Outer-Momentum Restarting in High-Dimensional Two-Phase Optimization

Kristi Topollai KT2664@NYU.EDU
Allan Ma ALLAN.MA@NYU.EDU
Tolga Dimlioglu TD2249@NYU.EDU
Sui Jiet Tay ST5494@NYU.EDU
Anna Choromanska AC5455@NYU.EDU
New York University, New York, USA

Abstract

Communication-efficient distributed optimizers such as DiLoCo reduce synchronization costs by letting workers perform many local updates before aggregating their progress with an outer momentum optimizer. Recent theory suggests that the outer optimizer acts on an effective spectrum induced by the inner optimization loop, and that the choice of outer momentum controls how progress from local updates is accumulated across communication rounds. We study periodic restarting of the outer momentum as a simple complementary mechanism for controlling this outer memory. In a linearized squared-loss model where prediction-space residuals evolve under the empirical NTK, we derive a mode-wise restart contraction showing that resets exploit phase cancellation by discarding stale momentum while preserving inner-loop progress. Toy experiments verify the predicted contraction behavior, and language-model pretraining experiments show that periodic restarts widen the stable range of outer learning rates and momentum values across communication periods.

1. Introduction

Modern language-model (LM) training is increasingly constrained by communication in distributed setups. In standard synchronous data-parallel training, workers exchange gradients or optimizer states at every step, which becomes costly when accelerators such as GPUs or TPUs are distributed across machines or clusters. This led to renewed interest in local and *two-phase* training methods, where workers perform several local optimization steps before synchronizing [3, 4, 15, 18, 19, 21]. In this view, training is split into an inner phase, which performs independent local progress, and an outer phase, which aggregates that progress across workers. Similar to Lookahead methods [26], DiLoCo brings this structure to large-scale LM training: workers run many inner steps, form a pseudo-gradient from their parameter displacement, and pass it to an outer optimizer [7, 8]. Empirically, DiLoCo and its variants can substantially reduce communication while maintaining training quality, and recent work suggests that some benefits arise even in the one-worker setting [7, 14]. This way, the outer optimizer determines how accumulated local progress is filtered, or damped at communication time.

Recent works on local SGD and federated optimization study how outer learning rates and momentum trade off optimization speed against stochastic noise [4, 15, 21]. The closest theoretical lens is provided in Agarwala [1], where it is shown that the inner loop induces an effective spectrum on which the outer optimizer operates. Their analysis shows that heavy-ball outer momentum can

© K. Topollai, A. Ma, T. Dimlioglu, S.J. Tay & A. Choromanska.

constrain directions on which the inner loop already makes strong progress, while Nesterov outer momentum mitigates this by making contraction depend on the effective progress of each mode [1]. This is consistent with recent empirical gains from Nesterov pseudo-gradient updates [14].

We study momentum restarting as a complementary mechanism for two-phase optimizers. Restarting has a long history in accelerated optimization, where resetting momentum can reduce oscillation or correct misalignment [9, 10, 16, 20], with related restart ideas being used in deep learning to improve stability [12, 23, 24]. In two-phase training, the outer momentum buffer stores pseudo-gradient history only at communication rounds, which means that outdated outer memory can affect how local progress is aggregated across synchronizations. We show that periodically resetting this buffer reduces fragility to the outer momentum and learning rate. Our theory explains the effect through a phase-dependent contraction, while our 150M-parameter LLaMA pretraining experiments show that restarts widen the stable region of outer-optimizer hyperparameters, a useful property in pretraining where optimizer tuning is already delicate [5, 25] and two-phase methods introduce an additional outer-loop hyperparameter layer.

2. Preliminaries: deterministic two-phase mode dynamics

We introduce a deterministic mode-wise model to isolate how the outer optimizer processes the progress accumulated by the inner loop. To do so, we follow the two-phase analysis of [1]. The object being analyzed is the training residual in prediction space, $R(\theta) = f(\theta) - y_{\mathrm{tr}}$ where $f(\theta)$ denotes the vector of model predictions on the training set and y_{tr} denotes the training targets. Under the standard linearized squared-loss/NTK approximation, the residual dynamics are governed by the empirical kernel H. After diagonalizing this kernel, the residual decomposes into independent scalar modes, so we use a deterministic scalar-mode surrogate for the residual dynamics rather than modeling the full nonlinear parameter trajectory [2, 13, 17].

We analyze one residual eigencoordinate at a time. Let $\lambda \geq 0$ denote the corresponding empirical-kernel eigenvalue. Within one communication round, the inner optimizer performs S gradient-descent steps with step size η where $0 \leq \eta\lambda \leq 1$. If x_t is the value of this residual mode at the beginning of outer round t, then after the inner steps,

$$x_t^{\mathrm{loc}} = (1-\eta\lambda)^S x_t, \qquad g_t := x_t - x_t^{\mathrm{loc}} = \sigma x_t, \qquad \sigma := 1-(1-\eta\lambda)^S. \tag{1}$$

Here x_t^{loc} denotes the value of the same residual mode after the worker has completed S local inner steps and g_t is the pseudo-gradient passed to the outer optimizer. The coefficient $\sigma \in [0,1]$ is the effective progress seen by the outer optimizer at communication time and large σ means that the inner loop has already made substantial progress on that residual mode.

For heavy-ball/EMA outer momentum, let m_t denote the outer momentum buffer. At outer round t, the pseudo-gradient g_t is accumulated through $m_{t+1} = \beta_{\mathrm{out}} m_t + (1-\beta_{\mathrm{out}}) g_t$, and the residual eigencoordinate is updated as $x_{t+1} = x_t - \nu m_{t+1}$. Since the inner loop gives $g_t = \sigma x_t$, the pair consisting of the residual mode and its outer momentum, $z_t = (x_t, m_t)^\top$, evolves linearly across communication rounds. Substituting $g_t = \sigma x_t$ into the two scalar updates yields

$$z_{t+1} = T_{\mathrm{HB}}(\sigma) z_t \qquad \text{where} \qquad T_{\mathrm{HB}}(\sigma) = \begin{pmatrix} 1-\nu(1-\beta_{\mathrm{out}})\sigma & -\nu\beta_{\mathrm{out}} \\ (1-\beta_{\mathrm{out}})\sigma & \beta_{\mathrm{out}} \end{pmatrix}. \tag{2}$$

Since $\det(T_{\mathrm{HB}}) = \beta_{\mathrm{out}}$, complex-regime eigenvalues have magnitude $\sqrt{\beta_{\mathrm{out}}}$, giving the σ-independent damping rate $r_\infty = -\frac{1}{2}\log\beta_{\mathrm{out}}$. This is the outer-momentum damping effect, that is, even high-progress modes with $\sigma \approx 1$ can be limited by the same slow HB envelope.

For Nesterov-style outer momentum, the buffer update is unchanged, but the outer step uses $x_{t+1} = x_t - \nu((1+\beta_{\text{out}})m_{t+1} - \beta_{\text{out}} m_t)$, which gives

$$T_{\text{NAG}}(\sigma) = \begin{pmatrix} 1-\nu(1-\beta_{\text{out}}^2)\sigma & -\nu\beta_{\text{out}}^2 \\ (1-\beta_{\text{out}})\sigma & \beta_{\text{out}} \end{pmatrix}, \qquad \det(T_{\text{NAG}}) = \beta_{\text{out}}\big(1-(1-\beta_{\text{out}})\nu\sigma\big). \tag{3}$$

3. Restarting Outer Momentum: An Old Trick Revisited

The preceding discussion explains why Nesterov outer momentum is effective in two-phase training. For HB, complex modes share the same envelope $\rho = \sqrt{\beta_{\text{out}}}$, so the outer update may remain slow even when the inner loop has made substantial progress. NAG addresses this spectrally as its contraction depends on $\nu\sigma$, making the outer optimizer more responsive to high-progress modes. We study a complementary mechanism: periodic restarting of the outer momentum buffer. Momentum restarting is classical in accelerated optimization, where resets can reduce oscillations or correct momentum misalignment [9, 10, 16, 20, 22]. In our setting, after every K outer rounds, we reset only the outer buffer, $m \leftarrow 0$. Unlike NAG, restarting does not modify the one-step spectrum; it modifies the duration over which outer memory is retained. In the complex regime, the residual coordinate oscillates within an envelope, so an appropriately chosen restart period can stop the trajectory near a phase cancellation, where the projected residual is much smaller than the envelope. Thus NAG and restarting act through distinct mechanisms: NAG changes the contraction factor, whereas restarting exploits favorable reset times. Starting a restart cycle with zero momentum, $m \leftarrow 0$, one cycle of K outer rounds maps the initial residual mode to

$$x_K = \chi_K(\sigma)x_0, \qquad \chi_K(\sigma) := \big[T_{\text{HB}}(\sigma)^K\big]_{11}, \qquad r_K(\sigma) := -\frac{1}{K}\log|\chi_K(\sigma)|.$$

Thus χ_K is simply the scalar contraction of the residual mode after one restart cycle, and r_K is its average per-round contraction rate.

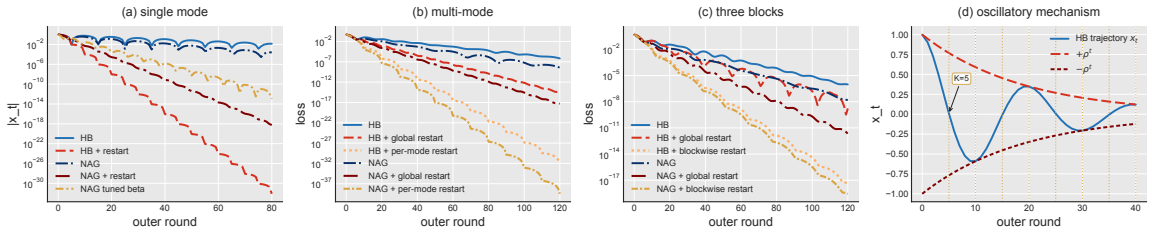


Figure 1: Restart mechanism across scalar, multi-mode, and blockwise settings. Periodic restart exploits phase cancellation in the outer momentum dynamics. In the scalar case, HB with restart can match a tuned no-restart NAG baseline; in heterogeneous spectra, global restart improves over no restart, while per-mode or blockwise periods capture additional gains. Panel (d) visualizes the underlying oscillatory mechanism: the residual coordinate moves inside the HB envelope $\pm\rho^t$, and restarting near a cancellation time makes the projected residual small.

Proposition 1 (Restart contraction) *Let $a = 1-\nu(1-\beta_{\text{out}})\sigma$. Then*

$$\chi_K = (a+\beta_{\text{out}})\chi_{K-1} - \beta_{\text{out}}\chi_{K-2}, \qquad \chi_0 = 1, \quad \chi_1 = a.$$

In the complex regime, write the eigenvalues of T_{HB} as $\rho e^{\pm i\varphi}$, where $\rho = \sqrt{\beta_{\text{out}}}$, $\cos\varphi = \frac{a+\beta_{\text{out}}}{2\rho}$. Then

$$\chi_K(\sigma) = \rho^K\Big[\cos(K\varphi) + C\sin(K\varphi)\Big], \qquad C = \frac{a-\beta_{\text{out}}}{2\rho\sin\varphi}.$$

For the high-momentum outer optimizers considered here, the complex regime is not a pathological case but the typical operating regime. The condition $\mathrm{tr}(T_{\mathrm{HB}})^2 < 4\det(T_{\mathrm{HB}})$ is equivalent to

$$\frac{1-\sqrt{\beta_{\mathrm{out}}}}{\nu(1+\sqrt{\beta_{\mathrm{out}}})} < \sigma < \frac{1+\sqrt{\beta_{\mathrm{out}}}}{\nu(1-\sqrt{\beta_{\mathrm{out}}})}.$$

When β_{out} is close to 1, this interval covers essentially all nontrivial effective progress values $\sigma \in (0,1]$ for $\nu = O(1)$: the lower endpoint is close to zero, while the upper endpoint is far larger than one. Thus, the oscillatory regime captured by Proposition 1 is the relevant one for typical outer momentum choices; the derivation is given in Appendix A.6. In this regime, the closed form decomposes the restarted dynamics into the usual HB envelope ρ^K and a phase-dependent projection term. Restart improves over the non-restarted HB envelope precisely when this projection term has magnitude below one.

Proposition 2 (K-period crossover) *In the complex regime, a K-period restart improves over the non-restarted HB envelope iff*

$$r_K(\sigma) > r_\infty \quad \Longleftrightarrow \quad |\chi_K(\sigma)| < \rho^K \quad \Longleftrightarrow \quad |\cos(K\varphi) + C\sin(K\varphi)| < 1.$$

Equivalently, writing $\cos(K\varphi) + C\sin(K\varphi) = \sqrt{1+C^2}\cos(K\varphi - \theta)$, $\theta = \arctan C$, *useful periods are those whose phase lands near a cancellation point,* $K_\ell \approx \left\lfloor \frac{\theta+\pi/2+\ell\pi}{\varphi} \right\rceil$, $\ell \in \mathbb{N}^+$.

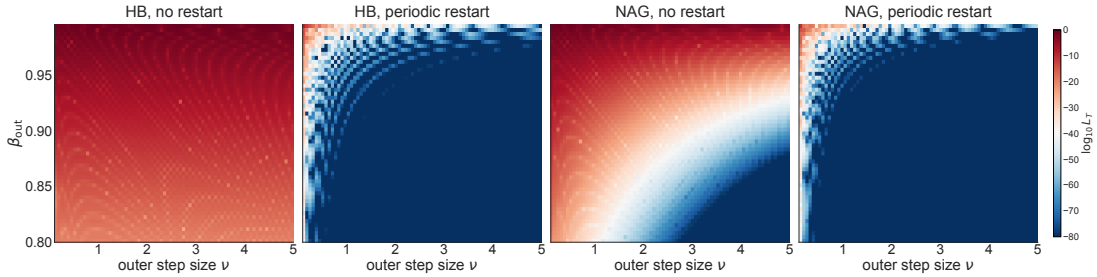


Figure 2: Robustness over outer hyperparameters. Each heatmap shows clipped $\log_{10}$ final loss over a grid of β_{out} and ν. The best periodic restart enlarges the good-hyperparameter region for HB and NAG.

In experiments, we select K from an admissible integer range by minimizing $|\chi_K(\sigma)|$, equivalently maximizing the restarted rate r_K. The phase interpretation clarifies why no single period is uniformly optimal as large-σ components rotate faster and favor shorter periods, whereas small-σ components rotate more slowly and favor longer periods. This tradeoff is illustrated in Figure 1, which compares global, per-mode, and blockwise restart periods as spectral heterogeneity increases. We apply the same projected-factor construction to NAG, and defer the NAG recurrence, closed form, and blockwise extension to the Appendix. Figure 2 further shows that these restart periods expand the region of outer learning rates and momentum values that yield stable performance.

4. Experiments

We now present our empirical results for LM pretraining. We follow the general DiLoCo training and pretrain a 150M-parameter LLaMA model in a 2-replica DiLoCo configuration using two H200 GPUs. All runs use a fixed training budget of approximately 3.3B tokens, following the Chinchilla scaling rule for a 150M-parameter model [11]. Additional details, results, ablations, and the full set of explored hyperparameter configurations are reported in Section B of Appendix.

Restarts widen the stable outer-hyperparameter region. Figure 3 shows validation perplexity across the outer learning rate ν and momentum β_{out}. Without restarts, large β_{out} creates a clear failure region, especially at longer communication periods. Periodic restarts largely remove this instability and expand the range of hyperparameters that achieve good performance. Thus the main benefit is robustness, since restarts make the outer optimizer less sensitive to ν and β_{out}.

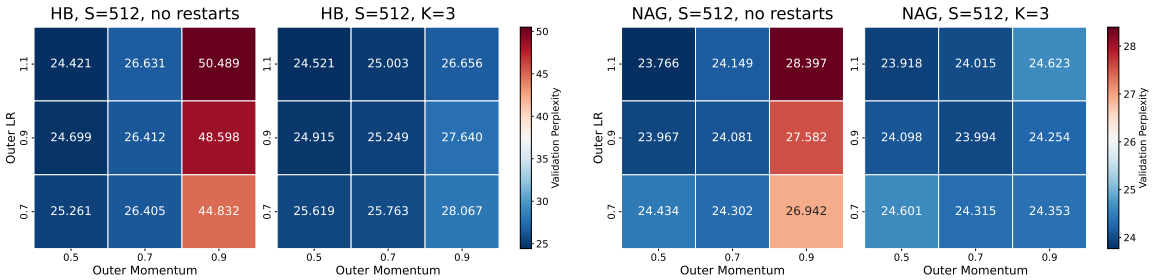


Figure 3: Validation perplexity over the outer hyperparameter grid at $S = 512$ for HB(left) and NAG(right), comparing standard DiLoCo against DiLoCo with momentum restart period $K = 3$. Periodic restarts reduce the high-β_{out} failure region while preserving peak performance.

Restarts reduce retuning sensitivity across communication periods. We next test how much retuning is needed as the communication period changes. We tune DiLoCo once at $S = 128$, obtaining $(\nu^*, \beta^*_{\text{out}}) = (0.9, 0.7)$ for NAG and $(1.1, 0.5)$ for HB. For no-restart runs, we keep $\nu = \nu^*$ and sweep β_{out}, while for runs with restart, we keep $(\nu, \beta_{\text{out}}) = (\nu^*, \beta^*_{\text{out}})$ and sweep only the restart period K. Thus, we compare tuning the outer momentum coefficient, against tuning the restart period. Figure 4 shows that not restarting exhibits sharp degradation at large momentum coefficients.

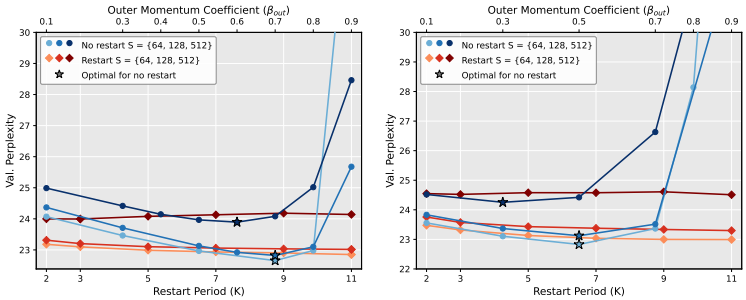


Figure 4: Final validation perplexity across communication periods $S \in \{64, 128, 512\}$, for NAG (left) and HB (right). No-restart curves fix $\nu = \nu^*$ and sweep β_{out}; restart curves fix $(\nu, \beta_{\text{out}}) = (\nu^*, \beta^*_{\text{out}})$ and sweep K.

By contrast, restart curves remain comparatively flat across K, and similar periods transfer across communication lengths. Hence K serves as a more forgiving hyperparameter than β_{out}. This effect is most pronounced for HB, where large β_{out} can diverge, while restarts remain stable.

5. Conclusion

We studied periodic outer-momentum restarts for two-phase optimization and found that they can widen the stable hyperparameter region by limiting unhelpful outer memory. Although our analysis focuses on fixed-period resets, the empirical results suggest that the method is fairly insensitive to the precise period, making periodic restarting a simple and robust baseline. This opens a broader question: can the outer loop decide for itself when its memory should be refreshed? Future work should develop adaptive restart rules, per-layer or blockwise periods, and larger-scale studies of how restart schedules vary with communication period and model size.

References

[1] Atish Agarwala. High dimensional theory of two-phase optimizers. *arXiv preprint arXiv:2603.26954*, 2026. URL https://arxiv.org/abs/2603.26954.

[2] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1904.11955.

[3] Zachary Charles and Jakub Konečný. Convergence and accuracy trade-offs in federated learning and meta-learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2575–2583. PMLR, 2021. URL https://proceedings.mlr.press/v130/charles21a.html.

[4] Zachary Charles and Keith Rush. Iterated vector fields and conservatism, with applications to federated learning. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 130–147. PMLR, 2022. URL https://proceedings.mlr.press/v167/charles22a.html.

[5] Zachary Charles, Gabriel Teston, Lucio Dery, Keith Rush, Nova Fallen, Zachary Garrett, Arthur Szlam, and Arthur Douillard. Communication-efficient language model training scales reliably and robustly: Scaling laws for DiLoCo. *arXiv preprint arXiv:2503.09799*, 2025. doi: 10.48550/arxiv.2503.09799. URL https://arxiv.org/abs/2503.09799.

[6] Zhaorui Dong, Yushun Zhang, Zhi-Quan Luo, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025. doi: 10.48550/arxiv.2505.02809. URL https://arxiv.org/abs/2505.02809.

[7] Arthur Douillard, Qixuan Feng, Andrei A. Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc'Aurelio Ranzato, Arthur Szlam, and Jiajun Shen. DiLoCo: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2023. doi: 10.48550/arxiv.2311.08105. URL https://arxiv.org/abs/2311.08105.

[8] Arthur Douillard, Yani Donchev, J. Keith Rush, Satyen Kale, Zachary Charles, Gabriel Teston, Zachary Garrett, Jiajun Shen, Ross McIlroy, David Lacey, Alexandre Ramé, Arthur Szlam, Marc'Aurelio Ranzato, and Paul R. Barham. Streaming DiLoCo with overlapping communication: Towards a distributed free lunch. In *Second Conference on Language Modeling*, 2025. doi: 10.48550/arxiv.2501.18512. URL https://openreview.net/forum?id=yYk3zK0X6Q.

[9] Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019. doi: 10.1093/imanum/drz007. URL https://arxiv.org/abs/1709.02300.

[10] Pontus Giselsson and Stephen Boyd. Monotonicity and restart in fast gradient methods. In *53rd IEEE Conference on Decision and Control*, pages 5058–5063. IEEE, 2014. doi:

10.1109/cdc.2014.7040179. URL https://web.stanford.edu/~boyd/papers/restart_fgm.html.

[11] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2176. URL https://arxiv.org/abs/2203.15556.

[12] Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware adam with momentum reset for stable LLM training. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arxiv.2501.06842. URL https://arxiv.org/abs/2501.06842.

[13] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1806.07572.

[14] Dominik Kallusky, Vinay Rao, Vishal Nandavanam, and Hao-Jun Michael Shi. SNOO: Step-K nesterov outer optimizer – the surprising effectiveness of nesterov momentum applied to pseudo-gradients. *arXiv preprint arXiv:2510.15830*, 2025. doi: 10.48550/arxiv.2510.15830. URL https://arxiv.org/abs/2510.15830.

[15] Ahmed Khaled, Satyen Kale, Arthur Douillard, Chi Jin, Rob Fergus, and Manzil Zaheer. Understanding outer optimizers in local SGD: Learning rates, momentum, and acceleration. *arXiv preprint arXiv:2509.10439*, 2025. doi: 10.48550/arxiv.2509.10439. URL https://arxiv.org/abs/2509.10439.

[16] Donghwan Kim and Jeffrey A. Fessler. Adaptive restart of the optimized gradient method for convex optimization. *Journal of Optimization Theory and Applications*, 178(1):240–263, 2018. doi: 10.1007/s10957-018-1287-4. URL https://doi.org/10.1007/s10957-018-1287-4.

[17] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1902.06720.

[18] Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtárik. From local SGD to local fixed-point methods for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6692–6701. PMLR, 2020. URL https://proceedings.mlr.press/v119/malinovskiy20a.html.

[19] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Pro-*

ceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

[20] Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015. doi: 10.1007/s10208-013-9150-3. URL https://doi.org/10.1007/s10208-013-9150-3.

[21] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

[22] Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/6712-sharpness-restart-and-acceleration.

[23] Kristi Topollai and Anna Choromanska. Understanding quantization of optimizer states in LLM pre-training: Dynamics of state staleness and effectiveness of state resets. *arXiv preprint arXiv:2603.16731*, 2026. URL https://arxiv.org/abs/2603.16731.

[24] Bao Wang, Tan M. Nguyen, Andrea L. Bertozzi, Richard G. Baraniuk, and Stanley J. Osher. Scheduled restart momentum for accelerated stochastic gradient descent. *SIAM Journal on Imaging Sciences*, 15(2):738–761, 2022. doi: 10.1137/21M1453311. URL https://doi.org/10.1137/21M1453311.

[25] Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. *arXiv preprint arXiv:2509.02046*, 2025. doi: 10.48550/arxiv.2509.02046. URL https://arxiv.org/abs/2509.02046.

[26] Michael R. Zhang, James Lucas, Jimmy Ba, and Geoffrey E. Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/90fd4f88f588ae64038134f1eeaa023f-Abstract.html.

[27] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.48550/arxiv.2402.16788. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ee0e45ff4de76cbfdf07015a7839f339-Abstract-Conference.html.

Appendix A. Derivations for the two-phase restart dynamics

A.1. From the quadratic model to the scalar pseudo-gradient

Consider the quadratic residual model

$$\ell(x) = \frac{1}{2}x^\top Hx, \qquad H \succeq 0.$$

Diagonalize $H = U\Lambda U^\top$. Since the dynamics are linear, each eigendirection evolves independently. For an eigenvalue λ, S inner GD steps with stepsize η give

$$x_t^{\mathrm{loc}} = (1-\eta\lambda)^S x_t.$$

The pseudo-gradient passed to the outer optimizer is the displacement produced by the inner loop,

$$g_t := x_t - x_t^{\mathrm{loc}} = \left[1-(1-\eta\lambda)^S\right] x_t = \sigma x_t, \qquad \sigma := 1-(1-\eta\lambda)^S.$$

Thus σ is the effective eigenvalue observed at communication time.

A.2. Heavy-ball/EMA outer transition

Throughout this section, t indexes outer communication rounds. One outer transition already includes the S inner steps, whose effect is absorbed into $\sigma = 1-(1-\eta\lambda)^S$.

The heavy-ball/EMA outer update is

$$m_{t+1} = \beta_{\mathrm{out}} m_t + (1-\beta_{\mathrm{out}})g_t, \qquad x_{t+1} = x_t - \nu m_{t+1}.$$

Substituting $g_t = \sigma x_t$ gives

$$m_{t+1} = (1-\beta_{\mathrm{out}})\sigma x_t + \beta_{\mathrm{out}} m_t, \qquad x_{t+1} = \big(1-\nu(1-\beta_{\mathrm{out}})\sigma\big)x_t - \nu\beta_{\mathrm{out}} m_t.$$

Hence, with $z_t = (x_t, m_t)^\top$,

$$z_{t+1} = T_{\mathrm{HB}}(\sigma) z_t, \qquad T_{\mathrm{HB}}(\sigma) = \begin{pmatrix} 1-\nu(1-\beta_{\mathrm{out}})\sigma & -\nu\beta_{\mathrm{out}} \\ (1-\beta_{\mathrm{out}})\sigma & \beta_{\mathrm{out}} \end{pmatrix}.$$

Let $a := 1-\nu(1-\beta_{\mathrm{out}})\sigma$. Then

$$\mathrm{tr}(T_{\mathrm{HB}}) = a + \beta_{\mathrm{out}},$$

and a direct computation gives

$$\det(T_{\mathrm{HB}}) = a\beta_{\mathrm{out}} - (-\nu\beta_{\mathrm{out}})\cdot(1-\beta_{\mathrm{out}})\sigma = \big(1-\nu(1-\beta_{\mathrm{out}})\sigma\big)\beta_{\mathrm{out}} + \nu\beta_{\mathrm{out}}(1-\beta_{\mathrm{out}})\sigma = \beta_{\mathrm{out}}.$$

The cancellation of the $\nu(1-\beta_{\mathrm{out}})\sigma$ terms is the algebraic origin of the σ-independent damping rate: the determinant of the outer transition is locked at β_{out}, regardless of how much progress σ the inner loop made.

Whenever $(a+\beta_{\mathrm{out}})^2 < 4\beta_{\mathrm{out}}$, the two eigenvalues are complex conjugates. Since their product is β_{out}, they can be written as

$$\lambda_\pm = \rho e^{\pm i\varphi}, \qquad \rho = \sqrt{\beta_{\mathrm{out}}}, \qquad \cos\varphi = \frac{\mathrm{tr}(T_{\mathrm{HB}})}{2\rho} = \frac{a+\beta_{\mathrm{out}}}{2\rho}.$$

The non-restarted asymptotic envelope is therefore ρ^t, giving the per-outer-round damping rate $r_\infty = -\log\rho = -\frac{1}{2}\log\beta_{\mathrm{out}}$.

A.3. Nesterov-style outer transition

For NAG, the same momentum buffer is used, but the outer step applies the lookahead combination

$$m_{t+1} = \beta_{\text{out}} m_t + (1-\beta_{\text{out}}) g_t, \qquad x_{t+1} = x_t - \nu\big((1+\beta_{\text{out}}) m_{t+1} - \beta_{\text{out}} m_t\big).$$

Using $g_t = \sigma x_t$,

$$(1+\beta_{\text{out}}) m_{t+1} - \beta_{\text{out}} m_t = (1-\beta_{\text{out}}^2)\sigma x_t + \beta_{\text{out}}^2 m_t.$$

Therefore

$$z_{t+1} = T_{\text{NAG}}(\sigma) z_t, \qquad T_{\text{NAG}}(\sigma) = \begin{pmatrix} 1-\nu(1-\beta_{\text{out}}^2)\sigma & -\nu\beta_{\text{out}}^2 \\ (1-\beta_{\text{out}})\sigma & \beta_{\text{out}} \end{pmatrix}.$$

Its determinant is

$$\begin{aligned} \det(T_{\text{NAG}}) &= \beta_{\text{out}}\big(1-\nu(1-\beta_{\text{out}}^2)\sigma\big) + \nu\beta_{\text{out}}^2(1-\beta_{\text{out}})\sigma \\ &= \beta_{\text{out}} - \nu\beta_{\text{out}}(1-\beta_{\text{out}})\sigma = \beta_{\text{out}}\big(1-(1-\beta_{\text{out}})\nu\sigma\big). \end{aligned}$$

This dependence on $\nu\sigma$ is the spectral advantage of NAG in the two-phase dynamics.

A.4. Restart contraction factor and recurrence

Suppose the outer momentum is reset every K outer rounds. At the beginning of a restart block, $m_0 = 0$, so

$$z_0 = (x_0, 0)^\top = x_0 e_1, \qquad e_1 = (1,0)^\top.$$

After K outer rounds of HB/EMA dynamics,

$$x_K = e_1^\top T_{\text{HB}}(\sigma)^K z_0 = \chi_K(\sigma) x_0, \qquad \chi_K(\sigma) := e_1^\top T_{\text{HB}}(\sigma)^K e_1.$$

The per-outer-round restarted rate is

$$r_K(\sigma) = -K^{-1} \log |\chi_K(\sigma)|.$$

The characteristic polynomial of T_{HB} is $q(\lambda) = \lambda^2 - (a+\beta_{\text{out}})\lambda + \beta_{\text{out}}$. By Cayley–Hamilton,

$$T_{\text{HB}}^2 = (a+\beta_{\text{out}}) T_{\text{HB}} - \beta_{\text{out}} I.$$

Multiplying by T_{HB}^{K-2} and projecting with $e_1^\top$ and e_1 gives

$$\chi_K = (a+\beta_{\text{out}})\chi_{K-1} - \beta_{\text{out}}\chi_{K-2}, \qquad \chi_0 = 1, \qquad \chi_1 = a.$$

This recurrence holds in all regimes, with real or complex eigenvalues; only its closed-form solution depends on the regime.

A.5. Closed form in the complex regime

In the complex regime, the eigenvalues are $\rho e^{\pm i\varphi}$, so any real sequence satisfying the recurrence has the form

$$\chi_K = \rho^K \big(A\cos(K\varphi) + B\sin(K\varphi)\big).$$

The initial condition $\chi_0 = 1$ gives $A = 1$. The condition $\chi_1 = a$ then gives

$$a = \rho\big(\cos\varphi + B\sin\varphi\big), \qquad B = \frac{a/\rho - \cos\varphi}{\sin\varphi} = \frac{a - \beta_{\text{out}}}{2\rho\sin\varphi},$$

where the second equality uses $\rho\cos\varphi = (a + \beta_{\text{out}})/2$. Thus

$$\boxed{\chi_K(\sigma) = \rho^K\left[\cos(K\varphi) + C\sin(K\varphi)\right]}, \qquad C = \frac{a - \beta_{\text{out}}}{2\rho\sin\varphi}.$$

The factor ρ^K is the same envelope as the non-restarted method, while the bracket is a phase-dependent projection onto the iterate coordinate.

A.6. When are we in the complex regime?

The closed form above and most of our analysis assumes the complex regime, $(a+\beta_{\text{out}})^2 < 4\beta_{\text{out}}$. Substituting $a = 1 - \nu(1-\beta_{\text{out}})\sigma$, this becomes

$$\big(1 + \beta_{\text{out}} - \nu(1-\beta_{\text{out}})\sigma\big)^2 < 4\beta_{\text{out}} \quad \Longleftrightarrow \quad |1 + \beta_{\text{out}} - \nu(1-\beta_{\text{out}})\sigma| < 2\sqrt{\beta_{\text{out}}}.$$

The right-hand inequality factors as $\nu(1-\beta_{\text{out}})\sigma < 1 + \beta_{\text{out}} + 2\sqrt{\beta_{\text{out}}} = (1+\sqrt{\beta_{\text{out}}})^2$; the left-hand one as $\nu(1-\beta_{\text{out}})\sigma > (1-\sqrt{\beta_{\text{out}}})^2$. Using $1-\beta_{\text{out}} = (1-\sqrt{\beta_{\text{out}}})(1+\sqrt{\beta_{\text{out}}})$, this simplifies to

$$\frac{1-\sqrt{\beta_{\text{out}}}}{\nu(1+\sqrt{\beta_{\text{out}}})} < \sigma < \frac{1+\sqrt{\beta_{\text{out}}}}{\nu(1-\sqrt{\beta_{\text{out}}})}.$$

For β_{out} close to 1, the lower bound is small (of order $(1-\beta_{\text{out}})/(4\nu)$) and the upper bound is large (of order $4/(\nu(1-\beta_{\text{out}}))$). For example, at $\beta_{\text{out}} = 0.9$ and $\nu = 1$ the regime is $\sigma \in (0.026, 38)$; at $\beta_{\text{out}} = 0.99$ and $\nu = 1$ it is $\sigma \in (0.0025, 398)$. Effectively all practically relevant $\sigma \in (0, 1]$ fall in the complex regime for the outer momenta typically used in DiLoCo.

In the boundary case $(a+\beta_{\text{out}})^2 = 4\beta_{\text{out}}$ (critical damping), the eigenvalues collide at ρ; outside the complex regime, they are real and distinct, so the iterate decays without oscillation and no K zeros the bracket. Outside the complex regime, the eigenvalues are real and distinct, so the iterate no longer exhibits the oscillatory phase-cancellation mechanism analyzed above. Thus, the restart benefit predicted by this phase argument is not expected in the same way in the real-eigenvalue regime, consistent with the standard intuition that restarting is most useful when momentum induces oscillatory transients.

A.7. Oracle restart period

The single-mode oracle period minimizes the restart contraction factor:

$$K^\star(\sigma) \in \arg\min_{K\in\mathbb{N}} |\chi_K(\sigma)|.$$

Since ρ^K varies smoothly with K, the main cancellation mechanism is the oscillatory bracket. Write

$$\cos(K\varphi) + C\sin(K\varphi) = \sqrt{1+C^2}\cos(K\varphi - \theta), \qquad \theta = \arctan C.$$

A zero of the bracket occurs when

$$K\varphi - \theta = \frac{\pi}{2} + \ell\pi, \qquad \ell \in \mathbb{Z}.$$

The first positive near-cancellation is therefore approximated by

$$K^\star(\sigma) \approx \left\lfloor \frac{\theta + \pi/2}{\varphi} \right\rceil,$$

where $\lfloor\cdot\rceil$ denotes rounding to the nearest positive integer. The rounding error costs at most a factor of $\sin(\varphi/2)\sqrt{1+C^2} \approx (\varphi/2)\sqrt{1+C^2}$ in the bracket amplitude, which is small for β_{out} close to 1 (since then φ is small). The phase interpretation also explains the qualitative dependence on σ: increasing σ increases the rotation frequency φ, so high-σ modes prefer shorter restart periods.

A.8. NAG restart factor and closed form

For NAG, we use the same projected-factor definition:

$$\chi_K^{\text{NAG}}(\sigma) := e_1^\top T_{\text{NAG}}(\sigma)^K e_1, \qquad r_K^{\text{NAG}}(\sigma) := -\frac{1}{K}\log|\chi_K^{\text{NAG}}(\sigma)|.$$

Let $a_{\text{N}} := 1 - \nu(1-\beta_{\text{out}}^2)\sigma$ and $D_{\text{N}} := \det(T_{\text{NAG}}) = \beta_{\text{out}}(1-(1-\beta_{\text{out}})\nu\sigma)$. The characteristic polynomial of T_{NAG} is $\lambda^2 - (a_{\text{N}} + \beta_{\text{out}})\lambda + D_{\text{N}}$, so by Cayley–Hamilton the same projection argument as in §A.4 gives

$$\chi_K^{\text{NAG}} = (a_{\text{N}} + \beta_{\text{out}})\,\chi_{K-1}^{\text{NAG}} - D_{\text{N}}\,\chi_{K-2}^{\text{NAG}}, \qquad \chi_0^{\text{NAG}} = 1, \qquad \chi_1^{\text{NAG}} = a_{\text{N}}.$$

When $D_{\text{N}} > 0$ and $(a_{\text{N}} + \beta_{\text{out}})^2 < 4D_{\text{N}}$, the eigenvalues are complex conjugates $\rho_{\text{N}} e^{\pm i\varphi_{\text{N}}}$ with

$$\rho_{\text{N}} = \sqrt{D_{\text{N}}}, \qquad \cos\varphi_{\text{N}} = \frac{a_{\text{N}} + \beta_{\text{out}}}{2\rho_{\text{N}}}.$$

Repeating the matching argument of §A.5 yields the same structural form as for HB:

$$\chi_K^{\text{NAG}}(\sigma) = \rho_{\text{N}}^K\left[\cos(K\varphi_{\text{N}}) + C_{\text{N}}\sin(K\varphi_{\text{N}})\right], \qquad C_{\text{N}} = \frac{a_{\text{N}} - \beta_{\text{out}}}{2\rho_{\text{N}}\sin\varphi_{\text{N}}}.$$

The structure is identical to the HB case, with three quantities replaced: $\beta_{\text{out}} \to D_{\text{N}}$ in the determinant (and hence $\rho \to \rho_{\text{N}}$), $a \to a_{\text{N}}$ in the trace, and the phase φ updated accordingly.

This makes the relationship between NAG and restarting explicit. NAG changes the spectral envelope through D_{N}, which now depends on $\nu\sigma$; restarting instead acts on the phase-dependent projection factor multiplying that envelope. The same projected-factor decomposition holds for both HB and NAG. Thus, restarting can also be applied on top of NAG: when an integer period K places the NAG phase $K\varphi_{\text{N}}$ near a cancellation point, the NAG envelope ρ_{N}^K is multiplied by a small oscillatory factor. In this sense, restarting can further contract the NAG dynamics through the same cancellation mechanism as in HB, but with NAG's phase φ_{N} instead of HB's phase φ. The two mechanisms are therefore complementary: NAG improves the spectral envelope, while restarting exploits favorable phases within that envelope.

A.9. Blockwise extension

The scalar theory in the main text is mode-wise. A direct per-mode restart period is useful analytically but unrealistic algorithmically. Here we describe a more practical relaxation: one restart period per parameter block.

We connect this blockwise view to the linearized squared-loss setting through the duality between the empirical NTK and the Gauss–Newton Hessian. If $J = [J_1, \dots, J_B]$ is the Jacobian partitioned by parameter blocks, then the empirical-kernel operator $JJ^\top/D$ and the parameter-space Gauss–Newton Hessian $J^\top J/D$ have the same nonzero spectrum. Thus the same effective-eigenvalue analysis can be read in parameter space, where blockwise curvature structure is meaningful. Motivated by evidence that Transformer Hessian spectra vary substantially across parameter blocks [27], and by work on near-block-diagonal Hessian structure in neural networks [6], assume

$$H_\theta^{\mathrm{GN}} = \frac{1}{D} J^\top J \approx \mathrm{blkdiag}(H_1, \dots, H_B), \qquad H_b = \frac{1}{D} J_b^\top J_b. \tag{4}$$

Equivalently, cross-block Gram terms $J_b^\top J_c/D$ are small for $b \neq c$.

Under this approximation, block b evolves under its local curvature. After S inner GD steps within outer round t,

$$x_{t,b}^{\mathrm{loc}} = (I - \eta H_b)^S x_{t,b}, \qquad g_{t,b} = x_{t,b} - x_{t,b}^{\mathrm{loc}} = \Sigma_b x_{t,b}, \qquad \Sigma_b := I - (I - \eta H_b)^S. \tag{5}$$

Thus one outer transition already includes the S inner steps through Σ_b. The HB/EMA outer update in block b is therefore the same two-state system as in the scalar case, with σ replaced by the block operator Σ_b:

$$\begin{pmatrix} x_{t+1,b} \\ m_{t+1,b} \end{pmatrix} = \begin{pmatrix} I - \nu_b(1-\beta_b)\Sigma_b & -\nu_b\beta_b I \\ (1-\beta_b)\Sigma_b & \beta_b I \end{pmatrix} \begin{pmatrix} x_{t,b} \\ m_{t,b} \end{pmatrix}. \tag{6}$$

Diagonalizing $H_b u_{b,j} = \lambda_{b,j} u_{b,j}$ recovers independent scalar dynamics inside the block, with

$$\sigma_{b,j} = 1 - (1 - \eta\lambda_{b,j})^S. \tag{7}$$

Thus a blockwise restart period K_b applies

$$x_{K_b,b} = \chi_{K_b}(\Sigma_b) x_{0,b}. \tag{8}$$

Writing $x_{0,b} = \sum_j \alpha_{b,j} u_{b,j}$, this becomes

$$x_{K_b,b} = \sum_j \chi_{K_b}(\sigma_{b,j}) \alpha_{b,j} u_{b,j}. \tag{9}$$

A natural blockwise oracle period is therefore

$$K_b^\star \in \arg\min_K \sum_j w_{b,j} |\chi_K(\sigma_{b,j})|^2, \tag{10}$$

where $w_{b,j}$ measures the residual energy in eigendirection j of block b. This gives a practical middle ground between one global period and an unrealistic per-eigenmode oracle.

High-momentum period heuristic. If the effective spectrum of block b is concentrated near $\bar{\sigma}_b$, and $\beta_b = 1 - \delta_b$ with $\delta_b \ll 1$, the scalar phase satisfies

$$\varphi_b \approx \sqrt{\nu_b \bar{\sigma}_b (1 - \beta_b)}. \tag{11}$$

Indeed, using $a = 1 - \nu_b(1 - \beta_b)\bar{\sigma}_b$ and $\rho = \sqrt{\beta_b}$, we have

$$\cos \varphi_b = \frac{a + \beta_b}{2\rho} = 1 - \frac{\nu_b \bar{\sigma}_b (1 - \beta_b)}{2} + O\big((1 - \beta_b)^2\big),$$

which gives (11). The phase shift $\theta = \arctan C$ is $O(\sqrt{1 - \beta_b})$, so the first cancellation period satisfies

$$K_b \approx \frac{\pi}{2\varphi_b} \approx \frac{\pi}{2\sqrt{\nu_b \bar{\sigma}_b (1 - \beta_b)}}. \tag{12}$$

Thus blocks with larger effective progress $\bar{\sigma}_b$ prefer shorter restart periods, while slower blocks prefer longer ones.

Appendix B. Experimental Details and Additional Results

In this appendix, we detail the additional experiments completed as well as present more detailed versions of plots from the main body. In addition to this, we show hyperparameter configurations used.

B.1. Details for Figure 1.

All panels use $\beta_{\text{out}} = 0.9$ and $\nu = 1$, except panel (a), which also shows a no-restart NAG baseline with β_{out} tuned to minimize the final loss at the same horizon. Panel (a) uses a single mode with $\sigma = 0.95$, horizon $T = 80$, and initial state $z_0 = (x_0, m_0) = (1, 0)$. Panel (b) uses the multi-mode spectrum $\sigma \in \{0.95, 0.85, 0.75, 0.60, 0.45, 0.30\}$, equal weights, and horizon $T = 120$; the per-mode restart curve is an oracle benchmark. Panel (c) uses three heterogeneous blocks with effective spectra concentrated in $[0.92, 1.00]$, $[0.55, 0.65]$, and $[0.18, 0.26]$, again with equal weighting and horizon $T = 120$.

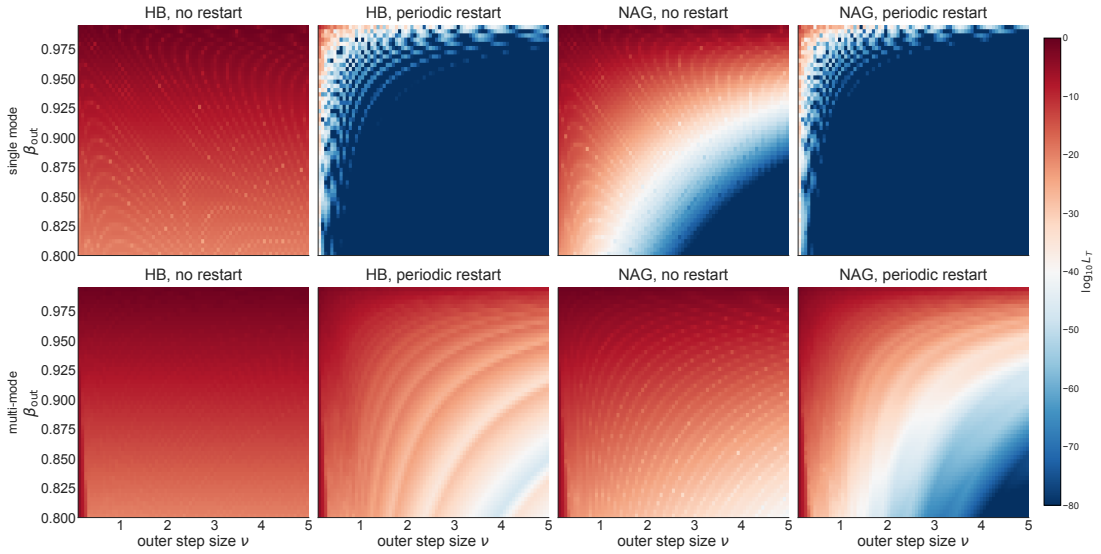


Figure 5: Robustness over outer hyperparameters. Each heatmap shows clipped $\log_{10}$ final loss over a grid of β_{out} and ν. The best periodic restart enlarges the good-hyperparameter region for HB and NAG.

B.2. Hyperparameter Sweeps for Llama-150M

We evaluate DiLoCo by pretraining Llama-150M on the C4 dataset from scratch using a 2-replica DiLoCo configuration with two H200 GPUs (128GB memory). We first tuned the inner optimizer's learning rate and carried the optimal learning rate of 1×10^{-3} forward to the rest of the experiments. A full grid search over all combinations of hyperparameters would be computationally prohibitive, so experiments were designed to isolate one or two axes of variation at a time while holding the others fixed.

In all experiments, the inner optimizer is AdamW with weight decay 0.1 and betas $(0.9, 0.999)$.

Table 1: Hyperparameter configuration for the DiLoCo runs.

Category	Hyperparameter	Value
Batching	Per-replica effective batch size	64
	Effective global batch size	128
Inner optimizer	Learning rate (AdamW)	$5\times10^{-4},\ 1\times10^{-3},\ 2\times10^{-3}$
	Communication period S	64, 128, 512, 1024, 2048
Outer optimizer	Type	Heavy-ball, Nesterov momentum
	Learning rate ν	0.1, 0.3, 0.5, 0.7, 0.9, 1.1
	Momentum β_{out}	0.1, 0.3, 0.5, 0.7, 0.9
	Restart period K	2, 3, 5, 7, 9, 11
Training schedule	LR scheduler	Cosine annealing
	Total inner steps	12,500
	Warmup steps	1,200

Table 2: Model, data, and evaluation configuration for DiLoCo experiments.

Category	Configuration	Value
Model	Architecture	`LlamaForCausalLM`[†]
	Hidden size	1024
	Intermediate size	2688
	Hidden layers	12
	Attention heads	16
	RMSNorm ϵ	10^{-5}
Data	Dataset	`allenai/c4`
	Tokenizer	`Mistral-7B-v0.1`[‡]
	Sequence length	2048
Runtime	Precision	`bf16-mixed`
	Compute	2 replicas, 2 H200 GPUs

[†]`PrimeIntellect/llama-150m-fresh` [‡]`mistralai/Mistral-7B-v0.1`

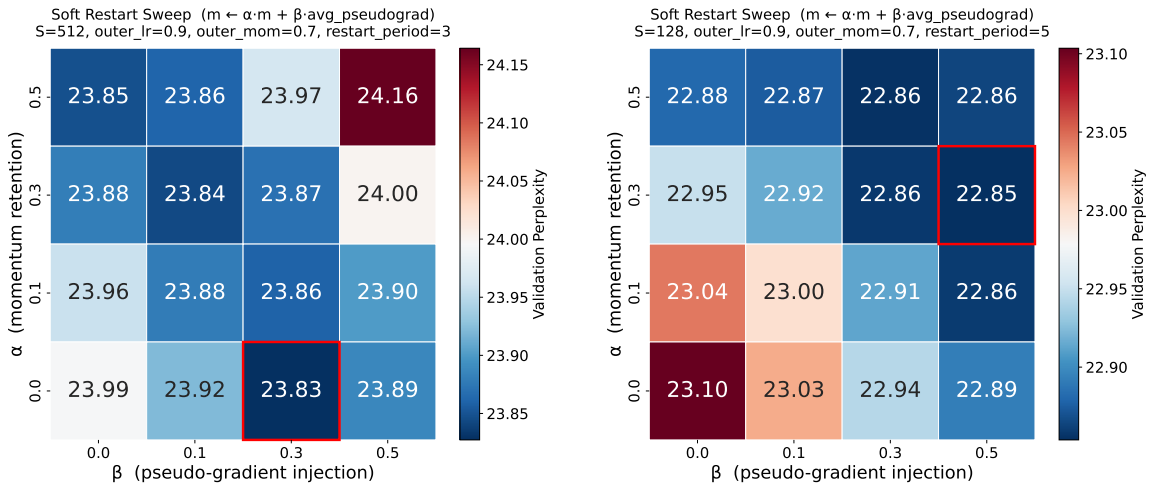
B.3. Soft Restarts

A soft restart applies a periodic buffer rewrite on top of the ordinary outer momentum update, occurring only at restart boundaries. Unlike a hard restart, which completely discards the accumulated momentum state by zeroing the buffer, the soft restart can be viewed as partially preserving the existing momentum while incorporating the current averaged DiLoCo pseudogradient. Specifically, at a restart boundary, we rewrite the buffer as $\widetilde{m}_t \leftarrow \alpha m_t + \beta \bar{g}_t$, where m_t denotes the momentum buffer after the ordinary outer optimizer update but before any restart rewrite, $\widetilde{m}_t$ denotes the momentum buffer after the possible soft-restart rewrite, and $\bar{g}_t$ denotes the averaged DiLoCo pseudogradient. Here, α governs how much of existing momentum memory we want to retain, while β governs the contribution of the current averaged pseudogradient to the rewritten buffer. We define

it as

$$\widetilde{m}_t = \begin{cases} \alpha m_t + \beta \bar{g}_t, & t \equiv 0 \pmod{R} \\ m_t, & \text{o.w.} \end{cases} \qquad \text{for } t \in \{1, \ldots, T\}.$$

where R denotes the restart period in outer steps. Below we present Figure 6 that illustrates how the optimal (α, β) landscape shifts substantially with the training configuration between $S = 512$ and $S = 128$. While carefully tuned soft restarts can marginally improve over hard restarts and sometimes improve over the optimal no-restart, the improvements are small in absolute perplexity terms and come at the cost of two additional hyperparameters whose optimal values appear to be sensitive to S and K. Given that the simpler hard restart already provides the robustness benefits described in Section 4, we do not find soft restarts to be worth the added tuning burden in practice.



(a) $S = 512$, $\nu = 0.9$, $\beta_{\text{out}} = 0.7$, $K = 3$.

(b) $S = 128$, $\nu = 0.9$, $\beta_{\text{out}} = 0.7$, $K = 5$.

Figure 6: Soft restart sweeps. Each cell shows final validation perplexity under the boundary update $m \leftarrow \alpha m + \beta \bar{g}$, where α controls momentum retention and β controls pseudo-gradient injection at the restart boundary.

B.4. Additional Robustness Results for DiLoCo

We use the term *robustness* to mean low sensitivity, or equivalently low variation, of the validation metric across the explored hyperparameter settings. Intuitively, robustness across hyperparameter choices is desirable because it reduces reliance on exhaustive tuning over a finely discretized hyperparameter grid to achieve good validation performance. Figure 7 extends the robustness comparison from Figure 4 to a wider range of communication periods, $S \in \{64, 128, 512, 1024, 2048\}$, reporting the final perplexity. Across all the communication frequencies, we see that the no-restart curve's optimal β_{out} shift downwards as S increases, as well as degrade sharply at high β_{out}, with divergence at $S = 2048$. On the other hand, the restart curve remains nearly flat for all tested periods K.

Figure 7: Final validation perplexity for **NAG** as outer optimizer, shown as a function of restart period K (bottom axis, red) and outer momentum coefficient β_{out} (top axis, blue) for $S \in \{64, 128, 512, 1024, 2048\}$. The blue curve sweeps β_{out} with no momentum restarts; the red curve fixes β_{out} and sweeps K. Dashed lines with annotated values indicate runs that diverged and exceeded the plot range. Without restarts, perplexity spikes sharply at high β_{out}, with catastrophic divergence at $S = 2048$; with restarts, the red curve remains nearly flat across all tested K, demonstrating that the restart period is a more forgiving hyperparameter than outer momentum.

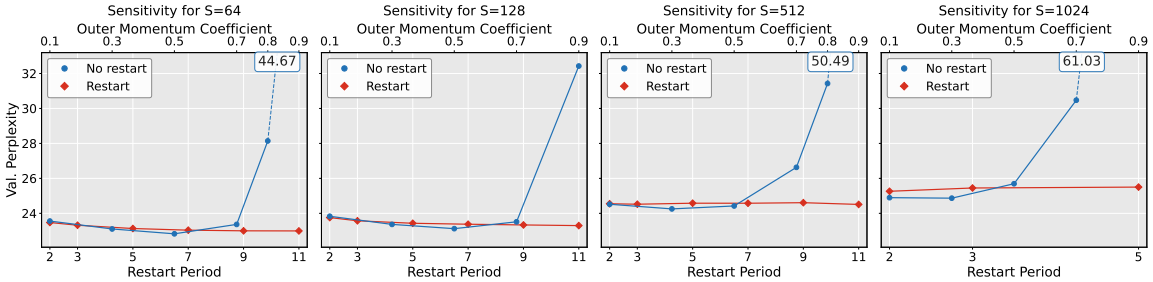


Figure 8: Same as Figure 7 but with **Heavy-Ball** (EMA) as the outer optimizer, for $S \in \{64, 128, 512, 1024\}$. Heavy-Ball without restarts is more sensitive to β_{out} than NAG, with earlier and more severe divergence as S increases. Periodic restarts again stabilize training across all communication periods, keeping perplexity flat across the sweep of K.

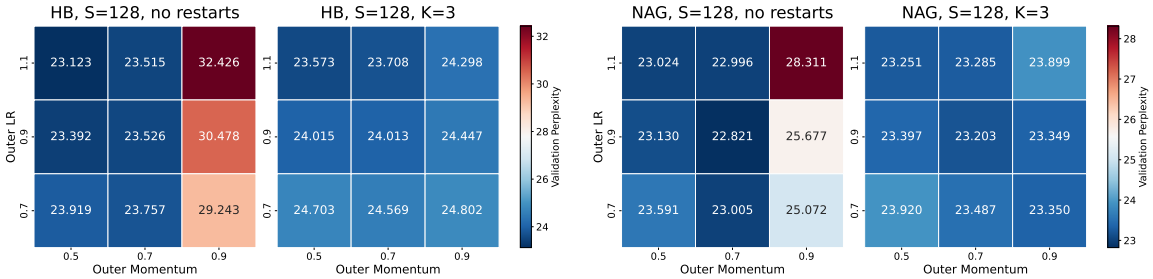


Figure 9: Validation perplexity over the outer hyperparameter grid at $S = 128$ for HB(left) and NAG(right), comparing standard DiLoCo against DiLoCo with momentum restart period $K = 3$. Periodic restarts reduce the high-β_{out} failure region while preserving peak performance.