DOSA: Dynamic Online State Allocation for Adaptive Optimizers via Per-Tensor Sketched Smoothness Tests

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Which parameters of an adaptive optimizer should retain exact second-moment state, and which can be compressed? Memory-efficient methods (e.g. Adafactor, GaLore, 8-bit Adam) commit at design time. On transformer LM training, AdamW second moments concentrate on a small fraction of coordinates, with one dominant tensor typically carrying most of the adaptive mass. **DOSA** treats state allocation as an online statistical decision: promote the candidate update rule with the largest lower confidence bound (LCB) on a smoothness-model predicted-descent score. The main result is an *LCB-certificate separation theorem*: under heavy-tail concentration, any uniform-width sketch meeting the same identification certificate as per-tensor adaptive widths requires asymptotically more memory. We formulate two decision rules that are connected by a horizon-parameterized *surrogate*: the one-step endpoint recovers the predicted-descent score used inside LCB-greedy, while a closed-form anchored-residual blend represents the long-horizon quadratic-surrogate endpoint. A swept blend improves validation perplexity over a factored-Adafactor baseline across model scales and datasets, and the closed-form long-horizon surrogate rule recovers about half the swept gain without tuning on DistilGPT2/WikiText-2.

1. Introduction

Memory-efficient adaptive optimizers (Adafactor [20], GaLore [26], 8-bit Adam [8]) commit at design time to a fixed compression of the second-moment state. The question here is whether that choice can be made *during training*: a local smoothness model scores the descent of each candidate update rule, and the optimizer promotes the candidate with the largest *lower confidence bound* (LCB) on this predicted-descent score.

This paper makes three main contributions. **(1) The DOSA framework** (§2): a single LCB primitive on smoothness-model predicted descent handles any online optimizer-design choice whose candidates can be enumerated. **(2) An LCB-certificate separation theorem** (Theorem 2): under heavy-tail concentration $\rho \in (0, 1]$, any uniform-width allocation [21] meeting the same identification certificate as per-tensor adaptive widths requires $\Omega(\rho|\mathcal{T}|)$ times more memory at constant confidence. For the residual/tail proxy used in experiments, we measure $\rho_v \geq 0.83$ on transformer LM training (§4.1). **(3) Two decision rules**: the one-step predicted-descent score used by LCB-greedy (Algorithm 1) and a closed-form anchored-residual-blend coefficient (§3.3) arise as the $H{=}1$ and $H{\to}\infty$ endpoints of a horizon-parameterized surrogate (Proposition 5); the LCB penalty is the separate statistical guard used for finite-sample identification. Swept blends improve validation perplexity across two model scales and two datasets; the closed-form long-horizon surrogate rule is evaluated on DistilGPT2/WikiText-2 and recovers 53% of the swept gain without tuning.

Related work. Sketched optimizer state [21] uses uniform widths; we refine to per-tensor with tail-mass dependence. Sketched second-order information [9, 19] and hybrid eigensubspace optimizers [15] target the preconditioner spectrum; AdaLoRA [25] adaptively allocates LoRA rank across weight matrices via sensitivity-importance scoring—DOSA differs in domain (optimizer state, not PEFT) and mechanism (LCB on predicted descent). Our identification condition has the same gap-based structure as fixed-confidence best-arm identification [12, 17]; BAI has been used for cross-run hyperparameter tuning [13] but not per-step within-run. Adaptive regularization and stochastic trust-region [2, 3, 5] use a similar realized-vs-predicted ratio for a single scalar. Hand-designed adaptive optimizers (LARS [23], LAMB [24], Lion [7], Sophia [14]) fix the rule at design time.

2. The DOSA framework

2.1. State and decision

A DOSA-instrumented optimizer carries weights $W_t = \bigoplus_T W_t^{(T)}$, per-tensor sketched moments $S^{(m)}m_t, S^{(v)}v_t$ of CountSketch widths $W_m^{(T)}, W_v^{(T)}$, an online smoothness estimator $\widehat{\beta}_t$, and a finite candidate set $\mathcal{U} = \{u^{(0)}, \dots, u^{(K-1)}\}$ in which $u^{(0)}$ is a cheap default (factored Adafactor) and each $u^{(i)}$ is a candidate perturbation (typically materialized exact state on a specific tensor). Every K_{id} steps Algorithm 1 scores each candidate by an empirical smoothness-model predicted descent and promotes the rule with the largest LCB.

Algorithm 1 DOSA identification step with frozen candidate state (every K_{id} training steps)

Require: W_t, $u_{t-1}, \mathcal{U}$, frozen sketches $\{S_t^{(m)}m_t, S_t^{(v)}v_t\}$, widths $\{W_m^{(T)}, W_v^{(T)}\}$, scoring batches M, confidences $\delta_g, \delta_\beta, \delta_S$.

1: Freeze the current sketches and construct all candidate directions from this frozen state.
2: **for** $i \in [K]$ **do**
3: $\quad \widetilde{d}^{(i)} \leftarrow$ rule $u^{(i)}$ applied to the frozen sketched moments.
4: **end for**
5: Draw fresh scoring batches $\{g^{(m)}\}_{m=1}^{M}$ at W_t and form $\bar{g} = M^{-1}\sum_m g^{(m)}$; update the local score $\widehat{\beta}_t$ using its own confidence event.
6: **for** $i \in [K]$ **do**
7: $\quad \widehat{\Delta^{\mathrm{pred}}}_i \leftarrow -\left\langle \bar{g}, \widetilde{d}^{(i)} \right\rangle - \widehat{\beta}_t \left\langle u_{t-1}, \widetilde{d}^{(i)} \right\rangle - \frac{\widehat{\beta}_t}{2}\left\|\widetilde{d}^{(i)}\right\|^2$; $\widehat{L}_i \leftarrow \widehat{\Delta^{\mathrm{pred}}}_i - r^{(i)}$ via Eq. 4.
8: **end for**
9: Promote $i^\star = \arg\max_i \widehat{L}_i$ and allocate any newly materialized state.
10: After the decision, update the sketches with the training/scoring gradient estimate for the next identification round.

2.2. The smoothness-model predicted descent

The quantity scored at each identification step is the smoothness-model predicted descent

$$\Delta^{\mathrm{pred}}(W; u, d, \beta) = -\langle \nabla f(W), d\rangle - \beta\langle u, d\rangle - \tfrac{\beta}{2}\|d\|^2, \tag{1}$$

obtained by adding correction d on top of an anchor direction u in the quadratic model. The anchor u supplies only the cross-term $-\beta\langle u, d\rangle$; we instantiate it with u_{t-1} in Algorithm 1 and with the cheap default $u^{(0)}$ in Proposition 5. Eq. (1) is the increment in the descent-lemma model score used inside

adaptive-regularization and trust-region methods [2, 5]. In deployment we use the secant estimator as a local smoothness *score*; the theory below certifies identification relative to this model score under the stated confidence events. The empirical estimate $\widehat{\Delta^{\mathrm{pred}}}_i$ replaces ∇f by a fresh scoring-batch average $\bar{g}$, β by the online estimator $\widehat{\beta}$, and $d^{(i)}$ by the frozen-sketch reconstructed $\widetilde{d}^{(i)}$. Freezing candidates before drawing scoring batches makes the inner-product gradient event conditionally valid.

Specialization via $\mathcal{U}$. Changing the candidate set $\mathcal{U}$ yields state allocation, step-size adaptation, clipping threshold, or optimizer-rule selection. We instantiate it with discrete state allocation (§3) and the continuous-coefficient anchored residual blend (§3.3).

3. Sample and sketch complexity

The framework converts a heuristic optimizer-design choice into a statistical identification problem: how large must M, how accurate must $\widehat{\beta}$, and how wide must $W_m^{(T)}, W_v^{(T)}$ be for LCB-greedy to identify the per-step-best update with prescribed confidence? We work with a differentiable loss f admitting sub-Gaussian stochastic gradient access $g = \nabla f(W) + \xi$ with proxy σ^2 (Assumption 1, Appendix A), and a local secant smoothness score satisfying Assumption 2. The decision gap is $\Delta_t^\star := \Delta_{i^\circ}^{\mathrm{pred}} - \max_{j \neq i^\circ} \Delta_j^{\mathrm{pred}}$ at $i^\circ = \arg\max_i \Delta_i^{\mathrm{pred}}$. The tail-mass profile is $\tau_x(W) := \|x_{\mathrm{tail}}(W)\|^2 / \|x\|^2$, where $x_{\mathrm{tail}}(W)$ zeroes the top-W coordinates of x by magnitude.

3.1. The joint LCB radius

We use the following certificate abstraction. Conditional on the optimizer history and the frozen candidate directions, assume the sketching/reconstruction procedure provides radii q_i such that

$$\mathbb{P}\left[\left\|\widetilde{d}^{(i)} - d^{(i)}\right\| \leq q_i \text{ for all } i \in [K]\right] \geq 1 - \delta_S. \tag{2}$$

This is the only sketch property needed for the LCB proof. For a particular implementation, q_i can be obtained from any valid CountSketch [6, 18], heavy-hitter, low-rank, or empirical reconstruction certificate. Related lower bounds for streaming norm estimation and ℓ_p primitives are due to Jayram & Woodruff [10], Jowhari et al. [11]; the separation theorem below is stated only in terms of the certified tensor difficulty that produces these q_i's.
Let

$$\varepsilon_{g,\mathrm{ip}}(\delta_g) := \sqrt{\frac{2\sigma^2 \log(2K/\delta_g)}{M}} \tag{3}$$

be the frozen-direction sub-Gaussian inner-product deviation [22, Ch. 2][4, Thm. 2.5]. Define

$$r^{(i)} = \varepsilon_{g,\mathrm{ip}} \left\|d^{(i)}\right\| + \varepsilon_\beta(\delta_\beta) R_\beta^{(i)} + \left[\|\bar{g}\| + \widehat{\beta}\left(\|u\| + \left\|d^{(i)}\right\|\right)\right] q_i + \frac{\widehat{\beta}}{2} q_i^2, \tag{4}$$

where $R_\beta^{(i)} = |\langle u, d^{(i)} \rangle| + \left\|d^{(i)}\right\|^2 / 2$. The sketch term uses the observed $\bar{g}$ and $\widehat{\beta}$, so no separate vector-norm concentration event is needed. The radius is random but certified: on the joint confidence event, $|\widehat{\Delta}_i^{\mathrm{pred}} - \Delta_i^{\mathrm{pred}}| \leq r^{(i)}$ for all i.

Corollary 1 (LCB identification certificate) *Assume Eq. (2), the gradient event (3), and Assumption 2 hold with failure probabilities summing to at most δ, and assume the smoothness score used in Eq. (4) is nonnegative. Let $i^\circ = \arg\max_i \Delta_i^{\mathrm{pred}}$ and $\Delta^\star = \Delta_{i^\circ}^{\mathrm{pred}} - \max_{j \neq i^\circ} \Delta_j^{\mathrm{pred}}$. On the joint confidence event, if the realized certified radii satisfy*

$$2 \max_i r^{(i)} < \Delta^\star, \tag{5}$$

then LCB-greedy, $i^\star = \arg\max_i(\widehat{\Delta}_i^{\mathrm{pred}} - r^{(i)})$, returns $i^\star = i^\circ$. Equivalently,

$$\mathbb{P}\left[i^\star = i^\circ \text{ or } 2 \max_i r^{(i)} \geq \Delta^\star \right] \geq 1 - \delta.$$

If Eq. (5) is enforced by deterministic upper bounds on the radii, then $\mathbb{P}[i^\star = i^\circ] \geq 1 - \delta$.

Upper-bounding the sketch part with uniform proxies $R_{\mathrm{score}} \geq \max_i\{\|\bar{g}\| + \widehat{\beta}(\|u\| + \|d^{(i)}\|)\}$ and $\bar{\beta} \geq \widehat{\beta}$, it suffices to enforce

$$q_i \leq \min\left\{ \frac{\Delta^\star}{8K R_{\mathrm{score}}}, \sqrt{\frac{\Delta^\star}{4K\bar{\beta}}} \right\} \tag{6}$$

after budgeting the gradient and smoothness radii. Constants are conservative; the certificate is monotone in q_i.

3.2. Per-tensor vs uniform sketching

Theorem 2 (Per-tensor versus uniform widths) *Fix a per-step gap $\Delta^\star > 0$ and constant confidence. Let L_d be a Lipschitz constant of the sketched-update map $(\widehat{m}, \widehat{v}) \mapsto \widetilde{d}^{(i)}$ on the supported domain (finite under the deployed second-moment clamp). Suppose the sketch certificate has the tensorwise form*

$$q_i^2 \leq C_{\mathrm{sk}} L_d^2 \frac{A_{T_i}}{W_{T_i}}, \tag{7}$$

where T_i is the tensor touched by candidate i, $A_T \geq 0$ is the certified reconstruction difficulty of tensor T, and W_T is the sketch width allocated to T. Assume this certified difficulty concentrates on one tensor T^*:

$$A_{T^*} \geq \rho \sum_{T \in \mathcal{T}} A_T, \qquad \rho \in (0, 1].$$

Then there exists a per-tensor adaptive allocation satisfying the LCB certificate with total width

$$W_{\mathrm{tot}}^{\mathrm{adap}} = O\left(\sum_T A_T \right) \cdot \Gamma,$$

where Γ collects the common factors from Eq. (6). Any uniform-width allocation $W_T = W_{\mathrm{tot}}/|\mathcal{T}|$ satisfying the same certificate requires

$$W_{\mathrm{tot}}^{\mathrm{unif}} = \Omega\left(\rho|\mathcal{T}| \sum_T A_T \right) \cdot \Gamma.$$

Consequently,

$$\frac{W_{\mathrm{tot}}^{\mathrm{unif}}}{W_{\mathrm{tot}}^{\mathrm{adap}}} = \Omega(\rho|\mathcal{T}|).$$

The difficulty A_T may come from a raw CountSketch certificate, a heavy-hitter residual certificate, or any sharper reconstruction guarantee; the experiments use residual/tail statistics as the certified-difficulty proxy.

3.3. Continuous-coefficient state allocation: the anchored residual blend

DOSA also supports continuous candidate sets. We blend a cheap structured anchor $v_T^{(0)}$ (factored Adafactor, 8-bit, low-rank) with a higher-fidelity unbiased correction $\widetilde{v}_T$ (CountSketch in our experiments):

$$\hat{v}_T(\alpha_T) = v_T^{(0)} + \alpha_T(\widetilde{v}_T - v_T^{(0)}), \qquad \alpha_T \in [0,1], \tag{8}$$

plugged into the AdamW-style preconditioner $u_T = -\eta g_T/\sqrt{\hat{v}_T(\alpha_T)+\varepsilon}$ (RMS-clipped). For $\|v_T\| > 0$, with $r_T = \left\|v_T - v_T^{(0)}\right\| / \|v_T\|$ and $e_T(W_T) = (\mathbb{E}\,\|\widetilde{v}_T - v_T\|^2)^{1/2}/\|v_T\|$, unbiased reconstruction gives the tensorwise MSE-optimal $\alpha_T^\star = r_T^2/(r_T^2 + e_T(W_T)^2)$—the classical bias-variance shrinkage optimum for combining a biased low-variance and an unbiased high-variance estimator (proof and edge cases in Appendix B).

Importance weights and EMA correction. Linearizing $u_{T,j}(v)$ gives weights $w_{T,j} = \eta^2 g_{T,j}^2/(4(v_{T,j}+\varepsilon)^3)$; the tensor-uniform aggregate is $\bar{\kappa}_T n_T := (\|v_T\|^2/n_T)\sum_{j\in T} w_{T,j}$, and the AdamW EMA shrinks the per-step sketch noise by $e_T^{\text{eff}}(W_T) = \sqrt{(1-\beta_2)/(1+\beta_2)}\, e_T(W_T)$. The long-horizon quadratic surrogate is then minimized by

$$\alpha_{\text{global}}^{\star,\kappa,\text{EMA}} = \frac{\sum_T \bar{\kappa}_T n_T r_T^2}{\sum_T \bar{\kappa}_T n_T \left(r_T^2 + e_T^{\text{eff}}(W_T)^2\right)}. \tag{9}$$

Proposition 5 identifies $H = 1$ with the one-step score inside LCB-greedy and Eq. (9) with the long-horizon quadratic-and-tensor-uniform surrogate—not the exact long-run realized loss decrease.

4. Experiments

4.1. Heavy-tailedness of AdamW moments on transformer LM

On DistilGPT2 (82M) and GPT-2-medium (354M) trained for 2,000 AdamW steps on WikiText-2, v_t is more concentrated than m_t at both scales (Appendix D, Fig. 1). DistilGPT2's v_t has Hoyer 0.88 with 85% of mass on 0.1% of coordinates (m_t: 0.64, 51%), consistent with squaring the gradient tail in $v_t = \beta_2 v + (1-\beta_2)g^2$.

Empirical concentration. At $W = 1024$, the dominant tensor (token embedding) carries $\rho_v = 0.83$ on DistilGPT2 and 0.99 on GPT-2-medium of the total v-sketch-variance mass; with $|\mathcal{T}| = 76$ and 292 moment tensors in these measurements, $\rho_v|\mathcal{T}| \gg 1$ on both, well inside the residual-concentration regime, and remains so throughout training (Appendix D, Fig. 2). The factored Adafactor relative residual $r_T \in [0.07, 1.21]$ across DistilGPT2 tensors yields the per-tensor variation that motivates the continuous-coefficient anchored residual blend.

Anchored residual blend. At $B = 0.1n$ with 3 seeds (Appendix E), the swept-best blend coefficient $\alpha = 0.75$ improves validation perplexity over the factored-Adafactor anchor by $\mathbf{3.27 \pm 0.82}$ on DistilGPT2/WikiText-2 (4.0σ), $\mathbf{3.96 \pm 0.46}$ on GPT-2-small (8.6σ), and $\mathbf{3.48 \pm 0.51}$ on WikiText-103 (6.9σ). The EMA-corrected long-horizon surrogate endpoint $\alpha_{\text{global}}^{\star,\kappa,\text{EMA}} \approx 0.16$ (no sweep) gains 1.72 ppl on DistilGPT2/WikiText-2, recovering $\mathbf{53}\%$ of the tuned-α gain.

References

[1] Bach, F. and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *NeurIPS*, 2013.

[2] Bellavia, S., Gurioli, G., Morini, B., and Toint, Ph. L. Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM Journal on Optimization*, 29(4):2881–2915, 2019.

[3] Blanchet, J., Cartis, C., Menickelly, M., and Scheinberg, K. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS Journal on Optimization*, 1(2):92–119, 2019.

[4] Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[5] Cartis, C., Gould, N. I. M., and Toint, Ph. L. Adaptive cubic regularisation methods for unconstrained optimization. *Mathematical Programming*, 127(2):245–295, 2011.

[6] Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. In *ICALP*, pp. 693–703, 2002.

[7] Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., Lu, Y., and Le, Q. V. Symbolic discovery of optimization algorithms (Lion). In *NeurIPS*, 2023.

[8] Dettmers, T., Lewis, M., Shleifer, S., and Zettlemoyer, L. 8-bit optimizers via block-wise quantization. In *ICLR*, 2022.

[9] Feinberg, V., Chen, X., Sun, Y., Anil, R., and Hazan, E. Sketchy: Memory-efficient adaptive regularization with frequent directions. In *NeurIPS*, 2023.

[10] Jayram, T. S. and Woodruff, D. P. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms*, 9(3):26:1–26:17, 2013.

[11] Jowhari, H., Sağlam, M., and Tardos, G. Tight bounds for ℓ_p samplers, finding duplicates in streams, and related problems. In *PODS*, pp. 49–58, 2011.

[12] Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *JMLR*, 17(1):1–42, 2016.

[13] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *JMLR*, 18(185):1–52, 2018.

[14] Liu, H., Li, Z., Hall, D., Liang, P., and Ma, T. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *ICLR*, 2024.

[15] Liu, L., Xu, Z., Zhang, Z., Kang, H., Li, Z., Liang, C., Chen, W., and Zhao, T. COSMOS: A hybrid adaptive optimizer for memory-efficient training of LLMs. *arXiv preprint arXiv:2502.17410*, 2025.

[16] Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate Bayesian inference. *JMLR*, 18(134):1–35, 2017.

[17] Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *JMLR*, 5:623–648, 2004.

[18] Minton, G. T. and Price, E. Improved concentration bounds for count-sketch. In *Proc. 25th ACM–SIAM Symposium on Discrete Algorithms (SODA)*, 2014.

[19] Pilanci, M. and Wainwright, M. J. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

[20] Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, 2018.

[21] Spring, R., Kyrillidis, A., Mohan, V., and Shrivastava, A. Compressing gradient optimizers via count-sketches. In *ICML*, 2019.

[22] Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[23] You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks (LARS). *arXiv preprint arXiv:1708.03888*, 2017.

[24] You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. Large batch optimization for deep learning: Training BERT in 76 minutes (LAMB). In *ICLR*, 2020.

[25] Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., and Zhao, T. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *ICLR*, 2023.

[26] Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *ICML* (Oral), 2024.

Appendix A. Setup, assumptions, and definitions

Assumption 1 (Sub-Gaussian noise) *For every $\lambda \in \mathbb{R}$ and $v \in \mathbb{R}^n$,*

$$\mathbb{E}[\exp(\lambda \langle \xi, v \rangle)] \leq \exp(\lambda^2 \sigma^2 \|v\|^2 / 2)$$

for an absolute constant σ.

Assumption 2 (Online smoothness score) *Let β_t denote the local smoothness score used to evaluate candidate corrections at W_t. The nonnegative secant estimator*

$$\widehat{\beta}_t^{\mathrm{sec}} := \|\bar{g}_t - \bar{g}_{t-1}\| \,/\, \|W_t - W_{t-1}\|$$

measures the previous training segment rather than the new candidate direction. We therefore use it as a modeling local smoothness score for Δ^{pred}, under the regularity condition that nearby candidate steps have comparable directional smoothness. For confidence $\delta_\beta \in (0,1)$ assume $\mathbb{P}[|\widehat{\beta}_t^{\mathrm{sec}} - \beta_t| > \varepsilon_\beta(\delta_\beta)] \leq \delta_\beta$. A coordinate-union bound gives the conservative choice $\varepsilon_\beta(\delta_\beta) = 2\sigma\sqrt{n \log(4n/\delta_\beta)/M_\beta} / \|W_t - W_{t-1}\|$.

Remark 3 (Effective dimension cancellation) *The bound contains an explicit $\sqrt{n}$ factor, but this cancels in the regime $\|W_t - W_{t-1}\| = \Theta(\mathrm{lr}\sqrt{n})$, i.e. when each coordinate update has $\Theta(\mathrm{lr})$ magnitude. It degrades near stationary points, under highly sparse updates, or under global rescaling; in those regimes the LCB becomes conservative by enlarging ε_β.*

Appendix B. Proofs

B.1. LCB certificate proof

Proof [Proof of Corollary 1] On the joint event, telescope the error for each frozen candidate:

$$\widehat{\Delta}_i^{\mathrm{pred}} - \Delta_i^{\mathrm{pred}} = A_i + B_i + C_i,$$

where

$$\begin{aligned} A_i &= -\left\langle \bar{g} - \nabla f(W),\, d^{(i)} \right\rangle, \\ B_i &= (\beta - \widehat{\beta}) \left(\left\langle u,\, d^{(i)} \right\rangle + \tfrac{1}{2} \left\| d^{(i)} \right\|^2 \right), \\ C_i &= \Delta^{\mathrm{pred}}(\bar{g}, \widehat{\beta}, \widetilde{d}^{(i)}) - \Delta^{\mathrm{pred}}(\bar{g}, \widehat{\beta}, d^{(i)}). \end{aligned}$$

Because Algorithm 1 freezes candidate directions before drawing the scoring batches, the standard sub-Gaussian inner-product event gives $|A_i| \le \varepsilon_{g,\mathrm{ip}} \left\| d^{(i)} \right\|$ simultaneously over $i \in [K]$. Assumption 2 gives $|B_i| \le \varepsilon_\beta R_\beta^{(i)}$. For C_i, let $e_i = \widetilde{d}^{(i)} - d^{(i)}$. Direct expansion yields

$$\begin{aligned} |C_i| &= \left| -\langle \bar{g},\, e_i \rangle - \widehat{\beta} \langle u,\, e_i \rangle - \frac{\widehat{\beta}}{2} \left(\left\| d^{(i)} + e_i \right\|^2 - \left\| d^{(i)} \right\|^2 \right) \right| \\ &\le \left[\|\bar{g}\| + \widehat{\beta}(\|u\| + \left\| d^{(i)} \right\|) \right] \|e_i\| + \frac{\widehat{\beta}}{2} \|e_i\|^2 . \end{aligned}$$

The sketch certificate (2) gives $\|e_i\| \le q_i$, so Eq. (4) bounds $|\widehat{\Delta}_i^{\mathrm{pred}} - \Delta_i^{\mathrm{pred}}|$ for all candidates. Hence

$$\widehat{L}_i = \widehat{\Delta}_i^{\mathrm{pred}} - r^{(i)} \le \Delta_i^{\mathrm{pred}} \quad \text{and} \quad \widehat{L}_i \ge \Delta_i^{\mathrm{pred}} - 2r^{(i)}.$$

If $2 \max_i r^{(i)} < \Delta^\star$, then for every $j \ne i^\circ$,

$$\widehat{L}_{i^\circ} \ge \Delta_{i^\circ}^{\mathrm{pred}} - 2r^{(i^\circ)} > \Delta_j^{\mathrm{pred}} \ge \widehat{L}_j,$$

so LCB-greedy selects i°. A union bound over the gradient, smoothness, and sketch events gives the stated eventwise guarantee with probability at least $1 - \delta$.

Proof [Proof of Theorem 2] Let Γ denote the common width multiplier needed to make Eq. (7) imply Eq. (6). For example, up to constants,

$$\Gamma = L_d^2 \max\left\{ \frac{R_{\mathrm{score}}^2 K^2}{(\Delta^\star)^2}, \frac{\bar{\beta} K}{\Delta^\star} \right\}.$$

An adaptive allocation chooses $W_T \asymp A_T \Gamma$, so

$$W_{\mathrm{tot}}^{\mathrm{adap}} = \sum_T W_T \asymp \left(\sum_T A_T \right) \Gamma.$$

A uniform allocation has $W_T = W_{\mathrm{tot}}/|\mathcal{T}|$. To satisfy the same certificate on the dominant tensor T^*, it must obey

$$\frac{W_{\mathrm{tot}}}{|\mathcal{T}|} \gtrsim A_{T^*}\Gamma.$$

Using $A_{T^*} \geq \rho \sum_T A_T$ gives

$$W_{\mathrm{tot}}^{\mathrm{unif}} \gtrsim \rho|\mathcal{T}| \left(\sum_T A_T\right) \Gamma.$$

Dividing by the adaptive budget proves the $\Omega(\rho|\mathcal{T}|)$ separation. The proof compares allocations under the stated LCB certificate and does not use an information-theoretic lower bound.

Proposition 4 (Optimal mixing coefficient) *If $\|v_T\| > 0$, define*

$$r_T = \frac{\left\|v_T - v_T^{(0)}\right\|}{\|v_T\|}, \qquad e_T(W_T) = \frac{\left(\mathbb{E}\,\|\widetilde{v}_T - v_T\|^2\right)^{1/2}}{\|v_T\|}.$$

If $r_T^2 + e_T(W_T)^2 > 0$, the MSE-minimizer of $\mathbb{E}\,\|\hat{v}_T(\alpha_T) - v_T\|^2$ over $\alpha_T \in [0,1]$ is

$$\alpha_T^\star = \frac{r_T^2}{r_T^2 + e_T(W_T)^2}, \qquad \mathbb{E}[\mathrm{MSE}_T(\alpha_T^\star)] = \|v_T\|^2 \frac{r_T^2 e_T(W_T)^2}{r_T^2 + e_T(W_T)^2}.$$

If $r_T = e_T(W_T) = 0$, every $\alpha_T \in [0,1]$ is optimal and the minimum MSE is zero. If $\|v_T\| = 0$, the normalized quantities above are undefined and the unnormalized quadratic form in the proof should be used instead.

Proof [Proof of Proposition 4] Let $b_T := v_T^{(0)} - v_T$ and $z_T := \widetilde{v}_T - v_T$. The estimator error is

$$\hat{v}_T(\alpha) - v_T = (1-\alpha)b_T + \alpha z_T.$$

Because the sketch correction is unbiased, $\mathbb{E}z_T = 0$, so the cross term vanishes:

$$\mathbb{E}\,\|\hat{v}_T(\alpha) - v_T\|^2 = (1-\alpha)^2\,\|b_T\|^2 + \alpha^2 \mathbb{E}\,\|z_T\|^2\,.$$

If $\|v_T\| > 0$, writing $\|b_T\| = r_T\,\|v_T\|$ and $(\mathbb{E}\,\|z_T\|^2)^{1/2} = e_T(W_T)\,\|v_T\|$ gives a one-dimensional quadratic in α. When $r_T^2 + e_T(W_T)^2 > 0$, differentiating gives the unique minimizer in Eq. (**??**), which already lies in $[0,1]$; the displayed value follows by substitution. When both normalized error terms vanish, the quadratic is identically zero and every α is optimal. If $\|v_T\| = 0$, the same conclusions follow from the unnormalized quadratic except that the normalized display is not used.

Appendix C. Horizon-endpoint surrogate: full statement and proof

Proposition 5 (Surrogate endpoints) *Let $\widetilde{S}_H(\alpha)$ be the cumulative* model-predicted *score obtained by summing local quadratic predicted-descent scores under a frozen blend coefficient α on a window $[t_0, t_0 + H]$. Under a local isotropic quadratic model, sub-Gaussian stationary gradient noise, unbiased independent CountSketch reconstruction noise, and fixed α on the window:*

(a) the $H = 1$ surrogate differs from $\Delta^{\text{pred}}(W_{t_0}; u^{(0)}, d(\alpha), \beta)$ by an α-independent constant, so it recovers the unpenalized one-step score used inside LCB-greedy;

(b) after Taylor-expanding the AdamW-style preconditioner $u_j = -\eta g_j/\sqrt{v_j + \varepsilon}$, omitting the signed linear-bias term, and using a tensor-uniform importance-weight approximation, the resulting quadratic error surrogate is minimized by Eq. (9).

Proof For $H = 1$, the local quadratic model gives

$$\widetilde{S}_1(\alpha) = -\left\langle \nabla f,\, u^{(0)} + d(\alpha) \right\rangle - \frac{\beta}{2}\left\| u^{(0)} + d(\alpha) \right\|^2 .$$

Expanding the square separates an α-independent term and the bracket

$$-\left\langle \nabla f,\, d(\alpha) \right\rangle - \beta\left\langle u^{(0)},\, d(\alpha) \right\rangle - \frac{\beta}{2}\left\| d(\alpha) \right\|^2 = \Delta^{\text{pred}}(W_{t_0}; u^{(0)}, d(\alpha), \beta),$$

so maximizing the one-step endpoint is equivalent to maximizing the DOSA model score.

For the long-horizon surrogate we work to leading order in the step size, the regime of the diffusion stationary-distribution approximation [1, 16]. Use the same analyzed preconditioner convention as above, $P_j(\alpha) = 1/\sqrt{\hat{v}_j(\alpha) + \varepsilon}$, so $P_j(\alpha)^2 = 1/(\hat{v}_j(\alpha) + \varepsilon)$. Expanding around v_j gives

$$P_j(\alpha)^2 = \frac{1}{v_j + \varepsilon} - \frac{\delta v_j(\alpha)}{(v_j + \varepsilon)^2} + \frac{\delta v_j(\alpha)^2}{(v_j + \varepsilon)^3} + O(\delta v_j^3).$$

The linear term produces an aggregate αL depending on the signed coordinate-wise anchor bias. The closed-form rule omits this signed term and keeps the MSE-style second-order contribution, so this step is a surrogate approximation rather than an exact optimality statement for the signed-bias objective. The retained quadratic term is

$$Q(\alpha) = \sum_j w_j \left[(1-\alpha)^2 (v_j^{(0)} - v_j)^2 + \alpha^2 \frac{1-\beta_2}{1+\beta_2} \mathbb{E}\zeta_j^2 \right].$$

Approximating w_j by its tensor average $n_T^{-1} \sum_{j \in T} w_{T,j}$ and writing $e_T^{\text{eff}} = \sqrt{(1-\beta_2)/(1+\beta_2)}\, e_T$ gives

$$\begin{aligned} Q(\alpha) &\approx \sum_T \left(\frac{\|v_T\|^2}{n_T} \sum_{j \in T} w_{T,j} \right) \left[(1-\alpha)^2 r_T^2 + \alpha^2 (e_T^{\text{eff}})^2 \right] \\ &= \sum_T \bar{\kappa}_T n_T \left[(1-\alpha)^2 r_T^2 + \alpha^2 (e_T^{\text{eff}})^2 \right]. \end{aligned}$$

Differentiating this scalar quadratic and solving $Q'(\alpha) = 0$ yields Eq. (9).

Appendix D. Heavy-tail aggregates table

Appendix E. Anchored residual blend: full results table

Appendix F. Per-tensor vs uniform sketched optimizer state (extended)

The per-tensor variation documented in Section 4.1 predicts that uniform-width sketching wastes budget on small tensors and starves the dense ones. We test this directly by comparing allocation policies under a fixed total state budget.

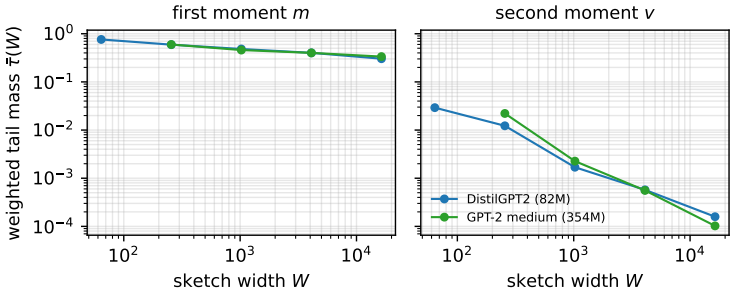

Figure 1: Weighted tail mass $\bar{\tau}(W)$ at step 2,000 on WikiText-2: v (right) has $\sim 4\times$ less tail mass than m (left) at every W, supporting the residual-concentration premise behind Theorem 2.

Table 1: Aggregate heavy-tail measurements of AdamW moments at step 2,000 on two GPT-2 scales. Bold marks the heavier-tailed second moment.

Model (n)	Mom.	Hoyer	Top-1%	Top-0.1%	$\tau(W{=}1024)$
DistilGPT2 (82M)	m_t	0.64	0.67	0.51	0.68
	v_t	**0.88**	**0.94**	**0.85**	**0.21**
GPT-2-med (354M)	m_t	0.42	0.40	0.24	0.86
	v_t	**0.72**	**0.84**	**0.77**	**0.35**

Setup. DistilGPT2 ($\sim$ 82M parameters) is trained on WikiText-2 for 5,000 steps under each policy at matched total budget B summed across m and v sketches. Uniform allocates $W_m^{(T)} = W_v^{(T)} = B/(2|\mathcal{T}|)$ to every tensor. Proportional sets $W^{(T)} \propto n^{(T)}$. Adaptive (sqrt) sets $W^{(T)} \propto \sqrt{n^{(T)}\tau_{1024}^{(T)}}$. We sweep $B \in \{0.01n, 0.025n, 0.05n, 0.1n, 0.25n\}$, single seed.
Adaptive (sqrt) achieves lower perplexity than uniform at every budget except the most constrained ($B = 0.01n$) and lower perplexity than proportional at four of five budget points. Single seed; gaps at moderate budgets are within typical seed variance for short LM training.

Appendix G. Cross-domain validation: CIFAR-100 / ResNet-18

The certificate separation predicts the largest gains when the certified residual mass concentrates on a small number of tensors. We train ResNet-18 on CIFAR-100 for 10,000 steps with the auto-tensor candidate generator and sweep $\mathrm{mpp} \in \{1, 2, 4, 8, 16\}$ at state-price 0.
The LCB concentrates state on a single tensor: the best DOSA configuration ($\mathrm{mpp} = 1$) makes one promotion — the largest convolutional weight — and the resulting 10% state fraction recovers 1.2pp over the factored baseline. Aggressive promotion flips the gain into a loss. The structural regime is therefore not a transformer artifact.

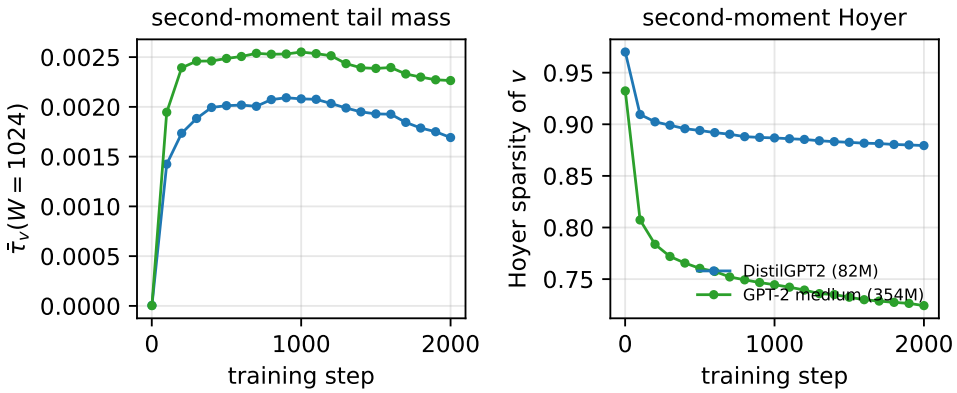

Figure 2: Heavy-tail concentration of the second moment v through training. Left: weighted tail mass $\bar{\tau}_v(W{=}1024)$ over 2,000 AdamW steps. Right: Hoyer sparsity of v. Both scales remain in the same residual-concentration regime at every measured step.

Table 2: Anchored residual blend $\hat{v}_T(\alpha) = v_T^{(0)} + \alpha(\widetilde{v}_T - v_T^{(0)})$ validation perplexity at fixed per-tensor sketch budget $B = 0.1n$ on WikiText-2 (5,000 steps, promotions suppressed). Anchor: factored Adafactor. Correction: CountSketch. Headline rows are 3 seeds (mean $\pm$ std); MSE-framework rows are single-seed pilot data. Bold: best per model.

Variant	val_ppl
Memory-efficient baselines (DistilGPT2, 3 seeds):	
`torch.optim.Adafactor` (vanilla)	51.27 ± 0.10
`bitsandbytes.AdamW8bit`	61.68 ± 0.45
Blend on DistilGPT2 (3 seeds for headline):	
Factored Adafactor anchor ($\alpha = 0$)	48.46 ± 0.52
Tuned $\alpha = 0.75$ **(uniform** W**)**	$\mathbf{45.19 \pm 0.64}$
$\alpha_T^\star$ at uniform W (per-step v-MSE, no EMA correction)	47.39
$\alpha = \alpha_{\text{global}}^{\star,\kappa,\text{EMA}} = 0.16$ *(Eq. 9, surrogate endpoint)*	46.74 ± 0.52
LCB-greedy on continuous-α grid (one-step score plus LCB penalty)	51.95 ± 0.71
Blend on GPT-2-small / WikiText-2 (3 seeds):	
Factored Adafactor anchor ($\alpha = 0$)	40.77 ± 0.40
Tuned $\alpha = 0.75$	$\mathbf{36.81 \pm 0.23}$
Blend on DistilGPT2 / WikiText-103 (3 seeds):	
Factored Adafactor anchor ($\alpha = 0$)	47.14 ± 0.35
Tuned $\alpha = 0.75$	$\mathbf{43.66 \pm 0.36}$

Table 3: Validation perplexity at fixed total state budget on DistilGPT2/WikiText-2 (5,000 steps, single seed). Lower is better. Bold: best per row.

Budget B	uniform	proportional	adaptive (linear)	adaptive (sqrt)
$0.01n$	**48.20**	48.91	49.31	48.24
$0.025n$	48.94	**47.86**	48.66	48.58
$0.05n$	48.87	48.43	**48.02**	48.31
$0.1n$	48.61	48.30	48.30	**47.73**
$0.25n$	48.41	48.30	48.30	**47.22**

Table 4: DOSA on CIFAR-100/ResNet-18 (10k steps, seed 0). Best DOSA configuration recovers 1.2pp over the factored `low_state_adamw` baseline at 10% optimizer state.

Variant	Val. acc.	State frac.	# prom.
`full_adamw`	0.5055	1.000	0
`cosa` (mpp=1)	**0.4704**	0.100	1
`cosa_horizon`	0.4674	0.100	100
`cosa` (mpp=2)	0.4562	0.100	2
`cosa` (mpp=4)	0.4596	0.100	4
`cosa` (mpp=8)	0.4525	0.100	8
`cosa` (mpp=16)	0.4520	0.100	11
`low_state_adamw`	0.4584	0.002	0