h^3_\mu &= \sum_{i=1}^{M} \phi^i_\mu h^{2,\text{out}}_{\mu,i} \\
\mathbb{R} \ni h^4_\mu &= \frac{1}{\sqrt{N}} w^{4\intercal} h^3_\mu \\
f_\mu &= \frac{1}{\gamma} h^4_\mu \\
\eta &= \eta_0 \gamma^2 \\
\eta_E &= \eta_0 \gamma^2 N \\
\eta_Q &= \eta_0 \gamma^2 \kappa \\
\gamma &= \gamma_0 \sqrt{N} \\
\eta_0, \gamma_0 &\sim O(1) \\
w^4_\alpha(0), \left[W^{2,\text{out}}_i(0)\right]_{\alpha\beta}, \left[W^{2,\text{in}}_i(0)\right]_{\alpha\beta}, W^1_{\alpha\beta}(0), Q_{\alpha\beta}(0) &\sim \mathcal{N}(0,1)
\end{aligned}
\tag{O.1}
$$

Again, we let $\boldsymbol{\theta} = \text{Vec}\{W^1, W^{2,\text{in}}_i, W^{2,\text{out}}_i, w^4, Q\}$, and define

$$
\tilde{\phi}^i_\mu := \kappa N \phi^i_\mu \tag{O.2}
$$

Following an analogous derivation as in Appendix L, which we omit for conciseness, we arrive at the following self-consistent set of DMFT equations.

$$\alpha^1(t) \sim \mathcal{N}(0, \Phi^0) \qquad \beta^3(t) \sim \mathcal{N}(0, \mathbb{I}) \qquad \beta^{1,\phi}(t) \sim \mathcal{GP}(0, G^{1,\phi})$$
$$\alpha^\phi(t) \sim \mathcal{GP}(0, \Phi^1) \qquad \alpha^{2,\text{in}}(t) \sim \mathcal{GP}(0, \Phi^1) \qquad \beta^{2,\text{in}}(t) \sim \mathcal{GP}(0, \tilde{\phi}\,\tilde{\phi}\, G^3)$$

$$\Phi^0_{\mu\nu} = \frac{1}{D} x_\mu x_\nu^\top, \quad \Phi^1_{\mu\nu}(t,s) = \langle \sigma(h^1_\mu(t)) \sigma(h^1_\nu(s)) \rangle_{\mathcal{Z}_N^{global}}, \quad \Phi^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle \sigma(h^{2,\text{in}}_\mu(t))\, \sigma(h^{2,\text{in}}_\nu(s)) \right\rangle_{\mathcal{Z}_{N_e}|\mathcal{K}_{global}}$$

$$\bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle \tilde{\phi}_\mu(s)\, \tilde{\phi}_\nu(t) \Phi^{2,\text{in}}_{\mu\nu}(s,t) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}} \qquad \bar{G}^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle G^{2,\text{in}}_{\mu\nu}(s,t) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}$$

$$\Phi^3_{\mu\nu}(s,t) = \left\langle h^3_\mu(t) h^3_\nu(s) \right\rangle_{\mathcal{Z}_N^{global}}, \quad G^1_{\mu\nu}(t,s) = \langle [\dot{\sigma}(h^1_\mu(t)) \odot z^1_\mu(t)] [\dot{\sigma}(h^1_\nu(s)) \odot z^1_\nu(s)] \rangle_{\mathcal{Z}_N^{global}}$$

$$G^{2,\text{in}}_{\mu\nu}(s,t) = \left\langle \left[\dot{\sigma}(h^{2,\text{in}}_\mu(t)) \odot z^{2,\text{in}}_\mu(t)\right] \left[\dot{\sigma}(h^{2,\text{in}}_\nu(s)) \odot z^{2,\text{in}}_\nu(s)\right] \right\rangle_{\mathcal{Z}_{N_e}|\mathcal{K}_{global}}$$

$$G^{1,\phi}_{\mu\nu}(s,t) = \left\langle g^{1,\phi}_\mu(s)\, g^{1,\phi}_\nu(t) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}, \quad G^3_{\mu\nu}(s,t) = \left\langle z^3_\mu(s)\, z^3_\nu(t) \right\rangle_{\mathcal{Z}_N^{global}}$$

$$A^{2,\text{in}}_{\mu\nu}(t,s) = \left\langle \frac{\partial \sigma(h^{2,\text{in}}_\mu(t))}{\partial \beta^{2,\text{in}\top}_\nu(s)} \right\rangle_{\mathcal{Z}_{N_e}|\mathcal{K}_{global}}, \qquad \bar{A}^{2,\text{in}}_{\mu\nu}(t,s) = \frac{\kappa}{\iota} \left\langle \tilde{\phi}_\mu(t) \tilde{\phi}_\nu(s) A^{2,\text{in}}_{\mu\nu}(t,s) \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}$$

$$A^{1,\phi}_{\mu\nu}(t,s) = \left\langle \frac{\partial \sigma(h^1_\mu(t))}{\partial \beta^{1,\phi\top}_\nu(s)} \right\rangle_{\mathcal{Z}_N^{global}}, B^{1,\phi}_{\mu\nu}(t,s) = \left\langle \frac{\partial g^{1,\phi}_\mu(t))}{\partial \alpha^{\phi\top}_\nu(s)} \right\rangle_{\mathcal{Z}_M|\mathcal{K}_{global}} . \bar{B}^1_{\mu\nu}(s,t) = \left\langle \frac{\partial g^{2,\text{in}}_\mu(t))}{\partial \alpha^{2,\text{in}\top}_\nu(s)} \right\rangle_{\mathcal{Z}_{N_e}|\mathcal{K}_{global}}$$

$$g^{1,\phi}_\mu(t) = \frac{\iota \eta_0 \gamma_0}{\kappa P} \int_0^t ds \sum_\nu \Delta_\nu(s) G^3_{\mu\nu}(s,t) \left[\Phi^{2,in}_{\mu\nu}(s,t) \tilde{\phi}_\nu(t) - \bar{\Phi}^{2,in}_{\mu\nu}(s,t)\right] \tilde{\phi}_\mu(s)$$

$$z^1_\mu(t) = \beta^{1,\phi}_\mu(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \left\{ \bar{B}^1_{\nu\mu}(s,t) + B^{1,\phi}_{\nu\mu}(s,t) + \Delta_\nu(s) \left[\frac{\iota}{\kappa} \bar{G}^{2,\text{in}}_{\mu\nu}(s,t) + G^{1,\phi}_{\mu\nu}(s,t)\right] \right\} \sigma(h^1_\nu(s))$$

$$z^{2,\text{in}}_\mu(t) = \beta^{2,\text{in}}_\mu(t) + \frac{\eta_0 \gamma_0}{P \kappa} \int_0^t ds \sum_\nu \Delta_\nu(s) G^3_{\mu\nu}(s,t)\, \tilde{\phi}_\mu(s)\, \tilde{\phi}_\nu(t)\, \sigma\left(h^{2,\text{in}}_\nu(s)\right), \quad j \in \{1, 2, ..., N_e\}$$

$$z^3_\mu(t) = \beta^3_\mu(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu(s)\, h^3_\nu(s),$$

$$h^1_\mu(t) = \alpha^1_\mu(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \Delta_\nu\, \Phi^0_{\mu\nu}\, (z^1_\mu(s) \cdot \dot{\sigma}(h^1_\mu(s)))$$

$$h^{2,\text{in}}_\mu(t) = \alpha^{2,\text{in}}_\mu(t) + \frac{\iota \eta_0 \gamma_0}{\kappa P} \int_0^t ds \sum_\nu \Delta_\nu(s) \Phi^1_{\mu\nu}(t,s)\, \left(\dot{\sigma}\left(h^{2,\text{in}}_\mu(t)\right) \odot z^{2,\text{in}}_\nu(s)\right), \quad j \in \{1, 2, ..., N_e\}$$

$$h^3_\mu(t) = \frac{\iota \eta_0 \gamma_0}{\kappa P} \int_0^t ds \sum_\nu \left[\bar{A}^{2,\text{in}}_{\mu\nu}(t,s) + \Delta_\nu(s) \bar{\Phi}^{2,\text{in}}_{\mu\nu}(s,t)\right] z^3_\nu(s)$$

$$\psi_\mu(t) = \alpha^\phi_\mu(t) + \frac{\eta_0 \gamma_0}{P} \int_0^t ds \sum_\nu \left[A^{1,\phi}_{\mu\nu}(s,t) + \Delta_\nu(s)\, \Phi^1_{\mu\nu}(s,t)\right] g^{1,\phi}_\nu(t)$$

$$\tilde{\phi}_\mu(t) = \frac{e^{\hat{\psi}_\mu(t)}}{\mathcal{S}_\mu(t)}, \qquad \hat{\psi}_\mu(t) = m_\mu(t) \psi_\mu(t), \qquad m_\mu(t) = \mathbf{1}(\psi_\mu(t) - \tau_\mu(t))$$

where $\tau_\mu(t)$ is chosen such that $\rho = \langle m_\mu(t) \rangle_{\mathcal{Z}_M|\mathcal{K}_{global}}$

$$\frac{df_\mu(t)}{dt} = \frac{\eta_0}{P} \sum_\nu \Delta_\nu \Bigg[G^1_{\mu\nu}(t,t) \Phi^0_{\mu\nu}(t,t) + \left[\frac{\iota}{\kappa} \bar{G}^{2,\text{in}}_{\mu\nu}(t,t) + G^{1,\phi}_{\mu\nu}(t,t)\right] \Phi^1_{\mu\nu}(t,t)$$
$$+ \frac{\iota}{\kappa} G^3_{\mu\nu}(t,t) \bar{\Phi}^{2,\text{in}}_{\mu\nu}(t,t) + \Phi^3_{\mu\nu}(t,t) \Bigg]$$

Comparison to μP. Limiting dynamics under μP admit a simpler mean-field structure than under MSSP. Under μP, the expert weights are initialized independently across experts. Consequently, there is no expert-hidden disorder shared by the expert population, and the limiting theory does not require a separate shared expert-hidden single-site process. The expert averages self-average directly over independently initialized experts. The corresponding DMFT is provided in App.O.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we tried to summarize and contextualize our contributions while properly acknowledging related work to the best of our ability.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly state the caveats of every reported result. We mostly discuss limitations and avenues for future work in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We detail all assumptions and the full theoretical analyses in the Theory part of the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We detail all necessary information to reproduce our experiments in the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: Upon acceptance, we plan to make open source code to reproduce all experiments publicly available on GitHub.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail all necessary information in the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: If not stated otherwise, all experiments include 2σ error bars from 4 random seeds (affecting initial weights and data shuffling). Due to computational constraints, we only train Transformer MoEs on several random seeds for a subset of the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: A single run at the largest GPT MoE model size finishes within less than 3 hours on a 8xH100 node. A GPT MoE coordinate check across widths using 3 random seeds takes around 24 hours on a 8xH100 node due to fine-grained logging.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research in this paper conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper aims to advance the understanding of deep learning practice. We do not foresee any societal impacts beyond the dual use considerations that apply to the whole field of machine learning research.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper studies pre-existing training procedures on established datasets. We do not release any data or high-risk models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the standard TinyImageNet (Le et al., 2015) and Dolma3 (Olmo et al., 2025) datasets following the standard practice. We also cite the Python assets PyTorch (Paszke et al., 2019) and OlMoE (Muennighoff et al., 2024) that we use as a basis for our experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.