UNDERSTANDING THE WEIGHT AVERAGING MECHANISM IN LLM TRAINING FOR POST-TRAINING QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are typically pretrained in high precision but increasingly deployed with low-precision post-training quantization (PTQ). Recent studies have shown that using weight averaging during pretraining can improve PTQ performance compared with learning-rate decay, suggesting that it might provide a simple way to improve the pretraining-to-quantization transition. But the mechanism behind weight averaging remains insufficiently explained. This leads to inconsistent and fragile performance gains, thereby preventing practitioners from applying such a technique confidently. As a response, we formulate weight averaging as a trade-off between retaining training progress and improving robustness under perturbation. We further derive a continuous family of averaging kernels that unifies conventional strategies and achieves the Pareto frontier between the two competing goals. Critically, a theoretical framework for performing weight averaging under PTQ is developed. It can be shown that coarser quantization is more susceptible to perturbations, whereas finer quantization could be less affected. Thus, our results could provide unified theoretical guidance for performing weight averaging under different PTQ conditions. Experiments validate both the predicted behavior and the proposed averaging strategy.

1 INTRODUCTION

Large language models (LLMs) are typically pretrained at relatively high numerical precision (Workshop et al., 2022; Team et al., 2025a;b). As model scale continues to grow, post-training quantization (PTQ) has become increasingly important for reducing deployment costs by converting pretrained models to lower precision (Xiao et al., 2023; Shao et al., 2026; Li et al., 2025). This mismatch between training and deployment raises a key challenge at the training-to-quantization transition: better performance before quantization does not necessarily translate into better performance afterward (Ouyang et al., 2025; Kumar et al., 2025; Chen et al., 2026; Catalan-Tatjer et al., 2026).

Prior work has shown that training dynamics affect both training performance and sensitivity to quantization (Ahmadian et al., 2023; Wang et al., 2025; Catalan-Tatjer et al., 2026; Luo et al., 2026). Late-stage learning-rate (LR) decay can further reduce training loss, but may move the model into sharper regions that are more sensitive (Watts et al., 2026; Microsoft AI Team, 2026; Yano et al., 2026). By contrast, under a stable LR, after completing T training steps, one can retain the last M weights, $\boldsymbol{\theta}_T, \ldots, \boldsymbol{\theta}_{T-M+1}$, where M denotes the number of retained weights. We index them by their age $d \in \{0, \ldots, M-1\}$, with $d = 0$ corresponding to the most recent weight, and average them as

$$\boldsymbol{\theta}_T^{\boldsymbol{w}} = \sum_{d=0}^{M-1} w_d \boldsymbol{\theta}_{T-d}, \qquad w_d \geq 0, \quad \sum_{d=0}^{M-1} w_d = 1, \tag{1}$$

where $\boldsymbol{w} = (w_0, \ldots, w_{M-1})$ is the averaging kernel, with different choices corresponding to different averaging strategies. Recent work shows that such weight averaging can achieve high-precision performance comparable to LR decay while improving PTQ performance (Sanyal et al., 2023; Sandler et al., 2023; Tian et al., 2026). However, existing evidence for the benefits of weight averaging at the pretraining-to-quantization transition remains largely empirical (Ajroldi et al., 2025; Catalan-Tatjer et al., 2026; Li et al., 2026c), and a theoretical understanding of how the averaging strategy

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

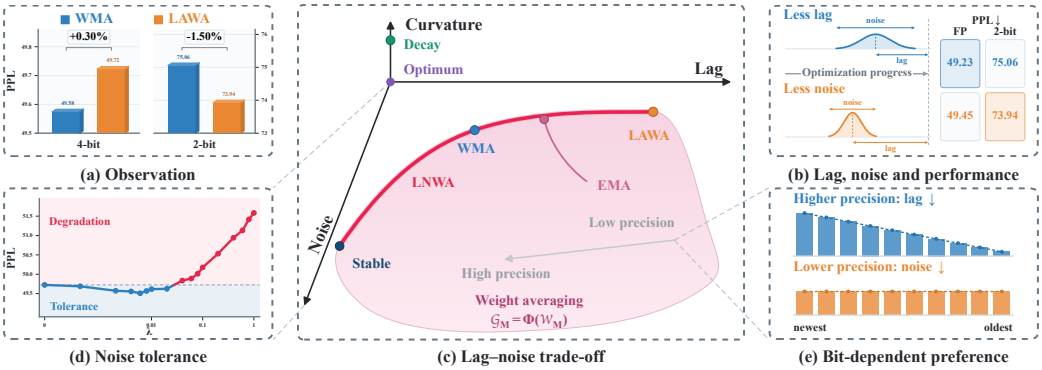


Figure 1: Overview of our work. (a) Different averaging strategies' rankings reverse between 4-bit and 2-bit PTQ. (b) Less lag favors better current performance (FP), while less noise favors greater robustness to quantization (2-bit). (c) We formulate weight averaging as a lag–noise trade-off and derive its Pareto frontier, LNWA. (d) Reduced lag can offset additional noise within a finite tolerance interval; beyond it, further noise leads to degradation. (e) The preferred balance reverses with precision, with lower precision favoring less noise and higher precision favoring less lag.

affects PTQ is still lacking. This gap leaves unclear why weight averaging benefits PTQ and when a particular averaging kernel is effective. As a consequence, blindly applying any weight averaging strategy now cannot guarantee improved downstream PTQ performance; and in some cases, it might be even worse. This is exemplified by our discovery that the same averaging strategy could outperform another at higher precision but instead fall behind at lower precision. Specifically, Fig. 1(a) compares linearly weighted WMA with uniformly weighted LAWA on OPT-125M, showing that their performance ranking reverses between 4-bit and 2-bit PTQ. In short, favorable empirical findings may not generalize across settings, and extensive trial-and-error is usually required. This strongly motivates a unified theoretical framework to understand how weight averaging in pretraining affects later quantization, thereby providing principled guidance for averaging strategy selection across settings.

Building upon prior art, we believe it is useful to evaluate a training trajectory through its genuine progress towards improved model abilities and its robustness to different kinds of noise and perturbations (Morales-Brotons et al., 2024; Block & Zhang, 2025). Specifically, we can decompose each retained weight as $\boldsymbol{\theta}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t$ at training step t, where the latent mean state $\boldsymbol{\mu}_t$ denotes the expectation of $\boldsymbol{\theta}_t$ across repeated training runs over training randomness, with $\boldsymbol{\mu}_T$ serving as an idealized noise-free final model, and $\boldsymbol{\epsilon}_t$ denotes the corresponding zero-mean fluctuation arising from mini-batch sampling, data ordering, and other sources of training randomness. Applying weight averaging in Eq. (1) to the decomposition yields

$$\boldsymbol{\theta}_T^{\boldsymbol{w}} - \boldsymbol{\mu}_T = -\underbrace{\left(\boldsymbol{\mu}_T - \sum_{d=0}^{M-1} w_d \boldsymbol{\mu}_{T-d}\right)}_{\text{temporal lag}} + \underbrace{\sum_{d=0}^{M-1} w_d \boldsymbol{\epsilon}_{T-d}}_{\text{spatial noise}}. \tag{2}$$

Thus, the deviation of the averaged model from the idealized final model $\boldsymbol{\mu}_T$ decomposes into temporal lag and spatial noise. Temporal lag captures the genuine progress sacrificed by weight averaging, which can impair the model's performance at the transition to PTQ. At the same time, weight averaging suppresses spatial noise by reducing the variance of stochastic fluctuations, thereby improving robustness to PTQ, as illustrated in Fig. 1(b). The averaging kernel thus determines the balance between these two effects. We then derive Lag–Noise Trade-off Weight Averaging (LNWA), a closed-form kernel family that parameterizes the Pareto frontier of the lag–noise trade-off and includes classical averaging kernels as special cases, as shown in Fig. 1(c).

Building on this formulation, we develop a theoretical framework for weight averaging strategy selection under PTQ. We first study how moving along the lag–noise frontier affects PTQ performance at a fixed precision, quantifying how strongly the temporal lag could be reduced before PTQ performance

degrades (Fig. 1(d)). Within this tolerance region, we then vary the quantization precision and show that lower precision favors less spatial noise, whereas higher precision favors less temporal lag (Fig. 1(e)); we further characterize the quantization precision threshold at which this preference reverses. Together, we extend conventional method-specific weight averaging selection into a unified theoretical framework under different PTQ conditions. Finally, experiments on LLMs from 125M to 1.3B parameters validate our theory. Our contributions can be summarized as follows:

1. We establish a theoretical framework for understanding weight averaging through a two-way decomposition in Eq. (2). Specifically, we separate its effects into temporal lag, which captures inherent training progress, and spatial noise, which captures the solution's robustness to perturbations. We show that these two objectives are in competitive tension, and a careful trade-off is necessary.
2. A new class of weight averaging strategies called LNWA is proposed based on the aforementioned framework. It operates on the Pareto frontier between temporal lag and spatial noise (Theorem 4.1) and smoothly interpolates between them. This class generalizes existing strategies (see Table 1) and can achieve any desired Pareto-optimal trade-off between the two objectives.
3. We further provide a theoretical analysis of how the lag–noise balance affects PTQ performance (Sections 4.2 and 4.3). We show that this balance depends critically on quantization precision (Theorems 4.2 and 4.3) and therefore should not be selected universally. These results provide principled guidance for performing weight averaging across PTQ settings.

2 Related Work

Learning-rate decay. Late-stage LR decay is widely used in LLM pretraining to further improve final performance (Hu et al., 2024; Hägele et al., 2024). Prior work has mainly focused on understanding why LR decay benefits pretraining (Wen et al., 2025; Li et al., 2026a;b; Liu et al., 2026). However, recent studies show that LR decay can also affect how well the resulting model transfers to subsequent stages (Chen et al., 2026). Similar effects have been observed across various downstream tasks (Wang et al., 2025; Yano et al., 2026; Microsoft AI Team, 2026; Catalan-Tatjer et al., 2026), suggesting that lower final pretraining loss does not necessarily imply a better transition to downstream adaptation.

Post-training quantization. PTQ converts a model to lower precision, reducing storage and inference costs (Chee et al., 2023). Various PTQ methods have been proposed (Nagel et al., 2020; Lin et al., 2024; Ashkboos et al., 2024; Dettmers et al., 2022; Egiazarian et al., 2024), and error compensation is one of the most well-established approaches (Yao et al., 2024; Zhang et al., 2025), with GPTQ as a widely used representative (Frantar et al., 2022). GPTQ reduces layerwise reconstruction error through second-order updates to the remaining unquantized weights (Frantar et al., 2022). Recent work has further improved PTQ through a range of transformation, clipping, and reconstruction techniques (Liu et al., 2025; Sun et al., 2024; Shao et al., 2024; Tseng et al., 2024; Zhang et al., 2024).

Weight averaging. Classical Polyak–Ruppert averaging established the variance-reduction benefit of averaging iterates (Polyak & Juditsky, 1992). Motivated by it, stochastic weight averaging (SWA) showed that averaging weights along the training trajectory can lead to wider optima and better generalization (Izmailov et al., 2018). Moreover, LAWA (Sanyal et al., 2023) uniformly averages recent weights sparsely, while EMA (Morales-Brotons et al., 2024) and WMA (Li et al., 2026c) assign larger coefficients to more recent weights through exponential and linear weighting, respectively. Recent work has applied LAWA to the pretraining-to-quantization transition, showing that the averaged models can match or outperform LR decay after PTQ (Catalan-Tatjer et al., 2026).

3 Preliminaries

Notation. Scalars are denoted by italic letters (e.g., f), vectors by bold lowercase letters (e.g., $\boldsymbol{f}$), matrices by bold uppercase letters (e.g., $\boldsymbol{F}$), and sets by calligraphic uppercase letters (e.g., $\mathcal{F}$).

Consider a fixed window of retained model weights $\boldsymbol{\theta}_T, \ldots, \boldsymbol{\theta}_{T-M+1} \in \mathbb{R}^P$ ($M \geq 2$), where T is the training steps, P is the number of parameters and age d counts backward from the newest retained weight at $d = 0$. Let $\boldsymbol{w} = (w_0, \ldots, w_{M-1})$ denote the averaging kernel, where w_d is the averaging

coefficient assigned to the weight $\boldsymbol{\theta}_{T-d}$. Its feasible set $\mathcal{W}_M$ and the resulting averaged model $\boldsymbol{\theta}_T^{\boldsymbol{w}}$ are defined as

$$\mathcal{W}_M = \left\{ \boldsymbol{w} \in \mathbb{R}^M : w_d \geq 0,\ \sum_{d=0}^{M-1} w_d = 1 \right\}, \qquad \boldsymbol{\theta}_T^{\boldsymbol{w}} = \sum_{d=0}^{M-1} w_d \boldsymbol{\theta}_{T-d}. \tag{3}$$

Thus $\boldsymbol{\theta}_T^{\boldsymbol{w}}$ lies in the convex hull of the retained weights. We first define the latent mean $\boldsymbol{\mu}_t$, which provides an idealized noise-free reference to separate training progress from spatial noise.

Definition 3.1 (Latent-mean Trajectory)**.** Let $\mathcal{F}_0$ be the sigma-field generated by the information shared across stochastic training runs, such as the model configuration, common initialization, and training setup. Let ξ index the run-specific stochasticity, such as data ordering and dropout. For the aligned state $\boldsymbol{\theta}_{T-d}(\xi)$, assume $\mathbb{E}[\|\boldsymbol{\theta}_{T-d}(\xi)\|_2^2 \mid \mathcal{F}_0] < \infty$. The conditional mean state $\boldsymbol{\mu}_t$ and the zero-mean spatial noise $\boldsymbol{\epsilon}_t$ at the training step t of an individual run are

$$\boldsymbol{\mu}_t := \mathbb{E}_\xi[\boldsymbol{\theta}_t(\xi) \mid \mathcal{F}_0], \qquad \boldsymbol{\epsilon}_t(\xi) := \boldsymbol{\theta}_t(\xi) - \boldsymbol{\mu}_t. \tag{4}$$

For an averaging kernel $\boldsymbol{w}$, write $\boldsymbol{\mu}_T^{\boldsymbol{w}} := \sum_{d=0}^{M-1} w_d \boldsymbol{\mu}_{T-d}$ for the mean of $\boldsymbol{\theta}_T^{\boldsymbol{w}}$, and write $\boldsymbol{\epsilon}_T^{\boldsymbol{w}} := \sum_d w_d \boldsymbol{\epsilon}_{T-d}$ for the averaged spatial noise.

Besides, we model how PTQ reduces the reconstruction error. Let $\boldsymbol{\theta}^{(\ell,i)} \in \mathbb{R}^{d_{\text{in}}}$ denote the parameter of output channel i in layer ℓ, extracted from $\boldsymbol{\theta}$. Quantization changes $\boldsymbol{\theta}^{(\ell,i)}$ by $\Delta\boldsymbol{\theta}^{(\ell,i)}$, so the output error is $\left(\Delta\boldsymbol{\theta}^{(\ell,i)}\right)^\top \boldsymbol{X}$ for input activations $\boldsymbol{X} \in \mathbb{R}^{d_{\text{in}} \times n}$. Its mean squared reconstruction error is

$$\frac{1}{n}\|\left(\Delta\boldsymbol{\theta}^{(\ell,i)}\right)^\top \boldsymbol{X}\|_2^2 = \left(\Delta\boldsymbol{\theta}^{(\ell,i)}\right)^\top \boldsymbol{G}^{(\ell,i)} \Delta\boldsymbol{\theta}^{(\ell,i)} = \|\Delta\boldsymbol{\theta}^{(\ell,i)}\|_{\boldsymbol{G}^{(\ell,i)}}^2, \tag{5}$$

where $\boldsymbol{G}^{(\ell,i)} := \boldsymbol{X}\boldsymbol{X}^\top/n$. GPTQ (Frantar et al., 2022) reduces errors by adjusting parameters that remain unquantized. We represent this procedure with two parameter groups: quantize the first group, use the second group to reduce the resulting output error, and then quantize the second group. For a parameter vector $\boldsymbol{z}$, $\text{quant}_b(\boldsymbol{z})$ denotes scale-and-zero-point quantization at precision b using round.

Definition 3.2 (Simplified GPTQ Model)**.** Partition $\boldsymbol{\theta}^{(\ell,i)}$ into $\boldsymbol{\theta}_{\text{f}}^{(\ell,i)} \in \mathbb{R}^{d_{\text{f}}}$ and $\boldsymbol{\theta}_{\text{r}}^{(\ell,i)} \in \mathbb{R}^{d_{\text{r}}}$, where $d_{\text{f}} + d_{\text{r}} = d_{\text{in}}$ and $d_{\text{f}}, d_{\text{r}} \geq 1$. Let $\boldsymbol{X}_{\text{f}}, \boldsymbol{X}_{\text{r}}$ be the corresponding row blocks of $\boldsymbol{X}$, and partition $\boldsymbol{G}^{(\ell,i)}$ accordingly. First quantize $\boldsymbol{q}_{\text{f}}^{(\ell,i)} = \text{quant}_b(\boldsymbol{\theta}_{\text{f}}^{(\ell,i)})$, with quantization error $\boldsymbol{e}_{\text{f}}^{(\ell,i)} := \boldsymbol{q}_{\text{f}}^{(\ell,i)} - \boldsymbol{\theta}_{\text{f}}^{(\ell,i)}$. Then use the remaining group to minimize the reconstruction error caused by $\boldsymbol{e}_{\text{f}}^{(\ell,i)}$:

$$\boldsymbol{c}_{\text{r}}^{(\ell,i),\star} = \underset{\boldsymbol{c}_{\text{r}} \in \mathbb{R}^{d_{\text{r}}}}{\text{argmin}} \frac{1}{n} \left\| \left(\boldsymbol{e}_{\text{f}}^{(\ell,i)}\right)^\top \boldsymbol{X}_{\text{f}} + \boldsymbol{c}_{\text{r}}^\top \boldsymbol{X}_{\text{r}} \right\|_2^2. \tag{6}$$

Then quantize the corrected remaining group:

$$\boldsymbol{q}_{\text{r}}^{(\ell,i)} = \text{quant}_b\left(\boldsymbol{\theta}_{\text{r}}^{(\ell,i)} + \boldsymbol{c}_{\text{r}}^{(\ell,i),\star}\right).$$

The quantized $\boldsymbol{\theta}^{(\ell,i)}$ is $\boldsymbol{q}^{(\ell,i)} = ((\boldsymbol{q}_{\text{f}}^{(\ell,i)})^\top, (\boldsymbol{q}_{\text{r}}^{(\ell,i)})^\top)^\top$. At precision b, applying these steps to every output channel of every linear layer gives the full-model simplified GPTQ $Q_b(\boldsymbol{\theta})$ (see Appendix A.5).

4 Lag–Noise Trade-off

In this section, we formulate weight averaging as a trade-off between temporal lag and spatial noise and derive its Pareto frontier (Section 4.1). We then connect lag and noise to PTQ and further characterize the tolerance-to-degradation transition at a fixed precision (Section 4.2) and the precision-dependent shift in the preferred lag–noise balance (Section 4.3).

4.1 Lag–noise decomposition and the LNWA frontier

Using the latent-mean trajectory in Definition 3.1, a retained weight can be written as $\boldsymbol{\theta}_{T-d} = \boldsymbol{\mu}_{T-d} + \boldsymbol{\epsilon}_{T-d}$. As illustrated in Fig. 2(a), across different data orders, the individual training trajectories fluctuate around a shared latent-mean trajectory. The latent mean captures the progress that is shared across runs, while the difference between an individual trajectory and the latent mean

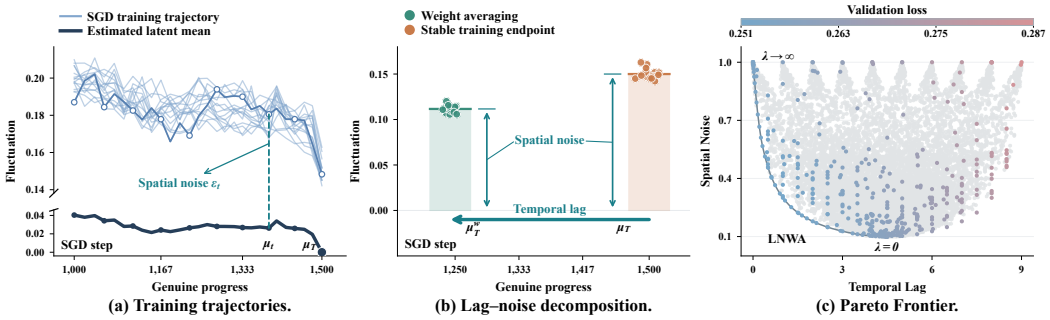


Figure 2: Illustrative visualizations with a two-layer MLP on MNIST. (a) SGD training trajectories with different data orders and their estimated latent mean. (b) Weight averaging reduces spatial noise but introduces temporal lag compared with the stable training endpoint. (c) LNWA characterizes the Pareto frontier between temporal lag and spatial noise. Colors indicate validation loss.

captures the fluctuation caused by run-specific randomness. Weight averaging therefore acts on both components. Let $\boldsymbol{v}_t := \boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}, \quad T - M + 2 \le t \le T$, denote the latent progress between adjacent retained weights. To quantify the effects of temporal lag and spatial noise in Eq. (2), define

$$m(\boldsymbol{w}) := \sum_{d=0}^{M-1} dw_d, \qquad R(\boldsymbol{w}) := \sum_{d=0}^{M-1} w_d^2.$$

Here, $m(\boldsymbol{w})$ is the mean age of the averaged weights, which measures temporal lag; $R(\boldsymbol{w})$ is the squared ℓ_2 norm of the averaging kernel, which determines the amount of spatial-noise variance retained after averaging. Lemma A.2 then gives

$$\boldsymbol{\theta}_T^{\boldsymbol{w}} - \boldsymbol{\mu}_T = -m(\boldsymbol{w})\boldsymbol{v}_T + \boldsymbol{\epsilon}_T^{\boldsymbol{w}} + \boldsymbol{e}_{\boldsymbol{w}}^{\mathrm{traj}}, \tag{7}$$

where $\boldsymbol{e}_{\boldsymbol{w}}^{\mathrm{traj}}$ is a bounded trajectory remainder, and $\mathrm{Cov}(\boldsymbol{\epsilon}_T^{\boldsymbol{w}} \mid \mathcal{F}_0) = R(\boldsymbol{w})\boldsymbol{\Sigma}$. Fig. 2(b) visualizes how averaging acts on spatial noise and temporal lag. Weight averaging reduces the fluctuation across training runs (spatial noise), while shifting the averaged latent mean $\boldsymbol{\mu}_T^{\boldsymbol{w}}$ away from the latest latent mean $\boldsymbol{\mu}_T$ along the direction of genuine progress (temporal lag). These two effects therefore form a fundamental trade-off in weight averaging. We formalize this trade-off by minimizing retained spatial noise under a fixed temporal-lag budget (Lemma A.3). As the budget varies, the resulting solutions trace the lag–noise Pareto frontier. Thus, we can parameterize it with a one-parameter family of averaging kernels termed Lag–Noise Trade-off Weight Averaging (LNWA).

Theorem 4.1 (LNWA Kernel)**.** *For $0 \le \lambda < \infty$, let*

$$\boldsymbol{w}_\lambda := \arg\min_{\boldsymbol{w} \in \mathcal{W}_M} \left\{ R(\boldsymbol{w}) + \lambda m(\boldsymbol{w})^2 \right\}. \tag{8}$$

For each finite λ, the unique minimizer has an active prefix $S_q = \{0, \dots, q-1\}$ with $2 \le q \le M$, and its coefficients are

$$(\boldsymbol{w}_\lambda)_d = \begin{cases} q^{-1} + \lambda m_q(\lambda)(\bar{d}_q - d), & 0 \le d < q, \\ 0, & q \le d < M, \end{cases} \tag{9}$$

where $m_q(\lambda) = \bar{d}_q/(1 + \lambda D_q)$, $\bar{d}_q = (q-1)/2$, and $D_q = q(q^2-1)/12$. Define $\boldsymbol{w}_\infty := \lim_{\lambda \to \infty} \boldsymbol{w}_\lambda = \boldsymbol{e}_0 = (1, 0, \dots, 0)^\top$, the Stable endpoint. The extended family $\{\boldsymbol{w}_\lambda\}_{0 \le \lambda \le \infty}$ parameterizes the complete Pareto frontier between temporal lag and spatial noise.

In particular, for $0 \le \lambda_1 < \lambda_2 \le \infty$,

$$m(\boldsymbol{w}_{\lambda_1}) > m(\boldsymbol{w}_{\lambda_2}), \qquad R(\boldsymbol{w}_{\lambda_1}) < R(\boldsymbol{w}_{\lambda_2}). \tag{10}$$

Thus, increasing λ moves the averaging kernel along the frontier toward larger spatial noise, while decreasing λ moves it toward larger temporal lag. Lemma A.4 characterizes how the number of retained weights assigned nonzero averaging coefficients changes with λ, while Lemma A.5 gives explicit formulas for the corresponding temporal lag and spatial noise. Fig. 2(c) visualizes the feasible

Table 1: Relation between classical averaging kernels and LNWA. LAWA and Stable are the two endpoints, and WMA is an exact interior point. EMA is exact when $M = 2$. For $M > 2$, λ_{EMA} gives a local approximation with error $\mathcal{O}(\eta^2 M)$, where $\eta = -\log\beta$.

Methods	Kernel w_d	Lag $m(w)$	Noise $R(w)$	Exact match
LAWA	$1/M$	$(M-1)/2$	$1/M$	✓
WMA	$2(M-d)/[M(M+1)]$	$(M-1)/3$	$2(2M+1)/[3M(M+1)]$	✓
EMA	$(1-\beta)\beta^d/(1-\beta^M)$	$\dfrac{\beta - M\beta^M + (M-1)\beta^{M+1}}{(1-\beta)(1-\beta^M)}$	$\dfrac{(1-\beta)(1+\beta^M)}{(1+\beta)(1-\beta^M)}$	$M=2$: ✓ $M>2$: $\mathcal{O}(\eta^2 M)$
Stable	$\boldsymbol{e}_0 = (1, 0, \ldots, 0)$	0	1	✓

lag–noise region induced by different averaging kernels, where LNWA traces its Pareto boundary. The distribution of validation loss across this region shows that the optimal averaging strategy depends on how these two factors are balanced. Several classical averaging kernels can also be viewed as special cases of the LNWA family: LAWA and Stable form its two endpoints (Proposition A.1); WMA is an exact interior kernel (Proposition A.2); and EMA admits an exact representation for $M = 2$ and a local approximation for $M > 2$ (Proposition A.4). Table 1 summarizes the lag–noise coordinates of these classical kernels.

4.2 Noise tolerance and quantitative degradation

As discussed above, increasing λ along the LNWA frontier reduces temporal lag while increasing spatial noise. In this section, we quantify how much lag reduction can compensate for the accompanying increase in noise and thereby improve PTQ performance at a fixed precision.

We measure the effect of an averaging kernel by the expected output reconstruction error (defined in Eq. (5)) of the quantized model (defined in Definition 3.2) relative to the latest noise-free state $\boldsymbol{\mu}_T$:

$$\mathcal{E}_b(\boldsymbol{w}) := \frac{1}{2}\mathbb{E}\Big[\big\|Q_b(\boldsymbol{\theta}_T^{\boldsymbol{w}}) - \boldsymbol{\mu}_T\big\|_{\boldsymbol{G}}^2\Big] = \mathcal{P}_0(\boldsymbol{w}) + \gamma_b \mathcal{Q}(\boldsymbol{w}). \tag{11}$$

Here $\mathcal{P}_0$ is the error before quantization, $\gamma_b\mathcal{Q}$ is the reconstruction error introduced by quantization, and γ_b is a precision-dependent coefficient. Lemma A.9 proves this decomposition. Next, we need to show that both $\mathcal{P}_0$ and $\mathcal{Q}$ depend on temporal lag and spatial noise. Before quantization, these two effects are readily separated:

$$\mathcal{P}_0(\boldsymbol{w}) := \frac{1}{2}\mathbb{E}\big[\|\boldsymbol{\theta}_T^{\boldsymbol{w}} - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2\big] = \frac{1}{2}\|\boldsymbol{\mu}_T^{\boldsymbol{w}} - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2 + N_G R(\boldsymbol{w}), \qquad N_G := \frac{\sigma^2}{2}\operatorname{Tr}(\boldsymbol{G}). \tag{12}$$

The first term measures the error associated with temporal lag, while the second is proportional to the spatial-noise variance retained by the averaging kernel. However, the contributions of temporal lag and spatial noise are less direct after quantization. To make these contributions explicit, we decompose the quantization error change into separate terms and evaluate their expected contributions using Lemma A.8. This yields the quantization term in Lemma A.9, which depends on spatial noise and indirectly on temporal lag through the averaged latent mean:

$$\mathcal{Q}(\boldsymbol{w}) = \|\boldsymbol{\mu}_T^{\boldsymbol{w}}\|_2^2 + P\sigma^2 R(\boldsymbol{w}), \qquad \gamma_b = \frac{c_G}{(2^b-1)^2}, \tag{13}$$

where $c_G > 0$ collects the constants associated with the simplified GPTQ model. Thus, we can analyze how the total reconstruction error $\mathcal{E}_b(\boldsymbol{w})$ changes along the LNWA frontier as temporal lag is reduced and more spatial noise is retained. Let $m_0 := (M-1)/2$ be the mean age of LAWA, and set $x := 1 - m(\boldsymbol{w}_\lambda)/m_0 \in [0, 1]$ to measure the fraction of this age removed. Thus $x = 0$ gives LAWA, while $x = 1$ selects only the most recent saved weight, corresponding to $\lambda = \infty$. Thus, increasing x reduces temporal lag and retains more spatial noise. Each x specifies a frontier kernel $\boldsymbol{w}_\lambda$; we write its averaged latent mean and retained noise factor as $\boldsymbol{\mu}(x) := \boldsymbol{\mu}_T^{\boldsymbol{w}_\lambda}$ and $R(x) := R(\boldsymbol{w}_\lambda)$. To identify the tolerance region, we compare $\mathcal{E}_b(x)$ with its value at LAWA, $\mathcal{E}_b(0)$. We first examine the error on a fixed-support segment, where the frontier formulas in Lemma A.5 show that the averaged latent mean is affine in x and $R(x)$ is quadratic. For $q \geq 2$ active weights, differentiating the reconstruction

error gives

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\mathcal{E}_b(x) = m_0^2\left[\boldsymbol{a}_q^\top \boldsymbol{G}\boldsymbol{a}_q + 2\gamma_b\|\boldsymbol{a}_q\|_2^2 + \frac{2N_G + 2\gamma_b P\sigma^2}{D_q}\right] > 0, \tag{14}$$

where $\boldsymbol{a}_q := m_0^{-1}\mathrm{d}\boldsymbol{\mu}(x)/\mathrm{d}x$ is constant on the segment and $D_q := q(q^2-1)/12$. The positive curvature increases the error slope within each fixed-support segment. Although the slope may jump at support boundaries, Lemma A.10 controls these jumps. The comparison with the starting error $\mathcal{E}_b(0)$ therefore depends on their cumulative effect along $[0, x]$. To track the sign of this comparison, we normalize the error difference by x:

$$F_b(x) := \frac{\mathcal{E}_b(x) - \mathcal{E}_b(0)}{x}, \qquad x > 0, \qquad F_b(0) := \mathcal{E}_b'(0^+). \tag{15}$$

For $x > 0$, $F_b(x)$ has the same sign as $\mathcal{E}_b(x) - \mathcal{E}_b(0)$. Its derivative accounts for the curvature accumulated over $[0, x]$ and the slope jumps at boundaries crossed along the way. Lemma A.11 shows that, for sufficiently large T, the positive curvature dominates any adverse cumulative contribution from these jumps. Eq. (69) therefore gives constants $0 < k_b \le K_b$ such that

$$\frac{k_b}{2} \le F_b'(x) \le \frac{K_b}{2}. \tag{16}$$

Since F_b is continuous, the positive lower bound makes it strictly increasing over the entire frontier. Lemma A.11 also gives $F_b(1) = \mathcal{E}_b(1) - \mathcal{E}_b(0) > 0$. Integrating the derivative bounds gives the quantitative comparison in the following theorem.

Theorem 4.2 (Finite Noise Tolerance and Degradation)**.** *Under Assumptions A.3–A.7, for all sufficiently large T and every evaluated precision $b \ge 2$, there exist constants $0 < k_b \le K_b$ and a unique threshold $\tau_b \in [0, 1)$. With $\ell = (M-1)/T$, $g = \|\boldsymbol{G}\|_2$, and $A_* = \|\boldsymbol{\mu}_T\|_2 + V\ell$, where V is the increment bound in Assumption A.6, the threshold satisfies*

$$0 \le \tau_b \le \frac{D_M V}{m_0 T}\left(\frac{gV\ell}{N_G} + \frac{2A_*}{P\sigma^2}\right).$$

(i) Tolerance: *for $0 < x \le \tau_b$,*

$$-\frac{K_b}{2}x(\tau_b - x) \le \mathcal{E}_b(x) - \mathcal{E}_b(0) \le -\frac{k_b}{2}x(\tau_b - x), \tag{17}$$

(ii) Degradation: *for $\tau_b < x \le 1$,*

$$\mathcal{E}_b(x) - \mathcal{E}_b(0) \ge \frac{k_b}{2}x(x - \tau_b) > 0. \tag{18}$$

Theorem 4.2 characterizes noise tolerance and degradation at a fixed precision. When $\tau_b > 0$, the recovery of recent progress compensates for the additional spatial noise and quantization error over $0 < x < \tau_b$; beyond τ_b, the reconstruction error exceeds its value at $x = 0$ and increases strictly. If $\tau_b = 0$, any positive movement causes degradation. This is because the model may be unable to tolerate any additional spatial noise at sufficiently low precision. The remaining question is: at what precision does the $\tau_b = 0$ regime arise and favor less spatial noise?

4.3 Precision determines the preferred balance

By Eqs. (11) and (13), lower precision increases the weight on the quantization term $\mathcal{Q}$. Lemma A.12 shows that the retained latent-mean norm cannot decrease as x increases. The retained noise factor also increases, so on every open support segment,

$$\mathcal{Q}'(x) = 2\boldsymbol{\mu}(x)^\top \boldsymbol{\mu}'(x) + P\sigma^2 R'(x) > 0, \qquad x > 0. \tag{19}$$

Thus, $\mathcal{Q}$ increases as the kernel shifts toward more recent weights. Lemma A.13 gives $\mathcal{P}_0'(0^+) < 0$, so $\mathcal{P}_0$ initially decreases. As precision varies, whether this decrease offsets the increase in $\gamma_b\mathcal{Q}$ determines which tolerance regime in Theorem 4.2 applies: $\tau_b > 0$ when $\mathcal{E}_b$ initially decreases, and $\tau_b = 0$ otherwise. To characterize where this change occurs, Lemma A.14 places the unquantized and evaluated-precision optima in the first support segment. Write $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}(0)$ and $\boldsymbol{u} = \boldsymbol{\mu}'(0^+) = m_0\boldsymbol{a}_M$. On this segment, both error components are quadratic:

$$\mathcal{P}_0(x) - \mathcal{P}_0(0) = \mathcal{P}_0'(0^+)x + \frac{\mathcal{P}_0''(0^+)}{2}x^2, \quad \mathcal{Q}(x) - \mathcal{Q}(0) = \mathcal{Q}'(0^+)x + \frac{\mathcal{Q}''(0^+)}{2}x^2. \tag{20}$$

Appendix A.6 gives $\mathcal{Q}'(0^+) \geq 0$, $a := \mathcal{P}_0''(0^+) > 0$, and $c := \mathcal{Q}''(0^+) > 0$. The unquantized optimum and the optimum at precision b are

$$x_0 = \frac{\boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}})}{a}, \tag{21}$$

and

$$x_b = \max\left\{0, \frac{\boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}}) - 2\gamma_b \bar{\boldsymbol{\mu}}^\top \boldsymbol{u}}{a + c\gamma_b}\right\} \in [0, x_0). \tag{22}$$

Eq. (22) shows that quantization can make $x_b = 0$, corresponding to $\tau_b = 0$ in Theorem 4.2. Lemma A.14 establishes that this x_b is the global optimum across all support segments, so we can quantify the precision threshold in the next theorem.

Theorem 4.3 (Precision Threshold for Lag–Noise Preference). *Under Assumptions A.3–A.10, for sufficiently large T, define the precision threshold*

$$b_{\text{crit}} := \log_2\left(1 + \sqrt{\frac{2c_G \bar{\boldsymbol{\mu}}^\top \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}})}}\right). \tag{23}$$

For every evaluated precision $b \geq b_{\min}$,

$$\begin{aligned} b \leq b_{\text{crit}} \quad &\Longrightarrow \quad x_b = 0, \quad \tau_b = 0 \qquad \textbf{\textit{(less spatial noise)}}, \\ b > b_{\text{crit}} \quad &\Longrightarrow \quad x_b > 0, \quad \tau_b > 0 \qquad \textbf{\textit{(less temporal lag)}}. \end{aligned} \tag{24}$$

Moreover, x_b is nondecreasing in b and strictly increasing whenever $b > b_{\text{crit}}$. If $b_{\text{crit}} < b_{\min}$, all evaluated precisions lie in the second regime.

Theorem 4.3 answers the question left at the end of Section 4.2: b_{crit} marks where zero tolerance gives way to a positive tolerance interval. This threshold characterizes the reversal in the preferred lag–noise balance, from denoising at lower precision to reducing temporal lag at higher precision.

5 Experiments

Experimental setup. We pretrain OPT-series models (125M, 350M, and 1.3B) (Zhang et al., 2022) from scratch on FineWeb-Edu (Penedo et al., 2024) with AdamW (Loshchilov & Hutter, 2017). Following prior work (Catalan-Tatjer et al., 2026), we save 10 weights at 500-step intervals near the end of training and average them. We then apply GPTQ (Frantar et al., 2022) and AWQ (Lin et al., 2024) at 4, 3, and 2 bits (denoted by W4, W3, W2) using the same C4 calibration set (Raffel et al., 2020), then evaluate the PTQ performance by perplexity (PPL; lower is better) on a held-out RefinedWeb set (Penedo et al., 2023). Complete settings and results are in Appendix C.

From noise tolerance to degradation. We evaluate LNWA across coefficients ranging from $\lambda = 0$ (LAWA) to $\lambda = 1$. We report the performance change $\Delta\text{NLL}(\lambda)$ relative to LAWA, where $\Delta\text{NLL}(\lambda) = \text{NLL}(\lambda) - \text{NLL}(0) = \log \frac{\text{PPL}(\lambda)}{\text{PPL}(0)}$, and negative values indicate an improvement.

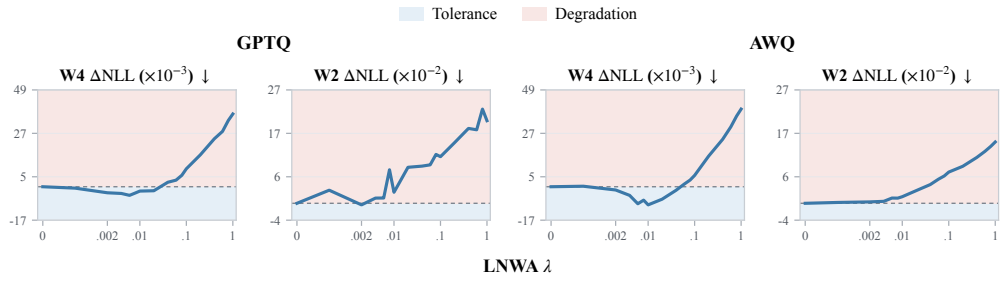


Figure 3: Comparison of ΔNLL across LNWA coefficients for OPT-125M under GPTQ and AWQ at 4-bit (W4) and 2-bit (W2).

Figure 3 shows that NLL can fall below the LAWA baseline at small λ, then rises above it at larger λ. This indicates that the benefit of reducing temporal lag can offset or outweigh the cost of retaining

additional spatial noise. Moving further along the frontier retains more noise and leads to performance degradation, consistent with the tolerance-to-degradation pattern characterized in Theorem 4.2.

Precision changes the preferred lag-noise balance. We compare LAWA, WMA, and LNWA in the small-λ range by fitting PPL against the LNWA-equivalent averaging coefficient. Kernel choices and regression details are given in Appendix C.1.

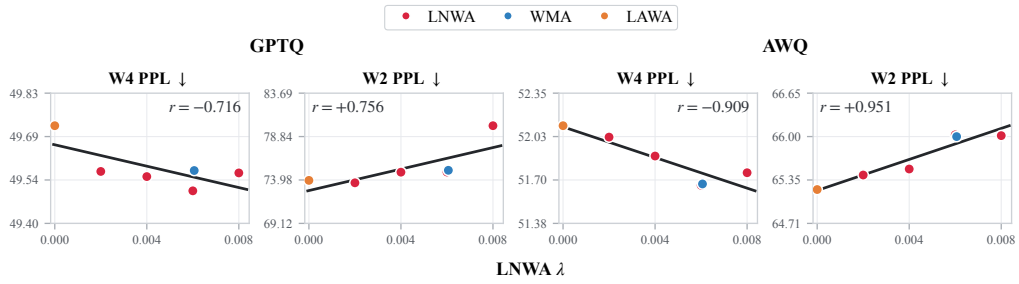


Figure 4: Comparison of PPL across LNWA-equivalent averaging coefficients for OPT-125M under GPTQ and AWQ at 4-bit (W4) and 2-bit (W2).

Figure 4 shows a positive correlation between PPL and the averaging coefficient at W2 and a negative correlation at W4. Smaller coefficients retain less spatial noise, whereas larger coefficients reduce temporal lag. These correlations therefore indicate that lower precision favors less noise and higher precision favors less lag, supporting a precision-dependent reversal in preference in Theorem 4.3.

Comparison with learning-rate decay. We finally compare LNWA with two LR decay schedules on OPT-1.3B. As shown in Figure 5, LNWA achieves lower average PPL (Avg.) than both decay schedules under GPTQ and AWQ, with the largest improvement at W2. These results support weight averaging as an effective strategy for the pretraining-to-quantization handoff.

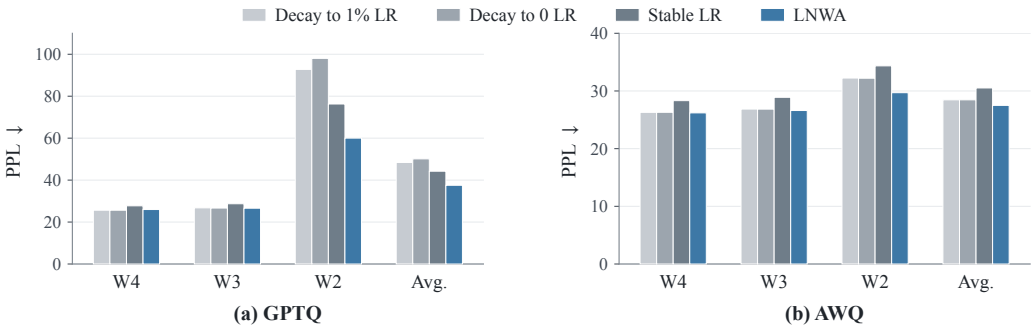


Figure 5: Comparison of PPL between LR decay, Stable LR, and LNWA for OPT-1.3B under GPTQ and AWQ at 4-bit (W4), 3-bit (W3), 2-bit (W2), and their average.

6 Conclusion

We developed a theoretical framework for weight averaging at the pretraining-to-quantization transition. Averaging suppresses spatial noise but introduces temporal lag by sacrificing recent optimization progress. We quantify this trade-off and derive LNWA, a closed-form family of averaging kernels that parameterizes its Pareto frontier. Through a simplified GPTQ model, we connect this trade-off to output reconstruction error and establish two main results. At a fixed precision, reduced lag can compensate for additional noise within a finite tolerance interval; beyond it, the error increases. Within this interval, we further show that lower precision favors less spatial noise, whereas higher precision favors less temporal lag. Experiments on LLMs support both results.

AI USE STATEMENT

Generative AI tools were used only to assist with literature search, coding, writing refinement, and consistency checking. The research motivation, theoretical development, experiments, and conclusions were independently developed and completed by the authors. All AI-assisted outputs were carefully reviewed and verified, and the authors take full responsibility for the contents of this paper.

REFERENCES

Arash Ahmadian, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Zhen Stephen Gou, Phil Blunsom, Ahmet Üstün, and Sara Hooker. Intriguing properties of quantization at scale. *Advances in Neural Information Processing Systems*, 36:34278–34294, 2023.

Niccolò Ajroldi, Antonio Orvieto, and Jonas Geiping. When, where and why to average weights? *arXiv preprint arXiv:2502.06761*, 2025.

Eduardo Alvarez, Omri Almog, Eric Chung, Simon Layton, Dusan Stosic, Ronny Krashinsky, and Kyle Aubrey. Introducing nvfp4 for efficient and accurate low-precision inference. *NVIDIA Technical Blog*, 24, 2025.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Aciq: Analytical clipping for integer quantization of neural networks. 2018.

Adam Block and Cyril Zhang. Ema without the lag: Bias-corrected iterate averaging schemes. *arXiv preprint arXiv:2508.00180*, 2025.

Albert Catalan-Tatjer, Niccolo Ajroldi, and Jonas Geiping. Training dynamics impact post-training quantization robustness. In *International Conference on Learning Representations*, volume 2026, pp. 64294–64319, 2026.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in neural information processing systems*, 36: 4396–4429, 2023.

Fan Chen, Audrey Huang, Noah Golowich, Sadhika Malladi, Adam Block, Jordan Ash, Akshay Krishnamurthy, and Dylan Foster. The coverage principle: How pre-training enables post-training. In *International Conference on Learning Representations*, volume 2026, pp. 112225–112293, 2026.

Junren Chen, Yueqi Wang, and Michael K Ng. Quantized low-rank multivariate regression with random dithering. *IEEE Transactions on Signal Processing*, 71:3913–3928, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

Sjoerd Dirksen, Weilin Li, and Johannes Maly. Subspace and doa estimation under coarse quantization. *IEEE Transactions on Information Theory*, 2025.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. *arXiv preprint arXiv:2401.06118*, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna B Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37:76232–76264, 2024.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Tanishq Kumar, Zachary Ankner, Benjamin Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *International Conference on Learning Representations*, volume 2025, pp. 71833–71864, 2025.

Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. *Advances in Neural Information Processing Systems*, 38:101211–101269, 2026a.

Binghui Li, Zilin Wang, Fengling Chen, Shiyang Zhao, Ruiheng Zheng, and Lei Wu. Optimal learning-rate schedules under functional scaling laws: Power decay and warmup-stable-decay. *arXiv preprint arXiv:2602.06797*, 2026b.

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. Gptaq: Efficient finetuning-free quantization for asymmetric calibration. *arXiv preprint arXiv:2504.02692*, 2025.

Yunshui Li, Yiyuan Ma, Shen Yan, Chaoyi Zhang, Jing Liu, Jianqiao Lu, Ziwen Xu, Mengzhao Chen, Minrui Wang, Shiyi Zhan, et al. Model merging in pre-training of large language models. *Advances in Neural Information Processing Systems*, 38:133668–133691, 2026c.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6: 87–100, 2024.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025.

Zihan Liu, Ruiheng Zheng, Shaobo Zhang, Changxin Tian, Kunlong Chen, Zhiqiang Zhang, and Lei Wu. Effective learning rate governs loss dynamics in language model pretraining. *arXiv preprint arXiv:2608.24814*, 2026.

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

Kairong Luo, Zhenbo Sun, Haodong Wen, Xinyu Shi, Jiarui Cui, Chenyi Dang, Kaifeng Lyu, and Wenguang Chen. How learning rate decay wastes your best data in curriculum-based llm pretraining. In *International Conference on Learning Representations*, volume 2026, pp. 147266–147297, 2026.

The Microsoft AI Team. Mai-thinking-1: Building a hill-climbing machine, 2026.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *arXiv preprint arXiv:2411.18704*, 2024.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, pp. 7197–7206. PMLR, 2020.

Xu Ouyang, Tao Ge, Thomas Hartvigsen, Zhisong Zhang, Haitao Mi, and Dong Yu. Low-bit quantization favors undertrained llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32338–32348, 2025.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172, 2023.

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

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Nolan Miller. Training trajectories, mini-batch losses and the curious role of the learning rate. *arXiv preprint arXiv:2301.02312*, 2023.

Sunny Sanyal, Atula Neerkaje, Jean Kaddour, Abhishek Kumar, and Sujay Sanghavi. Early weight averaging meets high learning rates for llm pre-training. *arXiv preprint arXiv:2306.03241*, 2023.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45472–45496, 2024.

Yuantian Shao, Yuanteng Chen, Peisong Wang, Jianlin Yu, Jing Lin, Zhihui Wei, Jian Cheng, et al. Dartquant: efficient rotational distribution calibration for llm quantization. *Advances in Neural Information Processing Systems*, 38:143936–143970, 2026.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, et al. Flatquant: Flatness matters for llm quantization. *arXiv preprint arXiv:2410.09426*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025a.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025b.

Changxin Tian, Jiapeng Wang, Qian Zhao, Kunlong Chen, Jia Liu, Ziqi Liu, Jiaxin Mao, Xin Zhao, Zhiqiang Zhang, and Jun Zhou. Wsm: decay-free learning rate schedule via checkpoint merging for llm pre-training. In *International Conference on Learning Representations*, volume 2026, pp. 134767–134795, 2026.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *Proceedings of Machine Learning Research*, 235:48630, 2024.

Xingjin Wang, Howe Tissue, Lu Wang, Linjing Li, and Daniel Dajun Zeng. Learning dynamics in continual pre-training for large language models. *arXiv preprint arXiv:2505.07796*, 2025.

R.A. Wannamaker, S.P. Lipshitz, J. Vanderkooy, and J.N. Wright. A theory of nonsubtractive dither. *IEEE Transactions on Signal Processing*, 48(2):499–516, 2000. doi: 10.1109/78.823976.

Ishaan Watts, Catherine Li, Sachin Goyal, Jacob Mitchell Springer, and Aditi Raghunathan. Sharpness-aware pretraining mitigates catastrophic forgetting. *arXiv preprint arXiv:2605.02105*, 2026.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *International Conference on Learning Representations*, volume 2025, pp. 42840–42885, 2025.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pp. 38087–38099. PMLR, 2023.

Kazuki Yano, Shun Kiyono, Sosuke Kobayashi, Sho Takase, and Jun Suzuki. Pre-training llm without learning rate decay enhances supervised fine-tuning. *arXiv preprint arXiv:2603.16127*, 2026.

Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. Exploring post-training quantization in llms from comprehensive study to low rank compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19377–19385, 2024.

Cheng Zhang, Jianyi Cheng, George A Constantinides, and Yiren Zhao. Lqer: Low-rank quantization error reconstruction for llms. *arXiv preprint arXiv:2402.02446*, 2024.

Cheng Zhang, Jeffrey TH Wong, Can Xiao, George Constantinides, and Yiren Zhao. Qera: an analytical framework for quantization error reconstruction. In *International Conference on Learning Representations*, volume 2025, pp. 12531–12560, 2025.

Jinjie Zhang and Rayan Saab. Spfq: A stochastic algorithm and its error analysis for neural network quantization. *Information and Inference: A Journal of the IMA*, 15(2):iaag011, 2026.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

APPENDIX

A Theoretical Proofs **14**

A.1 Setup . . . 14

A.2 Temporal lag and spatial noise . . . 15

A.3 Lag–noise Pareto frontier and closed-form LNWA kernel . . . 16

A.4 Classical averaging kernels as LNWA special cases . . . 20

A.5 PTQ tolerance and degradation . . . 23

A.6 Precision-dependent averaging preference . . . 36

B Key Assumptions and Limitation **40**

B.1 Empirical Assessment of Key Assumptions . . . 40

B.2 Limitation . . . 42

C Additional Experimental Results **43**

C.1 Experimental Settings . . . 43

C.2 Experimental Results . . . 45

A THEORETICAL PROOFS

This section presents the detailed proofs of all theoretical results in Section 4.

A.1 SETUP

Definition A.1 (Finite Convex Weight Averaging)**.** Let $\boldsymbol{\theta}_T, \boldsymbol{\theta}_{T-1}, \ldots, \boldsymbol{\theta}_{T-M+1} \in \mathbb{R}^P$ be a window of $M \geq 2$ retained model weights. Age $d = 0$ denotes the newest weights; T denotes the training steps. Define

$$\mathcal{W}_M = \left\{ \boldsymbol{w} \in \mathbb{R}^M : w_d \geq 0,\ \sum_{d=0}^{M-1} w_d = 1 \right\}, \qquad \boldsymbol{\theta}_T^{\boldsymbol{w}} = \sum_{d=0}^{M-1} w_d \boldsymbol{\theta}_{T-d}.$$

Here, $\mathcal{W}_M$ is the set of feasible averaging kernels, $\boldsymbol{w} \in \mathcal{W}_M$ is an averaging kernel, and $\boldsymbol{\theta}_T^{\boldsymbol{w}}$ lies in the finite convex hull of the retained weights.

Definition A.2 (Latent-mean Trajectory)**.** Let $\mathcal{F}_0$ be the sigma-field generated by the information shared across stochastic training runs, such as the model configuration, common initialization, and training setup. Let ξ index the run-specific stochasticity, such as data ordering and dropout. For the aligned state $\boldsymbol{\theta}_{T-d}(\xi)$, assume $\mathbb{E}[\|\boldsymbol{\theta}_{T-d}(\xi)\|_2^2 \mid \mathcal{F}_0] < \infty$. The conditional mean state $\boldsymbol{\mu}_t$ and the zero-mean spatial noise $\boldsymbol{\epsilon}_t$ at the training step t of an individual run are

$$\boldsymbol{\mu}_{T-d} := \mathbb{E}_\xi[\boldsymbol{\theta}_{T-d}(\xi) \mid \mathcal{F}_0], \qquad \boldsymbol{\epsilon}_{T-d}(\xi) := \boldsymbol{\theta}_{T-d}(\xi) - \boldsymbol{\mu}_{T-d}.$$

A stochastic training run therefore satisfies $\boldsymbol{\theta}_{T-d} = \boldsymbol{\mu}_{T-d} + \boldsymbol{\epsilon}_{T-d}$ and $\mathbb{E}[\boldsymbol{\epsilon}_{T-d} \mid \mathcal{F}_0] = \mathbf{0}$.

Throughout Appendix A, all expectations and covariances condition on $\mathcal{F}_0$, and averaging kernels are fixed given $\mathcal{F}_0$.

For ages $d, e \in \{0, \ldots, M-1\}$, define the conditional covariance blocks of the fluctuations around the latent-mean trajectory:

$$\boldsymbol{\Gamma}_{de} := \mathrm{Cov}(\boldsymbol{\epsilon}_{T-d}, \boldsymbol{\epsilon}_{T-e} \mid \mathcal{F}_0). \tag{25}$$

Within the retained stable-LR window, we use a common covariance for spatial noise and assume that fluctuations at different retained positions are uncorrelated.

Assumption A.1 (Common Covariance and Uncorrelated Spatial Noise)**.** Conditional on $\mathcal{F}_0$, the spatial-noise vectors have finite second moments and satisfy, for $d, e \in \{0, \dots, M-1\}$,

$$\mathbb{E}[\boldsymbol{\epsilon}_{T-d} \mid \mathcal{F}_0] = \mathbf{0}, \qquad \mathbb{E}[\boldsymbol{\epsilon}_{T-d}\boldsymbol{\epsilon}_{T-e}^\top \mid \mathcal{F}_0] = \begin{cases} \boldsymbol{\Sigma}, & d = e, \\ \mathbf{0}, & d \neq e, \end{cases} \qquad \boldsymbol{\Sigma} \succeq 0.$$

Thus $\boldsymbol{\Gamma}_{dd} = \boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}_{de} = \mathbf{0}$ for $d \neq e$.

Define the forward latent increment $\boldsymbol{v}_{T-j+1} = \boldsymbol{\mu}_{T-j+1} - \boldsymbol{\mu}_{T-j}, 1 \leq j < M$, and the cumulative old-weight mass $s_j(\boldsymbol{w}) = \sum_{d=j}^{M-1} w_d$, we have

$$\boldsymbol{b}(\boldsymbol{w}) = \sum_{j=1}^{M-1} s_j(\boldsymbol{w})\boldsymbol{v}_{T-j+1}, \qquad \boldsymbol{\epsilon}_T^{\boldsymbol{w}} = \sum_{d=0}^{M-1} w_d \boldsymbol{\epsilon}_{T-d}, \qquad R(\boldsymbol{w}) = \sum_{d=0}^{M-1} w_d^2.$$

Generally, we have

$$\mathrm{Cov}(\boldsymbol{\epsilon}_T^{\boldsymbol{w}} \mid \mathcal{F}_0) = \sum_{d,e} w_d w_e \boldsymbol{\Gamma}_{de}.$$

Under Assumption A.1, this reduces to

$$\mathrm{Cov}(\boldsymbol{\epsilon}_T^{\boldsymbol{w}} \mid \mathcal{F}_0) = R(\boldsymbol{w})\boldsymbol{\Sigma}.$$

Accordingly, $R(\boldsymbol{w})^{-1}$ is a variance-equivalent effective sample size.

A.2 Temporal Lag and Spatial Noise

Lemma A.1 (Lag-noise Decomposition)**.** *Every $\boldsymbol{w} \in \mathcal{W}_M$ satisfies*

$$\boldsymbol{\theta}_T^{\boldsymbol{w}} = \boldsymbol{\mu}_T - \boldsymbol{b}(\boldsymbol{w}) + \boldsymbol{\epsilon}_T^{\boldsymbol{w}}. \tag{26}$$

Under Assumption A.1, its residual covariance is

$$\mathrm{Cov}(\boldsymbol{\epsilon}_T^{\boldsymbol{w}} \mid \mathcal{F}_0) = R(\boldsymbol{w})\boldsymbol{\Sigma}.$$

Proof. The increments can be written as $\boldsymbol{\mu}_{T-d} = \boldsymbol{\mu}_T - \sum_{j=1}^{d} \boldsymbol{v}_{T-j+1}$. Substituting this identity into $\sum_d w_d(\boldsymbol{\mu}_{T-d} + \boldsymbol{\epsilon}_{T-d})$ gives

$$\sum_{d=0}^{M-1} w_d \sum_{j=1}^{d} \boldsymbol{v}_{T-j+1} = \sum_{j=1}^{M-1} \left(\sum_{d=j}^{M-1} w_d \right) \boldsymbol{v}_{T-j+1}.$$

This proves Eq. (26). The conditional cross-covariances of distinct $\boldsymbol{\epsilon}_{T-d}$ vanish by Assumption A.1, hence $\mathrm{Cov}(\boldsymbol{\epsilon}_T^{\boldsymbol{w}} \mid \mathcal{F}_0) = \sum_d w_d^2 \boldsymbol{\Sigma} = R(\boldsymbol{w})\boldsymbol{\Sigma}$. □

Assumption A.2 (Bounded Adjacent Progress Variation)**.** Adjacent latent increments in the retained discrete window obey

$$\|\boldsymbol{v}_{T-j+1} - \boldsymbol{v}_{T-j+2}\|_2 \leq \kappa, \qquad 2 \leq j < M,$$

for a finite $\kappa \geq 0$.

Under Assumption A.2, the next lemma reduces the trajectory displacement to a scalar lag coordinate with a controlled remainder.

Lemma A.2 (Scalar Lag with Controlled Remainder)**.** *Under Assumption A.2, define $m(\boldsymbol{w}) = \sum_{d=0}^{M-1} d w_d$. Then*

$$\boldsymbol{\theta}_T^{\boldsymbol{w}} = \boldsymbol{\mu}_T - m(\boldsymbol{w})\boldsymbol{v}_T + \boldsymbol{\epsilon}_T^{\boldsymbol{w}} + \boldsymbol{e}_{\boldsymbol{w}}^{\mathrm{traj}}, \qquad \|\boldsymbol{e}_{\boldsymbol{w}}^{\mathrm{traj}}\|_2 \leq \frac{\kappa}{2} \sum_{d=0}^{M-1} w_d d(d-1). \tag{27}$$

When $\boldsymbol{v}_{T-j+1} = \boldsymbol{v}_T$ throughout the sampled window, $\boldsymbol{e}_{\boldsymbol{w}}^{\mathrm{traj}} = \mathbf{0}$.

Proof. The cumulative weights satisfy

$$\sum_{j=1}^{M-1} s_j(\boldsymbol{w}) = \sum_{d=0}^{M-1} d\, w_d = m(\boldsymbol{w}).$$

Define the trajectory deviation $\boldsymbol{\delta}_j^{\mathrm{traj}} := \boldsymbol{v}_{T-j+1} - \boldsymbol{v}_T$ for $1 \le j < M$. Then

$$\boldsymbol{b}(\boldsymbol{w}) = m(\boldsymbol{w})\boldsymbol{v}_T + \sum_{j=1}^{M-1} s_j(\boldsymbol{w})\boldsymbol{\delta}_j^{\mathrm{traj}}.$$

Adjacent variation implies $\|\boldsymbol{\delta}_j^{\mathrm{traj}}\|_2 \le (j-1)\kappa$. Non-negativity of the weights and

$$\sum_{j=1}^{M-1} (j-1)s_j(\boldsymbol{w}) = \frac{1}{2}\sum_{d=0}^{M-1} w_d d(d-1)$$

give Eq. (27) after setting

$$\boldsymbol{e}_{\boldsymbol{w}}^{\mathrm{traj}} = -\sum_{j=1}^{M-1} s_j(\boldsymbol{w})\boldsymbol{\delta}_j^{\mathrm{traj}}.$$

□

A.3 Lag–noise Pareto frontier and closed-form LNWA kernel

The decomposition above motivates the kernel coordinates.

Definition A.3 (Lag–noise Coordinates)**.** For any $\boldsymbol{w} \in \mathcal{W}_M$, define

$$\Phi(\boldsymbol{w}) = \big(m(\boldsymbol{w}), R(\boldsymbol{w})\big), \qquad \mathcal{G}_M = \{\Phi(\boldsymbol{w}) : \boldsymbol{w} \in \mathcal{W}_M\}.$$

Here, $\Phi(\boldsymbol{w})$ denotes the lag–noise coordinates of the averaging kernel $\boldsymbol{w}$, and $\mathcal{G}_M$ is the set of all feasible lag–noise coordinates.

Write $\mathrm{Eff}(\mathcal{G}_M)$ for the set of Pareto-efficient points in $\mathcal{G}_M$: a point belongs to this set if no feasible kernel has both a smaller or equal temporal lag and a smaller or equal spatial noise, with at least one strict inequality.

To identify the efficient set, first fix a lag budget ℓ:

$$\boldsymbol{w}^\star(\ell) = \arg\min_{\boldsymbol{w}\in\mathcal{W}_M} R(\boldsymbol{w}) \quad \text{s.t.} \quad m(\boldsymbol{w}) \le \ell, \qquad 0 \le \ell \le \frac{M-1}{2}. \tag{28}$$

Lemma A.3 (Lag-noise Pareto Frontier)**.** *For $0 < \ell < (M-1)/2$, Eq. (28) has a unique solution supported on the q most recent saved weights, where*

$$q = q(\ell) = \min\{M, \lceil 3\ell + 1\rceil\} \in \{2, \dots, M\}. \tag{29}$$

Its coordinates are

$$\big(\boldsymbol{w}^\star(\ell)\big)_d = \begin{cases} q^{-1} - \gamma_q(d - \bar{d}_q), & 0 \le d < q, \\ 0, & q \le d < M, \end{cases} \tag{30}$$

where

$$\bar{d}_q = \frac{q-1}{2}, \quad D_q = \sum_{d=0}^{q-1}(d-\bar{d}_q)^2 = \frac{q(q^2-1)}{12}, \quad \gamma_q = \frac{\bar{d}_q - \ell}{D_q}, \tag{31}$$

and

$$R^\star(\ell) = \frac{1}{q} + \frac{(\bar{d}_q - \ell)^2}{D_q}.$$

Proof. The feasible set is compact, and the objective is strictly convex. Hence the minimizer is unique. With $\boldsymbol{d} = (0, 1, \dots, M-1)^\top$, the Lagrangian is

$$\mathcal{L}_{\mathrm{KKT}}(\boldsymbol{w}, \tau, \nu, \boldsymbol{\alpha}) = \|\boldsymbol{w}\|_2^2 + \tau(\boldsymbol{d}^\top\boldsymbol{w} - \ell) + \nu(\mathbf{1}^\top\boldsymbol{w} - 1) - \boldsymbol{\alpha}^\top\boldsymbol{w}.$$

Slater's condition holds for $\ell > 0$, so the KKT conditions are

$$2\boldsymbol{w} + \tau \boldsymbol{d} + \nu \mathbf{1} - \boldsymbol{\alpha} = 0, \qquad \tau(\boldsymbol{d}^\top \boldsymbol{w} - \ell) = 0, \qquad \alpha_d w_d = 0, \tag{32}$$

with $\tau \geq 0$ and $\boldsymbol{\alpha} \geq 0$. If $\tau = 0$, they give the uniform simplex minimizer, which is infeasible because $(M-1)/2 > \ell$. Thus $\tau > 0$ and the lag constraint is active. Stationarity and complementary slackness in Eq. (32) give

$$\begin{aligned} w_d > 0 &\implies \alpha_d = 0 \implies w_d = -\frac{\nu + \tau d}{2} > 0, \\ w_d = 0 &\implies -\frac{\nu + \tau d}{2} = -\frac{\alpha_d}{2} \leq 0. \end{aligned} \tag{33}$$

Combining the two cases yields

$$w_d = \left[-\frac{\nu + \tau d}{2} \right]_+ . \tag{34}$$

Because $\tau > 0$, the expression inside the positive part decreases with d. Hence the support is a prefix $S_q = \{0, \ldots, q-1\}$. For $d \in S_q$, the positive-part operator is inactive and

$$w_d = -\frac{\nu}{2} - \frac{\tau}{2} d. \tag{35}$$

Using $\sum_{d=0}^{q-1} w_d = 1$ and $\sum_{d=0}^{q-1} d = q\bar{d}_q$ in Eq. (35) gives

$$1 = -\frac{q\nu}{2} - \frac{\tau q \bar{d}_q}{2}, \qquad -\frac{\nu}{2} = \frac{1}{q} + \frac{\tau}{2}\bar{d}_q.$$

Substitution back into Eq. (35) yields

$$w_d = \frac{1}{q} - \frac{\tau}{2}(d - \bar{d}_q), \qquad d \in S_q.$$

Therefore, with $\boldsymbol{c}_q = (0, 1, \ldots, q-1)^\top - \bar{d}_q \mathbf{1}_q$,

$$\boldsymbol{w}^\star_{S_q} = q^{-1}\mathbf{1}_q - \frac{\tau}{2}\boldsymbol{c}_q.$$

Finally, the lag constraint is active, so

$$\begin{aligned} \ell &= \sum_{d=0}^{q-1} d\, w_d \\ &= \bar{d}_q - \frac{\tau}{2} \sum_{d=0}^{q-1} d(d - \bar{d}_q) \\ &= \bar{d}_q - \frac{\tau}{2} D_q. \end{aligned}$$

Hence $\tau/2 = (\bar{d}_q - \ell)/D_q = \gamma_q$, which proves Eqs. (30)–(31). The last-active and first-inactive conditions reduce to

$$\frac{q-2}{3} < \ell \leq \frac{q-1}{3} \quad (q < M), \qquad \ell > \frac{M-2}{3} \quad (q = M),$$

which gives Eq. (29). Finally, $\mathbf{1}_q^\top \boldsymbol{c}_q = 0$ and $\|\boldsymbol{c}_q\|_2^2 = D_q$, so

$$R(\boldsymbol{w}^\star(\ell)) = \left\| q^{-1}\mathbf{1}_q - \frac{\tau}{2}\boldsymbol{c}_q \right\|_2^2 = \frac{1}{q} + \frac{(\bar{d}_q - \ell)^2}{D_q}.$$

□

We next parameterize the budget-optimal kernels by a single coefficient.

Theorem A.1 (Lag-noise Trade-off Weight Averaging (LNWA) Kernel)**.** *For $0 \le \lambda < \infty$, define the Lag-noise Trade-off Weight Averaging (*LNWA*) kernel by*

$$\boldsymbol{w}_\lambda = \arg\min_{\boldsymbol{w}\in\mathcal{W}_M} \{R(\boldsymbol{w}) + \lambda m(\boldsymbol{w})^2\}. \tag{36}$$

The finite-λ family satisfies

$$\{\Phi(\boldsymbol{w}_\lambda) : 0 \le \lambda < \infty\} = \mathrm{Eff}(\mathcal{G}_M) \setminus \{\Phi(\boldsymbol{e}_0)\}. \tag{37}$$

Its unique solution has an active prefix of size $q \ge 2$ and

$$(\boldsymbol{w}_\lambda)_d = \begin{cases} q^{-1} + \lambda m_q(\lambda)(\bar{d}_q - d), & 0 \le d < q, \\ 0, & q \le d < M, \end{cases} \tag{38}$$

where

$$m_q(\lambda) = m(\boldsymbol{w}_\lambda) = \frac{\bar{d}_q}{1 + \lambda D_q}, \qquad \bar{d}_q = \frac{q-1}{2}, \qquad D_q = \sum_{d=0}^{q-1} (d - \bar{d}_q)^2 = \frac{q(q^2-1)}{12}. \tag{39}$$

Proof. Any kernel with $m(\boldsymbol{w}) > (M-1)/2$ is dominated by the uniform kernel. For $0 \le \ell \le (M-1)/2$, uniqueness in Eq. (28) implies

$$\mathrm{Eff}(\mathcal{G}_M) = \{\Phi(\boldsymbol{w}^\star(\ell)) : 0 \le \ell \le (M-1)/2\}. \tag{40}$$

Indeed, a dominator of $\boldsymbol{w}^\star(\ell)$ contradicts its optimality or uniqueness; conversely, $\boldsymbol{w}^\star(m(\boldsymbol{w}))$ dominates every feasible $\boldsymbol{w}$ that is not the budget minimizer.

For finite λ, Eq. (36) is strictly convex. By the same stationarity and complementary-slackness argument as Eqs. (33)–(34), its KKT conditions give

$$w_d = \Big[-\frac{\nu}{2} - \lambda m d\Big]_+, \qquad m = m(\boldsymbol{w}_\lambda), \tag{41}$$

so the active set is a prefix. Moreover, for sufficiently small $t > 0$,

$$\begin{aligned} &R((1-t)\boldsymbol{e}_0 + t\boldsymbol{e}_1) + \lambda m((1-t)\boldsymbol{e}_0 + t\boldsymbol{e}_1)^2 \\ &\quad = 1 - 2t + (\lambda+2)t^2 < 1 = R(\boldsymbol{e}_0) + \lambda m(\boldsymbol{e}_0)^2, \end{aligned} \tag{42}$$

so $\boldsymbol{e}_0$ cannot solve Eq. (36); hence $q \ge 2$. On the active prefix $S_q = \{0, \dots, q-1\}$, Eq. (41) becomes

$$w_d = -\frac{\nu}{2} - \lambda m d, \qquad d \in S_q. \tag{43}$$

Using $\sum_{d=0}^{q-1} w_d = 1$ and $\sum_{d=0}^{q-1} d = q\bar{d}_q$ gives

$$1 = -\frac{q\nu}{2} - \lambda m q \bar{d}_q, \qquad -\frac{\nu}{2} = \frac{1}{q} + \lambda m \bar{d}_q.$$

Substituting this identity into Eq. (43) yields

$$w_d = \frac{1}{q} + \lambda m(\bar{d}_q - d), \qquad d \in S_q.$$

Therefore,

$$\begin{aligned} m &= \sum_{d=0}^{q-1} d\, w_d \\ &= \bar{d}_q + \lambda m \sum_{d=0}^{q-1} d(\bar{d}_q - d) \\ &= \bar{d}_q - \lambda m D_q, \end{aligned}$$

where

$$\sum_{d=0}^{q-1} d(\bar{d}_q - d) = q\bar{d}_q^2 - \sum_{d=0}^{q-1} d^2 = -D_q.$$

Consequently,

$$m = \frac{\bar{d}_q}{1 + \lambda D_q},$$

which proves Eqs. (38) and (39).

For $0 < \ell < (M-1)/2$, set

$$\lambda = \frac{\gamma_q}{\ell} = \frac{\bar{d}_q - \ell}{D_q \ell}.$$

Then Eqs. (30) and (38) coincide, so every interior budget solution lies on the LNWA path. The budget endpoint $\ell = (M-1)/2$ is the uniform kernel $\boldsymbol{w}_0$, whereas $\ell = 0$ has the unique solution $\boldsymbol{e}_0$. Eq. (42) excludes $\boldsymbol{e}_0$ for every finite λ. Together with Eq. (40), these facts prove Eq. (37). □

Lemma A.4 (Global Structure of the LNWA)**.** *The* LNWA *path has endpoints*

$$\boldsymbol{w}_0 = \frac{1}{M}\mathbf{1}, \qquad \lim_{\lambda \to \infty} \boldsymbol{w}_\lambda = \boldsymbol{e}_0. \tag{44}$$

For $q \geq 3$, define

$$\lambda_q^{\mathrm{knot}} = \frac{6}{q(q-1)(q-2)}.$$

If $M > 2$, all M coordinates are active on $[0, \lambda_M^{\mathrm{knot}})$; for each $q \in \{3, \ldots, M-1\}$, exactly q coordinates are active on $[\lambda_{q+1}^{\mathrm{knot}}, \lambda_q^{\mathrm{knot}})$; and exactly two coordinates are active on $[\lambda_3^{\mathrm{knot}}, \infty)$. If $M = 2$, both coordinates remain active for every finite λ. Defining $\boldsymbol{w}_\infty := \boldsymbol{e}_0$ therefore completes the frontier:

$$\{\Phi(\boldsymbol{w}_\lambda) : 0 \leq \lambda \leq \infty\} = \mathrm{Eff}(\mathcal{G}_M).$$

Moreover, for $0 \leq \lambda_1 < \lambda_2 \leq \infty$,

$$m(\boldsymbol{w}_{\lambda_2}) < m(\boldsymbol{w}_{\lambda_1}), \qquad R(\boldsymbol{w}_{\lambda_2}) > R(\boldsymbol{w}_{\lambda_1}). \tag{45}$$

Proof. At $\lambda = 0$, Eq. (36) reduces to minimizing $R(\boldsymbol{w})$ over $\mathcal{W}_M$, whose unique solution is $\boldsymbol{w}_0 = M^{-1}\mathbf{1}$. This proves the first identity in Eq. (44).

Suppose the active prefix is $S_q = \{0, \ldots, q-1\}$. To verify that its size is exactly q, we check its last active coefficient and the affine continuation to the next coordinate, denoted by $\widetilde{w}_q$. Eqs. (38)–(39) give

$$\begin{aligned} \widetilde{w}_q &:= q^{-1} + \lambda m_q(\lambda)(\bar{d}_q - q), \\ (\boldsymbol{w}_\lambda)_{q-1} > 0 &\iff \lambda < \lambda_q^{\mathrm{knot}} \qquad (q \geq 3), \\ \widetilde{w}_q \leq 0 &\iff \lambda \geq \lambda_{q+1}^{\mathrm{knot}} \qquad (q < M). \end{aligned}$$

These two boundary conditions give $[\lambda_{q+1}^{\mathrm{knot}}, \lambda_q^{\mathrm{knot}})$; at each knot, the oldest active coefficient reaches zero.

When $M > 2$, the two-coordinate segment is $\lambda \in [\lambda_3^{\mathrm{knot}}, \infty)$; when $M = 2$, it is the entire finite path $\lambda \in [0, \infty)$. On this segment, Theorem A.1 gives

$$\boldsymbol{w}_\lambda = \left(\frac{\lambda+1}{\lambda+2}, \frac{1}{\lambda+2}, 0, \ldots, 0\right). \tag{46}$$

Taking $\lambda \to \infty$ in Eq. (46) gives

$$\lim_{\lambda \to \infty} \boldsymbol{w}_\lambda = \boldsymbol{e}_0.$$

This proves the second identity in Eq. (44).

To prove Eq. (45), first suppose $0 \leq \lambda_1 < \lambda_2 < \infty$. Optimality at the two coefficients gives

$$R(\boldsymbol{w}_{\lambda_1}) + \lambda_1 m(\boldsymbol{w}_{\lambda_1})^2 \leq R(\boldsymbol{w}_{\lambda_2}) + \lambda_1 m(\boldsymbol{w}_{\lambda_2})^2, \tag{47}$$

$$R(\boldsymbol{w}_{\lambda_2}) + \lambda_2 m(\boldsymbol{w}_{\lambda_2})^2 \leq R(\boldsymbol{w}_{\lambda_1}) + \lambda_2 m(\boldsymbol{w}_{\lambda_1})^2. \tag{48}$$

Adding Eqs. (47) and (48) yields

$$(\lambda_2 - \lambda_1)\left[m(\boldsymbol{w}_{\lambda_2})^2 - m(\boldsymbol{w}_{\lambda_1})^2\right] \leq 0.$$

Thus,

$$m(\boldsymbol{w}_{\lambda_2}) \le m(\boldsymbol{w}_{\lambda_1}).$$

Equality would force $R(\boldsymbol{w}_{\lambda_2}) = R(\boldsymbol{w}_{\lambda_1})$ by Eqs. (47)–(48), so strict convexity would give $\boldsymbol{w}_{\lambda_2} = \boldsymbol{w}_{\lambda_1}$. But Eq. (43) gives $(\boldsymbol{w}_\lambda)_0 - (\boldsymbol{w}_\lambda)_1 = \lambda m(\boldsymbol{w}_\lambda)$, with $m(\boldsymbol{w}_\lambda) > 0$ for every finite λ. Thus the same kernel cannot occur at distinct coefficients, and the inequality is strict. Eq. (47) then gives

$$R(\boldsymbol{w}_{\lambda_1}) - R(\boldsymbol{w}_{\lambda_2}) \le \lambda_1 \left[m(\boldsymbol{w}_{\lambda_2})^2 - m(\boldsymbol{w}_{\lambda_1})^2 \right] \le 0.$$

This is strict when $\lambda_1 > 0$. For $\lambda_1 = 0$, it is strict because $\boldsymbol{w}_0$ uniquely minimizes R and $\boldsymbol{w}_{\lambda_2} \neq \boldsymbol{w}_0$. Thus,

$$R(\boldsymbol{w}_{\lambda_2}) > R(\boldsymbol{w}_{\lambda_1}).$$

Finally, Eq. (44) gives $m(\boldsymbol{w}_\infty) = 0$ and $R(\boldsymbol{w}_\infty) = 1$. Every finite-$\lambda_1$ kernel has at least two positive coordinates, so

$$m(\boldsymbol{w}_\infty) = 0 < m(\boldsymbol{w}_{\lambda_1}),$$
$$R(\boldsymbol{w}_\infty) = 1 > R(\boldsymbol{w}_{\lambda_1}).$$

□

For the PTQ analysis below, we record the closed-form coordinates of each fixed-support segment of the LNWA frontier.

Lemma A.5 (Closed-Form Coordinates of the LNWA Frontier)**.** *On an active prefix of size $q \ge 2$,*

$$m(\boldsymbol{w}_\lambda)^2 = \frac{\bar{d}_q^2}{(1 + D_q\lambda)^2}, \qquad R(\boldsymbol{w}_\lambda) = \frac{1}{q} + \frac{D_q \bar{d}_q^2 \lambda^2}{(1 + D_q\lambda)^2}. \tag{49}$$

Proof. Eq. (39) gives the first identity. Using $\sum_{d=0}^{q-1}(\bar{d}_q - d) = 0$ and $\sum_{d=0}^{q-1}(\bar{d}_q - d)^2 = D_q$ in Eq. (38) gives

$$\begin{aligned} R(\boldsymbol{w}_\lambda) &= \sum_{d=0}^{q-1} \left[q^{-1} + \lambda m_q(\lambda)(\bar{d}_q - d) \right]^2 \\ &= \frac{1}{q} + \frac{2\lambda m_q(\lambda)}{q} \sum_{d=0}^{q-1} (\bar{d}_q - d) + \lambda^2 m_q(\lambda)^2 \sum_{d=0}^{q-1} (\bar{d}_q - d)^2 \\ &= \frac{1}{q} + D_q \lambda^2 m_q(\lambda)^2 = \frac{1}{q} + \frac{D_q \bar{d}_q^2 \lambda^2}{(1 + D_q\lambda)^2}. \end{aligned}$$

□

Theorem A.1 derives the LNWA kernel and its finite-λ Pareto path. Lemma A.4 then characterizes the global structure of this path, including its endpoints, support transitions, and monotonicity in λ. Lemma A.5 gives its closed-form lag–noise coordinates on each fixed-support segment.

A.4 Classical averaging kernels as LNWA special cases

This subsection shows that classical averaging kernels can be viewed as special cases of the LNWA framework. We first present its two endpoints, LAWA (uniform averaging) and Stable (the latest weight at the end of the stable-LR training trajectory), followed by WMA (linear weighting) and EMA (exponential weighting).

Proposition A.1 (The two endpoints)**.** *The $\lambda = 0$ endpoint of the extended* LNWA *path is* LAWA*, whereas its $\lambda = \infty$ endpoint is Stable. Their lag–noise coordinates are therefore*

$$\Phi(\boldsymbol{w}^{\mathrm{LAWA}}) = \left(\frac{M-1}{2}, \frac{1}{M} \right), \qquad \Phi(\boldsymbol{w}^{\mathrm{Stable}}) = (0, 1).$$

This proposition is established by Lemma A.4. We next locate the linearly weighted rule on the interior of the frontier.

Proposition A.2 (WMA on the LNWA Frontier)**.** *The* WMA *kernel* $\boldsymbol{w}^{\mathrm{WMA}} \in \mathcal{W}_M$ *has coordinates* $(\boldsymbol{w}^{\mathrm{WMA}})_d = 2(M-d)/[M(M+1)]$*. Its lag–noise coordinates and corresponding* LNWA *coefficient are*

$$m_{\mathrm{WMA}} = \frac{M-1}{3}, \qquad R_{\mathrm{WMA}} = \frac{2(2M+1)}{3M(M+1)}, \qquad \lambda_{\mathrm{WMA}} = \frac{6}{M(M^2-1)}.$$

Consequently, $\boldsymbol{w}^{\mathrm{WMA}} = \boldsymbol{w}_{\lambda_{\mathrm{WMA}}}$.

Proof. Substituting the WMA coefficients into the definitions of m and R gives

$$m_{\mathrm{WMA}} = \sum_{d=0}^{M-1} d\,(\boldsymbol{w}^{\mathrm{WMA}})_d = \frac{2}{M(M+1)}\left(M\sum_{d=0}^{M-1} d - \sum_{d=0}^{M-1} d^2\right) = \frac{M-1}{3},$$
$$R_{\mathrm{WMA}} = \sum_{d=0}^{M-1} \left((\boldsymbol{w}^{\mathrm{WMA}})_d\right)^2 = \frac{4}{M^2(M+1)^2}\sum_{d=0}^{M-1}(M-d)^2 = \frac{2(2M+1)}{3M(M+1)}.$$

On the full-support face, both kernels are affine in d:

$$(\boldsymbol{w}_\lambda)_d = \frac{1}{M} + \lambda m(\boldsymbol{w}_\lambda)(\bar{d}_M - d),$$
$$(\boldsymbol{w}^{\mathrm{WMA}})_d = \frac{2}{M+1} - \frac{2}{M(M+1)}d. \tag{50}$$

Matching the coefficients of d gives

$$\lambda_{\mathrm{WMA}} m_{\mathrm{WMA}} = \frac{2}{M(M+1)} \qquad \Longrightarrow \qquad \lambda_{\mathrm{WMA}} = \frac{6}{M(M^2-1)}.$$

For this coefficient, the intercept of the LNWA kernel is

$$\frac{1}{M} + \lambda_{\mathrm{WMA}} m_{\mathrm{WMA}} \bar{d}_M = \frac{1}{M} + \frac{2}{M(M+1)}\frac{M-1}{2} = \frac{2}{M+1},$$

which equals the WMA intercept in Eq. (50). Hence $\boldsymbol{w}^{\mathrm{WMA}} = \boldsymbol{w}_{\lambda_{\mathrm{WMA}}}$. □

The next two propositions give the exact coordinates of EMA and its LNWA approximation.

Proposition A.3 (Exact Lag-Noise Coordinates of EMA)**.** *For* $M \geq 2$ *and* $0 < \beta < 1$*, the* EMA *kernel is*

$$(\boldsymbol{w}^{\mathrm{EMA}})_d = \frac{(1-\beta)\beta^d}{1-\beta^M},$$

with exact coordinates

$$m_{\mathrm{EMA}} = \frac{\beta - M\beta^M + (M-1)\beta^{M+1}}{(1-\beta)(1-\beta^M)},$$
$$R_{\mathrm{EMA}} = \frac{(1-\beta)(1+\beta^M)}{(1+\beta)(1-\beta^M)}.$$

Proof. The finite geometric series gives

$$\sum_{d=0}^{M-1} (\boldsymbol{w}^{\mathrm{EMA}})_d = \frac{1-\beta}{1-\beta^M}\sum_{d=0}^{M-1}\beta^d = 1.$$

Substitution into the definitions of m and R yields

$$m_{\mathrm{EMA}} = \frac{1-\beta}{1-\beta^M}\beta\frac{\mathrm{d}}{\mathrm{d}\beta}\left(\frac{1-\beta^M}{1-\beta}\right) = \frac{\beta - M\beta^M + (M-1)\beta^{M+1}}{(1-\beta)(1-\beta^M)},$$
$$R_{\mathrm{EMA}} = \left(\frac{1-\beta}{1-\beta^M}\right)^2 \sum_{d=0}^{M-1}\beta^{2d} = \left(\frac{1-\beta}{1-\beta^M}\right)^2 \frac{1-\beta^{2M}}{1-\beta^2} = \frac{(1-\beta)(1+\beta^M)}{(1+\beta)(1-\beta^M)}.$$

□

Proposition A.4 (EMA on the LNWA Frontier). *For $M = 2$,* EMA *is exactly an* LNWA *kernel for every $0 < \beta < 1$, with $\lambda = (1-\beta)/\beta$. For $M > 2$, let $\eta = -\log\beta$. For any $0 < \rho \le 1$ such that $\eta(M-1) \le \rho$, define*

$$\lambda_{\text{EMA}} := \frac{\eta}{M\bar{d}_M - \eta D_M}.$$

Then the approximation residual satisfies

$$\varepsilon_{\text{EMA}} := \left\|\boldsymbol{w}^{\text{EMA}} - \boldsymbol{w}_{\lambda_{\text{EMA}}}\right\|_\infty \le \frac{e^{\rho/2}\eta^2}{2M}\left[\frac{(M-1)^2}{4} + \left(1+\frac{\rho}{2}\right)\frac{M^2-1}{12}\right] = \mathcal{O}(\eta^2 M). \quad (51)$$

Thus, EMA *is an exact special case of* LNWA *when $M = 2$ and an approximate special case with the above residual bound when $M > 2$.*

Proof. For $M = 2$, direct substitution gives

$$\boldsymbol{w}^{\text{EMA}} = \left(\frac{1}{1+\beta}, \frac{\beta}{1+\beta}\right) = \left(\frac{\lambda+1}{\lambda+2}, \frac{1}{\lambda+2}\right) = \boldsymbol{w}_\lambda, \qquad \lambda = \frac{1-\beta}{\beta}.$$

For $M > 2$, $\eta(M-1) \le \rho \le 1$ gives

$$M\bar{d}_M - \eta D_M = \frac{M(M-1)}{2}\left(1 - \frac{\eta(M+1)}{6}\right) \ge \frac{M(M-1)}{2}\left(1 - \frac{\rho(M+1)}{6(M-1)}\right) > 0,$$

so λ_{EMA} is well defined and positive. It remains to verify that the corresponding LNWA kernel has full support. By Lemma A.4, this is equivalent to $\lambda_{\text{EMA}} < \lambda_M^{\text{knot}}$. Substituting the definitions of λ_{EMA} and λ_M^{knot} gives

$$\lambda_{\text{EMA}} < \lambda_M^{\text{knot}} \iff \eta(M-1) < 2.$$

Because $\eta(M-1) \le \rho \le 1$, this condition holds. Substituting λ_{EMA} into Eq. (39) with $q = M$ gives

$$\lambda_{\text{EMA}} m_M(\lambda_{\text{EMA}}) = \frac{\lambda_{\text{EMA}}\bar{d}_M}{1 + \lambda_{\text{EMA}} D_M} = \frac{\eta}{M}. \quad (52)$$

Substituting Eq. (52) into Eq. (38) with $q = M$ gives

$$(\boldsymbol{w}_{\lambda_{\text{EMA}}})_d = \frac{1}{M} - \frac{\eta}{M}(d - \bar{d}_M).$$

Define

$$r_d := e^{-\eta(d-\bar{d}_M)} - \left[1 - \eta(d-\bar{d}_M)\right], \qquad \bar{r} := \frac{1}{M}\sum_{j=0}^{M-1} r_j.$$

Since $|\eta(d-\bar{d}_M)| \le \rho/2$, we can bound each r_d; averaging these bounds then gives the bound for $\bar{r}$:

$$0 \le r_d \le \frac{e^{\rho/2}\eta^2(d-\bar{d}_M)^2}{2}, \qquad 0 \le \bar{r} \le \frac{e^{\rho/2}\eta^2 D_M}{2M}.$$

Using $\sum_{j=0}^{M-1}(j - \bar{d}_M) = 0$, the normalized EMA coefficient is exactly

$$(\boldsymbol{w}^{\text{EMA}})_d = \frac{e^{-\eta(d-\bar{d}_M)}}{\sum_{j=0}^{M-1} e^{-\eta(j-\bar{d}_M)}} = \frac{1 - \eta(d-\bar{d}_M) + r_d}{M(1+\bar{r})}.$$

Therefore,

$$\begin{aligned}\left|(\boldsymbol{w}^{\text{EMA}})_d - (\boldsymbol{w}_{\lambda_{\text{EMA}}})_d\right| &= \frac{\left|r_d - \left[1-\eta(d-\bar{d}_M)\right]\bar{r}\right|}{M(1+\bar{r})} \\ &\le \frac{e^{\rho/2}\eta^2}{2M}\left[(d-\bar{d}_M)^2 + \left(1+\frac{\rho}{2}\right)\frac{D_M}{M}\right].\end{aligned}$$

Taking the maximum over d and using $|d - \bar{d}_M| \le (M-1)/2$ and $D_M/M = (M^2-1)/12$ proves Eq. (51). □

Together, the preceding results establish the lag-noise decomposition, characterize the LNWA frontier, and relate classical averaging methods to LNWA. We next analyze how weight averaging affects PTQ, leading to our two main theorems.

A.5 PTQ TOLERANCE AND DEGRADATION

We now ask how far an averaging kernel can move toward the newest weight before quantization performance begins to degrade. We measure performance by the expected output reconstruction error relative to the latest latent mean. We first compute the error introduced by the two quantization stages of simplified GPTQ model. This gives an exact expression for the expected reconstruction error. We then study this expression along the LNWA frontier and locate the transition from noise tolerance to degradation.

We use the latent-mean decomposition and write the mean of an averaged state as

$$\boldsymbol{\mu}_T^{\boldsymbol{w}} := \sum_{d=0}^{M-1} w_d \boldsymbol{\mu}_{T-d} = \boldsymbol{\mu}_T - \boldsymbol{b}(\boldsymbol{w}), \qquad \boldsymbol{\theta}_T^{\boldsymbol{w}} = \boldsymbol{\mu}_T^{\boldsymbol{w}} + \boldsymbol{\epsilon}_T^{\boldsymbol{w}}.$$

The term $-\boldsymbol{b}(\boldsymbol{w})$ is the displacement due to temporal lag, and $\boldsymbol{\epsilon}_T^{\boldsymbol{w}}$ is the retained spatial noise. Both will enter the output reconstruction error through the channel matrices defined below.

Assumption A.3 (Bounded Isotropic Spatial Noise)**.** In addition to Assumption A.1, assume

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_P, \qquad \sigma > 0, \qquad |[\boldsymbol{\epsilon}_{T-d}]_j| \le \sigma K_\epsilon \quad (0 \le d < M,\ 1 \le j \le P)$$

almost surely, for a fixed finite K_ϵ.

Thus the averaged spatial noise has

$$\mathbb{E}\boldsymbol{\epsilon}_T^{\boldsymbol{w}} = \boldsymbol{0}, \qquad \mathrm{Cov}(\boldsymbol{\epsilon}_T^{\boldsymbol{w}}) = \sigma^2 R(\boldsymbol{w}) \boldsymbol{I}_P.$$

We measure the effect of a parameter perturbation by the reconstruction error it induces on the fixed calibration inputs. For each linear layer ℓ, let i index its output channels and $n^{(\ell)}$ denote the number of calibration samples. Let $\boldsymbol{X}^{(\ell)} \in \mathbb{R}^{d_{\mathrm{in}}^{(\ell)} \times n^{(\ell)}}$ contain its fixed calibration inputs. The channel vector $\boldsymbol{\theta}^{(\ell,i)}$ is the transpose of row i of that layer's parameter matrix. A change $\boldsymbol{z}^{(\ell,i)}$ causes mean squared reconstruction error

$$\frac{1}{n^{(\ell)}} \|(\boldsymbol{z}^{(\ell,i)})^\top \boldsymbol{X}^{(\ell)}\|_2^2 = (\boldsymbol{z}^{(\ell,i)})^\top \boldsymbol{G}^{(\ell,i)} \boldsymbol{z}^{(\ell,i)}, \qquad \boldsymbol{G}^{(\ell,i)} := \frac{\boldsymbol{X}^{(\ell)} (\boldsymbol{X}^{(\ell)})^\top}{n^{(\ell)}}.$$

Let $\mathcal{I}$ be the set of layer–channel pairs included in the model, and stack their parameter vectors in a fixed order. The sum of their reconstruction errors is $\boldsymbol{z}^\top \boldsymbol{G} \boldsymbol{z}$, where

$$\boldsymbol{G} = \mathrm{diag}_{(\ell,i)\in\mathcal{I}} \boldsymbol{G}^{(\ell,i)} = \begin{pmatrix} \boldsymbol{G}^{(\ell_1,i_1)} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{G}^{(\ell_2,i_2)} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}, \qquad \|\boldsymbol{z}\|_{\boldsymbol{G}}^2 := \boldsymbol{z}^\top \boldsymbol{G} \boldsymbol{z}.$$

The global matrix is block diagonal across channels. The stacked parameter dimension is $P = \sum_{(\ell,i)\in\mathcal{I}} d_{\mathrm{in}}^{(\ell)}$.

Write $Q_b(\boldsymbol{\theta}; s)$ for the model produced by the simplified GPTQ model at precision b and scale s, whose channel operations are specified below. We evaluate each averaging kernel using the expected reconstruction error

$$\mathcal{E}_b(\boldsymbol{w}) := \frac{1}{2} \mathbb{E}\left[\|Q_b(\boldsymbol{\theta}_T^{\boldsymbol{w}}; s(\boldsymbol{w})) - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2 \right]. \tag{53}$$

This sums the squared channel-output differences between the quantized averaged model and the latest latent mean. It includes the temporal lag and spatial noise already present before quantization, together with the change caused by quantization. We assume that every channel matrix $\boldsymbol{G}^{(\ell,i)}$ is positive definite.

The simplified GPTQ model quantizes one group of coordinates and uses the other to compensate for its output error. Partition each channel into two nonempty groups f and r. Reorder the stacked coordinates so that all f entries precede all r entries, and apply the same permutation to every global vector and matrix. We retain the symbols $\boldsymbol{\theta}$ and $\boldsymbol{G}$ in this order. Then

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_{\mathrm{f}} \\ \boldsymbol{\theta}_{\mathrm{r}} \end{pmatrix}, \qquad \boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_{\mathrm{f,f}} & \boldsymbol{G}_{\mathrm{f,r}} \\ \boldsymbol{G}_{\mathrm{r,f}} & \boldsymbol{G}_{\mathrm{r,r}} \end{pmatrix}.$$

To simplify the theoretical analysis, we do not model group-wise quantization scales. Instead, we assign the averaged state a single scale, computed before rounding from its RMS magnitude.

Assumption A.4 (Fixed RMS Scale)**.** For each averaging kernel $\boldsymbol{w}$, the scale is

$$s(\boldsymbol{w})^2 := \frac{1}{P}\mathbb{E}\|\boldsymbol{\theta}_T^{\boldsymbol{w}}\|_2^2 = \frac{\|\boldsymbol{\mu}_T^{\boldsymbol{w}}\|_2^2}{P} + \sigma^2 R(\boldsymbol{w}).$$

For each group $\mathrm{g} \in \mathrm{f}, \mathrm{r}$, we use a fixed range multiplier $H_{\mathrm{g}} > 0$, shared across kernels and precisions. Given the RMS scale $s(\boldsymbol{w})$, the corresponding parameter-space and normalized quantization steps are

$$\Delta_{b,\mathrm{g}}(\boldsymbol{w}) := \frac{2H_{\mathrm{g}}s(\boldsymbol{w})}{2^b - 1}, \qquad \bar{\Delta}_{b,\mathrm{g}} := \frac{\Delta_{b,\mathrm{g}}(\boldsymbol{w})}{s(\boldsymbol{w})} = \frac{2H_{\mathrm{g}}}{2^b - 1}.$$

Accordingly, the 2^b reconstruction levels for group g cover the range

$$[-H_{\mathrm{g}}s(\boldsymbol{w}),\ H_{\mathrm{g}}s(\boldsymbol{w})].$$

Deterministic rounding generally leads to less tractable rounding-error statistics (Alvarez et al., 2025; Zhang & Saab, 2026). To simplify the analysis, following Wannamaker et al. (2000), we instead use nonsubtractive triangular dither and define

$$\mathrm{round}(t) := \left\lfloor t + U + V + \frac{1}{2} \right\rfloor, \qquad U, V \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[-1/2, 1/2].$$

Here $\mathrm{Unif}[a, b]$ denotes the uniform distribution on $[a, b]$, and $\lfloor\cdot\rfloor$ denotes the floor operator, so $\mathrm{round}(t)$ returns an integer code.

For an input $\boldsymbol{z}$, a scale $s > 0$, and group range H_{g}, write $\Delta = s\bar{\Delta}_{b,\mathrm{g}}$. Quantization assigns the integer code and maps it back to parameter space as

$$k_j = \mathrm{clip}_{[0,2^b-1]}\left(\mathrm{round}\left(\frac{z_j}{\Delta} + \frac{2^b - 1}{2}\right)\right), \qquad [\mathrm{quant}_b(\boldsymbol{z}; s, H_{\mathrm{g}})]_j = \Delta\left(k_j - \frac{2^b - 1}{2}\right).$$

Here clip restricts the integer code to the displayed range. We next derive the compensation and write the full-model simplified GPTQ. At the common scale s, the first group has rounding error

$$\boldsymbol{e}_{\mathrm{f}} := \mathrm{quant}_b(\boldsymbol{\theta}_{\mathrm{f}}; s, H_{\mathrm{f}}) - \boldsymbol{\theta}_{\mathrm{f}}.$$

The remaining group is adjusted to minimize the output reconstruction error caused by this first rounding:

$$\boldsymbol{c}_{\mathrm{r}}^\star := \operatorname*{argmin}_{\boldsymbol{c}_{\mathrm{r}}} \begin{pmatrix} \boldsymbol{e}_{\mathrm{f}} \\ \boldsymbol{c}_{\mathrm{r}} \end{pmatrix}^\top \boldsymbol{G} \begin{pmatrix} \boldsymbol{e}_{\mathrm{f}} \\ \boldsymbol{c}_{\mathrm{r}} \end{pmatrix}. \tag{54}$$

Differentiating this quadratic objective gives

$$\boldsymbol{G}_{\mathrm{r,r}}\boldsymbol{c}_{\mathrm{r}}^\star + \boldsymbol{G}_{\mathrm{r,f}}\boldsymbol{e}_{\mathrm{f}} = \boldsymbol{0}.$$

Since $\boldsymbol{G}_{\mathrm{r,r}} \succ 0$, the compensation is

$$\boldsymbol{c}_{\mathrm{r}}^\star = -\boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}, \qquad \boldsymbol{K} := \boldsymbol{G}_{\mathrm{r,r}}^{-1}\boldsymbol{G}_{\mathrm{r,f}}.$$

The matrix $\boldsymbol{K}$ specifies how first-group rounding changes the remaining group. Write $B_K := \|\boldsymbol{K}\|_\infty$, where the matrix infinity norm is the maximum absolute row sum. The second rounding, applied after compensation, has error

$$\boldsymbol{\eta}_{\mathrm{r}} := \mathrm{quant}_b(\boldsymbol{\theta}_{\mathrm{r}} - \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}; s, H_{\mathrm{r}}) - (\boldsymbol{\theta}_{\mathrm{r}} - \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}).$$

Thus the complete model-level map is

$$Q_b(\boldsymbol{\theta}; s) = \begin{pmatrix} \boldsymbol{\theta}_{\mathrm{f}} + \boldsymbol{e}_{\mathrm{f}} \\ \boldsymbol{\theta}_{\mathrm{r}} - \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}} + \boldsymbol{\eta}_{\mathrm{r}} \end{pmatrix}, \qquad \boldsymbol{e}_{\mathrm{f}} = \mathrm{quant}_b(\boldsymbol{\theta}_{\mathrm{f}}; s, H_{\mathrm{f}}) - \boldsymbol{\theta}_{\mathrm{f}}. \tag{55}$$

We assume that the quantizer codebook has a sufficiently wide range to represent the parameter values in both quantization stages.

Assumption A.5 (Rounding Inputs within the Quantizer Range)**.** Let $b_{\min} = 2$. For every $\boldsymbol{w} \in \mathcal{W}_M$ and integer $b \geq b_{\min}$, the two rounding inputs satisfy

$$\frac{\|(\boldsymbol{\theta}_T^{\boldsymbol{w}})_{\mathrm{f}}\|_\infty}{s(\boldsymbol{w})} < H_{\mathrm{f}}\left(1 - \frac{1}{2^b - 1}\right), \qquad \frac{\|(\boldsymbol{\theta}_T^{\boldsymbol{w}})_{\mathrm{r}} - \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}\|_\infty}{s(\boldsymbol{w})} < H_{\mathrm{r}}\left(1 - \frac{1}{2^b - 1}\right)$$

almost surely over the training distribution and the first-stage rounding randomness.

Lemma A.6 (A Sufficient Quantization Range)**.** *Under Assumptions A.3 and A.4, define*

$$Z_{\max} := \sqrt{P} + K_\epsilon \sqrt{M}.$$

If

$$H_{\mathrm{f}} > \frac{3}{2} Z_{\max}, \qquad H_{\mathrm{r}} > \frac{3}{2} \left(Z_{\max} + B_K H_{\mathrm{f}} \right),$$

then Assumption A.5 holds.

Proof. For any $\boldsymbol{w} \in \mathcal{W}_M$, $R(\boldsymbol{w}) \geq 1/M$. Hence $s(\boldsymbol{w}) \geq \sigma/\sqrt{M}$ and $\|\boldsymbol{\mu}_T^{\boldsymbol{w}}\|_2 / s(\boldsymbol{w}) \leq \sqrt{P}$. For each coordinate,

$$\begin{aligned} \frac{|[\boldsymbol{\theta}_T^{\boldsymbol{w}}]_j|}{s(\boldsymbol{w})} &\leq \frac{\|\boldsymbol{\mu}_T^{\boldsymbol{w}}\|_2}{s(\boldsymbol{w})} + \frac{\sum_d w_d |[\boldsymbol{\epsilon}_{T-d}]_j|}{s(\boldsymbol{w})} \\ &\leq \sqrt{P} + \frac{\sigma K_\epsilon}{s(\boldsymbol{w})} \\ &\leq \sqrt{P} + K_\epsilon \sqrt{M} \\ &= Z_{\max}. \end{aligned}$$

Because $b \geq b_{\min} = 2$, $1 - 1/(2^b - 1) \geq 2/3$. The first-stage input therefore satisfies

$$\frac{\|(\boldsymbol{\theta}_T^{\boldsymbol{w}})_{\mathrm{f}}\|_\infty}{s(\boldsymbol{w})} \leq Z_{\max} < \frac{2}{3} H_{\mathrm{f}} \leq H_{\mathrm{f}} \left(1 - \frac{1}{2^b - 1} \right).$$

For any scalar input z satisfying the corresponding margin

$$\frac{|z|}{s(\boldsymbol{w})} < H_{\mathrm{g}} \left(1 - \frac{1}{2^b - 1} \right),$$

we have

$$\frac{1}{2} < \frac{z}{\Delta_{b,\mathrm{g}}} + \frac{2^b - 1}{2} < 2^b - \frac{3}{2}.$$

Since $-1 \leq U + V \leq 1$, it follows that

$$0 < \frac{z}{\Delta_{b,\mathrm{g}}} + \frac{2^b - 1}{2} + U + V + \frac{1}{2} < 2^b.$$

Its floor therefore belongs to $\{0, \ldots, 2^b - 1\}$, so clipping is inactive. The first-stage margin established above thus ensures that the perturbation has magnitude at most $\Delta_{b,\mathrm{f}}$, and nearest-integer rounding adds at most $\Delta_{b,\mathrm{f}}/2$. Thus

$$\|\boldsymbol{e}_{\mathrm{f}}\|_\infty \leq \frac{3}{2} \Delta_{b,\mathrm{f}}(\boldsymbol{w}), \qquad \frac{\|\boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}\|_\infty}{s(\boldsymbol{w})} \leq \frac{3 B_K H_{\mathrm{f}}}{2^b - 1}.$$

The second input consequently satisfies

$$\begin{aligned} \frac{\|(\boldsymbol{\theta}_T^{\boldsymbol{w}})_{\mathrm{r}} - \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}\|_\infty}{s(\boldsymbol{w})} &\leq Z_{\max} + \frac{3 B_K H_{\mathrm{f}}}{2^b - 1} \\ &\leq Z_{\max} + B_K H_{\mathrm{f}} \\ &< \frac{2}{3} H_{\mathrm{r}} \\ &\leq H_{\mathrm{r}} \left(1 - \frac{1}{2^b - 1} \right). \end{aligned}$$

This is the second inequality in Assumption A.5; the first was established above. Hence that assumption holds for both quantization stages. □

We next derive the parameter error induced by the simplified GPTQ procedure and its corresponding output reconstruction error.

$$\boldsymbol{C}_{\mathrm{f}} := \boldsymbol{G}_{\mathrm{f,f}} - \boldsymbol{G}_{\mathrm{f,r}} \boldsymbol{K}.$$

Lemma A.7 (Parameter and Reconstruction Errors of the Simplified GPTQ Model)**.** *The Schur complement $\boldsymbol{C}_{\mathrm{f}}$ is positive definite. The quantization-induced global parameter error is*

$$\boldsymbol{\delta}_b(\boldsymbol{\theta}; s) := Q_b(\boldsymbol{\theta}; s) - \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{e}_{\mathrm{f}} \\ -\boldsymbol{K}\boldsymbol{e}_{\mathrm{f}} + \boldsymbol{\eta}_{\mathrm{r}} \end{pmatrix}.$$

Its output reconstruction error satisfies

$$\boldsymbol{\delta}_b(\boldsymbol{\theta}; s)^\top \boldsymbol{G} \boldsymbol{\delta}_b(\boldsymbol{\theta}; s) = \boldsymbol{e}_{\mathrm{f}}^\top \boldsymbol{C}_{\mathrm{f}} \boldsymbol{e}_{\mathrm{f}} + \boldsymbol{\eta}_{\mathrm{r}}^\top \boldsymbol{G}_{\mathrm{r,r}} \boldsymbol{\eta}_{\mathrm{r}}. \tag{56}$$

Proof. Because $\boldsymbol{G} \succ 0$, its principal block $\boldsymbol{G}_{\mathrm{r,r}}$ is positive definite. Let J be the global reconstruction objective in Eq. (54). For a fixed first-group error, expanding and differentiating it gives

$$\begin{aligned} J(\boldsymbol{c}_{\mathrm{r}}) &= \boldsymbol{e}_{\mathrm{f}}^\top \boldsymbol{G}_{\mathrm{f,f}} \boldsymbol{e}_{\mathrm{f}} + 2\boldsymbol{c}_{\mathrm{r}}^\top \boldsymbol{G}_{\mathrm{r,f}} \boldsymbol{e}_{\mathrm{f}} + \boldsymbol{c}_{\mathrm{r}}^\top \boldsymbol{G}_{\mathrm{r,r}} \boldsymbol{c}_{\mathrm{r}}, \\ \nabla_{\boldsymbol{c}_{\mathrm{r}}} J &= 2\boldsymbol{G}_{\mathrm{r,f}} \boldsymbol{e}_{\mathrm{f}} + 2\boldsymbol{G}_{\mathrm{r,r}} \boldsymbol{c}_{\mathrm{r}}. \end{aligned}$$

Setting this derivative to zero recovers $\boldsymbol{c}_{\mathrm{r}}^\star = -\boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}$.

Completing the square decomposes the reconstruction error into the residual error after optimal compensation and the additional error from a non-optimal compensation:

$$J(\boldsymbol{c}_{\mathrm{r}}) = \underbrace{\boldsymbol{e}_{\mathrm{f}}^\top \boldsymbol{C}_{\mathrm{f}} \boldsymbol{e}_{\mathrm{f}}}_{\text{residual error}} + \underbrace{(\boldsymbol{c}_{\mathrm{r}} + \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}})^\top \boldsymbol{G}_{\mathrm{r,r}} (\boldsymbol{c}_{\mathrm{r}} + \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}})}_{\text{compensation error}}. \tag{57}$$

This identity holds for every first-group rounding error $\boldsymbol{e}_{\mathrm{f}}$. At the optimum, $\boldsymbol{c}_{\mathrm{r}}^\star = -\boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}$, so the second term in Eq. (57) vanishes and

$$\boldsymbol{e}_{\mathrm{f}}^\top \boldsymbol{C}_{\mathrm{f}} \boldsymbol{e}_{\mathrm{f}} = \begin{pmatrix} \boldsymbol{e}_{\mathrm{f}} \\ -\boldsymbol{K}\boldsymbol{e}_{\mathrm{f}} \end{pmatrix}^\top \boldsymbol{G} \begin{pmatrix} \boldsymbol{e}_{\mathrm{f}} \\ -\boldsymbol{K}\boldsymbol{e}_{\mathrm{f}} \end{pmatrix} > 0 \qquad (\boldsymbol{e}_{\mathrm{f}} \neq \boldsymbol{0}).$$

Hence $\boldsymbol{C}_{\mathrm{f}} \succ 0$. The second rounding introduces the additional error $\boldsymbol{\eta}_{\mathrm{r}}$, so that

$$\boldsymbol{c}_{\mathrm{r}} = \boldsymbol{c}_{\mathrm{r}}^\star + \boldsymbol{\eta}_{\mathrm{r}},$$

therefore

$$\boldsymbol{c}_{\mathrm{r}} + \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}} = \boldsymbol{\eta}_{\mathrm{r}},$$

Substituting this expression into Eq. (57) gives

$$J = \boldsymbol{e}_{\mathrm{f}}^\top \boldsymbol{C}_{\mathrm{f}} \boldsymbol{e}_{\mathrm{f}} + \boldsymbol{\eta}_{\mathrm{r}}^\top \boldsymbol{G}_{\mathrm{r,r}} \boldsymbol{\eta}_{\mathrm{r}},$$

which proves Eq. (56). □

Lemma A.7 gives the parameter and reconstruction errors of the simplified GPTQ model. We use this result to analyze the quantization-induced contribution to $\mathcal{E}_b(\boldsymbol{w})$. Let

$$\boldsymbol{a} := \boldsymbol{\theta} - \boldsymbol{\mu}_T = \begin{pmatrix} \boldsymbol{a}_{\mathrm{f}} \\ \boldsymbol{a}_{\mathrm{r}} \end{pmatrix}.$$

we have

$$Q_b(\boldsymbol{\theta}; s) - \boldsymbol{\mu}_T = \boldsymbol{a} + \boldsymbol{\delta}_b(\boldsymbol{\theta}; s),$$

and the change in reconstruction error introduced by quantization is

$$\begin{aligned} &\frac{1}{2} \|\boldsymbol{a} + \boldsymbol{\delta}_b(\boldsymbol{\theta}; s)\|_{\boldsymbol{G}}^2 - \frac{1}{2} \|\boldsymbol{a}\|_{\boldsymbol{G}}^2 \\ &= \boldsymbol{a}_{\mathrm{f}}^\top \boldsymbol{C}_{\mathrm{f}} \boldsymbol{e}_{\mathrm{f}} + \boldsymbol{a}_{\mathrm{f}}^\top \boldsymbol{G}_{\mathrm{f,r}} \boldsymbol{\eta}_{\mathrm{r}} + \boldsymbol{a}_{\mathrm{r}}^\top \boldsymbol{G}_{\mathrm{r,r}} \boldsymbol{\eta}_{\mathrm{r}} + \frac{1}{2} \boldsymbol{e}_{\mathrm{f}}^\top \boldsymbol{C}_{\mathrm{f}} \boldsymbol{e}_{\mathrm{f}} + \frac{1}{2} \boldsymbol{\eta}_{\mathrm{r}}^\top \boldsymbol{G}_{\mathrm{r,r}} \boldsymbol{\eta}_{\mathrm{r}}. \end{aligned} \tag{58}$$

Eq. (58) separates the quantization-induced change into linear and quadratic rounding terms. We first fix the input and take the expectation over the rounding randomness, then take the expectation over the spatial noise. The resulting moments give the exact expected contribution of each term. Partition $\boldsymbol{\mu}_T = (\boldsymbol{\mu}_{T,\mathrm{f}}^\top, \boldsymbol{\mu}_{T,\mathrm{r}}^\top)^\top$ in the same way as $\boldsymbol{\theta}$. Substituting $\boldsymbol{a} = \boldsymbol{\theta} - \boldsymbol{\mu}_T$ into Eq. (58) and expanding the quadratic forms gives

$$\mathbb{E}\left[\frac{1}{2} \|\boldsymbol{a} + \boldsymbol{\delta}_b(\boldsymbol{\theta}; s)\|_{\boldsymbol{G}}^2 - \frac{1}{2} \|\boldsymbol{a}\|_{\boldsymbol{G}}^2\right] = \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4, \tag{59}$$

Here the four terms are

$$\mathcal{T}_1 := \mathbb{E}[\boldsymbol{\theta}_{\mathrm{f}}^\top \boldsymbol{C}_{\mathrm{f}} \boldsymbol{e}_{\mathrm{f}}] + \mathbb{E}[\boldsymbol{\theta}_{\mathrm{f}}^\top \boldsymbol{G}_{\mathrm{f,r}} \boldsymbol{\eta}_{\mathrm{r}}] + \mathbb{E}[\boldsymbol{\theta}_{\mathrm{r}}^\top \boldsymbol{G}_{\mathrm{r,r}} \boldsymbol{\eta}_{\mathrm{r}}],$$
$$\mathcal{T}_2 := -\boldsymbol{\mu}_{T,\mathrm{f}}^\top \boldsymbol{C}_{\mathrm{f}} \mathbb{E}[\boldsymbol{e}_{\mathrm{f}}] - \boldsymbol{\mu}_{T,\mathrm{f}}^\top \boldsymbol{G}_{\mathrm{f,r}} \mathbb{E}[\boldsymbol{\eta}_{\mathrm{r}}] - \boldsymbol{\mu}_{T,\mathrm{r}}^\top \boldsymbol{G}_{\mathrm{r,r}} \mathbb{E}[\boldsymbol{\eta}_{\mathrm{r}}],$$
$$\mathcal{T}_3 := \frac{1}{2} \sum_j (\boldsymbol{C}_{\mathrm{f}})_{jj} \mathbb{E}[(\boldsymbol{e}_{\mathrm{f}})_j^2] + \frac{1}{2} \sum_j (\boldsymbol{G}_{\mathrm{r,r}})_{jj} \mathbb{E}[(\boldsymbol{\eta}_{\mathrm{r}})_j^2],$$
$$\mathcal{T}_4 := \frac{1}{2} \sum_{j \neq k} (\boldsymbol{C}_{\mathrm{f}})_{jk} \mathbb{E}[(\boldsymbol{e}_{\mathrm{f}})_j (\boldsymbol{e}_{\mathrm{f}})_k] + \frac{1}{2} \sum_{j \neq k} (\boldsymbol{G}_{\mathrm{r,r}})_{jk} \mathbb{E}[(\boldsymbol{\eta}_{\mathrm{r}})_j (\boldsymbol{\eta}_{\mathrm{r}})_k].$$

$\mathcal{T}_1$ captures the coupling between the input parameters and rounding errors, while $\mathcal{T}_2$ contains rounding errors multiplied by the fixed reference $\boldsymbol{\mu}_T$. The remaining terms are quadratic in the rounding errors: $\mathcal{T}_3$ collects coordinatewise squares, and $\mathcal{T}_4$ collects cross-coordinate products. The following lemma gives the conditional moments needed to evaluate these terms.

Lemma A.8 (Moments of Triangular-Dither Rounding)**.** *Under Assumption A.5, let $\boldsymbol{z}$ be either the first-stage input $\boldsymbol{\theta}_{\mathrm{f}}$ or the compensated second-stage input $\boldsymbol{\theta}_{\mathrm{r}} - \boldsymbol{K}\boldsymbol{e}_{\mathrm{f}}$, with its corresponding fixed step $\Delta > 0$. Let e_j denote the parameter-space error introduced by $\mathrm{quant}_b(\cdot)$ in coordinate j. For any integrable coefficient $\phi(\boldsymbol{z})$, the conditional moments corresponding to $\mathcal{T}_1, \ldots, \mathcal{T}_4$ are*

$$\mathbb{E}[\phi(\boldsymbol{z}) e_j \mid \boldsymbol{z}] = 0,$$
$$\mathbb{E}[e_j \mid \boldsymbol{z}] = 0,$$
$$\mathbb{E}[e_j^2 \mid \boldsymbol{z}] = \frac{\Delta^2}{4},$$
$$\mathbb{E}[e_j e_k \mid \boldsymbol{z}] = 0 \qquad (j \neq k).$$

Here the conditional expectation averages over the fresh rounding draws, which are independent across coordinates.

Proof. Condition on the vector input $\boldsymbol{z}$ and fix a coordinate j. Write $z = z_j$. In parameter space, let u, v be independent uniform variables on $[-\Delta/2, \Delta/2]$. Fix v and put $t = z + v$. Let a and $a + \Delta$ be the neighboring grid points with $a \leq t < a + \Delta$, and set $r = t - a$. Nearest rounding of $t + u$ returns a with probability $1 - r/\Delta$ and $a + \Delta$ with probability r/Δ. If q denotes the output, direct calculation gives

$$\mathbb{E}[q - t \mid z, v] = (-r)\left(1 - \frac{r}{\Delta}\right) + (\Delta - r)\frac{r}{\Delta} = 0, \tag{60}$$
$$\mathbb{E}[(q - t)^2 \mid z, v] = r^2\left(1 - \frac{r}{\Delta}\right) + (\Delta - r)^2\frac{r}{\Delta} = r(\Delta - r).$$

The full error is $e = q - z = (q - t) + v$. Taking the expectation over v therefore yields

$$\mathbb{E}[e \mid z] = \mathbb{E}[v] = 0.$$

As v ranges uniformly over one grid cell, r also ranges uniformly over $[0, \Delta)$, irrespective of z and the grid translation. By Eq. (60), the mixed term satisfies

$$\mathbb{E}[v(q - t) \mid z] = \mathbb{E}[v\,\mathbb{E}[q - t \mid z, v] \mid z] = 0.$$

Consequently,

$$\begin{aligned}
\mathbb{E}[e^2 \mid z] &= \mathbb{E}[(q - t)^2 \mid z] + \mathbb{E}[v^2] + 2\mathbb{E}[v(q - t) \mid z] \\
&= \mathbb{E}[r(\Delta - r)] + \mathbb{E}[v^2] + 0 \\
&= \frac{1}{\Delta}\int_0^\Delta r(\Delta - r)\,\mathrm{d}r + \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} v^2\,\mathrm{d}v \\
&= \frac{\Delta^2}{6} + \frac{\Delta^2}{12} = \frac{\Delta^2}{4}.
\end{aligned}$$

This calculation holds for every coordinate of $\boldsymbol{z}$. Since $\phi(\boldsymbol{z})$ is fixed conditional on $\boldsymbol{z}$,

$$\mathbb{E}[\phi(\boldsymbol{z}) e_j \mid \boldsymbol{z}] = \phi(\boldsymbol{z})\mathbb{E}[e_j \mid \boldsymbol{z}] = 0.$$

Independent rounding draws across coordinates give

$$\mathbb{E}[e_j e_k \mid \boldsymbol{z}] = \mathbb{E}[e_j \mid \boldsymbol{z}]\mathbb{E}[e_k \mid \boldsymbol{z}] = 0 \qquad (j \neq k).$$

This proves all four identities in the lemma. □

For the first group, condition on the full input $\boldsymbol{\theta}$. For the remaining group, condition on both $\boldsymbol{\theta}$ and the first-stage error $\boldsymbol{e}_{\mathrm{f}}$; its compensated input is then fixed and its rounding draws are fresh. With $d_{\mathrm{f}} := \dim \boldsymbol{\theta}_{\mathrm{f}}$ and $d_{\mathrm{r}} := \dim \boldsymbol{\theta}_{\mathrm{r}}$, Lemma A.8 gives

$$\begin{aligned} \mathbb{E}[\boldsymbol{e}_{\mathrm{f}} \mid \boldsymbol{\theta}] = \mathbf{0}, \qquad \mathbb{E}[\boldsymbol{e}_{\mathrm{f}}\boldsymbol{e}_{\mathrm{f}}^{\top} \mid \boldsymbol{\theta}] = \frac{\Delta_{b,\mathrm{f}}(\boldsymbol{w})^2}{4}\boldsymbol{I}_{d_{\mathrm{f}}}, \\ \mathbb{E}[\boldsymbol{\eta}_{\mathrm{r}} \mid \boldsymbol{\theta}, \boldsymbol{e}_{\mathrm{f}}] = \mathbf{0}, \quad \mathbb{E}[\boldsymbol{\eta}_{\mathrm{r}}\boldsymbol{\eta}_{\mathrm{r}}^{\top} \mid \boldsymbol{\theta}, \boldsymbol{e}_{\mathrm{f}}] = \frac{\Delta_{b,\mathrm{r}}(\boldsymbol{w})^2}{4}\boldsymbol{I}_{d_{\mathrm{r}}}. \end{aligned} \tag{61}$$

Eq. (61) gives

$$\mathbb{E}[\boldsymbol{\delta}_b(\boldsymbol{\theta}; s(\boldsymbol{w})) \mid \boldsymbol{\theta}] = \begin{pmatrix} \mathbb{E}[\boldsymbol{e}_{\mathrm{f}} \mid \boldsymbol{\theta}] \\ -\boldsymbol{K}\mathbb{E}[\boldsymbol{e}_{\mathrm{f}} \mid \boldsymbol{\theta}] + \mathbb{E}[\mathbb{E}[\boldsymbol{\eta}_{\mathrm{r}} \mid \boldsymbol{\theta}, \boldsymbol{e}_{\mathrm{f}}] \mid \boldsymbol{\theta}] \end{pmatrix} = \mathbf{0}.$$

Thus the mixed contribution in Eq. (58) has zero expectation. In the four-term expansion, $\mathcal{T}_1 = \mathcal{T}_2 = \mathcal{T}_4 = 0$, and

$$\begin{aligned} \mathcal{T}_3 &= \frac{1}{2}\sum_j (\boldsymbol{C}_{\mathrm{f}})_{jj}\frac{\Delta_{b,\mathrm{f}}(\boldsymbol{w})^2}{4} + \frac{1}{2}\sum_j (\boldsymbol{G}_{\mathrm{r,r}})_{jj}\frac{\Delta_{b,\mathrm{r}}(\boldsymbol{w})^2}{4} \\ &= \frac{\Delta_{b,\mathrm{f}}(\boldsymbol{w})^2}{8}\operatorname{Tr}(\boldsymbol{C}_{\mathrm{f}}) + \frac{\Delta_{b,\mathrm{r}}(\boldsymbol{w})^2}{8}\operatorname{Tr}(\boldsymbol{G}_{\mathrm{r,r}}). \end{aligned}$$

This evaluates the quantization contribution. The next lemma combines it with the temporal lag and spatial noise present before quantization.

Lemma A.9 (Exact Expected Reconstruction Error)**.** *Under Assumptions A.3, A.4, and A.5, the simplified GPTQ model satisfies, for every $\boldsymbol{w} \in \mathcal{W}_M$ and integer $b \geq b_{\min}$,*

$$\mathcal{E}_b(\boldsymbol{w}) = \mathcal{P}_0(\boldsymbol{w}) + \gamma_b \mathcal{Q}(\boldsymbol{w}), \tag{62}$$

where the reconstruction error before quantization is

$$\mathcal{P}_0(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{\mu}_T^{\boldsymbol{w}} - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2 + N_G R(\boldsymbol{w}), \qquad N_G := \frac{\sigma^2}{2}\operatorname{Tr}(\boldsymbol{G}).$$

The quantization contribution separates into the parameter energy

$$\mathcal{Q}(\boldsymbol{w}) := \mathbb{E}\|\boldsymbol{\theta}_T^{\boldsymbol{w}}\|_2^2 = Ps(\boldsymbol{w})^2 = \|\boldsymbol{\mu}_T^{\boldsymbol{w}}\|_2^2 + P\sigma^2 R(\boldsymbol{w})$$

and the precision-dependent coefficient

$$\gamma_b := \frac{c_G}{(2^b-1)^2}, \qquad c_G := \frac{H_{\mathrm{f}}^2 \operatorname{Tr}(\boldsymbol{C}_{\mathrm{f}}) + H_{\mathrm{r}}^2 \operatorname{Tr}(\boldsymbol{G}_{\mathrm{r,r}})}{2P} > 0.$$

Proof. The latent-mean decomposition and zero-mean spatial noise give

$$\frac{1}{2}\mathbb{E}\|\boldsymbol{\theta}_T^{\boldsymbol{w}} - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2 = \frac{1}{2}\|\boldsymbol{\mu}_T^{\boldsymbol{w}} - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2 + \frac{\sigma^2}{2}\operatorname{Tr}(\boldsymbol{G})R(\boldsymbol{w}) = \mathcal{P}_0(\boldsymbol{w}).$$

Set $\boldsymbol{a} = \boldsymbol{\theta}_T^{\boldsymbol{w}} - \boldsymbol{\mu}_T$ and use the conditional zero mean of $\boldsymbol{\delta}_b$. Expanding the definition of $\mathcal{E}_b$ and then applying Eq. (56) gives

$$\begin{aligned} \mathcal{E}_b(\boldsymbol{w}) &= \frac{1}{2}\mathbb{E}\|Q_b(\boldsymbol{\theta}_T^{\boldsymbol{w}}; s(\boldsymbol{w})) - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2 \\ &= \frac{1}{2}\mathbb{E}\|\boldsymbol{a} + \boldsymbol{\delta}_b(\boldsymbol{\theta}_T^{\boldsymbol{w}}; s(\boldsymbol{w}))\|_{\boldsymbol{G}}^2 \\ &= \mathcal{P}_0(\boldsymbol{w}) + \mathbb{E}\big[\boldsymbol{a}^{\top}\boldsymbol{G}\mathbb{E}[\boldsymbol{\delta}_b(\boldsymbol{\theta}_T^{\boldsymbol{w}}; s(\boldsymbol{w})) \mid \boldsymbol{\theta}_T^{\boldsymbol{w}}]\big] + \frac{1}{2}\mathbb{E}\|\boldsymbol{\delta}_b(\boldsymbol{\theta}_T^{\boldsymbol{w}}; s(\boldsymbol{w}))\|_{\boldsymbol{G}}^2 \\ &= \mathcal{P}_0(\boldsymbol{w}) + \frac{1}{8}\left[\Delta_{b,\mathrm{f}}(\boldsymbol{w})^2 \operatorname{Tr}(\boldsymbol{C}_{\mathrm{f}}) + \Delta_{b,\mathrm{r}}(\boldsymbol{w})^2 \operatorname{Tr}(\boldsymbol{G}_{\mathrm{r,r}})\right] \\ &= \mathcal{P}_0(\boldsymbol{w}) + \frac{H_{\mathrm{f}}^2 \operatorname{Tr}(\boldsymbol{C}_{\mathrm{f}}) + H_{\mathrm{r}}^2 \operatorname{Tr}(\boldsymbol{G}_{\mathrm{r,r}})}{2P(2^b-1)^2} Ps(\boldsymbol{w})^2 \\ &= \mathcal{P}_0(\boldsymbol{w}) + \gamma_b \mathcal{Q}(\boldsymbol{w}). \end{aligned}$$

This proves the identity for every averaging kernel in the stated model. □

Eq. (62) connects the expected reconstruction error to the lag-induced displacement of the averaged mean and the retained spatial noise. To make this dependence explicit along the LNWA frontier, we use mean age to parameterize temporal lag, derive the corresponding retained noise factor, and track the averaged latent mean. These relations determine the transition from tolerance to degradation.

Recall that the LNWA frontier is the family $\boldsymbol{w}_\lambda$. The LAWA kernel $\boldsymbol{w}^{\mathrm{LAWA}}$ has mean age $m_0 := m(\boldsymbol{w}^{\mathrm{LAWA}}) = (M-1)/2$. We measure the reduction from LAWA by

$$y := m_0 - m(\boldsymbol{w}_\lambda), \qquad 0 \le y \le m_0.$$

Thus $y = 0$ corresponds to LAWA, while $y = m_0$ corresponds to the most recent saved weight. The frontier coordinates in Lemma A.5 show that mean age decreases strictly as λ increases. Hence each $y < m_0$ determines a unique $\lambda(y)$. To simplify notation, we write $\boldsymbol{w}(y) := \boldsymbol{w}_{\lambda(y)}$ for that kernel. We also write $m(y) := m(\boldsymbol{w}(y)) = m_0 - y$ and $R(y) := R(\boldsymbol{w}(y))$ for its mean age and retained noise factor. For notational consistency, primes below denote derivatives along this parameterized frontier. In particular,

$$\mathcal{E}_b'(y) := \frac{\mathrm{d}}{\mathrm{d}y}\mathcal{E}_b(\boldsymbol{w}(y)), \qquad \mathcal{E}_b''(y) := \frac{\mathrm{d}^2}{\mathrm{d}y^2}\mathcal{E}_b(\boldsymbol{w}(y)),$$

and analogously for $\mathcal{P}_0$ and $\mathcal{Q}$. After the later change of variable to x, primes correspondingly denote derivatives with respect to x. Thus mean weight age decreases linearly with y; the accompanying change in spatial noise is determined next from the frontier kernel.

On a segment with $q \ge 2$ active saved weights, the positive coefficients have ages $0, \dots, q-1$. Uniform averaging over these weights has mean age $\bar{d}_q = (q-1)/2$. The reduction relative to this local baseline is

$$z_q := \bar{d}_q - m(y) = y - \frac{M-q}{2}.$$

Lemma A.5 then gives the kernel coefficients

$$w_d(y) = \begin{cases} \dfrac{1}{q} + \dfrac{z_q(\bar{d}_q - d)}{D_q}, & 0 \le d < q, \\ 0, & d \ge q, \end{cases} \tag{63}$$

where $D_q := q(q^2-1)/12$. Squaring and summing the coefficients therefore gives the spatial noise factor

$$R(y) = \sum_{d=0}^{q-1} w_d(y)^2 = \frac{1}{q} + \frac{z_q^2}{D_q},$$

whose derivatives are

$$R'(y) = \frac{2z_q}{D_q}, \qquad R''(y) = \frac{2}{D_q}.$$

Finally, the original frontier parameter is recovered from $z_q = D_q \lambda m(y)$ as

$$\lambda(y) = \frac{z_q}{D_q(m_0 - y)}. \tag{64}$$

The endpoint is $y = m_0$, with $\lambda(m_0) = \infty$. Write $\boldsymbol{\mu}(y) := \boldsymbol{\mu}_T^{\boldsymbol{w}(y)}$ for the latent mean of this averaged state. It is affine within each support segment because the averaging coefficients are affine there. Using the latent-mean increments defined in Appendix A.1, its derivative is

$$
\begin{aligned}
\boldsymbol{a}_q := \boldsymbol{\mu}'(y) &= \frac{1}{D_q}\sum_{d=0}^{q-1}(\bar{d}_q - d)\boldsymbol{\mu}_{T-d} \\
&= \frac{1}{D_q}\sum_{d=0}^{q-1}(\bar{d}_q - d)\left(\boldsymbol{\mu}_{T-q+1} + \sum_{j=d+1}^{q-1}\boldsymbol{v}_{T-j+1}\right) \\
&= \frac{1}{D_q}\sum_{j=1}^{q-1}\left(\sum_{d=0}^{j-1}(\bar{d}_q - d)\right)\boldsymbol{v}_{T-j+1} \\
&= \frac{1}{D_q}\sum_{j=1}^{q-1}\left(j\bar{d}_q - \frac{j(j-1)}{2}\right)\boldsymbol{v}_{T-j+1} \\
&= \sum_{j=1}^{q-1}\frac{j(q-j)}{2D_q}\boldsymbol{v}_{T-j+1}.
\end{aligned}
$$

The coefficients are nonnegative and satisfy

$$\sum_{j=1}^{q-1}\frac{j(q-j)}{2D_q} = 1.$$

Thus $\boldsymbol{a}_q$ is a convex combination of the retained mean increments; its norm is at most $\bar{v} := \max_{1\le j<M}\|\boldsymbol{v}_{T-j+1}\|_2$. We use this increment bound below to control changes in the averaged mean at support boundaries.

On a fixed-support segment, $\boldsymbol{\mu}'(y) = \boldsymbol{a}_q$ is constant and $R''(y) = 2/D_q$. The two components have second derivatives

$$\mathcal{P}_0''(y) = \boldsymbol{a}_q^\top \boldsymbol{G}\boldsymbol{a}_q + \frac{2N_G}{D_q} > 0,$$

$$\mathcal{Q}''(y) = 2\|\boldsymbol{a}_q\|_2^2 + \frac{2P\sigma^2}{D_q} > 0.$$

Consequently,

$$
\begin{aligned}
\mathcal{E}_b''(y) &= \mathcal{P}_0''(y) + \gamma_b\mathcal{Q}''(y) \\
&= \boldsymbol{a}_q^\top \boldsymbol{G}\boldsymbol{a}_q + 2\gamma_b\|\boldsymbol{a}_q\|_2^2 + \frac{2N_G + 2\gamma_b P\sigma^2}{D_q} > 0.
\end{aligned} \tag{65}
$$

These results describe movement within one support segment. When the oldest positive averaging coefficient reaches zero, the active set changes and $\boldsymbol{\mu}'(y)$ can jump. We next relate these changes to the evolution of the latent mean over the retained window.

Across nearby saved weights in a stable training phase, latent-mean increments are small and change gradually with normalized training time. The following regularity condition bounds their magnitudes and adjacent variation.

Assumption A.6 (Local Regularity of the Latent Mean)**.** Measure training time as t/T. The M consecutive retained weights span the normalized length $\ell := (M-1)/T$, with spacing $h := 1/T = \ell/(M-1)$. The forward latent increments satisfy

$$\|\boldsymbol{v}_{T-j+1}\|_2 \le Vh \quad (1 \le j < M), \qquad \|\boldsymbol{v}_{T-j+1} - \boldsymbol{v}_{T-j+2}\|_2 \le Ah^2 \quad (2 \le j < M),$$

where $V, A \ge 0$ are uniform over the considered late-training windows and independent of the spacing h.

Summing the d increments between $\boldsymbol{\mu}_{T-d}$ and $\boldsymbol{\mu}_T$ gives

$$\|\boldsymbol{\mu}_{T-d} - \boldsymbol{\mu}_T\|_2 \le Vdh.$$

Denote the mean under LAWA by $\bar{\boldsymbol{\mu}} := M^{-1}\sum_{d=0}^{M-1}\boldsymbol{\mu}_{T-d}$, and let $D_\mu := \max_d\|\boldsymbol{\mu}_{T-d} - \boldsymbol{\mu}_T\|_2$. Averaging the preceding bound yields

$$\|\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_T\|_2 \le \frac{1}{M}\sum_{d=0}^{M-1}Vdh = \frac{V\ell}{2}.$$

The second bound in Assumption A.6 verifies Assumption A.2 with $\kappa = Ah^2$. The first bound controls D_μ and the LAWA mean displacement. Collecting these consequences gives

$$\kappa = \frac{A\ell^2}{(M-1)^2}, \qquad D_\mu \le V\ell, \qquad \|\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_T\|_2 \le \frac{V\ell}{2}. \tag{66}$$

At a support transition, both the averaging kernel and the averaged latent mean remain continuous, although the mean slope may change. The following lemma bounds that change and the cumulative effect of the transitions.

Lemma A.10 (Slope Changes at Support Boundaries)**.** *Under Assumption A.2, let y_j, $1 \le j \le M-2$, be the internal support boundaries in increasing order, and write $\Delta\boldsymbol{\mu}'(y_j) := \boldsymbol{\mu}'(y_j+) - \boldsymbol{\mu}'(y_j-)$. At every such boundary,*

$$\|\Delta\boldsymbol{\mu}'(y_j)\|_2 \le \frac{\kappa}{2}, \qquad R'(y_j-) = R'(y_j+).$$

For $0 < y \le m_0$, their positions satisfy

$$\frac{\sum_{y_j \le y} y_j}{y^2} \le H_M, \tag{67}$$

where

$$H_2 = 0, \qquad H_M = \frac{12(M-2)(M-1)}{(3M-5)^2} \quad (M \ge 3).$$

Proof. At a transition from $q+1$ to q active weights, subtracting the two coefficient derivatives in Eq. (63) gives

$$\Delta w'_d = \frac{12(q-1-3d)}{q(q-1)(q+1)(q+2)} \quad (0 \le d < q), \qquad \Delta w'_q = \frac{6}{(q+1)(q+2)}.$$

Set

$$K_{qj} := \frac{6j(j+1)(q-j)}{q(q-1)(q+1)(q+2)} \quad (1 \le j < q), \qquad K_{q0} = K_{qq} = 0.$$

The corresponding change in the mean slope is

$$\Delta\boldsymbol{\mu}'(y_j) = \sum_{r=1}^{q-1} K_{qr}(\boldsymbol{\mu}_{T-r-1} - 2\boldsymbol{\mu}_{T-r} + \boldsymbol{\mu}_{T-r+1}).$$

Expanding the second differences reproduces the coefficient changes: the endpoint coefficients are K_{q1} and $K_{q,q-1}$, and an interior coefficient is $K_{q,j-1} - 2K_{qj} + K_{q,j+1}$. Summing the coefficients gives

$$\sum_{j=1}^{q-1} K_{qj} = \frac{6}{q(q-1)(q+1)(q+2)} \sum_{j=1}^{q-1} j(j+1)(q-j) = \frac{1}{2}.$$

Assumption A.2 therefore gives

$$\|\Delta\boldsymbol{\mu}'(y_j)\|_2 \le \kappa \sum_{j=1}^{q-1} K_{qj} = \frac{\kappa}{2}.$$

The oldest coefficient becomes zero at $z_{q+1} = (q+2)/6$. Substituting this in the definition of z_{q+1} gives $y_j = (M+2j-1)/6$ for $1 \le j \le M-2$. At the same transition, $z_q = (q-1)/6$, and hence

$$R'(y_j-) = \frac{2z_{q+1}}{D_{q+1}} = \frac{4}{q(q+1)} = \frac{2z_q}{D_q} = R'(y_j+).$$

Between consecutive boundaries the numerator in Eq. (67) is constant, and its ratio to y^2 decreases. At the jth boundary the ratio is

$$\frac{\sum_{i=1}^{j} y_i}{y_j^2} = \frac{6j(M+j)}{(M+2j-1)^2}.$$

This expression increases for $1 \le j \le M-2$, treating j as a real variable, the sign of its derivative is the sign of $M(M-1) - 2j > 0$. Its last value is H_M. There are no internal boundaries when $M = 2$.

□

For each precision b, write the curvature on a segment with q active weights as

$$a_{q,b} := \boldsymbol{a}_q^\top \boldsymbol{G} \boldsymbol{a}_q + 2\gamma_b \|\boldsymbol{a}_q\|_2^2 + \frac{2N_G + 2\gamma_b P\sigma^2}{D_q}, \qquad 2 \le q \le M,$$

as given by Eq. (65), and define

$$a_{\min,b} := \min_{2\le q\le M} a_{q,b}, \qquad a_{\max,b} := \max_{2\le q\le M} a_{q,b}.$$

In particular,

$$a_{\min,b} \ge \frac{2N_G + 2\gamma_b P\sigma^2}{D_M} > 0.$$

Write $A_\mu := \max_{0\le d<M} \|\boldsymbol{\mu}_{T-d}\|_2$ and $g := \|\boldsymbol{G}\|_2$. The gradient with respect to the averaged mean satisfies

$$\|\nabla_{\boldsymbol{\mu}} \mathcal{E}_b\|_2 = \|\boldsymbol{G}(\boldsymbol{\mu} - \boldsymbol{\mu}_T) + 2\gamma_b \boldsymbol{\mu}\|_2 \le g D_\mu + 2\gamma_b A_\mu =: G_{\mathcal{E},b}.$$

On each fixed-support segment, differentiation with respect to y gives

$$\mathcal{E}_b'(y) = \big[\boldsymbol{G}(\boldsymbol{\mu}(y) - \boldsymbol{\mu}_T) + 2\gamma_b \boldsymbol{\mu}(y)\big]^\top \boldsymbol{\mu}'(y) + (N_G + \gamma_b P\sigma^2) R'(y).$$

At a transition point y_j, $\boldsymbol{\mu}(y)$ is continuous and Lemma A.10 gives $R'(y_j-) = R'(y_j+)$. Thus the R' term has no jump, and

$$\begin{aligned} |\Delta \mathcal{E}_b'(y_j)| &= \Big| \big[\boldsymbol{G}(\boldsymbol{\mu}(y_j) - \boldsymbol{\mu}_T) + 2\gamma_b \boldsymbol{\mu}(y_j)\big]^\top \Delta \boldsymbol{\mu}'(y_j) \Big| \\ &\le \frac{G_{\mathcal{E},b}\kappa}{2}. \end{aligned} \tag{68}$$

Thus the adjacent-increment bound controls the effect of changing the active set on the slope of the expected reconstruction error.

We now locate where the expected reconstruction error along the frontier changes from tolerance to degradation. Define its average change from the starting kernel by

$$F_b(y) := \frac{\mathcal{E}_b(\boldsymbol{w}(y)) - \mathcal{E}_b(\boldsymbol{w}(0))}{y} \quad (y > 0), \qquad F_b(0) := \mathcal{E}_b'(0+).$$

Thus $F_b(y) < 0$ marks a kernel with lower error than the starting kernel, while $F_b(y) > 0$ marks degradation. To determine whether this sign can change more than once, we bound the slope of F_b across both support segments and their boundaries.

For a point t away from the support boundaries, its slope equals the initial right slope plus the curvature accumulated within the support segments and the slope jumps at all preceding boundaries:

$$\mathcal{E}_b'(t) = \mathcal{E}_b'(0+) + \int_0^t \mathcal{E}_b''(s)\,\mathrm{d}s + \sum_{y_j<t} \Delta \mathcal{E}_b'(y_j).$$

Integrating this slope over $t \in [0, y]$ and exchanging the order of integration gives

$$\begin{aligned} &\mathcal{E}_b(\boldsymbol{w}(y)) - \mathcal{E}_b(\boldsymbol{w}(0)) \\ &= y\mathcal{E}_b'(0+) + \int_0^y \int_0^t \mathcal{E}_b''(s)\,\mathrm{d}s\,\mathrm{d}t + \int_0^y \sum_{y_j<t} \Delta \mathcal{E}_b'(y_j)\,\mathrm{d}t \\ &= y\mathcal{E}_b'(0+) + \int_0^y (y-t)\mathcal{E}_b''(t)\,\mathrm{d}t + \sum_{y_j<y} (y-y_j)\Delta \mathcal{E}_b'(y_j). \end{aligned}$$

Each curvature value at t contributes over the remaining interval of length $y - t$, while a slope jump at y_j contributes over the remaining interval of length $y - y_j$.

For $y > 0$ away from a support boundary, the quotient rule gives

$$y^2 F_b'(y) = y\mathcal{E}_b'(y) - \big[\mathcal{E}_b(\boldsymbol{w}(y)) - \mathcal{E}_b(\boldsymbol{w}(0))\big].$$

Substituting the two preceding identities cancels the initial-slope terms. The remaining terms simplify as

$$\begin{aligned} y^2 F_b'(y) &= y\int_0^y \mathcal{E}_b''(t)\,\mathrm{d}t - \int_0^y (y-t)\mathcal{E}_b''(t)\,\mathrm{d}t \\ &\quad + y\sum_{y_j<y} \Delta\mathcal{E}_b'(y_j) - \sum_{y_j<y}(y-y_j)\Delta\mathcal{E}_b'(y_j) \\ &= \int_0^y t\,\mathcal{E}_b''(t)\,\mathrm{d}t + \sum_{y_j<y} y_j \Delta\mathcal{E}_b'(y_j). \end{aligned}$$

The segment curvatures and Eqs. (67) and (68) bound its two terms by

$$\frac{a_{\min,b}y^2}{2} \le \int_0^y t\,\mathcal{E}_b''(t)\,\mathrm{d}t \le \frac{a_{\max,b}y^2}{2}, \qquad \left|\sum_{y_j<y} y_j\Delta\mathcal{E}_b'(y_j)\right| \le \frac{H_M G_{\mathcal{E},b}\kappa y^2}{2}.$$

Hence, with

$$c_b := a_{\min,b} - H_M G_{\mathcal{E},b}\kappa, \qquad C_b := a_{\max,b} + H_M G_{\mathcal{E},b}\kappa,$$

we obtain

$$\frac{c_b}{2} \le F_b'(y) \le \frac{C_b}{2}. \tag{69}$$

The bounds extend to one-sided derivatives at support boundaries. It remains to establish $c_b > 0$ and show that the error eventually exceeds its starting value. Eq. (66) gives $D_\mu \le V\ell$; the triangle inequality then yields

$$A_\mu = \max_{0\le d<M} \|\boldsymbol{\mu}_{T-d}\|_2 \le \|\boldsymbol{\mu}_T\|_2 + D_\mu \le A_* := \|\boldsymbol{\mu}_T\|_2 + V\ell.$$

Substituting these bounds gives

$$\begin{aligned} c_b &\ge \frac{2N_G + 2\gamma_b P\sigma^2}{D_M} - H_M(gD_\mu + 2\gamma_b A_\mu)\kappa \\ &\ge \left[\frac{2N_G}{D_M} - \frac{H_M g V A}{(M-1)^2}\ell^3\right] + 2\gamma_b\left[\frac{P\sigma^2}{D_M} - \frac{H_M A_* A}{(M-1)^2}\ell^2\right]. \end{aligned} \tag{70}$$

The possible loss from boundary jumps decreases with the window length, whereas both positive curvature terms remain bounded below.

The same trajectory bounds also determine the comparison between the two endpoints. The corresponding means and noise factors are

$$(\boldsymbol{\mu}(0), R(0)) = (\bar{\boldsymbol{\mu}}, 1/M), \qquad (\boldsymbol{\mu}(m_0), R(m_0)) = (\boldsymbol{\mu}_T, 1).$$

Here $y = 0$ gives LAWA, while $y = m_0$ selects the most recent saved weight. Write $c_M := 1 - 1/M$. The mean-norm difference satisfies

$$\begin{aligned} \big|\|\boldsymbol{\mu}_T\|_2^2 - \|\bar{\boldsymbol{\mu}}\|_2^2\big| &= \big|(\boldsymbol{\mu}_T + \bar{\boldsymbol{\mu}})^\top(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}})\big| \\ &\le 2A_*\|\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}}\|_2 \le A_* V\ell. \end{aligned}$$

Substituting these endpoint means and noise factors into $\mathcal{E}_b$ therefore yields

$$\begin{aligned} &\mathcal{E}_b(\boldsymbol{w}(m_0)) - \mathcal{E}_b(\boldsymbol{w}(0)) \\ &\quad = N_G c_M - \frac{1}{2}\|\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2 + \gamma_b\left(\|\boldsymbol{\mu}_T\|_2^2 - \|\bar{\boldsymbol{\mu}}\|_2^2 + P\sigma^2 c_M\right) \\ &\quad \ge \left[N_G c_M - \frac{gV^2\ell^2}{8}\right] + \gamma_b\left[P\sigma^2 c_M - A_* V\ell\right]. \end{aligned} \tag{71}$$

We retain a fixed number of consecutive saved weights as T increases. Their normalized time span therefore tends to zero.

Assumption A.7 (Fixed Retained Count under Time Refinement)**.** Consider a family of training trajectories indexed by $T \to \infty$ on the normalized time interval $[0, 1]$. The number $M \geq 2$ of consecutive retained weights is fixed across the family. The model dimension P, calibration matrix $\boldsymbol{G}$, and spatial-noise level $\sigma > 0$ are fixed across the family. The constants V, A in Assumption A.6 are uniform in T, and $\sup_T \|\boldsymbol{\mu}_T\|_2 < \infty$. Thus $\ell = (M-1)/T \to 0$.

Lemma A.11 (Finite-Window Criterion and Its Consequences)**.** *Under Assumptions A.3–A.6, let* $A_* := \|\boldsymbol{\mu}_T\|_2 + V\ell$ *and* $c_M := 1 - 1/M$*. If the normalized window length satisfies*

$$0 < \ell \leq \min\left\{ \left(\frac{N_G(M-1)^2}{H_M g V A D_M}\right)^{1/3}, \left(\frac{P\sigma^2(M-1)^2}{2H_M A_* A D_M}\right)^{1/2}, \frac{2\sqrt{N_G c_M}}{\sqrt{g}\,V}, \frac{P\sigma^2 c_M}{2A_* V} \right\},$$

where a term with zero denominator is $+\infty$*, then the expected reconstruction error obeys*

$$\begin{aligned} c_b &\geq \frac{N_G + \gamma_b P\sigma^2}{D_M} > 0, \\ \mathcal{E}_b(\boldsymbol{w}(m_0)) - \mathcal{E}_b(\boldsymbol{w}(0)) &\geq \frac{(N_G + \gamma_b P\sigma^2)c_M}{2} > 0. \end{aligned} \tag{72}$$

Under Assumption A.7, there is a T_0*, independent of* b*, such that the displayed short-window criterion and hence Eq. (72) hold for every* $T \geq T_0$*.*

Proof. The displayed condition is equivalent to the four inequalities

$$\frac{H_M g V A \ell^3}{(M-1)^2} \leq \frac{N_G}{D_M}, \qquad \frac{H_M A_* A \ell^2}{(M-1)^2} \leq \frac{P\sigma^2}{2D_M},$$
$$\frac{gV^2\ell^2}{8} \leq \frac{N_G c_M}{2}, \qquad A_* V \ell \leq \frac{P\sigma^2 c_M}{2}.$$

The first row inserted into Eq. (70) gives

$$c_b \geq \frac{N_G}{D_M} + \gamma_b \frac{P\sigma^2}{D_M} = \frac{N_G + \gamma_b P\sigma^2}{D_M} > 0.$$

The second row inserted into Eq. (71) gives

$$\mathcal{E}_b(\boldsymbol{w}(m_0)) - \mathcal{E}_b(\boldsymbol{w}(0)) \geq \frac{N_G c_M}{2} + \gamma_b \frac{P\sigma^2 c_M}{2} > 0.$$

These are the two bounds in Eq. (72).

To show that the four inequalities eventually hold, set $B_\mu := \sup_T \|\boldsymbol{\mu}_T\|_2 < \infty$. Since M is fixed and $\ell = (M-1)/T \to 0$, we have $A_* \leq B_\mu + V$ for all sufficiently large T. The left sides of the four inequalities therefore tend to zero, while their right sides are fixed and positive because $N_G = \sigma^2 \operatorname{Tr}(\boldsymbol{G})/2 > 0$, $P\sigma^2 > 0$, and $c_M > 0$. Hence all four hold for the same T_0. They contain no b, so this T_0 is independent of precision. □

Eq. (69) bounds the change in F_b along the frontier. Lemma A.11 makes its lower slope bound positive and places the final frontier point above the starting error. These facts locate the transition from tolerance to degradation. Normalize the age reduction as $x := y/m_0 = 1 - m(y)/m_0$, with $x = 0$ at the starting kernel and $x = 1$ at the most recent saved weight.

Theorem A.2 (Finite Noise Tolerance and Degradation)**.** *Under Assumptions A.3–A.7, for sufficiently large* T*, consider the expected reconstruction error* $\mathcal{E}_b = \mathcal{P}_0 + \gamma_b \mathcal{Q}$ *at a fixed precision* $b \geq b_{\min}$*. Write* $\mathcal{E}_b(x) := \mathcal{E}_b(\boldsymbol{w}(m_0 x))$ *and set* $k_b = m_0^2 c_b > 0$*,* $K_b = m_0^2 C_b$*. There is a unique threshold* $\tau_b \in [0, 1)$*. It is zero if* $\mathcal{E}_b'(0+) \geq 0$*; otherwise, it is the unique positive solution of* $\mathcal{E}_b(\tau_b) = \mathcal{E}_b(0)$*. It satisfies the explicit upper bound*

$$0 \leq \tau_b \leq \frac{D_M V}{m_0 T}\left(\frac{gV\ell}{N_G} + \frac{2A_*}{P\sigma^2}\right).$$

(i) Tolerance: *for* $0 < x \leq \tau_b$*,*

$$-\frac{K_b}{2}x(\tau_b - x) \leq \mathcal{E}_b(x) - \mathcal{E}_b(0) \leq -\frac{k_b}{2}x(\tau_b - x). \tag{73}$$

(ii) Degradation: *for* $\tau_b < x \le 1$,

$$\mathcal{E}_b(x) - \mathcal{E}_b(0) \ge \frac{k_b}{2} x(x - \tau_b) > 0. \tag{74}$$

Moreover, $\mathcal{E}_b$ *is strictly increasing beyond* τ_b.

Proof. Recall that $F_b(y)$ is the error change from $\boldsymbol{w}(0)$ divided by y, with $F_b(0)$ equal to the initial right slope in y. Since $y = m_0 x$, $\mathcal{E}_b'(0+) = m_0 F_b(0)$, so the two initial slopes have the same sign.

The frontier kernels make $\mathcal{E}_b(\boldsymbol{w}(y))$ continuous and piecewise quadratic, so F_b is continuous. Eq. (69) and Lemma A.11 give $F_b'(y) \ge c_b/2 > 0$ on each support segment, including the one-sided bounds at its boundaries. Thus F_b is strictly increasing. The endpoint bound in that lemma gives $F_b(m_0) > 0$.

If $F_b(0) < 0$, there is a unique zero $y_b \in (0, m_0)$; set $\tau_b = y_b/m_0$. If $F_b(0) \ge 0$, set $y_b = \tau_b = 0$. This establishes the threshold claimed in the theorem.

We next bound its location. At $y = 0$, $R'(0) = 0$ and $\boldsymbol{\mu}'(0+) = \boldsymbol{a}_M$, so

$$F_b(0) = \left[\boldsymbol{G}(\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_T) + 2\gamma_b \bar{\boldsymbol{\mu}}\right]^\top \boldsymbol{a}_M, \qquad |F_b(0)| \le (gD_\mu + 2\gamma_b A_\mu)\|\boldsymbol{a}_M\|_2.$$

The convex-combination representation of $\boldsymbol{a}_M$ gives $\|\boldsymbol{a}_M\|_2 \le V/T$; also $D_\mu \le V\ell$ and $A_\mu \le A_*$. For $y_b > 0$,

$$-F_b(0) = F_b(y_b) - F_b(0) = \int_0^{y_b} F_b'(t)\,\mathrm{d}t \ge \int_0^{y_b} \frac{c_b}{2}\,\mathrm{d}t = \frac{c_b y_b}{2}.$$

Using $c_b \ge (N_G + \gamma_b P\sigma^2)/D_M$ from Lemma A.11 then gives

$$\begin{aligned}
\tau_b &\le \frac{2|F_b(0)|}{m_0 c_b} \\
&\le \frac{2D_M\|\boldsymbol{a}_M\|_2}{m_0} \frac{gD_\mu + 2\gamma_b A_\mu}{N_G + \gamma_b P\sigma^2} \\
&\le \frac{2D_M V}{m_0 T}\left(\frac{gV\ell}{N_G} + \frac{2A_*}{P\sigma^2}\right).
\end{aligned}$$

Assumption A.7 fixes M and bounds A_*, so for sufficiently large T this preliminary bound is smaller than $(M+1)/[3(M-1)]$, the end of the first support segment in x. On that segment, $\mathcal{E}_b(\boldsymbol{w}(y)) - \mathcal{E}_b(\boldsymbol{w}(0)) = F_b(0)y + a_{M,b}y^2/2$, hence

$$0 = F_b(y_b) = F_b(0) + \frac{a_{M,b}}{2} y_b.$$

Using $a_{M,b} \ge 2(N_G + \gamma_b P\sigma^2)/D_M$ now gives the upper bound:

$$\begin{aligned}
\tau_b &= \frac{-2F_b(0)}{m_0 a_{M,b}} \\
&\le \frac{D_M\|\boldsymbol{a}_M\|_2}{m_0} \frac{gD_\mu + 2\gamma_b A_\mu}{N_G + \gamma_b P\sigma^2} \\
&\le \frac{D_M V}{m_0 T}\left(\frac{gV\ell}{N_G} + \frac{2A_*}{P\sigma^2}\right).
\end{aligned}$$

When $y_b = 0$, the upper bound is immediate.

For $y_b > 0$ and $0 \le y \le y_b$, Eq. (69) and $F_b(y_b) = 0$ give

$$-\frac{C_b}{2}(y_b - y) \le -\int_y^{y_b} F_b'(t)\,\mathrm{d}t = F_b(y) - F_b(y_b) = F_b(y) \le -\frac{c_b}{2}(y_b - y).$$

Multiplying by y and setting $y = m_0 x$, $y_b = m_0 \tau_b$ proves Eq. (73). If $y_b = 0$, that bound concerns only $x = 0$. For $y > y_b$,

$$F_b(y) = F_b(y_b) + \int_{y_b}^{y} F_b'(t)\,\mathrm{d}t \ge F_b(y_b) + \frac{c_b}{2}(y - y_b) \ge \frac{c_b}{2}(y - y_b);$$

the last inequality also holds when $y_b = 0$, since then $F_b(0) \geq 0$. Therefore

$$\mathcal{E}_b(\boldsymbol{w}(y)) - \mathcal{E}_b(\boldsymbol{w}(0)) = yF_b(y) \geq \frac{c_b}{2} y(y - y_b),$$

which proves Eq. (74).

Finally, for $y > y_b$ away from support boundaries,

$$\frac{\mathrm{d}}{\mathrm{d}y}\mathcal{E}_b(\boldsymbol{w}(y)) = F_b(y) + yF_b'(y) \geq \frac{c_b}{2}(2y - y_b) > 0.$$

The same inequality holds for both one-sided derivatives at a boundary. Multiplying by m_0 yields $\mathrm{d}\mathcal{E}_b(x)/\mathrm{d}x \geq k_b(2x - \tau_b)/2 > 0$. Continuity across the finitely many boundaries then gives strict increase throughout $(\tau_b, 1]$. □

The corresponding threshold in the original frontier parameter is $\lambda_{\mathrm{tol},b} = \lambda(m_0\tau_b)$, obtained from Eq. (64). The next subsection characterizes precision-dependent averaging preferences.

A.6 Precision-dependent averaging preference

The first main theorem compares the expected reconstruction error $\mathcal{E}_b$ with its value at LAWA for a fixed precision. We now determine how quantization precision shifts the preferred kernel and identify the precision at which its lag–noise preference changes.

Recall that $x = 1 - m/m_0$ increases from uniform averaging toward the most recent saved weight. Increasing x reduces mean weight age and retains more spatial noise. Write $\boldsymbol{\mu}(x) = \boldsymbol{\mu}_T^{\boldsymbol{w}(m_0 x)}$ and $R(x) = R(\boldsymbol{w}(m_0 x))$. For any frontier function in this subsection, an argument x means evaluation at $\boldsymbol{w}(m_0 x)$; in particular, $\mathcal{E}_b(x) = \mathcal{E}_b(\boldsymbol{w}(m_0 x))$. The exact identity in Lemma A.9 separates the unquantized error and the quantization contribution:

$$\mathcal{E}_b(x) = \mathcal{P}_0(x) + \gamma_b \mathcal{Q}(x),$$

where $\gamma_b := c_G/(2^b - 1)^2$. Its two precision-independent components are

$$\mathcal{P}_0(x) = \frac{1}{2}\|\boldsymbol{\mu}(x) - \boldsymbol{\mu}_T\|_{\boldsymbol{G}}^2 + N_G R(x), \qquad \mathcal{Q}(x) = \|\boldsymbol{\mu}(x)\|_2^2 + P\sigma^2 R(x).$$

Here $\mathcal{P}_0$ is the unquantized reconstruction error, and $\gamma_b \mathcal{Q}$ is the quantization contribution under the simplified GPTQ model. The coefficients N_G and $c_G > 0$ were defined in the preceding subsection. Precision enters $\mathcal{E}_b$ through γ_b, which increases as precision decreases. The effect of this increase depends on how $\mathcal{Q}$ varies with x.

We assume that, during a stable training phase, latent-mean increments are aligned with one another and with the latent mean at the start of the retained window.

Assumption A.8 (Pairwise Alignment of Latent-Mean Increments)**.** The latent increments in the retained window satisfy

$$\boldsymbol{v}_{T-i+1}^\top \boldsymbol{v}_{T-j+1} \geq 0 \qquad (1 \leq i, j < M).$$

Assumption A.9 (Alignment with the Oldest Retained Latent Mean)**.** The latent increments also satisfy

$$\boldsymbol{\mu}_{T-M+1}^\top \boldsymbol{v}_{T-j+1} \geq 0 \qquad (1 \leq j < M).$$

Assumption A.8 says that the progress increments do not oppose one another; Assumption A.9 says that they do not oppose the oldest retained latent mean. Their magnitudes and directions may differ. The next lemma uses both sets of inner products to show that favoring more recent saved weights cannot decrease the latent-mean norm.

Lemma A.12 (Latent-Mean Norm along LNWA)**.** *Under Assumptions A.8 and A.9, every open* LNWA *support segment satisfies*

$$\boldsymbol{\mu}(x)^\top \boldsymbol{\mu}'(x) \geq 0.$$

Proof. Express each retained latent mean using the oldest mean and the progress made since that position:

$$\boldsymbol{\mu}_{T-d} = \boldsymbol{\mu}_{T-M+1} + \sum_{j=d+1}^{M-1} \boldsymbol{v}_{T-j+1}.$$

Averaging these expressions, with $s_j(\boldsymbol{w}) = \sum_{d=j}^{M-1} w_d$ as in Appendix A.1, gives

$$\boldsymbol{\mu}_T^{\boldsymbol{w}} = \boldsymbol{\mu}_{T-M+1} + \sum_{j=1}^{M-1} \big[1 - s_j(\boldsymbol{w})\big] \boldsymbol{v}_{T-j+1}. \tag{75}$$

The coefficient $1 - s_j(\boldsymbol{w})$ is the averaging mass assigned to saved weights that include the jth increment. These coefficients are nonnegative, so Assumptions A.8 and A.9 give

$$(\boldsymbol{\mu}_T^{\boldsymbol{w}})^\top \boldsymbol{v}_{T-j+1} = \boldsymbol{\mu}_{T-M+1}^\top \boldsymbol{v}_{T-j+1} + \sum_{i=1}^{M-1} \big[1 - s_i(\boldsymbol{w})\big] \boldsymbol{v}_{T-i+1}^\top \boldsymbol{v}_{T-j+1} \geq 0.$$

On a segment with an active prefix of size $q \geq 2$, differentiating the kernel in Eqs. (38)–(39) with respect to $x = 1 - m/m_0$ yields

$$\begin{aligned} \boldsymbol{\mu}'(x) &= \frac{m_0}{D_q} \sum_{d=0}^{q-1} (\bar{d}_q - d) \boldsymbol{\mu}_{T-d} \\ &= \frac{m_0}{2D_q} \sum_{j=1}^{q-1} j(q-j) \boldsymbol{v}_{T-j+1}. \end{aligned}$$

The second equality follows from Eq. (75) and $\sum_{d=0}^{j-1} (\bar{d}_q - d) = j(q-j)/2$. Each coefficient in this last sum is positive. Taking its inner product with $\boldsymbol{\mu}(x)$ and applying the preceding inequality proves $\boldsymbol{\mu}(x)^\top \boldsymbol{\mu}'(x) \geq 0$. □

Alignment describes the direction of recent progress. We further assume that the latent mean continues to change near the end of training.

Assumption A.10 (Nonvanishing Local Progress)**.** Let $h = 1/T$ be the spacing in Assumption A.6 and $v_* := \|\boldsymbol{v}_T\|_2$. For constant $\nu > 0$ independent of h, the newest increment satisfies, for all sufficiently small h,

$$v_* \geq \nu h.$$

Together with Assumptions A.6 and A.7, Assumption A.10 yields the quantitative local-progress bound in Lemma A.13.

Lemma A.13 (Local Progress Bound)**.** *Under Assumptions A.6, A.7, and A.10, for sufficiently large T, let $\rho_T := ((M-2)A/\nu)h$. Then*

$$\max_{1 \leq j < M} \|\boldsymbol{v}_{T-j+1} - \boldsymbol{v}_T\|_2 \leq \rho_T v_*, \qquad \rho_T \longrightarrow 0. \tag{76}$$

Proof. Eq. (66) gives $\kappa = Ah^2$. By the adjacent-increment bound and Assumption A.10,

$$\|\boldsymbol{v}_{T-j+1} - \boldsymbol{v}_T\|_2 \leq (j-1)\kappa \leq (M-2)Ah^2 \leq \frac{(M-2)A}{\nu} h\, v_* = \rho_T v_*.$$

Since M, A, and ν are fixed and $h \to 0$, we have $\rho_T \to 0$, proving Eq. (76). □

In the interior of the first support segment, all M averaging coefficients are positive. Its right endpoint is

$$x_{\mathrm{F}} := \frac{M+1}{3(M-1)}, \qquad Y := m_0 x_{\mathrm{F}} = \frac{M+1}{6}.$$

For $M = 2$, this segment extends to $x = 1$. Let $\boldsymbol{u} := m_0 \boldsymbol{a}_M$ denote the change in the averaged mean per unit increase of x on this segment. The frontier formulas give

$$\boldsymbol{\mu}(x) = \bar{\boldsymbol{\mu}} + x\boldsymbol{u}, \qquad R(x) = \frac{1}{M} + \frac{m_0^2 x^2}{D_M}, \qquad 0 \leq x \leq x_{\mathrm{F}}.$$

The first derivatives at $x = 0$ are

$$\mathcal{P}_0'(0^+) = \boldsymbol{u}^\top \boldsymbol{G}(\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_T), \qquad \mathcal{Q}'(0^+) = 2\bar{\boldsymbol{\mu}}^\top \boldsymbol{u}.$$

Here $R'(0^+) = 0$, so spatial noise has no first-order contribution. To determine the signs of these slopes, write the two vectors in the first expression as averages of the retained increments:

$$\frac{\boldsymbol{u}}{m_0} = \sum_{j=1}^{M-1} \frac{j(M-j)}{2D_M} \boldsymbol{v}_{T-j+1},$$
$$\frac{\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}}}{m_0} = \sum_{j=1}^{M-1} \frac{2(M-j)}{M(M-1)} \boldsymbol{v}_{T-j+1}.$$

Both sets of nonnegative coefficients sum to one. By Lemma A.13, each average is within $\rho_T v_*$ of $\boldsymbol{v}_T$ in Euclidean norm. Thus $\boldsymbol{G} \succ 0$ and $\rho_T \to 0$ give, for sufficiently large T,

$$-\frac{\mathcal{P}_0'(0^+)}{m_0^2} = \left(\frac{\boldsymbol{u}}{m_0}\right)^\top \boldsymbol{G} \left(\frac{\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}}}{m_0}\right)$$
$$\geq \left[\lambda_{\min}(\boldsymbol{G}) - \|\boldsymbol{G}\|_2(2\rho_T + \rho_T^2)\right] v_*^2 > 0.$$

For the other slope, Assumptions A.8 and A.9 give $\bar{\boldsymbol{\mu}}^\top \boldsymbol{v}_{T-j+1} \geq 0$ by Eq. (75). Since the first average above has nonnegative coefficients,

$$\mathcal{Q}'(0^+) = \frac{m_0}{D_M} \sum_{j=1}^{M-1} j(M-j) \bar{\boldsymbol{\mu}}^\top \boldsymbol{v}_{T-j+1} \geq 0.$$

Hence increasing x initially lowers the unquantized error and changes the quantization contribution at the nonnegative rate $\gamma_b \mathcal{Q}'(0^+)$. Their second derivatives are constant on this segment:

$$a := \mathcal{P}_0''(0^+) = \boldsymbol{u}^\top \boldsymbol{G} \boldsymbol{u} + \frac{2N_G m_0^2}{D_M} > 0, \qquad c := \mathcal{Q}''(0^+) = 2\|\boldsymbol{u}\|_2^2 + \frac{2P\sigma^2 m_0^2}{D_M} > 0.$$

Consequently, the following quadratic expansions are exact throughout the first support segment:

$$\begin{aligned} \mathcal{P}_0(x) - \mathcal{P}_0(0) &= \mathcal{P}_0'(0^+)x + \frac{\mathcal{P}_0''(0^+)}{2}x^2, \\ \mathcal{Q}(x) - \mathcal{Q}(0) &= \mathcal{Q}'(0^+)x + \frac{\mathcal{Q}''(0^+)}{2}x^2. \end{aligned} \tag{77}$$

Under the assumptions of Theorem A.2, the upper bound on τ_b lies below x_{F} for sufficiently large T. Hence Eq. (77) applies throughout the tolerance interval. We next use these quadratics to locate the minimizers of $\mathcal{P}_0$ and $\mathcal{E}_b$ and show that they are global.

Lemma A.14 (Optima on the First Support Segment)**.** *Under Assumptions A.3–A.10, for sufficiently large T, the unique minimizer of $\mathcal{P}_0$ on $[0,1]$ is $x_0 = \boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}})/a \in (0, x_{\mathrm{F}})$. For every evaluated precision $b \geq b_{\min}$, the unique minimizer of $\mathcal{E}_b$ on $[0,1]$ is*

$$x_b = \max\left\{0, \frac{\boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}}) - 2\gamma_b \bar{\boldsymbol{\mu}}^\top \boldsymbol{u}}{a + c\gamma_b}\right\} \in [0, x_0). \tag{78}$$

Proof. The preceding estimates give $\mathcal{P}_0'(0^+) < 0$ and $\mathcal{Q}'(0^+) \geq 0$. We next show that the unquantized error increases beyond the first segment. Set $\bar{v} := \max_{1 \leq j < M} \|\boldsymbol{v}_{T-j+1}\|_2$. Assumptions A.6 and A.7 give $\bar{v} \leq Vh$ and, for sufficiently large T,

$$\sqrt{g}\,(M-1)\bar{v} \leq \sqrt{g}\,V\ell \leq 2\sqrt{N_G/M}.$$

In the mean-age coordinate y, the increment bound and the convex coefficients in $\boldsymbol{a}_q$ imply

$$\|\boldsymbol{\mu}(y) - \boldsymbol{\mu}_T\|_2 \leq \bar{v}\, m(y) \leq m_0 \bar{v}, \qquad \|\boldsymbol{a}_q\|_2 \leq \bar{v}.$$

The derivative $R'(y)$ is continuous and increasing across the support segments. For $y \geq Y$, it is at least $R'(Y) = 2Y/D_M = 4/[M(M-1)]$. Consequently, on every such segment,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}y}\mathcal{P}_0(\boldsymbol{w}(y)) &= \boldsymbol{a}_q^\top \boldsymbol{G}(\boldsymbol{\mu}(y) - \boldsymbol{\mu}_T) + N_G R'(y) \\ &\geq -gm_0\bar{v}^2 + \frac{4N_G}{M(M-1)} \\ &\geq \frac{2N_G}{M(M-1)} > 0.\end{aligned}$$

The final inequality uses the displayed bound on $\bar{v}$ and $m_0 = (M-1)/2$. The same estimates apply to the one-sided derivatives at support boundaries. For $M = 2$, they apply to the left derivative at $Y = m_0$.

On the first segment, $\mathcal{P}_0'(0^+) < 0$ and $\mathcal{P}_0'' = a > 0$. Its derivative is positive at x_{F} by the preceding bound, so it vanishes once at $x_0 = \boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}})/a \in (0, x_{\mathrm{F}})$. Thereafter $\mathcal{P}_0$ increases on every segment. Continuity at their boundaries makes x_0 its unique global minimum.

For $\mathcal{E}_b$, the initial slope is $\boldsymbol{u}^\top \boldsymbol{G}(\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_T) + 2\gamma_b \bar{\boldsymbol{\mu}}^\top \boldsymbol{u}$. If it is nonnegative, Theorem A.2 gives $\tau_b = 0$ and strict increase for $x > 0$, so $x_b = 0$. If it is negative, Lemma A.12 gives $\mathcal{Q}'(x) = 2\boldsymbol{\mu}(x)^\top \boldsymbol{\mu}'(x) + P\sigma^2 R'(x) > 0$ for $x > 0$ within each segment. In particular, $\mathcal{E}_b'(x_0) = \gamma_b \mathcal{Q}'(x_0) > 0$. Solving the affine derivative of Eq. (77) therefore gives the unique zero in $(0, x_0)$ stated in Eq. (78). Beyond the first segment, both $\mathcal{P}_0'$ and $\mathcal{Q}'$ are positive. Hence this zero is the unique global minimizer of $\mathcal{E}_b$. □

Lemma A.14 gives the global minimizer at each evaluated precision. In Eq. (78), precision enters only through γ_b: lower precision increases this coefficient and gives more weight to the quantization cost, while higher precision reduces it. This changes the optimal trade-off between denoising and reducing temporal lag. The following theorem identifies the precision at which that preference changes and proves how the optimum responds.

Theorem A.3 (Precision-Dependent Averaging Preference)**.** *Under Assumptions A.3–A.10, for sufficiently large T, define*

$$b_{\mathrm{crit}} := \log_2\left(1 + \sqrt{\frac{2c_G \bar{\boldsymbol{\mu}}^\top \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}})}}\right). \tag{79}$$

For every evaluated precision $b \geq b_{\min}$,

$$\begin{aligned} b \leq b_{\mathrm{crit}} \quad &\Longrightarrow \quad x_b = 0, \quad \tau_b = 0 \qquad \textbf{\textit{(less spatial noise)}}, \\ b > b_{\mathrm{crit}} \quad &\Longrightarrow \quad x_b > 0, \quad \tau_b > 0 \qquad \textbf{\textit{(less temporal lag)}}. \end{aligned}$$

The optimal position x_b is nondecreasing in b and strictly increasing in the regime $b > b_{\mathrm{crit}}$. If $b_{\mathrm{crit}} < b_{\min}$, all evaluated precisions lie in the second regime.

Proof. By Lemma A.14, x_b is determined by the sign of $\mathcal{E}_b'(0^+)$. Since

$$\mathcal{E}_b'(0^+) = \boldsymbol{u}^\top \boldsymbol{G}(\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_T) + \frac{2c_G \bar{\boldsymbol{\mu}}^\top \boldsymbol{u}}{(2^b - 1)^2},$$

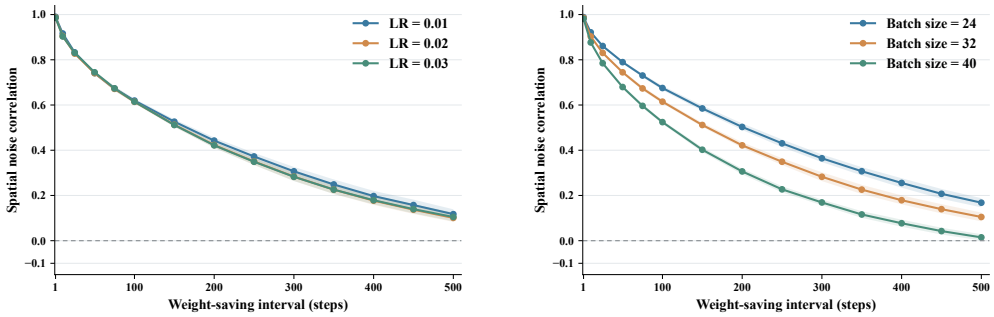
this slope is nonnegative exactly when $b \leq b_{\mathrm{crit}}$. The equivalence also holds when $\bar{\boldsymbol{\mu}}^\top \boldsymbol{u} = 0$, in which case $b_{\mathrm{crit}} = 0$ and every evaluated precision has a negative initial slope. Theorem A.2 then gives $\tau_b = 0$ in the nonnegative-slope regime and $\tau_b > 0$ in the negative-slope regime.

Where the maximum in Eq. (78) is positive, differentiation with respect to γ_b gives

$$\frac{\mathrm{d}}{\mathrm{d}\gamma_b}\left(\frac{\boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}}) - 2\gamma_b \bar{\boldsymbol{\mu}}^\top \boldsymbol{u}}{a + c\gamma_b}\right) = -\frac{2a\bar{\boldsymbol{\mu}}^\top \boldsymbol{u} + c\boldsymbol{u}^\top \boldsymbol{G}(\boldsymbol{\mu}_T - \bar{\boldsymbol{\mu}})}{(a + c\gamma_b)^2} < 0.$$

As γ_b strictly decreases with b, x_b is nondecreasing overall and strictly increasing above the threshold. □

Together, these results characterize the optimal kernel's shift in lag–noise preference: lower precision favors denoising, whereas higher precision favors reducing temporal lag. Eq. (79) gives the explicit precision threshold for this transition.



(a) Learning rate variation (b) Batch size variation

Figure 6: Spatial-noise correlation as the interval between saved weights increases. Left: learning-rate variation at batch size 32. Right: batch-size variation at learning rate 0.03.

B Key Assumptions and Limitation

B.1 Empirical Assessment of Key Assumptions

Our assumptions concern the latent mean, which is defined as a conditional expectation over training randomness. The law of large numbers gives, at each fixed training step t,

$$\widehat{\boldsymbol{\mu}}_t^{(n)} := \frac{1}{n}\sum_{r=1}^{n} \boldsymbol{\theta}_t^{(r)} \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[\boldsymbol{\theta}_t \mid \mathcal{F}_0] = \boldsymbol{\mu}_t.$$

Repeated runs let us estimate the latent mean and spatial noise. Since rerunning LLM pretraining many times is costly, we use a small neural network to compare trajectories across runs and empirically assess our key assumptions.

Experimental setup. We train a two-layer ReLU multilayer perceptron (MLP) on MNIST dataset. Images are average-pooled to 14×14 and standardized, and the network has dimensions $196 \to 64 \to 10$. We train the full network with SGD and analyze its 650 output-layer weights and biases, denoted by $\boldsymbol{\theta}_t$. To reflect the single-epoch data consumption of LLM pretraining, we sample without replacement throughout training. We split MNIST into 58,000 training and 2,000 validation images, train a shared prefix on 16,000 images, and then branch into 16 runs that traverse the remaining data in different orders. Across the 16 branches, we estimate the latent mean and spatial noise by

$$\widehat{\boldsymbol{\mu}}_t^{(16)} = \frac{1}{16}\sum_{r=1}^{16} \boldsymbol{\theta}_t^{(r)}, \qquad \widehat{\boldsymbol{\epsilon}}_t^{(r)} = \boldsymbol{\theta}_t^{(r)} - \widehat{\boldsymbol{\mu}}_t^{(16)}.$$

The shared prefix uses 500 updates at batch size 32 and learning rate 0.03. Each branch then runs for 1,000 updates. We vary the continuation learning rate over $\{0.01, 0.02, 0.03\}$ at batch size 32 and the batch size over $\{24, 32, 40\}$ at learning rate 0.03.

Spatial noise decorrelation. Figure 6 measures the correlation between spatial noise at different training positions. Let $\mathcal{T} := \{500, \ldots, 1000\}$ denote the updates in the analysis window. For a saving interval h, let $\mathcal{T}_h := \{t \in \mathcal{T} : t - h \geq 0\}$ denote the endpoints for which both $\widehat{\boldsymbol{\epsilon}}_{t-h}$ and $\widehat{\boldsymbol{\epsilon}}_t$ are available. We evaluate the correlation as

$$\widehat{\rho}_\epsilon(h) := \frac{\displaystyle\sum_{t\in\mathcal{T}_h}\sum_{r=1}^{16}\langle \widehat{\boldsymbol{\epsilon}}_{t-h}^{(r)}, \widehat{\boldsymbol{\epsilon}}_t^{(r)}\rangle}{\sqrt{\displaystyle\sum_{t\in\mathcal{T}_h}\sum_{r=1}^{16}\|\widehat{\boldsymbol{\epsilon}}_{t-h}^{(r)}\|_2^2}\sqrt{\displaystyle\sum_{t\in\mathcal{T}_h}\sum_{r=1}^{16}\|\widehat{\boldsymbol{\epsilon}}_t^{(r)}\|_2^2}},$$

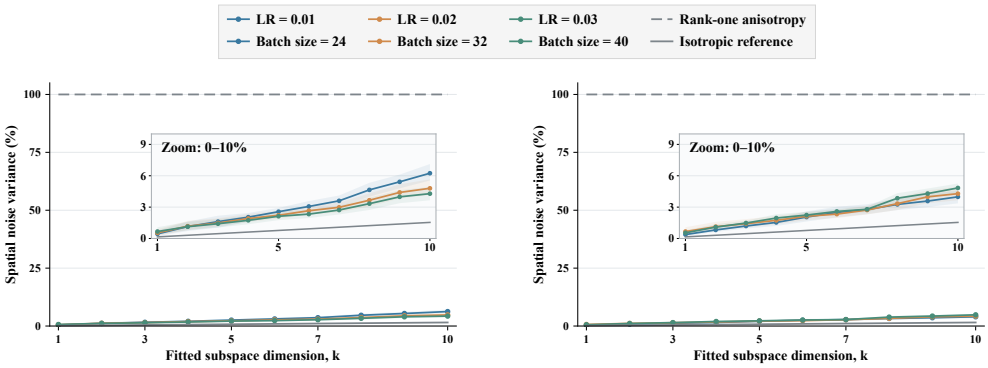


Figure 7: Spatial-noise variance captured by the first k fitted directions. Left: learning-rate variation at batch size 32. Right: batch-size variation at learning rate 0.03.

The correlation decreases steadily with h. At $h = 500$, it is 0.101–0.117 across learning rates and decreases from 0.168 to 0.015 as the batch size increases from 24 to 40. This supports the weak temporal dependence used in Assumption A.1 for sparsely saved weights.

Spatial noise dispersion. Figure 7 uses eight branches to identify the principal directions of spatial noise and the remaining eight to evaluate them. Let $\mathcal{R}_{\text{fit}}$ and $\mathcal{R}_{\text{eval}}$ denote these two sets. We recompute the latent-mean estimate within each set and obtain the corresponding residuals $\widehat{\boldsymbol{\epsilon}}^{(r)}_{t,\text{fit}}$ and $\widehat{\boldsymbol{\epsilon}}^{(r)}_{t,\text{eval}}$. From the fitting set, we form the empirical noise covariance

$$\widehat{\boldsymbol{\Sigma}}_{\text{fit}} := \frac{1}{|\mathcal{T}|\,|\mathcal{R}_{\text{fit}}|} \sum_{t\in\mathcal{T}} \sum_{r\in\mathcal{R}_{\text{fit}}} \widehat{\boldsymbol{\epsilon}}^{(r)}_{t,\text{fit}} \big(\widehat{\boldsymbol{\epsilon}}^{(r)}_{t,\text{fit}}\big)^{\top}.$$

Let $\boldsymbol{U}_k \in \mathbb{R}^{650\times k}$ contain the orthonormal eigenvectors associated with its k largest eigenvalues. We then measure how much spatial-noise variance these directions capture on the evaluation branches:

$$\widehat{V}_{\epsilon}(k) := 100 \times \frac{\displaystyle\sum_{t\in\mathcal{T}} \sum_{r\in\mathcal{R}_{\text{eval}}} \|\boldsymbol{U}_k^{\top} \widehat{\boldsymbol{\epsilon}}^{(r)}_{t,\text{eval}}\|_2^2}{\displaystyle\sum_{t\in\mathcal{T}} \sum_{r\in\mathcal{R}_{\text{eval}}} \|\widehat{\boldsymbol{\epsilon}}^{(r)}_{t,\text{eval}}\|_2^2} \,\%.$$

The parameter space has 650 dimensions. If the noise were concentrated in only a few directions, the first ten directions would explain a large fraction of its variance; rank-one noise, for example, would be captured entirely by the first direction. Instead, the first ten directions explain only 3.98%–6.24% of the variance. Thus the spatial noise is not concentrated in a small number of directions but dispersed across many parameter directions. For analytical simplicity, Assumption A.3 models this dispersed spatial noise with an isotropic covariance.

Alignment between progress increments. Let $t_0 < \cdots < t_9$ be the ten selected positions and define the run-specific increments $\Delta\boldsymbol{\theta}^{(r)}_j := \boldsymbol{\theta}^{(r)}_{t_j} - \boldsymbol{\theta}^{(r)}_{t_{j-1}}$. Figure 8 estimates the inner product between latent-mean increments using distinct runs,

$$\widehat{A}_{ij} := \frac{1}{16\cdot 15} \sum_{r\neq s} \big\langle \Delta\boldsymbol{\theta}^{(r)}_i, \Delta\boldsymbol{\theta}^{(s)}_j \big\rangle, \qquad 1 \le i < j \le 9.$$

Each increment is first computed within one branch. In $\widehat{A}_{ij}$, the increment at position i from branch r is paired with the increment at position j from a different branch $s \neq r$, and the result is averaged over all $16 \cdot 15$ ordered branch pairs. Because the two branches have independent zero-mean

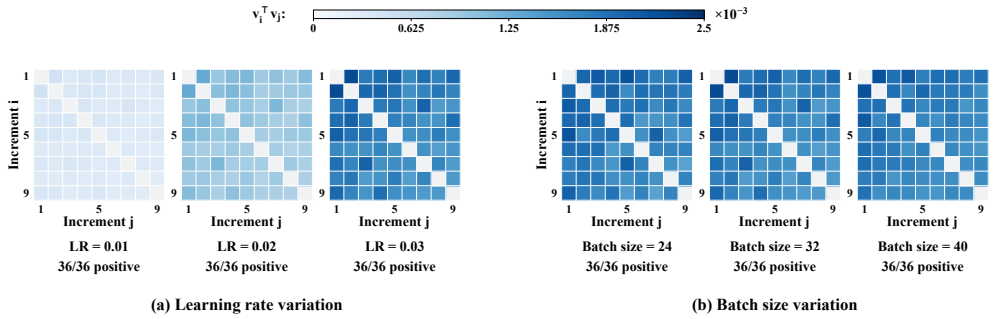


Figure 8: Cross-run estimates of the pairwise inner products between nine consecutive latent-mean progress increments. Left: learning-rate variation. Right: batch-size variation.

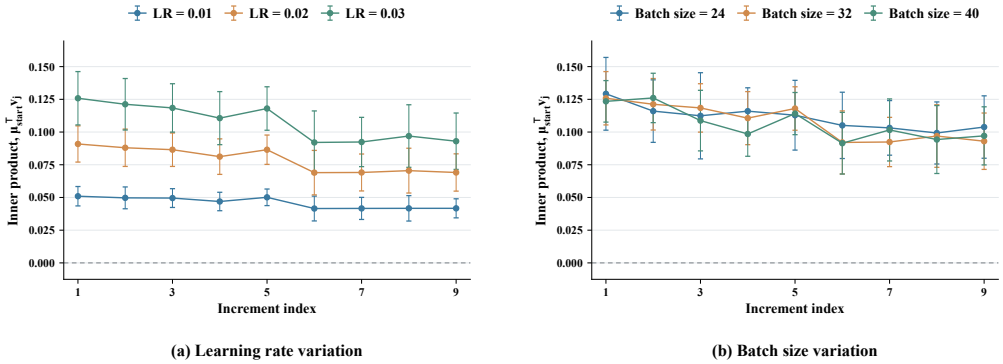


Figure 9: Cross-run estimates of the inner products between the window-start latent mean and nine subsequent progress increments. Left: learning-rate variation. Right: batch-size variation.

fluctuations, their noise terms vanish in expectation. Hence $\widehat{A}_{ij}$ estimates the inner product between the corresponding latent-mean increments. All 36 pairs are positive in every setting, showing that recent progress tends to remain aligned and supporting Assumption A.8.

Alignment with the window-start mean. Figure 9 measures the inner products between the latent mean at the start of the analysis window and the nine subsequent progress increments. It uses the similar cross-run estimate

$$\widehat{B}_j := \frac{1}{16 \cdot 15} \sum_{r \neq s} \langle \boldsymbol{\theta}_{t_0}^{(r)}, \Delta \boldsymbol{\theta}_j^{(s)} \rangle, \qquad 1 \leq j \leq 9.$$

All nine values are positive in every setting, supporting Assumption A.9.

B.2 Limitation

The main limitation of our work is that the theoretical guarantees are established under an idealized setting. However, bridging tractable theory and practical quantization pipelines is a broader challenge for the community rather than our uniqueness. In particular, our training analysis makes assumptions about the latent mean and spatial noise. We provide an empirical assessment of the key assumptions in Section B.1, which offers support for the theoretical setting. Our quantization analysis adopts a simplified GPTQ model: we use randomized dither rounding rather than deterministic nearest rounding, which has well-established statistical properties and has been used in quantization theory (Wannamaker et al., 2000; Chen et al., 2023; Dirksen et al., 2025); we also use a global scale rather than group-wise scales and consider only a two-stage quantization and compensation process. Such simplifications are common and widely accepted in the community (Banner et al., 2018; Nagel

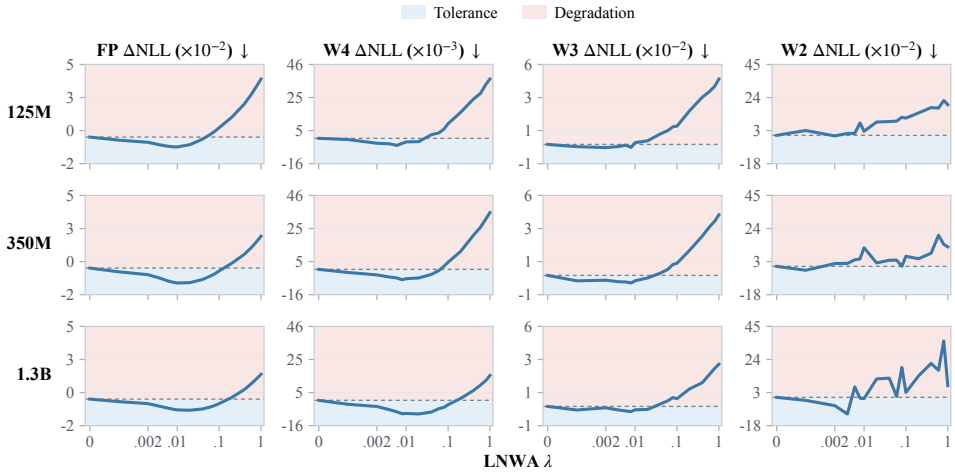


Figure 10: Comparison of ΔNLL across LNWA coefficients for OPT-125M, OPT-350M, and OPT-1.3B under GPTQ at 4-bit (W4), 3-bit (W3), 2-bit (W2), and full precision (FP).

et al., 2020; Chee et al., 2023). Future work will further develop the theory toward more practical quantization settings.

C Additional Experimental Results

C.1 Experimental Settings

Pretraining settings. We pretrain OPT-series models (Zhang et al., 2022) (125M, 350M, and 1.3B) from scratch on FineWeb-Edu (Penedo et al., 2024), with the model-specific configurations summarized in Table 2. Each training example is a sequence of length 512, and the global batch size denotes the number of such sequences processed per optimizer update. We use AdamW (Loshchilov & Hutter, 2017) with $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$ throughout training. The LR is linearly warmed up to its peak value during the first 1% of training and kept constant for the following 89%. Over the final 10%, we either continue training at the peak LR or transition to a cosine decay schedule, yielding the stable-LR and LR decay baselines, respectively. For weight averaging, we save the last $M = 10$ weights from each stable-LR trajectory at intervals of 500 optimizer steps (Catalan-Tatjer et al., 2026). All experiments are run on 4 NVIDIA RTX PRO 6000 GPUs. The code and configurations are available in the *Supplementary Material*.

Table 2: Pretraining settings used in the experiments. Global batch size is measured in sequences per optimizer update, and gradient clipping is by global norm.

Model size	125M	350M	1.3B
Training tokens	2B	4B	20B
Batch size	256	512	2560
Sequence length	512	512	512
Peak learning rate	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}
LR warmup ratio	1%	1%	1%
LR decay ratio	10%	10%	10%
Weight decay	0.1	0.1	0.1
Gradient clipping	0.5	0.5	0.5

Post-training quantization settings. For both GPTQ (Frantar et al., 2022) and AWQ (Lin et al., 2024), we use the same calibration set, consisting of 128 length-512 sequences deterministically

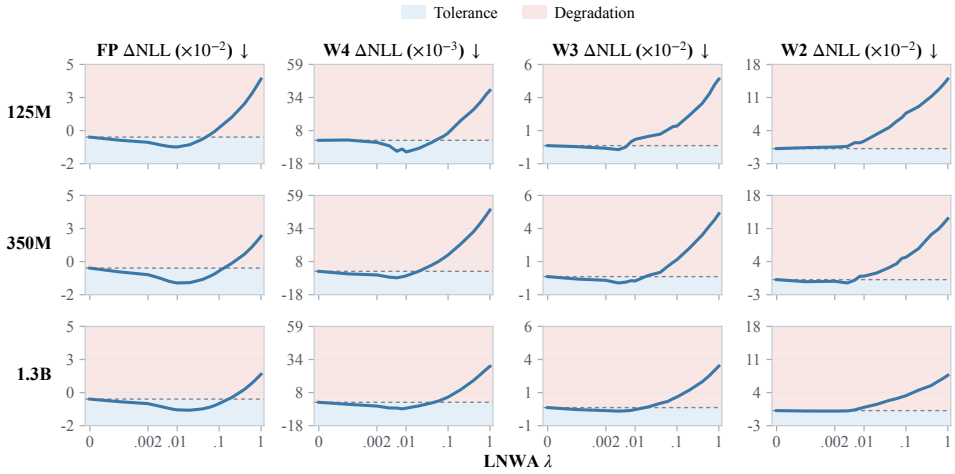


Figure 11: Comparison of ΔNLL across LNWA coefficients for OPT-125M, OPT-350M, and OPT-1.3B under AWQ at 4-bit (W4), 3-bit (W3), 2-bit (W2), and full precision (FP).

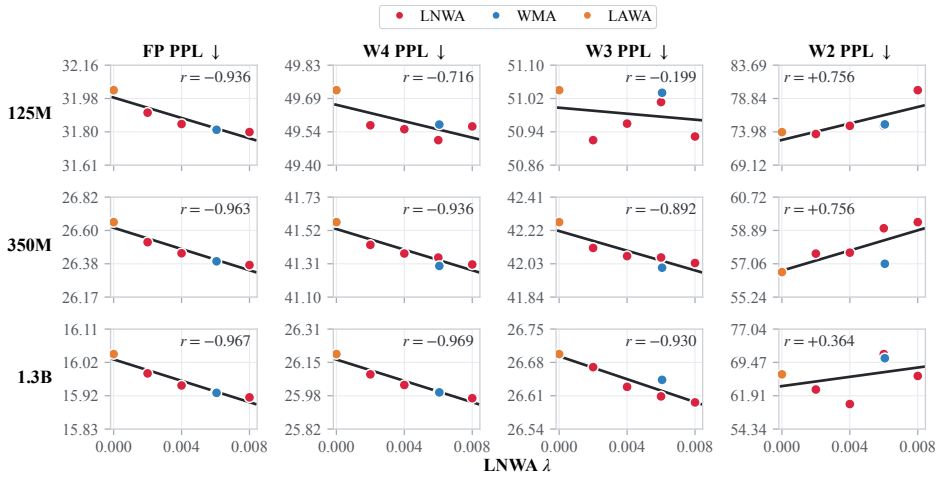


Figure 12: Comparison of PPL across LNWA-equivalent averaging coefficients for OPT-125M, OPT-350M, and OPT-1.3B under GPTQ at 4-bit (W4), 3-bit (W3), 2-bit (W2), and full precision (FP).

sampled from a fixed English C4 training split (Raffel et al., 2020). Both methods use a group size of 128 and are evaluated at 4-bit (W4), 3-bit (W3), and 2-bit (W2) precision. For GPTQ, we use 1% Hessian damping, disable activation ordering, and set the column block size to 128. For AWQ, we use asymmetric affine quantization with zero points, with scaling exponent $\alpha = 0.5$ and clipping ratio $c = 0.8$. We evaluate every quantized endpoint on 1,024 blocks from a fixed held-out RefinedWeb split. All remaining quantization settings are kept identical across model sizes.

Averaging kernels and coefficient grids. The tolerance-to-degradation experiment evaluates the same coefficients $\lambda \in [0, 1]$ at every model size and precision. The precision-preference comparison uses LAWA, WMA, and LNWA with $\lambda \in \{0.002, 0.004, 0.006, 0.008\}$. The LNWA-equivalent coefficients of the fixed baselines are 0 for LAWA and approximately 0.006061 for WMA. Appendix A.4 derives these relations.

For each model–precision panel, we form $n = 6$ pairs (x_i, y_i) from LAWA, WMA, and LNWA with $\lambda \in \{0.002, 0.004, 0.006, 0.008\}$, where x_i is the endpoint's LNWA-equivalent coefficient and

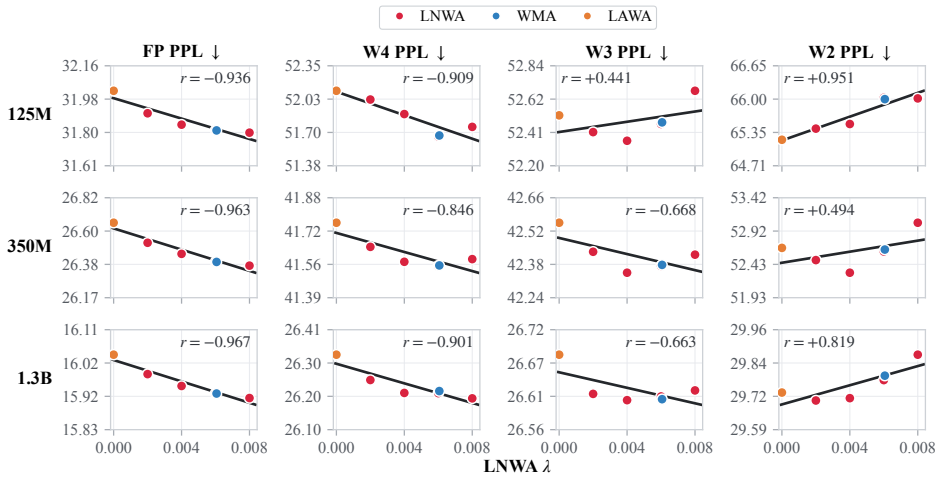


Figure 13: Comparison of PPL across LNWA-equivalent averaging coefficients for OPT-125M, OPT-350M, and OPT-1.3B under AWQ at 4-bit (W4), 3-bit (W3), 2-bit (W2), and full precision (FP).

y_i is its measured PPL. The black line is the ordinary least-squares fit

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_{i=1}^{n} \big[y_i - (a + bx_i)\big]^2,$$

and each panel reports the Pearson correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \in [-1, 1].$$

Here $\bar{x}$ and $\bar{y}$ are the sample means of the n coefficient and PPL values, respectively. Lower PPL indicates better performance. A negative correlation associates larger λ, and hence less lag, with lower PPL. A positive correlation associates smaller λ, and hence less noise, with lower PPL. The magnitude $|r|$ measures the strength of the linear association across the six endpoints.

C.2 Experimental Results

From noise tolerance to degradation. Figures 10 and 11 extend Figure 3 to OPT-125M, OPT-350M, and OPT-1.3B, including full precision (FP) and 4-bit (W4), 3-bit (W3), and 2-bit (W2) quantization under GPTQ and AWQ, respectively. FP is evaluated on the original FineWeb-Edu validation set, whereas W4/W3/W2 are evaluated on the fixed held-out RefinedWeb validation set.

At small λ, NLL remains close to or below the LAWA baseline; at larger λ, it rises above the baseline. Reducing temporal lag can initially offset or outweigh the cost of retaining additional spatial noise. Further movement along the frontier retains more noise and degrades performance. These results across model scales and quantization methods further support the tolerance-to-degradation transition predicted by Theorem 4.2.

Precision changes the preferred averaging kernel. Figures 12 and 13 extend Figure 4 to OPT-125M, OPT-350M, and OPT-1.3B, including FP, W4, W3, and W2, with GPTQ and AWQ, respectively. At every model scale, PPL is positively correlated with the averaging coefficient at W2 and negatively correlated at W4 under both methods. Since smaller coefficients retain less spatial noise and larger coefficients reduce temporal lag, the reversal indicates a preference for less noise at lower precision and less lag at higher precision. This further supports the precision-dependent preference predicted by Theorem 4.3.

Comparison with learning-rate decay. We compare continued stable-learning-rate training, cosine decay to zero, cosine decay to 1% of the maximum learning rate, and LNWA on OPT-125M.

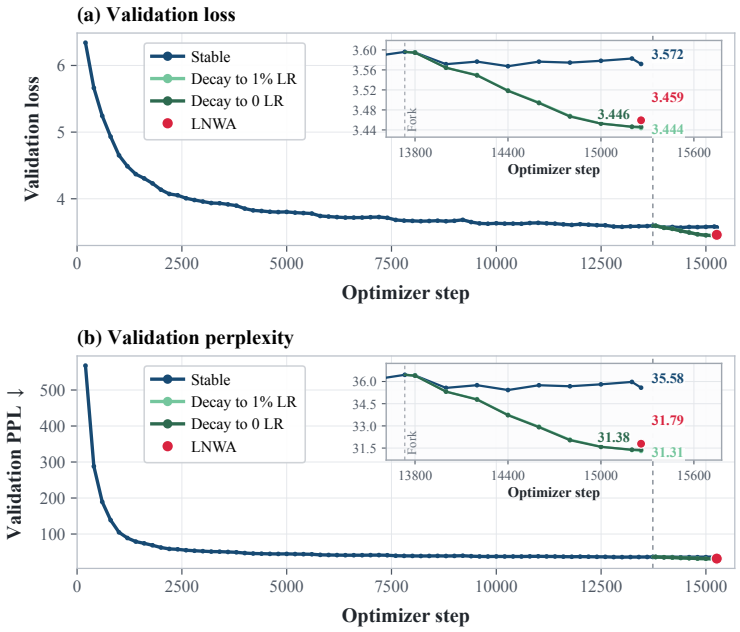


Figure 14: Comparison of validation loss and perplexity under Stable LR, LR decay, and LNWA for OPT-125M.

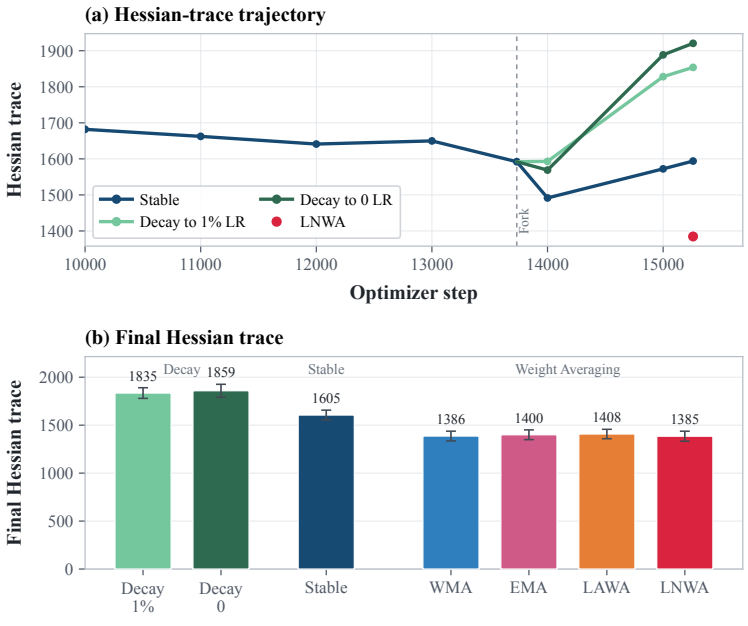


Figure 15: Comparison of Hessian-trace trajectories and final Hessian trace between LR decay, Stable LR, and weight averaging for OPT-125M.

Validation loss and perplexity are recorded every 200 optimizer steps, including the shared starting weight and the final weights. The LNWA endpoint uses $\lambda = 0.008$ and is shown at the final step.

Figure 14 shows that LNWA reaches a validation PPL of 31.79, compared with 35.58 for continued stable-LR training and 31.31 and 31.38 for decay to 1% LR and zero, respectively.

Figure 15 shows that the final Hessian trace is 1,385 for LNWA, compared with 1,605 for stable LR and 1,835 and 1,859 for the two decay schedules. Weight averaging therefore achieves validation performance close to LR decay while maintaining lower local curvature.